The Annals of Statistics
2012, Vol. 40, No. 2, 639–665
DOI: 10.1214/11-AOS961

KULLBACK–LEIBLER AGGREGATION AND MISSPECIFIED GENERALIZED LINEAR MODELS[1]

BY PHILIPPE RIGOLLET

Princeton University

In a regression setup with deterministic design, we study the pure aggregation problem and introduce a natural extension from the Gaussian distribution to distributions in the exponential family. While this extension bears strong connections with generalized linear models, it does not require identifiability of the parameter or even that the model on the systematic component is true. It is shown that this problem can be solved by constrained and/or penalized likelihood maximization and we derive sharp oracle inequalities that hold both in expectation and with high probability. Finally all the bounds are proved to be optimal in a minimax sense.

1. Introduction. The last decade has witnessed a growing interest in the general problem of *aggregation*, which turned out to be a flexible way to capture many statistical learning setups. Originally introduced in the regression framework by Nemirovski (2000) and Juditsky and Nemirovski (2000) as an extension of the problem of model selection, aggregation became a mature statistical field with the papers of Tsybakov (2003) and Yang (2004) where optimal rates of aggregation were derived. Subsequent applications to density estimation [Rigollet and Tsybakov (2007)] and classification [Belomestny and Spokoiny (2007)] constitute other illustrations of the generality and versatility of aggregation methods.

The general problem of aggregation can be described as follows. Consider a finite family $\mathcal{H}$ (hereafter called *dictionary*) of candidates for a certain statistical task. Assume also that the dictionary $\mathcal{H}$ belongs to a certain linear space so that linear combinations of functions in $\mathcal{H}$ remain plausible candidates. Given a subset $\mathcal{C}$ of the linear span $\operatorname{span}(\mathcal{H})$ of $\mathcal{H}$, the goal of aggregation is to mimic the best element of $\mathcal{C}$.

One salient feature of aggregation as opposed to standard statistical modeling is that it does not rely on an underlying model. Indeed, the goal is not to estimate the parameters of an underlying "true" model but rather to construct an estimator that mimics the performance of the best model in a given class, whether this model is true or not. From a statistical analysis standpoint, this difference is significant

Received December 2010; revised December 2011.

[1]Supported in part by NSF Grants DMS-09-06424, DMS-10-53987 and AFOSR Grant A9550-08-1-0195.

MSC2010 subject classifications. Primary 62G08; secondary 62J12, 68T05, 62F11.

Key words and phrases. Aggregation, regression, classification, oracle inequalities, finite sample bounds, generalized linear models, logistic regression, minimax lower bounds.

since performance cannot be measured in terms of parameters: there is no true parameter. Rather, a stochastic optimization point of view is adopted. If $R(\cdot)$ denotes a convex risk function, the goal pursued in aggregation is to construct an *aggregate estimator* $\hat{h}$ such that

$$\mathbb{E}R(\hat{h}) \le C \min_{f \in \mathcal{C}} R(f) + \varepsilon, \tag{1.1}$$

where ε is a small term that characterizes the performance of the given aggregate $\hat{h}$. As illustrated below, the remainder term ε is an explicit function of the size M of the dictionary and the sample size n that shows the interplay between these two fundamental parameters. Such oracle inequalities with optimal remainder term ε were originally derived by Yang (2000) and Catoni (2004) for model selection in the problems of density estimation and Gaussian regression, respectively. They used a method, called *progressive mixture*, that was later extended to more general stochastic optimization problems in Juditsky, Rigollet and Tsybakov (2008). However, only bounds in expectation have been derived for this estimator and it is argued in Audibert (2008) that this estimator cannot achieve optimal remainder terms with high probability. In the same paper, Audibert suggests a different estimator that satisfies such an oracle inequality with high probability at the cost of large constants in the remainder term. One contribution (Theorem 3.2) of the present paper is to develop a new estimator that enjoys this desirable property with small constants. We also study two other aggregation problems: linear and convex aggregation.

When the model is misspecified, the minimum risk satisfies $\min_{f \in \mathcal{C}} R(f) > 0$, and it is therefore important to obtain a leading constant $C = 1$ in (1.1). Many oracle inequalities with leading constant term $C > 1$ can be found in the literature for related problems. Yang (2004) derives oracle inequalities with $C > 1$ but where the class $\mathcal{C} = \mathcal{C}_n$ actually depends on the sample size n so that $\min_{f \in \mathcal{C}_n} R(f)$ goes to 0 as n goes to infinity under additional regularity assumptions. In this paper, we focus on the so-called *pure* aggregation setup as defined by Nemirovski (2000) and Tsybakov (2003) where the class $\mathcal{C}$ is fixed and remains very general. As a result, we are only seeking oracle inequalities that have leading constant $C = 1$. Because they hold for finite M and n, such oracle inequalities are truly finite sample results.

The pure aggregation framework departs from the original problem of aggregation, where the goal was to achieve adaptation by mimicking the best of given estimators built from an independent sample. Thus a typical aggregation procedure consists in splitting the sample in two parts, using the first part to construct estimators and the second to aggregate them [see, e.g., Lecué (2007), Rigollet and Tsybakov (2007)]. This procedure relies heavily on the fact that the observations are identically distributed, which is not the case in the fixed design regression framework studied in the rest of the paper. It is worth mentioning that in the case of model selection aggregation for Gaussian regression with fixed design, the dictionary can be taken to be a family of projection or even affine estimators built from

the same sample. This specific case has been investigated in more detail by Alquier and Lounici (2011), Dalalyan and Salmon (2011), Rigollet and Tsybakov (2011), but is beyond the scope of this paper. Nevertheless, pure aggregation, where the dictionary $\mathcal{H}$ is deterministic, has grown into a field of its own [see, e.g., Bunea, Tsybakov and Wegkamp (2007), Juditsky and Nemirovski (2000), Juditsky, Rigollet and Tsybakov (2008), Lounici (2007), Nemirovski (2000), Tsybakov (2003)]. In the case of regression with fixed design studied in this paper, the dictionary can be thought of as a family of functions with minimal conditions that is expected to have good approximation properties.

Pure aggregation turns out to be a stochastic optimization problem, where the goal is to minimize an unknown risk function R over a certain set $\mathcal{C}$. This paper is devoted to the case where the risk function is given by the Kullback–Leibler divergence, and three constraint sets that were introduced in Nemirovski (2000) are investigated.

We consider an extension of aggregation for Gaussian regression that encompasses distributions for responses in a one-parameter exponential family, with particular focus on the family of Bernoulli distributions in order to cover binary classification. A natural measure of risk in this problem is related to the Kullback–Leibler divergence between the distribution of the actual observations and that of observations generated from a given model. In a way, this extension is close to generalized linear models [see, e.g., McCullagh and Nelder (1989)], which are optimally solved by maximum likelihood estimation [see, e.g., Fahrmeir and Kaufmann (1985)]. However, in the present aggregation framework, it is not assumed that there is one true model but we prove that maximum likelihood estimators still perform almost as well as the optimal solution of a suitable stochastic optimization problem. This generalized framework encompasses logistic regression as a particular case.

Throughout the paper, for any $x \in \mathbb{R}^n$, let x_j denote its jth coordinate. In other words, any vector $x \in \mathbb{R}^n$ can be written $x = (x_1, \ldots, x_n)$. Similarly an $n \times M$ matrix H has coordinates $H_{i,j}$, $1 \le i \le n$, $1 \le j \le M$. The derivative of a function $b : \mathbb{R} \to \mathbb{R}$ is denoted by b'. For any real-valued function f, we denote by $\|f\|_\infty = \sup_x |f(x)| \in [0, \infty]$, its sup-norm. Finally, for any two real numbers x and y, we use the notation $x \wedge y = \min(x, y)$ and $x \vee y = \max(x, y)$.

The paper is organized as follows. In the next section, we define the problem of Kullback–Leibler aggregation, in the context of misspecified generalized linear models. In particular, we exhibit a natural measure of performance that suggests the use of constrained likelihood maximization to solve it. Exact oracle inequalities, both in expectation and with high probability, are gathered in Section 3 and their optimality for finite M and n is assessed in Section 4. These oracle inequalities for the case of large M are illustrated on a logistic regression problem, similar to the problem of training a boosting algorithm, in Section 5. Finally, Section 6 contains the proofs of the main results together with useful properties on the concentration and the moments of sums of random variables with distribution in an exponential family.

2. Kullback–Leibler aggregation.

2.1. *Setup and notation.* Let $x_1, \dots, x_n$ be n given points in a space $\mathcal{X}$ and consider the equivalence relation $\sim$ on the space of functions $f : \mathcal{X} \to \mathbb{R}$ that is defined such that $f \sim g$ if and only if $f(x_i) = g(x_i)$ for all $i = 1, \dots, n$. Denote by $Q_{1:n}$ the quotient space associated to this equivalence relation and define the norm $\|\cdot\|$ by

$$\|f\|^2 = \frac{1}{n}\sum_{i=1}^{n} f^2(x_i), \qquad f \in Q_{1:n}.$$

Note that $\|\cdot\|$ is a norm on the quotient space but only a seminorm on the whole space of functions $f : \mathcal{X} \to \mathbb{R}$. In what follows, it will be useful to define the inner product associated to $\|\cdot\|$ by

$$\langle f, g\rangle = \frac{1}{n}\sum_{i=1}^{n} f(x_i)g(x_i).$$

Using this inner product, we can also denote the average of a function f by $\langle f, \mathbb{1}\rangle$, where $\mathbb{1}(\cdot)$ is the function in $Q_{1:n}$ that is identically equal to 1.

Recall that a random variable $Y \in \mathbb{R}$ has distribution in a (one-parameter) *canonical exponential family* if it admits a density with respect to a reference measure on $\mathbb{R}$ given by

$$p(y;\theta) = \exp\left\{\frac{y\theta - b(\theta)}{a} + c(y)\right\}. \tag{2.1}$$

A detailed treatment of exponential families of distributions together with examples can be found in Barndorff-Nielsen (1978), Brown (1986), McCullagh and Nelder (1989) and in Lehmann and Casella (1998). Several examples are also presented in Section 5 of the present paper. It can be easily shown that if Y admits a density given by (2.1), then

$$\mathbb{E}[Y] = b'(\theta) \quad \text{and} \quad \operatorname{var}[Y] = ab''(\theta). \tag{2.2}$$

We assume hereafter that the distribution of Y is not degenerate so that (2.2) ensures that b is strictly convex and b' is onto its image space.

