CONSISTENCY OF M-ESTIMATORS FOR NON-IDENTICALLY DISTRIBUTED DATA: THE CASE OF FIXED-DESIGN DISTRIBUTIONAL REGRESSION

Axel Bücher, Johan Segers, Torben Staud

LIDAM Discussion Paper ISBA
2025 / 21

ISBA
Voie du Roman Pays 20 - L1.04.01
B-1348 Louvain-la-Neuve
Email : lidam-library@uclouvain.be
https://uclouvain.be/en/research-institutes/lidam/isba/publication

Consistency of M-estimators for non-identically distributed data: the case of fixed-design distributional regression

Axel Bücher* Johan Segers† Torben Staud‡

November 17, 2025

Abstract

This paper explores strong and weak consistency of M-estimators for non-identically distributed data, extending prior work. Emphasis is given to scenarios where data is viewed as a triangular array, which encompasses distributional regression models with non-random covariates. Primitive conditions are established for specific applications, such as estimation based on minimizing empirical proper scoring rules or conditional maximum likelihood. A key motivation is addressing challenges in extreme value statistics, where parameter-dependent supports can cause criterion functions to attain the value $-\infty$, hindering the application of existing theorems.

Keywords: Block-Maxima Method, Conditional Maximum Likelihood, Distributional Regression, Minimum Scoring Rule Estimation, Non-Random Covariates.

1 Introduction

M-estimators are a broad class of estimators that arise from maximizing an objective function that corresponds to a sample average. More specifically, given observations $z_1, \dots, z_n$, they arise from (approximately) maximizing a function of the form

$$\eta \mapsto M_n(\eta) = \frac{1}{n} \sum_{i=1}^{n} m_\eta(z_i) \tag{1.1}$$

over a given parameter set H, where m_η is a known function taking values in $[-\infty, \infty)$. The framework is broad enough to cover as special cases maximum likelihood estimators from classical parametric statistics or empirical risk minimizers from supervised learning problems. Asymptotic theory, including weak or strong consistency as well as asymptotic normality, has been studied extensively, in particular for the case of independent and

*Ruhr-Universität Bochum, Fakultät für Mathematik. Email: axel.buecher@rub.de
†KU Leuven, Department of Mathematics, and UCLouvain, LIDAM/ISBA. Email: jjjsegers@kuleuven.be
‡Ruhr-Universität Bochum, Fakultät für Mathematik. Email: torben.staud@rub.de

identically distributed observations. For an overview, we refer to Section 5 in van der Vaart (1998) and the references therein.

The present paper is concerned with strong (and weak) consistency of M-estimators in situations where the data is not identically distributed, thereby providing extensions of Sections 5.1 and 5.2 in van der Vaart (1998). Such situations commonly occur in (distributional) regression models in the fixed-design setting (Kneib et al., 2023; Klein, 2024), that is, with non-random covariates, a classical example being the ordinary least squares estimator in linear models. Respective asymptotic theory is usually based on viewing the data as a triangular array, and we adopt this perspective as well; see, for instance, Liese and Vajda (1994), Berlinet et al. (2000), and Salibián-Barrera (2006) for models where the parameter only enters through a location term. A general framework for deriving strong consistency is offered by Lachout et al. (2005), but their assumptions are high-level. It is one the main purposes of this paper to provide more primitive conditions and model assumptions.

A specific need has motivated this paper: common parametric distributional regression models in extreme value statistics have parameter-dependent supports; see Philip et al. (2020) for applications in extreme-event attribution and Section 4.2 below for mathematical details. Consequently, the criterion function of any likelihood-based estimator may attain the value $-\infty$, hindering, for example, the application of Theorem 2.2 in Lachout et al. (2005). Despite acknowledgment of this issue (see, e.g., the discussion of Assumption C.1 in Padoan et al., 2010), a formal consistency proof for conditional maximum likelihood estimators in extreme value statistics remains absent to the best of our knowledge.

General consistency results for M-estimators in triangular arrays are presented in Section 2 and adapted to distributional regression models with non-random covariates in Section 3. Specific applications are detailed in Section 4, including estimators based on optimizing empirical proper scoring rules (Gneiting and Raftery, 2007; Dawid et al., 2016) and (conditional) maximum likelihood estimators. Additionally, we examine conditional maximum likelihood estimators for heavy-tailed regression models for block maxima. The corresponding consistency result, outlined in Section 4.3, can be viewed as an extension of the main finding in Dombry (2015) to a conditional framework, focusing on the heavy-tailed case, as in Bücher and Segers (2018). Furthermore, Section 5 provides adaptations of Theorem 2.2 from Lachout et al. (2005) suitable for settings where the criterion function may attain the value $-\infty$. These adaptations could be collectively termed 'argmax theorems without uniform convergence', with the main proof idea going back to Wald (1949). Finally, all proofs are provided in Section 6.

Throughout, the arrow $\rightsquigarrow$ denotes weak convergence of probability measures on a metric space. For positive integer n, we write $[n] = \{1, \ldots, n\}$.

2 Consistency of M-estimators in triangular arrays

As motivated in the introduction, we consider a data-generating process corresponding to a sufficiently regular triangular array of random variables.

Assumption 1 (Data-generating process)**.** *The metric space $(\mathscr{Z}, d_{\mathscr{Z}})$ is separable, and $(Z_{n,i} : n \in \mathbb{N}, i \in [n])$ is a rowwise independent triangular array of random variables in $\mathscr{Z}$, all defined on the same probability space. The distributions $P_{n,i}$ of $Z_{n,i}$ satisfy $n^{-1}\sum_{i=1}^{n} P_{n,i} \rightsquigarrow P_Z$ as $n \to \infty$ for some distribution P_Z on $\mathscr{Z}$.*

The condition that the random variables $Z_{n,i}$ are all defined on the same probability space is only needed for strong consistency (Lemma 2.1). For weak consistency (Lemma 2.2), it is sufficient that within each row n, the variables $(Z_{n,i} : i \in [n])$ are defined on the same probability space. For simplicity, we will disregard this distinction henceforth.

The subsequent condition pertains to the known criterion function $m(\eta, z) = m_\eta(z)$ informally introduced in (1.1).

Assumption 2 (Criterion function)**.** *The metric space (H, d_H) is compact and the function $m : H \times \mathscr{Z} \to [-\infty, \infty)$ is upper semicontinuous. For all $\eta \in H$, the function m_η is P_Z-quasi-integrable.*

While compactness of H is vital for our proofs, the results remain applicable even when H is not compact, provided suitable adjustments are made. Specifically, one can either additionally show that the estimator eventually maps into a compact set, or replace H with an appropriate compactification. The latter may present additional challenges, potentially resolvable through an observation-grouping technique as outlined in Exercise 5.25 of van der Vaart (1998).

For $\eta \in H$ fixed, we will write $m_\eta : \mathscr{Z} \to [-\infty, \infty)$ for the function $z \mapsto m(\eta, z)$. The random and asymptotic criterion functions $M_n, M : H \to [-\infty, \infty)$ are defined as

$$M_n(\eta) = \frac{1}{n}\sum_{i=1}^{n} m_\eta(Z_{n,i}), \tag{2.1}$$

$$M(\eta) = \int_{\mathscr{Z}} m_\eta(z)\, P_Z(\mathrm{d}z) = P_Z m_\eta. \tag{2.2}$$

Informally, $M_n(\eta)$ can be considered as an estimator for $M(\eta)$, such that any (approximate) maximizer of M_n can be regarded as an (approximate) maximizer of M. The following assumption concerns the latter maximization problem.

Assumption 3 (True parameter)**.** *The function M has a unique point of finite maximum $\eta_0 \in H$:*

$$\forall \eta \in H \setminus \{\eta_0\} : \qquad -\infty \le M(\eta) < M(\eta_0) < \infty.$$

The function m_{η_0} is continuous P_Z-almost everywhere.

The aforementioned proximity of $M_n(\eta)$ to $M(\eta)$ necessitates a suitable (weak or strong) law of large numbers. Notably, for triangular arrays, strong consistency demands non-standard results which require more than mere (uniform) integrability (Chandra and Goswami, 1992, p. 215). A sufficient condition is L^2-stochastic dominance, as detailed in Hu et al. (1989): a family of random variables $(X_\alpha)_{\alpha \in A}$ on the real line is said to be L^p*-stochastically dominated*, with $1 \le p < \infty$, if there exists a nonnegative random variable W with $\mathsf{E}[W^p] \le \infty$ such that

$$\forall t \in [0, \infty) : \qquad \sup_{\alpha \in A} \mathsf{P}(|X_\alpha| > t) \le \mathsf{P}(W > t). \tag{2.3}$$

The family is called *uniformly integrable* if

$$\lim_{c \to \infty} \sup_{\alpha \in A} \mathsf{E}[|X_\alpha| \mathbb{1}_{\{|X_\alpha| > c\}}] = 0. \tag{2.4}$$

By Markov's inequality, a sufficient condition for L^p-stochastic dominance is the existence of $\delta > 0$ such that $\sup_{\alpha \in A} \mathsf{E}[|X_\alpha|^{p+\delta}] < \infty$, and a sufficient condition for uniform integrability is L^1-stochastic dominance. For $B \subseteq H$, put $m_B(z) = \sup_{\eta \in B} m_\eta(z)$. Let $\overline{B}(\eta, \delta)$ be the closed ball in H with center $\eta \in H$ and radius $\delta > 0$. Let $(x)^+ = \max(x, 0)$ denote the positive part of $x \in \mathbb{R}$.

Assumption 4 (L^2-stochastic dominance). *(i) There exists $n_0 \in \mathbb{N}$ such that the family $\{m_{\eta_0}(Z_{n,i}) : n \ge n_0, i \in [n]\}$ is L^2-stochastically dominated.*

(ii) For every $\eta \in H \setminus \{\eta_0\}$ there exists $\delta > 0$ and $n_0 \in \mathbb{N}$ such that $P_Z m_{\overline{B}(\eta,\delta)} < \infty$ and the family $\{(m_{\overline{B}(\eta,\delta)}(Z_{n,i}))^+ : n \ge n_0, i \in [n]\}$ is L^2-stochastically dominated.

Often, the function m is upper bounded on $H \times \mathscr{Z}$, that is,

$$\sup_{\eta \in H, z \in \mathscr{Z}} m_\eta(z) < \infty. \tag{2.5}$$

Under this condition, the L^2-stochastic dominance condition in (ii) is immediate, while the condition in (i) becomes a condition on the lower tail of $m_{\eta_0}(Z_{n,i})$ only.

Theorem 2.1 (Strong consistency). *If Assumptions 1, 2, 3 and 4 hold, then every estimator sequence $\hat{\eta}_n$ defined on the same probability space as the array $(Z_{n,i})_{n,i}$ that satisfies $M_n(\hat{\eta}_n) \ge M_n(\eta_0) - \mathrm{o}(1)$ almost surely as $n \to \infty$ is strongly consistent for η_0.*

Weak consistency of $\hat{\eta}_n$ can be derived under a weaker integrability condition than L^2-stochastic dominance. Again, if m satisfies Eq. (2.5), item (ii) in the next condition is immediate and item (i) only concerns the lower tail of $m_{\eta_0}(Z_{n,i})$.

Assumption 5 (Uniform integrability). *(i) There exists $n_0 \in \mathbb{N}$ such that the family $\{m_{\eta_0}(Z_{n,i}) : n \ge n_0, i \in [n]\}$ is uniformly integrable.*

(ii) For every $\eta \in H \setminus \{\eta_0\}$ there exists $\delta > 0$ and $n_0 \in \mathbb{N}$ such that $P_Z m_{\overline{B}(\eta,\delta)} < \infty$ and such that the family $\{(m_{\overline{B}(\eta,\delta)}(Z_{n,i}))^+ : n \ge n_0, i \in [n]\}$ is uniformly integrable.

Theorem 2.2 (Weak consistency). *If Assumptions 1, 2, 3 and 5 hold, then every estimator sequence $\hat{\eta}_n$ such that $\hat{\eta}_n$ is defined on the same probability space as the tuple $(Z_{n,i})_{i \in [n]}$ and that satisfies $M_n(\hat{\eta}_n) \ge M_n(\eta_0) - \mathsf{o}_{\mathsf{P}}(1)$ as $n \to \infty$ is weakly consistent for η_0.*

Remark 2.3. By the mere definition of weak consistency, the conclusion in Lemma 2.2 does not depend on the dependence between rows in the triangular array. This is different in Lemma 2.1, and it is perhaps surprising that the conclusion in Lemma 2.1 is true *for any* form of stochastic dependence/independence between rows. The price to be paid for this generality is a stronger integrability condition.

