Minimum divergence estimators, Maximum Likelihood and the generalized bootstrap

Michel Broniatowski

▶ **To cite this version:**

Michel Broniatowski. Minimum divergence estimators, Maximum Likelihood and the generalized bootstrap. Entropy, 2021, Entropy, 23,2 (2). hal-02936714v1

HAL Id: hal-02936714

https://hal.science/hal-02936714v1

Submitted on 2 Nov 2020 (v1), last revised 24 Dec 2022 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Minimum divergence estimators, Maximum Likelihood and the generalized bootstrap

Michel Broniatowski
LPSM, CNRS UMR 8001, Sorbonne-Université Paris, France

November 2, 2020

Abstract

This paper is an attempt to set a justification for making use of some dicrepancy indexes, starting from the classical Maximum Likelihood definition, and adapting the corresponding basic principle of inference to situations where minimization of those indexes between a model and some extension of the empirical measure of the data appears as its natural extension. This leads to the so called generalized bootstrap setting for which minimum divergence inference seems to replace Maximum Likelihood one.

Keywords: Statistical divergences; Maximum likelihood; Conditional limit theorem; Bahadur efficiency; Minimum divergence estimator

1 Motivation and context

Divergences between probability measures are widely used in Statistics and Data Science in order to perform inference under models of various kinds, parametric or semi parametric, or even in non parametric settings. The corresponding methods extend the likelihood paradigm and insert inference in some minimum "distance" framing, which provides a convenient description for the properties of the resulting estimators and tests, under the model or under misspecification. Furthermore they pave the way to a large number of competitive methods, which allows for trade-off between efficiency and robustness, among others. Many families of such divergences have been proposed, some of them stemming from classical statistics (such as the Chi-square), while others have their origin in other fields such as Information theory. Some measures of discrepancy involve regularity of the corresponding probability measures while others seem to be restricted to measures on finite or countable spaces, at least when using them as inferential tools, henceforth in situations when the elements of a model have to be confronted with a dataset. The choice of a specific discrepancy measure in specific context is somehow arbitrary in many cases, although the resulting conclusion of the inference might differ accordingly, above all under misspecification; however the need for such approaches is clear when aiming at robustness.

This paper considers a specific class of divergences, which contains most of the classical inferential tools, and which is indexed by a single scalar parameter. This class of divergences belongs to the Csiszar-Ali-Silvey-Arimoto family of divergences (see [16]), and is usually referred to as the power divergence class, which has been considered By Cressie and Read [21]; however this denomination is also shared by other discrepancy measures of some different nature [3]; see [12] for a comprehensive description of those various inferential tools with a discussion on their relations. We will use the acronym CR for the class of divergences under consideration in this paper.

We have tried to set a justification for those discrepancy indexes, starting from the classical Maximum Likelihood definition, and adapting the corresponding basic principle of inference to situations where those indexes appear as its natural extension. This leads to the so called generalized bootstrap setting for which minimum divergence inference seems to replace Maximum Likelihood one.

The contents of this approach can be summarized as follows.

Section 2 states that the MLE is obtained as a proxy of the minimizer of the Kullback-Leibler divergence between the generic law of the observed variable and the model, which is the large deviation limit for the empirical distribution. This limit statement is nothing but the continuation of the classical ML paradigm, namely to make the dataset more "probable " under the fitted distribution in the model, or, equivalently, to fit the most "likely" distribution in the model to the dataset.

Section 3 states that given a divergence pseudo distance ϕ in CR the Minimum Divergence Estimator (MDE) is obtained as a proxy of the minimizer of the large deviation limit for some bootstrap version of the empirical distribution, which establishes that the MDE is MLE for bootstrapped samples defined in relation with the divergence. This fact is based on the strong relation which associates to any CR ϕ -divergence a specific RV W (see Section 1.1.2) ; this link is the cornerstone for the interpretation of the minimum ϕ-divergence estimators as MLE's for specific bootstrapped sampling schemes where W has a prominent rôle. Some specific remark explores the link between MDE and MLE in exponential families. As a by product we also introduce a bootstrapped estimator of the divergence pseudo-distance ϕ between the distribution of the data and the model.

In Section 4 we specify the bootstrapped estimator of the divergence which can be used in order to perform an optimal test of fit. Due to the type of asymptotics handle in this paper, optimality is studied in terms of Bahadur efficiency. It is shown that tests of fit based on such estimators enjoy Bahadur optimality with respect to other bootstrap plans when the bootstrap is performed under the distribution associated with the divergence criterion itself.

The discussion held in this paper pertains to parametric estimation in a model $\mathcal{P}_\Theta$ whose elements P_θ are probability measures defined on the same finite space $\mathcal{Y} := \{d_1, .., d_K\}$, and $\theta \in \Theta$ an index space; we assume identifiability, namely different values of θ induce different probability laws P_θ's. Also all the entries of P_θ will be positive for all θ in Θ.

1.1 Notation

1.1.1 Divergences

We consider regular *divergence functions* φ which are non negative convex functions with values in $\overline{\mathbb{R}^+}$ which belong to $C^2(\mathbb{R})$ and satisfy $\varphi(1) = \varphi'(1) = 0$ and $\varphi''(1) = 1$; see [16] and [10] for properties and extensions . An important class of such functions is defined through the power divergence functions

$$\varphi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)} \tag{1}$$

defined for all real $\gamma \neq 0, 1$ with $\varphi_0(x) := -\log x + x - 1$ (the likelihood divergence function) and $\varphi_1(x) := x \log x - x + 1$ (the Kullback-Leibler divergence function). This class is usually referred to as the Cressie-Read family of divergence functions (see [21]). It is a very simple class of functions (with the limits in $\gamma \to 0, 1$) which allows to represent nearly all commonly used statistical criterions. Parametric inference in commonly met situations including continuous models or some non regular models can be performed with them; see[6]. The L_1 divergence function $\varphi(x) := |x - 1|$ is not captured by the CR family of functions. When undefined the function φ is declared to assume value $+\infty$.

Associated with a divergence function φ, ϕ is the *divergence pseudo-distance* between a probability measure and a finite signed measure; see [12].

For $P := (p_1, .., p_K)$ and $Q := (q_1, .., q_K)$ in $\mathbb{S}^K$, the simplex of all probability measures on $\mathcal{Y}$, define, whenever Q and P have non null entries

$$\phi(Q, P) := \sum_{k=1}^{K} p_k \varphi\left(\frac{q_k}{p_k}\right).$$

Indexing this pseudo-distance by γ and using φ_γ as divergence function yields the likelihood divergence $\phi_0(Q, P) := -\sum p_k \log\left(\frac{q_k}{p_k}\right)$, the Kullback-Leibler divergence $\phi_1(Q, P) := \sum q_k \log\left(\frac{q_k}{p_k}\right)$, the Hellinger divergence $\phi_{1/2}(Q, P) := \frac{1}{2}\sum p_k \left(\sqrt{\left(\frac{q_k}{p_k}\right)} - 1\right)^2$, the modified (or Neyman) χ^2 divergence $\phi_{-1}(Q, P) := \frac{1}{2}\sum p_k \left(\left(\frac{q_k}{p_k}\right) - 1\right)^2 \left(\frac{q_k}{p_k}\right)^{-1}$. The χ^2 divergence $\phi_2(Q, P) := \frac{1}{2}\sum p_k \left(\left(\frac{q_k}{p_k}\right) - 1\right)^2$ is defined between signed measures; see [5] for definitions in more general setting, and [6] for the advantage to extend the definition to possibly signed measures in the context of parametric inference for non regular models. Also the present discussion which is restricted to finite spaces $\mathcal{Y}$ can be extended to general spaces.

