HAL
open science

Learning by mirror averaging

Anatoli B. Juditsky, Philippe Rigollet, Alexandre Tsybakov

To cite this version:

Anatoli B. Juditsky, Philippe Rigollet, Alexandre Tsybakov. Learning by mirror averaging. Annals of Statistics, 2008, 36 (5), pp.2183-2206. 10.1214/07-AOS546 . hal-00014097v3

HAL Id: hal-00014097

https://hal.science/hal-00014097v3

Submitted on 3 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ccsd-00014097, version 3 - 3 May 2006

Learning by mirror averaging

Anatoli Juditsky * Philippe Rigollet† Alexandre Tsybakov †

May 3, 2006

Abstract

Given a collection of M different estimators or classifiers, we study the problem of model selection type aggregation, i.e., we construct a new estimator or classifier, called aggregate, which is nearly as good as the best among them with respect to a given risk criterion. We define our aggregate by a simple recursive procedure which solves an auxiliary stochastic linear programming problem related to the original non-linear one and constitutes a special case of the mirror averaging algorithm. We show that the aggregate satisfies sharp oracle inequalities under some general assumptions. The results allow one to construct in an easy way sharp adaptive nonparametric estimators for several problems including regression, classification and density estimation.

Mathematics Subject Classifications: Primary 62G08, Secondary 62C20, 62G05, 62G20.
Key Words: learning, aggregation, oracle inequalities, mirror averaging, model selection, stochastic optimization.
Short title: Learning by mirror averaging.

1 Introduction

Several problems in statistics and machine learning can be stated as follows: given a collection of M different estimators , construct a new estimator which is nearly as good as the best among them with respect to a given risk criterion. This target is called model selection (MS) type aggregation, and it can be described in terms of the following stochastic optimization problem.

Let $(\mathcal{Z}, \mathfrak{F})$ be a measurable space and let Θ be the simplex

$$\Theta = \left\{ \theta \in \mathbb{R}^M : \sum_{j=1}^M \theta^{(j)} = 1,\ \theta^{(j)} \geq 0,\ j = 1, \ldots, M \right\}.$$

Here and throughout the paper we suppose that $M \geq 2$ and we denote by $z^{(j)}$ the jth component of a vector $z \in \mathbb{R}^M$. We denote by $\left[z^{(j)}\right]_{j=1}^M$ the vector $z = (z^{(1)}, \ldots, z^{(M)})^\top \in \mathbb{R}^M$.

Let Z be a random variable with values in $\mathcal{Z}$. The distribution of Z is denoted by P and the corresponding expectation by E. Suppose that P is unknown and that we observe n i.i.d. random variables $Z_1, \ldots, Z_n$ with values in $\mathcal{Z}$ having the same distribution as Z. The distribution (respectively, expectation) w.r.t. the sample $Z_1, \ldots, Z_n$ is denoted by P_n (respectively, by E_n).

Consider a measurable function $Q : \mathcal{Z} \times \Theta \to \mathbb{R}$ and the corresponding average risk function

$$A(\theta) = EQ(Z, \theta),$$

*Laboratoire de Modélisation et Calcul - Université Grenoble I. Email: anatoli.iouditski@imag.fr
†Laboratoire de Probabilités et Modèles aléatoires, UMR 7599, Université Paris 6, case 188, 4, pl. Jussieu, F-75252 Paris Cedex 5, France. Email: {rigollet, tsybakov}@ccr.jussieu.fr

assuming that this expectation exists for all $\theta \in \Theta$. Stochastic optimization problems that are usually studied in this context consist in minimization of A on some subsets of Θ, given the sample $Z_1, \dots, Z_n$. Note that since the distribution of Z is unknown, direct (deterministic) minimization of A is not possible.

For $j \in \{1, \dots, M\}$, denote by e_j the jth coordinate unit vector in $\mathbb{R}^M$: $e_j = (0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{R}^M$, where 1 appears in jth position.

The stochastic optimization problem associated to MS aggregation is $\min_{\theta \in \{e_1, \dots, e_M\}} A(\theta)$. The aim of MS aggregation is to "mimic the oracle" $\min_{1 \le j \le M} A(e_j)$, i.e. , to construct an estimator $\tilde{\theta}_n$ measurable w.r.t. $Z_1, \dots, Z_n$ and called aggregate, such that

$$E_n A(\tilde{\theta}_n) \le \min_{1 \le j \le M} A(e_j) + \Delta_{n,M} \,, \tag{1.1}$$

where $\Delta_{n,M} > 0$ is a remainder term that should be as small as possible.

As an example, one may consider the loss function of the form $Q(z, \theta) = \ell(z, \theta^\top H)$ where $\ell : \mathcal{Z} \times \mathbb{R} \to \mathbb{R}$ and $H = (h_1, \dots, h_M)$ is a vector of preliminary estimators (classifiers) constructed from a training sample which is supposed to be frozen in our considerations (thus, h_j can be viewed as fixed functions). The value $A(e_j) = E\ell(Z, h_j)$ is the loss corresponding to h_j. Inequality (1.1) can then be interpreted as follows: the aggregate $\tilde{\theta}_n^\top H$, i.e. the convex combination of initial estimators (classifiers) h_j, with the vector of mixture coefficients $\tilde{\theta}_n$ measurable w.r.t. $Z_1, \dots, Z_n$, is nearly as good as the best among $h_1, \dots, h_M$. The word "nearly" here means that the value $\min_{1 \le j \le M} A(e_j)$ is reproduced up to a reasonably small remainder term $\Delta_{n,M}$. Lower bounds can be established showing that, under some assumptions, the smallest possible value of $\Delta_{n,M}$ in a minimax sense has the form

$$\Delta_{n,M} = \frac{C \log M}{n} \,, \tag{1.2}$$

with some constant $C > 0$ [cf. Tsybakov (2003)].

Besides being in themselves precise finite sample results, oracle inequalities of the type (1.1) are very useful in adaptive nonparametric estimation. They allow one to prove that the aggregate estimator $\tilde{\theta}_n^\top H$ is adaptive in a minimax asymptotic sense (and even sharp minimax adaptive is several cases: for more discussion see, e.g., Nemirovski (2000)).

The aim of this paper is to obtain bounds of the form (1.1) – (1.2) under some general conditions on the loss function Q. For two special cases (density estimation with the Kullback-Leibler (KL) loss, and regression model with squared loss) such bounds have been proved earlier in the benchmark works of Catoni (1997, 1999, 2004) and Yang (2000). They independently obtained the bound for density estimation with the KL loss, and Catoni (1999, 2004) solved the problem for the regression model with squared loss. Bunea and Nobel (2005) suggested another proof of the regression result of Catoni (1999, 2004) improving it in the case of bounded response, and obtained some inequalities with suboptimal remainder terms under weaker conditions. For a problem which is different but close to ours (MS aggregation in the Gaussian white noise model with squared loss) Nemirovski (2000, p.226) established an inequality similar to (1.1), with a suboptimal remainder term. Leung and Barron (2004) improved upon this result to achieve the optimality.

Several other works provided less precise bounds than (1.1) – (1.2), with $K \min_{1 \le j \le M} A(e_j)$ where $K > 1$, instead of $\min_{1 \le j \le M} A(e_j)$ in (1.1) and with a remainder term which is sometimes larger than the optimal one (1.2): a detailed account can be found in the survey of Boucheron, Bousquet and Lugosi (2005) or in the lecture notes of Massart (2006). We mention here only some recent work where aggregation of arbitrary estimators is considered: Wegkamp (2003), Bartlett, Boucheron and Lugosi (2002), Lugosi and Wegkamp (2004), Yang (2004), Zhang (2003), Bunea, Tsybakov and Wegkamp (2004), Samarov and Tsybakov (2005). These results are useful for statistical applications, especially if K is close to 1. However, the inequalities with $K > 1$ do not provide valid bounds for the excess risk $E_n A(\tilde{\theta}_n) - \min_{1 \le j \le M} A(e_j)$, i.e., they do not show that $\tilde{\theta}_n$ approximately solves the stochastic optimization problem.

Here we study the aggregate $\hat{\theta}_n$ which is defined by a simple recursive procedure. The procedure solves an auxiliary stochastic linear programming problem related to the original non-linear one $\min_{1\le j\le M} A(e_j)$, and it constitutes a special case of the mirror averaging algorithm of Juditsky, Nazin, Tsybakov and Vayatis (2005). In particular cases, it yields the methods described by Catoni (2004) and Yang (2000). We prove that the mirror averaging aggregate $\hat{\theta}_n$ satisfies oracle inequalities (1.1) – (1.2) under some general assumptions on Q such as, for example, exponential concavity. We show that these assumptions are fulfilled for several statistical models including regression, classification and density estimation.