3 Consistency of M-estimators in conditional models with non-random covariates

Throughout this section, we consider M-estimators in conditional models that are based on n observations from a tuple (x, Y), with the distribution of the response Y depending on a non-random covariate x. In many related situations such as nonparametric regression or classical linear regression, deriving asymptotic results is most naturally framed by viewing the data as a triangular array. We adopt this perspective and impose conditions that enable us to apply the general results from the previous section.

Specifically, we assume that (x, Y) takes values in $\mathbb{X} \times \mathbb{Y}$, where $(\mathbb{X}, d_{\mathbb{X}})$ and $(\mathbb{Y}, d_{\mathbb{Y}})$ are separable metric spaces equipped with the respective Borel σ-fields $\mathscr{B}(\mathbb{X})$ and $\mathscr{B}(\mathbb{Y})$, respectively. The distribution of the response variable is assumed to depend on x, which will formally be described by the notion of a Markov kernel, that is, by a map $Q : \mathbb{X} \times \mathscr{B}(\mathbb{Y}) \to [0, 1]$ such that $Q(x, \cdot\,)$ is a probability measure on $\mathbb{Y}$ for each $x \in \mathbb{X}$ and such that $x \mapsto Q(x, B)$ is measurable for each $B \in \mathscr{B}(\mathbb{Y})$. Despite the fact that x is deterministic, it helps to think of $Q(x, \cdot\,)$ as the conditional distribution of the response variable taking values in $\mathbb{Y}$ given the covariate x; an alternative notation to keep in mind would be $P_{Y|X=x}(\,\cdot\,)$ instead of $Q(x, \cdot\,)$.

In a nutshell, we will apply the framework of Section 2 to $\mathscr{X} = \mathbb{X} \times \mathbb{Y}$ and $Z_{n,i} = (x_{n,i}, Y_{n,i})$ in Assumption 6 below. The same approach could be followed to handle random covariates, putting $Z_{n,i} = (X_{n,i}, Y_{n,i})$; we do not develop this further in this paper.

Assumption 6 (Data-generating process)**.** *For sample size $n \in \mathbb{N}$, we observe a finite sequence $(x_{n,i}, Y_{n,i})_{i\in[n]}$, where $x_{n,1}, \dots, x_{n,n}$ are deterministic elements in $\mathbb{X}$ and where $Y_{n,1}, \dots, Y_{n,n}$ are independent $\mathbb{Y}$-valued random variables such that $Y_{n,i} \sim Q_n(x_{n,i}, \mathrm{d}y)$ for all $i \in [n]$ and for some Markov kernel $Q_n : \mathbb{X} \times \mathscr{B}(\mathbb{Y}) \to [0, 1]$. Moreover:*

- *(a) The observations $x_{n,1}, \dots, x_{n,n}$ are regular in the sense that there exists a non-degenerate probability distribution P_X on $\mathbb{X}$ such that $n^{-1} \sum_{i=1}^{n} \delta_{x_{n,i}} \rightsquigarrow P_X$ as $n \to \infty$.*
- *(b) The distribution of $Y_{n,1}, \dots, Y_{n,n}$ is regular in the sense that there exists a Markov kernel $Q : \mathbb{X} \times \mathscr{B}(\mathbb{Y}) \to [0, 1]$ such that $Q_n(\,\cdot\,, \mathrm{d}y)$ converges continuously to $Q(\,\cdot\,, \mathrm{d}y)$ in the weak topology, that is, for any sequence $x_n \to x$ in $\mathbb{X}$, the measures $Q_n(x_n, \mathrm{d}y)$ converge weakly to $Q(x, \mathrm{d}y)$. Moreover, for the limiting kernel Q, the function $x \mapsto Q(x, \mathrm{d}y)$ is weakly continuous, i.e., if $x_n \to x$ in $\mathbb{X}$, then $Q(x_n, \mathrm{d}y) \rightsquigarrow Q(x, \mathrm{d}y)$ as $n \to \infty$.*

A typical example for part (a) of Assumption 6, often studied in non-parametric regression, is the simple uniform design $x_{i,n} = i/n$, which could in practice for instance correspond to rescaled time. It has also been imposed in Liese and Vajda (1994); Berlinet et al. (2000) for studying M-estimators in general mean regression models. Moreover, it is comparable to standard assumptions for consistency statements on the ordinary least squares estimator in linear mean regression models with non-random design matrix $\boldsymbol{X}_n = (x_{n,1}^\top, \dots, x_{n,n}^\top)^\top$, which is typically assumed to satisfy $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_{n,i} x_{n,i}^\top = \boldsymbol{S}$

for some matrix $\boldsymbol{S} \in \mathbb{R}^{k\times k}$. Under Assumption 6(a), we would have $\boldsymbol{S} = \int xx^\top P_X(\mathrm{d}x)$, provided P_X has finite cross-moments of some order $p > 2$.

An important special case for part (b) of Assumption 6 arises when $Q_n = Q$ does not depend on n. In this case, the condition simplifies to requiring only weak continuity of the mapping $x \mapsto Q(x, \mathrm{d}y)$; this covers most parametric distributional regression models like $Q(x, \cdot\,) = \mathscr{N}(f_\theta(x), \sigma^2)$ for some smooth, parametric regression function f_θ and some $\sigma^2 > 0$. More specific examples will be worked out in Section 4.

We start by showing that Assumption 6 implies Assumption 1. Consistency of M-estimators is then an immediate corollary.

Lemma 3.1. *Suppose that Assumption 6 is met. Then Assumption 1 is met with* $\mathscr{X} = \mathbb{X} \times \mathbb{Y}$, $Z_{n,i} = (x_{n,i}, Y_{n,i})$, $P_{n,i}(\mathrm{d}(x, y)) = Q_n(x, \mathrm{d}y)\,\delta_{x_{n,i}}(\mathrm{d}x)$ *and* $P_Z = P_{X,Y}$, *where*

$$P_{X,Y}(\mathrm{d}(x, y)) = Q(x, \mathrm{d}y)P_X(\mathrm{d}x) \tag{3.1}$$

and where δ_x *is the Dirac-measure at* x.

Corollary 3.2. *Suppose that Assumption 6 holds. If, additionally, Assumptions 2, 3 and 4 are met for* $\mathscr{X} = \mathbb{X} \times \mathbb{Y}$, $Z_{n,i} = (x_{n,i}, Y_{n,i})$ *and* $P_Z = P_{X,Y}$, *then every estimator sequence* $\hat\eta_n$ *defined on the same probability space as the array* $(x_{n,i}, Y_{n,i})_{n,i}$ *that satisfies* $M_n(\hat\eta_n) \geq M_n(\eta_0) - \mathrm{o}(1)$ *almost surely as* $n \to \infty$ *is strongly consistent for* η_0.

Corollary 3.3. *Suppose that Assumption 6 holds. If, additionally, Assumptions 2, 3 and 5 are met for* $\mathscr{X} = \mathbb{X} \times \mathbb{Y}$, $Z_{n,i} = (x_{n,i}, Y_{n,i})$ *and* $P_Z = P_{X,Y}$, *then every estimator sequence* $\hat\eta_n$ *such that* $\hat\eta_n$ *is defined on the same probability space as* $(x_{n,i}, Y_{n,i})_{i\in[n]}$ *and that satisfies* $M_n(\hat\eta_n) \geq M_n(\eta_0) - \mathsf{o}_\mathsf{P}(1)$ *as* $n \to \infty$ *is weakly consistent for* η_0.

4 Applications to optimum score and conditional maximum likelihood estimators

As outlined in the introduction, the results from the previous sections can be applied to various common estimators. We start by dealing with optimum scoring rule estimation, then adapt these results to (conditional) maximum likelihood estimation, and finally examine a pseudo-maximum likelihood estimator in a specific heavy-tailed regression model for block maxima. While the dependence of the distribution of $Y_{n,i}$ on n only enters through $x_{n,i}$ in the first two applications (i.e., $Q_n = Q$ in Assumption 6), a more general dependence is needed for the latter application.

4.1 Optimum score estimators with non-random covariates

We specialize the set-up of Section 3 to the case of optimum score estimators. Recall that $\mathbb{X}$ and $\mathbb{Y}$ are separable metric spaces, while the parameter space H is compact.

Assumption 7 (Parametric distributional regression model and sampling scheme)**.** *Let* $\{x_{n,i} \colon n \in \mathbb{N}, i \in [n]\} \subseteq \mathbb{X}$ *be a deterministic triangular array satisfying Assumption 6(a)* *and let $\{Y_{n,i}\}_{i,n}$ be a rowwise independent $\mathbb{Y}$-valued triangular array such that $Y_{n,i} \sim P_{\bar\theta(x_{n,i},\eta_0)}$, where $(P_\theta : \theta \in \Theta)$ is a known parametric family of distributions, $\bar\theta : \mathbb{X} \times H \to \Theta$ is a known link function and $\eta_0 \in H$ is the true unknown parameter of interest. The function $\bar\theta$ is continuous while the map $\theta \mapsto P_\theta$ is one-to-one (i.e., the parameter θ is identifiable) and weakly continuous. Finally, η_0 is identifiable in the sense that, for the limit distribution P_X in Assumption 6(a), we have*

$$\forall \eta \in H \setminus \{\eta_0\} : \qquad P_X\left(\left\{x \in \mathbb{X} : \bar\theta(x,\eta) \neq \bar\theta(x,\eta_0)\right\}\right) > 0. \tag{4.1}$$

In the notation of Section 3, the Markov kernel $Q_n(x,\,\cdot\,) = Q(x,\,\cdot\,) = P_{\bar\theta(x,\eta_0)}(\,\cdot\,)$ does not depend on n. Moreover, the distribution $P_{X,Y}$ from Eq. (3.1) is

$$P_{X,Y}(\mathrm{d}(x,y)) = P_{\bar\theta(x,\eta_0)}(\mathrm{d}y)\, P_X(\mathrm{d}x). \tag{4.2}$$

Let $\mathscr{P}$ be a convex set of probability measures on $\mathbb{Y}$. Given an outcome $y \in \mathbb{Y}$ we wish to attach a score to a probabilistic forecast $P \in \mathscr{P}$, with higher scores being better. A *strictly proper scoring rule* is a map $S : \mathscr{P} \times \mathbb{Y} \to [-\infty,\infty]$ with the following properties:

(i) For every $P \in \mathscr{P}$, the map $y \mapsto S(P,y)$ is Borel measurable.

(ii) For every $P, P' \in \mathscr{P}$, the integral $S(P',P) = \int_{\mathbb{Y}} S(P',y)\, P(\mathrm{d}y)$ exists in $[-\infty,\infty]$.

(iii) For every $P, P' \in \mathscr{P}'$, we have $S(P,P) \geq S(P',P)$, with strict inequality if $P \neq P'$.

We refer to Gneiting and Raftery (2007); Dawid et al. (2016) for further context and references.

Assumption 8 (Scoring rule)**.** *$S : \mathscr{P} \times \mathbb{Y} \to [-\infty,\infty]$ is a strictly proper scoring rule on $\mathscr{P}$, a convex set of probability measures on $\mathbb{Y}$ containing P_θ for all $\theta \in \Theta$, and satisfies $S(P_\theta,y) < \infty$ for all $(\theta,y) \in \Theta \times \mathbb{Y}$. The map $(\theta,y) \mapsto S(P_\theta,y)$ is upper semicontinuous and the map $(x,y) \mapsto S(P_{\bar\theta(x,\eta_0)},y)$ is continuous $P_{X,Y}$-almost everywhere.*

For $\eta \in H$, define $m_\eta : \mathbb{X} \times \mathbb{Y} \to [-\infty,\infty)$ by $m_\eta(x,y) = S(P_{\bar\theta(x,\eta)},y)$. The random and limiting criterion functions in Eqs. (2.1) and (2.2) become

$$\begin{aligned} M_n(\eta) &= \frac{1}{n}\sum_{i=1}^{n} S(P_{\bar\theta(x_{n,i},\eta)}, Y_{n,i}), \qquad (4.3)\\ M(\eta) &= \int_{\mathbb{X}}\int_{\mathbb{Y}} S(P_{\bar\theta(x,\eta)},y)\, P_{\bar\theta(x,\eta_0)}(\mathrm{d}y)\, P_X(\mathrm{d}x) = \int_{\mathbb{X}} S(P_{\bar\theta(x,\eta)}, P_{\bar\theta(x,\eta_0)})\, P_X(\mathrm{d}x). \end{aligned}$$

We will prove below that the function $\eta \mapsto M(\eta)$ is well defined (i.e., for every $\eta \in H$, the map $x \mapsto S(P_{\bar\theta(x,\eta)}, P_{\bar\theta(x,\eta_0)})$ is quasi-integrable with respect to P_X) as a consequence of the following stochastic dominance condition; see Eq. (2.3).