The conjugate divergence function of φ is defined through

$$\widetilde{\varphi}(x) := x\varphi\left(\frac{1}{x}\right) \tag{2}$$

and the corresponding divergence pseudo-distance $\widetilde{\phi}(P,Q)$ is

$$\widetilde{\phi}(P,Q) := \sum_{k=1}^{K} q_k \widetilde{\varphi}\left(\frac{p_k}{q_k}\right)$$

which satisfies

$$\widetilde{\phi}(P,Q) = \phi(Q,P)$$

whenever defined, and equals $+\infty$ otherwise. When $\varphi = \varphi_\gamma$ then $\widetilde{\varphi} = \varphi_{1-\gamma}$ as follows by substitution. Pairs $(\varphi_\gamma, \varphi_{1-\gamma})$ are therefore *conjugate pairs*. Inside the Cressie-Read family, the Hellinger divergence function is self-conjugate.

For $P = P_\theta$ and $Q \in \mathbb{S}^K$ we denote $\phi(Q,P)$ by $\phi(Q,\theta)$ (resp $\phi(\theta,Q)$, or $\phi(\theta',\theta)$, etc according to the context).

1.1.2 Weights

This paragraph introduces the special link which connects CR divergences with specific random variables, which we call weights. Those will be associated to the dataset and define what is usually referred to as a generalized bootstrap procedure. This is the setting which allows for an interpretation of the MDE's as generalized bootstrapped MLE's.

For a given real valued random variable (RV) W denote

$$M(t) := \log E \exp tW \tag{3}$$

its cumulant generating function which we assume to be finite in a non void neighborhood of 0 . The Fenchel Legendre transform of M (also called the Chernoff function) is defined through

$$\varphi^W(x) = M^*(x) := \sup_t tx - M(t). \tag{4}$$

The function $x \to \varphi^W(x)$ is non negative, is C^∞ and convex. We also assume that $EW = 1$ together with $VarW = 1$ which implies $\varphi^W(1) = \left(\varphi^W\right)'(1) = 0$ and $\left(\varphi^W\right)''(1) = 1$. Hence $\varphi^W(x)$ is a divergence function with corresponding divergence pseudo-distance ϕ^W. Associated with φ^W is the conjugate divergence $\widetilde{\phi^W}$ with divergence function $\widetilde{\varphi^W}$, which therefore satisfies $\phi^W(Q,P) = \widetilde{\phi^W}(P,Q)$ whenever neither P nor Q have null entries.

It is of interest to note that the classical power divergences φ_γ can be represented through (4) for $\gamma \leq 1$ or $\gamma \geq 2$. A first proof of this lays in the fact that when W has a distribution in a Natural Exponential Family (NEF) with power variance function $\alpha = 2 - \gamma$, then the Legendre transform φ^W of its cumulant generating function M is indeed of the form (1). See [15] and [2] for NEF's and power variance functions, and [9] for relation to the bootstrap. A general result of a different nature , including the former ones, can be seen in [11], Theorem 20. Correspondence between the various values of γ and the distribution of the respective weights can be found in [11], Example 39, and it can be summarized as presented now.

For $\gamma < 0$ the RV W is constructed as follows: Let Z be an auxiliary RV with density f_Z and support $[0, \infty)$ of a stable law with parameter triplet $\left(-\frac{\gamma}{1-\gamma}, 0, \frac{(1-\gamma)^{-\gamma//(1-\gamma)}}{\gamma}\right)$ in terms of the "form B notation" on p 12 in [24], and

$$f_W(y) := \frac{\exp\left(-y/(1-\gamma)\right)}{\exp(1/\gamma)} f_Z(y) 1_{[0,\infty)}(y).$$

For $\gamma = 0$ (which amounts to consider the limit as $\gamma \to 0$ in (1)) then W has a standard exponential distribution $E(1)$ on $[0, \infty)$.

For $\gamma \in (0, 1)$ then W has a compound Gamma-Poisson distribution $C\left(POI(\theta), GAM(\alpha, \beta)\right)$ where $\theta = 1/\gamma$, $\alpha = 1/(1-\gamma)$ and $\beta = \gamma/(1-\gamma)$.

For $\gamma = 1$ then W has a Poisson distribution with parameter $1, POI(1)$.

For $\gamma = 2$ then W has normal distribution with expectation and variance equal to 1.

For $\gamma > 2$ then the RV W is constructed as follows: Let Z be an auxiliary RV with density f_Z and support $(-\infty, \infty)$ of a stable law with parameter triplet $\left(\frac{\gamma}{\gamma-1}, 0, \frac{(\gamma-1)^{-\gamma//(\gamma-1)}}{\gamma}\right)$ in terms of the "form B notation" on p 12 in [24], and

$$f_W(y) := \frac{\exp\left(y/(\gamma-1)\right)}{\exp(1/\gamma)} f_Z(-y) \qquad ,y \in \mathbb{R}.$$

2 Maximum likelihood under finitely supported distributions and simple sampling

2.1 Standard derivation

Let $X_1, ...X_n$ be a set of n independent random variables with common probability measure P_{θ_T} and consider the Maximum Likelihood estimator of θ_T . A common way to define the ML paradigm is as follows: For any θ consider independent random variables $(X_{1,\theta}, ...X_{n,\theta})$ with probability measure P_θ , thus *sampled in the same way as the X_i's*, but under some alternative θ. Define θ_{ML} as the value of the parameter θ for which the probability that, up to a permutation of the order of the $X_{i,\theta}$'s, the probability that $(X_{1,\theta}, ...X_{n,\theta})$ coincides with $X_1, ...X_n$ is maximal, conditionally on the observed sample $X_1, ...X_n$. In formula, let σ denote a random permutation of the indexes $\{1, 2, ..., n\}$ and θ_{ML} is defined through

$$\theta_{ML} := \arg\max_\theta \frac{1}{n!} \sum_{\sigma \in \mathfrak{S}} P_\theta \left(\left(X_{\sigma(1),\theta}, ..., X_{\sigma(n),\theta}\right) = (X_1, ...X_n) \middle| (X_1, ...X_n) \right) \tag{5}$$

where the summation is extended on all equally probable permutations of $\{1, 2, ..., n\}$.

Denote

$$P_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$$

and

$$P_{n,\theta} := \frac{1}{n}\sum_{i=1}^{n}\delta_{X_{i,\theta}}$$

the empirical measures pertaining respectively to $(X_1, ...X_n)$ and $(X_{1,\theta}, ...X_{n,\theta})$

An alternative expression for θ_{ML} is

$$\theta_{ML} := \arg\max_{\theta} P_\theta\left(P_{n,\theta} = P_n \middle| P_n\right). \tag{6}$$

An explicit enumeration of the above expression $P_\theta\left(P_{n,\theta} = P_n \middle| P_n\right)$ involves the quantities

$$n_j := card\left\{i : X_i = d_j\right\}$$

for $j = 1, ..., K$ and yields

$$P_\theta\left(P_{n,\theta} = P_{n,X} \middle| P_{n,X}\right) = \frac{n! P_\theta\left(d_j\right)^{n_j}}{\prod_{j=1}^{K} n_j!} \tag{7}$$

as follows from the classical multinomial distribution. Optimizing on θ in (7) yields

$$\begin{aligned}\theta_{ML} &= \arg\max_{\theta}\sum_{j=1}^{K}\frac{n_j}{n}\log P_\theta\left(d_j\right)\\ &= \arg\max_{\theta}\frac{1}{n}\sum_{i=1}^{n}\log P_\theta\left(X_i\right).\end{aligned}$$