Our results have a connection to the theory of on-line prediction of individual deterministic sequences [cf. Cesa-Bianchi and Lugosi (2006)]. A general problem considered there is for an agent to compete against the observed predictions of a group of experts, so that the agent's error is close to that of the best expert. The results and the methods are similar to ours, in particular, oracle inequalities can be obtained for deterministic bounded or binary sequences of observations [Vovk (1990), Kivinen and Warmuth (1999)]. However, they cannot be meaningfully applied to our stochastic setup because they deal with the cumulative loss rather than with the expected loss, and the algorithms of deterministic prediction do not have the averaging step [cf. (2.3), (2.6) below]. For more references and discussion we refer to Cesa-Bianchi and Lugosi (2006).

2 The algorithm

We first recall the definition of a particular version of the mirror averaging algorithm. For $\beta > 0$ define the function $W_\beta : \mathbb{R}^M \to \mathbb{R}$ by

$$W_\beta(z) \triangleq \beta \log\left(\frac{1}{M}\sum_{j=1}^{M} \mathrm{e}^{-z^{(j)}/\beta}\right), \qquad z = (z^{(1)}, \dots, z^{(M)}). \tag{2.1}$$

The gradient of W_β is given by

$$\nabla W_\beta(z) = \left[-\frac{\mathrm{e}^{-z^{(j)}/\beta}}{\sum_{k=1}^{M} \mathrm{e}^{-z^{(k)}/\beta}}\right]_{j=1}^{M}.$$

Consider now $Q(z,\theta)$ which is convex and differentiable in θ for all $z \in \mathcal{Z}$, with gradient $\nabla_\theta Q(z,\theta)$. The following algorithm is a particular case of the general mirror averaging algorithm of Juditsky, Nazin, Tsybakov and Vayatis (2005).

- Fix the initial values $\bar{\theta}_0 \in \Theta$ and $\zeta_0 = 0 \in \mathbb{R}^M$.
- For $i = 1, \dots, n-1$, do the recursive update

$$\begin{aligned} \zeta_i &= \zeta_{i-1} + \nabla_\theta Q(Z_i, \bar{\theta}_{i-1}), \\ \bar{\theta}_i &= -\nabla W_\beta(\zeta_i). \end{aligned} \tag{2.2}$$

- Output at iteration n the average:

$$\tilde{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \bar{\theta}_{i-1}. \tag{2.3}$$

The name mirror averaging is due to the fact that (2.2) does a stochastic gradient descent in the dual space with further "mirroring" to the primal space (so-called mirror descent, cf. Ben Tal and Nemirovski (1999)) and averaging: for more details and discussion see Juditsky, Nazin, Tsybakov and Vayatis (2005). They show that under some assumptions the following oracle inequality holds

$$E_n A(\tilde{\theta}_n) \le \min_{\theta\in\Theta} A(\theta) + C_0\sqrt{\frac{\log M}{n}}\,, \tag{2.4}$$

where $C_0 > 0$ is a constant depending only on β and on the supremum norm of the gradient $\nabla_\theta Q(\cdot,\cdot)$.

Note that in (2.4) the minimum is taken over the whole simplex Θ, so an inequality of the type (1.1) holds as well, but for large n the remainder term in (2.4) is of larger order than the optimal one given in (1.2).

To improve upon this, consider another version of the mirror averaging method. If A is a convex function, we can bound it from above by a linear function:

$$A(\theta) \le \sum_{j=1}^{M} \theta^{(j)} A(e_j) \triangleq \tilde{A}(\theta), \quad \forall\, \theta \in \Theta,$$

where

$$\tilde{A}(\theta) = \mathbb{E}\,\tilde{Q}(Z,\theta) \quad \text{with} \;\; \tilde{Q}(Z,\theta) \triangleq \theta^\top u(Z), \;\; u(Z) \triangleq \Big(Q(Z,e_1),\ldots,Q(Z,e_M)\Big)^\top .$$

Note that

$$\tilde{A}(e_j) = A(e_j), \quad j = 1,\ldots,M.$$

Since Θ is a simplex, the minimum of the linear function $\tilde{A}$ is attained at one of its vertices. Therefore,

$$\min_{\theta\in\Theta} \tilde{A}(\theta) = \min_{1\le j\le M} A(e_j).$$

Thus, the problem of MS aggregation can be solved using a mirror averaging algorithm for the linear stochastic programming problem of minimization of $\tilde{A}$ on Θ. It is defined as follows.

Consider the vector

$$u_i \triangleq \Big(Q(Z_i,e_1),\ldots,Q(Z_i,e_M)\Big)^\top = u(Z_i) = \nabla_\theta \tilde{Q}(Z_i,\theta)\,,$$

and the iterations:

- Fix the initial values $\theta_0 \in \Theta$ and $\zeta_0 = 0 \in \mathbb{R}^M$.
- For $i = 1,\ldots,n-1$, do the recursive update

$$\begin{aligned} \zeta_i &= \zeta_{i-1} + u_i\,, \\ \theta_i &= -\nabla W_\beta(\zeta_i)\,. \end{aligned} \tag{2.5}$$

- Output at iteration n the average

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \theta_{i-1}\,. \tag{2.6}$$

Remark that we define the algorithm (2.5) – (2.6) for a general function Q, not necessarily for a convex one. Note also that $\hat{\theta}_n$ is measurable w.r.t. the subsample $(Z_1, \ldots, Z_{n-1})$. The components $\theta_i^{(j)}$ of the vector θ_i from (2.5) can be written in the form

$$\theta_i^{(j)} = \frac{\exp\left(-\beta^{-1}\sum_{m=1}^{i} Q(Z_m, e_j)\right)}{\sum_{k=1}^{M}\exp\left(-\beta^{-1}\sum_{m=1}^{i} Q(Z_m, e_k)\right)}, \quad j = 1, \ldots, M.$$

Since (2.5) – (2.6) is a particular case of the mirror averaging method (2.2) – (2.3) corresponding to a linear function $\tilde{A}$, inequality (2.4) remains valid with A replaced by $\tilde{A}$. But we show below that in fact $\hat{\theta}_n$ satisfies a stronger inequality, i.e., one with the optimal remainder term (1.2).

Finally, note that W_β defined in (2.1) is not the only possible choice: other functions W_β satisfying the conditions described in Juditsky, Nazin, Tsybakov and Vayatis (2005) can be used to construct the updates (2.5).

3 Main results

In this section we give two theorems. They establish results in the form of oracle inequalities satisfied by $\hat{\theta}_n$. Theorem 3.1 requires a more conservative assumption on the loss functions Q than Theorem 3.2. This assumption is easier to check, and it often leads to a sharper bound but not for such models as nonparametric density estimation with the L_2 loss which will be treated using Theorem 3.2. In some cases (for example, in regression with Gaussian noise) Theorem 3.1 yields a suboptimal remainder term, while Theorem 3.2 does the correct job. In both theorems it is supposed that the values $A(e_1), \ldots, A(e_M)$ are finite. We will also need the following definition.

Definition 3.1 *A function $T : \mathbb{R}^M \to \mathbb{R}$ is* exponentially concave *if the composite function* $\exp \circ T$ *is concave.*

It is straightforward to see that exponential concavity of a function $-T$ implies that T is convex. Furthermore, if $-T/\beta$ is exponentially concave for some $\beta > 0$, then $-T/\beta'$ is exponentially concave for all $\beta' > \beta$. Let Q_1 be the function on $\mathcal{Z} \times \Theta \times \Theta$ defined by $Q_1(z, \theta, \theta') = Q(z, \theta) - Q(z, \theta')$ for all $z \in \mathcal{Z}$ and all $\theta, \theta' \in \Theta$.