For any $g \in Q_{1:n}$, let P_g denote the distribution of n independent random variables $Y_1, \dots, Y_n \in \mathcal{Y} \subset \mathbb{R}$ such that Y_i has density given by $p(y;\theta_i)$ where $\theta_i = [b']^{-1} \circ g(x_i)$ so that Y_i has expectation $g(x_i)$.

In this paper, we assume that we observe n independent random variables $Y_1, \dots, Y_n \in \mathcal{Y}$ with joint distribution $\mathbb{P} = P_f$ for some unknown f. We denote by $\mathbb{E}$ the corresponding expectation.

2.2. *Aggregation and misspecified generalized linear models.* When $\mathcal{X} \subset \mathbb{R}^d$, generalized linear models (GLMs) assume that the distribution of the observation Y_i belongs to a given exponential family with expectation $\mathbb{E}[Y_i] = f(x_i)$, $i = 1, \ldots, n$, and that $l \circ f(x) = \beta^\top x$ where $l : \check{\mathcal{Y}} \to \mathbb{R}$ is a *link function* and $\beta \in \mathbb{R}^d$ is the unknown parameter of interest. A canonical choice for the link function is $l = [b']^{-1}$ and in the rest of the paper, we study only this choice. In particular, this canonical choice implies that $\theta_i = \beta^\top x_i$. While GLMs allow more choices for the distribution of the response variable, the modeling assumption $\theta_i = \beta^\top x_i$ is quite strong and may be violated in practice. Aggregation offers a nice setup to study the performance of estimators of f even when this model is misspecified.

Aggregation for the regression problem was introduced by Nemirovski (2000) and further developed by Tsybakov (2003) where the author considers a regression problem with random design that has known distribution. We now recall the main ideas of aggregation applied to the regression problem, with emphasis on its difference with the linear regression model. In the framework of the previous section, consider a finite dictionary $\mathcal{H} = \{h_1, \ldots, h_M\}$ such that $\|h_j\|$ is finite and for any $\lambda \in \mathbb{R}^M$, let h_λ denote the linear combination of h_j's defined by

$$\mathsf{h}_\lambda = \sum_{j=1}^{M} \lambda_j h_j. \tag{2.3}$$

Assume that we observe n independent random couples (x_i, Y_i), $i = 1, \ldots, n$, such that $\mathbb{E}[Y_i] = f(x_i)$. The goal of *aggregation* is to solve the following optimization problem:

$$\min_{\lambda \in \Lambda} \|\mathsf{h}_\lambda - f\|^2, \tag{2.4}$$

where Λ is a given subset of $\mathbb{R}^M$ and f is unknown. Previous papers on aggregation in the regression problem have focused on three choices for the set Λ corresponding to the three different problems of aggregation originally introduced by Nemirovski (2000). Optimal rates of aggregation for these three problems in the Gaussian regression setup can be found in Tsybakov (2003).

MODEL SELECTION AGGREGATION. The goal is to mimic the best h_j in the dictionary $\mathcal{H}$. Therefore, we can choose Λ to be the finite set $\mathcal{V} = \{e_1, \ldots, e_M\}$ formed by the M vectors in the canonical basis of $\mathbb{R}^M$. The optimal rate of model selection aggregation in the Gaussian case is $(\log M)/n$.

LINEAR AGGREGATION. The goal is to mimic the best linear combination of the h_j's in the dictionary $\mathcal{H}$. Therefore, we can choose Λ to be whole space $\mathbb{R}^M$. The optimal rate of linear aggregation in the Gaussian case is M/n.

CONVEX AGGREGATION. The goal is to mimic the best convex combination of the h_j's in the dictionary $\mathcal{H}$. Therefore, we can choose Λ to be the flat simplex

of $\mathbb{R}^M$, denoted by Λ_1^+ and defined by

$$\Lambda_1^+ = \left\{ \lambda \in \mathbb{R}^M : \lambda_j \geq 0, \, j = 1, \ldots, M, \, \sum_{j=1}^{M} \lambda_j = 1 \right\}. \tag{2.5}$$

The optimal rate of convex aggregation in the Gaussian case is $(M/n) \wedge \sqrt{\log(1 + M/\sqrt{n})/n}$.

In practice, the regression function f is unknown and it is impossible to perfectly solve (2.4). Our goal is therefore to recover an approximate solution of this problem in the following sense. We wish to construct an estimator $\hat{\lambda}_n$ such that

$$\|\mathsf{h}_{\hat{\lambda}_n} - f\|^2 - \min_{\lambda \in \Lambda} \|\mathsf{h}_\lambda - f\|^2 \tag{2.6}$$

is as small as possible. An inequality that provides an upper bound on the (random) quantity in (2.6) in a certain probabilistic sense is called *oracle inequality*.

Observe that this is not a linear model since we do not assume that the function f is of the form h_λ for some $\lambda \in \mathbb{R}^M$. Rather, the bias term $\min_{\lambda \in \Lambda} \|\mathsf{h}_\lambda - f\|^2$ may not vanish and the goal is to mimic the linear combination with the smallest bias term.

The notion of Kullback–Leibler aggregation defined in the next subsection broadens the scope of the above problem of aggregation to encompass other distributions for Y.

2.3. *Kullback–Leibler aggregation.* Recall that the ubiquitous squared norm $\|\cdot\|^2$ as a measure of performance for regression problems takes its roots in the Gaussian regression model. The Kullback–Leibler divergence between two probability distributions P and Q is defined by

$$\mathcal{K}(P\|Q) = \begin{cases} \displaystyle\int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P, & \text{if } P \ll Q, \\ \infty, & \text{otherwise.} \end{cases}$$

Denote by P_f the joint distribution of the observations $Y_i, i = 1, \ldots, n$. If P_f denotes an n-variate Gaussian distribution with mean $(f(x_1), \ldots, f(x_n))^\top$ and variance $\sigma^2 I_n$, where I_n denotes the $n \times n$ identity matrix, then $\mathcal{K}(P_f\|P_g) = \frac{n}{2\sigma^2}\|f - g\|^2$. In order to allow an easier comparison between the results of this paper and the literature, consider a normalized Kullback–Leibler divergence defined by $\bar{\mathcal{K}}(P_f\|P_g) = \mathcal{K}(P_f\|P_g)/n$. In the Gaussian regression setup, the quantity of interest in (2.6) can be written

$$\bar{\mathcal{K}}(P_f\|P_{\mathsf{h}_{\hat{\lambda}_n}}) - \min_{\lambda \in \Lambda} \bar{\mathcal{K}}(P_f\|P_{\mathsf{h}_\lambda}), \tag{2.7}$$

up to a multiplicative constant term equal to $2\sigma^2$. Nevertheless, the quantity in (2.7) is meaningful for other distributions in the exponential family.

Given a subset Λ of $\mathbb{R}^M$, the goal of *Kullback–Leibler aggregation* (in short, KL-aggregation) is to construct an estimator $\hat{\lambda}_n$ such that the *excess-KL*, defined by

$$\mathcal{E}_{\mathrm{KL}}(\mathsf{h}_{\hat{\lambda}_n}, \Lambda, \mathcal{H}) = \bar{\mathcal{K}}(P_f \| P_{b' \circ \mathsf{h}_{\hat{\lambda}_n}}) - \inf_{\lambda \in \Lambda} \bar{\mathcal{K}}(P_f \| P_{b' \circ \mathsf{h}_\lambda}), \tag{2.8}$$

is as small as possible.

Whereas KL-aggregation is a purely finite sample problem, it bears connections with the asymptotic theory of model misspecification as defined in White (1982), following LeCam (1953) and Akaike (1973). White (1982) proves that if the regression function f is not of the form $f = b' \circ \mathsf{h}_\lambda$ for some λ in the set of parameters Λ, then under some identifiability and regularity conditions, the maximum likelihood estimator converges to λ^* defined by

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \mathcal{K}(P_f \| P_{b' \circ \mathsf{h}_\lambda}).$$

Upper bounds on the excess-KL can be interpreted as finite sample versions of those original results.

Note that assuming that Y_i admits a density of the form (2.1) with known cumulant function $b(\cdot)$ is a strong assumption unless Y_i has Bernoulli distribution, in which case identification of this distribution is trivial from the context of the statistical experiment. We emphasize here that model misspecification pertains only to the systematic component.

3. Main results. Let $\mathcal{Z} = \{(x_1, Y_1), \ldots, (x_n, Y_n)\}$ be n independent observations and assume that for each i, the density of Y_i is of the form $p(y_i; \theta_i)$ as defined in (2.1) where $\theta_i = [b']^{-1} \circ f(x_i)$. Then, we can write for any $\lambda \in \mathbb{R}^M$,

$$\mathcal{K}(P_f \| P_{b' \circ \mathsf{h}_\lambda}) = -\frac{n}{a}(\langle f, \mathsf{h}_\lambda \rangle - \langle b \circ \mathsf{h}_\lambda, \mathbb{1} \rangle) - \sum_{i=1}^n \mathbb{E}[c(Y_i)] + \mathrm{Ent}(P_f), \tag{3.1}$$

where $\mathrm{Ent}(P_f)$ denotes the entropy of P_f and is defined by

$$\mathrm{Ent}(P_f) = \sum_{i=1}^n \mathbb{E}\big[\log\big(p\big(Y_i; [b']^{-1} \circ f(x_i)\big)\big)\big].$$

Note that the term $-\sum_{i=1}^n \mathbb{E}[c(Y_i)] + \mathrm{Ent}(P_f)$ does not depend on λ.

For estimators of the form $\hat{\theta}_i = \mathsf{h}_\lambda(x_i)$, maximizing the log-likelihood is equivalent to maximizing

$$\ell_n(\lambda) = \sum_{i=1}^n \{Y_i \mathsf{h}_\lambda(x_i) - \langle b \circ \mathsf{h}_\lambda, \mathbb{1} \rangle\} \tag{3.2}$$

over a certain set Λ that depends on the problem at hand.

We now give bounds for the problem of KL-aggregation for the choices of Λ corresponding to the three problems of aggregation introduced in the previous section. All proofs are gathered in Section 6 and rely on the following conditions, which can be easily checked given the cumulant function b.

CONDITION 1. The set of admissible parameters is $\Theta = \mathbb{R}$ and there exists a positive constant B^2 such that

$$\sup_{\theta \in \Theta} b''(\theta) \le B^2. \tag{3.3}$$

CONDITION 2. We say that the couple $(\mathcal{H}, \Lambda)$ satisfies Condition 2 if there exists a positive constant κ^2 such that

$$b''(\mathsf{h}_\lambda(x)) \ge \kappa^2,$$

uniformly for all $x \in \mathcal{X}$ and all $\lambda \in \Lambda$.

