A refined Jensen's inequality in Hilbert spaces and empirical approximations.

S. Leorato

University of Rome Tor Vergata

Abstract

Let $f : \mathfrak{X} \to \mathbb{R}$ be a convex mapping and $\mathfrak{X}$ a Hilbert space. In this paper we prove the following refinement of Jensen's inequality:

$$\mathbb{E}(f\,|X \in A\,) \geq \mathbb{E}(f\,|X \in B\,)$$

for every A, B such that $\mathbb{E}(X\,|X \in A\,) = \mathbb{E}(X\,|X \in B\,)$ and $B \subset A$. Expectations of Hilbert space valued random elements are defined by means of the Pettis integrals. Our result generalizes a result of Karlin and Novikov (1963), who derived it for $\mathfrak{X} = \mathbb{R}$. The inverse implication is also true if P is an absolutely continuous probability measure. A convexity criterion based on the Jensen-type inequalities follows and we study its asymptotic accuracy when the empirical distribution function based on a $n-$dimensional sample approximates the unknown distribution function. Some statistical applications are addressed, such as nonparametric estimation and testing for convex regression functions or other functionals.

Key words: Jensen's inequality, supporting hyperplane, empirical measure, convex regression function, linearly ordered classes of sets, Pettis integral.

MSC: 60E15,62G08

1 Introduction

Convexity has numerous implications in almost every field from pure to applied mathematics (including statistics, economics, information theory, utility theory among the others) and this explains the huge amount of monographs and papers on the subject.

One of the most interesting results, establishing a *liaison* between convexity and probability, is the so-called Jensen's inequality.

Let $f : I \to \mathbb{R}$ be a continuous convex function and $(I, \mathcal{B}, P)$ be any probability space on the interval $I \subseteq \mathbb{R}$. Then, Jensen's inequality states that

$$\mathbb{E}(f(X)) \geq f(\mathbb{E}X) \tag{1.1}$$

where $\mathbb{E}$ is the expectation with respect to the measure P. The same inequality holds, almost surely, if the expectation in (1.1) is replaced by the conditional expectation $\mathbb{E}(\cdot\,|\mathcal{F})$ for some sub $\sigma-$field $\mathcal{F} \subset \mathcal{B}$ (see for instance Doob, 1953).

The converse is also true: if (1.1) holds for every probability space $(I, \mathcal{B}, P)$, then f is convex on I.

Extensions to higher-dimensional spaces, finite or infinite, have also been proved (see Perlman (1974)).

In this paper we prove a refined version of Jensen's inequality and we introduce a characterization criterion for convexity. Our framework is the following: let $\mathfrak{X}$ be a Hilbert space and $(\Omega, \mathcal{A}, P)$ a probability space. Let $X : \Omega \to \mathfrak{X}$ a $\mathfrak{X}-$valued random element and P_X the measure induced by X on $(\mathfrak{X}, \mathcal{B}(\mathfrak{X}))$. Let also $f : \mathfrak{X} \to \mathbb{R}$ be lower semicontinuous and convex. Then for every two convex closed subsets $A \supset B$ with the same *baricenter* (*i.e.* such that $\mathbb{E}(X\,|X \in A) = \mathbb{E}(X\,|X \in B)$), it holds

$$\mathbb{E}(f(X)\,|X \in A) \geq \mathbb{E}(f(X)\,|X \in B). \tag{1.2}$$

The converse is also true, provided that P is an absolutely continuous probability measure.

Expectations $\mathbb{E}(X\,|A)$ for random elements of infinite-dimensional spaces are defined by means of the Pettis integral. More details are given in Section 3.

A motivation for this criterion might be the following: suppose you have a sample of n observations from 2 random variables $(X_i, W_i), \quad i = 1, \ldots, n$, assumed to be linked by some unknown functional relationship, $W = f(X)$. A typical example is when W is the regression function of a response variable Y: $W = \mathbb{E}(Y\,|X) = f(X)$. One of the relevant questions in applications is wether the function f is convex (concave). We could think of a data-driven criterion for checking convexity of f based on the empirical version of Jensen's inequality (1.1):

$$\frac{1}{n}\sum_{i=1}^{n} W_i \geq f\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right). \tag{1.3}$$

Since f is unknown, we are not able to derive the right hand side in (1.3), unless $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ coincides with X_l for some *unique* $1 \leq l \leq n$, in which case we can reasonably write $f(\bar{X}_n) = W_l$.

Even ignoring that (a *naïve* approximation of $\bar{W}$ could be given by some convex combination of the nearest observed points) the difference $\frac{1}{n}\sum_{i=1}^{n} W_i - \bar{W}$ captures only information on convexity around $\bar{X}$. In fact, Jensen's inequality implies convexity only if it holds for all possible probability measures on $(\Omega, \mathcal{A})$, that cannot be done in this context, since the data are sampled from a fixed (although unknown) probability distribution.

On the other hand, inequality (1.2) always makes sense, provided that two convex subsets $B \subset A$ satisfying

$$\frac{\sum_{i\in A} X_i}{P_n(A)} = \frac{\sum_{i\in B} X_i}{P_n(B)}$$

do exist (we will actually impose less restrictive conditions on the empirical conditional expectations, allowing for the their absolute difference to be greater than zero in finite samples).

Ever since the pioneering paper by Jensen (1906), hundreds of papers have been devoted to generalizations and refinements of (1.1).

Extensions to multi-functions of infinite-dimensional spaces can be found in Perlman (1974) or in Kozek and Suchanecki (1980), as well as in Zapala (2001) for the conditional version. In order to define convexity, since higher-dimensional spaces are not totally ordered, they introduce the so-called closed cone ordering. To and Wing (1975) prove a conditional Jensen's inequality for Bochner-integrable functions on a Banach space.

Roselli and Willem (2002) extend integral inequality (1.1) to an arbitrary nonnegative measure (not necessarily integrating to one). In Mercer (2003) a variant of Jensen's inequality is proved, namely $f\left(x_1 + x_n - \sum w_k x_k\right) \leq f(x_1) + f(x_n) - \sum w_k f(x_k)$, which is further generalized in Abramovich *et al.* (2003). In Sanchez *et al.* (2005) a weakened version of (1.1) is used in order to introduce the concept of $\varepsilon-$convexity of a function defined on a convex subset of a real Banach space. Recently, Merkle (2005) has proved a version of Jensen's inequality for medians. Dragomir devoted also several papers to refinements and sharpenings of (1.1) (see f.i. Dragomir, 1992).

A result similar to (1.2) (for $\mathfrak{X} \subseteq \mathbb{R}$) can be found in Karlin and Novikoff (1963), who studied the inequality $\int f dQ \leq \int f dP$, P and Q being two different probability measures.

Later, Spiegelman (1980) rediscovered their result. However, the approach used here is completely different and is suitable to extensions to higher dimensional spaces. The main tools of our proof are the supporting hyperplane theorem and the duality lemma (the Fenchel-Legendre transform).

As far as to the author's knowledge, neither inequality (1.2) for functionals of infinite-dimensional spaces, nor the statistical implications of the characterization have been addressed so far.

We underline that if X is an absolutely continuous random variable (or vector), then inequality (1.2) implies (1.1) by trivially chosing $B = \{\mathbb{E}\left(X \,|X \in A\right)\}$, since the conditional distribution $P\left(\cdot \,|X \in B\right)$ reduces to the degenerate distribution at point $\bar{X} = \mathbb{E}\left(X \,|X \in A\right)$.

We first consider the easiest case $\mathfrak{X} \subset \mathbb{R}$. We prove the refined Jensen's inequality for arbitrary probability measures (continuous or discrete) although we show that the inverse implication fails for arbitrary discrete distributions. Section 3 extends the main result to Hilbert spaces with inner product $\langle \cdot, \cdot \rangle$. A characterization of convexity of lower-semicontinuous functionals $f : \mathfrak{X} \to \mathbb{R}$ is then derived, based on the refined Jensen's inequality. In Section 4 we consider the empirical version of inequality (1.2). The last section is devoted to the motivation of the Jensen-type convexity criterion and its empirical approximation in a statistical framework, by the illustration of some possible applications.

The first application consists in tests for convexity of regression models. In Sec-

tion 5.1-5.2 tests for convexity of more general functions are also taken into account (asymptotic upper and lower bounds for probabilities related to errors of I and II type are derived). Section 5.3 motivates the convexity criterion in a minimum divergence inference setting, both for testing and estimation of convex regression functions: mainly, the convenience of our characterization of convexity stems from the fact that it reduces to a set of linear inequality contraints on the distribution function, that are typically the most manageable in a minimum divergence approach. All the applications addressed enter the range of inference under shape restrictions that is a living matter in nonparametric statistics: limiting to the last three years only, we mention, for example, the papers by Abrevaya and Jiang (2005), Baraud *et al.* (2005), Birke and Dette (2007a,2007b), Dette *et al.* (2006), Dümbgen *et al.* (2006), Hall and Van Keilegom (2005), Hall and Yatchew (2005), Orbe *et al.* (2006), Reboul (2005).

Throughout the paper the expression $\mathbb{E}(\cdot\,|A)$ must be read as $\mathbb{E}(\cdot\,|X \in A)$.

2 Real convex functions.

Let $\mathfrak{X}$ be a bounded closed interval of the real line and f a lower semicontinuous function. Let $P = P_X$ be a probability measure induced on $(\mathfrak{X}, \mathcal{B})$ by the mapping $X : \Omega \to \mathfrak{X}$ (we can assume, without loss of generality, that $\mathfrak{X}$ coincides with the support of P). When dealing with a discrete probability distribution, we will assume that the support of P is a set Λ satisfying $ConvexHull(\Lambda) = \mathfrak{X}$.

Theorem 1 *Let P be a probability measure on $(\mathfrak{X}, \mathcal{B})$ and $f : \mathfrak{X} \to \mathbb{R}$ lower semicontinuous. Consider the following statements.*

(i) f is convex in $\mathfrak{X}$

(ii) For every closed convex $A \in \mathcal{B}$ and for every closed and convex $B \subset A$, such that $\mathbb{E}(X\,|A) = \mathbb{E}(X\,|B)$, it holds

$$\mathbb{E}(f(X)\,|A) \geq \mathbb{E}(f(X)\,|B). \tag{2.4}$$

The equality occurs only if f is linear in A.

If P is absolutely continuous then (i) and (ii) are equivalent. If P is discrete, then (i) implies (ii).

Proof $(i) \Rightarrow (ii)$. If f is linear in all A, then the identity $\mathbb{E}(f\,|A) = \mathbb{E}(f\,|B)$ trivially follows by linearity of the expectation operator. Let f be strictly convex almost everywhere in A. The proof follows almost straightforwardly by considering that every convex and closed subset $A \subset \mathfrak{X}$ is either an interval or a single point $\{x\}$. In both cases, if f is stricly convex in A, A can be written as $\{x : f(x) \leq ax + b\}$ for some constants $a, b \in \mathbb{R}$.

Then for every two convex subsets A, B of $\mathfrak{X}$ satisfying $\mathbb{E}(X\,|A) = \mathbb{E}(X\,|B)$ and $B \subset A$, we can find the coefficients a_A, b_A, a_B, b_B, such that

$$A = \{x : f(x) \leq a_A x + b_A\}, \quad B = \{x : f(x) \leq a_B x + b_B\}. \tag{2.5}$$

Consequently

$$A \setminus B = A \cap B^c = \{x : a_B x + b_B < f(x) \leq a_A x + b_A\}.$$

Since the identity $\mathbb{E}(X|A) = \frac{P(A)-P(B)}{P(A)}\mathbb{E}(X|A \cap B^c) + \frac{P(B)}{P(A)}\mathbb{E}(X|B)$ implies $\mathbb{E}(X|A \cap B^c) = \mathbb{E}(X|A) = \mathbb{E}(X|B)$, then we have that

$$\begin{aligned}\mathbb{E}(f|A \cap B^c) &= \int_{A\setminus B} f(x)\frac{p(x)}{P(A \setminus B)}dx > \int_{A\setminus B}(a_B x + b_B)\frac{p(x)}{P(A \setminus B)}dx \\ &= a_B\mathbb{E}(X|A \setminus B) + b_B = a_B\mathbb{E}(X|B) + b_B \\ &= \int_B (a_B x + b_B)\frac{p(x)}{P(B)} \geq \int_B f(x)\frac{p(x)}{P(B)} = \mathbb{E}(f|B).\end{aligned}$$

The result finally follows from

$$\mathbb{E}(f|A) = \frac{P(A)-P(B)}{P(A)}\mathbb{E}(f|A \cap B^c) + \frac{P(B)}{P(A)}\mathbb{E}(f|B) > \mathbb{E}(f|B).$$

Suppose now that f is not strictly convex in A, that is, there exists a subset of A, with positive probability, where f is linear. Then, it might not be possible to find the coefficients a_A, b_A, a_B, b_B as in the above case, such that (2.5) holds. Nevertheless, since f is convex, there exist at least two couples of lines (h_1^A, h_2^A) and (h_1^B, h_2^B) such that

$$A = \left\{x : f(x) \leq \min\left(h_1^A(x), h_2^A(x)\right)\right\}, \quad B = \left\{x : f(x) \leq \min\left(h_1^B(x), h_2^B(x)\right)\right\}.$$

The proof of inequality (2.4) is omitted because it follows the lines of the proof of Theorem 4.