Theorem 3.1 *Assume that Q_1 can be decomposed into the sum of two functions $Q_1 = Q_2 + Q_3$ such that:*

- *The mapping $\theta \mapsto -Q_2(z, \theta, \theta')/\beta$ is exponentially concave on the simplex Θ, for all $z \in \mathcal{Z}$, $\theta' \in \Theta$, and $Q_2(z, \theta, \theta) = 0$ for all $z \in \mathcal{Z}, \theta \in \Theta$.*
- *There exists a function R on $\mathcal{Z}$ integrable w.r.t. P and such that $-Q_3(z, \theta, \theta') \le R(z)$, for all $z \in \mathcal{Z}$, $\theta, \theta' \in \Theta$.*

Then the aggregate $\hat{\theta}_n$ satisfies, for any $M \ge 2, n \ge 1$, the following oracle inequality

$$E_{n-1}A(\hat{\theta}_n) \le \min_{1 \le j \le M} A(e_j) + \frac{\beta \log M}{n} + E\big[R(Z)\big] .$$

Theorem 3.2 *Assume that for some $\beta > 0$ there exists a Borel function $\Psi_\beta : \Theta \times \Theta \to \mathbb{R}_+$ such that the mapping $\theta \mapsto \Psi_\beta(\theta, \theta')$ is concave on the simplex Θ for any fixed $\theta' \in \Theta$, $\Psi_\beta(\theta, \theta) = 1$ and*

$E\exp(-Q_1(Z,\theta,\theta')/\beta) \le \Psi_\beta(\theta,\theta')$ *for all* $\theta,\theta' \in \Theta$. *Then the aggregate* $\hat\theta_n$ *satisfies, for any* $M \ge 2, n \ge 1$, *the following oracle inequality*

$$E_{n-1}A(\hat\theta_n) \le \min_{1\le j\le M} A(e_j) + \frac{\beta \log M}{n}\,.$$

Proofs of both theorems are based on the following lemma. Introduce the discrete random variable ω with values in the set $\{e_1,\dots,e_M\}$ and with the distribution $\mathbb{P}$ defined conditionally on $(Z_1,\dots,Z_{n-1})$ by $\mathbb{P}[\omega = e_j] = \hat\theta_n^{(j)}$ where $\hat\theta_n^{(j)}$ is the jth component of $\hat\theta_n$. The expectation corresponding to $\mathbb{P}$ is denoted by $\mathbb{E}$.

Lemma 3.1 *For any measurable function* Q *and any* $\beta > 0$ *we have*

$$E_{n-1}A(\hat\theta_n) \le \min_{1\le j\le M} A(e_j) + \frac{\beta \log M}{n} + S_1. \tag{3.1}$$

where

$$S_1 \triangleq \beta E_n \log\Bigl(\mathbb{E}\exp\Bigl[-\frac{Q_1(Z_n,\omega,\mathbb{E}[\omega])}{\beta}\Bigr]\Bigr).$$

PROOF. By definition of $W_\beta(\cdot)$, for $i = 1,\dots,n$,

$$W_\beta(\zeta_i) - W_\beta(\zeta_{i-1}) = \beta\log\left(\frac{\sum_{j=1}^M \mathrm{e}^{-\zeta_i^{(j)}/\beta}}{\sum_{j=1}^M \mathrm{e}^{-\zeta_{i-1}^{(j)}/\beta}}\right) = \beta\log\bigl(-v_i^\top \nabla W_\beta(\zeta_{i-1})\bigr) = \beta\log\bigl(v_i^\top\theta_{i-1}\bigr)\,, \tag{3.2}$$

where

$$v_i = \left[\exp\left(-\frac{u_i^{(j)}}{\beta}\right)\right]_{j=1}^M.$$

Taking expectations on both sides of (3.2), summing up over i, using the fact that (θ_{i-1}, Z_i) has the same distribution as (θ_{i-1}, Z_n) for $i = 1,\dots,n$, and applying the Jensen inequality, we get

$$\begin{aligned}
\frac{E_n\bigl[W_\beta(\zeta_n) - W_\beta(\zeta_0)\bigr]}{n} &= \frac{\beta}{n}\sum_{i=1}^n E_n\log\Bigl(\sum_{j=1}^M \theta_{i-1}^{(j)}\exp\Bigl[-\frac{Q(Z_i,e_j)}{\beta}\Bigr]\Bigr)\\
&= \frac{\beta}{n}\sum_{i=1}^n E_n\log\Bigl(\sum_{j=1}^M \theta_{i-1}^{(j)}\exp\Bigl[-\frac{Q(Z_n,e_j)}{\beta}\Bigr]\Bigr)\\
&\le \beta E_n\log\Bigl(\sum_{j=1}^M \hat\theta_n^{(j)}\exp\Bigl[-\frac{Q(Z_n,e_j)}{\beta}\Bigr]\Bigr) \triangleq S\,.
\end{aligned} \tag{3.3}$$

Since $Q_1(z,\omega,\mathbb{E}[\omega]) = Q(z,\omega) - Q(z,\mathbb{E}[\omega])$, and $\mathbb{E}[\omega] = \hat\theta_n$, the RHS of (3.3) can be written in the form

$$\begin{aligned}
S &= \beta E_n\log\Bigl(\mathbb{E}\exp\Bigl[-\frac{Q(Z_n,\omega)}{\beta}\Bigr]\Bigr)\\
&= \beta E_n\log\Bigl(\exp\Bigl[-\frac{Q(Z_n,\mathbb{E}[\omega])}{\beta}\Bigr]\Bigr) + \beta E_n\log\Bigl(\mathbb{E}\exp\Bigl[-\frac{Q_1(Z_n,\omega,\mathbb{E}[\omega])}{\beta}\Bigr]\Bigr)\\
&= -E_{n-1}A(\hat\theta_n) + S_1
\end{aligned} \tag{3.4}$$

We now bound from below the LHS of equation (3.3). As in Juditsky, Nazin, Tsybakov and Vayatis (2005), denote by βV the Fenchel-Legendre dual of $W_\beta(-z)$:

$$\beta V(\theta) = \sup_{z\in\mathbb{R}^M} [-z^\top\theta - W_\beta(z)].$$

Then, clearly,

$$-W_\beta(\zeta_n) \le \beta V(\theta) + \zeta_n^\top\theta, \quad \forall\,\theta\in\Theta\,,$$

and

$$V(\theta) = \log M + \sum_{j=1}^M \theta^{(j)}\log\theta^{(j)} \le \log M\,, \quad \forall\,\theta\in\Theta\,.$$

The last two inequalities and the fact that $W_\beta(\zeta_0) = W_\beta(0) = 0$ imply

$$\frac{E_n\big[W_\beta(\zeta_n) - W_\beta(\zeta_0)\big]}{n} \ge -\frac{\beta\log M}{n} - \min_{\theta\in\Theta}\frac{E_n\big[\zeta_n^\top\theta\big]}{n} = -\frac{\beta\log M}{n} - \min_{1\le j\le M} A(e_j). \tag{3.5}$$

Combining (3.3), (3.4) and (3.5) gives the lemma. ∎

In view of Lemma 3.1, to prove Theorems 3.1 and 3.2 it remains to give appropriate upper bounds for S_1.

Proof of Theorem 3.1. Since $Q_1 = Q_2 + Q_3$, with $-Q_3(z,\theta,\theta') \le R(z)$ for all $z\in\mathcal{Z}, \theta,\theta'\in\Theta$, the quantity S_1 can be bounded from above as follows

$$S_1 \le \beta E_n \log\left(\mathbb{E}\exp\Big[-\frac{Q_2(Z_n,\omega,\mathbb{E}[\omega])}{\beta}\Big]\right) + E_n\big[R(Z_n)\big]\,.$$

Now since $-Q_2(z,\cdot)/\beta$ is exponentially concave on Θ for all $z\in\mathcal{Z}$, the Jensen inequality yields

$$\mathbb{E}\exp\Big[-\frac{Q_2(Z_n,\omega,\mathbb{E}[\omega])}{\beta}\Big] \le \exp\Big[-\frac{Q_2(Z_n,\mathbb{E}[\omega],\mathbb{E}[\omega])}{\beta}\Big] = 1$$

Therefore $S_1 \le E_n\big[R(Z_n)\big]$. This and Lemma 3.1 imply the result of the theorem. ∎

Proof of Theorem 3.2. Using the Jensen inequality twice, with the concave functions $\log(\cdot)$ and $\Psi_\beta(\cdot,\mathbb{E}[\omega])$, we get

$$\begin{aligned} S_1 &\le E_{n-1}\log\left(E\,\mathbb{E}\exp\Big[-\frac{Q_1(Z,\omega,\mathbb{E}[\omega])}{\beta}\Big]\right)\\ &= E_{n-1}\log\left(\mathbb{E}\,E\exp\Big[-\frac{Q_1(Z,\omega,\mathbb{E}[\omega])}{\beta}\Big]\right)\\ &\le E_{n-1}\log\Big(\mathbb{E}\Psi_\beta(\omega,\mathbb{E}[\omega])\Big)\\ &\le E_{n-1}\log\Big(\Psi_\beta(\mathbb{E}[\omega],\mathbb{E}[\omega])\Big) = 0, \end{aligned} \tag{3.6}$$

where the first equality is due to the Fubini theorem. Theorem 3.2 follows now from (3.6) and Lemma 3.1.

4 Examples

In this section we apply Theorems 3.1 and 3.2 to three common statistical problems (regression, classification and density estimation). All the loss functions considered below are twice differentiable. The following proposition gives a simple sufficient condition for exponential concavity.

Proposition 4.1 *Let g be a twice differentiable function on Θ with gradient $\nabla g(\theta)$ and Hessian matrix $\nabla^2 g(\theta)$, $\theta \in \Theta$. If there exists $\beta > 0$ such that for any $\theta \in \Theta$, the matrix*

$$\beta \nabla^2 g(\theta) - \nabla g(\theta)(\nabla g(\theta))^\top ,$$

is positive semi-definite then $-g(\cdot)/\beta$ is exponentially concave on the simplex Θ.

