Some non-asymptotic results on resampling in high dimension, I: Confidence regions, II: Multiple tests

Sylvain Arlot, Gilles Blanchard, Etienne Roquain

Submitted to the Annals of Statistics

SOME NON-ASYMPTOTIC RESULTS ON RESAMPLING IN HIGH DIMENSION, I: CONFIDENCE REGIONS

By Sylvain Arlot*,†, Gilles Blanchard*,‡ and Etienne Roquain*,§

CNRS ENS, Weierstrass Institut and University of Paris 6

We study generalized bootstrap confidence regions for the mean of a random vector whose coordinates have an unknown dependency structure. The random vector is supposed to be either Gaussian or to have a symmetric and bounded distribution. The dimensionality of the vector can possibly be much larger than the number of observations and we focus on a non-asymptotic control of the confidence level, following ideas inspired by recent results in learning theory. We consider two approaches, the first based on a concentration principle (valid for a large class of resampling weights) and the second on a direct resampled quantile, specifically using Rademacher weights. Several intermediate results established in the approach based on concentration principles are of self-interest. We also discuss the question of accuracy when using Monte-Carlo approximations of the resampled quantities.

1. Introduction.

1.1. *Goals and motivations.* Let $\mathbf{Y} := (\mathbf{Y}^1, \dots, \mathbf{Y}^n)$ be a sample of $n \geq 2$ i.i.d. observations of an integrable random vector in $\mathbb{R}^K$, with dimensionality K possibly much larger than n, and unknown dependency structure of the coordinates. Let $\mu \in \mathbb{R}^K$ denote the common mean of the $\mathbf{Y}^i$; our goal is to find a non-asymptotic $(1-\alpha)$-confidence region $\mathcal{G}(\mathbf{Y}, 1-\alpha)$ for μ, of the form:

$$\mathcal{G}(\mathbf{Y}, 1-\alpha) = \left\{ x \in \mathbb{R}^K \mid \phi\left(\overline{\mathbf{Y}} - x\right) \leq t_\alpha(\mathbf{Y}) \right\} \quad , \tag{1}$$

where $\phi : \mathbb{R}^K \to \mathbb{R}$ is a fixed in advance function (measuring a kind of distance, for example an ℓ_p-norm for $p \in [1, \infty]$), $\alpha \in (0,1)$, $t_\alpha : \left(\mathbb{R}^K\right)^n \to \mathbb{R}$ is a possibly data-dependent threshold, and

*This work was supported in part by the IST and ICT Programmes of the European Community, successively under the PASCAL (IST-2002-506778) and PASCAL2 (ICT-216886) networks of excellence.

†Research mostly carried out at Univ Paris-Sud (Laboratoire de Mathematiques, CNRS - UMR 8628).

‡This research was in part carried out while the second author held an invited position at the statistics department of the University of Chicago, which is warmly acknowledged.

§Research mostly carried out at the French institute INRA-Jouy and at the Free University of Amsterdam.

AMS 2000 subject classifications: Primary 62G15; secondary 62G09.

Keywords and phrases: confidence regions, high dimensional data, non-asymptotic error control, resampling, cross-validation, concentration inequalities, resampled quantile.

$\overline{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}^i \in \mathbb{R}^K$ is the empirical mean of the sample $\mathbf{Y}$.

The point of view developed in the present work focuses on the following goal:

- obtaining *non-asymptotic* results, valid for any fixed K and n, with K possibly much larger than the number of observations n,
- while avoiding any specific assumption on the dependency structure of the coordinates of $\mathbf{Y}^i$ (although we will consider some general assumptions over the distribution of $\mathbf{Y}$, namely symmetry and boundedness or Gaussianity).

In the Gaussian case, a traditional parametric method based on the direct estimation of the covariance matrix to derive a confidence region would not be appropriate in the situation where $K \gg n$, unless the covariance matrix is assumed to belong to some parametric model of lower dimension, which we explicitly don't want to posit here. In this sense, the approach followed here is closer in spirit to non-parametric or semiparametric statistics.

This point of view is motivated by some practical applications, especially neuroimaging [26, 8, 18]. In a magnetoencephalography (MEG) experiment, each observation $\mathbf{Y}^i$ is a two or three dimensional brain activity map, obtained as a difference between brain activities in the presence or absence of some stimulation. The activity map is typically composed of about 15 000 points; the data can also be a time series of length between 50 and 1 000 of such maps. The dimensionality K can thus range from 10^4 to 10^7. Such observations are repeated $n = 15$ up to 4 000 times, but this upper bound is seldom attained [32]; in typical cases, one has $n \leq 100 \ll K$. In such data, there are strong dependencies between locations (the 15 000 points are obtained by pre-processing data of 150 sensors), and these dependencies are spatially highly non-homogeneous, as noted by [26]. Moreover, there may be long-distance correlations, for example depending on neural connections inside the brain, so that a simple parametric model of the dependency structure is generally not adequate. Another motivating example is given by microarray data [14], where it is common to observe samples of limited size (say, less than 100) of a vector in high dimension (say, more than 20 000, each dimension corresponding to a specific gene), and where the dependency structure may be quite arbitrary.

1.2. *Two approaches to our goal.* The ideal threshold t_α in (1) is obviously the $(1-\alpha)$ quantile of the distribution of $\phi\left(\overline{\mathbf{Y}} - \mu\right)$. However, this quantity depends on the unknown dependency structure of the coordinates of $\mathbf{Y}^i$ and is therefore itself unknown.

The approach studied in this work is to use a (generalized) resampling scheme in order to estimate t_α. The heuristics of the resampling method (introduced in [11], generalized to exchangeable weighted bootstrap by [23, 28]) is that the distribution of the unobservable variable $\overline{\mathbf{Y}} - \mu$ is "mimicked" by the distribution, conditionally to $\mathbf{Y}$, of the resampled empirical mean of the centered data. This last quantity is an observable variable, and we denote it as follows:

$$\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle} := \frac{1}{n}\sum_{i=1}^{n}(W_i - \overline{W})\mathbf{Y}^i = \frac{1}{n}\sum_{i=1}^{n} W_i(\mathbf{Y}^i - \overline{\mathbf{Y}}) = \overline{\left(\mathbf{Y} - \overline{\mathbf{Y}}\right)}^{\langle W\rangle}\,, \tag{2}$$

where $(W_i)_{1\leq i\leq n}$ are real random variables independent of $\mathbf{Y}$ called the *resampling weights*, and $\overline{W} = n^{-1}\sum_{i=1}^n W_i$. We emphasize that the weight family $(W_i)_{1\leq i\leq n}$ itself *need not be independent.*

We define in more detail several specific resampling weights in Section 2.4, inspired both from traditional resampling methods [23, 28] and from recent statistical learning theory. Let us give two typical examples reflecting these two sources:

- Efron's bootstrap weights: W is a multinomial random vector with parameters $(n; n^{-1}, \ldots, n^{-1})$. This is the standard bootstrap.
- Rademacher weights: W_i are i.i.d. Rademacher variables, that is, $W_i \in \{-1, 1\}$ with equal probabilities. They are closely related to symmetrization techniques in learning theory.

It is useful to notice at this point that, to the extent that we only consider resampled data after empirical centering, shifting all weights by the same (but possibly random) offset $C > 0$ does not change the resampled quantity introduced in (2). Hence, to reconcile the intuition of traditional resampling with what could possibly appear as unfamiliar weights, one could always assume that the weights are translated to enforce (for example) weight positivity, or the condition $n^{-1}\sum_{i=1}^n W_i = 1$ (though of course in general both conditions can't be ensured at the same time simply by translation). For example, Rademacher weights can be interpreted as a resampling scheme where each $\mathbf{Y}^i$ is independently discarded or "doubled" with equal probability.

Following the general resampling idea, we investigate two distinct approaches in order to obtain non-asymptotic confidence regions:

- Approach 1 ("concentration approach", developed in Section 2):
  The expectations of $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ and $\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)$ can be precisely compared, and the processes $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ and $\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)\right]$ concentrate well around their respective expectations, where $\mathbb{E}_W$ denotes the expectation operator with respect to the distribution of W (that is, conditionally to $\mathbf{Y}$).
- Approach 2 ("direct quantile approach", developed in Section 3):
  The $1-\alpha$ quantile of the distribution of $\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)$ conditionally to $\mathbf{Y}$ is close to the $1-\alpha$ quantile of $\phi\left(\overline{\mathbf{Y}} - \mu\right)$.

Regarding the second approach, we will restrict ourselves specifically to Rademacher weights in our analysis, and rely heavily on a symmetrization principle.

1.3. *Relation to previous work.* Using resampling to construct confidence regions is a vast field of study in statistics (see for instance [11, 16, 15, 9, 4, 27]). Available results are however mostly asymptotic, based on the celebrated fact that the bootstrap process is asymptotically close to the original empirical process [31]. Because we focus on a non-asymptotic viewpoint, this asymptotic approach is not adapted to the goals we have fixed. Note also that the non-

asymptotic viewpoint can be used as a basis for an asymptotic analysis in the situation where the dimension K grows with n, a setting which is typically not covered by standard asymptotics.

The "concentration approach" mentioned in the previous section is inspired by recent results coming from learning theory, and relates in particular the notion of Rademacher complexity [20]. This notion has been extended in the recent work of Fromont [13] to more general resampling schemes, and this latter work has had a strong influence on the present one.

On the other hand, what we called the "quantile approach" in the previous section is strongly related to exact randomization tests (which are based on an invariance of the null distribution under a given transformation; the underlying idea can be traced back to Fisher's permutation test [12]). Namely, we will only consider symmetric distributions: this is a specific instance of an invariance with respect to a transformation and will allow us to make use of distribution-preserving randomization via sign-flipping. Here, the main difference with traditional exact randomization tests is that, since our goal is to derive a confidence region, the vector of the means is unknown and therefore, so is the exact invariant transformation. Our contribution to this point is essentially to show that the true vector of the means can be replaced by the empirical one in the randomization, for the price of additional terms of smaller order in the threshold thus obtained. To our knowledge, this gives the first non-asymptotic approximation result on resampled quantiles with an unknown distribution mean.

Finally, we contrast the setting studied here to a strand of research studying adaptive confidence regions (in a majority of cases, ℓ_2 balls) in nonparametric Gaussian regression. A seminal paper on this topic is [22], and recent work includes [21, 17, 29] (in an asymptotical point of view) and [5, 3, 19, 6] (which present non-asymptotic results). Related to this setting and ours is [10], where adaptive tests for zero mean are developed for symmetric distributions, using the randomization by sign-flipping. The setting considered in these papers is that of regression on a fixed design in high dimension (or in the Gaussian sequence model), with one observation per point and i.i.d. noise. This corresponds (in our notation) to $n = 1$, while the K coordinates are assumed independent. Despite some similarities, the problem considered here has a different nature: in the above works, the focus is on the adaptivity with respect to some properties of the true mean vector, materialized by a family of models (e.g. linear subspaces or Besov balls in the Gaussian sequence setting); usually an adaptive estimator performing implicit or explicit model selection relative to this collection is studied, and a crucial question for obtaining confidence regions is that of estimating empirically the bias of this estimator, while the noise dependence structure is known. In the present paper, we do not consider the problem of model selection, but the focus is on evaluating the estimation error under an unknown noise dependence structure (for the "naive" unbiased estimator given by the empirical mean).

1.4. *Notation.* We first define some notation that will be useful throughout the paper.

- A boldface letter indicates a matrix. This will almost exclusively concern the $K \times n$ data matrix $\mathbf{Y}$. A superscript index such as $\mathbf{Y}^i$ indicates the i-th column of a matrix.
- If $\mu \in \mathbb{R}^K$, $\mathbf{Y} - \mu$ is the matrix obtained by subtracting μ from each (column) vector of

$\mathbf{Y}$. Similarly, for a vector $W \in \mathbb{R}^n$ and $c \in \mathbb{R}$, we denote $W - c := (W_i - c)_{1 \le i \le n} \in \mathbb{R}^n$.

- If X is a random variable, $\mathcal{D}(X)$ is its distribution and $\mathrm{Var}(X)$ is its variance. We use the notation $X \sim Y$ to indicate that X and Y have the same distribution. The support of $\mathcal{D}(X)$ is moreover denoted by $\operatorname{supp} \mathcal{D}(X)$.
- We denote by $\mathbb{E}_W[\cdot]$, the expectation operator over the distribution of the weight vector W only, that is, conditional to $\mathbf{Y}$. We use a similar notation $\mathbb{P}_W$ for the corresponding probability operator and $\mathbb{E}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y}}$ for the same operations conditional to W. Since $\mathbf{Y}$ and W are always assumed to be independent, the operators $\mathbb{E}_W$ and $\mathbb{E}_{\mathbf{Y}}$ commute by Fubini's theorem.
- The vector $\sigma = (\sigma_k)_{1 \le k \le K}$ is the vector of the standard deviations of the data: $\forall k, 1 \le k \le K$, $\sigma_k := \mathrm{Var}^{1/2}(\mathbf{Y}_k^1)$.
- $\overline{\Phi}$ is the standard Gaussian upper tail function: if $X \sim \mathcal{N}(0,1)$, $\forall x \in \mathbb{R}$, $\overline{\Phi}(x) := \mathbb{P}(X \ge x)$.
- We define the mean of the weight vector $\overline{W} := \frac{1}{n}\sum_{i=1}^n W_i$, the empirical mean vector $\overline{\mathbf{Y}} := \frac{1}{n}\sum_{i=1}^n \mathbf{Y}^i$, and the resampled empirical mean vector $\overline{\mathbf{Y}}^{\langle W \rangle} := \frac{1}{n}\sum_{i=1}^n W_i \mathbf{Y}^i$.
- We use the operator $|\cdot|$ to denote the cardinality of a set.
- For two positive sequences $(u_n)_n$ and $(v_n)_n$, we denote $u_n = \Theta(v_n)$ when $(u_n v_n^{-1})_n$ stays bounded away from zero and $+\infty$.