Consider now the Kullback-Leibler distance between P_θ and P_n which is non commutative and defined through

$$\begin{aligned}KL\left(P_n, \theta\right) &:= \sum_{j=1}^{k}\varphi\left(\frac{n_j/n}{P_\theta\left(d_j\right)}\right)P_\theta\left(d_j\right)\\ &= \sum_{j=1}^{k}\left(n_j/n\right)\log\frac{n_j/n}{P_\theta\left(d_j\right)}\end{aligned} \tag{8}$$

where

$$\varphi_1(x) := x\log x - x + 1 \tag{9}$$

which is the Kullback-Leibler divergence function. Minimizing the Kullback-Leibler distance $KL(P_n, \theta)$ upon θ yields

$$\begin{aligned}\theta_{KL} &= \arg\min_{\theta} KL(P_n, \theta) \\ &= \arg\min_{\theta} -\sum_{j=1}^{K} \frac{n_j}{n} \log P_\theta(d_j) \\ &= \arg\max_{\theta} \sum_{j=1}^{K} \frac{n_j}{n} \log P_\theta(d_j) \\ &= \theta_{ML}.\end{aligned}$$

Introduce the *conjugate divergence function* $\widetilde{\varphi} = \varphi_0$ of φ_1 , inducing the modified Kullback-Leibler, or so-called Likelihood divergence pseudo-distance KL_m which therefore satisfies

$$KL_m(\theta, P_n) = KL(P_n, \theta).$$

We have seen that minimizing the Kullback-Leibler divergence $KL(P_n, \theta)$ amounts to minimizing the Likelihood divergence $KL_m(\theta, P_n)$ and produces the ML estimate of θ_T.

2.2 Asymptotic derivation

We assume that

$$\lim_{n\to\infty} P_n = P_{\theta_T} \quad \text{a.s.}$$

This holds for example when the X_i's are drawn as an iid sample with common law P_{θ_T} which we may assume in the present context. From an asymptotic standpoint, Kullback-Leibler divergence is related to the way P_n keeps away from P_θ when θ is not equal to the true value of the parameter θ_T generating the observations X_i's and is closely related with the type of sampling of the X_i's. In the present case, when i.i.d. sampling of the $X_{i,\theta}$'s under P_θ are performed, Sanov Large Deviation theorem leads to

$$\lim_{n\to\infty} \frac{1}{n} \log P_\theta\left(P_{n,\theta} = P_n \middle| P_n\right) = -KL(\theta_T, \theta). \tag{10}$$

This result can easily be obtained from (7) using Stirling formula to handle the factorial terms and the law of large numbers which states that for all j's, n_j/n tends to $P_{\theta_T}(d_j)$ as n tends to infinity. Comparing with (8) we note that the MLE θ_{ML} is a proxy of the minimizer of the natural estimator θ_T of $KL(\theta_T, \theta)$ in θ, substituting the unknown measure generating the X_i's by its empirical counterpart P_n . Alternatively as will be used in the sequel, θ_{ML} minimizes upon θ the Likelihood divergence $KL_m(\theta, \theta_T)$ between P_θ and P_{θ_T} substituting the unknown measure P_{θ_T} generating the X_i's by its empirical counterpart P_n . Summarizing we have obtained:

The ML estimate can be obtained from a LDP statement as given in (10), optimizing in θ in the estimator of the LDP rate where the plug-in method of the empirical measure of the data is used instead of the unknown measure P_{θ_T}. Alternatively it holds

$$\theta_{ML} := \arg\min_{\theta} \widehat{KL_m}\left(\theta, \theta_T\right) \tag{11}$$

with

$$\widehat{KL_m}\left(\theta, \theta_T\right) := KL_m\left(\theta, P_n\right).$$

This principle will be kept throughout this paper: the estimator is defined as maximizing the probability that the simulated empirical measure be close to the empirical measure as observed on the sample, conditionally on it, following the same sampling scheme. This yields a maximum likelihood estimator, and its properties are then obtained when randomness is introduced as resulting from the sampling scheme.

3 Bootstrap and weighted sampling

The sampling scheme which we consider is commonly used in connection with the bootstrap and is referred to as the *weighted* or *generalized bootstra*p, sometimes called *wild bootstrap*, first introduced by Newton and Mason [17].

Let $X_1, ..., X_n$ with common distribution P_{θ_T} on $\mathcal{Y} := \{d_1, .., d_K\}$.

Consider a collection $W_1, ..., W_n$ of independent copies of W, whose distribution satisfies the conditions stated in Section 1. The weighted empirical measure P_n^W is defined through

$$P_n^W := \frac{1}{n}\sum_{i=1}^{n} W_i \delta_{X_i}.$$

This empirical measure need not be a probability measure, since its mass may not equal 1. Also it might not be positive, since the weights may take negative values. Therefore P_n^W can be identified with a random point in $\mathbb{R}^K$. The measure P_n^W converges almost surely to P_{θ_T} when the weights W_i's satisfy the hypotheses stated in Section 1. Indeed general results pertaining to this sampling procedure state that under regularity, functionals of the measure P_n^W are asymptotically distributed as are the same functionals of P_n when the X_i's are i.i.d. Therefore the weighted sampling procedure mimics the i.i.d. sampling fluctuation in a two steps procedure: choose n values of X_i such that they asymptotically fit to P_{θ_T}, which means

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} = P_{\theta_T}$$

a.s. and then play the W_i's on each of the x_i's. Then get P_n^W, a proxy to the random empirical measure P_n .

We also consider the normalized weighted empirical measure

$$\mathfrak{P}_n^W := \sum_{i=1}^{n} Z_i \delta_{X_i} \tag{12}$$

where

$$Z_i := \frac{W_i}{\sum_{j=1}^{n} W_j} \tag{13}$$

whenever $\sum_{j=1}^{n} W_j \neq 0$, and

$$\mathfrak{P}_n^W = \infty$$

when $\sum_{j=1}^{n} W_j = 0$, where $\mathfrak{P}_n^W = \infty$ means $\mathfrak{P}_n^W(d_k) = \infty$ for all d_k in $\mathcal{Y}$.

3.1 A conditional Sanov type result for the weighted empirical measure

We now state a conditional Sanov type result for the family of random measures $\mathfrak{P}_n^W$. It follows readily from a companion result pertaining to P_n^W and enjoys a simple form when the weights W_i are associated to power divergences, as defined in Section 1.1.2. We quote the following results, referring to [11].

Consider a set Ω in $\mathbb{R}^K$ such that

$$cl\Omega = cl(Int\Omega) \tag{14}$$

which amounts to a regularity assumption (obviously met when Ω is an open set), which allows for the replacement of the usual lim inf and lim sup by standard limits in usual LDP statements. We denote by P^W the probability measure of the random family of iid weights W_i.

It then holds

Proposition 1 *(Theorem 9 in [11])The weighted empirical measure P_n^W satisfies a conditional Large Deviation Principle in $\mathbb{R}^K$ namely, denoting P the a.s. limit of P_n,*

$$\lim_{n\to\infty} \frac{1}{n} \log P^W \left(P_n^W \in \Omega \middle| X_1^n \right) = -\phi^W(\Omega, P)$$

where $\phi^W(\Omega, P) := \inf_{Q\in\Omega} \phi^W(Q, P)$.

