Dipartimento di Economia Politica

Materiali di discussione

\\ 681 \\

Adaptive Minimax Estimation over Sparse ℓ_q - Hulls

Zhan Wang[1]
Sandra Paterlini[2]
Fuchang Gao[3]
Yuhong Yang[4]

January 2012

1-4 School of Statistics, University of Minnesota
USA
e-mail: wangx607@stat.umn.edu;
yyang@stat.umn.edu

2 Department of Economics,
RECent & CEFIN University of Modena and Reggio E.
Italy
e-mail: sandra.paterlini@unimore.it

3 Department of Mathematics,
University of Idaho
USA
e-mail: fuchang@uidaho.edu

ISSN: 2039-1439 a stampa
ISSN: 2039-1447 on line

Adaptive Minimax Estimation over Sparse ℓ_q-Hulls

Zhan Wang[1], **Sandra Paterlini** *[2], **Fuchang Gao**[3], **and Yuhong Yang** †[1]

[1] *School of Statistics, University of Minnesota, USA,*

e-mail: wangx607@stat.umn.edu; yyang@stat.umn.edu

[2] *Department of Economics,RECent & CEFIN University of Modena and Reggio E., Italy,*

e-mail: sandra.paterlini@unimore.it

[3] *Department of Mathematics, University of Idaho, USA,*

e-mail: fuchang@uidaho.edu

Abstract: Given a dictionary of M_n initial estimates of the unknown true regression function, we aim to construct linearly aggregated estimators that target the best performance among all the linear combinations under a sparse q-norm ($0 \leq q \leq 1$) constraint on the linear coefficients. Besides identifying the optimal rates of aggregation for these ℓ_q-aggregation problems, our multi-directional (or universal) aggregation strategies by model mixing or model selection achieve the optimal rates simultaneously over the full range of $0 \leq q \leq 1$ for general M_n and upper bound t_n of the q-norm. Both random and fixed designs, with known or unknown error variance, are handled, and the ℓ_q-aggregations examined in this work cover major types of aggregation problems previously studied in the literature. Consequences on minimax-rate adaptive regression under ℓ_q-constrained true coefficients ($0 \leq q \leq 1$) are also provided.

Our results show that the minimax rate of ℓ_q-aggregation ($0 \leq q \leq 1$) is basically determined by an effective model size, which is a sparsity index that depends on q, t_n, M_n, and the sample size n in an easily interpretable way based on a classical model selection theory that deals with a large number of models. In addition, in the fixed design case, the model selection approach is seen to yield optimal rates of convergence not only in expectation but also with exponential decay of deviation probability. In contrast, the model mixing approach can have leading constant one in front of the target risk in the oracle inequality while not offering optimality in deviation probability.

Keywords and phrases: minimax risk, adaptive estimation, sparse ℓ_q-constraint, linear combining, aggregation, model mixing, model selection.

*Sandra Paterlini's research was partially supported by Fondazione Cassa di Risparmio di Modena.

†Yuhong Yang's research was partially supported by NSF grant DMS-1106576.

1. Introduction

The idea of sharing strengths of different estimation procedures by combining them instead of choosing a single one has led to fruitful and exciting research results in statistics and machine learning. In statistics, the theoretical advances have centered on optimal risk bounds that require almost no assumption on the behaviors of the individual estimators to be integrated (see, e.g., [64, 67, 22, 24, 42, 52, 69, 58] for early representative work). 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 aims 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 aims at improving over the performance of all the candidate procedures in certain ways. Whatever the goal is, for the purpose of estimating a target function (e.g., the true 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), 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 the proportion of the initial data size 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, although combining procedures for adaptation and selecting the best one share the same goal of achieving the best performance offered by the candidate procedures, the former usually wins when model selection uncertainty is high (see, e.g., [74]). Theoretically, one only needs to pay a relatively small price for aggregation for adaptation ([66, 24, 58]). 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 [42, 52, 69, 58]). 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, [58]), aggregation to achieve the best performance of a linear combination of no more than a given number of initial estimates ([19]) and also under an additional constraint on the ℓ_1-norm of these coefficients ([49]). Interestingly, combining for adaptation has a fundamental role for combining for improvement: it serves as an effective tool in constructing multi-directional (or universal) aggregation methods that simultaneously achieve the best performance in multiple specific directions of aggregation for improvement. This strategy was taken in section 3 of [69], where aggregations of subsets of estimates are then aggregated to be suitably aggressive and conservative in an adaptive way. Other uses of subset models for universal aggregation have been handled in [19, 54].

The goal of this paper 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 [58], of the best linear combination of the initial estimates under the constraint that the q-norm ($0 \leq q \leq 1$) 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 ($0 \leq q \leq 1$) and t_n ($t_n > 0$). From the work in [42, 69, 58, 4], it is known that by suitable aggregation methods, the squared L_2 risk is no larger than that of the best linear combination of the initial M_n estimates with the ℓ_1-norm of the coefficients bounded by 1 plus the order $(\log(M_n/\sqrt{n})/n)^{1/2}$ when $M_n \geq \sqrt{n}$ or M_n/n when $M_n < \sqrt{n}$. Two important features are evident here: 1) When M_n is large, its effect on the risk enlargement is only through a logarithmic fashion; 2) No assumption is needed at all on how the initial estimates are possibly correlated. The strong result comes from the ℓ_1-constraint on the coefficients.

Indeed, in the last decade of the twentieth century, the fact that ℓ_1-type of constraints induce sparsity has been used in different ways for statistical estimation to attain relatively fast rates of convergence as a means to overcome the curse of dimensionality. Among the most relevant ones, Barron [9] studied the use of ℓ_1-constraint in construction of estimators for fast convergence with neural nets; Tibshirani [57] introduced the Lasso; Chen, Donoho and Saunders [25] proposed the basis pursuit with over complete bases. Theoretical advantages have also been pointed out. Barron [8] showed that for estimating a high-dimensional function that has integrable Fourier transform or a neural net representation, accurate approximation error is achievable. Together with model selection over finite dimensional neural network models, relatively fast rates of convergence, e.g., $[(d \log n)/n]^{1/2}$, where d is the input dimension, are obtained (see, e.g., [9] with parameter discretization, section III.B in [71] and section 4.2 in [11] with continuous models). Donoho and Johnstone [30] identified how the ℓ_q-constraint ($q > 0$) on the mean vector affects estimation accuracy under ℓ_p loss ($p \geq 1$) in an illustrative Gaussian sequence model. For function estimation, Donoho [28] studied sparse estimation with unconditional orthonormal bases and related the essential rate of convergence to a sparsity index. In that direction, for a special case of function classes with unconditional basis defined basically in terms of bounded q-norm on the coefficients of the orthonormal expansion, the rate of convergence $(\log n/n)^{1-q/2}$ was given in [71] (section 5). The same rate also appeared in the earlier work of Donoho and Johnstone [30] in some asymptotic settings. Note that when $q = 1$, this is exactly the same rate of the risk regret for ℓ_1-aggregation when M_n is of order n^κ for $1/2 \leq \kappa < \infty$.

General model selection theories on function estimation intend to work with general and possibly complicatedly dependent terms. Considerable research has been built upon subset selection as a natural way to pursue sparse and flexible estimation. When exponentially many or more models are entertained, optimality theories that handle a small number of models (e.g., [56, 48]) are no longer suitable. General theories were then developed for estimators based on criteria that add an additional penalty to the AIC type criteria, where the additional penalty term prevents substantial overfitting that often occurs when working with exponentially many models by standard information criteria, such as AIC and BIC. A masterpiece of work with tremendous breadth and depth is Barron, Birgé and Massart [11], and some other general results in specific contexts of density estimation and regression with fixed or random design are in [71, 65, 18, 5, 6, 15].

These model selection theories are stated for nonparametric scenarios where none of the finite-dimensional approximating models is assumed to hold but they are used as suitable sieves to deliver good estimators when the size of the sieve is properly chosen (see, e.g., [55, 59, 17] for non-adaptive sieve theories). If one makes the assumption that a subset model of at most k_n terms holds (ℓ_0-constraint), then the general risk bounds mentioned in the previous paragraph immediately give the order $k_n \log\left(M_n/k_n\right)/n$ for the risk of estimating the target function under quadratic type losses.

Thus, the literature shows that both ℓ_0- and ℓ_1-constraints result in fast rates of convergence (provided that M_n is not too large and k_n is relatively small), with hard-sparsity directly coming from that only a small number of terms is involved in the true model under the ℓ_0-constraint, and soft-sparsity originating from the fact that there can only be a few large coefficients under the ℓ_1-constraint. In this work, with new approximation error bounds in $\ell_{q,t_n}^{M_n}$-hulls (defined in section 2.1) for $0 < q \le 1$, from a theoretical standpoint, we will see that model selection or model combining with all subset models in fact simultaneously exploits the advantage of sparsity induced by ℓ_q-constraints for $0 \le q \le 1$ to the maximum extent possible.

Clearly, all subset selection is computationally infeasible when the number of terms M_n is large. To overcome this difficulty, an interesting research direction is based on greedy approximation, where terms are added one after another sequentially (see, e.g., [12]). Some general theoretical results are given in the recent work of [40], where a theory on function estimation via penalized squared error criteria is established and is applicable to several greedy algorithms. The associated risk bounds yield optimal rate of convergence for sparse estimation scenarios. For aggregation methods based on exponential weighting under fixed design, practical algorithms based on Monte Carlo methods have been given in [27, 54].

Considerable recent research has focused on ℓ_1-regularization, producing efficient algorithms and related theories. Interests are both on risk of regression estimation and on variable selection. Some estimation risk bounds are in [13, 37, 43, 44, 50, 51, 62, 60, 76, 75, 77, 73].

The ℓ_q-constraint, despite being non-convex for $0 < q < 1$, poses an easier optimization challenge than the ℓ_0-constraint, which is known to define a NP-hard optimization problem and be hardly tractable for large dimensions. Although a few studies have devoted to the algorithmic developments of the ℓ_q-constraint optimization problem, such as multi-stage convex relaxation algorithm ([78]) and the DC programming approach ([33]), little work has been done with respect to the theoretical analysis of the ℓ_q-constrained framework.

Sparse model estimation by imposing the ℓ_q-constraint 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 do not only promote sparsity but 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 ([29]).

Our ℓ_q-aggregation risk upper bounds require no assumptions on dependence of the initial estimates in the dictionary and the true regression function is arbitrary (except that it has a known sup-norm upper bound in the random design case). The results readily give minimax rate optimal estimators for a regression function that is representable as a linear combination of the predictors subject to ℓ_q-constraints on the linear coefficients.

Two recent and interesting results are closely related to our work, both under fixed design only. Raskutti, Wainwright and Yu [53] derived in-probability minimax rates of convergence for estimating the regression functions in $\ell_{q,t_n}^{M_n}$-hulls with minimal conditions for the full range of $0 \leq q \leq 1$. In addition, in an informative contrast, they have also handled the quite different problem of estimating the coefficients under necessarily much stronger conditions. Rigollet and Tsybakov [54] nicely showed that exponential mixing of least squares estimators by an algorithm of Leung and Barron [46] over subset models achieves universal aggregation of five different types of aggregation, which involve ℓ_0- and/or ℓ_1-constraints. Furthermore, they implemented a MCMC based algorithm with favorable numerical results. As will be seen, in this context of regression under fixed design, our theoretical results are broader with improvements in several different ways.

Our theoretical work emphasizes adaptive minimax estimation under the mean squared risk.

Building upon effective estimators and powerful risk bounds for model selection or aggregation for adaptation, we propose several aggregation/combining strategies and derive the corresponding oracle inequalities or index of resolvability bounds. Upper bounds for ℓ_q-aggregations and for linear regression with ℓ_q-constraints are then readily obtained by evaluating the index of resolvability for the specific situations, incorporating an approximation error result that follows from a new and precise metric entropy calculation on function classes of $\ell_{q,t_n}^{M_n}$-hulls. Minimax lower bounds that match the upper rates are also provided in this work. Whatever the relationships between the dictionary size M_n, the sample size n, and upper bounds on the ℓ_q-constraints, our estimators automatically take advantage of the best sparse ℓ_q-representation of the regression function in a proper sense.

By using classical model selection theory, we have a simple explanation of the minimax rates, by considering the effective model size m_*, which provides the best possible trade-off between the approximation error, the estimation error, and the additional price due to searching over not pre-ordered terms. The optimal rate of risk regret for ℓ_q-aggregation, under either hard or soft sparsity (or both together), can then be unifyingly expressed as

$$REG(m_*) = 1 \wedge \frac{m_*\left(1 + \log \frac{M_n}{m_*}\right)}{n},$$

which can then be interpreted as the log number of models of size m_* divided by the sample size ($\wedge 1$), as was previously suggested for the hard sparsity case $q = 0$ (e.g., Theorem 1 of [71], Theorems 1 and 4 of [65]).

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. In addition, we derive metric entropy and approximation error bounds for $\ell_{q,t_n}^{M_n}$-hulls that play an important role in determining the minimax rate of convergence and adaptation. In section 3, we derive optimal rates of ℓ_q-aggregation and show that our methods achieve multi-directional aggregation. We also briefly talk about ℓ_q-combination of procedures. In section 4, we derive the minimax rate for linear regression with ℓ_q-constrained coefficients also under random design. In section 5, we handle ℓ_q-regression/aggregation under fixed design with known or unknown variance. A discussion is then reported in section 6. In section 7, oracle inequalities are given for the random design. Proofs of the results are provided in section 8. We note that some upper and lower bounds in the last two sections may be of independent interest.

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 or model selection methods 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 $\mathcal{X} \subset \mathbb{R}^d$ with a probability distribution P_X. We assume the regression model

$$Y_i = f_0(\mathbf{X}_i) + \varepsilon_i, \qquad i = 1, \ldots n, \tag{2.1}$$

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, $\|f_j\| \leq 1$, $1 \leq j \leq M_n$. 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 $\mathbb{R}^{M_n}$. 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. But 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

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

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 $C = 1$ under some conditions. We aim to obtain the optimal order of convergence for the risk regret term.

2.2. Two starting estimators

A key step of our strategy is the construction of candidate estimators using subsets of the initial estimates. The following two estimators (T- and AC-estimators) were chosen because of the relatively mild assumptions for them to work with respect to the squared L_2 risk. Under the data generating model (2.1) and i.i.d. observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, suppose we are given $(g_1, \ldots, g_m)$ terms for the regression problem.

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

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

(T-estimator) Birgé [15] 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 in [15]. 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*

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

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 $\mathcal{G} = \{g = \sum_{j=1}^{m} \vartheta_j g_j : \|\vartheta\|_1 \leq s\}$ for some $s > 0$. Audibert and Catoni proposed a sophisticated AC-estimator $\hat{f}_s^{(AC)}$ ([4], page 25). The following proposition is a direct result from Theorem 4.1 in [4] under the following conditions.

AC1. There exists a constant $H > 0$ such that $\sup_{g, g' \in \mathcal{G}, \mathbf{x} \in \mathcal{X}} |g(\mathbf{x}) - g'(\mathbf{x})| = H < \infty$.

AC2. There exists a constant $\sigma' > 0$ such that $\sup_{\mathbf{x} \in \mathcal{X}} E\left((Y - g^*(\mathbf{X}))^2 | \mathbf{X} = \mathbf{x}\right) \leq (\sigma')^2 < \infty$, where $g^* = \inf_{g \in \mathcal{G}} \|g - f_0\|^2$.

Proposition 2. *Suppose Assumptions* AC1 *and* AC2 *hold. For any* $s > 0$*, we can construct an*

AC-estimator $\hat{f}_s^{(AC)}$ *such that*

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

where c *is a pure constant.*

Note that under the assumption $\|f_0\|_\infty \leq L$, 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 [68]) and Catoni's algorithm (Catoni [24]), 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 $0 < \underline{\sigma} \leq \sigma \leq \overline{\sigma} < \infty$;

Y2. The error density function $h(x)$ has a finite fourth moment and 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 $S_0 \leq S \leq S_0^{-1}$,

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

We can construct an estimator $\hat{f}^Y$ which aggregates $\check{f}_1, \ldots, \check{f}_N$ by the ARM algorithm as described below.

Step 1. Split the data into two parts $Z^{(1)} = (\mathbf{X}_i, Y_i)_{i=1}^{n_1}$, $Z^{(2)} = (\mathbf{X}_i, Y_i)_{i=n_1+1}^{n}$. Take $n_1 = \lceil n/2 \rceil$.

Step 2. Estimate σ^2 for each $\check{f}_k$ using the data $Z^{(1)}$,

$$\hat{\sigma}_k^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} \left(Y_i - \check{f}_k(\mathbf{X}_i)\right)^2, \text{ for } 1 \le k \le N.$$

Clip the estimate $\hat{\sigma}_k^2$ into the range $[\underline{\sigma}^2, \overline{\sigma}^2]$ if needed.

Step 3. Evaluate predictions for each k. For $n_1 + 1 \le l \le n$, predict Y_l by $\check{f}_k(\mathbf{X}_l)$ and compute

$$E_{k,l} = \frac{\prod_{i=n_1+1}^{l} h\left((Y_i - \check{f}_k(\mathbf{X}_i))/\hat{\sigma}_k\right)}{\hat{\sigma}_k^{l-n_1}}.$$

Step 4. Compute the final estimate $\hat{f}^Y = \sum_{k=1}^N W_k \check{f}_k$ with

$$W_k = \frac{1}{n - n_1} \sum_{l=n_1+1}^{n} W_{k,l} \quad \text{and} \quad W_{k,l} = \frac{\pi_k E_{k,l}}{\sum_{j=1}^N \pi_j E_{j,l}},$$

where π_k are prior probabilities such that $\sum_{k=1}^N \pi_k = 1$.

Proposition 3. *(Yang [69], 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* $S_0 = \underline{\sigma}/\overline{\sigma}, R_0 = L$*).*

Remark 1. If σ is known or other estimators of σ are available, the data splitting is not required, and the ARM algorithm consists of only Steps 3 and 4.

(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. $\frac{E(\varepsilon_i^2 \exp(\alpha|\varepsilon_i|))}{E(\exp(\alpha|\varepsilon_i|))} \le V_\alpha$.

The estimator built using Catoni's algorithm is $\hat{f}^C = \sum_{k=1}^N W_k \check{f}_k$ with

$$W_k = \frac{1}{n} \sum_{l=1}^{n} \frac{\pi_k \left(\prod_{i=1}^{l} q_k(Y_i|\mathbf{X}_i)\right)}{\sum_{j=1}^N \pi_j \left(\prod_{i=1}^{l} q_j(Y_i|\mathbf{X}_i)\right)}, \text{ and } \; q_k(Y_i|\mathbf{X}_i) = \sqrt{\frac{\lambda_C}{2\pi}} \exp\left\{-\frac{\lambda_C}{2}(Y_i - \check{f}_k(\mathbf{X}_i))^2\right\},$$

where $\lambda_C = \min\{\frac{\alpha}{2L}, (U_\alpha(17L^2 + 3.4V_\alpha))^{-1}\}$, and π_k is the prior for $\check{f}_k$, $1 \le k \le N$, such that $\sum_{k=1}^N \pi_k = 1$.

Proposition 4. *(Catoni [24], 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* $\check{f}_1, \dots, \check{f}_N$ *by Catoni's algorithm has the risk*

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

Remark 2. In the risk bound above, the multiplying constant in front of $\|\check{f}_k - f_0\|^2$ is one, which can be important sometimes. Catoni [24] provided results under weaker assumptions than C1 and C2. In particular, ε_i and $\mathbf{X}_i$ do not have to be independent.

2.4. 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 [14, 72].

For each $1 \le m \le M_n$ and each subset $J_m \subset \{1, 2, \dots, M_n\}$ of size m, define $\mathcal{F}_{J_m} = \{\sum_{j \in J_m} \theta_j f_j : \theta \in \mathbb{R}^m\}$. 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 $\mathcal{F}$.

Theorem 1. *(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

$$\log N_\epsilon \le \begin{cases} c_q \left(t_n \epsilon^{-1}\right)^{\frac{2q}{2-q}} \log(1 + M_n^{\frac{1}{q} - \frac{1}{2}} t_n^{-1} \epsilon) & \text{if } \epsilon > t_n M_n^{\frac{1}{2} - \frac{1}{q}}, \\ c_q M_n \log(1 + M_n^{\frac{1}{2} - \frac{1}{q}} t_n \epsilon^{-1}) & \text{if } \epsilon \le t_n M_n^{\frac{1}{2} - \frac{1}{q}}. \end{cases} \tag{2.2}$$

(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

$\|\theta^m\|_1 \leq t_n$ *such that the sparse approximation error is upper bounded as follows*

$$\|f_{\theta^m} - f_0\|^2 - d^2(f_0; \mathcal{F}_q(t_n)) \leq 2^{2/q-1} t_n^2 m^{1-2/q}. \tag{2.3}$$

The metric entropy estimate (2.2) is the best possible. Indeed, if f_j, $1 \leq j \leq M_n$, are orthonormal functions, then (2.2) is sharp in order for any ϵ satisfying that ϵ/t_n is bounded away from 1 (see [45]). Also note that if we let ν_0 be the discrete measure $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, where $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ are fixed points in a fixed design, then $\|g\|_{L^2(\nu_0)} = (\frac{1}{n}\sum_{i=1}^n |g(\mathbf{x}_i)|^2)^{1/2}$. Thus, part (i) of Theorem 1 implies Lemma 3 of [53], 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 useful to derive the exact minimax rates for some of the possible situations in terms of M_n, q, and t_n.

With the tools provided in Yang and Barron [72], given fixed q and t_n, one can derive minimax rates of convergence for ℓ_q-aggregation problems and also for linear regression with ℓ_q-constraints. However, the goal for this work is to obtain adaptive estimators that simultaneously work for $\mathcal{F}_q(t_n)$ with any choice of $0 \leq q \leq 1$ and t_n, and more.

2.5. 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 1, classical model selection theory can provide key insight on what to expect regarding the minimax rate of convergence for estimating a function in $\ell_{q,t_n}^{M_n}$-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

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

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

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

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 ratio m^*/M_n is called a sparsity index in [71] (section III.D) that characterizes, up to a log factor, how much sparse estimation by model selection improves the estimation accuracy based on nested models only. 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 [72]). 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 [72]). 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., [10, 71, 11, 65, 18, 6]). 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 $0 \leq m \leq M_n$. 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 $m^* \leq M_n/2$. Define

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

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., $t_n^2 m^{1-2/q} + \frac{\sigma^2 m}{n}$). The best choice will typically make the approximation error $t_n^2 m^{1-2/q}$ of the same order as $\frac{m(1+\log\frac{M_n}{m})}{n}$. 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 oracle inequality at the best model of this size gives the minimax rate of convergence, as will be seen. 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., [71], 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 $0 \le q \le 1$.

The aforementioned rates of convergence for estimating functions in $\ell_{q,t_n}^{M_n}$-hulls for $0 \le q \le 1$ will be confirmed, and our estimators will achieve the rates adaptively in some generality. From the insight gained above, to construct a multi-directional (or universal) aggregation method that works for all $0 \le q \le 1$, it suffices to aggregate the estimates from the subset models for adaptation, which will automatically lead to simultaneous optimal performance in $\ell_{q,t_n}^{M_n}$-hulls.

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, let $\mathcal{F}_{J_m}$ be as defined in section 2.4, and let $\mathcal{F}^L_{J_m,s} = \{f_\theta = \sum_{j\in J_m} \theta_j f_j : \|\theta\|_1 \le s, \|f_\theta\|_\infty \le L\}$ 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 combination of $s \in \mathbb{N}$ and function class $\mathcal{F}^L_{J_m,s}$.

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

Let $\mathcal{F}_q^L(t_n) = \mathcal{F}_q(t_n) \cap \{f : \|f\|_\infty \le L\}$ for $0 \le q \le 1$. In the previous section, we have defined $m_* = m_*(q, t_n)$ to be the effective model size for $0 < q \le 1$. Now, for ease of presentation, we extend the definition to

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

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 $0 < q \le 1$) and the hard sparsity one (ℓ_0-constraint), and this smaller size defines 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. In particular, Theorems 2 and 3 provide upper and lower bounds to determine the order of the risk regret for ℓ_q-aggregation of estimates. The specific behaviors of $REG(m_*^{\mathcal{F}})$ for the three different cases will be precisely discussed later.

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 will be 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 [69] (page 36) for multi-directional aggregation and in Lounici [49] (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 [49].

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

Theorem 2. *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* $0 < q \leq 1$, *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*

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

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

For all these cases, C_0 *and* C_2 *do not depend on* n, f_0, t_n, q, k_n, M_n; C_1 *does not depend on* n, f_0, t_n, 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 3. When $q = 1$, our theorem covers some important previous aggregation results. With $t_n = 1$, Juditsky and Nemirovski [42] obtained the optimal result for large M_n; Yang [69] 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 [58] presented the optimal rate for both large and small M_n, 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 [4] have improved over [69] and [58] by giving an optimal risk bound. Even when $q = 1$, our result is more general in that t_n is allowed to be arbitrary. Note also that in some specific cases, the induced sparsity with ℓ_1-constraint was explored earlier in e.g., [30, 9, 71]. The latter two papers dealt with nonparametric situations with mild assumptions on the terms in the approximation systems. In particular, when the true function has a finite-order linear expression, the estimators achieve the minimax optimal rate $\sqrt{(\log n)/n}$ when M_n grows polynomially fast in n.

Remark 4. The upper rate for $q = 0$ as well as its interpretation is not new in the literature (see, e.g., Theorem 1 of [71], Theorems 1 and 4 of [65]): by noticing that there are $\binom{M_n}{k_n}$ subsets of size k_n and that $\log \binom{M_n}{k_n} \le k_n\,(1 + \log(M_n/k_n))$, the rate for $q = 0$, which directly imposes hard sparsity on the maximum number of relevant terms, is just the log number of models of size k_n divided by the sample size.

Remark 5. 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, if $t_n \le e^{cn} \wedge e^{cm_*(1+\log(M_n/m_*))}$ for a pure constant c, then $\log(1 + t_n)/n$ is upper bounded by a multiple of $REG(m_*^{\mathcal{F}_q(t_n)})$. Then, under the condition that the approximation errors involved in the risk bounds are of the same order, **AC-** strategies have the same upper bound orders as **T-** strategies. For case 2, the same is true if for some $s \le e^{cn} \wedge e^{ck_n(1+\log(M_n/k_n))}$, the ℓ_1 norm constraint does not enlarge the approximation error order.

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.

Next, we show that the upper rates in Theorem 2 cannot be generally improved by giving a theorem stating that the lower bounds of the risk are of the same order in some situations, as is typically done in the literature on aggregation of estimates. The following theorem implies that the estimator by our strategies is indeed minimax adaptive for ℓ_q-aggregation of estimates. 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 3. *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 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 $M_n/\left(1 + \log \frac{M_n}{k_n}\right) \leq bn$ for some constant $b > 0$ and the orthonormal basis is specially chosen.

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

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

Thus, under the condition that $\left(\sup_{1 \leq j \leq M_n} \|f_j\|_\infty\right) t_n$ is upper bounded, $\sup_{f_\theta \in \mathcal{F}_q(t_n)} \|f_\theta\|_\infty \leq L$ 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. 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 can be expressed in different ways:

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

The second expression is transparent in interpretation: due to the sparsity condition, we only need to consider models of the effective size m_* and the risk goes with the searching price $\frac{m_*}{n} SER\left(m_*\right)$ (the estimation error of m_* parameters is being dominated in order). The last expression means that we can do better than searching over the models of the ideal model size m^*, which has the risk $\frac{m^*}{n} SER\left(m^*\right)$. The minimax risk is deflated by a factor of $SER\left(m^*\right)^{\frac{q}{2}}$, which becomes larger as $q \to 1$, pointing out that the factor $SER(m^*)$ has to be downsized more as the ℓ_q-ball becomes larger. When $m^* = M_n$ (the full model), $SER(m^*)$ reduces to 1. When $m^* \leq (1 + \log(M_n/m^*))^{q/2}$ 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 $k_n = M_n \leq n$, 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 2, we deduce that for both the **T-** strategies and **AC-** strategies, under the assumption $\sup_j \|f_j\|_\infty \leq L$, our estimator satisfies

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

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 [58], 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 $q = 1$).

3.3. ℓ_q-combination of procedures

Suppose we start with a collection of estimation procedures $\Delta = \{\delta_1, \ldots, \delta_{M_n}\}$ instead of a dictionary of estimates. Let $\hat{f}_j$ be the estimator of the unknown true regression function based on the procedure δ_j, $1 \leq j \leq M_n$, at a certain sample size. Our goal is to combine the estimators

$\{\hat{f}_j : 1 \le j \le M_n\}$ to achieve the best performance in

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

We split the data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ into three parts: $Z^{(1)} = (\mathbf{X}_i, Y_i)_{i=1}^{n_1}$, $Z^{(2)} = (\mathbf{X}_i, Y_i)_{i=n_1+1}^{n_1+n_2}$ and $Z^{(3)} = (\mathbf{X}_i, Y_i)_{i=n_1+n_2+1}^{n}$. Use the data $Z^{(1)}$ to obtain estimators $\hat{f}_1, \ldots, \hat{f}_{M_n}$ and use the data $Z^{(2)}$ to construct T-estimators or AC-estimators based on subsets of $\hat{f}_1, \ldots, \hat{f}_{M_n}$. The data $Z^{(3)}$ are used to construct the final estimator $\hat{f}_\Delta$ by aggregating the T-estimators or AC-estimators and the null model using Catoni's or the ARM algorithm as done in the previous section. For simplicity, assume n is a multiple of 4 and choose $n_1 = n/2$, $n_2 = n/4$. Upper bounds for combining procedures by our strategy are obtained similarly. The only difference is that $d^2(f_0; \mathcal{F})$ is replaced by the risk of the best constrained linear combination of the estimators $\hat{f}_{1,n/2}, \ldots, \hat{f}_{M_n,n/2}$, where we add the second subscript $n/2$ to emphasize that the estimators are constructed with a reduced sample size. For example, by **T-** strategies, we have that for any $0 < q \le 1$ and $t_n > 0$,

$$R(\hat{f}_\Delta; f_0; n) \le C_0 \inf_{\theta \in B_q(t_n; M_n)} E \left\| f_0 - \sum_{j=1}^{M_n} \theta_j \hat{f}_{j,n/2} \right\|^2 + C_1 \cdot REG(m_*^{\mathcal{F}_q(t_n)}),$$

and again such risk bounds simultaneously hold for $0 \le q \le 1$ and $t_n > 0$.

Note that these risk bounds involve the accuracies of the candidate procedures at a reduced sample size $n/2$ due to data splitting to come up with the estimates to be aggregated. Ideally, we want to have $C_0 = 1$ and $\hat{f}_{j,n/2}$ replaced by $\hat{f}_{j,n}$. At this time, we are unaware of any such risk bound that holds for combining general estimators (in fixed design case, Leung and Barron's algorithm does not involve data splitting, but it works only for least squares estimators). Because of this, the theoretical attractiveness that the constant C_0 being 1 in the aggregation stage, unfortunately, disappears since the remaining parts in the risk bounds also depend on the data splitting and there seems to be no reason to expect with certainty that an aggregation method with $C_0 = 1$ has a better risk, even asymptotically, than another one with $C_0 > 1$. Therefore, for combining general statistical procedures, it is unclear how useful $C_0 = 1$ is even from a theoretical perspective. (It seems that there is one scenario that one can argue otherwise: the candidate estimates are truly provided. In the application of combining forecasts sequentially, the candidate forecasts may be provided by other experts/commercial companies and the statistician does not have access to the data based on which the forecasts are built. In this context, since no data splitting is needed, $C_0 = 1$ leads to a theoretical advantage compared to $C_0 > 1$.) For this reason, in our view, results with $C_0 > 1$ (but not too large) are also important for combining procedures. Indeed, such results often have strengths in other aspects such as allowing heavy tail distributions for the errors and allowing dependence of the observations.

Nonetheless, regardless of the degree of practical relevance, limiting attention to the aggregation step and pursuing $C_0 = 1$ in that local goal is certainly not without a theoretical appeal.

Some additional interesting results on combining procedures are in [3, 15, 20, 26, 27, 35, 36, 39, 38, 63, 68].