$(ii) \Rightarrow (i)$. Suppose now that P is continuous. Let us assume that f is not convex. Then we can find a convex subset $A \in \mathcal{B}$ where f is locally strictly concave. Therefore, we can choose a subset $B \subset A$ satisfying $\mathbb{E}(X|A) = \mathbb{E}(X|B)$ and apply the first part of the Theorem to the convex function $-f(X) : A \to \mathbb{R}$, to get

$$\mathbb{E}(f(X)|A) < \mathbb{E}(f(X)|B)$$

which contradicts the hypothesis.

Note that if P is absolutely continuous the subset B always exists: a trivial choice is the set $B = \{\mathbb{E}(X|A)\} \subset A$ for which, inequality (2.4) reduces to the classical Jensen's inequality $\mathbb{E}(f(X)|A) < f(\mathbb{E}(X|A)) = \mathbb{E}(f(X)|B)$.

The inverse implication $(ii) \Rightarrow (i)$ doesn't hold for arbitrary discrete probability measures. The reason is that even though the function f is convex on the support Λ of the probability distribution, there might not yet be convex subsets A, B such that $\mathbb{E}(X|A) = \mathbb{E}(X|B)$. For example, let X be a random

variable taking values $\Lambda = \{0, 1, 5, 8\}$ all with equal probability $1/4$. Let f be a mapping from $\mathfrak{X} = ConvHull(\Lambda) = [0, 8]$ to $\mathbb{R}$. Clearly, in this case $\mathbb{E}X = 3, 5 \notin \Lambda$. Moreover, if we consider the class of all sets of the form $C \cap \Lambda$ where C is a convex subset of $\mathfrak{X}$, we obtain

$$\mathcal{C} = \Big\{\{0\}; \{1\}; \{5\}; \{8\}; \{0, 1\}; \{1, 5\}; \{5, 8\}; \{0, 1, 5\}; \{1, 5, 8\}; \{0, 1, 5, 8\}\Big\}.$$

It is easily seen that, for every $A \in \mathcal{C}$, there is no proper subset $B \subset A$, $B \in \mathcal{C}$ such that $\mathbb{E}(X\,|A) = \mathbb{E}\,(X\,|B)$.

Remark 1 *As we already mentioned in the Introduction, the implication $(i) \Rightarrow (ii)$ can also be derived by Lemma b in Karlin and Novikoff (1963). In fact, if $A = [a_1, a_2]$ and $B = [b_1, b_2]$ are such that $a_1 < b_1 < b_2 < a_2$ and such that $\mathbb{E}(X\,|A) = \mathbb{E}(X\,|B)$ holds for some P, then we easily find that the distribution functions $P_A = \frac{P(x)}{\Pr(A)}$ and $P_B = \frac{P(x)}{\Pr(B)}$ satisfy:*

$$\begin{aligned} dP_A(x) - dP_B(x) = dP_A(x) > 0 \quad & x \in [a_1, b_1) \\ dP_A(x) - dP_B(x) =< 0 \qquad\qquad & x \in [b_1, b_2] \\ dP_A(x) - dP_B(x) = dP_A(x) > 0 \quad & x \in (b_2, a_2] \end{aligned}$$

and thus

$$\int_{a_1}^{a_2} f(x)\,(dP_A(x) - dP_B(x)) \geq 0,$$

which yields the result.

Remark 2 *Yet another proof of the sufficiency part (for absolutely continuous probability measures) can be attained by approximating f through a spline function g:*

$$g(x) = f(x_{i+1}) - \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}\,(x_{i+1} - x) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}\,(x - x_i) + f(x_i) \tag{2.6}$$

for $x \in [x_i, x_{i+1})$. The grid points x_i, $i = 0, 1, 2, \ldots, 2k$ must be chosen in order that $\mathbb{E}\,(X\,|C_j) = \mathbb{E}\,(X\,|A)$, $j = 1, \ldots k$, where $C_j = [x_j, x_{j+1}) \cup [x_{2k-j-1}, x_{2k-j})$. Let $A = \bigcup_{j=1}^{k} C_j$ and $B = \bigcup_{j=l}^{k} C_j$. In view of

$$\mathbb{E}\left(g(X)\,\middle|\bigcup_{j=l}^{k} C_j\right) = \mathbb{E}\,(g(X)\,|B) \leq \mathbb{E}\,(g(X)\,|A \setminus B) = \mathbb{E}\left(g(X)\,\middle|\bigcup_{j=1}^{l-1} C_j\right).$$

it is enough to prove the chain of implications $\mathbb{E}\,(g\,|C_j) \geq \mathbb{E}\,(g\,|C_{j+1})$ in order to get $\mathbb{E}\,(g\,|A \setminus B) \geq \mathbb{E}\,(g\,|B)$. The proof of this inequality follows from convexity and stepwise linearity of g.

Finally, the case of an arbitrary function f is derived by allowing the sets C_j to be arbitrarily small and by taking into account that $g \geq f$ and therefore

$$\mathbb{E}(f\,|B) \leq \mathbb{E}(f\,|A \setminus B) + R$$

where $p_j = P\{[x_j, x_{j+1})\}$ *and the term*

$$R = \sum_{j=1}^{l-1} \frac{\Pr(C_j)}{\Pr(A \backslash B)} \left[\frac{p_j}{\Pr(C_j)} \mathbb{E}\Big((\max\{f(x_j), f(x_{j+1})\} - f) \,\Big|\, [x_j, x_{j+1}) \Big) + \right.$$
$$\left. + \frac{p_{2k-j-1}}{\Pr(C_j)} \mathbb{E}\Big((\max\{f(x_{2k-j-1}), f(x_{2k-j})\} - f) \,\Big|\, [x_{2k-j-1}, x_{2k-j}) \Big) \right]$$

can be made arbitrarily small in view of the continuity of f*.*

The advantage of our proof with respect to the one in Karlin and Novikov (1963) as well as to the proof sketched in Remark 2 is that it can be adapted to higher dimensions of $\mathfrak{X}$, because it doesn't require $\mathfrak{X}$ to be totally ordered.

3 Functionals of a Hilbert space.

Let $\mathfrak{X}$ be a closed Hilbert space with inner product $\langle \cdot, \cdot \rangle$. Denote by $\mathfrak{X}^*$ the dual space (of continuous linear functionals) induced by $\langle \cdot, \cdot \rangle$. Let X be a $\mathfrak{X}-$valued random element defined on a probability space $(\Omega, \mathcal{A}, P)$ with range in $\mathfrak{X}$.

The definition of the expectation $\mathbb{E}X$ (and conditional expectations $\mathbb{E}(\cdot\,|A)$) is made in terms of Pettis integrals. $X : \Omega \to \mathfrak{X}$ is Pettis integrable if: (a) for all $x^* \in \mathfrak{X}^*$ $x^*(X)$ is $\mathcal{A}-$measurable; (b) for all $x^* \in \mathfrak{X}^*$, the integral $\int x^*(X) dP$ exists; (c) for all $B \in \mathcal{A}$, there exists a $x_B \in \mathfrak{X}$ (not necessarily unique) such that for all $x^* \in \mathfrak{X}^*$, $x^*(x_B) = \int_B x^*(X) dP$ (see also Perlman (1974)). The Pettis integral x_B for all $B \in \mathcal{A}$ is unique for $\mathfrak{X}$ a Hilbert space. For $B = \Omega$, then $x_B = \mathbb{E}X$. The conditional expectations $\mathbb{E}(X\,|A)$, $A \in \mathcal{B}(\mathfrak{X})$ are defined by $\mathbb{E}(X\,|A) = \frac{1}{P(B)} x_B$, where $B = X^{-1}(A) = \{\omega : X(\omega) \in A\}$.

The fact that $\mathfrak{X}$ is convex and closed ensures that $\mathbb{E}X \in \mathfrak{X}$ and $\mathbb{E}(X\,|A) \in A$ for all convex closed $A \in \mathcal{B}(\mathfrak{X})$. A convex and close support for f guarantees Jensen's inequality to hold under the weaker assumption of lower semicontinuity for f.

Finally, we say that X is absolutely continuous if the probability law induced on $(\mathbb{R}, \mathcal{B})$ by the linear tranformation $x^*(X) : \Omega \to \mathbb{R}$ is absolutely continuous for every $x^* \in \mathfrak{X}^*$.

In order to prove the main theorem, we need to invoke the following results, which are stated here for the convenience of the reader. Although both the theorems hold if $\mathfrak{X}$ is a locally convex Hausdorff topological vector spaces, for the purposes of this paper we assume $\mathfrak{X}$ to be a Hilbert space.

Theorem 2 (Supporting hyperplane theorem) *Every non-empty convex set* $\mathcal{C} \subset \mathfrak{X}$ *has a supporting hyperplane in* $x \in \mathcal{C}$ *if and only if* x *doesn't belong to the relative interior of* $\mathcal{C}$*.*

As a consequence, every closed and convex set can be writen as $\bigcap_j \{x : h_j(x) \leq 0\}$ where $h_j \in \mathfrak{X}^*$ is a continuous linear functional (hyperplane).

For every continuous linear functional $p \in \mathfrak{X}^*$ the Fenchel-Legendre transform of f (the adjoint functional) is defined by

$$f^*(p) = \sup_{x \in \mathfrak{X}} \langle p, x \rangle - f(x).$$

Theorem 3 (Duality Lemma, see Dembo and Zeitouni (1998) p.152) *If f is convex and lower semicontinuous, then*

$$f(x) = \sup_{p \in \mathfrak{X}^*} \langle x, p \rangle - f^*(p). \tag{3.7}$$

In what follows, we shall write, to shorten the notation, $P(A) = P(\omega : X(\omega) \in A)$, for every $A \in \mathcal{B}(\mathfrak{X})$.

Theorem 4 *Let $f : \mathfrak{X} \to \mathbb{R}$ be lower semicontinuous. If X is absolutely continuous the following statements are equivalent:*

(i) f is convex in $\mathfrak{X}$
(ii) For every closed and bounded convex subsets $A, B \subseteq \mathfrak{X}$ satisfying $\mathbb{E}(X\,|A) = \mathbb{E}(X\,|B)$ and $B \subset A$,

$$\mathbb{E}(f\,|A) \geq \mathbb{E}(f\,|B). \tag{3.8}$$

If X is not absolutely continuous then (i) implies (ii).

Proof. The proof of the implication $(ii) \Rightarrow (i)$ is the same as the corresponding proof for the case of real convex funtions.

$(i) \Rightarrow (ii)$. Let A, B be two closed and bounded convex subsets of $\mathfrak{X}$ satisfying $B \subset A$ and $\mathbb{E}(X\,|A) = \mathbb{E}(X\,|B)$. We first consider the case where both A and B can be written as

$$A = \{x \in \mathfrak{X} : f(x) \leq h_A(x)\} \quad B = \{x \in \mathfrak{X} : f(x) \leq h_B(x)\},$$

where h_A and h_B belong to $\mathfrak{X}^*$.

In this case, the proof of inequality (3.8) is just the same as in Theorem 1.

Let us now assume that

$$A = \bigcap_{j \in J_A} \{x : f(x) \leq h_j(x)\} \quad B = \bigcap_{j \in J_B} \{x : f(x) \leq h_j(x)\} \tag{3.9}$$

for some families of hyperplanes $\{h_j,\ j \in J_A\}$ and $\{h_j,\ j \in J_B\}$ respectively. It follows,

$$A \setminus B = \left\{x \in \mathfrak{X} : \min_{j \in J_A} h_j \geq f(x) \geq h_l \quad \text{for some } l \in J_B\right\}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left(f\,|B\right) &= \frac{1}{P(B)}\int_{X^{-1}(B)} f(X)(\omega)dP(\omega) \leq \frac{1}{P(B)}\int_{X^{-1}(B)} \min_{j\in J_B} h_j(X)dP \\
&\leq \text{(apply Jensen's inequality to the concave function } \varphi(x) = \min_j h_j(x)) \\
&\leq \min_{j\in J_B}\int_{X^{-1}(B)} \frac{h_j(X)}{P(B)}dP = \min_{j\in J_B}\int_{X^{-1}(A\setminus B)} \frac{h_j(X)}{P(A\setminus B)}dP \\
&\leq \int_{X^{-1}(A\setminus B)} \frac{h_l(X)}{P(A\setminus B)}dP \leq \int_{X^{-1}(A\setminus B)} \frac{f(X)}{P(A\setminus B)}dP = \mathbb{E}\left(f\,|A\setminus B\right).
\end{aligned}$$

In order to complete the proof, we must show that every closed and bounded convex subset can be approximated arbitrarily well by a set of the form (3.9). We recall that every closed and convex subset of $\mathfrak{X}$ can be written as

$$C = \bigcap_{j\in J}\{x : h_j \geq 0\} = \left\{x : \min_j h_j(x) \geq 0\right\}, \tag{3.10}$$

where $h_j \in \mathfrak{X}^*$ for all j.