As a direct consequence of the former result, it holds, for any $\Omega \subset \mathbb{S}^K$ satisfying (14), where $\mathbb{S}^K$ designates the simplex of all pm's on $\mathcal{Y}$

Theorem 2 *(Theorem 12 in [11])The normalized weighted empirical measure $\mathfrak{P}_n^W$ satisfies a conditional Large Deviation Principle in $\mathbb{S}^K$*

$$\lim_{n\to\infty} \frac{1}{n} \log P^W \left(\mathfrak{P}_n^W \in \Omega \middle| X_1^n \right) = -\inf_{m\neq 0} \phi^W(m\Omega, P). \tag{15}$$

A flavour of the simple proofs of Proposition 1 and Theorem 2 is presented in the Appendix; see [11] for a detailed treatment; see also Theorem 3.2 and Corollary 3.3 in [23] where Theorem 2 is proved in a more abstract setting. Note that the mapping $Q \to \inf_{m\neq 0} \phi^W(mQ, P)$ is indeed a divergence in the simplex $\mathbb{S}^K$ for all pm P defined on $\mathcal{Y}$ with positive entries.

We will be interested in the pm's in Ω which minimize the RHS in the above display. The case when ϕ^W is a power divergence, namely $\phi^W = \phi_\gamma$ for some real $\gamma \in (-\infty, 1] \cup [2, \infty)$ enjoys a special property with respect to the pm's Q achieving the infimum (upon Q in Ω) in (15). It holds

Proposition 3 *(Lemma 14 in [11])Assume that ϕ^W is a power divergence. Then*

$$Q \in \arg\inf \left\{ \inf_{m\neq 0} \phi^W(mQ, P), Q \in \Omega \right\}$$

and

$$Q \in \arg\inf \left\{ \phi^W(Q, P), Q \in \Omega \right\}$$

are equivalent statements.

Indeed Proposition 3 holds as a consequence of the following results, to be used later on.

Lemma 4 *For Q and P two pm's such that the involved expressions are finite, it holds*

(i) For $\gamma \in (0, 1)$ it holds $\inf_{m\neq 0} \phi_\gamma(mQ, P) = (1-\gamma)\, \phi_\gamma(Q, P)$.

(ii) For $\gamma < 0$ and $\gamma > 1$ it holds $\inf_{m\neq 0} \phi_\gamma(mQ, P) = \frac{1}{\gamma}\left[1 - (1 + \gamma(\gamma-1)\phi_\gamma(Q, P))^{-1/(\gamma-1)}\right]$.

(iii) $\inf_{m\neq 0} \phi_1(mQ, P) = 1 - \exp(-KL(Q, P)) = 1 - \exp(-\phi_1(Q, P))$.

(iv)$\inf_{m\neq 0} \phi_0(mQ, P) = KL_m(Q, P) = \phi_0(Q, P)$

The weighted empirical measure P_n^W has been used in the weighted bootstrap (or wild bootstrap) context, although it is not a pm. However, conditionally upon the sample points, its produces statistical estimators $T(P_n^W)$ whose weak behavior (conditionally upon the sample) converges to the same limit as does $T(P_n)$ when normalized on the classical CLT range; see eg Newton and Mason [17]. Large deviation theorem for the weighted empirical measure P_n^W has been obtained by [1]; for other contributions in line with those, see [18] and [23]. Normalizing the weights produces families of exchangeable weights Z_i, and the normalized weighted empirical measure $\mathfrak{P}_n^W$ is the cornerstone for the so-called non parametric Bayesian bootstrap, initiated by [22], and further developed by [19] among others. Note however that in this context the RV's W_i's are chosen as distributed as standard exponential variables. The link with spacings from a uniform distribution and the corresponding reproducibility of the Dirichlet distributions are the basic ingredients which justify the non parametric bootstrap approach; in the present context, the choice of the distribution of the W_i's is a natural extension of this paradigm, at least when those W_i's are positive RV's.

3.2 Maximum Likelihood for the generalized bootstrap

We will consider maximum likelihood in the same spirit as developed in Section 2.2, here in the context of the normalized weighted empirical measure; it amounts to justify minimum divergence estimators as appropriate MLE's under such bootstrap procedure.

We thus consider the same statistical model $\mathcal{P}_\Theta$ and keep in mind the ML principle as seen as resulting from a maximization of the conditional probability of getting simulated observations close to the initially observed data. Similarly as in Section 2 fix an arbitrary θ and simulate $X_{1,\theta},..,X_{n,\theta}$ with distribution P_θ. Define accordingly $P^W_{n,\theta}$ and $\mathfrak{P}^W_{n,\theta}$ making use of iid RV's $W_1,..,W_n$. Now the event $\mathfrak{P}^W_{n,\theta}(k) = n_k/n$ has probability 0 in most cases (for example when W has a continuous distribution), and therefore we are led to consider events of the form $\mathfrak{P}^W_{n,\theta} \in V_\varepsilon(P_n)$, meaning $\max_k \left|\mathfrak{P}^W_{n,\theta}(d_k) - P_n(d_k)\right| \leq \varepsilon$ for some positive ε; notice that $V_\varepsilon(P_n)$ defined through $V_\varepsilon(P_n) := \left\{Q \in \mathbb{S}^K : \max_k |Q(d_k) - P_n(d_k)| \leq \varepsilon\right\}$ has non void interior.

For such a configuration consider

$$P^W\left(\mathfrak{P}^w_{n,\theta} \in V_\varepsilon(P_n)\middle| X_{1,\theta},..,X_{n,\theta},X_1,..,X_n\right) \tag{16}$$

where the $X_{i,\theta}$ are randomly drawn iid under P_θ. Obviously for θ far away from θ_T the sample $(X_{1,\theta},..,X_{n,\theta})$ is realized "far away " from $(X_1,..,X_n)$, which has been generated under the truth, namely P_{θ_T} , and the probability in (16) is small, whatever the weights, for small ε.

We will now consider (16) asymptotically on n, since, in contrast with the first derivation of the standard MLE in Section 2.1, we cannot perform the same calculation for each n, which was based on multinomial counts. Note that we obtained a justification for the usual MLE through the asymptotic Sanov LDP, leading to the KL divergence and finally back to the MLE through an approximation step of this latest.

We first state

Theorem 5 *With the above notation the following conditioned LDP result holds, for some $\alpha < 1 < \beta$*

$$-\inf_{m\neq 0}\phi^W(mV_{\alpha\epsilon}(P_{\theta_T}),\theta) \leq \lim_{n\to\infty}\frac{1}{n}\log P^W\left(\mathfrak{P}^W_{n,\theta} \in V_\epsilon(P_n)|X_{1,\theta},...,X_{n,\theta},X_1,..,X_n\right) \tag{17}$$

$$\leq -\inf_{m\neq 0}\phi^W(mV_{\beta\epsilon}(P_{\theta_T}),\theta)$$

where $\phi^W(V_{c\epsilon}(\theta_T),\theta) = \inf_{\mu\in V_{c\epsilon}(P_{\theta_T}))}\phi^W(\mu,\theta)$.

The above result follows from Theorem 15 together with the a.s. convergence of P_n to P_{θ_T} in $\mathbb{S}^K$.