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

$$Y = f_0(\mathbf{X}) + \varepsilon = \sum_{j=1}^{M_n} \theta_j X_j + \varepsilon. \tag{4.1}$$

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 $1 \leq m \leq M_n \wedge n$ 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 $\mathcal{G}_{J_m,s}^L = \{\sum_{j \in J_m} \theta_j x_j : \|\theta\|_1 \leq s, \|f_\theta\|_\infty \leq L\}$. 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 $\mathcal{F}^L = \{f : f \in \mathcal{F}, \|f\|_\infty \leq L\}$ 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 $\mathcal{F}_0^L(k_n; M_n)$. For ease of presentation, define $\Psi^{\mathcal{F}}$ as follows:

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

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

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

In addition, for lower bound results, let $\underline{\Psi}^{\mathcal{F}_q^L(t_n;M_n)}$ $(0 \leq q \leq 1)$ and $\underline{\Psi}^{\mathcal{F}_q^L(t_n;M_n) \cap \mathcal{F}_0^L(k_n;M_n)}$ $(0 < q \leq 1)$ 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 $0 < q \leq 1$ 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.

Theorem 4. *Suppose $A_{\mathbf{E-G}}$ holds for the **E-G** strategy respectively, and* $\sup_{1 \leq j \leq M_n} \|X_j\|_\infty \leq 1$. *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*.*

Remark 8. The constants C_1 and C_2 may depend on L, p_0, σ^2, $\overline{\sigma}^2/\underline{\sigma}^2$, α, U_α, V_α when relevant.

Remark 9. The rate $\left(\frac{\log n}{n}\right)^{1-q/2}$ for $0 < q < 1$ has appeared in related regression or normal mean problems, e.g., in [30] (Theorem 3), [72] (section 5), [40] (section 6), and [41]. For function classes defined in terms of infinite order orthonormal expansion with bounded q-norm of the coefficients and with ℓ_2-norm of the tail coefficients decaying at a polynomial order, the rate of convergence $(\log n/n)^{1-q/2}$ is derived in [71] (page 1588) (when the tail of the coefficients decays fast, the rate is improved to $(1/n)^{1-q/2}$). Note that only the upper rates are given there.

4.2. Lower bounds

To derive lower bounds, we make the following near orthogonality assumption on sparse sub-collections of the predictors. Such an assumption, similar to the sparse Riesz condition (SRC) (Zhang [78]) under fixed design, is used only for lower bounds but not for upper bounds.

Assumption SRC: For some $\gamma > 0$, there exit 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 5. *Suppose the noise ε follows a normal distribution with mean 0 and variance $0 < \sigma^2 < \infty$.*

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

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

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

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

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

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

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

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

Remark 10. Note that in (i), at the transition from $m_* > 1$ to $m_* = 1$, i.e., $nt_n^2\tau \approx 1 + \log\frac{M_n}{(nt_n^2\tau)^{q/2}}$, we see continuity:

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

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. Normality of the errors is not essential at all for the lower bounds. With some additional efforts, one can show that these lower rates are also valid under Assumption Y2, which we will not give here.

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 Theorem 6 below) are held constant in the relevant expressions.

Theorem 6. *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

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

Remark 11. When considering jointly the ℓ_q-constraint for a fixed $0 < q \le 1$ and $q = 0$, since the associated function classes are not nested, one cannot immediately deduct the optimal rate of convergence for their intersection. In our problem, the simple rule works: when the upper bound k_n of the ℓ_0-constraint is smaller than the effective model size m_*, the additional ℓ_q-constraint does reduce the parameter searching space, but this reduction is not essential and the rate is equal to the rate for $q = 0$. In contrast, when the effective model size m_* is smaller than k_n, the ℓ_0-constraint does reduce the parameter searching space determined by the ℓ_q-constraint, but not essential from the uniform estimation standpoint and the rate is then $m_* \log(1 + M_n/m_*)/n$. Clearly, both rates can be interpreted as the log number of models of size k_n or m_* over the sample size.

5. Adaptive minimax estimation under fixed design

Consider the linear regression model (4.1) under fixed design, $Y_i = f_0(\mathbf{x}_i) + \varepsilon_i$, $i = 1, ..., n$, where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,M_n})' \in \mathcal{X} \subset \mathbb{R}^{M_n}$ are fixed, $1 \le i \le n$, and the random errors ε_i are i.i.d. $N(0, \sigma^2)$. Suppose $\max_{1\le j\le M_n} \sum_{i=1}^n x_{i,j}^2/n \le 1$. Let $f_0^n = (f_0(\mathbf{x}_1), \ldots, f_0(\mathbf{x}_n))'$. For any function f: $\mathcal{X} \to \mathbb{R}$, define the norm $\|\cdot\|_n$ by $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f^2(\mathbf{x}_i)$. Our goal is to estimate the regression mean f_0^n through a linear combination of the predictors with the coefficients θ satisfying a ℓ_q-constraint ($0 \le q \le 1$). For an estimate $\hat{f}$ of f_0, define its average squared error to be

$$ASE(\hat{f}) = \|\hat{f} - f_0\|_n^2.$$

We consider subset selection based estimators. Let $J_m \subset \{1, 2, \ldots, M_n\}$ be a model of size m ($1 \le m \le M_n$). Our strategy is to choose a model using a model selection criterion, and the resulting least squares estimator is used for f_0^n. The loss of a given model J_m is $ASE(\hat{f}_{J_m}) = \|\hat{Y}_{J_m} - f_0^n\|_n^2$ (with a slight abuse of notation), where $\hat{Y}_{J_m} = (\hat{Y}_{1,J_m}, \ldots, \hat{Y}_{n,J_m})'$ is the projection onto the column span of the design matrix of model J_m. The alternative strategy of model mixing will be taken as well. Although our estimators do not directly consider the ℓ_q-constraint, it will be shown to automatically adapt to the sparsity of f_0 in terms of ℓ_q-representation by the dictionary.

For a function class $\mathcal{F}$, for the fixed design, define the approximation error $d_n^2(f_0; \mathcal{F}) = \inf_{f\in\mathcal{F}} \|f - f_0\|_n^2$. We will consider both σ known and σ unknown cases. As will be seen, the results are quite different in some aspects, and an understanding on what the different assumptions can lead to is important to reach a deeper insight on the theoretical issues.

5.1. When σ is known

For a model J_m of size m ($1 \leq m \leq M_n$), the ABC criterion proposed in Yang (1999) is

$$ABC(J_m) = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i,J_m})^2 + 2r_{J_m}\sigma^2 + \lambda\sigma^2 C_{J_m},$$

where λ is a pure constant, r_{J_m} is the rank of the design matrix of J_m, and C_{J_m} is the model index descriptive complexity. Let r_{M_n} denote the rank of the full model J_{M_n}, which is assumed to be at least 1.

Let $\bar{J}$ denote the model that gives the full projection matrix $I_{n\times n}$ (since the ASE at the design points is the loss of interest, this identity projection is permitted). We define $ABC(\bar{J}) = 2n\sigma^2 + \lambda\sigma^2 C_{\bar{J}}$. Let J_0 denote the null model that only includes the intercept and define $ABC(J_0) = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2\sigma^2 + \lambda\sigma^2 C_{J_0}$, where $\bar{Y} = \sum_{i=1}^{n} Y_i/n$. The model index descriptive complexity C_J satisfies $C_J > 0$ and $\sum_J e^{-C_J} \leq 1$, where the summation is over all the candidate models being considered.

The subset models of size $1 \leq m \leq M_n \wedge n$, the models J_0 and $\bar{J}$ are considered with the complexity $C_{J_m} = -\log 0.85 + \log((M_n - 1) \wedge n) + \log\binom{M_n}{m}$ for a subset model with $m < M_n$, $C_{J_{M_n}} = -\log 0.05$ for the full model J_{M_n}, $C_{J_0} = -\log 0.05$ for the null model J_0, and $C_{\bar{J}} = -\log 0.05$ for the full projection model $\bar{J}$. Note that for the purpose of estimating f_0^n, there is no problem with duplication in the list of candidate models.

Let Γ_n denote the set of all the models considered and the model chosen by the ABC criterion is

$$\hat{J} = \arg\min_{J\in\Gamma_n} ABC(J).$$

The ABC estimator $\hat{f}_{\hat{J}}$ is the fitted value $\hat{Y}_{\hat{J}}$. Let $\bar{f}_J = \mathcal{P}_J f_0^n$ be the projection of f_0^n into the column space of the design matrix of model J.

For ease of presentation, define $\Phi^{\mathcal{F}}$ as follows:

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

$$\Phi^{\mathcal{F}_0(k_n;M_n)} = \frac{\sigma^2 k_n \left(1 + \log \frac{M_n}{k_n}\right)}{n} \wedge \frac{\sigma^2 r_{M_n}}{n},$$
$$\Phi^{\mathcal{F}_q(t_n;M_n)\cap\mathcal{F}_0(k_n;M_n)} = \Phi^{\mathcal{F}_q(t_n;M_n)} \wedge \Phi^{\mathcal{F}_0(k_n;M_n)}.$$

In addition, for lower bound results, let $\underline{\Phi}^{\mathcal{F}_q(t_n;M_n)}$ $(0 \le q \le 1)$ and $\underline{\Phi}^{\mathcal{F}_q(t_n;M_n)\cap\mathcal{F}_0(k_n;M_n)}$ $(0 < q \le 1)$ be the same as $\Phi^{\mathcal{F}_q(t_n;M_n)}$ and $\Phi^{\mathcal{F}_q(t_n;M_n)\cap\mathcal{F}_0(k_n;M_n)}$, respectively, except that when $0 < q \le 1$ and $m_* = 1$, $\underline{\Phi}^{\mathcal{F}_q(t_n;M_n)}$ takes the value $t_n^2 \wedge \frac{\sigma^2 r_{M_n}}{n}$ instead of $(t_n^2 \vee \frac{\sigma^2}{n}) \wedge \frac{\sigma^2 r_{M_n}}{n}$ and $\underline{\Phi}^{\mathcal{F}_q(t_n;M_n)\cap\mathcal{F}_0(k_n;M_n)}$ is modified the same way. In the fixed design case, the ranks of the design matrices are certainly relevant in risk bounds (see, e.g., [65, 54]).

Theorem 7. *When $\lambda \ge 5.1 \log 2$, the ABC estimator $\hat{f}_{\hat{J}}$ simultaneously has the following properties.*

(i) For $\mathcal{F} = \mathcal{F}_q(t_n; M_n)$ with $0 < q \le 1$, or $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ with $1 \le k_n \le M_n$, or $\mathcal{F} = \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)$ with $0 < q \le 1$ and $1 \le k_n \le M_n$, we have

$$\sup_{f_0 \in \mathcal{F}} E(ASE(\hat{f}_{\hat{J}})) \le B\Phi^{\mathcal{F}},$$

where the constant B depends only on q and λ for the first and third cases of $\mathcal{F}$, and depends only on λ for the second case.

(ii) In general, for an arbitrary f_0^n, we have

$$E(ASE(\hat{f}_{\hat{J}})) \le B \left(\left\| \bar{f}_{J_{M_n}} - f_0^n \right\|_n^2 + \inf_{J_m: 1 \le m < M_n} \left(\left\| \bar{f}_{J_m} - \bar{f}_{J_{M_n}} \right\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log(M_n \wedge n)}{n} + \frac{\sigma^2 \log \binom{M_n}{m}}{n} \right) \wedge \frac{\sigma^2 r_{M_n}}{n} \right) \wedge B \left(\left(\| \bar{f}_{J_0} - f_0^n \|_n^2 + \frac{\sigma^2}{n} \right) \wedge \sigma^2 \right),$$

where the constant B depends only on λ.

Remark 12. In (i), the case $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ does not require $\max_{1 \le j \le M_n} \sum_{i=1}^n x_{i,j}^2 / n \le 1$.

Remark 13. In pursuing the best performance in each case of $\mathcal{F}$, the general risk bound in (ii) reduces to $B\Phi^{\mathcal{F}}$ plus the approximation error $d_n^2(f_0; \mathcal{F}) = \inf_{f \in \mathcal{F}} \|f - f_0\|_n^2$.

For the lower bound results, as before, additional conditions are needed. Let Ξ denote the design matrix of the full model J_{M_n}.

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 $\|\theta\|_0 \leq \min(2\gamma, M_n)$, we have

$$\underline{a}\|\theta\|_2 \leq \frac{1}{\sqrt{n}}\|\Xi\theta\|_2 \leq \overline{a}\|\theta\|_2.$$

This condition is slightly weaker than Assumption 2 in [53], which was used to derive minimax lower bounds for $0 < q \leq 1$.

Theorem 8. *Suppose the noise ε follows a normal distribution with mean 0 and variance $0 < \sigma^2 < \infty$. For $\mathcal{F} = \mathcal{F}_q(t_n; M_n)$ with $0 < q \leq 1$, or $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ with $1 \leq k_n \leq M_n$, or $\mathcal{F} = \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)$ with $0 < q \leq 1$ and $1 \leq k_n \leq M_n$, under Assumption SRC$'$ with $\gamma = m_*$, or k_n, or $k_n \wedge m_*$ respectively, we have*

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}} E(ASE(\hat{f})) \geq B^{'} \underline{\Phi}^{\mathcal{F}},$$

where the estimator $\hat{f}$ is over all estimators, and the constant $B^{'}$ depends only on $\underline{a}$ and $\overline{a}$ for the second case of $\mathcal{F}$ and additionally on q for the first and third cases of $\mathcal{F}$.

Remark 14. If SRC$'$ is not satisfied on the set of all the predictors but is satisfied on a subset of M_0 predictors, as long as $\log \frac{M_n}{m_*}$, $\log \frac{M_n}{k_n}$, and $\log \frac{M_n}{m_* \wedge k_n}$ are of the same order as $\log \frac{M_0}{m_*}$, $\log \frac{M_0}{k_n}$, and $\log \frac{M_0}{m_* \wedge k_n}$, respectively, we get the same risk lower rates. When M_n is really large, this relaxation of SRC$'$ can be much less stringent for application.

For the case $q = 0$, the achievability of the upper rate is a direct consequence of [65]. The lower rates for $q = 0$ and/or 1 are given in [54], where the satisfiability of the SRC$^{'}$ is also worked out. Raskutti et al. [53], under the assumption that the rank of the full design matrix is n, derived the minimax rates of convergence $t_n^q(\log(M_n)/n)^{1-q/2}$ for $0 < q < 1$ in an in-probability sense for linear regression with fixed design with the ℓ_q-constraint when $M_n \gg n$ and $M_n/(t_n^q n^{q/2}) \geq M_n^\kappa$ with some $\kappa \in (0, 1)$. From our result, the ABC estimator simultaneously achieves the minimax rates of convergence for all $0 \leq q \leq 1$ and for all $M_n \geq 2$ and t_n no smaller than order $n^{-1/2}$, and also under the joint constraints when $q = 0$ and $0 < q \leq 1$. We also need to point out that we only work on estimating the regression mean in this work, but [53] showed that, under additional conditions, these upper rates are also valid for the estimation of the parameter θ under the squared error and verified their minimaxity. Concurrent work by Ye and Zhang [73] also derived performance bounds on the coefficient estimation that are optimal in a sense of uniformity over the different designs.

In application, the assumption that $f_0 \in \mathcal{F}_q(t_n; M_n)$ or $f_0 \in \mathcal{F}_0(k_n; M_n)$ may sometimes be too strong to be appropriate. Thus, risk bounds that permit model mis-specification, i.e., $f_0 \notin \mathcal{F}_q(t_n; M_n)$, are desirable. Part (ii) in the upper bound theorem (Theorem 7) shows that the ABC estimator handles model mis-specification. Indeed, for the different ℓ_q-constraints, the risk of the ABC estimator is upper bounded by a multiple of $d_n^2(f_0; \mathcal{F}_q(t_n; M_n))$ plus the earlier upper bounds, respectively. Therefore, model mis-specification or not, our estimator is minimax rate adaptive over the $\ell_{q,t_n}^{M_n}$-hulls without any knowledge about the values of q, t_n and k_n (as long as t_n is not trivially small).

One limitation of this result, from one theoretical point, is that the factor is larger than one in front of $d_n^2(f_0; \mathcal{F})$. When the initial estimates need to be obtained based on the same data available, the multiplying factor being one no longer necessarily has any essential advantage. However, striving for the right constant is theoretically attractive when the elements in the dictionary are observed or truly provided by others.

In that direction, recently, Rigollet and Tsybakov [54], by considering an estimator based on the mixing-least-square-estimators algorithm of Leung and Barron [46] with some specific choice of prior probabilities on the models, have provided in-expectation optimal upper bounds for ℓ_0- and/or ℓ_1-aggregation. With the power of the oracle inequality (or the index of resolvability bound), their estimator is shown to be adaptive over ℓ_0- and ℓ_1-hulls. Their results do not address ℓ_q-aggregation for $0 < q < 1$. We next show that we can have an estimator that handles all $0 \leq q \leq 1$ in generality.

The mixed least squares estimator by the mixing algorithm of Leung and Barron (2006) is given by

$$\hat{f}^{MLS} = \sum_{J \in \Gamma_n} w_J \hat{Y}_J \qquad \text{with} \qquad w_J = \frac{\pi_J \exp\{-\widehat{R}_J/(4\sigma^2)\}}{\sum_{J' \in \Gamma_n} \pi_{J'} \exp\{-\widehat{R}_{J'}/(4\sigma^2)\}},$$

where $\widehat{R}_J = n\|Y - \hat{Y}_J\|_n^2 + 2r_J\sigma^2 - n\sigma^2$ is the unbiased risk estimate for $\hat{Y}_J$. Let the prior on model J be chosen as $\pi_{J_m} = 0.85 \left(((M_n - 1) \wedge n) \binom{M_n}{m} \right)^{-1}$ for $1 \leq m \leq (M_n - 1) \wedge n$, and

$\pi_{J_{M_n}} = \pi_{J_0} = \pi_{\bar{J}} = 0.05$.

Theorem 9. *Suppose $0 < \sigma < \infty$ is known. For any $M_n \geq 1$, $n \geq 1$, the estimator $\hat{f}^{MLS}$ simultaneously has the following properties.*

(i) For any $0 < q \leq 1$, $t_n > 0$,

$$\begin{aligned} E(ASE(\hat{f}^{MLS})) &\leq d_n^2(f_0; \mathcal{F}_q(t_n; M_n)) \\ &+ B_1 \begin{cases} \frac{\sigma^2 r_{M_n}}{n}, & \text{if } m_* = M_n \wedge n, \\ \sigma^{2-q} t_n^q \left(\frac{1+\log \frac{M_n}{(nt_n^2\tau)^{q/2}}}{n} \right)^{1-q/2} \wedge \frac{\sigma^2 r_{M_n}}{n}, & \text{if } 1 < m_* < M_n \wedge n, \end{cases} \end{aligned}$$

and

$$\begin{aligned} E(ASE(\hat{f}^{MLS})) &\leq \left(d_n^2(f_0; \mathcal{F}_q(t_n; M_n)) + \frac{B_1 \left(\sigma^2 (1 + \log M_n) \wedge \sigma^2 r_{M_n} \right)}{n} \right) \\ &\quad \wedge \left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\widetilde{B}_1 \sigma^2}{n} \right), \text{ if } m_* = 1 . \end{aligned}$$

(ii) For $1 \leq k_n \leq M_n$,

$$E(ASE(\hat{f}^{MLS})) \leq d_n^2(f_0; \mathcal{F}_0(k_n; M_n)) + B_2 \left(\frac{\sigma^2 k_n \left(1 + \log \frac{M_n}{k_n}\right)}{n} \wedge \frac{\sigma^2 r_{M_n}}{n} \right).$$

(iii) For any $0 < q \leq 1$, $t_n > 0$, and $1 \leq k_n \leq M_n$,

$$\begin{aligned} E(ASE(\hat{f}^{MLS})) &\leq d_n^2(f_0; \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)) \\ &+ B_3 \begin{cases} \frac{\sigma^2 k_n \left(1+\log \frac{M_n}{k_n}\right)}{n} \wedge \frac{\sigma^2 r_{M_n}}{n} & \text{if } m_* > k_n, \\ \sigma^{2-q} t_n^q \left(\frac{1+\log \frac{M_n}{(nt_n^2\tau)^{q/2}}}{n} \right)^{1-q/2} \wedge \frac{\sigma^2 r_{M_n}}{n} & \text{if } 1 < m_* \leq k_n. \end{cases} \end{aligned}$$

and

$$\begin{aligned} E(ASE(\hat{f}^{MLS})) &\leq \left(d_n^2(f_0; \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)) + \frac{B_3 \left(\sigma^2 (1 + \log M_n) \wedge \sigma^2 r_{M_n} \right)}{n} \right) \\ &\quad \wedge \left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\widetilde{B}_3 \sigma^2}{n} \right), \text{ if } m_* = 1 . \end{aligned}$$

(iv) For every f_0, we have

$$E(ASE(\hat{f}^{MLS})) \leq B_4 \sigma^2.$$

For these cases, the constants $\widetilde{B}_1$, B_2, $\widetilde{B}_3$ and B_4 are pure constants, and B_1 and B_3 depend on q.

Remark 15. From (ii) above, by taking $k_n = 1$, we have

$$E(ASE(\hat{f}^{MLS})) \leq \inf_{1\leq j\leq M_n} \|f_j^n - f_0^n\|_n^2 + B_2\left(\frac{\sigma^2\,(1+\log M_n)}{n} \wedge \frac{\sigma^2 r_{M_n}}{n}\right),$$

where $f_j^n = (x_{1,j}, \ldots, x_{n,j})'$. Thus, we have achieved aggregation for adaptation as well under the fixed design.

The risk upper bounds above when q is restricted to be either 0 or 1 or under both constraints are already given in Theorem 6.1 of [54]. The first four cases given there are clearly reproduced here (note that their cases 3 and 1 are just special case and immediate consequence, respectively, of their case 4, given in our bound in (ii)). Their case 5, a sparse aggregation with k_n estimates as studied in [69] (page 36) and [49] (called D-convex aggregation) is implied by our bound in (iii) with q taken to be 1. In the case $q = 1$, a minor difference is that if $\|\bar{f}_{J_0} - f_0^n\|_n^2$ happens to be of a smaller order than $t_n\left(\frac{1+\log\frac{M_n}{(nt_n^2)^{1/2}}}{n}\right)^{1/2} \wedge \frac{r_{M_n}}{n}$, then our risk bound in (iii) yields a faster rate of convergence. In addition, our inclusion of the full projection model among the candidates guarantees that the risk of our estimator is always bounded, which is not true for the estimator in [54]. Our main contribution here is to handle adaptive ℓ_q-aggregation for the whole range of q between 0 and 1. Note that the upper bounds in the above theorem have already been shown to be minimax-rate optimal under the conditions in Theorem 8.

5.2. A comment on the model selection and model mixing approaches

From the risk bounds in the previous subsection, we see that the model mixing approach leads to the optimal constant 1 in front of the approximation error $d_n^2(f_0;\mathcal{F})$ for the three choices of $\mathcal{F}$, which is not the case for the model selection based estimator. However, the model selection approach may also have its own advantages.

From the proof of Theorem 7 and proof of Theorem 1 in [65], besides the given risk bounds, we also have a general in-probability bound of the form: for any $x > 0$, there are constants c, c'

(absolute constants) and $c^{''}$ (depending on λ and σ^2) such that

$$P\left(\frac{ASE(\hat{f}_{\hat{J}}) + \frac{\lambda\sigma^2 C_{\hat{J}}}{n}}{R_n(f_0)} \geq c + x\right) \leq c^{'} \exp\left(-c^{''} x\right),$$

where $R_n(f_0) = \inf_{J\in\Gamma_n}\left(\left\|\bar{f}_J - f_0^n\right\|_n^2 + \frac{\sigma^2 r_J}{n} + \frac{\lambda\sigma^2 C_J}{n}\right)$ is an index of resolvability, which specializes to the upper bounds in (i) and (ii) of Theorem 7, respectively in those situations. Thus, we know that not only $ASE(\hat{f}_{\hat{J}})$ is at order $R_n(f_0)$ with upper deviation probability exponentially small (in x), but also the complexity of the selected model, $\frac{\lambda\sigma^2 C_{\hat{J}}}{n}$, is upper bounded in probability in the same way as well. In particular, for estimating a linear regression function with the soft or hard (or both) constraint(s) on the coefficients, the ABC estimator converges at rate $\frac{m_*\left(1+\log\frac{M_n}{m_*}\right)}{n} \wedge \frac{r_{M_n}}{n}$ both in expectation and with upper deviation probability exponentially small, where m_* is the corresponding effective model size in each case. Furthermore, the rank (the actual number of free-parameters) of the model selected by ABC is right at order $m_* \wedge r_{M_n}$ with exception probability exponentially small.

For model mixing estimators based on exponential weighting, however, to our knowledge, no result has shown that their losses are generally at the optimal rate in probability. In fact, a negative result is given in [2] that shows that an exponential weighting based estimator *optimal* for aggregation for adaptation (i.e., its risk regret, or the expected excessive loss, is of order $\frac{\log M_n}{n}$) is necessarily *sub-optimal* in probability (with a non-vanishing probability its excessive loss is at least at the much larger order of $\sqrt{\frac{\log M_n}{n}}$) in certain settings.

Thus, we tend to believe that both the model selection and model mixing approaches have their own theoretical strengths in different ways.

5.3. *When σ is unknown*

Needless to say, the assumption that σ is fully known is unrealistic. When σ is unknown but is upper bounded by a known constant $\overline{\sigma} > 0$, similar results for rate of convergence can be obtained with a model selection rule different from ABC.

For this situation, Yang [65] proposed the ABC′ criterion:

$$ABC'(J_m) = \left(1 + \frac{2r_{J_m}}{n - r_{J_m}}\right)\left(\sum_{i=1}^{n}(Y_i - \hat{Y}_{i,J_m})^2 + \lambda\overline{\sigma}^2 C_{J_m}\right),$$

which is a modification of Akaike's FPE criterion [1]. We define $ABC'(\bar{J}) = (1+2n)\,\lambda\overline{\sigma}^2 C_{\bar{J}}$ and $ABC'(J_0) = \left(1 + \frac{2}{n-1}\right)\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + \lambda\overline{\sigma}^2 C_{J_0}\right)$. The list of candidate models and complexity assignments need to be different for the different situations, as described below.

1. When $M_n \le n/2$, all the subset models, J_0 and $\bar{J}$ are considered with the complexity $C_{J_m} = -\log 0.85 + \log(M_n - 1) + \log\binom{M_n}{m}$ for a subset model with $m < M_n$, $C_{J_{M_n}} = C_{J_0} = C_{\bar{J}} = -\log 0.05$.
2. When $M_n > n/2$ and $r_{M_n} \ge n/2$, we only consider models with size $m \le n/2$, the model J_0 and the model $\bar{J}$. Then we assign the complexity $C_{J_m} = -\log 0.8 + \log(\lfloor n/2 \rfloor) + \log\binom{M_n}{m}$ for a subset model, $C_{J_0} = C_{\bar{J}} = -\log 0.1$.
3. When $M_n > n/2$ and $r_{M_n} < n/2$, we only consider models with size $m \le n/2$, the full model J_{M_n}, the null model J_0, and the model $\bar{J}$. We assign the complexity $C_{J_m} = -\log 0.85 + \log(\lfloor n/2 \rfloor) + \log\binom{M_n}{m}$ for a subset model, $C_{J_{M_n}} = C_{J_0} = C_{\bar{J}} = -\log 0.05$.

In any of the cases above, let Γ'_n denote the set of all the models considered. The model chosen by the ABC′ is

$$\hat{J}' = \arg\min_{J \in \Gamma'_n} ABC'(J),$$

producing the ABC′ estimator $\hat{f}_{\hat{J}'} = \hat{Y}_{\hat{J}'}$.

Theorem 10. *When $\lambda \ge 40\log 2$, the ABC′ estimator $\hat{f}_{\hat{J}'}$ simultaneously has the following properties.*

(i) For $\mathcal{F} = \mathcal{F}_q(t_n; M_n)$ with $0 < q \le 1$, or $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ with $1 \le k_n \le M_n$, or $\mathcal{F} = \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)$ with $0 < q \le 1$ and $1 \le k_n \le M_n$, we have

$$\sup_{f_0 \in \mathcal{F}} E(ASE(\hat{f}_{\hat{J}'})) \le B\Phi^{\mathcal{F}},$$

where the constant B depends only on q, λ, $\overline{\sigma}$, σ for the first and third cases of $\mathcal{F}$, and depends only on λ, $\overline{\sigma}$, σ for the second case.

(ii) In general, for an arbitrary f_0^n, we have

$$\begin{aligned} & E(ASE(\hat{f}_{\hat{J}'})) \\ \leq \; & B\left(\|\bar{f}_{J_{M_n}} - f_0^n\|_n^2 + \inf_{J_m: 1 \leq m < M_n}\left(\|\bar{f}_{J_m} - \bar{f}_{J_{M_n}}\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log(M_n \wedge n)}{n}\right.\right. \\ & \left.\left. + \frac{\sigma^2 \log\binom{M_n}{m}}{n}\right) \wedge \frac{\sigma^2 r_{M_n}}{n}\right) \wedge B\left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n}\right) \wedge \sigma^2\right), \end{aligned}$$

where the constant B depends only on λ, $\overline{\sigma}$, σ.

Remark 16. For the results in (i), as seen before, when f_0 is not in the respective class of linear combinations, an obvious modification is needed by adding a multiple of the approximation error $d_n^2(f_0; \mathcal{F})$ in the risk bound.

When $0 < \sigma < \infty$ is fully unknown, a model selection method by Baraud, Giraud and Huet [7] can be used to obtain results on ℓ_q-regression.

They consider a different modification of the FPE criterion [1]:

$$BGH(J_m) = \left(1 + \frac{pen(J_m)}{n - r_{J_m}}\right)\left(\sum_{i=1}^{n}(Y_i - \hat{Y}_{i,J_m})^2\right),$$

where $pen(J_m)$ is a penalty assigned to the model J_m. They devise a new form for $pen(J_m)$ (Section 4.1 in [7]) to yield a nice oracle inequality (Corollary 1) that does not require any knowledge of σ, but at the expense of excluding some large models in the consideration. When $M_n \leq (n-7) \wedge \varsigma n$ for some $0 < \varsigma < 1$, we consider all subset models in the model selection process. When M_n is large, we consider only subset models with $n - r_{J_m} \geq 7$ and $m \vee \log\binom{M_n}{m} \leq \varsigma n$ for a fixed $0 < \varsigma < 1$. Combining the tools developed in this and their papers, we have the following result.

Theorem 11. *The BGH estimator $\hat{f}_{\hat{J}}$ has the following properties.*

(i) When $M_n \leq (n-7) \wedge \varsigma n$, for $\mathcal{F} = \mathcal{F}_q(t_n; M_n)$ with $0 < q \leq 1$, or $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ with $1 \leq k_n \leq M_n$, or $\mathcal{F} = \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)$ with $0 < q \leq 1$ and $1 \leq k_n \leq M_n$, we have

$$\sup_{f_0 \in \mathcal{F}} E(ASE(\hat{f}_{\hat{J}})) \leq B\Phi^{\mathcal{F}},$$

where the constant B depends only on q and ς for the first and third cases of $\mathcal{F}$, and depends on ς for the second case.

(ii) For a general M_n, if m_* satisfies $m_* \leq n-7$ and $m_* \vee \log \binom{M_n}{m_*} \leq \varsigma n$, we have

$$\sup_{f_0 \in \mathcal{F}_q(t_n; M_n)} E(ASE(\hat{f}_{\hat{j}})) \leq B \begin{cases} \sigma^{2-q} t_n^q \left(\frac{1+\log \frac{M_n}{(n t_n^2 \tau)^{q/2}}}{n} \right)^{1-q/2} & \text{if } m_* > 1, \\ t_n^2 \vee \frac{\sigma^2}{n} & \text{if } m_* = 1, \end{cases}$$

where B depends only on q and ς. If k_n satisfies $k_n \leq n-7$ and $k_n \vee \log \binom{M_n}{k_n} \leq \varsigma n$, we have

$$\sup_{f_0 \in \mathcal{F}_0(k_n; M_n)} E(ASE(\hat{f}_{\hat{j}})) \leq B' \frac{\sigma^2 k_n \left(1 + \log \frac{M_n}{k_n}\right)}{n},$$

where B' is a constant that depends only on ς.

Remark 17. As before, when f_0 is not in the respective class, a multiple of the approximation error $d_n^2(f_0; \mathcal{F}) = \inf_{f \in \mathcal{F}} \|f - f_0\|_n^2$ needs to be added in the aggregation risk bound.

From the above theorem, we see that when σ is fully unknown, as long as $M_n \leq (n-7) \wedge \varsigma n$ for some $0 < \varsigma < 1$, similar risk bounds to those in Theorem 10 for ℓ_q-regression hold. However, when M_n is larger, the previous risk bounds are seriously compromised: 1) the possible improvement in risk due to low rank of the full model is no longer guaranteed; 2) the previous upper rates determined by the effective model size m_* or k_n are valid only when those model sizes are not excluded from consideration by the BGH criterion; 3) The risk is no longer guaranteed to be always uniformly bounded. Indeed, due to the restriction on the model sizes to be considered, the final risk here can be arbitrarily large. It turns out that this last aspect is not due to technical deficiency in the analysis, but it is a necessary price to pay for not knowing σ at all (see [61]).