Several properties may be assumed for the function $\phi : \mathbb{R}^K \to \mathbb{R}$ used to define confidence regions of the form (1):

- Subadditivity: $\forall x, x' \in \mathbb{R}^K, \quad \phi(x + x') \le \phi(x) + \phi(x')$.
- Positive-homogeneity: $\forall x \in \mathbb{R}^K, \forall \lambda \in \mathbb{R}^+, \quad \phi(\lambda x) = \lambda \phi(x)$.
- Boundedness by the ℓ_p-norm, $p \in [1, \infty]$: $\forall x \in \mathbb{R}^K$, $|\phi(x)| \le \|x\|_p$, where $\|x\|_p$ is equal to $(\sum_{k=1}^K |x_k|^p)^{1/p}$ if $p < \infty$ and $\max_k\{|x_k|\}$ for $p = +\infty$. Notice also that all the results of the paper are still valid with any normalization of the ℓ_p-norm (in particular, it can be taken equal to $(K^{-1}\sum_{k=1}^K |x_k|^p)^{1/p}$, so that the ℓ_p-norm of a vector with equal coordinates does not depend on the dimensionality K).

Finally, we define the following possible assumptions on the generating distribution of $\mathbf{Y}$:

(GA) The Gaussian assumption: the $\mathbf{Y}^i$ are Gaussian vectors.

(SA) The symmetric assumption: the $\mathbf{Y}^i$ are symmetric with respect to μ, that is, $(\mathbf{Y}^i - \mu) \sim (\mu - \mathbf{Y}^i)$.

(BA)(p, M) The boundedness assumption: $\|\mathbf{Y}^i - \mu\|_p \le M$ a.s.

In this paper, we primarily focus on the Gaussian framework (GA), where the corresponding results will be more accurate. In the sequel, when considering (GA) and the assumption that ϕ is bounded by the ℓ_p-norm for some $p \ge 1$, we will additionally always assume that we know some upper bound on the ℓ_p-norm of σ. The question of finding an upper bound for $\|\sigma\|_p$ based on the data is discussed in Section 4.1.

2. Confidence region using concentration.

2.1. *Main result.* We consider here a general *resampling weight vector* W, that is, a $\mathbb{R}^n$-valued random vector $W = (W_i)_{1\leq i\leq n}$ independent of $\mathbf{Y}$ satisfying the following properties: for all $i \in \{1,\dots,n\}$ $\mathbb{E}\left[W_i^2\right] < \infty$ and $n^{-1}\sum_{i=1}^n \mathbb{E}\left|W_i - \overline{W}\right| > 0$.

We will mainly consider in this section an *exchangeable resampling weight vector*, that is, a resampling weight vector W such that $(W_i)_{1\leq i\leq n}$ has an exchangeable distribution (in other words, invariant under any permutation of the indices). Several examples of exchangeable resampling weight vectors are given below in Section 2.4, where we also address the question of how to choose between different possible distributions of W. An extension of our results to non-exchangeable weight vectors is proposed in Section 2.5.1.

Four constants that depend only on the distribution of W appear in the results below (the fourth one is defined only for a particular class of weights). They are defined as follows and computed for classical resamplings in Table 1:

$$A_W := \mathbb{E}\left|W_1 - \overline{W}\right| \tag{3}$$

$$B_W := \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \left(W_i - \overline{W}\right)^2\right)^{\frac{1}{2}}\right] \tag{4}$$

$$C_W := \left(\frac{n}{n-1}\mathbb{E}\left[\left(W_1 - \overline{W}\right)^2\right]\right)^{\frac{1}{2}} \tag{5}$$

$$D_W := a + \mathbb{E}\left|\overline{W} - x_0\right| \quad \text{if } \forall i,\ |W_i - x_0| = a \text{ a.s. (with } a > 0, x_0 \in \mathbb{R}). \tag{6}$$

Note that these quantities are positive for an exchangeable resampling weight vector W and satisfy:

$$0 < A_W \leq B_W \leq C_W\sqrt{1 - 1/n}\ .$$

Moreover, if the weights are i.i.d., we have $C_W = \mathrm{Var}(W_1)^{\frac{1}{2}}$. We can now state the main result of this section:

Theorem 2.1 *Fix $\alpha \in (0,1)$ and $p \in [1,\infty]$. Let $\phi : \mathbb{R}^K \to \mathbb{R}$ be any function which is subadditive, positive-homogeneous and bounded by the ℓ_p-norm, and let W be an exchangeable resampling weight vector.*

1. *If $\mathbf{Y}$ satisfies (GA), then*

$$\phi\left(\overline{\mathbf{Y}} - \mu\right) < \frac{\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right)\right]}{B_W} + \|\sigma\|_p\, \overline{\Phi}^{-1}(\alpha/2)\left[\frac{C_W}{nB_W} + \frac{1}{\sqrt{n}}\right] \tag{7}$$

holds with probability at least $1-\alpha$. The same bound holds for the lower deviations, that is, with inequality (7) reversed and the additive term replaced by its opposite.

2. If $\mathbf{Y}$ *satisfies* (SA) *and* $(BA)(p, M)$ *for some* $M > 0$*, then*

$$\phi\left(\overline{\mathbf{Y}} - \mu\right) < \frac{\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right)\right]}{A_W} + \frac{2M}{\sqrt{n}}\sqrt{\log(1/\alpha)} \tag{8}$$

holds with probability at least $1-\alpha$*. If moreover the weight vector satisfies the assumption of* (6)*, then*

$$\phi\left(\overline{\mathbf{Y}} - \mu\right) > \frac{\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right)\right]}{D_W} - \frac{M}{\sqrt{n}}\sqrt{1 + \frac{A_W^2}{D_W^2}}\sqrt{2\log(1/\alpha)} \tag{9}$$

holds with probability at least $1-\alpha$*.*

Inequalities (7) and (8) give regions of the form (1) that are confidence regions of level at least $1-\alpha$. They require to know some upper bound on $\|\sigma\|_p$ (resp. M), or a good estimate of it. We address this question in Section 4.1.

In order to get some insight about these bounds, it is useful to compare them with an elementary inequality. In the Gaussian case, it is true for each coordinate $k, 1 \le k \le K$ that the following inequality holds with probability $1-\alpha$: $|\overline{\mathbf{Y}}_k - \mu_k| < \frac{\sigma_k}{\sqrt{n}}\overline{\Phi}^{-1}(\alpha/2)$. By applying a simple union bound over the coordinates and using that ϕ is positive-homogenous and bounded by the ℓ_p-norm, we conclude that the following inequality holds with probability at least $1-\alpha$:

$$\phi(\overline{\mathbf{Y}} - \mu) < \frac{\|\sigma\|_p}{\sqrt{n}}\overline{\Phi}^{-1}\left(\frac{\alpha}{2K}\right) =: t_{\mathrm{Bonf}}(\alpha)\,, \tag{10}$$

which is a minor variation on the well-known Bonferroni bound. By comparison, the main term in the remainder part of (7) takes a similar form, but with K being replaced by 1: the remainder term is *dimension-independent*. Naturally, the "dimension complexity" has not disappeared, but will be taken into account in the main resampled term instead. When K is large, the bound (7) can improve over the Bonferroni threshold if there are strong dependencies between the coordinates, resulting in a significantly smaller resampling term.

For illustration, consider an extreme example where all pairwise coordinate correlations are exactly 1, that is, the random vector $\mathbf{Y}$ is made of K copies of the same random variable so that there is in fact no dimension complexity. Take $\phi(X) = \sup_i X_i$ (corresponding to a uniform one-sided confidence bound for the mean components). Then the resampled quantity in (7) is equal to zero and the obtained bound is close to optimal (up to the two following points: the level is divided by a factor 2 and there is an additional term of order $\frac{1}{n}$). By comparison, the Bonferroni bound divides the level by a factor K, resulting in a significantly worse threshold. In passing, this example illustrates that the order $n^{-1/2}$ of the remainder term cannot be improved.

If we now interpret the bound (7) from an asymptotic point of view (with $K(n)$ depending on n and $\|\sigma\|_p = \Theta(1)$), the rate of convergence to zero cannot be faster than $n^{-\frac{1}{2}}$ (which

corresponds to the standard parametric rate when K is fixed), but it can be potentially slower, for example if K increases exponentially with n. In the latter case, the rate of convergence of the Bonferroni threshold is always strictly slower than $n^{-\frac{1}{2}}$. In general, as far as the order in n is concerned, the resampled threshold converges at least as fast as Bonferroni's, but whether it is stricly faster once again depends on the coordinate dependency structure.

However, if the coordinates are only "weakly dependent", the threshold (7) can be more conservative than Bonferroni's by a multiplicative factor, while the Bonferroni threshold can sometimes be essentially optimal (for instance, with $\phi = ||\cdot||_\infty$, all the coordinates independent and with small α). This motivates the next result, where we assume more generally that an alternate analysis of the problem can lead to deriving a *deterministic* threshold t_α such that $\mathbb{P}(\phi\left(\overline{\mathbf{Y}} - \mu\right) > t_\alpha) \leq \alpha$. In this case, we would ideally like to take the "best of two approaches" and consider the minimum of t_α and the resampling-based thresholds considered above. In the Gaussian case, the following proposition establishes that we can combine the concentration threshold corresponding to (7) with t_α to obtain a threshold that is very close to the minimum of the two.

Proposition 2.2 *Fix $\alpha, \delta \in (0,1)$, $p \in [1,\infty]$ and take ϕ and W as in Theorem 2.1. Suppose that $\mathbf{Y}$ satisfies (GA) and that $t_{\alpha(1-\delta)}$ is a real number such that $\mathbb{P}\left(\phi\left(\overline{\mathbf{Y}} - \mu\right) > t_{\alpha(1-\delta)}\right) \leq \alpha(1-\delta)$. Then with probability at least $1-\alpha$, $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ is less than or equal to the minimum between $t_{\alpha(1-\delta)}$ and*

$$\frac{\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right)\right]}{B_W} + \frac{\|\sigma\|_p}{\sqrt{n}}\overline{\Phi}^{-1}\left(\frac{\alpha(1-\delta)}{2}\right) + \frac{\|\sigma\|_p C_W}{nB_W}\overline{\Phi}^{-1}\left(\frac{\alpha\delta}{2}\right). \tag{11}$$

The important point to notice in Proposition 2.2 is that, since the last term of (11) becomes negligible with respect to the rest when n grows large, we can choose δ to be quite small (typically $\delta = \Theta(1/n)$), and obtain a threshold very close to the minimum between t_α and the threshold corresponding to (7). Therefore, this result is more subtle than just considering the minimum of two thresholds each taken at level $1 - \frac{\alpha}{2}$, as would be obtained by a direct union bound.

The proof of Theorem 2.1 involves results which are of self interest: the comparison between the expectations of the two processes $\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right)\right]$ and $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ and the concentration of these processes around their means. These two issues are correspondingly examined in the two next sections (2.2 and 2.3). In Section 2.4, we give some elements for an appropriate choice of resampling weight vectors among several classical examples. The last section (2.5) tackles the practical issue of computation time.

2.2. *Comparison in expectation.* In this section, we compare $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right)\right]$ and $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right)\right]$. We note that these expectations exist in the Gaussian (GA) and the bounded (BA) cases provided

that ϕ is measurable and bounded by a ℓ_p-norm. Otherwise, in particular in Propositions 2.3 and 2.4, we assume that these expectations exist.

In the Gaussian case, these quantities are equal up to a factor that depends only on the distribution of W:

Proposition 2.3 *Let* $\mathbf{Y}$ *be a sample satisfying* (GA) *and let* W *be a resampling weight vector. Then, for any measurable positive-homogeneous function* $\phi : \mathbb{R}^K \to \mathbb{R}$*, we have the following equality:*

$$B_W \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right)\right] = \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right)\right] \quad . \tag{12}$$

If the weights are such that $\sum_{i=1}^n (W_i - \overline{W})^2 = n$ *, then the above equality holds for any function* ϕ *(and* $B_W = 1$*).*

For some classical weights, we give bounds or exact expressions for B_W on Table 1. In general, we can compute the value of B_W by simulation. Note that in a non-Gaussian framework, the constant B_W is still of interest, in an asymptotical sense: Theorem 3.6.13 in [31] uses the limit of B_W when n goes to infinity as a normalizing constant.

