Mirror averaging, aggregation and model selection

Alexandre Tsybakov
(joint work with Anatoli Juditsky, Philippe Rigollet)

Several problems in statistics and machine learning can be stated as follows: given a collection of M different estimators (classifiers), construct a new estimator (classifier) 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

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

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

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 $Z_1, \dots, Z_n$ with values in $\mathcal{Z}$ having the same distribution as Z. The distribution (respectively, expectation) w.r.t. the sample $Z_1, \dots, Z_n$ 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

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

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" $\min_{1 \leq j \leq M} A(e_j)$, i.e. , to construct an estimator $\tilde{\theta}_n$ measurable w.r.t. $Z_1, \dots, Z_n$ and called aggregate, such that

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

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

$$\Delta_{n,M} = \frac{C \log M}{n}\,, \tag{2}$$

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

The main application of oracle inequalities of the type (1) is 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) – (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 has been proved earlier in the benchmark works of Catoni (2004) and Yang (2000). They independently obtained the bound for density estimation with the KL loss, and Catoni (2004) solved the problem for the regression model with squared loss. Bunea and Nobel (2005) suggested another proof of the regression result of Catoni (2004) improving it in the case of bounded response, and obtained some inequalities with suboptimal remainder terms under weaker conditions.

Here we study the recursive aggregate $\hat{\theta}_n$ which is defined in the following way. Set $\beta > 0$, define the vector

$$u_i \triangleq \Big(Q(Z_i, e_1), \ldots, Q(Z_i, e_M)\Big)^\top$$

and consider 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

$$\begin{aligned} \zeta_i &= \zeta_{i-1} + u_i\,, \\ \theta_i &= \left[\frac{\mathrm{e}^{-\zeta_i^{(j)}/\beta}}{\sum_{k=1}^M \mathrm{e}^{-\zeta_i^{(k)}/\beta}}\right]_{j=1}^M . \end{aligned} \tag{3}$$

- Output at iteration n the average

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n \theta_{i-1}\,. \tag{4}$$

Note that $\hat{\theta}_n$ is measurable w.r.t. the subsample $(Z_1, \ldots, Z_{n-1})$. Recursions (3) – (4) represent 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 the following results.

Theorem 1. *Let B be a measurable subset of $\mathcal{Z}$. Assume that $\beta > 0$ is such that the mapping $\theta \mapsto \exp(-Q(z,\theta)/\beta)$ is concave on the simplex Θ, for all $z \in B$. Assume also that there exists two functions $L_Q(\cdot)$ and $R_Q(\cdot)$ on $\mathcal{Z} \setminus B$, with values in $\mathbb{R}$ such that for all $z \in \mathcal{Z} \setminus B$ and all $\theta \in \Theta$ we have $L_Q(z) \le Q(z,\theta) \le$*

$R_Q(z)$. *Then the aggregate* $\hat{\theta}_n$ *satisfies, for any* $M \geq 2, n \geq 1$, *the following oracle inequality*

$$E_{n-1}A(\hat{\theta}_n) \leq \min_{1\leq j\leq M} A(e_j) + \frac{\beta \log M}{n} + E\big[\big(R_Q(Z) - L_Q(Z)\big)\mathbb{1}_{\{Z\notin B\}}\big]\,,$$

where $\mathbb{1}_{\{\cdot\}}$ *denotes the indicator function.*

Theorem 2. *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* $\theta, \theta' \in \Theta$. *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* $E\exp(-Q_1(Z,\theta,\theta')/\beta) \leq \Psi_\beta(\theta,\theta')$ *for all* $\theta, \theta' \in \Theta$. *Then the aggregate* $\hat{\theta}_n$ *satisfies, for any* $M \geq 2, n \geq 1$, *the following oracle inequality*

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

We show that the assumptions of Theorems 1 and 2 are fulfilled for several statistical models including regression, classification and density estimation. This allows one to construct in an easy way sharp adaptive nonparametric estimators for the above mentioned statistical problems.

References

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

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

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

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

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

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