The Annals of Statistics
2014, Vol. 42, No. 1, 211–224
DOI: 10.1214/13-AOS1190

OPTIMAL LEARNING WITH *Q*-AGGREGATION

BY GUILLAUME LECUÉ[1] AND PHILIPPE RIGOLLET[2]

CNRS, Ecole Polytechnique and Princeton University

We consider a general supervised learning problem with strongly convex and Lipschitz loss and study the problem of model selection aggregation. In particular, given a finite dictionary functions (learners) together with the prior, we generalize the results obtained by Dai, Rigollet and Zhang [*Ann. Statist.* **40** (2012) 1878–1905] for Gaussian regression with squared loss and fixed design to this learning setup. Specifically, we prove that the Q-aggregation procedure outputs an estimator that satisfies optimal oracle inequalities both in expectation and with high probability. Our proof techniques somewhat depart from traditional proofs by making most of the standard arguments on the Laplace transform of the empirical process to be controlled.

1. Introduction and main results. Let $\mathcal{X}$ be a probability space and let $(X, Y) \in \mathcal{X} \times \mathbb{R}$ be a random couple. Broadly speaking, the goal of statistical learning is to predict Y given X. To achieve this goal, we observe a dataset $\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ that consists of n independent copies of (X, Y) and use these observations to construct a function (learner) $f : \mathcal{X} \to \mathbb{R}$ such that $f(X)$ is close to Y in a certain sense. More precisely, the prediction quality of a (possibly data dependent) function $\hat{f}$ is measured by a *risk function* $R : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$ associated to a *loss function* $\ell : \mathbb{R}^2 \to \mathbb{R}$ in the following way:

$$R(\hat{f}) = \mathbb{E}\big[\ell\big(Y, \hat{f}(X)\big)|\mathcal{D}\big].$$

We focus hereafter on loss functions ℓ that are *convex* in their second argument. Moreover, for the sake of simplicity, throughout this article we restrict ourselves to functions f and random variables (X, Y) for which $|Y| \le b$ and $|f(X)| \le b$ almost surely, for some fixed $b \ge 0$. For any real-valued measurable f on $\mathcal{X}$, for which this quantity is finite, we define $\|f\|_2 = \sqrt{\mathbb{E}[f(X)^2]}$.

We are given a finite set $\mathcal{F} = \{f_1, \dots, f_M\}$ of measurable functions from $\mathcal{X}$ to $\mathbb{R}$. This set is called a *dictionary*. The elements in $\mathcal{F}$ may have been constructed using an independent, frozen, dataset at some previous step or may simply be good

Received July 2013; revised November 2013.

[1]Supported by French Agence Nationale de la Recherche ANR Grant "PROGNOSTIC" ANR-09-JCJC-0101-01.

[2]Supported in part by NSF Grants DMS-09-06424, DMS-13-17308, CAREER-DMS-10-53987, a Howard B. Wentz Jr. award and a gift from the Bendheim Center for Finance.

MSC2010 subject classifications. Primary 68Q32; secondary 62G08, 62G05.

Key words and phrases. Learning theory, empirical risk minimization, aggregation, empirical processes theory.

candidates for the learning task at hand. To focus our contribution on the aggregation problem, we restrict our attention to the case where $\mathcal{F}$ consists of deterministic functions and, because of the diversity of dictionaries that can be considered, we do not want to assume anything on the dictionary except boundedness.

The aim of model selection aggregation [7, 8, 16, 30] is to use the data $\mathcal{D}$ to construct a function $\hat{f}$ having an *excess-risk* $R(\hat{f}) - \min_{f\in\mathcal{F}} R(f)$ as small as possible. Namely, we seek the smallest deterministic *residual term* $\Delta_n(\mathcal{F}) > 0$ such that the excess risk is bounded above by $\Delta_n(\mathcal{F})$, either in expectation or with high probability, or, in this instance, in both. In the high probability case, such bounds are called *oracle inequalities*. This problem was introduced and studied in [7, 16]. Many results have been obtained in aggregation theory during the last decade, for instance, in [2], the suboptimality in deviation of the Gibbs aggregates is proved, in [3], several procedures related to Gibbs aggregates are proved to be optimal (in expectation) even under moment assumptions, in [6], a "universal" aggregation method is constructed to solve several type of aggregation problems in the Gaussian regression model. Other construction of optimal aggregation procedures in various setups can also be found in [18, 19, 23, 30–33].

From a minimax standpoint, it has been proved that $\Delta_n(\mathcal{F}) = C(\log M)/n$, $C > 0$ is the smallest residual term that one can hope for the regression problem with quadratic loss [30]. An estimator $\hat{f}$ achieving such a rate (up to some multiplying constant) is called an optimal aggregate. One of the first procedures proved to achieve this optimal rate is a progressive mixture rule of Gibbs estimators (cf. [3, 7, 19, 33]). The optimality of this procedure holds for any "exponentially concave" loss function (cf. Theorem 4.2 in [19]) but only in expectation (cf. [2]).

The aim of this paper is to construct optimal aggregates (both in expectation and deviation) under general conditions on the loss function ℓ and for a random design. We also want these procedures to have the ability to take into account some prior information on the dictionary unlike the existing optimal aggregation procedures that have been constructed in this setup so far (cf. [2, 23]).