PROOF. Since g is twice differentiable $\exp(-g(\cdot)/\beta)$ is also twice differentiable with Hessian matrix

$$\mathcal{H}(\theta) = \frac{1}{\beta} \exp\left(-\frac{g(\theta)}{\beta}\right) \left[\frac{\nabla g(\theta)(\nabla g(\theta))^\top}{\beta} - \nabla^2 g(\theta)\right]. \tag{4.1}$$

For any $\lambda \in \mathbb{R}^M$, $\theta \in \Theta$, we have

$$\lambda^\top \mathcal{H}(\theta)\lambda = \frac{1}{\beta} \exp\left(-\frac{g(\theta)}{\beta}\right) \left[\frac{(\lambda^\top \nabla g(\theta))^2}{\beta} - \lambda^\top [\nabla^2 g(\theta)]\lambda\right] \leq 0.$$

Hence $\exp(-g(\cdot)/\beta)$ has a negative semi-definite Hessian and is therefore concave. ∎

4.1 Applications of Theorem 3.1

We begin with the models that satisfy assumptions of Theorem 3.1.

1. Regression with squared loss. Let $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ where $\mathcal{X}$ is a complete separable metric space equipped with its Borel σ-algebra. Consider a random variable $Z = (X, Y)$ with $X \in \mathcal{X}$ and $Y \in \mathbb{R}$. Assume that the conditional expectation $f(X) = E(Y|X)$ exists and define $\xi = Y - E(Y|X)$, so that

$$Y = f(X) + \xi, \tag{4.2}$$

where $X \in \mathcal{X}$ is a random variable with probability distribution P_X, $Y \in \mathbb{R}$, $f : \mathcal{X} \to \mathbb{R}$ is the regression function and ξ is a real valued random variable satisfying $E(\xi|X) = 0$. Assume that $E(Y^2) < \infty$ and $\|f\|_\infty \leq L$ for some finite constant $L > 0$ where $\|\cdot\|_\infty$ denotes the $L_\infty(P_X)$-norm. We dispose of M functions $f_1, \ldots, f_M$ such that $\|f_j\|_\infty \leq L$, $j = 1, \ldots, M$. Define $\|f\|^2_{2,P_X} = \int_{\mathcal{X}} f^2(x) P_X(\mathrm{d}x)$. Our goal is to construct an aggregate that mimics the oracle $\min_{1\leq j\leq M} \|f_j - f\|^2_{2,P_X}$. The aggregate is based on the i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ where (X_i, Y_i) have the same distribution as (X, Y). For this model, with $z = (x, y) \in \mathcal{X} \times \mathbb{R}$, define the loss function

$$Q(z, \theta) = \big(y - \theta^\top H(x)\big)^2, \ \forall \theta \in \Theta ,$$

with $H(x) = \big(f_1(x), \ldots, f_M(x)\big)^\top$. It yields for all $z \in \mathcal{Z}, \theta, \theta' \in \Theta$,

$$Q_1(z, \theta, \theta') = Q(z, \theta) - Q(z, \theta') = 2y(\theta' - \theta)^\top H(x) + \big[\theta^\top H(x)\big]^2 - \big[\theta'^\top H(x)\big]^2$$

Consider positive constants b and B and assume that $\beta > (b/B)^2$. We now decompose Q_1 into the sum $Q_1 = Q_2 + Q_3$ where

$$Q_2(z, \theta, \theta') = 2y \mathbb{1}_{\{|y|<B\beta\}}(\theta' - \theta)^\top H(x) + \big[\theta^\top H(x)\big]^2 - \big[\theta'^\top H(x)\big]^2 + \frac{y^2}{B\beta}\big[(\theta' - \theta)^\top H(x)\big]^2 \mathbb{1}_{\{b\sqrt{\beta}<|y|<B\beta\}},$$

and

$$Q_3(z,\theta,\theta') = 2y\mathbb{1}_{\{|y|\geq B\beta\}}(\theta'-\theta)^\top H(x) - \frac{y^2}{B\beta}\big[(\theta'-\theta)^\top H(x)\big]^2\mathbb{1}_{\{b\sqrt{\beta}<|y|<B\beta\}}.$$

We have

$$-Q_3(z,\theta,\theta') \leq 4L|y|\mathbb{1}_{\{|y|\geq B\beta\}} + \frac{4L^2y^2}{B\beta}\mathbb{1}_{\{b\sqrt{\beta}<|y|<B\beta\}} \triangleq R_\beta(y)\,. \tag{4.3}$$

On the other hand, $Q_2(z,\theta,\theta)=0, \forall\theta\in\Theta, z\in\mathcal{Z}$ and we can prove that the mapping $\theta\mapsto -Q_2(z,\theta,\theta')/\beta$ is exponentially concave for any $z\in\mathcal{Z}, \theta'\in\Theta$ when b and B are properly chosen. For all $\theta\in\Theta$ and $z=(x,y)$ the gradient and Hessian of Q_2 are respectively given by

$$\nabla_\theta Q_2 = \nabla_\theta Q_2(z,\theta,\theta') = -2\big(y\mathbb{1}_{\{|y|<B\beta\}} - \theta^\top H(x)\big)H(x) - 2\frac{y^2}{B\beta}\mathbb{1}_{\{b\sqrt{\beta}<|y|<B\beta\}}\big[(\theta'-\theta)^\top H(x)\big]H(x)$$

and

$$\nabla^2_{\theta\theta}Q_2 = \nabla^2_{\theta\theta}Q_2(z,\theta,\theta') = 2H(x)H(x)^\top + 2\frac{y^2}{B\beta}\mathbb{1}_{\{b\sqrt{\beta}<|y|<B\beta\}}H(x)H(x)^\top\,.$$

We now prove that Proposition 4.1 applies for $g(\theta)=Q_2(z,\theta,\theta')$, for all $z=(x,y)\in\mathcal{Z}$ and $\theta'\in\Theta$. For any $\lambda\in\mathbb{R}^M$, any $\theta,\theta'\in\Theta$, and any $z\in\mathcal{Z}$,

$$\big(\lambda^\top\nabla_\theta Q_2\big)^2 \leq \Big(2|y|\mathbb{1}_{\{|y|<B\beta\}} + 2L + \frac{4Ly^2}{B\beta}\mathbb{1}_{\{b\sqrt{\beta}<|y|<B\beta\}}\Big)^2\big[\lambda^\top H(x)\big]^2.$$

Note now that $|y|\leq B\beta$ implies that $y^2/B\beta\leq|y|$. Hence

$$\begin{aligned}\big(\lambda^\top\nabla_\theta Q_2\big)^2 &\leq \Big(2|y|\mathbb{1}_{\{|y|\leq b\sqrt{\beta}\}} + 2L + (4L+2)|y|\mathbb{1}_{\{b\sqrt{\beta}<|y|<B\beta\}}\Big)^2\big[\lambda^\top H(x)\big]^2\\ &\leq \Big(8b^2\beta + 8L^2 + 2(4L+2)^2|y|^2\mathbb{1}_{\{b\sqrt{\beta}<|y|<B\beta\}}\Big)\big[\lambda^\top H(x)\big]^2\end{aligned}$$

Therefore

$$\frac{\big(\lambda^\top\nabla_\theta Q_2\big)^2}{\beta} - \lambda^\top(\nabla^2_{\theta\theta}Q_2)\lambda \leq \Big(8b^2 + \frac{8L^2}{\beta} - 2 + \Big[2(4L+2)^2 - \frac{2}{B}\Big]\frac{|y|^2}{\beta}\mathbb{1}_{\{b\sqrt{\beta}<|y|<B\beta\}}\Big)\big[\lambda^\top H(x)\big]^2$$

If we choose $B\leq(4L+2)^{-2}$ and $LB<b<1/4$, the above quadratic form is smaller than or equal to 0 and Proposition 4.1 applies for any $\beta>(b/B)^2$. Now, since $A(\theta)=EQ(Z,\theta)=E\big(Y-\theta^\top H(X)\big)^2 = \|f-\theta^\top H\|^2_{2,P_X} + E(\xi^2)$ for all $\theta\in\Theta$, we obtain the following corollary of Theorem 3.1.