6. Discussion

Since early 1990s, sparse estimation has been recognized as an important tool for multi-dimensional function estimation. Emergence of high-dimensional statistical problems in the information age has prompted an increasing attention on the topic from theoretical, computational and applied perspectives. We focus only on a theoretical standpoint in the discussion below.

To our knowledge, several lines of research on sparse function estimation in 1990s produced theoretical foundations that still provide essential understandings on ways to explore sparsity and associated price to pay when pursuing sparse estimation from minimax perspectives. It has been discovered that for some function classes, sparse representations (in contrast to traditional full approximation) result in faster rates of convergence, which alleviate the curse of dimensionality when the problem size is large. Such function classes include, for example, Besov classes (e.g., [31]), Jones-Barron classes ([9, 42]) and may also be defined directly in terms of sparse approximation (e.g., [71], Section III.D). Regarding methods to achieve the optimal sparse estimation, wavelet thresholding with one or more orthonormal dictionaries and model selection with a descriptive complexity penalty term added to the sum of negative maximized likelihood (or a general contrast function) and a multiple of the model dimension have yielded successful theoretical advancements. Oracle inequalities/index of resolvability bounds have been derived that readily give minimax-rate adaptive estimators for various scenarios. In linear representation, ℓ_1-constraints on the coefficients have been long known to be associated with fast rate of convergence for both orthogonal and non-orthogonal bases by model selection or aggregation methods, as mentioned in the introduction of this paper.

It is worth noticing that these research works usually target nonparametric settings. In the past few years, the situation of a large number of naturally observed predictors has attracted much attention, shifting the focus to much simpler linear modeling. As pointed out earlier, the work in the 1990s on model selection has direct implications for the high-dimensional linear regression. For example, if the sum of the absolute values of the linear coefficients is bounded (ℓ_1-constraint), then the rate of convergence is bounded by $(\log n/n)^{1/2}$ as long as M_n increases only polynomially in n. If only k_n terms have non-zero coefficients (ℓ_0-constraint), then the rate of convergence is of order $k_n(1 + \log(M_n/k_n))/n$ based on model selection with mild conditions on the predictors. However, such subset selection based estimators pose computational challenges in real applications.

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 [34] 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 sparsity natures of the target function in terms of linear approximations subject to ℓ_q-constraints.

Donoho and Johnstone [30] derived insightful general asymptotic minimax risk expressions for estimating the mean vector in ℓ_q-balls ($0 < q < \infty$) under ℓ_p loss ($p \geq 1$) in a Gaussian sequence framework. The work by Raskutti et al. [53] and by Rigollet and Tsybakov [54] are directly related to our work in the fixed design case. The former successfully obtains optimal non-adaptive in-probability loss bounds for their main scenario that M_n is much larger than n for general $0 \leq q \leq 1$ when the true regression function is assumed to be in the $\ell_{q,t_n}^{M_n}$-hull. In contrast, our estimators are adaptive and the risk bounds hold without restrictions on M_n or the "norm" parameter t_n, also allowing the true regression function to be really arbitrary. The work of Rigollet and Tsybakov [54] nicely shows the adaptive aggregation capability of model mixing over ℓ_0 and ℓ_1-balls. Our results are valid over the whole range of $0 \leq q \leq 1$. For lower bounds, our formulation is somewhat different from theirs. In addition, unlike those results, we have also provided results when the error variance is unknown but upper bounded by a known constant or fully unknown. Furthermore, our model selection based estimators have optimal convergence rates also in terms of upper deviation probability, which may not hold for the model mixing estimators. We need to point out that both [53] and [54] have given results on related problems that we do not address in this work.

In our results, the effective model size m_* (as defined in Section 2.5) 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 $1 \wedge \frac{m_*\left(1+\log\left(\frac{M_n}{m_*}\right)\right)}{n}$.

Risk bounds for selection/mixing least squares estimators from a countable collection of linear models (such as given in [65, 46]), together with sparse approximation error bounds, are essential for our approach to devise minimax optimal sparse estimation for fixed design. When the predictors are taken as some initial estimates, the selection/mixing methods can be regarded as aggregation methods with the risk bounds as aggregation risk bounds. In a strict sense, however, these results are not totally satisfactory for at least two reasons. First, the evaluation of performance only at the design points that have been seen already has limited value: i) The strengths of the candidate procedures may not be reflected at all on such a measure; ii) A small ASE on the design points does not mean good behaviors on future predictor values. Second, when the initial estimates are not given (which is almost always the case), to combine arbitrary estimators, data splitting is typically necessary to come up with the candidate estimates and use the rest of the sample for weight assignment. Then, the final risk bounds, unfortunately, depend on how the data are split. In contrast, for the random design case, this is not an issue. We have also seen that because ASE cares only about the performance at the design points, given the i.i.d. normal error assumption, there is absolutely no condition needed on the true regression function, as pointed out in a remark to Theorem 1 in [65]. For random design, however, we have made the sup-norm bound assumption, but the risk bounds guarantee optimal future performance as long as the sampling distribution is unchanged.

Regarding aggregation, we notice that 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 a 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 [58] 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 [69]). 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 $0 < q \leq 1$.

One interesting observation is that aggregation for adaptation is essentially a special case of ℓ_q-aggregation, yet our way of achieving the simultaneous ℓ_q-aggregation is by methods of aggregation for adaptation through model selection/combination.

Aggregation of estimates and regression estimation problems are closely related. For aggregation, besides that the predictors to be aggregated are from some initial estimations (and thus are not directly observed), the emphases are: i) One is unwilling to make assumptions on relationships between the initial estimates so that they can have arbitrary dependence; ii) One is unwilling to make specific assumptions on the true regression function beyond that it is uniformly bounded and hence allow model mis-specification. In this game, there is little interest on the true or optimal coefficients in the representation of the regression function in terms of the initial estimates.

Obviously, there are other directions of aggregation that one may pursue. The ℓ_q-aggregation strategy that relies on aggregating subset choices of the initial estimates, as in [69], while producing the most general aggregation risk bounds so far, follows a global aggregation paradigm, i.e., the linear coefficients are globally determined. It is conceivable that sometimes localized weights may provide better estimation/prediction performance (see, e.g., [70]). Much more work is needed here to result in practically effective localized aggregation methods.

Aggregation of estimates, as an important step in combining statistical procedures, has proven to bring theoretically elegant and practically feasible methods for regression estimation/prediction. It is an important vehicle to share strengths of different function estimation methodologies to produce adaptively optimal and robust estimators that work well under minimal conditions. Aggregation by mixing certainly cannot replace model selection when selection of an estimator among candidates or a set of predictors is essential for interpretation or business/operational decisions.

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.

7. General oracle inequalities for random design

Consider the setting in Section 3.2.

Theorem 12. *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 theorem, 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 [10], $IR(f_0)$ 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. (i) For the **T-C** strategy,

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

For the **T-Y** strategy,

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

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

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

For the **AC-Y** strategy,

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

□

Remark 18. Similar oracle inequalities hold for the estimator $\hat{f}_A$ under the linear regression setting with random design: $d^2(f_0; \mathcal{F}_{J_m})$ is replaced by $d^2(f_0; \mathcal{G}_{J_m})$, and $\sum_{j\in J_m} \theta_j f_j$ is replaced by $\sum_{j\in J_m} \theta_j x_j$ in the above theorem.

8. Proofs

Proof of Theorem 1.

Proof. (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., [32], page 98). Thus, there are 2^{k-1} functions g_j, $1 \le j \le 2^{k-1}$, such that

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

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 $\sum_{i=1}^{M_n} |c_i| \le \epsilon_k$. We define a random function U, such that

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

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

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

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 the 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 $|L^*|t_n^q/m < \sum |c_j|^q \le t_n^q$. So, $|L^*| < m$. Also,

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

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 $\mathbb{P}(U = D\text{sign}(c_j)f_j) = |c_j|/D$, $j \notin L^*$ and $\mathbb{P}(U = 0) = 1 - \sum_{j\notin L^*} |c_j|/D$. Thus, $\mathbb{E}U = w^*$, where $\mathbb{E}$ denotes expectation with respect to the randomness $\mathbb{P}$ (just introduced). Also, $\|U\| \le D\sup_{1\le j\le M_n} \|f_j\| \le D$. 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,

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

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

□

Proof of Theorem 2.

Proof. To derive the upper bounds, we only need to examine the index of resolvability for each strategy. The natures of the constants in Theorem 2 follow from Theorem 12.

(i) For **T-** strategies, according to Theorem 1 and the general oracle inequalities in Theorem 12, 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

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

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 1, we have

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

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

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

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

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

The upper bound for this case then follows.

For $\mathcal{F} = \mathcal{F}_0(k_n)$, by evaluating the index of resolvability from Theorem 12 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 $\|\theta\|_1 \le s$ ($s > 0$), $\|\theta\|_q \le t_n$ ($0 \le q \le 1$, $t_n > 0$) and $\|f_\theta\|_\infty \le L$ are imposed on the coefficients. Notice that $\|\theta\|_1 \le \|\theta\|_q$ when $0 < q \le 1$, 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.

□

Global metric entropy and local metric entropy. The tools developed in Yang and Barron [72] allow us to derive minimax lower bounds for ℓ_q-aggregation of estimates or regression under ℓ_q-constraints. Both global and local entropies of the regression function classes are relevant. The following lower bound result slightly generalizes Lemma 1 in [69].

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

$$Y = f_0(\mathbf{X}) + \sigma \cdot \varepsilon, \tag{8.1}$$

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

$$\begin{aligned} \underline{M}_\alpha^{\text{loc}}(\sigma\epsilon_n) &\geq n\epsilon_n^2 + 2\log 2, & (8.2)\\ \overline{M}(\sqrt{2}\sigma\bar{\epsilon}_n) &= n\bar{\epsilon}_n^2, & (8.3)\\ \underline{M}(\sigma\underline{\epsilon}_n) &= 4n\bar{\epsilon}_n^2 + 2\log 2. & (8.4) \end{aligned}$$

Proposition 5. *(Yang and Barron [72]) The minimax risk for estimating f_0 from model (8.1) 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 [72], but not in the concrete forms. The second lower bound is given in [69]. 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, the mutual information between Θ and the observations $(\mathbf{X}_i, Y_i)_{i=1}^n$ is upper bounded by $\frac{n}{2}\epsilon_n^2$ (see Yang and Barron [72], Sections 7 and 3.2) and an application of Fano's inequality 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 $|N|$ denote 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 [72], 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 [72], the same lower bound holds for $\underline{\mathcal{F}}$ as well.

$\square$

Proof of Theorem 3.

Proof. 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 $B_q(t_n; M_n) = \{\theta : \|\theta\|_q \leq t_n\}$. 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 $\|\theta\|_q \leq {M_n}^{1/q-1/2}\|\theta\|_2$ for $0 < q \leq 2$, for $\epsilon \leq \sqrt{M_n/n}$, taking $f_0^* \equiv 0$, we have

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

where the last equality holds because when $\epsilon \leq \sqrt{M_n/n}$, for $\|\theta\|_2 \leq \epsilon$, $\|\theta\|_q \leq t_n$ is always satisfied. Consequently, for $\epsilon \leq \sqrt{M_n/n}$, the $(\epsilon/2)$-packing of $\mathcal{B}(f_0^*; \epsilon)$ under the L_2 distance is equivalent to the $(\epsilon/2)$-packing of $B_\epsilon = \{\theta : \|\theta\|_2 \leq \epsilon\}$ 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 $M_{1/2}^{\text{loc}}(\epsilon)$ of $\mathcal{F}_q(t)$ under the L_2 distance is at least $\underline{M}_{1/2}^{\text{loc}}(\epsilon) = M_n \log 2$ for $\epsilon \leq \sqrt{M_n/n}$. The minimax lower bound for the case of $\widetilde{m}^* = M_n$ then directly follows from Proposition 5.

When $1 < \widetilde{m}_* \leq \widetilde{m}^* < M_n$, the use of global entropy is handy. Applying the minimax lower bound in terms of global entropy in Proposition 5, with the metric entropy order for larger ϵ (which is tight in our case of orthonormal functions in the dictionary) from Theorem 1, 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 1 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 5.

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 [53] (the result there actually also holds when requiring the pairwise Hamming distance to be strictly larger than $k/2$; see also the derivation of a metric entropy lower bound in [45]), 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 5 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 [45], 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 5. 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 5. To prove the results for the case $M_n/\left(1+\log \frac{M_n}{k_n}\right) \le bn$, as in [58], 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 5, 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 $\sqrt{k_n\left(1+\log \frac{M_n}{k_n}\right)/n}$.

□

Proof of Theorem 4.

Proof. For linear regression with random design, we assume the true regression function 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 $1 \leq m \leq M_n \wedge n$, according to the general oracle inequalities in Theorem 12 , the adaptive estimator $\hat{f}_A$ has

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

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 $1 < m_* \leq m^* < M_n \wedge n$, the similar argument of Theorem 2 leads to an upper bound of order $1 \wedge \frac{m_*}{n} \left(1 + \log \frac{M_n}{m_*}\right)$. Since $(nt_n^2)^{q/2} \left(1 + \log \frac{M_n}{(nt_n^2)^{q/2}}\right)^{-q/2} \leq m_* \leq 4(nt_n^2)^{q/2} \left(1 + \log \frac{M_n}{2(nt_n^2)^{q/2}}\right)^{-q/2}$, 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 $\|f_0\|^2 + \frac{1}{n} \leq t_n^2 + \frac{1}{n} \leq 2(t_n^2 \vee \frac{1}{n})$ 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 2.

(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 2. For $\mathcal{F} = \mathcal{F}_0^L(k_n; M_n)$, the approximation error is $\inf_{s \geq 1} \left(\inf_{\{\theta: \|\theta\|_1 \leq s, \|\theta\|_0 \leq k_n, \|f_\theta\|_\infty \leq L\}} \|f_\theta - f_0\|^2 + 2c_3 \frac{\log(1+s)}{n} \right) \leq \inf_{\{\theta: \|\theta\|_1 \leq \alpha_n, \|\theta\|_0 \leq k_n, \|f_\theta\|_\infty \leq L\}} \|f_\theta - f_0\|^2 + 2c_3 \frac{\log(1+\alpha_n)}{n} = 2c_3 \frac{\log(1+\alpha_n)}{n}$ if $f_0 \in \mathcal{F}_0^L(k_n; M_n)$. The upper bound then follows.

□

Proof of Theorem 5.

Proof. Without loss of generality, we assume $\sigma^2 = 1$ for the error variance. First, we give a simple fact. Let $B_k(\eta) = \{\theta : \|\theta\|_2 \leq \eta, \|\theta\|_0 \leq k\}$ and $\mathcal{B}_k(f_0;\epsilon) = \{f_\theta : \|f_\theta\| \leq \epsilon, \|\theta\|_0 \leq k\}$ (take $f_0 = 0$). Then, under Assumption SRC with $\gamma = k$, the $\frac{\underline{a}}{2\overline{a}}$-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 3. 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 5. The key is to realize that in the derivation of the metric entropy lower bound for $\{\theta : \|\theta\|_q \leq t_n\}$ in [45], 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 $m^* \leq \left(1 + \log \frac{M_n}{m^*}\right)^{q/2}$, observe that $\mathcal{F}_q(t_n; M_n) \supset \{\beta x_j : |\beta| \leq t_n\}$ for any $1 \leq j \leq M_n$. The use of the local entropy result in Proposition 5 readily gives the desired result.

(ii) As in the proof of Theorem 3, without loss of generality, we can assume $k_n \leq M_n/2$. Together with the simple fact given at the beginning of the proof, for $B_{k_n}(\epsilon/\overline{a}) = \{\theta : \|\theta\|_2 \leq \epsilon/\overline{a}, \|\theta\|_0 \leq k_n\}$, with $\eta' = \epsilon/(\overline{a}\sqrt{k_n})$, we know $B_{k_n}(\epsilon/\overline{a})$ contains the set

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

For $\theta_1 = \eta' I_1, \theta_2 = \eta' I_2$ both in the above set, by Assumption SRC, $\|f_{\theta_1} - f_{\theta_2}\|^2 \geq \underline{a}^2 \eta'^2 d_{HM}(I_1, I_2) \geq \underline{a}^2\epsilon^2/(2\overline{a}^2)$ 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 3 (case 2), we know the local entropy $M^{loc}_{\underline{a}/(\sqrt{2}\overline{a})}(\epsilon)$ of $\mathcal{F}_0(k_n; M_n) \cap \{f_\theta : \|\theta\|_2 \leq a_n\}$ with $a_n \geq \epsilon$ is lower bounded by $\frac{k_n}{2} \log \frac{2(M_n - k_n)}{k_n}$. Then, under the condition $a_n \geq C\sqrt{k_n\left(1 + \log \frac{M_n}{k_n}\right)/n}$ 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 3 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 $\|\theta\|_0 \leq k_n$, $\|\theta\|_q \leq k_n^{1/q-1/2}\|\theta\|_2 \leq k_n^{1/q-1/2}\sqrt{Ck_n\left(1+\log\frac{M_n}{k_n}\right)/n}$ for θ with $\|\theta\|_2 \leq \sqrt{Ck_n\left(1+\log\frac{M_n}{k_n}\right)/n}$. 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.

□

Proof of Theorem 6.

Proof. (i) We only need to derive the lower bound part. Under the assumptions that $\sup_j \|X_j\|_\infty \leq L_0 < \infty$ for some constant $L_0 > 0$, for a fixed $t_n = t > 0$, we have $\forall f_\theta \in \mathcal{F}_q(t_n; M_n)$, $\|f_\theta\|_\infty \leq \sup_j \|X_j\|_\infty \cdot \sum_{j=1}^{M_n} |\theta_j| \leq L_0\|\theta\|_1 \leq L_0\|\theta\|_q \leq L_0 t$. Then the conclusion follows directly from Theorem 5 (Part (i)). Note that when t_n is fixed, the case $m_* = 1$ needs not to be separately considered.

(ii) For the upper rate part, we use the **AC-C** upper bound. For f_θ with $\|\theta\|_\infty \leq L_0$, clearly, we have $\|\theta\|_1 \leq M_n L_0$, and consequently, since $\log(1+M_nL_0)$ 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 Theorem 4. Under the assumptions that $\sup_j \|X_j\|_\infty \leq L_0 < \infty$ and $k_n\sqrt{\left(1+\log\frac{M_n}{k_n}\right)/n} \leq \sqrt{K_0}$, we know that $\forall f_\theta \in \mathcal{F}_0(k_n; M_n)\bigcap\{f_\theta : \|\theta\|_2 \leq a_n\}$ 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

$$\|\sum_{j=1}^{M_n}\theta_j x_j\|_\infty \leq L_0\|\theta\|_1 \leq L_0\sqrt{k_n}a_n = CL_0k_n\sqrt{\frac{1+\log\frac{M_n}{k_n}}{n}} \leq C\sqrt{K_0}L_0.$$

Then the functions in $\mathcal{F}_0(k_n; M_n)\bigcap\{f : \|\theta\|_2 \leq a_n\}$ have sup-norm uniformly bounded. Note that for bounded a_n, $\|\theta\|_2 \leq a_n$ implies that $\|\theta\|_\infty \leq a_n$. Thus, the extra restriction $\|\theta\|_\infty \leq L_0$ does not affect the minimax lower rate established in part (ii) of Theorem 5.

(iii) The upper and lower rates follow similarly from Theorems 4 and 5. The details are thus skipped.

□

Now we turn to the setup in Section 5 with σ^2 known.

Proposition 6. *(Yang [65], Theorem 1) When $\lambda \geq 5.1\log 2$, we have*

$$E(ASE(\hat{f}_{\hat{J}})) \leq B \inf_{J\in\Gamma_n}\left(\left\|\bar{f}_J - f_0^n\right\|_n^2 + \frac{\sigma^2 r_J}{n} + \frac{\lambda\sigma^2 C_J}{n}\right),$$

where $B > 0$ is a constant that depends only on λ.

Proof of Theorem 7.

Proof. The general case (ii) is easily derived based on our estimation procedure and Proposition 6.

To prove (i), when $\mathcal{F} = \mathcal{F}_q(t_n; M_n)$, according to the upper bound in (ii) and Theorem 1, when $f_0^n \in \mathcal{F}_q(t_n; M_n)$, for any $1 \leq m \leq (M_n-1)\wedge n$, there exists a subset J_m and $f_{\theta^m} \in \mathcal{F}_{J_m}$ such that

$$\begin{aligned}
& E(ASE(\hat{f}_{\hat{J}})) \\
\leq\ & B\left(\left(\|f_{\theta^m} - f_0^n\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2\log(M_n\wedge n)}{n} + \frac{\sigma^2\log\binom{M_n}{m}}{n}\right)\wedge\frac{\sigma^2 r_{M_n}}{n}\right) \\
& \wedge B\left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n}\right)\wedge\sigma^2\right) \\
\leq\ & B\left(\left(2^{2/q-1}t_n^2 m^{1-2/q} + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2\log(M_n\wedge n)}{n} + \frac{\sigma^2\log\binom{M_n}{m}}{n}\right)\wedge\frac{\sigma^2 r_{M_n}}{n}\right) \\
& \wedge B\left(\left(t_n^2 + \frac{\sigma^2}{n}\right)\wedge\sigma^2\right).
\end{aligned}$$

Since $\log\binom{M_n}{m} \leq m\left(1+\log\frac{M_n}{m}\right)$ and $\log M_n \leq m\left(1+\log\frac{M_n}{m}\right)$, then for models with size $1 \leq m \leq (M_n-1)\wedge n$, we have

$$\begin{aligned}
E(ASE(\hat{f}_{\hat{J}})) \leq\ & B'\left(\left(t_n^2 m^{1-2/q} + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 m\left(1+\log\frac{M_n}{m}\right)}{n}\right)\wedge\frac{\sigma^2 r_{M_n}}{n}\right) \\
& \wedge B\left(\left(t_n^2 + \frac{\sigma^2}{n}\right)\wedge\sigma^2\right),
\end{aligned}$$

where B' only depends on q and λ.

When $m_* = m^* = M_n \wedge n$, the full model J_{M_n} leads to an upper bound of order $\frac{\sigma^2 r_{M_n}}{n}$. When $1 < m_* \leq m^* < M_n \wedge n$, we get the desired upper bounds by evaluating the risk bounds choosing J_{m_*} and J_{M_n}. When $m_* = 1$, models J_0 and J_{M_n} result in the desired upper bound.

The arguments for cases $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ and $\mathcal{F} = \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)$ are similar to those of Theorem 2 and above with r_{J_m} replacing m in the upper bounds.

□

Proof of Theorem 8.

Proof. Without loss of generality, assume the error variance $\sigma^2 = 1$. Let $P_f(y^n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{1}{2}(y_i - f(\mathbf{x}_i)^2)\big)$ denote the joint density of $Y^n = (Y_1, ..., Y_n)^{'}$, where the components are independent with mean $f(\mathbf{x}_i)$ and variance 1, $1 \leq i \leq n$. Then the Kullback-Leibler distance between $P_{f_1}(y^n)$ and $P_{f_2}(y^n)$ is

$$D\left(P_{f_1}(y^n) \parallel P_{f_2}(y^n)\right) = \frac{1}{2}\sum_{i=1}^n \left(f_1(\mathbf{x}_i) - f_2(\mathbf{x}_i)\right)^2.$$

To prove the lower bounds, instead of the global L_2 distance on the regression functions, we need to work with the distance $d(f_1, f_2) = \sqrt{\sum_{i=1}^n \left(f_1(\mathbf{x}_i) - f_2(\mathbf{x}_i)\right)^2}$.

First consider the case $\mathcal{F} = \mathcal{F}_q(t_n; M_n)$. Let $B_k(\eta) = \{\theta : \|\theta\|_2 \leq \eta, \|\theta\|_0 \leq k\}$ and $\mathcal{B}_k(f_0; \epsilon) = \{f_\theta : \|f_\theta\|_n \leq \epsilon, \|\theta\|_0 \leq k\}$ $(f_0 = 0)$. Then, under Assumption SRC$'$ with $\gamma = k$, the $\frac{\underline{a}}{2\bar{a}}$-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 $\eta = \frac{\epsilon}{\bar{a}}$. When $\gamma = m_*$, the proof is the same as the proof of Theorem 5.

Now consider the case $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ and again assume $k_n \leq M_n/2$ as in the proof of Theorem 5. When Assumption SRC$'$ holds with $\gamma = k_n$, the lower bound is of order $\frac{k_n(1+\log M_n/k_n)}{n}$ as before in the random design case. The proof for the last case $\mathcal{F} = \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)$ is similarly done as in the proof of Theorem 5.

□

Proof of Theorem 9.

Proof. According to Corollary 6 from [46], we have

$$E(ASE(\hat{f}^{MLS})) \leq \inf_{J \in \Gamma_n} \left(\|\bar{f}_J - f_0^n\|_n^2 + \frac{\sigma^2 r_J}{n} + \frac{4\sigma^2 \log(1/\pi_J)}{n} \right),$$

which is basically the same as Proposition 6 with $B = 1$. Thus, the rest of the proof is basically the same as that of Theorem 7.

□

To prove Theorem 10 , we need an oracle inequality, which improves Theorem 4 of [65], where only a convergence in probability result is given. Suppose that only the subset models J_m with rank $r_{J_m} \leq n/2$ are considered (which is automatically satisfied when $M_n \leq n/2$). Let Γ denote these models. (More generally, a risk bound similar to the following holds if we consider models with size no more than $(1-\rho)n$ for any small $\rho > 0$.) Let C_J be the descriptive complexity of the model J in Γ.

Proposition 7. *When $\lambda \geq 40 \log 2$, the selected model $\hat{J}'$ by ABC′ satisfies*

$$E(ASE(\hat{f}_{\hat{J}'})) \leq B \inf_{J \in \Gamma} \left(\left\|\bar{f}_J - f_0^n\right\|_n^2 + \frac{\sigma^2 r_J}{n} + \frac{\lambda \sigma^2 C_J}{n} \right),$$

where B is a constant that depends on λ, $\overline{\sigma}^2$, and σ^2.

Remark 19. If we add models with rank $r_J > n/2$ into the competition, as long as the complexity assignment over all the models is valid (i.e., satisfying the summability condition), if we can show that for these added models, $ABC'(J)$ are also upper and lower bounded with high probabilities as in (8.5) and (8.6), then the risk bound in the proposition continues to hold.

Proof. Let $e_n = (\varepsilon_1, \ldots, \varepsilon_n)'$. For ease in writing, we simplify $\|\cdot\|_n^2$ to $\|\cdot\|^2$ in this proof. From page 495 in [65], for each candidate model J, we have

$$ABC'(J) = \|A_J f_0^n\|^2 + r_J \left(\frac{2}{n - r_J} \left(\|Y_n - \hat{Y}_J\|^2 + \lambda \overline{\sigma}^2 C_J \right) - \sigma^2 \right) + \lambda \overline{\sigma}^2 C_J + 2\mathrm{rem}_1(J) + \mathrm{rem}_2(J),$$

where $\|A_J f_0^n\|^2 = \|\overline{f}_J - f_0^n\|^2$, $\text{rem}_1(J) = e_n'(f_0^n - M_J f_0^n)$ and $\text{rem}_2(J) = r_J - e_n' M_J e_n$. Note also that $\|Y_n - \hat{Y}_J\|^2 + \lambda\overline{\sigma}^2 C_J = \|A_J f_n\|^2 + (n - r_J)\sigma^2 + \big(e_n' A_J e_n - (n - r_J)\sigma^2\big) + 2e_n' A_J f_n + \lambda\overline{\sigma}^2 C_J$. Let

$$T(J) = \|A_J f_0^n\|^2 + (n - r_J)\sigma^2 + \lambda\overline{\sigma}^2 C_J, \text{ and } nR_n(J) = \|A_J f_0^n\|^2 + r_J\sigma^2 + \lambda\overline{\sigma}^2 C_J.$$

As is shown in the proof of Theorem 1, [65], if $\lambda > h(\tau_1, \tau_2) = \max(\sup_{\xi\geq 0}((2(\log 2)\xi)^{1/2}/\tau_1 - \xi), \sup_{\rho\geq 0}(\rho/\tau_2 - 1)2(\log 2)/(\rho - \log(\rho+1)))$ for some constants τ_1 and τ_2 with $2\tau_1 + \tau_2 < 1$, then for any $\delta > 0$, with probability no less than $1 - 5\delta$, $|\text{rem}_1(J)| \leq \tau_1(nR_n(J) + g_1(\delta))$, $|\text{rem}_2(J)| \leq \tau_2(nR_n(J) + g_2(\delta))$, and $|e_n' A_J e_n - (n - r_J)\sigma^2| \leq \tau_2(T(J) + g_2(\delta))$, where $g_1(\delta) = g_2(\delta) = \lambda \log_2(1/\delta)$. Then with probability no less than $1 - 5\delta$, we have

$$\begin{aligned}
ABC'(J) \;\geq\;& \|A_J f_0^n\|^2 + r_J\left(\frac{2(T(J) - \tau_2(T(J) + g_2(\delta)) - 2\tau_1(nR_n(J) + g_1(\delta)))}{n - r_J} - \sigma^2\right) \\
& -2\tau_1(nR_n(J) + g_1(\delta)) - \tau_2(nR_n(J) + g_2(\delta)) + \lambda\overline{\sigma}^2 C_J \\
\geq\;& \|A_J f_0^n\|^2 + r_J\left(\frac{2(1 - (2\tau_1 + \tau_2))T(J)}{n - r_J} - \frac{2(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{n - r_J} - \sigma^2\right) \\
& -(2\tau_1 + \tau_2)nR_n(J) - (2\tau_1 g_1(\delta) + \tau_2 g_2(\delta)) + \lambda\overline{\sigma}^2 C_J \\
\geq\;& \|A_J f_0^n\|^2 + r_J(1 - (4\tau_1 + 2\tau_2))\sigma^2 - \frac{2r_J(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{n - r_J} \\
& -(2\tau_1 + \tau_2)nR_n(J) - (2\tau_1 g_1(\delta) + \tau_2 g_2(\delta)) + \lambda\overline{\sigma}^2 C_J \\
\geq\;& (1 - (6\tau_1 + 3\tau_2))nR_n(J) - \frac{n + r_J}{n - r_J}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta)). \qquad (8.5)
\end{aligned}$$

Suppose $6\tau_1 + 3\tau_2 < 1$. Let J_n be the candidate model that minimizes $R_n(J)$. Then with exception probability less than 5δ, we have

$$\begin{aligned}
ABC'(J_n) \;\leq\;& \|A_{J_n} f_0^n\|^2 + r_{J_n}\left(\frac{2(1 + (2\tau_1 + \tau_2))T(J_n)}{n - r_{J_n}} - \sigma^2\right) + (2\tau_1 + \tau_2)nR_n(J_n) \\
& + \frac{n + r_{J_n}}{n - r_{J_n}}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta)) + \lambda\overline{\sigma}^2 C_{J_n}.
\end{aligned}$$

Since $T(J_n)/(n - r_{J_n}) = (1 + r_{J_n}/(n - r_{J_n}))R_n(J_n) + (1 - r_{J_n}/(n - r_{J_n}))\sigma^2 \leq 2R_n(J_n) + \sigma^2$, then

$$ABC'(J_n) \leq (5 + 14\tau_1 + 7\tau_2)nR_n(J_n) + \frac{n + r_{J_n}}{n - r_{J_n}}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta)). \qquad (8.6)$$

Thus, for any $\delta > 0$, when the sample size is large enough, we have that with probability no less than $1-5\delta$,

$$
\begin{aligned}
nR_n(\hat{J}') &\le \frac{ABC'(\hat{J}') + \frac{n+r_{\hat{J}'}}{n-r_{\hat{J}'}}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{1-(6\tau_1+3\tau_2)} \\
&\le \frac{ABC'(J_n) + \frac{n+r_{\hat{J}'}}{n-r_{\hat{J}'}}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{1-(6\tau_1+3\tau_2)} \\
&\le \frac{(5+14\tau_1+7\tau_2)nR_n(J_n) + \frac{n+r_{J_n}}{n-r_{J_n}}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta)) + \frac{n+r_{\hat{J}'}}{n-r_{\hat{J}'}}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{1-(6\tau_1+3\tau_2)}.
\end{aligned}
$$

Thus, with probability at least $1-5\delta$,

$$
\begin{aligned}
R_n(\hat{J}')/R_n(J_n) &\le \frac{5+14\tau_1+7\tau_2}{1-(6\tau_1+3\tau_2)} + \frac{\left(\frac{n+r_{J_n}}{n-r_{J_n}} + \frac{n+r_{\hat{J}'}}{n-r_{\hat{J}'}}\right)(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{(1-(6\tau_1+3\tau_2))nR_n(J_n)} \\
&\le \frac{5+14\tau_1+7\tau_2}{1-(6\tau_1+3\tau_2)} + \frac{\left(\frac{n+r_{J_n}}{n-r_{J_n}} + \frac{n+r_{\hat{J}'}}{n-r_{\hat{J}'}}\right)(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{(1-(6\tau_1+3\tau_2))\sigma^2} \\
&\le \frac{5+14\tau_1+7\tau_2}{1-(6\tau_1+3\tau_2)} + \frac{6(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{(1-(6\tau_1+3\tau_2))\sigma^2}.
\end{aligned}
$$

Let

$$
\widetilde{W} = b_n^{-1}\left(\frac{R_n(\hat{J}')}{R_n(J_n)} - \frac{5+14\tau_1+7\tau_2}{1-(6\tau_1+3\tau_2)}\right) \text{ and } b_n = \frac{6(2\tau_1+\tau_2)\lambda}{(1-(6\tau_1+3\tau_2))\sigma^2}.
$$

Then $P\left(\widetilde{W} \ge -\log_2 \delta\right) \le 5\delta$ for $0<\delta<1$. Since $E(\widetilde{W}^+) = \int_0^\infty P(\widetilde{W} \ge t)dt \le 5\int_0^\infty 2^{-t}dt = 5/\ln 2$ and $R_n(J_n) \le (\overline{\sigma}^2/\sigma^2)\inf_{J\in\Gamma} R_n(f_0;J)$ where $R_n(f_0;J) = \|\overline{f}_J - f_0^n\|_n^2 + r_J\sigma^2/n + \lambda\sigma^2 C_J/n$, then we have

$$
\begin{aligned}
\frac{E\left(R_n(\hat{J}')\right)}{\inf_{J\in\Gamma} R_n(f_0;J)} &= \frac{E\left(R_n(\hat{J}')\right)}{R_n(J_n)} \cdot \frac{R_n(J_n)}{\inf_{J\in\Gamma} R_n(f_0;J)} \\
&\le \left(\frac{5+14\tau_1+7\tau_2}{1-(6\tau_1+3\tau_2)} + \frac{30(2\tau_1+\tau_2)\lambda}{(\ln 2)(1-(6\tau_1+3\tau_2))\sigma^2}\right)\cdot\left(\frac{\overline{\sigma}^2}{\sigma^2}\right).
\end{aligned}
$$

So $E(ASE(\hat{f}_{\hat{J}'})) \le B\inf_{J\in\Gamma} R_n(f_0;J)$, where the constant B depends on τ_1, τ_2, $\overline{\sigma}$, and σ. Minimizing $h(\tau_1,\tau_2)$ over $\tau_1>0$ and $\tau_2>0$ in the region $6\tau_1+3\tau_2<1$, one finds a minimum value less than $40\log 2$. Thus, the results of the theorem hold when $\lambda \ge 40\log 2$.