Let f^* be the adjoint (dual) function of f: for every $p \in \mathfrak{X}^*$,

$$f^*(p) = \sup_{x\in\mathfrak{X}}\langle p, x\rangle - f(x).$$

From the duality Lemma (see Dembo and Zeitouni (1998), p. 153), the convexity of f implies that $f(x) = \sup_{p\in\mathfrak{X}^*}\langle x, p\rangle - f^*(p)$. For every $M \geq 1$ and every fixed $p \in \mathfrak{X}^*$, we can define the set

$$\begin{aligned}
C_p(M) &= \left\{x \in \mathfrak{X} : M\min_{j\in J} h_j(x) \geq f(x) - \langle x, p\rangle + f^*(p)\right\} \\
&= \left\{x \in \mathfrak{X} : \min_{j\in J}\langle x, Mh_j + p\rangle - f^*(p) \geq f(x)\right\} \\
&= \left\{x \in \mathfrak{X} : \min_{j\in J} h'_j(x) - f^*(p) \geq f(x)\right\} \subset C
\end{aligned} \tag{3.11}$$

with $h'_j = Mh_j + p \in \mathfrak{X}^*$ and $f^*(p)$ constant with x. Since for every fixed $p \in \mathfrak{X}^*$, $f(x) + f^*(p) - \langle x, p\rangle \geq 0$, then the sequence $C_p(M)$ is increasing with M, and $\lim_M C_p(M) = \cup_M C_p(M) = C$. Then for every $\varepsilon > 0$ and every $p \in \mathfrak{X}^*$, we can find M^* and $C_p^* := C_p(M^*)$ such that $P(C_p^*) \geq (1-\varepsilon)P(C)$. Moreover, for every $j \in J$,

$$\begin{aligned}
&\left|\mathbb{E}\left(\langle X, h_j'\rangle - f^*(p)\,|C\right) - \mathbb{E}\left(\langle X, h_j'\rangle - f^*(p)\,\middle|C_p^*\right)\right| \\
&= \frac{P\left(C \setminus C_p^*\right)}{P(C)}\left|\mathbb{E}\left(\langle X, h_j'\rangle\,\middle|C \setminus C_p^*\right) - \mathbb{E}\left(\langle X, h_j'\rangle\,\middle|C_p^*\right)\right| \\
&< \varepsilon\left\langle\left[\mathbb{E}\left(X\,\middle|C \setminus C_p^*\right) - \mathbb{E}\left(X\,\middle|C_p^*\right)\right], h_j'\right\rangle \\
&< \varepsilon\|h_j'\|_*\left\|\mathbb{E}\left(X\,\middle|C \setminus C_p^*\right) - \mathbb{E}\left(X\,\middle|C_p^*\right)\right\| < 2\varepsilon\|h_j'\|_* \max_{x\in C}\|x\| = K\varepsilon\|h_j'\|_*
\end{aligned}$$

where $\|\cdot\|_*$ is the norm induced by $\langle\cdot,\cdot\rangle$ on the space $\mathfrak{X}^*$. Then, for every $\varepsilon > 0$, we can find a couple $M^* \in \mathbb{R}, p^* \in \mathfrak{X}^*$ such that for the set $C^* = C_{p^*}^*$, $P(C^*) > (1-\varepsilon)P(C)$ and $\|h_j'\|_* = \|M^*h_j + p^*\|_* < c$ for some fixed constant $c > 0$ independent on ε.

It then follows that

$$\left|\mathbb{E}\left(\langle X, h_j'\rangle\,|C\right) - \mathbb{E}\left(\langle X, h_j'\rangle\,|C^*\right)\right| < Kc\varepsilon. \tag{3.12}$$

A similar bound holds for the difference

$$|\mathbb{E}\,(f\,|C\,) - \mathbb{E}\,(f\,|C^*)| < K'\varepsilon,$$

where K' depends on $\|h_j'\|_* < c$ and on the finite value $\max_{x\in C}|f(x)|$.

In light of all this, for every two closed and bounded convex subsets A, B of the form

$$A = \left\{x : \min_{j\in J_A} h_j(x) \geq 0\right\} \quad \text{and} \quad B = \left\{x : \min_{j\in J_B} h_j(x) \geq 0\right\}, \tag{3.13}$$

such that $B \subset A$ and $\mathbb{E}\,(X\,|A) = \mathbb{E}\,(X\,|B)$, for all $\varepsilon > 0$, we can find the two approximating subsets A^* and B^* satisfying

$$P(A^*) > (1-\varepsilon)(A) \quad P(B^*) > (1-\varepsilon)P(B)$$

$$\begin{aligned}
|\mathbb{E}\,(\langle X, h'\rangle\,|A\,) - \mathbb{E}\,(\langle X, h'\rangle\,|A^*)| < K_A\varepsilon \\
|\mathbb{E}\,(\langle X, h'\rangle\,|B\,) - \mathbb{E}\,(\langle X, h'\rangle\,|B^*)| < K_B\varepsilon
\end{aligned} \tag{3.14}$$

for a vector $h' \in \mathfrak{X}^*$ such that $\|h'\|_* < const$, thus implying

$$|\mathbb{E}\,(\langle X, h'\rangle\,|A^*\,) - \mathbb{E}\,(\langle X, h'\rangle\,|B^*)| < \varepsilon(K_A + K_B).$$

Moreover, by

$$\begin{aligned}
&|\mathbb{E}\,(\langle X, h'\rangle\,|A^*\,) - \mathbb{E}\,(\langle X, h'\rangle\,|B^*)| \\
&\qquad = \tfrac{P(A^*\setminus B^*)}{P(A^*)}\,|\mathbb{E}\,(\langle X, h'\rangle\,|A^* \setminus B^*\,) - \mathbb{E}\,(\langle X, h'\rangle\,|B^*)|
\end{aligned}$$

and for $P(A \setminus B) > \varepsilon > 0$, we derive

$$
\begin{aligned}
&|\mathbb{E}\left(\langle X, h'\rangle | A^* \setminus B^*\right) - \mathbb{E}\left(\langle X, h'\rangle | B^*\right)| \\
&\qquad \leq \frac{\varepsilon(K_A + K_B)P(A^*)}{P(A^* \setminus B^*)} < \varepsilon \frac{P(A)(K_A + K_B)}{(1-\varepsilon)P(A) - P(B)}.
\end{aligned} \tag{3.15}
$$

The same reasoning promptly leads to similar bounds for f:

$$
|\mathbb{E}\left(f\,|A\right) - \mathbb{E}\left(f\,|A^*\right)| \leq \varepsilon K'_A, \quad |\mathbb{E}\left(f\,|B\right) - \mathbb{E}\left(f\,|B^*\right)| \leq \varepsilon K'_B \tag{3.16}
$$

where, as well as in (3.14), the constants K'_A and K'_B do not depend on ε. Since we have that

$$
\begin{aligned}
A^* \setminus B^* &= \left\{ x \in \mathfrak{X}^* : \min_{j \in J_A} h'_j(x) \geq f(x) + f^*(p^*) > \min_{j \in J_B} h'_j(x) \right\} \\
&= \left\{ x \in \mathfrak{X}^* : \min_{j \in J_A} h'_j(x) \geq f(x) + f^*(p^*) > h'_l(x) \ \text{ for some } l \in J_B \right\},
\end{aligned}
$$

then we can repeat the arguments used for the sets (3.9). We thus get

$$
\begin{aligned}
&\mathbb{E}\left(f\,|B^*\right) + f^*(p^*) \\
&\quad = \frac{1}{P(B^*)} \int_{X^{-1}(B^*)} f(X) dP + f^*(p^*) \leq \frac{1}{P(B^*)} \int_{X^{-1}(B^*)} \min_{j \in J_B} h'_j(X) dP \\
&\quad \leq \int_{X^{-1}(B^*)} \frac{h'_l(X)}{P(B^*)} dP = \mathbb{E}\left(h'_l(X)\,|B^*\right) \\
&\quad = \mathbb{E}\left(h'_l(X)\,|B^*\right) - \mathbb{E}\left(h'_l(X)\,|A^* \setminus B^*\right) + \mathbb{E}\left(h'_l(X)\,|A^* \setminus B^*\right) \\
&\quad < |\mathbb{E}\left(h'_l(X)\,|B^*\right) - \mathbb{E}\left(h'_l(X)\,|A^* \setminus B^*\right)| + \mathbb{E}\left(f\,|A^* \setminus B^*\right) + f^*(p^*)
\end{aligned}
$$

and, by using (3.15),

$$
\mathbb{E}\left(f\,|B^*\right) < \mathbb{E}\left(f\,|A^* \setminus B^*\right) + \varepsilon \frac{P(A)(K_A + K_B)}{(1-\varepsilon)P(A) - P(B)}.
$$

It then follows that $\mathbb{E}\left(f\,|A^*\right) - \mathbb{E}\left(f\,|B^*\right) \geq -\varepsilon(K_A + K_B)$ and this yelds

$$
\begin{aligned}
&\mathbb{E}\left(f\,|A\right) - \mathbb{E}\left(f\,|B\right) \\
&\quad = \mathbb{E}\left(f\,|A\right) - \mathbb{E}\left(f\,|A^*\right) + \mathbb{E}\left(f\,|B^*\right) - \mathbb{E}\left(f\,|B\right) + \mathbb{E}\left(f\,|A^*\right) - \mathbb{E}\left(f\,|B^*\right) \\
&\quad > -\varepsilon(K'_A + K'_B) - \varepsilon \frac{P(A)(K_A + K_B)}{(1-\varepsilon)P(A) - P(B)}.
\end{aligned}
$$

The arbitrariness of ε completes the result.

Definition 1 *Let $\mathfrak{M}$ be a family of subsets of a set A. We say that $\mathfrak{M}$ is linearly ordered by inclusion if for every two sets $B_1, B_2 \in \mathfrak{M}$, either $B_1 \subseteq B_2$ or $B_2 \subseteq B_1$.*

In other words, if $\mathfrak{M}$ is linearly ordered by inclusion, then it can be represented as

$$\mathfrak{M} = \{B_\alpha, \alpha \in \Gamma\}$$

for Γ a totally ordered net of indices and with $B_{\alpha_1} \subset B_{\alpha_2}$ if $\alpha_1 < \alpha_2$.

Let $\mathcal{C}(\mathfrak{X})$ be the family of all convex closed subsets of $\mathfrak{X}$ and, for every $A \in \mathcal{C}(\mathfrak{X})$, let $\mathfrak{M}(A)$ denote the largest family of convex closed subsets B of A, linearly ordered by inclusion and such that $\mathbb{E}(X\,|A) = \mathbb{E}(X\,|B)$.

Corollary 1 *Let $f : \mathfrak{X} \to \mathbb{R}$ be a lower semicontinuous mapping and $\mathfrak{X}$ a convex closed Hilbert space and let X be an absolutely continuous $\mathfrak{X}-$valued random element. Then f is convex in* $\mathfrak{X}$ if and only if

$$\min_{A\in\mathcal{C}(\mathfrak{X})} \min_{B\in\mathfrak{M}(A)} \mathbb{E}(f\,|A) - \mathbb{E}(f\,|B) \geq 0. \tag{3.17}$$

We close this section by re-writing the above corollary and condition (3.17) in a different way, suitable to the purposes of the next section.

