ROBUST ESTIMATION OF A REGRESSION FUNCTION IN EXPONENTIAL FAMILIES

YANNICK BARAUD AND JUNTONG CHEN

ABSTRACT. We observe n pairs of independent random variables $X_1 = (W_1, Y_1), \ldots, X_n = (W_n, Y_n)$ and assume, although this might not be true, that for each $i \in \{1, \ldots, n\}$, the conditional distribution of Y_i given W_i belongs to a given exponential family with real parameter $\theta_i^\star = \boldsymbol{\theta}^\star(W_i)$ the value of which is an unknown function $\boldsymbol{\theta}^\star$ of the covariate W_i. Given a model $\overline{\boldsymbol{\Theta}}$ for $\boldsymbol{\theta}^\star$, we propose an estimator $\widehat{\boldsymbol{\theta}}$ with values in $\overline{\boldsymbol{\Theta}}$ the construction of which is independent of the distribution of the W_i. We show that $\boldsymbol{\theta}^\star$ possesses the properties of being robust to contamination, outliers and model misspecification. We establish non-asymptotic exponential inequalities for the upper deviations of a Hellinger-type distance between the true distribution of the data and the estimated one based on $\widehat{\boldsymbol{\theta}}$. We deduce a uniform risk bound for $\widehat{\boldsymbol{\theta}}$ over the class of Hölderian functions and we prove the optimality of this bound up to a logarithmic factor. Finally, we provide an algorithm for calculating $\widehat{\boldsymbol{\theta}}$ when $\boldsymbol{\theta}^\star$ is assumed to belong to functional classes of low or medium dimensions (in a suitable sense) and, on a simulation study, we compare the performance of $\widehat{\boldsymbol{\theta}}$ to that of the MLE and median-based estimators. The proof of our main result relies on an upper bound, with explicit numerical constants, on the expectation of the supremum of an empirical process over a VC-subgraph class. This bound can be of independent interest.

1. INTRODUCTION

In order to motivate the statistical problem we wish to solve here, let us start with a preliminary example.

Example 1 (Logit regression)**.** We study a cohort of n patients with respective clinical characteristics $W_1, \ldots, W_n$ with values in $\mathbb{R}^d$. For the sake of simplicity we shall assume that d is small compared to n even though this situation might not be the practical one. We associate the label $Y_i = 1$ to

Date: November 5, 2021.

2010 *Mathematics Subject Classification.* Primary 62J12, 62F35, 62G35, 62G05; Secondary 60G99.

Key words and phrases. Generalized linear model, logit (logistic) regression, Poisson regression, robust estimation, supremum of an empirical process.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 811017.

the patient i if he/she develops the disease D and $Y_i = -1$ otherwise. A classical model for studying the effect of the clinical characteristic W on the probability of developing the disease D is the logit one

$$\mathbb{P}\left[Y = y|W\right] = \frac{1}{1+\exp\left[-y\left\langle w^{\star}, W\right\rangle\right]} \in (0,1) \quad \text{for } y \in \{-1,+1\} \tag{1}$$

where $w^{\star}$ is an unknown vector and $\langle\cdot,\cdot\rangle$ the inner product of $\mathbb{R}^d$. The problem is to estimate $w^{\star}$ on the basis of the observations (W_i, Y_i) for $i \in \{1,\ldots,n\}$.

A common way of solving this problem is to use the Maximum Likelihood Estimator (MLE for short). In exponential families, the MLE is known to enjoy many nice properties but it also suffers from several defects. First of all, it is not difficult to see that it might not exist. This is in particular the case when a hyperplane separates the two subsets of $\mathbb{R}^d$ given by $\mathcal{W}_+ = \{W_i,\, Y_i = +1\}$ and $\mathcal{W}_- = \{W_i,\, Y_i = -1\}$, i.e. when there exists a unit vector $w_0 \in \mathbb{R}^d$ such that $\langle w, w_0\rangle > 0$ for all $w \in \mathcal{W}_+$ and $\langle w, w_0\rangle < 0$ for $w \in \mathcal{W}_-$. In this case, the conditional likelihood function at λw_0 with $\lambda > 0$ writes as

$$\prod_{i=1}^{n}\frac{1}{1+\exp\left[-\lambda Y_i\left\langle w_0, W_i\right\rangle\right]} = \prod_{i=1}^{n}\frac{1}{1+\exp\left[-\lambda\left|\left\langle w_0, W_i\right\rangle\right|\right]} \underset{\lambda\to+\infty}{\longrightarrow} 1$$

hence the maximal value 1 is not reached. For a thorough study of the existence of the MLE in the logit model we refer to Candès and Sur (2020) as well as the references therein.

Another issue with the use of the MLE lies in the fact that it is not robust and we illustrate its instability in our simulation study. Robustness is nevertheless an important property in practice since, to return to our Example 1, it may happen that our database contains a few corrupted data that correspond to mislabelled patients (some patients might have developed a disease which is not D but has similar symptoms). A natural question arises: how can we provide a suitable estimation of $w^{\star}$ despite the presence of possible corrupted data?

This is the kind of issue we want to solve here. Our approach is not, however, restricted to the logit model but applies more generally whenever the conditional distribution of Y given W belongs to a one-parameter exponential family. More precisely, the statistical framework is the following. We observe n pairs of independent random variables $(W_1, Y_1), \ldots, (W_n, Y_n)$ with values in $\mathscr{W} \times \mathscr{Y}$. For each $i \in \{1,\ldots,n\}$, the conditional distribution of Y_i given W_i is presumed to belong to a one-parameter exponential family with parameter $\theta_i^{\star} = \boldsymbol{\theta}^{\star}(W_i) \in \mathbb{R}$, which is an unknown (regression) function $\boldsymbol{\theta}^{\star}$ of the covariate W_i. Given a model $\overline{\boldsymbol{\Theta}}$ for $\boldsymbol{\theta}^{\star}$, i.e. a class of functions on $\mathscr{W}$, our aim is to estimate $\boldsymbol{\theta}^{\star}$ from the observation of $(W_1, Y_1), \ldots, (W_n, Y_n)$.

The estimator $\widehat{\boldsymbol{\theta}}$ we propose is a ρ-estimator and it enjoys the following properties:

- the estimator $\widehat{\boldsymbol{\theta}}$ is rate optimal (up to a logarithmic factor) when $\boldsymbol{\theta}^{\star}$ belongs to a VC-class of functions $\overline{\boldsymbol{\Theta}}$ or a Hölderian class of smoothness;
- the estimator $\widehat{\boldsymbol{\theta}}$ still performs well when $\boldsymbol{\theta}^{\star}$ does not belong to $\overline{\boldsymbol{\Theta}}$ but lies close enough to it;
- the performance of the estimator $\widehat{\boldsymbol{\theta}}$ remains stable when the data set $(W_1, Y_1), \ldots, (W_n, Y_n)$ is contaminated or contains outliers;
- even in the unfavourable situation where the conditional distributions of the Y_i given W_i are not exactly described by our model but lie, for most of them, close enough to an unknown element $\overline{P}$ that does belong to it, the conditional distribution based on the exponential family and $\widehat{\boldsymbol{\theta}}$ would still provide a good estimator of $\overline{P}$.

These results are based on a general non-asymptotic risk bound for $\widehat{\boldsymbol{\theta}}$ that involves explicit constants and holds under the only assumption that the model $\overline{\boldsymbol{\Theta}}$ is VC-subgraph and the data $(W_1, Y_1), \ldots, (W_n, Y_n)$ are independent.

The work presented here is different from the study of ρ-estimators conducted in Baraud and Birgé (2018)[Section 9] for estimating a regression function (seen as the parameter of interest in the conditional distribution of Y given W). In Baraud and Birgé (2018), the authors studied a regression model in which the errors are assumed to be i.i.d., homoscedastic with a density with respect to the Lebesgue measure. In the present paper, the errors are typically heteroscedastic, independent but not i.i.d. and they may not admit a density with respect the Lebesgue measure. This is the case in the logistic and Poisson regression settings for example. Actually, new results had to be established in order to analyze further the behaviour of ρ-estimators in the statistical setting we consider here. The proof of our main result combines the theory of ρ-estimation – see Baraud *et al.* (2017) and Baraud and Birgé (2018) – and an original result that establishes the fact that the family of functions on $\mathscr{W} \times \mathscr{Y}$ of the form

$$(w, y) \mapsto S(y)\boldsymbol{\eta}(w) - A(\boldsymbol{\eta}(w)) \quad \text{with} \quad \boldsymbol{\eta} \in \boldsymbol{\Gamma}$$

is VC-subgraph when S is an arbitrary function on $\mathscr{Y}$, A a convex function defined on an interval I of positive length and $\boldsymbol{\Gamma}$ a VC-subgraph class of functions defined on $\mathscr{W}$ with values in I. The proof of our main result also relies on an upper bound with explicit constants (see Theorem 2) on the expectation of the supremum of an empirical process over a VC-class of functions. Since we are not aware of such a result (with explicit constants) in the literature, this bound can be of independent interest.

Besides our theoretical guarantees on the performance of the estimator $\widehat{\boldsymbol{\theta}}$, we carry out a simulation study in order to compare it with the MLE and median-based estimators. The simulation study addresses both the situations where the data are generated from the model and when it is contaminated or contains an outlier. To our knowledge, it is the first time that ρ-estimators are implemented numerically and their performance is studied on simulated data.

There exist only few papers in the literature that tackle the estimation problem in such exponential families and establish risk bounds for the proposed estimators of $\boldsymbol{\theta}^\star$. When $\mathscr{W} = [0,1]$, Kolaczyk and Nowak (2005) proposed an estimation of $\boldsymbol{\theta}^\star$ by piecewise polynomials. When the exponential family is given in its canonical form and the natural parameter is a smooth function of the mean, they propose estimators that achieve, up to extra logarithmic factors, the classical rate $n^{-\alpha/(2\alpha+1)}$ over Besov balls with regularity $\alpha > 0$ for a Hellinger-type loss. Brown *et al.* (2010) considered one-parameter exponential families which possess the property that the variances of the distributions are quadratic functions of their means. These families include as special cases the binomial, gamma and Poisson distributions, among others, and have been studied earlier by Antoniadis and Sapatinas (2001). When the exponential family is parametrized by its mean, Brown *et al.* (2010) used a variance stabilizing transformation in order to turn the original problem of estimating the function $\boldsymbol{\theta}^\star$ into that of estimating a regression function in the homoscedastic Gaussian regression framework. They established uniform rates of convergence with respect to some $\mathbb{L}_2$-loss over classes of functions $\boldsymbol{\theta}^\star$ that belong to Besov balls and are bounded from above and below by positive numbers. Finally, in the case of the Poisson family parametrized by its mean, Kroll (2019) proposed an estimator of $\boldsymbol{\theta}^\star$ which is based on a model selection procedure. For some $\mathbb{L}_2$-loss, he proved that his estimator achieved the minimax rate of convergence over Sobolev-type ellipsoids.

A common feature of these papers lies in the fact that they make strong assumptions on the regression function and the distribution of the covariates W_i while our approach does not require any assumption about these quantities. They assume that the distribution of the W_i is known or partly known and the regression function is smooth enough. Besides, none of these papers addresses the problem of model misspecification nor proposes an estimator to solve it.

Our paper is organized as follows. We describe our statistical framework in Section 2 and present there several examples to which our approach applies. The construction of the estimator and our main result about its risk are presented in Section 3. We also explain why the deviation inequality we derive guarantees the desired robustness property of the estimator. Uniform risk bounds over Hölderian classes are established in Section 4. We also show that the minimax rates may differ from the usual ones established in the Gaussian case. Section 5 is devoted to the description of our algorithm and the simulation study. Our bound on the expectation of the supremum of an empirical process over a VC-subgraph class can be found in Section 6 as well as its proof. Section 7 is devoted to the other proofs.

2. The statistical setting

We observe n pairs of independent, but not necessarily i.i.d., random variables $X_1 = (W_1, Y_1), \dots, X_n = (W_n, Y_n)$ with values in a measurable product space $(\mathscr{X}, \mathcal{X}) = (\mathscr{W} \times \mathscr{Y}, \mathcal{W} \otimes \mathcal{Y})$. We assume that for each $i \in \{1, \dots, n\}$, the conditional probability of Y_i given W_i exists and is given by the value at W_i of a measurable function $Q_i^\star$ from $(\mathscr{W}, \mathcal{W})$ into the set $\mathscr{R}$ of all probabilities on $(\mathscr{Y}, \mathcal{Y})$. We equip $\mathscr{R}$ with the Borel σ-algebra $\mathcal{T}$ associated to the total variation distance (which induces the same topology as the Hellinger one). By doing so, the mapping $w \mapsto h^2(Q_i^\star(w), Q)$ on $(\mathscr{W}, \mathcal{W})$ is measurable whatever the probability $Q \in \mathscr{R}$ and $i \in \{1, \dots, n\}$.

With a slight abuse of language, we also refer to $Q_i^\star$ as the *conditional distribution of* Y_i *given* W_i although this distribution is actually $Q_i^\star(W_i)$. Apart from independence of the W_i, $1 \leqslant i \leqslant n$, we assume nothing about their respective distributions P_{W_i} which can therefore be arbitrary.

Let $\overline{\mathscr{Q}} \subset \mathscr{R}$ be an exponential family on the measured space $(\mathscr{Y}, \mathcal{Y}, \nu)$ where ν is an arbitrary σ-finite (positive) measure. We assume that $\overline{\mathscr{Q}} = \{Q_\theta,\ \theta \in I\}$ is indexed by a natural parameter θ that belongs to some non-trivial interval $I \subset \mathbb{R}$ (i.e. $\mathring{I} \neq \varnothing$). This means that for all $\theta \in I$, the distribution Q_θ admits a density (with respect to ν) of the form

$$q_\theta : y \mapsto e^{S(y)\theta - A(\theta)} \quad \text{with} \quad A(\theta) = \log\left[\int_{\mathscr{Y}} e^{\theta S(y)} d\nu(y)\right], \tag{2}$$

where S is a real-valued measurable function on $(\mathscr{Y}, \mathcal{Y})$ which does not coincide with a constant ν-a.e. We also recall that the function A is infinitely differentiable on the interior $\mathring{I}$ of I and strictly convex on I. It is of course possible to parametrize $\overline{\mathscr{Q}}$ in a different way (i.e. with a non-natural parameter) by performing a variable change $\gamma = v(\theta)$ where v is a continuous and strictly monotone function on I. We shall see in Section 3.3 that our main result remains unchanged under such a transformation and we therefore choose, for the sake of simplicity, to introduce it under a natural parametrization first.

Given a class of functions $\overline{\boldsymbol{\Theta}}$ from $\mathscr{W}$ into I, we presume that there exists $\boldsymbol{\theta}^\star$ in $\overline{\boldsymbol{\Theta}}$ such that the conditional distribution $Q_i^\star(W_i)$ is of the form $Q_{\boldsymbol{\theta}^\star(W_i)}$ for all $i \in \{1, \dots, n\}$. We refer to $\boldsymbol{\theta}^\star$ as the *regression function*. Even though our estimator is based on these assumptions, we keep in mind that our statistical model might be slightly misspecified: the conditional distributions

$Q_i^\star(W_i)$ might not be exactly of the form $Q_{\boldsymbol{\theta}^\star(W_i)}$ but only close to a distribution of this form and, even if they were, the set $\overline{\boldsymbol{\Theta}}$ might not contain $\boldsymbol{\theta}^\star$ but only provide a suitable approximation of it.

For $i \in \{1,\ldots,n\}$, let $\mathscr{Q}_{\mathscr{W}}$ be the set of all measurable mappings (conditional probabilities) from $(\mathscr{W}, \mathcal{W})$ into $(\mathscr{R}, \mathcal{T})$. We set $\boldsymbol{\mathscr{Q}}_{\mathscr{W}} = \mathscr{Q}_{\mathscr{W}}^n$ so that the n-tuple $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$ belongs to $\boldsymbol{\mathscr{Q}}_{\mathscr{W}}$. We endow the space $\boldsymbol{\mathscr{Q}}_{\mathscr{W}}$ with the Hellinger-type (pseudo) distance $\mathbf{h}$ defined as follows. For $\mathbf{Q} = (Q_1, \ldots, Q_n)$ and $\mathbf{Q}' = (Q_1', \ldots, Q_n')$ in $\boldsymbol{\mathscr{Q}}_{\mathscr{W}}$,

$$\mathbf{h}^2(\mathbf{Q}, \mathbf{Q}') = \mathbb{E}\left[\sum_{i=1}^n h^2\left(Q_i(W_i), Q_i'(W_i)\right)\right] = \sum_{i=1}^n \int_{\mathscr{W}} h^2\left(Q_i(w), Q_i'(w)\right) dP_{W_i}(w) \tag{3}$$

where h denotes the Hellinger distance. In particular, $\mathbf{h}(\mathbf{Q}, \mathbf{Q}') = 0$ implies that for all $i \in \{1,\ldots,n\}$, $Q_i = Q_i'$ P_{W_i}-a.s. We recall that the Hellinger distance between two probabilities $P = p \cdot \mu$ and $R = r \cdot \mu$ dominated by a measure μ on a measurable space $(E, \mathcal{E})$ is given by

$$h(P, R) = \left[\frac{1}{2}\int_E \left(\sqrt{p} - \sqrt{r}\right)^2 d\mu\right]^{1/2},$$

the result being independent of the choice of the dominating measure μ.