Assumption 9 (L^2-stochastic dominance for scoring rules)**.** *(i) The family of random variables $\{S(P_\theta,Y) : \theta \in \Theta, Y \sim P_\theta\}$ is L^2-stochastically dominated.*

(ii) For every $\eta \in H \setminus \{\eta_0\}$ there exists $\delta > 0$ such that the family of random variables

$$\Big\{ \sup_{\bar\eta \in \overline{B}(\eta,\delta)} \big(S(P_{\bar\theta(x,\bar\eta)}, Y)\big)^+ : x \in \mathbb{X}, Y \sim P_{\bar\theta(x,\eta_0)} \Big\} \tag{4.4}$$

is L^2-stochastically dominated.

Theorem 4.1. *If Assumptions 7, 8, and 9 are met, then every estimator $\hat\eta_n$ satisfying $M_n(\hat\eta_n) \geq M_n(\eta_0) - o_{a.s.}(1)$ is strongly consistent for η_0.*

For the corresponding result about weak consistency, it is sufficient to replace the words "L^2-stochastically dominated" in Assumption 9 by "uniformly integrable" and replace the $o_{a.s.}(1)$ term in Lemma 4.1 by $o_{\mathsf{P}}(1)$.

Example 4.2 (Optimum energy score estimation)**.** For positive integer d and for $0 < \beta < 2$, let $\mathscr{P}_\beta(\mathbb{R}^d)$ be the set of probability measures P on $\mathbb{R}^d$ such that $\int_{\mathbb{R}^d} \|y\|^\beta \, P(\mathrm{d}y) < \infty$, where $\|\cdot\|$ denotes the Euclidean norm. The *energy score* (Székely, 2003) $\mathrm{ES}_\beta : \mathscr{P}_\beta(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathrm{ES}_\beta(P, y) = \frac{1}{2} \mathsf{E}\big[\|Y - Y'\|^\beta\big] - \mathsf{E}\big[\|Y - y\|^\beta\big],$$

where the random vectors Y and Y' are independent and have distribution P. The energy score is strictly proper relative to $\mathscr{P}_\beta(\mathbb{R}^d)$ thanks to Székely (2003, Theorem 1); see also Gneiting and Raftery (2007, Sections 4.3 and 5.1). The special case $d = 1$ and $\beta = 1$ yields the Continuous Ranked Probability Score (CRPS). The energy score is itself a special case of the *kernel score* introduced in Dawid (2007), see also Gneiting and Raftery (2007, Section 5.1).

Suppose that Assumption 7 is met with $\mathbb{Y} = \mathbb{R}^d$ and with $P_\theta \in \mathscr{P}_\beta(\mathbb{R}^d)$ for all $\theta \in \Theta$. Further assume that there exists $\delta > 0$ such that

$$\sup_{\theta \in \Theta} \int_{\mathbb{R}^d} \|y\|^{2\beta+\delta} \, P_\theta(\mathrm{d}y) < \infty. \tag{4.5}$$

Then every optimum energy score estimator $\hat\eta_n$ satisfying $M_n(\hat\eta_n) \geq M_n(\eta_0) - o_{a.s.}(1)$ is strongly consistent for η_0, where

$$M_n(\eta) = \frac{1}{n} \sum_{i=1}^n \mathrm{ES}_\beta(P_{\bar\theta(x_{n,i},\eta)}, Y_{n,i}), \qquad \eta \in H,$$

See Section 6.3 for the proof. A weak consistency statement only requires Eq. (4.5) with the exponent $2\beta + \delta$ replaced by $\beta + \delta$; we omit further details.

4.2 Conditional maximum likelihood estimation

Conditional maximum likelihood estimators arise as a special case of optimum score estimators in Section 4.1. Specifically, we will work under Assumption 7 and additionally assume that the parametric model $(P_\theta : \theta \in \Theta)$ is dominated by a σ-finite measure ν on $\mathbb{Y}$ with densities $p_\theta = \mathrm{d}P_\theta/\mathrm{d}\nu$ for $\theta \in \Theta$. As scoring rule in Assumption 8 we consider the *logarithmic score*

$$\mathrm{LogS}(P, y) = \log\Big(\frac{\mathrm{d}P}{\mathrm{d}\nu}(y)\Big) \qquad P \in \mathscr{P},\, y \in \mathbb{Y}, \tag{4.6}$$

which is strictly proper with respect to the family $\mathscr{P}$ of probability measures on $\mathbb{Y}$ dominated by ν; see Gneiting and Raftery (2007, Section 4.1) for context and references. If $P = P_\theta$ for some $\theta \in \Theta$, we also write

$$\mathrm{LogS}(P, y) = \log p_\theta(y) = \ell_\theta(y),$$

such that $m_\eta(x, y) = \ell_{\bar{\theta}(x,\eta)}(y)$ and such that the criterion function M_n in Eq. (4.3) is equal to the conditional log-likelihood function

$$M_n(\eta) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\theta(x_{n,i},\eta)}(Y_{n,i}).$$

Note that y does not need to belong to the support of P_θ, so that the value $\ell_\theta(y) = -\infty$ is allowed; since $\ell_\theta(y) = +\infty$ can only occur on a ν-null set, we can and will exclude it altogether. A maximiser $\hat{\eta}_n$ of $\eta \mapsto M_n(\eta)$ is a *conditional maximum likelihood estimator*. Under the following set of conditions, its consistency is an immediate corollary of Lemma 4.1.

Assumption 10. *Assumption 7 holds and the parametric model $(P_\theta : \theta \in \Theta)$ is dominated by a σ-finite measure ν on $\mathbb{Y}$ with densities $p_\theta = \mathrm{d}P_\theta/\mathrm{d}\nu$ for $\theta \in \Theta$. Moreover:*

(i) the map $(\theta, y) \mapsto p_\theta(y)$ takes values in $[0, \infty)$ and is upper semicontinuous;

(ii) the map $(x, y) \mapsto p_{\bar{\theta}(x,\eta_0)}(y)$ is continuous $P_{X,Y}$-almost everywhere;

(iii) the family $\{\ell_\theta(Y) : \theta \in \Theta, Y \sim P_\theta\}$ is L^2-stochastically dominated;

(iv) for every $\eta \in H \setminus \{\eta_0\}$ there exists $\delta > 0$ such that $\{\sup_{\bar{\eta}\in\overline{B}(\eta,\delta)}(\ell_{\bar{\theta}(x,\bar{\eta})}(Y))^+ : x \in \mathbb{X}, Y \sim P_{\bar{\theta}(x,\eta_0)}\}$ is L^2-stochastically dominated.

The most complicated condition is (iv). A simple and useful sufficient condition is $\sup_{\theta\in\Theta, y\in\mathbb{Y}} p_\theta(y) < \infty$; see, for instance, Lemmas 4.4 and 4.5 below.

Corollary 4.3. *If Assumption 10 holds, then every estimator sequence $\hat{\eta}_n$ such that $M_n(\hat{\eta}_n) \geq M_n(\eta_0) - \mathrm{o}_{a.s.}(1)$ as $n \to \infty$ is strongly consistent for η_0.*

As before, weak consistency already holds if L^2-stochastic dominance in Assumption 10(iii) and (iv) is replaced by uniform integrability and if the $\mathrm{o}_{a.s.}(1)$ term in Lemma 4.3 is replaced by an $\mathrm{o}_{\mathsf{P}}(1)$ term.

Lemma 4.3 establishes consistency for estimators in models with covariate-dependent parameters, which are central in extreme value analysis and its applications. A key challenge in such models is that the support of the distribution depends on the parameter itself, rendering classical maximum likelihood estimation theory inapplicable.

Example 4.4 (Generalized extreme value distribution with covariate-dependent parameters)**.** The density function of the generalized extreme value (GEV) distribution on $\mathbb{Y} = \mathbb{R}$ with location-scale-shape parameter vector $\theta = (\mu, \sigma, \xi) \in \mathbb{R} \times (0, \infty) \times \mathbb{R}$ is

$$p_\theta(y) = \begin{cases} \frac{1}{\sigma}\left(1 + \xi\frac{y-\mu}{\sigma}\right)^{-1-1/\xi} \exp\left\{-\left(1 + \xi\frac{y-\mu}{\sigma}\right)^{-1/\xi}\right\}, & \text{if } \xi \neq 0,\ 1 + \xi\frac{y-\mu}{\sigma} > 0, \\ \frac{1}{\sigma}e^{-(y-\mu)/\sigma} \exp\left(-e^{-(y-\mu)/\sigma}\right), & \text{if } \xi = 0, \\ 0, & \text{otherwise.} \end{cases}$$

The support of the GEV distribution is

$$\{y \in \mathbb{R} : \sigma + \xi(y - \mu) > 0\} = \begin{cases} (\mu - \sigma/\xi, \infty) & \text{if } \xi > 0, \\ \mathbb{R} & \text{if } \xi = 0, \\ (-\infty, \mu - \sigma/\xi) & \text{if } \xi < 0, \end{cases}$$

and depends on the parameter vector θ. Let $\mathbb{X}$ and H be compact metric spaces and let the link function $\bar{\theta} : \mathbb{X} \times H \to \mathbb{R} \times (0, \infty) \times \mathbb{R}$ be continuous and such that $\Theta := \bar{\theta}(\mathbb{X} \times H) \subseteq \mathbb{R} \times (0, \infty) \times (-1, \infty)$, that is, the shape parameter is restricted to $\xi > -1$. In contrast to the general theory in the preceding sections, we assume that the covariate space $\mathbb{X}$ is compact rather than just separable, and this in order to deal with the L^2-stochastic dominance condition in Assumption 10, which involves suprema over $\theta \in \Theta$ and $x \in \mathbb{X}$. Under the set-up of Assumption 7 with true $\eta_0 \in H$, the conditions in Assumption 10 hold. As a consequence, the conditional maximum likelihood estimator in H is strongly consistent for the regression parameter η_0. Note that distributional regression models for the GEV-distribution are very common in extreme-event attribution; see Philip et al. (2020) and the references therein.

Example 4.5 (Generalized Pareto distribution with covariate-dependent parameters)**.** The generalized Pareto (GP) density with scale-shape parameter vector $\theta = (a, \xi) \in (0, \infty) \times \mathbb{R}$ on $\mathbb{Y} = (0, \infty)$ has density function

$$p_\theta(y) = \begin{cases} a^{-1}\left(1 + \xi y/a\right)^{-1-1/\xi} & \text{if } \xi \neq 0,\ 1 + \xi y/a > 0, \\ a^{-1}e^{-y/a} & \text{if } \xi = 0, \\ 0 & \text{otherwise.} \end{cases}$$

As for the GEV distribution, the support of the GP distribution depends on the parameter:

$$\{y \in (0, \infty) : a + \xi y > 0\} = \begin{cases} (0, \infty) & \text{if } \xi \geq 0, \\ (0, -a/\xi) & \text{if } \xi < 0. \end{cases}$$

As in Lemma 4.4, we restrict attention to the case $\xi > -1$. If $\mathbb{X}$ and H are compact metric spaces and if $\bar{\theta} : \mathbb{X} \times H \to (0, \infty) \times (-1, \infty)$ is a continuous link function, then in the set-up of Assumption 7, Lemma 4.3 yields the strong consistency of the conditional maximum likelihood estimator for the true regression parameter η_0 within H.