Theorem 5 *Let $f : \mathfrak{X} \to \mathbb{R}$ be a lower semicontinuous mapping and $\mathfrak{X}$ a convex closed Hilbert space and let $X : \Omega \to \mathfrak{X}$ be an absolutely continuous random element on the probability space $(\Omega, \mathcal{A}, P)$. For every $x \in \mathfrak{X}$, let $\mathfrak{M}_x = \{B_\alpha, \alpha \in \Gamma\}$ be the largest family of subsets of $\mathfrak{X}$, linearly ordered by inclusion, satisfying $\mathbb{E}(X\,|B) = x$, for all $B \in \mathfrak{M}_x$. Let also $\Gamma_m = \{\alpha_1, \ldots, \alpha_m\}$ be an increasing sequence of subsets of Γ whose limit is dense in Γ. Then the condition (3.17) is equivalent to.*

$$\lim_{m\to\infty} \min_{x\in\mathfrak{X}} \min_{\alpha_i\in\Gamma_m} \left(\mathbb{E}\left(f\,\middle|B_{\alpha_{i+1}}\right) - \mathbb{E}(f\,|B_{\alpha_i})\right) \geq 0. \tag{3.18}$$

Proof For every $m \geq 1$, let us denote by $\mathfrak{M}_x^{(m)}$ the sub-family of $\mathfrak{M}_x$ defined by the subsets B_{α_i}, for $\alpha_i \in \Gamma_m$. We first show that, for every fixed $m \geq 1$,

$$\min_{x\in\mathfrak{X}} \min_{\alpha_i\in\Gamma_m} \mathbb{E}\left(f \mid B_{\alpha_{i+1}}\right) - \mathbb{E}(f\,|B_{\alpha_i}) \geq \min_{A\in\mathcal{C}(\mathfrak{X})} \min_{B\in\mathfrak{M}(A)} \mathbb{E}(f\,|A) - \mathbb{E}(f\,|B) \tag{3.19}$$

namely, that (3.17) implies (3.18). In fact,

$$\begin{aligned}
&\min_{A\in\mathcal{C}(\mathfrak{X})} \min_{B\in\mathfrak{M}(A)} \mathbb{E}(f\,|A) - \mathbb{E}(f\,|B) \leq \min_{x} \min_{A\in\mathfrak{M}_x^{(m)}} \min_{B_{\alpha_i}\in\mathfrak{M}_x^{(m)}} \mathbb{E}(f\,|A) - \mathbb{E}(f\,|B_{\alpha_i}) \\
&= \min_{x} \min_{\alpha_j\in\Gamma_m} \min_{\alpha_i\in\Gamma_m,\, \alpha_i<\alpha_j} \mathbb{E}\left(f\,\middle|B_{\alpha_j}\right) - \mathbb{E}(f\,|B_{\alpha_i}) \\
&= \min_{x} \min_{\alpha_j\in\Gamma_m} \min_{\alpha_i\in\Gamma_m,\, \alpha_i<\alpha_j} \sum_{h=i}^{j-1} \left(\mathbb{E}\left(f\,\middle|B_{\alpha_{h+1}}\right) - \mathbb{E}(f\,|B_{\alpha_h})\right) \\
&\leq \min_{x} \min_{j\in\Gamma_m} \left(\mathbb{E}\left(f\,\middle|B_{\alpha_j}\right) - \mathbb{E}\left(f\,\middle|B_{\alpha_{j-1}}\right)\right),
\end{aligned}$$

where the last inclusion follows from $\min_{i\leq j} g(i) \leq g(j)$, with

$$g(i) = \textstyle\sum_{h=i}^{j-1} \left(\mathbb{E}(f\,|B_{h+1}) - \mathbb{E}(f\,|B_h)\right).$$

Now the inverse implication. Let $A \in \mathcal{C}(\mathfrak{X})$ be such that $\mathbb{E}(X\,|A) = x$. Then $\mathfrak{M}(A) \subseteq \mathfrak{M}_x$ and $A \in \mathfrak{M}_x$. Let $\alpha_h \in \Gamma_m$ satisfy $B_{\alpha_h} \subseteq A \subset B_{\alpha_{h+1}}$:

$$\min_{\substack{B_{\alpha_i} \in \mathfrak{M}(A):\\ \alpha_i \in \Gamma_m}} \mathbb{E}\,(f\,|A) - \mathbb{E}\,(f\,|B_{\alpha_i})$$

$$= \min_{B_{\alpha_i} \in \mathfrak{M}_x^{(m)}} \left(\sum_{j\geq i}^{h-1} \mathbb{E}\left(f\,\middle|B_{\alpha_{j+1}}\right) - \mathbb{E}\left(f\,\middle|B_{\alpha_j}\right) + \mathbb{E}\,(f\,|A) - \mathbb{E}\,(f\,|B_{\alpha_h}) \right)$$

$$= \min_{B_{\alpha_i} \in \mathfrak{M}_x^{(m)}} \left(\sum_{j\geq i}^{h} \mathbb{E}\left(f\,\middle|B_{\alpha_{j+1}}\right) - \mathbb{E}\left(f\,\middle|B_{\alpha_j}\right) + \mathbb{E}\,(f\,|A) - \mathbb{E}\,(f\,|B_{\alpha_h}) \right)$$

$$\geq \min_{B_{\alpha_i} \in \mathfrak{M}_x^{(m)}} \left[\sum_{j\geq i}^{h} \mathbb{E}\left(f\,\middle|B_{\alpha_{j+1}}\right) - \mathbb{E}\left(f\,\middle|B_{\alpha_j}\right) - \int_{B_{\alpha_h}} |f| dP \left| \tfrac{1}{P(A)} - \tfrac{1}{P(B_{\alpha_h})} \right| - \int_{A\setminus B_{\alpha_h}} |f| \tfrac{dP}{P(A)} \right]$$

$$\geq \min_{B_{\alpha_i} \in \mathfrak{M}_x^{(m)}} \left(\sum_{j\geq i}^{h} \mathbb{E}\left(f\,\middle|B_{\alpha_{j+1}}\right) - \mathbb{E}\left(f\,\middle|B_{\alpha_j}\right) - \min_{x\in A} |f(x)| \tfrac{P\left(A\setminus B_{\alpha_h}\right)}{P(A)} \right)$$

$$\geq \min_{B_{\alpha_i} \in \mathfrak{M}_x^{(m)}} \left(\sum_{j\geq i}^{h} \mathbb{E}\left(f\,\middle|B_{\alpha_{j+1}}\right) - \mathbb{E}\left(f\,\middle|B_{\alpha_j}\right) - \min_{x\in B_{\alpha_h}} |f(x)| \tfrac{P\left(B_{\alpha_{h+1}}\setminus B_{\alpha_h}\right)}{P(B_{\alpha_h})} \right).$$

Then, for every A in $\mathcal{C}(\mathfrak{X})$ and for every $m \geq 1$,

$$\min_{\alpha_i\in\Gamma_h} \mathbb{E}\,(f\,|A) - \mathbb{E}\,(f\,|B_{\alpha_i})$$

$$\geq \min_x \min_{\alpha_h\in\Gamma_m} \sum_{j\geq i}^{h} \left[\mathbb{E}\left(f\,\middle|B_{\alpha_{j+1}}\right) - \mathbb{E}\left(f\,\middle|B_{\alpha_j}\right) \right] - \min_{x\in B_{\alpha_{h+1}}} |f(x)| \frac{P(B_{\alpha_{h+1}}\setminus B_{\alpha_h})}{P(B_{\alpha_h})}.$$

Finally, taking into account that $\mathfrak{M}(A) = \mathfrak{M}_x \cap A = \{B_\alpha \cap A, B_{\alpha\in\mathfrak{M}_x}\}$ and since Γ_m is dense in Γ implies $P(B_{\alpha_{h+1}} \setminus B_{\alpha_h}) \to 0$, then, taking the limit for $m \to \infty$, we obtain

$$\min_A \min_{B\in\mathfrak{M}(A)} \mathbb{E}\,(f\,|A) - \mathbb{E}\,(f\,|B) = \lim_{m\to\infty} \min_A \min_{\substack{B_{\alpha_i} \in \mathfrak{M}_x^{(m)}:\\ B_{\alpha_i} \subset A}} \mathbb{E}\,(f\,|A) - \mathbb{E}\,(f\,|B_{\alpha_i})$$

$$\geq \lim_{m\to\infty} \min_x \min_{\substack{B_{\alpha_i} \in \mathfrak{M}_x^{(m)}:\\ B_{\alpha_i} \subset A}} \mathbb{E}\left(f\,\middle|B_{\alpha_{i+1}}\right) - \mathbb{E}\,(f\,|B_{\alpha_i}).$$

4 Jensen-type inequality for empirical measures

For the whole section, we set $\mathfrak{X} \subset \mathbb{R}$ and we denote by $P = P_X$ the absolutely continuous probability measure induced on $(\mathfrak{X}, \mathcal{B})$ by the mapping $X : \Omega \to \mathfrak{X}$. Let $\hat{P}_n = \hat{P}_{X,n}$ be the empirical measure associated to the random sample $(X_1, \ldots, X_n) \sim P$:

$$\hat{P}_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(X_i).$$

In this section we are going to focus on the convexity criterion (3.18) when $\hat{P}_n$ is used in place of P. The aim is to prove the *consistency* of an appropriate plug-in procedure, in a sense which will be better specified later.

The framework of the rest of this section is the following. We observe a random sample $\{(X_i, W_i), i = 1, \ldots, n\}$, where the random variables $X_1, \ldots, X_n$ are independent and identically distributed with probability law P, while W_i is linked to X_i by an unknown functional relationship, $W_i = f(X_i)$, for all $i \geq 1$. We don't know the mapping $f : \mathfrak{X} \to \mathbb{R}$, but we want to assess whether f is convex or not in a bounded interval. Although the convexity criterion is true for functionals of an arbitrary Hilbert space, we are limiting ourselves to the case where f is a real function and $\mathfrak{X} \subset \mathbb{R}$ is a bounded interval. Most of the results below can be adjusted, with some effort, to apply to a higher-dimensional domain.

The idea is to define a sample version of the convexity criterion (3.18), approximating the probability P, which is unknown, by the empirical measure.

On one hand, since, for every fixed sample $(X_1, \ldots, X_n)$, $\hat{P}_n$ is a discrete distribution, then as we pointed out in Section 2, the whole set of refined Jensen inequalities in (3.18) (as well as (3.17)) is not *sufficient* to convexity. Indeed this failure is due to the fact that there might not be enough closed convex subsets A and B such that $\hat{\mathbb{E}}_n(X\,|A) = \hat{\mathbb{E}}_n\,(X\,|B\,)$. In fact, the families $\mathfrak{M}_x$ will be empty for most of the $x \in \mathfrak{X}$ when the expectations are done with respect to $\hat{P}_n$.

On the other hand, as soon as n tends to infinity, $\hat{P}_n$ approaches the *true* measure. Therefore, we should be able to replace the too restrictive conditions $\hat{\mathbb{E}}_n(X\,|A) = \hat{\mathbb{E}}_n\,(X\,|B\,)$ with some milder versions, guaranteeing

$$\lim_{n\to\infty} \left\| \hat{\mathbb{E}}_n(X\,|A) - \hat{\mathbb{E}}_n\,(X\,|B) \right\| = 0.$$

We first prove the following lemma.

Lemma 1 *Let $\mathfrak{X} \subseteq \mathbb{R}$. For any fixed $n \geq 1$, for every $x \in \mathfrak{X}$ and every convex closed A such that $\mathbb{E}(X\,|A) = x$, there exists a convex closed A_n with $\hat{\mathbb{E}}_n(X\,|A_n\,) = x + O_P\left(n^{-1/2}\right)$ $P-$a.s..*

Proof. Let $A = [\mathsf{A}_1, \mathsf{A}_2]$ and let $A_n = [X_{(l)}, X_{(k)}]$, where $X_{(j)}$ is the $j-$th order

statistic and with

$$X_{(l)} \leq \mathsf{A}_1 < X_{(l+1)} < \ldots < X_{(k-1)} < \mathsf{A}_2 \leq X_{(k)}.$$

We need to show that the difference between $\hat{\mathbb{E}}_n(X\,|X \in A_n\,)$ and $\mathbb{E}(X\,|X \in A\,)$ is in absolute value almost surely an $O_P(n^{-1/2})$. The proof can be completed by applying Theorem 1 p. 678 – and subsequently Theorem 2 p. 94 – in Shorack and Wellner (1986) to the randomly trimmed mean

$$\hat{\mathbb{E}}_n\left(X\,|X \in A_n\,\right) = \frac{1}{k-l+1}\sum_{i=l}^{k} X_{(i)}.$$

Before continuing the analysis, we need to introduce some notation. Let $\{\mathcal{U}_n\} = \{U_1, \ldots, U_{m_n}\}$ form a partition for $\mathfrak{X}$, such that:

$$c \cdot m_n^{-1} \leq \min_j \|U_j\| \leq \max_j \|U_j\| \leq C \cdot m_n^{-1}, \tag{4.20}$$

for some constants $C > c > 0$, with all U_j convex and closed.

