Optimal Rates of Aggregation

Alexandre B. Tsybakov

Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6, 4 pl. Jussieu,
75252 Paris Cedex 05, France,
tsybakov@ccr.jussieu.fr

Abstract. We study the problem of aggregation of M arbitrary estimators of a regression function with respect to the mean squared risk. Three main types of aggregation are considered: model selection, convex and linear aggregation. We define the notion of optimal rate of aggregation in an abstract context and prove lower bounds valid for any method of aggregation. We then construct procedures that attain these bounds, thus establishing optimal rates of linear, convex and model selection type aggregation.

1 Introduction

Consider the regression model

$$Y_i = f(X_i) + \xi_i, \quad i = 1, \dots, n, \tag{1}$$

where $X_1, \dots, X_n$ are i.i.d. random vectors with values in a Borel subset $\mathcal{X}$ of $\mathbf{R}^d$, ξ_i are i.i.d. zero-mean random variables in $\mathbf{R}$ such that $(\xi_1, \dots, \xi_n)$ is independent of $(X_1, \dots, X_n)$ and $f : \mathcal{X} \to \mathbf{R}$ is an unknown regression function. The problem is to estimate the function f from the data $D_n = ((X_1, Y_1), \dots, (X_n, Y_n))$.

Denote P_f and P^X the probability distributions of D_n and of X_1 respectively. For an estimator $\hat{f}_n$ of f based on the sample D_n, define the L_2-risk

$$R(\hat{f}_n, f) = E_f \|\hat{f}_n - f\|^2$$

where E_f denotes the expectation w.r.t. the measure P_f and, for a Borel function $g : \mathcal{X} \to \mathbf{R}$,

$$\|g\| = \left(\int_{\mathcal{X}} g^2(x) P^X(dx) \right)^{1/2}.$$

Suppose that we have $M \geq 2$ arbitrary estimators $f_{n,1}, \dots f_{n,M}$ of the function f based on the sample D_n. The aim of aggregation is to construct a new estimate of f (called *aggregate*) that mimics in a certain sense the behavior of the best among the estimators $f_{n,j}$. We will consider the following three well-known aggregation problems (cf. Nemirovski (2000)).

Problem (L). (*Linear aggregation.*) Find an aggregate estimator $\tilde{f}_n$ which is at least as good as the best linear combination of $f_{n,1}, \dots, f_{n,M}$, up to a small remainder term, i.e.

$$R(\tilde{f}_n, f) \leq \inf_{\lambda \in \mathbf{R}^M} R(f^*_\lambda, f) + \Delta^{\mathbf{L}}_{n,M}$$

for every f belonging to a large class of functions $\mathcal{F}$, where

$$f_\lambda^* = \sum_{j=1}^{M} \lambda_j f_{n,j}, \quad \lambda = (\lambda_1, \ldots, \lambda_M),$$

and $\Delta_{n,M}^{\mathbf{L}}$ is a remainder term that does not depend on f.

Problem (C). (*Convex aggregation.*) Find an aggregate estimator $\tilde{f}_n$ which is at least as good as the best convex combination of $f_{n,1}, \ldots, f_{n,M}$, up to a small remainder term, i.e.

$$R(\tilde{f}_n, f) \le \inf_{\lambda \in \Lambda^M} R(f_\lambda^*, f) + \Delta_{n,M}^{\mathbf{C}}$$

for every f belonging to a large class of functions $\mathcal{F}$, where

$$\Lambda^M = \left\{ \lambda \in \mathbf{R}^M : \ \lambda_j \ge 0, \sum_{j=1}^{M} \lambda_j \le 1 \right\}$$

and $\Delta_{n,M}^{\mathbf{C}}$ is a remainder term that does not depend on f.

Problem (MS). (*Model selection aggregation.*) Find an aggregate estimator $\tilde{f}_n$ which is at least as good as the best among $f_{n,1}, \ldots, f_{n,M}$, up to a small remainder term, i.e.

$$R(\tilde{f}_n, f) \le \min_{1 \le j \le M} R(f_{n,j}, f) + \Delta_{n,M}^{\mathbf{MS}}$$

for every f belonging to a large class of functions $\mathcal{F}$, where $\Delta_{n,M}^{\mathbf{MS}}$ is a remainder term that does not depend on f.

