Scaling by subsampling for big data, with applications to statistical learning

Patrice Bertail, Mohammed Bouchouia, Ons Jelassi, Jessica Tressou, Mélanie Zetlaoui

Scaling by subsampling for big data, with applications to statistical learning

Patrice Bertail[a] and Mohammed Bouchouia [b] and Ons Jelassi [b] and Jessica Tressou [c] and Mélanie Zetlaoui [a]

[a] MODAL'X, UMR 9023, UPL - Université Paris Nanterre, CNRS, F92000 Nanterre, France;
[b] LTCI, Telecom Paris, Institut Polytechnique de Paris, Palaiseau, France;
[c] UPS - AgroParisTech - INRAE, UMR MIA-Paris, Paris, France, University of Tokyo, Graduate School of Agriculture and Life

ARTICLE HISTORY
Compiled April 11, 2023

ABSTRACT
Handling large datasets and calculating complex statistics on huge datasets require important computing resources. Using subsampling methods to calculate statistics of interest on small samples is often used in practice to reduce computational complexity, for instance using the divide and conquer strategy. In this article, we recall some results on subsampling distributions and derive a precise rate of convergence for these quantities and the corresponding quantiles. We also develop some standardization techniques based on subsampling unstandardized statistics in the framework of large datasets. It is argued that using several subsampling distributions with different subsampling sizes brings a lot of information on the behavior of statistical learning procedures: subsampling allows to estimate the rate of convergence of different algorithms, to estimate the variability of complex statistics, to estimate confidence intervals for out-of-sample errors and interpolate their values at larger scales. These results are illustrated on simulations, but also on two important datasets, frequently analyzed in the statistical learning community, EMNIST (recognition of digits) and VeReMi (analysis of Network Vehicular Reference Misbehavior).

KEYWORDS
Scaling, big data, Subsampling, Convergence rate estimation, Confidence intervals in statistical learning, Out-of sample error, EMNIST digits VeReMi

1. Introduction

The capacity to collect data has increased faster than our ability to analyze big datasets with a huge number of individuals. The standard statistical tools or statistical learning algorithms, like machine-learning procedures, maximum likelihood estimations, or general methods based on contrast minimization, are time-consuming in terms of optimization despite their polynomial complexities or require to access the data too many times. For these reasons, these approaches may be unsuitable for big data problems. To remedy the apparent intractability problem of learning from databases of explosive sizes and break the current computational barriers, we propose in this paper to use some variations of subsampling techniques studied in Politis et al. (1999) and Bertail

CONTACT M. Zetlaoui Email: melanie.zetlaoui@gmail.com

et al. (1999). Such approaches have been implemented in many applied problems and developed for instance in Kleiner et al. (2014). In the framework of big data, they are also at the core of some recent developments in survey sampling methods (Clémençon et al., 2014; Bertail et al., 2015, 2017) applied to statistical learning procedures.

The universal validity of the subsampling methods is proved in Politis and Romano (1994) and further developed in Bertail et al. (1999, 2004) for general converging or diverging statistics. A good survey on the power of subsampling methodology is given in Politis et al. (1999). More precisely, the subsampling distribution constructed with a much smaller size than the original one is a correct approximation of the distribution of the statistic of interest (possibly with an unknown rate of convergence), if the latter has a non-degenerate distribution, that is continuous at some point of interest. As mentioned in Hall (2003), such methods, close to bootstrap and subsampling ideas, were already proposed in the works of Mahalanobis in the 1940s (see the reprint Mahalanobis, 1958) but were abandoned because of the cost of paper: at that time, calculations were carried out by hand. They have also been developed by Bretagnolle (1983) and Bickel and Yahav (1988). See also the discussions about interpolations and extrapolations in Bertail and Politis (2001), Bertail (1997), when the computer capacities were not sufficient to handle even moderate sample sizes. Such methods are themselves related to well-known numerical methods (see for instance Isaacson and Keller, 1966; Har-Peled, 2011).

Most of these methods based on subsampling rely on an adequate standardization of the statistics of interest. Such standardization may be hard to obtain for complicated procedures including statistical learning procedures. It is even more complicated to extrapolate to very large sample sizes. Indeed, the extrapolation of the distribution of a statistic from smaller scales to a large one requires the knowledge of the rate of convergence τ_n of the procedure of interest or at least an estimator of the latter, with n being the size of the dataset. In many situations, this task is difficult because the rate itself depends on the true generating process of the data.

In this paper, we present a variant of the subsampling distribution estimation methodology studied by Bertail et al. (1999, 2004) to derive a consistent estimator of the rate $\tau_{\cdot}$ and study its rate of convergence. We prove the asymptotic validity of the method to construct asymptotically valid confidence intervals and give some precise rate of convergence, assuming that the centering of the subsampling distribution satisfies some concentration inequality. What differs from previous works is the way the subsampling distribution of the statistic of interest is centered. It allows precise control of the rate of the approximation. In Bertail et al. (1999, 2004), the centering quantity was the statistic of interest computed on the whole data set (which may not be achievable in practice for big data, when the size n is very large), and the associated rate of convergence was of the order of $1/\log n$. Here, the centering quantity is computed as the mean of the statistics of interest obtained on subsamples of size b_n. This yields for subsampling distribution, the expected rate of convergence of order $\sqrt{b_n/n}$ plus a bias term of order (in regular cases) $1/\sqrt{b_n}$ which is the rate at which the true distribution based on a sample of size b_n approaches the asymptotic distribution. In that case, we obtain an optimal rate of $n^{-1/4}$ for the choice $b_n = n^{-1/2}$. Moreover, we show that the mean of the subsampling distribution (eventually approximated by Monte Carlo simulations) yields attractive hyper-efficiency properties for some slow algorithms (with a rate slower than $\sqrt{n}$). This estimator satisfies precisely the required concentration inequality. We then use the extrapolation of this estimation for large datasets to construct confidence intervals for many procedures that are difficult to analyze otherwise. We show how this idea may be used in the case of ma-

chine learning procedures to obtain confidence intervals for general risks. By enabling the construction of confidence intervals for the test error rates or out-of-sample error rates of various algorithms, it becomes easier to compare them. We also show how it is possible to practically integrate the dynamic aspect of the big data environments (especially in the case of streaming data flows).

These subsampling techniques are then implemented with potentially time-consuming procedures (k-nearest neighbors, random forest, neural nets,...) first on simulated data and next on two real databases, the EMNIST dataset, and the VeReMi dataset. The implementation of simulated data is performed with the software R on a standard machine and the implementation of real data illustrations with Python on an Intel(R) Xeon(R) Gold 6154 CPU @ 3.00GHz machine with 144 cores and 250GB RAM. We estimate the rate of convergence of several algorithms and obtain confidence intervals for the out-of-sample risks of several standard algorithms. An interesting by-product of the study is that using different subsampling sizes also allows for detecting the instability of the procedures.

The article is organized as follows. Section 2 presents the state of the art of the subsampling methods and introduces some estimators of the convergence rate of the sample statistic distribution. Section 3 presents the two main results and a discussion on the choice of the subsampling sizes. First, we prove the asymptotic validity of the method to construct asymptotically valid confidence intervals. Then, we also prove that the mean of the randomized subsampling distribution yields attractive hyper-efficiency properties for some slow algorithms. Section 4 presents applications on simulated data and on the VeReMi and EMNIST-digits datasets. Finally, the proofs are deferred to the Appendix section.

2. Subsampling methods for big data

2.1. Definition

In Politis and Romano (1994) (see also Politis et al. (1999) for more a developed framework and references) a general subsampling methodology has been proposed for the construction of large-sample confidence regions for an unknown parameter $\theta = \theta(P) \in \mathbb{R}^q$ under very minimal conditions. Considering $\underline{X}_n = (X_1, ..., X_n)$ an i.i.d. sample, the construction of confidence intervals for θ requires an approximation to the sampling distribution under P, generally unknown, of a standardized statistic $T_n = T_n(\underline{X}_n)$. This statistic is assumed to be consistent for θ at some *known* rate τ_n. For example, in the statistical learning methodology, θ may be the Bayes Risk and T_n the estimated risk linked to a given algorithm (see section 3.3 for illustration). In the framework of prediction, θ may be a value to predict and T_n a predictor.

To fix some notations, assume that there is a non-degenerate asymptotic distribution for the centered re-normalized statistic $\tau_n(T_n - \theta)$, denoted by $K(x, P)$, continuous in x, such that for any real number x,

$$K_n(x, P) \equiv \Pr{}_P\{\tau_n(T_n - \theta) \leq x\} \underset{n\to\infty}{\longrightarrow} K(x, P). \qquad (\mathbf{A1})$$

Recall that, in the framework of resampling techniques, we are interested in estimating the distribution of the original statistics $K_n(x, P)$ or more generally the sequence of distributions $K_m(x, P)$, $m \leq n$. Indeed for big data, the statistic T_n itself may be difficult to compute on the whole database (see the applications to the EMNIST and

reduced EMNIST dataset in part 4.2).
Then the subsampling distribution with subsampling size b_n, is defined by

$$K_{b_n}(x \mid \underline{X}_n, \tau_\cdot) \equiv q^{-1} \sum_{i=1}^{q} 1\{\tau_{b_n}(T_{b_n,i} - T_n) \leq x\}, \tag{1}$$

where $q = \binom{n}{b_n}$ and $T_{b_n,i}$ is a value of the statistic of interest, calculated on a subset of size b_n chosen from $\underline{X}_n$. It was shown in Politis and Romano (1994) that the the subsampling methodology is asymptotically valid. Precisely, under the following assumptions on b_n

$$b_n \underset{n\to\infty}{\longrightarrow} \infty, \qquad \frac{b_n}{n} \underset{n\to\infty}{\longrightarrow} 0, \tag{\textbf{A2}}$$

and

$$\frac{\tau_{b_n}}{\tau_n} \underset{n\to\infty}{\longrightarrow} 0. \tag{\textbf{A3}}$$

we have

$$K_{b_n}(x \mid \underline{X}_n, \tau_\cdot) - K_n(x, P) \underset{n\to\infty}{\longrightarrow} 0,$$

uniformly in x over neighborhoods of continuity points of $K(x, P)$. The same result holds if n is replaced by any sequence m_n and the corresponding subsampling size b_n satisfies conditions (A2) and (A3).
Remark(On conditions **A2-A3**) *The key point for this result is based on the fact that, when T_n is replaced by θ in equation (1), one obtains a U-statistic of degree b_n whose variance is of order $\frac{b_n}{n}$. Then the condition **A2** ensures that the mean of this U-statistic $K_{b_n}(x, P)$ converges to a limiting distribution and that the variance of the U-statistic converges to 0. The condition **A3** allows replacing the true value of the parameter by T_n. In that case, **A3** ensures that this replacement does not affect the limiting distribution. When choosing an adequate re-centering that may differ from T_n (for instance the mean or median of the subsampling distribution), the condition **A3** may be completely dropped as discussed below.*

For large databases, computing $q = \binom{n}{b_n}$ values of the statistics $T_{b_n,i}$ may be unfeasible. In this case, it is recommended to use its Monte-Carlo approximation

$$K_{b_n}^{(B)}(x \mid \underline{X}_n, \tau_\cdot) = B^{-1} \sum_{j=1}^{B} 1\{\tau_{b_n}(T_{b_n,j} - T_n) \leq x\}, \tag{2}$$

where $\{T_{b_n,j}\}_{j=1,\ldots,B}$ are now the values of the statistic calculated on B subsamples of size b_n taken without replacement from the original sample. It can be easily shown by incomplete U-statistics arguments that if B is large then the error induced by the Monte-Carlo step on the subsampling distribution is only of size

$$K_{b_n}^{(B)}(x \mid \underline{X}_n, \tau_\cdot) - K_{b_n}(x \mid \underline{X}_n, \tau_\cdot) = O_P\left(\frac{1}{\sqrt{B}}\right),$$

where the notation $O_P(.)$ refers to stochastic boundedness. We recall that $Z_B = O_P(a_B)$ means that, for any $\varepsilon > 0$, there exists a finite $\delta_\varepsilon > 0$ and a finite $N_\varepsilon > 0$ such that, $\forall B > N_\varepsilon$, $\Pr(|Z_B/a_B| > \delta_\varepsilon) < \varepsilon$.

Then, if the error of $K_{b_n}(x \mid \underline{X}_n, \tau_.)$ on the true distribution is controlled, it is always possible to find a value of B (eventually linked to n) such that the Monte-Carlo approximation is negligible. MacDiarmid's inequality applied to the indicator functions yields a precise control of the error at any probability error level δ.

The subsampling method is generally based on the centering by T_n computed on the whole database. This centering may not be adapted for big data since the calculation of T_n itself may be too complicated: the exact size may be unknown or the complexity of the algorithm and the cost induced by retrieving all the information may be too high. In the subsampling method, the main reason for using the centering by T_n is simply due to the fact that under the condition **A3**, the convergence rate of T_n, τ_n, is faster than τ_{b_n}. Indeed :

$$\begin{aligned}\tau_{b_n}(T_{b_n,j} - T_n) &= \tau_{b_n}(T_{b_n,j} - \theta) + \tau_{b_n}(\theta - T_n) \\ &= \tau_{b_n}(T_{b_n,j} - \theta) + O_P\left(\frac{\tau_{b_n}}{\tau_n}\right) \\ &= \tau_{b_n}(T_{b_n,j} - \theta) + o_P(1).\end{aligned}$$

This suggests using any centering whose convergence rate is faster than τ_{b_n}.

This is, for instance, the case if one constructs a subsampling distribution without any centering or standardization with a subsampling size $m_n >> b_n$ such that $\frac{b_n}{m_n} \to 0$ and $\frac{\tau_{b_n}}{\tau_{m_n}} \to 0$. In this case we have that $\frac{1}{B}\sum_{j=1}^{B} T_{m_n,j}$ which is a proxy of $\frac{1}{q}\sum_{j=1}^{q} T_{m_n,j}$ (with an error of size $1/\sqrt{B}$) converges to θ at a rate at least as fast as τ_{m_n} (provided that the expectation of these quantities exists). The same results hold if one chooses the median rather than the mean of the $T_{m_n,j}$ $'s$ (when considering the mean this amounts to recenter at the median of means) as considered in Laforgue et al. (2019).

2.2. Rate of convergence for subsampling distributions

To our knowledge, the precise rate of convergence of subsampling distributions and their Monte-Carlo approximations has not been investigated except in some very specific cases (and especially the mean). Actually, in general, this requires more precise control of the centering factor and of the modulus of continuity of the asymptotic distribution as well as some control on the rate of approximation to the asymptotic distribution.

In the following, we denote by $\widehat{\theta}_n$ any centering such that

$$\tau_{b_n}(\theta - \widehat{\theta}_n) = o_P(1). \tag{3}$$

For reasons that appear clearly in the proofs and for further applications in statistical learning, we assume that there is a concentration inequality for this estimator of the following form: for some $\eta > 0$, there exists some universal constants $C_i > 0$, $i = 1, 2$ such that

$$\Pr\left(\tau_{b_n}|\widehat{\theta}_n - \theta| > \eta\right) \leq C_1 \exp(-\frac{n}{b_n} C_2 \eta^2). \tag{**A4**}$$

In general, because the right rate of convergence of $\widehat{\theta}_n$ is much more rapid than τ_{b_n}, such concentration inequality holds for regular statistics. This is the case of the mean or the moments for instance, under the existence of some exponential moments. We will show later that the mean of the statistics computed on (all or a sufficiently large number of) subsamples of well-chosen size b_n satisfies such a requirement under minimal variance assumptions (see Laforgue et al. (2019) for other concentration results for the median of means).