Corollary 4.1 *Consider the regression model* (4.2) *where* $X\in\mathcal{X}$, $Y\in\mathbb{R}$, $f:\mathcal{X}\to\mathbb{R}$ *and* $\xi=Y-f(X)$ *is a real valued random variable satisfying* $E(\xi|X)=0$. *Assume also that* $E\big(Y^2\big)<\infty$ *and* $\|f_j\|_\infty\leq L$, $j=1,\ldots,M$, *for some finite constant* $L>0$. *Then for any positive constants* $B\geq(4L+2)^{-2}, LB<b<1/4$ *and any* $\beta\geq(b/B)^2$, *the aggregate estimator* $\tilde{f}_n(x)=\hat{\theta}_n^\top H(x)$, $x\in\mathcal{X}$, *where* $\hat{\theta}_n$ *is obtained by the mirror averaging algorithm, satisfies*

$$E_{n-1}\|\tilde{f}_n - f\|^2_{2,P_X} \leq \min_{1\leq j\leq M}\|f_j - f\|^2_{2,P_X} + \frac{\beta\log M}{n} + E\big[R_\beta(Y)\big], \tag{4.4}$$

where

$$R_\beta(y) = 4L|y|\mathbb{1}_{\{|y|\geq B\beta\}} + \frac{4L^2y^2}{B\beta}\mathbb{1}_{\{b\sqrt{\beta}<|y|<B\beta\}}\,.$$

This result improves an inequality obtained by Bunea and Nobel (2005). We note that the aggregate $\tilde{f}_n$ as in Corollary 4.1 is of the form suggested by Catoni (1999, 2004). If there exists a constant $L_0 > 0$ such that $|Y| \le L_0$ a.s., the last summand disappears for $\beta > 16L_0^2$, and (4.4) follows from Catoni (1999, 2004), under a more restrictive assumption on β.

An advantage of Corollary 4.1 is that no heavy assumption on the moments of ξ is needed to get reasonable bounds. Thus, the second moment assumption on Y is enough for a bound with the $n^{-1/2}$ rate. Indeed, choosing $\beta \sim (n/\log M)^{2/(2+s)}$, $s > 0$, in Corollary 4.1, we immediately get the following result.

Corollary 4.2 *Consider the regression model* (4.2) *where* $X \in \mathcal{X}$, $Y \in \mathbb{R}$, $f : \mathcal{X} \to \mathbb{R}$ *and* $\xi = Y - f(X)$ *is a real valued random variable satisfying* $E(\xi|X) = 0$. *Assume also that* $E(|Y|^s) \le m_s < \infty$ *for some* $s \ge 2$ *and* $\|f_j\|_\infty \le L$, $j = 1, \dots, M$, *for some finite constant* $L > 0$. *Then there exist constants* $C_1 > 0$ *and* $C_2 = C_2(m_s, L, C_1) > 0$ *such that the aggregate estimator* $\tilde{f}_n(x) = \hat{\theta}_n^\top H(x)$, $x \in \mathcal{X}$, *where* $\hat{\theta}_n$ *is obtained by the mirror averaging algorithm with* $\beta = C_1(n/\log M)^{2/(2+s)}$, *satisfies*

$$E_{n-1}\|\tilde{f}_n - f\|_{2,P_X}^2 \le \min_{1\le j\le M} \|f_j - f\|_{2,P_X}^2 + C_2 \left(\frac{\log M}{n}\right)^{s/(2+s)} . \tag{4.5}$$

2. Classification. Consider the problem of binary classification. Let $(\mathcal{X}, \mathcal{F})$ be a measurable space, and set $\mathcal{Z} = \mathcal{X} \times \{-1, 1\}$. Consider $Z = (X, Y)$ where X is a random variable with values in $\mathcal{X}$ and Y is a random label with values in $\{-1, 1\}$. For a fixed convex twice differentiable function $\varphi : \mathbb{R} \to \mathbb{R}_+$, define the φ-risk of a real valued classifier $h : \mathcal{X} \to [-1, 1]$ as $E\varphi(-Yh(X))$. In our framework, we dispose of M such classifiers $h_1, \dots, h_M$ and the goal is to mimic the oracle $\min_{1\le j\le M} E\varphi(-Yh_j(X))$ based on the i.i.d. sample $(X_1, Y_1), \dots, (X_n, Y_n)$ where (X_i, Y_i) have the same distribution as (X, Y). For any $z = (x, y) \in \mathcal{X} \times \{-1, 1\}$, we define the loss function

$$Q(z, \theta) = \varphi(-y\theta^\top H(x)) \ge 0, \ \forall\, \theta \in \Theta ,$$

where $H(x) = \big(h_1(x), \dots, h_M(x)\big)^\top$. For such a function and for all $\theta \in \mathbb{R}^M, z = (x, y) \in \mathcal{X} \times \{-1, 1\}$ we have

$$\nabla_\theta Q_1(z, \theta, \theta') = -y\varphi'\big(-y\theta^\top H(x)\big)H(x) \quad \text{and} \quad \nabla^2_{\theta\theta} Q_1(z, \theta, \theta') = \varphi''\big(-y\theta^\top H(x)\big)H(x)H(x)^\top .$$

Thus, from Proposition 4.1 the mapping $\theta \mapsto -Q_1(z, \theta, \theta')/\beta$ is exponentially concave for all z and θ' if $\beta \ge \beta_\varphi$ where β_φ is such that $\big[\varphi'(x)\big]^2 \le \beta_\varphi \varphi''(x)$, $\forall\, |x| \le 1$. Now, since $A(\theta) = EQ(Z, \theta)$ and $Q(Z, \theta) = \varphi(-Y\theta^\top H(X))$, $\forall\, \theta \in \mathbb{R}^M$, $Z = (X, Y)$, we obtain the following corollary of Theorem 3.1 applied with $Q_2 = Q_1$ and $Q_3 \equiv 0$.

Corollary 4.3 *Consider the binary classification problem as described above. Assume that the convex function* φ *is such that* $\big[\varphi'(x)\big]^2 \le \beta_\varphi \varphi''(x)$, $\forall\, |x| \le 1$. *Then the aggregate classifier* $\tilde{h}_n(x) = \hat{\theta}_n^\top H(x)$, $x \in \mathcal{X}$, *where* $\hat{\theta}_n$ *is obtained by the mirror averaging algorithm with* $\beta \ge \beta_\varphi$, *satisfies*

$$E_n\varphi(-Y_n\tilde{h}_n(X_n)) \le \min_{1\le j\le M} E\varphi(-Y\tilde{h}_j(X)) + \frac{\beta \log M}{n} . \tag{4.6}$$

For example, inequality (4.6) holds with the exponential Boosting loss $\varphi_1(x) = \mathrm{e}^x$, for which $\beta_{\varphi_1} = \mathrm{e}$ and for the Logit-Boosting loss $\varphi_2(x) = \log_2(1 + \mathrm{e}^x)$ (in that case $\beta_{\varphi_2} = \mathrm{e}\log 2$). For the squared loss $\varphi_3(x) = (1 - x)^2$ and the 2-norm soft margin loss $\varphi_4(x) = \max\{0, 1 - x\}^2$ inequality (4.6) is satisfied with $\beta \ge 2$.

3. Nonparametric density estimation with Kullback-Leibler (KL) loss. Let X be a random variable with values in a measurable space $(\mathcal{X}, \mathcal{F})$. Assume that the distribution of X admits a density p w.r.t. a σ-finite measure μ on $(\mathcal{X}, \mathcal{F})$. Assume also that we dispose of M probability densities p_j w.r.t. μ on $(\mathcal{X}, \mathcal{F})$ (estimators of p) and of an i.i.d. sample $X_1, \ldots, X_n$ where X_i take values in $\mathcal{X}$, and have the same distribution as X. Define the KL divergence between two probability densities p and q w.r.t. μ as

$$\mathcal{K}(p, q) \triangleq \int_{\mathcal{X}} \log\left(\frac{p(x)}{q(x)}\right) p(x)\mu(\mathrm{d}x)\,,$$

if the probability distribution corresponding to p is absolutely continuous w.r.t. the one corresponding to q, and $\mathcal{K}(p, q) = \infty$ otherwise. We assume that the entropy integral $\int p(x) \log p(x)\mu(\mathrm{d}x)$ is finite.

Our goal is to construct an aggregate that mimics the KL oracle defined by $\min_{1\le j\le M} \mathcal{K}(p, p_j)$. For $x \in \mathcal{X}, \theta \in \Theta$, we introduce the corresponding loss function

$$Q(x, \theta) = -\log(\theta^\top H(x)),$$

where $H(x) = \big(p_1(x), \ldots, p_M(x)\big)^\top$. We set $Z = X$. Then $A(\theta) = EQ(X, \theta) = -\int \log(\theta^\top H(x))p(x)\mu(\mathrm{d}x)$, where the integral is finite if all the divergences $\mathcal{K}(p, p_j)$ are finite. In particular, $A(e_j) = \mathcal{K}(p, p_j) - \int p(x) \log p(x)\mu(\mathrm{d}x)$. Since, for all $x \in \mathcal{X}$, $\exp(-Q_1(x, \theta, \theta')/\beta) = \big(\theta^\top H(x)\big)^{1/\beta}\big(\theta'^\top H(x)\big)^{-1/\beta}$, the mapping $\theta \mapsto -Q_1(x, \theta, \theta')/\beta$ is exponentially concave on Θ for any $\beta \ge 1$. Hence, we can apply Theorem 3.1, again with $Q_2 = Q_1$ and $Q_3 \equiv 0$ and we obtain the following corollary.