Note that the optimal residuals for model selection aggregation are of the order $1/n$ as opposed to the standard parametric rate $1/\sqrt{n}$. This *fast* rate essentially comes from the strong convexity of the quadratic loss. In what follows, we show that indeed, strong convexity is sufficient to obtain fast rates. It is known that rates of optional order $1/n$ cannot be achieved if the loss function is only assumed to be convex. Indeed, it follows from [21], Theorem 2, that if the loss is linear then the best achievable residual term is at least of the order $\sqrt{(\log|\mathcal{F}|)/n}$. Recall that a function g is said to be strongly convex on a nonempty convex set $C \subset \mathbb{R}$ if there exists a constant c such that

$$g\bigl(\alpha a + (1-\alpha)a'\bigr) \le \alpha g(a) + (1-\alpha)g\bigl(a'\bigr) - \frac{c}{2}\alpha(1-\alpha)\bigl(a-a'\bigr)^2$$

for any $a, a' \in C, \alpha \in (0,1)$. In this case, c is called *modulus of strong convexity*. For technical reasons, we will also need to assume that the loss function is Lipschitz. We now introduce the set of assumptions that are sufficient for our approach.

ASSUMPTION 1. The loss function ℓ is such that for any $f, g \in [-b, b]$, we have

$$|\ell(Y, f) - \ell(Y, g)| \le C_b |f - g| \qquad \text{a.s.}$$

Moreover, almost surely, the function $\ell(Y, \cdot)$ is *strongly convex* with modulus of strong convexity C_ℓ on $[-b, b]$.

A central quantity that is used for the construction of aggregates is the empirical risk defined by

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(Y_i, f(X_i)\big) \tag{1.1}$$

for any real-valued function f defined over $\mathcal{X}$. A natural aggregation procedure consists in taking the function in $\mathcal{F}$ that minimizes the empirical risk. This procedure is called empirical risk minimization (ERM). It has been proved that ERM is suboptimal for the aggregation problem (cf. Proposition 2.1 in [19] or Chapter 3.5 in [7], Theorem 1.1 in [24], Theorem 3 in [22], Theorem 2 in [26] and Theorem 2.1 in [29]). Somehow, this procedure does not take advantage of the convexity of the loss since the class of functions on which the empirical risk is minimized to construct the ERM is $\mathcal{F}$, a finite set. As it turns out, the performance of ERM relies critically on the convexity of the class of functions on which the empirical risk is minimized [24, 26]. Therefore, a natural idea is to "improve the geometry" of $\mathcal{F}$ by taking its convex hull $\operatorname{conv}(\mathcal{F})$ and then by minimizing the empirical risk over it. However, this procedure is also suboptimal [9, 23]. The weak point of this procedure lies in the metric complexity of the problem: taking the convex hull of $\mathcal{F}$ indeed "improves the geometry" of $\mathcal{F}$ but it also increases by too much its complexity. The complexity of the convex hull of a set can be much larger than the complexity of the set itself and this leads to a failure of this naive convexification trick. Nevertheless, a compromise between geometry and complexity was stricken in [2] and [23] where optimal aggregates have been successfully constructed. In [2], this improvement is achieved by minimizing the empirical risk over a carefully chosen star-shaped subset of the convex hull of $\mathcal{F}$. In [23], a better geometry was achieved by taking the convex hull of an appropriate subset of $\mathcal{F}$ and then by minimizing the empirical risk over it.

In this paper, we show that a third procedure, called Q-aggregation, and that was introduced in [9, 27] for fixed design Gaussian regression, also leads to optimal rates of aggregation. Unlike the above two procedures that rely on finding an appropriate constraint for ERM, Q-aggregation is based on a penalization of the empirical risk but the constraint set is kept to be the convex hull of $\mathcal{F}$. Let Θ denote the flat simplex of $\mathbb{R}^M$ defined by

$$\Theta = \left\{ (\theta_1, \dots, \theta_M) \in \mathbb{R}^M : \theta_j \ge 0, \sum_{j=1}^{M} \theta_j = 1 \right\}$$

and for any $\theta \in \Theta$, define the convex combination $f_\theta = \sum_{j=1}^M \theta_j f_j$. For any fixed ν, the Q-functional is defined for any $\theta \in \Theta$ by

$$Q(\theta) = (1-\nu)R_n(f_\theta) + \nu \sum_{j=1}^M \theta_j R_n(f_j). \tag{1.2}$$

We keep the terminology Q-aggregation from [9] in purpose. Indeed, Q stands for *quadratic* and while do not employ a quadratic loss, we exploit strong convexity in the same manner as in [9] and [27]. Indeed the first term in Q acts as a regularization of the linear interpolation of the empirical risk, and is therefore a strongly convex regularization.

We consider the following aggregation procedure. Unlike the procedures introduced in [2, 23], the Q-aggregation procedure allows us to put a prior weight given by a prior probability $\pi = (\pi_1, \dots, \pi_M)$ on each element of the dictionary $\mathcal{F}$. This feature turns out to be crucial for applications [1, 10–15, 28, 29]. Let $\beta > 0$ be the *temperature* parameter and $0 < \nu < 1$. Consider any vector of weights $\hat\theta \in \Theta$ defined by

$$\hat\theta \in \operatorname*{argmin}_{\theta\in\Theta} \left[(1-\nu)R_n(f_\theta) + \nu \sum_{j=1}^M \theta_j R_n(f_j) - \frac{\beta}{n} \sum_{j=1}^M \theta_j \log \pi_j \right]. \tag{1.3}$$

It comes out of our analysis that $f_{\hat\theta}$ achieves an optimal rate of aggregation if β satisfies

$$\beta > \max\left[\frac{8C_b^2(1-\nu)^2}{\mu}, 4\sqrt{3}bC_b(1-\nu), \frac{C_b\nu(\nu C_b + 4\mu b)}{\mu} \right], \tag{1.4}$$

where $\mu = \min(\nu, 1-\nu)(C_\ell)/10$.