As in Götze and Rakauskas (2001), we need to control the (deterministic) convergence rate of the true distribution to the asymptotic distribution (this actually plays the role of the bias in approximating the true distribution). In general, this rate is given by some Berry-Esseen theorems or Edgeworth expansions. For this we assume that, for any n large enough, uniformly in x,

$$K_n(x,P) - K(x,P) = O\left(\frac{1}{n^\beta}\right), \quad \text{for some } \beta > 0. \tag{\textbf{A5}}$$

In regular cases, Berry-Esseen theorems yield typically a rate of approximation $n^{-1/2}$ with $\beta = 1/2$ but for instance, for symmetric statistics (or asymptotically χ^2 distribution), we rather expect $\beta = 1$.

We finally assume that we can locally control the modulus of continuity of the asymptotic distribution at points of continuity $K(.,P)$, by some increasing function null at 0. To simplify we assume that the distribution is Lipschitz in neighborhoods of continuity points (similar results but with different rates can be obtained under Hölder assumptions). There exists $L > 0$, such that, at any point x of continuity of $K(.,P)$ and any y in a neighborhood of x.

$$|K(y,P) - K(x,P)| \leq L|y-x|. \tag{\textbf{A6}}$$

Since most of the time, the asymptotic distribution is differentiable with a bounded derivative, the distribution is Lipschitz and this hypothesis is trivially satisfied.

For simplicity, we use the same notation as before and now define the subsampling distribution with the centering $\widehat{\theta}_n$ as

$$K_{b_n}(x \mid \underline{X}_n, \tau_.) \equiv q^{-1} \sum_{i=1}^{q} 1\{\tau_{b_n}(T_{b_n,i} - \widehat{\theta}_n) \leq x\},$$

and its Monte-Carlo approximation

$$K_{b_n}^{(B)}(x \mid \underline{X}_n, \tau_.) = B^{-1} \sum_{j=1}^{B} 1\{\tau_{b_n}(T_{b_n,j} - \widehat{\theta}_n) \leq x\}.$$

The following result gives a new rate of convergence for $K_{b_n}(x \mid \underline{X}_n, \tau_.)$ and $K_{b_n}^{(B)}(x \mid \underline{X}_n, \tau_.)$ for the centering $\widehat{\theta}_n$. This rate will be denoted by

$$\delta_\beta(n) = \sqrt{\frac{b_n}{n}} + \frac{1}{b_n^\beta}.$$

Theorem 1 *Assume that conditions **A1** to **A6** hold, then we have uniformly over*

the neighborhood of points of continuity of $K(x,P)$ (uniformly over $\mathbb{R}$ if $K(x,P)$ is continuous)

$$K_{b_n}(x \mid \underline{X}_n, \tau_\cdot) - K_{b_n}(x,P) = O_P\left(\sqrt{\frac{b_n}{n}}\right),$$

$$K_{b_n}^{(B)}(x \mid \underline{X}_n, \tau_\cdot) - K_{b_n}(x,P) = O_P\left(\sqrt{\frac{b_n}{n}}\right) + O_P\left(\frac{1}{\sqrt{B}}\right).$$

Moreover, we have

$$K_{b_n}(x \mid \underline{X}_n, \tau_\cdot) - K(x,P) = O_P\left(\delta_\beta(n)\right),$$

and

$$K_{b_n}(x \mid \underline{X}_n, \tau_\cdot) - K_n(x,P) = O_P\left(\delta_\beta(n)\right);$$

if in addition $\frac{n}{b_n} = 0(B)$, then the above hold upon replacing $K_{b_n}(x \mid \underline{X}_n, \tau_\cdot)$ with its Monte-Carlo approximation $K_{b_n}^{(B)}(x \mid \underline{X}_n, \tau_\cdot)$.

This result gives some precise results on how to choose the subsampling size and the number of replications by optimizing the rate of the approximation. Typically for $\beta = 1/2$, we can choose the optimal value $b_n = n^{1/2}$ and $B = n^{1/2}$ which actually will drastically reduce the computation costs in comparison to bootstrap procedures. In this case, $\delta_\beta(n) = n^{-1/4}$. For $\beta = 1$ (the symmetric case), we can choose $b_n = n^{1/3}$ and $B = n^{2/3}$, and in this case, we get a better rate $\delta_\beta(n) = n^{-1/3}$. It can be seen that the smaller the bias in **A5**, the better the approximation, but in this case, the Monte Carlo step will require more computations.

2.3. Estimating the convergence rate

The main drawback of this approach is the knowledge of the standardization (or rate) τ_n. However, this rate may be easily estimated at least when the rate of convergence is of the form $\tau_n = n^\alpha L(n)$ as shown in Bertail et al. (1999)). Here α is an unknown real and L is a normalized slowly varying function, that is, such that $L(1) = 1$ and for any $\lambda > 0$, $\lim_{x\to\infty} \frac{L(\lambda x)}{L(x)} = 1$ (see Bingham et al. (1987)). For simplicity, we will now assume that $\tau_n = n^\alpha$. The general case $\tau_n = n^\alpha L(n)$ may be treated similarly with additional assumptions on the slowly varying function (see Bertail et al. (1999)). In any case, the estimator proposed in Bertail et al. (1999) may be used in our framework. We now propose a simplified approach that is attractive in practice.

First, construct the subsampling without any standardization and denote for simplicity

$$K_{b_n}(x \mid \underline{X}_n) \equiv K_{b_n}(x \mid \underline{X}_n, 1)$$

the subsampling distribution of the root $(T_n - \theta)$. Then we have

$$K_{b_n}(x\, \tau_{b_n}^{-1} \mid \underline{X}_n) = K_{b_n}(x \mid \underline{X}_n, \tau_\cdot) \tag{4}$$

Let denote now $F^{-1}(t)$ the quantile transformation, i.e., $F^{-1}(t) = \inf\{x : F(x) \geq t\}$ for a given distribution F on the real line and a number $t \in (0,1)$. The following Lemma extends Lemma 1 of Bertail and Politis (2001) by providing a rate of convergence for quantiles of subsampling distribution under natural assumptions on the limiting distribution.

Lemma 1. *Assume that $K(x,P)$ is strictly increasing at least in the neighborhood of the quantile of interest $K^{-1}(t,P)$. Then under the conditions of Theorem 1, we have*

$$K_{b_n}^{-1}(t \mid \underline{X}_n, \tau_\cdot) = K^{-1}(t,P) + O_P(\delta_\beta(n)^{-1}).$$

Now, it is easy to see with the rate of convergence obtained in Theorem 1 that we have by Lemma 1,

$$\begin{aligned} K_{b_n}^{-1}(t \mid \underline{X}_n, \tau_\cdot) &= \tau_{b_n}\, K_{b_n}^{-1}(t \mid \underline{X}_n) \qquad (5) \\ &= K^{-1}(t,P)\left(1 + O_P\left(\delta_\beta(n)\right)\right), \qquad (6) \end{aligned}$$

yielding

$$\log\left(|K_{b_n}^{-1}(t \mid \underline{X}_n)|\right) = \log\left(|K^{-1}(t,P)|\right) - \alpha \log(b_n) + O_P\left(\delta_\beta(n)\right).$$

If two different subsampling sizes satisfy the conditions stated before and are such that $b_{n_1} = b_n$, $b_{n_1}/b_{n_2} = e$, then one gets

$$\log\left(|K_{b_{n_1}}^{-1}(t \mid \underline{X}_n)|\right) - \log\left(|K_{b_{n_2}}^{-1}(t \mid \underline{X}_n)|\right) = \alpha + O_P\left(\delta_\beta(n)\right),$$

uniformly in a neighborhood of t. Using sample sizes of the same order avoids the complicated constructions used in Bertail et al. (2004) and suggests that the parameter α may be simply estimated by averaging this quantity over several subsampling distributions. The rate that we obtain here is clearly better than the one obtained in Bertail et al. (2004) : this is due to the concentration hypothesis (**A4**) and hypothesis (**A6**). Again for the regular case $\beta = 1/2$, we can choose the optimal value $b_{n_1} = n^{1/2}$ and obtain a rate for α equal to $n^{-1/4}$ which clearly improves over the $\log(n)^{-1}$ rate obtained in Bertail et al. (2004). Due to the "bias" term (involved by the asymptotic approximation rate in **A5**), this rate can not be improved.

Computing these two subsampling distributions mainly requires the computation of $B \times\ b_{n_1}(1+e)$ values of the statistic of interest (whose computation may be easily distributed). Thus the resampling size should be chosen large enough not to perturb too much the subsampling distributions but sufficiently small so that the cost of computing these quantities is small in comparison to the global cost of computing a single statistic over the whole database. Similarly as before, if we choose $B = n^{1/2}$ then we may replace the true subsampling distribution with the Monte-Carlo approximation without changing the rate. However, for some very large datasets, one may wish to choose a smaller B to have a lower computation cost, at the price of a worse approximation.

Notice that the previously described estimator of the rate parameter α involves the difference of two log-subsample quantiles, which has the drawback of depending on

the subsample centering. Such centering can be removed by considering an estimator that involves the differencing of subsample quantiles before applying logarithms. For any $0 < t_1 < 1/2 < t_2 < 1$, at which the asymptotic distribution is strictly increasing, we have

$$\begin{aligned}&\log\left(K_{b_n}^{-1}(t_2 \mid \underline{X}_n) - K_{b_n}^{-1}(t_1 \mid \underline{X}_n)\right)\\ &= \log\left(K^{-1}(t_2, P) - K^{-1}(t_1, P)\right) - \alpha \log(b_n) + O_P\left(\delta_\beta(n)\right).\end{aligned}$$

For $b_{n_1} = b_n \to \infty$ and $b_{n_2} = b_{n_1}/e$, it follows that

$$\log\left(\frac{K_{b_{n_1}}^{-1}(t_2 \mid \underline{X}_n) - K_{b_{n_1}}^{-1}(t_1 \mid \underline{X}_n)}{K_{b_{n_2}}^{-1}(t_2 \mid \underline{X}_n) - K_{b_{n_2}}^{-1}(t_1 \mid \underline{X}_n)}\right) = \alpha + O_P\left(\delta_\beta(n)\right). \tag{7}$$

By looking simply at two subsampling distributions, it is thus possible to estimate the parameter α at a rate which is at least $O_P\left(\delta_\beta(n)\right)$. One may choose for instance $t_1 = 0.75$ and $t_2 = 0.25$, corresponding to the log of interquartiles.

3. Subsampling with estimated rates : some rate of convergence

3.1. Confidence intervals based on subsampling

For a given estimator of τ_n, typically $\widehat{\tau}_n = n^{\widehat{\alpha}}$, we will use

$$\widehat{K}_n(x, P) = \Pr{}_P\{\ \widehat{\tau}_n(T_n - \theta) \le x\}.$$

Theorem 2 *Assume **A1** holds for $\tau_n = n^\alpha$, for some $\alpha > 0$ and some $K(x, P)$ continuous in x; assume also that assumption **A2** holds. Let $\hat{\alpha} = \alpha + o_P((\log n)^{-1})$, and $\widehat{\tau}_n = n^{\hat{\alpha}}$. Then*

$$\Delta_n = \sup_x |K_{b_n}(x \mid \underline{X}_n, \widehat{\tau}_\cdot) - \widehat{K}_n(x, P)| = o_P(1).$$

Let $\gamma \in (0, 1)$, and let $c_n(1-\gamma) = K_{b_n}^{-1}(1-\gamma \mid \underline{X}_n, \widehat{\tau}_\cdot)$ be the $(1-\gamma)^{th}$ quantile of the subsampling distribution $K_{b_n}(x \mid \underline{X}_n, \widehat{\tau}_\cdot)$. Then

$$\Pr{}_P\{\hat{\tau}_n(T_n - \theta) \ge c_n(1-\gamma)\} \underset{n\to\infty}{\longrightarrow} \gamma. \tag{8}$$

Thus with an asymptotic coverage probability of $1-\gamma$, we have

$$-\hat{\tau}_n^{-1} c_n(1-\gamma) \le \theta - T_n$$

and by symmetry,

$$\theta - T_n \le -\hat{\tau}_n^{-1} c_n(\gamma).$$

*If in addition conditions **A1-A6** hold, if we choose an estimator of α which satisfies*

$$\hat{\alpha} = \alpha + O_P\left(\delta_\beta(n)^{-1}\right)$$

then the rate of the subsampling approximation with the estimated rate becomes

$$\Delta_n = O_P\left(log(n)\delta_\beta(n)^{-1}\right).$$

Remark: Notice that estimating the rate as we did before, only results in a loss of $log(n)$ in the subsampling distribution with an estimated rate. For $\beta = 1/2$ corresponding to smooth parameters, the optimal rate will be $log(n)/n^{1/4}$. For moderate sample sizes, the subsampling approximation in Theorem 2 may be unsatisfactory. But, for very large datasets, it can still lead to an acceptable approximation, as will be seen in our applications.

Recall that $K_{b_n}^{-1}(1-\gamma \mid \underline{X}_n, \widehat{\tau}_\cdot)$ is the $(1-\gamma)^{\text{th}}$ quantile of the rescaled subsampling distribution. Just like in Hall (1986) assume that B is such that $(B+1)\times\gamma$ is an integer (thus if $\gamma = 5\%$ or $\gamma = 1\%$, $B = 999$ or $B = 9999$ is fine), the $(1-\gamma)^{\text{th}}$ quantile is defined uniquely and equal to $\tau_{b_n}(T_{b_n}^{((B+1)(1-\gamma))} - \widehat{\theta}_n)$ where $T_{b_n}^{((B+1)(1-\gamma))}$ is the $(B+1)(1-\gamma)$ largest value over the B subsampled values. It then follows that the lower bound for $\theta - T_n$ is given by

$$\theta - T_n \geq -\frac{\widehat{\tau}_{b_n}}{\hat{\tau}_n}(T_{b_n}^{((B+1)(1-\gamma))} - T_n). \tag{9}$$

A straightforward application of this result is to compare the generalization capability of statistical learning algorithms (see section 3.3) when n is so large that the estimated risk T_n of most algorithms, even with polynomial complexity, may be hardly computed in a reasonable time.

However, in some cases, it may be difficult to compute the statistics T_n on the whole database. The preceding result also allows to build confidence intervals for θ based on a single subsampling realization say $\widehat{\theta}_{m_n} = T_{m_n}(X_{i_1}, ..., X_{i_{m_n}})$ based on a different size m_n such that $n >> m_n >> b_n$. For this, assume that $(B+1)\times\gamma/2$ is an integer. In that case, by combining (8) and (9), a two-sided confidence interval for θ is simply given by

$$\widehat{\theta}_{m_n} - \frac{\widehat{\tau}_{b_n}}{\hat{\tau}_{m_n}}\left(T_{b_n}^{((B+1)(1-\gamma/2))} - \widehat{\theta}_{m_n}\right) \leq \theta \leq \widehat{\theta}_{m_n} - \frac{\widehat{\tau}_{b_n}}{\hat{\tau}_{m_n}}\left(T_{b_n}^{((B+1)\gamma/2)} - \widehat{\theta}_{m_n}\right).$$

This interval suffers of course from lower precision due to information loss in not being based on the whole data set. However, sometimes we can find an estimator $\widehat{\theta}_n$ that can be computed on the whole database. For instance, the mean of a given subsampling distribution may be an adequate candidate in some cases (see part 3.2) with a rate that satisfies conditions (3) and **A4**. In that case, a rescaled confidence interval is simply given by

$$\widehat{\theta}_{n} - \frac{\widehat{\tau}_{b_n}}{\hat{\tau}_n}\left(\widehat{\theta}_{b_n}^{((B+1)(1-\gamma/2))} - \widehat{\theta}_{n}\right) \leq \theta \leq \widehat{\theta}_{n} - \frac{\widehat{\tau}_{b_n}}{\hat{\tau}_n}\left(\widehat{\theta}_{b_n}^{((B+1)\gamma/2)} - \widehat{\theta}_{n}\right).$$

In our simulation studies and our applications, regarding the estimation of risks of a learning algorithm ($\hat{\theta}_n$ will be itself an estimated risk), the variability of the data may be so high that somehow there is very little difference between confidence intervals computed with T_n or a subsampling estimator $\widehat{\theta}_{m_n}$. To explain this phenomenon, recall that in statistical learning, the generalization error (see below) of an algorithm (or of

an estimation procedure) can be decomposed into three terms

- the square of the bias, linked to erroneous assumptions in the learning algorithm (for instance a linear assumption while the model is not linear);
- the variance, measuring the variability of the procedure or underlying estimator;
- the irreducible error of the model, that is, the variance of the intrinsic error.