4.3 Fitting heavy-tailed regression models to block maxima

We apply the triangular set-up of Section 3 to show the consistency of the maximum likelihood estimator in a regression model for the Fréchet distribution fitted to heavy-tailed block maxima in the presence of covariates. The distribution of the block maxima depends on the block size and becomes Fréchet only in the limit as the block size tends to infinity. This feature motivates the dependence of the Markov kernel on n in Assumption 1. This section's main result, presented in Lemma 4.8, can be regarded as a version of the main result in Dombry (2015), restricted to the heavy-tailed case but extended to the conditional setting.

Let F_0 be a cdf on $\mathbb{R}$ in the max-domain of attraction of the Fréchet distribution with shape parameter $\alpha \in (0, \infty)$: there exists a sequence $a_r > 0$ such that for all $y \in \mathbb{R}$ we have

$$\lim_{r\to\infty} F_0^r(a_r y) = \Phi_\alpha(y) = \begin{cases} \exp\left(-y^{-\alpha}\right), & \text{if } y > 0, \\ 0, & \text{if } y \le 0. \end{cases} \tag{4.7}$$

For all x in some separable metric space $\mathbb{X}$, let F_x be a cdf on $\mathbb{R}$ such that the map $(x, y) \mapsto F_x(y)$ is Borel measurable. Assume that there exists $c : \mathbb{X} \to (0, \infty)$ such that

$$\lim_{y\to\infty} \sup_{x\in\mathbb{X}} \left| \frac{\log F_x(y)}{\log F_0(y)} - c(x) \right| = 0. \tag{4.8}$$

In words, the tail $1 - F_x(y) \sim -\log F_x(y)$ can be approximated uniformly well by the tail of F_0, up to a covariate-dependent multiplicative constant $c(x)$. The assumption mirrors that of heteroscedastic extremes in Einmahl et al. (2016).

Similar as in Dombry (2015), we observe n independent block maxima $M_{n,1}, \ldots, M_{n,n}$ together with n covariates $x_{n,1}, \ldots, x_{n,n} \in \mathbb{X}$. The block maxima take the form

$$M_{n,i} = \max_{t\in[r_n]} \xi_{n,i;t}, \tag{4.9}$$

where $r_n \in \mathbb{N}$ is the block size and $\xi_{n,i;1}, \ldots, \xi_{n,i;r_n}$ is a potentially unobserved independent random sample drawn from $F_{x_{n,i}}$. The cdf of the rescaled block maximum $M_{n,i}/a_{r_n}$ is thus

$$\mathsf{P}\left(M_{n,i}/a_{r_n} \le y\right) = F_{x_{n,i}}^{r_n}(a_{r_n} y).$$

Asymptotically, this distribution is Fréchet with covariate-dependent scale parameter.

Lemma 4.6. *Under Eqs.* (4.7) *and* (4.8)*, we have, for all* $y > 0$*,*

$$\lim_{r\to\infty} \sup_{x\in\mathbb{X}} \left| F_x^r(a_r y) - \Phi_\alpha\big(y/\sigma(x)\big) \right| = 0 \qquad \textit{with} \quad \sigma(x) = c(x)^{1/\alpha}. \tag{4.10}$$

Proof. Fix $y > 0$. For $x \in \mathbb{X}$ and $r \in \mathbb{N}$, we have

$$F_x^r(a_r y) = \exp\left\{ r \log F_0(a_r y) \cdot \frac{\log F_x(a_r y)}{\log F_0(a_r y)} \right\} = \left(F_0^r(a_r y)\right)^{\frac{\log F_x(a_r y)}{\log F_0(a_r y)}}.$$

By Eq. (4.8), the big exponent converges to $c(x)$ uniformly in $x \in \mathbb{X}$, so by Eq. (4.7), the whole expression converges to $[\exp(-y^{-\alpha})]^{c(x)} = \exp\left(-c(x) y^{-\alpha}\right)$, uniformly in $x \in \mathbb{X}$; this is Eq. (4.10). □

Remark 4.7. A converse statement to Lemma 4.6 holds as well in the sense that if $c(x)$ is bounded away from zero and infinity, then (4.10) implies the existence of F_0 such that Eqs. (4.7) and (4.8) hold. To see this, simply put $F_0 = F_{x_0}$ for some arbitrary $x_0 \in \mathbb{X}$. Eq. (4.7) is then immediate, while Eq. (4.8) requires some careful analysis, of which we omit the details for brevity.

We now suppose that the scale function $\sigma(x)$ in Eq. (4.10) can be described by a parameter $\beta \in B$. A typical case is the loglinear model $\sigma_\beta(x) = \exp\big(\beta^\top x\big)$ where $\mathbb{X}, B \subseteq \mathbb{R}^d$. The aim is to estimate the triple (α, β, a_{r_n}) on the basis of the observed covariate-maxima pairs $(x_{n,i}, M_{n,i})$ for $i \in [n]$. Under Eq. (4.9), the distribution of $M_{n,i}$ is approximately Fréchet with shape parameter α and scale parameter $a_{r_n}\sigma_\beta(x_{n,i})$, so we can hope to estimate the parameters consistently by (conditional) maximum likelihood.

The probability density function of the Fréchet distribution Φ_α in (4.7) is

$$p_{1,\alpha}(y) = \begin{cases} \alpha y^{-\alpha-1} \exp\left(-y^{-\alpha}\right) & \text{if } y > 0, \\ 0 & \text{if } y \le 0. \end{cases}$$

The density of the Fréchet distribution with scale-shape parameter vector $(\tau, \alpha) \in (0,\infty)^2$ is thus

$$p_{\tau,\alpha}(y) = \tau^{-1} p_{1,\alpha}(y/\tau) = \frac{\alpha}{\tau}(y/\tau)^{-\alpha-1} \exp\left(-(y/\tau)^{-\alpha}\right)$$

The rescaled loglikelihood based on the observed covariate-maxima pairs is then

$$L_n(\tau, \beta, \alpha) := \frac{1}{n}\sum_{i=1}^{n} \log p_{\tau\sigma_\beta(x_{n,i}),\alpha}(M_{n,i}). \tag{4.11}$$

Let $\alpha_0 \in (0,\infty)$ and $\beta_0 \in B$ denote the true shape and regression parameters, respectively, in the sense that for every $y \in (0,\infty)$ we have, as in Lemma 4.6,

$$\lim_{n\to\infty} \max_{i\in[n]} \left|\mathsf{P}\left(M_{n,i}/a_{r_n} \le y\right) - \Phi_{\alpha_0}\left(y/\sigma_{\beta_0}(x_{n,i})\right)\right| = 0.$$

Theorem 4.8. *Let $A \subseteq (0,\infty)$ be compact and let $B, \mathbb{X}$ be compact metric spaces. Let $(M_{n,i} : n \in \mathbb{N}, i \in [n])$ be as in Eq. (4.9) with distributions F_x as in Eqs. (4.7) and (4.8), where $\alpha_0 \in A$ denotes the true parameter and where $c(x) = \sigma_{\beta_0}(x)^{\alpha_0}$ for some continuous map $(x,\beta) \mapsto \sigma_\beta(x) \in (0,\infty)$ and some $\beta_0 \in B$. Assume that $\log n = o(r_n)$, that $n^{-1}\sum_{i=1}^n x_{n,i} \rightsquigarrow P_X$, and that β_0 is identifiable in the sense that*

$$\forall(\beta,\gamma) \in (B \times (0,\infty)) \setminus \{(\beta_0, 1)\} : \qquad P_X\left(\{x \in \mathbb{X} : \gamma\sigma_\beta(x) \ne \sigma_{\beta_0}(x)\}\right) > 0. \tag{4.12}$$

Let $0 < \gamma_- < 1 < \gamma_+ < \infty$. Then any random sequence $\hat\eta_n = (\hat\alpha_n, \hat\beta_n, \hat\tau_n)$ in $A \times B \times [\gamma_- a_{r_n}, \gamma_+ a_{r_n}]$ such that $L_n(\hat\eta_n) \ge L_n(\alpha_0, \beta_0, a_{r_n}) + \mathrm{o}_{a.s.}(1)$ satisfies

$$\left(\hat\alpha_n, \hat\beta_n, \hat\tau_n/a_{r_n}\right) \to (\alpha_0, \beta_0, 1), \qquad n \to \infty, \quad a.s.$$

Remark 4.9 (Block size)**.** The condition $\log n = o(r_n)$ requires that the block size r_n grows faster than logarithmically with the number of block maxima n. It serves to ensure that with large probability, all block maxima are bounded away from zero, see Lemma 6.8. The same condition has been imposed in (Dombry, 2015, Theorem 2).

Remark 4.10 (Scaling sequence)**.** The scaling sequence a_{r_n} is unknown, and so is the search interval $[\gamma_- a_{r_n}, \gamma_+ a_{r_n}]$ for the overall scale parameter τ of the conditional maximum likelihood estimator $\hat{\eta}_n$. This issue can be solved by taking γ_- and γ_+ sufficiently small and large, respectively, and replacing a_{r_n} by the median (or any other fixed quantile) of $M_{n,1}, \ldots, M_{n,n}$. This median is of the form $(\kappa + \mathrm{o}_{a.s.}(1))\, a_{r_n}$ with $\kappa > 0$ a model-dependent constant, more precisely the corresponding quantile of the $\sigma_\beta(X)$-scale mixture of the Fréchet distribution with shape parameter α_0, where X has distribution P_X.

5 Argmax theorems without uniform convergence

The Argmax continuous mapping theorem in van der Vaart and Wellner (2023, Theorem 3.2.2) states conditions under which the argmax of a sequence of random criterion functions converges weakly to the argmax of the weak limit, in an appropriate function space, of those criterion functions. In Theorem 2.2 in Lachout et al. (2005), such functional convergence is not required, and an upper bound on suprema over certain sets suffices. The two theorems below provide slightly modified versions of the latter theorem, versions that are at the basis of our Lemma 2.1 and Lemma 2.2.

Let $(\mathbb{E}, \mathscr{E})$ be a measurable space. Let $\mathbb{P}_n$, for $n \in \mathbb{N}$, be a sequence of random probability measures on $\mathbb{E}$ and let P be a non-random probability measure on $\mathbb{E}$, respectively; we think of $\mathbb{P}_n$ as being close to P for large n. Write integrals of extended real-valued functions $f : \mathbb{E} \to [-\infty, \infty]$ as $\mathbb{P}_n f$ and Pf. Let (H, d_H) be a compact metric space and let $m : H \times \mathbb{E} \to [-\infty, \infty)$ be a measurable function. We frequently write $m_\eta(z) = m(\eta, z)$ for $\eta \in H$ and $z \in \mathbb{E}$. The integrals $M(\eta) = Pm_\eta$ and $M_n(\eta) = \mathbb{P}_n m_\eta$ are assumed to exist in $[-\infty, \infty]$ for all $\eta \in H$; condition (A1) below will in fact imply that $M(\eta)$ is not equal to ∞. Recall that $m_B(z) = \sup_{\eta \in B} m_\eta(z)$ for $B \subseteq H$ and $z \in E$, and that $B(\eta, \delta)$ and $\overline{B}(\eta, \delta)$ are the open and closed balls, respectively, in H with center $\eta \in H$ and radius $\delta > 0$.

Theorem 5.1. *Assume the above set-up with the following conditions:*

(A1) The function $M : H \to [-\infty, \infty]$ has a unique point of finite maximum $\eta_0 \in H$, that is,

$$-\infty \le M(\eta) < M(\eta_0) < \infty \qquad \textit{for all } \eta \in H \setminus \{\eta_0\}.$$

(A2) For every $\eta \in H$, there exists $V_\eta \in \mathscr{E}$ with $P(V_\eta) = 0$ such that for all $z \in \mathbb{E} \setminus V_\eta$, the function $\bar{\eta} \mapsto m_{\bar{\eta}}(z)$ is upper semicontinuous at η.