On the basis of the observations $X_1, \ldots, X_n$, we build an estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^\star$ with values in $\overline{\boldsymbol{\Theta}}$ and evaluate its performance by the quantity $\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\widehat{\boldsymbol{\theta}}})$ with the notations

$$\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star) \quad \text{and} \quad \mathbf{Q}_{\boldsymbol{\theta}} = (Q_{\boldsymbol{\theta}}, \ldots, Q_{\boldsymbol{\theta}})$$

when $\boldsymbol{\theta}$ is a measurable function from $\mathscr{W}$ into I. More precisely,

$$\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\widehat{\boldsymbol{\theta}}}) = \sum_{i=1}^n \int_{\mathscr{W}} h^2\left(Q_i^\star(w), Q_{\widehat{\boldsymbol{\theta}}(w)}\right) dP_{W_i}(w).$$

We write $P = Q \cdot P_W$ when P is the distribution of a random variable $(W, Y) \in \mathscr{W} \times \mathscr{Y}$ where the first marginal distribution is P_W and the conditional distribution of Y given W is Q. For $P = Q \cdot P_W$ and $P' = Q' \cdot P_W$ the squared Hellinger distance between P and P' writes as

$$h^2(P, P') = \int_{\mathscr{W}} h^2\left(Q(w), Q'(w)\right) dP_W(w).$$

Setting for $i \in \{1,\ldots,n\}$ and $\boldsymbol{\theta} : \mathscr{W} \to I$, $P_i^\star = Q_i^\star \cdot P_{W_i}$ and $P_{i,\boldsymbol{\theta}} = Q_{\boldsymbol{\theta}} \cdot P_{W_i}$ we deduce that

$$\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}}) = \sum_{i=1}^n h^2\left(P_i^\star, P_{i,\boldsymbol{\theta}}\right)$$

and $\mathbf{h}^2(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}})$ therefore equals $\mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\theta}}) = \sum_{i=1}^n h^2\left(P_i^\star, P_{i,\boldsymbol{\theta}}\right)$ where $\mathbf{P}^\star = \otimes_{i=1}^n P_i^\star$ is the true distribution of the observed data $\boldsymbol{X} = (X_1, \ldots, X_n) =$

while $\mathbf{P}_{\boldsymbol{\theta}} = \bigotimes_{i=1}^n P_{i,\boldsymbol{\theta}} = \bigotimes_{i=1}^n (Q_{\boldsymbol{\theta}} \cdot P_{W_i})$ is the joint distribution of independent random variables $(W_1, Y_1), \ldots, (W_n, Y_n)$ for which the conditional distribution of Y_i given W_i is given by $Q_{\boldsymbol{\theta}(W_i)} \in \overline{\mathscr{Q}}$ for all i. This shows that the quantity $\mathbf{h}(\mathbf{Q}^\star, \mathbf{Q}_{\boldsymbol{\theta}}) = \mathbf{h}(\mathbf{P}^\star, \mathbf{P}_{\boldsymbol{\theta}})$ may also be interpreted as a distance between the probability distributions $\mathbf{P}^\star$ and $\mathbf{P}_{\boldsymbol{\theta}}$ and not only as a (pseudo) distance between the conditional ones $\mathbf{Q}^\star$ and $\mathbf{Q}_{\boldsymbol{\theta}}$. More generally, given two measurable functions $\boldsymbol{\theta}, \boldsymbol{\theta}' : \mathscr{W} \to I$, the quantity $\mathbf{h}(\mathbf{Q}_{\boldsymbol{\theta}}, \mathbf{Q}_{\boldsymbol{\theta}'})$ also writes as $\mathbf{h}(\mathbf{P}_{\boldsymbol{\theta}}, \mathbf{P}_{\boldsymbol{\theta}'})$. Note that, unlike $\mathbf{Q}_{\widehat{\boldsymbol{\theta}}}$, $\mathbf{P}_{\widehat{\boldsymbol{\theta}}}$ is not an estimator (of $\mathbf{P}^\star$) since it depends on the marginal distributions $P_{W_1}, \ldots, P_{W_n}$ which are unknown.

2.1. **Examples.** Let us present here some typical statistical settings to which our approach applies.

Example 2 (Gaussian regression with known variance)**.** Given n independent random variables $W_1, \ldots, W_n$ with values in $\mathscr{W}$, let

$$Y_i = \boldsymbol{\theta}^\star(W_i) + \sigma\varepsilon_i \quad \text{for all } i \in \{1, \ldots, n\}$$

where the ε_i are i.i.d. standard real-valued Gaussian random variables, σ is a known positive number and $\boldsymbol{\theta}^\star$ an unknown regression function with values in $I = \mathbb{R}$. In this case, $\overline{\mathscr{Q}}$ is the set of all Gaussian distributions with variance σ^2 and for all $\theta \in I = \mathbb{R}$, $Q_\theta = \mathcal{N}(\theta, \sigma^2)$ has a density with respect to $\nu = \mathcal{N}(0, \sigma^2)$ on $(\mathscr{Y}, \mathcal{Y}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ which is of the form (2) with $A(\theta) = \theta^2/(2\sigma^2)$ and $S(y) = y/\sigma^2$ for all $y \in \mathbb{R}$.

Example 3 (Binary regression)**.** The pairs of random variables (W_i, Y_i) with $i \in \{1, \ldots, n\}$ are independent with values in $\mathscr{W} \times \{0, 1\}$ and

$$\mathbb{P}\left[Y_i = y | W_i\right] = \frac{\exp\left[y\boldsymbol{\theta}^\star(W_i)\right]}{1 + \exp\left[\boldsymbol{\theta}^\star(W_i)\right]} \quad \text{for all } y \in \{0, 1\}. \tag{4}$$

This means that the conditional distribution of Y_i given W_i is Bernoulli with mean $\left(1 + \exp\left[-\boldsymbol{\theta}^\star(W_i)\right]\right)^{-1}$ for some regression function $\boldsymbol{\theta}^\star$ with values in $I = \mathbb{R}$. This model is equivalent to the logit one presented in Example 1 by changing $Y_i \in \{0, 1\}$ into $Y_i' = 2Y_i - 1 \in \{-1, 1\}$ for all i. The exponential family $\overline{\mathscr{Q}}$ consists of the Bernoulli distribution Q_θ with mean $1/[1 + e^{-\theta}] \in (0, 1)$ and $\theta \in I = \mathbb{R}$. For all $\theta \in \mathbb{R}$, Q_θ admits thus a density with respect to the counting measure ν on $\mathscr{Y} = \{0, 1\}$ of the form (2) with $A(\theta) = \log(1+e^\theta)$ and $S(y) = y$ for all $y \in \mathscr{Y}$.

Example 4 (Poisson regression)**.** The exponential family $\overline{\mathscr{Q}}$ is the set of all Poisson distributions Q_θ with mean e^θ, $\theta \in I = \mathbb{R}$. Taking for ν the Poisson distribution with mean 1, the density of Q_θ with respect to ν takes the form (2) with $S(y) = y$ for all $y \in \mathbb{N}$ and $A(\theta) = e^\theta - 1$ for all $\theta \in \mathbb{R}$. The conditional distribution of Y_i given W_i is presumed to be Poisson with mean $\exp\left[\boldsymbol{\theta}^\star(W_i)\right]$ for some regression function $\boldsymbol{\theta}^\star$ with values in $I = \mathbb{R}$.

Example 5 (Exponential multiplicative regression)**.** The random variables $W_1, \ldots, W_n$ are independent and

$$Y_i = \frac{Z_i}{\boldsymbol{\theta}^\star(W_i)} \quad \text{for all } i \in \{1, \ldots, n\} \tag{5}$$

where the Z_i are i.i.d. with exponential distribution of parameter 1 and independent of the W_i. The conditional distribution of Y_i given W_i is then exponential with mean $1/\boldsymbol{\theta}^\star(W_i) \in I = (0, +\infty)$. Exponential distributions parametrized by $\theta \in I$ admit densities with respect to the Lebesgue measure on $\mathbb{R}_+$ of the form (2) with $S(y) = -y$ for all $y \in \mathscr{Y} = \mathbb{R}_+$ and $A(\theta) = -\log\theta$.

3. The main results

3.1. **The estimation procedure.** As mentioned in the introduction, our approach is based on ρ-estimation. We do not recall here the basic ideas that underline the construction of these estimators and rather refer the reader to Baraud and Birgé (2018). Let ψ be the function defined on $[0, +\infty]$ by

$$\psi(x) = \frac{x-1}{x+1} \qquad \text{for } x \in [0, +\infty) \qquad \text{and} \qquad \psi(+\infty) = 1. \tag{6}$$

In order to avoid measurability issues, we restrict ourselves to a finite or countable subset $\boldsymbol{\Theta}$ of $\overline{\boldsymbol{\Theta}}$ and define

$$\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_{i=1}^{n} \psi\left(\sqrt{\frac{q_{\boldsymbol{\theta}'}(X_i)}{q_{\boldsymbol{\theta}}(X_i)}}\right) \qquad \text{for} \qquad \boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}, \tag{7}$$

with the conventions $0/0 = 1$ and $a/0 = +\infty$ for all $a > 0$. Then, we set

$$\boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\theta}') \qquad \text{for all} \qquad \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{8}$$

and choose $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{X})$ as any (measurable) element of the random (and non-void) set

$$\mathscr{E}(\boldsymbol{X}) = \left\{\boldsymbol{\theta} \in \boldsymbol{\Theta} \;\text{ such that }\; \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}) \leqslant \inf_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}') + 1\right\}. \tag{9}$$

The random variable $\widehat{\boldsymbol{\theta}}(\boldsymbol{X})$ is our estimator of the regression function $\boldsymbol{\theta}^\star$ and $\mathbf{Q}_{\widehat{\boldsymbol{\theta}}} = (Q_{\widehat{\boldsymbol{\theta}}}, \ldots, Q_{\widehat{\boldsymbol{\theta}}})$.

The construction of the estimator is only based on the choices of the exponential family given by (2) and the subset $\boldsymbol{\Theta}$ of $\overline{\boldsymbol{\Theta}}$. In particular, the estimator does not depend on the distributions P_{W_i} of the W_i which may therefore be unknown.

In the right-hand side of (9), the additive constant 1 plays no magic role and can be replaced by any smaller positive number. Whenever possible, the choice of an estimator $\widehat{\boldsymbol{\theta}}$ satisfying $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) = \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta})$ should be preferred. Since $\mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\inf_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}') \geqslant 0$ and any element $\widetilde{\boldsymbol{\theta}}$ that satisfies $\boldsymbol{\upsilon}(\boldsymbol{X},\widetilde{\boldsymbol{\theta}}) \leqslant 1$ belongs to the set $\mathscr{E}(\boldsymbol{X})$. It can therefore be chosen as a ρ-estimator.

The fact that we build our estimator on a finite or countable subset $\boldsymbol{\Theta}$ of $\overline{\boldsymbol{\Theta}}$ is not restrictive as we shall see. Besides, this assumption is consistent with the practice of calculating an estimator on a computer that can handle a finite number of values only.

3.2. **The main assumption and the performance of the estimator $\widehat{\boldsymbol{\theta}}$.** Let us make the following assumption:

Assumption 1. *The class $\overline{\boldsymbol{\Theta}}$ is VC-subgraph on $\mathscr{W}$ with dimension not larger than $V \geqslant 1$.*

We refer the reader to van der Vaart and Wellner (1996)[Section 2.6.2] for the definition of VC-subgraph classes and their properties. In the present paper, we mainly use the facts that Assumption 1 is satisfied when $\overline{\boldsymbol{\Theta}}$ is a linear space $\mathcal{V}$ with finite dimension $d \geqslant 1$, in which case $V = d+1$ and that it is also satisfied when $\overline{\boldsymbol{\Theta}}$ is of the form $\{F(\boldsymbol{\beta}),\ \boldsymbol{\beta} \in \mathcal{V}\}$ where F is a monotone function on the real-line. In this latter case, the VC-dimension of $\overline{\boldsymbol{\Theta}}$ is not larger than that of $\mathcal{V}$. We set

$$c_1 = 150, \quad c_2 = 1.1 \times 10^6, \quad c_3 = 5014 \tag{10}$$

and, for $\mathbf{Q} \in \mathscr{Q}_{\mathscr{W}}$ and $\mathbf{A} \subset \mathscr{Q}_{\mathscr{W}}$,

$$\mathbf{h}(\mathbf{Q}, \mathbf{A}) = \inf_{\mathbf{Q}' \in \mathbf{A}} \mathbf{h}\left(\mathbf{Q}, \mathbf{Q}'\right).$$

The risk of our estimator satisfies the following properties.

Theorem 1. *Let $\xi > 0$. Under Assumption 1, whatever the conditional probabilities $\mathbf{Q}^\star = (Q_1^\star, \ldots, Q_n^\star)$ of the Y_i given W_i and the distributions of the W_i, the estimator $\widehat{\boldsymbol{\theta}}$ defined in Section 3.1 satisfies, with a probability at least $1 - e^{-\xi}$,*

$$\mathbf{h}^2\left(\mathbf{Q}^\star, \mathbf{Q}_{\widehat{\boldsymbol{\theta}}}\right) \leqslant c_1 \mathbf{h}^2(\mathbf{Q}^\star, \mathscr{Q}) + c_2 V \left[9.11 + \log_+\left(\frac{n}{V}\right)\right] + c_3 (1.5 + \xi) \tag{11}$$

where $\mathscr{Q} = \{\mathbf{Q}_{\boldsymbol{\theta}} = (Q_{\boldsymbol{\theta}}, \ldots, Q_{\boldsymbol{\theta}}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and $\log_+ = \max(0, \log)$.

The constants c_1, c_2 and c_3 are numerical constants. They are independent of the choice of the exponential family. When the model $\mathscr{Q}$ is exact, the bound we get only depends on the VC-dimension of $\overline{\boldsymbol{\Theta}}$.

It is clear that (11) also holds true for $\overline{\mathscr{Q}} = \{\mathbf{Q}_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\}$ in place of $\mathscr{Q}$ when $\mathscr{Q}$ is dense in $\overline{\mathscr{Q}}$ with respect to the Hellinger-type distance $\mathbf{h}$. This is the case when $\boldsymbol{\Theta}$ is dense in $\overline{\boldsymbol{\Theta}}$ for the topology of pointwise convergence. We do not comment on our result any further in this direction and rather refer to Baraud and Birgé (2018) Section 4.2. From now on, we assume for the sake of simplicity that $\mathscr{Q}$ is dense in $\overline{\mathscr{Q}}$, doing as if $\overline{\boldsymbol{\Theta}} = \boldsymbol{\Theta}$. In the remaining part of this section, C will denote a positive numerical constant that may vary from line to line.