When the sample is only assumed to have a symmetric distribution, we obtain the following inequalities:

Proposition 2.4 *Let* $\mathbf{Y}$ *be a sample satisfying* (SA)*,* W *an exchangeable resampling weight vector and* $\phi : \mathbb{R}^K \to \mathbb{R}$ *any subadditive, positive-homogeneous function.*

(i) We have the general following lower bound:

$$A_W \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right)\right] \leq \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right)\right] \quad . \tag{13}$$

(ii) If the weight vector satisfies the assumption of (6)*, we have the following upper bound:*

$$D_W \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right)\right] \geq \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right)\right] \quad . \tag{14}$$

The bounds (13) and (14) are tight (that is, $A_W/D_W \to 1$ as $n \to \infty$) for some classical weights, see Table 1. When $\mathbf{Y}$ is not assumed to have a symmetric distribution and $\overline{W} = 1$ a.s., Proposition 2 of [13] showed that (13) holds with A_W replaced by $\mathbb{E}(W_1 - \overline{W})_+$. Therefore, assumption (SA) allows us to get a tighter result (for instance twice sharper with Efron or Rademacher weights). It can be shown (see [1], Chapter 9) that this factor 2 is unavoidable in general for a fixed n when (SA) is not satisfied, although it is unnecessary when n goes to infinity. We conjecture that an inequality close to (13) holds under an assumption less restrictive than (SA) (for instance, concerning an appropriate measure of skewness of the distribution of $\mathbf{Y}^1$).

2.3. *Concentration around the expectation.* In this section, we present concentration results for the two processes $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ and $\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)\right]$.

Proposition 2.5 *Let $p \in [1,\infty]$, $\mathbf{Y}$ a sample satisfying (GA) and $\phi : \mathbb{R}^K \to \mathbb{R}$ be any subadditive function, bounded by the ℓ_p-norm.*

(i) For all $\alpha \in (0,1)$, with probability at least $1-\alpha$ the following holds:

$$\phi\left(\overline{\mathbf{Y}} - \mu\right) < \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right)\right] + \frac{\|\sigma\|_p \,\overline{\Phi}^{-1}(\alpha/2)}{\sqrt{n}} \ , \tag{15}$$

and the same bound holds for the corresponding lower deviations.

(ii) Let W be an exchangeable resampling weight vector. Then, for all $\alpha \in (0,1)$, with probability at least $1-\alpha$ the following holds:

$$\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)\right] < \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)\right] + \frac{\|\sigma\|_p \, C_W \overline{\Phi}^{-1}(\alpha/2)}{n} \ , \tag{16}$$

and the same bound holds for the corresponding lower deviations.

The bound (15) with a remainder in $n^{-1/2}$ is classical; this order in n cannot be improved, as seen for example by taking $K = 1$ and ϕ the identity function. The bound (16) is more interesting because it illustrates one of the key properties of resampling, the "stabilization effect": the resampled expectation concentrates much faster to its expectation than the original quantity. This effect is known and has been studied asymptotically (in fixed dimension) using Edgeworth expansions (see [15]); here we demonstrate its validity non-asympotically in a specific case (see also Section 4.2 below for additional discussion).

In the bounded case, the next proposition is a minor variation of a result by Fromont. It is a consequence of McDiarmid's inequality [25]; we refer the reader to [13] (Proposition 1) for a proof.

Proposition 2.6 *Let $p \in [1,\infty]$, $M > 0$, $\mathbf{Y}$ a sample satisfying $(BA)(p,M)$ and $\phi : \mathbb{R}^K \to \mathbb{R}$ be any subadditive function, bounded by the ℓ_p-norm.*

(i) For all $\alpha \in (0,1)$, with probability at least $1-\alpha$ the following holds:

$$\phi\left(\overline{\mathbf{Y}} - \mu\right) < \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right)\right] + \frac{M}{\sqrt{n}}\sqrt{\log\left(1/\alpha\right)} \ , \tag{17}$$

and the same bound holds for the corresponding lower deviations.

(ii) Let W be an exchangeable resampling weight vector. Then, for all $\alpha \in (0,1)$, with probability at least $1-\alpha$ the following holds:

$$\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)\right] < \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)\right] + \frac{A_W M}{\sqrt{n}}\sqrt{\log\left(1/\alpha\right)} \ , \tag{18}$$

and the same bound holds for the corresponding lower deviations.

Efron	$2\left(1-\frac{1}{n}\right)^n = A_W \leq B_W \leq \sqrt{\frac{n-1}{n}} \quad C_W = 1$
Efron, $n \to +\infty$	$\frac{2}{e} = A_W \leq B_W \leq 1 = C_W$
Rademacher	$1-\frac{1}{n} = A_W \leq B_W \leq \sqrt{1-\frac{1}{n}} \quad C_W = 1 \leq D_W \leq 1+\frac{1}{\sqrt{n}}$
Rad., $n \to +\infty$	$A_W = B_W = C_W = D_W = 1$
rho(q)	$A_W = 2\left(1-\frac{q}{n}\right) \quad B_W = \sqrt{\frac{n}{q}-1}$
	$C_W = \sqrt{\frac{n}{n-1}}\sqrt{\frac{n}{q}-1} \quad D_W = \frac{n}{2q} + \left\lvert 1-\frac{n}{2q}\right\rvert$
rho($n/2$)	$A_W = B_W = D_W = 1 \quad C_W = \sqrt{\frac{n}{n-1}}$
Leave-one-out	$\frac{2}{n} = A_W \leq B_W = \frac{1}{\sqrt{n-1}} \quad C_W = \frac{\sqrt{n}}{n-1} \quad D_W = 1$
regular V-fcv	$A_W = \frac{2}{V} \leq B_W = \frac{1}{\sqrt{V-1}} \quad C_W = \sqrt{n}(V-1)^{-1} \quad D_W = 1.$

TABLE 1
Resampling constants for some classical resampling weight vectors.

2.4. *Resampling weight vectors.* In this section, we consider the question of choosing some appropriate exchangeable resampling weight vector W when using Theorem 2.1 or Corollary 2.2. We define the following resampling weight vectors:

1. **Rademacher**: W_i i.i.d. Rademacher variables, that is, $W_i \in \{-1, 1\}$ with equal probabilities.
2. **Efron** (Efron's bootstrap weights): W has a multinomial distribution with parameters $(n; n^{-1}, \ldots, n^{-1})$.
3. **Random hold-out (q)** (rho(q) for short), $q \in \{1, \ldots, n\}$: $W_i = \frac{n}{q}\mathbb{1}_{i\in I}$, where I is uniformly distributed on subsets of $\{1, \ldots, n\}$ of cardinality q. These weights may also be called cross validation weights, or leave-$(n-q)$-out weights. A classical choice is $q = n/2$ (assuming n is even). When $q = n-1$, these weights are called **leave-one-out** weights. Note that this resampling scheme is a particular case of subsampling.

As noticed in the introduction, the first example is common in learning theory while the second is classical in the framework of the resampling literature [23, 28]. Random hold-out weights have the particularity to be related to both: they are non negative and satisfy $\sum_i W_i = n$ a.s., and they come from a data-splitting idea (choosing I amounts to choose a subsample), upon which the cross-validation idea has been built. This analogy motivates the "V-fold cross-validation weights" (defined in Sect. 2.5), in order to reduce the computational complexity of the procedures proposed here.

For these classical weights, exact or approximate values for the quantities A_W, B_W, C_W and D_W (defined by equations (3) to (6)) can be easily derived (see Table 1). Proofs are given in Section 5.3, where several other weights are considered. Now, to use Theorem 2.1 or Corollary 2.2, we have to choose a particular resampling weight vector. In the Gaussian case, we propose the following accuracy and complexity criteria:

- first, relation (7) suggests that the quantity $C_W B_W^{-1}$ can be proposed as *accuracy* index for W. Namely, this index enters directly in the deviation term of the upper bound (while we know from Proposition 2.3 that the expectation term is exact) so that the smaller this index is, the sharper the bound.

Resampling	$C_W B_W^{-1}$ (accuracy)	$\lvert\operatorname{supp}\mathcal{D}(W)\rvert$ (complexity)
Efron	$\leq \frac{1}{2}\left(1-\frac{1}{n}\right)^{-n} \xrightarrow[n\to\infty]{} \frac{e}{2}$	$\binom{2n-1}{n-1} = \Theta(n^{-\frac{1}{2}}4^n)$
Rademacher	$\leq n/(n-1) \xrightarrow[n\to\infty]{} 1$	2^n
rho $(n/2)$	$= \sqrt{\frac{n}{n-1}} \xrightarrow[n\to\infty]{} 1$	$\binom{n}{n/2} = \Theta(n^{-1/2}2^n)$
Leave-one-out	$= \sqrt{\frac{n}{n-1}} \xrightarrow[n\to\infty]{} 1$	n
regular V-fcv	$= \sqrt{\frac{n}{V-1}}$	V

TABLE 2
Choice of the resampling weight vectors: accuracy-complexity trade-off.

- secondly, an upper bound on the computational burden to compute exactly the resampling quantity is given by the cardinality of the support of $\mathcal{D}(W)$, thus providing a *complexity* index. These two criteria are estimated in Table 2 for classical weights. For any exchangeable weight vector W, we have $C_W B_W^{-1} \geq [n/(n-1)]^{1/2}$ and the cardinality of the support of $\mathcal{D}(W)$ is larger than n. Therefore, the *leave-one-out weights* satisfy the best accuracy-complexity trade-off among exchangeable weights.

2.5. *Practical computation of the thresholds.* In practice, the exact computation of the resampling quantity $\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)\right]$ can still be too complex for the weights defined above. In this section, we consider two possible ways to address this issue. First, it is possible to use non-exchangeable weights with a lower complexity index and for which the exact computation is tractable. Alternatively, we propose to use a Monte-Carlo approximation, as is often done in practice to compute resampled quantities. In both cases, the thresholds have to be made slightly larger in order to keep a rigourous non-asymptotic control on the level. This is detailed in the two paragraphs below.

2.5.1. *V-fold cross-validation weights.* In order to reduce the computation complexity, we can use "piecewise exchangeable" weights: consider a regular partition $(B_j)_{1\leq j\leq V}$ of $\{1,\dots,n\}$ (where $V \in \{2,\dots,n\}$ and V divides n), and define the weights $W_i = \frac{V}{V-1}\mathbb{1}_{i\notin B_J}$ with J uniformly distributed on $\{1,\dots,V\}$. These weights are called the *(regular) V-fold cross validation* weights (V-fcv for short).

By applying our previous results to the process $(\widetilde{\mathbf{Y}}^j)_{1\leq j\leq V}$ where $\widetilde{\mathbf{Y}}^j := \frac{V}{n}\sum_{i\in B_j}\mathbf{Y}^i$ is the empirical mean of $\mathbf{Y}$ on block B_j, we can show that Theorem 2.1 can be extended to (regular) V-fold cross validation weights with the following resampling constants:

$$A_W = \frac{2}{V} \quad B_W = \frac{1}{\sqrt{V-1}} \quad C_W = \frac{\sqrt{n}}{V-1} \quad D_W = 1 \ .$$

Additionally, when V does not divide n and the blocks are no longer regular, Theorem 2.1 can also be generalized, but the constants have more complex expressions (see Section 10.7.5 in [1] for details). With V-fcv weights, the complexity index is only V, but we lose a factor $[(n-1)/(V-1)]^{1/2}$ in the accuracy index. With regard to the accuracy/complexity tradeoff, the

most accurate cross-validation weights are leave-one-out ($V = n$), whereas the 2-fcv weights are the best from the computational viewpoint (but also the less accurate). The choice of V is thus a trade-off between these two terms and depends on the particular constraints of each problem.

However, it is worth noting that as far as the bound of inequality (7) is concerned, it is not necessarily indispensable to aim for an accuracy index close to 1. Namely, this will result in a corresponding deviation term or order n^{-1}, while there is additionally another unavoidable deviation term or order $n^{-\frac{1}{2}}$ in the bound. This suggests that an accuracy index of order $o(n^{\frac{1}{2}})$ would actually be sufficient (as n grows large). In other words, using V-fcv with V "large" (for instance, $V = \Theta(\log(n))$) would result in only a negligible loss of overall accuracy as compared to leave-one-out. Of course, this discussion is specific to the form of the bound (7). We cannot formally exclude that a different approach could lead to a different conclusion, unless it can be proved that the deviation terms in (7) cannot be significantly improved, which is an issue we don't address here.

2.5.2. *Monte-Carlo approximation.* When using a Monte-Carlo approximation to evaluate $\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)\right]$, we draw randomly a number B of i.i.d. weight vectors $W^1, \ldots, W^B$ and compute $\frac{1}{B}\sum_{j=1}^{B}\phi\left(\overline{\mathbf{Y}}^{\langle W^j-\overline{W^j}\rangle}\right)$. This method is quite standard in the bootstrap literature and can be improved in several ways (see for instance [15], appendix II).

On the one hand, the number B of draws of W should be taken small enough so that B times the computational cost of evaluating $\phi\left(\overline{\mathbf{Y}}^{\langle W^j-\overline{W^j}\rangle}\right)$ is still tractable. On the other hand, the number B should be taken large enough to make the Monte-Carlo approximation accurate. In our framework, this is quantified more precisely by the following proposition (for bounded weights).