For every $\varepsilon > 0$ and for every $U_j \in \{\mathcal{U}_n\}$, $1 \leq j \leq m_n$, let $\hat{\mathfrak{M}}_\varepsilon(j) = \{B_1^{(j)}, B_2^{(j)}, \ldots, B_{\hat{k}_\varepsilon(j)}^{(j)}\}$ be the largest family of convex closed subsets of $\mathfrak{X}$, linearly ordered by inclusion and satisfying:

$$\hat{P}_n(B_1^{(j)}) > \varepsilon, \qquad \hat{P}_n(B_i^{(j)}) < \hat{P}_n(B_{i+1}^{(j)}) - \varepsilon \tag{4.21}$$

and

$$\hat{\mathbb{E}}_n\left(X\,\middle|B_i^{(j)}\right) \in U_j \quad \text{for all } i \leq \hat{k}_\varepsilon(j),\ 1 \leq j \leq m_n. \tag{4.22}$$

Then we have the following result.

Theorem 6 *Let $f : \mathfrak{X} \to \mathbb{R}$ be lower-semicontinuous and define the events $F_n(\alpha, \varepsilon)$, for $\varepsilon > 0$, $\alpha \neq 0$:*

$$F_n(\alpha,\varepsilon) := \left\{\omega : \min_{1\leq j\leq m_n} \min_{1\leq i\leq \hat{k}_\varepsilon(j)} \hat{\mathbb{E}}_n\left(f\,\middle|B_{i+1}^{(j)}\right) - \hat{\mathbb{E}}_n\left(f\,\middle|B_i^{(j)}\right) \geq \alpha\right\}. \tag{4.23}$$

(i) If f is strictly concave on a convex subset $I_0 \subseteq \mathfrak{X}$ with $1 \geq \Pr(I_0) = p_0 > 0$, then for every $\varepsilon, \alpha > 0$, and for every partition $\{\mathcal{U}_n\}$ such that (4.20) holds and such that

$$\lim_n m_n n^{-1/2} > 0. \tag{4.24}$$

we have that

$$\Pr\left\{F_n(\alpha,\varepsilon)\right\} \leq \frac{1}{\varepsilon}\,\mathrm{e}^{-O_P(n)} \tag{4.25}$$

(ii) If f is convex everywhere in $\mathfrak{X}$, then, for every $\varepsilon, \alpha > 0$, and for every partition $\{\mathcal{U}_n\}$ such that (4.20) and (4.24) hold, then

$$\Pr\left\{F_n(-\alpha,\varepsilon)^c\right\} \leq \frac{m_n}{\varepsilon}\,\mathrm{e}^{-O_P(n)}\,. \tag{4.26}$$

Proof. *(i).* Without loss of generality, we can assume that $I_0 = \mathfrak{X}$. With this setting, f is strictly concave in every set $B_i^{(j)}$ appearing in the minima of (4.23).

Let also j^* be the index corresponding to the set U_{j^*} satisfying $\mathbb{E}\,(X\,|I_0\,) \in U_{j^*}$.

Then a trivial bound of the left hand side of (4.25) is given by

$$\Pr\left\{F_n(\alpha,\varepsilon)\right\} \leq \Pr\left\{\omega : \min_{1\leq i\leq \hat{k}_\varepsilon^*} \hat{\mathbb{E}}_n\left(f\,\middle|B_{i+1}^{(j^*)}\right) - \hat{\mathbb{E}}_n\left(f\,\middle|B_i^{(j^*)}\right) \geq \alpha\right\}$$

where $\hat{k}_\varepsilon^* = \hat{k}_\varepsilon(j^*)$. Henceforth, we omit the superscript (j^*) in the sets $B_i^{(j^*)}$.

Let us now focus on the difference $\hat{\mathbb{E}}_n\,(f\,|B_{i+1}\,) - \hat{\mathbb{E}}_n\,(f\,|B_i\,)$. Since $\hat{\mathbb{E}}_n\,(f\,|A\,) = \frac{1}{\hat{P}_n(A)}\hat{\mathbb{E}}_n(f\cdot 1_A)$, simple manipulations yield:

$$\begin{aligned}
\hat{\mathbb{E}}_n\,(f\,|A\,) - \mathbb{E}\,(f\,|A\,) &= \left(\hat{\mathbb{E}}_n - \mathbb{E}\right)\frac{f1_A}{P(A)} + \hat{\mathbb{E}}_n\frac{f1_A}{P(A)}\left(\frac{P(A)}{\hat{P}_n(A)} - 1\right) \\
&= \left(\hat{\mathbb{E}}_n - \mathbb{E}\right)\tfrac{f1_A}{P(A)} + \left(\tfrac{P(A)}{\hat{P}_n(A)} - 1\right)\mathbb{E}\tfrac{f1_A}{P(A)} + \left(\tfrac{P(A)}{\hat{P}_n(A)} - 1\right)\left(\hat{\mathbb{E}}_n - \mathbb{E}\right)\tfrac{f1_A}{P(A)} \\
&= \left(\hat{\mathbb{E}}_n - \mathbb{E}\right)\left(\tfrac{f1_A}{P(A)} - \mathbb{E}\,(f\,|A\,)\,\tfrac{1_A}{P(A)}\right) \\
&\qquad + \left(\tfrac{P(A)}{\hat{P}_n(A)} - 1\right)\left(\hat{\mathbb{E}}_n - \mathbb{E}\right)\left(\tfrac{f1_A}{P(A)} - \mathbb{E}\,(f\,|A\,)\,\tfrac{1_A}{P(A)}\right) \\
&= (\hat{\mathbb{E}}_n - \mathbb{E})g_A - \left[\frac{P(A)}{\hat{P}_n(A)} - 1\right](\hat{\mathbb{E}}_n - \mathbb{E})g_A, \qquad (4.27)
\end{aligned}$$

where $g_A := \left(\frac{f1_A}{P(A)} - \mathbb{E}\,(f\,|A\,)\,\frac{1_A}{P(A)}\right)$.

Therefore, the left hand side of (4.25) can be bounded by

$$
\begin{aligned}
\Pr\left\{F_n(\alpha,\varepsilon)\right\} &\leq \Pr\left\{\omega : \min_{1\leq i\leq \hat{k}^*_\varepsilon} \hat{\mathbb{E}}_n\left(f\,|B_{i+1}\right) - \hat{\mathbb{E}}_n\left(f\,|B_i\right) \geq \alpha\right\} \\
&\leq \Pr\left\{\min\nolimits_{i\leq \hat{k}^*_\varepsilon}\left[(\hat{\mathbb{E}}_n - \mathbb{E})\bar{g}(i)\left[1+\max\nolimits_i\left(\tfrac{P(B_i)}{\hat{P}_n(B_i)}-1\right)\right] + \mathbb{E}\left(f\,|B_{i+1}\right) - \mathbb{E}\left(f\,|B_i\right)\right] \geq \alpha\right\} \\
&\leq \min\nolimits_i \Pr\left\{\left[1+\max\nolimits_i\left(\tfrac{P(B_i)}{\hat{P}_n(B_i)}-1\right)\right](\hat{\mathbb{E}}_n - \mathbb{E})\bar{g}(i) \geq \alpha - \max\nolimits_i(\mathbb{E}\left(f\,|B_{i+1}\right) - \mathbb{E}\left(f\,|B_i\right))\right\} \\
&= \min\nolimits_i \Pr\left\{(1+\alpha)\left(\hat{\mathbb{E}}_n - \mathbb{E}\right)\bar{g}(i) \geq \alpha - \max\nolimits_i(\mathbb{E}\left(f\,|B_{i+1}\right) - \mathbb{E}\left(f\,|B_i\right))\right\} \\
&\qquad \times \Pr\left\{\max\nolimits_i\left(\tfrac{P(B_i)}{\hat{P}_n(B_i)}-1\right) \leq \alpha\right\} + \Pr\left\{\max\nolimits_i\left(\tfrac{P(B_i)}{\hat{P}_n(B_i)}-1\right) > \alpha\right\} \\
&\leq \min\nolimits_i \Pr\left\{(\hat{\mathbb{E}}_n - \mathbb{E})\bar{g}(i) \geq \tfrac{\alpha + \min_i(\mathbb{E}(f|B_i) - \mathbb{E}(f|B_{i+1}))}{1+\alpha}\right\} \\
&\qquad + \Pr\left\{\max\nolimits_i\left(\tfrac{P(B_i)}{\hat{P}_n(B_i)}-1\right) > \alpha\right\} \qquad (4.28)
\end{aligned}
$$

where $\bar{g}(i) := g_{B_{i+1}} - g_{B_i}$.

We consider the two terms of (4.28) separately.

In order to bound the term $\Pr\left\{\max_i\left(\frac{P(B_i)}{\hat{P}_n(B_i)}-1\right) > \alpha\right\}$, we remark that, for every $i \leq \hat{k}^*_\varepsilon$, the quantity $P\left(B_{i+1}\right)/\hat{P}_n\left(B_{i+1}\right)$ writes

$$
\frac{P\left(B_{i+1}\right)}{\hat{P}_n\left(B_{i+1}\right)} - 1 = \sum_{h=1}^{i+1} \alpha_h^{(i,n)}\left[\frac{P\left(B_h \setminus B_{h-1}\right)}{\hat{P}_n\left(B_h \setminus B_{h-1}\right)} - 1\right]
$$

where $\alpha_h^{(i,n)} = \frac{\hat{P}_n(B_h \setminus B_{h-1})}{\hat{P}_n(B_{i+1})}$ satisfies $\sum_{h=1}^{i+1} \alpha_h^{(i,n)} = 1$.

Thus

$$
\max_{i\leq \hat{k}^*_\varepsilon}\left[\frac{P\left(B_{i+1}\right)}{\hat{P}_n\left(B_{i+1}\right)} - 1\right] \leq \max_{h\leq \hat{k}^*_\varepsilon}\left[\frac{P\left(B_h \setminus B_{h-1}\right)}{\hat{P}_n\left(B_h \setminus B_{h-1}\right)} - 1\right]. \qquad (4.29)
$$

We can apply inequality (10.3.2) in Shorack and Wellner (1986) to the ratios in (4.29): since $n\hat{P}_n\left(B_h \setminus B_{h-1}\right) \sim Binomial\left(n, P\left(B_h \setminus B_{h-1}\right)\right)$, conditionally on $\hat{\mathfrak{M}}_\varepsilon := \hat{\mathfrak{M}}_\varepsilon(j^*)$, then

$$
\begin{aligned}
\Pr\left\{\max_i\left(\frac{P(B_i)}{\hat{P}_n(B_i)} - 1\right) \geq \alpha \,\middle|\, \hat{\mathfrak{M}}_\varepsilon\right\} &\leq \Pr\left\{\max_i\left(\frac{P\left(B_i \setminus B_{i-1}\right)}{\hat{P}_n\left(B_i \setminus B_{i-1}\right)} - 1\right) \geq \alpha \,\middle|\, \hat{\mathfrak{M}}_\varepsilon\right\} \\
&\leq \sum_{i\leq \hat{k}^*_\varepsilon} \Pr\left\{\frac{P\left(B_i \setminus B_{i-1}\right)}{\hat{P}_n\left(B_i \setminus B_{i-1}\right)} \geq \alpha + 1 \,\middle|\, \hat{\mathfrak{M}}_\varepsilon\right\} \\
&\leq \hat{k}^*_\varepsilon \max_{i\leq \hat{k}^*_\varepsilon} \exp\left\{-nP\left(B_i \setminus B_{i-1}\right) h\left(1/\left(1+\alpha\right)\right)\right\} \leq \hat{k}^*_\varepsilon e^{-n\varepsilon h(1/(1+\alpha))} \qquad (4.30)
\end{aligned}
$$

where $h(t) = t\log t - t + 1$. In light of $h(1+\delta) \geq \frac{\delta^2}{2}(1-\delta)$, we have

$$h\left(1/\left(1+\alpha\right)\right) = h\left(1 + \frac{-\alpha}{1+\alpha}\right) \geq \frac{\alpha^2}{2(1+\alpha)^2}\left(1 + \frac{\alpha}{1+\alpha}\right) \geq \frac{\alpha^2}{2(1+\alpha)^2}$$

and we get that,

$$\Pr\left\{\max_i \left(\frac{P\left(B_i\right)}{\hat{P}_n\left(B_i\right)} - 1\right) \geq \alpha \left| \hat{\mathfrak{M}}_\varepsilon \right.\right\} \leq \hat{k}_\varepsilon^* \exp\left\{-\frac{n\varepsilon\alpha^2}{2(1+\alpha)^2}\right\}.$$

Since bounds (4.21) imply that $\hat{k}_\varepsilon \leq \varepsilon^{-1}$, independently on $\hat{\mathfrak{M}}_\varepsilon$, we derive

$$\Pr\left\{\max_i \left(\frac{P\left(B_i\right)}{\hat{P}_n\left(B_i\right)} - 1\right) \geq \alpha\right\} \leq \varepsilon^{-1} \exp\left\{-\frac{n\varepsilon\alpha^2}{2(1+\alpha)^2}\right\}. \tag{4.31}$$