Conditions 1 and 2 are discussed in the light of several examples in Section 5. Condition 1 is used only to ensure that the distributions of Y_i have uniformly bounded variances and sub-Gaussian tails, whereas Condition 2 is a strong convexity condition that depends not only on the cumulant function b but also on the aggregation problem at hand that is characterized by the couple $(\mathcal{H}, \Lambda)$.

3.1. *Model selection aggregation.* Recall that the goal of model selection aggregation is to mimic a function h_j such that $\mathcal{K}(P_f \| P_{b' \circ h_j}) \le \mathcal{K}(P_f \| P_{b' \circ h_k})$ for all $k \ne j$. A natural candidate would be the function in the dictionary that maximizes the function ℓ_n defined in (3.2) either over the finite set $\mathcal{V} = \{e_1, \dots, e_M\}$ formed by the M vectors in the canonical basis of $\mathbb{R}^M$ or over its convex hull. However, it has been established [see, e.g., Juditsky, Rigollet and Tsybakov (2008), Lecué (2007), Lecué and Mendelson (2009), Rigollet and Tsybakov (2012)] that such a choice is suboptimal in general. Lecué and Mendelson (2009) proved that the maximum likelihood estimator on the flat simplex Λ_1^+ defined in Section 3.3 is also suboptimal for the problem of model selection. As a consequence, we resort to a compromise between these two ideas and maximize a partially interpolated log-likelihood. Define $\hat{\lambda} \in \Lambda_1^+$ to be such that

$$\hat{\lambda} \in \operatorname*{arg\,max}_{\lambda \in \Lambda_1^+} \left\{ \sum_{j=1}^M \lambda_j \ell_n(e_j) + \ell_n(\lambda) \right\}. \tag{3.4}$$

Note that the criterion maximized in the above equation is the sum of the log-likelihood and a linear interpolation of the values of the log-likelihood at the vertices of the flat simplex. As argued above, both of these terms are needed. Indeed, using only the linear interpolation would lead us to choose $\hat{\lambda}$ to be one of the vertices of the simplex which, as mentioned above, is a suboptimal choice.

THEOREM 3.1. *Assume that Condition 1 holds and that* $(\mathcal{H}, \Lambda_1^+)$ *satisfies Condition 2. Recall that* $\mathcal{V} = \{e_1, \dots, e_M\}$ *is the finite set formed by the* M *vectors in the canonical basis of* $\mathbb{R}^M$. *Then, the aggregate* $\mathsf{h}_{\hat{\lambda}}$ *with* $\hat{\lambda}$ *defined in* (3.4) *satisfies*

$$\mathbb{E}[\mathcal{E}_{\mathrm{KL}}(\mathsf{h}_{\hat{\lambda}}, \mathcal{V}, \mathcal{H})] \le \frac{8B^2}{\kappa^2} \frac{\log M}{n}. \tag{3.5}$$

A similar result for $\mathsf{h}_{\tilde{\lambda}}$ where $\tilde{\lambda}$ are exponential weights was obtained by Dalalyan and Tsybakov (2007) for a different class of regression problems with deterministic design under the squared loss. For random design, Juditsky, Rigollet and Tsybakov (2008) obtained essentially the same results for the mirror averaging algorithm. Also for random design, Lecué and Mendelson (2009) proposed a different estimator to solve this problem and give for the first time a bound with high probability with the optimal remainder term. Such a result was claimed by Audibert (2008) for a different estimator when the design is random. Despite this recent effervescence, no bounds that hold with high probability have been derived for the deterministic design case considered here and the estimator proposed by Lecué and Mendelson (2009) is based on a sample splitting argument that does not extend to deterministic design. The next theorem aims at giving such an inequality for the aggregate $\mathsf{h}_{\hat{\lambda}}$.

THEOREM 3.2. *Assume that Condition 1 holds and that* $(\mathcal{H}, \Lambda_1^+)$ *satisfies Condition 2. Recall that* $\mathcal{V} = \{e_1, \dots, e_M\}$ *is the finite set formed by the* M *vectors in the canonical basis of* $\mathbb{R}^M$. *Then, for any* $\delta > 0$, *with probability* $1 - \delta$, *the aggregate* $\mathsf{h}_{\hat{\lambda}}$ *with* $\hat{\lambda}$ *defined in* (3.4) *satisfies*

$$\mathcal{E}_{\mathrm{KL}}(\mathsf{h}_{\hat{\lambda}_n}, \mathcal{V}, \mathcal{H}) \le \frac{8B^2}{\kappa^2} \frac{\log(M/\delta)}{n}. \tag{3.6}$$

The proofs of both theorems are gathered in Section 6.2.

3.2. *Linear aggregation.* Let $\Lambda \subset \mathbb{R}^M$ be a closed convex set or $\mathbb{R}^M$ itself. The *maximum likelihood aggregate* over $\Lambda \subset \mathbb{R}^M$ is uniquely defined as a function in the quotient space $Q_{1:n}$ by the linear combination $\mathsf{h}_{\hat{\lambda}_n}$ with coefficients given by

$$\hat{\lambda}_n \in \arg\max_{\lambda \in \Lambda} \ell_n(\lambda). \tag{3.7}$$

Note that both $\hat{\lambda}_n$ and $\lambda^* \in \arg\min_{\lambda \in \Lambda} \mathcal{K}(P_f \| P_{b' \circ \mathsf{h}_\lambda})$ exist as soon as Λ is a closed convex set [see Ekeland and Témam (1999), Chapter II, Proposition 1.2]. Likewise, from the same proposition, we find that if $\Lambda = \mathbb{R}^M$, Condition 2 entails that both $\hat{\lambda}_n$ and λ^* exist. Indeed, under Condition 2, the function b is convex coercive and thus both functionals

$$\mathsf{h}_\lambda \mapsto -\sum_{i=1}^n \{Y_i \mathsf{h}_\lambda(x_i) - \langle b \circ \mathsf{h}_\lambda, \mathbb{1} \rangle\} \quad \text{and} \quad \mathsf{h}_\lambda \mapsto -\langle f, \mathsf{h}_\lambda \rangle + \langle b \circ \mathsf{h}_\lambda, \mathbb{1} \rangle$$

are convex coercive. Thus, the aggregates h_{λ^*} and $\mathsf{h}_{\hat{\lambda}_n}$ are uniquely defined as functions in the quotient space $Q_{1:n}$, even though λ^* and $\hat{\lambda}_n$ may not be unique.

We first extend the original results of Nemirovski (2000) and Tsybakov (2003) by providing bounds on the expected excess-KL, $\mathbb{E}[\mathcal{E}_{\mathrm{KL}}(\mathsf{h}_{\hat{\lambda}_n}, \Lambda, \mathcal{H})]$ where Λ is either a closed convex set or $\Lambda = \mathbb{R}^M$, which corresponds to the problem of linear aggregation.

THEOREM 3.3. *Let Λ be a closed convex subset of $\mathbb{R}^M$ or $\mathbb{R}^M$ itself, such that $(\mathcal{H}, \Lambda)$ satisfies Condition 2. If the marginal variances satisfy $\mathbb{E}[Y_i - f(x_i)]^2 \le \sigma^2$ for any $i = 1, \dots, n$, then the maximum likelihood aggregate $\mathsf{h}_{\hat{\lambda}_n}$ over Λ satisfies*

$$\begin{aligned} \mathbb{E}[\mathcal{E}_{\mathrm{KL}}(\mathsf{h}_{\hat{\lambda}_n}, \Lambda, \mathcal{H})] &\le \frac{2\sigma^2}{a\kappa^2}\frac{D}{n}, \\ \mathbb{E}\|\mathsf{h}_{\hat{\lambda}_n} - \mathsf{h}_{\lambda^*}\|^2 &\le \frac{4\sigma^2}{\kappa^4}\frac{D}{n}, \end{aligned} \tag{3.8}$$

where $D \le M$ is the dimension of $\operatorname{span}(\mathcal{H})$ *and $\lambda^* \in \arg\min_{\lambda\in\Lambda} \mathcal{K}(P_f \| P_{b'\circ \mathsf{h}_\lambda})$.*

Vectors $\lambda^* \in \arg\min_{\lambda\in\Lambda} \mathcal{K}(P_f \| P_{b'\circ \mathsf{h}_\lambda})$ are oracles since they cannot be computed without the knowledge of P_f. The oracle distribution $P_{b'\circ \mathsf{h}_{\lambda^*}}$ corresponds to the distribution of the form $P_{b'\circ \mathsf{h}_\lambda}$, $\lambda \in \Lambda$, that is the closest to the true distribution P_f in terms of Kullback–Leibler divergence. Introducing this oracle allows us to assess the performance of the maximum likelihood aggregate, without assuming that P_f is of the form $P_{b'\circ \mathsf{h}_\lambda}$ for some $\lambda \in \Lambda$. Note also that from (2.2), the bounded variance condition $\mathbb{E}[Y_i - f(x_i)]^2 \le \sigma^2$ is a direct consequence of Condition 1 with $\sigma^2 = aB^2$.

Theorem 3.3 is valid in expectation. The following theorem shows that these bounds are not only valid in expectation but also with high probability.

THEOREM 3.4. *Let Λ be a closed convex subset of $\mathbb{R}^M$ or $\mathbb{R}^M$ itself and such that $(\mathcal{H}, \Lambda)$ satisfies Condition 2. Moreover, let Condition 1 hold and let D be the dimension of the linear span of the dictionary $\mathcal{H} = \{h_1, \dots, h_M\}$. Then, for any $\delta > 0$, with probability $1 - \delta$, the maximum likelihood aggregate $\mathsf{h}_{\hat{\lambda}_n}$ over Λ satisfies*

$$\begin{aligned} \mathcal{E}_{\mathrm{KL}}(\mathsf{h}_{\hat{\lambda}_n}, \Lambda, \mathcal{H}) &\le \frac{8B^2}{\kappa^2}\frac{D}{n}\log\biggl(\frac{4}{\delta}\biggr), \\ \|\mathsf{h}_{\hat{\lambda}_n} - \mathsf{h}_{\lambda^*}\|^2 &\le \frac{16aB^2}{\kappa^4}\frac{D}{n}\log\biggl(\frac{4}{\delta}\biggr), \end{aligned} \tag{3.9}$$

where $\lambda^* \in \arg\min_{\lambda\in\Lambda} \mathcal{K}(P_f \| P_{b'\circ \mathsf{h}_\lambda})$.