Clearly,

$$\min_{1 \le j \le M} R(f_{n,j}, f) \ge \inf_{\lambda \in \Lambda^M} R(f_\lambda^*, f) \ge \inf_{\lambda \in \mathbf{R}^M} R(f_\lambda^*, f). \tag{2}$$

The smallest possible remainder terms $\Delta_{n,M}^{\mathbf{L}}$, $\Delta_{n,M}^{\mathbf{C}}$ and $\Delta_{n,M}^{\mathbf{MS}}$ characterize the price to pay for aggregation. We will see that they satisfy a relation that is in a sense inverse to (2): the largest price $\Delta_{n,M}$ is to be paid for linear aggregation and the smallest one for model selection aggregation. Convex aggregation has an intermediate price.

Aggregation of arbitrary estimators for regression with random design (1) under the L_2-risk has been studied by several authors, mostly in the case of model selection (Yang (2000), Catoni (2001), Wegkamp (2000), Györfi, Kohler, Krzyżak and Walk (2002), Birgé (2002)) and convex aggregation (Nemirovski (2000), Juditsky and Nemirovski (2000), Yang (2001)). Linear aggregation for the Gaussian white noise model is discussed by Nemirovski (2000). Aggregation procedures are typically based on sample splitting. The initial sample D_n is divided into two independent subsamples D_m^1 and D_l^2 of sizes m and l respectively where $m \gg l$ and $m + l = n$. The first subsample D_m^1 is used to construct estimators $f_{n,1}, \ldots, f_{n,M}$ and the second subsample D_l^2 is used to aggregate them,

i.e. to construct $\tilde{f}_n$ (thus, $\tilde{f}_n$ is measurable w.r.t. the whole sample D_n). In this paper we will not consider sample splitting schemes but rather deal with an idealized framework (following Nemirovski (2000), Juditsky and Nemirovski (2000)) where the first subsample is fixed and thus instead of the estimators $f_{n,1}, \dots, f_{n,M}$ we have fixed functions $f_1, \dots, f_M$. The problem is to find linear, convex and model selection aggregates of $f_1, \dots, f_M$ based on the sample D_n that would converge with the fastest possible rate (i.e. with the smallest possible remainder terms $\Delta_{n,M}$) in a minimax sense. A partial solution of this problem for the case of convex aggregation has been given by Juditsky and Nemirovski (2000) and Yang (2001). Here we solve the problem for all the three types of aggregation, in particular, improving these results concerning convex aggregation. The main goal of this paper is to find optimal rates of aggregation in the sense of a general definition given below.

2 Main definition and lower bounds

We start with the following definition that covers more general framework than the one considered in the paper.

Definition 1. *Let H be a given abstract index set and let $\mathcal{F}$, $\mathcal{F}'$ be a given classes of Borel functions on $\mathcal{X}$.*

A sequence of positive numbers ψ_n is called **optimal rate of aggregation for** $(H, \mathcal{F}, \mathcal{F}')$ *if*

- *for any family of Borel functions $\{f_\lambda, \lambda \in H\}$ indexed by H and contained in $\mathcal{F}'$ there exists an estimator $\tilde{f}_n$ of f (aggregate) such that*

$$\sup_{f \in \mathcal{F}} \left[R(\tilde{f}_n, f) - \inf_{\lambda \in H} \|f_\lambda - f\|^2 \right] \le C\psi_n, \tag{3}$$

  for some constant $C < \infty$ and any integer n,
  and
- *there exists a family of Borel functions $\{f_\lambda, \lambda \in H\}$ indexed by H and contained in $\mathcal{F}'$ such that for all estimators T_n of f we have*

$$\sup_{f \in \mathcal{F}} \left[R(T_n, f) - \inf_{\lambda \in H} \|f_\lambda - f\|^2 \right] \ge c\psi_n, \tag{4}$$

  for some constant $c > 0$ and any integer n.