From the above result it appears that as $\varepsilon \to 0$, by continuity it holds

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log P^W\left(\mathfrak{P}^W_{n,\theta}\in V_\epsilon(P_n)|X_{1,\theta},...,X_{n,\theta},X_1,..,X_n\right)=-\inf_{m\neq 0}\phi^W(mP_{\theta_T},\theta). \tag{18}$$

The ML principle amounts to maximize $P^W\left(\mathfrak{P}^W_{n,\theta}\in V_\epsilon(P_n)|X_{1,\theta},...,X_{n,\theta},X_1,..,X_n\right)$ upon θ. Whenever Θ is a compact set we may insert this optimization in (17) which yields, following (18)

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log\sup_\theta P^W\left(\mathfrak{P}^W_{n,\theta}\in V_\epsilon(P_n)|X_{1,\theta},...,X_{n,\theta},X_1,..,X_n\right)=-\inf_{\theta\in\Theta}\inf_{m\neq 0}\phi^W(mP_{\theta_T},\theta).$$

By Proposition 3 the argument of the infimum upon θ in the RHS of the above display coincides with the corresponding argument of $\phi^W(\theta_T,\theta)$, which obviously gets θ_T. This justifies to consider a proxy of this minimization problem as a "ML" estimator based on normalized weighted data.

A further interpretation of the MDE as a Maximum a posteriory estimator (MAP) in the context of non parametric bayesian procedures may also be proposed; this is postponed to a next paper.

Since

$$\phi^W(\theta_T,\theta)=\overset{\sim}{\phi}{}^W(\theta,\theta_T)$$

the ML estimator is obtained as in the conventional case by plug in the LDP rate. Obviously the "best" plug in consists in the substitution of P_{θ_T} by P_n, the empirical measure of the sample, since P_n achieves the best rate of convergence to P_{θ_T} when confronted to any bootstrapped version, which adds "noise" to the sampling. We may therefore call

$$\begin{aligned}\theta^W_{ML}&:=\arg\inf_{\theta\in\Theta}\overset{\sim}{\phi}{}^W(\theta,P_n):=\arg\inf_{\theta\in\Theta}\sum_{k=1}^K P_n(d_k)\widetilde{\varphi}\left(\frac{P_\theta(d_k)}{P_n(d_k)}\right) \\ &=\arg\inf_{\theta\in\Theta}\sum_{k=1}^K P_\theta(d_k)\varphi\left(\frac{P_n(d_k)}{P_\theta(d_k)}\right)\end{aligned} \tag{19}$$

the MLE for the bootstrap sampling; here $\overset{\sim}{\phi}{}^W$ (with divergence function $\widetilde{\varphi}$) is the conjugate divergence of ϕ^W (with divergence function φ) .Since $\phi^W=\phi_\gamma$ for some γ, it holds $\overset{\sim}{\phi}{}^W=\phi_{1-\gamma}$.

Obviously we can also plug in the normalized weighted empirical measure, which also is a proxy of P_{θ_T} for each run of the weights. This produces a boot-

strap estimate of θ_T through

$$\theta_B^W := \arg\inf_{\theta\in\Theta} \overset{\sim}{\phi}^W(\theta, \mathfrak{P}_n^W) := \arg\inf_{\theta\in\Theta} \sum_{k=1}^K \mathfrak{P}_n^W(d_k)\widetilde{\varphi}\left(\frac{P_\theta(d_k)}{\mathfrak{P}_n^W(d_k)}\right) \tag{20}$$
$$= \arg\inf_{\theta\in\Theta} \sum_{k=1}^K P_\theta(d_k)\varphi\left(\frac{\mathfrak{P}_n^W(d_k)}{P_\theta(d_k)}\right)$$

where $\mathfrak{P}_n^W$ is defined in (12), assuming n large enough such that this the sum of the W_i's not zero. Whenever W has positive probability to assume value 0, these estimators are defined for large n in order that $\mathfrak{P}_n^W(d_k)$ be positive for all k. Since $E(W) = 1$, this occurs for large samples.

When $\mathcal{Y}$ is not a finite space then an equivalent construction can be developed based on the variational form of the divergence; see [6].

Remark 6 *We may also consider cases when the MLE defined through θ_{ML}^W defined in (19) coincide with the standard MLE θ_{ML} under iid sampling, and when their bootstrapped counterparts θ_B^W defined in (20) coincides with the bootstrapped standard MLE θ_{ML}^b defined through the likelihood estimating equation where the factor $1/n$ is substituted by the weight Z_i . It is proved in Theorem 5 of [8] that whenever $\mathcal{P}_\Theta$ is an exponential family with natural parametrization $\theta \in \mathbb{R}^d$ and sufficient statistics T*

$$P_\theta(d_j) = \exp\left[T(d_j)'\theta - C(\theta)\right], \qquad 1 \le j \le K$$

where the Hessian matrix of $C(\theta)$ is definite positive, then for all divergence pseudo distance ϕ satisfying regularity conditions (including therefore the present cases), θ_{ML}^W equals θ_{ML}, the classical MLE in $\mathcal{P}_\Theta$ defined as the solution of the normal equation

$$\frac{1}{n}\sum T(X_i) = \nabla C(\theta_{ML})$$

irrespectively upon ϕ. Therefore on regular exponential families, and under iid sampling, all minimum divergence estimators coincide with the MLE (which is indeed one of them). The proof of this result is based on the variational form of the estimated of divergence $Q \to \phi(Q, P)$, which coincides with the plug in version in (19) when the common support of all distribution in $\mathcal{P}_\Theta$ is finite. Following verbatim the proof of Theorem 5 in [8] substituting P_n by $\mathfrak{P}_n^W$ it results that θ_B^W equals the weighted MLE (standard generalized bootstrapped MLE θ_{ML}^b) defined through the normal equation

$$\sum_{i=1}^n Z_i T(X_i) = \nabla C(\theta_{ML}^b).$$

where the Z_i's are defined in (13). This fact holds for any choice of the weights, irrespectively on the choice of the divergence function φ with the only restriction

that it satisfies the mild conditions (RC) in [8]. It results that for those models any generalized bootstrapped MDE coincides with the corresponding bootstrapped MLE.

Remark 7 *The estimators θ_{ML}^{W} defined in (19) have been considered for long irrespectively of the present approach; see e.g. [20]. Their statistical properties in various contexts have been studied for general support spaces $\mathcal{Y}$ in [6] for parametric models, and for various semi parametric models in [7] and [4].*

Example 8 *A-In the case when W is a RV with standard exponential distribution, then the normalized weighted empirical measure $\mathfrak{P}_n^W$ is a realization of the a posteriori distribution for the non informative prior on the non parametric distribution of X. See [22]. In this case $\varphi(x) = -\log x + x - 1$ and $\widetilde{\varphi}(x) = x \log x - x + 1$; the resulting estimator is the minimum Kullback-Leibler one.*

B-When W has a standard Poisson distribution then the couple $(\varphi, \widetilde{\varphi})$ is reverse wrt the above one, and the resulting estimator is the minimum modified Kullback-Leibler one. which takes the usual weighted form of the standard generalized bootstrap MLE

$$\theta_B^{POI(1)} := \arg\sup_{\theta} \sum_{k=1}^{K} \left(\frac{\sum_{i=1}^{n} W_i 1_k(X_i)}{\sum_{i=1}^{n} W_i} \right) \log P_\theta(k)$$

which is defined for n large enough so that $\sum_{i=1}^{n} W_i \neq 0$. Also in this case θ_{ML}^{W} coincides with the standard MLE.

C-In case when W has an Inverse Gaussian distribution IG(1,1) then $\varphi(x) = \varphi_{-1}(x) = \frac{1}{2}(x-1)^2/x$ for $x > 0$ and the ML estimator minimizes the Pearson Chi-square divergence with generator function $\varphi_2(x) = \frac{1}{2}(x-1)^2$ which is defined on $\mathbb{R}$.