(A3) For all $\eta \in H \setminus \{\eta_0\}$ there exists $\delta > 0$ such that $Pm_{\overline{B}(\eta,\delta)} < \infty$ and such that, for all $\rho \in (0, \delta]$, we have

$$\limsup_{n\to\infty} \mathbb{P}_n m_{\overline{B}(\eta,\rho)} \le Pm_{\overline{B}(\eta,\rho)}, \qquad \textit{a.s.}$$

(A4) We have

$$\liminf_{n\to\infty} M_n(\eta_0) \geq M(\eta_0), \qquad a.s.$$

Then every estimator $\hat{\eta}_n$ *satisfying* $M_n(\hat{\eta}_n) \geq M_n(\eta_0) - \mathrm{o}_{a.s.}(1)$ *is strongly consistent for the estimation of* η_0*, that is,* $\lim_{n\to\infty} \hat{\eta}_n = \eta_0$ *a.s.*

The following result is a variation of the previous theorem for weak rather than strong consistency.

Theorem 5.2. *Assume the above set-up with Assumptions (A1) and (A2) as in Lemma 5.1, and with the following two assumptions:*

(A3') For all $\eta \in H \setminus \{\eta_0\}$ *there exists* $\delta > 0$ *such that* $Pm_{\overline{B}(\eta,\delta)} < \infty$ *and such that, for all* $\rho \in (0, \delta]$*, we have*

$$\mathbb{P}_n m_{\overline{B}(\eta,\rho)} \leq P m_{\overline{B}(\eta,\rho)} + \mathrm{o}_{\mathsf{P}}(1), \qquad n \to \infty.$$

(A4') We have

$$M_n(\eta_0) \geq M(\eta_0) - \mathrm{o}_{\mathsf{P}}(1), \qquad n \to \infty.$$

Then every estimator $\hat{\eta}_n$ *satisfying* $M_n(\hat{\eta}_n) \geq M_n(\eta_0) - \mathrm{o}_{\mathsf{P}}(1)$ *is weakly consistent for the estimation of* η_0*, that is,* $\lim_{n\to\infty} \mathsf{P}\left[d_H(\hat{\eta}_n, \eta_0) \geq \delta_0\right] = 0$ *for every* $\delta_0 > 0$.

Remark 5.3 (Comparison with Theorem 2.2 in Lachout et al. (2005))**.** As mentioned above, Lemma 5.1 is a slightly modified version of Theorem 2.2 in Lachout et al. (2005). Translating their notation to our setup, the most important differences are as follows. While they require m_η to be real-valued, we permit $m_\eta(z) = -\infty$, thus accommodating conditional maximum likelihood estimators when the support is parameter-dependent. Their estimator $\hat{\eta}_n$ must satisfy $M_n(\hat{\eta}_n) \geq \sup_{\eta\in H} M_n(\eta) - \mathrm{o}_{a.s.}(1)$, while we only require $M_n(\hat{\eta}_n) \geq M_n(\eta_0) - \mathrm{o}_{a.s.}(1)$. Finally, they allow for the case where the criterion function M may have multiple maxima, and they show almost sure convergence of the estimator to the set of maximizers, as in Theorem 5.14 in van der Vaart (1998). The latter conclusion coincides with ours in case the set of maximizers is a singleton, as required in condition (A1).

6 Auxiliary results and proofs

6.1 Proofs for Section 2

Let $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{Z_{n,i}}$ denote the empirical distribution of the nth row of variables in Assumption 1.

Lemma 6.1 (Strong law of large numbers)**.** *Under Assumption 1, if* $f : \mathscr{Z} \to \mathbb{R}$ *is Borel measurable and continuous* P_Z*-almost everywhere, and if the family* $\{f(Z_{n,i}) : n \geq n_0, i \in [n]\}$ *is* L^2*-stochastically dominated for some* $n_0 \in \mathbb{N}$*, then* $P_Z|f| < \infty$ *and* $\mathbb{P}_n f \to P_Z f$ *almost surely as* $n \to \infty$*.*

Proof of Lemma 6.1. First we show $P_Z|f| < \infty$. Let W be a nonnegative random variable such that $\mathsf{E}[W^2] < \infty$ and such that, for some $n_0 \in \mathbb{N}$,

$$\forall n \geq n_0, i \in [n], t \geq 0 : \qquad P_{n,i}\{|f| > t\} \leq \mathsf{P}(W > t).$$

Then $P_{n,i}f^2 \leq \mathsf{E}[W^2]$ and thus also $P_n f^2 \leq \mathsf{E}[W^2]$ where $P_n = n^{-1}\sum_{i=1}^n P_{n,i}$. It follows that $P_Z|f| = \lim_{k\to\infty} P_Z(|f| \wedge k)$ is finite, since

$$P_Z(|f| \wedge k) = \lim_{n\to\infty} P_n(|f| \wedge k) \leq \lim_{n\to\infty}(P_n f^2)^{1/2} \leq (\mathsf{E}[W^2])^{1/2}.$$

Next we show the almost sure convergence of $\mathbb{P}_n f$. We have

$$\mathsf{E}[\mathbb{P}_n f] = \frac{1}{n}\sum_{i=1}^n P_{n,i}f = P_n f$$

and thus

$$\mathbb{P}_n f - \mathsf{E}[\mathbb{P}_n f] = \frac{1}{n}\sum_{i=1}^n \epsilon_{n,i} \qquad \text{where} \quad \epsilon_{n,i} = f(Z_{n,i}) - P_{n,i}f.$$

The random variables $\epsilon_{n,i}$ are rowwise independent and the array $\{\epsilon_{n,i} : n \geq n_0, i \in [n]\}$ is L^2-stochastically dominated by $W + (\mathsf{E}[W^2])^{1/2}$; indeed, $|\epsilon_{n,i}| \leq |f(Z_{n,i})| + |P_{n,i}f|$ and $|P_{n,i}f| \leq (P_{n,i}f^2)^{1/2} \leq (\mathsf{E}[W^2])^{1/2}$ for all $n \geq n_0$ and $i \in [n]$. By Theorem 2 in Hu et al. (1989), we find $\mathbb{P}_n f - \mathsf{E}[\mathbb{P}_n f] \to 0$ almost surely as $n \to \infty$.

It remains to show that $P_n f \to P_Z f$ as $n \to \infty$. Since $P_n \rightsquigarrow P_Z$ as $n \to \infty$ and since f is continuous P_Z-almost everywhere, the continuous mapping theorem (van der Vaart and Wellner, 2023, Theorem 1.11.1) implies that $f_\# P_n \rightsquigarrow f_\# P_Z$ as $n \to \infty$; in words, the distribution of f under P_n converges to the one of f under P_Z. Moreover, by L^2-stochastic dominance, $\sup_{n\geq n_0} P_n f^2 \leq \sup_{n\geq n_0, i\in[n]} P_{n,i}f^2 < \infty$. By Example 2.21 in van der Vaart (1998), the first moment of f under P_n converges to the first moment of f under P_Z. □

Corollary 6.2 (Varadarajan's theorem for triangular arrays, Varadarajan, 1958a)**.** *Under Assumption 1, we have* $\mathsf{P}(\mathbb{P}_n \rightsquigarrow P_Z \text{ as } n \to \infty) = 1$.

Proof of Lemma 6.2. By Lemma 6.3, it is sufficient to show that $\mathbb{P}_n f \to P_Z f$ almost surely as $n \to \infty$ for every bounded and continuous function $f : \mathscr{X} \to \mathbb{R}$. But this is an immediate consequence of Lemma 6.1. □

Lemma 6.3. *Let $P, P_1, P_2, \ldots$ be probability measures on a metric space $\mathbb{M}$ equipped with its Borel σ-field. If $\mathbb{M}$ is separable, there exists a countable set $\mathscr{D} \subseteq \mathrm{Lip}_b(\mathbb{M})$ such that $P_n g \to Pg$ for all $g \in \mathscr{D}$ already implies $P_n \rightsquigarrow P$.*

Proof. Fix $f \in \mathrm{Lip}_b(\mathbb{M})$. By the Portmanteau theorem, it is sufficient to show that $P_n f \to Pf$. By the proof of Theorem 3.1 in Varadarajan (1958b), we may assume that $\mathbb{M}$ is totally bounded. Arguments in the proof of Theorem 11.4.1 in Dudley (2002) yield the separability of $(\mathrm{Lip}_b(\mathbb{M}), \|\cdot\|_\infty)$. Thus, there exists a countable and dense subset

$\mathscr{D} \subseteq (\mathrm{Lip}_b(\mathbb{M}), \|\cdot\|_\infty)$. Let $\varepsilon > 0$ and choose a $g \in \mathscr{D}$ with $\|f - g\|_\infty < \varepsilon/3$. Further, by assumption there exists n_0 such that for all $n \geq n_0$ we have $|P_n g - Pg| \leq \varepsilon/3$. Hence,

$$|P_n f - Pf| \leq |P_n f - P_n g| + |P_n g - Pg| + |Pg - Pf| \leq \varepsilon$$

for all $n \geq n_0$ which implies the assertion. □

Lemma 6.4 (Portmanteau for unbounded upper semicontinuous functions)**.** *Under Assumption 1, if $f : \mathscr{X} \to \mathbb{R}$ is upper semicontinuous and quasi-integrable with respect to P_Z, and if the family $\{(f^+(Z_{n,i}) : n \geq n_0, i \in [n]\}$ is L^2-stochastically dominated for some $n_0 \in \mathbb{N}$, then*

$$\mathsf{P}\left[\limsup_{n\to\infty} \mathbb{P}_n f \leq P_Z f\right] = 1.$$

Proof of Lemma 6.4. Nothing has to be shown if $P_Z f = \infty$, so assume that $P_Z f < \infty$. Write $f = g_1 + g_2$ where $g_1 = \max(f, 0) = f^+ \geq 0$ and $g_2 = \min(f, 0) \leq 0$, and note that both g_1 and g_2 are upper semicontinuous. Since $\mathbb{P}_n f = \mathbb{P}_n g_1 + \mathbb{P}_n g_2$ and $P_Z f = P_Z g_1 + P_Z g_2$, it is sufficient to show that

$$\mathsf{P}\left[\limsup_{n\to\infty} \mathbb{P}_n g_j \leq P_Z g_j\right] = 1 \tag{6.1}$$

for $j = 1$ and $j = 2$ separately.

For $j = 2$, the claim in Eq. (6.1) follows immediately from Lemma 6.2 and item (v) in the Portmanteau Theorem 1.3.4 in van der Vaart and Wellner (2023), as g_2 is upper semicontinuous and bounded above by 0.

For $j = 1$, we may repeat the arguments in the second paragraph of the proof of Lemma 6.1 to find that $\mathbb{P}_n g_1 - \mathsf{E}[\mathbb{P}_n g_1] = \mathrm{o}(1)$ almost surely as $n \to \infty$. Since $\mathsf{E}[\mathbb{P}_n g_1] = P_n g_1$ with $P_n = n^{-1}\sum_{i=1}^n P_{n,i}$, it remains to show that

$$\limsup_{n\to\infty} P_n g_1 \leq P_Z g_1. \tag{6.2}$$

By assumption, the family $\{(g_1)_\# P_{n,i} : n \geq n_0, i \in [n]\}$ is L^2-stochastically dominated, and therefore so is the family $\{(g_1)_\# P_n : n \geq n_0\}$. As a consequence, $\sup_{n\geq n_0} P_n[g_1^2] < \infty$, so that g_1 is uniformly integrable under P_n, i.e., $\lim_{k\to\infty}\sup_{n\geq n_0} P_n[g_1 \mathbb{1}_{\{g_1 > k\}}] = 0$. But then we have for all $k > 0$ the inequalities

$$\begin{aligned}\limsup_{n\to\infty} P_n g_1 &\leq \limsup_{n\to\infty} P_n[\min(g_1, k)] + \sup_{n\geq n_0} P_n[(g_1 - k)^+] \\ &\leq P_Z[\min(g_1, k)] + \sup_{n\geq n_0} P_n[g_1 \mathbb{1}_{\{g_1 > k\}}],\end{aligned}$$

where, in the last inequality, we used again item (v) in Theorem 1.3.4 in van der Vaart and Wellner (2023). Letting $k \to \infty$ yields Eq. (6.2), as required. □

The following permanence property of upper semicontinuous functions can be easily proven using a subsequence argument but is also part of a special case of Bergman's maximum theorem; see for instance Lemma 17.30 in Aliprantis and Border (2006).