From a statistical point of view, just consider the target model $Y_i = f(X_i) + \epsilon_i$ for some class of functions $f \in \mathcal{F}$. The total L_2-loss (or generalization error) of a predictor $\hat{f}(x)$ at a point x is

$$E(Y - \hat{f}(x))^2 = (E\hat{f}(x) - f(x))^2 + Var(\hat{f}(x)) + Var(\epsilon)$$

For big data, this last error is generally big due to the great variability of the observations whereas, at the same time, the bias and variance of the estimation procedure are small due to the large dataset.

3.2. Improving rates by means of subsampling estimators

In the particular case when the recentering estimator $\widehat{\theta}_n$ is chosen to be the mean of the subsampling distribution, we show in the following theorem that there is a concentration phenomenon at a speed that is sufficient to ensure conditions (3) and **A4**. This result may be seen as a generalization of the hyper-efficiency results for pooled estimators computed on partitioned data obtained in Banerjee et al. (2019). For slow procedures (i.e., procedures with a convergence rate slower than $\sqrt{n}$), we prove this super-efficiency phenomenon that is, a convergence rate faster than the original rate and close to $\sqrt{n}$, for the mean of subsampling distribution under a simple variance condition.

For simplicity, define the mean estimator $\widehat{\theta}_n^q = q^{-1} \sum_{1 \leq i \leq q} T_{b_n,i}$ where q is either $\binom{n}{b_n}$ or some deterministic B. In that case, the B subsamples are chosen uniformly over all possible subsamples of size b_n.

Notice that, when we consider all subsamples, $\widehat{\theta}_n^q$ is nothing else than a U-statistic with a kernel $T_{b_n}(.)$ of degree b_n . Recall that if T_{b_n} is bounded by some constant M (which will the case in the statistical learning procedures that we will study later) and if the variance exists say $Var(\tau_{b_n} T_{b_n}) \leq C < \infty$, then Hoeffding proved a Bernstein type inequality (see also Arcones (1995)) which becomes here

$$\Pr(|\widehat{\theta}_n^q - E(T_{b_n})| > \varepsilon) \leq 2 \exp\left(- \frac{\frac{n}{b_n} \varepsilon^2 M^2}{2C/\tau_{b_n}^2 + \frac{2}{3} M \varepsilon} \right)$$

yielding (by changing ε into $\varepsilon \sqrt{\frac{b_n}{n \tau_{b_n}^2}}$)

$$\Pr(\sqrt{\frac{n}{b_n}} \tau_{b_n} |\widehat{\theta}_n^q - E(T_{b_n})| > \varepsilon) \leq 2 \exp\left(- \frac{\varepsilon^2 M^2}{2C + \frac{2}{3} M \varepsilon \tau_{b_n} \sqrt{\frac{b_n}{n}}} \right)$$

Thus if b_n is chosen such that $\tau_{b_n} \sqrt{\frac{b_n}{n}}$ is bounded (which will be the case for the choice of b_n controlling the bias and is always true when b_n is chosen very small), then we

will get an Hoeffding bound of type (A4). Moreover by a straightforward inversion of the Bernstein inequality, we have

$$|\widehat{\theta}_n^q - E(T_{b_n})| = O_P(\sqrt{\frac{b_n}{n}}\frac{1}{\tau_{b_n}} + \frac{b_n}{n})$$

which can be better than the rate of convergence of the original statistics τ_n. We state this result under an unbiased condition (to avoid lengthy discussions on the form of the bias).

Theorem 3 *Assume that the statistics of interest T_n is an unbiased bounded estimator of θ has rate $\tau_n \leq \sqrt{n}$, under the assumptions **A1-A2** and assuming that the asymptotic variance is bounded uniformly in n, that is,*

$$Var(\tau_n T_n) \leq C.$$

*Then the rate of convergence of $\widehat{\theta}_n^q$, for $q = \binom{n}{b_n}$ or $q = B$ with $B \geq n/b_n$, is at least $\tau_{b_n}\sqrt{\frac{n}{b_n}}$. Moreover, this estimator satisfies the concentration inequality **A4**.*

Remark *In particular, if $\tau_n = n^{1/2}$ then the subsampling mean estimator $\widehat{\theta}_n^q$ has the same rate $n^{1/2}$. But if $\tau_n = n^\alpha, \alpha < 1/2$ then this estimator has a better rate of convergence given by $n^{1/2}/b_n^{1/2-\alpha}$. We can even choose $b_n = \log(n)$ and thus a rate close to $\sqrt{n}$ asymptotically. Of course, there is no free lunch. In that case, we need more computations of the subsampling estimator (at least $n/\log(n)$ which may be too much in practice). Moreover, it is emphasized in Banerjee et al. (2019) that this estimator may not be locally regular which may be a drawback for some applications where uniformity is needed.*

3.3. Some subsampling results tailored to statistical learning

3.3.1. Subsampling prediction error

Let $D_n = \{(X_i, Y_i),\ i = 1, ..., n\}$ be i.i.d. with distribution P defined on $(\Omega, \mathcal{A}, \mathbf{P})$, taking their values in some measurable product space $\mathcal{X} \times \mathcal{Y}$. The (X_i, Y_i)'s correspond to independent copies of a generic r.v. (X, Y). A predictor is a measurable function $\phi : \mathcal{X} \to \mathcal{Y}$, $x \mapsto y = \phi(x)$. To measure the risk of a predictor, we introduce the loss function $L : \mathcal{Y}^2 \to \mathbb{R}$ which is assumed to be bounded by some constant M. The Bayes classifier is the one obtained by minimizing the expectation of the loss function over all classifiers :

$$\phi^* = \arg\min_{\phi \in \mathcal{F}} E_P L(Y, \phi(X)).$$

However, in the estimation procedure, we will only minimize over a given class of function $\mathcal{F}$ corresponding to a specific algorithm. Moreover, since P is unknown, we will approximate the expected loss by the empirical one and consider the estimator $\widehat{\phi}_n$

defined by

$$\widehat{\phi}_n \underset{def}{=} \widehat{\phi}_n(\mathcal{F}) = \arg\min_{\phi\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \phi(X_i)).$$

One goal of statistical learning is to evaluate the generalization capability of the algorithm measured by the discrepancy between the optimal risk $\theta^* = E_P L(Y, \phi^*(X))$ and the one evaluated on the resulting predictor $\widehat{\phi}_n$, say

$$\Delta_n = E_P(L(Y, \widehat{\phi}_n(X))|D_n) - E_P L(Y, \phi^*(X)).$$

Constructing confidence intervals for this quantity called **prediction error** can allow us to distinguish between different algorithms. In the following, it is assumed that Δ_n (recentered thus at 0) converges asymptotically to a distribution $K(x, P)$ at a rate $\tau_n = \tau_n(P)$, which is clearly unknown in most situations (even when one is able to obtain concentration inequalities). Notice that Δ_n which plays the role of T_n in our first part is not really a statistics in the usual sense since we have taken expectation over (X, Y). However, using subsampling ideas we can still get a subsampling approximation of the distribution of this quantity

Following the subsampling ideas exposed before, for any subsampling set of size b_n, $D_{b_n}^{(j)} = \{(X_i, Y_i), i \in s_{b_n}^{(j)}\}$, with $s_{b_n}^{(j)} \subset \{1, ..., n\}$ for $j = 1, .., q$, we define the subsampling counterpart of $E_P L(Y, \widehat{\phi}_n(X)|D_n)$ by

$$\mathcal{E}_{b_n}^{(j)} = E_P(L(Y, \widehat{\phi}_{b_n}^{(j)}(X))|D_{b_n}^{(j)}),$$

evaluated at the estimator $\widehat{\phi}_{b_n}^{(j)} = \arg\min_{\phi\in\mathcal{F}} \frac{1}{b_n}\sum_{i\in s_{b_n}^{(j)}} L(Y_i, \phi(X_i))$.

Since $\mathcal{E}_{b_n}^{(j)}$ depends on the true distribution, we will estimate it by its empirical version computed on the set $\overline{D}_{b_n}^{(j)} = \{(X_i, Y_i), i \in \overline{s_{b_n}^{(j)}}\}$

$$\widehat{\mathcal{E}_{b_n}^{(j)}} = \frac{1}{n - b_n}\sum_{i\in\overline{s_{b_n}^{(j)}}} L(Y_i, \widehat{\phi}_{b_n}^{(j)}(X_i)).$$

Let's now define $\widehat{\theta}_n^q = \underset{1\le j\le q}{\text{Mean}}(\widehat{\mathcal{E}_{b_n}^{(j)}})$ and the subsampling distribution of the risk, on all $q = \binom{n}{b_n}$ subsamples,

$$K_{b_n}(x \mid \underline{X}_n, \tau_\cdot) = q^{-1}\sum_{j=1}^{q} 1\{\tau_{b_n}(\widehat{\mathcal{E}_{b_n}^{(j)}} - \widehat{\theta}_n^q) \le x\}.$$

As before, we introduce the approximate subsampling distribution based only on B

simulations where now $\widehat{\theta}_n^B = \underset{1\leq j\leq B}{\text{Mean}}(\widehat{\mathcal{E}_{b_n}^{(j)}})$ defined by

$$K_{b_n}^{(B)}(x \mid \underline{X}_n, \tau_.) = B^{-1}\sum_{j=1}^{B} 1\{\tau_{b_n}(\widehat{\mathcal{E}_{b_n}^{(j)}} - \widehat{\theta}_n^B) \leq x\},$$

where the $\widehat{\mathcal{E}_{b_n}^{(j)}}$, $j = 1, ..., B$ are taken at random uniformly on the set of all subsamples. We expect $K_{b_n}^{(B)}(x \mid \underline{X}_n, \tau_.)$ to be an estimator of

$$K_{b_n}(x) = \Pr\left(\tau_{b_n}(\mathcal{E}_{b_n}^{(j)} - \theta^*) \leq x\right),$$

which is itself asymptotically close to the distribution of $\Pr_P(\tau_n\Delta_n \leq x)$.

Then, we apply the same rate estimation procedure as before to compute an estimator of the convergence rate $\tau_.$, say $\widehat{\tau}_.$. Applying the same arguments as in Theorem 2 yields the following result.

Corollary 1 *Assume that:*

$$K_{b_n}^{(B)}(x \mid \underline{X}_n, \tau_.) \xrightarrow[n\to\infty]{\Pr} K(x, P).$$

Moreover, under the same hypotheses as in Theorem 2, we have, with an estimated rate of convergence,

$$K_{b_n}^{(B)}(x \mid \underline{X}_n, \widehat{\tau}_.) \xrightarrow[n\to\infty]{\Pr} K(x, P)$$

yielding a confidence interval for Δ_n of level $1-\gamma$ given by

$$\hat{\tau}_n^{-1}c_n(\gamma/2) \leq \Delta_n \leq \hat{\tau}_n^{-1}c_n(1-\gamma/2)$$

where $c_n(t)$ is the quantile of order t of the distribution $K_{b_n}^{(B)}(x \mid \underline{X}_n, \widehat{\tau}_.)$.

Remark *By the same arguments as in the proof of corollary 1, $\widehat{\theta}_n^q = \underset{1\leq i\leq q}{Mean}(\widehat{\mathcal{E}_{b_n}^{(i)}})$ satisfies a concentration inequality around the true risk, provided that the variance of the risk estimator has an asymptotic variance uniformly bounded. This is automatically satisfied since the loss function L is assumed to be bounded. The problem reduces to obtaining a concentration result for*

$$K_{b_n}^{(B)}(x\ ) = B^{-1}\sum_{j=1}^{B} 1\{\tau_{b_n}(\mathcal{E}_{b_n}^{(j)} - \theta^*) \leq x\},$$

which is simply a U-statistic of degree b_n.

3.3.2. Subsampling out-of-sample error or generalization error

It is known in statistical theory that conditional risk or measurement of prediction error may be too optimistic and could lead to the choice of an algorithm that overfits the data. For this reason, it is often recommended to split the data (of size n) into two parts, the training set called $D_{n,Tr}$ of size $n_{Train} = p_0 n$, for some $p_0 > 1/2$ and the testing set say $D_{n,Test}$ of size $n_{Test} = (1-p_0)n$. To avoid unnecessary problems and truncation, we assume that these sizes are integers. In our simulations we will choose $p_0 = 0.7$. The training and test sets are generally selected at random (practically using first a random permutation of the data and selecting the n_{Train} first units for the training set). In this framework, we are interested in the **out-sample error** estimator or **generalization error** given by

$$\mathcal{T}_n = \frac{1}{(1-p_0)n} \sum_{j \in D_{n,Test}} L(Y_j, \widehat{\phi}_{n_{Train}}(X_j)))$$

where now $\widehat{\phi}_{n_{Train}}$ is determined on the training set $D_{n,Tr}$ that is

$$\widehat{\phi}_{n_{Train}} = \arg\min_{\phi \in \mathcal{F}} \frac{1}{p_0 n} \sum_{i \in D_{n,Tr}} L(Y_i, \phi(X_i)).$$

$\mathcal{T}_n$ may be interpreted as an estimator of the predictive performance of the algorithm. For a convergent algorithm, that is, an algorithm which yields a loss asymptotically close to the Bayesian risk, we expect this quantity to converge to the unconditional full risk $\theta = E_P L(Y_i, \phi^*(X_i))$. In this framework, assuming that $\mathcal{T}_n - \theta$ is converging asymptotically to a non-degenerate distribution, we can apply straightforwardly the results of parts 2 and 3. Indeed if we choose a subsample $D_{b_n}^{(k)}$ size b_n for some $k = 1, ..., \begin{pmatrix} n \\ b_n \end{pmatrix}$, we can divide as well this set into a training set $D_{b_n,Tr}^{(k)}$ of size $p_0 b_n$ and a test set $D_{b_n,Test}^{(k)}$ of size $(1-p_0)b_n$ to construct the corresponding estimator

$$\mathcal{T}_{b_n}^{(k)} = \frac{1}{(1-p_0)b_n} \sum_{j \in D_{b_n,Test}^{(k)}} L(Y_j, \widehat{\phi}_{b_n Train}^{(k)}(X_j)))$$

where now $\widehat{\phi}_{b_n Train}^{(k)}$ is given by

$$\widehat{\phi}_{n_{Train}} = \arg\min_{\phi \in \mathcal{F}} \frac{1}{p_0 n} \sum_{i \in D_{b_n,Tr}} L(Y_i, \phi(X_i)).$$

The subsampling distribution is then the usual one and we can obtain confidence intervals for θ but also improve the convergence rate of $\mathcal{T}_n$ when it is not converging at rate $\sqrt{n}$.

3.4. How to choose the "optimal" subsampling sizes

The choice of the subsampling size is a delicate subject that has been discussed in very few papers including Bickel and Sakov (2008); Götze and Rakauskas (2001); Bickel et al. (2010); Politis et al. (1999) Chap.9. Our preceding results were essentially

asymptotic. For instance, for $\beta = 1/2$ the optimal choice is of the form $b_n = C\sqrt{n}$ for some constant C. But in practice the choice of the constant is crucial... The main idea underlying most propositions is to construct several subsampling distributions by using two different subsampling sizes say b_n and kb_n for $k \in]0,1[$ (we recommend due to our preceding results $k = 1/e$). It is easy to see that when the subsampling distribution is a convergent estimator of the true distribution then the distance d between the subsampling distribution and the true one is stochastically equivalent to $d(K_{b_n}, K_{kb_n})$.