For the first term in (4.28), we need to consider the differences $\mathbb{E}\left(X\left|B_{i+1}\right.\right) - \mathbb{E}\left(X\left|B_i\right.\right)$:

$$\begin{aligned}
&\max_i \left\|\mathbb{E}\left(X\left|B_{i+1}\right.\right) - \mathbb{E}\left(X\left|B_i\right.\right)\right\| \\
&\quad \leq 2\max_i \left\|\hat{\mathbb{E}}_n\left(X\left|B_{i+1}\right.\right) - \mathbb{E}\left(X\left|B_{i+1}\right.\right)\right\| + \left\|\hat{\mathbb{E}}_n\left(X\left|B_{i+1}\right.\right) - \hat{\mathbb{E}}_n\left(X\left|B_{i+1}\right.\right)\right\| \\
&\quad \leq O_P\left(n^{-1/2}\right) + O_P\left(m_n^{-1}\right) \leq O_P(n^{-1/2})
\end{aligned} \tag{4.32}$$

where the term $O_P(n^{-1/2})$ comes from the same arguments leading to (4.27), with $f(x)$ replaced by the identity function. In fact the function $h_B(x) = \frac{x1_B(x)}{P(B)} - \mathbb{E}(X\left|B\right.)\frac{1_B(x)}{P(B)}$ is bounded above for every subset $B \in \hat{\mathfrak{M}}(j^*)$. Moreover, the class $\hat{\mathfrak{M}}(j^*)$ converges to the class $\mathfrak{M}(\overline{x}_0)$, with $\overline{x}_0 = \mathbb{E}\left(X\left|I_0\right.\right)$, that is a VC-class with VC-order equal to 1 (see Dudley (1984)). The term $O_P\left(m_n^{-1}\right)$ easily follows from the definition of $\hat{\mathfrak{M}}(j^*)$: $\hat{\mathbb{E}}_n\left(X\left|B_i\right.\right) \in U_{j^*}$ for all $i \leq \hat{k}_\varepsilon^*$ and thus

$$\left\|\hat{\mathbb{E}}_n\left(X\left|B_{i+1}\right.\right) - \hat{\mathbb{E}}_n\left(X\left|B_i\right.\right)\right\| \leq \sup_{(x_1,x_2)\in U_{j^*}} |x_1 - x_2| = \|U_{j^*}\|_{L_1} \leq O_P(m_n^{-1})$$

by (4.20). The inequality
$\left\|\mathbb{E}\left(X\left|B_{i+1}\setminus B_i\right.\right) - \mathbb{E}\left(X\left|B_i\right.\right)\right\| = \frac{P(B_{i+1})}{P(B_{i+1}\setminus B_i)}\left\|\mathbb{E}\left(X\left|B_{i+1}\right.\right) - \mathbb{E}\left(X\left|B_i\right.\right)\right\| \leq \frac{O_P(n^{-1/2})}{\varepsilon}$
yields that

$$\mathbb{E}\left(X\left|B_i\right.\right) - \varepsilon^{-1}O_P(n^{-1/2}) \leq \mathbb{E}\left(X\left|B_{i+1}\setminus B_i\right.\right) \leq \mathbb{E}\left(X\left|B_i\right.\right) + \varepsilon^{-1}O_P(n^{-1/2}).$$

Having in mind that $-f$ is stricly convex we can rewrite the sets B_i and $B_{i+1}\setminus B_i$ as

$$B_i = \{x : -f(x) \leq a_i x + b_i\},\ B_{i+1}\setminus B_i = \{x : a_i x + b_i < -f(x) \leq a_{i+1}x + b_{i+1}\}$$

where $a_i, a_{i+1}, b_i, b_{i+1}$ are bounded constants, as in the Proof of Theorem 1. Then, repeating the same arguments as before, we obtain that

$$\begin{aligned}\frac{1}{P\left(B_{i+1} \setminus B_i\right)} \int_{B_{i+1} \setminus B_i} (-f(x)) dP(x) &\geq \frac{1}{P\left(B_{i+1} \setminus B_i\right)} \int_{B_{i+1} \setminus B_i} (a_i x + b_i) dP(x) \\ &= a_i \mathbb{E}\left(X \,|B_{i+1} \setminus B_i\right) + b_i. \end{aligned} \tag{4.33}$$

If $a_i \geq 0$, in light of the above considerations, the right hand side of (4.33) is

$$a_i \mathbb{E}\left(X \,|B_{i+1} \setminus B_i\right) + b_i \geq a_i \mathbb{E}\left(X \,|B_i\right) + b_i - a_i O_P\left(n^{-1/2} \varepsilon^{-1}\right).$$

If instead $a_i \leq 0$, then $a_i \mathbb{E}\left(X \,|B_{i+1} \setminus B_i\right) + b_i \geq a_i \mathbb{E}\left(X \,|B_i\right) + b_i + a_i O_P\left(n^{-1/2} \varepsilon^{-1}\right)$. In both cases,

$$\mathbb{E}\left(-f \,|B_{i+1} \setminus B_i\right) \geq a_i \mathbb{E}\left(X \,|B_{i+1} \setminus B_i\right) + b_i \geq \mathbb{E}\left(-f \,|B_i\right) - O_P\left(n^{-1/2} \varepsilon^{-1}\right). \tag{4.34}$$

Using the inequality (4.34), we can finally work on the first term in (4.28):

$$\begin{aligned}&\min_i \Pr\left\{ (\hat{\mathbb{E}}_n - \mathbb{E}) \bar{g}(i) \geq \frac{\alpha + \min_i \left(\mathbb{E}\left(-f \,|B_{i+1}\right) - \mathbb{E}\left(-f \,|B_i\right)\right)}{1+\alpha} \right\} \\ &\quad \leq \min_i \Pr\left\{ n(\hat{\mathbb{E}}_n - \mathbb{E}) \bar{g}(i) \geq \frac{n\alpha - O_P\left(n^{1/2}\right)}{1+\alpha} \right\} \end{aligned} \tag{4.35}$$

Now consider that $\mathbb{E}\bar{g}(i) = 0$ for every i and that

$$\begin{aligned}\mathbb{E}\bar{g}(i)^2 &= \mathbb{E} g_{B_{i+1}}^2 + \mathbb{E} g_{B_i}^2 = \frac{Var\left(f \,|B_{i+1}\right)}{P(B_{i+1})} + \frac{Var\left(f \,|B_i\right)}{P(B_i)} \\ &\leq \frac{4}{P(B_i)} \max_i \mathbb{E}\left(f^2 \,|B_i\right) \leq \frac{4M^2}{\varepsilon},\end{aligned}$$

where $M = \sup_{x \in I_0} |f(x)| < \infty$ because f is bounded on bounded sets. We can finally apply Bernstein's inequality to (4.35), which yields

$$\min_i \Pr\left\{ n(\hat{\mathbb{E}}_n - \mathbb{E})\bar{g}(i) \geq \frac{n\alpha - O_P\left(n^{1/2}\right)}{1+\alpha} \right\}$$
$$\leq \min_i \exp\left\{ \frac{-n^2\alpha^2 + O_P\left(n^{3/2}\right)}{(1+\alpha)^2\left(n\mathbb{E}\bar{g}(i)^2 + n\alpha \sup_x \bar{g}(i)(x)/3(1+\alpha)\right)} \right\}$$
$$\leq \exp\left\{ \frac{-3\varepsilon\alpha^2 n}{4(1+\alpha)[(1+\alpha)M^2 + \alpha M]} + O_P(n^{1/2}) \right\}. \tag{4.36}$$

By unifying (4.31) and (4.36) we get

$$\Pr\{F_n(\alpha,\varepsilon)\} \leq \exp\left\{ \frac{-3\varepsilon\alpha^2 n + O_P(n^{1/2})}{4(1+\alpha)[(1+\alpha)M^2 + \alpha M]} \right\} + \hat{k}_\varepsilon^* \exp\left\{ -\frac{n\varepsilon\alpha^2}{2(1+\alpha)^2} \right\}$$
$$\leq (\hat{k}_\varepsilon^* + 1) \exp\left\{ \frac{-n\varepsilon\alpha^2}{4(1+\alpha)[(1+\alpha)M^2 + \alpha M]} + O(n^{1/2}) \right\}.$$

(ii) By using arguments similar to those of part (i), we can immediately write that

$$\Pr\{F_n(-\alpha,\varepsilon)^c\} = \Pr\left\{ \max_j \max_i \hat{\mathbb{E}}_n\left(f\left|B_i^{(j)}\right.\right) - \hat{\mathbb{E}}_n\left(f\left|B_{i+1}^{(j)}\right.\right) \geq \alpha \right\}$$
$$\leq \sum_{j \leq m_n} \Pr\left\{ \max_i \hat{\mathbb{E}}_n\left(f\left|B_i^{(j)}\right.\right) - \hat{\mathbb{E}}_n\left(f\left|B_{i+1}^{(j)}\right.\right) \right\}$$
$$\leq \sum_{j \leq m_n} \Pr\left\{ \max_i \left(\hat{\mathbb{E}}_n - \mathbb{E}\right)\bar{g}(i,j) \geq \frac{\alpha + \min_i[\mathbb{E}\left(f\left|B_{i+1}^{(j)}\right.\right) - \mathbb{E}\left(f\left|B_i^{(j)}\right.\right)]}{1+\alpha} \right\}$$
$$+ \sum_{j \leq m_n} \Pr\left\{ \max_i \left(\frac{P(B_i^{(j)})}{\hat{P}_n(B_i^{(j)})} - 1 \right) > \alpha \right\} \tag{4.37}$$

where $\bar{g}(i,j) := g_{B_{i+1}^{(j)}} - g_{B_i^{(j)}}$.

By (4.31), for all $1 \leq j \leq m_n$, we have that

$$\Pr\left\{ \max_i \left(\frac{P(B_i^{(j)})}{\hat{P}_n(B_i^{(j)})} - 1 \right) > \alpha \right\} \leq \varepsilon^{-1} \exp\left\{ -\frac{n\varepsilon\alpha^2}{2(1+\alpha)^2} \right\}.$$

Repeating the same arguments used in (4.32), (4.33) and (4.34), we get, for every $j \leq m_n$ and by Bernstein's inequality,

$$\Pr\left\{\max_i \left(\hat{\mathbb{E}}_n - \mathbb{E}\right)\bar{g}(i,j) \geq \frac{\alpha + \min_i[\mathbb{E}\left(f\left|B_{i+1}^{(j)}\right.\right) - \mathbb{E}\left(f\left|B_i^{(j)}\right.\right)]}{1+\alpha}\right\}$$

$$\leq \Pr\left\{\max_i n(\hat{\mathbb{E}}_n - \mathbb{E})\bar{g}(i,j) \geq \frac{n\alpha - O_P\left(n^{1/2}\right)}{1+\alpha}\right\}$$

$$\leq \varepsilon^{-1}\max_i \Pr\left\{n(\hat{\mathbb{E}}_n - \mathbb{E})\bar{g}(i,j) \geq \frac{n\alpha - O_P\left(n^{1/2}\right)}{1+\alpha}\right\}$$

$$\leq \varepsilon^{-1}\exp\left\{-\frac{3n\varepsilon\alpha^2 + O_P(n^{1/2})}{4(1+\alpha)[(1+\alpha)M^2 + \alpha M]}\right\} \leq \varepsilon^{-1}e^{-\frac{n\varepsilon\alpha^2}{4M^2(1+\alpha)}}.$$

Finally, by unifying the bounds for the two terms in (4.37), we get

$$\Pr\left\{F_n(-\alpha,\varepsilon)^c\right\} \leq m_n\varepsilon^{-1}e^{-O_P(n)}.$$

5 Statistical applications.

5.1 Convexity of regression models.

The results of the previous section and the characterization (3.18) can suggest new statistics for testing the convexity of functions. In fact, Theorem 6 permits to control for the asymptotic I and II type errors based on the inequalities in (4.23).

However, we must underline that the upper bounds in Theorem 6 hold in the ideal situation when the observed variable is $W = f(X)$, that is, when the functional relationship is exact. In most of concrete situations, the former identity occurs with an error term or a noise. This error term, clearly, must be taken into account when looking for exponential bounds on the probabilities either of $F_n(\alpha,\varepsilon)$ or $F_n(-\alpha,\varepsilon)^c$.