In this paper we are interested in the following index sets:

$$H = \begin{cases} \{1, \dots, M\} & \text{for Problem (MS)}, \\ \Lambda^M & \text{for Problem (C)}, \\ \mathbf{R}^M & \text{for Problem (L)}, \end{cases}$$

and we consider $\mathcal{F} = \mathcal{F}_0$ defined by

$$\mathcal{F}_0 = \{f : \ \|f\|_\infty \le L\}, \tag{5}$$

where $\|\cdot\|_\infty$ denotes the L_∞ norm associated with the measure P^X and $L < \infty$ is an unknown constant. We take also $\mathcal{F}' = \mathcal{F}_0$ for Problems (MS) and (C), and $\mathcal{F}' = L_2(\mathcal{X}, P^X)$ for Problem (L).

Optimal rates of aggregation for $(\{1, \dots, M\}, \mathcal{F}_0, \mathcal{F}_0)$, for $(\Lambda^M, \mathcal{F}_0, \mathcal{F}_0)$ and for $(\mathbf{R}^M, \mathcal{F}_0, L_2(\mathcal{X}, P^X))$ will be called for brevity optimal rates of model selection, convex and linear aggregation respectively.

In the rest of the paper the notation f_λ for a vector $\lambda = (\lambda_1, \dots, \lambda_M) \in \mathbf{R}^M$ is understood in the following sense:

$$f_\lambda = \sum_{j=1}^{M} \lambda_j f_j.$$

In this section we prove lower bounds of the type (4) for model selection, convex and linear aggregation. The proofs will be based on the following lemma on minimax lower bounds which can be obtained, for example, by combining Theorems 2.2 and 2.5 in Tsybakov (2003).

Lemma 1. *Let $\mathcal{C}$ be a finite set of functions on $\mathcal{X}$ such that $N = \mathrm{card}(\mathcal{C}) \geq 2$,*

$$\|f - g\|^2 \geq 4\psi_n > 0, \quad \forall\ f, g \in \mathcal{C}, \quad f \neq g,$$

and the Kullback divergences $K(P_f, P_g)$ between the measures P_f and P_g satisfy

$$K(P_f, P_g) \leq (1/16) \log N, \quad \forall\ f, g \in \mathcal{C}.$$

Then

$$\inf_{T_n} \sup_{f \in \mathcal{C}} R(T_n, f) \geq c_1 \psi_n,$$

where $\inf_{T_n}$ denotes the infimum over all estimators and $c_1 > 0$ is a constant.

Throughout the paper we denote by c_i finite positive constants. Introduce the following assumptions.

(A1) The errors ξ_i are i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ random variables, $0 < \sigma < \infty$.

(A2) There exists a cube $S \subset \mathcal{X}$ such that P^X admits a bounded density $\mu(\cdot)$ on S w.r.t. the Lebesgue measure and $\mu(x) \geq \mu_0 > 0$ for all $x \in S$.

(A3) There exists a constant c_0 such that $\log M \leq c_0 n$.

(A4) There exists a constant c_0 such that $M \leq c_0 n$.

Theorem 1. *Under assumptions (A1)–(A3) we have*

$$\sup_{f_1, \dots, f_M \in \mathcal{F}_0} \inf_{T_n} \sup_{f \in \mathcal{F}_0} \left[R(T_n, f) - \min_{1 \leq j \leq M} \|f_j - f\|^2 \right] \geq c \psi_n^{\mathbf{MS}}(M)$$

for some constant $c > 0$ and any integer n, where $\inf_{T_n}$ denotes the infimum over all estimators and

$$\psi_n^{\mathbf{MS}}(M) = \frac{\log M}{n}\ .$$

Proof. Let $\{\varphi_j\}_{j=1}^M$ be an orthogonal system of functions in $L_2(S, dx)$ for the cube S given in assumption (A2) and satisfying $\|\varphi_j\|_\infty \le A < \infty$ for $j = 1, \dots, M$. Such functions can be constructed, for example, by taking $\varphi_j(x) = A_1 \cos(ajx_1 + b)$ for $x \in S$ and for suitably chosen constants A_1, a and b, where x_1 is the first coordinate of x. Define the functions