The idea is then to find the largest b_n, which minimizes this quantity. Several distances (Kolmogorov distance, Wasserstein metrics, etc...) may be used.

Of course, for large datasets, such a method is very computationally expensive. We recommend only choosing a limited range of values for b_n and discretizing this range so as to compute the distance $d(K_{b_n}, K_{kb_n})$ only on a limited number of points and to select the ones which minimize this quantity.

Other empirical approaches have been proposed to deal with the problem of the high volatility of subsampling distributions for too large subsampling sizes (see for instance Politis et al. (1999) chap. 9). The idea is simply to look at the quantiles of subsampling distributions and to find the largest value such that the quantile remains stable. Some empirical arguments to understand the principle underlying this idea are given in Bertail (2011). Indeed a subsampling distribution may simply be seen as a U-statistic but with varying kernel of increasing size b_n and its quantiles the inverse of a U-process. The main tools for studying the behavior of subsampling distribution are Hoeffding decomposition of the U-statistic and empirical process theory as considered in Arcones and Gine (1993); Heilig and Nolan (2001). The difficulty in choosing the subsampling size is that, in comparison to U-statistics with a fixed degree, the linear part of the U-statistic is not always the dominating part in the Hoeffding decomposition. For rather small or moderate b_n, it can be shown that the U-statistic is asymptotically normal with a convergence rate of order $\sqrt{\frac{n}{b_n}}$. However, when b_n becomes too large, the remainder in the Hoeffding decomposition dominates and the U-statistic behaves very erratically. This explains why the quantiles of subsampling distributions behave erratically as soon as b_n is too large.

3.5. Subsampling in a growing environment

In many fields, data are now collected online. As a consequence, the size of the database may evolve quickly in time. Then, one may wish to update previous estimations without accessing the whole database again. How is it possible to update the subsampling distributions taking into account the new incoming data, when the size of the database is large and increases? To solve this problem, we present a very simple sequential algorithm.

The idea is as follows: assume that at time t, we have obtained a subsample without replacement of size b_n (uniformly) from the original data. That is the probability of a given subsample is $\binom{n}{b_n}^{-1}$. At the time $t+1$, the new sample size is $n+1$. Then for this newcomer, proceed as follows:

- Draw a Bernoulli r.v. Z with parameter $1 - b_n/(n+1)$
- keep the original subsample if $Z = 1$, with probability $1 - b_n/(n+1)$,
- else with probability $b_n/(n+1)$, choose one element of the current subsample (without replacement, uniformly with probability $1/b_n$) and replace it with this

newcomer.

If several newcomers arrive at the same time, then use sequentially the same algorithm by increasing the size of the population. Notice that this algorithm may be easily implemented sequentially to update all the subsamples already obtained by Monte Carlo simulations at some given time.

The arguments below show that the resulting algorithm is the realization of subsampling without replacement from the total new population.

It may be simply proved by recurrence (McLeod and Bellhouse, 1983). Indeed, assume that the probability of the original sample is $\binom{n}{b_n}^{-1}$ then

- if $Z = 1$, the probability of the new sample is $\binom{n}{b_n}^{-1} \times (1 - \frac{b_n}{n+1}) = \binom{n+1}{b_n}^{-1}$
- if $Z = 0$, the probability of the new sample is $\binom{n}{b_n}^{-1} \times \frac{b_n}{n+1}(\frac{1}{b_n} + \frac{n-b_n}{b_n}) = \binom{n+1}{b_n}^{-1}$.

It follows that the corresponding subsample at any step is actually a subsample obtained without replacement from the total current population.

If we want to increase the size of the subsample, starting from a subsample of size b_n in a population of size n then we simply draw uniformly without replacement from the $n - b_n$ remaining outside sample, observations with probability $1/(n - b_n)$. It may be sometimes easier (for instance using Apache Spark) to use sampling with replacement. It is known in that case that when b_n is small enough such that $\frac{b_n}{\sqrt{n}} \to 0$, then the probability to draw the same individual twice converges to 0. Indeed, when $\frac{b_n}{\sqrt{n}} \to 0$, by Stirling formula, we have $\frac{\binom{n}{b_n}}{n^{b_n}} \to 1$, so that with and without replacement samplings are asymptotically equivalent under this condition.

4. Some empirical results

4.1. Simulation results

In this simulation section, the implementations were executed under R on a standard PC with a 5GHz Intel processor and 2G of Ram. The purpose is to show that we might gain a lot in terms of computation times for a large database and that using our confidence intervals for risks in a simple pattern recognition framework makes sense.

4.1.1. Maximum likelihood estimation for a simple logistic model

The purpose of this example is to explore the feasibility and the computer performance of the procedures described before in an estimation framework. We consider here a very simple toy model to highlight some inherent difficulties with subsampling. Consider a linear logistic regression model with parameter $\theta = (\beta_0,\ \beta) \in \mathbb{R} \times \mathbb{R}^d$. Let X be a d-dimensional marginal vector of the input random variables. The linear logistic regression model related to a pair (X, Y) can be written as

$$\mathbb{P}_\theta\{Y = +1 \mid X\} = \frac{\exp(\beta_0 + \beta^T X)}{1 + \exp(\beta_0 + \beta^T X)}.$$

In high-dimension *i.e.*, when d is very large and for very large n, the computation of the full parametric maximum likelihood estimator (MLE) of θ may be difficult to

obtain in a reasonable time. We assume that $d << n$ but also that the subsampling sizes which will be used are such that $d << b_n$.

For unbalanced populations (a lot of 1's in comparison with 0's and vice versa), the probability to get a subsample with only unit values (or zeros) is high and the MLE will not be convergent (a similar problem appears if the labels are fully separated). This is by no means contradictory with the asymptotic validity of subsampling in this case: it has been shown in Le Cam (1990) that the true variance of the MLE in a finite population is $+\infty$. Subsampling simply reproduces this fact on a smaller scale. In that case, one should condition on the fact that the ratio of the numbers of 1's by the number of 0's is not too small (or not too close to 1). Else, the subsample should be eliminated. We fix this ratio to 3% in our simulations.

In the following, we simulate the toy logistic model

$$Y_i = \begin{cases} 1 & \text{if } 3X_i + \varepsilon_i > 0 \\ 0 & \text{else} \end{cases}$$

with $X_i \sim N(0,1)$ and ε_i independent logistic random variables with mean 0 and variance 1. We choose respectively $n = 10^6$ and $n = 10^7$. We present the results for two randomly generated data set of large size, n (with one regressor variable here) and Table 1 shows the averages and sample variances of $B = 999$ subsample copies of the MLE (i.e., for fitting a logistic model) based on subsample draws of size b_n and b_n/e (when estimating the rate of convergence). In this example, the statistics T_n to which we apply our results is thus the MLE of the parameter β

Even for reasonable sizes, our estimation procedure proved to be useful. For instance in R, with 1 GB of memory, the usual libraries (sampleSelection, glm) fail to estimate the model with a size of $n = 10^7$ observations (for capacity reasons), whereas it takes only $12s$ to get subsampling based confidence bound with $B = 999$ replications of the procedure and $b_n = n^{1/3}$. Here, it is not required to estimate the rate of convergence since the rate $\tau_n = n^{1/2}$ is known, but we did it and present the result with an estimated rate with $J = 2$ subsampling distributions as proposed in this paper, one with size b_n and the other with size b_n/e. If we estimate the rate of convergence with $J = 29$ subsampling distributions based on subsampling sizes equal to $n^{1/3+j/(3(J-1))}, j = 0, \ldots, 28$, as done in the paper Bertail et al. (1999), then one gets quite similar results in terms of variance of the MLE but with 999×29 simulations : it then takes around 14 times more time to complete these tasks on the same computer. We notice that the mean trick (pooled estimator) does not improve the MLE theoretically but allows a quicker and more feasible computation of the recentering factor.

The mean of the estimations of β (and the variances) over the $B = 999$ repetitions with the subsampling procedure are given in Table 1 for different values of b_n $n^{1/3}, n^{1/2}, n^{2/3}$ given in the table and on the whole sample with the corresponding total execution times. $\hat{var}^{1/2}$ gives a rescaled estimator of the variance (to compare with the estimator of the variance on the whole database). That is, we compute the variance of the subsampling distribution say $s^2_{b_n,n}$ and then extrapolate the variance to the original size by computing $\tau_{b_n} s^2_{b_n,n}/\tau_n$. When the rate is unknown the same trick is applied with an estimated rate. The estimated variance $\hat{var}(\beta_n)^{1/2}$) is here obtained using the Hessian of the likelihood function at the MLE using the glm function: 10^7 is the order of the largest size for which we have been able to obtain estimator of this quantity in R with this simple model.

Table 1. MLE variance estimations for a logistic model with $n = 10^6, 10^7$.

	subsample ($B = 999$ replications)				whole sample
n	b_n	$\overline{\hat{\beta}_{b_n}}$	$\hat{var}^{1/2}$	time	$\hat{\beta}_n$ ($\hat{var}(\beta_n)^{1/2}$) time
10^6	$n^{1/3} \approx 100$	3.19	0.0064	13 s	2.992
	$n^{1/2} \approx 1000$	3.022	0.0063	36 s	(0.0061)
	$n^{2/3} \approx 10000$	2.996	0.0060	3.26 mn	28.75 s
10^7	$n^{1/3} \approx 215$	3.10	0.0020	41 s	2.998
	$n^{1/2} \approx 3162$	3.009	0.0020	1.25 mn	(0.0019)
	$n^{2/3} \approx 46415$	2.998	0.0019	12 mn	4.69 mn

Notice that even with a size of $n^{1/3}$, we were able to get the correct order for the variance. In estimating $\beta = 3$, the bias in Table 1 can be large for small subsampling sizes but almost vanishes for $n^{2/3}$. For this size, we get the same order as the one on the MLE on the whole database : but in terms of computation, $n^{2/3}$ is already too big, since in that case we were able to proceed with the MLE on the whole database in less than 5 minutes (whereas it takes 12 minutes to replicate 999 times the procedure on the $n^{2/3}$ sample size). But for $n^{1/2}$, we get a gain of 4 (and 12 for $n^{1/3}$) for a similar accuracy: of course, this strongly depends on the degree of accuracy that one wishes to obtain on the parameter of interest and on the capacity of the computer.

4.1.2. Estimation of the out-of-sample error with k-nearest-neighbor algorithm.

Consider a simple pattern recognition framework. Assume that $\{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ is a sample of i.i.d. random pairs taking their values in some measurable product space $\mathcal{X} \times \{-1, +1\}$. In this standard binary classification framework, the multidimensional r.v.'s X are used to predict the binary label Y. The distribution P can also be described by the pair (F, η) where $F(dx)$ denotes the marginal distribution of the input variable X and $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$, $x \in \mathcal{X}$, is the *conditional distribution*. The goal is to build a measurable classifier $\phi : \mathcal{X} \mapsto \{-1, +1\}$ with minimum risk defined by

$$L(Y, \phi(X)) \stackrel{def}{=} \mathbb{I}\{\phi(X) \neq Y\}, \tag{10}$$

where $\mathbb{I}\{.\}$ is the indicator function. It is well-known that the *Bayes classifier* $\phi^*(x) = 2\mathbb{I}\{\eta(x) > 1/2\} - 1$ is a solution of the risk minimization problem over the collection of all classifiers defined on the input space $\mathcal{X}$.

It is now possible to apply the different subsampling procedures to different classes of functions (or algorithms) to estimate their prediction capability as described in part 3.3. We focus here on the out-of-sample errors. In this simulation part, we propose the use of $\mathcal{F}_1$: a parametric logit model, $\mathcal{F}_2$: the k-nearest neighbor method. We will also consider random forest models, SVM and neural networks in the case studies of part 4.2.

It is known that, under some regularity assumptions, the four methods are consistent so that, asymptotically, the approximation error Δ_n converges to 0 a.s.. For $\mathcal{F}_1$, it is known that the rate of convergence of the empirical risk is $\sqrt{n}$; however for the other algorithms, even if some bounds exist on the generalization capability, the rate of convergence is not clear. Our method will allow us to evaluate the different algorithms and the rate of convergence of the algorithms.

Considering the same simulated data as in 4.1.1, we now use the subsampling method to estimate the out-of-sample errors of k-nearest neighbor (KNN) method and logit models on several subsampling sizes and compare them to that obtained on the full database. We consider a training set equal to $0.7n$ and a test set of size $0.3n$ (similar results have been obtained for other test sets). The computation times in Table 2 clearly show the computation gains. A striking result is for $n = 10^7$ because it takes almost 5 hours to get an estimator of this quantity on the whole sample whereas the subsampling method takes (at worst) 15 minutes with $n^{2/3}$. Table 2 presents the average of out-of-sample errors (over B=999 subsamples) of the subsampling distribution. It seems that even with a size of order $n^{1/3}$ we still get a good approximation in less than 45 seconds. With the subsampling method by using an extrapolated variance as described in 4.1.1, we are also able to estimate the variance of the out-of-sample error (given in parenthesis in the table over the estimation on the whole sample).

Table 2. Estimation of the out-of-sample error by subsampling and on the whole sample - KNN model

KNN	subsample ($B = 999$ replications)			whole sample	
n	b_n	out-of-samp. error	time	out-of-samp. err	time
10^6	$n^{1/3}$	0.1177	4.79 s	0.1158	5.252 mn
	$n^{1/2}$	0.1165	5.76 s	(0.008)	
	$n^{2/3}$	0.1167	43.5 s		
10^7	$n^{1/3}$	0.1166	44.7 s	0.1141	4h57mn
	$n^{1/2}$	0.1163	50.7 s	(0.006)	
	$n^{2/3}$	0.1161	15.35 mn		

To evaluate the performance of our confidence intervals, we have repeated the preceding procedure M=500 times. First, for independent realizations (L=999), we evaluate the risk over the whole dataset and average this risk to have an idea of the true risk (average in Table 3). Then for m=1 to M=500 different samples, we construct confidence intervals for the risks of two methods: the one associated with the logistic regression and the one associated with KNN. We give in Table 3 the average of the lower (Lower) and upper (Upper) bounds of the confidence intervals. The coverage (cover) probability of our intervals is obtained by computing the number of times the "true risk" associated with the procedure belongs to the confidence intervals.

Table 3 shows that the intervals constructed with an interpolated variance using the average as re-centering factor is quite large in terms of errors even for quite large sample sizes and seem to be a little bit too conservative. However in this example we obtain narrower and clearly more accurate results for the Logit model which seems reasonable, since the data itself is generated in this true model.

Table 3. Risk, confidence intervals and coverage probability: KNN and logit models

Method	*KNN*				*Logit*			
size	*Lower*	*Upper*	*average*	*cover*	*Lower*	*Upper*	*average*	*cover*
$n = 10^6$	0.032	0.207	0.118	0.98	0.012	0.110	0.066	0.96
$n = 10^7$	0.046	0.185	0.115	0.96	0.017	0.108	0.061	0.96

These simulations show that it is possible to compare in a reasonable time the out-of-sample errors for several competing methods (with confidence intervals). We should mention that the computational time is not reasonable for neural networks because

they require a long time to be trained and the parameters should be well-calibrated for each simulation. Despite this problem we also have implemented also this method in the case studies below.

4.2. Two case studies on real data sets

In this section, the implementation is performed in Python using the libraries NumPy, SciPy, Sklearn, Tensorflow, on an Intel(R) Xeon(R) Gold 6154 CPU @ 3.00GHz machine with 144 cores and 250GB RAM.

4.2.1. Tested models

Subsampling techniques are implemented on potentially time-consuming procedures (Decision Tree **DT**(Breiman et al., 1984), Random Forest **RF**(Breiman, 2001), Support Vector Machine **SVM**(Chang and Lin, 2011), Neural nets (3 types: **NeuralNet** which is a fully connected multi-layer perceptron with one hidden layer (Rumelhart et al., 1986), **NeuralNet3** which is a deeper multi-layer perceptron with three hidden layers, **ConvNet** (LeCun et al., 1989) which is a special architecture of a neural net that takes account of the hierarchical pattern in data and is commonly used in computer vision), a logit model **Logit**(McCullagh and Nelder, 1983), and K-nearest neighbors **KNN**.