We see that the price to pay to obtain bounds with high probability is essentially the same as for the bounds in expectation up to an extra multiplicative term of order $\log(1/\delta)$.

3.3. *Convex aggregation.* In this subsection, we assume that $\Lambda \subset \Lambda_1^+$ is a closed convex set. Note that both a maximum likelihood estimator $\hat{\lambda}_n$ and an oracle $\lambda^* \in \arg\min_{\lambda\in\Lambda} \mathcal{K}(P_f \| P_{b'\circ \mathsf{h}_\lambda})$ exist.

Recall that if $(\mathcal{H}, \Lambda)$ satisfies Condition 2, Theorems 3.3 and 3.4 also hold. The following theorems ensure a better rate for the maximum likelihood aggregate $\mathsf{h}_{\hat{\lambda}_n}$ over Λ when D, and thus M, becomes much larger than n. It extends the problem of convex aggregation defined by Nemirovski (2000), Juditsky and Nemirovski (2000) and Tsybakov (2003) to the case where the distribution of the response variables is not restricted to be Gaussian.

THEOREM 3.5. *Let Λ be any closed convex subset of the flat simplex Λ_1^+ defined in (2.5). Let Condition 1 hold and assume that the dictionary $\mathcal{H}$ consists of functions satisfying $\|h_j\| \le R$, for any $j = 1, \dots, M$ and some $R > 0$. Then, the maximum likelihood aggregate $\mathsf{h}_{\hat{\lambda}_n}$ over Λ satisfies*

$$\mathbb{E}[\mathcal{E}_{\mathrm{KL}}(\mathsf{h}_{\hat{\lambda}_n}, \Lambda, \mathcal{H})] \le RB\sqrt{\frac{\log M}{an}}. \tag{3.10}$$

Moreover, if $(\mathcal{H}, \Lambda)$ satisfies Condition 2, then

$$\mathbb{E}\|\mathsf{h}_{\hat{\lambda}_n} - \mathsf{h}_{\lambda^*}\|^2 \le \frac{2RB}{\kappa^2}\sqrt{\frac{a\log M}{n}},$$

where $\lambda^* \in \arg\min_{\lambda\in\Lambda} \mathcal{K}(P_f \| P_{b'\circ \mathsf{h}_\lambda})$.

The bounds of Theorem 3.5 also have a counterpart with high probability as shown in the next theorem.

THEOREM 3.6. *Let Λ be any closed convex subset of the flat simplex Λ_1^+ defined in (2.5). Fix $M \ge 3$, let Condition 1 hold and assume that the dictionary $\mathcal{H}$ consists of functions satisfying $\|h_j\| \le R$, for any $j = 1, \dots, M$ and some $R > 0$. Then, for any $\delta > 0$, with probability $1 - \delta$, the maximum likelihood aggregate $\mathsf{h}_{\hat{\lambda}_n}$ over Λ satisfies*

$$\mathcal{E}_{\mathrm{KL}}(\mathsf{h}_{\hat{\lambda}_n}, \Lambda, \mathcal{H}) \le RB\sqrt{\frac{2\log(M/\delta)}{an}}. \tag{3.11}$$

Moreover, if $(\mathcal{H}, \Lambda)$ satisfies Condition 2, then on the same event of probability $1 - \delta$, it holds

$$\|\mathsf{h}_{\hat{\lambda}_n} - \mathsf{h}_{\lambda^*}\|^2 \le \frac{2RB}{\kappa^2}\sqrt{\frac{2a\log(M/\delta)}{n}}, \tag{3.12}$$

where $\lambda^* \in \arg\min_{\lambda\in\Lambda} \mathcal{K}(P_f \| P_{b'\circ \mathsf{h}_\lambda})$.

This explicit logarithmic dependence in the dimension M illustrates the benefit of the ℓ_1 constraint for high-dimensional problems. Raskutti, Wainwright and Yu (2011) have obtained essentially the same result as Theorem 3.6 for the special case of Gaussian linear regression. While their proof technique yields significantly larger constants, they also cover the case of aggregation over ℓ_q balls for $q < 1$ explicitly. However, their result is limited to the linear regression model where the regression function f is of the form $f = \mathsf{h}_{\lambda^*}$ for some $\lambda^* \in \Lambda_1$, where Λ_1 denotes the unit ℓ_1 ball of $\mathbb{R}^M$.

Most of the existing bounds for convex aggregation hold for the expected excess-KL. Many papers provide bounds with high probability [see, e.g., Koltchinskii (2011), Massart (2007), Mitchell and van de Geer (2009) and references therein] but they typically do not hold for the excess-KL itself but for a quantity related to

$$\bar{\mathcal{K}}(P_f \| P_{b' \circ \mathsf{h}_{\hat{\lambda}_n}}) - C \min_{\lambda \in \Lambda} \bar{\mathcal{K}}(P_f \| P_{b' \circ \mathsf{h}_\lambda}),$$

where $C > 1$ is a constant. When the quantity $\min_{\lambda \in \Lambda} \bar{\mathcal{K}}(P_f \| P_{b' \circ \mathsf{h}_\lambda})$ is not small enough, such bounds can become uninformative. A notable exception is Nemirovski et al. [(2008), Proposition 2.2] where the authors derive a result similar to Theorem 3.6 under a different but similar set of assumptions. Most importantly, their bounds do not hold for the maximum likelihood estimator but for the output of a recursive stochastic optimization algorithm.

3.4. *Discussion.* As mentioned before, it is worth noticing that the technique employed in proving the bounds in expectation of the previous subsection yield bounds with high probability at almost no extra cost.

We finally mention the question of *persistence* posed by Greenshtein and Ritov (2004) and further studied by Greenshtein (2006) and Bartlett, Mendelson and Neeman (2012). In these papers, the goal is to find performance bounds that explicitly depend on n, M and the radius R of the ℓ_1 ball $R\Lambda_1$ when the functions of the dictionary are scaled to have unit norm. Clearly, this is essentially the same problem as ours if we choose the dictionary to be $\{0, Rh_1, \dots, Rh_M, -Rh_1, \dots, -Rh_M\}$. More precisely, allowing M and R to depend on n, persistence asks the question of which regime gives remainder terms that converge to 0. While we do not pursue directly this question, we can obtain such bounds for deterministic design and show that the constrained maximum likelihood estimator on a closed convex subset of the ℓ_1 ball is persistent as long as $R = R(n) = o(\sqrt{n/\log(M)})$. The original result of Greenshtein and Ritov (2004) in this sense allows only $R = o([n/\log(M)]^{1/4})$ but when the design is random with unknown distribution. The use of deterministic design in the present paper makes the prediction task much easier. Indeed, a significant amount of work to prove persistence has been made toward describing general conditions on the distribution of the design to ensure persistence at a rate $R = o(\sqrt{n/\log(M)})$, as in Greenshtein (2006) and Bartlett, Mendelson and Neeman (2012).

4. Optimal rates of aggregation. In Section 3, we have derived upper bounds for the excess-risk both in expectation and with high probability under appropriate conditions. The bounds in expectation can be summarized as follows. For a given $\Lambda \subseteq \mathbb{R}^M$, there exists an estimator T_n such that its excess-KL satisfies

$$\mathbb{E}[\bar{\mathcal{K}}(P_f \| P_{T_n})] - \inf_{\lambda \in \Lambda} \bar{\mathcal{K}}(P_f \| P_{b' \circ \mathsf{h}_\lambda}) \le C \Delta_{n,M}(\Lambda),$$

where $C > 0$ and

$$\Delta_{n,M}(\Lambda) = \begin{cases} \dfrac{D}{n} \wedge \dfrac{\log M}{n}, & \\ \qquad \text{if } \Lambda = \mathcal{V} & \text{(model selection aggregation)}, \\ \dfrac{D}{n}, & \\ \qquad \text{if } \Lambda \subseteq \mathbb{R}^M & \text{(linear aggregation)}, \\ \dfrac{D}{n} \wedge \sqrt{\dfrac{\log M}{n}}, & \\ \qquad \text{if } \Lambda = \Lambda_1^+ & \text{(convex aggregation)}. \end{cases} \tag{4.1}$$

Here $D \le M \wedge n$ is the dimension of the linear span of the dictionary $\mathcal{H}$ and $\Lambda \subseteq \mathbb{R}^M$ means that Λ is either a closed convex subset of $\mathbb{R}^M$ or $\mathbb{R}^M$ itself. Note that for model selection aggregation, the estimator that achieves this rate is given by $T_n = b' \circ \mathsf{h}_{\tilde{\lambda}_n} \mathbb{I}(D \ge \log M) + b' \circ \mathsf{h}_{\hat{\lambda}_n} \mathbb{I}(D \le \log M)$, where $\tilde{\lambda}_n$ is defined in (3.4), $\mathsf{h}_{\hat{\lambda}_n}$ is the maximum likelihood aggregate over Λ_1^+ and $\mathbb{I}(\cdot)$ denotes the indicator function. Obviously, the lower bound for linear aggregation does not hold for *any* closed convex subset of $\mathbb{R}^M$ since $\{0\}$ is such a set and clearly $\Delta_{n,M}(\{0\}) \equiv 0$. We will prove the lower bound on the unit ℓ_∞ ball defined by $\Lambda_\infty = \{x \in \mathbb{R}^M : \max_{1 \le j \le M} |x_j| \le 1\}$.

For linear and model selection aggregation, these rates are known to be optimal in the Gaussian case where the design is random but with known distribution [Tsybakov (2003)] and where the design is deterministic [Rigollet and Tsybakov (2011)]. For convex aggregation, it has been established by Tsybakov (2003) [see also Rigollet and Tsybakov (2011)] that the optimal rate for Gaussian regression is of order $\sqrt{\log(1 + eM/\sqrt{n})/n}$, which is equivalent to the upper bounds obtained in Theorems 3.5–3.6 of the present paper when $M \gg \sqrt{n}$ but is smaller in general. To obtain better upper bounds, one may resort to more complicated, combinatorial procedures such as the ones derived in the papers cited above but the full description of this idea goes beyond the scope of this paper. Note that in the case of bounded regression with quadratic risk and random design, Lecué (2012) recently proved that the constrained empirical risk minimizer attains the optimal rate $\sqrt{\log(1 + eM/\sqrt{n})/n}$ without any modification.