$$f_j(x) = \gamma \sqrt{\frac{\log M}{n}} \varphi_j(x) I(x \in S), \quad j = 1, \dots, M,$$

where $I(\cdot)$ denotes the indicator function and γ is a positive constant to be chosen. In view of assumption (A3), $\{f_1, \dots, f_M\} \subset \mathcal{F}_0$ if γ is small enough. Thus, it suffices to prove the lower bound of the theorem for $f \in \{f_1, \dots, f_M\}$. But for such f we have $\min_{1 \le j \le M} \|f_j - f\|^2 = 0$, and to finish the proof of the theorem it is sufficient to bound from below by $c\psi_n^{\mathbf{MS}}(M)$ the quantity $\sup_{f \in \{f_1, \dots, f_M\}} R(T_n, f)$ uniformly over all estimators T_n. This is done by applying Lemma 1. Using assumption (A2) and orthogonality of the system $\{f_j\}_{j=1}^M$ on S we get, for $j \ne k$,

$$\|f_j - f_k\|^2 \asymp \int_S (f_j(x) - f_k(x))^2 dx = \int_S f_j^2(x) dx + \int_S f_k^2(x) dx \asymp \frac{\gamma^2 \log M}{n} . \quad (6)$$

Since ξ_j's are $\mathcal{N}(0, \sigma^2)$ random variables, the Kullback divergence $K(P_{f_j}, P_{f_k})$ between P_{f_j} and P_{f_k} satisfies

$$K(P_{f_j}, P_{f_k}) = \frac{n}{2\sigma^2} \|f_j - f_k\|^2, \quad j = 1, \dots, M. \quad (7)$$

In view of (6) and (7), one can choose γ small enough to have $K(P_{f_j}, P_{f_k}) \le (1/16) \log M$, $j, k = 1, \dots, M$. To finish the proof it remains to use this inequality, (6) and Lemma 1.

Theorem 2. *Under assumptions (A1)–(A3) we have*

$$\sup_{f_1, \dots, f_M \in \mathcal{F}_0} \inf_{T_n} \sup_{f \in \mathcal{F}_0} \left[R(T_n, f) - \min_{\lambda \in \Lambda^M} \|f_\lambda - f\|^2 \right] \ge c\psi_n^{\mathbf{C}}(M)$$

for some constant $c > 0$ and any integer n, where $\inf_{T_n}$ denotes the infimum over all estimators and

$$\psi_n^{\mathbf{C}}(M) = \begin{cases} M/n & \text{if } M \le \sqrt{n}, \\ \sqrt{\frac{1}{n} \log\left(\frac{M}{\sqrt{n}} + 1\right)} & \text{if } M > \sqrt{n}. \end{cases}$$

Proof. Consider first the case where $M > \sqrt{n}$. Let the functions $\{\varphi_j\}_{j=1}^M$ be as in the proof of Theorem 1. Set

$$f_j(x) = \gamma \varphi_j(x) I(x \in S), \quad j = 1, \dots, M, \quad (8)$$

for some constant γ to be chosen later. Define an integer

$$m = \left\lceil c_2 \Big[n / \log\Big(\frac{M}{\sqrt{n}} + 1 \Big) \Big]^{1/2} \right\rceil \tag{9}$$

for a constant $c_2 > 0$ chosen in such a way that $M \geq 6m$. Denote by $\mathcal{C}$ the finite set of such convex combinations of $f_1, \ldots, f_M$ that m of the coefficients λ_j are equal to $1/m$ and the remaining $M - m$ coefficients are zero. For every pair of functions $g_1, g_2 \in \mathcal{C}$ we have

$$\|g_1 - g_2\|^2 \leq c_3 \gamma^2 / m. \tag{10}$$

Clearly, $\mathcal{C} \subset \mathcal{F}_0$ for γ small enough and $\min_{\lambda \in \Lambda^M} \|f_\lambda - f\|^2 = 0$ for any $f \in \mathcal{C}$. Therefore, to prove the theorem for $M > \sqrt{n}$ it is sufficient to bound from below by $c\sqrt{\frac{1}{n} \log\Big(\frac{M}{\sqrt{n}} + 1 \Big)}$ the supremum $\sup_{f \in \mathcal{C}} R(T_n, f)$ uniformly over all estimators T_n. In fact, we will show that the required lower bound holds already for the quantity $\sup_{f \in \mathcal{N}} R(T_n, f)$ where $\mathcal{N}$ is a subset of $\mathcal{C}$ of cardinality $\mathrm{card}(\mathcal{N})$ satisfying