□

Proposition 7 may not provide optimal risk rate when r_{M_n} is small, or when r_{M_n} is larger than $n/2$ (in which case the risk bound on $E(ASE(\hat{J}'))$ can be arbitrarily large because the approximation errors can be arbitrarily large when the models are restricted to be of size $n/2$ or smaller). The issue can be resolved by considering the full model J_{M_n} and the full projection model $\bar{J}$ in the candidate model list, as described before Theorem 10 .

Proof of Theorem 10 :

Proof. Observe that for the full projection model $\bar{J}$, with the chosen $C_{\bar{J}}$, we have that

$$(1-(6\tau_1+3\tau_2))nR_n(\bar{J}) \le ABC'(\bar{J}) \le \xi nR_n(\bar{J}) = \xi\left(n\sigma^2+\lambda\overline{\sigma}^2 C_{\bar{J}}\right)$$

for some constant $\xi > 0$ that depends only on λ, $\overline{\sigma}^2$ and σ^2. From the remark after Proposition 7, we have the following risk bounds for the three situations. Below B and B' are constants depending only on λ, $\overline{\sigma}^2$, and σ^2.

1. When $M_n \le n/2$, we have the general risk bound

$$\begin{aligned}
& E(ASE(\hat{f}_{\hat{J}'})) \\
\le\ & B'\Bigg(\inf_{J_m:1\le m<M_n}\left(\left\|\bar{f}_{J_m}-f_0^n\right\|_n^2+\frac{\sigma^2 r_{J_m}}{n}+\frac{\lambda\sigma^2 C_{J_m}}{n}\right)\wedge\left(\left\|\bar{f}_{J_{M_n}}-f_0^n\right\|_n^2+\frac{\sigma^2 r_{M_n}}{n}\right) \\
& \wedge R_n(\bar{J})\wedge R_n(J_0)\Bigg) \\
\le\ & B'\Bigg(\|\bar{f}_{J_{M_n}}-f_0^n\|_n^2+\inf_{J_m:1\le m<M_n}\Bigg(\|\bar{f}_{J_m}-\bar{f}_{J_{M_n}}\|_n^2+\frac{\sigma^2 r_{J_m}}{n}+\frac{\sigma^2\log(M_n-1)}{n} \\
& +\frac{\sigma^2\log\binom{M_n}{m}}{n}\Bigg)\wedge\frac{\sigma^2 r_{M_n}}{n}\Bigg)\wedge B'\left(\left(\|\bar{f}_{J_0}-f_0^n\|_n^2+\frac{\sigma^2}{n}\right)\wedge\sigma^2\right).
\end{aligned}$$

For $f_0^n \in \mathcal{F}_q(t_n;M_n)$, from above, by an argument similar to that in Theorem 7, for any $1\le m<M_n$, there exists a subset J_m and $f_{\theta^m}\in\mathcal{F}_{J_m}$ such that

$$\begin{aligned}
E(ASE(\hat{f}_{\hat{J}'})) \ \le\ & B'\left(\left(t_n^2 m^{1-2/q}+\frac{\sigma^2 r_{J_m}}{n}+\frac{\sigma^2 m\left(1+\log\frac{M_n}{m}\right)}{n}\right)\wedge\frac{\sigma^2 r_{M_n}}{n}\right) \\
& \wedge B'\left(\left(t_n^2+\frac{\sigma^2}{n}\right)\wedge\sigma^2\right). \qquad (8.7)
\end{aligned}$$

When $m_*=m^*=M_n$, the full model J_{M_n} leads to an upper bound of order $\frac{\sigma^2 r_{M_n}}{n}$. When $1<m_*<M_n$, we get the desired upper bound by taking the smaller value of the index of resolvability at J_{m_*} and J_{M_n}. When $m_*=1$, the smaller value of the index of resolvability at J_0 and J_{M_n} results in the given upper bound.

The arguments for cases $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ and $\mathcal{F} = \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)$ are similar to those of Theorem 7.

2. When $M_n > n/2$ and $r_{M_n} \geq n/2$, evaluating the index of resolvability gives

$$\begin{aligned}
& E(ASE(\hat{f}_{\hat{J}'})) \\
\leq\ & B\left(\inf_{J_m: 1 \leq m \leq n/2}\left(\left\|\bar{f}_{J_m} - f_0^n\right\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\lambda\sigma^2 C_{J_m}}{n}\right) \wedge R_n(\bar{J}) \wedge R_n(J_0)\right) \\
\leq\ & B' \inf_{J_m: 1 \leq m \leq n/2}\left(\|\bar{f}_{J_m} - f_0^n\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log\lfloor n/2 \rfloor}{n} + \frac{\sigma^2 \log\binom{M_n}{m}}{n}\right) \\
& \wedge B'\left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n}\right) \wedge \sigma^2\right).
\end{aligned}$$

In this case, for the full model, clearly, we have $\|\bar{f}_{J_{M_n}} - f_0^n\|_n^2 + \frac{\sigma^2 r_{M_n}}{n} \geq \frac{1}{2}\sigma^2$, which cannot be better than the model $\bar{J}$ up to a constant factor. We next show that adding the models with size $n/2 < m < M_n$ does not help either in terms of the rate in the risk bound. If $r_{J_m} \geq r_{M_n}/2$, then obviously $\|\bar{f}_{J_m} - f_0^n\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log\lfloor n/2 \rfloor}{n} + \frac{\sigma^2 \log\binom{M_n}{m}}{n} \geq \frac{1}{4}\sigma^2$. For $r_{J_m} < r_{M_n}/2$, if $n/2 < m \leq M_n/2$, then there exists a smaller model with size $\widetilde{m} \leq n/2$ that has the same approximation error and rank, but smaller complexity $C_{J_{\widetilde{m}}}$ (i.e., $C_{J_{\widetilde{m}}} \leq C_{J_m}$), where $C_{J_m} = \log(n \wedge M_n) + \log\binom{M_n}{m}$ when $m > n/2$. If $m > M_n/2$ (and $r_{J_m} < r_{M_n}/2$), then due to the monotonicity of the function $\binom{M_n}{m}$ in $m \geq M_n/2$, since there must be more than $r_{M_n}/2$ terms left out in the model, we must have $\log\binom{M_n}{m} \geq \log\binom{M_n}{M_n - \lfloor r_{M_n}/2 \rfloor} \geq \lfloor r_{M_n}/2 \rfloor \log\frac{M_n}{\lfloor r_{M_n}/2 \rfloor}$, which is at least of order n under the condition $r_{M_n} \geq n/2$. Putting the above facts together, we conclude that adding the models with size $n/2 < m \leq M_n$ does not affect the validity of the risk bound given in part (ii) of Theorem 10 (note that $\log\lfloor n/2 \rfloor$ is of the same order as $\log(M_n \wedge n)$ in our case). Then, the general risk upper bound becomes (with B' enlarged by an absolute constant factor)

$$\begin{aligned}
& B' \inf_{J_m: 1 \le m < M_n} \left(\|\bar{f}_{J_m} - f_0^n\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log(M_n \wedge n)}{n} + \frac{\sigma^2 \log \binom{M_n}{m}}{n} \right) \\
& \wedge B' \left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n} \right) \wedge \sigma^2 \right) \wedge B' \left(\|\bar{f}_{J_{M_n}} - f_0^n\|_n^2 + \frac{\sigma^2 r_{M_n}}{n} \right) \\
\le \; & B' \left(\|\bar{f}_{J_{M_n}} - f_0^n\|_n^2 + \inf_{J_m: 1 \le m < M_n} \left(\|\bar{f}_{J_m} - \bar{f}_{J_{M_n}}\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log(M_n \wedge n)}{n} \right.\right. \\
& \left.\left. + \frac{\sigma^2 \log \binom{M_n}{m}}{n} \right) \wedge \frac{\sigma^2 r_{M_n}}{n} \right) \wedge B' \left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n} \right) \wedge \sigma^2 \right).
\end{aligned}$$

For $f_0^n \in \mathcal{F}_q(t_n; M_n)$ and any $1 \le m < M_n$, there exists a subset J_m and $f_{\theta^m} \in \mathcal{F}_{J_m}$ such that the inequality (8.7) holds. When $m_* = m^* = M_n \wedge n$, the full projection model $\bar{J}$ leads to an upper bound of order σ^2. When $1 < m_* < M_n \wedge n$, we get the desired upper bounds by choosing J_{m_*} and $\bar{J}$ to evaluate the index of resolvability. When $m_* = 1$, models J_0 and $\bar{J}$ result in the desired upper bound.

3. When $M_n > n/2$ and $r_{M_n} < n/2$, the full model is already included, and, similarly as above, the models with $n/2 < m < M_n$ can be included in the minimization set of the general risk bound. Indeed, if $r_{M_n} = 1$, the statement is trivial. If $r_{J_m} \ge r_{M_n}/2$, then $\|\bar{f}_{J_m} - f_0^n\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log\lfloor n/2 \rfloor}{n} + \frac{\sigma^2 \log \binom{M_n}{m}}{n} \ge \left\|\bar{f}_{J_{M_n}} - f_0^n\right\|_n^2 + \frac{\sigma^2 J_{M_n}}{2n}$, which means that the model cannot beat the full model up to a constant factor. For $r_{J_m} < r_{M_n}/2$, if $m > M_n/2$, then we again have $\log \binom{M_n}{m} \ge \log \binom{M_n}{M_n - \lfloor r_{M_n}/2 \rfloor} \ge \lfloor r_{M_n}/2 \rfloor \log \frac{M_n}{\lfloor r_{M_n}/2 \rfloor}$. Thus there exists a model in Γ_n' with the same rank of $r_{J_m} \le n/2$ and approximation error, and its complexity is at most at the same order as J_m. Then with the same arguments for the case of $r_{M_n} \ge n/2$, we again conclude that adding the models with size $n/2 < m \le M_n$ does not affect the validity of the risk bound

given in part (ii) of Theorem 10. Thus, the general risk bound is

$$\begin{aligned}
& E(ASE(\hat{f}_{\hat{j}'})) \\
\leq\ & B\left\{\left(\|\bar{f}_{J_{M_n}} - f_0^n\|_n^2 + \inf_{J_m: 1\leq m\leq n/2}\left(\|\bar{f}_{J_m} - \bar{f}_{J_{M_n}}\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{2\lambda\sigma^2 C_{J_m}}{n}\right)\wedge\right.\right. \\
& \left.\left.\frac{\sigma^2 r_{M_n}}{n}\right)\right\} \wedge B\left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n}\right)\wedge\sigma^2\right) \\
\leq\ & B'\left(\|\bar{f}_{J_{M_n}} - f_0^n\|_n^2 + \inf_{J_m: 1\leq m< M_n}\left(\|\bar{f}_{J_m} - \bar{f}_{J_{M_n}}\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2\log(M_n\wedge n)}{n}\right.\right. \\
& \left.\left. + \frac{\sigma^2\log\binom{M_n}{m}}{n}\right)\wedge\frac{\sigma^2 r_{M_n}}{n}\right)\wedge B'\left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n}\right)\wedge\sigma^2\right).
\end{aligned}$$

For $f_0^n \in \mathcal{F}_q(t_n; M_n)$ and any $1 \leq m < M_n$, there exists a subset J_m and $f_{\theta^m} \in \mathcal{F}_{J_m}$ such that the inequality (8.7) holds. When $m_* = m^* = M_n \wedge n$, the full model J_{M_n} leads to an upper bound of order $\frac{\sigma^2 r_{M_n}}{n}$. When $1 < m_* < M_n \wedge n$, we get the desired upper bounds by choosing J_{m_*} and J_{M_n} when evaluating the index of resolvability. When $m_* = 1$, taking models J_0 and J_{M_n} results in the desired upper bound.

□

Proof of Theorem 11:

Proof. The proof is similar to that of Theorem 10 except that we use the oracle inequality (4.7) in [7] instead of that in Proposition 7 (and there is no need to consider the different scenarios). Note that if $M_n \leq (n-7)\wedge\varsigma n$, then $m \vee \log\binom{M_n}{m} < \varsigma n$ for all $1 \leq m \leq M_n$. Thus all subset models are allowed by the BGH criterion. When M_n is larger, however, the conditions required in Corollary 1 of [7] may invalidate the choice of m_* or k_n when it is too large, hence the upper bound assumption on m_* and k_n. We skip the details of the proof.

□

Acknowledgments

We thank Yannick Baraud for helpful discussions and for pointing out the work of Audibert and Catoni (2010) that helped us to remove a logarithmic factor in some of our previous aggregation risk bounds and for his suggestion of handling the case of fully unknown error variance for fixed design. Sandra Paterlini conducted part of this research while visiting the School of Mathematics, University of Minnesota.

References

[1] Akaike, H. (1970) Statistical prediction identification. *Ann. Inst. Statist. Math.* **22**, 203-217.

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

[3] Audibert, J.-Y. (2009) Fast learning rates in statistical inference through aggregation. *Annals of Statistics* **37** 1591-1646.

[4] Audibert, J.-Y and Catoni, O. (2010) Risk bounds in linear regression through PAC-Bayesian truncation. Available at arXiv:1010.0072.

[5] Baraud, Y. (2000) Model selection for regression on a fixed design. *Probability Theory Related Fields* **117** 467-493.

[6] Baraud, Y. (2002) Model selection for regression on a random design. *ESAIM Probab.Statist.* **6** 127-146.

[7] Baraud, Y., Giraud, C. and Huet, S. (2009) Gaussian model selection with an unknown variance. *Ann. Statist.* **37** 630-672.

[8] Barron, A. R. (1993) Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39** 930-945.

[9] Barron, A .R. (1994) Approximation and estimation bounds for artificial neural networks. *Machine Learning* **14** 115-133.

[10] Barron, A. R. and Cover, T. M. (1991) Minimum complexity density estimation. *IEEE Transactions on Information Theory* **37** 1034-1054.

[11] Barron, A. R., Birgé, L. and Massart, P. (1999) Risk bounds for model selection via penalization. *Probability Theory and Related Fields* **113** 301-413.

[12] Barron, A. R., Cohen, A., Dahmen, W., and Devore, R. (2008). Approximation and learning by greedy algorithms. *Annals of Statistics* **36**, 64-94.

[13] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37** 1705-1732.

[14] Birgé, L. (1986) On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Related Fields* **97** 113-150.

[15] Birgé, L. (2004) Model selection for Gaussian regression with random design. *Bernoulli* **10** 1039-1051.

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

[17] Birgé, L. and Massart, P. (1998) Minimum Contrast Estimators on Sieves: Exponential Bounds and Rates of Convergence. *Bernoulli* **4** 329-375.

[18] Birgé, L. and Massart, P. (2001) Gaussian model selection. *Journal of European Math. Society* **3** 203-268.

[19] Bunea, F., Tsybakov, A.B. and Wegkamp, M. H. (2007) Aggregation for Gaussian regression. *Annals of Statistics* **35** 1674-1697.

[20] Bunea, F. and Nobel, A. (2008) Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Transactions on Information Theory* **54** 1725 - 1735.

[21] Candes, E. and Tao, T. (2007) The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* **35** 2313-2351.

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

[23] Catoni, O. (1999) 'Universal' aggregation rules with exact bias bounds. Preprint 510, Laboratoire de Probabilitès et Modèles Aléatoires, Univ. Paris 6 and Paris 7. Available at

http://www.proba.jussieu.fr/mathdoc/preprints/index.html#1999.

[24] Catoni, O. (2004) *Statistical Learning Theory and Stochastic Optimization. Ecole d'Eté de Probabilités de Saint-Flour XXXI–2001, Lectures Notes in Mathematics* **1851**, Springer, New York.

[25] Chen, D., Donoho, D. L. and Saunders, M. A. (1998) Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing* **20** 33-61.

[26] Dalalyan, A. S. and Tsybakov, A. B. (2007) Aggregation by exponential weighting and sharp oracle inequalities. *Lecture Notes in Computer Science* **4539** 97-111.

[27] Dalalyan, A. S. and Tsybakov, A. B. (2009) Sparse regression learning by aggregation and Langevin Monte Carlo. Available at ArXiv:0903.1223.

[28] Donoho, L. (1993) Unconditional Bases are optimal bases for data compression and for statistical estimation. *Appl. Computational Harmonic analysis* **1** 100-115.

[29] Donoho, D. L. (2006) Compressed sensing. *IEEE Transactions on Information Theory* **52** 1289-1306.

[30] Donoho, D. L. and Johnstone, I. M. (1994) Minimax risk over ℓ_p-balls for ℓ_q-error. *Prob. Theory and Related Fields* **99** 277-303.

[31] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996) Density estimation by wavelet thresholding. *Annals of Statistics* **24** 508-539.

[32] Edmunds, D. E. and Triebel H. (1998) Function spaces, entropy numbers, and differential operators. *Cambridge Tracts in Mathematics* **120**. Cambridge University Press.

[33] Gasso, G., Rakotomamonjy, A. and Canu, S. (2009) Recovering sparse signals with non-convex penalties and DC programming. *IEEE Trans. Signal Processing* **57** 4686-4698.

[34] van de Geer, S.A. and Bühlmann, P. (2009) On the conditions used to prove oracle results for the Lasso. *Electron. J. Statist.* **3** 1360-1392.

[35] Giraud, C. (2008) Mixing least-squares estimators when the variance is unknown. *Bernoulli* **14** 1089-1107.

[36] Goldenshluger, A. (2009) A universal procedure for aggregating estimators. *Annals of Statistics* **37** 542-568.

[37] Greenshtein, E. and Ritov, Y., (2004) Persistency in high dimensional linear predictor-selection and the virtue of overparametrization. *Bernoulli* **10** 971-988.

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

[39] Györfi, L. and Ottucsák, Gy. (2007) Sequential prediction of unbounded stationary time series. *IEEE Transactions on Information Theory* **53** 1866-1872.

[40] Huang, C., Cheang, G. H. L. and Barron, A. R. (2008) Risk of penalized least squares, greedy selection and L_1-penalization for flexible function libraries. Manuscript.

[41] Ing, C-K. (2010) On the optimal convergence rate of orthogonal greedy algorithms in high-dimensional sparse regression models. Presented at *The Eighth ICSA International Conference: Frontiers of Interdisciplinary and Methodological Statistical Research*, Guangzhou, China, December 19-22, 2010.

[42] Juditsky, A. and Nemirovski, A. (2000) Functional aggregation for nonparametric estimation. *Annals of Statistics* **28** 681-719.

[43] Koltchinskii, V. (2009) The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799-828.

[44] Koltchinskii, V. (2009) Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.* **45** 7-57.

[45] Kühn, T. (2001) A lower estimate for entropy numbers. *Journal of Approximation Theory* **110** 120-124.

[46] Leung, G. and Barron, A. R. (2006) Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* **52** 3396-3410.

[47] Lecué, G. and Mendelson S. (2010) Optimality of the aggregate with exponential weights for low temperatures. Manuscript.

[48] Li, K.C. (1987) Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: discrete index set. *Annals of Statistics* **15** 958-975.

[49] Lounici, K. (2007) Generalized mirror averaging and D-convex aggregation. *Mathematical methods of statistics* **16** 246-259.

[50] Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the lasso. *Annals of Statistics* **34** 1436-1462.

[51] Meinshausen, N. and Yu, B. (2009) Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* **37** 246-270.

[52] Nemirovski, A. (2000) *Topics in Non-parametric Statistics.* In: P. Bernard, editor, *Ecole d'Eté de Probabilités de Saint-Flour 1998* volume XXVIII of *Lecture Notes in Mathematics* **1738** Springer, New York.

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

[54] Rigollet, P. and Tsybakov, A. B. (2010) Exponential screening and optimal rates of sparse estimation *Annals of Statistics* **39** 731-771.

[55] Shen, X. and Wong, W.H. (1994) Convergence rates of sieve estimates. *Annals of Statistics* **22**, 580-615.

[56] Shibata, R. (1981) An optimal selection of regression variables. *Biometrika* **68** 45-54.

[57] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B* **58** 267-288.

[58] Tsybakov, A.B. (2003) Optimal rates of aggregation. In: *Proceedings of 16th Annual Conference on Learning Theory (COLT)* and *7th Annual Workshop on Kernel Machines.* Lecture Notes in Artificial Intelligence, **2777** 303-313. Springer-Verlag, Heidelberg.

[59] van de Geer, S.A. (1995) The method of sieves and minimum contrast estimators. *Math. Methods Statist.* **4**, 20-38.

[60] van de Geer, S.A. (2008) High-dimensional generalized linear models and the Lasso. *Annals of Statistics* **36** 614-645.

[61] Verzelen, N. (2010) Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. Arxiv preprint arXiv:1008.0526.

[62] Wainwright, M.J. (2009) Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Information Theory* **55**

2183-2202.

[63] Wegkamp, M. (2003) Model selection in nonparametric regression. *Annals of Statistics* **31** 252-273.

[64] Yang, Y. (1996) "Minimax optimal density estimation", Ph.D. Dissertation, Department of Statistics, Yale University, May 1996.

[65] Yang, Y. (1999) Model selection for nonparametric regression. *Statistica Sinica*, **9** 475-499.

[66] Yang, Y. (2000) Mixing strategies for density estimation. *Annals of Statistics*, **28** 75-87.

[67] Yang, Y. (2000) Combining different procedures for adaptive regression. *Journal of Multivariate Analysis* **74** 135-161.

[68] Yang, Y. (2001) Adaptive regression by mixing. *Journal of American Statistical Association* **96** 574-588.

[69] Yang, Y. (2004) Aggregating regression procedures to improve performance. *Bernoulli* **10** 25-47. An older version of the paper is Pre-print #1999-17 of Department of Statistics at Iowa State University.

[70] Yang, Y. (2008) Localized model selection for regression. *Econometric Theory* **24**, 472-492.

[71] Yang, Y. and Barron, A.R. (1998) An asymptotic property of model selection criteria. *IEEE Trans. Inform. Theory* **44** 95-116.

[72] Yang, Y. and Barron, A.R. (1999) Information theoretic determination of minimax rates of convergence. *Annals of Statistics*, **27** 1564-1599.

[73] Ye, F. and Zhang, C.H. (2010) Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research* **11** 3519-2540.

[74] Yuan, Z. and Yang, Y. (2005) Combining linear regression models: When and how? *Journal of the American Statistical Association* **100** 1202-1214.

[75] Zhang, C.H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** 894-942.

[76] Zhang, C.H. and Huang, J. (2008) The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics* **36** 1567-1594.

[77] Zhang, T. (2009) Some sharp performance bounds for least squares regression with L_1 regu-

larization. *Annals of Statistics* **37** 2109-2144.

[78] ZHANG, T. (2010) Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research* **11** 1081-1107.

"Materiali di Discussione" LATER PUBLISHED ELSEWHERE

N. 546 - M. Murat and B. Pistoresi, *Emigrants and immigrants networks in FDI*, Applied Economics letters, April 2008, **http://www.informaworld.com/ /content~content=a789737803~db=all~order=author** (electronic publication), **WP No. 546 (December 2006).**

N. 545 - M.Brunetti and C. Torricelli, *The Population Ageing in Italy:Facts and Impact on Household Portfolios*, in M. Balling & E. Gnan & F. Lierman (eds.), Money, Finance and Demography: The Consequences of Ageing, Vienna, Suerf (2007), **WP No. 545 (November 2006).**

N. 532 – M: Montanari, Between European Integration and Regional Autonomy: The Case of Italy from an Economic Perspective, Constitutional Political Economy, Vol. 17, 4, pp. 277-301 (2006), **WP No. 532 (March 2006).**

N. 529 - M. Montanari, *Knocking on the EU's door: the Political Economy of EU-Ukraine Relations*, Journal of Contemporary European Research, Vol.3,1, pp.64-78 (2007), **WP No. 529 (February 2006).**

N. 518 - M.Brunetti and C. Torricelli, *Economic Activity and Recession Probabilities: information content and predictive power of the term spread in Italy*, Applied Economics (2009), **WP No. 518 (December 2005).**

N. 517 - M. Murat and S. Paba (2006), *I distretti industriali tra immigrazioni e internazionalizzazione produttiva*, in B. Quintieri (ed.) I distretti italiani dal locale al globale, Rubbettino (2006), **WP No. 517 (December 2005).**

N. 491 - V. Moriggia, S. Muzzioli and C. Torricelli, *On the no arbitrage condition in option implied trees*, European Journal of Operational Research (2009), **WP No. 491 (May 2005).**

N. 482 - G. Di Lorenzo and G. Marotta, *A less effective monetary transmission in the wake of EMU? Evidence from lending rates passthrough***,** ICFAI Journal of Monetary Economics, Vol. 4, 2, pp. 6-31 (2006), **WP No. 482 (February 2005).**

N. 472 - M.Brunetti and C. Torricelli, *The internal and cross market efficiency in index option markets: an investigation of the Italian market*, Applied Financial Economics, Vol. 17, 1, pp. 25-33 (2007), **WP No. 472 (November 2004).**

N. 466 - G. Marotta, *La finanza del settore non profit tra ritardi nei pagamenti e Basilea 2*, Banca Impresa Società , Vol. XXIV, 1, pp. 35-51 (2005), **WP No. 466 (September 2004).**

N. 453 - Pederzoli and C. Torricelli, *Capital requirements and Business Cycle Regimes: Forward-looking modelling of Default Probabilities* , Journal of Banking and Finance, Vl. 29, 12, pp. 3121-3140 (2005), **WP No. 453 (February 2004).**

N. 448 - V. Moriggia, S. Muzzioli, C. Torricelli, *Call and put implied volatilities and the derivation of option implied trees*, Frontiers In Finance and Economics, vol.4, 1, pp. 35-64 (2007), **WP No. 448 (November 2003).**

N. 436 - M.Brunetti and C. Torricelli, *Put-Call Parity and cross-market efficiency in the Index Options Markets: evidence from the Italian market*, International Review of Financial Analysis, Vl.14, 5, pp. 508-532 (2005), **WP No. 436 (July 2003).**

N. 429 - G. Marotta, *When do trade credit discounts matter? Evidence from Italian Firm-Level Data*, Applied Economics, Vol. 37, 4, pp. 403-416 (2005), **WP No. 429 (February 2003).**

N. 426 - A. Rinaldi and M. Vasta, *The Structure of Italian Capitalism, 1952-1972: New Evidence Using the Interlocking Directorates Technique*, Financial History Review, vol, 12, 2, pp. 173-198 (2005), **WP No. 426 (January 2003).**

N. 417 - A. Rinaldi, *The Emilian Model Revisited: Twenty Years After*, Business History, vol. 47, 2, pp. 244-226 (2005), **WP No. 417 (September 2002).**

N. 375 - G. Marotta, *La direttiva comunitaria contro i ritardi nei pagamenti tra imprese. Alcune riflessioni sul caso italiano*, Banca, Impresa, Società, Vol. XX, 3, pp. 451-71 (2001), **WP No. 375 (September 2001).**

N. 303 - G. Marotta and M. Mazzoli, *Fattori di mutamento nella domanda di prestiti ed effetti sulla trasmissione della politica monetaria*, in P. ALESSANDRINI (ed.) Il sistema finanziario italiano tra globalizzazione e localismo, Bologna, Il Mulino, pp. 223-260 (2001), **WP No. 303 (April 2000)**

N. 131 - G. Marotta, *Does trade credit redistribution thwart monetary policy? Evidence from Italy*, Applied Economics, Vol. 29, December, pp. 1619-29 (1997), **WP No. 131 (1996).**

N. 121 - G. Marotta, *Il credito commerciale in Italia: una nota su alcuni aspetti strutturali e sulle implicazioni di politica monetaria*, L'Industria, Vol. XVIII, 1, pp. 193-210 (1997), **WP No. 121 (1995)**

N. 105 - G. Marotta, *Credito commerciale e "lending view"*, Giornale degli Economisti e Annali di Economia, Vol. LIV, 1-3, gennaio-marzo, pp. 79-102; anche in G. Vaciago (a cura di) *Moneta e finanza*, Bologna, Il Mulino (1995), **WP No. 105 (1994).**

Adaptive Minimax Estimation over Sparse ℓ_q-Hulls

Zhan Wang[1], Sandra Paterlini [*2], Fuchang Gao[3], and Yuhong Yang [†1]

[1] *School of Statistics, University of Minnesota, USA,*

e-mail: wangx607@stat.umn.edu*;* yyang@stat.umn.edu

[2] *Department of Economics,RECent & CEFIN University of Modena and Reggio E., Italy,*

e-mail: sandra.paterlini@unimore.it

[3] *Department of Mathematics, University of Idaho, USA,*

e-mail: fuchang@uidaho.edu

Abstract: Given a dictionary of M_n initial estimates of the unknown true regression function, we aim to construct linearly aggregated estimators that target the best performance among all the linear combinations under a sparse q-norm ($0 \leq q \leq 1$) constraint on the linear coefficients. Besides identifying the optimal rates of aggregation for these ℓ_q-aggregation problems, our multi-directional (or universal) aggregation strategies by model mixing or model selection achieve the optimal rates simultaneously over the full range of $0 \leq q \leq 1$ for general M_n and upper bound t_n of the q-norm. Both random and fixed designs, with known or unknown error variance, are handled, and the ℓ_q-aggregations examined in this work cover major types of aggregation problems previously studied in the literature. Consequences on minimax-rate adaptive regression under ℓ_q-constrained true coefficients ($0 \leq q \leq 1$) are also provided.

Our results show that the minimax rate of ℓ_q-aggregation ($0 \leq q \leq 1$) is basically determined by an effective model size, which is a sparsity index that depends on q, t_n, M_n, and the sample size n in an easily interpretable way based on a classical model selection theory that deals with a large number of models. In addition, in the fixed design case, the model selection approach is seen to yield optimal rates of convergence not only in expectation but also with exponential decay of deviation probability. In contrast, the model mixing approach can have leading constant one in front of the target risk in the oracle inequality while not offering optimality in deviation probability.

Keywords and phrases: minimax risk, adaptive estimation, sparse ℓ_q-constraint, linear combining, aggregation, model mixing, model selection.