Let us now (11) in a slightly different form. We have seen in Section 2 that the quantity $\mathbf{h}\left(\mathbf{Q}^{\star}, \mathbf{Q}_{\boldsymbol{\theta}}\right)$ with $\boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}$, which involves the conditional probabilities of $\mathbf{P}^{\star}$ and $\mathbf{P}_{\boldsymbol{\theta}}$ with respect to the W_i, can also be interpreted in terms of the Hellinger(-type) distance between these two product probabilities. Inequality (11) therefore implies that

$$\mathbb{P}\left[C\mathbf{h}^2\left(\mathbf{P}^{\star}, \mathbf{P}_{\widehat{\boldsymbol{\theta}}}\right) > \mathbf{h}^2(\mathbf{P}^{\star}, \overline{\mathscr{P}}) + V\left[1 + \log_+\left(\frac{n}{V}\right)\right] + \xi\right] \leqslant e^{-\xi}, \tag{12}$$

where $\overline{\mathscr{P}} = \{\mathbf{P}_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\}$. Integrating this inequality with respect to $\xi > 0$ leads to the following risk bound for our estimator $\widehat{\boldsymbol{\theta}}$

$$C\mathbb{E}\left[\mathbf{h}^2\left(\mathbf{P}^{\star}, \mathbf{P}_{\widehat{\boldsymbol{\theta}}}\right)\right] \leqslant \mathbf{h}^2(\mathbf{P}^{\star}, \overline{\mathscr{P}}) + V\left[1 + \log_+\left(\frac{n}{V}\right)\right]. \tag{13}$$

In order to comment upon (13), let us start with the ideal situation where $\mathbf{P}^{\star}$ belongs to $\overline{\mathscr{P}}$, i.e. $\mathbf{P}^{\star} = \mathbf{P}_{\boldsymbol{\theta}^{\star}}$ with $\boldsymbol{\theta}^{\star} \in \overline{\boldsymbol{\Theta}}$, in which case (13) leads to

$$C\mathbb{E}\left[\mathbf{h}^2\left(\mathbf{P}_{\boldsymbol{\theta}^{\star}}, \mathbf{P}_{\widehat{\boldsymbol{\theta}}}\right)\right] \leqslant V\left[1 + \log_+\left(\frac{n}{V}\right)\right]. \tag{14}$$

Up to the logarithmic factor, the right-hand side of this inequality is of the expected order of magnitude V for the quantity $\mathbf{h}^2(\mathbf{P}_{\boldsymbol{\theta}^{\star}}, \mathbf{P}_{\widehat{\boldsymbol{\theta}}})$: in typical situations V is of the same order as the number of parameters that are used to parametrize $\overline{\boldsymbol{\Theta}}$.

When the true distribution $\mathbf{P}^{\star}$ writes as $\mathbf{P}_{\boldsymbol{\theta}^{\star}}$ but the regression function $\boldsymbol{\theta}^{\star}$ does not belong to $\overline{\boldsymbol{\Theta}}$, or when the conditional distributions of the Y_i given W_i do not belong to our exponential family, inequality (13) shows that, as compared to (14), the bound we get involves the approximation term $\mathbf{h}^2(\mathbf{P}^{\star}, \overline{\mathscr{P}})$ that accounts for the fact that our statistical model is misspecified. However, as long as this quantity remains small enough as compared to $V\left[1 + \log_+(n/V)\right]$, our risk bound will be of the same order as that given by (14) when the model is exact. This property accounts for the stability of our estimation procedure under misspecification. In order to be more specific, let us assume that

$$\mathbf{P}^{\star} = \bigotimes_{i=1}^{n}\left[(1-\alpha_i)P_{i,\overline{\boldsymbol{\theta}}} + \alpha_i R_i\right], \quad \sum_{i=1}^{n}\alpha_i \leqslant \frac{n}{2} \tag{15}$$

where R_i is an arbitrary distribution on $\mathscr{X}$ and α_i a number in $[0,1]$ for all $i \in \{1, \ldots, n\}$. Such a distribution $\mathbf{P}^{\star}$ allows us to model different form of robustness including robustness to the presence of contaminating data as well as outliers. In the case of contamination, $P_{W_i} = P_W$, $\alpha_i = \alpha \in (0, 1/2]$ and $R_i = R \neq P_{\overline{\boldsymbol{\theta}}} = Q_{\overline{\boldsymbol{\theta}}} \cdot P_W$ for all $i \in \{1, \ldots, n\}$ and one observes a n-sample a portion $(1-\alpha)$ of which is drawn according to a distribution $P_{\overline{\boldsymbol{\theta}}}$ that belongs to our model $\overline{\mathscr{P}}$ while the remaining part of the data is drawn according to a contaminating distribution R. In the second case, the data set contains the outliers $\{a_i,\ i \in J\}$ for some subset $J \subset \{1,\dots,n\}$ with $J \neq \varnothing$ so that $\mathbf{P}^\star$ is of the form (15) with $\alpha_i = 1\!\!1_{i\in J}$ for all $i \in \{1,\dots,n\}$ and $R_i = \delta_{a_i}$ for all $i \in J$. In all cases, using the classical inequality $h^2 \leqslant D$ where D denotes the total variation distance between probabilities, we get

$$\mathbf{h}^2(\mathbf{P}^\star, \overline{\mathscr{P}}) \leqslant \mathbf{h}^2(\mathbf{P}^\star, \mathbf{P}_{\overline{\boldsymbol{\theta}}}) \leqslant \sum_{i=1}^n D(P_i^\star, P_{\overline{\boldsymbol{\theta}}}) \leqslant \sum_{i=1}^n \alpha_i \tag{16}$$

which means that whenever $\sum_{i=1}^n \alpha_i$ remains small as compared to $V(1+\log_+(n/V))$, the performance of the estimator remains almost the same as if $\mathbf{P}^\star$ were equal to $\mathbf{P}_{\overline{\boldsymbol{\theta}}}$. The estimator $\widehat{\boldsymbol{\theta}}$ therefore possesses some stability properties with respect to contamination and the presence of outliers.

3.3. **From a natural to a general exponential family.** In Section 2, we focused on an exponential family $\overline{\mathscr{Q}}$ parametrized by its natural parameter. However statisticians often write exponential families $\overline{\mathscr{Q}}$ under the general form $\overline{\mathscr{Q}} = \{R_\gamma = r_\gamma \cdot \nu,\ \gamma \in J\}$ with

$$r_\gamma : y \mapsto e^{u(\gamma)S(y)-B(\gamma)} \quad \text{for } \gamma \in J. \tag{17}$$

In (17), J denotes a (non-trivial) interval of $\mathbb{R}$ and u a continuous and strictly monotone function from J onto I so that $B = A \circ u$. In the exponential family $\overline{\mathscr{Q}} = \{R_\gamma,\ \gamma \in J\} = \{Q_\theta,\ \theta \in I\}$, the probabilities R_γ are associated to the probabilities Q_θ by the formula $R_\gamma = Q_{u(\gamma)}$.

With this new parametrization, we could alternatively write our statistical model $\overline{\boldsymbol{\mathscr{Q}}}$ as

$$\overline{\boldsymbol{\mathscr{Q}}} = \left\{\mathbf{R}_{\boldsymbol{\gamma}} = (R_{\gamma}, \dots, R_{\gamma}),\ \boldsymbol{\gamma} \in \overline{\boldsymbol{\Gamma}}\right\} \tag{18}$$

where $\overline{\boldsymbol{\Gamma}}$ is a class of functions $\boldsymbol{\gamma}$ from $\mathscr{W}$ into J. Starting from such a statistical model and presuming that $\mathbf{Q}^\star = \mathbf{R}_{\boldsymbol{\gamma}^\star}$ for some function $\boldsymbol{\gamma}^\star \in \overline{\boldsymbol{\Gamma}}$, we can build an estimator $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}^\star$ as follows: given a finite or countable subset $\boldsymbol{\Gamma}$ of $\overline{\boldsymbol{\Gamma}}$ we set $\widehat{\boldsymbol{\gamma}} = u^{-1}(\widehat{\boldsymbol{\theta}})$ where $\widehat{\boldsymbol{\theta}}$ is any estimator obtained by applying the procedure described in Section 3.1 under the natural parametrization of the exponential family $\overline{\mathscr{Q}}$ and using the finite or countable model $\boldsymbol{\Theta} = \{\boldsymbol{\theta} = u \circ \boldsymbol{\gamma},\ \boldsymbol{\gamma} \in \boldsymbol{\Gamma}\}$.

Since our model $\overline{\boldsymbol{\mathscr{Q}}}$ for the conditional probabilities $\mathbf{Q}^\star$ is unchanged (only its parametrization changes), it would be interesting to establish a result on the performance of the estimator $\mathbf{R}_{\widehat{\boldsymbol{\gamma}}} = \mathbf{Q}_{\widehat{\boldsymbol{\theta}}}$ which is independent of the parametrization. A nice feature of the VC-subgraph property lies in the fact that it is preserved by composition with a monotone function: since u is monotone, if $\overline{\boldsymbol{\Gamma}}$ is VC-subgraph with dimension not larger than V, so is $\overline{\boldsymbol{\Theta}}$ and our Theorem 1 applies. The following corollary is therefore straightforward.

Corollary 1. *Let $\xi > 0$. If the statistical model $\overline{\boldsymbol{\mathscr{Q}}}$ is under the general form (18) and $\overline{\boldsymbol{\Gamma}}$ is VC-subgraph with dimension not larger than $V \geqslant 1$, whatever the conditional probabilities $\mathbf{Q}^\star = (Q_1^\star, \dots, Q_n^\star)$ of the Y_i given W_i*

and the distributions of the W_i, the estimator $\widehat{\boldsymbol{\gamma}}$ satisfies with a probability at least $1-e^{-\xi}$,

$$\mathbf{h}^2\left(\mathbf{Q}^\star,\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}\right)\leqslant c_1\mathbf{h}^2(\mathbf{Q}^\star,\boldsymbol{\mathscr{Q}})+c_2V\left[9.11+\log_+\left(\frac{n}{V}\right)\right]+c_3\left(1.5+\xi\right) \tag{19}$$

where $\boldsymbol{\mathscr{Q}}=\{\mathbf{R}_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma}\in\boldsymbol{\Gamma}\}$. In particular,

$$\mathbb{E}\left[\mathbf{h}^2\left(\mathbf{Q}^\star,\mathbf{R}_{\widehat{\boldsymbol{\gamma}}}\right)\right]\leqslant C'\left[\mathbf{h}^2(\mathbf{Q}^\star,\boldsymbol{\mathscr{Q}})+V\left[1+\log_+\left(\frac{n}{V}\right)\right]\right], \tag{20}$$

for some numerical constant $C'>0$.

A nice feature of our approach lies in the fact that (19) holds for all exponential families simultaneously and all ways of parametrizing them.

4. Uniform risk bounds

Throughout this section, we assume that the W_i are i.i.d. with common distribution P_W and that $\mathbf{Q}^\star=\mathbf{R}_{\boldsymbol{\gamma}^\star}$ belongs to a statistical model of the (general) form given by (18) where $\overline{\boldsymbol{\Gamma}}$ is a class of smooth functions. More precisely, we assume that for some $\alpha\in(0,1]$ and $M>0$, $\overline{\boldsymbol{\Gamma}}=\mathcal{H}_\alpha(M)$ is the set of functions $\boldsymbol{\gamma}$ on $[0,1]$ with values in J that satisfy the Hölder condition

$$|\boldsymbol{\gamma}(x)-\boldsymbol{\gamma}(y)|\leqslant M|x-y|^\alpha\quad\text{for all } x,y\in[0,1]. \tag{21}$$

Our aim is both to estimate the regression function $\boldsymbol{\gamma}^\star$ under the assumption that it belongs to $\mathcal{H}_\alpha(M)$ and to evaluate the minimax risk over this set, i.e. the quantity

$$\begin{aligned}\mathcal{R}_n(\mathcal{H}_\alpha(M))&=\inf_{\widetilde{\boldsymbol{\gamma}}}\sup_{\boldsymbol{\gamma}^\star\in\mathcal{H}_\alpha(M)}\mathbb{E}\left[h^2\left(R_{\boldsymbol{\gamma}^\star},R_{\widetilde{\boldsymbol{\gamma}}}\right)\right]\\&=\inf_{\widetilde{\boldsymbol{\gamma}}}\sup_{\boldsymbol{\gamma}^\star\in\mathcal{H}_\alpha(M)}\mathbb{E}\left[\int_{\mathscr{W}}h^2\left(R_{\boldsymbol{\gamma}^\star(w)},R_{\widetilde{\boldsymbol{\gamma}}(w)}\right)dP_W(w)\right]\end{aligned} \tag{22}$$

where the infimum runs among all estimators $\widetilde{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}^\star$ based on the n-sample $X_1,\ldots,X_n$. In this section we establish the following facts:

- (i) the quantity $\mathcal{R}_n(\mathcal{H}_\alpha(M))$ may not be of order $n^{-2\alpha/(2\alpha+1)}$;
- (ii) there exists a sufficient condition on the parametrization of exponential families under which $\mathcal{R}_n(\mathcal{H}_\alpha(M))$ is of order $n^{-2\alpha/(2\alpha+1)}$;
- (iii) the ρ-estimator is minimax, at least in all exponential families that satisfy this condition.

4.1. Parametrizing by the mean.

A common parametrization of an exponential family $\overline{\mathscr{Q}}=\{R_\gamma,\ \gamma\in J\}$ is by the means of the distributions, i.e. $\gamma=\int_{\mathscr{Y}}ydR_\gamma(y)$. This is typically the case for the Bernoulli, Gaussian and Poisson families for example. For such a parametrization, the minimax rate over the Hölderian ball $\mathcal{H}_\alpha(M)$ defined by (21) may not be of order $n^{-2\alpha/(2\alpha+1)}$.

Proposition 1. *Let* $\alpha \in (0,1]$, $M>0$, P_W *be the uniform distribution on* $[0,1]$ *and* $\overline{\mathscr{Q}}$ *the set of Poisson distributions* R_γ *with means* $\gamma \in J=(0,+\infty)$. *For all* $n \geqslant 1$,

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) \geqslant \frac{(1-e^{-1})}{144}\left[\left(\frac{3M^{1/\alpha}}{2^{4+\alpha+3/\alpha}n}\right)^{\frac{\alpha}{1+\alpha}} \bigwedge \frac{M}{8} \bigwedge \left(1+\frac{\sqrt{3}}{2}\right)\right].$$

In the Poisson case, the rate for $\mathcal{R}_n(\mathcal{H}_\alpha(M))$ is therefore at least of order $n^{-\alpha/(1+\alpha)}$, hence much slower than the one we would get in the Gaussian case (also parametrized with its mean), namely $n^{-2\alpha/(2\alpha+1)}$. We conclude that, depending on the exponential family, the parametrization by the mean may lead to different minimax rates.

4.2. **Uniform risk bounds over Hölder classes.** As seen above, the parametrization of the exponential family influences the minimax rate of convergence. In this section, we introduce the following assumptions on the parametrization.

Assumption 2. *There exists a constant* $\kappa > 0$ *such that*

$$h(R_\gamma, R_{\gamma'}) \leqslant \kappa \left|\gamma-\gamma'\right| \quad \text{for all } \gamma,\gamma' \in J \tag{23}$$

and for a (non trivial) compact interval $K \subset J$, *there exists a constant* $c_K > 0$ *such that*

$$h(R_\gamma, R_{\gamma'}) \geqslant c_K \left|\gamma-\gamma'\right| \quad \text{for all } \gamma,\gamma' \in K. \tag{24}$$

For example, this assumption is satisfied in the following case.

Proposition 2. *Let* $\overline{\mathscr{Q}} = \{Q_\theta,\ \theta \in I\}$ *be a natural exponential family defined by (2) where* I *is an open interval. If the function* v *satisfies*

$$v'(\theta) = \sqrt{\frac{A''(\theta)}{8}} > 0 \quad \text{for all } \theta \in I, \tag{25}$$

when parametrized by the parameter $\gamma = v(\theta)$, *the exponential family* $\overline{\mathscr{Q}} = \{R_\gamma = Q_{v^{-1}(\gamma)},\ \gamma \in J\}$ *satisfies Assumption 2 with* $\kappa = 1$ *and for all choices of a (non-trivial) compact subset* K *of* J.

It is well-known that the functions $v_1 : \theta \mapsto (1/\sqrt{2})\arcsin(1/\sqrt{1+e^{-\theta}})$, $v_2 : \theta \mapsto (1/\sqrt{2})e^{\theta/2}$ on $\mathbb{R}$ and $v_3 : \theta \mapsto (\sqrt{8})^{-1}\log\theta$ on $(0,+\infty)$ satisfy (25) in the cases of Examples 3, 4 and 5 respectively.

As a consequence of Assumption 2, by integration with respect to P_W, we obtain that for all functions $\boldsymbol{\gamma},\boldsymbol{\gamma}'$ on $\mathscr{W}$ with values in J,

$$h^2\left(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}\right) \leqslant \kappa^2 \left\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\right\|_2^2 = \kappa^2 \int_{\mathscr{W}} \left(\boldsymbol{\gamma}-\boldsymbol{\gamma}'\right)^2 dP_W \tag{26}$$

and there exists a constant $c_K > 0$ such that for all functions $\boldsymbol{\gamma}, \boldsymbol{\gamma}'$ on $\mathscr{W}$ with values in K,

$$h^2\left(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}\right) \geqslant c_K^2 \left\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\right\|_2^2.$$

Assumption 2 makes the Hellinger-type distance $h\left(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}\right)$ and the $\mathbb{L}_2(P_W)$-one between $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ comparable.