$$\log(\mathrm{card}(\mathcal{N})) \geq c_4 m \log\left(\frac{M}{m} + 1 \right) \tag{11}$$

and such that for every two functions $g_1, g_2 \in \mathcal{N}$ we have

$$\|g_1 - g_2\|^2 \geq c_5 \gamma^2 / m.$$

The existence of such a subset $\mathcal{N}$ of $\mathcal{C}$ follows, for example, from Lemma 4 of Birgé and Massart (2001). Now, using (7) – (11) and the definition of m we get that, for any $g_1, g_2 \in \mathcal{N}$,

$$K(P_{g_1}, P_{g_2}) \leq c_6 \gamma^2 n / m \leq c_7 \gamma^2 \log(\mathrm{card}(\mathcal{N})).$$

Finally, we choose γ small enough to have $c_7 \gamma^2 < 1/16$ and we apply Lemma 1 to get the result.

Consider now the case $M \leq \sqrt{n}$. Define the functions f_j by (8) and introduce a finite set of functions

$$\mathcal{C}_1 = \left\{ f = \frac{1}{\sqrt{n}} \sum_{j=1}^{M} \omega_j f_j : \omega \in \Omega \right\} \tag{12}$$

where Ω is the set of all vectors ω of length M with binary coordinates $\omega_j \in \{0, 1\}$. Since $M \leq \sqrt{n}$ we have $\mathcal{C}_1 \subset \mathcal{F}_0$ for γ small enough and $\mathcal{C}_1 \subset \{f_\lambda : \lambda \in \Lambda^M\}$. Therefore, similarly to the previous proofs, it is sufficient to bound from below $\inf_{T_n} \sup_{f \in \mathcal{C}_1} R(T_n, f)$. Using assumption (A2) we get that, for any $g_1, g_2 \in \mathcal{C}_1$,

$$\|g_1 - g_2\|^2 \leq c_8 \gamma^2 M / n. \tag{13}$$

If $M < 8$ we have $\psi_n^{\mathbf{C}}(M) \asymp 1/n$, and the lower bound of the theorem can be easily deduced from testing between two hypotheses: $f_1 \equiv 0$ and $f_2(x) =$

$n^{-1/2} I(x \in S)$. For $M \geq 8$ it follows from the Varshamov-Gilbert bound (see e.g. Tsybakov (2003), Ch.2) that there exists a subset $\mathcal{N}_1$ of $\mathcal{C}_1$ such that $\mathrm{card}(\mathcal{N}_1) \geq 2^{M/8}$ and

$$\|g_1 - g_2\|^2 \geq c_9 \gamma^2 M/n. \tag{14}$$

for any $g_1, g_2 \in \mathcal{N}_1$. Using (7) and (13) we get, for any $g_1, g_2 \in \mathcal{N}_1$,

$$K(P_{g_1}, P_{g_2}) \leq c_{10} \gamma^2 M \leq c_{11} \gamma^2 \log(\mathrm{card}(\mathcal{N}_1)),$$

and by choosing γ small enough, we can finish the proof in the same way as in the case $M > \sqrt{n}$.

Note that Theorem 2 generalizes the lower bounds for convex aggregation given by Juditsky and Nemirovski (2000) and Yang (2001). Juditsky and Nemirovski (2000) considered the case of very large M (satisfying $M \geq n/\log n$) and they proved the lower bound with the rate $\sqrt{n^{-1} \log M}$ which coincides in order with $\psi_n^{\mathbf{C}}(M)$ in this zone. Yang (2001) obtained the lower bounds for convex aggregation with polynomial M, i.e. $M \asymp n^\tau$ for $0 < \tau < \infty$. His bounds also follow as a special case from Theorem 2.