[*] Sandra Paterlini's research was partially supported by Fondazione Cassa di Risparmio di Modena.

[†] Yuhong Yang's research was partially supported by NSF grant DMS-1106576.

1. Introduction

The idea of sharing strengths of different estimation procedures by combining them instead of choosing a single one has led to fruitful and exciting research results in statistics and machine learning. In statistics, the theoretical advances have centered on optimal risk bounds that require almost no assumption on the behaviors of the individual estimators to be integrated (see, e.g., [64, 67, 22, 24, 42, 52, 69, 58] for early representative work). 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 aims 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 aims at improving over the performance of all the candidate procedures in certain ways. Whatever the goal is, for the purpose of estimating a target function (e.g., the true 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), 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 the proportion of the initial data size 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, although combining procedures for adaptation and selecting the best one share the same goal of achieving the best performance offered by the candidate procedures, the former usually wins when model selection uncertainty is high (see, e.g., [74]). Theoretically, one only needs to pay a relatively small price for aggregation for adaptation ([66, 24, 58]). 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 [42, 52, 69, 58]). 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, [58]), aggregation to achieve the best performance of a linear combination of no more than a given number of initial estimates ([19]) and also under an additional constraint on the ℓ_1-norm of these coefficients ([49]). Interestingly, combining for adaptation has a fundamental role for combining for improvement: it serves as an effective tool in constructing multi-directional (or universal) aggregation methods that simultaneously achieve the best performance in multiple specific directions of aggregation for improvement. This strategy was taken in section 3 of [69], where aggregations of subsets of estimates are then aggregated to be suitably aggressive and conservative in an adaptive way. Other uses of subset models for universal aggregation have been handled in [19, 54].

The goal of this paper 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 [58], of the best linear combination of the initial estimates under the constraint that the q-norm ($0 \leq q \leq 1$) 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 ($0 \leq q \leq 1$) and t_n ($t_n > 0$). From the work in [42, 69, 58, 4], it is known that by suitable aggregation methods, the squared L_2 risk is no larger than that of the best linear combination of the initial M_n estimates with the ℓ_1-norm of the coefficients bounded by 1 plus the order $(\log(M_n/\sqrt{n})/n)^{1/2}$ when $M_n \geq \sqrt{n}$ or M_n/n when $M_n < \sqrt{n}$. Two important features are evident here: 1) When M_n is large, its effect on the risk enlargement is only through a logarithmic fashion; 2) No assumption is needed at all on how the initial estimates are possibly correlated. The strong result comes from the ℓ_1-constraint on the coefficients.

Indeed, in the last decade of the twentieth century, the fact that ℓ_1-type of constraints induce sparsity has been used in different ways for statistical estimation to attain relatively fast rates of convergence as a means to overcome the curse of dimensionality. Among the most relevant ones, Barron [9] studied the use of ℓ_1-constraint in construction of estimators for fast convergence with neural nets; Tibshirani [57] introduced the Lasso; Chen, Donoho and Saunders [25] proposed the basis pursuit with over complete bases. Theoretical advantages have also been pointed out. Barron [8] showed that for estimating a high-dimensional function that has integrable Fourier transform or a neural net representation, accurate approximation error is achievable. Together with model selection over finite dimensional neural network models, relatively fast rates of convergence, e.g., $[(d \log n)/n]^{1/2}$, where d is the input dimension, are obtained (see, e.g., [9] with parameter discretization, section III.B in [71] and section 4.2 in [11] with continuous models). Donoho and Johnstone [30] identified how the ℓ_q-constraint ($q > 0$) on the mean vector affects estimation accuracy under ℓ_p loss ($p \geq 1$) in an illustrative Gaussian sequence model. For function estimation, Donoho [28] studied sparse estimation with unconditional orthonormal bases and related the essential rate of convergence to a sparsity index. In that direction, for a special case of function classes with unconditional basis defined basically in terms of bounded q-norm on the coefficients of the orthonormal expansion, the rate of convergence $(\log n/n)^{1-q/2}$ was given in [71] (section 5). The same rate also appeared in the earlier work of Donoho and Johnstone [30] in some asymptotic settings. Note that when $q = 1$, this is exactly the same rate of the risk regret for ℓ_1-aggregation when M_n is of order n^{κ} for $1/2 \leq \kappa < \infty$.

General model selection theories on function estimation intend to work with general and possibly complicatedly dependent terms. Considerable research has been built upon subset selection as a natural way to pursue sparse and flexible estimation. When exponentially many or more models are entertained, optimality theories that handle a small number of models (e.g., [56, 48]) are no longer suitable. General theories were then developed for estimators based on criteria that add an additional penalty to the AIC type criteria, where the additional penalty term prevents substantial overfitting that often occurs when working with exponentially many models by standard information criteria, such as AIC and BIC. A masterpiece of work with tremendous breadth and depth is Barron, Birgé and Massart [11], and some other general results in specific contexts of density estimation and regression with fixed or random design are in [71, 65, 18, 5, 6, 15].

These model selection theories are stated for nonparametric scenarios where none of the finite-dimensional approximating models is assumed to hold but they are used as suitable sieves to deliver good estimators when the size of the sieve is properly chosen (see, e.g., [55, 59, 17] for non-adaptive sieve theories). If one makes the assumption that a subset model of at most k_n terms holds (ℓ_0-constraint), then the general risk bounds mentioned in the previous paragraph immediately give the order $k_n \log(M_n/k_n)/n$ for the risk of estimating the target function under quadratic type losses.

Thus, the literature shows that both ℓ_0- and ℓ_1-constraints result in fast rates of convergence (provided that M_n is not too large and k_n is relatively small), with hard-sparsity directly coming from that only a small number of terms is involved in the true model under the ℓ_0-constraint, and soft-sparsity originating from the fact that there can only be a few large coefficients under the ℓ_1-constraint. In this work, with new approximation error bounds in $\ell_{q,t_n}^{M_n}$-hulls (defined in section 2.1) for $0 < q \leq 1$, from a theoretical standpoint, we will see that model selection or model combining with all subset models in fact simultaneously exploits the advantage of sparsity induced by ℓ_q-constraints for $0 \leq q \leq 1$ to the maximum extent possible.

Clearly, all subset selection is computationally infeasible when the number of terms M_n is large. To overcome this difficulty, an interesting research direction is based on greedy approximation, where terms are added one after another sequentially (see, e.g., [12]). Some general theoretical results are given in the recent work of [40], where a theory on function estimation via penalized squared error criteria is established and is applicable to several greedy algorithms. The associated risk bounds yield optimal rate of convergence for sparse estimation scenarios. For aggregation methods based on exponential weighting under fixed design, practical algorithms based on Monte Carlo methods have been given in [27, 54].

Considerable recent research has focused on ℓ_1-regularization, producing efficient algorithms and related theories. Interests are both on risk of regression estimation and on variable selection. Some estimation risk bounds are in [13, 37, 43, 44, 50, 51, 62, 60, 76, 75, 77, 73].

The ℓ_q-constraint, despite being non-convex for $0 < q < 1$, poses an easier optimization challenge than the ℓ_0-constraint, which is known to define a NP-hard optimization problem and be hardly tractable for large dimensions. Although a few studies have devoted to the algorithmic developments of the ℓ_q-constraint optimization problem, such as multi-stage convex relaxation algorithm ([78]) and the DC programming approach ([33]), little work has been done with respect to the theoretical analysis of the ℓ_q-constrained framework.

Sparse model estimation by imposing the ℓ_q-constraint 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 do not only promote sparsity but 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 ([29]).

Our ℓ_q-aggregation risk upper bounds require no assumptions on dependence of the initial estimates in the dictionary and the true regression function is arbitrary (except that it has a known sup-norm upper bound in the random design case). The results readily give minimax rate optimal estimators for a regression function that is representable as a linear combination of the predictors subject to ℓ_q-constraints on the linear coefficients.

Two recent and interesting results are closely related to our work, both under fixed design only. Raskutti, Wainwright and Yu [53] derived in-probability minimax rates of convergence for estimating the regression functions in $\ell_{q,t_n}^{M_n}$-hulls with minimal conditions for the full range of $0 \leq q \leq 1$. In addition, in an informative contrast, they have also handled the quite different problem of estimating the coefficients under necessarily much stronger conditions. Rigollet and Tsybakov [54] nicely showed that exponential mixing of least squares estimators by an algorithm of Leung and Barron [46] over subset models achieves universal aggregation of five different types of aggregation, which involve ℓ_0- and/or ℓ_1-constraints. Furthermore, they implemented a MCMC based algorithm with favorable numerical results. As will be seen, in this context of regression under fixed design, our theoretical results are broader with improvements in several different ways.

Our theoretical work emphasizes adaptive minimax estimation under the mean squared risk.

Building upon effective estimators and powerful risk bounds for model selection or aggregation for adaptation, we propose several aggregation/combining strategies and derive the corresponding oracle inequalities or index of resolvability bounds. Upper bounds for ℓ_q-aggregations and for linear regression with ℓ_q-constraints are then readily obtained by evaluating the index of resolvability for the specific situations, incorporating an approximation error result that follows from a new and precise metric entropy calculation on function classes of $\ell_{q,t_n}^{M_n}$-hulls. Minimax lower bounds that match the upper rates are also provided in this work. Whatever the relationships between the dictionary size M_n, the sample size n, and upper bounds on the ℓ_q-constraints, our estimators automatically take advantage of the best sparse ℓ_q-representation of the regression function in a proper sense.

By using classical model selection theory, we have a simple explanation of the minimax rates, by considering the effective model size m_*, which provides the best possible trade-off between the approximation error, the estimation error, and the additional price due to searching over not pre-ordered terms. The optimal rate of risk regret for ℓ_q-aggregation, under either hard or soft sparsity (or both together), can then be unifyingly expressed as

$$REG(m_*) = 1 \wedge \frac{m_* \left(1 + \log \frac{M_n}{m_*}\right)}{n},$$

which can then be interpreted as the log number of models of size m_* divided by the sample size ($\wedge 1$), as was previously suggested for the hard sparsity case $q = 0$ (e.g., Theorem 1 of [71], Theorems 1 and 4 of [65]).

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. In addition, we derive metric entropy and approximation error bounds for $\ell_{q,t_n}^{M_n}$-hulls that play an important role in determining the minimax rate of convergence and adaptation. In section 3, we derive optimal rates of ℓ_q-aggregation and show that our methods achieve multi-directional aggregation. We also briefly talk about ℓ_q-combination of procedures. In section 4, we derive the minimax rate for linear regression with ℓ_q-constrained coefficients also under random design. In section 5, we handle ℓ_q-regression/aggregation under fixed design with known or unknown variance. A discussion is then reported in section 6. In section 7, oracle inequalities are given for the random design. Proofs of the results are provided in section 8. We note that some upper and lower bounds in the last two sections may be of independent interest.

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 or model selection methods 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 $\mathcal{X} \subset \mathbb{R}^d$ with a probability distribution P_X. We assume the regression model

$$Y_i = f_0(\mathbf{X}_i) + \varepsilon_i, \qquad i = 1, \ldots n, \tag{2.1}$$

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 $F_n = \{f_1, f_2, \dots, f_{M_n}\}$ be a dictionary of M_n initial estimates of f_0. In this paper, unless stated otherwise, $\|f_j\| \leq 1$, $1 \leq j \leq M_n$. 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 $\mathbb{R}^{M_n}$. 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 $\theta = (\theta_1, \dots, \theta_{M_n})'$, 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. But 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

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

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 $C = 1$ under some conditions. We aim to obtain the optimal order of convergence for the risk regret term.

2.2. Two starting estimators

A key step of our strategy is the construction of candidate estimators using subsets of the initial estimates. The following two estimators (T- and AC-estimators) were chosen because of the relatively mild assumptions for them to work with respect to the squared L_2 risk. Under the data generating model (2.1) and i.i.d. observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, suppose we are given $(g_1, \ldots, g_m)$ terms for the regression problem.

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

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

(T-estimator) Birgé [15] 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 in [15]. 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*

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

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 $\mathcal{G} = \{g = \sum_{j=1}^m \vartheta_j g_j : \|\vartheta\|_1 \leq s\}$ for some $s > 0$. Audibert and Catoni proposed a sophisticated AC-estimator $\hat{f}_s^{(AC)}$ ([4], page 25). The following proposition is a direct result from Theorem 4.1 in [4] under the following conditions.

AC1. There exists a constant $H > 0$ such that $\sup_{g, g' \in \mathcal{G}, \mathbf{x} \in \mathcal{X}} |g(\mathbf{x}) - g'(\mathbf{x})| = H < \infty$.

AC2. There exists a constant $\sigma' > 0$ such that $\sup_{\mathbf{x} \in \mathcal{X}} E\left((Y - g^*(\mathbf{X}))^2 | \mathbf{X} = \mathbf{x}\right) \leq (\sigma')^2 < \infty$, where $g^* = \inf_{g \in \mathcal{G}} \|g - f_0\|^2$.

Proposition 2. *Suppose Assumptions* AC1 *and* AC2 *hold. For any* $s > 0$*, we can construct an*

AC-estimator $\hat{f}_s^{(AC)}$ such that

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

where c is a pure constant.

Note that under the assumption $\|f_0\|_\infty \leq L$, 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 [68]) and Catoni's algorithm (Catoni [24]), 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 $0 < \underline{\sigma} \leq \sigma \leq \overline{\sigma} < \infty$;

Y2. The error density function $h(x)$ has a finite fourth moment and 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 $S_0 \leq S \leq S_0^{-1}$,

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

We can construct an estimator $\hat{f}^Y$ which aggregates $\check{f}_1, \ldots, \check{f}_N$ by the ARM algorithm as described below.

Step 1. Split the data into two parts $Z^{(1)} = (\mathbf{X}_i, Y_i)_{i=1}^{n_1}$, $Z^{(2)} = (\mathbf{X}_i, Y_i)_{i=n_1+1}^{n}$. Take $n_1 = \lceil n/2 \rceil$.

Step 2. Estimate σ^2 for each $\check{f}_k$ using the data $Z^{(1)}$,

$$\hat{\sigma}_k^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} \left(Y_i - \check{f}_k(\mathbf{X}_i)\right)^2, \text{ for } 1 \le k \le N.$$

Clip the estimate $\hat{\sigma}_k^2$ into the range $[\underline{\sigma}^2, \overline{\sigma}^2]$ if needed.

Step 3. Evaluate predictions for each k. For $n_1 + 1 \le l \le n$, predict Y_l by $\check{f}_k(\mathbf{X}_l)$ and compute

$$E_{k,l} = \frac{\prod_{i=n_1+1}^{l} h\left((Y_i - \check{f}_k(\mathbf{X}_i))/\hat{\sigma}_k\right)}{\hat{\sigma}_k^{l-n_1}}.$$

Step 4. Compute the final estimate $\hat{f}^Y = \sum_{k=1}^N W_k \check{f}_k$ with

$$W_k = \frac{1}{n - n_1} \sum_{l=n_1+1}^{n} W_{k,l} \quad \text{and} \quad W_{k,l} = \frac{\pi_k E_{k,l}}{\sum_{j=1}^N \pi_j E_{j,l}},$$

where π_k are prior probabilities such that $\sum_{k=1}^N \pi_k = 1$.

Proposition 3. *(Yang [69], 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* $S_0 = \underline{\sigma}/\overline{\sigma}, R_0 = L$*).*

Remark 1. If σ is known or other estimators of σ are available, the data splitting is not required, and the ARM algorithm consists of only Steps 3 and 4.

(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. $\frac{E(\varepsilon_i^2 \exp(\alpha|\varepsilon_i|))}{E(\exp(\alpha|\varepsilon_i|))} \le V_\alpha$.

The estimator built using Catoni's algorithm is $\hat{f}^C = \sum_{k=1}^N W_k \check{f}_k$ with

$$W_k = \frac{1}{n} \sum_{l=1}^{n} \frac{\pi_k \left(\prod_{i=1}^{l} q_k(Y_i|\mathbf{X}_i)\right)}{\sum_{j=1}^N \pi_j \left(\prod_{i=1}^{l} q_j(Y_i|\mathbf{X}_i)\right)}, \text{ and } q_k(Y_i|\mathbf{X}_i) = \sqrt{\frac{\lambda_C}{2\pi}} \exp\left\{-\frac{\lambda_C}{2}(Y_i - \check{f}_k(\mathbf{X}_i))^2\right\},$$

where $\lambda_C = \min\{\frac{\alpha}{2L}, (U_\alpha(17L^2 + 3.4V_\alpha))^{-1}\}$, and π_k is the prior for $\check{f}_k$, $1 \le k \le N$, such that $\sum_{k=1}^N \pi_k = 1$.

Proposition 4. *(Catoni [24], 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* $\check{f}_1, \dots, \check{f}_N$ *by Catoni's algorithm has the risk*

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

Remark 2. In the risk bound above, the multiplying constant in front of $\|\check{f}_k - f_0\|^2$ is one, which can be important sometimes. Catoni [24] provided results under weaker assumptions than C1 and C2. In particular, ε_i and $\mathbf{X}_i$ do not have to be independent.

2.4. 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 [14, 72].

For each $1 \le m \le M_n$ and each subset $J_m \subset \{1, 2, \dots, M_n\}$ of size m, define $\mathcal{F}_{J_m} = \{\sum_{j\in J_m} \theta_j f_j : \theta \in \mathbb{R}^m\}$. 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 $\mathcal{F}$.

Theorem 1. *(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

$$\log N_\epsilon \le \begin{cases} c_q \left(t_n\epsilon^{-1}\right)^{\frac{2q}{2-q}} \log(1 + M_n^{\frac{1}{q}-\frac{1}{2}} t_n^{-1}\epsilon) & \text{if } \epsilon > t_n M_n^{\frac{1}{2}-\frac{1}{q}}, \\ c_q M_n \log(1 + M_n^{\frac{1}{2}-\frac{1}{q}} t_n\epsilon^{-1}) & \text{if } \epsilon \le t_n M_n^{\frac{1}{2}-\frac{1}{q}}. \end{cases} \tag{2.2}$$

(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*

$$\|f_{\theta^m} - f_0\|^2 - d^2(f_0; \mathcal{F}_q(t_n)) \le 2^{2/q-1} t_n^2 m^{1-2/q}. \tag{2.3}$$

The metric entropy estimate (2.2) is the best possible. Indeed, if f_j, $1 \le j \le M_n$, are orthonormal functions, then (2.2) is sharp in order for any ϵ satisfying that ϵ/t_n is bounded away from 1 (see [45]). Also note that if we let ν_0 be the discrete measure $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, where $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ are fixed points in a fixed design, then $\|g\|_{L^2(\nu_0)} = (\frac{1}{n}\sum_{i=1}^n |g(\mathbf{x}_i)|^2)^{1/2}$. Thus, part (i) of Theorem 1 implies Lemma 3 of [53], 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 useful to derive the exact minimax rates for some of the possible situations in terms of M_n, q, and t_n.

With the tools provided in Yang and Barron [72], given fixed q and t_n, one can derive minimax rates of convergence for ℓ_q-aggregation problems and also for linear regression with ℓ_q-constraints. However, the goal for this work is to obtain adaptive estimators that simultaneously work for $\mathcal{F}_q(t_n)$ with any choice of $0 \le q \le 1$ and t_n, and more.

2.5. *An insight from the sparse approximation bound based on classical model selection theory*

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

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

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

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

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 ratio m^*/M_n is called a sparsity index in [71] (section III.D) that characterizes, up to a log factor, how much sparse estimation by model selection improves the estimation accuracy based on nested models only. 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 [72]). 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 [72]). 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., [10, 71, 11, 65, 18, 6]). 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 $0 \leq m \leq M_n$. 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 $m^* \leq M_n/2$. Define

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

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., $t_n^2 m^{1-2/q} + \frac{\sigma^2 m}{n}$). The best choice will typically make the approximation error $t_n^2 m^{1-2/q}$ of the same order as $\frac{m(1+\log\frac{M_n}{m})}{n}$. 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 oracle inequality at the best model of this size gives the minimax rate of convergence, as will be seen. 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., [71], 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 $0 \le q \le 1$.

The aforementioned rates of convergence for estimating functions in $\ell_{q,t_n}^{M_n}$-hulls for $0 \le q \le 1$ will be confirmed, and our estimators will achieve the rates adaptively in some generality. From the insight gained above, to construct a multi-directional (or universal) aggregation method that works for all $0 \le q \le 1$, it suffices to aggregate the estimates from the subset models for adaptation, which will automatically lead to simultaneous optimal performance in $\ell_{q,t_n}^{M_n}$-hulls.

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, let $\mathcal{F}_{J_m}$ be as defined in section 2.4, and let $\mathcal{F}^L_{J_m,s} = \{f_\theta = \sum_{j\in J_m} \theta_j f_j : \|\theta\|_1 \le s, \|f_\theta\|_\infty \le L\}$ 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 combination of $s \in \mathbb{N}$ and function class $\mathcal{F}^L_{J_m,s}$.

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

Let $\mathcal{F}_q^L(t_n) = \mathcal{F}_q(t_n) \cap \{f : \|f\|_\infty \le L\}$ for $0 \le q \le 1$. In the previous section, we have defined $m_* = m_*(q, t_n)$ to be the effective model size for $0 < q \le 1$. Now, for ease of presentation, we extend the definition to

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

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 $0 < q \le 1$) and the hard sparsity one (ℓ_0-constraint), and this smaller size defines 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. In particular, Theorems 2 and 3 provide upper and lower bounds to determine the order of the risk regret for ℓ_q-aggregation of estimates. The specific behaviors of $REG(m_*^{\mathcal{F}})$ for the three different cases will be precisely discussed later.

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 will be 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 [69] (page 36) for multi-directional aggregation and in Lounici [49] (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 [49].

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

Theorem 2. *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* $0 < q \leq 1$, *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*

$$\begin{aligned} &R(\hat{f}_{F_n}; f_0; n) \leq C_1 REG(m_*^{\mathcal{F}}) + \\ &C_0 \begin{cases} d^2(f_0; \mathcal{F}_q^L(t_n)) + \frac{C_2\sigma^2 \log(1+t_n)}{n} & \textit{for case 1,} \\ \inf_{s\geq 1}\left(\inf_{\{\theta: \|\theta\|_1 \leq s, \|\theta\|_0 \leq k_n, \|f_\theta\|_\infty \leq L\}} \|f_\theta - f_0\|^2 + \frac{C_2\sigma^2 \log(1+s)}{n}\right) & \textit{for case 2,} \\ d^2(f_0; \mathcal{F}_q^L(t_n) \cap \mathcal{F}_0^L(k_n)) + \frac{C_2\sigma^2 \log(1+t_n)}{n} & \textit{for case 3.} \end{cases} \end{aligned}$$

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

For all these cases, C_0 *and* C_2 *do not depend on* n, f_0, t_n, q, k_n, M_n*;* C_1 *does not depend on* n, f_0, t_n, 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 3. When $q = 1$, our theorem covers some important previous aggregation results. With $t_n = 1$, Juditsky and Nemirovski [42] obtained the optimal result for large M_n; Yang [69] 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 [58] presented the optimal rate for both large and small M_n, 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 [4] have improved over [69] and [58] by giving an optimal risk bound. Even when $q = 1$, our result is more general in that t_n is allowed to be arbitrary. Note also that in some specific cases, the induced sparsity with ℓ_1-constraint was explored earlier in e.g., [30, 9, 71]. The latter two papers dealt with nonparametric situations with mild assumptions on the terms in the approximation systems. In particular, when the true function has a finite-order linear expression, the estimators achieve the minimax optimal rate $\sqrt{(\log n)/n}$ when M_n grows polynomially fast in n.

Remark 4. The upper rate for $q = 0$ as well as its interpretation is not new in the literature (see, e.g., Theorem 1 of [71], Theorems 1 and 4 of [65]): by noticing that there are $\binom{M_n}{k_n}$ subsets of size k_n and that $\log \binom{M_n}{k_n} \leq k_n \left(1 + \log(M_n/k_n)\right)$, the rate for $q = 0$, which directly imposes hard sparsity on the maximum number of relevant terms, is just the log number of models of size k_n divided by the sample size.

Remark 5. 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, if $t_n \leq e^{cn} \wedge e^{cm_*(1+\log(M_n/m_*))}$ for a pure constant c, then $\log(1 + t_n)/n$ is upper bounded by a multiple of $REG(m_*^{\mathcal{F}_q(t_n)})$. Then, under the condition that the approximation errors involved in the risk bounds are of the same order, **AC-** strategies have the same upper bound orders as **T-** strategies. For case 2, the same is true if for some $s \leq e^{cn} \wedge e^{ck_n(1+\log(M_n/k_n))}$, the ℓ_1 norm constraint does not enlarge the approximation error order.

Remark 6. For case 2, the boundedness assumption of $\|f_j\| \leq 1$, $1 \leq j \leq M_n$ 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.

Next, we show that the upper rates in Theorem 2 cannot be generally improved by giving a theorem stating that the lower bounds of the risk are of the same order in some situations, as is typically done in the literature on aggregation of estimates. The following theorem implies that the estimator by our strategies is indeed minimax adaptive for ℓ_q-aggregation of estimates. 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 $\|f_0\|_\infty \leq L$ 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 3. *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 $0 < q \leq 1$, one can find a regression function f_0 (that may depend on $\mathcal{F}$) such that*

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

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 $\|f_0\| \leq L$ 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 and L for cases 1 and 3 and on L for case 2.*

(iii) *With the additional knowledge $\|f_0\|_\infty \leq L$ 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 $0 < q \leq 1$, if $\sup_{f_\theta \in \mathcal{F}_q(t_n)} \|f_\theta\|_\infty \leq L$; 2) for $\mathcal{F} = \mathcal{F}_0(k_n)$, if $\sup_{1 \leq j \leq M_n} \|f_j\|_\infty \leq L < \infty$ 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 $M_n/\left(1+\log\frac{M_n}{k_n}\right) \leq bn$ for some constant $b>0$ and the orthonormal basis is specially chosen.

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

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

Thus, under the condition that $\left(\sup_{1\leq j\leq M_n}\|f_j\|_\infty\right)t_n$ is upper bounded, $\sup_{f_\theta\in\mathcal{F}_q(t_n)}\|f_\theta\|_\infty \leq L$ 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. 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 can be expressed in different ways:

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

The second expression is transparent in interpretation: due to the sparsity condition, we only need to consider models of the effective size m_* and the risk goes with the searching price $\frac{m_*}{n}SER\left(m_*\right)$ (the estimation error of m_* parameters is being dominated in order). The last expression means that we can do better than searching over the models of the ideal model size m^*, which has the risk $\frac{m^*}{n} SER\left(m^*\right)$. The minimax risk is deflated by a factor of $SER\left(m^*\right)^{\frac{q}{2}}$, which becomes larger as $q \to 1$, pointing out that the factor $SER(m^*)$ has to be downsized more as the ℓ_q-ball becomes larger. When $m^* = M_n$ (the full model), $SER(m^*)$ reduces to 1. When $m^* \leq (1 + \log(M_n/m^*))^{q/2}$ 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 $k_n = M_n \leq n$, 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 2, we deduce that for both the **T-** strategies and **AC-** strategies, under the assumption $\sup_j \|f_j\|_\infty \leq L$, our estimator satisfies

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

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 [58], 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 $q = 1$).

3.3. ℓ_q-combination of procedures

Suppose we start with a collection of estimation procedures $\Delta = \{\delta_1, \ldots, \delta_{M_n}\}$ instead of a dictionary of estimates. Let $\hat{f}_j$ be the estimator of the unknown true regression function based on the procedure δ_j, $1 \leq j \leq M_n$, at a certain sample size. Our goal is to combine the estimators

$\{\hat{f}_j : 1 \le j \le M_n\}$ to achieve the best performance in

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

We split the data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ into three parts: $Z^{(1)} = (\mathbf{X}_i, Y_i)_{i=1}^{n_1}$, $Z^{(2)} = (\mathbf{X}_i, Y_i)_{i=n_1+1}^{n_1+n_2}$ and $Z^{(3)} = (\mathbf{X}_i, Y_i)_{i=n_1+n_2+1}^{n}$. Use the data $Z^{(1)}$ to obtain estimators $\hat{f}_1, \ldots, \hat{f}_{M_n}$ and use the data $Z^{(2)}$ to construct T-estimators or AC-estimators based on subsets of $\hat{f}_1, \ldots, \hat{f}_{M_n}$. The data $Z^{(3)}$ are used to construct the final estimator $\hat{f}_\Delta$ by aggregating the T-estimators or AC-estimators and the null model using Catoni's or the ARM algorithm as done in the previous section. For simplicity, assume n is a multiple of 4 and choose $n_1 = n/2$, $n_2 = n/4$. Upper bounds for combining procedures by our strategy are obtained similarly. The only difference is that $d^2(f_0;\mathcal{F})$ is replaced by the risk of the best constrained linear combination of the estimators $\hat{f}_{1,n/2}, \ldots, \hat{f}_{M_n,n/2}$, where we add the second subscript $n/2$ to emphasize that the estimators are constructed with a reduced sample size. For example, by **T-** strategies, we have that for any $0 < q \le 1$ and $t_n > 0$,

$$R(\hat{f}_\Delta; f_0; n) \le C_0 \inf_{\theta \in B_q(t_n;M_n)} E \left\| f_0 - \sum_{j=1}^{M_n} \theta_j \hat{f}_{j,n/2} \right\|^2 + C_1 \cdot REG(m_*^{\mathcal{F}_q(t_n)}),$$

and again such risk bounds simultaneously hold for $0 \le q \le 1$ and $t_n > 0$.

Note that these risk bounds involve the accuracies of the candidate procedures at a reduced sample size $n/2$ due to data splitting to come up with the estimates to be aggregated. Ideally, we want to have $C_0 = 1$ and $\hat{f}_{j,n/2}$ replaced by $\hat{f}_{j,n}$. At this time, we are unaware of any such risk bound that holds for combining general estimators (in fixed design case, Leung and Barron's algorithm does not involve data splitting, but it works only for least squares estimators). Because of this, the theoretical attractiveness that the constant C_0 being 1 in the aggregation stage, unfortunately, disappears since the remaining parts in the risk bounds also depend on the data splitting and there seems to be no reason to expect with certainty that an aggregation method with $C_0 = 1$ has a better risk, even asymptotically, than another one with $C_0 > 1$. Therefore, for combining general statistical procedures, it is unclear how useful $C_0 = 1$ is even from a theoretical perspective. (It seems that there is one scenario that one can argue otherwise: the candidate estimates are truly provided. In the application of combining forecasts sequentially, the candidate forecasts may be provided by other experts/commercial companies and the statistician does not have access to the data based on which the forecasts are built. In this context, since no data splitting is needed, $C_0 = 1$ leads to a theoretical advantage compared to $C_0 > 1$.) For this reason, in our view, results with $C_0 > 1$ (but not too large) are also important for combining procedures. Indeed, such results often have strengths in other aspects such as allowing heavy tail distributions for the errors and allowing dependence of the observations.

Nonetheless, regardless of the degree of practical relevance, limiting attention to the aggregation step and pursuing $C_0 = 1$ in that local goal is certainly not without a theoretical appeal.

Some additional interesting results on combining procedures are in [3, 15, 20, 26, 27, 35, 36, 39, 38, 63, 68].

4. Linear regression with ℓ_q-constrained coefficients under random design

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

$$Y = f_0(\mathbf{X}) + \varepsilon = \sum_{j=1}^{M_n} \theta_j X_j + \varepsilon. \tag{4.1}$$

As previously defined, for a function $f(x_1, \dots, x_{M_n}) : \mathcal{X} \to \mathbb{R}$, the L_2-norm $\|f\|$ is the square root of $Ef^2(X_1, \dots, X_{M_n})$, 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 $1 \leq m \leq M_n \wedge n$ 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 $\mathcal{G}^L_{J_m,s} = \{\sum_{j\in J_m} \theta_j x_j : \|\theta\|_1 \leq s, \|f_\theta\|_\infty \leq L\}$. 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 $\mathcal{G}^L_{J_m,s}$ instead of $\mathcal{F}_{J_m}$ and $\mathcal{F}^L_{J_m,s}$.