This procedure was studied in the case of fixed design in [9], where it is shown that greedy algorithms similar to the Frank–Wolfe algorithm, can be employed to solve the optimization problem in (1.3). In particular, such algorithms can yield solutions $\hat\theta$ that are very sparse: they can have a little as two nonzero coordinates. In this case, and when the prior π is uniform, this two-step procedure recovers the STAR algorithm of Audibert [2]. Furthermore, unlike the STAR algorithm, the greedy algorithm of [9] was shown to (i) allow to handle any prior π and (ii) yield better constants as well as better numerical performance for a larger number of iterations (see [9] for more details). Similar algorithms can be employed in the present case and it follows trivially from [9], Proposition 4.1, that n iterations suffice to obtain an optimization error of the same order as the statistical error. Going down to two iterations as in [9], Theorem 4.2, requires a more delicate analysis, similar to the one employed in [9], but is beyond the scope of this paper.

THEOREM A. *Let $\mathcal{F}$ be a finite dictionary of cardinality M and (X, Y) be a random couple of $\mathcal{X} \times \mathbb{R}$ such that $|Y| \le b$ and $\max_{f\in\mathcal{F}} |f(X)| \le b$ a.s. for*

some $b > 0$. Assume that Assumption 1 holds and that β satisfies (1.4). Then, for any $x > 0$, with probability greater than $1 - \exp(-x)$

$$R(f_{\hat{\theta}}) \le \min_{j=1,\dots,M}\left[R(f_j) + \frac{\beta}{n}\log\left(\frac{1}{\pi_j}\right)\right] + \frac{2\beta x}{n}.$$

Moreover,

$$\mathbb{E}\big[R(f_{\hat{\theta}})\big] \le \min_{j=1,\dots,M}\left[R(f_j) + \frac{\beta}{n}\log\left(\frac{1}{\pi_j}\right)\right].$$

If π is the uniform distribution, that is $\pi_j = 1/M$ for all $j = 1, \dots, M$, then we recover in Theorem A the classical optimal rate of aggregation $(\log M)/n$ and the estimator $\hat{\theta}$ is just the one minimizing the Q-functional defined in (1.2). In particular, no temperature parameter β is needed for its construction. As a result, in this case, the parameter b need not be known for the construction of the Q-aggregation procedure.

2. Preliminaries to the proof of Theorem A. An important part of our analysis is based upon concentration properties of empirical processes. While our proofs are similar to those employed in [27] and [9], they contain genuinely new arguments. In particular, this learning setting, unlike the denoising setting considered in [9, 27] allows us to employ various new tools such as symmetrization and contraction. A classical tool to quantify the concentration of measure phenomenon is given by Bernstein's inequality for bounded variables. In terms of Laplace transform, Bernstein's inequality [5], Theorem 1.10, states that if $Z_1, \dots, Z_n$ are n i.i.d. real-valued random variables such that for all $i = 1, \dots, n$,

$$|Z_i| \le c \qquad \text{a.s.} \quad \text{and} \quad \mathbb{E}Z_i^2 \le v,$$

then for any $0 < \lambda < 1/c$,

$$\mathbb{E}\exp\left[\lambda\left(\sum_{i=1}^{n}\{Z_i - \mathbb{E}Z_i\}\right)\right] \le \exp\left(\frac{nv\lambda^2}{2(1-c\lambda)}\right). \tag{2.1}$$

Bernstein's inequality usually yields a bound of order $\sqrt{n}$ for the deviations of a sum around its mean. As mentioned above, such bounds are not sufficient for our purposes and we thus consider the following concentration result.

PROPOSITION 1. *Let $Z_1, \dots, Z_n$ be i.i.d. real-valued random variables and let $c_0 > 0$. Assume that $|Z_1| \le c$ a.s. Then, for any $0 < \lambda < (2c_0)/(1 + 2c_0c)$,*

$$\mathbb{E}\exp\left[n\lambda\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}Z_i - c_0\mathbb{E}Z_i^2\right)\right] \le 1$$

and

$$\mathbb{E}\exp\left[n\lambda\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}Z_i - Z_i - c_0\mathbb{E}Z_i^2\right)\right] \le 1.$$

PROOF. It follows from Bernstein's inequality (2.1) that for any $0 < \lambda < (2c_0)/(1 + 2c_0c)$,

$$\mathbb{E}\exp\left[n\lambda\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}Z_i - c_0\mathbb{E}Z_i^2\right)\right]$$
$$\leq \exp\left(\frac{n\mathbb{E}Z_1^2\lambda^2}{2(1-c\lambda)}\right)\exp\left[-n\lambda c_0\mathbb{E}Z_1^2\right] \leq 1.$$

The second inequality is obtained by replacing Z_i by $-Z_i$. □

We also recall the following exponential bound for Rademacher processes: let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables and $a_1, \ldots, a_n$ be some real numbers then, by Hoeffding's inequality,

$$\mathbb{E}\exp\left(\sum_{i=1}^{n} \varepsilon_i a_i\right) \leq \exp\left(\frac{1}{2}\sum_{i=1}^{n} a_i^2\right). \tag{2.2}$$

We will also use a slightly modified version of the symmetrization inequality: let $\mathcal{F}$ be a function class, A_f, $f \in \mathcal{F}$ be a given function on $\mathcal{F}$ and Φ be a convex nondecreasing function then

$$\mathbb{E}\Phi\Big(\sup_{f\in\mathcal{F}}[Pf - P_n f - A_f]\Big) \leq \mathbb{E}\Phi\Big(2\sup_{f\in\mathcal{F}}[P_{n,\varepsilon} f - A_f]\Big), \tag{2.3}$$

where P is a measure, P_n its associated empirical measure and $P_{n,\varepsilon}$ the symmetrized empirical measure defined by

$$Pf = \mathbb{E}f(Z), \qquad P_n f = \frac{1}{n}\sum_{i=1}^{n} f(Z_i) \quad \text{and} \quad P_{n,\varepsilon} f = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i),$$

where $Z, Z_1, \ldots, Z_n$ are i.i.d. random variables distributed according to P and $\varepsilon_1, \ldots, \varepsilon_n$ are independent Rademacher independent of $Z, Z_1, \ldots, Z_n$. The proof of (2.3) follows the same line as the symmetrization inequality (cf. e.g., Theorem 2.1 in [20]).