Lemma 6.5 (Upper semicontinuity)**.** *Let $\mathscr{Z}$ and H be metric spaces and let $f : H \times \mathscr{Z} \to [-\infty, \infty]$ be upper semicontinuous. If $K \subseteq H$ is non-empty and compact, then the function $f_K : \mathscr{Z} \to [-\infty, \infty]$ defined by*

$$f_K(z) = \sup_{\eta \in K} f(\eta, z), \qquad z \in \mathscr{Z},$$

is upper semicontinuous.

Proof of Lemma 2.1. The result follows from Lemma 5.1, with $\mathbb{E} = \mathscr{Z}$, $P = P_Z$ and $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{Z_{n,i}}$. We only need to check Conditions (A1)–(A4).

- Assumption (A1) is part of Assumption 3.
- Assumption (A2) follows from Assumption 2, since slicing an upper semicontinuous function on a product space to a single variable produces an upper semicontinuous function again.
- Finiteness of the integral in Assumption (A3) is a consequence of the first part of Assumption 4(ii). The inequality involving the limsup in Assumption (A3) follows from Lemma 6.4 applied to $m_{\overline{B}(\eta,\rho)}$: the latter function is upper semicontinuous by Assumption 2 and Lemma 6.5, and its positive part is L^2-stochastically dominated by Assumption 4(ii).
- Assumption (A4) is a consequence of Lemma 6.1 applied to $f = m_{\eta_0}$ in combination with Assumption 3 and Assumption 4(i). □

Proposition 6.6 (Weak law of large numbers for triangular arrays)**.** *For some $n_0 \in \mathbb{N}$, consider a triangular array $\{X_{ni} : n \geq n_0, i \in [n]\}$ of rowwise independent random variables. If the family is uniformly integrable, see Eq.* (2.4)*, then*

$$\frac{1}{n} \sum_{i=1}^n (X_{ni} - \mathsf{E}[X_{ni}]) \to 0 \qquad \textit{in probability, as } n \to \infty.$$

Lemma 6.6 is a direct consequence of the main theorem in Gut (1992); uniform integrability in the Cesàro sense is already sufficient. According to Chandra and Goswami (1992, p. 215), uniform integrability is *not* enough to conclude the *strong* law of large numbers.

Lemma 6.7. *Under Assumption 1, if $f : \mathscr{Z} \to \mathbb{R}$ is upper semicontinuous and quasi-integrable with respect to P_Z, and if the family $\{f(Z_{n,i})^+ : n \geq n_0, i \in [n]\}$ is uniformly integrable, then $\mathbb{P}_n f \leq P_Z f + o_{\mathsf{P}}(1)$ as $n \to \infty$.*

Proof of Lemma 6.7. If $P_Z f = \infty$, there is nothing to prove, so assume $P_Z f < \infty$. As in the proof of Lemma 6.4, we apply the same decomposition $f = g_1 + g_2$ with $g_1 = f^+ \geq 0$ and $g_2 = \min(f, 0) \leq 0$. It is sufficient to show that $\mathbb{P}_n g_j \leq P_Z g_j + o_{\mathsf{P}}(1)$ for $j \in \{1, 2\}$.

For $j = 2$, the desired inequality is even true almost surely, by the same argument as in the proof of Lemma 6.4, thanks to the Pormanteau Theorem 1.3.4 in van der Vaart and Wellner (2023): we have $\mathbb{P}_n \rightsquigarrow P_Z$ almost surely by Lemma 6.2 (which only relies on Assumption 1), while g_2 is upper semicontinuous and bounded above (by 0).

For $j = 1$, we apply Lemma 6.6 to the uniformly integrable triangular array $X_{ni} = g_1(Z_{n,i})$ for $n \geq n_0$ and $i \in [n]$: we have

$$\mathbb{P}_n g_1 = \frac{1}{n}\sum_{i=1}^{n} g_1(Z_{n,i}) = \frac{1}{n}\sum_{i=1}^{n} X_{ni},$$
$$\frac{1}{n}\sum_{i=1}^{n} \mathsf{E}[X_{ni}] = \frac{1}{n}\sum_{i=1}^{n} P_{n,i} g_1 = P_n g_1.$$

As the difference between the two quantities converges to zero in probability thanks to Lemma 6.6, it remains to show that $\limsup_{n\to\infty} P_n g_1 \leq P_Z g_1$. But this follows from the fact that $P_n \rightsquigarrow P$ (Assumption 1) and the upper semicontinuity of g_1, by an application of the Portmanteau Theorem 1.3.4 in van der Vaart and Wellner (2023). □

Proof of Lemma 2.2. This time, we apply Lemma 5.2.

Assumptions (A1) and (A2) are verified by the same arguments as in the proof of Lemma 2.1.

The integrability condition in Assumption (A3') is a consequence of the first part of Assumption 4(ii). The inequality involving the convergence in probability in Assumption (A3') follows from Lemma 6.7 applied to $f = m_{\overline{B}(\eta,\delta)}$: this function is upper semicontinuous by Assumption 2 and Lemma 6.5, and its positive part is uniformly integrable with respect to the distributions $P_{n,i}$ by Assumption 5(ii).

Assumption (A4') is a consequence of Lemma 6.6 applied to $X_{ni} = m_{\eta_0}(Z_{n,i})$, which are uniformly integrable by Assumption 5(i): we have

$$M_n(\eta_0) = \frac{1}{n}\sum_{i=1}^{n} m_{\eta_0}(Z_{n,i}) = \frac{1}{n}\sum_{i=1}^{n} X_{n,i}$$

with expectation

$$\frac{1}{n}\sum_{i=1}^{n} \mathsf{E}[X_{n,i}] = \frac{1}{n}\sum_{i=1}^{n} P_{n,i} m_{\eta_0} = P_n m_{\eta_0}.$$

By Theorem 2.20 in van der Vaart (1998), $P_n m_{\eta_0}$ converges to $M(\eta_0) = P_Z m_{\eta_0}$ since $P_n \rightsquigarrow P_Z$ (Assumption 1), m_{η_0} is continuous P_Z-almost everywhere (Assumption 3) and $(m_{\eta_0})_{\#} P_n$ is uniformly integrable (as a consequence of Assumption 5(i) and the identity $P_n = n^{-1}\sum_{k=1}^{n} P_{n,k}$). □

6.2 Proofs for Section 3

Proof of Lemma 3.1. We only need to show that $P_n \rightsquigarrow P_{X,Y}$, where $P_n = n^{-1}\sum_{i=1}^{n} P_{n,i}$. By Lemma 6.3, it is sufficient to show that $P_n f \to P_{X,Y} f$ as $n \to \infty$, for every bounded and uniformly continuous function $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$. For that purpose, let

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_{n,i}}, \qquad h_n(x) = \int_{\mathbb{Y}} f(x,y)\, Q_n(x, \mathrm{d}y), \qquad h(x) = \int_{\mathbb{Y}} f(x,y)\, Q(x, \mathrm{d}y),$$

so that $P_n f = \mu_n h_n$ and $P_{X,Y} f = P_X h$. The proof will be finished once we show $\mu_n h_n \to P_X h$ as $n \to \infty$. Note that, by Assumption 6(a), we have $\mu_n \rightsquigarrow P_X$ as $n \to \infty$.

First, we have that $h_n - h$ converges continuously to zero. Indeed let $x_n \to \bar{x} \in \mathbb{X}$ and $\varepsilon > 0$. By uniform continuity of f and Assumption 6(b), we may choose $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have $\|f(x_n, \cdot) - f(\bar{x}, \cdot)\|_\infty < \varepsilon/4$ and $|\{Q_n(x_n, \cdot) - Q(\bar{x}, \cdot)\}[f(\bar{x}, \cdot)]| \vee |\{Q(x_n, \cdot) - Q(\bar{x}, \cdot)\}[f(\bar{x}, \cdot)]| < \varepsilon/4$. Then, for all $n \geq n_0$,

$$
\begin{aligned}
&|h_n(x_n) - h(x_n)| \\
&\leq \int_{\mathbb{Y}} |f(x_n, y) - f(\bar{x}, y)| \, Q_n(x_n, \mathrm{d}y) + \int_{\mathbb{Y}} |f(x_n, y) - f(\bar{x}, y)| \, Q(x_n, \mathrm{d}y) \\
&\qquad + \Big| \int_{\mathbb{Y}} f(\bar{x}, y) \{Q_n(x_n, \mathrm{d}y) - Q(\bar{x}, \mathrm{d}y)\} \Big| + \Big| \int_{\mathbb{Y}} f(\bar{x}, y) \{Q(x_n, \mathrm{d}y) - Q(\bar{x}, \mathrm{d}y)\} \Big|,
\end{aligned}
$$

which is strictly smaller than ε. By the extended continuous mapping theorem [Theorem 1.11.1 in van der Vaart and Wellner (2023)], this gives $h_n(Z_n) - h(Z_n) \rightsquigarrow 0$ for $Z_n \sim \mu_n$, and since $|h_n| \vee |h| \leq |f|$ and since f is bounded, the moments also converge to zero, that is, $\mu_n[h_n - h] \to 0$. Thus, we are left with verifying $\mu_n h \to P_X h$. By Assumption 6(a) this reduces to checking that h is bounded and continuous. The first part was already verified and the continuity follows by similar but simpler arguments as in the last display. □

6.3 Proofs for Section 4

Proof of Lemma 4.1. We apply Lemma 3.2 and check Assumptions 2, 3, 4 and 6, starting with the latter.

Regarding Assumption 6, the Markov kernel is $Q_n(x, \mathrm{d}y) = P_{\bar{\theta}(\eta_0, x)}(\mathrm{d}y)$. This kernel does not depend on n and is weakly continuous as a function of x by Assumption 7.

Regarding Assumption 2, the compactness of H was assumed from the start. By Assumption 8, the objective function $(\eta, x, y) \mapsto m_\eta(x, y) = S(P_{\bar{\theta}(x,\eta)}, y)$ attains values in $[-\infty, \infty)$. Furthermore, it is upper semicontinuous as the link function $\bar{\theta}$ is continuous and the map $(\theta, y) \mapsto S(P_\theta, y)$ is upper semicontinuous by Assumptions 7 and 8, respectively. Finally, since $\{S(P_\theta, Y) : \theta \in \Theta, Y \sim P_\theta\}$ is L^2-stochastically dominated[1] by Assumption 9, we have $\sup_{\theta \in \Theta} |S(P_\theta, P_\theta)| < \infty$, which in turn implies that m_η is $P_{X,Y}$-integrable, with $P_{X,Y}$ in Eq. (4.2), as $S(P_{\bar{\theta}(x,\eta)}, P_{\bar{\theta}(x,\eta_0)}) \leq S(P_{\bar{\theta}(x,\eta_0)}, P_{\bar{\theta}(x,\eta_0)})$.

Regarding Assumption 3, the continuity of m_{η_0} is stipulated in Assumption 8. As argued in the proof of Assumption 2, m_η is $P_{X,Y}$-integrable, which implies that $M(\eta_0) = \int_{\mathbb{X}} S(P_{\bar{\theta}(x,\eta_0)}, P_{\bar{\theta}(x,\eta_0)}) \, P_X(\mathrm{d}x) < \infty$. Moreover, since S is a strictly proper scoring rule, the identifiability of the parametric model in Assumption 7 implies that

$$
\begin{aligned}
&P_X \left(\left\{ x \in \mathbb{X} : S(P_{\bar{\theta}(x,\eta)}, P_{\bar{\theta}(x,\eta_0)}) < S(P_{\bar{\theta}(x,\eta_0)}, P_{\bar{\theta}(x,\eta_0)}) \right\} \right) \\
&\qquad\qquad\qquad\qquad = P_X \left(\left\{ x \in \mathbb{X} : \bar{\theta}(x, \eta) \neq \bar{\theta}(x, \eta_0) \right\} \right) > 0,
\end{aligned}
$$

[1] Or uniformly integrable in the case of weak consistency.

from which

$$
\begin{aligned}
M(\eta) &= \int_{\mathbb{X}} S(P_{\bar\theta(x,\eta)}, P_{\bar\theta(x,\eta_0)})\, P_X(\mathrm{d}x) \\
&< \int_{\mathbb{X}} S(P_{\bar\theta(x,\eta_0)}, P_{\bar\theta(x,\eta_0)})\, P_X(\mathrm{d}x) = M(\eta_0).
\end{aligned}
$$

Finally, Assumption 4 follows from Assumption 9 upon setting $Z_{n,i} = (x_{n,i}, Y_{n,i})$:

(i) Since $Y_{n,i} \sim P_{\bar\theta(x_{n,i},\eta_0)}$, the family $\{m_{\eta_0}(Z_{n,i}) = S(P_{\bar\theta(x_{n,i},\eta_0)}, Y_{n,i})\}_{n\geq 1, i\in[n]}$ is contained in the family $\{S(P_\theta, Y) : \theta \in \Theta, Y \sim P_\theta\}$. The L^2-stochastic dominance of the latter in Assumption 9(i) then implies the same property for the former, yielding Assumption 4(i).