In this section, we prove that these rates are minimax optimal under weaker conditions that are also satisfied by the Bernoulli distribution. The notion of optimality for aggregation employed here is a natural extension of the one introduced

TABLE 1
Exponential families of distributions and constants in Conditions 1 and 2 where H_∞ is defined in (4.3). [Source: McCullagh and Nelder (1989)]

	Θ	$\mathbb{E}(Y)$	a	$b(\theta)$	$b''(\theta)$	B^2	κ^2
Normal	$\mathbb{R}$	θ	σ^2	$\frac{\theta^2}{2}$	1	1	1
Bernoulli	$\mathbb{R}$	$\frac{e^\theta}{1+e^\theta}$	1	$\log(1+e^\theta)$	$\frac{e^\theta}{(1+e^\theta)^2}$	$\frac{1}{4}$	$\frac{e^{H_\infty}}{(1+e^{H_\infty})^2}$
Gamma	$(-\infty, 0)$	$-\frac{1}{\theta}$	$\frac{1}{\alpha}$	$-\log(-\theta)$	$1/\theta^2$	∞	$\frac{1}{H_\infty^2}$
Negative binomial	$(0, \infty)$	$\frac{r}{1-e^\theta}$	1	$r\log(\frac{e^\theta}{1-e^\theta})$	$\frac{re^\theta}{(1-e^\theta)^2}$	∞	$\frac{re^{H_\infty}}{(1-e^{H_\infty})^2}$
Poisson	$\mathbb{R}$	e^θ	1	e^θ	e^θ	∞	e^{-H_∞}

by Tsybakov (2003). Before stating the main result of this section, we need to introduce the following definition. Fix $\kappa^2 > 0$ and let $\Gamma(\kappa^2)$ be the level set of the function b'' defined by

$$\Gamma(\kappa^2) = \{\theta \in \mathbb{R} : b''(\theta) \geq \kappa^2\}. \tag{4.2}$$

In the Gaussian case, it is clear from Table 1 that $\Gamma(\kappa^2) = \mathbb{R}$ for any $\kappa^2 \leq 1$. For the cumulant function of the Bernoulli distribution, when $\kappa^2 < 1/4$, $\Gamma(\kappa^2)$ is a compact symmetric interval given by

$$\left[2\log\left(\frac{1-\sqrt{1-4\kappa^2}}{2\kappa}\right), 2\log\left(\frac{1+\sqrt{1-4\kappa^2}}{2\kappa}\right)\right].$$

Furthermore, we have $\Gamma(1/4) = \{0\}$ and $\Gamma(\kappa^2) = \varnothing$, for $\kappa^2 > 1/4$. In the next theorem, we assume that for a given $\kappa^2 > 0$, $\Gamma(\kappa^2)$ is convex. This is clearly the case when the cumulant function b is such that b'' is quasi-concave, that is, satisfies for any $\theta, \theta' \in \mathbb{R}, u \in [0,1]$, $b''(u\theta + (1-u)\theta') \geq \min[b''(\theta), b''(\theta')]$. This assumption is satisfied for the Gaussian and Bernoulli distributions.

Let $\bar{\mathcal{D}}$ denote the class of dictionaries $\mathcal{H} = \{h_1, \ldots, h_M\}$ such that $\|h_j\|_\infty \leq 1$, $j = 1, \ldots, M$. Moreover, for any convex set $\Lambda \subseteq \mathbb{R}^M$, denote by $I(\Lambda)$ the interval $[-H_\infty, H_\infty]$, where

$$H_\infty = H_\infty(\Lambda) = \sup_{\mathcal{H}\in\bar{\mathcal{D}}} \sup_{\lambda\in\Lambda} \sup_{x\in\mathcal{X}} |\mathsf{h}_\lambda(x)| \in [0, \infty]. \tag{4.3}$$

For example, we have

$$I(\Lambda) = \begin{cases} [-1,1], & \text{if } \Lambda = \mathcal{V} \quad \text{(model selection aggregation)}, \\ \mathbb{R}, & \text{if } \Lambda = \mathbb{R}^M \quad \text{(linear aggregation)}, \\ [-1,1], & \text{if } \Lambda = \Lambda_1^+ \quad \text{(convex aggregation)}. \end{cases}$$

To state the minimax lower bounds properly, we use the notation

$$\mathfrak{E}_{\mathrm{KL}}(T_n, \Lambda, f, \mathcal{H}) = \mathbb{E}[\bar{\mathcal{K}}(P_f \| P_{T_n})] - \inf_{\lambda\in\Lambda} \bar{\mathcal{K}}(P_f \| P_{b'\circ\mathsf{h}_\lambda}),$$

that makes the dependence in the regression function f explicit. Finally, we denote by E_f the expectation with respect to the distribution P_f.

THEOREM 4.1. *Fix $M \geq 2, n \geq 1, D \geq 1, \kappa^2 > 0$, and assume that Condition 1 holds. Moreover, assume that for a given set $\Lambda \subseteq \mathbb{R}^M$, we have $I(\Lambda) \subset \Gamma(\kappa^2)$. Then, there exists a dictionary $\mathcal{H} \in \bar{\mathcal{D}}$, with rank less than D, and positive constants c^*, δ such that*

$$\inf_{T_n} \sup_{\lambda \in \Lambda} P_{b' \circ \mathsf{h}_\lambda} \left[\mathfrak{E}_{\mathrm{KL}}(T_n, \Lambda, b' \circ \mathsf{h}_\lambda, \mathcal{H}) > c_* \frac{\kappa^2}{2a} \Delta^*_{n,M}(\Lambda) \right] \geq \delta \tag{4.4}$$

and

$$\inf_{T_n} \sup_{\lambda \in \Lambda} E_{b' \circ \mathsf{h}_\lambda} [\mathfrak{E}_{\mathrm{KL}}(T_n, \Lambda, b' \circ \mathsf{h}_\lambda, \mathcal{H})] \geq \delta c_* \frac{\kappa^2}{2a} \Delta^*_{n,M}(\Lambda), \tag{4.5}$$

where the infimum is taken over all estimators and where

$$\Delta^*_{n,M}(\Lambda) = \begin{cases} \dfrac{D}{n} \wedge \dfrac{\log M}{n}, & \text{if } \Lambda = \mathcal{V}, \\ \dfrac{D}{n}, & \text{if } \Lambda \supset \Lambda_\infty(1), \\ \dfrac{D}{n} \wedge \sqrt{\dfrac{\log(1 + eM/\sqrt{n})}{n}}, & \text{if } \Lambda = \Lambda_1^+. \end{cases} \tag{4.6}$$

This theorem covers the Gaussian and the Bernoulli case for which Condition 1 is satisfied. Lower bounds for aggregation in the Gaussian case have already been proved in Rigollet and Tsybakov [(2011), Section 6] in a weaker sense. Indeed, we enforce here that $\mathcal{H} \in \bar{\mathcal{D}}$ and has rank bounded by D, whereas Rigollet and Tsybakov (2011) use unbounded dictionaries with rank that may exceed D by a logarithmic multiplicative factor.

Observe that from (4.5), the least favorable regression functions are of the form $f = b' \circ \mathsf{h}_\lambda, \lambda \in \Lambda$, as it is the case for Gaussian aggregation [see, e.g., Tsybakov (2003)].

A consequence of Theorem 4.1 is that the rates of convergence obtained in Section 3, both in expectation and with high probability, cannot be improved without further assumptions except for the logarithmic term of convex aggregation. The proof of Theorem 4.1 is provided in the supplementary material [Rigollet (2012)].

5. Examples.

5.1. *Examples of exponential families.* This subsection is a reminder of the versatility of exponential families of distributions and its goal is to illustrate Conditions 1 and 2 on some examples. Most of the material can be found, for example,

in McCullagh and Nelder (1989). The form of the density described in (2.1) is usually referred to as natural form. We now recall that it already encompasses many different distributions. Table 1 gives examples of distributions that have such a density. For distributions with several parameters, it is assumed that all parameters but θ are known. For the Normal and Gamma distributions, the reference measure is the Lebesgue measure whereas for the Bernoulli, Negative binomial and Poisson distributions, the reference measure is the counting measure on $\mathbb{Z}$. For all these distributions, the cumulant function $b(\cdot)$ is twice continuously differentiable.

Observe first that only the Normal and Bernoulli distributions satisfy Condition 1. Indeed, all other distributions in the table do not have sub-Gaussian tails and therefore, we cannot use Lemma 6.1 to control the deviations and moments of the sum of independent random variables. Therefore, only Theorem 3.3 applies to the remaining distributions even though direct computation of the moments can yield results of the same type as Theorems 3.5 and 3.6 but with bounds that are larger by orders of magnitude.

Another important message of Table 1 is that the constant κ^2 can depend on the constant H_∞ defined in (4.3). Consequently the L_2 distance $\|\mathsf{h}_{\hat{\lambda}_n} - \mathsf{h}_{\lambda^*}\|^2$ is affected by the constant κ^2 and thus by H_∞. However, the constant B^2 does not depend on H_∞. Therefore, the bounds on the excess-KL presented in Theorems 3.5 and 3.6 hold without extra assumption of the dictionary. For the Normal distribution, $\kappa^2 = B^2 = 1$ regardless of the value H_∞, which makes it a particular case.

5.2. *Bounds for logistic regression with a large dictionary.* Let us now focus on the Bernoulli distribution. Recall that in the setup of binary classification, we observe a collection of independent random couples $(x_1, Y_1), \dots, (x_n, Y_n)$ such that $Y_i \in \{0, 1\}$ has Bernoulli distribution with parameter $f(x_i), i = 1, \dots, n$. As shown in the survey by Boucheron, Bousquet and Lugosi (2005), there exists a tremendous amount of work in this topic and we will focus on the so-called boosting type algorithms. A dictionary of base classifiers $\mathcal{H} = \{h_1, \dots, h_M\}$, that is, functions taking values in $[-1, 1]$, is given and training a boosting algorithm consists in combining them in such a way that $\mathsf{h}_\lambda(x_i)$ predicts $f(x_i)$ well.