D-When W follows a normal distribution with expectation and variance 1, then the resulting divergence is the Pearson Chi-square divergence $\varphi_2(x)$ and the resulting estimator minimizes the Neyman Chi-square divergence with $\varphi(x) = \varphi_{-1}(x)$.

E-When W has a Compound Poisson Gamma distribution $C(POI(2), \Gamma(2,1))$ distribution then the corresponding divergence is $\varphi_{1/2}(x) = 2(\sqrt{x} - 1)^2$ which is self conjugate, whence the ML estimator is the minimum Hellinger distance one.

4 Optimal weighting in relation with the bootstrapped estimator of the divergence

The definition of the divergence estimator (20) opens to the definition of various bootstrap estimators for the divergence pseudo distance between P_{θ_T} and the model $\mathcal{P}$. Indeed the choice of the distribution of the weights W_i's in $\mathfrak{P}_n^W$ needs not be related to the divergence function ϕ; for example we might define

some $\mathfrak{P}_n^V$ in order to define an estimator of $\phi^W(Q,P)$ through (20) with $\mathfrak{P}_n^W$ substituted by $\mathfrak{P}_n^V$ where

$$Z_i := \frac{V_i}{\sum_{j=1}^n V_j}$$

and the vector $(V_1,..,V_n)$ is not related in any way with ϕ^W. The resulting generalized bootstrapped estimate of $\phi^W(Q,P)$ may also be used as a test statistics in order to assess whether $\theta = \theta_T$ for example. It may seem a natural insight that the choice when $(V_1,..,V_n)$ has same distribution as $(W_1,..,W_n)$ should bear some optimality property. The rôle of the present section is to explore this question, for specific choices of the distribution of the V_i's.

4.1 Comparing bootstrapped statistics

Let ϕ^W be a power divergence defined by some weight W through (4). We assume that θ_T is known and we measure the divergence between $\mathfrak{P}_n^W$ and P_{θ_T} as a bootstrapped version of the corresponding distance between P_n and P_{θ_T}, where the distance is suited to the distribution of the weights. We compare the decay to 0 of this same distance with the corresponding decay substituting $\mathfrak{P}_n^W$ by $\mathfrak{P}_n^V$ for some competing family of weights $(V_1,..,V_n)$. Both RV's W and V are assumed to have distributions such that the Legendre transform of their cumulant generating functions belong to the Cressie Read family of divergences. The divergence ϕ^W is associated to the generator φ_γ and, respectively, V is associated to a generator $\varphi_{\gamma'}$ by the corresponding formula (4). For brevity we

restrict the discussion to RV's W and V which are associated to divergence functions φ_γ and $\varphi_{\gamma'}$ with $\gamma \in (0,1)$, as other cases are similar, making use of the corresponding formulas from Lemma 4.

The bootstrap distance between P_{θ_T} and the bootstrapped dataset will be defined as $\widetilde{\phi_\gamma}\left(\theta_T, \mathfrak{P}_n^W\right)$.

Looking at case (i) in Proposition 3 we denote by φ the generator of the divergence $Q \to \inf_{m\neq 0} \phi_\gamma(mQ,P)$ and ψ the generator of the divergence $Q \to \inf_{m\neq 0} \phi_{\gamma'}(mQ,P)$ from which

$$\varphi(x) = \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma}$$

$$\psi(x) = \frac{x^{\gamma'} - \gamma' x + \gamma' - 1}{\gamma'}.$$

Also for clearness we denote

$$\Phi(Q,P) := \inf_{m\neq 0} \phi_\gamma(mQ,P)$$

and

$$\Psi(Q,P) := \inf_{m\neq 0} \phi_{\gamma'}(mQ,P)$$

For $\gamma \in (0,1)$ due to Proposition 3 (i) and Lemma 4, for Q and P in $\mathbb{S}^K$ with non null entries, we define the conjugate divergence $\widetilde{\Phi}(Q,P) := \Phi(P,Q)$ and the generator of $Q \to \widetilde{\Phi}(Q,P)$ writes

$$\widetilde{\varphi}(x) := (\gamma - 1)\, \varphi_{1-\gamma}(x)$$

we will denote accordingly $\widetilde{\psi}(x) := (\gamma' - 1)\, \varphi_{1-\gamma'}(x)$ the generator of $\widetilde{\Psi}(Q,P)$, the conjugate divergence of $\Psi(Q,P)$.

In order to simplify the notation , for any event A, $P^W_{X_1^n}(A)$ denotes the probability of A conditioned upon $(X_1,..,X_n)$.

By Theorem 2

$$\lim_{n\to\infty} \frac{1}{n} \log P^W_{X_1^n}\left(\widetilde{\phi_\gamma}\left(\theta_T, \mathfrak{P}^V_n\right) > t\right) = -\inf\left\{\Psi(Q,\theta_T), Q : \widetilde{\phi_\gamma}(\theta_T, Q) > t\right\}. \tag{21}$$

We prove

Proposition 9 *For $\gamma \in (0,1)$ and any $\gamma' \notin (1,2)$, for any positive t,*

$$\lim_{n\to\infty} \frac{1}{n} \log P^W_{X_1^n}\left(\widetilde{\phi_\gamma}\left(\theta_T, \mathfrak{P}^W_n\right) > t\right) = -t(1-\gamma) \qquad ((\text{i}))$$

while

$$\lim_{n\to\infty} \frac{1}{n} \log P^W_{X_1^n}\left(\widetilde{\phi_\gamma}\left(\theta_T, \mathfrak{P}^V_n\right) > t\right) \geq -t(1-\gamma). \qquad ((\text{ii}))$$

Proof: By Theorem 2, since $\Phi(Q,\theta) = (1-\gamma)\phi_\gamma(Q,\theta)$, it holds for any $t > 0$

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log P^W_{X_1^n}\left(\widetilde{\phi_\gamma}\left(\theta_T, \mathfrak{P}^W_n\right) > t\right) \\
&= -\inf\left\{\Phi(Q,\theta_T), Q : \widetilde{\phi_\gamma}(\theta_T, Q) > t\right\} \\
&= -\inf\left\{\Phi(Q,\theta_T), Q : \phi_\gamma(Q,\theta_T) > t\right\} \\
&= -\inf\left\{\Phi(Q,\theta_T), Q : \Phi(Q,\theta_T) > t(1-\gamma)\right\} \\
&= -t(1-\gamma)
\end{aligned}$$

which proves (i).

Using (21), for any $t > 0$

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \log P^W_{X_1^n}\left(\widetilde{\phi_\gamma}\left(\theta_T, \mathfrak{P}^V_n\right) > t\right) &= -\inf\left\{\Psi(Q,\theta_T), Q : \phi_\gamma(Q,\theta_T) > t\right\} \\
&= -\inf\left\{\Psi(Q,\theta_T), Q : \Phi(Q,\theta_T) > t(1-\gamma)\right\}
\end{aligned}$$

from which (ii) holds whenever there exists some R in $\mathbb{S}^K$ satisfying both

$$\Phi(R,\theta_T) > t(1-\gamma)$$

and

$$\Psi\left(R,\theta_T\right)\leq t(1-\gamma).$$

For $K\geq 2$, let $R:=(r_1,..,r_K)$ such that $r_i=ap_i$ for $i=1,..,K-1$ where $P_{\theta_T}:=(p_1,..,p_K)\in\mathbb{S}^K$ with non null entries. Assume $a<1$. Then $a\to\Phi\left(R,P\right)$ is decreasing on $(0,1)$, $\lim_{a\to 0}\Phi\left(R,P\right)=+\infty$ and $\lim_{a\to 1}\Phi\left(R,P\right)=0$; thus there exists $a_\varphi(t)$ such that for $a\in\left(0,a_\varphi(t)\right)$, it holds $\Phi\left(R,P\right)>t(1-\gamma)$. In the same way there exists $a_\psi(t)$ such that for $a\in(a_\psi(t),1)$ it holds $\Psi\left(R,P\right)<t(1-\gamma)$. Hence for $a\in\left(\min\left(a_\varphi(t),a_\psi(t)\right),\max\left(a_\varphi(t),a_\psi(t)\right)\right)$, there exists some R which satisfies the claim.