Note that ConvNet is only used on MNIST data set as it is mainly used on image processing and that no hierarchical pattern exists between the features of VeReMi dataset.

The hyperparameters of all tested models were not specifically optimized to do the task on both data sets. We kept the default values found in sklearn that are based on recommendations from the original authors.

4.2.2. Description of the datasets

Vehicular Reference Misbehavior (VeReMi)

The VeReMi (Vehicular Reference Misbehavior) data set (see van der Heijden et al. (2018)) contains data about the detection of misbehavior in vehicular networks, a particularly sensitive topic in cooperative autonomous driving. The purpose of this application is to compare the relative risks of several algorithms and their confidence intervals. The VeReMi data set comprises $N = 424,810$ individuals and only 12 variables.

EMNIST digits

The well-known EMNIST data set studied for instance in LeCun et al. (1995) contains binary images of handwritten digits and the corresponding true digits; the well-known purpose is to find an algorithm that correctly recognizes the digit, see Cohen et al. (2017). The EMNIST data set comprises $N = 240,000$ images and 784 variables (28×28 pixels for each image).

A smaller version of this database is the MNIST digits popularized by Lecun. The data set comprises $N = 60,000$ images. This version of the database is used to check whether the estimation of the convergence rate is similar when based on a smaller data set. We refer to this database as "Lecun MNIST data set".

4.2.3. Comparison of the performances of the tested models

Computing Out-of-sample errors

Figures 1, 2 and 3 provide the estimation of the out-of-sample errors of several models with 90% confidence intervals as well as the computing times for each of them, according to subsampling sizes, for the VeReMi and EMNIST digits data sets. The methodology is described in sections 3.1 and 3.2: B is set to 999, b_n is ranging from $N^{1/3}$ to $N^{2/3}$ similarly to what we did in the previous section, with the estimation of $J = 29$ subsampling distributions. For VeReMi, we observe the superiority of the two tree-based approaches (RF and DT), and the lower performance of the KNN approach in terms of errors, but DT would clearly be preferred as its computing time does not explode as that of RF.

For the EMNIST digit, most confidence intervals overlap for subsampling sizes up to $b_n = 4,000$ although ConvNet and SVM show the lowest errors. From this error plot, we can not conclude that their superiority is significant. In terms of computing times, SVM would clearly be preferred to ConvNet as the computing time for the latter is greater than 6 hours for $b_n = 4,000$.

For Lecun MNIST data set, only 6 models were compared, and SVM and RF have the best performances in terms of errors, with SVM showing lower computing times. Note that the computing times on this data set 4 times smaller are also 4 times better.

The RF, NeuralNet and SVM models have a higher execution time than other models for all subsampling sizes. In particular, for the VeReMi data set where the number of variables is very small, NeuralNet has a relatively smaller execution time compared to Random Forests due to the reduced number of parameters in the network. This changes for the EMNIST digit, where the number of features is bigger and NeuralNet takes longer to train. The simpler models KNN, DT, and Logit take relatively smaller execution times. Overall all execution times increase with subsampling sizes with different rates for each model.

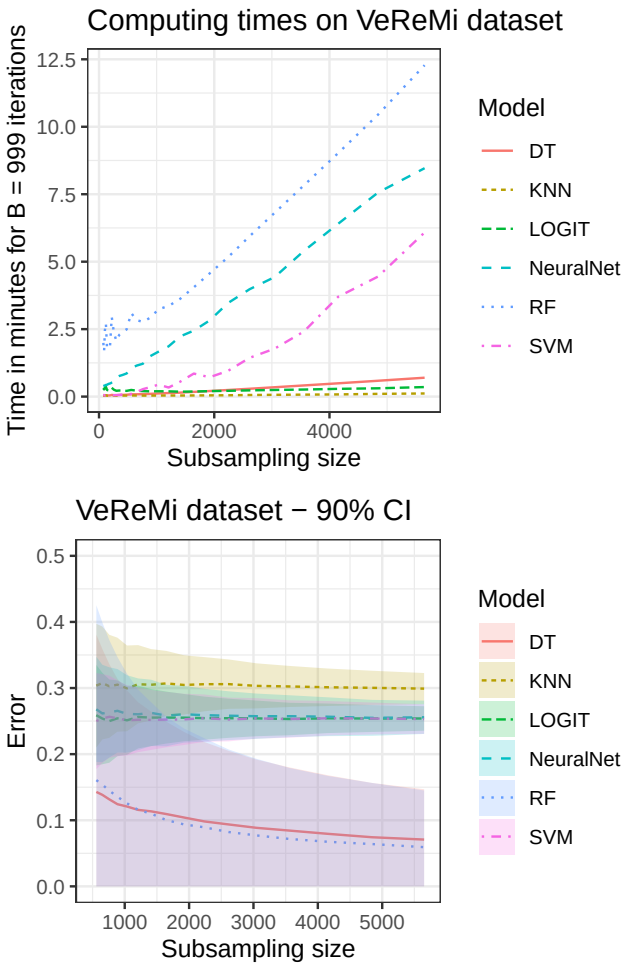

Figure 1. Comparison of Out-of-sample errors and their associated computing times from 6 different models according to the subsampling size, VeReMi data set.
For errors, 90% confidence intervals are provided.

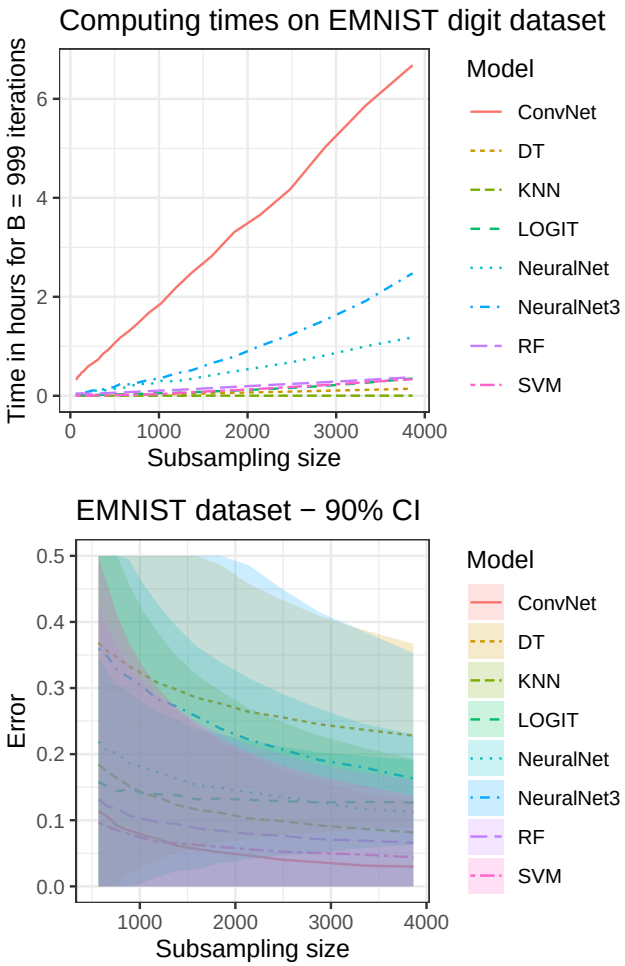

Figure 2. Comparison of Out-of-sample errors and their associated computing times from 8 different models according to the subsampling size, EMNIST digit data set.
For errors, 90% confidence intervals are provided.

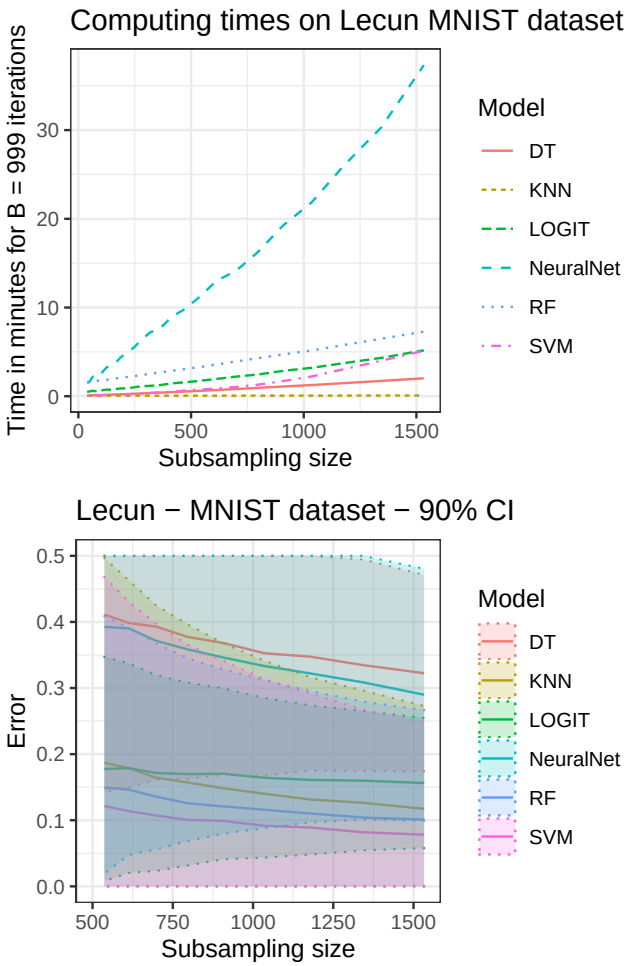

Figure 3. Comparison of Out-of-sample errors and their associated computing times from 6 different models according to the subsampling size. Lecun MNIST dataset.
For errors, 90% confidence intervals are provided.

Figures 4 shows the extrapolation to the full dataset sizes of the out-of-sample errors. It requires the computation of the estimated convergence rate presented in the next section, that is, the computation of the $J = 29$ subsampling distributions (the results with only $J = 2$ are quite similar).

For VeReMi, RF, and DT perform the best, similarly to each other. For EMNIST, Convnet clearly outperforms the other models, SVM is next. For Lecun MNIST, SVM, and RF do not differ much and we can see that NeuralNet behaves better on the $N = 240,000$ EMNIST data set than the $60,000$ Lecun MNIST data set.

In comparison with the Out-of-sample errors on the full dataset (Table 4, 5), the order of best to worst performing methods matches the one with the extrapolated error. Except for KNN on the VeREMi data set where the introduction of the full data helped achieve much better performance.

Estimation of the convergence rates

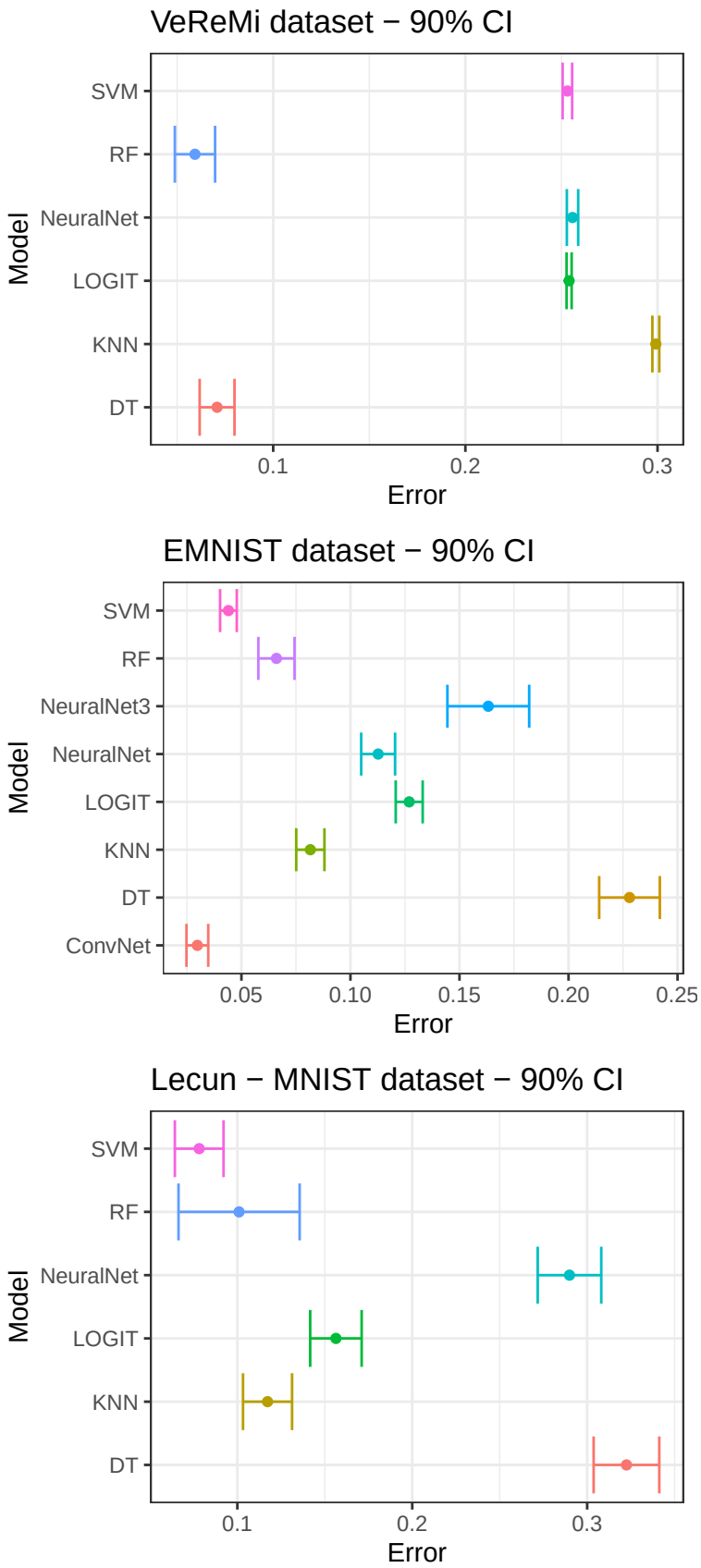

Figure 4. Comparison of Out-of-sample errors from 6 or 8 different models, with 90% confidence intervals, extrapolated to the full dataset size.
For VeReMi $N = 424,810$, for EMNIST digit, $N = 240,000$, for Lecun MNIST, $N = 60,000$

Table 4. Summary statistics of Out-of-sample errors from different models on the full EMNIST data set size

	Logit	RF	SVM	DT	KNN	NeuralNet
count	10	10	10	10	10	10
mean	0.062435	0.019450	0.011373	0.080225	0.018935	0.028037
std	0.000742	0.000503	0.000431	0.001078	0.000345	0.000926
min	0.061417	0.019000	0.010604	0.078000	0.018542	0.026792
10%	0.061623	0.019056	0.010792	0.078844	0.018617	0.027110
50%	0.062375	0.019208	0.011385	0.080719	0.018844	0.028010
90%	0.063517	0.020296	0.011827	0.081125	0.019500	0.028950
max	0.063667	0.020333	0.011958	0.081125	0.019500	0.029812

Table 5. Summary statistics of Out-of-sample errors from different models on the full VeReMi dataset size

	logit	RF	SVM	DT	Knn	NeuralNet
count	10	10	10	10	10	10
mean	0.254134	0.000194	0.254134	0.001743	0.030807	0.254228
std	0.001391	0.000050	0.001391	0.000208	0.000526	0.001340
min	0.251442	0.000141	0.251442	0.001565	0.029849	0.251677
10%	0.252957	0.000152	0.252957	0.001587	0.029933	0.252948
50%	0.253878	0.000177	0.253878	0.001671	0.030920	0.254172
90%	0.255643	0.000262	0.255643	0.002120	0.031286	0.255632
max	0.255738	0.000294	0.255738	0.002130	0.031402	0.255738

We estimate here the rate of convergence τ_n as n^α using 2 and 29 different values of subsampling size: in terms of estimation, the rate is of the same order. See figures 5 to 11 to understand why: just pick two points (not too close) among all subsampling sizes and it somehow gives the right slope. However, it seems that using more subsampling distributions as in Bertail et al. (2004) gives more precise results (we thus only give these estimators in Table 6 and 7). As described at the end of section 2.3, we consider here a regression of log range, on the log of the subsampling size. We propose 3 different ranges, either the inter-quartile-range (percentiles at 25% and 75%) denoted IQR, or an inter-percentile range with length 80% denoted IPR80 (percentiles at 10% and 90%), or with length 90% denoted IPR90 (percentiles at 5% and 95%). Individual graphs showing the slope α of each model/IPR choice are postponed to the appendix (see Figures B1,B2,B3,B4,B5,B6,B7) and results are summarized in Tables 6 and 7. For both data sets, we observe that the results are quite stable when changing the IPR approach as can be seen from the figures in the appendix, or using only 2 subsampling distributions. For VeReMi, we observe that RF and DT have faster convergence rates ($\alpha \approx$-.65) than the 4 other models ($\alpha \approx -1/2$). These models also reached the lowest Out-of-sample errors. For EMNIST, we tested two more models: a Neural Net with 3 layers and a Convolution Net to see how the rates of convergence would be impacted. We observe that here, Logit, DT and NeuralNet3 models have $\alpha \approx -1/2$ while KNN and RF have $\alpha \approx -2/3$, and SVM and ConvNet have $\alpha \approx -3/4$, NeuralNet being closer to $\alpha \approx -.6$. Again, the models with faster convergence rates are also those with lower out-of-sample errors. However, we see that the convergence rates do depend on the learning problem as we observe for instance that DT is very efficient in the VeReMi case (12 features) and not in the EMNIST case (784 features).