Corollary 4.4 *Consider the density estimation problem with the KL loss as described above, such that $\int p(x)|\log p(x)|\mu(\mathrm{d}x) < \infty$. Then the aggregate estimator $\tilde{p}_n(x) = \hat{\theta}_n^\top H(x)$, $x \in \mathcal{X}$, where $\hat{\theta}_n$ is obtained by the mirror averaging algorithm with $\beta = 1$, satisfies*

$$E_{n-1}\mathcal{K}(p, \tilde{p}_n) \le \min_{1\le j\le M} \mathcal{K}(p, p_j) + \frac{\log M}{n}\,.$$

We note that the KL aggregate $\tilde{p}_n$ as in Corollary 4.4 coincides with the "progressive mixture rule" considered by Catoni (1997, 1999, 2004) and Yang (2000) and the oracle inequality of Corollary 4.4 is the one obtained in those papers. Extension of Corollary 4.4 to $\beta \ge 1$ is straightforward but the oracle inequality for the corresponding aggregate ("Gibbs estimator", cf. Catoni (2004)) is less interesting because it has obviously a larger remainder term.

4.2 Applications of Theorem 3.2

4. Regression with squared loss and finite exponential moment. We consider here the regression model described in Corollary 4.1 under the additional assumption that, conditionally on X, the regression residual ξ admits an exponential moment, i.e., there exist positive constants b and D such that, P_X-a.s.,

$$E\big(\exp(b|\xi|)|X\big) \le D.$$

Since $E(\xi|X) = 0$, this assumption is equivalent to the existence of positive constants b_0 and σ^2 such that, P_X-a.s.,

$$E\big(\exp(t\xi)|X\big) \le \exp(\sigma^2 t^2/2), \quad \forall\ |t| \le b_0, \tag{4.7}$$

[cf. Petrov (1995), p. 56].

In this case, application of Corollary 4.1 leads to suboptimal rates because of the term $E\big[R_\beta(Y)\big]$ in (4.4). We show now that using Theorem 3.2 we can obtain an oracle inequality with optimal rate $(\log M)/n$.

To apply Theorem 3.2, we analyze the mapping $\theta \mapsto E\exp\big(-Q_1(Z,\theta,\theta')/\beta\big)$. For the regression model with squared loss as described above, we have $Z=(X,Y)$, $Q(Z,\theta)=(Y-\theta^\top H(X))^2$, and

$$\begin{aligned} E\exp\big(-Q_1(Z,\theta,\theta')/\beta\big) &= E\exp\bigg(-\frac{1}{\beta}\Big[(Y-H^\top(X)\theta)^2-(Y-H^\top(X)\theta')^2\Big]\bigg) \\ &= E\exp\bigg(-\frac{1}{\beta}\Big[-2\xi(U(X,\theta)-U(X,\theta'))+U^2(X,\theta)-U^2(X,\theta')\Big]\bigg), \end{aligned}$$

where $U(X,\theta)\triangleq f(X)-H^\top(X)\theta$. Since $|2(U(X,\theta)-U(X,\theta'))|=2|(\theta-\theta')^\top H(X)|\le 4L$, conditioning on X and using (4.7) we get that, for any $\beta\ge 4L/b_0$,

$$E\exp\big(-Q_1(Z,\theta,\theta')/\beta\big)\le \Psi_\beta(\theta,\theta')$$

where

$$\Psi_\beta(\theta,\theta')\triangleq E\exp\bigg(\frac{2\sigma^2}{\beta^2}\big[(\theta-\theta')^\top H(X)\big]^2-\frac{1}{\beta}\big[U^2(X,\theta)-U^2(X,\theta')\big]\bigg).$$

Clearly, $\Psi_\beta(\theta,\theta)=1$. Thus, to apply Theorem 3.2 it suffices now to specify $\beta_0>0$ such that the mapping

$$\theta\mapsto \bar{Q}(x,\theta,\theta')\triangleq\left(-\frac{1}{\beta}+\frac{2\sigma^2}{\beta^2}\right)(\theta^\top H(x))^2-\frac{4\sigma^2}{\beta^2}(H(x)^\top\theta)(H(x)^\top\theta')+\frac{2}{\beta}f(x)(H(x)^\top\theta)$$

is exponentially concave for all $\beta\ge\beta_0,\ \theta'\in\Theta$ and almost all $x\in\mathcal{X}$. Note that

$$\begin{aligned} \nabla_\theta\bar{Q}(x,\theta,\theta') &= \left(-2\gamma(H^\top(x)\theta)-\frac{4\sigma^2}{\beta^2}(H^\top(x)\theta')+\frac{2}{\beta}f(x)\right)H(x), \\ \nabla^2_{\theta\theta}\bar{Q}(x,\theta,\theta') &= -2\gamma H(x)H^\top(x) \end{aligned}$$

where $\gamma=\frac{1}{\beta}-\frac{2\sigma^2}{\beta^2}$. Proposition 4.1 implies that $\bar{Q}$ is exponentially concave in θ if $\nabla^2_{\theta\theta}\bar{Q}(x,\theta,\theta')+\nabla_\theta\bar{Q}(x,\theta,\theta')(\nabla_\theta\bar{Q}(x,\theta,\theta'))^\top\ge 0$. Using the inequalities $|H^\top(x)\theta|\le L$, $|H^\top(x)\theta'|\le L$, $|f(x)|\le L$ we obtain that the latter property holds for $\beta\ge\beta_0\triangleq 2\sigma^2+8L^2$. Thus, Theorem 3.2 applies for $\beta\ge\max(2\sigma^2+8L^2,\,4L/b_0)$ and we have proved the following result.

Corollary 4.5 *Consider the regression model* (4.2) *where $X\in\mathcal{X}$, $Y\in\mathbb{R}$, $f:\mathcal{X}\to\mathbb{R}$ and the random variable $\xi=Y-f(X)$ is such that there exist positive constants b_0 and σ^2 for which (4.7) holds P_X-a.s. Assume also that $\|f\|_\infty\le L$ and $\|f_j\|_\infty\le L$, $j=1,\dots,M$, for some finite constant $L>0$. Then for any $\beta\ge\max(2\sigma^2+8L^2,\,4L/b_0)$ the aggregate estimator $\tilde{f}_n(x)=\hat{\theta}_n^\top H(x)$, $x\in\mathcal{X}$, where $\hat{\theta}_n$ is obtained by the mirror averaging algorithm, satisfies*

$$E_{n-1}\|\tilde{f}_n-f\|^2_{2,P_X}\le\min_{1\le j\le M}\|f_j-f\|^2_{2,P_X}+\frac{\beta\log M}{n}\,. \tag{4.8}$$

To see how good the constants are, we may compare this corollary with the results obtained in other papers for the particular case where ξ is conditionally Gaussian given X. In this case we have $b_0=\infty$ and Corollary 4.5 yields the following result.

Corollary 4.6 *Consider the regression model* (4.2) *where $X\in\mathcal{X}$, $Y\in\mathbb{R}$, $f:\mathcal{X}\to\mathbb{R}$ and, conditionally on X, the random variable $\xi=Y-f(X)$ is Gaussian with zero mean and variance bounded by σ^2. Assume also that $\|f\|_\infty\le L$ and $\|f_j\|_\infty\le L$, $j=1,\dots,M$, for some finite constant $L>0$. Then for any $\beta\ge 2\sigma^2+8L^2$ the aggregate estimator $\tilde{f}_n(x)=\hat{\theta}_n^\top H(x)$, $x\in\mathcal{X}$, where $\hat{\theta}_n$ is obtained by the mirror averaging algorithm, satisfies (4.8).*

This corollary improves upon the best known bound [Catoni (2004, p.89)] where the condition on β is $\beta \geq 18.01\sigma^2 + 70.4L^2$.

5. Nonparametric density estimation with the L_2 loss. Let μ be a σ-finite measure on the measurable space $(\mathcal{X}, \mathcal{F})$. In this whole example, densities are understood w.r.t μ and $\|\cdot\|_\infty$ denotes the $L_\infty(\mu)$-norm. Assume that we dispose of M probability densities p_j, $\|p_j\|_\infty \leq L$, $j = 1, \ldots, M$, and of an i.i.d. sample $X_1, \ldots, X_n$ where X_i take values in $\mathcal{X}$, and are distributed as a random variable X with unknown probability density p such that $\|p\|_\infty \leq L$ for some positive constant L. Our goal is to mimic the oracle defined by $\min_{1\leq j\leq M} \|p_j - p\|_2^2$ where $\|p\|_2^2 = \int p^2(x)\mu(\mathrm{d}x)$.