To summarize the meaning of Proposition 9, one can say that it inlights the necessary fit between the divergence and the law of the weights when exploring the asymptotic behavior of the bootstrapped empirical measure. It can also be captured stating that given a divergence ϕ_γ there exists an optimal bootstrap in the sense that the chances for $\widetilde{\phi_\gamma}\left(\theta_T,\mathfrak{P}_n^W\right)$ to be large are minimal; the "noise" caused by the weights is tampered down when those are fitted to the divergence, hence in no way in an arbitrary way.

4.2 Bahadur efficiency of minimum divergence tests under generalized bootstrap

In [13] Efron and Tibshirani suggest the bootstrap as a valuable approach for testing, based on bootstrapped samples. We show that bootstrap testing for parametric models based on appropriate divergence statistics enjoys maximal Bahadur efficiency with respect to any bootstrap test statistics.

The standard approach to Bahadur efficiency can be adapted for the present generalized Bootstrapped tests as follows.

Consider the test of some null hypothesis H0: $\theta_T=\theta$ versus a simple hypothesis H1 $\theta_T=\theta'$.

We consider two competitive statistics for this problem. The first one is based on the bootstrap estimate of $\overset{\sim}{\phi}^W\left(\theta,\theta_T\right)$ is defined through

$$T_{n,X}:=\widetilde{\Phi}\left(\theta,\mathfrak{P}_{n,X}^W\right)=T\left(\mathfrak{P}_{n,X}^W\right)$$

which allows to reject H0 for large values since $\lim_{n\to\infty}T_{n,X}=0$ whenever H0 holds. In the above display we have emphasized in $\mathfrak{P}_{n,X}^W$ the fact that we have used the RV X_i's. Let

$$L_n(t):=P^W(T_{n,X}>t|\,X_1,..,X_n).$$

We use P^W to emphasize the fact that the hazard is due to the weights. Consider now a set of RV's $Z_1,..,Z_n$ extracted from a sequence such that $\lim_{n\to\infty}P_{n,Z}=P_{\theta'}$ a.s ; we have denoted $P_{n,Z}$ the empirical measure of $(Z_1,..,Z_n)$; accordingly define $\mathfrak{P}_{n,Z}^{W'}$, the normalized weighted empirical measure of the Z_i

's making use of weights $(W_1', .., W_n')$ which are iid copies of $(W_1, .., W_n)$, drawn independently from $(W_1, .., W_n)$. Define accordingly

$$T_{n,Z} := \widetilde{\Phi}\left(\theta, \mathfrak{P}_{n,Z}^{W'}\right) = T\left(\mathfrak{P}_{n,Z}^{W'}\right).$$

Define

$$L_n(T_{n,Z}) := P^W(T_{n,W} > T_{n,Z} | X_1, .., X_n)$$

which is a RV (as a function of $T_{n,Z}$) . It holds

$$\lim_{n\to\infty} T_{n,Z} = \widetilde{\Phi}(\theta, \theta') \quad \text{a.s}$$

and therefore the Bahadur slope for the test with statistics T_n is $\Phi(\theta', \theta)$ as follows from

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \log L_n(T_{n,Z}) &= -\inf\left\{\Phi(Q, \theta_T) : \widetilde{\Phi}(\theta, Q) > \widetilde{\Phi}(\theta, \theta')\right\} \\ &= -\inf\{\Phi(Q, \theta_T) : \Phi(Q, \theta) > \Phi(\theta', \theta)\} \\ &= -\Phi(\theta', \theta)\end{aligned}$$

if $\theta_T = \theta$. Under H0 the rate of decay of the $p-$value corresponding to a sampling under H1 is captured through the divergence $\Phi(\theta', \theta)$.

Consider now a competitive test statistics $S\left(\mathfrak{P}_{n,X}^W\right)$ and evaluate its Bahadur slope. Similarly as above it holds, assuming continuity of the functional S on $\mathbb{S}^K$

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \log P^W\left(S\left(\mathfrak{P}_{n,X}^W\right) > S\left(\mathfrak{P}_{n,Z}^{W'}\right) \middle| X_1, .., X_n\right) &= -\inf\{\Phi(Q, \theta_T) : S(Q) > S(\theta')\} \\ &\geq -\Phi(\theta', \theta_T)\end{aligned}$$

as follows from the continuity of $Q \to \Phi(Q, \theta_T)$. Hence the Bahadur slope of the test based on $S\left(\mathfrak{P}_{n,X}^W\right)$ is larger or equal $\Phi(\theta', \theta)$.

We have proved that the chances under H0 for the statistics $T_{n,X}$ to exceed a value obtained under H1 are (asymptotically) less that the corresponding chances associated with any other statistics based on the same bootstrapped sample; as such it is most specific on this scale with respect to any competing ones. Namely

Proposition 10 *Under the weighted sampling the test statistics $T\left(\mathfrak{P}_{n,X}^W\right)$ is most efficient among all tests which are empirical versions of continuous functionals on $\mathbb{S}^K$.*

5 Appendix

5.0.1 A heuristic derivation of the conditional LDP for the normalized weighted empirical measure

The following sketch of proof gives the core argument which yields to Proposition 1; a proof adapted to a more abstract setting can be found in [23], following their Theorem 3.2 and Corollary 3.3, but we find it useful to present a proof which reduces to simple arguments. We look at the probability of the event

$$P_n^W \in V(R) \tag{22}$$

for a given vector R in $\mathbb{R}^K$, where $V(R)$ denotes a neighborhood of R, therefore defined through

$$(Q \in V(R)) \Longleftrightarrow (Q(d_l) \approx R(d_l); 1 \leq l \leq k)$$

We denote by P the distribution of the RV X so that P_n converges to P a.s.

Evaluating loosely the probability of the event defined in (22) yields, denoting $P_{X_1^n}$ the conditional distribution given $(X_1, .., X_n)$

$$\begin{aligned}
P_{X_1^n}\left(P_n^W \in V(R)\right) &= P_{X_1^n}\left(\bigcap_{l=1}^{K}\left(\frac{1}{n}\sum_{i=1}^{n} W_i \delta_{X_i}(d_l) \approx R(d_l)\right)\right) \\
&= P_{X_1^n}\left(\bigcap_{l=1}^{K}\left(\frac{1}{n}\sum_{i=1}^{n_l} W_{i,l} \approx R(d_l)\right)\right) \\
&= \prod_{l=1}^{K} P_{X_1^n}\left(\frac{1}{n_l}\sum_{i=1}^{n_l} W_{i,l} \approx \frac{n}{n_l} R(d_l)\right) \\
&= \prod_{l=1}^{K} P_{X_1^n}\left(\frac{1}{n_l}\sum_{i=1}^{n_l} W_{i,l} \approx \frac{R(d_l)}{P(d_l)}\right)
\end{aligned}$$

where we used repeatedly the fact that the r.v's W are i.i.d. In the above display, from the second line on, the r.v's are independent copies of W_1 for all i and l. In the above displays n_l is the number of X_i's which equal d_l, and the $W_{i,l}$ are the weights corresponding to these X_i's. Note that we used the convergence of n_l/n to $P(d_l)$ in the last display.