On the other hand, there is a central case when the noise can be ignored, since it "integrates out", namely when $f(X) = \mu(Y\,|X\,)$ is the regression function. If in fact we are interested in testing convexity of the function $f(X) = \mathbb{E}(Y\,|X\,)$, by using the above characterization, X, Y being two a.c. random variables, in principle, we should check that

$$\mathbb{E}\,(W\,|A\,) - \mathbb{E}\,(W\,|B\,) \geq 0$$

for all subsets $B \subset A$ such that $\mathbb{E}\,(X\,|A\,) = \mathbb{E}\,(X\,|B\,)$. The key point is that, by the definition of conditional expectations, $\int_A W dP = \int_A Y dP$ for all $A \in \mathcal{B}(\mathfrak{X})$. This allows us replace Y to W, in the (3.18) or (3.17), because of

$$\mathbb{E}\,(Y\,|A\,) - \mathbb{E}\,(Y\,|B\,) = \mathbb{E}\,(W\,|A\,) - \mathbb{E}\,(W\,|B\,).$$

In general, assume that we observe a sample $\{(X_i, Y_i), i \geq 1\}$ with $Y_i = f(X_i) + U_i$, where U is a random variable that can be interpreted loosely speaking as an error term and that we want to test for convexity of the function $f(X) = W$. Here we are not necessarily assuming $f(X) = \mu(X) = \mathbb{E}(Y\,|X\,)$. We can consider the following situations:

1. $\mathbb{E}\,(U\,|X\,) = 0$. It corresponds to the case $f(X) = \mathbb{E}\,(Y\,|X\,)$ described above. As we have already pointed out, this position ensures the identity

$$\mathbb{E}\,(Y\,|X \in B\,) = \mathbb{E}\,(f(X)\,|X \in B\,) + \mathbb{E}\,(U\,|X \in B\,) = \mathbb{E}\,(f\,|X \in B\,)$$

for all $B \in \mathcal{B}(\mathfrak{X})$.

2. $\mathbb{E}\,(U\,|X\,) \neq 0$, but $U = \gamma X + \nu$ for some $\gamma \in \mathfrak{X}^*$ and for ν such that $\mathbb{E}\,(\nu\,|X\,) = 0$. Then, for every $A \in \mathcal{C}(\mathfrak{X})$ and $B \in \mathfrak{M}(A)$, one has

$$\begin{aligned}\mathbb{E}\,(Y\,|A) - \mathbb{E}\,(Y\,|B\,) &= \mathbb{E}\,(f\,|A) - \mathbb{E}\,(f\,|B\,) + \gamma\,[\mathbb{E}\,(X\,|A) - \mathbb{E}\,(X\,|B)] \\ &= \mathbb{E}\,(f\,|A) - \mathbb{E}\,(f\,|B\,)\end{aligned}$$

3. $U = g(X) + \nu$ where g is convex in x and $\mathbb{E}(\nu\,|X\,) = 0$. In this case,

$$\begin{aligned}\mathbb{E}\,(Y\,|A) - \mathbb{E}\,(Y\,|B\,) &= \mathbb{E}\,(f\,|A) - \mathbb{E}\,(f\,|B\,) + \mathbb{E}\,(U\,|A) - \mathbb{E}\,(U\,|B\,) \\ &\geq \mathbb{E}\,(f\,|A) - \mathbb{E}\,(f\,|B\,)\end{aligned}$$

4. $U = g(X) + \nu$ where g is concave in x and $\mathbb{E}(\nu\,|X\,) = 0$. Then,

$$\mathbb{E}\,(Y\,|A) - \mathbb{E}\,(Y\,|B\,) \leq \mathbb{E}\,(f\,|A) - \mathbb{E}\,(f\,|B\,)\,.$$

The following Corollary of Theorem 6 follows straightforwardly.

Corollary 2 *Let $Y_i = f(X_i) + U_i$ and let*

$$G_n(\alpha, \varepsilon) = \left\{\omega : \min_j \min_i \sum_h \frac{Y_h 1\{X_h \in B_{i+1}^{(j)}\}}{\hat{P}_n\left(B_{i+1}^{(j)}\right)} - \sum_h \frac{Y_h 1\{X_h \in B_i^{(j)}\}}{\hat{P}_n\left(B_i^{(j)}\right)}\right\},$$

where the families of subsets $B_i^{(j)}$ satisfy the conditions of Theorem 6.

(i) Let U satisfy cases [1-2]. If f is convex almost everywhere, then $\Pr\{G_n(\alpha, \varepsilon)^c\} \leq m_n \varepsilon^{-1} e^{-O_P(n)}$. If f is concave in $I_0 \subseteq \mathfrak{X}$, with $P(I_0) \geq p_0 > 0$, then $\Pr\{G_n(\alpha, \varepsilon)\} \leq \varepsilon^{-1} e^{-O_P(n)}$.

(ii) Let U satisfy case [3]. Then, if f is convex almost everywhere, then $\Pr\{G_n(\alpha, \varepsilon)^c\} \leq m_n \varepsilon^{-1} e^{-O_P(n)}$.

(iii) Let U satisfy case [4]. If f is concave in $I_0 \subseteq \mathfrak{X}$, with $P(I_0) \geq p_0 > 0$, then $\Pr\{G_n(\alpha, \varepsilon)\} \leq \varepsilon^{-1} e^{-O_P(n)}$.

Example 1 (Conditional quantile function.) *Let $f(X) = Q(\alpha\,|X\,)$ be the conditional $\alpha-$quantile of Y. Then the model $Y_i = f(X_i) + U_i$ is known as a regression quantile model. The condition $\mathbb{E}\,(U\,|X\,) = 0$ is not satisfied in general, but still in some cases it is possible to apply Corollary 2.*
(i) Exponential distribution *If the conditional distribution of Y for given*

$X = x$ is an exponential distribution with parameter $\lambda(x)$, then
$\mathbb{E}(U\,|X\,) = \mathbb{E}\,(Y\,|X\,) - Q(\alpha\,|X\,) = (1 + \log(1-\alpha))/\lambda(x)$
is proportional to $\mathbb{E}(Y\,|X\,)$. Therefore, if $1/\lambda(x)$ is convex (resp. concave) and $\alpha < 1 - e^{-1}$, then both $Q(\alpha\,|X\,)$ and $\mathbb{E}(U\,|X\,) = (1 - \log(1-\alpha))/\lambda(x)$ are also convex (resp. concave).
(ii) Pareto distribution *Let the distribution of Y conditional to X be Pareto with parameters ν, m, $\nu > 1$, $m > 0$ (where m is the lower bound of the support of the distribution and ν is the shape parameter). If ν is invariant with X, while $m = m(x)$, then $\mathbb{E}\,(Y\,|X\,) = \frac{\nu m(X)}{\nu - 1}$ and for every $\alpha \in (0,1)$, $Q(\alpha\,|X\,) = \frac{m(X)}{(1-\alpha)^{1/\nu}}$, so that $\mathbb{E}(U\,|X\,)$ is proportional to $m(X)$.*
(iii) Gaussian distribution. *Let $f(y\,|x)$ be a Gaussian density with mean $\mu(x)$ and constant variance. Then $Q(\alpha\,|X\,) = \mu(X) + \sigma\zeta_\alpha$, ζ_α being the $\alpha-$quantile of a $N(0,1)$ random variable. Thus, under the homoskedasticity assumption, $\mathbb{E}(U\,|X\,) = \sigma\zeta_\alpha$ and this example enters in the case [2] above.*
(iv) Gumbel distribution. *If Y has a Gumbel distribution conditional to X, with parameters $\mu = \mu(X)$ and β (not depending on X), then $\mathbb{E}\,(Y\,|X\,) = \mu(X) + \beta\gamma$, where $\gamma \approx 0.57721$ is the Euler-Mascheroni constant. Moreover, $Q\,(\alpha\,|X\,) = \mu(X) - \beta\ln(\ln(1/\alpha))$. Thus, if the scale parameter β doesn't depend on X, again the case [2] is satisfied. If instead β is some convex or concave function of X, we are in cases [3] or [4] depending on the sign of $\gamma - \ln(\ln(1/\alpha))$.*
(iv) Weibull distribution. *Let $f(y\,|x) = \frac{k}{\lambda}\left(\frac{y}{\lambda}\right)^{k-1} e^{-(y/\lambda)^k}$, with $\lambda = \lambda(x)$. Then $\mathbb{E}(U\,|X\,) = \mathbb{E}\,(Y\,|X\,) - Q(\alpha\,|X\,) = \lambda\left[\Gamma\left(1 + \frac{1}{k}\right) - (\ln(1/(1-\alpha)))^{1/k}\right]$. In particular, the case of a conditional Rayleigh(β) distribution corresponds to the case $k = 2$, $\lambda = \sqrt{2}\beta$.*

Example 2 (Conditional Mode function) *Similar arguments can be used if $f(X) = Mode\,(Y\,|X\,)$. For example, the case of Y having a chi-squared distribution with $k = k(X)$ degrees of freedom is coherent with case [2], since the mode, for $k \geq 2$, is $k-2 = \mathbb{E}\,(Y \mid X) - 2$. More generally, if the distribution of Y conditional to $X = x$ is $Gamma(k(x), \theta(x))$, then the difference between the conditional expectation and the mode is just $\theta(x)$, so that cases [2], [3] or [4] are encountered for θ being respectively at most linear, convex or concave in x.*

Example 3 (Conditional expected shortfall) *The expected shortfall, or tail conditional mean, is one of the most popular measures of risk used in finance. Unlike the Value at Risk, it has the advantage of being a coherent risk measure. Recently, conditional versions of the expected shortfall have been also been proposed. One of the possible ways to define the conditional expected shortfall at a fixed quantile α is:*

$$\tau\,(\alpha\,|x) = \int_{-\infty}^{Q(\alpha|x)} y dF(y\,|Y \leq Q(\alpha\,|x), x).$$

Thus, $\tau(\alpha\,|x)$ is the conditional expectation of the variable $\frac{Y1\{Y \leq c\}}{P(Y \leq c)}$, for $c = Q(\alpha\,|x)$. This example then enters in Corollary 2 case [1], with $f(X) =$

$\mathbb{E}\left(\frac{Y1\{Y\leq c\}}{P(Y\leq c)}\,|X\right)$ *and* $\mathbb{E}\left(f\,|X\in A\right)=\mathbb{E}\left(\frac{Y1\{Y\leq c\}}{P(Y\leq c)}\,|X\in A\right)$ *for all* $A\in\mathcal{B}(\mathfrak{X})$.

5.2 Functionals of the distribution function.

The range of possible statistical applications founded on the results of the previous sections is not limited to testing the convexity of regression models (even in broad sense). There are cases when the object of investigation are functionals of the distribution function. When the aim is to test the convexity of a given transformation, $H(F)$, a typical way to proceed is to replace F by some estimate and then control for convexity by looking at second derivatives of $H(\hat{F})$. Since however the empirical distribution function $\hat{F}_n$ is discontinuous, $H(\hat{F}_n)$ is not differentiable, so this type of approach clearly requires a different estimator from the empirical distribution function to be chosen. The fact that our criterion focuses on differences between expectations makes it possible to construct tests for convexity of $H(F)$ by using the empirical distribution function to estimate F, because discontinuities are simply averaged out. A similar idea is exploited by Hall and Van Keilegom (2005) for testing the monotonicity of the hazard rate (via the convexity of the cumulant hazard rate, $H(F(x))=-\log(1-F(x))$). Instead of finding a smooth estimator for F, replace it into the functional H and look at the signs of the second derivatives, they propose the following statistic, that is obtained from $\hat{H}(x)=H(\hat{F}_n(x))$ by integration:

$$\int\int_{x,y:x-y,x+y\in I}\max\{0,2\hat{H}(x)-\hat{H}(x+y)-\hat{H}(x-y)\}^r w(x,y)dxdy$$

where I is the interval where the hazard rate is suspected to be monotone and w is some weight function.

The main advantage of this type of approach is that the statistics reduce to transformations of the empirical process and thus the asymptotic properties can be in general achieved by invoking some versions of the functional Glivenko-Cantelli or Donsker Theorems.

To give an example, assume that $H:[0,1]\to\mathbb{R}$ is a continuous function. Let $G=H\circ F:\mathfrak{X}\to\mathbb{R}$ and $\hat{G}=H(\hat{F}_n)$.