Table 6. Estimation of the convergence rates of the 6 different models, with 3 different methods - the VeReMi data set

Model	IQR	IPR80	IPR90
NeuralNet	-0.487	-0.500	-0.499
Logit	-0.493	-0.513	-0.498
KNN	-0.497	-0.506	-0.493
SVM	-0.510	-0.494	-0.504
RF	-0.595	-0.636	-0.635
DT	-0.611	-0.624	-0.630

Table 7. Estimation of the convergence rates of the 8 different models, with 3 different methods - the EMNIST data set

Model	IQR	IPR80	IPR90
Logit	-0.558	-0.560	-0.562
NeuralNet3	-0.560	-0.514	-0.525
DT	-0.567	-0.542	-0.533
NeuralNet	-0.591	-0.587	-0.578
KNN	-0.657	-0.655	-0.673
RF	-0.688	-0.675	-0.687
ConvNet	-0.722	-0.753	-0.750
SVM	-0.773	-0.754	-0.757

Table 8 also compares the EMNIST IQR results with the Lecun MNIST IQR results. The smaller dataset shows somehow lower rates than the larger one.

Table 8. Estimation of the convergence rates of the 6 different models (IQR method), comparison of EMNIST data set ($N = 240,000$) and Lecun MNIST data set ($N = 60,000$)

Model	EMNIST	Lecun MNIST
Logit	-0.558	-0.564
RF	-0.688	-0.656
DT	-0.567	-0.518
KNN	-0.657	-0.641
NeuralNet	-0.591	-0.427
SVM	-0.773	-0.682

Appendix A. Appendix A

This appendix is dedicated to the proof of the theorems and the corollary of this article.

A.1. Proof of Theorem 1 from section 2.2

Introduce the U-statistic

$$V_{b_n}(x) = q^{-1} \sum_{i=1}^{q} 1\{\tau_{b_n}(T_{b_n,i} - \theta) \leq x\}.$$

Then, we have the simple decomposition

$$\begin{aligned}
&\Pr\left(|K_{b_n}(x \mid \underline{X}_n, \tau_\cdot) - K_{b_n}(x, P)| > \varepsilon\right) \\
&\leq \Pr(|K_{b_n}(x \mid \underline{X}_n, \tau_\cdot) - V_{b_n}(x)| > \varepsilon/2) + \Pr(|V_{b_n}(x) - K_{b_n}(x, P)| > \varepsilon/2).
\end{aligned}$$

Since $E_P[V_{b_n}(x)] = K_{b_n}(x, P)$ and V_{b_n} is a U-statistic of degree b_n with kernel bounded by 1, we have by Hoeffding's inequality

$$\Pr(|V_{b_n}(x) - K_{b_n}(x, P)| > \varepsilon) \leq 2 \exp\left(-\frac{n}{b_n}\varepsilon^2/2\right).$$

Now, we can write using the same argument (twice), for any $\eta > 0$,

$$\begin{aligned}
\Pr(|K_{b_n}(x \mid \underline{X}_n, \tau_\cdot) - V_{b_n}(x)| > \varepsilon/2) &= \Pr(|V_{b_n}(x - \tau_{b_n}|\widehat{\theta}_n - \theta|\,) - V_{b_n}(x)| > \varepsilon/2) \\
&\leq \Pr\left(\tau_{b_n}|\widehat{\theta}_n - \theta| > \eta\right) + \Pr(|V_{b_n}(x - \eta\,) - V_{b_n}(x)| > \varepsilon/2) \\
&\leq \Pr\left(\tau_{b_n}|\widehat{\theta}_n - \theta| > \eta\right) + \Pr(|V_{b_n}(x - \eta\,) - K_{b_n}(x - \eta, P\,)| > \varepsilon/6) + \\
&\Pr(|K_{b_n}(x, P) - V_{b_n}(x)| > \varepsilon/6) + \Pr(K_{b_n}(x - \eta, P\,) - K_{b_n}(x, P)| > \varepsilon/6) \\
&\leq \Pr\left(\tau_{b_n}|\widehat{\theta}_n - \theta| > \eta\right) + 4\exp(-\frac{n}{b_n}\varepsilon^2/72) \\
&+ \Pr(|K_{b_n}(x - \eta, P\,) - K_{b_n}(x, P)| > \varepsilon/6)
\end{aligned}$$

But since $K_{b_n}(x, P)$ is supposed to be continuous at x (at least asymptotically), for n large enough, the last term is 0 for a well chosen η. More precisely, under **A5** and the Lipschitz condition **A6**, we have

$$\begin{aligned}
|K_{b_n}(x - \eta, P\,) - K_{b_n}(x, P)| &\leq |K_{b_n}(x - \eta, P\,) - K(x - \eta, P)| + \\
&|K_{b_n}(x, P) - K(x, P)| + |K(x - \eta, P) - K(x, P)| \\
&\leq O(b_n^{-\beta}) + L\eta
\end{aligned}$$

It follows that if we choose η such that $\eta < \varepsilon/12L$, for n large enough the last term vanishes. Now for this choice, we get that, by hypothesis (**A4**), for some non-negative

constants M_1 and M_2, we also have an exponential inequality for $K_{b_n}(x \mid \underline{X}_n, \tau_.)$ of the form

$$\Pr\left(|K_{b_n}(x \mid \underline{X}_n, \tau_.) - K_{b_n}(x, P)| > \varepsilon\right) \leq M_1 \exp(-\frac{n}{b_n}\varepsilon^2/M_2)$$

This proves the first result.
Now, the distribution $K_{b_n}^{(B)}(x \mid \underline{X}_n, \tau_.)$ is obtained (conditionally to the data) by sampling with replacement over all possible subsamples. According to this resampling plan, the $K_{b_n}^{(B)}(x \mid \underline{X}_n, \tau_.)$ concentrates around its mean $K_{b_n}(x, P)$, by Hoeffding's inequality, at a rate $1/\sqrt{B}$. Thus by combining with the preceding results, we get an error of size $O_P\left(\sqrt{\frac{b_n}{n}}\right) + O_P\left(\frac{1}{\sqrt{B}}\right)$. Notice that, when $B >> \frac{n}{b_n}$, we get that the final error is of order $O_P\left(\sqrt{\frac{b_n}{n}}\right)$.

Now for the last propositions, just notice that we have the decomposition

$$K_{b_n}(x \mid \underline{X}_n, \tau_.) - K(x, P) = K_{b_n}(x \mid \underline{X}_n, \tau_.) - K_{b_n}(x, P) + K_{b_n}(x, P) - K(x, P)$$

and

$$K_{b_n}(x \mid \underline{X}_n, \tau_.) - K_n(x, P) = K_{b_n}(x \mid \underline{X}_n, \tau_.) - K(x, P) + K(x, P) - K_n(x, P)$$

Now use assumption **A5** and the preceding results to conclude.

A.2. Proof of Lemma 1 from section 2.3

For any $\epsilon > 0$, we know from Theorem 1 in section 2.2, that there exists some $L = L_\varepsilon$,

$$\Pr{}_P\{|K_{b_n}(x \mid \underline{X}_n, \tau_.) - K(x, P)| \geq L/\delta_\beta(n)\} \leq \varepsilon \tag{A1}$$

uniformly in x. Put $\eta_n = \frac{L}{\delta_\beta(n)}$ and define the quantile $z = K_{b_n}^{-1}(t - \eta_n \mid \underline{X}_n, \tau_.)$, then $K_{b_n}(z \mid \underline{X}_n, \tau_.) \geq t - \eta_n$. Combining this with (A1) implies that $z \geq K^{-1}(t - 2\eta_n, P)$.with probability at most ε. Similarly, define $y = K^{-1}(t, P)$, then we have $y \geq K_{b_n}^{-1}(t - \eta_n | \ \underline{X}_n, \tau_.)$ with probability at most ε. Hence, for any t and any $\epsilon > 0$, we have the inequality :

$$\Pr{}_P\left(K^{-1}(t - 2\eta_n, P) \leq K_{b_n}^{-1}(t - \eta_n \mid \underline{X}_n, \tau_.) \leq K^{-1}(t, P)\right) \leq 2\varepsilon. \tag{A2}$$

This clearly yields , for $\epsilon \to 0^+$, that

$$K_{b_n}^{-1}(t \mid \underline{X}_n, \tau_.) = K^{-1}(t, P) + o_P(1).$$

But, by assumption **A6** and using the Inverse function theorem for Lipschitz, strictly increasing continuous functions that, there exists an $L_\varepsilon^{'}$ such that

$$|K^{-1}(t - 2\eta_n, P) - K^{-1}(t, P)| \leq L_\varepsilon^{'}\eta_n$$

Using again (A2) twice, we get that

$$\begin{aligned} K_{b_n}^{-1}(t \mid \underline{X}_n, \tau_\cdot) &= K^{-1}(t, P) + O_P(\eta_n) \\ &= K^{-1}(t, P) + O_P(\delta_\beta(n)^{-1}), \end{aligned}$$

uniformly in the neighborhood of t.

A.3. Proof of Theorem 2 from section 3.1

The proof follows the same lines as Bertail et al. (1999). For any x, consider

$$\begin{aligned} K_{b_n}(x \mid \underline{X}_n, \widehat{\tau}_\cdot) &\equiv q^{-1} \sum_{i=1}^{q} 1\{b_n^{\hat{\alpha}}(T_{b_n,i} - \widehat{\theta}_n) \leq x\} \\ &= q^{-1} \sum_{i=1}^{q} 1\{b_n^{\hat{\alpha}}(T_{b_n,i} - \theta) - b_n^{\hat{\alpha}}(\widehat{\theta}_n - \theta) \leq x\} \end{aligned}$$

and define the correctly recentered U-statistic

$$U_n(x) = q^{-1} \sum_{i=1}^{q} 1\{b_n^{\hat{\alpha}}(T_{b_n,i} - \theta) \leq x\}$$

and the event

$$E_n = \{b_n^{\hat{\alpha}} |\widehat{\theta}_n - \theta| \leq \epsilon\},$$

for some $\epsilon > 0$.
Since $\hat{\alpha} = \alpha + o_P((\log n)^{-1})$, we have as well $n^{\hat{\alpha}} = n^{\alpha}(1 + o_P(1))$ and $b_n^{\hat{\alpha}} = b_n^{\alpha}(1 + o_P(1))$.
Notice for the last statement of the theorem, that if

$$\hat{\alpha} = \alpha + O_P(\delta_\beta(n)^{-1}),$$

we get

$$n^{\hat{\alpha}} = n^{\alpha}(1 + O_P(log(n)/\delta_\beta(n)))$$

and

$$b_n^{\hat{\alpha}} = b_n^{\alpha}(1 + O_P(log(b_n)/\delta_\beta(n)))).$$

Conditions **A2** imply that $\Pr(E_n) \underset{n\to\infty}{\longrightarrow} 1$; hence, with probability tending to one, we get that

$$U_n(x - \epsilon) \leq K_{b_n}(x \mid \underline{X}_n, \widehat{\tau}_\cdot) \leq U_n(x + \epsilon).$$

Let us show that $U_n(x)$ converges to $K(x, P)$ in probability. For this, introduce the

U-statistic with varying kernel defined by :

$$V_n(x) = q^{-1} \sum_{i=1}^{q} 1\{b_n^\alpha (T_{b_n,i} - \theta) \leq x\},$$

which is the equivalent of $U_n(x)$, with the true rate rather than the estimated one. Recall that since $V_n(x)$ is a U-statistic of degree b_n, such that $\frac{b_n}{n} \to 0$, by Hoeffding's inequality, we have $V_n(x) = K(x, P) + O_P(1/\sqrt{(b_n/n)})$ as $n \to \infty$ in probability. Now, for any $\epsilon_1 > 0$, we have that, with probability tending to 1,

$$U_n(x) = q^{-1} \sum_{i=1}^{q} 1 \left\{ b_n^\alpha (T_{b_n,i} - \theta) \leq \frac{b_n^\alpha}{b_n^{\hat{\alpha}}} x \right\} \leq V_n(x + \epsilon_1).$$

A similar argument shows that we also have $U_n(x) \geq V_n(x - \epsilon_1)$ with probability tending to one. But we have $V_n(x+\epsilon_1) \to K(x+\epsilon_1, P)$ and $V_n(x-\epsilon_1) \to K(x-\epsilon_1, P)$ in probability. Therefore, letting $\epsilon_1 \to 0$, we have that $U_n(x) \to K(x, P)$ in probability as required.

Proving that we have

$$\widehat{K}_n(x, P) - K(x, P) \to 0 \text{ as n} \to \infty.$$

follows now by the same arguments as before by recalling that

$$\begin{aligned} \widehat{K}_n(x, P) &= \Pr\left(\tau_n(T_n - \theta) \leq x \frac{\tau_n}{\widehat{\tau}_n} \right) \\ &= \Pr(\tau_n(T_n - \theta) \leq x(1 + o_P(1)) \end{aligned}$$

and using the continuity of the limiting distribution.

The second part of the theorem is a straightforward consequence of the uniform convergence of $K_{b_n}(x \mid \underline{X}_n, \widehat{\tau}_\cdot) - \widehat{K}_n(x, P)$ to 0, over the neighborhood of any continuity point of the true limiting distribution.

The last result is obtained by the same arguments, by just replacing ϵ by $\epsilon log(n)\delta_\beta(n)^{-1}$. In that case, the probability of the event E_n is controlled by assumption **A4**. All the approximations $o_P(1)$ then becomes $O_P(log(n)\delta_\beta(n)^{-1})$ using the same arguments as in Theorem 1. Similarly ϵ_1 can be replaced by $\epsilon_1 log(b_n)\delta_\beta(n)^{-1}$ which is anyway smaller than $O_P(log(n)\delta_\beta(n)^{-1})$.