Now for each l in $\{1, 2, .., K\}$ by the Cramer LDP for the empirical mean, it holds

$$\frac{1}{n_l}\log P\left(\frac{1}{n_l}\sum_{i=1}^{n_l} W_{i,l} \approx \frac{R(d_l)}{P(d_l)}\right) \approx -\varphi^W\left(\frac{R(d_l)}{P(d_l)}\right)$$

i.e.

$$\frac{1}{n}\log P\left(\frac{1}{n_l}\sum_{i=1}^{n_l} W_{i,l} \approx \frac{R(l)}{P(l)}\right) \approx -\frac{R(d_l)}{P(d_l)}\varphi^W\left(\frac{R(d_l)}{P(d_l)}\right)$$

as follows from the classical Cramer LDP, and therefore

$$\begin{aligned}
&\frac{1}{n}\log P_{X_1^n}\left(P_n^W \in V(R)\right)\\
&\approx \frac{1}{n}\log P_{X_1^n}\left(\bigcap_{l=1}^{K}\left(\frac{1}{n}\sum_{i=1}^{n_l} W_{i,l} \approx R(d_l)\right)\right)\\
&\to -\sum_{l=1}^{K}\varphi^W\left(\frac{R(d_l)}{P(d_l)}\right)P(d_l) = -\phi^W\left(R,P\right)
\end{aligned}$$

as $n \to \infty$.

A precise derivation of Proposition 1 involves two arguments: firstly for a set $\Omega \subset \mathbb{R}^K$ a covering procedure by small balls allowing to use the above derivation locally, and the regularity assumption (14) which allows to obtain proper limits in the standard LDP statement.

The argument leading from Proposition 1 to Theorem 2 can be summarized now.

For some subset Ω in $\mathbb{S}^K$ with non void interior it holds

$$\left(\mathfrak{P}_n^W \in \Omega\right) = \bigcup_{m\neq 0}\left(\left(P_n^W \in m\Omega\right) \cap \left(\sum_{i=1}^{n} W_i = m\right)\right)$$

and $\left(P_n^W \in m\Omega\right) \subset \left(\sum_{i=1}^{n} W_i = m\right)$ for all $m \neq 0$. Therefore

$$P_{X_1^n}\left(\mathfrak{P}_n^W \in \Omega\right) = P_{X_1^n}\left(\bigcup_{m\neq 0}\left(P_n^W \in m\Omega\right)\right).$$

Making use of Theorem 1

$$\lim_{n\to\infty}\frac{1}{n}\log P_{X_1^n}\left(\mathfrak{P}_n^W \in \Omega\right) = -\phi^W\left(\bigcup_{m\neq 0} m\Omega, P\right).$$

Now

$$\phi^W\left(\bigcup_{m\neq 0} m\Omega, P\right) = \inf_{m\neq 0}\inf_{Q\in\Omega}\phi^W\left(mQ,P\right).$$

We have sketched the arguments leading to Theorem 2; see [11] for details.

References

[1] Barbe, P., Bertail, P. The Weighted Bootstrap. In *Lecture Notes in Statistics* ; Springer-Verlag, New York,1995.

[2] Bar-Lev, S. K.; Enis, P. Reproducibility and natural exponential families with power variance functions. *Ann. Statist.* **1986**, *14*, 1507–1522.

[3] Basu, A; Harris, Ian R.; Hjort, N. L.; Jones, M. C. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **1998**, *85* , 549–559.

[4] Broniatowski, M.; Decurninge, A. Estimation for models defined by conditions on their L-moments. *IEEE Trans. Inform. Theory* **2016**, *62*, 5181–5198.

[5] Broniatowski, M.; Keziou, A. Minimization of ϕ-divergences on sets of signed measures. *Studia Sci. Math. Hungar.* **2006**, *43*, 403–442.

[6] Broniatowski, M.; Keziou, A. Parametric estimation and tests through divergences and the duality technique. *J. Multivariate Anal.* **2009**, *100*, 16–36.

[7] Broniatowski, M.; Keziou, A. Divergences and duality for estimation and test under moment condition models. *J. Statist. Plann. Inference* **2012**, *142*, 2554–2573.

[8] Broniatowski, Michel Minimum divergence estimators, maximum likelihood and exponential families. Statist. Probab. Lett. 93 (2014), 27–33. 62F10 (62B10)

[9] Broniatowski, M. A weighted bootstrap procedure for divergence minimization problems. In *Analytical methods in statistics*; Springer Proc. Math. Stat., 193, Springer, Cham, 2017; pp. 1–22

[10] Broniatowski, M.; Stummer, W. Some universal insights on divergences for statistics, machine learning and artificial intelligence. In *Geometric structures of information*; Signals Commun. Technol., Springer, Cham, 2019; pp. 149–211

[11] Broniatowski, M.; Stummer, W., A precise bare simulation approach to the minimization of some distances. *Foundations*. Preprint under preparation. **2020**.

[12] Broniatowski, M.; Vajda, I. Several applications of divergence criteria in continuous families. *Kybernetika* **2012** , *48*, 600–636.

[13] Efron, B.; Tibshirani, R. J. *An introduction to the bootstrap* ; Chapman and Hall, New York, 1993; pp. xvi+436 .

[14] Grendar M., and Judge, G. Asymptotic equivalence of empirical likelihood and Bayesian MAP. *Ann. Statist.* **2009**, *37*, 2445-2457.

[15] Letac, G.; Mora, M. Natural real exponential families with cubic variance functions. *Ann. Statist.* **1990**, *18*, 1–37.

[16] Liese, F., Vajda, I.*Convex statistical distances;* Teubner-Texte zur Mathematik [Teubner Texts in Mathematics], 95. BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1987.

[17] Mason D. M.; Newton, M. A. A rank statistic approach to the consistency of a general bootstrap, *Ann. Statist.* **1992**, *20*, 1611-1624.

[18] Najim, J. A Cramer type theorem for weighted random variables. *Electron. J. Probab.* **2002**, *7*.

[19] Newton, Michael A.; Raftery, Adrian E. Approximate Bayesian inference with the weighted likelihood bootstrap. With discussion and a reply by the authors. *J. Roy. Statist. Soc.* **1994**, *56*, 3–48.

[20] Pardo, Leandro Statistical inference based on divergence measures. Statistics: Textbooks and Monographs, 185. Chapman & Hall/CRC, Boca Raton, FL, 2006. xx+492 pp.

[21] Read, T. R. C., Cressie, N. A. C. *Goodness-of-fit statistics for discrete multivariate data*; Springer Series in Statistics. Springer-Verlag, New York, 1988; pp.xii+211.

[22] Rubin, Donald B. The Bayesian bootstrap. *Ann. Statist* **1981**, *9*, 130–134.

[23] Trashorras, J; Wintenberger, O. Large deviations for bootstrapped empirical measures. *Bernoulli* **2014**, *20*, 1845–1878.

[24] Zolotarev, Vladimir M. *Modern theory of summation of random variables.* Modern Probability and Statistics ; VSP, Utrecht, 1997; pp. x+412.