Proposition 2.7 *Let $B \geq 1$ and $W^1, \ldots, W^B$ be i.i.d. exchangeable resampling weight vectors such that $W_1^1 - \overline{W^1} \in [c_1, c_2]$ a.s. Let $p \in [1, \infty]$, $\phi : \mathbb{R}^K \to \mathbb{R}$ be any subadditive function, bounded by the ℓ_p-norm. If $\mathbf{Y}$ is a fixed sample, for every $\beta \in (0,1)$,*

$$\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)\right] \leq \frac{1}{B}\sum_{j=1}^{B}\phi\left(\overline{\mathbf{Y}}^{\langle W^j-\overline{W^j}\rangle}\right) + (c_2-c_1)\sqrt{\frac{\log(\beta^{-1})}{2B}}\,\|\widetilde{\sigma}\|_p \tag{19}$$

holds with probability at least $1-\beta$, where $\widetilde{\sigma}$ denotes the vector of average absolute deviations to the median, $\widetilde{\sigma} := \left(\left(\frac{1}{n}\sum_{i=1}^{n}|\mathbf{Y}_k^i - M_k|\right)\right)_{1\leq k\leq K}$ (M_k denoting a median of $(\mathbf{Y}_k^i)_{1\leq i\leq n}$).

As a consequence, Proposition 2.7 proposes an explicit correction of the concentration thresholds taking into account B bounded weight vectors. For instance, with Rademacher weights, we can use (19) with $c_2 - c_1 = 2$ and $\beta = \gamma\alpha$ ($\gamma \in (0,1)$). Then, in the thresholds built upon Theorem 2.1 and Proposition 2.2, one can replace $\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)\right]$ by its Monte-Carlo approximation at the price of changing α into $(1-\gamma)\alpha$, and adding $B_W^{-1}\sqrt{\frac{2\log((\gamma\alpha)^{-1})}{B}}\,\|\widetilde{\sigma}\|_p$ to the threshold.

As n grows large, this remainder term is negligible in front of the main one when B is (for instance) of order n^2. In practical applications, B can be chosen as a function of $\mathbf{Y}$ because (19) holds conditionally to the observed sample. Therefore, we can use the following strategy: first, compute a rough estimate $t_{\mathrm{est},\alpha}$ of the final threshold (for instance, if $\phi = \|\cdot\|_\infty$ and $\mathbf{Y}$ is Gaussian, take the Bonferroni threshold (10)). Then, choose $B \gg t_{\mathrm{est},\alpha}^2 \, \|\widetilde{\sigma}\|_p^2 \log\left((\gamma\alpha)^{-1}\right)$.

3. Confidence region using resampled quantiles.

3.1. *Main result.* In this section, we consider a different approach to construct confidence regions, directly based on the estimation of the quantile via resampling. Once again, since we aim at a non-asymptotic result for $K \gg n$, the standard asymptotic approaches cannot be applied here. For this reason, we base the proposed results on ideas coming from exact randomized tests and consider here the case where $\mathbf{Y}^1$ has a symmetric distribution and where W is an i.i.d Rademacher weight vector, that is, weights are i.i.d. with $\mathbb{P}(W_i = 1) = \mathbb{P}(W_i = -1) = 1/2$.

The resampling idea applied here is to approximate the quantiles of the distribution $\mathcal{D}\left(\phi\left(\overline{\mathbf{Y}} - \mu\right)\right)$ by the quantiles of the corresponding resampling-based distribution:

$$\mathcal{D}\left(\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right)\middle|\mathbf{Y}\right) = \mathcal{D}\left(\phi\left(\overline{\left(\mathbf{Y} - \overline{\mathbf{Y}}\right)}^{\langle W\rangle}\right)\middle|\mathbf{Y}\right) \quad .$$

For this, we take advantage of the symmetry of each $\mathbf{Y}^i$ around its mean. Let us define for a function ϕ the resampled empirical quantile by:

$$q_\alpha(\phi, \mathbf{Y}) := \inf\left\{x \in \mathbb{R} \,\middle|\, \mathbb{P}_W\left(\phi(\overline{\mathbf{Y}}^{\langle W\rangle}) > x\right) \leq \alpha\right\} \quad . \tag{20}$$

The following lemma, close in spirit to exact test results, easily derives from the "symmetrization trick", that is, from taking advantage of the distribution invariance of the data via sign-flipping.

Lemma 3.1 *Let $\mathbf{Y}$ be a data sample satisfying assumption (SA) and $\phi : \mathbb{R}^K \to \mathbb{R}$ be a measurable function. Then the following holds:*

$$\mathbb{P}\left(\phi(\overline{\mathbf{Y}} - \mu) > q_\alpha\left(\phi, \mathbf{Y} - \mu\right)\right) \leq \alpha \quad . \tag{21}$$

Of course, since $q_\alpha\left(\phi, \mathbf{Y} - \mu\right)$ still depends on the unknown μ, we cannot use this threshold to get a confidence region of the form (1). It is in principle possible to build a confidence region directly from Lemma 3.1 by using the duality beween tests and confidence regions, but this would be difficult to compute, and not of the desired form (1). Therefore, following the general philosophy of resampling, we propose to replace the true mean μ by the empirical mean $\overline{\mathbf{Y}}$ in the quantile $q_\alpha\left(\phi, \mathbf{Y} - \mu\right)$. The main technical result of this section gives a non-asymptotic bound on the price to perform this operation:

Theorem 3.2 *Fix $\delta, \alpha_0 \in (0,1)$. Let $\mathbf{Y}$ be a data sample satisfying assumption (SA). Let $f : \left(\mathbb{R}^K\right)^n \to [0,\infty)$ be a nonnegative function. Let $\phi : \mathbb{R}^K \to \mathbb{R}$ be a nonnegative, subadditive,and positive-homogeneous function. Denote $\widetilde{\phi}(x) := \max\left(\phi(x), \phi(-x)\right)$. The following holds:*

$$\mathbb{P}\left(\phi(\overline{\mathbf{Y}} - \mu) > q_{\alpha_0(1-\delta)}\left(\phi, \mathbf{Y} - \overline{\mathbf{Y}}\right) + \gamma_1(\alpha_0\delta) f(\mathbf{Y})\right) \;\leq\; \alpha_0 + \mathbb{P}\left(\widetilde{\phi}(\overline{\mathbf{Y}} - \mu) > f(\mathbf{Y})\right) \;, \tag{22}$$

where $\gamma_1(\eta) := \frac{2\overline{\mathcal{B}}(n,\frac{\eta}{2}) - n}{n}$ and $\overline{\mathcal{B}}(n,\eta) := \max\left\{k \in \{0,\dots,n\} \middle| 2^{-n}\sum_{i=k}^{n}\binom{n}{i} \geq \eta\right\}$ is the upper quantile function of a Binomial $(n, \frac{1}{2})$ variable.

In the above result, the resampled quantile term $q_{\alpha_0(1-\delta)}\left(\phi, \mathbf{Y} - \overline{\mathbf{Y}}\right)$ should be interpreted as the main term of the threshold, and the rest, involving the function f, a remainder term. In the usual resampling philosophy, one would only consider the main term at the target level, that is, $\alpha_0 = \alpha$ and $\delta = 0$. Here, the additional remainder terms are introduced to account rigorously for the validity of the result in a non-asymptotic setting. These remainder terms have two effects: first, the resampled quantile in the main term is computed at a "shrunk" error level $\alpha_0(1-\delta) < \alpha$, and secondly, there is an additional additive term in the threshold itself.

The role of the parameters δ, α_0 and f is to strike a balance between these effects. Generally speaking, f should be an available upper bound on a quantile of $\widetilde{\phi}(\overline{\mathbf{Y}} - \mu)$ at a level $\alpha_1 \ll \alpha_0$. On the left-hand side, f appears in the threshold with the factor γ_1, which can be more explicitly bounded by

$$\gamma_1(\alpha_0\delta) \leq \left(\frac{2\log\left(\frac{2}{\alpha_0\delta}\right)}{n}\right)^{1/2} , \tag{23}$$

using Hoeffding's inequality. The above result therefore transforms a possibly coarse "a priori" bound f on quantiles into a more accurate quantile bound based on a main term estimated by resampling and a remainder term based on f multiplied by a small factor.

In order to get a clearer insight, let us consider an example of specific choices for the parameters δ, α_0 and f in the Gaussian case. First, choose $\delta = \Theta(n^{-\gamma})$ and $\frac{\alpha_0}{\alpha} = 1 - \Theta(n^{-\gamma})$ for some $\gamma > 0$, say $\gamma = 1$. This way, the main term is the resampled quantile at level $\alpha_0(1-\delta) = \alpha(1-\Theta(n^{-\gamma}))$. For the choice of f, let us pick Bonferroni's threshold (10) at level $\alpha_1 = (\alpha - \alpha_0) = \Theta(n^{-\gamma})$, so that the overall probability control in (22) is really at the target level α. Then $f_{\mathrm{Bonf}}(\mathbf{Y}) \leq \Theta((\log(Kn^\gamma)/n)^{\frac{1}{2}})$, and, using (23), we conclude that the remainder term is bounded by $\Theta(\log(Kn^\gamma)/n)$. This is indeed a remainder term with respect to the main term which is of order at least $\Theta(n^{-\frac{1}{2}})$ as n grows (assuming that the dimension $K(n)$ grows sub-exponentially with n).

There are other possibilities to choose f depending on the context: the Bonferroni threshold can be adapted correspondingly to the non-Gaussian case when an upper bound on the tail of each coordinate is available. This still makes the remainder term directly dependent on K, and

a possibly more interesting idea is to recycle the results of Section 2 (when the data is either Gaussian or bounded and symmetric) and plug in the thresholds derived there for the function f.

Finally, if the a priori bound on the quantiles is too coarse, it is possible to iterate the process and estimate smaller quantiles more accurately using resampling again. Namely, by iteration of Theorem 3.2, we obtain the following corollary:

Corollary 3.3 *Fix J a positive integer, $(\alpha_i)_{i=0,\dots,J-1}$ a finite sequence in $(0,1)$ and $\delta \in (0,1)$. Consider $\mathbf{Y}$, f, ϕ and $\widetilde{\phi}$ as in Theorem 3.2. Then the following holds:*

$$\mathbb{P}\left(\phi(\overline{\mathbf{Y}} - \mu) > q_{\alpha_0(1-\delta)}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + \sum_{i=1}^{J-1} \gamma_i q_{\alpha_i(1-\delta)}(\widetilde{\phi}, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma_J f(\mathbf{Y})\right) \leq \sum_{i=0}^{J-1} \alpha_i + \mathbb{P}\left(\widetilde{\phi}(\overline{\mathbf{Y}} - \mu) > f(\mathbf{Y})\right) \quad , \tag{24}$$

where, for $k \geq 1$, $\gamma_k := n^{-k} \prod_{i=0}^{k-1} \left(2\overline{\mathcal{B}}\left(n, \frac{\alpha_i \delta}{2}\right) - n\right)$.

The rationale behind this result is that the sum appearing inside the probability in (24) should be interpreted as a series of corrective terms of decreasing order of magnitude, because we expect the sequence γ_k to be sharply decreasing. From (23), this will be the case if the levels are such that $\alpha_i \gg \exp(-n)$.

The conclusion is that, even if the *a priori* available bound f on small quantiles is not sharp, its contribution to the threshold can be made small in comparison to the (more accurate) resampling terms. The counterpart to pay is the loss in the level and the additional terms in the threshold; for large n, these terms decay very rapidly, but for small n, they may still result in a non negligible contribution; in this case a precise tuning for the parameters $J, (\alpha_i), \delta$ and f is of much more importance and also more delicate.

At this point, we should also mention that the remainder terms given by Theorem 3.2 and Corollary 3.3 are certainly overestimated, even if f is very well chosen. This makes the theoretical thresholds slightly too conservative in general (particularly for small values of n). From simulations not reported here (see [2] and Section 4.3 below), it even appears that the remainder terms could be (almost) unnecessary in standard situations, even for n relatively small. Proving this fact rigorously in a non-asymptotic setting, possibly with some additional assumption on the distribution of $\mathbf{Y}$, remains an open issue. Another interesting open problem would be to obtain a self-contained result based on the symmetry assumption (SA) alone (or a negative result proving that (SA) is not sufficient for a distribution-free result of this form).

3.2. *Practical computation of the resampled quantile.* Since the above results use Rademacher weight vectors, the exact computation of the quantile q_α requires in principle 2^n iterations and

thus is too complex as n becomes large. Parallel to what was proposed for the concentration-based thresholds in Section 2.5, one can, as a first solution, consider a block-wise Rademacher resampling scheme, or, equivalently, applying the previous method to a block-averaged sample, at the price of a possibly substantial loss in accuracy.