Our analysis also relies upon some geometric argument. Indeed, the strong convexity of the loss function in Assumption 1 implies the 2-convexity of the risk in the sense of [4]; cf. (2.4) for an explicit definition of the 2-convexity of a function $R(\cdot)$. This translates into a lower bound on the gain obtained when applying Jensen's inequality to the risk function R.

PROPOSITION 2. *Let (X, Y) be a random couple in $\mathcal{X} \times \mathbb{R}$ and $\mathcal{F} = \{f_1, \ldots, f_M\}$ be a finite dictionary in $L_2(\mathcal{X}, P_X)$ such that $|f_j(X)| \leq b$, $\forall j = 1, \ldots, M$ and $|Y| \leq b$ a.s. Assume that, almost surely, the function $\ell(Y, \cdot)$ is*

strongly convex with modulus of strong convexity C_ℓ on $[-b, b]$. Then it holds that, for any $\theta \in \Theta$,

$$R\left(\sum_{j=1}^{M} \theta_j f_j\right) \le \sum_{j=1}^{M} \theta_j R(f_j) - \frac{C_\ell}{2} \sum_{j=1}^{M} \theta_j \left\| f_j - \sum_{j=1}^{M} \theta_j f_j \right\|_2^2. \tag{2.4}$$

PROOF. Define the random function $\ell(\cdot) = \ell(Y, \cdot)$. By strong convexity and [17], Theorem 6.1.2, it holds almost surely that for any a, a' in $[-b, b]$,

$$\ell(a) \ge \ell(a') + (a - a')\ell'(a') + \frac{C_\ell}{2}(a - a')^2$$

for any $\ell'(a')$ in the subdifferential of ℓ at a'. Plugging $a = f_j(X)$, $a' = f_\theta(X)$, we get almost surely

$$\begin{aligned}&\ell\big(Y, f_j(X)\big)\\ &\quad \ge \ell\big(Y, f_\theta(X)\big) + \big(f_j(X) - f_\theta(X)\big)\ell'\big(f_\theta(X)\big) + \frac{C_\ell}{2}\big[f_j(X) - f_\theta(X)\big]^2.\end{aligned}$$

Now, multiplying both sides by θ_j and summing over j, we get almost surely,

$$\sum_j \theta_j \ell\big(Y, f_j(X)\big) \ge \ell\big(Y, f_\theta(X)\big) + \frac{C_\ell}{2} \sum_j \theta_j \big[f_j(X) - f_\theta(X)\big]^2.$$

To complete the proof, it remains to take the expectation. □

3. Proof of Theorem A. Let $x > 0$ and assume that Assumption 1 holds throughout this section. We start with some notation. For any $\theta \in \Theta$, define

$$\ell_\theta(y, x) = \ell\big(y, f_\theta(x)\big) \quad \text{and} \quad \mathsf{R}(\theta) = \mathbb{E}\ell_\theta(Y, X) = \mathbb{E}\ell\big(Y, f_\theta(X)\big),$$

where we recall that $f_\theta = \sum_{j=1}^{M} \theta_j f_j$ for any $\theta \in \mathbb{R}^M$. Let $0 < \nu < 1$. Let $(e_1, \dots, e_M)$ is the canonical basis of $\mathbb{R}^M$ and for any $\theta \in \mathbb{R}^M$ define

$$\tilde{\ell}_\theta(y, x) = (1 - \nu)\ell_\theta(y, x) + \nu \sum_{j=1}^{M} \theta_j \ell_{e_j}(y, x) \quad \text{and} \quad \tilde{\mathsf{R}}(\theta) = \mathbb{E}\tilde{\ell}_\theta(Y, X).$$

We also consider the functions

$$\theta \in \mathbb{R}^M \mapsto K(\theta) = \sum_{j=1}^{M} \theta_j \log\left(\frac{1}{\pi_j}\right)$$

and

$$\theta \in \mathbb{R}^M \mapsto V(\theta) = \sum_{j=1}^{M} \theta_j \|f_j - f_\theta\|_2^2.$$

Let $\mu > 0$. Consider any oracle $\theta^* \in \Theta$ such that

$$\theta^* \in \operatorname*{argmin}_{\theta\in\Theta}\biggl(\tilde{\mathsf{R}}(\theta) + \mu V(\theta) + \frac{\beta}{n}K(\theta)\biggr).$$

We start with a geometrical aspect of the problem. The two following results follow from the strong convexity of the loss function ℓ.