Theorem 3. *Under assumptions (A1), (A2), (A4) we have*

$$\sup_{f_1, \dots, f_M \in \mathcal{F}_0} \inf_{T_n} \sup_{f \in \mathcal{F}_0} \left[R(T_n, f) - \min_{\lambda \in \mathbf{R}^M} \|f_\lambda - f\|^2 \right] \geq c \psi_n^{\mathbf{L}}(M)$$

for some constant $c > 0$ and any integer n, where $\inf_{T_n}$ denotes the infimum over all estimators and

$$\psi_n^{\mathbf{L}}(M) = \frac{M}{n} \ .$$

Proof. Assume w.l.o.g. that there exist disjoint subsets $S_1, \dots, S_M$ of S such that the Lebesgue measure of S_j is $1/M$. Define the functions $f_j(x) = \gamma I(x \in S_j),\ j = 1, \dots, M$, for a constant $\gamma > 0$, and the set

$$\mathcal{C}_2 = \left\{ f = \sqrt{\frac{M}{n}} \sum_{j=1}^{M} \omega_j f_j : \omega \in \Omega \right\}$$

with Ω as in (12). Assumption (A4) guarantees that $\mathcal{C}_2 \subset \mathcal{F}_0$ for γ small enough. Since the functions f_j are mutually orthogonal and $\int f_j^2(x) dx = \gamma^2/M$, the rest of the proof is identical to the part of the proof of Theorem 2 after (13) (with $\mathcal{C}_1$ replaced by $\mathcal{C}_2$), and it is therefore omitted.

3 Attainability of the lower bounds

In this section we show that the lower bounds of Theorems 1–3 give optimal rates of aggregation. We start with the problem of linear aggregation (Problem (L)) in which case we construct an aggregate attaining in order the lower bound of Theorem 3.

Denote by $\mathcal{L}$ the linear span of $f_1, \dots, f_M$. Let $\varphi_1, \dots, \varphi_{M'}$ with $M' \le M$ be an orthonormal basis of $\mathcal{L}$ in $L_2(\mathcal{X}, P^X)$. Consider a linear aggregate

$$\tilde{f}_n^{\mathbf{L}}(x) = \sum_{j=1}^{M'} \hat{\lambda}_j \varphi_j(x), \quad x \in \mathcal{X}, \tag{15}$$

where

$$\hat{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \varphi_j(X_i).$$

Theorem 4. *Let* $\mathbf{E}(\xi_i) = 0$, $\mathbf{E}(\xi_i^2) \le \sigma^2 < \infty$. *Then*

$$R(\tilde{f}_n^{\mathbf{L}}, f) - \min_{\lambda \in \mathbf{R}^M} \|f_\lambda - f\|^2 \le \frac{(\sigma^2 + L^2)M}{n}$$

for any integers $M \ge 2$, $n \ge 1$ *and any* $f, f_1, \dots, f_M \in \mathcal{F}_0$, *where* L *is the constant in (5).*

Proof. We have $\min_{\lambda \in \mathbf{R}^M} \|f_\lambda - f\|^2 = \|f_\lambda^* - f\|^2$ where $f_\lambda^* = \sum_{j=1}^{M'} \lambda_j^* \varphi_j$, $\lambda_j^* = (f, \varphi_j)$, and $(\cdot, \cdot)$ is the scalar product in $L_2(\mathcal{X}, P^X)$. Now,

$$\|\tilde{f}_n^{\mathbf{L}} - f\|^2 = \sum_{j=1}^{M'} (\hat{\lambda}_j - \lambda_j^*)^2 + \|f_\lambda^* - f\|^2,$$

and to finish the proof it suffices to note that $\mathbf{E}(\hat{\lambda}_j) = \lambda_j^*$, $\mathbf{E}\Big[(\hat{\lambda}_j - \lambda_j^*)^2\Big] = \mathrm{Var}(\hat{\lambda}_j) \le (\sigma^2 + L^2)/n$.

Theorems 3 and 4 imply the following result.