Proposition 3. *Let $\alpha \in (0,1]$ and $M > 0$. If P_W is the uniform distribution on $[0,1]$ and Assumption 2 holds true for a compact interval K of length $2\overline{L} > 0$, then*

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) \geqslant \frac{c_K^2}{48}\left[\left(\frac{3M^{1/\alpha}}{2^{2\alpha+4+1/\alpha}\kappa^2 n}\right)^{\frac{2\alpha}{1+2\alpha}} \wedge \left(\frac{M^2}{4}\right) \wedge \overline{L}^2\right].$$

This result says that in all exponential families for which Assumption 2 is satisfied, the order of magnitude of the minimax rate over $\mathcal{H}_\alpha(M)$ cannot be smaller than $n^{-2\alpha/(2\alpha+1)}$, at least when P_W is the uniform distribution on $[0,1]$.

The following uniform risk bound of the performance of the ρ-estimator over $\mathcal{H}_\alpha(M)$ shows that this rate is optimal and the ρ-estimator minimax (up to a logarithmic factor). Note that the upper bound holds without any assumption on P_W and under the weaker assumptions that only (23) is satisfied.

Proposition 4. *Assume that (23) is satisfied. Let $\alpha \in (0,1]$, $M > 0$ and $\overline{\boldsymbol{S}}$ be the set of functions with values in the interval J which are piecewise constant on each element of a partition $\{I_j, j \in \{1,\ldots,D\}\}$ of $[0,1]$ into $D \geqslant 1$ intervals of lengths $1/D$. For*

$$D = D(\alpha, M, n) = \min\left\{k \in \mathbb{N},\ \left(\frac{\kappa^2 M^2 n}{1+\log n}\right)^{\frac{2\alpha}{1+2\alpha}} \leqslant k\right\},$$

the ρ-estimator $\widehat{\boldsymbol{\gamma}}$ based on (any) countable and dense subset $\boldsymbol{S}$ of $\overline{\boldsymbol{S}}$ (with respect to supremum norm) satisfies

$$\sup_{\boldsymbol{\gamma}^\star \in \mathcal{H}_\alpha(M)} \mathbb{E}\left[h^2\left(R_{\boldsymbol{\gamma}^\star}, R_{\widehat{\boldsymbol{\gamma}}}\right)\right] \leqslant 2C'\left[\left(\frac{(\kappa M)^{1/\alpha}\log(en)}{n}\right)^{\frac{2\alpha}{1+2\alpha}} + \frac{3\log(en)}{2n}\right]$$

where C' is the numerical constant appearing in (20).

5. Calculation of ρ-estimators and simulation study

In this section, we study the performance of the ρ-estimator $\widehat{\boldsymbol{\theta}}$ of the regression function $\boldsymbol{\theta}^\star$ in the cases of Examples 3, 4, 5 which correspond respectively to the logit regression, Poisson and exponential distributions parametrized by their natural parameters.

The models. The function space $\overline{\boldsymbol{\Theta}}$ consists of functions $\boldsymbol{\theta}$ on $\mathscr{W} = \mathbb{R}^5$ with values in I and for $w = (w_1, \ldots, w_5) \in \mathscr{W}$ the value $\boldsymbol{\theta}(w)$ has the following form:

— In the Bernoulli model, $I = \mathbb{R}$ and

$$\boldsymbol{\theta}(w) = \eta_0 + \sum_{j=1}^{5} \eta_j w_j \quad \text{with } \eta = (\eta_0, \ldots, \eta_5) \in \mathbb{R}^6. \tag{27}$$

— In the Poisson model, $I = \mathbb{R}$ and

$$\boldsymbol{\theta}(w) = \log\log\left[1 + \exp\left(\eta_0 + \sum_{j=1}^{5} \eta_j w_j\right)\right] \quad \text{with } \eta = (\eta_0, \ldots, \eta_5) \in \mathbb{R}^6. \tag{28}$$

— In the exponential model, $I = (0, +\infty)$ and

$$\boldsymbol{\theta}(w) = \log\left[1 + \exp\left(\eta_0 + \sum_{j=1}^{5} \eta_j w_j\right)\right] \quad \text{with } \eta = (\eta_0, \ldots, \eta_5) \in \mathbb{R}^6. \tag{29}$$

For all theses cases, the set $\overline{\boldsymbol{\Theta}}$ is VC-subgraph with dimension not larger than 7. For the calculation of the estimator on a computer, we do as if $\overline{\boldsymbol{\Theta}}$ was countable and consequently take $\boldsymbol{\Theta} = \overline{\boldsymbol{\Theta}}$.

The competitors. We compare the performance of $\widehat{\boldsymbol{\theta}}$ to that of the MLE and, in cases of Examples 4 and 5, to a median-based estimator $\widehat{\boldsymbol{\theta}}_0$. The estimator $\widehat{\boldsymbol{\theta}}_0$ is defined as any minimizer over $\overline{\boldsymbol{\Theta}}$ of the criterion

$$\boldsymbol{\theta} \mapsto \sum_{i=1}^{n} |Y_i - m(\boldsymbol{\theta}(W_i))|$$

where $m(\theta)$ is the median (or an approximation of it) of the distribution Q_θ for $\theta \in I$. We take $m(\theta) = e^\theta + 1/3 - 0.02 e^{-\theta}$ for the Poisson distribution with parameter e^θ and $m(\theta) = (\log 2)/\theta$ for the exponential one with parameter θ.

5.1. **Calculation of the ρ-estimator.** As mentioned in Section 3, we call ρ-estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{X})$ any element of the random set

$$\mathscr{E}(\boldsymbol{X}) = \left\{\boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ such that } \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}) \leqslant \inf_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}') + 1\right\},$$

where

$$\boldsymbol{\upsilon}(\boldsymbol{X}, \boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \mathbf{T}(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\theta}') = \sup_{\boldsymbol{\theta}' \in \boldsymbol{\Theta}} \sum_{i=1}^{n} \psi\left(\sqrt{\frac{q_{\boldsymbol{\theta}'}(X_i)}{q_{\boldsymbol{\theta}}(X_i)}}\right) \quad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

We recall that $\widehat{\boldsymbol{\theta}}$ is (necessarily) a ρ-estimator if $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leqslant 1$. To calculate $\widehat{\boldsymbol{\theta}}$ we use the iterative Algorithm 1 described below. We stop it either when the condition $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leqslant 1$ is met (then $\widehat{\boldsymbol{\theta}}$ is a ρ-estimator) or after $L = 100$ iterations. Since ρ-estimators are not unique, there is no reason for the algorithm to converge to a point and we are not expecting the algorithm to do so. Since the algorithm is based on the test statistic $\mathbf{T}(\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\theta}')$, that provides a robust test between the probabilities $\mathbf{P}_{\boldsymbol{\theta}}$ and $\mathbf{P}_{\boldsymbol{\theta}'}$, we expect the algorithm to get closer to the truth as we iterate it. As we shall see, only few iterations are in general necessary to meet the condition $\boldsymbol{\upsilon}(\boldsymbol{X},\widehat{\boldsymbol{\theta}}) \leqslant 1$ and when it is not the case, the estimator obtained after $L = 100$ iterations provides a suitable estimation of the parameter. To find a maximizer of the mapping $\boldsymbol{\theta} \mapsto \mathbf{T}(\boldsymbol{X},\widehat{\boldsymbol{\theta}},\boldsymbol{\theta})$ at each iteration, we use the CMA (Covariance Matrix Adaptation) method which turns out to be more stable than the gradient descent method. For more details about the CMA method, we refer the reader to Hansen (2016).

Algorithm 1 Searching for the ρ-estimator

Input:

$\boldsymbol{X} = (X_1, \cdots, X_n)$: the data

$\boldsymbol{\theta}_0$: the starting point

Output: $\widehat{\boldsymbol{\theta}}$

Initialize $l = 0$, $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$;

while $\boldsymbol{\upsilon}(\boldsymbol{X},\widehat{\boldsymbol{\theta}}) > 1$ and $l \le L$ **do**

$\quad l \leftarrow l + 1$

$\quad \boldsymbol{\theta}_1 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\, \mathbf{T}(\boldsymbol{X},\widehat{\boldsymbol{\theta}},\boldsymbol{\theta})$

$\quad \widehat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}_1$

end while

Return $\widehat{\boldsymbol{\theta}}$.

To intialize the process we choose the value of $\boldsymbol{\theta}_0$ as follows. In the case of Bernoulli regression, we take for $\boldsymbol{\theta}_0$ the function on $\mathbb{R}^d$ that minimizes on $\overline{\boldsymbol{\Theta}}$ the penalized criterion (that can be found in the e1071 R-package)

$$\boldsymbol{\theta} \mapsto 10 \sum_{i=1}^{n} \left(1 - (2Y_i - 1)\boldsymbol{\theta}(W_i)\right)_+ + \frac{1}{2}\sum_{i=1}^{d} \left|\boldsymbol{\theta}(e_i) - \boldsymbol{\theta}(0)\right|^2,$$

where $e_1, \ldots, e_d$ denotes the canonical basis of $\mathbb{R}^d$ (with $d = 6$). The e1071 R-package is used for the purpose of classifying the Y_i from the W_i. For the other exponential families we choose for $\boldsymbol{\theta}_0$ the median-based estimator $\widehat{\boldsymbol{\theta}}_0$.

5.2. **Comparisons of the estimators when the model is exact.** Throughout this section, we assume that the data $X_1, \ldots, X_n$ are i.i.d. with distribution $P_{\boldsymbol{\theta}^\star} = Q_{\boldsymbol{\theta}^\star} \cdot P_W$, $\boldsymbol{\theta}^\star \in \overline{\boldsymbol{\Theta}}$, and we evaluate the risk

$$R_n(\widetilde{\boldsymbol{\theta}}) = \mathbb{E}\left[h^2\left(P_{\boldsymbol{\theta}^\star}, P_{\widetilde{\boldsymbol{\theta}}}\right)\right] = \mathbb{E}\left[\int_{\mathscr{W}} h^2\left(Q_{\boldsymbol{\theta}^\star}(w), Q_{\widetilde{\boldsymbol{\theta}}(\boldsymbol{X})}(w)\right) dP_W(w)\right]$$

of an estimator $\widetilde{\boldsymbol{\theta}}(\boldsymbol{X})$ by the Monte Carlo method on the basis of 500 replications. For this simulation study $n = 500$. We recall that for a natural exponential family,

$$h^2(Q_\theta, Q_{\theta'}) = 1 - \exp\left[A\left(\frac{\theta+\theta'}{2}\right) - \frac{A(\theta)+A(\theta')}{2}\right] \tag{30}$$

where A is given in (2).

Bernoulli model. We consider the function $\boldsymbol{\theta}^\star = \boldsymbol{\theta}$ given by (27) with $\eta = (1, \ldots, 1) \in \mathbb{R}^6$. The distribution P_W is $(P_W^{(1)} + P_W^{(2)} + P_W^{(3)})/3$ where $P_W^{(1)}, P_W^{(2)}$ and $P_W^{(3)}$ are respectively the uniform distributions on the cubes

$$[-a, a]^5, \quad [b-0.25, b+0.25]^5 \quad \text{and} \quad [-b-0.25, -b+0.25]^5$$

with $a = 0.25$ and $b = 2$.

Poisson model. In this case $\boldsymbol{\theta}^\star = \boldsymbol{\theta}$ given by (28) with $\eta = (0.7, 3, 4, 10, 2, 5)$. The distribution P_W is $P_{W,1}^{\otimes 2} \otimes P_{W,2} \otimes P_{W,3}^{\otimes 2}$ where $P_{W,1}, P_{W,2}$ and $P_{W,3}$ are the uniform distributions on $[0.2, 0.25]$, $[0.2, 0.3]$ and $[0.1, 0.2]$ respectively.

Exponential model. We set $\boldsymbol{\theta}^\star = \boldsymbol{\theta}$ given by (29) with $\eta = (0.07, 3, 4, 6, 2, 1)$. The distribution P_W is $P_{W,1}^{\otimes 3} \otimes P_{W,2}^{\otimes 2}$ where $P_{W,1}$ and $P_{W,2}$ are the uniform distributions on $[0, 0.01]$ and $[0, 0.1]$ respectively.

In order to compare the performance of the ρ-estimator to the two other competitors we proceed as follows: we compute the (estimated) risk $R_n(\widehat{\boldsymbol{\theta}})$ of $\widehat{\boldsymbol{\theta}}$. We then use this quantity as a benchmark and given another estimator $\widetilde{\boldsymbol{\theta}}$ we compute the quantity

$$\mathcal{E}(\widetilde{\boldsymbol{\theta}}) = \frac{R_n(\widetilde{\boldsymbol{\theta}}) - R_n(\widehat{\boldsymbol{\theta}})}{R_n(\widehat{\boldsymbol{\theta}})} \quad \text{so that} \quad R_n(\widetilde{\boldsymbol{\theta}}) = \left(1 + \mathcal{E}(\widetilde{\boldsymbol{\theta}})\right) R_n(\widehat{\boldsymbol{\theta}}). \tag{31}$$

Note that large positive values of $\mathcal{E}(\widetilde{\boldsymbol{\theta}})$ indicate a significant superiority of our estimator as compared to $\widetilde{\boldsymbol{\theta}}$ and negative values inferiority. The respective values of $R_n(\widehat{\boldsymbol{\theta}})$ and $\mathcal{E}(\widetilde{\boldsymbol{\theta}})$ are displayed in Table 1.

TABLE 1. Values of $R_n(\widehat{\boldsymbol{\theta}})$ and $\mathcal{E}(\widetilde{\boldsymbol{\theta}})$ when the model is well-specified

	$R_n(\widehat{\boldsymbol{\theta}})$	$\mathcal{E}(\text{MLE})$	$\mathcal{E}(\widehat{\boldsymbol{\theta}}_0)$
Logit	0.0015	<+0.1%	–
Poisson	0.0015	<+0.1%	+450%
Exponential	0.0015	<+0.1%	+110%

We observe the following facts:

- When the model is correct, the risks of the MLE and $\widehat{\boldsymbol{\theta}}$ are the same (the value of $\mathcal{E}(\text{MLE})$ is not larger than 1/1000). In fact, a look at the simulations shows that the ρ-estimator coincides most of the time with the MLE, a fact which is consistent with the result proved in Baraud *et al.* (2017) (Section 5) that states the following: under (strong enough) assumptions, the MLE is a ρ-estimator when the statistical model is regular, exact and n is large enough. Our simulations indicate that the result actually holds under weaker assumptions.

- Both the MLE and the ρ-estimator outperform the median-based estimator $\widehat{\boldsymbol{\theta}}_0$.
- The quantities $R_n(\widehat{\boldsymbol{\theta}})$ are of order 0.0015 in all three cases. This fact can be explained as follows. In a regular statistical model $\mathscr{M}_0 = \{P_{\boldsymbol{\eta}},\ \boldsymbol{\eta} \in S\}$ parametrized with a parameter $\boldsymbol{\eta} \in S \subset \mathbb{R}^d$, the asymptotic normality properties of the MLE $\widehat{\boldsymbol{\eta}}_n$ together with the local equivalence of the Hellinger distance with the Euclidean one imply that when the data are i.i.d. with distribution $P_{\boldsymbol{\eta}^\star} \in \mathscr{M}_0$,
$$n\mathbb{E}\left[h^2(P_{\widehat{\boldsymbol{\eta}}_n}, P_{\boldsymbol{\eta}^\star})\right] \underset{n\to+\infty}{\longrightarrow} \frac{d}{8}.$$
In our simulation, conditionally to W, the distribution of Y is given by an exponential family parametrized by $d = 6$ parameters and the number of data being $n = 500$, we expect a risk of order $d/(8n) = 0.0015$, which is exactly what we obtained.
- The above result provides evidence that the algorithm we use does calculate the ρ-estimator as expected.
- In all the simulations we carried out, the algorithm required at most **two iterations** before the stopping condition $\boldsymbol{\upsilon}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leqslant 1$ was met.

5.3. **Comparisons of the estimators in presence of outliers.** We now work with $n = 501$ independent random variables $X_1, \ldots, X_n$. The 500 first variables $X_1, \ldots, X_{n-1}$ are i.i.d. with distribution $P_{\boldsymbol{\theta}^\star}$ and simply follow the framework of the previous section. The last observation is chosen as follows. In the Bernoulli model $W_n = 1000(1, 1, 1, 1, 1)$ and $Y_n = -1$, for the Poisson case $W_n = 0.1(1, 1, 1, 1, 1)$ and $Y_n = 200$ and for the exponential case $W_n = 5 \times 10^{-3}(1, 1, 1, 10, 10)$ and $Y_n = 1000$. The results are displayed in Table 2 on the basis of 500 replications.