PROPOSITION 3. *When $\mu \le (1-\nu)C_\ell/2$, the function $\theta \mapsto H(\theta) = \tilde{\mathsf{R}}(\theta) + \mu V(\theta) + (\beta/n)K(\theta)$ is convex over the convex set Θ.*

PROOF. Let $\theta, \beta \in \Theta$ and $0 \le \alpha \le 1$. It follows from some computation that

$$V\bigl(\alpha\theta + (1-\alpha)\beta\bigr) = (1-\alpha)V(\beta) + \alpha V(\theta) + \alpha(1-\alpha)\|f_\theta - f_\beta\|_2^2.$$

It follows from the strong convexity of $\ell(y,\cdot)$ that

$$R\bigl(\alpha\theta + (1-\alpha)\beta\bigr) \le (1-\alpha)R(\beta) + \alpha R(\theta) - \frac{C_\ell}{2}\alpha(1-\alpha)\|f_\theta - f_\beta\|_2^2.$$

Therefore, when $\mu \le (1-\nu)C_\ell/2$, we have

$$H\bigl(\alpha\theta + (1-\alpha)\beta\bigr) \le (1-\alpha)H(\beta) + \alpha H(\theta). \qquad \square$$

PROPOSITION 4. *Let $\mu \le (1-\nu)C_\ell/2$. For any $\theta \in \Theta$,*

$$\begin{aligned}&\tilde{\mathsf{R}}(\theta) - \tilde{\mathsf{R}}(\theta^*)\\ &\quad\ge \mu\bigl(V(\theta^*) - V(\theta)\bigr) + \frac{\beta}{n}\bigl(K(\theta^*) - K(\theta)\bigr) + \biggl(\frac{(1-\nu)C_\ell}{2} - \mu\biggr)\|f_\theta - f_{\theta^*}\|_2^2.\end{aligned}$$

PROOF. Since $\theta \mapsto H(\theta) = \tilde{\mathsf{R}}(\theta) + \mu V(\theta) + (\beta/n)K(\theta)$ is convex over the convex set Θ and θ^* is a minimizer of H over Θ, then there exists a subgradient $\nabla H(\theta^*)$ such that for any $\theta \in \Theta$ it holds, $\langle \nabla H(\theta^*), \theta - \theta^*\rangle \ge 0$. It yields

$$\begin{aligned}&\bigl\langle \nabla\tilde{\mathsf{R}}(\theta^*), \theta - \theta^*\bigr\rangle\\ &\quad\ge \mu\bigl\langle \nabla V(\theta^*), \theta^* - \theta\bigr\rangle + (\beta/n)\bigl\langle \nabla K(\theta^*), \theta^* - \theta\bigr\rangle \qquad (3.1)\\ &\quad= \mu\bigl(V(\theta^*) - V(\theta)\bigr) - \mu\|f_\theta - f_{\theta^*}\|_2^2 + (\beta/n)\bigl(K(\theta^*) - K(\theta)\bigr).\end{aligned}$$

It follows from the strong convexity of $\ell(y,\cdot)$ that

$$\begin{aligned}&\tilde{\mathsf{R}}(\theta) - \tilde{\mathsf{R}}(\theta^*)\\ &\quad\ge \bigl\langle \nabla\tilde{\mathsf{R}}(\theta^*), \theta - \theta^*\bigr\rangle + \frac{(1-\nu)C_\ell}{2}\|f_\theta - f_{\theta^*}\|_2^2\\ &\quad\ge \mu\bigl(V(\theta^*) - V(\theta)\bigr) + \frac{\beta}{n}\bigl(K(\theta^*) - K(\theta)\bigr) + \biggl(\frac{(1-\nu)C_\ell}{2} - \mu\biggr)\|f_\theta - f_{\theta^*}\|_2^2,\end{aligned}$$

where the second inequality follows from the previous display. $\square$

Let $\mathbf{H}$ be the $M \times M$ matrix with entries $\mathbf{H}_{j,k} = \|f_j - f_k\|_2^2$ for all $1 \le j, k \le M$. Let s and x be positive numbers and consider the random variable

$$Z_n = (P - P_n)(\tilde{\ell}_{\hat{\theta}} - \tilde{\ell}_{\theta^*}) - \mu \sum_{j=1}^{M} \hat{\theta}_j \|f_j - f_{\theta^*}\|_2^2 - \mu \hat{\theta} \mathbf{H} \theta^* - \frac{1}{s} K(\hat{\theta}).$$

PROPOSITION 5. *Assume that* $10\mu \le \min(1 - \nu, \nu) C_\ell$ *and* $\beta \ge 2n/s$. *Then it holds*

$$R(\hat{\theta}) \le \min_{1 \le j \le M} \left[R(e_j) + \frac{\beta}{n} \log\left(\frac{1}{\pi_j}\right) \right] + 2Z_n.$$

PROOF. First note that the following equalities hold:

$$\sum_{j=1}^{M} \hat{\theta}_j \|f_j - f_{\theta^*}\|_2^2 = V(\hat{\theta}) + \|f_{\hat{\theta}} - f_{\theta^*}\|_2^2 \tag{3.2}$$

and

$$\hat{\theta} \mathbf{H} \theta^* = V(\hat{\theta}) + V(\theta^*) + \|f_{\theta^*} - f_{\hat{\theta}}\|_2^2. \tag{3.3}$$

It follows from the definition of $\hat{\theta}$ that

$$\tilde{\mathsf{R}}(\hat{\theta}) - \tilde{\mathsf{R}}(\theta^*) \le (P - P_n)(\tilde{\ell}_{\hat{\theta}} - \tilde{\ell}_{\theta^*}) + \frac{\beta}{n}\big(K(\theta^*) - K(\hat{\theta})\big). \tag{3.4}$$