This part of the paper is mostly inspired by Friedman, Hastie and Tibshirani (2000) who propose a statistical view of boosting following an original remark of Breiman (1999). Specifically, they offer an interpretation of the original AdaBoost algorithm introduced in Freund and Schapire (1996) as a sequential optimization procedure that fits an extended additive model for a particular choice of the loss function. Then they propose to directly maximize the Bernoulli log-likelihood using quasi-Newton optimization and derive a new algorithm called LogitBoost. Even though we do not detail how maximization of the likelihood is performed, LogitBoost aims at solving the same problem as the one studied here. One difference here is that while extended additive models assume that there exists $\lambda \in \Lambda \subset \mathbb{R}^M$ such that the regression function is of the form $f = [b']^{-1} \circ \mathsf{h}_\lambda$,

KL-aggregation does not. The paper of Friedman, Hastie and Tibshirani (2000) focuses on the optimization side of the problem and does not contain finite sample results. A recent attempt to compensate for a lack of statistical analysis can be found in Mease and Wyner (2008) and the many discussions that it produced. We propose to contribute to this discussion by illustrating some statistical aspects of LogitBoost based on the rates derived in Section 3 and in particular, how its performance depends on the size of the dictionary.

Given a convex subset $\Lambda \subset \mathbb{R}^M$ and a convex function $\varphi : \mathbb{R} \to \mathbb{R}$, training a boosting algorithm, and more generally a large margin classifier, consists in minimizing the risk function defined by

$$R_\varphi(\mathsf{h}_\lambda) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\varphi(-\tilde{Y}_i \mathsf{h}_\lambda(x_i))]$$

over $\lambda \in \Lambda$, where $\tilde{Y}_i = 2Y_i - 1 \in \{-1, 1\}$. It is not hard to show that minimizing the Kullback–Leibler divergence $\mathcal{K}(P_f \| P_{b' \circ \mathsf{h}_\lambda})$, is equivalent to choosing

$$\varphi(x) = \frac{\log(1 + e^x)}{\log 2}, \tag{5.1}$$

up to the normalizing constant $\log 2$ that appears to ensure that $\varphi(0) = 1$. For the choice of φ defined in (5.1), we have

$$R_\varphi(\mathsf{h}_\lambda) - \min_{\lambda \in \Lambda} R_\varphi(\mathsf{h}_\lambda) = \frac{1}{\log 2}\mathcal{E}_{\mathrm{KL}}(\mathsf{h}_\lambda, \Lambda, \mathcal{H}).$$

In boosting algorithms, the size of the dictionary M is much larger than the sample size n so that the results of Theorems 3.3 and 3.4 are useless and it is necessary to constrain λ to be in the rescaled flat simplex $R\Lambda_1^+$ so that $H_\infty = R$. Given that for the Bernoulli distribution, we have $a = 1$, $B^2 = 1/4$, the constants in the main theorems can be explicitly computed and in fact, they remain low. We can therefore apply Theorems 3.5 and 3.6 to obtain the following corollary that gives oracle inequalities for the φ-risk R_φ, both in expectation and with high probability. We focus on the case where M is (much) larger than n as it is usually the case in boosting.

COROLLARY 5.1. *Consider the boosting problem with a given dictionary of base classifiers and let φ be the convex function defined in (5.1). Then, the maximum likelihood aggregate $\mathsf{h}_{\hat{\lambda}_n}$ over the rescaled flat simplex $R\Lambda_1^+$, $R > 0$, defined in (3.7) satisfies*

$$\mathbb{E}[R_\varphi(\mathsf{h}_{\hat{\lambda}_n})] \le \min_{\lambda \in R\Lambda_1^+} R_\varphi(\mathsf{h}_\lambda) + \frac{R}{2\log 2}\sqrt{\frac{\log M}{n}}.$$

Moreover, for any $\delta > 0$, with probability $1 - \delta$, it holds

$$R_\varphi(\mathsf{h}_{\hat{\lambda}_n}) \le \min_{\lambda \in R\Lambda_1^+} R_\varphi(\mathsf{h}_\lambda) + \frac{R}{2\log 2}\sqrt{\frac{2\log(M/\delta)}{n}}.$$

6. Proof of the main results. In this section, we prove the main theorems. We begin by recalling some properties of exponential families of distributions. While similar results can be found in the literature, the results presented below are tailored to our needs. In particular, the constants in the upper bounds are explicit and kept as small as possible. In this section, for any $\omega \in \ell_2(\mathbb{R})$, denote by $|\omega|_2$ its ℓ_2-norm.

6.1. *Some useful results on canonical exponential families.* Let $Y \in \mathbb{R}$ be a random variable with distribution in a canonical exponential family that admits a density with respect to a reference measure on $\mathbb{R}$ given by

$$p(y;\theta) = \exp\left\{\frac{y\theta - b(\theta)}{a} + c(y)\right\}, \qquad \theta \in \mathbb{R}. \tag{6.1}$$

It can be easily shown [see, e.g., Lehmann and Casella (1998), Theorem 5.10] that the moment generating function of Y is given by

$$\mathbb{E}[e^{tY}] = e^{(b(\theta+at)-b(\theta))/a}. \tag{6.2}$$

Using (6.2) we can derive the Chernoff-type bounds presented in the following lemma.

LEMMA 6.1. *Let $\omega = (\omega_1, \dots, \omega_n) \in \mathbb{R}^n$ be a vector of deterministic weights. Let $Y_1, \dots, Y_n$ be independent random variables such that Y_i has density $p(\cdot;\theta_i)$ defined in (6.1), $\theta_i \in \mathbb{R}$, $i = 1, \dots, n$, and define the weighted sum $S_n^\omega = \sum_{i=1}^n \omega_i Y_i$. Assume that Condition 1 holds. Then the following inequalities hold*:

$$\mathbb{E}\big[\exp\big(s|S_n^\omega - \mathbb{E}(S_n^\omega)|\big)\big] \le \exp\left(\frac{s^2 B^2 a|\omega|_2^2}{2}\right), \tag{6.3}$$

$$\mathbb{P}[|S_n^\omega - \mathbb{E}(S_n^\omega)| > t] \le 2\exp\left(-\frac{t^2}{2aB^2|\omega|_2^2}\right), \tag{6.4}$$

and for any $r \ge 0$, we have

$$\mathbb{E}|S_n^\omega - \mathbb{E}(S_n^\omega)|^r \le C_r|\omega|_2^r, \tag{6.5}$$

where $C_r = r(2aB^2)^{r/2}\Gamma(r/2)$ and $\Gamma(\cdot)$ denotes the Gamma function.

PROOF. Using, respectively, (6.2), (2.2) and (3.3), we get

$$\begin{aligned}\mathbb{E}\big[\exp\big(s\big(S_n^\omega - \mathbb{E}(S_n^\omega)\big)\big)\big] &= \exp\left(\frac{1}{a}\sum_{i=1}^n [b(\theta_i + as\omega_i) - b(\theta_i) - as\omega_i b'(\theta_i)]\right) \\ &\le \exp\left(\frac{s^2 B^2 a|\omega|_2^2}{2}\right).\end{aligned}$$

The same inequality holds with s replaced by $-s$ so (6.3) holds.

The proof of (6.4) follows from (6.3) together with a Chernoff bound. Next, note that

$$\mathbb{E}|S_n^\omega - \mathbb{E}(S_n^\omega)|^r = \int_0^\infty \mathbb{P}\big(|S_n^\omega - \mathbb{E}(S_n^\omega)| > t^{1/r}\big)\,\mathrm{d}t \le 2\int_0^\infty \exp\Big(-\frac{t^{2/r}}{2aB^2|\omega|_2^2}\Big)\,\mathrm{d}t,$$

where we used (6.4) in the last inequality. Using a change of variable, it is not hard to see that this bound yields (6.5). □

6.2. *Proof of Theorems 3.1 and 3.2.* According to (3.1), minimizing $\lambda \mapsto \mathcal{K}(P_f \| P_{b'\circ \mathsf{h}_\lambda})$ is equivalent to maximizing $\lambda \mapsto L(\lambda)$ where

$$L(\lambda) = \langle f, \mathsf{h}_\lambda\rangle - \langle b \circ \mathsf{h}_\lambda, \mathbb{1}\rangle. \tag{6.6}$$

Note that for any $\Lambda \subset \mathbb{R}^M$, the set of optimal solutions Λ^* satisfies

$$\Lambda^* = \arg\min_{\lambda\in\Lambda} \mathcal{K}(P_f \| P_{b'\circ \mathsf{h}_\lambda}) = \arg\max_{\lambda\in\Lambda} L(\lambda).$$

Moreover, for any $\lambda \in \Lambda$, $\lambda^* \in \Lambda^*$, we have

$$L(\lambda^*) - L(\lambda) = a\mathcal{E}_{\mathrm{KL}}(\mathsf{h}_\lambda, \Lambda, \mathcal{H}). \tag{6.7}$$

For any fixed $\lambda \in \Lambda_1^+$, define the following quantities:

$$S_n(\lambda) = \sum_{j=1}^M \lambda_j \ell_n(e_j) + \ell_n(\lambda),$$

$$S(\lambda) = n\sum_{j=1}^M \lambda_j L(e_j) + nL(\lambda)$$

and observe that $S(\lambda) = \mathbb{E}[S_n(\lambda)]$ and that for any $\lambda \in \Lambda_1^+$,

$$S_n(\lambda) - S(\lambda) = 2\sum_{i=1}^n \big(Y_i - f(x_i)\big)\mathsf{h}_\lambda(x_i).$$

Let $\beta > 0$ be a parameter to be chosen later. By definition of $\hat\lambda$, we have for any $\lambda \in \Lambda_1^+$ that

$$S(\hat\lambda) \ge S(\lambda) - \Delta_n(\lambda) - \beta \log M, \tag{6.8}$$

where $\Delta_n(\lambda) = 2\sum_{i=1}^n (Y_i - f(x_i))\mathsf{h}_{\hat\lambda-\lambda}(x_i) - \beta\log M$. The following lemma is useful to control the term $\Delta_n(\lambda)$ both in expectation and with high probability.