TABLE 2. Values of $R_n(\widehat{\boldsymbol{\theta}})$ and $\mathcal{E}(\widetilde{\boldsymbol{\theta}})$ in presence of an outlier

	$R_n(\widehat{\boldsymbol{\theta}})$	$\mathcal{E}$(MLE)	$\mathcal{E}(\widehat{\boldsymbol{\theta}}_0)$
Logit	0.0015	+13000%	–
Poisson	0.0019	+1900%	+330%
Exponential	0.0018	+6000%	+78%

We observe the following facts:

- the risks of the ρ-estimator are quite similar to those given in Table 1 despite the presence of an outlier among the data set;
- the MLE behaves poorly;
- the performance of $\widehat{\boldsymbol{\theta}}$ remains much better than that of the median-based estimator $\widehat{\boldsymbol{\theta}}_0$.

Let us now display the quartiles of the distribution of the number of iterations that have been necessary to compute the ρ-estimator.

Table 3. Quartiles for the number of iterations in presence of outliers

	1st Quartile	Median	3rd Quartile	Maximum
Logit	3	3	3	6
Poisson	2	2	2	3
Exponential	2	2	2	3

Table (3) shows that the computation of the ρ-estimator requires only a few iterations of the algorithm.

5.4. **Comparisons of the estimators when the data are contaminated.** We now set $n = 500$ and define $P_{\boldsymbol{\theta}^\star}$ and P_W as in Section 5.2. We now assume that $X_1, \ldots, X_n$ are i.i.d. with distribution $P^\star = 0.95P_{\boldsymbol{\theta}^\star} + 0.05R$ for some (contaminating) distribution R on $\mathscr{W} \times \mathscr{Y}$ with first marginal given by P_W. We restrict ourselves to the Poisson and exponential cases (we exclude the Bernoulli model since the Bernoulli distribution remains stable under the contamination by another Bernoulli distribution). In the Poisson case, we choose for R the distribution of the random variable $(W, 80 + B)$ where the conditional distribution of B given $W = (w_1, \ldots, w_5)$ is Bernoulli with mean $(1 + \exp[-(w_1 - w_2 - w_4 + w_5)])^{-1}$. In the case of the exponential distribution $R = P_W \otimes \mathcal{U}([50, 60])$ where $\mathcal{U}([50, 60])$ denotes the uniform distribution on $[50, 60]$.

We measure the performance of an estimator $\widetilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^\star$ by means of the quantity

$$\overline{R}_n(\widetilde{\boldsymbol{\theta}}) = \mathbb{E}\left[h^2\left(P^\star, P_{\widetilde{\boldsymbol{\theta}}}\right)\right]$$

that we evaluate by Monte Carlo on the basis of 500 replications. We compare the performance of $\widehat{\boldsymbol{\theta}}$ to a competitor $\widetilde{\boldsymbol{\theta}}$ by evaluating the quantity

$$\overline{\mathcal{E}}(\widetilde{\boldsymbol{\theta}}) = \frac{\overline{R}_n(\widetilde{\boldsymbol{\theta}}) - \overline{R}_n(\widehat{\boldsymbol{\theta}})}{\overline{R}_n(\widehat{\boldsymbol{\theta}})}. \tag{32}$$

The results are displayed in Table 4.

Let us now comment these results.

TABLE 4. Values of $\overline{R}_n(\widehat{\boldsymbol{\theta}})$ and $\overline{\mathcal{E}}(\widetilde{\boldsymbol{\theta}})$ under contamination (5%)

	$\overline{R}_n(\widehat{\boldsymbol{\theta}})$	$\overline{\mathcal{E}}(\text{MLE})$	$\overline{\mathcal{E}}(\widehat{\boldsymbol{\theta}}_0)$
Poisson	0.028	+760%	+11%
Exponential	0.040	+320%	–17%

- With our choices of the contaminating distributions R, the (squared) Hellinger distance between the true distribution $P^\star$ of the data and the model is of order $h^2(P^\star, P_{\boldsymbol{\theta}^\star}) \approx 0.025$. As expected, we get that $\overline{R}_n(\widehat{\boldsymbol{\theta}}) \geqslant 0.025 \approx h^2(P^\star, P_{\boldsymbol{\theta}^\star})$. Note that the situation is extreme in the sense that the approximation error is much larger than estimation error that can be achieved when the model is well specified (which is about 0.0015). This means that the model is "very" misspecified.
- The MLE behaves poorly.
- In the exponential case, the median-based estimator $\widehat{\boldsymbol{\theta}}_0$ outperforms the ρ-estimator while the opposite situation occurs in the Poisson case.

TABLE 5. Quartiles for the number of iterations when the data are contaminated

	1st Quartile	Median	3rd Quartile	Maximum
Poisson	5	5	5	100
Exponential	5	10	30	100

In Table 5, we observe that the number of iterations for calculating the ρ-estimator increases substantially as compared to the two previous situations. We note that for some simulations the algorithm was iterated 100 times (which corresponds to the maximal number of iterations that we allow) and the stopping condition $\boldsymbol{v}(\boldsymbol{X}, \widehat{\boldsymbol{\theta}}) \leqslant 1$ was not met. Despite this fact, the estimator that we get at the final iteration, hence after 100 iterations, performs well since the values of the risks $\overline{R}_n(\widehat{\boldsymbol{\theta}})$ are of the same order as $h^2(P^\star, P_{\boldsymbol{\theta}^\star})$ and comparable to the median-based estimator $\widehat{\boldsymbol{\theta}}_0$.

6. An upper bound on the expectation of the supremum of an empirical process over a VC-subgraph class

The aim of this section is to prove the following result.

Theorem 2. *Let $X_1, \dots, X_n$ be n independent random variables with values in $(\mathscr{X}, \mathcal{X})$ and $\mathscr{F}$ an at most countable VC-subgraph class of functions with* *values in* $[-1,1]$ *and VC-dimension not larger than* $V \geqslant 1$*. If*

$$Z(\mathscr{F}) = \sup_{f\in\mathscr{F}} \left|\sum_{i=1}^{n}(f(X_i) - \mathbb{E}\left[f(X_i)\right])\right| \quad \textit{and} \quad \sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[f^2(X_i)\right] \leqslant \sigma^2 \leqslant 1,$$

then

$$\mathbb{E}\left[Z(\mathscr{F})\right] \leqslant 4.74\sqrt{nV\sigma^2\mathscr{L}(\sigma)} + 90V\mathscr{L}(\sigma), \tag{33}$$

with $\mathscr{L}(\sigma) = 9.11 + \log(1/\sigma^2)$*.*

Let us now turn to the proof. It follows from classical symmetrisation arguments that $\mathbb{E}\left[Z(\mathscr{F})\right] \leqslant 2\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]$, where $\overline{Z}(\mathscr{F}) = \sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|$ and $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. Rademacher random variables. It is therefore enough to prove that

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leqslant 2.37\sqrt{nV\sigma^2\mathscr{L}(\sigma)} + 45V\mathscr{L}(\sigma). \tag{34}$$

Given a probability P and a class of functions $\mathscr{G}$ on $(E,\mathcal{E})$ we denote by $N_r(\epsilon,\mathscr{G},P)$ the smallest cardinality of an ϵ-net for the $\mathbb{L}_r(E,\mathcal{E},P)$-norm $\|\cdot\|_{r,P}$, i.e. the minimal cardinality of a subset $\mathscr{G}[\epsilon]$ of $\mathscr{G}$ that satisfies for all $g\in\mathscr{G}$

$$\inf_{\overline{g}\in\mathscr{G}[\epsilon]}\|g-\overline{g}\|_{r,P} = \inf_{\overline{g}\in\mathscr{G}[\epsilon]}\left(\int_E |g-\overline{g}|^r\,dP\right)^{1/r} \leqslant \epsilon.$$

We start with the following lemma.

Lemma 1. *Whatever the probability* P *on* $(\mathscr{X},\mathcal{X})$*,* $\epsilon\in(0,2)$ *and* $r\geqslant 1$

$$N_r(\epsilon,\mathscr{F},P) \leqslant e(V+1)(2e)^V\left(\frac{2}{\epsilon}\right)^{rV}.$$

Proof of Lemma 1. Let λ be the Lebesgue measure on $([-1,1],\mathscr{B}([-1,1]))$ and Q the product probability $P\otimes(\lambda/2)$ on $(E,\mathcal{E}) = (\mathscr{X}\times[-1,1],\mathcal{X}\times\mathscr{B}([-1,1]))$. Given two elements $f,g\in\mathscr{F}$ and $x\in\mathscr{X}$

$$\begin{aligned}\int_{[-1,1]}\left|1\!\!1_{f(x)>t} - 1\!\!1_{g(x)>t}\right|dt &= \int_{[-1,1]}\left(1\!\!1_{f(x)>t\geqslant g(x)} + 1\!\!1_{g(x)>t\geqslant f(x)}\right)dt\\ &= |f(x)-g(x)|\end{aligned}$$

and, setting $C_f = \{(x,t)\in\mathscr{X}\times[-1,1],\ f(x)>t\}$ the subgraph of f and similarly C_g that of g, we deduce from Fubini's theorem that

$$\begin{aligned}\|f-g\|_{1,P} &= \int_{\mathscr{X}}|f-g|\,dP = 2\int_{\mathscr{X}\times[-1,1]}\left|1\!\!1_{C_f}(x,t) - 1\!\!1_{C_g}(x,t)\right|dQ\\ &= 2\left\|1\!\!1_{C_f} - 1\!\!1_{C_g}\right\|_{1,Q}.\end{aligned}$$

Since the functions $f,g\in\mathscr{F}$ take their values in $[-1,1]$,

$$\|f-g\|_{r,P}^r = \int_{\mathscr{X}}|f-g|^r\,dP \leqslant 2^{r-1}\int_{\mathscr{X}}|f-g|\,dP \leqslant 2^r\left\|1\!\!1_{C_f} - 1\!\!1_{C_g}\right\|_{1,Q}$$

and consequently, for all $\varepsilon>0$

$$N_r(\epsilon,\mathscr{F},P)\leqslant N_1((\varepsilon/2)^r,\mathscr{G},Q)\quad\text{with}\quad\mathscr{G}=\{\mathbb{1}_{C_f},\ f\in\mathscr{F}\}.$$

Since $\mathscr{F}$ is VC-subgraph with VC-dimension not larger than V, the class $\mathscr{G}$ is by definition VC with dimension not larger than V and the result follows from Corollary 1 in Haussler (1995). □

The proof of Theorem 2 is based on a chaining argument. It follows from the monotone convergence theorem that it is actually enough to prove (34) with $\mathscr{F}_J$, $J\geqslant 1$, in place of $\mathscr{F}$ where $(\mathscr{F}_J)_{J\geqslant 1}$ is a sequence of finite subsets of $\mathscr{F}$ which is increasing for the inclusion and satisfies $\bigcup_{J\geqslant 1}\mathscr{F}_J=\mathscr{F}$. We may therefore assume with no loss of generality that $\mathscr{F}$ is finite.

Let q be some positive number in $(0,1)$ to be chosen later on and $P_{\boldsymbol{X}}$ the empirical distribution $n^{-1}\sum_{i=1}^n\delta_{X_i}$. We shall denote by $\mathbb{E}_\varepsilon$ the expectation with respect to the Rademacher random variables ε_i, hence conditionally on $\boldsymbol{X}=(X_1,\ldots,X_n)$. We set

$$\widehat{\sigma}^2=\widehat{\sigma}^2(\boldsymbol{X})=\sup_{f\in\mathscr{F}}\|f\|^2_{2,\boldsymbol{X}}=\sup_{f\in\mathscr{F}}\left[\frac{1}{n}\sum_{i=1}^n f^2(X_i)\right]\in[0,1].$$

For each positive integer k, let $\mathscr{F}_k=\mathscr{F}_k(\boldsymbol{X})$ be a minimal $(q^k\widehat{\sigma})$-net for $\mathscr{F}$ with respect to the $\mathbb{L}_2(\mathscr{X},\mathcal{X},P_{\boldsymbol{X}})$-norm denoted $\|\cdot\|_{2,\boldsymbol{X}}$. In particular, we can associate to a function $f\in\mathscr{F}$ a sequence $(f_k)_{k\geqslant 1}$ with $f_k\in\mathscr{F}_k$ satisfying $\|f-f_k\|_{2,\boldsymbol{X}}\leqslant q^k\widehat{\sigma}$ for all $k\geqslant 1$. Actually, since $\mathscr{F}$ is finite $f_k=f$ for all k large enough. Besides, it follows from Lemma 1 with the choices $r=2$ and $P=P_{\boldsymbol{X}}$ that for all $k\geqslant 1$ we can choose $\mathscr{F}_k$ in such a way that $\log[\operatorname{Card}\mathscr{F}_k]$ is not larger than $h(q^k\widehat{\sigma})$ where

$$h(\epsilon)=\log\left[e(V+1)(2e)^V\right]+2V\log\left(\frac{2}{\epsilon}\right)\quad\text{for all }\epsilon\in(0,1]. \tag{35}$$

For $f\in\mathscr{F}$, the following (finite) decomposition holds

$$\begin{aligned}\sum_{i=1}^n\varepsilon_i f(X_i)&=\sum_{i=1}^n\varepsilon_i f_1(X_i)+\sum_{i=1}^n\varepsilon_i\sum_{k=1}^{+\infty}[f_{k+1}(X_i)-f_k(X_i)]\\&=\sum_{i=1}^n\varepsilon_i f_1(X_i)+\sum_{k=1}^{+\infty}\left[\sum_{i=1}^n\varepsilon_i\left(f_{k+1}(X_i)-f_k(X_i)\right)\right].\end{aligned}$$

Setting $\mathscr{F}_k^2=\{(f_k,f_{k+1}),\ f\in\mathscr{F}\}$ for all $k\geqslant 1$, we deduce that

$$\begin{aligned}\overline{Z}(\mathscr{F})\leqslant&\sup_{f\in\mathscr{F}_1}\left|\sum_{i=1}^n\varepsilon_i f(X_i)\right|\\&+\sum_{k=1}^{+\infty}\sup_{(f_k,f_{k+1})\in\mathscr{F}_k^2}\left|\sum_{i=1}^n\varepsilon_i\left[f_{k+1}(X_i)-f_k(X_i)\right]\right|\end{aligned}$$

and consequently,

$$\mathbb{E}_\varepsilon\left[\overline{Z}(\mathscr{F})\right] \leqslant \mathbb{E}_\varepsilon\left[\sup_{f\in\mathscr{F}_1}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] + \sum_{k=1}^{+\infty}\mathbb{E}_\varepsilon\left[\sup_{(f_k,f_{k+1})\in\mathscr{F}_k^2}\left|\sum_{i=1}^n \varepsilon_i\left[f_k(X_i)-f_{k+1}(X_i)\right]\right|\right].$$

Given a finite set $\mathscr{G}$ of functions on $\mathscr{X}$ and setting $-\mathscr{G}=\{-g,\ g\in\mathscr{G}\}$ and $v^2=\max_{g\in\mathscr{G}}\|g\|_{2,\boldsymbol{X}}^2$, we shall repeatedly use the inequality

$$\mathbb{E}\left[\sup_{g\in\mathscr{G}}\left|\sum_{i=1}^n \varepsilon_i g(X_i)\right|\right] = \mathbb{E}\left[\sup_{g\in\mathscr{G}\cup(-\mathscr{G})}\sum_{i=1}^n \varepsilon_i g(X_i)\right] \leqslant \sqrt{2n\log(2\operatorname{Card}\mathscr{G})v^2}$$

that can be found in Massart (2007)[inequality (6.3)]. Since $\max_{f\in\mathscr{F}_1}\|f\|_{2,\boldsymbol{X}}^2 \leqslant \widehat{\sigma}^2$, $\log(\operatorname{Card}\mathscr{F}_1)\leqslant h(q\widehat{\sigma})$, $\log(\operatorname{Card}\mathscr{F}_k^2)\leqslant h(q^k\widehat{\sigma})+h(q^{k+1}\widehat{\sigma})$ and

$$\begin{aligned}\sup_{(f_k,f_{k+1})\in\mathscr{F}_k^2}\|f_k-f_{k+1}\|_{2,\boldsymbol{X}}^2 &\leqslant \sup_{f\in\mathscr{F}}\sup_{(f_k,f_{k+1})\in\mathscr{F}_k^2}\left(\|f-f_k\|_{2,\boldsymbol{X}}+\|f-f_{k+1}\|_{2,\boldsymbol{X}}\right)^2\\ &\leqslant (1+q)^2 q^{2k}\widehat{\sigma}^2\end{aligned}$$

we deduce that

$$\mathbb{E}_\varepsilon\left[\sup_{f\in\mathscr{F}_1}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] \leqslant \widehat{\sigma}\sqrt{2n\left(\log 2+h(q\widehat{\sigma})\right)},$$

and for all $k\geqslant 1$

$$\begin{aligned}&\mathbb{E}_\varepsilon\left[\sup_{(f,g)\in\mathscr{F}_k^2}\left|\sum_{i=1}^n \varepsilon_i\left[g(X_i)-f(X_i)\right]\right|\right]\\ &\qquad\leqslant \widehat{\sigma}(1+q)q^k\sqrt{2n\left(\log 2+h(q^k\widehat{\sigma})+h(q^{k+1}\widehat{\sigma})\right)}.\end{aligned}$$