Then we use (3.2) and (3.3) together with (3.4) to get

$$\begin{aligned} \tilde{\mathsf{R}}(\hat{\theta}) - \tilde{\mathsf{R}}(\theta^*) &\le 2\mu V(\hat{\theta}) + \mu V(\theta^*) + 2\mu \|f_{\hat{\theta}} - f_{\theta^*}\|_2^2 \\ &\quad + \frac{1}{s} K(\hat{\theta}) + \frac{\beta}{n}\big(K(\theta^*) - K(\hat{\theta})\big) + Z_n. \end{aligned} \tag{3.5}$$

Together with Proposition 4, it yields

$$\left(\frac{(1-\nu)C_\ell}{2} - 3\mu\right) \|f_{\hat{\theta}} - f_{\theta^*}\|_2^2 \le 3\mu V(\hat{\theta}) + \frac{1}{s} K(\hat{\theta}) + Z_n.$$

We plug the above inequality into (3.5) to obtain

$$\begin{aligned} \tilde{\mathsf{R}}(\hat{\theta}) - \tilde{\mathsf{R}}(\theta^*) &\le \left(1 + \frac{2\mu}{(1-\nu)C_\ell/2 - 3\mu}\right)\left(\frac{1}{s} K(\hat{\theta}) + Z_n\right) \\ &\quad + \frac{\beta}{n}\big(K(\theta^*) - K(\hat{\theta})\big) + \mu V(\theta^*) \\ &\quad + \left(2\mu + \frac{6\mu^2}{(1-\nu)C_\ell/2 - 3\mu}\right) V(\hat{\theta}). \end{aligned}$$

Thanks to the 2-convexity of the risk (cf. Proposition 2), we have

$$\tilde{\mathsf{R}}(\hat{\theta}) \ge \mathsf{R}(\hat{\theta}) + \nu(C_\ell/2) V(\hat{\theta}).$$

Therefore, it follows from (3.6) that

$$\begin{aligned}\mathsf{R}(\hat\theta) \le \tilde{\mathsf{R}}(\theta^*) + \mu V(\theta^*) + \frac{\beta}{n}K(\theta^*) + \biggl(1 + \frac{4\mu}{(1-\nu)C_\ell - 6\mu}\biggr)Z_n \\ + \biggl(2\mu + \frac{12\mu^2}{(1-\nu)C_\ell - 6\mu} - \nu\frac{C_\ell}{2}\biggr)V(\hat\theta) \\ + \biggl(\frac{1}{s} + \frac{4\mu}{s((1-\nu)C_\ell - 6\mu)} - \frac{\beta}{n}\biggr)K(\hat\theta).\end{aligned} \tag{3.6}$$

Note now that $10\mu \le \min(\nu, 1-\nu)C_\ell$ implies that

$$\frac{4\mu}{(1-\nu)C_\ell - 6\mu} \le 1$$

and

$$2\mu + \frac{12\mu^2}{(1-\nu)C_\ell - 6\mu} - \nu\frac{C_\ell}{2} \le 0.$$

Moreover, together, the two conditions of the proposition yield

$$\frac{1}{s} + \frac{4\mu}{s((1-\nu)C_\ell - 6\mu)} - \frac{\beta}{n} \le 0.$$

Therefore, it follows from the above three displays that

$$\begin{aligned}\mathsf{R}(\hat\theta) &\le \min_{\theta\in\Theta}\biggl[\tilde{\mathsf{R}}(\theta) + \mu V(\theta) + \frac{\beta}{n}K(\theta)\biggr] + 2Z_n \\ &\le \min_{j=1,\dots,M}\biggl[\mathsf{R}(e_j) + \frac{\beta}{n}\log\biggl(\frac{1}{\pi_j}\biggr)\biggr] + 2Z_n.\end{aligned}$$

□

To complete our proof, it remains to prove that $\mathbf{P}[Z_n > (\beta x)/n] \le \exp(-x)$ and $\mathbb{E}[Z_n] \le 0$ under suitable conditions on μ and β. Using, respectively, a Chernoff bound and Jensen's inequality, it is easy to see that both conditions follow if we prove that $\mathbb{E}\exp(nZ_n/\beta) \le 1$. It follows from the excess loss decomposition:

$$\tilde\ell_{\hat\theta}(y,x) - \tilde\ell_{\theta^*}(y,x) = (1-\nu)\bigl(\ell_{\hat\theta}(y,x) - \ell_{\theta^*}(y,x)\bigr) + \nu\sum_{j=1}^M(\hat\theta_j - \theta_j^*)\ell_{e_j}(y,x)$$

and the Cauchy–Schwarz inequality implies that it is enough to prove that

$$\begin{aligned}\mathbb{E}\exp\biggl[s\biggl((1-\nu)(P-P_n)(\ell_{\hat\theta} - \ell_{\theta^*}) \\ - \mu\sum_{j=1}^M\hat\theta_j\|f_j - f_{\theta^*}\|_2^2 - \frac{1}{s}K(\hat\theta)\biggr)\biggr] \le 1\end{aligned} \tag{3.7}$$

and

$$\mathbb{E}\exp\left[s\left(\nu(P-P_n)\left(\sum_{j=1}^{M}(\hat{\theta}_j-\theta_j^*)\ell_{e_j}\right)-\mu\hat{\theta}\mathbf{H}\theta^*-\frac{1}{s}K(\hat{\theta})\right)\right]\le 1 \tag{3.8}$$

for some $s \ge 2n/\beta$ and assume this condition holds in the rest of the proof.