A possibly better way to address this issue is by Monte-Carlo quantile approximation, on which we focus now. Let $\mathbf{W}$ denote a $n \times B$ matrix of i.i.d. Rademacher weights (independent of all other variables), and define

$$\widetilde{q}_\alpha(\phi, \mathbf{Y}, \mathbf{W}) := \inf\left\{ x \in \mathbb{R} \,\middle|\, \frac{1}{B}\sum_{j=1}^{B} \mathbb{1}\left\{ \phi\left(\overline{\mathbf{Y}}^{\langle \mathbf{W}^j \rangle} \right) \geq x \right\} \leq \alpha \right\},$$

that is, $\widetilde{q}_\alpha$ is defined just as q_α except that the true distribution $\mathbb{P}_W$ of the Rademacher weight vector is replaced by the empirical distribution constructed from the columns of $\mathbf{W}$, $\widetilde{\mathbb{P}}_{\mathbf{W}} = B^{-1}\sum_{j=1}^{B} \delta_{\mathbf{W}^j}$; note that the strict inequality $\phi\left(\overline{\mathbf{Y}}^{\langle \mathbf{W} \rangle}\right) > x$ in (20) was replaced by $\phi\left(\overline{\mathbf{Y}}^{\langle \mathbf{W}^j \rangle}\right) \geq x$ for technical reasons. The following result then holds:

Proposition 3.4 *Consider the same conditions as in Theorem 3.2 except the function f can now be a function of both $\mathbf{Y}$ and $\mathbf{W}$. We have:*

$$\begin{aligned} \mathbb{P}_{\mathbf{Y},\mathbf{W}}\left(\phi(\overline{\mathbf{Y}} - \mu) > \widetilde{q}_{\alpha_0(1-\delta)}\left(\phi, \mathbf{Y} - \overline{\mathbf{Y}}, \mathbf{W} \right) + \gamma(\mathbf{W}, \alpha_0\delta) f(\mathbf{Y}, \mathbf{W}) \right) \\ \leq \widetilde{\alpha}_0 + \mathbb{P}_{\mathbf{Y},\mathbf{W}}\left(\widetilde{\phi}(\overline{\mathbf{Y}} - \mu) > f(\mathbf{Y}, \mathbf{W}) \right) \ , \end{aligned}$$

where $\widetilde{\alpha}_0 := \dfrac{\lfloor B\alpha_0 \rfloor + 1}{B+1} \leq \alpha_0 + \dfrac{1}{B+1}$ *and* $\gamma(\mathbf{W}, \eta) := \max\left\{ y \geq 0 \,\middle|\, \dfrac{1}{B}\sum_{j=1}^{B} \mathbb{1}\left\{ \left|\overline{\mathbf{W}^j}\right| \geq y \right\} \geq \eta \right\}$

is the $(1-\eta)$-quantile of $\left|\overline{W}\right|$ under the empirical distribution $\widetilde{\mathbb{P}}_{\mathbf{W}}$.

Note that for practical purposes, we can choose $f(\mathbf{W}, \mathbf{Y})$ to depend on $\mathbf{Y}$ only and use another type of bound to control the last term on the right-hand side, as in the earlier discussion. The above result tells us that if we replace in Theorem 3.2 the true quantile by an empirical quantile based on B i.i.d. weight vectors, and the factor γ_1 is similarly replaced by an empirical quantile of $|\overline{W}|$, then we lose at most $(B+1)^{-1}$ in the corresponding covering probability. Furthermore, it can be seen easily that if α_0 is taken to be a positive multiple of $(B+1)^{-1}$, then there is no loss in the final covering probability (that is, $\widetilde{\alpha}_0 = \alpha_0$).

4. Discussions and concluding remarks.

4.1. *Estimating $\|\sigma\|_p$.* In the concentration approach and in the Gaussian case, the derived thresholds depend explicitly on the ℓ_p-norm of the vector of standard deviations $\sigma = (\sigma_k)_k$ (an upper bound on this quantity can be used as well). While we have left aside the problem of

determining this parameter if no prior information is available, it is possible to estimate σ by its empirical counterpart

$$\widehat{\sigma} := \left(\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{Y}_k^i - \overline{\mathbf{Y}}_k \right)^2} \right)_{1 \leq k \leq K} .$$

Interestingly, the quantity $\|\widehat{\sigma}\|_p$ enjoys the same type of concentration property as the resampled expectations considered in Section 2.3, so that we can derive, by a similar argument, a *dimension-free* confidence bound for $\|\sigma\|_p$:

Proposition 4.1 *Assume that* $\mathbf{Y}$ *satisfies* (GA). *Then, with probability at least* $1 - \delta$,

$$\|\sigma\|_p \leq \left(C_n - \frac{1}{\sqrt{n}} \overline{\Phi}^{-1} \left(\frac{\delta}{2} \right) \right)^{-1} \|\widehat{\sigma}\|_p \quad , \tag{25}$$

where $C_n = \sqrt{\frac{2}{n}} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}$.

It can easily be checked via Stirling's formula that $C_n = 1 - O(n^{-1})$, so that replacing $\|\sigma\|_p$ by the above upper bound does not make the corresponding thresholds significantly more conservative.

A similar question holds for the parameter M in the bounded case. In practical applications, an absolute bound on the possible data values is often known (for instance from physical or biological constraints). It can also be estimated, but it seems harder to obtain a rigourous non-asymptotic control on the level of the resulting threshold in the general bounded case.

A different and potentially more important problem arises if the vector of variances σ is not constant. Since the confidence regions proposed in this paper are isotropic, they will — inevitably — tend to be conservative when the variances of the coordinates are very different. The standard way to address this issue is to consider studentized data. While this would solve this heteroscedasticity issue, it also voids the assumption of independent datapoints – a crucial assumption in all of our proofs. Therefore, generalizing our results to studentized observations is an important, but probably challenging, direction for future work.

4.2. *Interpretation and use of ϕ-confidence regions.* We have built high-dimensional confidence regions taking the form of "ϕ-balls" (where ϕ can be any ℓ^p-norm with $p \geq 1$, but more general choices are possible, such as $\phi(x) = \sup_k (x_k)_+$). Such confidence regions in very high dimension are certainly quite difficult to visualize and one can ask how they have to be interpreted. In our opinion, the most intuitive and interesting interpretation again comes from learning theory, by regarding ϕ as a type of loss function. In this sense, a ϕ-confidence region is an upper confidence bound on some relevant loss measure of the estimator $\mathbf{Y}$ to the target μ . Additionally, in the particular case when $\phi = \sup_k (x_k)_+$ or $\phi = \|\cdot\|_\infty$, the corresponding regions can be interpreted as simultaneous confidence intervals over all coordinate means.

The results presented here can also provide confidence intervals for the ℓ^p-risk (that is, the averaged ϕ-loss) for the estimator $\overline{\mathbf{Y}}$ of the mean vector μ. Indeed, combining (12) and Proposition 2.5 (ii), we derive that for a Gaussian sample $\mathbf{Y}$ and any $p \in [1, \infty]$, the following upper bound holds with probability at least $1 - \alpha$:

$$\mathbb{E}\left\|\overline{\mathbf{Y}} - \mu\right\|_p < \frac{\mathbb{E}_W\left[\left\|\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right\|_p\right]}{B_W} + \frac{\|\sigma\|_p C_W}{nB_W}\overline{\Phi}^{-1}(\alpha/2) \ , \tag{26}$$

and a similar lower bound holds. It is worth noticing that the rate $C_W/(nB_W)$ is close to n^{-1} for most of the weights, meaning that resampling provides a much better estimate of $\mathbb{E}\left\|\overline{\mathbf{Y}} - \mu\right\|_p$ than $\left\|\overline{\mathbf{Y}} - \mu\right\|_p$ itself. This stabilization effect of resampling is well-known in standard asymptotical settings (see for instance [15]).

The ℓ^p-risk is also related to the leave-one-out estimation of the prediction risk. Indeed, consider using $\overline{\mathbf{Y}}$ for *predicting* a new data point $\mathbf{Y}^{n+1} \sim \mathbf{Y}^1$ (independent of $\mathbf{Y} = (\mathbf{Y}^1, \ldots, \mathbf{Y}^n)$). The corresponding ℓ^p-prediction risk is given by $\mathbb{E}\left\|\overline{\mathbf{Y}} - \mathbf{Y}^{n+1}\right\|_p$. In the Gaussian setting, this prediction risk is proportional to the ℓ^p-risk: $\mathbb{E}\left\|\overline{\mathbf{Y}} - \mu\right\|_p = (n+1)^{\frac{1}{2}}\mathbb{E}\left\|\overline{\mathbf{Y}} - \mathbf{Y}^{n+1}\right\|_p$, so that the previous resampling estimator of the ℓ^p-risk also leads to an estimator of the prediction risk. In particular, using leave-one-out weights and noting $\overline{\mathbf{Y}}^{(-i)}$ the mean of the $(\mathbf{Y}^j, j \neq i, 1 \leq j \leq n)$, our results prove that the leave-one-out estimator

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\overline{\mathbf{Y}}^{(-i)} - \mathbf{Y}^i\right\|_p$$

correctly estimates the prediction risk (up to the factor $(1 - 1/n^2)^{\frac{1}{2}} \simeq 1$).

Finally, another important field of application is hypothesis testing. When $\phi = \sup_k (x_k)_+$ or $\phi = \|\cdot\|_\infty$, the thresholds derived here can be used to derive *multiple testing* procedures for the value of the mean of each coordinate. This question is extensively developed in the companion paper [2]. It is also possible to take advantage of the generality of our results, where ϕ is allowed to be any ℓ^p-norm with $p \geq 1$, for single global hypothesis testing. The confidence regions can be used straightforwardly to test several single global hypotheses, such as $\mu = \mu^\star$ against $\|\mu - \mu^\star\|_p \geq R > 0$. Depending on particular features of the problem, having the choice between different functions ϕ allows to take into account specific forms of alternative hypotheses in the construction of the threshold.

4.3. *Simulation study.* In the companion paper [2] (Section 4), a simulation study compares the thresholds built in this paper and Bonferroni's threshold, using $\phi = \|\cdot\|_\infty$, considering Gaussian data with different levels of correlations, and assuming the coordinate variance σ to be constant and known. Without entering into details, its general conclusions are the following.

First, all of the thresholds proposed in the present paper can improve on Bonferroni's when the correlations are strong enough. Even though our thresholds are seen to be more conservative than the "ideal" one (that is, the true quantile), they all exhibit adaptivity to the correlations, as expected from their construction. However, when the vector coordinates are close to being independent, the proposed thresholds are somewhat more conservative than Bonferroni's (the latter being essentially optimal in this case).

The second observation made on the simulations is that the quantile approach generally appears to be less conservative than the concentration approach. But the remaining advantage of the concentration approach is that it can be combined with Bonferroni's threshold (using Proposition 2.2) so that one can almost take "the less conservative of the two" and only suffer a negligible loss if the Bonferroni threshold turns out to be better. Remember also that the concentration threshold can be of use for the remainder terms of the quantile threshold.

Finally, we also tested the resampled quantile without remainder term (that is, taking the raw resampled quantile of the empirically centered data at the desired level, without modification). Although this threshold is not theoretically justified in the present work, it appeared to be very close to the ideal threshold in the performed simulations. This supports the conjecture that the remainder terms in the theoretical threshold could either be made significantly smaller, or possibly even completely dropped off in some cases.

4.4. *Comparing non-asymptotic and asymptotic approaches.* Although simulations have shown that the various thresholds proposed here can outperform Bonferroni's when significant correlations are present, we have also noticed that these thresholds are generally noticeably more conservative than the ideal ones (the true quantiles), especially for small values of n. Moreover, taking into account other sources of error such as the estimation of $\|\sigma\|_p$ as above, or Monte-Carlo approximations, will result in even more conservative thresholds. The main reason for this additional conservativeness is that our control on the level is *non-asymptotic*, that is, valid for every fixed K and n. In this sense, it would be somewhat unfair to compare the thresholds proposed here to those of "traditional" resampling theory, that are only proved to be valid asymptotically in n and for fixed K. The non-asymptotic results derived here can nevertheless also be used for an asymptotic analysis, in a setting where $K(n)$ is a function of n, and possibly rapidly (say, exponentially) growing. This type of situation seems to have been only scarcely touched by existing asymptotic approaches. In this sense, in practical situations we can envision to "cheat" somewhat and replace the theoretical thresholds by their leading component (under some mild assumptions on the growth of $K(n)$) as n tends to infinity. From a theoretical point of view, an interesting avenue for future endeavors is to prove that the thresholds considered here, while certainly not second-order correct, are at least asymptotically optimal under various dependency conditions.

5. Proofs.

5.1. *Confidence regions using concentration.* In this section, we prove all the statements of Section 2 except computations of resampling weight constants (made in Section 5.3).