(ii) Let $\eta \in H \setminus \{\eta_0\}$ and let $\delta > 0$ be such that the requirement in Assumption 9(ii) holds. Let $\mathscr{S}_{\eta,\delta}$ be the L^2-stochastically dominated family in Eq. (4.4). Let W be a nonnegative, square-integrable random variable whose survival function dominates that of all random variables in $\mathscr{S}_{\eta,\delta}$. For all $x \in \mathbb{X}$, we have

$$
\int_{\mathbb{Y}} \sup_{\bar\eta \in \overline{B}(\eta,\delta)} \big(S(P_{\bar\theta(x,\bar\eta)}, y)\big)^+ \, P_{\bar\theta(x,\eta_0)}(\mathrm{d}y) \leq \mathsf{E}[W] < \infty,
$$

and thus, by definition of $P_Z = P_{X,Y}$ in Eq. (4.2),

$$
P_Z\big(m_{\overline{B}(\eta,\delta)}\big)^+ = \int_{\mathbb{X}} \int_{\mathbb{Y}} \sup_{\bar\eta \in \overline{B}(\eta,\delta)} \big(S(P_{\bar\theta(x,\bar\eta)}, y)\big)^+ \, P_{\bar\theta(x,\eta_0)}(\mathrm{d}y)\, P_X(\mathrm{d}x) < \infty.
$$

It follows that $m_{\overline{B}(\eta,\delta)}$ is P_Z-quasi-integrable and that $P_Z m_{\overline{B}(\eta,\delta)} < +\infty$. Further, since $Y_{n,i} \sim P_{\bar\theta(x_{n,i},\eta_0)}$, the family

$$
\left\{ \Big(m_{\overline{B}(\eta,\delta)}(Z_{n,i})\Big)^+ = \sup_{\bar\eta \in \overline{B}(\eta,\delta)} \Big(S(P_{\bar\theta(x_{n,i},\bar\eta)}, Y_{n,i})\Big)^+ : n \geq 1, i \in [n] \right\}
$$

is contained in $\mathscr{S}_{\eta,\delta}$ and is therefore L^2-stochastically dominated too. □

Proof of Lemma 4.2. We apply Lemma 4.1 and need to check Assumptions 8 and 9. Regarding Assumption 8 recall that, as noted above, ES_β is a strictly proper scoring rule on $\mathscr{P}_\beta(\mathbb{R}^d)$. Next, writing $m := \sup_{\theta\in\Theta} \int_{\mathbb{R}^d} \|y\|^\beta \, P_\theta(\mathrm{d}y)$, which is finite by Eq. (4.5), we have

$$
\mathrm{ES}_\beta(P_\theta, y) \leq \frac{1}{2}\, \mathsf{E}_{P_\theta \otimes P_\theta}[\|Y - Y'\|^\beta] \leq 2^{\beta-1}\, \mathsf{E}_{P_\theta}[\|Y\|^\beta] = 2^{\beta-1} m
$$

for all $\theta \in \Theta$ and $y \in \mathbb{R}^d$. As a consequence, $x \mapsto \mathrm{ES}_\beta(P_{\bar\theta(x,\eta)}, P_{\bar\theta(x,\eta_0)})$ is quasi-integrable with respect to P_X, for every $\eta \in H$. Moreover, the map $(\theta, y) \mapsto \mathrm{ES}_\beta(P_\theta, y)$ is continuous by the uniform integrability of $y \mapsto \|y\|^\theta$ with respect to $(P_\theta : \theta \in \Theta)$. Since $(x, \eta) \mapsto \bar\theta(x, \eta)$ is continuous, the map $(x, y) \mapsto \mathrm{ES}_\beta(P_{\bar\theta(x,\eta_0)}, y)$ is then continuous too. We have hence shown Assumption 8. Finally, Assumption 9 follows from the Markov inequality and the assumption in Eq. (4.5). □

Proof of Lemma 4.3. We apply Lemma 4.1, and need to check Assumptions 7, 8, and 9 for $S(P, y) = \mathrm{LogS}(P, y)$ from Eq. (4.6). Assumption 7 is part of the conditions in Assumption 10. Next, regarding Assumption 8, the fact that S is strictly proper follows from Gneiting and Raftery (2007, Section 4.1). We have $S(P_\theta, y) = \log p_\theta(y) < \infty$ by Assumption 10(i), and the two continuity requirements follow from (i) and (ii). Finally, Assumption 9(i) and (ii) exactly correspond to Assumption 10(iii) and (iv), respectively. □

Proof of Lemma 4.4. The map $(\theta, y) \mapsto p_\theta(y) \in [0, \infty)$ is continuous upon imposing the restriction $\xi > -1$, i.e., on the domain $(\mathbb{R} \times (0, \infty) \times (-1, \infty)) \times \mathbb{R}$; a key point is that $-1 - 1/\xi > 0$ for $-1 < \xi < 0$. Hence, the same is true for $(\eta, x, y) \mapsto m_\eta(x, y) = \log p_{\bar\theta(x,\eta)}(y) \in [-\infty, \infty)$, by the continuity of the link function $\bar\theta$ and the assumption on its image Θ. This settles the continuity requirements in Assumption 10.

As $\bar\theta$ is continuous and $\mathbb{X}$ and H are compact, the image Θ is compact too. Let Θ_j be the projection of Θ onto its jth coordinate ($j = 1, 2, 3$). We have

$$\begin{aligned} \sup_{\eta \in H, x \in \mathbb{X}, y \in \mathbb{R}} p_{\bar\theta(x,\eta)}(y) = \sup_{\theta \in \Theta, y \in \mathbb{R}} p_\theta(y) &= \sup_{\theta \in \Theta, y \in \mathbb{R}} \frac{1}{\sigma} p_{0,1,\xi}\left(\frac{y-\mu}{\sigma}\right) \\ &\le \sup_{\sigma \in \Theta_2, \xi \in \Theta_3, z \in \mathbb{R}} \frac{1}{\sigma} p_{0,1,\xi}(z) < \infty, \end{aligned} \tag{6.3}$$

where we applied Proposition 1 in Dombry (2015) in the last step, upon noting that Θ_2 and Θ_3 are compact subsets of $(0, \infty)$ and $(-1, \infty)$, respectively. We find that $m_\eta(x, y) = S(P_{\bar\theta(x,\eta)}, y) = \log p_{\bar\theta(x,\eta)}(y)$ is bounded above uniformly in (η, x, y). This deals with the L^2-stochastic dominance of the positive part in Assumption 10(iv).

Regarding the L^2-stochastic dominance in Assumption 10(iii), it is sufficient to treat the negative part $(-\ell_\theta(y))^+$ of the loglikelihood and show that $\{(-\ell_\theta(Y))^+ : \theta \in \Theta, Y \sim p_\theta\}$ is L^2-stochastically dominated, since the positive part of the loglikelihood is already uniformly bounded above by Eq. (6.3). Let $\theta \in \Theta$ and $Y \sim p_\theta$. Write $Z = (Y - \mu)/\sigma$, with density $p_{0,1,\xi}$. Since $1 + \xi Z > 0$ a.s., it is sufficient to control $(-\ell_\theta(y))^+$ for (θ, y) such that $1 + \xi z > 0$ with $z = (y - \mu)/\sigma$, a restriction on y (or z) which we will assume from now on without further mentioning. Write

$$u(z) = u_\xi(z) = \exp\left(-\int_0^z \frac{1}{1+\xi t} \mathrm{d}t\right) = \begin{cases} (1+\xi z)^{-1/\xi} & \text{if } \xi \neq 0, \\ e^{-z} & \text{if } \xi = 0, \end{cases}$$

with values in $(0, \infty)$. As $\xi > -1$ for all $\theta \in \Theta$, we have

$$\begin{aligned} (-\ell_\theta(y))^+ &= (\log(\sigma) + u(z) - (\xi + 1)\log(u(z)))^+ \\ &\le (\log(\sigma))^+ + u(z) + (\xi + 1)\left(-\log(u(z))\right)^+ . \end{aligned}$$

Let $C_2 = C_2(\Theta) = \max_{\sigma \in \Theta_2} (\log \sigma)^+ \in [0, \infty)$ and $C_3 = C_3(\Theta) = \max_{\xi \in \Theta_3}(\xi + 1) \in (0, \infty)$, with Θ_j as above. It follows that

$$(-\ell_\theta(y))^+ \le C_2 + u(z) + C_3 \left(-\log(u(z))\right)^+ .$$

The cdf of p_θ is $F_\theta(y) = e^{-u(z)}$. Since $u(Z) = -\log F_\theta(Y)$ has a unit-exponential distribution, the family $\{(-\ell_\theta(Y))^+ : \theta \in \Theta, Y \sim p_\theta\}$ is L^2-stochastically dominated, as required. $\square$

Proof of Lemma 4.5. As for the GEV distribution, the restriction of the shape parameter to $\xi > -1$ ensures the continuity of the map $(\theta, y) \mapsto p_\theta(y)$ on $(-1,\infty) \times (0,\infty)$, which implies the continuity requirements in Assumption 10. The same restriction also implies the bound $p_\theta(y) \le a^{-1}$ (note that if $-1 < \xi < 0$, then $-1 - 1/\xi > 0$), so that on the compact subset $\Theta = \bar{\theta}(\mathbb{X} \times H)$ of the parameter space, the GP density is uniformly bounded above. As for the GEV distribution, this implies the L^2-stochastic dominance required in Assumption 10(iv).

To show the L^2-stochastic dominance in Assumption 10(iii), note that for $y > 0$ such that $1 + \xi y/a > 0$, we have

$$\ell_\theta(y) = \log p_\theta(y) = -\log a + (\xi + 1)\log(1 - F_\theta(y))$$

where F_θ is the cdf of p_θ, given by

$$F_\theta(y) = \int_0^y p_\theta(t)\,\mathrm{d}t = \begin{cases} 1 - (1 + \xi y/a)^{-1/\xi}, & \text{if } \xi \neq 0, \\ 1 - \exp(-y/a), & \text{if } \xi = 0. \end{cases}$$

If $Y \sim p_\theta$, the distribution of $-\log(1 - F_\theta(Y))$ is unit-exponential. The L^2-stochastic dominance of $\{\ell_\theta(Y) : \theta \in \Theta, Y \sim p_\theta\}$ then follows from the fact that Θ is a compact subset of $(0,\infty) \times (-1,\infty)$. $\square$

Proof of Lemma 4.8. We apply Lemma 3.2 to $Y_{n,i} = (M_{n,i} \vee 1)/a_{r_n}$ and $\mathbb{Y} = (0,\infty)$. The reason to take the maximum with 1 has to do with the L^2-stochastic dominance condition, see below. We have

$$\mathsf{P}(Y_{n,i} \le y) = \mathsf{P}(M_{n,i} \vee 1 \le a_{r_n} y) = \begin{cases} F_{x_{n,i}}^{r_n}(a_{r_n} y), & \text{if } a_{r_n} y \ge 1, \\ 0, & \text{otherwise.} \end{cases}$$

The sequence of Markov kernels is thus given by

$$Q_n(x, (0,y)) = F_x^{r_n}(a_{r_n} y)\, \mathbb{1}_{[1/a_{r_n},\infty)}(y), \qquad y \in (0,\infty), \tag{6.4}$$

and the limit kernel by

$$Q(x, (0,y)) = \Phi_{\alpha_0}(y/\sigma_{\beta_0}(x)).$$

The latter is continuous as a function of x, by continuity of the map $(x,\beta) \mapsto \sigma_\beta(x)$, while the continuous convergence of Q_n to Q follows from Eq. (4.10): for $y \in (0,\infty)$ and every sequence $x_n \to x$ in $\mathbb{X}$, we have

$$\begin{aligned} &\left| Q_n(x_n, (0,y)) - Q(x, (0,y)) \right| \\ &\le \left| F_{x_n}^{r_n}(a_{r_n} y)\, \mathbb{1}_{[1/a_{r_n},\infty)}(y) - \Phi_{\alpha_0}(y/\sigma_{\beta_0}(x_n)) \right| + \left| \Phi_{\alpha_0}(y/\sigma_{\beta_0}(x_n)) - \Phi_{\alpha_0}(y/\sigma_{\beta_0}(x)) \right|, \end{aligned}$$

and both terms on the right-hand side converge to zero, the former by the uniformity in $x \in \mathbb{X}$ in Eq. (4.10) and the fact that $a_r \to \infty$ as $r \to \infty$.