We begin by proving (3.7). To that end, define the symmetrized empirical process by $h \mapsto P_{n,\varepsilon}h = n^{-1}\sum_{i=1}^{n}\varepsilon_i h(Y_i, X_i)$ where $\varepsilon_1, \dots, \varepsilon_n$ are n i.i.d. Rademacher random variables independent of the (X_i, Y_i)'s. Moreover, take s and μ such that

$$s \le \frac{\mu n}{[2C_b(1-\nu)]^2}. \tag{3.9}$$

It yields

$$\begin{aligned}
&\mathbb{E}\exp\left[s\left((1-\nu)(P-P_n)(\ell_{\hat{\theta}}-\ell_{\theta^*})-\mu\sum_{j=1}^{M}\hat{\theta}_j\|f_j-f_{\theta^*}\|_2^2-\frac{1}{s}K(\hat{\theta})\right)\right]\\
&\quad\le \mathbb{E}\exp\left[s\max_{\theta\in\Theta}\left((1-\nu)(P-P_n)(\ell_\theta-\ell_{\theta^*})\right.\right.\\
&\qquad\qquad\left.\left.-\mu\sum_{j=1}^{M}\theta_j\|f_j-f_{\theta^*}\|_2^2-\frac{1}{s}K(\theta)\right)\right]\\
&\quad\le \mathbb{E}\exp\left[s\max_{\theta\in\Theta}\left(2(1-\nu)P_{n,\varepsilon}(\ell_\theta-\ell_{\theta^*})\right.\right.\\
&\qquad\qquad\left.\left.-\mu\sum_{j=1}^{M}\theta_j\|f_j-f_{\theta^*}\|_2^2-\frac{1}{s}K(\theta)\right)\right] \qquad (3.10)\\
&\quad\le \mathbb{E}\exp\left[s\max_{\theta\in\Theta}\left(2C_b(1-\nu)P_{n,\varepsilon}(f_\theta-f_{\theta^*})\right.\right.\\
&\qquad\qquad\left.\left.-\mu\sum_{j=1}^{M}\theta_j\|f_j-f_{\theta^*}\|_2^2-\frac{1}{s}K(\theta)\right)\right], \qquad (3.11)
\end{aligned}$$

where (3.10) follows from the slightly modified version of the symmetrization inequality in (2.3) and (3.11) follows from the contraction principle [25], Theorem 4.12, applied to contractions

$$\varphi_i(t_i) = C_b^{-1}\big[\ell\big(Y_i, f_{\theta^*}(X_i)-t_i\big)-\ell\big(Y_i, f_{\theta^*}(X_i)\big)\big]$$

and $T \subset \mathbb{R}^n$ is defined by

$$T = \big\{t \in \mathbb{R}^n : t_i = f_{\theta^*}(X_i) - f_\theta(X_i), \theta \in \Theta\big\}.$$

Next, using the fact that the maximum of a linear function over a polytope is attained at a vertex, we get

$$\begin{aligned}
&\mathbb{E}\exp\left[s\left((1-\nu)(P-P_n)(\ell_{\hat\theta}-\ell_{\theta^*})-\mu\sum_{j=1}^M\hat\theta_j\|f_j-f_{\theta^*}\|_2^2-\frac{1}{s}K(\hat\theta)\right)\right]\\
&\quad\le\sum_{k=1}^M\pi_k\mathbb{E}\mathbb{E}_\varepsilon\exp\bigl[s\bigl(2C_b(1-\nu)P_{n,\varepsilon}(f_k-f_{\theta^*})-\mu\|f_k-f_{\theta^*}\|_2^2\bigr)\bigr]\\
&\quad\le\sum_{k=1}^M\pi_k\mathbb{E}\exp\Biggl[\frac{[2C_b(1-\nu)s)]^2}{2n}\\
&\qquad\qquad\times\left(P_n-\frac{2\mu n}{[2C_b(1-\nu)]^2s}P\right)(f_k-f_{\theta^*})^2\Biggr] \qquad (3.12)\\
&\quad\le\sum_{k=1}^M\pi_k\mathbb{E}\exp\Biggl[\frac{(2C_b(1-\nu)s)^2}{2n}\\
&\qquad\qquad\times\left((P_n-P)(f_k-f_{\theta^*})^2-\frac{1}{4b^2}P(f_k-f_{\theta^*})^4\right)\Biggr], \qquad (3.13)
\end{aligned}$$

where (3.12) follows from (2.2) and (3.13) follows from (3.9). Together with the above display, Proposition 1 yields (3.7) as long as

$$s<\frac{n}{2\sqrt{3}bC_b(1-\nu)}. \qquad (3.14)$$

We now prove (3.8). We have

$$\begin{aligned}
&\mathbb{E}\exp\left[s\left(\nu(P-P_n)\left(\sum_{j=1}^M(\hat\theta_j-\theta_j^*)\ell_{e_j}\right)-\mu\hat\theta\mathbf{H}\theta^*-\frac{1}{s}K(\hat\theta)\right)\right]\\
&\quad\le\sum_{j=1}^M\theta_j^*\sum_{k=1}^M\pi_k\mathbb{E}\exp\bigl[s\bigl(\nu(P-P_n)(\ell_{e_k}-\ell_{e_j})-\mu\|f_j-f_k\|_2^2\bigr)\bigr]\\
&\quad\le\sum_{j=1}^M\theta_j^*\sum_{k=1}^M\pi_k\mathbb{E}\exp\left[s\nu\left((P-P_n)(\ell_{e_k}-\ell_{e_j})-\frac{\mu}{\nu C_b^2}P(\ell_{e_j}-\ell_{e_k})^2\right)\right]\le 1,
\end{aligned}$$

where the last inequality follows from Proposition 1 when

$$s<\frac{2\mu n}{C_b\nu(\nu C_b+4\mu b)}. \qquad (3.15)$$

It is now straightforward to see that the conditions of Proposition 5, the ones of (3.9), (3.14) and (3.15) are fulfilled when

$$s=\frac{2n}{\beta},\qquad \mu=\min(\nu,1-\nu)\frac{C_\ell}{10}$$

and

$$\beta > \max\left[\frac{8C_b^2(1-\nu)^2}{\mu}, 4\sqrt{3}bC_b(1-\nu), \frac{C_b\nu(\nu C_b + 4\mu b)}{\mu}\right].$$

REFERENCES

[1] ALQUIER, P. and LOUNICI, K. (2011). PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Stat.* **5** 127–145. MR2786484

[2] AUDIBERT, J.-Y. (2007). Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA.