Proof of corollary 1 from section 3.3.1

Recall that $\widehat{\theta}_n^q = \frac{1}{q} \sum_{1 \leq j \leq q} \widehat{\mathcal{E}_{b_n}^{(j)}}$. Notice first that the value $\widehat{\mathcal{E}_{b_n}^{(j)}}$ is close to $\mathcal{E}_{b_n}^{(j)}$ at a rate $\sqrt{n - b_n}$ that can be controlled by standard arguments on sums. Indeed, by Hoeffding's

inequality, we have that, for some constant $M > 0$,

$$\Pr\left(|\widehat{\mathcal{E}_{b_n}^{(j)}} - \mathcal{E}_{b_n}^{(j)}| > x\right) = E_{D_{b_n}^{(j)}} P_{\overline{D}_{b_n}^{(j)}}\left(\widehat{\mathcal{E}_{b_n}^{(j)}} - \mathcal{E}_{b_n}^{(j)} > x | D_{b_n}^{(j)}\right)$$
$$\leq 2\exp\left(-\frac{2x^2(n-b_n)}{M^2}\right)$$

so that we have

$$\Pr\left(\sup_{j=1,\ldots,B} |\widehat{\mathcal{E}_{b_n}^{(j)}} - \mathcal{E}_{b_n}^{(j)}| > x\right) \leq 2B\exp\left(-\frac{2x^2(n-b_n)}{M^2}\right).$$

Now, notice that the subsampling distribution may be written

$$K_{b_n}^{(B)}(x \mid \underline{X}_n, \tau_\cdot) = B^{-1}\sum_{j=1}^{B} 1\{\tau_{b_n}(\mathcal{E}_{b_n}^{(j)} - \theta^* + a_n^{(j)}) \leq x\}$$

with $a_n^{(j)} = \widehat{\mathcal{E}_{b_n}^{(j)}} - \mathcal{E}_{b_n}^{(j)} + \theta^* - \widehat{\theta}_n^q$.
As in the proof of Theorem 1, consider the event $E_{b_n} = \{\tau_{b_n} \sup_{j=1,\ldots,B} |\widehat{\mathcal{E}_{b_n}^{(j)}} - \mathcal{E}_{b_n}^{(j)}| < \varepsilon\}$. Then, by the preceding Hoeffding's inequality, using the fact that $B = n^\gamma$, we get :

$$\Pr(E_{b_n}^c) \leq \exp\left(-\frac{2\varepsilon^2(n-b_n)}{\tau_{b_n}^2 M^2} + \gamma\ln(n)\right), \tag{A3}$$

which goes to 0 under our assumptions on b_n. It follows that $\Pr(E_{b_n}) \to 1$ as $n \to \infty$. As in the proof of Theorem 3, we have a Bernstein inequality for $\tau_{b_n}|\widehat{\theta}_n^q - \theta^*|$. Now, apply the same arguments as in Theorem 3, with $T_{b_n,i} = \mathcal{E}_{b_n}^{(i)}$ to get that, some constants $C1, C_2$ and for our b_n, for any fixed $\varepsilon > 0$

$$\Pr\left(\tau_{b_n}\sqrt{\frac{n}{b_n}}|\widehat{\theta}_n^q - \theta^*| > \varepsilon\right) \leq C_1\exp\left(-C_2\varepsilon^2\right) + \Pr(E_{b_n}^c)$$

From this and equation (A3), we get that $|\widehat{\theta}_n^q - \theta^*| = O_P\left(\tau_{b_n}^{-1}\sqrt{\frac{b_n}{n}}\right)$.

This result and again equation (A3) imply that the $\tau_{b_n}a_n^{(j)}$'s are uniformly small. Using the continuity of the limiting distribution as in the proof of Theorem 1, it follows that it is sufficient to study the subsampling distribution

$$K_{b_n}^{(B)}(x) = B^{-1}\sum_{j=1}^{B} 1\{\tau_{b_n}(\mathcal{E}_{b_n}^{(j)} - \theta^*) \leq x\}.$$

which is exactly the one of the U-statistics of Theorem 1, so that the preceding results apply.

Appendix B. Appendix B

B.1. Detailed results on VeReMi dataset

B.1.1. IQR

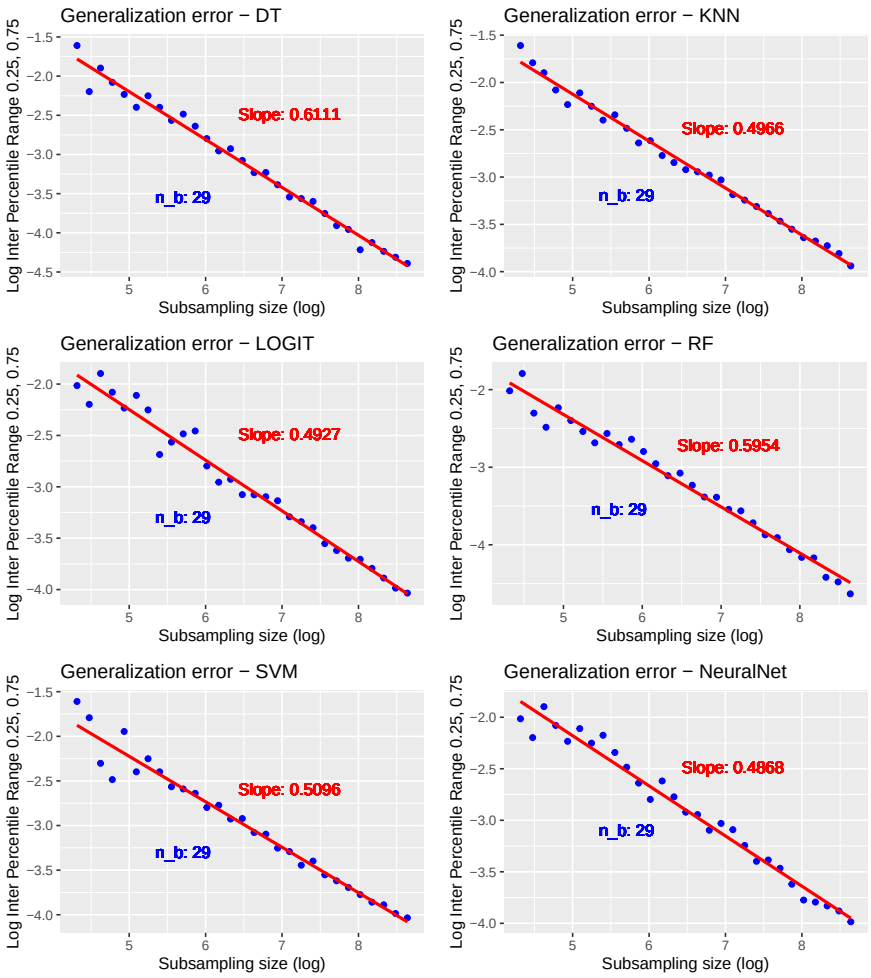

Figure B1. Estimation of convergence rate, based on IQR, VeReMi dataset, $N = 424,810$

B.1.2. IPR80

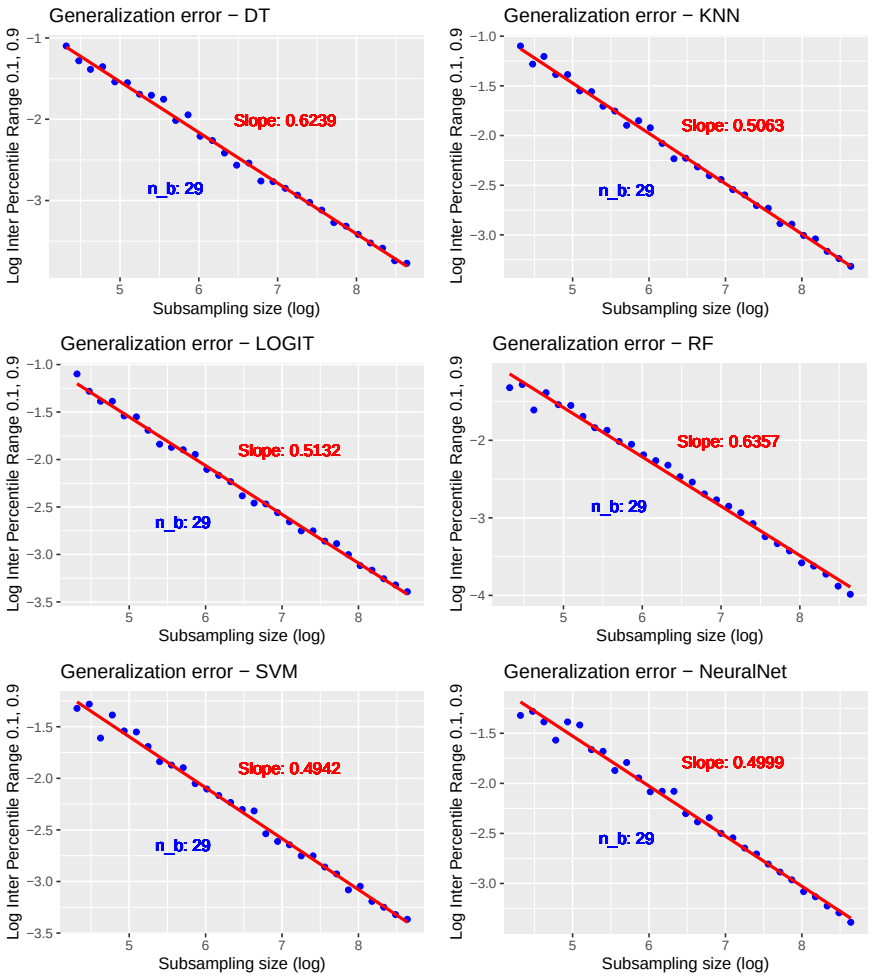

Figure B2. Estimation of convergence rate, based on IPR80, VeReMi dataset, $N = 424,810$

B.1.3. IPR90

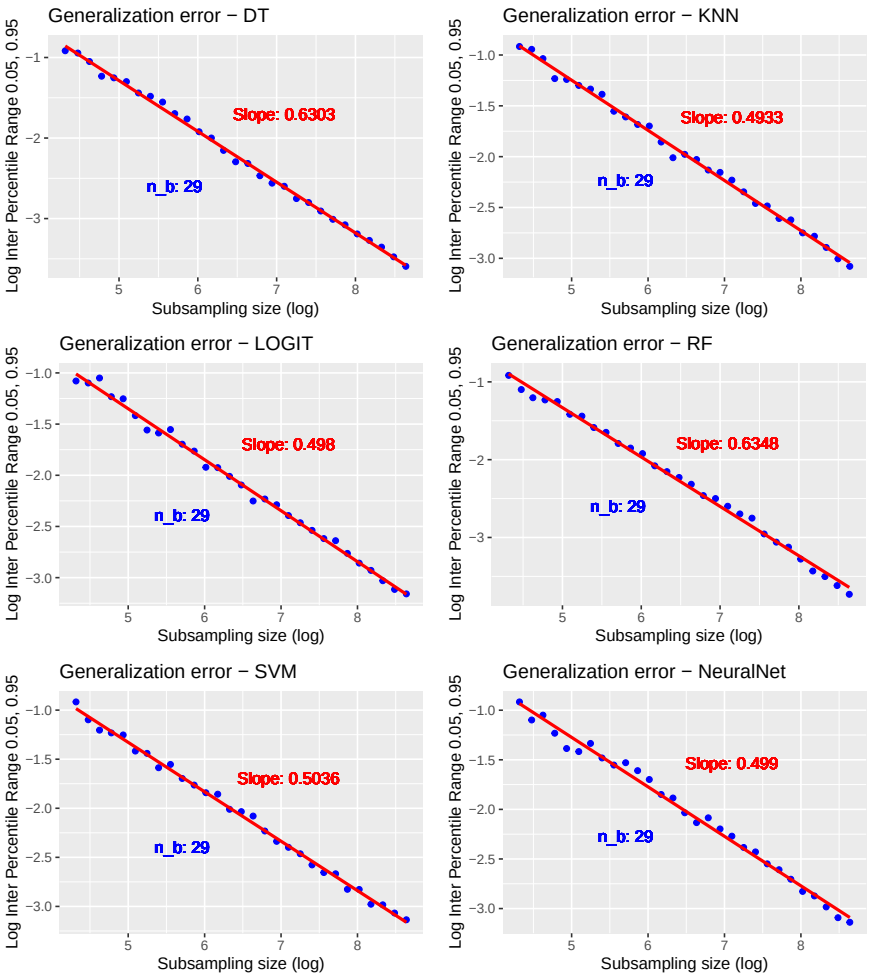

Figure B3. Estimation of convergence rate on IPR90, VeReMi dataset, $N = 424,810$

B.2. Detailed results on EMNIST digit dataset

B.2.1. IQR

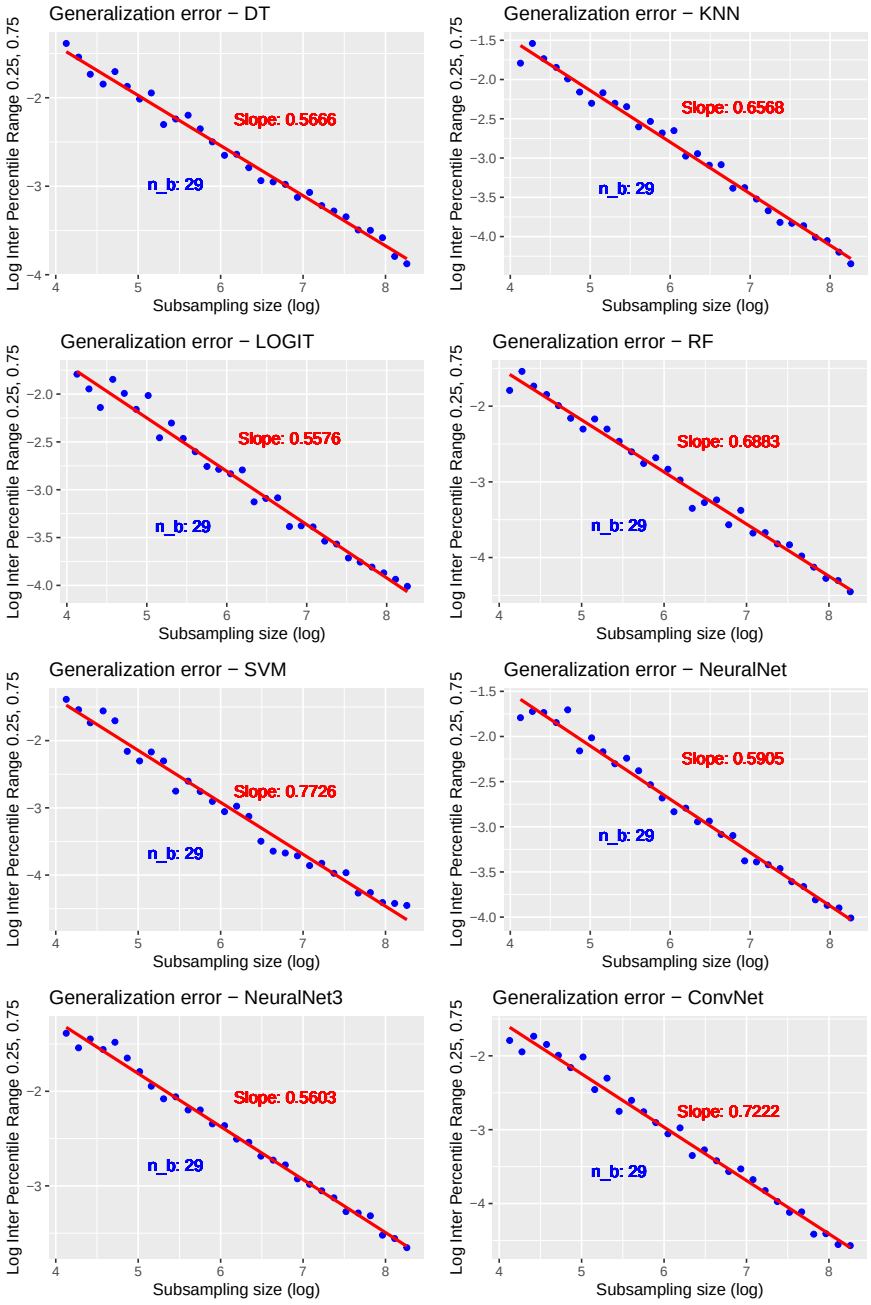

Figure B4. Estimation of convergence rate, based on IQR, EMNIST digit dataset, $N = 240,000$