LEMMA 6.2. *Under Condition 1, for any $\lambda \in \Lambda_1^+$ we have*

$$\mathbb{E}\left[\exp\left(\frac{\Delta_n(\lambda)}{\beta} - \frac{2B^2an}{\beta^2}\sum_{j=1}^M \hat\lambda_j \|h_j - \mathsf{h}_\lambda\|^2\right)\right] \le 1.$$

PROOF. For any $\lambda \in \Lambda_1^+$, $j = 1, \dots, M$, define Υ_j by

$$\Upsilon_j(\lambda) = \frac{2B^2 an}{\beta^2} \|h_j - \mathsf{h}_\lambda\|^2.$$

Jensen's inequality and the fact that $\log M = \sum_{j=1}^M \hat{\lambda}_j (\log M)$ yield

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\Delta_n(\lambda)}{\beta} - \sum_{j=1}^M \hat{\lambda}_j \Upsilon_j(\lambda)\right)\right] \\
&\quad \le \mathbb{E}\left[\sum_{j=1}^M \hat{\lambda}_j \exp\left(\frac{2}{\beta}\sum_{i=1}^n \big(Y_i - f(x_i)\big)\big(h_j(x_i) - \mathsf{h}_\lambda(x_i)\big) - \log M - \Upsilon_j(\lambda)\right)\right] \\
&\quad \le \frac{1}{M}\sum_{j=1}^M \mathbb{E}\left[\exp\left(\frac{2}{\beta}\sum_{i=1}^n \big(Y_i - f(x_i)\big)\big(h_j(x_i) - \mathsf{h}_\lambda(x_i)\big) - \Upsilon_j(\lambda)\right)\right].
\end{aligned}$$

Now, from (6.3), which holds under Condition 1, we have for any $\lambda \in \Lambda_1^+$, $j = 1, \dots, M$, that

$$\mathbb{E}\left[\exp\left(\frac{2}{\beta}\sum_{i=1}^n \big(Y_i - f(x_i)\big)\big(h_j(x_i) - \mathsf{h}_\lambda(x_i)\big)\right)\right] \le \exp(\Upsilon_j(\lambda)),$$

and the result of the lemma follows from the previous two displays. □

Take any $\bar{\lambda} \in \arg\max_{\lambda \in \Lambda_1^+} S(\lambda)$ and observe that Condition 2 together with a second-order Taylor expansion of the function $S(\cdot)$ around $\bar{\lambda}$ gives for any $\lambda \in \Lambda_1^+$

$$S(\lambda) \le S(\bar{\lambda}) + [\nabla_\lambda S(\bar{\lambda})]^\top (\lambda - \bar{\lambda}) - \frac{n\kappa^2}{2}\|\mathsf{h}_\lambda - \mathsf{h}_{\bar{\lambda}}\|^2,$$

where $\nabla_\lambda S(\bar{\lambda})$ denotes the gradient of $\lambda \mapsto S(\lambda)$ at $\bar{\lambda}$. Since $\bar{\lambda}$ is a maximizer of $\lambda \mapsto S(\lambda)$ over the set Λ_1^+ to which λ also belongs, we find that $\nabla_\lambda S(\bar{\lambda})^\top (\lambda - \bar{\lambda}) \le 0$ so that, together with (6.8), the previous display yields

$$\frac{n\kappa^2}{2}\|\mathsf{h}_{\hat{\lambda}} - \mathsf{h}_{\bar{\lambda}}\|^2 \le S(\bar{\lambda}) - S(\hat{\lambda}) \le \Delta_n(\bar{\lambda}) + \beta \log M. \tag{6.9}$$

PROOF OF THEOREM 3.1. Using the convexity inequality $t \le e^t - 1$ for any $t \in \mathbb{R}$, Lemma 6.2 yields

$$\mathbb{E}[\Delta_n(\bar{\lambda})] \le \beta \mathbb{E}\sum_{j=1}^M \hat{\lambda}_j \Upsilon_j(\bar{\lambda}) = \beta \mathbb{E}\sum_{j=1}^M \hat{\lambda}_j \Upsilon_j(\hat{\lambda}) + \frac{2B^2 an}{\beta}\sum_{j=1}^M \mathbb{E}\|\mathsf{h}_{\hat{\lambda}} - \mathsf{h}_{\bar{\lambda}}\|^2.$$

The previous display combined with (6.9) gives

$$S(\bar{\lambda}) - \mathbb{E}[S(\hat{\lambda})] \le \beta \mathbb{E}\sum_{j=1}^M \hat{\lambda}_j \Upsilon_j(\hat{\lambda}) + \frac{4B^2 a}{\beta\kappa^2}\big[S(\bar{\lambda}) - \mathbb{E}[S(\hat{\lambda})]\big] + \beta \log M.$$

It implies that for $\beta \geq 8B^2a/\kappa^2$

$$S(\bar{\lambda}) - \mathbb{E}[S(\hat{\lambda})] \leq 2\beta \mathbb{E} \sum_{j=1}^{M} \hat{\lambda}_j \Upsilon_j(\hat{\lambda}) + 2\beta \log M. \tag{6.10}$$

Observe now that a second-order Taylor expansion of the function $L(\cdot)$ around $\hat{\lambda}$, together with Condition 2, gives for any $\lambda \in \Lambda_1^+$

$$L(\lambda) \leq L(\hat{\lambda}) + [\nabla_\lambda L(\hat{\lambda})]^\top (\lambda - \hat{\lambda}) - \frac{\kappa^2}{2} \|\mathsf{h}_\lambda - \mathsf{h}_{\hat{\lambda}}\|^2.$$

Thus

$$\sum_{j=1}^{M} \hat{\lambda}_j L(e_j) \leq L(\hat{\lambda}) - \frac{\kappa^2}{2} \sum_{j=1}^{M} \hat{\lambda}_j \|h_j - \mathsf{h}_{\hat{\lambda}}\|^2.$$

It follows that

$$S(\hat{\lambda}) = n \sum_{j=1}^{M} \hat{\lambda}_j L(e_j) + nL(\hat{\lambda}) \leq 2nL(\hat{\lambda}) - \frac{n\kappa^2}{2} \sum_{j=1}^{M} \hat{\lambda}_j \|h_j - \mathsf{h}_{\hat{\lambda}}\|^2.$$

Combined with (6.10), the above inequality yields

$$S(\bar{\lambda}) - 2n\mathbb{E}[L(\hat{\lambda})] \leq \left(2\beta - \frac{\kappa^2\beta^2}{4B^2a}\right) \mathbb{E} \sum_{j=1}^{M} \hat{\lambda}_j \Upsilon_j(\hat{\lambda}) + 2\beta \log M \leq 2\beta \log M$$

for $\beta \geq 8B^2a/\kappa^2$. Note that for any $j = 1, \ldots, M$, $S(\bar{\lambda}) \geq S(e_j) = 2nL(e_j)$ so that from (6.7), we get

$$a\mathbb{E}[\mathcal{E}_{\mathrm{KL}}(\mathsf{h}_{\hat{\lambda}}, \mathcal{V}, \mathcal{H})] = \max_{1 \leq j \leq M} L(e_j) - \mathbb{E}[L(\hat{\lambda})] \leq \frac{\beta}{n} \log M. \qquad \square$$

PROOF OF THEOREM 3.2. From Lemma 6.2 and a Chernoff bound, we get for any $\lambda \in \Lambda_1^+$ and any $\delta > 0$ that

$$\mathbb{P}\left[\Delta_n(\lambda) - \frac{2B^2an}{\beta} \sum_{j=1}^{M} \hat{\lambda}_j \|h_j - \mathsf{h}_\lambda\|^2 > \beta \log(1/\delta)\right] \leq \delta.$$

Thus, the event $\mathcal{A}_\lambda(\delta) = \{\Delta_n(\lambda) \leq \frac{2B^2an}{\beta} \sum_{j=1}^{M} \hat{\lambda}_j \|h_j - \mathsf{h}_\lambda\|^2 + \beta \log(1/\delta)\}$ has probability greater than $1 - \delta$. Theorem 3.2 follows by applying the same steps as in the proof of Theorem 3.1 but on the event $\mathcal{A}_{\bar{\lambda}}(\delta)$ instead of in expectation. $\square$

6.3. *Proofs of Theorems 3.3–3.6.* The following lemma exploits the strong convexity property stated in Condition 2.

LEMMA 6.3. *Let $\phi_1, \dots, \phi_D$ be an orthonormal basis of the linear span of the dictionary $\mathcal{H}$. Let Λ be a closed convex subset of $\mathbb{R}^M$ or $\mathbb{R}^M$ itself and assume that $(\mathcal{H}, \Lambda)$ satisfies Condition 2. Denote by λ^* any maximizer of the function $\lambda \mapsto L(\lambda)$ over the set Λ. Then any maximum likelihood estimator $\hat{\lambda}_n$ satisfies*

$$\frac{\kappa^2}{2}\|\mathsf{h}_{\hat{\lambda}_n} - \mathsf{h}_{\lambda^*}\|^2 \le L(\lambda^*) - L(\hat{\lambda}_n) \le \frac{2}{\kappa^2}\sum_{j=1}^{D} \zeta_j^2, \tag{6.11}$$

where $\zeta_j = \frac{1}{n}\sum_{i=1}^n Y_i \phi_j(x_i) - \langle f, \phi_j \rangle$, $j = 1, \dots, D$. Moreover, if $\Lambda \subset \Lambda_1^+$ is a closed convex set, then $\hat{\lambda}_n$ satisfies

$$\frac{\kappa^2}{2}\|\mathsf{h}_{\hat{\lambda}_n} - \mathsf{h}_{\lambda^*}\|^2 \le L(\lambda^*) - L(\hat{\lambda}_n) \le \max_{1 \le j \le M} |\xi_j|, \tag{6.12}$$

where $\xi_j = \frac{1}{n}\sum_{i=1}^n Y_i h_j(x_i) - \langle f, h_j \rangle$, $j = 1, \dots, M$.

PROOF. A second-order Taylor expansion of the function $L(\cdot)$ around λ^* gives for any $\lambda \in \Lambda$

$$L(\lambda) \le L(\lambda^*) + [\nabla_\lambda L(\lambda^*)]^\top (\lambda - \lambda^*) - \frac{\kappa^2}{2}\|\mathsf{h}_\lambda - \mathsf{h}_{\lambda^*}\|^2,$$

where we used Condition 2 and where $\nabla_\lambda L(\lambda^*)$ denotes the gradient of $\lambda \mapsto L(\lambda)$ at λ^*. Since λ^* is a maximizer of $\lambda \mapsto L(\lambda)$ over the set Λ to which λ also belongs, we find that $\nabla_\lambda L(\lambda^*)^\top (\lambda - \lambda^*) \le 0$ so that

$$L(\lambda^*) - L(\lambda) \ge \frac{\kappa^2}{2}\|\mathsf{h}_\lambda - \mathsf{h}_{\lambda^*}\|^2 \tag{6.13}$$

for any $\lambda \in \Lambda$, which gives the left inequalities in (6.11) and (6.12).