[3] AUDIBERT, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *Ann. Statist.* **37** 1591–1646. MR2533466

[4] BARTLETT, P. L., JORDAN, M. I. and MCAULIFFE, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032

[5] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2012). *Concentration Inequalities with Applications*. Clarendon Press, Oxford.

[6] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101

[7] CATONI, O. (2004). *Statistical Learning Theory and Stochastic Optimization. Lecture Notes in Math.* **1851**. Springer, Berlin. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. MR2163920

[8] CATONI, O. (2007). *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **56**. IMS, Beachwood, OH. MR2483528

[9] DAI, D., RIGOLLET, P. and ZHANG, T. (2012). Deviation optimal learning using greedy Q-aggregation. *Ann. Statist.* **40** 1878–1905. MR3015047

[10] DALALYAN, A. S., INGSTER, Y. and TSYBAKOV, A. (2014). Statistical inference in compound functional models. *Probab. Theory Related Fields*. To appear.

[11] DALALYAN, A. S. and SALMON, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.* **40** 2327–2355. MR3059085

[12] DALALYAN, A. S. and TSYBAKOV, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *Learning Theory. Lecture Notes in Computer Science* **4539** 97–111. Springer, Berlin. MR2397581

[13] DALALYAN, A. S. and TSYBAKOV, A. B. (2008). Aggregation by exponential weighting, sharp pac-Bayesian bounds and sparsity. *J. Mach. Learn. Res.* **72** 39–61.

[14] DALALYAN, A. S. and TSYBAKOV, A. B. (2010). Mirror averaging with sparsity priors. *Bernoulli* **18** 914–944. MR2948907

[15] DALALYAN, A. S. and TSYBAKOV, A. B. (2012). Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.* **78** 1423–1443. MR2926142

[16] EMERY, M., NEMIROVSKI, A. and VOICULESCU, D. (2000). *Lectures on Probability Theory and Statistics. Lecture Notes in Math.* **1738**. Springer, Berlin. MR1775638

[17] HIRIART-URRUTY, J.-B. and LEMARÉCHAL, C. (2001). *Fundamentals of Convex Analysis. Grundlehren Text Editions*. Springer, Berlin. Abridged version of *Convex Analysis and Minimization Algorithms*. I [Springer, Berlin, 1993; MR1261420 (95m:90001)] and II [ibid.; MR1295240 (95m:90002)]. MR1865628

[18] JUDITSKY, A. and NEMIROVSKI, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.* **28** 681–712. MR1792783

[19] JUDITSKY, A., RIGOLLET, P. and TSYBAKOV, A. B. (2008). Learning by mirror averaging. *Ann. Statist.* **36** 2183–2206. MR2458184

[20] KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Math*. **2033**. Springer, Heidelberg. MR2829871

[21] LECUÉ, G. (2007). Optimal rates of aggregation in classification under low noise assumption. *Bernoulli* **13** 1000–1022. MR2364224

[22] LECUÉ, G. (2007). Suboptimality of penalized empirical risk minimization in classification. In *Learning Theory*. *Lecture Notes in Computer Science* **4539** 142–156. Springer, Berlin. MR2397584

[23] LECUÉ, G. and MENDELSON, S. (2009). Aggregation via empirical risk minimization. *Probab. Theory Related Fields* **145** 591–613. MR2529440

[24] LECUÉ, G. and MENDELSON, S. (2010). Sharper lower bounds on the performance of the empirical risk minimization algorithm. *Bernoulli* **16** 605–613. MR2730641

[25] LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*: *Isoperimetry and Processes*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (3)] **23**. Springer, Berlin. MR1102015

[26] LEE, W. S., BARTLETT, P. L. and WILLIAMSON, R. C. (1996). The importance of convexity in learning with squared loss. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory* 140–146. ACM Press, New York.

[27] RIGOLLET, P. (2012). Kullback–Leibler aggregation and misspecified generalized linear models. *Ann*. *Statist*. **40** 639–665. MR2933661

[28] RIGOLLET, P. and TSYBAKOV, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann*. *Statist*. **39** 731–771. MR2816337

[29] RIGOLLET, P. and TSYBAKOV, A. B. (2012). Sparse estimation by exponential weighting. *Statist*. *Sci*. **27** 558–575. MR3025134

[30] TSYBAKOV, A. B. (2003). Optimal rate of aggregation. In *Computational Learning Theory and Kernel Machines* (*COLT*-2003). *Lecture Notes in Artificial Intelligence* **2777** 303–313. Springer, Heidelberg.

[31] TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann*. *Statist*. **32** 135–166. MR2051002

[32] YANG, Y. (2000). Combining different procedures for adaptive regression. *J*. *Multivariate Anal*. **74** 135–161. MR1790617

[33] YANG, Y. (2000). Mixing strategies for density estimation. *Ann*. *Statist*. **28** 75–87. MR1762904

CNRS, CMAP
ECOLE POLYTECHNIQUE
PALAISEAU, 91120
FRANCE
E-MAIL: guillaume.lecue@univ-mlv.fr

DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: rigollet@princeton.edu