The corresponding loss function is defined, for any $x \in \mathcal{X}, \theta \in \Theta$, by

$$Q(x, \theta) = \theta^\top G\theta - 2\theta^\top H(x)\,, \tag{4.9}$$

where $H(x) = \big(p_1(x), \ldots, p_M(x)\big)^\top$ and G is an $M \times M$ positive semi-definite matrix with elements $G_{jk} = \int p_j p_k \mathrm{d}\mu \leq L$. We set $Z = X$. Then $A(\theta) = EQ(X, \theta) = \|p - \theta^\top H\|_2^2 - \|p\|_2^2$. We now want to check conditions of Theorem 3.2, i.e., to show that for the loss function (4.9), the mapping $\theta \mapsto E\exp\big(-Q_1(X, \theta, \theta')/\beta\big)$ is concave on Θ, for any $\theta' \in \Theta$ and for $\beta \geq \beta_0$ with some $\beta_0 > 0$ that will be specified below. Note first that

$$Q_1(x, \theta, \theta') = Q(x, \theta) - Q(x, \theta') = (\theta - \theta')^\top G(\theta + \theta') - 2(\theta - \theta')^\top H(x). \tag{4.10}$$

Fix $\theta' \in \Theta$. Concavity of the above mapping can be checked by considering its Hessian $\tilde{\mathcal{H}}$ which, in view of (4.1), satisfies

$$\lambda^\top \tilde{\mathcal{H}}(\theta)\lambda = \frac{1}{\beta^2} E\bigg\{ \exp\left(-\frac{Q_1(X, \theta, \theta')}{\beta}\right) \Big[(\lambda^\top \nabla_\theta Q_1(X, \theta, \theta'))^2 - \beta\lambda^\top \nabla^2_{\theta\theta} Q_1(X, \theta, \theta')\lambda\Big]\bigg\}, \quad \forall\, \lambda \in \mathbb{R}^M,\ \theta \in \Theta\,.$$

Note that for any $x \in \mathcal{X}, \theta \in \Theta$ we have

$$\nabla_\theta Q_1(x, \theta, \theta') = 2G\theta - 2H(x) \quad \text{and} \quad \nabla^2_{\theta\theta} Q_1(x, \theta, \theta') = 2G\,.$$

By (4.10) this yields, for any $\lambda \in \mathbb{R}^M$, $\theta, \theta' \in \Theta$,

$$\begin{aligned} \lambda^\top \tilde{\mathcal{H}}(\theta)\lambda &= -\frac{2}{\beta^2} E\left\{\exp\left(-\frac{(\theta - \theta')^\top G(\theta + \theta') - 2(\theta - \theta')^\top H(X)}{\beta}\right)\big[\beta\lambda^\top G\lambda - 2(\lambda^\top(G\theta - H(X)))^2\big]\right\} \\ &\leq -\frac{2}{\beta^2}\exp\left(-\frac{(\theta - \theta')^\top G(\theta + \theta')}{\beta}\right) F(\lambda, \theta, \theta'), \end{aligned} \tag{4.11}$$

where

$$F(\lambda, \theta, \theta') = E\left\{\exp\left(\frac{2(\theta - \theta')^\top H(X)}{\beta}\right)\big[\beta\lambda^\top G\lambda - 4(\lambda^\top G\theta)^2 - 4(\lambda^\top H(X))^2\big]\right\}$$

Observe that by the Cauchy inequality

$$(\lambda^\top G\theta)^2 \leq \lambda^\top G\lambda\, \theta^\top G\theta \leq L\lambda^\top G\lambda, \ \ \forall \theta \in \Theta. \tag{4.12}$$

Further,

$$E(\lambda^\top H(X))^2 = \int \big(\lambda^\top H(x)\big)^2 p(x)\mu(\mathrm{d}x) \leq L\int \big(\lambda^\top H(x)\big)^2 \mu(\mathrm{d}x) = L\lambda^\top G\lambda. \tag{4.13}$$

Using (4.12) and (4.13) and the fact that $\|\theta-\theta'\|_1 \le 2$ where $\|\cdot\|_1$ stands for the $\ell_1(\mathbb{R}^M)$ norm, we obtain

$$\begin{aligned} F(\lambda,\theta,\theta') &\ge (\beta-4L)\lambda^\top G\lambda E\exp\left(\frac{2(\theta-\theta')^\top H(X)}{\beta}\right) - 4E\left\{\exp\left(\frac{2(\theta-\theta')^\top H(X)}{\beta}\right)(\lambda^\top H(X))^2\right\} \\ &\ge (\beta-4L)\lambda^\top G\lambda\,\exp\left(-\frac{4L}{\beta}\right) - 4L\lambda^\top G\lambda\,\exp\left(\frac{4L}{\beta}\right) \ge 0 \end{aligned}$$

provided that

$$\frac{\beta-4L}{4L}\exp\left(-\frac{8L}{\beta}\right) \ge 1.$$

Note that the last inequality is guaranteed for $\beta \ge \beta_0 = 12L$. We conclude that for $\beta \ge 12L$ the Hessian $\tilde{\mathcal{H}}$ in (4.11) is negative semi-definite and therefore the mapping $\theta \mapsto E\exp\big(-Q_1(X,\theta,\theta')/\beta\big)$ is concave on Θ for any fixed $\theta' \in \Theta$. Thus we have proved the following corollary of Theorem 3.2.

Corollary 4.7 *Consider the density estimation problem with the L_2 loss as described above. Then, for any $\beta \ge 12L$, the aggregate estimator $\tilde{p}_n(x) = \hat{\theta}_n^\top H(x)$, $x \in \mathcal{X}$, where $\hat{\theta}_n$ is obtained by the mirror averaging algorithm, satisfies*

$$E_{n-1}\|\tilde{p}_n - p\|_2^2 \le \min_{1\le j\le M}\|p_j - p\|_2^2 + \frac{\beta\log M}{n}\,.$$

6. Parametric estimation with Kullback-Leibler (KL) loss. Let $\mathcal{P} = \{P_a,\, a \in \mathcal{A}\}$ be a family of probability measures on a measurable space $(\mathcal{X},\mathcal{F})$ dominated by a σ-finite measure μ on $(\mathcal{X},\mathcal{F})$. Here $\mathcal{A} \subset \mathbb{R}^m$ is a bounded set of parameters. The densities relative to μ are denoted by $p(x,a) = (\mathrm{d}P_a/\mathrm{d}\mu)(x)$, $x \in \mathcal{X}$. Let X be a random variable with values in $\mathcal{X}$ distributed according to P_{a^*} where $a^* \in \mathcal{A}$ is the unknown true value of the parameter.

In the aggregation framework, we dispose of M values $a_1,\dots,a_M \in \mathcal{A}$ (preliminary estimators of a) and of an i.i.d. sample $X_1,\dots,X_n$ where X_i take values in $\mathcal{X}$, and have the same distribution as X. Our goal is to construct an aggregate $\tilde{a}_n$ that mimics the parametric KL oracle defined by $\min_{1\le j\le M} K(a^*,a_j)$ where

$$K(a,b) \triangleq \mathcal{K}\big(p(\cdot,a),p(\cdot,b)\big)\,, \quad \forall\, a,b \in \mathcal{A}.$$

For $x \in \mathcal{X}, \theta \in \Theta$, we introduce the corresponding loss function

$$Q(x,\theta) = -\log p(x,\theta^\top H),$$

where $H = (a_1,\dots,a_M)^\top$. We set $Z = X$. Then $A(\theta) = EQ(X,\theta) = -\int \log\big(p(x,\theta^\top H)\big)p(x,a^*)\mu(\mathrm{d}x)$, $A(e_j) = K(a^*,a_j) - \int p(x,a^*)\log\big(p(x,a^*)\big)\mu(\mathrm{d}x)$. Since, for all $x \in \mathcal{X}$, $\exp(-Q(x,\theta)/\beta) = \big(p(x,\theta^\top H)\big)^{1/\beta}$, to apply Theorem 3.2 we need the following assumption.

Assumption 4.1 *For some $\beta > 0$ and for any $a \in \mathcal{A}$ there exists a Borel function $\Psi_\beta : \Theta\times\Theta \to \mathbb{R}_+$ such that $\theta \mapsto \Psi_\beta(\theta,\theta')$ is concave on the simplex Θ for all $\theta' \in \Theta$, $\Psi_\beta(\theta,\theta) = 1$ and*

$$\int\left(\frac{p(x,H^\top\theta)}{p(x,H^\top\theta')}\right)^{1/\beta} p(x,a)\mu(\mathrm{d}x) \le \Psi_\beta(\theta,\theta')$$

for all $\theta,\theta' \in \Theta$.