5.1.1. *Comparison in expectation.*

Proof of Proposition 2.3. Denoting by $\mathbf{\Sigma}$ the common covariance matrix of the $\mathbf{Y}^i$, we have $\mathcal{D}\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle} \,\middle|\, W\right) = \mathcal{N}(0, (n^{-1}\sum_{i=1}^n (W_i-\overline{W})^2) n^{-1}\mathbf{\Sigma})$, and the result follows because $\mathcal{D}(\overline{\mathbf{Y}}-\mu) = \mathcal{N}(0, n^{-1}\mathbf{\Sigma})$ and ϕ is positive-homogeneous. This last assumption is of course unnecessary if it holds that $\sum_{i=1}^n (W_i-\overline{W})^2 = n$ a.s. ■

Proof of Proposition 2.4. By independence between W and $\mathbf{Y}$, exchangeability of W and the positive homogeneity of ϕ, for every realization of $\mathbf{Y}$ we have:

$$A_W\phi\left(\overline{\mathbf{Y}}-\mu\right) = \phi\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left|W_i-\overline{W}\right|\left(\mathbf{Y}^i-\mu\right)\,\middle|\,\mathbf{Y}\right]\right) \ .$$

Then, by convexity of ϕ,

$$A_W\phi\left(\overline{\mathbf{Y}}-\mu\right) \le \mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n \left|W_i-\overline{W}\right|\left(\mathbf{Y}^i-\mu\right)\right)\,\middle|\,\mathbf{Y}\right] \ .$$

We integrate with respect to $\mathbf{Y}$, and use the symmetry of the $\mathbf{Y}^i$ with respect to μ and again the independence between W and $\mathbf{Y}$ to show finally that

$$\begin{aligned} A_W\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}-\mu\right)\right] &\le \mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n \left|W_i-\overline{W}\right|\left(\mathbf{Y}^i-\mu\right)\right)\right] \\ &= \mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n \left(W_i-\overline{W}\right)\left(\mathbf{Y}^i-\mu\right)\right)\right] = \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)\right] . \end{aligned}$$

The point (ii) is proved via the following chain of inequalities:

$$\begin{aligned} \mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}^{\langle W-\overline{W}\rangle}\right)\right] &\le \mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n (W_i-x_0)(\mathbf{Y}^i-\mu)\right)\right] + \mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n (x_0-\overline{W})(\mathbf{Y}^i-\mu)\right)\right] \\ &= \mathbb{E}\left[\phi\Big(\frac{1}{n}\sum_{i=1}^n |W_i-x_0|(\mathbf{Y}^i-\mu)\Big)\right] + \mathbb{E}\left[\phi\Big(\frac{1}{n}\sum_{i=1}^n |x_0-\overline{W}|(\mathbf{Y}^i-\mu)\Big)\right] \\ &\le (a+\mathbb{E}|\overline{W}-x_0|)\mathbb{E}\left[\phi(\overline{\mathbf{Y}}-\mu)\right] . \end{aligned}$$

In the second line, we used as earlier the symmetry of the $\mathbf{Y}^i$ with respect to μ together with the independence of W and $\mathbf{Y}$. In the last inequality we used the assumption $|W_i-x_0|=a$ a.s. and the positive-homogeneity of ϕ. ■

5.1.2. *Concentration inequalities.*

Proof of Proposition 2.5. We use here concentration principles following closely the approach in [24], Section 3.2.4 . The essential ingredient is the Gaussian concentration theorem of Cirel'son, Ibragimov and Sudakov ([7] and recalled in [24], Theorem 3.8), stating that if F is a Lipschitz function on $\mathbb{R}^N$ with constant L, then for the standard Gaussian measure on $\mathbb{R}^N$ we have $\mathbb{P}\left(F \geq \mathbb{E}\left[F\right] + t\right) \leq 2\overline{\Phi}(t/L)$.

Let us denote by $\mathbf{A}$ a square root of the common covariance matrix of the $\mathbf{Y}^i$. If ζ_i is a K-dimensional, standard normal vector, then $\mathbf{A}\zeta_i$ has the same distribution as $\mathbf{Y}^i - \mu$. We let for all $\zeta \in \left(\mathbb{R}^K\right)^n$, $T_1(\zeta) := \phi\left(\frac{1}{n}\sum_{i=1}^n \mathbf{A}\zeta_i\right)$ and $T_2(\zeta) := \mathbb{E}\left[\phi\left(\frac{1}{n}\sum_{i=1}^n (W_i - \overline{W})\mathbf{A}\zeta_i\right)\right]$. If we endow $\left(\mathbb{R}^K\right)^n$ with the standard Gaussian measure, then T_1 , resp. T_2 , has the same distribution as $\phi\left(\overline{\mathbf{Y}} - \mu\right)$, resp. $\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right)$.

From the Gaussian concentration theorem recalled above, in order to reach the conclusion we therefore only need to establish that T_1 (resp. T_2) is a Lipschitz function with constant $\left\|\sigma\right\|_p/\sqrt{n}$ (resp. $\left\|\sigma\right\|_p C_W/n$) with respect to the Euclidean norm $\left\|\cdot\right\|_{2,Kn}$ on $\left(\mathbb{R}^K\right)^n$. Let $\zeta, \zeta' \in \left(\mathbb{R}^K\right)^n$ and denote by $(a_k)_{1\leq k\leq K}$ the rows of $\mathbf{A}$. Using that ϕ is 1-Lipschitz with respect to the ℓ_p-norm (because it is subadditive and bounded by the ℓ_p-norm), we get

$$\left|T_1(\zeta) - T_1(\zeta')\right| \leq \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{A}(\zeta_i - \zeta_i')\right\|_p = \left\|\left(\left\langle a_k, \frac{1}{n}\sum_{i=1}^n (\zeta_i - \zeta_i')\right\rangle\right)_k\right\|_p \ .$$

For each coordinate k, by Cauchy-Schwartz's inequality and since $\left\|a_k\right\|_2 = \sigma_k$, we deduce

$$\left|\left\langle a_k, \frac{1}{n}\sum_{i=1}^n (\zeta_i - \zeta_i')\right\rangle\right| \leq \sigma_k \left\|\frac{1}{n}\sum_{i=1}^n (\zeta_i - \zeta_i')\right\|_2 \ .$$

Therefore, we get

$$\left|T_1(\zeta) - T_1(\zeta')\right| \leq \left\|\sigma\right\|_p \left\|\frac{1}{n}\sum_{i=1}^n (\zeta_i - \zeta_i')\right\|_2 \leq \frac{\left\|\sigma\right\|_p}{\sqrt{n}} \left\|\zeta - \zeta'\right\|_{2,Kn} \ ,$$

using the convexity of $x \in \mathbb{R}^K \mapsto \left\|x\right\|_2^2$, and we obtain (i). For T_2, we use the same method as for T_1 to obtain:

$$\begin{aligned} \left|T_2(\zeta) - T_2(\zeta')\right| &\leq \left\|\sigma\right\|_p \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i')\right\|_2 \\ &\leq \frac{\left\|\sigma\right\|_p}{n} \sqrt{\mathbb{E}\left\|\sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i')\right\|_2^2} \ . \end{aligned} \tag{27}$$

Note that since $\left(\sum_{i=1}^n (W_i - \overline{W})\right)^2 = 0$, we have $\mathbb{E}(W_1 - \overline{W})(W_2 - \overline{W}) = -C_W^2/n$. We now develop $\left\|\sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i')\right\|_2^2$ in the Euclidean space $\mathbb{R}^K$:

$$\mathbb{E}\left\| \sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i') \right\|_2^2 = C_W^2 \left(1 - n^{-1}\right) \sum_{i=1}^n \|\zeta_i - \zeta_i'\|_2^2 - \frac{C_W^2}{n} \sum_{i \neq j} \left\langle \zeta_i - \zeta_i', \zeta_j - \zeta_j' \right\rangle$$
$$= C_W^2 \sum_{i=1}^n \|\zeta_i - \zeta_i'\|_2^2 - \frac{C_W^2}{n} \left\| \sum_{i=1}^n (\zeta_i - \zeta_i') \right\|_2^2 .$$

Consequently,

$$\mathbb{E}\left\| \sum_{i=1}^n \left(W_i - \overline{W}\right) (\zeta_i - \zeta_i') \right\|_2^2 \leq C_W^2 \sum_{i=1}^n \|\zeta_i - \zeta_i'\|_2^2 = C_W^2 \|\zeta - \zeta'\|_{2,Kn}^2 . \tag{28}$$

Combining expression (27) and (28), we find that T_2 is $\|\sigma\|_p C_W/n$-Lipschitz. ■

Remark 5.1 *The proof of Proposition 2.5 is still valid under the weaker assumption (instead of exchangeability of W) that $\mathbb{E}\left[(W_i - \overline{W})(W_j - \overline{W})\right]$ can only take two possible values depending on whether or not $i = j$.*

5.1.3. *Main results.*

Proof of Theorem 2.1. The case (BA)(p, M) and (SA) is obtained by combining Proposition 2.4 and 2.6. The (GA) case is a straightforward consequence of Proposition 2.3 and the proof of Proposition 2.5 (considering the Lipschitz function $T_1 - T_2$). ■

Proof of Proposition 2.2. From Proposition 2.5 (i), with probability at least $1 - \alpha(1 - \delta)$, $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ is less than or equal to the minimum between $t_{\alpha(1-\delta)}$ and $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right)\right] + \frac{\|\sigma\|_p \overline{\Phi}^{-1}(\alpha(1-\delta)/2)}{\sqrt{n}}$ (since both of these thresholds are deterministic). In addition, Proposition 2.3 and Proposition 2.5 (ii) give that with probability at least $1-\alpha\delta$, $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}} - \mu\right)\right] \leq \frac{\mathbb{E}_W\left[\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right)\right]}{B_W} + \frac{\|\sigma\|_p C_W}{B_W n} \overline{\Phi}^{-1}(\alpha\delta/2)$. The result follows by combining the two last expressions. ■

5.1.4. *Monte-Carlo approximation.*

Proof of Proposition 2.7. The idea of the proof is to apply McDiarmid's inequality (see [25]) conditionally to $\mathbf{Y}$. For any realizations W and W' of the resampling weight vector and any $\nu \in \mathbb{R}^k$,

$$\left|\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle}\right) - \phi\left(\overline{\mathbf{Y}}^{\langle W' - \overline{W'}\rangle}\right)\right| \leq \phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W}\rangle} - \overline{\mathbf{Y}}^{\langle W' - \overline{W'}\rangle}\right) \leq \frac{c_2 - c_1}{n} \left\| \left(\sum_{i=1}^n \left| \mathbf{Y}_k^i - \nu_k \right| \right)_k \right\|_p$$

since ϕ is sub-additive, bounded by the ℓ_p-norm and $W_i - \overline{W} \in [c_1, c_2]$ a.s.

The sample $\mathbf{Y}$ being deterministic, we can take ν_k equal to a median M_k of $(\mathbf{Y}_k^i)_{1\leq i\leq n}$. Since $W^1, \dots, W^B$ are independent, McDiarmid's inequality gives (19). ■

5.1.5. *Estimation of the variance.*

Proof of Proposition 4.1. We use the same notation and approach based on Gaussian concentration as in the proof of Proposition 2.5. Writing $\mathbf{Y}^i - \mu = \mathbf{A}\zeta_i$, we upper bound the Lipschitz constant of $\|\widehat{\sigma}\|_p$ as a function of $\zeta = (\zeta_1, \dots, \zeta_n)$: given $\zeta, \zeta' \in \left(\mathbb{R}^K\right)^n$, we have

$$\begin{aligned}\|\widehat{\sigma}(\zeta)\|_p - \|\widehat{\sigma}(\zeta')\|_p \leq \|\widehat{\sigma}(\zeta) - \widehat{\sigma}(\zeta')\|_p &\leq \left\| \left(\frac{1}{n} \sum_{i=1}^n \left\langle a_k, (\zeta_i - \overline{\zeta}) - (\zeta_i' - \overline{\zeta'}) \right\rangle^2 \right)_k^{\frac{1}{2}} \right\|_p \\ &\leq \frac{\|\sigma\|_p}{\sqrt{n}} \left(\sum_{i=1}^n \left\| (\zeta_i - \overline{\zeta}) - (\zeta_i' - \overline{\zeta'}) \right\|_2^2 \right)^{\frac{1}{2}} ;\end{aligned}$$

We then additionally have

$$\sum_{i=1}^n \left\| (\zeta_i - \overline{\zeta}) - (\zeta_i' - \overline{\zeta'}) \right\|_2^2 = \sum_{i=1}^n \|\zeta_i - \zeta_i'\|_2^2 - n \left\| \overline{\zeta} - \overline{\zeta'} \right\|_2^2 \leq \|\zeta - \zeta'\|_{2,Kn}^2 .$$

allowing to conclude that $\|\widehat{\sigma}(\zeta)\|_p$ has Lispchitz constant $\frac{\|\sigma\|_p}{\sqrt{n}}$. Concerning the expectation, observe that for each coordinate k, the variable $\sqrt{n}\widehat{\sigma}_k/\sigma_k$ has the same distribution as the square root of a $\chi^2(n-1)$ variable. Elementary calculations for the expectation of such a variable lead to $\mathbb{E}[\widehat{\sigma}_k] = C_n \sigma_k$. We finally conclude that with probability at least $1-\delta$, the following inequality holds:

$$C_n \|\sigma\|_p = \|\mathbb{E}[\widehat{\sigma}]\|_p \leq \mathbb{E}\left[\|\widehat{\sigma}\|_p\right] \leq \|\widehat{\sigma}\|_p + \frac{\|\sigma\|_p}{\sqrt{n}} \overline{\Phi}^{-1}\left(\frac{\delta}{2}\right) ;$$

solving this inequality in $\|\sigma\|_p$ yields the result. ■

5.2. *Quantiles.* Remember the following inequality coming from the definition of the quantile q_α: for any fixed $\mathbf{Y}$

$$\mathbb{P}_W\left(\phi\left(\overline{\mathbf{Y}^{\langle W\rangle}}\right) > q_\alpha(\phi, \mathbf{Y})\right) \leq \alpha \leq \mathbb{P}_W\left(\phi\left(\overline{\mathbf{Y}^{\langle W\rangle}}\right) \geq q_\alpha(\phi, \mathbf{Y})\right) . \tag{29}$$