4.1. Upper bounds

We give upper bounds for the risk of our estimator assuming $f_0 \in \mathcal{F}^L_q(t_n; M_n)$, $\mathcal{F}^L_0(k_n; M_n)$, or $\mathcal{F}^L_q(t_n; M_n) \cap \mathcal{F}^L_0(k_n; M_n)$, where $\mathcal{F}^L = \{f : f \in \mathcal{F}, \|f\|_\infty \leq L\}$ for a positive constant L. Let $\alpha_n = \sup_{f \in \mathcal{F}^L_0(k_n;M_n)} \inf\{\|\theta\|_1 : f_\theta = f\}$ be the maximum smallest ℓ_1-norm needed to represent the functions in $\mathcal{F}^L_0(k_n; M_n)$. For ease of presentation, define $\Psi^{\mathcal{F}}$ as follows:

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

$$\Psi^{\mathcal{F}^L_0(k_n;M_n)} = \sigma^2 \left(1 \wedge \frac{k_n \left(1 + \log \frac{M_n}{k_n}\right)}{n} \right),$$

$$\Psi^{\mathcal{F}^L_q(t_n;M_n) \cap \mathcal{F}^L_0(k_n;M_n)} = \Psi^{\mathcal{F}^L_q(t_n;M_n)} \wedge \Psi^{\mathcal{F}^L_0(k_n;M_n)}.$$

In addition, for lower bound results, let $\underline{\Psi}^{\mathcal{F}^L_q(t_n;M_n)}$ $(0 \leq q \leq 1)$ and $\underline{\Psi}^{\mathcal{F}^L_q(t_n;M_n)\cap\mathcal{F}^L_0(k_n;M_n)}$ $(0 < q \leq 1)$ be the same as $\Psi^{\mathcal{F}^L_q(t_n;M_n)}$ and $\Psi^{\mathcal{F}^L_q(t_n;M_n)\cap\mathcal{F}^L_0(k_n;M_n)}$, respectively, except that when $0 < q \leq 1$ and $m_* = 1$, $\underline{\Psi}^{\mathcal{F}^L_q(t_n;M_n)}$ 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 $\underline{\Psi}^{\mathcal{F}^L_q(t_n;M_n)\cap\mathcal{F}^L_0(k_n;M_n)}$ is modified the same way.

Theorem 4. *Suppose $A_{\mathbf{E-G}}$ holds for the **E-G** strategy respectively, and* $\sup_{1\leq j\leq M_n} \|X_j\|_\infty \leq 1$. *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* $0 < q \leq 1$, *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* $0 < q \leq 1$, *we have*

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

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* $0 < q \leq 1$, *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* $0 < q \leq 1$, *we have*

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

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

Remark 8. The constants C_1 and C_2 may depend on L, p_0, σ^2, $\overline{\sigma}^2/\underline{\sigma}^2$, α, U_α, V_α when relevant.

Remark 9. The rate $\left(\frac{\log n}{n}\right)^{1-q/2}$ for $0 < q < 1$ has appeared in related regression or normal mean problems, e.g., in [30] (Theorem 3), [72] (section 5), [40] (section 6), and [41]. For function classes defined in terms of infinite order orthonormal expansion with bounded q-norm of the coefficients and with ℓ_2-norm of the tail coefficients decaying at a polynomial order, the rate of convergence $(\log n/n)^{1-q/2}$ is derived in [71] (page 1588) (when the tail of the coefficients decays fast, the rate is improved to $(1/n)^{1-q/2}$). Note that only the upper rates are given there.

4.2. Lower bounds

To derive lower bounds, we make the following near orthogonality assumption on sparse sub-collections of the predictors. Such an assumption, similar to the sparse Riesz condition (SRC) (Zhang [78]) under fixed design, is used only for lower bounds but not for upper bounds.

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

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

Theorem 5. *Suppose the noise ε follows a normal distribution with mean 0 and variance $0 < \sigma^2 < \infty$.*

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

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

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

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

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

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

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

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

Remark 10. Note that in (i), at the transition from $m_* > 1$ to $m_* = 1$, i.e., $nt_n^2\tau \approx 1 + \log\frac{M_n}{(nt_n^2\tau)^{q/2}}$, we see continuity:

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

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. Normality of the errors is not essential at all for the lower bounds. With some additional efforts, one can show that these lower rates are also valid under Assumption Y2, which we will not give here.

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 Theorem 6 below) are held constant in the relevant expressions.

Theorem 6. *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 $0 < \sigma \leq \overline{\sigma} < \infty$. Also assume there exists a known constant $L_0 > 0$ such that $\sup_{1 \leq j \leq M_n} \|X_j\|_\infty \leq L_0 < \infty$.*

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

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

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

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

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

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

Remark 11. When considering jointly the ℓ_q-constraint for a fixed $0 < q \leq 1$ and $q = 0$, since the associated function classes are not nested, one cannot immediately deduct the optimal rate of convergence for their intersection. In our problem, the simple rule works: when the upper bound k_n of the ℓ_0-constraint is smaller than the effective model size m_*, the additional ℓ_q-constraint does reduce the parameter searching space, but this reduction is not essential and the rate is equal to the rate for $q = 0$. In contrast, when the effective model size m_* is smaller than k_n, the ℓ_0-constraint does reduce the parameter searching space determined by the ℓ_q-constraint, but not essential from the uniform estimation standpoint and the rate is then $m_* \log(1 + M_n/m_*)/n$. Clearly, both rates can be interpreted as the log number of models of size k_n or m_* over the sample size.

5. Adaptive minimax estimation under fixed design

Consider the linear regression model (4.1) under fixed design, $Y_i = f_0(\mathbf{x}_i) + \varepsilon_i$, $i = 1, ..., n$, where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,M_n})' \in \mathcal{X} \subset \mathbb{R}^{M_n}$ are fixed, $1 \le i \le n$, and the random errors ε_i are i.i.d. $N(0, \sigma^2)$. Suppose $\max_{1 \le j \le M_n} \sum_{i=1}^n x_{i,j}^2/n \le 1$. Let $f_0^n = (f_0(\mathbf{x}_1), \ldots, f_0(\mathbf{x}_n))'$. For any function f: $\mathcal{X} \to \mathbb{R}$, define the norm $\|\cdot\|_n$ by $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f^2(\mathbf{x}_i)$. Our goal is to estimate the regression mean f_0^n through a linear combination of the predictors with the coefficients θ satisfying a ℓ_q-constraint ($0 \le q \le 1$). For an estimate $\hat{f}$ of f_0, define its average squared error to be

$$ASE(\hat{f}) = \|\hat{f} - f_0\|_n^2.$$

We consider subset selection based estimators. Let $J_m \subset \{1, 2, \ldots, M_n\}$ be a model of size m ($1 \le m \le M_n$). Our strategy is to choose a model using a model selection criterion, and the resulting least squares estimator is used for f_0^n. The loss of a given model J_m is $ASE(\hat{f}_{J_m}) = \|\hat{Y}_{J_m} - f_0^n\|_n^2$ (with a slight abuse of notation), where $\hat{Y}_{J_m} = (\hat{Y}_{1,J_m}, \ldots, \hat{Y}_{n,J_m})'$ is the projection onto the column span of the design matrix of model J_m. The alternative strategy of model mixing will be taken as well. Although our estimators do not directly consider the ℓ_q-constraint, it will be shown to automatically adapt to the sparsity of f_0 in terms of ℓ_q-representation by the dictionary.

For a function class $\mathcal{F}$, for the fixed design, define the approximation error $d_n^2(f_0; \mathcal{F}) = \inf_{f \in \mathcal{F}} \|f - f_0\|_n^2$. We will consider both σ known and σ unknown cases. As will be seen, the results are quite different in some aspects, and an understanding on what the different assumptions can lead to is important to reach a deeper insight on the theoretical issues.

5.1. When σ is known

For a model J_m of size m ($1 \le m \le M_n$), the ABC criterion proposed in Yang (1999) is

$$ABC(J_m) = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i,J_m})^2 + 2r_{J_m}\sigma^2 + \lambda\sigma^2 C_{J_m},$$

where λ is a pure constant, r_{J_m} is the rank of the design matrix of J_m, and C_{J_m} is the model index descriptive complexity. Let r_{M_n} denote the rank of the full model J_{M_n}, which is assumed to be at least 1.

Let $\bar{J}$ denote the model that gives the full projection matrix $I_{n\times n}$ (since the ASE at the design points is the loss of interest, this identity projection is permitted). We define $ABC(\bar{J}) = 2n\sigma^2 + \lambda\sigma^2 C_{\bar{J}}$. Let J_0 denote the null model that only includes the intercept and define $ABC(J_0) = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + 2\sigma^2 + \lambda\sigma^2 C_{J_0}$, where $\bar{Y} = \sum_{i=1}^{n} Y_i/n$. The model index descriptive complexity C_J satisfies $C_J > 0$ and $\sum_J e^{-C_J} \le 1$, where the summation is over all the candidate models being considered.

The subset models of size $1 \le m \le M_n \wedge n$, the models J_0 and $\bar{J}$ are considered with the complexity $C_{J_m} = -\log 0.85 + \log((M_n - 1) \wedge n) + \log\binom{M_n}{m}$ for a subset model with $m < M_n$, $C_{J_{M_n}} = -\log 0.05$ for the full model J_{M_n}, $C_{J_0} = -\log 0.05$ for the null model J_0, and $C_{\bar{J}} = -\log 0.05$ for the full projection model $\bar{J}$. Note that for the purpose of estimating f_0^n, there is no problem with duplication in the list of candidate models.

Let Γ_n denote the set of all the models considered and the model chosen by the ABC criterion is

$$\hat{J} = \arg\min_{J\in\Gamma_n} ABC(J).$$

The ABC estimator $\hat{f}_{\hat{J}}$ is the fitted value $\hat{Y}_{\hat{J}}$. Let $\bar{f}_J = \mathcal{P}_J f_0^n$ be the projection of f_0^n into the column space of the design matrix of model J.

For ease of presentation, define $\Phi^{\mathcal{F}}$ as follows:

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

$$\Phi^{\mathcal{F}_0(k_n;M_n)} = \frac{\sigma^2 k_n\left(1+\log\frac{M_n}{k_n}\right)}{n} \wedge \frac{\sigma^2 r_{M_n}}{n},$$
$$\Phi^{\mathcal{F}_q(t_n;M_n)\cap\mathcal{F}_0(k_n;M_n)} = \Phi^{\mathcal{F}_q(t_n;M_n)} \wedge \Phi^{\mathcal{F}_0(k_n;M_n)}.$$

In addition, for lower bound results, let $\underline{\Phi}^{\mathcal{F}_q(t_n;M_n)}$ $(0 \le q \le 1)$ and $\underline{\Phi}^{\mathcal{F}_q(t_n;M_n)\cap\mathcal{F}_0(k_n;M_n)}$ $(0 < q \le 1)$ be the same as $\Phi^{\mathcal{F}_q(t_n;M_n)}$ and $\Phi^{\mathcal{F}_q(t_n;M_n)\cap\mathcal{F}_0(k_n;M_n)}$, respectively, except that when $0 < q \le 1$ and $m_* = 1$, $\underline{\Phi}^{\mathcal{F}_q(t_n;M_n)}$ takes the value $t_n^2 \wedge \frac{\sigma^2 r_{M_n}}{n}$ instead of $(t_n^2 \vee \frac{\sigma^2}{n}) \wedge \frac{\sigma^2 r_{M_n}}{n}$ and $\underline{\Phi}^{\mathcal{F}_q(t_n;M_n)\cap\mathcal{F}_0(k_n;M_n)}$ is modified the same way. In the fixed design case, the ranks of the design matrices are certainly relevant in risk bounds (see, e.g., [65, 54]).

Theorem 7. *When $\lambda \ge 5.1\log 2$, the ABC estimator $\hat{f}_{\hat{J}}$ simultaneously has the following properties.*

(i) For $\mathcal{F} = \mathcal{F}_q(t_n;M_n)$ with $0 < q \le 1$, or $\mathcal{F} = \mathcal{F}_0(k_n;M_n)$ with $1 \le k_n \le M_n$, or $\mathcal{F} = \mathcal{F}_q(t_n;M_n)\cap \mathcal{F}_0(k_n;M_n)$ with $0 < q \le 1$ and $1 \le k_n \le M_n$, we have

$$\sup_{f_0\in\mathcal{F}} E(ASE(\hat{f}_{\hat{J}})) \le B\Phi^{\mathcal{F}},$$

where the constant B depends only on q and λ for the first and third cases of $\mathcal{F}$, and depends only on λ for the second case.

(ii) In general, for an arbitrary f_0^n, we have

$$E(ASE(\hat{f}_{\hat{J}})) \le B\left(\left\|\bar{f}_{J_{M_n}} - f_0^n\right\|_n^2 + \inf_{J_m:1\le m<M_n}\left(\left\|\bar{f}_{J_m} - \bar{f}_{J_{M_n}}\right\|_n^2 + \frac{\sigma^2 r_{J_m}}{n}\right.\right.$$
$$\left.\left.+\frac{\sigma^2\log(M_n\wedge n)}{n} + \frac{\sigma^2\log\binom{M_n}{m}}{n}\right) \wedge \frac{\sigma^2 r_{M_n}}{n}\right) \wedge B\left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n}\right)\wedge\sigma^2\right),$$

where the constant B depends only on λ.

Remark 12. In (i), the case $\mathcal{F} = \mathcal{F}_0(k_n;M_n)$ does not require $\max_{1\le j\le M_n}\sum_{i=1}^n x_{i,j}^2/n \le 1$.

Remark 13. In pursuing the best performance in each case of $\mathcal{F}$, the general risk bound in (ii) reduces to $B\Phi^{\mathcal{F}}$ plus the approximation error $d_n^2(f_0;\mathcal{F}) = \inf_{f\in\mathcal{F}}\|f - f_0\|_n^2$.

For the lower bound results, as before, additional conditions are needed. Let Ξ denote the design matrix of the full model J_{M_n}.

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 $\|\theta\|_0 \leq \min(2\gamma, M_n)$, we have

$$\underline{a}\|\theta\|_2 \leq \frac{1}{\sqrt{n}}\|\Xi\theta\|_2 \leq \overline{a}\|\theta\|_2.$$

This condition is slightly weaker than Assumption 2 in [53], which was used to derive minimax lower bounds for $0 < q \leq 1$.

Theorem 8. *Suppose the noise ε follows a normal distribution with mean 0 and variance $0 < \sigma^2 < \infty$. For $\mathcal{F} = \mathcal{F}_q(t_n; M_n)$ with $0 < q \leq 1$, or $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ with $1 \leq k_n \leq M_n$, or $\mathcal{F} = \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)$ with $0 < q \leq 1$ and $1 \leq k_n \leq M_n$, under Assumption SRC$'$ with $\gamma = m_*$, or k_n, or $k_n \wedge m_*$ respectively, we have*

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}} E(ASE(\hat{f})) \geq B^{'}\underline{\Phi}^{\mathcal{F}},$$

where the estimator $\hat{f}$ is over all estimators, and the constant $B^{'}$ depends only on $\underline{a}$ and $\overline{a}$ for the second case of $\mathcal{F}$ and additionally on q for the first and third cases of $\mathcal{F}$.

Remark 14. If SRC$'$ is not satisfied on the set of all the predictors but is satisfied on a subset of M_0 predictors, as long as $\log \frac{M_n}{m_*}$, $\log \frac{M_n}{k_n}$, and $\log \frac{M_n}{m_* \wedge k_n}$ are of the same order as $\log \frac{M_0}{m_*}$, $\log \frac{M_0}{k_n}$, and $\log \frac{M_0}{m_* \wedge k_n}$, respectively, we get the same risk lower rates. When M_n is really large, this relaxation of SRC$'$ can be much less stringent for application.

For the case $q = 0$, the achievability of the upper rate is a direct consequence of [65]. The lower rates for $q = 0$ and/or 1 are given in [54], where the satisfiability of the SRC$^{'}$ is also worked out. Raskutti et al. [53], under the assumption that the rank of the full design matrix is n, derived the minimax rates of convergence $t_n^q(\log(M_n)/n)^{1-q/2}$ for $0 < q < 1$ in an in-probability sense for linear regression with fixed design with the ℓ_q-constraint when $M_n \gg n$ and $M_n/(t_n^q n^{q/2}) \geq M_n^\kappa$ with some $\kappa \in (0, 1)$. From our result, the ABC estimator simultaneously achieves the minimax rates of convergence for all $0 \leq q \leq 1$ and for all $M_n \geq 2$ and t_n no smaller than order $n^{-1/2}$, and also under the joint constraints when $q = 0$ and $0 < q \leq 1$. We also need to point out that we only work on estimating the regression mean in this work, but [53] showed that, under additional conditions, these upper rates are also valid for the estimation of the parameter θ under the squared error and verified their minimaxity. Concurrent work by Ye and Zhang [73] also derived performance bounds on the coefficient estimation that are optimal in a sense of uniformity over the different designs.

In application, the assumption that $f_0 \in \mathcal{F}_q(t_n; M_n)$ or $f_0 \in \mathcal{F}_0(k_n; M_n)$ may sometimes be too strong to be appropriate. Thus, risk bounds that permit model mis-specification, i.e., $f_0 \notin \mathcal{F}_q(t_n; M_n)$, are desirable. Part (ii) in the upper bound theorem (Theorem 7) shows that the ABC estimator handles model mis-specification. Indeed, for the different ℓ_q-constraints, the risk of the ABC estimator is upper bounded by a multiple of $d_n^2(f_0; \mathcal{F}_q(t_n; M_n))$ plus the earlier upper bounds, respectively. Therefore, model mis-specification or not, our estimator is minimax rate adaptive over the $\ell_{q,t_n}^{M_n}$-hulls without any knowledge about the values of q, t_n and k_n (as long as t_n is not trivially small).

One limitation of this result, from one theoretical point, is that the factor is larger than one in front of $d_n^2(f_0; \mathcal{F})$. When the initial estimates need to be obtained based on the same data available, the multiplying factor being one no longer necessarily has any essential advantage. However, striving for the right constant is theoretically attractive when the elements in the dictionary are observed or truly provided by others.

In that direction, recently, Rigollet and Tsybakov [54], by considering an estimator based on the mixing-least-square-estimators algorithm of Leung and Barron [46] with some specific choice of prior probabilities on the models, have provided in-expectation optimal upper bounds for ℓ_0- and/or ℓ_1-aggregation. With the power of the oracle inequality (or the index of resolvability bound), their estimator is shown to be adaptive over ℓ_0- and ℓ_1-hulls. Their results do not address ℓ_q-aggregation for $0 < q < 1$. We next show that we can have an estimator that handles all $0 \leq q \leq 1$ in generality.

The mixed least squares estimator by the mixing algorithm of Leung and Barron (2006) is given by

$$\hat{f}^{MLS} = \sum_{J \in \Gamma_n} w_J \hat{Y}_J \qquad \text{with} \qquad w_J = \frac{\pi_J \exp\{-\widehat{R}_J/(4\sigma^2)\}}{\sum_{J' \in \Gamma_n} \pi_{J'} \exp\{-\widehat{R}_{J'}/(4\sigma^2)\}},$$

where $\widehat{R}_J = n\|Y - \hat{Y}_J\|_n^2 + 2r_J\sigma^2 - n\sigma^2$ is the unbiased risk estimate for $\hat{Y}_J$. Let the prior on model J be chosen as $\pi_{J_m} = 0.85\left(((M_n - 1) \wedge n)\binom{M_n}{m}\right)^{-1}$ for $1 \leq m \leq (M_n - 1) \wedge n$, and

$\pi_{J_{M_n}} = \pi_{J_0} = \pi_{\bar{J}} = 0.05$.

Theorem 9. *Suppose $0 < \sigma < \infty$ is known. For any $M_n \geq 1$, $n \geq 1$, the estimator $\hat{f}^{MLS}$ simultaneously has the following properties.*

(i) For any $0 < q \leq 1$, $t_n > 0$,

$$\begin{aligned} E(ASE(\hat{f}^{MLS})) &\leq d_n^2(f_0; \mathcal{F}_q(t_n; M_n)) \\ &+ B_1 \begin{cases} \frac{\sigma^2 r_{M_n}}{n}, & \text{if } m_* = M_n \wedge n, \\ \sigma^{2-q} t_n^q \left(\frac{1 + \log \frac{M_n}{(nt_n^2 \tau)^{q/2}}}{n} \right)^{1-q/2} \wedge \frac{\sigma^2 r_{M_n}}{n}, & \text{if } 1 < m_* < M_n \wedge n, \end{cases} \end{aligned}$$

and

$$\begin{aligned} E(ASE(\hat{f}^{MLS})) &\leq \left(d_n^2(f_0; \mathcal{F}_q(t_n; M_n)) + \frac{B_1 \left(\sigma^2 (1 + \log M_n) \wedge \sigma^2 r_{M_n}\right)}{n} \right) \\ &\quad \wedge \left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\widetilde{B}_1 \sigma^2}{n} \right), \ \text{if } m_* = 1\ . \end{aligned}$$

(ii) For $1 \leq k_n \leq M_n$,

$$E(ASE(\hat{f}^{MLS})) \leq d_n^2(f_0; \mathcal{F}_0(k_n; M_n)) + B_2 \left(\frac{\sigma^2 k_n \left(1 + \log \frac{M_n}{k_n}\right)}{n} \wedge \frac{\sigma^2 r_{M_n}}{n} \right).$$

(iii) For any $0 < q \leq 1$, $t_n > 0$, and $1 \leq k_n \leq M_n$,

$$\begin{aligned} E(ASE(\hat{f}^{MLS})) &\leq d_n^2(f_0; \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)) \\ &+ B_3 \begin{cases} \frac{\sigma^2 k_n \left(1 + \log \frac{M_n}{k_n}\right)}{n} \wedge \frac{\sigma^2 r_{M_n}}{n} & \text{if } m_* > k_n, \\ \sigma^{2-q} t_n^q \left(\frac{1 + \log \frac{M_n}{(nt_n^2 \tau)^{q/2}}}{n} \right)^{1-q/2} \wedge \frac{\sigma^2 r_{M_n}}{n} & \text{if } 1 < m_* \leq k_n. \end{cases} \end{aligned}$$

and

$$\begin{aligned} E(ASE(\hat{f}^{MLS})) &\leq \left(d_n^2(f_0; \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)) + \frac{B_3 \left(\sigma^2 (1 + \log M_n) \wedge \sigma^2 r_{M_n}\right)}{n} \right) \\ &\quad \wedge \left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\widetilde{B}_3 \sigma^2}{n} \right), \ \text{if } m_* = 1\ . \end{aligned}$$

(iv) For every f_0, we have

$$E(ASE(\hat{f}^{MLS})) \leq B_4 \sigma^2.$$

For these cases, the constants $\widetilde{B}_1$, B_2, $\widetilde{B}_3$ and B_4 are pure constants, and B_1 and B_3 depend on q.

Remark 15. From (ii) above, by taking $k_n = 1$, we have

$$E(ASE(\hat{f}^{MLS})) \leq \inf_{1 \leq j \leq M_n} \|f_j^n - f_0^n\|_n^2 + B_2 \left(\frac{\sigma^2 (1 + \log M_n)}{n} \wedge \frac{\sigma^2 r_{M_n}}{n} \right),$$

where $f_j^n = (x_{1,j}, \dots, x_{n,j})'$. Thus, we have achieved aggregation for adaptation as well under the fixed design.

The risk upper bounds above when q is restricted to be either 0 or 1 or under both constraints are already given in Theorem 6.1 of [54]. The first four cases given there are clearly reproduced here (note that their cases 3 and 1 are just special case and immediate consequence, respectively, of their case 4, given in our bound in (ii)). Their case 5, a sparse aggregation with k_n estimates as studied in [69] (page 36) and [49] (called D-convex aggregation) is implied by our bound in (iii) with q taken to be 1. In the case $q = 1$, a minor difference is that if $\|\bar{f}_{J_0} - f_0^n\|_n^2$ happens to be of a smaller order than $t_n \left(\frac{1 + \log \frac{M_n}{(nt_n^2)^{1/2}}}{n} \right)^{1/2} \wedge \frac{r_{M_n}}{n}$, then our risk bound in (iii) yields a faster rate of convergence. In addition, our inclusion of the full projection model among the candidates guarantees that the risk of our estimator is always bounded, which is not true for the estimator in [54]. Our main contribution here is to handle adaptive ℓ_q-aggregation for the whole range of q between 0 and 1. Note that the upper bounds in the above theorem have already been shown to be minimax-rate optimal under the conditions in Theorem 8.

5.2. A comment on the model selection and model mixing approaches

From the risk bounds in the previous subsection, we see that the model mixing approach leads to the optimal constant 1 in front of the approximation error $d_n^2(f_0; \mathcal{F})$ for the three choices of $\mathcal{F}$, which is not the case for the model selection based estimator. However, the model selection approach may also have its own advantages.

From the proof of Theorem 7 and proof of Theorem 1 in [65], besides the given risk bounds, we also have a general in-probability bound of the form: for any $x > 0$, there are constants c, c'

(absolute constants) and $c^{''}$ (depending on λ and σ^2) such that

$$P\left(\frac{ASE(\hat{f}_{\hat{J}}) + \frac{\lambda\sigma^2 C_{\hat{J}}}{n}}{R_n(f_0)} \geq c + x\right) \leq c^{'} \exp\left(-c^{''} x\right),$$

where $R_n(f_0) = \inf_{J\in\Gamma_n}\left(\left\|\bar{f}_J - f_0^n\right\|_n^2 + \frac{\sigma^2 r_J}{n} + \frac{\lambda\sigma^2 C_J}{n}\right)$ is an index of resolvability, which specializes to the upper bounds in (i) and (ii) of Theorem 7, respectively in those situations. Thus, we know that not only $ASE(\hat{f}_{\hat{J}})$ is at order $R_n(f_0)$ with upper deviation probability exponentially small (in x), but also the complexity of the selected model, $\frac{\lambda\sigma^2 C_{\hat{J}}}{n}$, is upper bounded in probability in the same way as well. In particular, for estimating a linear regression function with the soft or hard (or both) constraint(s) on the coefficients, the ABC estimator converges at rate $\frac{m_*\left(1+\log\frac{M_n}{m_*}\right)}{n} \wedge \frac{r_{M_n}}{n}$ both in expectation and with upper deviation probability exponentially small, where m_* is the corresponding effective model size in each case. Furthermore, the rank (the actual number of free-parameters) of the model selected by ABC is right at order $m_* \wedge r_{M_n}$ with exception probability exponentially small.

For model mixing estimators based on exponential weighting, however, to our knowledge, no result has shown that their losses are generally at the optimal rate in probability. In fact, a negative result is given in [2] that shows that an exponential weighting based estimator *optimal* for aggregation for adaptation (i.e., its risk regret, or the expected excessive loss, is of order $\frac{\log M_n}{n}$) is necessarily *sub-optimal* in probability (with a non-vanishing probability its excessive loss is at least at the much larger order of $\sqrt{\frac{\log M_n}{n}}$) in certain settings.

Thus, we tend to believe that both the model selection and model mixing approaches have their own theoretical strengths in different ways.

5.3. When σ is unknown

Needless to say, the assumption that σ is fully known is unrealistic. When σ is unknown but is upper bounded by a known constant $\overline{\sigma} > 0$, similar results for rate of convergence can be obtained with a model selection rule different from ABC.

For this situation, Yang [65] proposed the ABC′ criterion:

$$ABC'(J_m) = \left(1 + \frac{2r_{J_m}}{n - r_{J_m}}\right)\left(\sum_{i=1}^{n}(Y_i - \hat{Y}_{i,J_m})^2 + \lambda\overline{\sigma}^2 C_{J_m}\right),$$

which is a modification of Akaike's FPE criterion [1]. We define $ABC'(\bar{J}) = (1+2n)\,\lambda\overline{\sigma}^2 C_{\bar{J}}$ and $ABC'(J_0) = \left(1 + \frac{2}{n-1}\right)\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + \lambda\overline{\sigma}^2 C_{J_0}\right)$. The list of candidate models and complexity assignments need to be different for the different situations, as described below.

1. When $M_n \le n/2$, all the subset models, J_0 and $\bar{J}$ are considered with the complexity $C_{J_m} = -\log 0.85 + \log(M_n - 1) + \log\binom{M_n}{m}$ for a subset model with $m < M_n$, $C_{J_{M_n}} = C_{J_0} = C_{\bar{J}} = -\log 0.05$.
2. When $M_n > n/2$ and $r_{M_n} \ge n/2$, we only consider models with size $m \le n/2$, the model J_0 and the model $\bar{J}$. Then we assign the complexity $C_{J_m} = -\log 0.8 + \log(\lfloor n/2 \rfloor) + \log\binom{M_n}{m}$ for a subset model, $C_{J_0} = C_{\bar{J}} = -\log 0.1$.
3. When $M_n > n/2$ and $r_{M_n} < n/2$, we only consider models with size $m \le n/2$, the full model J_{M_n}, the null model J_0, and the model $\bar{J}$. We assign the complexity $C_{J_m} = -\log 0.85 + \log(\lfloor n/2 \rfloor) + \log\binom{M_n}{m}$ for a subset model, $C_{J_{M_n}} = C_{J_0} = C_{\bar{J}} = -\log 0.05$.

In any of the cases above, let Γ'_n denote the set of all the models considered. The model chosen by the ABC′ is

$$\hat{J}' = \arg\min_{J \in \Gamma'_n} ABC'(J),$$

producing the ABC′ estimator $\hat{f}_{\hat{J}'} = \hat{Y}_{\hat{J}'}$.

Theorem 10. *When $\lambda \ge 40\log 2$, the ABC′ estimator $\hat{f}_{\hat{J}'}$ simultaneously has the following properties.*

(i) For $\mathcal{F} = \mathcal{F}_q(t_n; M_n)$ with $0 < q \le 1$, or $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ with $1 \le k_n \le M_n$, or $\mathcal{F} = \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)$ with $0 < q \le 1$ and $1 \le k_n \le M_n$, we have

$$\sup_{f_0 \in \mathcal{F}} E(ASE(\hat{f}_{\hat{J}'})) \le B\Phi^{\mathcal{F}},$$

where the constant B depends only on q, λ, $\overline{\sigma}$, σ for the first and third cases of $\mathcal{F}$, and depends only on λ, $\overline{\sigma}$, σ for the second case.

(ii) In general, for an arbitrary f_0^n, we have

$$
\begin{aligned}
& E(ASE(\hat{f}_{\hat{J}'})) \\
\leq \; & B\left(\|\bar{f}_{J_{M_n}} - f_0^n\|_n^2 + \inf_{J_m: 1 \leq m < M_n}\left(\|\bar{f}_{J_m} - \bar{f}_{J_{M_n}}\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log(M_n \wedge n)}{n}\right.\right. \\
& \left.\left. + \frac{\sigma^2 \log \binom{M_n}{m}}{n}\right) \wedge \frac{\sigma^2 r_{M_n}}{n}\right) \wedge B\left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n}\right) \wedge \sigma^2\right),
\end{aligned}
$$

where the constant B depends only on λ, $\overline{\sigma}$, σ.

Remark 16. For the results in (i), as seen before, when f_0 is not in the respective class of linear combinations, an obvious modification is needed by adding a multiple of the approximation error $d_n^2(f_0; \mathcal{F})$ in the risk bound.

When $0 < \sigma < \infty$ is fully unknown, a model selection method by Baraud, Giraud and Huet [7] can be used to obtain results on ℓ_q-regression.

They consider a different modification of the FPE criterion [1]:

$$
BGH(J_m) = \left(1 + \frac{pen(J_m)}{n - r_{J_m}}\right)\left(\sum_{i=1}^{n}(Y_i - \hat{Y}_{i,J_m})^2\right),
$$

where $pen(J_m)$ is a penalty assigned to the model J_m. They devise a new form for $pen(J_m)$ (Section 4.1 in [7]) to yield a nice oracle inequality (Corollary 1) that does not require any knowledge of σ, but at the expense of excluding some large models in the consideration. When $M_n \leq (n-7) \wedge \varsigma n$ for some $0 < \varsigma < 1$, we consider all subset models in the model selection process. When M_n is large, we consider only subset models with $n - r_{J_m} \geq 7$ and $m \vee \log \binom{M_n}{m} \leq \varsigma n$ for a fixed $0 < \varsigma < 1$. Combining the tools developed in this and their papers, we have the following result.

Theorem 11. *The BGH estimator $\hat{f}_{\hat{J}}$ has the following properties.*

(i) When $M_n \leq (n-7) \wedge \varsigma n$, for $\mathcal{F} = \mathcal{F}_q(t_n; M_n)$ with $0 < q \leq 1$, or $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ with $1 \leq k_n \leq M_n$, or $\mathcal{F} = \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)$ with $0 < q \leq 1$ and $1 \leq k_n \leq M_n$, we have

$$
\sup_{f_0 \in \mathcal{F}} E(ASE(\hat{f}_{\hat{J}})) \leq B\Phi^{\mathcal{F}},
$$

where the constant B depends only on q and ς for the first and third cases of $\mathcal{F}$, and depends on ς for the second case.