Corollary 1. *Under assumptions (A1), (A2), (A4) the sequence* $\psi_n^{\mathbf{L}}(M)$ *is optimal rate of linear aggregation.*

Consider now the problem of convex aggregation (Problem (C)). If $M \le \sqrt{n}$ the lower bound of Theorem 2 is identical to the linear aggregation case, so we can use the linear aggregate $\tilde{f}_n^{\mathbf{L}}$ defined in (15) that attains this bound in view of Theorem 4. For $M > \sqrt{n}$ we use a different procedure. To define this procedure, consider first the Kullback divergence based model selection aggregate (Catoni (2001), Yang (2000)). This aggregate, for the problem of model selection with N Borel functions $g_1, \dots, g_N$ on $\mathcal{X}$, is defined by

$$\tilde{g}_{n,N}(x) = \frac{1}{n+1} \sum_{k=0}^{n} p_{k,N}(x)$$

where

$$p_{k,0}(x) = \frac{1}{N} \sum_{j=1}^{N} g_j(x)$$

and, for $k = 1, \dots, N$,

$$p_{k,N}(x) = \frac{\sum_{j=1}^{N} g_j(x) \prod_{i=1}^{k} \exp(-(Y_i - g_j(X_i))^2/2\sigma^2)}{\sum_{j=1}^{N} \prod_{i=1}^{k} \exp(-(Y_i - g_j(X_i))^2/2\sigma^2)} \ .$$

As shown by Catoni (2001), for any integers $M \geq 2$, $n \geq 1$ and any functions $f, g_1, \dots, g_M \in \mathcal{F}_0$,

$$R(\tilde{g}_{n,N}, f) \leq \min_{1 \leq j \leq N} \|g_j - f\|^2 + C_0 \frac{\log N}{n} \tag{16}$$

where $C_0 < \infty$ is a constant that depends only on L and σ^2.

Now, define m by (9) and denote by $\mathcal{C}'$ the set of all sub-convex combinations of $f_1, \dots, f_M$ with weights equal to integer multiples of $1/m$. We have

$$\mathrm{card}(\mathcal{C}') = \sum_{j=1}^{m} \binom{M+j-1}{j} \leq \left(\frac{e(M+m)}{m} \right)^m \tag{17}$$

(cf., e.g., Devroye, Györfi and Lugosi (1996, p.218)). Let $N = \mathrm{card}(\mathcal{C}')$, let $g_1, \dots, g_N$ be the elements of $\mathcal{C}'$, and denote by $\tilde{f}_{n,m}^{\mathbf{MS}}$ the corresponding model selection aggregate $\tilde{g}_{n,N}$. Then (16) takes the form

$$R(\tilde{f}_{n,m}^{\mathbf{MS}}, f) \leq \min_{g \in \mathcal{C}'} \|g - f\|^2 + C_0 \frac{\log(\mathrm{card}(\mathcal{C}'))}{n} \ , \tag{18}$$

which holds for every $f \in \mathcal{F}_0$.

Finally, define a compound method of convex aggregation by

$$\tilde{f}_n^{\mathbf{C}} = \begin{cases} \tilde{f}_n^{\mathbf{L}} & \text{if } \ M \leq \sqrt{n}, \\ \tilde{f}_{n,m}^{\mathbf{MS}} & \text{if } \ M > \sqrt{n}. \end{cases}$$

This definition emphasizes an intermediate character of convex aggregation: it switches from linear to model selection aggregates. If $M > \sqrt{n}$ we are in a "sparse case": convex aggregation oracle concentrates only on a relatively small number of functions f_j, and optimal rate is attained on a model selection type procedure. On the contrary, for $M \leq \sqrt{n}$ convex aggregation oracle does not concentrate on the boundary of the set Λ^M, and therefore it essentially behaves as a linear oracle giving solution to unrestricted minimization problem.

Theorem 5. *Under assumption (A1) we have*

$$R(\tilde{f}_n^{\mathbf{C}}, f) - \min_{\lambda \in \Lambda^M} \|f_\lambda - f\|^2 \leq C \psi_n^{\mathbf{C}}(M)$$

for some constant $C < \infty$, any integers $M \geq 2$, $n \geq 1$ and any $f, f_1, \dots, f_M \in \mathcal{F}_0$.