Next, from the definition of $\hat{\lambda}_n$, we have

$$L(\hat{\lambda}_n) \ge L(\lambda^*) + T_n(\lambda^* - \hat{\lambda}_n), \tag{6.14}$$

where

$$T_n(\mu) = \frac{1}{n}\sum_{i=1}^n Y_i \mathsf{h}_\mu(x_i) - \langle f, \mathsf{h}_\mu \rangle, \qquad \mu \in \mathbb{R}^M.$$

Writing $\mathsf{h}_\mu = \sum_{j=1}^D \nu_j \phi_j$, $\nu \in \mathbb{R}^D$, we find that

$$T_n(\mu) = \sum_{j=1}^D \nu_j \left(\frac{1}{n}\sum_{i=1}^n Y_i \phi_j(x_i) - \langle f, \phi_j \rangle \right) = \sum_{j=1}^D \nu_j \zeta_j.$$

Define the random variable $V_n = \sup_{\mu\in\mathbb{R}^M:\,\|\mathsf{h}_\mu\|>0}\{|T_n(\mu)|/\|\mathsf{h}_\mu\|\}$, so that V_n satisfies

$$V_n = \sup_{\substack{\nu\in\mathbb{R}^M\\ \nu\neq 0}} \frac{|\sum_{j=1}^D \nu_j\zeta_j|}{(\sum_{j=1}^D \nu_j^2)^{1/2}} = \left(\sum_{j=1}^D \zeta_j^2\right)^{1/2}.$$

Since $T_n(\lambda^* - \hat\lambda_n) \geq -V_n\|\mathsf{h}_{\lambda^*-\hat\lambda_n}\|$, it yields together with (6.14) that

$$L(\hat\lambda_n) \geq L(\lambda^*) - \|\mathsf{h}_{\lambda^*-\hat\lambda_n}\|\left(\sum_{j=1}^D \zeta_j^2\right)^{1/2}. \tag{6.15}$$

Combining (6.15) and (6.13) with $\lambda = \hat\lambda_n$, we get (6.11).

We now turn to the proof of (6.12). From (6.14), and the Hölder inequality, we have

$$L(\lambda^*) - L(\hat\lambda_n) \leq \left(\sum_{j=1}^M |\hat\lambda_{n,j} - \lambda_j^*|\right)\max_{1\leq j\leq M}|\xi_j| \leq \max_{1\leq j\leq M}|\xi_j|.$$

Combined with (6.13), this inequality yields (6.12). □

In view of (6.7), to complete the proof of Theorems 3.3–3.6, it is sufficient to bound from above the quantities appearing on the right-hand side of (6.11) and (6.12). This is done using results from Section 6.1 and by observing that the random variables ζ_j and ξ_j are of the form

$$\zeta_j = S_n^{\omega^{(\zeta_j)}} - \mathbb{E}\big(S_n^{\omega^{(\zeta_j)}}\big), \qquad \omega_i^{(\zeta_j)} = \frac{\phi_j(x_i)}{n}, \qquad |\omega^{(\zeta_j)}|_2 = \frac{1}{\sqrt{n}} \tag{6.16}$$

and

$$\xi_j = S_n^{\omega^{(\xi_j)}} - \mathbb{E}\big(S_n^{\omega^{(\xi_j)}}\big), \qquad \omega_i^{(\xi_j)} = \frac{h_j(x_i)}{n}, \qquad |\omega^{(\xi_j)}|_2 \leq \frac{R}{\sqrt{n}}, \tag{6.17}$$

if $\max_{1\leq j\leq M}\|h_j\| \leq R$.

PROOF OF THEOREM 3.3. Since the random variables $Y_i, i = 1,\dots,n$, are mutually independent, we have

$$\mathbb{E}[\zeta_j^2] = \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^n Y_i\phi_j(x_i)\right) \leq \frac{\sigma^2}{n^2}\sum_{i=1}^n \phi_j^2(x_i) = \frac{\sigma^2}{n}.$$

Together with (6.7) and (6.11), this bound completes the proof of Theorem 3.3. □

PROOF OF THEOREM 3.4. For any $s, t > 0$, we have

$$\mathbb{P}\left[\sum_{j=1}^{D} \zeta_j^2 > t\right] = \mathbb{P}\left[\frac{1}{D}\sum_{j=1}^{D} \zeta_j^2 > \frac{t}{D}\right] \le e^{-st/D}\mathbb{E}\big[e^{(s/D)\sum_{j=1}^{D}\zeta_j^2}\big]$$

$$\le e^{-st/D}\frac{1}{D}\sum_{j=1}^{D}\mathbb{E}[e^{s\zeta_j^2}] \le e^{-st/D}\frac{1}{D}\sum_{j=1}^{D}\sum_{p=0}^{\infty}\frac{s^p}{p!}\mathbb{E}[\zeta_j^{2p}],$$

where we used, respectively: the Markov inequality, the Jensen inequality and Fatou's lemma. Observe now that (6.5), which holds under Condition 1, and (6.16) yield

$$\mathbb{E}[\zeta_j^{2p}] \le C_{2p}|\omega^{(\zeta_j)}|_2^{2p} = \frac{C_{2p}}{n^p} = 2(p!)\left(\frac{2aB^2}{n}\right)^p.$$

Therefore, the last two displays with $s = n/(4aB^2)$ yield

$$\mathbb{P}\left(\sum_{j=1}^{D}\zeta_j^2 > t\right) \le 4e^{-nt/(4aB^2D)}.$$

Theorem 3.4 follows by taking $t = \frac{4aB^2D}{n}\log(4/\delta)$ in the previous display together with (6.7) and (6.11). □

Before completing the proof of Theorems 3.5 and 3.6, observe that (6.3) and (6.17) imply that for any $j = 1, \dots, M$, the random variable $|\xi_j|$ is sub-Gaussian with variance proxy $\sigma^2 = (RB)^2 a/n$, that is,

$$\mathbb{E}\big[e^{s|\xi_j|}\big] \le e^{s^2\sigma^2/2} = e^{s^2(RB)^2 a/(2n)}. \tag{6.18}$$

PROOF OF THEOREM 3.5. It follows from Lemma 2.3 in Massart (2007) with the above choice of variance proxy that

$$\mathbb{E}\Big[\max_{1\le j\le M}|\xi_j|\Big] \le RB\sqrt{\frac{a\log M}{n}}.$$

Combined with (6.7) and (6.12) the previous inequality completes the proof of Theorem 3.5. □

PROOF OF THEOREM 3.6. Using, respectively, a union bound, a Chernoff bound and (6.18), we find

$$\mathbb{P}\Big(\max_{1\le j\le M}|\xi_j| > t\Big) \le M\exp\left(\frac{nt^2}{2(RB)^2 a}\right).$$

Together with (6.7) and (6.12), this bound completes the proof of Theorem 3.6 by taking $t = RB\sqrt{\frac{2a\log(M/\delta)}{n}}$. □

Acknowledgments. The author would like to thank Ramon van Handel, Guillaume Lecué and Vivian Viallon for helpful comments and suggestions.

SUPPLEMENTARY MATERIAL

Minimax lower bounds (DOI: 10.1214/11-AOS961SUPP; .pdf). Under some convexity and tail conditions, we prove minimax lower bounds for the three problems of Kullback–Leibler aggregation: model selection, linear and convex. The proof consists in three steps: first, we identify a subset of admissible estimators, then we reduce the problem to a usual problem of regression function estimation under the mean squared error criterion and finally, we use standard minimax lower bounds to complete the proof.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) 267–281. Akad. Kiadó, Budapest. MR0483125

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

AUDIBERT, J. Y. (2008). Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems* 20 (Y. S. J. Platt D. Koller and S. Roweis, eds.) 41–48. MIT Press, Cambridge, MA.

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester. MR0489333

BARTLETT, P. L., MENDELSON, S. and NEEMAN, J. (2012). ℓ_1-regularized linear regression: Persistence and oracle inequalities. *Probab. Theory Related Fields*. To appear.

BELOMESTNY, D. and SPOKOINY, V. (2007). Spatial aggregation of local likelihood estimates with applications to classification. *Ann. Statist.* **35** 2287–2311. MR2363972

BOUCHERON, S., BOUSQUET, O. and LUGOSI, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM Probab. Stat.* **9** 323–375. MR2182250

BREIMAN, L. (1999). Prediction games and arcing algorithms. *Neural Comput.* **11** 1493–1517.

BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **9**. IMS, Hayward, CA. MR0882001

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

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

DALALYAN, A. and SALMON, J. (2011). Sharp oracle inequalities for aggregation of affine estimators. Available at arXiv:1104.3969.

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

EKELAND, I. and TÉMAM, R. (1999). *Convex Analysis and Variational Problems*. *Classics in Applied Mathematics* **28**. SIAM, Philadelphia, PA. MR1727362

FAHRMEIR, L. and KAUFMANN, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13** 342–368. MR0773172

FREUND, Y. and SCHAPIRE, R. E. (1996). Experiments with a new boosting algorithm. In *International Conference on Machine Learning* 148–156. MR1473055

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *Ann. Statist.* **28** 337–407. MR1790002

GREENSHTEIN, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint. *Ann. Statist.* **34** 2367–2386. MR2291503

GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

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

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

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

LECAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.* **1** 277–329. MR0054913

LECUÉ, G. (2007). Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.* **35** 1698–1721. MR2351102

LECUÉ, G. (2012). Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli*. To appear.

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

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York. MR1639875

LOUNICI, K. (2007). Generalized mirror averaging and D-convex aggregation. *Math. Methods Statist.* **16** 246–259. MR2356820

MASSART, P. (2007). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

MEASE, D. and WYNER, A. (2008). Evidence contrary to the statistical view of boosting. *J. Mach. Learn. Res.* **9** 131–156.

MITCHELL, C. and VAN DE GEER, S. (2009). General oracle inequalities for model selection. *Electron. J. Stat.* **3** 176–204. MR2485876

NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1998). *Lecture Notes in Math.* **1738** 85–277. Springer, Berlin. MR1775640

NEMIROVSKI, A., JUDITSKY, A., LAN, G. and SHAPIRO, A. (2008). Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** 1574–1609. MR2486041

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274

RIGOLLET, P. (2012). Supplement to "Kullback–Leibler aggregation and misspecified generalized linear models." DOI:10.1214/11-AOS961SUPP.

RIGOLLET, P. and TSYBAKOV, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16** 260–280. MR2356821

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

RIGOLLET, P. and TSYBAKOV, A. (2012). Sparse estimation by exponential weighting. *Statist. Sci.* To appear.

TSYBAKOV, A. B. (2003). Optimal rates of aggregation. In *COLT* (B. Schölkopf and M. K. Warmuth, eds.). *Lecture Notes in Computer Science* **2777** 303–313. Springer, Berlin.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163

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

YANG, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli* **10** 25–47. MR2044592

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