(ii) For a general M_n, if m_* satisfies $m_* \le n-7$ and $m_* \vee \log \binom{M_n}{m_*} \le \varsigma n$, we have

$$\sup_{f_0 \in \mathcal{F}_q(t_n;M_n)} E(ASE(\hat{f}_{\hat{j}})) \le B \begin{cases} \sigma^{2-q} t_n^q \left(\frac{1+\log \frac{M_n}{(nt_n^2\tau)^{q/2}}}{n} \right)^{1-q/2} & \text{if } m_* > 1, \\ t_n^2 \vee \frac{\sigma^2}{n} & \text{if } m_* = 1, \end{cases}$$

where B depends only on q and ς. If k_n satisfies $k_n \le n-7$ and $k_n \vee \log \binom{M_n}{k_n} \le \varsigma n$, we have

$$\sup_{f_0 \in \mathcal{F}_0(k_n;M_n)} E(ASE(\hat{f}_{\hat{j}})) \le B' \frac{\sigma^2 k_n \left(1 + \log \frac{M_n}{k_n}\right)}{n},$$

where B' is a constant that depends only on ς.

Remark 17. As before, when f_0 is not in the respective class, a multiple of the approximation error $d_n^2(f_0;\mathcal{F}) = \inf_{f\in\mathcal{F}} \|f - f_0\|_n^2$ needs to be added in the aggregation risk bound.

From the above theorem, we see that when σ is fully unknown, as long as $M_n \le (n-7) \wedge \varsigma n$ for some $0 < \varsigma < 1$, similar risk bounds to those in Theorem 10 for ℓ_q-regression hold. However, when M_n is larger, the previous risk bounds are seriously compromised: 1) the possible improvement in risk due to low rank of the full model is no longer guaranteed; 2) the previous upper rates determined by the effective model size m_* or k_n are valid only when those model sizes are not excluded from consideration by the BGH criterion; 3) The risk is no longer guaranteed to be always uniformly bounded. Indeed, due to the restriction on the model sizes to be considered, the final risk here can be arbitrarily large. It turns out that this last aspect is not due to technical deficiency in the analysis, but it is a necessary price to pay for not knowing σ at all (see [61]).

6. Discussion

Since early 1990s, sparse estimation has been recognized as an important tool for multi-dimensional function estimation. Emergence of high-dimensional statistical problems in the information age has prompted an increasing attention on the topic from theoretical, computational and applied perspectives. We focus only on a theoretical standpoint in the discussion below.

To our knowledge, several lines of research on sparse function estimation in 1990s produced theoretical foundations that still provide essential understandings on ways to explore sparsity and associated price to pay when pursuing sparse estimation from minimax perspectives. It has been discovered that for some function classes, sparse representations (in contrast to traditional full approximation) result in faster rates of convergence, which alleviate the curse of dimensionality when the problem size is large. Such function classes include, for example, Besov classes (e.g., [31]), Jones-Barron classes ([9, 42]) and may also be defined directly in terms of sparse approximation (e.g., [71], Section III.D). Regarding methods to achieve the optimal sparse estimation, wavelet thresholding with one or more orthonormal dictionaries and model selection with a descriptive complexity penalty term added to the sum of negative maximized likelihood (or a general contrast function) and a multiple of the model dimension have yielded successful theoretical advancements. Oracle inequalities/index of resolvability bounds have been derived that readily give minimax-rate adaptive estimators for various scenarios. In linear representation, ℓ_1-constraints on the coefficients have been long known to be associated with fast rate of convergence for both orthogonal and non-orthogonal bases by model selection or aggregation methods, as mentioned in the introduction of this paper.

It is worth noticing that these research works usually target nonparametric settings. In the past few years, the situation of a large number of naturally observed predictors has attracted much attention, shifting the focus to much simpler linear modeling. As pointed out earlier, the work in the 1990s on model selection has direct implications for the high-dimensional linear regression. For example, if the sum of the absolute values of the linear coefficients is bounded (ℓ_1-constraint), then the rate of convergence is bounded by $(\log n/n)^{1/2}$ as long as M_n increases only polynomially in n. If only k_n terms have non-zero coefficients (ℓ_0-constraint), then the rate of convergence is of order $k_n(1 + \log(M_n/k_n))/n$ based on model selection with mild conditions on the predictors. However, such subset selection based estimators pose computational challenges in real applications.

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 [34] 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 sparsity natures of the target function in terms of linear approximations subject to ℓ_q-constraints.

Donoho and Johnstone [30] derived insightful general asymptotic minimax risk expressions for estimating the mean vector in ℓ_q-balls ($0 < q < \infty$) under ℓ_p loss ($p \geq 1$) in a Gaussian sequence framework. The work by Raskutti et al. [53] and by Rigollet and Tsybakov [54] are directly related to our work in the fixed design case. The former successfully obtains optimal non-adaptive in-probability loss bounds for their main scenario that M_n is much larger than n for general $0 \leq q \leq 1$ when the true regression function is assumed to be in the $\ell_{q,t_n}^{M_n}$-hull. In contrast, our estimators are adaptive and the risk bounds hold without restrictions on M_n or the "norm" parameter t_n, also allowing the true regression function to be really arbitrary. The work of Rigollet and Tsybakov [54] nicely shows the adaptive aggregation capability of model mixing over ℓ_0 and ℓ_1-balls. Our results are valid over the whole range of $0 \leq q \leq 1$. For lower bounds, our formulation is somewhat different from theirs. In addition, unlike those results, we have also provided results when the error variance is unknown but upper bounded by a known constant or fully unknown. Furthermore, our model selection based estimators have optimal convergence rates also in terms of upper deviation probability, which may not hold for the model mixing estimators. We need to point out that both [53] and [54] have given results on related problems that we do not address in this work.

In our results, the effective model size m_* (as defined in Section 2.5) 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 $1 \wedge \frac{m_*\left(1+\log\left(\frac{M_n}{m_*}\right)\right)}{n}$.

Risk bounds for selection/mixing least squares estimators from a countable collection of linear models (such as given in [65, 46]), together with sparse approximation error bounds, are essential for our approach to devise minimax optimal sparse estimation for fixed design. When the predictors are taken as some initial estimates, the selection/mixing methods can be regarded as aggregation methods with the risk bounds as aggregation risk bounds. In a strict sense, however, these results are not totally satisfactory for at least two reasons. First, the evaluation of performance only at the design points that have been seen already has limited value: i) The strengths of the candidate procedures may not be reflected at all on such a measure; ii) A small ASE on the design points does not mean good behaviors on future predictor values. Second, when the initial estimates are not given (which is almost always the case), to combine arbitrary estimators, data splitting is typically necessary to come up with the candidate estimates and use the rest of the sample for weight assignment. Then, the final risk bounds, unfortunately, depend on how the data are split. In contrast, for the random design case, this is not an issue. We have also seen that because ASE cares only about the performance at the design points, given the i.i.d. normal error assumption, there is absolutely no condition needed on the true regression function, as pointed out in a remark to Theorem 1 in [65]. For random design, however, we have made the sup-norm bound assumption, but the risk bounds guarantee optimal future performance as long as the sampling distribution is unchanged.

Regarding aggregation, we notice that 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 a 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 [58] 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 [69]). 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 $0 < q \leq 1$.

One interesting observation is that aggregation for adaptation is essentially a special case of ℓ_q-aggregation, yet our way of achieving the simultaneous ℓ_q-aggregation is by methods of aggregation for adaptation through model selection/combination.

Aggregation of estimates and regression estimation problems are closely related. For aggregation, besides that the predictors to be aggregated are from some initial estimations (and thus are not directly observed), the emphases are: i) One is unwilling to make assumptions on relationships between the initial estimates so that they can have arbitrary dependence; ii) One is unwilling to make specific assumptions on the true regression function beyond that it is uniformly bounded and hence allow model mis-specification. In this game, there is little interest on the true or optimal coefficients in the representation of the regression function in terms of the initial estimates.

Obviously, there are other directions of aggregation that one may pursue. The ℓ_q-aggregation strategy that relies on aggregating subset choices of the initial estimates, as in [69], while producing the most general aggregation risk bounds so far, follows a global aggregation paradigm, i.e., the linear coefficients are globally determined. It is conceivable that sometimes localized weights may provide better estimation/prediction performance (see, e.g., [70]). Much more work is needed here to result in practically effective localized aggregation methods.

Aggregation of estimates, as an important step in combining statistical procedures, has proven to bring theoretically elegant and practically feasible methods for regression estimation/prediction. It is an important vehicle to share strengths of different function estimation methodologies to produce adaptively optimal and robust estimators that work well under minimal conditions. Aggregation by mixing certainly cannot replace model selection when selection of an estimator among candidates or a set of predictors is essential for interpretation or business/operational decisions.

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.

7. General oracle inequalities for random design

Consider the setting in Section 3.2.

Theorem 12. *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 theorem, 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 [10], $IR(f_0)$ 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. (i) For the **T-C** strategy,

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

For the **T-Y** strategy,

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

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

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

For the **AC-Y** strategy,

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

□

Remark 18. Similar oracle inequalities hold for the estimator $\hat{f}_A$ under the linear regression setting with random design: $d^2(f_0; \mathcal{F}_{J_m})$ is replaced by $d^2(f_0; \mathcal{G}_{J_m})$, and $\sum_{j\in J_m} \theta_j f_j$ is replaced by $\sum_{j\in J_m} \theta_j x_j$ in the above theorem.

8. Proofs

Proof of Theorem 1.

Proof. (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., [32], page 98). Thus, there are 2^{k-1} functions g_j, $1 \le j \le 2^{k-1}$, such that

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

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 $\sum_{i=1}^{M_n} |c_i| \le \epsilon_k$. 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

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

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 the 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 $|L^*|t_n^q/m < \sum |c_j|^q \le t_n^q$. So, $|L^*| < m$. Also,

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

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 $\mathbb{P}(U = D\text{sign}(c_j)f_j) = |c_j|/D$, $j \notin L^*$ and $\mathbb{P}(U = 0) = 1 - \sum_{j\notin L^*} |c_j|/D$. Thus, $\mathbb{E}U = w^*$, where $\mathbb{E}$ denotes expectation with respect to the randomness $\mathbb{P}$ (just introduced). Also, $\|U\| \le D \sup_{1\le j\le M_n} \|f_j\| \le D$. 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,

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

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

□

Proof of Theorem 2.

Proof. To derive the upper bounds, we only need to examine the index of resolvability for each strategy. The natures of the constants in Theorem 2 follow from Theorem 12.

(i) For **T-** strategies, according to Theorem 1 and the general oracle inequalities in Theorem 12, 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

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

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

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

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 $1 \leq m_* \leq m^* < M_n \wedge n$, for $1 \leq m < M_n$, and from Theorem 1, we have

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

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

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

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

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

The upper bound for this case then follows.

For $\mathcal{F} = \mathcal{F}_0(k_n)$, by evaluating the index of resolvability from Theorem 12 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.

□

Global metric entropy and local metric entropy. The tools developed in Yang and Barron [72] allow us to derive minimax lower bounds for ℓ_q-aggregation of estimates or regression under ℓ_q-constraints. Both global and local entropies of the regression function classes are relevant. The following lower bound result slightly generalizes Lemma 1 in [69].

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

$$Y = f_0(\mathbf{X}) + \sigma \cdot \varepsilon, \tag{8.1}$$

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

$$\underline{M}_\alpha^{\text{loc}}(\sigma\epsilon_n) \geq n\epsilon_n^2 + 2\log 2, \tag{8.2}$$

$$\overline{M}(\sqrt{2}\sigma\bar{\epsilon}_n) = n\bar{\epsilon}_n^2, \tag{8.3}$$

$$\underline{M}(\sigma\underline{\epsilon}_n) = 4n\bar{\epsilon}_n^2 + 2\log 2. \tag{8.4}$$

Proposition 5. *(Yang and Barron [72]) The minimax risk for estimating f_0 from model (8.1) 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 [72], but not in the concrete forms. The second lower bound is given in [69]. 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, the mutual information between Θ and the observations $(\mathbf{X}_i, Y_i)_{i=1}^n$ is upper bounded by $\frac{n}{2}\epsilon_n^2$ (see Yang and Barron [72], Sections 7 and 3.2) and an application of Fano's inequality 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 $|N|$ denote 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 [72], 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 [72], the same lower bound holds for $\underline{\mathcal{F}}$ as well.

□

Proof of Theorem 3.

Proof. 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 $B_q(t_n; M_n) = \{\theta : \|\theta\|_q \leq t_n\}$. 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 $\|\theta\|_q \leq {M_n}^{1/q-1/2}\|\theta\|_2$ for $0 < q \leq 2$, for $\epsilon \leq \sqrt{M_n/n}$, taking $f_0^* \equiv 0$, we have

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

where the last equality holds because when $\epsilon \leq \sqrt{M_n/n}$, for $\|\theta\|_2 \leq \epsilon$, $\|\theta\|_q \leq t_n$ is always satisfied. Consequently, for $\epsilon \leq \sqrt{M_n/n}$, the $(\epsilon/2)$-packing of $\mathcal{B}(f_0^*; \epsilon)$ under the L_2 distance is equivalent to the $(\epsilon/2)$-packing of $B_\epsilon = \{\theta : \|\theta\|_2 \leq \epsilon\}$ 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 $M_{1/2}^{\text{loc}}(\epsilon)$ of $\mathcal{F}_q(t)$ under the L_2 distance is at least $\underline{M}_{1/2}^{\text{loc}}(\epsilon) = M_n \log 2$ for $\epsilon \leq \sqrt{M_n/n}$. The minimax lower bound for the case of $\widetilde{m}^* = M_n$ then directly follows from Proposition 5.

When $1 < \widetilde{m}_* \leq \widetilde{m}^* < M_n$, the use of global entropy is handy. Applying the minimax lower bound in terms of global entropy in Proposition 5, with the metric entropy order for larger ϵ (which is tight in our case of orthonormal functions in the dictionary) from Theorem 1, 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 1 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 5.

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 [53] (the result there actually also holds when requiring the pairwise Hamming distance to be strictly larger than $k/2$; see also the derivation of a metric entropy lower bound in [45]), 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 5 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 [45], 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 5. 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 $\|\theta\|_2 \leq \epsilon$ with ϵ being a multiple of $\sqrt{\frac{k_n\left(1+\log\frac{M_n}{k_n}\right)}{n}}$. Then together with $\|\theta\|_0 \leq k_n$, we have $\|\theta\|_1 \leq \sqrt{k_n}\|\theta\|_2$, which is bounded by assumption. The lower bound conclusion then follows from the last part of Proposition 5. To prove the results for the case $M_n/\left(1+\log\frac{M_n}{k_n}\right) \leq bn$, as in [58], we consider the special dictionary $F_n = \{f_i : 1 \leq i \leq M_n\}$ 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 5, 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 $\|f_\theta\| \leq \eta\sqrt{k_n}$ and $\|f_\theta\|_\infty \leq \eta\sqrt{M_n}$. 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 $\frac{M_n}{1+\log\frac{M_n}{k_n}} \leq bn$). 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 $\sqrt{k_n\left(1+\log\frac{M_n}{k_n}\right)/n}$.

□

Proof of Theorem 4.

Proof. For linear regression with random design, we assume the true regression function 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 Theorem 12 , the adaptive estimator $\hat{f}_A$ has

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

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 2 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 2.

(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 2. 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.

□

Proof of Theorem 5.

Proof. 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 $\frac{\underline{a}}{2\overline{a}}$-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 3. 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 5. The key is to realize that in the derivation of the metric entropy lower bound for $\{\theta : \|\theta\|_q \le t_n\}$ in [45], 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 5 readily gives the desired result.

(ii) As in the proof of Theorem 3, 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 3 (case 2), we know the local entropy $M^{loc}_{\underline{a}/(\sqrt{2}\overline{a})}(\epsilon)$ 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 3 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 $\|\theta\|_0 \leq k_n$, $\|\theta\|_q \leq k_n^{1/q-1/2}\|\theta\|_2 \leq k_n^{1/q-1/2}\sqrt{Ck_n\left(1+\log\frac{M_n}{k_n}\right)/n}$ for θ with $\|\theta\|_2 \leq \sqrt{Ck_n\left(1+\log\frac{M_n}{k_n}\right)/n}$. 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.

□

Proof of Theorem 6.

Proof. (i) We only need to derive the lower bound part. Under the assumptions that $\sup_j \|X_j\|_\infty \leq L_0 < \infty$ for some constant $L_0 > 0$, for a fixed $t_n = t > 0$, we have $\forall f_\theta \in \mathcal{F}_q(t_n; M_n)$, $\|f_\theta\|_\infty \leq \sup_j \|X_j\|_\infty \cdot \sum_{j=1}^{M_n} |\theta_j| \leq L_0\|\theta\|_1 \leq L_0\|\theta\|_q \leq L_0 t$. Then the conclusion follows directly from Theorem 5 (Part (i)). Note that when t_n is fixed, the case $m_* = 1$ needs not to be separately considered.

(ii) For the upper rate part, we use the **AC-C** upper bound. For f_θ with $\|\theta\|_\infty \leq L_0$, clearly, we have $\|\theta\|_1 \leq M_n L_0$, 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 Theorem 4. Under the assumptions that $\sup_j \|X_j\|_\infty \leq L_0 < \infty$ and $k_n\sqrt{\left(1+\log\frac{M_n}{k_n}\right)/n} \leq \sqrt{K_0}$, we know that $\forall f_\theta \in \mathcal{F}_0(k_n; M_n)\bigcap\{f_\theta : \|\theta\|_2 \leq a_n\}$ 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

$$\|\sum_{j=1}^{M_n}\theta_j x_j\|_\infty \leq L_0\|\theta\|_1 \leq L_0\sqrt{k_n}a_n = CL_0k_n\sqrt{\frac{1+\log\frac{M_n}{k_n}}{n}} \leq C\sqrt{K_0}L_0.$$

Then the functions in $\mathcal{F}_0(k_n; M_n)\bigcap\{f : \|\theta\|_2 \leq a_n\}$ have sup-norm uniformly bounded. Note that for bounded a_n, $\|\theta\|_2 \leq a_n$ implies that $\|\theta\|_\infty \leq a_n$. Thus, the extra restriction $\|\theta\|_\infty \leq L_0$ does not affect the minimax lower rate established in part (ii) of Theorem 5.

(iii) The upper and lower rates follow similarly from Theorems 4 and 5. The details are thus skipped.

□

Now we turn to the setup in Section 5 with σ^2 known.

Proposition 6. *(Yang [65], Theorem 1) When $\lambda \geq 5.1\log 2$, we have*

$$E(ASE(\hat{f}_{\hat{J}})) \leq B \inf_{J\in\Gamma_n}\left(\left\|\bar{f}_J - f_0^n\right\|_n^2 + \frac{\sigma^2 r_J}{n} + \frac{\lambda\sigma^2 C_J}{n}\right),$$

where $B>0$ is a constant that depends only on λ.

Proof of Theorem 7.

Proof. The general case (ii) is easily derived based on our estimation procedure and Proposition 6.

To prove (i), when $\mathcal{F} = \mathcal{F}_q(t_n; M_n)$, according to the upper bound in (ii) and Theorem 1, when $f_0^n \in \mathcal{F}_q(t_n; M_n)$, for any $1 \leq m \leq (M_n-1)\wedge n$, there exists a subset J_m and $f_{\theta^m} \in \mathcal{F}_{J_m}$ such that

$$\begin{aligned}
&E(ASE(\hat{f}_{\hat{J}}))\\
\leq\ & B\left(\left(\|f_{\theta^m} - f_0^n\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2\log(M_n\wedge n)}{n} + \frac{\sigma^2\log\binom{M_n}{m}}{n}\right)\wedge\frac{\sigma^2 r_{M_n}}{n}\right)\\
&\wedge B\left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n}\right)\wedge\sigma^2\right)\\
\leq\ & B\left(\left(2^{2/q-1}t_n^2 m^{1-2/q} + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2\log(M_n\wedge n)}{n} + \frac{\sigma^2\log\binom{M_n}{m}}{n}\right)\wedge\frac{\sigma^2 r_{M_n}}{n}\right)\\
&\wedge B\left(\left(t_n^2 + \frac{\sigma^2}{n}\right)\wedge\sigma^2\right).
\end{aligned}$$

Since $\log\binom{M_n}{m} \leq m\left(1+\log\frac{M_n}{m}\right)$ and $\log M_n \leq m\left(1+\log\frac{M_n}{m}\right)$, then for models with size $1\leq m\leq (M_n-1)\wedge n$, we have

$$\begin{aligned}
E(ASE(\hat{f}_{\hat{J}})) \leq\ & B'\left(\left(t_n^2 m^{1-2/q} + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 m\left(1+\log\frac{M_n}{m}\right)}{n}\right)\wedge\frac{\sigma^2 r_{M_n}}{n}\right)\\
&\wedge B\left(\left(t_n^2 + \frac{\sigma^2}{n}\right)\wedge\sigma^2\right),
\end{aligned}$$

where B' only depends on q and λ.

When $m_* = m^* = M_n \wedge n$, the full model J_{M_n} leads to an upper bound of order $\frac{\sigma^2 r_{M_n}}{n}$. When $1 < m_* \leq m^* < M_n \wedge n$, we get the desired upper bounds by evaluating the risk bounds choosing J_{m_*} and J_{M_n}. When $m_* = 1$, models J_0 and J_{M_n} result in the desired upper bound.

The arguments for cases $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ and $\mathcal{F} = \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)$ are similar to those of Theorem 2 and above with r_{J_m} replacing m in the upper bounds.

□

Proof of Theorem 8.

Proof. Without loss of generality, assume the error variance $\sigma^2 = 1$. Let $P_f(y^n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{1}{2}(y_i - f(\mathbf{x}_i)^2)\big)$ denote the joint density of $Y^n = (Y_1, ..., Y_n)'$, where the components are independent with mean $f(\mathbf{x}_i)$ and variance 1, $1 \leq i \leq n$. Then the Kullback-Leibler distance between $P_{f_1}(y^n)$ and $P_{f_2}(y^n)$ is

$$D\left(P_{f_1}(y^n) \parallel P_{f_2}(y^n)\right) = \frac{1}{2}\sum_{i=1}^n \left(f_1(\mathbf{x}_i) - f_2(\mathbf{x}_i)\right)^2.$$

To prove the lower bounds, instead of the global L_2 distance on the regression functions, we need to work with the distance $d(f_1, f_2) = \sqrt{\sum_{i=1}^n \left(f_1(\mathbf{x}_i) - f_2(\mathbf{x}_i)\right)^2}$.

First consider the case $\mathcal{F} = \mathcal{F}_q(t_n; M_n)$. Let $B_k(\eta) = \{\theta : \|\theta\|_2 \leq \eta, \|\theta\|_0 \leq k\}$ and $\mathcal{B}_k(f_0; \epsilon) = \{f_\theta : \|f_\theta\|_n \leq \epsilon, \|\theta\|_0 \leq k\}$ $(f_0 = 0)$. Then, under Assumption SRC$'$ with $\gamma = k$, the $\frac{\underline{a}}{2\overline{a}}$-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 $\eta = \frac{\epsilon}{\overline{a}}$. When $\gamma = m_*$, the proof is the same as the proof of Theorem 5.

Now consider the case $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ and again assume $k_n \leq M_n/2$ as in the proof of Theorem 5. When Assumption SRC$'$ holds with $\gamma = k_n$, the lower bound is of order $\frac{k_n(1+\log M_n/k_n)}{n}$ as before in the random design case. The proof for the last case $\mathcal{F} = \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)$ is similarly done as in the proof of Theorem 5.

□

Proof of Theorem 9.

Proof. According to Corollary 6 from [46], we have

$$E(ASE(\hat{f}^{MLS})) \leq \inf_{J \in \Gamma_n} \left(\|\bar{f}_J - f_0^n\|_n^2 + \frac{\sigma^2 r_J}{n} + \frac{4\sigma^2 \log(1/\pi_J)}{n} \right),$$

which is basically the same as Proposition 6 with $B = 1$. Thus, the rest of the proof is basically the same as that of Theorem 7.

□

To prove Theorem 10 , we need an oracle inequality, which improves Theorem 4 of [65], where only a convergence in probability result is given. Suppose that only the subset models J_m with rank $r_{J_m} \leq n/2$ are considered (which is automatically satisfied when $M_n \leq n/2$). Let Γ denote these models. (More generally, a risk bound similar to the following holds if we consider models with size no more than $(1-\rho)n$ for any small $\rho > 0$.) Let C_J be the descriptive complexity of the model J in Γ.

Proposition 7. *When $\lambda \geq 40 \log 2$, the selected model $\hat{J}'$ by ABC' satisfies*

$$E(ASE(\hat{f}_{\hat{J}'})) \leq B \inf_{J \in \Gamma} \left(\left\|\bar{f}_J - f_0^n\right\|_n^2 + \frac{\sigma^2 r_J}{n} + \frac{\lambda \overline{\sigma}^2 C_J}{n} \right),$$

where B is a constant that depends on λ, $\overline{\sigma}^2$, and σ^2.

Remark 19. If we add models with rank $r_J > n/2$ into the competition, as long as the complexity assignment over all the models is valid (i.e., satisfying the summability condition), if we can show that for these added models, $ABC'(J)$ are also upper and lower bounded with high probabilities as in (8.5) and (8.6), then the risk bound in the proposition continues to hold.

Proof. Let $e_n = (\varepsilon_1, \ldots, \varepsilon_n)'$. For ease in writing, we simplify $\|\cdot\|_n^2$ to $\|\cdot\|^2$ in this proof. From page 495 in [65], for each candidate model J, we have

$$ABC'(J) = \|A_J f_0^n\|^2 + r_J \left(\frac{2}{n - r_J} \left(\|Y_n - \hat{Y}_J\|^2 + \lambda \overline{\sigma}^2 C_J \right) - \sigma^2 \right) + \lambda \overline{\sigma}^2 C_J + 2\mathrm{rem}_1(J) + \mathrm{rem}_2(J),$$

where $\|A_J f_0^n\|^2 = \|\overline{f}_J - f_0^n\|^2$, $\mathrm{rem}_1(J) = e_n'(f_0^n - M_J f_0^n)$ and $\mathrm{rem}_2(J) = r_J - e_n' M_J e_n$. Note also that $\|Y_n - \hat{Y}_J\|^2 + \lambda\overline{\sigma}^2 C_J = \|A_J f_n\|^2 + (n - r_J)\sigma^2 + \big(e_n' A_J e_n - (n - r_J)\sigma^2\big) + 2e_n' A_J f_n + \lambda\overline{\sigma}^2 C_J$. Let

$$T(J) = \|A_J f_0^n\|^2 + (n - r_J)\sigma^2 + \lambda\overline{\sigma}^2 C_J, \text{ and } nR_n(J) = \|A_J f_0^n\|^2 + r_J\sigma^2 + \lambda\overline{\sigma}^2 C_J.$$

As is shown in the proof of Theorem 1, [65], if $\lambda > h(\tau_1, \tau_2) = \max(\sup_{\xi \geq 0}((2(\log 2)\xi)^{1/2}/\tau_1 - \xi), \sup_{\rho \geq 0}(\rho/\tau_2 - 1)2(\log 2)/(\rho - \log(\rho + 1)))$ for some constants τ_1 and τ_2 with $2\tau_1 + \tau_2 < 1$, then for any $\delta > 0$, with probability no less than $1 - 5\delta$, $|\mathrm{rem}_1(J)| \leq \tau_1(nR_n(J) + g_1(\delta))$, $|\mathrm{rem}_2(J)| \leq \tau_2(nR_n(J) + g_2(\delta))$, and $|e_n' A_J e_n - (n - r_J)\sigma^2| \leq \tau_2(T(J) + g_2(\delta))$, where $g_1(\delta) = g_2(\delta) = \lambda \log_2(1/\delta)$. Then with probability no less than $1 - 5\delta$, we have

$$\begin{aligned}
ABC'(J) \;\geq\;& \|A_J f_0^n\|^2 + r_J\left(\frac{2(T(J) - \tau_2(T(J) + g_2(\delta)) - 2\tau_1(nR_n(J) + g_1(\delta)))}{n - r_J} - \sigma^2\right) \\
& -2\tau_1(nR_n(J) + g_1(\delta)) - \tau_2(nR_n(J) + g_2(\delta)) + \lambda\overline{\sigma}^2 C_J \\
\geq\;& \|A_J f_0^n\|^2 + r_J\left(\frac{2(1 - (2\tau_1 + \tau_2))T(J)}{n - r_J} - \frac{2(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{n - r_J} - \sigma^2\right) \\
& -(2\tau_1 + \tau_2)nR_n(J) - (2\tau_1 g_1(\delta) + \tau_2 g_2(\delta)) + \lambda\overline{\sigma}^2 C_J \\
\geq\;& \|A_J f_0^n\|^2 + r_J(1 - (4\tau_1 + 2\tau_2))\sigma^2 - \frac{2r_J(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{n - r_J} \\
& -(2\tau_1 + \tau_2)nR_n(J) - (2\tau_1 g_1(\delta) + \tau_2 g_2(\delta)) + \lambda\overline{\sigma}^2 C_J \\
\geq\;& (1 - (6\tau_1 + 3\tau_2))nR_n(J) - \frac{n + r_J}{n - r_J}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta)). \qquad (8.5)
\end{aligned}$$

Suppose $6\tau_1 + 3\tau_2 < 1$. Let J_n be the candidate model that minimizes $R_n(J)$. Then with exception probability less than 5δ, we have

$$\begin{aligned}
ABC'(J_n) \;\leq\;& \|A_{J_n} f_0^n\|^2 + r_{J_n}\left(\frac{2(1 + (2\tau_1 + \tau_2))T(J_n)}{n - r_{J_n}} - \sigma^2\right) + (2\tau_1 + \tau_2)nR_n(J_n) \\
& + \frac{n + r_{J_n}}{n - r_{J_n}}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta)) + \lambda\overline{\sigma}^2 C_{J_n}.
\end{aligned}$$

Since $T(J_n)/(n - r_{J_n}) = (1 + r_{J_n}/(n - r_{J_n}))R_n(J_n) + (1 - r_{J_n}/(n - r_{J_n}))\sigma^2 \leq 2R_n(J_n) + \sigma^2$, then

$$ABC'(J_n) \leq (5 + 14\tau_1 + 7\tau_2)nR_n(J_n) + \frac{n + r_{J_n}}{n - r_{J_n}}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta)). \qquad (8.6)$$

Thus, for any $\delta > 0$, when the sample size is large enough, we have that with probability no less than $1-5\delta$,

$$\begin{aligned}
nR_n(\hat{J}') &\leq \frac{ABC'(\hat{J}') + \frac{n+r_{\hat{J}'}}{n-r_{\hat{J}'}}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{1-(6\tau_1+3\tau_2)} \\
&\leq \frac{ABC'(J_n) + \frac{n+r_{\hat{J}'}}{n-r_{\hat{J}'}}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{1-(6\tau_1+3\tau_2)} \\
&\leq \frac{(5+14\tau_1+7\tau_2)nR_n(J_n) + \frac{n+r_{J_n}}{n-r_{J_n}}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta)) + \frac{n+r_{\hat{J}'}}{n-r_{\hat{J}'}}(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{1-(6\tau_1+3\tau_2)}.
\end{aligned}$$

Thus, with probability at least $1-5\delta$,

$$\begin{aligned}
R_n(\hat{J}')/R_n(J_n) &\leq \frac{5+14\tau_1+7\tau_2}{1-(6\tau_1+3\tau_2)} + \frac{\left(\frac{n+r_{J_n}}{n-r_{J_n}} + \frac{n+r_{\hat{J}'}}{n-r_{\hat{J}'}}\right)(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{(1-(6\tau_1+3\tau_2))nR_n(J_n)} \\
&\leq \frac{5+14\tau_1+7\tau_2}{1-(6\tau_1+3\tau_2)} + \frac{\left(\frac{n+r_{J_n}}{n-r_{J_n}} + \frac{n+r_{\hat{J}'}}{n-r_{\hat{J}'}}\right)(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{(1-(6\tau_1+3\tau_2))\sigma^2} \\
&\leq \frac{5+14\tau_1+7\tau_2}{1-(6\tau_1+3\tau_2)} + \frac{6(2\tau_1 g_1(\delta) + \tau_2 g_2(\delta))}{(1-(6\tau_1+3\tau_2))\sigma^2}.
\end{aligned}$$

Let

$$\widetilde{W} = b_n^{-1}\left(\frac{R_n(\hat{J}')}{R_n(J_n)} - \frac{5+14\tau_1+7\tau_2}{1-(6\tau_1+3\tau_2)}\right) \text{ and } b_n = \frac{6(2\tau_1+\tau_2)\lambda}{(1-(6\tau_1+3\tau_2))\sigma^2}.$$

Then $P\left(\widetilde{W} \geq -\log_2\delta\right) \leq 5\delta$ for $0<\delta<1$. Since $E(\widetilde{W}^+) = \int_0^\infty P(\widetilde{W} \geq t)dt \leq 5\int_0^\infty 2^{-t}dt = 5/\ln 2$ and $R_n(J_n) \leq (\overline{\sigma}^2/\sigma^2)\inf_{J\in\Gamma} R_n(f_0;J)$ where $R_n(f_0;J) = \|\overline{f}_J - f_0^n\|_n^2 + r_J\sigma^2/n + \lambda\sigma^2 C_J/n$, then we have

$$\begin{aligned}
\frac{E\left(R_n(\hat{J}')\right)}{\inf_{J\in\Gamma} R_n(f_0;J)} &= \frac{E\left(R_n(\hat{J}')\right)}{R_n(J_n)} \cdot \frac{R_n(J_n)}{\inf_{J\in\Gamma} R_n(f_0;J)} \\
&\leq \left(\frac{5+14\tau_1+7\tau_2}{1-(6\tau_1+3\tau_2)} + \frac{30(2\tau_1+\tau_2)\lambda}{(\ln 2)(1-(6\tau_1+3\tau_2))\sigma^2}\right)\cdot\left(\frac{\overline{\sigma}^2}{\sigma^2}\right).
\end{aligned}$$

So $E(ASE(\hat{f}_{\hat{J}'})) \leq B\inf_{J\in\Gamma} R_n(f_0;J)$, where the constant B depends on τ_1, τ_2, $\overline{\sigma}$, and σ. Minimizing $h(\tau_1,\tau_2)$ over $\tau_1>0$ and $\tau_2>0$ in the region $6\tau_1+3\tau_2<1$, one finds a minimum value less than $40\log 2$. Thus, the results of the theorem hold when $\lambda \geq 40\log 2$.