Proof. In view of Theorem 4, it suffices to consider the case $M > \sqrt{n}$. Let $\lambda^0 = (\lambda_1^0, \dots, \lambda_M^0)$ be the weights of a convex aggregation oracle, i.e. a vector λ^0 satisfying $\|f_{\lambda^0} - f\| = \min_{\lambda \in \Lambda^M} \|f_\lambda - f\|$. Applying the argument of Nemirovski (2000, p.192–193) to the vector λ^0 instead of $\lambda(z)$ and putting there $K = m$ we find

$$\sum_{i=1}^{J} p_i \|h_i - f\|^2 \le \min_{\lambda \in \Lambda^M} \|f_\lambda - f\|^2 + \frac{L^2}{m}$$

where $J = \mathrm{card}(\mathcal{C}')$, $(p_1, \dots, p_J)$ is a probability vector (i.e. $p_1 + \cdots + p_J = 1$, $p_j \ge 0$) that depends on λ^0 and the functions h_i are the elements of the set $\mathcal{C}'$. This immediately implies that

$$\min_{g \in \mathcal{C}'} \|g - f\|^2 \le \min_{\lambda \in \Lambda^M} \|f_\lambda - f\|^2 + \frac{L^2}{m} \ . \tag{19}$$

Combining (17)–(19) we obtain

$$\begin{aligned} R(\tilde{f}_{n,m}^{\mathbf{MS}}, f) &\le \min_{\lambda \in \Lambda^M} \|f_\lambda - f\|^2 + \frac{L^2}{m} + c_{12} \frac{m}{n} \left[\log\left(\frac{M}{m}\right) + 1 \right] \\ &\le \min_{\lambda \in \Lambda^M} \|f_\lambda - f\|^2 + C \sqrt{\frac{1}{n} \log\left(\frac{M}{\sqrt{n}} + 1\right)} \end{aligned}$$

for a constant $C < \infty$.

Theorems 2 and 5 imply the following result.

Corollary 2. *Under assumptions (A1)–(A3) the sequence $\psi_n^{\mathbf{C}}(M)$ is optimal rate of convex aggregation.*

Finally, for the Problem (MS), the attainability of the lower bound of Theorem 1 follows immediately from (16) with $N = M$ and $g_j = f_j$. Thus, we have the following corollary.

Corollary 3. *Under assumptions (A1)–(A3) the sequence $\psi_n^{\mathbf{MS}}(M)$ is optimal rate of model selection aggregation.*

References

1. Birgé, L.: Model selection for Gaussian regression with random design. Prépublication n. 783, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 - Paris 7 (2002).
2. Birgé, L., Massart, P.: Gaussian model selection. J. Eur. Math. Soc. **3** (2001) 203–268.
3. Catoni, O.: *Statistical Learning Theory and Stochastic Optimization.* Ecole d'Eté de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics, Springer, N.Y. (to appear).
4. Devroye, L., Györfi, L., Lugosi, G.: *A Probabilistic Theory of Pattern Recognition.* Springer, N.Y. (1996).

5. Györfi, L., Kohler, M., Krzyżak, A., Walk, H.: *A Distribution-Free Theory of Nonparametric Regression*. Springer, N.Y.(2002).
6. Juditsky, A., Nemirovski, A.: Functional aggregation for nonparametric estimation. Annals of Statistics **28** (2000) 681–712.
7. Nemirovski, A.: *Topics in Non-parametric Statistics*. Ecole d'Eté de Probabilités de Saint-Flour XXVIII - 1998, Lecture Notes in Mathematics, v. 1738, Springer, N.Y. (2000).
8. Tsybakov, A.: *Introduction à l'estimation non-paramétrique.* (2003) Springer (to appear).
9. Wegkamp, M.: Model selection in nonparametric regression. Annals of Statistics **31** (2003) (to appear).
10. Yang, Y.: Combining different procedures for adaptive regression. J.of Multivariate Analysis **74** (2000) 135–161.
11. Yang, Y.: Aggregating regression procedures for a better performance (2001). Manuscript.