Setting $g: u\mapsto\sqrt{\log 2+h(u)+h(qu)}$ on $(0,1]$ and using the fact that g is decreasing (since h is) we deduce that

$$\begin{aligned}&\mathbb{E}_\varepsilon\left[\overline{Z}(\mathscr{F})\right]\\ &\leqslant \widehat{\sigma}\sqrt{2n}\left[\sqrt{\log 2+h(q\widehat{\sigma})}+(1+q)\sum_{k\geqslant 1}q^k\sqrt{\log 2+h(q^k\widehat{\sigma})+h(q^{k+1}\widehat{\sigma})}\right]\\ &\leqslant \widehat{\sigma}\sqrt{2n}\left[g(\widehat{\sigma})+(1+q)\sum_{k\geqslant 1}q^k g(q^k\widehat{\sigma})\right]\\ &\leqslant \sqrt{2n}\left[\frac{1}{1-q}\int_{q\widehat{\sigma}}^{\widehat{\sigma}}g(u)du+\frac{1+q}{1-q}\sum_{k\geqslant 1}\int_{q^{k+1}\widehat{\sigma}}^{q^k\widehat{\sigma}}g(u)du\right]\\ &\leqslant \sqrt{2n}\frac{1+q}{1-q}\int_0^{\widehat{\sigma}}g(u)du.\end{aligned}$$

The mapping g being positive and decreasing, the function $G : y \mapsto \int_0^y g(u)du$ is increasing and concave. Taking the expectation with respect to $\boldsymbol{X}$ on both sides of the previous inequality and using Jensen's inequality we get

$$\begin{aligned}\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] &\leqslant \sqrt{2n}\frac{1+q}{1-q}\mathbb{E}\left[G(\widehat{\sigma})\right] \leqslant \sqrt{2n}\frac{1+q}{1-q}G\left(\mathbb{E}\left[\widehat{\sigma}\right]\right)\\ &\leqslant \sqrt{2n}\frac{1+q}{1-q}G\left(\sqrt{\mathbb{E}\left[\widehat{\sigma}^2\right]}\right). \end{aligned} \tag{36}$$

By symmetrization and contraction arguments (see Theorem 4.12 in Ledoux and Talagrand (1991)),

$$\begin{aligned}\mathbb{E}\left[n\widehat{\sigma}^2\right] &\leqslant \mathbb{E}\left[\sup_{f\in\mathscr{F}}\sum_{i=1}^n\left(f^2(X_i)-\mathbb{E}\left[f^2(X_i)\right]\right)\right] + \sup_{f\in\mathscr{F}}\sum_{i=1}^n\mathbb{E}\left[f^2(X_i)\right]\\ &\leqslant 2\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n\varepsilon_i f^2(X_i)\right|\right] + n\sigma^2\\ &\leqslant 8\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n\varepsilon_i f(X_i)\right|\right] + n\sigma^2 = 8\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] + n\sigma^2 \end{aligned} \tag{37}$$

and we infer from (36) that

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leqslant \sqrt{2n}\frac{1+q}{1-q}G\left(B\right) \quad \text{with} \quad B = \sqrt{\sigma^2 + \frac{8\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]}{n}} \wedge 1. \tag{38}$$

The following lemma provides an evaluation of G.

Lemma 2. *Let a, b, y_0 be positive numbers and $y \in [y_0, 1]$,*

$$\int_0^y \sqrt{a + b\log(1/u)du} \leqslant \left(1 + \frac{b}{2a}\right) y\sqrt{a + b\log(1/y_0)}.$$

Proof. Using an integration by parts and the fact that

$$\frac{d}{du}\sqrt{a + b\log(1/u)} = -\frac{b}{2u\sqrt{a + b\log(1/u)}}$$

we get

$$\begin{aligned}\int_0^y \sqrt{a + b\log(1/u)}du &= \left[u\sqrt{a + b\log(1/u)}\right]_0^y + \frac{1}{2}\int_0^y \frac{b}{\sqrt{a + b\log(1/u)}}du\\ &\leqslant y\sqrt{a + b\log(1/y)} + \frac{by}{2\sqrt{a + b\log(1/y)}}\\ &= y\sqrt{a + b\log(1/y)}\left[1 + \frac{b}{2\left(a + b\log(1/y)\right)}\right] \end{aligned}$$

and the conclusion follows from the fact that $y_0 \leqslant y \leqslant 1$. □

Since for all $y \in (0, 1]$, $g(y) = \sqrt{a + b\log(1/y)}$ with

$$a = \log[2e^2(V+1)^2] + 2V\log(8e/q) \quad \text{and} \quad b = 4V$$

we may apply Lemma 2 with $y_0 = \sigma$ and $y = B$ and deduce from (38) that

$$\begin{aligned}\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] &\leqslant \sqrt{2n}\frac{1+q}{1-q}\left(1+\frac{b}{2a}\right)B\sqrt{a+b\log(1/\sigma)}\\ &\leqslant \sqrt{2n}\frac{1+q}{1-q}\left(1+\frac{b}{2a}\right)\sqrt{\sigma^2+\frac{8\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]}{n}}\sqrt{a+b\log(1/\sigma)}.\end{aligned}$$

Solving the inequality $\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leqslant A\sqrt{2n\sigma^2+16\mathbb{E}\left[\overline{Z}(\mathscr{F})\right]}$ with

$$A = \frac{1+q}{1-q}\left(1+\frac{b}{2a}\right)\sqrt{a+b\log(1/\sigma)}$$

we get that

$$\mathbb{E}\left[\overline{Z}(\mathscr{F})\right] \leqslant 8A^2+\sqrt{64A^4+2A^2n\sigma^2} \leqslant 16A^2+A\sqrt{2n\sigma^2}. \tag{39}$$

Finally, we conclude by using the inequalities

$$\begin{aligned}\frac{b}{2a} &= \frac{4V}{2\left[\log[2e^2(V+1)^2]+2V\log(8e/q)\right]} \leqslant \frac{1}{\log(8e/q)}\\ \frac{a}{b} &= \frac{\log[2e^2(V+1)^2]+2V\log(8e/q)}{4V}\\ &= \frac{\log(8e/q)}{2}+\frac{\log[2e^2(V+1)^2]}{4V} \leqslant \frac{\log(8e/q)}{2}+\frac{\log[8e^2]}{4}\\ &= \log\left(\frac{8^{3/4}e}{\sqrt{q}}\right)\end{aligned}$$

which, with our choice $q = 0.0185$, give

$$\begin{aligned}A &\leqslant \frac{1+q}{1-q}\left(1+\frac{1}{\log(8e/q)}\right)\sqrt{4V\left(\log\left(\frac{8^{3/4}e}{\sqrt{q}}\right)+\log\frac{1}{\sigma}\right)}\\ &\leqslant 2.37\sqrt{V\left(4.555+\log\frac{1}{\sigma}\right)}\end{aligned}$$

and together with (39) lead to (34).

7. Proofs

7.1. **Proof of Theorem 1.** We recall that the function ψ defined by (6) satisfies Assumption 2 of Baraud and Birgé (2018) with $a_0 = 4$, $a_1 = 3/8$ and $a_2^2 = 3\sqrt{2}$ (see their with Proposition 3). Theorem 1 is a consequence of Theorem 1 of Baraud and Birgé (2018). Set $\boldsymbol{\mu} = \bigotimes_{i=1}^n \mu_i$ with $\mu_i = P_{W_i} \otimes \nu$ for all $i \in \{1, \ldots, n\}$, denote by $\boldsymbol{\mathcal{P}}$ the following families of densities (with respect to $\boldsymbol{\mu}$) on $\mathscr{X}^n = (\mathscr{W} \times \mathscr{Y})^n$

$$\boldsymbol{\mathcal{P}} = \{\mathbf{p}_{\boldsymbol{\theta}} : \mathbf{x} = (x_1, \ldots, x_n) \mapsto q_{\boldsymbol{\theta}}(x_1)\ldots q_{\boldsymbol{\theta}}(x_n),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

and by $\mathscr{P}$ the corresponding ρ-model, i.e. the countable set $\{\mathbf{P} = \mathbf{p}_{\boldsymbol{\theta}} \cdot \boldsymbol{\mu},\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ with representation $(\boldsymbol{\mu}, \boldsymbol{\mathcal{P}})$.

Let us first prove

Proposition 5. *Under Assumption 1, the class of functions* $\mathcal{P} = \{q_{\boldsymbol{\theta}} : (w, y) \mapsto q_{\boldsymbol{\theta}(w)}(y),\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\}$ *on* $\mathscr{X} = \mathscr{W} \times \mathscr{Y}$ *is VC-subgraph with dimension not larger than* $9.41V$*.*

Proof. The exponential function being monotone, it suffices to prove that the family

$$\mathscr{F} = \left\{f : (w, y) \mapsto S(y)\boldsymbol{\theta}(w) - A(\boldsymbol{\theta}(w)),\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\right\}$$

is VC-subgraph on $\mathscr{X} = \mathscr{W} \times \mathscr{Y}$ with dimension not larger than $9.41V$. The function A being convex and continuous on I, the mapping defined on I by $\theta \mapsto S(y)\theta - A(\theta)$ is continuous and concave for all fixed $y \in \mathscr{Y}$. In particular, for $u \in \mathbb{R}$ the level set $\{\theta \in I,\ S(y)\theta - A(\theta) > u\}$ is an open subinterval of I of the form $(\underline{a}(y, u), \overline{a}(y, u))$ where $\underline{a}(y, u)$ and $\overline{a}(y, u)$ belong to the closure $\overline{I}$ of I in $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. For $\boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}$, let us set

$$\begin{aligned} C^+_{\boldsymbol{\theta}} &= \{(w, b, b') \in \mathscr{W} \times \overline{I}^2,\ \boldsymbol{\theta}(w) > b\} \\ C^-_{\boldsymbol{\theta}} &= \{(w, b, b') \in \mathscr{W} \times \overline{I}^2,\ \boldsymbol{\theta}(w) < b'\} \end{aligned}$$

and define $\mathscr{C}^+$ (respectively $\mathscr{C}^-$) as the class of all subsets $C^+_{\boldsymbol{\theta}}$ (respectively $C^-_{\boldsymbol{\theta}}$) when $\boldsymbol{\theta}$ varies among $\overline{\boldsymbol{\Theta}}$.

Let us prove that $\mathscr{C}^+$ is a VC-class of sets on $\mathscr{Z} = \mathscr{W} \times \overline{I}^2$ with dimension not larger than V. If $\mathscr{C}^+$ shatters the finite subset $\{z_1, \ldots, z_k\}$ of $\mathscr{Z}$ with $z_i = (w_i, b_i, b'_i)$ for $i \in \{1, \ldots, k\}$, necessarily the b_i belong to $\mathbb{R}$ for all $i \in \{1, \ldots, k\}$. Consequently, the class of subgraphs

$$\widetilde{\mathscr{C}}^+ = \left\{\{(w, b) \in \mathscr{W} \times \mathbb{R},\ \boldsymbol{\theta}(w) > b\},\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}}\right\}$$

shatters the points $\widetilde{z}_1 = (w_1, b_1), \ldots, \widetilde{z}_k = (w_k, b_k)$ in $\mathscr{W} \times \mathbb{R}$. This is possible only for $k \leqslant V$ since, by Assumption 1, $\overline{\boldsymbol{\Theta}}$ is VC-subgraph on $\mathscr{W}$ with dimension V.

Arguing similarly we obtain that $\mathscr{C}^-$ is also VC on $\mathscr{Z}$ with dimension not larger than V. In particular, it follows from van der Vaart and Wellner (2009) Theorem 1.1 that the class of subsets

$$\mathscr{C}^+ \bigwedge \mathscr{C}^- = \{C^+ \cap C^-,\quad C^+ \in \mathscr{C}^+,\ C^- \in \mathscr{C}^-\}$$

is VC on $\mathscr{Z}$ with dimension smaller than $9.41V$.

Let us now conclude the proof. If the class of subgraphs of $\mathscr{F}$ shatter the points $(w_1, y_1, u_1), \ldots, (w_k, y_k, u_k)$ in $\mathscr{W} \times \mathscr{Y} \times \mathbb{R}$, this means that for all subsets J of $\{1, \ldots, k\}$, there exists a function $\boldsymbol{\theta} = \boldsymbol{\theta}(J) \in \overline{\boldsymbol{\Theta}}$ such that

$$\begin{aligned} j \in J \quad &\Longleftrightarrow \quad S(y_j)\boldsymbol{\theta}(w_j) - A\left(\boldsymbol{\theta}(w_j)\right) > u_j \iff \boldsymbol{\theta}(w_j) \in (\underline{a}(y_j, u_j), \overline{a}(y_j, u_j)) \\ &\Longleftrightarrow \quad z_j = (w_j, \underline{a}(y_j, u_j), \overline{a}(y_j, u_j)) \in C^+_{\boldsymbol{\theta}} \cap C^-_{\boldsymbol{\theta}}. \end{aligned}$$

Hence, the class

$$\mathscr{C} = \left\{ C^+_{\boldsymbol{\theta}} \cap C^-_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \overline{\boldsymbol{\Theta}} \right\} \subset \mathscr{C}^+ \bigwedge \mathscr{C}^-$$

shatters $\{z_1, \ldots, z_k\}$ in $\mathscr{Z}$. This is possible for $k \leqslant 9.41V$ only and proves the fact that $\mathscr{F}$ is VC-subgraph with dimension not larger than $9.41V$. □

The result below provides an upper bound on the ρ-dimension function $(\mathbf{P}, \overline{\mathbf{P}}) \mapsto D^{\mathscr{P}}(\mathbf{P}, \overline{\mathbf{P}})$ of $\mathscr{P}$.