Corollary 4.8 *Consider the parametric estimation problem with the KL loss as described above and let* $\int p(x,a^*)|\log p(x,a^*)|\mu(\mathrm{d}x) < \infty$. *Suppose that Assumption 4.1 is fulfilled for some* $\beta > 0$. *Then the aggregate estimator* $\tilde{a}_n = \hat{\theta}_n^\top H$ *of the parameter* a^*, *where* $\hat{\theta}_n$ *is obtained by the mirror averaging algorithm, satisfies*

$$E_{n-1}K(a^*,\tilde{a}_n) \le \min_{1\le j\le M} K(a^*,a_j) + \frac{\beta\log M}{n}\,. \tag{4.14}$$

Aggregation procedures can be used to construct pointwise adaptive locally parametric estimators in nonparametric regression [cf. Belomestny and Spokoiny (2004)]. In this case inequality (4.14) can be applied to prove the corresponding adaptive risk bounds. We now check that Assumption 4.1 is satisfied for several standard parametric families.

- **Univariate Gaussian distribution.** Let μ be the Lebesgue measure on $\mathbb{R}$ and let $p(x,a) = (\sigma\sqrt{2\pi})^{-1}\exp(-(x-a)^2/(2\sigma^2))$ be the univariate Gaussian density with mean $a \in \mathcal{A} = [-L,L]$ and known variance $\sigma^2 > 0$. Replacing $f(x)$ by a^* and $H(x)$ by H in the proof of Corollary 4.6, and following exactly the same argument as there we find that Assumption 4.1 is satisfied for any $\beta \ge \beta_0 = 2\sigma^2 + 8L^2$. Hence, (4.14) also holds for such β. Note that in this case $K(a^*,a) = (a^*-a)^2/(2\sigma^2)$.

- **Bernoulli distribution.** Let μ be the discrete measure on $\{0,1\}$ such that $\mu(0) = \mu(1) = 1$ and let $p(x,a) = a\mathbb{1}_{\{x=0\}} + (1-a)\mathbb{1}_{\{x=1\}}$ be the density of a Bernoulli random variable with parameter $a \in \mathcal{A} = (0,1)$. Then

$$\int\left(\frac{p(x,H^\top\theta)}{p(x,H^\top\theta')}\right)^{1/\beta} p(x,a)\mu(\mathrm{d}x) = \left(\frac{H^\top\theta}{H^\top\theta'}\right)^{1/\beta} a + \left(\frac{1-H^\top\theta}{1-H^\top\theta'}\right)^{1/\beta}(1-a) \triangleq \Psi_\beta(\theta,\theta')\,.$$

  This function is concave in θ for any $\theta' \in \Theta$ if $\beta \ge 1$ and obviously $\Psi_\beta(\theta,\theta) = 1$. Therefore Assumption 4.1 is satisfied and Corollary 4.8 applies with $\beta = 1$.

- **Poisson distribution.** Let μ be the counting measure on the set of the nonnegative integers $\mathbb{N}$: $\mu(k) = 1,\ \forall\, k \in \mathbb{N}$, and let $p(x,a) = \sum_{k=0}^{\infty}\frac{a^k}{k!}\mathrm{e}^{-a}\mathbb{1}_{\{x=k\}}$ be the density of a Poisson random variable with parameter $a \in \mathcal{A} = [\ell,L]$ where $0 < \ell < L < \infty$. Then

$$\int\left(\frac{p(x,H^\top\theta)}{p(x,H^\top\theta')}\right)^{1/\beta} p(x,a)\mu(\mathrm{d}x) = \exp\left[a\left(\frac{H^\top\theta}{H^\top\theta'}\right)^{1/\beta} - a - \frac{H^\top(\theta-\theta')}{\beta}\right] \triangleq \Psi_\beta(\theta,\theta')\,. \tag{4.15}$$

  Clearly, $\Psi_\beta(\theta,\theta) = 1$ and it is not hard to show that Ψ_β in (4.15) is concave as a function of θ for any $\theta' \in \Theta$, provided that $\beta \ge 1 + L(1+L/\ell)(L/\ell)^{1/(2L+1)}$. Therefore Assumption 4.1 is satisfied and Corollary 4.8 applies with $\beta \ge \beta_0 = 1 + L(1+L/\ell)(L/\ell)^{1/(2L+1)}$.

Acknowledgement. We would like to thank J.-Y.Audibert for the remark that the moment conditions in Corollaries 4.1 and 4.2 can be improved.

References

[1] Bartlett, P.L., Boucheron, S. and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, **48**, 85 – 113.

[2] Belomestny, D. and Spokoiny, V. (2004). Local likelihood modeling via stagewise aggregation. Preprint n. 1000. WIAS-Berlin.

[3] Ben-Tal, A. and Nemirovski, A.S. (1999). The conjugate barrier mirror descent method for non-smooth convex optimization, *MINERVA Optim. Center Report.*, Haifa: Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology. `http://iew3.technion.ac.il/Labs/Opt/opt/Pap/CP_MD.pdf`.

[4] Boucheron, S., Bousquet, O. and Lugosi, G. (2005). Theory of classification: some recent advances. *ESAIM Probability & Statistics* **9** 323-375.

[5] Bunea, F. and Nobel, A. (2005). Sequential procedures for aggregating arbitrary estimators of a conditional mean. Manuscript. `http://www.stat.fsu.edu/~flori`.

[6] Bunea, F., Tsybakov, A. and Wegkamp, M. (2004). Aggregation for regression learning. Preprint n.948, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7. arXiv:math.ST/0410214, 8 Oct. 2004 and `http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2004`.

[7] Catoni, O. (1997). A mixture approach to universal model selection. Preprint LMENS-97-30, Ecole Normale Supérieure, `http://www.dmi.ens.fr/preprints/Index.97.html`.

[8] Catoni, O. (1999). "Universal" aggregation rules with exact bias bounds. Preprint n.510, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7 `http://www.proba.jussieu.fr/mathdoc/preprints/index.html#1999`.

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

[10] Cesa-Bianchi, N. and Lugosi, G. (2006) *Prediction, Learning, and Games.* Cambridge Univ. Press, to appear.

[11] Juditsky, A., Nazin, A., Tsybakov, A. and Vayatis, N. (2005). Recursive aggregation of estimators via the mirror descent algorithm with averaging. *Problems of Information Transmission*, **41**, n.4. `www.proba.jussieu.fr/pageperso/vayatis/publication.html`.

[12] Kivinen, J. and Warmuth, M.K. (1999). Averaging expert predictions. In: *Proc. Fourth. European Conf. on Computational Learning Theory*, H.U. Simon and P. Fischer, eds. Lecture Notes in Artificial Intelligence, vol. 1572. Springer, Berlin, 153 – 167.

[13] Leung, G. and Barron, A.R. (2004). Information theory and mixing least-squares regressions. To appear in *IEEE Trans. Inform. Theory.*

[14] Lugosi, G. and Wegkamp, M. (2004). Complexity regularization via localized random penalties. *Ann. Statist.*, **32**, 1679-1697.

[15] Massart, P. (2006). Saint-Flour Lecture Notes. *Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003.* Lecture Notes in Mathematics, Springer, New York, to appear.

[16] Nemirovski, A. (2000). Topics in Non-parametric Statistics. In: *Ecole d'Eté de Probabilités de Saint-Flour XXVIII - 1998.* Lecture Notes in Mathematics, vol. 1738, Springer, New York, 85-277.

[17] Petrov, V.V. (1995) *Limit Theorems of Probability Theory.* Clarendon Press, Oxford.

[18] Samarov, A. and Tsybakov, A. (2005). Aggregation of density estimators and dimension reduction. To appear in *Festschrift in Honor of Kjell Doksum.*

[19] Tsybakov, A. (2003). Optimal rates of aggregation. In: *Computational Learning Theory and Kernel Machines*, (B. Schölkopf and M. Warmuth, eds.), Lecture Notes in Artificial Intelligence, v.2777. Springer, Heidelberg, 303-313.

[20] Vovk,V. (1990) Aggregating strategies. In: *Proc. 3rd Annual Workshop on Computational Learning Theory*, Morgan Kaufman, San Mateo, 372-383.

[21] Wegkamp, M. (2003). Model selection in nonparametric regression. *Ann. Statist.*, **31**, 252-273.

[22] Yang, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.*, **28**, 75-87.

[23] Yang, Y. (2004). Aggregating regression procedures for a better performance. *Bernoulli*, **10**, 25 – 47.

[24] Zhang, T. (2003). From epsilon-entropy to KL-complexity: analysis of minimum information complexity density estimation. Tech. Report RC22980, IBM T.J.Watson Research Center.