□

Proposition 7 may not provide optimal risk rate when r_{M_n} is small, or when r_{M_n} is larger than $n/2$ (in which case the risk bound on $E(ASE(\hat{J}'))$ can be arbitrarily large because the approximation errors can be arbitrarily large when the models are restricted to be of size $n/2$ or smaller). The issue can be resolved by considering the full model J_{M_n} and the full projection model $\bar{J}$ in the candidate model list, as described before Theorem 10 .

Proof of Theorem 10 :

Proof. Observe that for the full projection model $\bar{J}$, with the chosen $C_{\bar{J}}$, we have that

$$(1-(6\tau_1+3\tau_2))nR_n(\bar{J}) \le ABC'(\bar{J}) \le \xi nR_n(\bar{J}) = \xi\left(n\sigma^2+\lambda\overline{\sigma}^2 C_{\bar{J}}\right)$$

for some constant $\xi>0$ that depends only on λ, $\overline{\sigma}^2$ and σ^2. From the remark after Proposition 7, we have the following risk bounds for the three situations. Below B and B' are constants depending only on λ, $\overline{\sigma}^2$, and σ^2.

1. When $M_n \le n/2$, we have the general risk bound

$$\begin{aligned}
& E(ASE(\hat{f}_{\hat{J}'})) \\
\le\ & B'\left(\inf_{J_m:1\le m<M_n}\left(\left\|\bar{f}_{J_m}-f_0^n\right\|_n^2+\frac{\sigma^2 r_{J_m}}{n}+\frac{\lambda\sigma^2 C_{J_m}}{n}\right)\wedge\left(\left\|\bar{f}_{J_{M_n}}-f_0^n\right\|_n^2+\frac{\sigma^2 r_{M_n}}{n}\right)\right. \\
& \left.\wedge R_n(\bar{J})\wedge R_n(J_0)\right) \\
\le\ & B'\left(\|\bar{f}_{J_{M_n}}-f_0^n\|_n^2+\inf_{J_m:1\le m<M_n}\left(\|\bar{f}_{J_m}-\bar{f}_{J_{M_n}}\|_n^2+\frac{\sigma^2 r_{J_m}}{n}+\frac{\sigma^2\log(M_n-1)}{n}\right.\right. \\
& \left.\left.+\frac{\sigma^2\log\binom{M_n}{m}}{n}\right)\wedge\frac{\sigma^2 r_{M_n}}{n}\right)\wedge B'\left(\left(\|\bar{f}_{J_0}-f_0^n\|_n^2+\frac{\sigma^2}{n}\right)\wedge\sigma^2\right).
\end{aligned}$$

For $f_0^n \in \mathcal{F}_q(t_n;M_n)$, from above, by an argument similar to that in Theorem 7, for any $1\le m<M_n$, there exists a subset J_m and $f_{\theta^m}\in\mathcal{F}_{J_m}$ such that

$$\begin{aligned}
E(ASE(\hat{f}_{\hat{J}'})) \ \le\ & B'\left(\left(t_n^2 m^{1-2/q}+\frac{\sigma^2 r_{J_m}}{n}+\frac{\sigma^2 m\left(1+\log\frac{M_n}{m}\right)}{n}\right)\wedge\frac{\sigma^2 r_{M_n}}{n}\right) \\
& \wedge B'\left(\left(t_n^2+\frac{\sigma^2}{n}\right)\wedge\sigma^2\right). \qquad (8.7)
\end{aligned}$$

When $m_*=m^*=M_n$, the full model J_{M_n} leads to an upper bound of order $\frac{\sigma^2 r_{M_n}}{n}$. When $1<m_*<M_n$, we get the desired upper bound by taking the smaller value of the index of resolvability at J_{m_*} and J_{M_n}. When $m_*=1$, the smaller value of the index of resolvability at J_0 and J_{M_n} results in the given upper bound.

The arguments for cases $\mathcal{F} = \mathcal{F}_0(k_n; M_n)$ and $\mathcal{F} = \mathcal{F}_q(t_n; M_n) \cap \mathcal{F}_0(k_n; M_n)$ are similar to those of Theorem 7.

2. When $M_n > n/2$ and $r_{M_n} \geq n/2$, evaluating the index of resolvability gives

$$
\begin{aligned}
& E(ASE(\hat{f}_{\hat{J}'})) \\
\leq\ & B\left(\inf_{J_m: 1\leq m\leq n/2}\left(\left\|\bar{f}_{J_m} - f_0^n\right\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\lambda\sigma^2 C_{J_m}}{n}\right) \wedge R_n(\bar{J}) \wedge R_n(J_0)\right) \\
\leq\ & B' \inf_{J_m: 1\leq m\leq n/2}\left(\|\bar{f}_{J_m} - f_0^n\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log\lfloor n/2\rfloor}{n} + \frac{\sigma^2 \log\binom{M_n}{m}}{n}\right) \\
& \wedge B'\left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n}\right) \wedge \sigma^2\right).
\end{aligned}
$$

In this case, for the full model, clearly, we have $\|\bar{f}_{J_{M_n}} - f_0^n\|_n^2 + \frac{\sigma^2 r_{M_n}}{n} \geq \frac{1}{2}\sigma^2$, which cannot be better than the model $\bar{J}$ up to a constant factor. We next show that adding the models with size $n/2 < m < M_n$ does not help either in terms of the rate in the risk bound. If $r_{J_m} \geq r_{M_n}/2$, then obviously $\|\bar{f}_{J_m} - f_0^n\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log\lfloor n/2\rfloor}{n} + \frac{\sigma^2 \log\binom{M_n}{m}}{n} \geq \frac{1}{4}\sigma^2$. For $r_{J_m} < r_{M_n}/2$, if $n/2 < m \leq M_n/2$, then there exists a smaller model with size $\widetilde{m} \leq n/2$ that has the same approximation error and rank, but smaller complexity $C_{J_{\widetilde{m}}}$ (i.e., $C_{J_{\widetilde{m}}} \leq C_{J_m}$), where $C_{J_m} = \log(n \wedge M_n) + \log\binom{M_n}{m}$ when $m > n/2$. If $m > M_n/2$ (and $r_{J_m} < r_{M_n}/2$), then due to the monotonicity of the function $\binom{M_n}{m}$ in $m \geq M_n/2$, since there must be more than $r_{M_n}/2$ terms left out in the model, we must have $\log\binom{M_n}{m} \geq \log\binom{M_n}{M_n - \lfloor r_{M_n}/2\rfloor} \geq \lfloor r_{M_n}/2\rfloor \log\frac{M_n}{\lfloor r_{M_n}/2\rfloor}$, which is at least of order n under the condition $r_{M_n} \geq n/2$. Putting the above facts together, we conclude that adding the models with size $n/2 < m \leq M_n$ does not affect the validity of the risk bound given in part (ii) of Theorem 10 (note that $\log\lfloor n/2\rfloor$ is of the same order as $\log(M_n \wedge n)$ in our case). Then, the general risk upper bound becomes (with B' enlarged by an absolute constant factor)

$$\begin{aligned}
& B' \inf_{J_m:1\le m<M_n} \left(\|\bar{f}_{J_m} - f_0^n\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log(M_n \wedge n)}{n} + \frac{\sigma^2 \log\binom{M_n}{m}}{n} \right) \\
& \wedge B' \left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n} \right) \wedge \sigma^2 \right) \wedge B' \left(\|\bar{f}_{J_{M_n}} - f_0^n\|_n^2 + \frac{\sigma^2 r_{M_n}}{n} \right) \\
\le \; & B' \left(\|\bar{f}_{J_{M_n}} - f_0^n\|_n^2 + \inf_{J_m:1\le m<M_n} \left(\|\bar{f}_{J_m} - \bar{f}_{J_{M_n}}\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log(M_n \wedge n)}{n} \right.\right. \\
& \left.\left. + \frac{\sigma^2 \log\binom{M_n}{m}}{n} \right) \wedge \frac{\sigma^2 r_{M_n}}{n} \right) \wedge B' \left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n} \right) \wedge \sigma^2 \right).
\end{aligned}$$

For $f_0^n \in \mathcal{F}_q(t_n; M_n)$ and any $1 \le m < M_n$, there exists a subset J_m and $f_{\theta^m} \in \mathcal{F}_{J_m}$ such that the inequality (8.7) holds. When $m_* = m^* = M_n \wedge n$, the full projection model $\bar{J}$ leads to an upper bound of order σ^2. When $1 < m_* < M_n \wedge n$, we get the desired upper bounds by choosing J_{m_*} and $\bar{J}$ to evaluate the index of resolvability. When $m_* = 1$, models J_0 and $\bar{J}$ result in the desired upper bound.

3. When $M_n > n/2$ and $r_{M_n} < n/2$, the full model is already included, and, similarly as above, the models with $n/2 < m < M_n$ can be included in the minimization set of the general risk bound. Indeed, if $r_{M_n} = 1$, the statement is trivial. If $r_{J_m} \ge r_{M_n}/2$, then $\|\bar{f}_{J_m} - f_0^n\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log\lfloor n/2\rfloor}{n} + \frac{\sigma^2 \log\binom{M_n}{m}}{n} \ge \left\|\bar{f}_{J_{M_n}} - f_0^n\right\|_n^2 + \frac{\sigma^2 J_{M_n}}{2n}$, which means that the model cannot beat the full model up to a constant factor. For $r_{J_m} < r_{M_n}/2$, if $m > M_n/2$, then we again have $\log\binom{M_n}{m} \ge \log\binom{M_n}{M_n - \lfloor r_{M_n}/2\rfloor} \ge \lfloor r_{M_n}/2\rfloor \log \frac{M_n}{\lfloor r_{M_n}/2\rfloor}$. Thus there exists a model in Γ_n' with the same rank of $r_{J_m} \le n/2$ and approximation error, and its complexity is at most at the same order as J_m. Then with the same arguments for the case of $r_{M_n} \ge n/2$, we again conclude that adding the models with size $n/2 < m \le M_n$ does not affect the validity of the risk bound

given in part (ii) of Theorem 10. Thus, the general risk bound is

$$
\begin{aligned}
& E(ASE(\hat{f}_{\hat{j}'})) \\
\leq\ & B\left\{\left(\|\bar{f}_{J_{M_n}} - f_0^n\|_n^2 + \inf_{J_m: 1\leq m \leq n/2}\left(\|\bar{f}_{J_m} - \bar{f}_{J_{M_n}}\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{2\lambda\sigma^2 C_{J_m}}{n}\right)\wedge\right.\right. \\
& \left.\left.\frac{\sigma^2 r_{M_n}}{n}\right)\right\} \wedge B\left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n}\right)\wedge \sigma^2\right) \\
\leq\ & B'\left(\|\bar{f}_{J_{M_n}} - f_0^n\|_n^2 + \inf_{J_m: 1\leq m < M_n}\left(\|\bar{f}_{J_m} - \bar{f}_{J_{M_n}}\|_n^2 + \frac{\sigma^2 r_{J_m}}{n} + \frac{\sigma^2 \log(M_n \wedge n)}{n}\right.\right. \\
& \left.\left. + \frac{\sigma^2 \log\binom{M_n}{m}}{n}\right) \wedge \frac{\sigma^2 r_{M_n}}{n}\right) \wedge B'\left(\left(\|\bar{f}_{J_0} - f_0^n\|_n^2 + \frac{\sigma^2}{n}\right)\wedge\sigma^2\right).
\end{aligned}
$$

For $f_0^n \in \mathcal{F}_q(t_n; M_n)$ and any $1 \leq m < M_n$, there exists a subset J_m and $f_{\theta^m} \in \mathcal{F}_{J_m}$ such that the inequality (8.7) holds. When $m_* = m^* = M_n \wedge n$, the full model J_{M_n} leads to an upper bound of order $\frac{\sigma^2 r_{M_n}}{n}$. When $1 < m_* < M_n \wedge n$, we get the desired upper bounds by choosing J_{m_*} and J_{M_n} when evaluating the index of resolvability. When $m_* = 1$, taking models J_0 and J_{M_n} results in the desired upper bound.

□

Proof of Theorem 11:

Proof. The proof is similar to that of Theorem 10 except that we use the oracle inequality (4.7) in [7] instead of that in Proposition 7 (and there is no need to consider the different scenarios). Note that if $M_n \leq (n-7) \wedge \varsigma n$, then $m \vee \log\binom{M_n}{m} < \varsigma n$ for all $1 \leq m \leq M_n$. Thus all subset models are allowed by the BGH criterion. When M_n is larger, however, the conditions required in Corollary 1 of [7] may invalidate the choice of m_* or k_n when it is too large, hence the upper bound assumption on m_* and k_n. We skip the details of the proof.

□

Acknowledgments

We thank Yannick Baraud for helpful discussions and for pointing out the work of Audibert and Catoni (2010) that helped us to remove a logarithmic factor in some of our previous aggregation risk bounds and for his suggestion of handling the case of fully unknown error variance for fixed design. Sandra Paterlini conducted part of this research while visiting the School of Mathematics, University of Minnesota.

References

[1] Akaike, H. (1970) Statistical prediction identification. *Ann. Inst. Statist. Math.* **22**, 203-217.

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

[3] Audibert, J.-Y. (2009) Fast learning rates in statistical inference through aggregation. *Annals of Statistics* **37** 1591-1646.

[4] Audibert, J.-Y and Catoni, O. (2010) Risk bounds in linear regression through PAC-Bayesian truncation. Available at arXiv:1010.0072.

[5] Baraud, Y. (2000) Model selection for regression on a fixed design. *Probability Theory Related Fields* **117** 467-493.

[6] Baraud, Y. (2002) Model selection for regression on a random design. *ESAIM Probab.Statist.* **6** 127-146.

[7] Baraud, Y., Giraud, C. and Huet, S. (2009) Gaussian model selection with an unknown variance. *Ann. Statist.* **37** 630-672.

[8] Barron, A. R. (1993) Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39** 930-945.

[9] Barron, A .R. (1994) Approximation and estimation bounds for artificial neural networks. *Machine Learning* **14** 115-133.

[10] Barron, A. R. and Cover, T. M. (1991) Minimum complexity density estimation. *IEEE Transactions on Information Theory* **37** 1034-1054.

[11] Barron, A. R., Birgé, L. and Massart, P. (1999) Risk bounds for model selection via penalization. *Probability Theory and Related Fields* **113** 301-413.

[12] Barron, A. R., Cohen, A., Dahmen, W., and Devore, R. (2008). Approximation and learning by greedy algorithms. *Annals of Statistics* **36**, 64-94.

[13] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37** 1705-1732.

[14] Birgé, L. (1986) On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Related Fields* **97** 113-150.

[15] Birgé, L. (2004) Model selection for Gaussian regression with random design. *Bernoulli* **10** 1039-1051.

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

[17] Birgé, L. and Massart, P. (1998) Minimum Contrast Estimators on Sieves: Exponential Bounds and Rates of Convergence. *Bernoulli* **4** 329-375.

[18] Birgé, L. and Massart, P. (2001) Gaussian model selection. *Journal of European Math. Society* **3** 203-268.

[19] Bunea, F., Tsybakov, A.B. and Wegkamp, M. H. (2007) Aggregation for Gaussian regression. *Annals of Statistics* **35** 1674-1697.

[20] Bunea, F. and Nobel, A. (2008) Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Transactions on Information Theory* **54** 1725 - 1735.

[21] Candes, E. and Tao, T. (2007) The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* **35** 2313-2351.

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

[23] Catoni, O. (1999) 'Universal' aggregation rules with exact bias bounds. Preprint 510, Laboratoire de Probabilitès et Modèles Aléatoires, Univ. Paris 6 and Paris 7. Available at

http://www.proba.jussieu.fr/mathdoc/preprints/index.html#1999.

[24] Catoni, O. (2004) *Statistical Learning Theory and Stochastic Optimization. Ecole d'Eté de Probabilités de Saint-Flour XXXI–2001, Lectures Notes in Mathematics* **1851**, Springer, New York.

[25] Chen, D., Donoho, D. L. and Saunders, M. A. (1998) Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing* **20** 33-61.

[26] Dalalyan, A. S. and Tsybakov, A. B. (2007) Aggregation by exponential weighting and sharp oracle inequalities. *Lecture Notes in Computer Science* **4539** 97-111.

[27] Dalalyan, A. S. and Tsybakov, A. B. (2009) Sparse regression learning by aggregation and Langevin Monte Carlo. Available at ArXiv:0903.1223.

[28] Donoho, L. (1993) Unconditional Bases are optimal bases for data compression and for statistical estimation. *Appl. Computational Harmonic analysis* **1** 100-115.

[29] Donoho, D. L. (2006) Compressed sensing. *IEEE Transactions on Information Theory* **52** 1289-1306.

[30] Donoho, D. L. and Johnstone, I. M. (1994) Minimax risk over ℓ_p-balls for ℓ_q-error. *Prob. Theory and Related Fields* **99** 277-303.

[31] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996) Density estimation by wavelet thresholding. *Annals of Statistics* **24** 508-539.

[32] Edmunds, D. E. and Triebel H. (1998) Function spaces, entropy numbers, and differential operators. *Cambridge Tracts in Mathematics* **120**. Cambridge University Press.

[33] Gasso, G., Rakotomamonjy, A. and Canu, S. (2009) Recovering sparse signals with non-convex penalties and DC programming. *IEEE Trans. Signal Processing* **57** 4686-4698.

[34] van de Geer, S.A. and Bühlmann, P. (2009) On the conditions used to prove oracle results for the Lasso. *Electron. J. Statist.* **3** 1360-1392.

[35] Giraud, C. (2008) Mixing least-squares estimators when the variance is unknown. *Bernoulli* **14** 1089-1107.

[36] Goldenshluger, A. (2009) A universal procedure for aggregating estimators. *Annals of Statistics* **37** 542-568.

[37] Greenshtein, E. and Ritov, Y., (2004) Persistency in high dimensional linear predictor-selection and the virtue of overparametrization. *Bernoulli* **10** 971-988.

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

[39] Györfi, L. and Ottucsák, Gy. (2007) Sequential prediction of unbounded stationary time series. *IEEE Transactions on Information Theory* **53** 1866-1872.

[40] Huang, C., Cheang, G. H. L. and Barron, A. R. (2008) Risk of penalized least squares, greedy selection and L_1-penalization for flexible function libraries. Manuscript.

[41] Ing, C-K. (2010) On the optimal convergence rate of orthogonal greedy algorithms in high-dimensional sparse regression models. Presented at *The Eighth ICSA International Conference: Frontiers of Interdisciplinary and Methodological Statistical Research*, Guangzhou, China, December 19-22, 2010.

[42] Juditsky, A. and Nemirovski, A. (2000) Functional aggregation for nonparametric estimation. *Annals of Statistics* **28** 681-719.

[43] Koltchinskii, V. (2009) The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799-828.

[44] Koltchinskii, V. (2009) Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.* **45** 7-57.

[45] Kühn, T. (2001) A lower estimate for entropy numbers. *Journal of Approximation Theory* **110** 120-124.

[46] Leung, G. and Barron, A. R. (2006) Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* **52** 3396-3410.

[47] Lecué, G. and Mendelson S. (2010) Optimality of the aggregate with exponential weights for low temperatures. Manuscript.

[48] Li, K.C. (1987) Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: discrete index set. *Annals of Statistics* **15** 958-975.

[49] Lounici, K. (2007) Generalized mirror averaging and D-convex aggregation. *Mathematical methods of statistics* **16** 246-259.

[50] MEINSHAUSEN, N. and BÜHLMANN, P. (2006) High-dimensional graphs and variable selection with the lasso. *Annals of Statistics* **34** 1436-1462.

[51] MEINSHAUSEN, N. and YU, B. (2009) Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* **37** 246-270.

[52] NEMIROVSKI, A. (2000) *Topics in Non-parametric Statistics.* In: P. Bernard, editor, *Ecole d'Eté de Probabilités de Saint-Flour 1998* volume XXVIII of *Lecture Notes in Mathematics* **1738** Springer, New York.

[53] RASKUTTI, G., WAINWRIGHT, M. and YU, B. (2011) Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory* **57** 6976-6994.

[54] RIGOLLET, P. and TSYBAKOV, A. B. (2010) Exponential screening and optimal rates of sparse estimation *Annals of Statistics* **39** 731-771.

[55] SHEN, X. AND WONG, W.H. (1994) Convergence rates of sieve estimates. *Annals of Statistics* **22**, 580-615.

[56] SHIBATA, R. (1981) An optimal selection of regression variables. *Biometrika* **68** 45-54.

[57] TIBSHIRANI, R. (1996) Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B* **58** 267-288.

[58] TSYBAKOV, A.B. (2003) Optimal rates of aggregation. In: *Proceedings of 16th Annual Conference on Learning Theory (COLT)* and *7th Annual Workshop on Kernel Machines*. Lecture Notes in Artificial Intelligence, **2777** 303-313. Springer-Verlag, Heidelberg.

[59] VAN DE GEER, S.A. (1995) The method of sieves and minimum contrast estimators. *Math. Methods Statist.* **4**, 20-38.

[60] VAN DE GEER, S.A. (2008) High-dimensional generalized linear models and the Lasso. *Annals of Statistics* **36** 614-645.

[61] VERZELEN, N. (2010) Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. Arxiv preprint arXiv:1008.0526.

[62] WAINWRIGHT, M.J. (2009) Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Information Theory* **55**

2183-2202.

[63] Wegkamp, M. (2003) Model selection in nonparametric regression. *Annals of Statistics* **31** 252-273.

[64] Yang, Y. (1996) "Minimax optimal density estimation", Ph.D. Dissertation, Department of Statistics, Yale University, May 1996.

[65] Yang, Y. (1999) Model selection for nonparametric regression. *Statistica Sinica*, **9** 475-499.

[66] Yang, Y. (2000) Mixing strategies for density estimation. *Annals of Statistics*, **28** 75-87.

[67] Yang, Y. (2000) Combining different procedures for adaptive regression. *Journal of Multivariate Analysis* **74** 135-161.

[68] Yang, Y. (2001) Adaptive regression by mixing. *Journal of American Statistical Association* **96** 574-588.

[69] Yang, Y. (2004) Aggregating regression procedures to improve performance. *Bernoulli* **10** 25-47. An older version of the paper is Pre-print #1999-17 of Department of Statistics at Iowa State University.

[70] Yang, Y. (2008) Localized model selection for regression. *Econometric Theory* **24**, 472-492.

[71] Yang, Y. and Barron, A.R. (1998) An asymptotic property of model selection criteria. *IEEE Trans. Inform. Theory* **44** 95-116.

[72] Yang, Y. and Barron, A.R. (1999) Information theoretic determination of minimax rates of convergence. *Annals of Statistics*, **27** 1564-1599.

[73] Ye, F. and Zhang, C.H. (2010) Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research* **11** 3519-2540.

[74] Yuan, Z. and Yang, Y. (2005) Combining linear regression models: When and how? *Journal of the American Statistical Association* **100** 1202-1214.

[75] Zhang, C.H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** 894-942.

[76] Zhang, C.H. and Huang, J. (2008) The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics* **36** 1567-1594.

[77] Zhang, T. (2009) Some sharp performance bounds for least squares regression with L_1 regu-

larization. *Annals of Statistics* **37** 2109-2144.

[78] Zhang, T. (2010) Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research* **11** 1081-1107.

"Materiali di Discussione" LATER PUBLISHED ELSEWHERE

N. 546 - M. Murat and B. Pistoresi, *Emigrants and immigrants networks in FDI*, Applied Economics letters, April 2008, **http://www.informaworld.com/ /content~content=a789737803~db=all~order=author** (electronic publication), **WP No. 546 (December 2006).**

N. 545 - M.Brunetti and C. Torricelli, *The Population Ageing in Italy:Facts and Impact on Household Portfolios*, in M. Balling & E. Gnan & F. Lierman (eds.), Money, Finance and Demography: The Consequences of Ageing, Vienna, Suerf (2007), **WP No. 545 (November 2006).**

N. 532 – M: Montanari, Between European Integration and Regional Autonomy: The Case of Italy from an Economic Perspective, Constitutional Political Economy, Vol. 17, 4, pp. 277-301 (2006), **WP No. 532 (March 2006).**

N. 529 - M. Montanari, *Knocking on the EU's door: the Political Economy of EU-Ukraine Relations*, Journal of Contemporary European Research, Vol.3,1, pp.64-78 (2007), **WP No. 529 (February 2006).**

N. 518 - M.Brunetti and C. Torricelli, *Economic Activity and Recession Probabilities: information content and predictive power of the term spread in Italy*, Applied Economics (2009), **WP No. 518 (December 2005).**

N. 517 - M. Murat and S. Paba (2006), *I distretti industriali tra immigrazioni e internazionalizzazione produttiva*, in B. Quintieri (ed.) I distretti italiani dal locale al globale, Rubbettino (2006), **WP No. 517 (December 2005).**

N. 491 - V. Moriggia, S. Muzzioli and C. Torricelli, *On the no arbitrage condition in option implied trees*, European Journal of Operational Research (2009), **WP No. 491 (May 2005).**

N. 482 - G. Di Lorenzo and G. Marotta, *A less effective monetary transmission in the wake of EMU? Evidence from lending rates passthrough***,** ICFAI Journal of Monetary Economics, Vol. 4, 2, pp. 6-31 (2006), **WP No. 482 (February 2005).**

N. 472 - M.Brunetti and C. Torricelli, *The internal and cross market efficiency in index option markets: an investigation of the Italian market*, Applied Financial Economics, Vol. 17, 1, pp. 25-33 (2007), **WP No. 472 (November 2004).**

N. 466 - G. Marotta, *La finanza del settore non profit tra ritardi nei pagamenti e Basilea 2*, Banca Impresa Società , Vol. XXIV, 1, pp. 35-51 (2005), **WP No. 466 (September 2004).**

N. 453 - Pederzoli and C. Torricelli, *Capital requirements and Business Cycle Regimes: Forward-looking modelling of Default Probabilities* , Journal of Banking and Finance, Vl. 29, 12, pp. 3121-3140 (2005), **WP No. 453 (February 2004).**

N. 448 - V. Moriggia, S. Muzzioli, C. Torricelli, *Call and put implied volatilities and the derivation of option implied trees*, Frontiers In Finance and Economics, vol.4, 1, pp. 35-64 (2007), **WP No. 448 (November 2003).**

N. 436 - M.Brunetti and C. Torricelli, *Put-Call Parity and cross-market efficiency in the Index Options Markets: evidence from the Italian market*, International Review of Financial Analysis, Vl.14, 5, pp. 508-532 (2005), **WP No. 436 (July 2003).**

N. 429 - G. Marotta, *When do trade credit discounts matter? Evidence from Italian Firm-Level Data*, Applied Economics, Vol. 37, 4, pp. 403-416 (2005), **WP No. 429 (February 2003).**

N. 426 - A. Rinaldi and M. Vasta, *The Structure of Italian Capitalism, 1952-1972: New Evidence Using the Interlocking Directorates Technique*, Financial History Review, vol, 12, 2, pp. 173-198 (2005), **WP No. 426 (January 2003).**

N. 417 - A. Rinaldi, *The Emilian Model Revisited: Twenty Years After*, Business History, vol. 47, 2, pp. 244-226 (2005), **WP No. 417 (September 2002).**

N. 375 - G. Marotta, *La direttiva comunitaria contro i ritardi nei pagamenti tra imprese. Alcune riflessioni sul caso italiano*, Banca, Impresa, Società, Vol. XX, 3, pp. 451-71 (2001), **WP No. 375 (September 2001).**

N. 303 - G. Marotta and M. Mazzoli, *Fattori di mutamento nella domanda di prestiti ed effetti sulla trasmissione della politica monetaria*, in P. ALESSANDRINI (ed.) Il sistema finanziario italiano tra globalizzazione e localismo, Bologna, Il Mulino, pp. 223-260 (2001), **WP No. 303 (April 2000)**

N. 131 - G. Marotta, *Does trade credit redistribution thwart monetary policy? Evidence from Italy*, Applied Economics, Vol. 29, December, pp. 1619-29 (1997), **WP No. 131 (1996).**

N. 121 - G. Marotta, *Il credito commerciale in Italia: una nota su alcuni aspetti strutturali e sulle implicazioni di politica monetaria*, L'Industria, Vol. XVIII, 1, pp. 193-210 (1997), **WP No. 121 (1995)**

N. 105 - G. Marotta, *Credito commerciale e "lending view"*, Giornale degli Economisti e Annali di Economia, Vol. LIV, 1-3, gennaio-marzo, pp. 79-102; anche in G. Vaciago (a cura di) *Moneta e finanza*, Bologna, Il Mulino (1995), **WP No. 105 (1994).**

RECENTLY PUBLISHED “Materiali di Discussione”

N. 680- *Credit, Venture Capital and Regional Economic Growth*, by Barbara Pistoresi e Valeria Venturelli [January 2012].

N. 679 - *Gender Health in Italy in a Capability Approach*, by Anita Chiarolanza, Antonella Picchio e Silvia Storchi [January 2012].

N. 679 - *Gender Health in Italy in a Capability Approach*, by Anita Chiarolanza, Antonella Picchio e Silvia Storchi [January 2012].

N. 678 - *Education systems facing the challenges of technological, organizational and social changes: US and Italy in a comparative perspective MIQUA project_Improving the quality of the pre-university education system*, by Piero Cipollone, Eugene García, Paola Mengoli, Mehmet "Dali" Öztürk, Annamaria Poggi, Donatella Poliandri and Margherita Russo [January 2012].

N. 677 - *The new maritime trade world geography. Opportunities for Italy?* by Margherita Paradisi [January 2012].

N. 676 - *Una lettura delle dinamiche dell’industria automobilistica europea in termini di Flying Geese e di Sophistication Index,* by Francesco Silvestri [January 2012]

N. 675 - *Living conditions and quality of work of Personal Care Assistants in a district area of Italy,* by Paola Bertolini, Giovanni Solinas and Roberta Salvioli [January 2012].

N. 674 - *Job Separations, Job Loss and Informality in the Russian Labor Market,* by Hartmut Lehmann, Tiziano Razzolini and Anzelika Zaiceva [January 2012]

N. 673 - *Fuzzy quantities and their ranking: a new proposal*, by Luca Anzilli and Gisella Facchinetti [December 2011].

N. 672 - *The Evolution of the Racial Gap in Education and the Legacy of Slavery*, by Graziella Bertocchi and Arcangelo Dimico [November 2011].

N. 671 - *Are defined contribution pension schemes socially sustainable? A conceptual map from a macroprudential perspective*, by Giuseppe Marotta [November 2011].

N. 670 - *Do immigrant students succeed? Evidence from Italy and France based on PISA 2006*, by Marina Murat [November 2011].

N. 669 - *Assessing the information content of option-based volatility forecasts using fuzzy regression methods,* by Silvia Muzzioli and Bernard De Baets **[**November 2011].

N. 668 - *Self-Commitment-Institutions and Cooperation in Overlapping Generations Games,* by Francesco Lancia and Alessia Russo [November 2011].

N. 667 - *Long-run Welfare under Externalities in Consumption, Leisure, and Production: A Case for Happy Degrowth vs. Unhappy Growth,* by Ennio Bilancini and Simone D’Alessandro [October 2011].

N. 666 - *Exports, imports and growth New evidence on Italy: 1863-2004,* by Barbara Pistoresi and Albero Rinaldi [October 2011].