Proposition 6. *Under Assumption 1, for all product probabilities* $\mathbf{P}, \overline{\mathbf{P}} = \otimes_{i=1}^n \overline{P}_i$ *on* $(\mathscr{X}^n, \mathcal{X}^{\otimes n})$ *with* $\overline{P}_i = \overline{p} \cdot \mu_i$ *for all* $i \in \{1, \ldots, n\}$,

$$D^{\mathscr{P}}(\mathbf{P}, \overline{\mathbf{P}}) \leqslant 10^3 V \left[9.11 + \log_+ \left(\frac{n}{V} \right) \right].$$

Proof. Given two product probabilities $\mathbf{R} = \otimes_{i=1}^n R_i$ and $\mathbf{R}' = \otimes_{i=1}^n R_i'$ on $(\mathscr{X}^n, \mathcal{X}^{\otimes n})$, we set $\mathbf{h}^2(\mathbf{R}, \mathbf{R}') = \sum_{i=1}^n h^2(R_i, R_i')$ and for $y > 0$,

$$\mathscr{F}_y = \left\{ \psi\left(\sqrt{\frac{q_{\boldsymbol{\theta}}}{\overline{p}}} \right) \middle|\ \boldsymbol{\theta} \in \boldsymbol{\Theta}, \mathbf{h}^2(\mathbf{P}, \overline{\mathbf{P}}) + \mathbf{h}^2(\mathbf{P}, \mathbf{p}_\theta \cdot \boldsymbol{\mu}) < y^2 \right\}.$$

It follows from Proposition 5 and Baraud *et al.* (2017)[Proposition 42] that $\mathscr{F}_y$ is VC-subgraph with dimension not larger than $\overline{V} = 9.41V$. Besides, by Proposition 3 in Baraud and Birgé (2018) we know that our function ψ satisfies their Assumption 2 and more precisely (11) which, together with the definition of $\mathscr{F}_y$, implies that $\sup_{f \in \mathscr{F}_y} n^{-1} \sum_{i=1}^n \mathbb{E}\left[f^2(X_i) \right] \leqslant \sigma^2(y) = (a_2^2 y^2/n) \wedge 1$. Applying Theorem 2 with $\mathscr{F} = \mathscr{F}_y$, we obtain that

$$\begin{aligned} w^{\mathscr{P}}(\mathbf{P}, \overline{\mathbf{P}}, y) &= \mathbb{E}\left[\sup_{f \in \mathcal{F}_y} \left| \sum_{i=1}^n f(X_i) - \mathbb{E}\left[f(X_i) \right] \right| \right] \\ &\leqslant 4.74 a_2 y \sqrt{\overline{V} \mathscr{L}(\sigma(y))} + 90 \overline{V} \mathscr{L}(\sigma(y)) \\ &= 14.55 a_2 y \sqrt{V \mathscr{L}(\sigma(y))} + 846.9 V \mathscr{L}(\sigma(y)). \end{aligned}$$

Let $D \geqslant a_1^2 V/(16 a_2^4) = 2^{-11} V$ to be chosen later on and $\beta = a_1/(4a_2)$. For $y \geqslant \beta^{-1} \sqrt{D}$,

$$\begin{aligned} \mathscr{L}(\sigma(y)) = 9.11 + \log_+ \left(\frac{n}{a_2^2 y^2} \right) &\leqslant 9.11 + \log_+ \left(\frac{n}{a_2^2 \beta^{-2} D} \right) \\ = 9.11 + \log_+ \left(\frac{a_1^2 n}{16 a_2^4 D} \right) &\leqslant 9.11 + \log_+ \left(\frac{n}{V} \right) = L. \end{aligned}$$

Hence for all $y \geqslant \beta^{-1}\sqrt{D}$,

$$\begin{aligned}
w^{\mathscr{P}}(\mathbf{P},\overline{\mathbf{P}},y) &\leqslant 14.55a_2y\sqrt{VL}+846.9VL \\
&= \frac{a_1y^2}{8}\left[\frac{8\times 14.55a_2\sqrt{VL}}{a_1y}+\frac{8\times 846.9VL}{a_1y^2}\right] \\
&\leqslant \frac{a_1y^2}{8}\left[\frac{8\times 14.55a_2\sqrt{VL}}{a_1\beta^{-1}\sqrt{D}}+\frac{8\times 846.9VL}{a_1\beta^{-2}D}\right] \\
&= \frac{a_1y^2}{8}\left[\frac{2\times 14.55\sqrt{VL}}{\sqrt{D}}+\frac{8\times 846.9a_1VL}{16a_2^2D}\right] \\
&= \frac{a_1y^2}{8}\left[\frac{29.1\sqrt{VL}}{\sqrt{D}}+\frac{37.5VL}{D}\right] \leqslant \frac{a_1y^2}{8}
\end{aligned}$$

for $D = 10^3VL > 2^{-11}V$. The result follows from the definition of the ρ-dimension given in Baraud and Birgé (2018)[Definition 4]. □

Let us now end the proof of Theorem 1. It follows from Baraud and Birgé (2018)[Theorem 1], that the ρ-estimator $\widehat{\mathbf{P}} = \mathbf{P}_{\widehat{\boldsymbol{\theta}}}$ built on the ρ-model $\mathscr{P}$, which coincides with that described in Section 3.1, satisfies for all $\overline{\mathbf{P}} \in \mathscr{P}$, with a probability at least $1-e^{-\xi}$,

$$\mathbf{h}^2(\mathbf{P}^\star,\widehat{\mathbf{P}}) \leqslant \gamma\mathbf{h}^2(\mathbf{P}^\star,\mathscr{P}) + \gamma'\left(\frac{D^{\mathscr{P}}(\mathbf{P}^\star,\overline{\mathbf{P}})}{4.7}+1.49+\xi\right)$$

with

$$\gamma = \frac{4(a_0+8)}{a_1}+2+\frac{84}{a_2^2} < 150 \quad \text{and} \quad \gamma' = \frac{4}{a_1}\left(\frac{35a_2^2}{a_1}+74\right) < 5014$$

and $D^{\mathscr{P}}(\mathbf{P}^\star,\overline{\mathbf{P}}) \leqslant 10^3V\left[9.11+\log_+(n/V)\right]$ by Proposition 6. Finally, the result follows from the facts that $\mathbf{h}^2(\mathbf{P}^\star,\widehat{\mathbf{P}}) = \mathbf{h}^2(\mathbf{Q}^\star,\mathbf{Q}_{\widehat{\boldsymbol{\theta}}})$ and $\mathbf{h}^2(\mathbf{P}^\star,\mathscr{P}) = \mathbf{h}^2(\mathbf{Q}^\star,\mathscr{Q})$.

7.2. **A preliminary result.** The following result holds.

Proposition 7. *Let g be a 1-Lipschitz function on $\mathbb{R}$ supported on $[0,1]$, N some positive integer and L some positive number. For $\varepsilon \in \{-1,1\}^{2^N}$ define the function G_ε as*

$$G_\varepsilon(x) = L\sum_{k=0}^{2^N-1}\varepsilon_{k+1}g\left(2^Nx-k\right) \quad \text{for all } x\in[0,1]. \tag{40}$$

Then, G_ε satisfies (21) with $\alpha \in (0,1]$ and $M > 0$ provided that $L \leqslant 2^{-[(N-1)\alpha+1]}M$.

Proof. For $k \in \Lambda = \{0,\dots,2^N-1\}$, we set $g_k : x \mapsto g(2^Nx-k)$. Since g is 1-Lipschitz and supported on $[0,1]$, the function g_k is 2^N-Lipschitz on $\mathbb{R}$ and supported on $I_k = [2^{-N}k, 2^{-N}(k+1)] \subset [0,1]$ for all $k \in \Lambda$. In particular, the intersection of the supports of g_k and $g_{k'}$ reduces to at most a singleton when $k \neq k'$.

Let $x < y$ be two points in $[0,1]$. If there exists $k \in \Lambda$ such that $x, y \in I_k$, using that $0 \leqslant y - x \leqslant 2^{-N}$ and the fact that $L2^{N\alpha} \leqslant L2^{(N-1)\alpha+1} \leqslant M$, we obtain that

$$\begin{aligned}|G_\varepsilon(y) - G_\varepsilon(x)| &= L\,|g_k(y) - g_k(x)| \leqslant L2^N(y-x) \\ &\leqslant L2^N(y-x)^{1-\alpha}(y-x)^\alpha \leqslant L2^{N\alpha}(y-x)^\alpha \leqslant M(y-x)^\alpha.\end{aligned}$$

If $x \in I_k$ and $y \in I_{k'}$ with $k' \geqslant k+1$,

$$\left(y - 2^{-N}k'\right) + \left(2^{-N}(k+1) - x\right) \leqslant 2^{-N+1} \wedge (y-x)$$

and since g vanishes at 0 and 1,

$$\begin{aligned}|G_\varepsilon(y) - G_\varepsilon(x)| &= L\,|\varepsilon_{k'+1}g_{k'}(y) - \varepsilon_{k+1}g_k(x)| \leqslant L\,|g_{k'}(y)| + L\,|g_k(x)| \\ &= L\left|g_{k'}(y) - g_{k'}(2^{-N}k')\right| + L\left|g_k(2^{-N}(k+1)) - g_k(x)\right| \\ &\leqslant L2^N\left[y - 2^{-N}k' + 2^{-N}(k+1) - x\right]^{1-\alpha+\alpha} \\ &\leqslant L2^N 2^{(1-\alpha)(-N+1)}(y-x)^\alpha = L2^{(N-1)\alpha+1}(y-x)^\alpha\end{aligned}$$

and the conclusion follows from the fact that $L \leqslant 2^{-[(N-1)\alpha+1]}M$. □

We shall often use the following version of Assouad's lemma.

Lemma 3 (Assouad's Lemma)**.** *Let $\mathcal{P}$ be a family of probabilities on a measurable space $(\mathscr{X}, \mathcal{X})$. Assume that for some integer $d \geqslant 1$, $\mathcal{P}$ contains a subset of the form $\mathcal{C} = \{P_\varepsilon,\ \varepsilon \in \{-1,1\}^d\}$ with the following properties:*

(i) there exists $\eta > 0$ such that for all $\varepsilon, \varepsilon' \in \{-1,1\}^d$

$$h^2(P_\varepsilon, P_{\varepsilon'}) \geqslant \eta\delta(\varepsilon,\varepsilon') \quad \textit{with} \quad \delta(\varepsilon,\varepsilon') = \sum_{j=1}^d \mathbb{1}_{\varepsilon_j \neq \varepsilon'_j}$$

(ii) there exists a constant $a \in [0, 1/2]$ such that

$$h^2(P_\varepsilon, P_{\varepsilon'}) \leqslant \frac{a}{n} \quad \textit{for all } \varepsilon, \varepsilon' \in \{-1,1\}^d \textit{ satisfying } \delta(\varepsilon,\varepsilon') = 1.$$

Then for all measurable mappings $\widehat{P} : \mathscr{X}^n \to \mathcal{P}$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{\mathbf{P}}\left[h^2(P, \widehat{P}(\boldsymbol{X}))\right] \geqslant \frac{d\eta}{8}\max\left\{1 - \sqrt{2a}, (1-a/n)^{2n}\right\}, \tag{41}$$

where $\mathbb{E}_{\mathbf{P}}$ denotes the expectation with respect to a random variable $\boldsymbol{X} = (X_1, \ldots, X_n)$ with distribution $\mathbf{P} = P^{\otimes n}$.

Proof. Given a probability P on $(\mathscr{X}, \mathcal{X})$, let $\overline{\varepsilon}$ be a minimizer over $\{-1,1\}^d$ of the mapping $\varepsilon \mapsto h^2(P, P_\varepsilon)$. By definition of $\overline{\varepsilon}$, for all $\varepsilon \in \{-1,1\}^d$

$$h^2(P_\varepsilon, P_{\overline{\varepsilon}}) \leqslant 2\left(h^2(P, P_\varepsilon) + h^2(P, P_{\overline{\varepsilon}})\right) \leqslant 4h^2(P, P_\varepsilon).$$

Hence by (i), for all $\varepsilon \in \{-1,1\}^d$

$$h^2(P_\varepsilon, P) \geqslant \frac{\eta}{4}\delta(\varepsilon,\overline{\varepsilon}) = \sum_{i=1}^{d}\left[\frac{1+\varepsilon_i}{2}\ell_i(P) + \frac{1-\varepsilon_i}{2}\ell_i'(P)\right]$$

with $\ell_i(P) = (\eta/4)\mathbb{1}_{\overline{\varepsilon}_i=-1}$ and $\ell_i'(P) = (\eta/4)\mathbb{1}_{\overline{\varepsilon}_i=+1}$ for $i \in \{1,\ldots,d\}$. The result follows by applying the version of Assouad's lemma that can be found in Birgé (1986) with $\beta_i = a/n$ for all $i \in \{1,\ldots,d\}$, $\alpha = \eta/4$ and the change of notation from $\varepsilon \in \{-1,1\}$ to $\varepsilon \in \{0,1\}$. □

7.3. **Proof of Proposition 1.** When the Poisson family is parametrized by the mean, given two functions $\boldsymbol{\gamma},\boldsymbol{\gamma}'$ mapping $\mathscr{W} = [0,1]$ into $J = (0,+\infty)$, The Hellinger-type distance $h^2(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'})$ writes as

$$h^2(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}) = \int_{\mathscr{W}}\left[1 - e^{-\left(\sqrt{\gamma(w)}-\sqrt{\gamma'(w)}\right)^2/2}\right] dP_W(w). \tag{42}$$

Using that for all $x \in [0,1]$, $(1-e^{-1})x \leqslant 1 - e^{-x} \leqslant x$, we deduce from (42) that

$$\frac{1}{2}(1-e^{-1})\left\|\sqrt{\boldsymbol{\gamma}} - \sqrt{\boldsymbol{\gamma}'}\right\|_2^2 \leqslant h^2(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}) \leqslant \frac{1}{2}\left\|\sqrt{\boldsymbol{\gamma}} - \sqrt{\boldsymbol{\gamma}'}\right\|_2^2 \tag{43}$$

whenever $\left\|\sqrt{\boldsymbol{\gamma}} - \sqrt{\boldsymbol{\gamma}'}\right\|_\infty \leqslant 1$.

Let N be some positive integer, L some positive number and g a 1-Lipschitz function supported on $[0,1]$ with values in $[-b,b]$. Let us set $\Lambda = \{0,\ldots,2^N-1\}$ and for $\varepsilon \in \{-1,1\}^{|\Lambda|}$, G_ε the function defined by (40) and $\boldsymbol{\gamma}_\varepsilon = L + G_\varepsilon$. Under our assumption on g, $\boldsymbol{\gamma}_\varepsilon$ takes its values in $[(1-b)L, (1+b)L]$ and by Proposition 7, $\boldsymbol{\gamma}_\varepsilon$ satisfies (21) provided that $L \leqslant 2^{-[(N-1)\alpha+1]}M$. Hence, under the conditions $L \leqslant 2^{-[(N-1)\alpha+1]}M$ and $b < 1$, $\boldsymbol{\gamma}_\varepsilon$ belongs to $\mathcal{H}_\alpha(M)$ for all $\varepsilon \in \{-1,1\}^{|\Lambda|}$. For all $\varepsilon,\varepsilon' \in \{-1,1\}^{|\Lambda|}$,

$$\frac{|G_\varepsilon - G_{\varepsilon'}|}{2\sqrt{(1+b)L}} \leqslant \left|\sqrt{\boldsymbol{\gamma}_\varepsilon} - \sqrt{\boldsymbol{\gamma}_{\varepsilon'}}\right| = \frac{|\boldsymbol{\gamma}_\varepsilon - \boldsymbol{\gamma}_{\varepsilon'}|}{\sqrt{\boldsymbol{\gamma}_\varepsilon} + \sqrt{\boldsymbol{\gamma}_{\varepsilon'}}} \leqslant \frac{|G_\varepsilon - G_{\varepsilon'}|}{2\sqrt{(1-b)L}},$$

and

$$\left|\sqrt{\boldsymbol{\gamma}_\varepsilon} - \sqrt{\boldsymbol{\gamma}_{\varepsilon'}}\right| \leqslant \sqrt{(1+b)L} - \sqrt{(1-b)L} = \left[\sqrt{1+b} - \sqrt{1-b}\right]\sqrt{L}.$$

In particular, $\left\|\sqrt{\boldsymbol{\gamma}_\varepsilon} - \sqrt{\boldsymbol{\gamma}_{\varepsilon'}}\right\|_\infty \leqslant 1$ for

$$L \leqslant \left(\sqrt{1+b} - \sqrt{1-b}\right)^{-2} = \frac{1+\sqrt{1-b^2}}{2b^2} = L_0$$

and writing R_ε for $R_{\boldsymbol{\gamma}_\varepsilon}$ for short, it follows from (43) that

$$\frac{(1-e^{-1})}{8(1+b)L}\left\|G_\varepsilon - G_{\varepsilon'}\right\|_2^2 \leqslant h^2(R_\varepsilon, R_{\varepsilon'}) \leqslant \frac{1}{8(1-b)L}\left\|G_\varepsilon - G_{\varepsilon'}\right\|_2^2. \tag{44}$$

Since P_W is the uniform distribution and the supports of the functions $g_k : x \mapsto g(2^N x - k)$ for $k \in \Lambda$ are disjoint, we obtain that for all $\varepsilon, \varepsilon' \in \{-1,1\}^{|\Lambda|}$

$$\begin{aligned}\|G_\varepsilon - G_{\varepsilon'}\|_2^2 &= L^2 \sum_{k\in\Lambda} \int_{I_k} \left(\varepsilon_{k+1} g_k(x) - \varepsilon'_{k+1} g_k(x)\right)^2 dx \\ &= L^2 \sum_{k\in\Lambda} \left|\varepsilon_{k+1} - \varepsilon'_{k+1}\right|^2 \int_{I_k} g_k^2(x)dx = 4L^2 2^{-N} \|g\|_2^2 \, \delta(\varepsilon, \varepsilon').\end{aligned}$$