B.2.2. IPR80

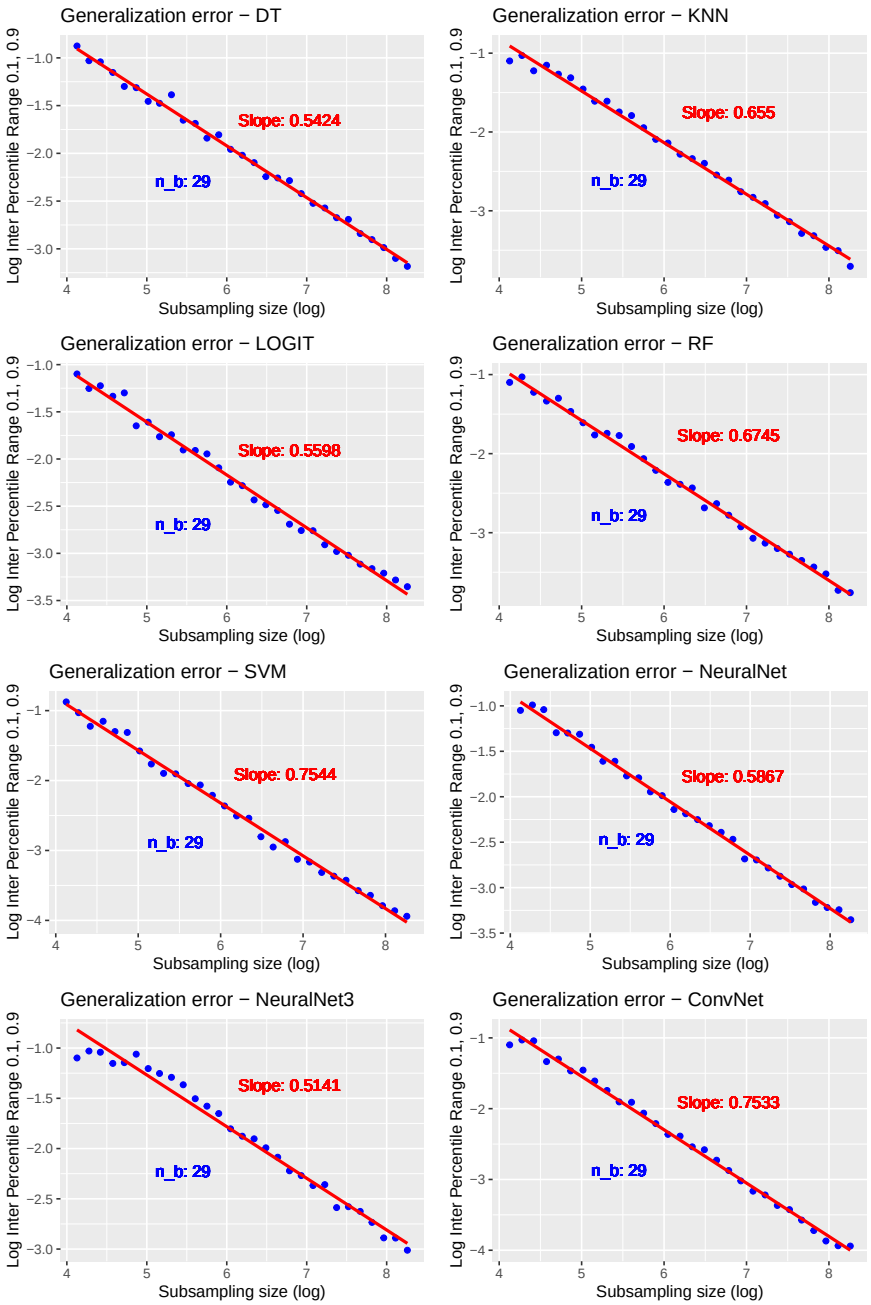

Figure B5. Estimation of convergence rate, based on IPR80, EMNIST digit dataset, $N = 240,000$

B.2.3. IPR90

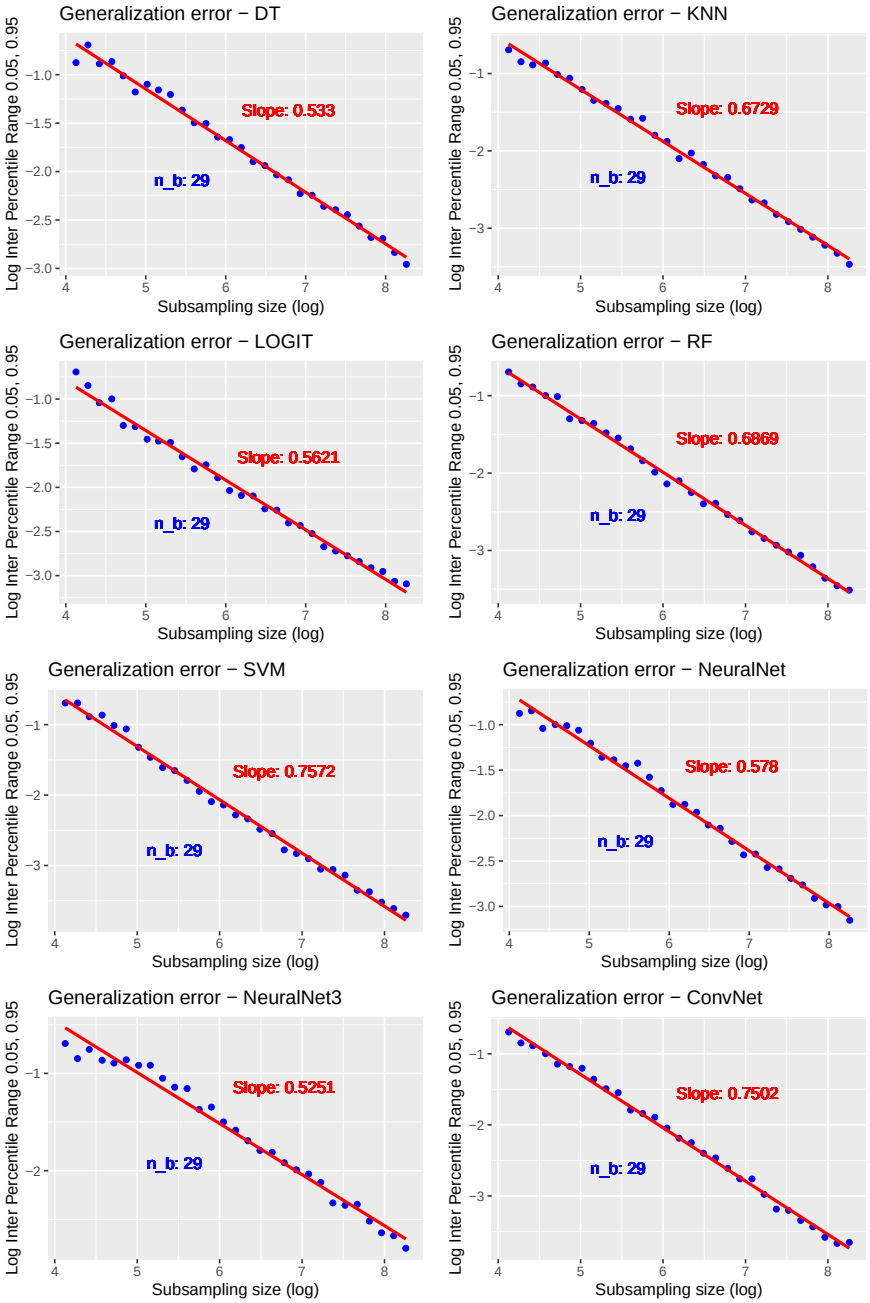

Figure B6. Estimation of convergence rate, based on IPR90, EMNIST digit dataset, $N = 240,000$

B.3. Detailed results on Lecun MNIST digit dataset

B.3.1. IQR

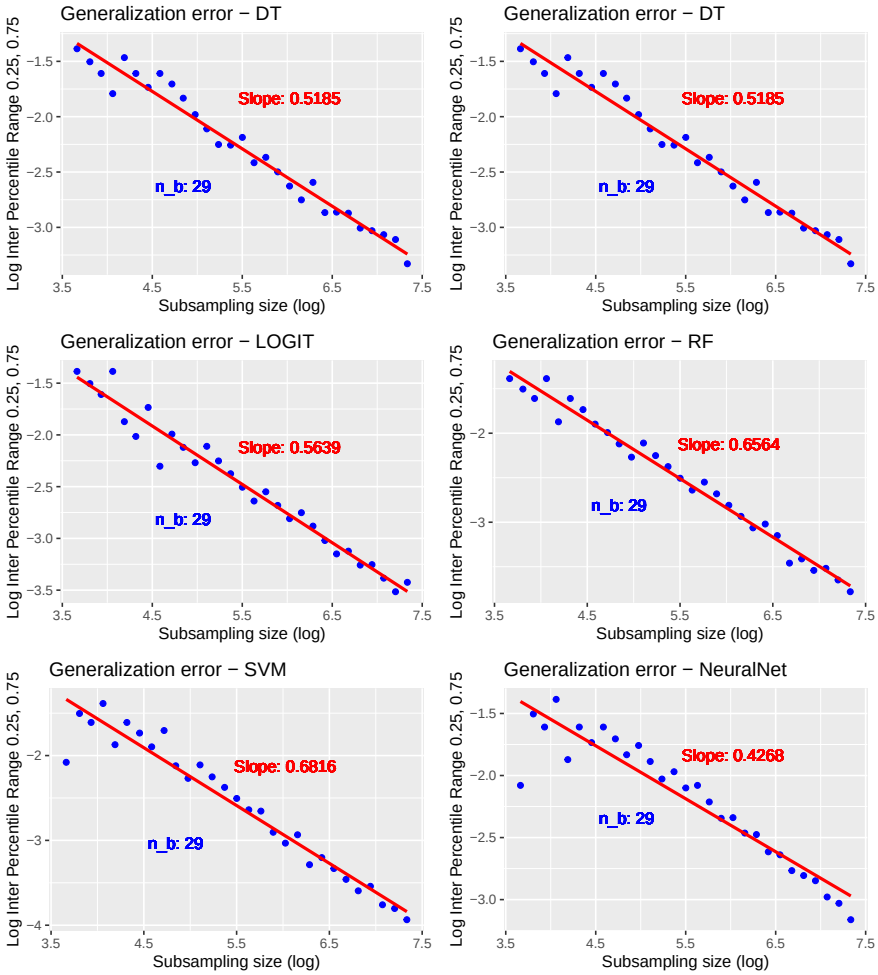

Figure B7. Estimation of convergence rate, based on IQR, Lecun MNIST digit dataset, $N = 60,000$

Compliance with ethical standards

Funding: This research has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01) as well as Teralab and the industrial chair "Machine Learning for Big Data".

Conflict of Interest: The authors declare that they have no conflict of interest.

References

Arcones MA (1995) A Bernstein-type inequality for U-statistics and U-processes. *Statistics and Probability Letters* 22(3):239–247

Arcones MA, Gine E (1993) Limit theorems for U-processes. *The Annals of Probability* 21(3):1494–1542

Banerjee M, Durot C, Sen B (2019) Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *The Annals of Statistics* 47(2):720–757

Bertail P (1997) Second-order properties of an extrapolated bootstrap without replacement under weak assumptions. *Bernoulli* 3(2):149–179

Bertail P (2011) Comments on: Subsampling weakly dependent time series and application to extremes. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research* 20(3):487–490

Bertail P, Politis DN (2001) Extrapolation of subsampling distribution estimators: The i.i.d. and strong mixing cases. *Canadian Journal of Statistics* 29(4):667–680

Bertail P, Politis DN, Romano JP (1999) On subsampling estimators with unknown rate of convergence. *Journal of the American Statistical Association* 94(446):569–579

Bertail P, Haefke C, Politis DN, White H (2004) Subsampling the distribution of diverging statistics with applications to finance. *Journal of Econometrics* 120(2):295–326

Bertail P, Chautru E, Clémençon S (2015) Tail index estimation based on survey data. *ESAIM: Probability and Statistics* 19:28 – 59

Bertail P, Chautru E, Clémençon S (2017) Empirical Processes in Survey Sampling with (Conditional) Poisson Designs. *Scandinavian Journal of Statistics* 44(1):97–111

Bickel PJ, Sakov A (2008) On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica* 18:967–985

Bickel PJ, Yahav JA (1988) Richardson extrapolation and the bootstrap. *Journal of the American Statistical Association* 83(402):387–393

Bickel PJ, Boley N, Brown JB, Huang H, Zhang NR (2010) Subsampling methods for genomic inference. *The Annals of Applied Statistics* 4(4):1660–1697

Bingham NH, Goldie CM, Teugels JL (1987) *Regular variation*. Cambridge University Press

Breiman L (2001) Random forests. *Machine learning* 45(1):5–32

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) *Classification and regression trees*. Routledge

Bretagnolle J (1983) Lois limites du bootstrap de certaines fonctionnelles. *Annales de l'Institut Henri Poincaré* 19(3):281–296

Chang CC, Lin CJ (2011) Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2(3):1–27

Clémençon S, Bertail P, Chautru E (2014) Scaling up m-estimation via sampling designs: The horvitz-thompson stochastic gradient descent. In: *2014 IEEE International Conference on Big Data (Big Data)*, pp 25–30

Cohen G, Afshar S, Tapson J, van Schaik A (2017) EMNIST: an extension of MNIST to handwritten letters. *CoRR* abs/1702.05373

Götze F, Rakauskas A (2001) Adaptive choice of bootstrap sample sizes. *Lecture Notes-Monograph Series* 36:286–309

Hall P (1986) On the Number of Bootstrap Simulations Required to Construct a Confidence Interval. *The Annals of Statistics* 14(4):1453 – 1462

Hall P (2003) A short prehistory of the bootstrap. *Statistical Science* 18(2):158–167

Har-Peled S (2011) *Geometric Approximation Algorithms*. American Mathematical Society, Boston, MA, USA

van der Heijden RW, Lukaseder T, Kargl F (2018) VeReMi: A dataset for comparable evaluation of misbehavior detection in vanets. In: *International Conference on Security and Privacy in Communication Systems*, Springer, pp 318–337

Heilig C, Nolan D (2001) Limit theorems for the infinite-degree u-process. *Statistica Sinica* 11(1):289–302

Isaacson E, Keller HB (1966) *Analysis of numerical methods*. New York: Wiley

Kleiner A, Talwalkar A, Sarkar P, Jordan MI (2014) A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(4):795 – 816

Laforgue P, Clémencon S, Bertail P (2019) On medians of (Randomized) pairwise means. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, *Proceedings of Machine Learning Research*, vol 97, pp 1272–1281

Le Cam L (1990) Maximum likelihood: An introduction. *International Statistical Review / Revue Internationale de Statistique* 58(2):153–171

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. *Neural computation* 1(4):541–551

LeCun Y, Jackel L, Bottou L, Brunot A, Cortes C, Denker J, Drucker H, Guyon I, Muller U, Sackinger E, et al. (1995) Comparison of learning algorithms for handwritten digit recognition. In: *International conference on artificial neural networks*, Perth, Australia, vol 60, pp 53–60

Mahalanobis PC (1958) Recent experiments in statistical sampling in the indian statistical institute. *Sankhyā: The Indian Journal of Statistics (1933-1960)* 20(3/4):329–398

McCullagh P, Nelder JA (1983) *Generalized linear models*. Routledge

McLeod AI, Bellhouse DR (1983) A convenient algorithm for drawing a simple random sample. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 32(2):182–184

Politis DN, Romano JP (1994) Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics* 22(4):2031–2050

Politis DN, Romano JP, Wolf M (1999) Subsampling. Springer New York, New York, NY

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088):533–536