Corollary 3 *Let the mapping H satisfy*

$$\|\hat{G}-G\|=\|H(\hat{F})-H(F)\|\leq c\|\hat{F}-F\|$$

for some positive constant c with $\|\cdot\|$ equal to the supremum norm. Let

$$\hat{F}(\alpha,\varepsilon)=\left\{\min_j\min_i\hat{\mathbb{E}}_n\left(\hat{G}\left|B_{i+1}^{(j)}\right.\right)-\hat{\mathbb{E}}_n\left(\hat{G}\left|B_i^{(j)}\right.\right)\right\}$$

where the families of subsets $B_i^{(j)}$ are found according to the conditions of Theorem 6. If G is strictly concave in a convex subset of $\mathfrak{X}$, then for every $\varepsilon,\alpha>0$,

$$\Pr\left\{\hat{F}(\alpha,\varepsilon)\right\}\leq\varepsilon^{-1}e^{-O_P(n)}.$$

If G is convex everywhere in $\mathfrak{X}$, then

$$1 - \Pr\left\{\hat{F}(-\alpha, \varepsilon)\right\} \leq m_n \varepsilon^{-1} e^{-O_P(n)}.$$

Proof. Since for every two subsets A, B

$$\begin{aligned}
&\hat{\mathbb{E}}_n\left(\hat{G}\,|A\right) - \hat{\mathbb{E}}_n\left(\hat{G}\,|B\right) \\
&\qquad = \hat{\mathbb{E}}_n\left(G\,|A\right) - \hat{\mathbb{E}}_n\left(G\,|B\right) + \hat{\mathbb{E}}_n\left(\hat{G} - G\,|A\right) - \hat{\mathbb{E}}_n\left(\hat{G} - G\,|B\right) \\
&\qquad \leq \hat{\mathbb{E}}_n\left(G\,|A\right) - \hat{\mathbb{E}}_n\left(G\,|B\right) + 2c\|\hat{F}_n - F\|,
\end{aligned}$$

then the proof will follow the lines of Theorem 6, once observed that

$$\Pr\left\{\hat{F}(\alpha, \varepsilon)\right\} \leq \Pr\left\{\min_j \min_i \hat{\mathbb{E}}_n\left(G\left|B_{i+1}^{(j)}\right.\right) - \hat{\mathbb{E}}_n\left(G\left|B_i^{(j)}\right.\right) \geq \alpha - 2c\|\hat{F}_n - F\|\right\},$$

$$\Pr\left\{\hat{F}(-\alpha, \varepsilon)^c\right\} \leq \Pr\left\{\max_j \max_i \hat{\mathbb{E}}_n\left(G\left|B_{i+1}^{(j)}\right.\right) - \hat{\mathbb{E}}_n\left(G\left|B_i^{(j)}\right.\right) \geq \alpha - 2c\|\hat{F}_n - F\|\right\}$$

and that $\|\hat{F} - F\| = O_P(n^{-1/2})$.

5.3 Divergence-based inference.

Minimum divergence procedures have been founding an increasing popularity among statisticians since the general notion of $\phi-$divergence was first introduced by Csiszár (see for instance Csiszár 1969,1975). According to his definition, a $\phi-$divergence is a measure of discrepancy between two probability distributions P, Q, indexed by a convex function ϕ. If P and Q have densities p, q, then the $\phi-$divergence writes:

$$I_\phi(Q, P) = \int \phi\left(\frac{q(x)}{p(x)}\right) p(x) dx.$$

Minimum divergence estimation procedures, entangle many of the most popular statistical methodologies, such as maximum likelihood or minimum χ^2. A wide class of parametric minimum divergence estimators may be found in Basu *et al.* (1998), Jones *et al.* (2001). Other diergence-based estimators are found in Menéndez *et al.* (1995,2001). Typically, the estimation procedure consists in minimizing the $\phi-$divergence between a given family Ω of distributions and the empirical distribution function, or a data-driven approximation of the unknown law generating the sample. In other words, the problem consists in finding the $\phi-$projection of the observed distribution on a given set Ω. In a hypothesis testing context, the divergence-based inference procedures find an even more natural application. The null/alternative hypotheses will be written in the following form:

$$H_0 : P \in \Omega \qquad vs \qquad H_1 : P \notin \Omega$$

and high values of $I_\phi(Q, \hat{P}_n)$ for all $Q \in \Omega$ correspond to the rejection region. Both for estimation or testing purposes, a crucial aspect is given by the

way the subset Ω can be characterized: a complex definition of Ω is generally related to a more complicated derivation of projections and of their asymptotic behaviour. One of the situations where minimum divergence methods are most easily implemented is when Ω is described by a set of linear constraints, namely constraints of the form $\mathbb{E}f_i = \int f_i dQ = 0$, for f_i belonging to a given class of functions. In this case, if the number of linear constraints is finite, the projection of P on the set Ω, with respect to $\phi-$divergences solves $\frac{q^*}{p} = (\phi')^{-1}\left(\sum_i c_i f_i\right)$ where the constants c_i are found by imposing that $Q^* \in \Omega$, that is $\int f_i dQ^* = 0$ for all f_i. We refer to the monograph by Liese and Vajda (1987) or the paper of Teboulle and Vajda (1993) for more details on minimum divergence under moment conditions. Estimation and testing methods for parametric models satisfying linear constraints can be found in Broniatowski and Keziou (2004). Inequality moment conditions can also be treated as unions of equality constraints.

In the situation where either we want to test the convexity of the regression function or we want to estimate $\mu(x)$ subject to the convexity contraint via minimum divergence methods, the set Ω is the subset of all probabilty measures $P = P(x, y)$ on $(\mathfrak{X} \times \mathfrak{Y}, \mathcal{B}(\mathfrak{X} \times \mathfrak{Y}))$ yielding a convex conditional mean function.

The convexity criterion Corollary 1 as well as Theorem 5 can be motivated in a divergence-based inference setting because they enable us to re-write the set Ω as a family of probability measures satisfying given moment conditions.

Following the notation of the previous section, let $\bar{g}_{A,B} := g_A - g_B$ be the functions defined above, with $g_A(x) = \left(\frac{f 1_A}{P(A)} - \mathbb{E}\left(f \,|A\right) \frac{1_A}{P(A)}\right)$, and with $A \in \mathcal{C}(\mathfrak{X})$ and $B \in \mathfrak{M}(A)$. Then, the set Ω can be written by means of (infinitely many) linear inequality constraints:

$$\Omega = \left\{ Q \in \mathcal{M}_{P_X} : \mathbb{E}_Q \, \bar{g}_{A,B} = \int \bar{g}_{A,B} dQ \geq 0, A \in \mathcal{C}(\mathfrak{X}), B \in \mathfrak{M}(A) \right\}. \quad (5.38)$$

A sieve procedure is typically used in order to approximate Ω by a sequence of sets defined by a finite number of constraints. The consistency of this procedure is ensured if the approximating sequence Ω_n converges monotonically to Ω (see Teboulle and Vajda, 1993). We can thus construct the appropriate approximating sequence by replacing the functions $g_{A,B}$ with the functions $\bar{g}(i, j) = g_{B_{i+1}^{(j)}} - g_{B_i^{(j)}}$, introduced above. In view of Theorem 5, the sequence Ω_n defined by constraints on $\bar{g}(i, j)$ is a decreasing sequence of probability measures on $(\mathfrak{X} \times \mathfrak{Y}, \mathcal{B}(\mathfrak{X} \times \mathfrak{Y}))$ and converges to Ω, and this yields

$$I_\phi(\Omega_n, P) = \min_{Q \in \Omega_n} I_\phi(Q, P) \to_n I_\phi(\Omega, P).$$

Minimum divergence between P and Ω_n is now easy to estimate, exploiting the above mentioned characterization of projection, by a plug-in version of the $\phi-$divergence or by alternative methods (see f.i. Broniatowski and Keziou (2004)).

References

[1] S. Abramovich, M.K. Bakula, M. Matic, J. Pečarič. A variant of Jensen-Steffensen's inequality and quasi-arithmetic means. *J. Math. Anal. Appl.* **307** (2005), 370–386.

[2] J. Abrevaya and W. Jiang. A nonparametric approach to measuring and testing curvature. *J. Bus. Econom. Statist.*, **23** (2005), no. 1, 1–19.

[3] Y. Baraud, S. Huet an B. Laurent. Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *Ann. Stat.*, **33** (2005), no. 1, 214–257.

[4] A. Basu, I.R. Harris, N.L. Hjort and M.C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85** (1998), no. 3, 549–559.

[5] M. Birke and H. Dette. Testing strict monotonicity in nonparametric regression. *Math. Methods Statist.* **16** (2007), no. 2, 110–123.

[6] M. Birke and H. Dette. Estimating a convex function in nonparametric regression. *Scand. J. Statist.* **34** (2007), no. 2, 384–404.

[7] M. Broniatowski and A. Keziou. Estimation and tests for models satisfying linear constraints with unknown parameter. *Prépublication de LSTA, Université Paris VI, Jussieu*, 2004.

[8] I. Csiszár. On generalized entropy. *Studia Sci. Math. Hungar.* **4** (1969), 401–419.

[9] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probability* **3** (1975), 146–158.

[10] A. Dembo and O. Zeitouni. *Large deviations techniques and applications.* Second edition. Applications of Mathematics (New York), 38. Springer-Verlag, New York, 1998.

[11] H. Dette, N. Neumeyer and K.F. Pilz. A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli*, **12** (2006), no. 3, 469–490.

[12] J.L. Doob. *Stochastic Processes.* Wiley, New York, 1953.

[13] S.S. Dragomir. Some refinements of Jensen's inequality. *J. Math. Anal. Appl.* **168** (1992), No.2, 518–522

[14] L. Dümbgen, V.I. Piterbarg and D. Zholud. On the limit distribution of multiscale test statistics for nonparametric curve estimation. *Math. Methods Statist.* **15** (2006), no. 1, 20–25.

[15] R.M. Dudley. A course on empirical processes. *École d'été de probabilités de Saint-Flour, XII—1982*, 1–142, Lecture Notes in Math., **1097**, Springer, Berlin, 1984.

[16] P. Hall and I. Van Keilegom. Testing for monotone increasing hazard rate. *Ann. Statist.* **33** (2005), no. 3, 1109–1137.

[17] P. Hall and A. Yatchew. Unified approach to testing functional hypotheses in semiparametric contexts. *J. Econometrics* **127** (2005), no. 2, 225–252.

[18] M.C. Jones, N.L. Hjort, I.R. Harris and A. Basu. A comparison of related density-based minimum divergence estimators. *Biometrika* **88** (2001), no. 3, 865–873.

[19] S. Karlin and A. Novikoff. Generalized convex inequalitites. *Pacific J. Math.*, **13** (1963), 1251–1279.

[20] A. Kozek and Z. Suchanecki. Multifunctions of faces for conditional expectations of selectors and Jensen's inequality. *J. Multivar. Anal.*, **10**(1980), 579–598.

[21] F. Liese and I. Vajda. Convex statistical distances. Teubner-Texte zur

Mathematik [Teubner Texts in Mathematics], 95. BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1987.
[22] M. Menéndez, D. Morales, L. Pardo and I. Vajda. Divergence-based estimation and testing of statistical models of classification. *J. Multivariate Anal.* **54** (1995), no. 2, 329–354.
[23] M. Menéndez, D. Morales, L. Pardo and I. Vajda. Minimum divergence estimators based on grouped data. *Ann. Inst. Statist. Math.* **53** (2001), no. 2, 277–288.
[24] A.McD. Mercer. A variant of Jensen's inequality. *J. Inequal. Pure Appl. Math.* **4** (2003), N.4, 2 pp. (electronic).
[25] M. Merkle. Jensen's inequality for medians. *Stat. Prob. Letters*, **71** (2005), 277–281.
[26] D. Morales, L. Pardo and I. Vajda. Some new statistics for testing hypotheses in parametric models. *J. Multivariate Anal.* **62** (1997), no. 1, 137–168.
[27] S. Orbe, E. Ferreira and J. Rodriguez-Poo. Nonparametric estimation of time varying parameters under shape restrictions. *J. Econometrics*, **126** (2005), no. 1, 53–77.
[28] M. Perlman. Jensen's inequality for a convex vector-valued function on an infinite-dimensional space. *J. Multivar. Anal.*, **4**(1974), 52–65.
[29] L. Reboul. Estimation of a function under shape restrictions. Applications to reliability. *Ann. Statist.* **33** (2005), no. 3, 1330–1356.
[30] J. Rooin. A refinement of Jensen's inequality. *J. Inequal. Pure Appl. Math.* **6** (2005), N. 2, Article 38, 4 pp. (electronic).
[31] P. Roselli and M. Willem. A convexity inequality. *Amer. Math. Monthly*, **109**, No. 1, 64–70.
[32] F.C. Sanchez, J.M.F. Castillo and P.L. Papini. Seven views on approximate convexity and the geometry of $K-$spaces. *J. London Math. Soc.* , **72** (2005), N.2, 457–477.
[33] G.R. Shorak and J.A. Wellner. *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986.
[34] C.H. Spiegelman. A univariate extension of Jensen's inequality. *J. Res. Nat. Bur. Standards*, **85** (1980), No. 5, 363–365.
[35] M. Teboulle and I. Vajda. Convergence of best ϕ-entropy estimates. *IEEE Trans. Inform. Theory* **39** (1993), no. 1, 297–301.
[36] T.O. To and K.W. Yip. A generalized Jensen's inequality. *Pacific J. Math.* **58** (1975), no. 1, 255–259.
[37] A. M. Zapała. Jensen's inequality for conditional expectations in Banach spaces. *Real Anal. Exchange* **26** (2000/01), no. 2, 541–552.