Let us denote by $P_{\boldsymbol{\gamma}} = R_{\boldsymbol{\gamma}} \cdot P_W$ the probability associated to $R_{\boldsymbol{\gamma}}$ and write P_ε for $P_{\boldsymbol{\gamma}_\varepsilon}$ for short. We deduce from (44) that provided that L and b satisfy

$$L \leqslant \left(2^{-[(N-1)\alpha+1]} M\right) \wedge \frac{1+\sqrt{1-b^2}}{2b^2} \wedge \frac{(1-b)2^{N-3}}{\|g\|_2^2 \, n} \tag{45}$$

the family of probabilities $\mathscr{C} = \{P_\varepsilon, \ \varepsilon \in \{-1,1\}^{|\Lambda|}\}$ is a subset of $\{P_{\boldsymbol{\gamma}}, \ \boldsymbol{\gamma} \in \mathcal{H}_\alpha(M)\}$ that fulfils the assumptions of Assouad's lemma (Lemma 3) with $d = |\Lambda| = 2^N$,

$$\eta = \frac{(1-e^{-1})L2^{-(N+1)} \|g\|_2^2}{1+b} \quad \text{and} \quad a = \frac{nL2^{-N} \|g\|_2^2}{1-b} \in [0, 1/8].$$

We derive from the equalities

$$h^2(R_\varepsilon, R_{\varepsilon'}) = \int_{\mathscr{W}} h^2\left(R_{\boldsymbol{\gamma}_\varepsilon(w)}, R_{\boldsymbol{\gamma}_{\varepsilon'}(w)}\right) dP_W(w) = h^2(P_\varepsilon, P_{\varepsilon'})$$

and (41) that

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) \geqslant \frac{(1-e^{-1}) \|g\|_2^2 \, L}{16(1+b)} \left(1 - \sqrt{2a}\right) \geqslant \frac{(1-e^{-1}) \|g\|_2^2 \, L}{32(1+b)}. \tag{46}$$

If $\|g\|_2^2 \, Mn > (1-b)/2$, we choose $N \geqslant 2$ such that

$$2^N \geqslant \left[\frac{2^{2+\alpha} \|g\|_2^2 \, Mn}{1-b}\right]^{\frac{1}{1+\alpha}} > 2^{N-1}.$$

Otherwise, we choose $N = 1$. Note that in any case,

$$2^{-[(N-1)\alpha+1]} M \leqslant \frac{(1-b)2^{N-3}}{n \, \|g\|_2^2}.$$

Besides, if $N \geqslant 2$

$$\begin{aligned}2^{-[(N-1)\alpha+1]} M = 2^{-1} M 2^{-(N-1)\alpha} &\geqslant 2^{-1} M \left[\frac{2^{2+\alpha} \|g\|_2^2 \, Mn}{1-b}\right]^{-\frac{\alpha}{1+\alpha}} \\ &= \left(\frac{(1-b)M^{\frac{1}{\alpha}}}{2^{3+\alpha+1/\alpha} \|g\|_2^2 \, n}\right)^{\frac{\alpha}{1+\alpha}} = L_1\end{aligned}$$

while for $2^{-[(N-1)\alpha+1]}M = M/2$ for $N=1$. Finally, we choose $L = L_0 \wedge L_1 \wedge (M/2)$, which satisfies (45), and we derive from (46) that

$$\begin{aligned}&\mathcal{R}_n(\mathcal{H}_\alpha(M))\\ &\geqslant \frac{(1-e^{-1})\|g\|_2^2}{32(1+b)}\left[\left(\frac{(1-b)M^{\frac{1}{\alpha}}}{2^{3+\alpha+1/\alpha}\|g\|_2^2 n}\right)^{\frac{\alpha}{1+\alpha}} \wedge \frac{M}{2} \wedge \frac{1+\sqrt{1-b^2}}{b^2}\right].\end{aligned}$$

The conclusion follows by taking $g(x) = x\mathbb{1}_{[0,1/2]} + (1-x)\mathbb{1}_{[1/2,1]}$ for which $b = 1/2$ and $\|g\|_2^2 = 1/12$.

7.4. **Proof of Proposition 2.** Since the statistical model $\overline{\mathscr{Q}} = \{R_\gamma = Q_{v^{-1}(\gamma)},\ \gamma \in J\}$ is regular with constant Fisher information equal to 8, by applying Theorem 7.6 page 81 in Ibragimov and Has'minskiĭ (1981) we obtain that

$$h^2\left(R_\gamma, R_{\gamma'}\right) \leqslant \left(\gamma' - \gamma\right)^2 \quad \text{for all } \gamma, \gamma' \in J$$

and for any compact subset K of J, there exists a constant $c_K > 0$

$$h^2\left(R_\gamma, R_{\gamma'}\right) \geqslant c_K^2\left(\gamma' - \gamma\right)^2 \quad \text{for all } \gamma, \gamma' \in K.$$

The result follows by substituting $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ to γ and γ' respectively and then integrating with respect to P_W.

7.5. **Proof of Proposition 3.** Let a_0 be the middle of the interval K of length $2\overline{L}$. Given $N \geqslant 1$, $L > 0$ and $\varepsilon \in \{-1,1\}^{2^N}$, we define $\boldsymbol{\gamma}_\varepsilon = a_0 + G_\varepsilon$ with G_ε as in Proposition 7. Provided that $L \leqslant \overline{L} \wedge L_0$ with $L_0 = 2^{-[(N-1)\alpha+1]}M$, the functions $\boldsymbol{\gamma}_\varepsilon$ takes their values in $K \subset J$ and satisfies (21) and consequently belongs to $\mathcal{H}_\alpha(M)$ for all $\varepsilon \in \{-1,1\}^{2^N}$. Let $R_\varepsilon = R_{\boldsymbol{\gamma}_\varepsilon}$ for all $\varepsilon \in \{-1,1\}^{2^N}$ and, as in the proof of Proposition 1, set $P_{\boldsymbol{\gamma}} = R_{\boldsymbol{\gamma}} \cdot P_W$ and $P_{\gamma_\varepsilon} = P_{\boldsymbol{\gamma}_\varepsilon}$ for short. Integrating the inequalities (23) and (24) with respect to P_W and using that for all $\varepsilon, \varepsilon' \in \{-1,1\}^{2^N}$, $\|G_\varepsilon - G_{\varepsilon'}\|_2 = \|\boldsymbol{\gamma}_\varepsilon - \boldsymbol{\gamma}_{\varepsilon'}\|_2$ we obtain that

$$c_K^2\|G_\varepsilon - G_{\varepsilon'}\|_2^2 \leqslant h^2\left(R_\varepsilon, R_{\varepsilon'}\right) \leqslant \kappa^2\|G_\varepsilon - G_{\varepsilon'}\|_2^2.$$

Since P_W is the uniform distribution on $[0,1]$, by arguing as in the proof of Proposition 1

$$\|G_\varepsilon - G_{\varepsilon'}\|_2^2 = 4L^2 2^{-N}\|g\|_2^2\delta(\varepsilon,\varepsilon') \quad \text{for all } \varepsilon, \varepsilon' \in \{-1,1\}^{2^N}$$

and consequently, provided that L satisfies

$$L \leqslant \overline{L} \wedge L_0 \wedge \left(4\kappa\|g\|_2\sqrt{2^{-(N-1)}n}\right)^{-1} \tag{47}$$

the family of probabilities $\mathscr{C} = \{P_\varepsilon,\ \varepsilon \in \{-1,1\}^{|\Lambda|}\}$ is a subset of $\mathcal{P} = \{P_{\boldsymbol{\gamma}},\ \boldsymbol{\gamma} \in \mathcal{H}_\alpha(M)\}$ that fulfils the assumptions of Lemma 3 with $d = 2^N$,

$$\eta = 4c_K^2 L^2 2^{-N}\|g\|_2^2 \quad \text{and} \quad a = 4n\kappa^2 L^2 2^{-N}\|g\|_2^2 \leqslant 1/8.$$

We derive from (41) that

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) \geqslant \frac{c_K^2\|g\|_2^2L^2}{2}\left(1-\sqrt{2a}\right) \geqslant \frac{c_K^2\|g\|_2^2L^2}{4}. \tag{48}$$

If $\kappa^2\|g\|_2^2M^2n > 1/8$, we choose $N \geqslant 2$ such that

$$2^N \geqslant \left(2^{2(2+\alpha)}\kappa^2\|g\|_2^2M^2n\right)^{1/(1+2\alpha)} > 2^{N-1}$$

and $N = 1$ otherwise. In any case, our choice of N satisfies

$$L_0 = 2^{-[(N-1)\alpha+1]}M \leqslant \left(4\kappa\|g\|_2\sqrt{2^{-(N-1)}n}\right)^{-1}.$$

When $N \geqslant 2$,

$$\begin{aligned} L_0^2 = 2^{-2\alpha(N-1)-2}M^2 &\geqslant \frac{M^2}{4}\left(2^{2(2+\alpha)}\kappa^2\|g\|_2^2M^2n\right)^{-\frac{2\alpha}{1+2\alpha}} \\ &= \left(\frac{M^{1/\alpha}}{2^{2\alpha+6+1/\alpha}\kappa^2\|g\|_2^2n}\right)^{\frac{2\alpha}{1+2\alpha}} = L_1^2, \end{aligned}$$

while $L_0 = M/2$ when $N = 1$. The choice $L = \overline{L} \wedge L_1 \wedge (M/2)$ satisfies (47) and we deduce from the equalities

$$h^2(R_\varepsilon, R_{\varepsilon'}) = \int_{\mathscr{W}} h^2\left(R_{\gamma_\varepsilon(w)}, R_{\gamma_{\varepsilon'}(w)}\right)dP_W(w) = h^2(P_\varepsilon, P_{\varepsilon'})$$

and (48) that

$$\mathcal{R}_n(\mathcal{H}_\alpha(M)) \geqslant \frac{c_K^2\|g\|_2^2}{4}\left[\left(\frac{M^{1/\alpha}}{2^{2\alpha+6+1/\alpha}\kappa^2\|g\|_2^2n}\right)^{\frac{2\alpha}{1+2\alpha}} \wedge \left(\frac{M^2}{4}\right) \wedge \overline{L}^2\right].$$

The conclusion follows by taking $g(x) = x\mathbb{1}_{[0,1/2]} + (1-x)\mathbb{1}_{[1/2,1]}$ which satisfies $\|g\|_2^2 = 1/12$.

7.6. **Proof of Proposition 4.** For $\boldsymbol{\gamma} \in \mathcal{H}_\alpha(M)$ and $j \in \{1,\dots,D\}$, let $\gamma_j = D\int_{I_j}\boldsymbol{\gamma}(w)dw$ and $\overline{\boldsymbol{\gamma}} = \sum_{j=1}^D\gamma_j\mathbb{1}_{I_j}$. Since $\boldsymbol{\gamma}$ takes its values in J, $\gamma_j \in J$ for all $j \in \{1,\dots,D\}$ and $\overline{\boldsymbol{\gamma}} = \sum_{j=1}^D\gamma_j\mathbb{1}_{I_j} \in \overline{\boldsymbol{S}}$. Since for all $w \in I_j$, $|\boldsymbol{\gamma}(w) - \overline{\boldsymbol{\gamma}}(w)| \leqslant \sup_{|w-w'|\leqslant 1/D}|\boldsymbol{\gamma}(w)-\boldsymbol{\gamma}(w')| \leqslant MD^{-\alpha}$ and $\boldsymbol{S}$ is dense in $\overline{\boldsymbol{S}}$ with respect to the supremum norm

$$\begin{aligned} \sup_{\boldsymbol{\gamma}\in\mathcal{H}_\alpha(M)}\inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{S}}}\|\boldsymbol{\gamma}-\overline{\boldsymbol{\gamma}}\|_2 &\leqslant \sup_{\boldsymbol{\gamma}\in\mathcal{H}_\alpha(M)}\inf_{\overline{\boldsymbol{\gamma}}\in\overline{\boldsymbol{S}}}\|\boldsymbol{\gamma}-\overline{\boldsymbol{\gamma}}\|_\infty \\ &= \sup_{\boldsymbol{\gamma}\in\mathcal{H}_\alpha(M)}\inf_{\overline{\boldsymbol{\gamma}}\in\boldsymbol{S}}\|\boldsymbol{\gamma}-\overline{\boldsymbol{\gamma}}\|_\infty \leqslant MD^{-\alpha}. \end{aligned}$$

Using (23) and the fact that the data $X_1,\dots,X_n$ are i.i.d., we deduce that for all functions $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ with values in J,

$$\mathbf{h}^2(\mathbf{R}_{\boldsymbol{\gamma}}, \mathbf{R}_{\boldsymbol{\gamma}'}) = nh^2(R_{\boldsymbol{\gamma}}, R_{\boldsymbol{\gamma}'}) \leqslant n\kappa^2\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_2^2 \leqslant n\kappa^2\|\boldsymbol{\gamma}-\boldsymbol{\gamma}'\|_\infty^2$$

and by applying Corollary 1 with $V = D+1$ we obtain that

$$\begin{aligned}\sup_{\gamma^\star\in\mathcal{H}_\alpha(M)} \mathbb{E}\left[h^2(R_{\gamma^\star}, R_{\widehat{\gamma}})\right] &\leqslant C'\left[\sup_{\gamma^\star\in\mathcal{H}_\alpha(M)} \inf_{\overline{\gamma}\in\boldsymbol{S}} h^2(R_{\gamma^\star}, R_{\overline{\gamma}}) + \frac{V}{n}\left[1+\log_+(n/V)\right]\right]\\ &\leqslant C'\left[\kappa^2 \sup_{\gamma^\star\in\mathcal{H}_\alpha(M)} \inf_{\overline{\gamma}\in\boldsymbol{S}} \|\boldsymbol{\gamma}^\star - \overline{\boldsymbol{\gamma}}\|_2^2 + \frac{V}{n}\left[1+\log_+(n/V)\right]\right]\\ &\leqslant C'\left[\kappa^2 M^2 D^{-2\alpha} + \frac{D+1}{n}\log(en)\right].\end{aligned}$$

Let us set $L_n = \log(en)$. With our choice of $D \geqslant 1$,

$$D - 1 < \left(\frac{\kappa^2 M^2 n}{L_n}\right)^{\frac{2\alpha}{1+2\alpha}} \leqslant D$$

hence $\kappa^2 M^2 D^{-2\alpha} \leqslant DL_n/n$, $D < 1 + (\kappa^2 M^2 n/L_n)^{\frac{2\alpha}{1+2\alpha}}$ and the result follows from the inequalities

$$\kappa^2 M^2 D^{-2\alpha} + \frac{(D+1)L_n}{n} \leqslant 2\frac{DL_n}{n} + \frac{L_n}{n} \leqslant 2\left[\frac{(\kappa M)^{1/\alpha} L_n}{n}\right]^{\frac{2\alpha}{1+2\alpha}} + \frac{3L_n}{n}.$$

References

Antoniadis, A. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with quadratic variance functions. *Biometrika*, 88(3):805–820.

Baraud, Y. and Birgé, L. (2018). Rho-estimators revisited: General theory and applications. *Ann. Statist.*, 46(6B):3767–3804.

Baraud, Y., Birgé, L., and Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517.

Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Relat. Fields*, 71(2):271–291.

Brown, L. D., Cai, T. T., and Zhou, H. H. (2010). Nonparametric regression in exponential families. *Ann. Statist.*, 38(4):2005–2046.

Candès, E. and Sur, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42.

Hansen, N. (2016). *The CMA Evolution Strategy: A Tutorial.*

Haussler, D. (1995). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Combin. Theory Ser. A*, 69(2):217–232.

Ibragimov, I. A. and Has'minskiĭ, R. Z. (1981). *Statistical Estimation. Asymptotic Theory*, volume 16. Springer-Verlag, New York.

Kolaczyk, E. D. and Nowak, R. D. (2005). Multiscale generalised linear models for nonparametric function estimation. *Biometrika*, 92(1):119–133.

Kroll, M. (2019). Non-parametric Poisson regression from independent and weakly dependent observations by model selection. *J. Statist. Plann. Inference*, 199:249–270.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin. Isoperimetry and processes.

Massart, P. (2007). *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

van der Vaart, A. and Wellner, J. A. (2009). A note on bounds for VC dimensions. In *High Dimensional Probability V: the Luminy volume*, volume 5 of *Inst. Math. Stat. Collect.*, pages 103–107. Inst. Math. Statist., Beachwood, OH.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer Series in Statistics. Springer-Verlag, New York.

Department of Mathematics (DMATH),
University of Luxembourg
Maison du nombre
6 avenue de la Fonte
L-4364 Esch-sur-Alzette
Grand Duchy of Luxembourg

Email address: yannick.baraud@uni.lu

Email address: juntong.chen@uni.lu