Proof of Lemma 3.1. We introduce the notation $\mathbf{Y} \bullet W = \mathbf{Y}.\mathrm{diag}(W)$ for the matrix obtained by multiplying the i-th column of $\mathbf{Y}$ by W_i, $i = 1, \dots, n$. We then have

$$\begin{aligned}\mathbb{P}_{\mathbf{Y}}\left(\phi(\overline{\mathbf{Y}} - \mu) > q_\alpha(\phi, \mathbf{Y} - \mu)\right) &= \mathbb{E}_W\left[\mathbb{P}_{\mathbf{Y}}\left(\phi(\overline{(\mathbf{Y} - \mu)}^{\langle W\rangle}) > q_\alpha(\phi, (\mathbf{Y} - \mu) \bullet W)\right)\right] \\ &= \mathbb{E}_{\mathbf{Y}}\left[\mathbb{P}_W\left(\phi\left(\overline{(\mathbf{Y} - \mu)}^{\langle W\rangle}\right) > q_\alpha(\phi, \mathbf{Y} - \mu)\right)\right] \leq \alpha \ . \end{aligned} \tag{30}$$

The first equality is due to the fact that the distribution of $\mathbf{Y}$ satisfies assumption (SA), hence the distribution of $(\mathbf{Y}-\mu)$ is invariant by multiplying by (arbitrary) signs $W \in \{-1,1\}^n$. In the second equality we used Fubini's theorem and the fact that for any arbitrary signs W as above $q_\alpha(\phi,(\mathbf{Y}-\mu)\bullet W) = q_\alpha(\phi,\mathbf{Y}-\mu)$; finally the last inequality comes from (29). ■

Proof of Theorem 3.2. Put $\gamma_1 = \gamma_1(\alpha_0\delta)$ for short and define the event

$$\mathcal{E} := \left\{\mathbf{Y} \mid q_{\alpha_0}(\phi,\mathbf{Y}-\mu) \le q_{\alpha_0(1-\delta)}(\phi,\mathbf{Y}-\overline{\mathbf{Y}}) + \gamma_1 f(\mathbf{Y})\right\};$$

then we have using (30) :

$$\begin{aligned} &\mathbb{P}\left(\phi(\overline{\mathbf{Y}}-\mu) > q_{\alpha_0(1-\delta)}(\phi,\mathbf{Y}-\overline{\mathbf{Y}}) + \gamma_1 f(\mathbf{Y})\right) \\ &\le \mathbb{P}\left(\phi(\overline{\mathbf{Y}}-\mu) > q_{\alpha_0}(\phi,\mathbf{Y}-\mu)\right) + \mathbb{P}\left(\mathbf{Y}\in\mathcal{E}^c\right) \le \alpha_0 + \mathbb{P}\left(\mathbf{Y}\in\mathcal{E}^c\right) \quad . \end{aligned} \tag{31}$$

We now concentrate on the event $\mathcal{E}^c$. Using the subadditivity of ϕ, and the fact that $\overline{(\mathbf{Y}-\mu)}^{\langle W\rangle} = \overline{(\mathbf{Y}-\overline{\mathbf{Y}})}^{\langle W\rangle} + \overline{W}(\overline{\mathbf{Y}}-\mu)$, we have for any fixed $\mathbf{Y}\in\mathcal{E}^c$:

$$\begin{aligned} \alpha_0 &\le \mathbb{P}_W\left(\phi(\overline{(\mathbf{Y}-\mu)}^{\langle W\rangle}) \ge q_{\alpha_0}(\phi,\mathbf{Y}-\mu)\right) \\ &\le \mathbb{P}_W\left(\phi(\overline{(\mathbf{Y}-\mu)}^{\langle W\rangle}) > q_{\alpha_0(1-\delta)}(\phi,\mathbf{Y}-\overline{\mathbf{Y}}) + \gamma_1 f(\mathbf{Y})\right) \\ &\le \mathbb{P}_W\left(\phi(\overline{(\mathbf{Y}-\overline{\mathbf{Y}})}^{\langle W\rangle}) > q_{\alpha_0(1-\delta)}(\phi,\mathbf{Y}-\overline{\mathbf{Y}})\right) + \mathbb{P}_W\left(\phi(\overline{W}(\overline{\mathbf{Y}}-\mu)) > \gamma_1 f(\mathbf{Y})\right) \\ &\le \alpha_0(1-\delta) + \mathbb{P}_W\left(\phi(\overline{W}(\overline{\mathbf{Y}}-\mu)) > \gamma_1 f(\mathbf{Y})\right). \end{aligned}$$

For the first and last inequalities we have used (29), and for the second inequality the definition of $\mathcal{E}^c$. From this we deduce that

$$\mathcal{E}^c \subset \left\{\mathbf{Y} \mid \mathbb{P}_W\left(\phi(\overline{W}(\overline{\mathbf{Y}}-\mu)) > \gamma_1 f(\mathbf{Y})\right) \ge \alpha_0\delta\right\}.$$

Now using the positive-homogeneity of ϕ, and the fact that both ϕ and f are nonnegative:

$$\begin{aligned} \mathbb{P}_W\left(\phi(\overline{W}(\overline{\mathbf{Y}}-\mu)) > \gamma_1 f(\mathbf{Y})\right) &= \mathbb{P}_W\left(\left|\overline{W}\right| > \frac{\gamma_1 f(\mathbf{Y})}{\phi(\mathrm{sign}(\overline{W})(\overline{\mathbf{Y}}-\mu))}\right) \\ &\le \mathbb{P}_W\left(\left|\overline{W}\right| > \frac{\gamma_1 f(\mathbf{Y})}{\widetilde{\phi}(\overline{\mathbf{Y}}-\mu)}\right) = 2\mathbb{P}_{B_n}\left(\frac{1}{n}(2B_n-n) > \frac{\gamma_1 f(\mathbf{Y})}{\widetilde{\phi}(\overline{\mathbf{Y}}-\mu)}\right), \end{aligned}$$

where B_n denotes a binomial $(n,\frac{1}{2})$ variable (independent of $\mathbf{Y}$). From the two last displays and the definition of γ_1, we conclude $\mathcal{E}^c \subset \left\{\mathbf{Y} \;\middle|\; \widetilde{\phi}(\overline{\mathbf{Y}}-\mu) > f(\mathbf{Y})\right\}$, which, put back in (31), leads to the desired conclusion. ■

Proof of Corollary 3.3. Define the function

$$g_0(\mathbf{Y}) = q_{(1-\delta)\alpha_0}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + \left(\sum_{i=1}^{J-1} \gamma_i q_{(1-\delta)\alpha_i}(\widetilde{\phi}, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma_J f(\mathbf{Y}) \right) ,$$

and for $k = 1, \ldots, J$,

$$g_k(\mathbf{Y}) = \gamma_k^{-1} \left(\sum_{i=k}^{J-1} \gamma_i q_{(1-\delta)\alpha_i}(\widetilde{\phi}, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma_J f(\mathbf{Y}) \right) ,$$

with the convention $g_J = f$. For $0 \leq k \leq J-1$, applying Theorem 3.2 with the function g_{k+1} yields the relation

$$\mathbb{P}_W \left(\phi(\overline{\mathbf{Y}} - \mu) > g_k(\mathbf{Y}) \right) \leq \alpha_k + \mathbb{P}_W \left(\phi(\overline{\mathbf{Y}} - \mu) > g_{k+1}(\mathbf{Y}) \right) .$$

Therefore we get

$$\mathbb{P}_W \left(\phi(\overline{\mathbf{Y}} - \mu) > g_0(\mathbf{Y}) \right) \leq \sum_{i=0}^{J-1} \alpha_i + \mathbb{P} \left(\widetilde{\phi}(\overline{\mathbf{Y}} - \mu) > f(\mathbf{Y}) \right) ,$$

as announced. ∎

Proof of Proposition 3.4. Let us first prove that an analogue of Lemma 3.1 holds with q_{α_0} replaced by $\widetilde{q}_{\alpha_0}$. First, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{W}} \mathbb{P}_{\mathbf{Y}} \left(\phi(\overline{\mathbf{Y}} - \mu) > \widetilde{q}_{\alpha_0}(\phi, \mathbf{Y} - \mu, \mathbf{W}) \right) \\
&\qquad = \mathbb{E}_{W'} \mathbb{E}_{\mathbf{W}} \mathbb{P}_{\mathbf{Y}} \left(\phi(\overline{(\mathbf{Y} - \mu)}^{\langle W' \rangle}) > \widetilde{q}_{\alpha_0}(\phi, (\mathbf{Y} - \mu) \bullet W', \mathbf{W}) \right) \\
&\qquad = \mathbb{E}_{\mathbf{Y}} \mathbb{P}_{\mathbf{W}, W'} \left(\phi(\overline{(\mathbf{Y} - \mu)}^{\langle W' \rangle}) > \widetilde{q}_{\alpha_0}(\phi, \mathbf{Y} - \mu, W' \bullet \mathbf{W}) \right) ,
\end{aligned}$$

where W' denotes a Rademacher vector independent of all other random variables and $W' \bullet \mathbf{W} = \mathrm{diag}(W').\mathbf{W}$ denotes the matrix obtained by multiplying the i-th row of $\mathbf{W}$ by W'_i, $i = 1, \ldots, n$. Note that $(W', W' \bullet \mathbf{W}) \sim (W', \mathbf{W})$. Therefore, by definition of the quantile $\widetilde{q}_{\alpha_0}$, the latter quantity is equal to

$$\mathbb{E}_{\mathbf{Y}} \mathbb{P}_{\mathbf{W}, W'} \left(\frac{1}{B} \sum_{j=1}^{B} \mathbb{1} \left\{ \phi(\overline{(\mathbf{Y} - \mu)}^{\langle \mathbf{W}^j \rangle}) \geq \phi(\overline{(\mathbf{Y} - \mu)}^{\langle W' \rangle}) \right\} \leq \alpha_0 \right) \leq \frac{\lfloor B\alpha_0 \rfloor + 1}{B + 1} ,$$

where the last step comes from Lemma 5.2 (see below).

The rest of the proof is similar to the one of Theorem 3.2, where $\mathbb{P}_W$ is replaced by the empirical distribution based on $\mathbf{W}$, $\widetilde{\mathbb{P}}_{\mathbf{W}} = \frac{1}{B}\sum_{j=1}^{B} \delta_{\mathbf{W}^j}$. Thus, (29) becomes for any fixed $\mathbf{Y}, \mathbf{W}$:

$$\widetilde{\mathbb{P}}_{\mathbf{W}}\left[\phi\left(\overline{\mathbf{Y}}^{\langle W\rangle}\right) > \widetilde{q}_{\alpha_0}(\phi, \mathbf{Y}, \mathbf{W})\right] \leq \alpha_0 \leq \widetilde{\mathbb{P}}_{\mathbf{W}}\left[\phi\left(\overline{\mathbf{Y}}^{\langle W\rangle}\right) \geq \widetilde{q}_{\alpha_0}(\phi, \mathbf{Y}, \mathbf{W})\right] .$$

Then, the role of $\mathcal{E}$ is taken by

$$\widetilde{\mathcal{E}} := \left\{\mathbf{Y}, \mathbf{W} \mid \widetilde{q}_{\alpha_0}(\phi, \mathbf{Y} - \mu, \mathbf{W}) \leq \widetilde{q}_{\alpha_0(1-\delta)}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}, \mathbf{W}) + \gamma f(\mathbf{Y}, \mathbf{W})\right\} ,$$

where we put $\gamma = \gamma(\mathbf{W}, \alpha_0\delta)$ for short. We then have similarly to (31):

$$\mathbb{P}_{\mathbf{Y},\mathbf{W}}\left(\phi(\overline{\mathbf{Y}} - \mu) > \widetilde{q}_{\alpha_0(1-\delta)}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma f(\mathbf{Y}, \mathbf{W})\right) \leq \frac{\lfloor B\alpha_0\rfloor + 1}{B+1} + \mathbb{P}_{\mathbf{Y},\mathbf{W}}\left(\widetilde{\mathcal{E}}^c\right) ,$$

and following further the proof of Theorem 3.2, we obtain

$$\widetilde{\mathcal{E}}^c \subset \left\{\mathbf{Y}, \mathbf{W} \,\middle|\, \widetilde{\mathbb{P}}_{\mathbf{W}}\left[|\overline{W}| > \frac{\gamma f(\mathbf{Y}, \mathbf{W})}{\widetilde{\phi}\left(\overline{\mathbf{Y}} - \mu\right)}\right] \geq \alpha_0\delta\right\} ,$$

which gives the result. ∎

We have used the following Lemma which essentially reproduces Lemma 1 of [30], with a minor strenghthening. While the proof was left to the reader in [30], because it was considered either elementary or common knowledge, we include a succinct proof below for completeness.

Lemma 5.2 (Minor variation of Lemma 1 of [30]) *Let Z_0, Z_1, …, Z_B be exchangeable real-valued random variables. Then for all $\alpha \in (0,1)$,*

$$\mathbb{P}\left(\frac{1}{B}\sum_{j=1}^{B} \mathbb{1}\left\{Z_j \geq Z_0\right\} \leq \alpha\right) \leq \frac{\lfloor B\alpha\rfloor + 1}{B+1} \leq \alpha + \frac{1}{B+1}.$$

The first inequality becomes an equality if $Z_i \neq Z_j$ a.s. For example, it is the case if the Z_is are i.i.d. variables from a distribution without atoms.

Proof of Lemma 5.2. Let U denote a random variable uniformly distributed in $\{0, \ldots, B\}$ and independent of the Z_is. We then have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{B}\sum_{j=1}^{B} \mathbb{1}\left\{Z_j \geq Z_0\right\} \leq \alpha\right) &= \mathbb{P}\left(\sum_{j=0}^{B} \mathbb{1}\left\{Z_j \geq Z_0\right\} \leq B\alpha + 1\right) \\
&= \mathbb{P}_U\mathbb{P}_{(Z_i)}\left(\sum_{j=0}^{B} \mathbb{1}\left\{Z_j \geq Z_U\right\} \leq B\alpha + 1\right) \\
&= \mathbb{P}_{(Z_i)}\mathbb{P}_U\left(\sum_{j=0}^{B} \mathbb{1}\left\{Z_j \geq Z_U\right\} \leq \lfloor B\alpha\rfloor + 1\right) \leq \frac{\lfloor B\alpha\rfloor + 1}{B+1} .
\end{aligned}$$

Note that the last inequality is an equality if the Z_is are a.s. distinct. ∎

5.3. *Exchangeable resampling computations.* In this section, we compute constants A_W, B_W, C_W and D_W (defined by (3) to (6)) for some exchangeable resamplings. This implies all the statements in Table 1. We first define several additional exchangeable resampling weights (normalized so that $\mathbb{E}[W_i] = 1$):

- **Bernoulli** (p), $p \in (0,1)$: pW_i i.i.d. with a Bernoulli distribution of parameter p. A classical choice is $p = \frac{1}{2}$.
- **Efron** (q), $q \in \{1 \dots, n\}$: $qn^{-1}W$ has a multinomial distribution with parameters $(q; n^{-1}, \dots, n^{-1})$. A classical choice is $q = n$.
- **Poisson** (μ), $\mu \in (0, +\infty)$: μW_i i.i.d. with a Poisson distribution of parameter μ. A classical choice is $\mu = 1$.

Notice that $\overline{\mathbf{Y}}^{\langle W - \overline{W} \rangle}$ and all the resampling constants are invariant under translation of the weights, so that Bernoulli $(1/2)$ weights are completely equivalent to Rademacher weights in this paper.

Lemma 5.3 *1. Let W be Bernoulli (p) weights with $p \in (0,1)$. Then we have* $2(1-p)\left(1 - \frac{1}{n}\right) = A_W \le B_W \le \sqrt{\frac{1}{p} - 1}\sqrt{1 - \frac{1}{n}}$, $C_W = \sqrt{\frac{1}{p} - 1}$, *and* $D_W \le \frac{1}{2p} + \left|\frac{1}{2p} - 1\right| + \sqrt{\frac{1-p}{np}}$.

2. Let W be Efron (q) weights with $q \in \{1, \dots, n\}$. Then we have $2\left(1 - \frac{1}{n}\right)^q = A_W \le B_W \le \sqrt{\frac{n-1}{q}}$ *and* $C_W = \sqrt{\frac{n}{q}}$.

3. Let W be Poisson (μ) weights with $\mu > 0$. Then we have $A_W \le B_W \le \frac{1}{\sqrt{\mu}}\sqrt{1 - \frac{1}{n}}$ *and* $C_W = \frac{1}{\sqrt{\mu}}$. *Moreover, if $\mu = 1$, we get* $\frac{2}{e} - \frac{1}{\sqrt{n}} \le A_W$.

4. Let W be Random hold-out (q) weights with $q \in \{1, \dots, n\}$. Then we have $A_W = 2\left(1 - \frac{q}{n}\right)$, $B_W = \sqrt{\frac{n}{q} - 1}$, $C_W = \sqrt{\frac{n}{n-1}}\sqrt{\frac{n}{q} - 1}$ *and* $D_W = \frac{n}{2q} + \left|1 - \frac{n}{2q}\right|$.

Proof of Lemma 5.3. We consider the following cases:

General case. We first only assume that W is exchangeable. Then, from the concavity of $\sqrt{\cdot}$ and the triangular inequality, we have

$$\begin{aligned} \mathbb{E}\left|W_1 - \mathbb{E}[W_1]\right| - \sqrt{\mathbb{E}\left(\overline{W} - \mathbb{E}[W_1]\right)^2} &\le \mathbb{E}\left|W_1 - \mathbb{E}[W_1]\right| - \mathbb{E}\left|\overline{W} - \mathbb{E}[W_1]\right| \\ &\le A_W \le B_W \le \sqrt{\frac{n-1}{n}} C_W \ . \end{aligned} \tag{32}$$

Independent weights. When we suppose that the W_i are i.i.d., we get

$$\mathbb{E}\left|W_1 - \mathbb{E}[W_1]\right| - \frac{\sqrt{\mathrm{Var}(W_1)}}{\sqrt{n}} \le A_W \qquad \text{and} \qquad C_W = \sqrt{\mathrm{Var}(W_1)} \ . \tag{33}$$

Bernoulli. First, we have $A_W = \mathbb{E}\left|W_1 - \overline{W}\right| = \mathbb{E}\left|\left(1-\frac{1}{n}\right)W_1 - X_{n,p}\right|$ with $X_{n,p} := \frac{1}{n}\left(W_2 + \cdots + W_n\right)$. Since W_1 and $X_{n,p}$ are independent and $X_{n,p} \in [0, (n-1)/(np)]$ a.s., we obtain

$$A_W = p\mathbb{E}\left[\left(1-\frac{1}{n}\right)\frac{1}{p} - X_{n,p}\right] + (1-p)\mathbb{E}\left[X_{n,p}\right] = 1 - \frac{1}{n} + (1-2p)\mathbb{E}\left[X_{n,p}\right] \quad .$$

The formula for A_W follows since $\mathbb{E}\left[X_{n,p}\right] = (n-1)/n$. Second, remark that the Bernoulli (p) weights are i.i.d. with $\mathrm{Var}(W_1) = p^{-1} - 1$, $\mathbb{E}[W_1] = 1$ and $\mathbb{E}\left|W_1 - 1\right| = p\left(p^{-1} - 1\right) + (1-p) = 2(1-p)$. Hence, (32) and (33) lead to the bounds for B_W and C_W. Finally, the Bernoulli (p) weights satisfy the assumption of (6) with $x_0 = a = (2p)^{-1}$. Then,

$$D_W = \frac{1}{2p} + \mathbb{E}\left|\overline{W} - \frac{1}{2p}\right| \leq \frac{1}{2p} + \left|1 - \frac{1}{2p}\right| + \mathbb{E}\left|\overline{W} - 1\right| \leq \frac{1}{2p} + \frac{1}{p}\left|\frac{1}{2} - p\right| + \sqrt{\frac{1-p}{np}} \quad .$$

Efron. We have $\overline{W} = 1$ a.s. so that $C_W = \sqrt{\frac{n}{n-1} \times \mathrm{Var}(W_1)} = \sqrt{n/q}$. If moreover $q \leq n$, then $W_i < 1$ implies $W_i = 0$ and $A_W = \mathbb{E}\left|W_1 - 1\right| = \mathbb{E}\left[W_1 - 1 + 2\mathbb{1}\left\{W_1 = 0\right\}\right] = 2\mathbb{P}(W_1 = 0) = 2\left(1 - \frac{1}{n}\right)^q$. The result follows from (32).

Poisson. These weights are i.i.d. with $\mathrm{Var}(W_1) = \mu^{-1}$, $\mathbb{E}[W_1] = 1$. Moreover, if $\mu \leq 1$, $W_i < 1$ implies $W_i = 0$ and $\mathbb{E}\left|W_1 - 1\right| = 2\mathbb{P}(W_1 = 0) = 2e^{-\mu}$. With (32) and (33), the result follows.

Random hold-out. These weights are such that $\{W_i\}_{1 \leq i \leq n}$ takes only two values, with $\overline{W} = 1$. Then, A_W, B_W and C_W can be directly computed. Moreover, they satisfy the assumption of (6) with $x_0 = a = n/(2q)$. The computation of D_W is straightforward. ■

Acknowledgements. We thank Pascal Massart for his particulary relevant comments and suggestions. We also would like to thank the two referees and the AE for their insight, leading in particular to a more rational organization of the paper.

REFERENCES

[1] S. Arlot. *Resampling and Model Selection.* PhD thesis, University Paris XI, Dec. 2007.

[2] S. Arlot, G. Blanchard, and É. Roquain. Some non-asymptotic results on resampling in high dimension, II: Multiple tests. *Ann. Statist.*, 2009. To appear.

[3] Y. Baraud. Confidence balls in Gaussian regression. *Ann. Statist.*, 32(2):528–551, 2004.

[4] R. Beran. The impact of the bootstrap on statistical algorithms and theory. *Statist. Sci.*, 18(2):175–184, 2003. Silver anniversary of the bootstrap.

[5] R. Beran and L. Dümbgen. Modulation of estimators and confidence sets. *Annals of Statistics*, 26:1826–1856, 1998.

[6] T. Cai and M. Low. Adaptive confidence balls. *Annals of Statistics*, 34(1):202–228, 2006.

[7] B. R. Cirel'son, I. A. Ibragimov, and V. N. Sudakov. Norms of Gaussian sample functions. In *Proceedings of the Third Japan–USSR Symposium on Probability Theory*, volume 550 of *Lecture notes in mathematics*, pages 20–41. Springer, 1976.

[8] F. Darvas, M. Rautiainen, D. Pantazis, S. Baillet, H. Benali, J. Mosher, L. Garnero, and R. Leahy. Investigations of dipole localization accuracy in MEG using the bootstrap. *NeuroImage*, 25:355–368, 2005.

[9] T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statist. Sci.*, 11(3):189–228, 1996.
[10] C. Durot and Y. Rozenholc. An adaptive test for zero mean. *Math. Methods Statist.*, 15(1):26–60, 2006.
[11] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.
[12] R. A. Fisher. *The Design of Experiments.* Oliver and Boyd, Edinburgh, 1935.
[13] M. Fromont. Model selection by bootstrap penalization for classification. *Machine Learning*, 66(2-3):165–207, 2006.
[14] Y. Ge, S. Dudoit, and T. P. Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77, 2003.
[15] P. Hall. *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York, 1992.
[16] P. Hall and E. Mammen. On general resampling algorithms and their performance in distribution estimation. *Ann. Statist.*, 22(4):2011–2030, 1994.
[17] M. Hoffmann and O. Lepski. Random rates in anisotropic regression. *Ann. Statist.*, 30(2):325–396, 2002.
[18] K. Jerbi, J.-P. Lachaux, K. N'Diaye, D. Pantazis, R. M. Leahy, L. Garnero, and S. Baillet. Coherent neural representation of hand speed in humans revealed by MEG imaging. *PNAS*, 104(18):7676–7681, 2007.
[19] A. Juditsky and S. Lambert-Lacroix. Nonparametric confidence set estimation. *Math. Methods Statist.*, 12:410–428, 2003.
[20] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inf. Theory*, 47(5):1902–1914, 2001.
[21] O. V. Lepski. How to improve the accuracy of estimation. *Math. Methods Statist.*, 8(4):441–486 (2000), 1999.
[22] K. Li. Honest confidence regions for nonparametric regression. *Ann. Statist.*, 17:1001–1008, 1989.
[23] D. M. Mason and M. A. Newton. A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.*, 20(3):1611–1624, 1992.
[24] P. Massart. *Concentration Inequalities and Model Selection (Lecture notes of the St-Flour probability summer school 2003)*, volume 1896 of *Lecture notes in Mathematics*. Springer, 2007.
[25] C. McDiarmid. On the method of bounded differences. In *Surveys in combinatorics*, volume 141 of *London Mathematical Society Lecture Notes*, pages 148–188. Cambridge University Press, 1989.
[26] D. Pantazis, T. E. Nichols, S. Baillet, and R. M. Leahy. A comparison of random field theory and permutation methods for statistical analysis of MEG data. *NeuroImage*, 25:383–394, 2005.
[27] D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York, 1999.
[28] J. Præstgaard and J. A. Wellner. Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, 21(4):2053–2086, 1993.
[29] J. Robins and A. van der Vaart. Adaptive nonparametric confidence sets. *Ann. Statist.*, 34:229–253, 2006.
[30] J. P. Romano and M. Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.*, 100(469):94–108, 2005.
[31] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.
[32] T. Waberski, R. Gobbele, W. Kawohl, C. Cordes, and H. Buchner. Immediate cortical reorganization after local anesthetic block of the thumb: source localization of somatosensory evoked potentials in human subjects. *Neurosci. Lett.*, 347:151–154, 2003.

Sylvain Arlot
CNRS ; Willow Project-Team
Laboratoire d'Informatique de l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
45, rue d'Ulm, 75230 Paris, France
E-mail: sylvain.arlot@ens.fr

Gilles Blanchard
Weierstrass institute for applied stochastics and analysis
Mohrenstrasse 39, 10117 Berlin, Germany, and
Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489 Berlin,
Germany
E-mail: blanchar@wias-berlin.de

Etienne Roquain
University of Paris 6,LPMA,
4, Place Jussieu, 75252 Paris cedex 05, France
E-mail: etienne.roquain@upmc.fr