We abbreviate $r = r_n$ and write $\ell_{\tau,\alpha} = \log p_{\tau,\alpha}$. We introduce the scaling sequence a_r into the objective function in Eq. (4.11), writing it as

$$L_n(\alpha, \beta, \tau) = \frac{1}{n} \sum_{i=1}^{n} \ell_{\tau\sigma_\beta(x_{n,i}),\alpha}(M_{n,i}) = \frac{1}{n} \sum_{i=1}^{n} \ell_{(\tau/a_r)\sigma_\beta(x_{n,i}),\alpha}(M_{n,i}/a_r) - \log a_r.$$

Define $H := A \times B \times [\gamma_-, \gamma_+]$ and, for $(\alpha, \beta, \gamma) \in H$,

$$\tilde{L}_n(\alpha, \beta, \gamma) := \frac{1}{n} \sum_{i=1}^{n} \ell_{\gamma\sigma_\beta(x_{n,i}),\alpha}(Y_{n,i}).$$

Thanks to Lemma 6.8, the event that there exists (random) $n_0 \in \mathbb{N}$ such that $Y_{n,i} = M_{n,i}/a_r$ for all $n \geq n_0$ and $i \in [n]$ has probability one. Write $\tilde{\eta}_n = (\hat{\alpha}_n, \hat{\beta}_n, \hat{\gamma}_n)$ with $\hat{\gamma}_n = \hat{\tau}_n/a_r$, and note that $\tilde{\eta}_n$ takes its values in H. By the assumption on $\hat{\eta}_n$, we have, almost surely,

$$\begin{aligned}
\tilde{L}_n(\tilde{\eta}_n) = \frac{1}{n} \sum_{i=1}^{n} \ell_{(\hat{\tau}_n/a_r)\sigma_{\hat{\beta}_n}(x_{n,i}),\hat{\alpha}_n}(Y_{n,i}) &= \frac{1}{n} \sum_{i=1}^{n} \ell_{(\hat{\tau}_n/a_r)\sigma_{\hat{\beta}_n}(x_{n,i}),\hat{\alpha}_n}(M_{n,i}/a_r) + \mathrm{o}(1) \\
&= L_n(\hat{\eta}_n) + \log a_r + \mathrm{o}(1) \\
&\geq L_n(\alpha_0, \beta_0, a_r) + \log a_r + \mathrm{o}(1) \\
&= \tilde{L}_n(\alpha_0, \beta_0, 1) + \mathrm{o}(1), \qquad n \to \infty.
\end{aligned}$$

The conclusion now follows if we can apply Lemma 3.2 with contrast function

$$\begin{aligned}
m_\eta(x, y) &= \ell_{\gamma\sigma_\beta(x),\alpha}(y) \\
&= \log \alpha - \alpha \log \gamma - \alpha \log \sigma_\beta(x) - (\alpha + 1) \log y - \left(\frac{y}{\gamma\sigma_\beta(x)} \right)^{-\alpha}.
\end{aligned}$$

with $\eta = (\alpha, \beta, \gamma) \in H$ and true parameter $\eta_0 = (\alpha_0, \beta_0, 1)$.

Assumption 6 has already been verified at the beginning of the proof.

The map $(\eta, x, y) \mapsto m_\eta(x, y)$ is continuous and bounded above, so m_η is $P_{X,Y}$-quasi-integrable, which is Assumption 2.

In Assumption 3, it only remains to show that $\eta_0 = (\alpha_0, \beta_0, 1)$ is the unique maximizer in H of the limit function

$$\tilde{L}(\eta) = P_{X,Y} m_\eta = \int_{\mathbb{X}} \int_0^\infty m_\eta(x, y) \, p_{\sigma_{\beta_0}(x),\alpha_0}(y) \, \mathrm{d}y \, P_X(\mathrm{d}x).$$

But this follows from the identifiablity of the Fréchet parameter (σ, α) together with the identifiability assumption in Eq. (4.12): for all $x \in \mathbb{X}$, we have

$$\int_0^\infty m_\eta(x, y) \, p_{\sigma_{\beta_0}(x),\alpha_0}(y) \, \mathrm{d}y \leq \int_0^\infty m_{\eta_0}(x, y) \, p_{\sigma_{\beta_0}(x),\alpha_0}(y) \, \mathrm{d}y$$

with strict inequality as soon as $\eta = (\alpha, \beta, \gamma)$ satisfies $(\gamma\sigma_\beta(x), \alpha) \neq (\sigma_{\beta_0}(x), \alpha_0)$ (van der Vaart, 1998, Lemma 5.35), which happens on a set of positive P_X-probability. Further, it is not hard to see that $\tilde{L}(\eta_0) > -\infty$, since if $Y \sim \Phi_\alpha$, then $Y^{-\alpha}$ has a unit-exponential distribution.

Finally, consider the L^2-stochastic dominance in Assumption 4. Part (ii) follows from the fact $(\eta, x, y) \mapsto m_\eta(x, y)$ is upper bounded, and part (i) follows from Lemma 6.9, observing that $Z_{n,i} = (x_{n,i}, Y_{n,i}) \sim Q_n(x, \mathrm{d}y)\delta_{x_{n,i}}(\mathrm{d}x)$ with Q_n from (6.4). □

Lemma 6.8. *Assume Eqs.* (4.7) *and* (4.8) *with* $\inf_{x\in\mathbb{X}} c(x) > 0$. *If* $\log n = o(r_n)$, *then for* $M_{n,i}$ *in Eq.* (4.9) *we have* $\min_{i\in[n]} M_{n,i} \to \infty$ *almost surely as* $n \to \infty$.

Proof of Lemma 6.8. Choose $y < \infty$ arbitrarily large. It is sufficient to show that the event $\{\min_{i\in[n]} M_{n,i} \le y \text{ infinitely often}\}$ has probability zero. Let $\varepsilon > 0$ be such that $c(x) \ge 2\varepsilon$ for all $x \in \mathbb{X}$. By enlarging y if necessary, we can ensure that

$$\sup_{x\in\mathbb{X}} \left| \frac{\log F_x(y)}{\log F_0(y)} - c(x) \right| \le \varepsilon$$

and thus

$$\forall x \in \mathbb{X}: \qquad F_x(y) \le F_0(y)^{c(x)-\varepsilon} \le F_0(y)^\varepsilon < 1,$$

as the upper endpoint of F_0 is infinity by Eq. (4.7). Write $\bar{p} = \sup_{x\in\mathbb{X}} F_x(y) < 1$ and let n be sufficiently large so that $\log n \le |\log \bar{p}|\, r_n/3$. Then

$$\begin{aligned}
\mathsf{P}\left(\min_{i\in[n]} M_{n,i} \le y\right) &\le \sum_{i=1}^n \mathsf{P}\left(M_{n,i} \le y\right) \\
&\le n \sup_{x\in\mathbb{X}} F_x^{r_n}(y) = n \exp\left(-r_n \left|\log \bar{p}\right|\right) \le n \exp\left(-3\log n\right) = n^{-2}.
\end{aligned}$$

The upper bound is summable in n, so that the Borel–Cantelli permits to conclude. □

Lemma 6.9. *Under the conditions of Lemma 4.8, there exists* $n_0 \in \mathbb{N}$ *such that the family*

$$\left\{m_{\eta_0}(x, Y) : Y \sim Q_n(x, \,\cdot\,), x \in \mathbb{X}, n \ge n_0\right\},$$

with Q_n *as in Eq.* (6.4), *is* L^2*-stochastically dominated.*

Proof of Lemma 6.9. Note that

$$m_{\eta_0}(x, y) = \log \alpha_0 - \log \sigma_{\beta_0}(x) - (\alpha_0 + 1) \log\left(\frac{y}{\sigma_\beta(x)}\right) - \left(\frac{y}{\sigma_\beta(x)}\right)^{-\alpha_0}.$$

Since L^2-stochastic dominance is preserved under addition, it is sufficient find $n_0 = n_0(j)$ such that

$$D_j := \left\{f_j(x, Y) : Y \sim Q_n(x, \,\cdot\,), x \in \mathbb{X}, n \ge n_0\right\}, \qquad j \in \{1, 2, 3\},$$

is L^2-stochastically dominated, where

$$f_1(x,y) = \log_+(y/\sigma_{\beta_0}(x)), \quad f_2(x,y) = \log_-(y/\sigma_{\beta_0}(x)), \quad f_3(x,y) = (y/\sigma_{\beta_0}(x))^{-\alpha_0},$$

and where $\log_+(x) = \max\{\log(x), 0\}$ and $\log_-(x) = \max\{-\log(x), 0\}$. For that purpose, writing $r = r_n$ and recalling that the support of Q_n is $[1/a_r, \infty)$, we need to bound

$$t \mapsto \sup_{n \ge n_0} \sup_{x \in \mathbb{X}} S_{nj}(x,t) \quad \text{with} \quad S_{nj}(x,t) := Q_n\Big(x, \big\{y \ge 1/a_r : f_j(x,y) > t\big\}\Big)$$

by a survival function associated with a random variable that is square-integrable.

We start by the hardest case, $j = 3$. In view of the fact that $\sigma_+ = \sup_{x \in \mathbb{X}} \sigma_{\beta_0}(x) \in (0, \infty)$, we have

$$\begin{aligned} S_{n3}(x,t) &= Q_n\Big(x, \big\{1/a_r \le y < \sigma_{\beta_0}(x) t^{-1/\alpha_0}\big\}\Big) \\ &\le Q_n\Big(x, \big\{1/a_r \le y < \sigma_+ t^{-1/\alpha_0}\big\}\Big) =: S^+_{n3}(x,t). \end{aligned}$$

For some constant $c > 1$ that will be specified below, write $b_r^- = (\tilde{a}_r/c)^{\alpha_0}$ and $b_r^+ = \tilde{a}_r^{\alpha_0}$, where $\tilde{a}_r = a_r \sigma_+$. We will distinguish three cases according to whether (n, x, t) satisfies $t \in [1, b_r^-]$, $t \in (b_r^-, b_r^+]$ or $t > b_r^+$.

In the last case, $t > b_r^+$, we have $1/a_r > \sigma_+ t^{-1/\alpha_0}$, which immediately implies $S^+_{n3}(x,t) = 0$. For all $1 \le t \le b_r^+$ we have $1/a_r \le \sigma_+ t^{-1/\alpha_0}$, so that we may write

$$\begin{aligned} S^+_{n3}(x,t) &= F_x^r(\tilde{a}_r t^{-1/\alpha_0}) \\ &= \exp\Big(- \big\{ - r \log F_0(\tilde{a}_r)\big\} \cdot \frac{\log F_0(\tilde{a}_r t^{-1/\alpha_0})}{\log F_0(\tilde{a}_r)} \cdot \frac{\log F_x(\tilde{a}_r t^{-1/\alpha_0})}{\log F_0(\tilde{a}_r t^{-1/\alpha_0})}\Big). \end{aligned} \tag{6.5}$$

We next consider the case $t \in [1, b_r^-]$. First, Eq. (4.7) implies $-r \log F_0(\tilde{a}_r) = (\sigma_+)^{-\alpha_0} + \mathrm{o}(1)$, whence there exists $n_1 \in \mathbb{N}$ such that $-r \log F_0(\tilde{a}_r) \ge (\sigma_+)^{-\alpha_0}/2$ for all $n \ge n_1$. Next, in view of Eq. (4.8) there exists c (wlog larger than 1) such that

$$\sup_{x \in \mathbb{X}} \Big| \frac{\log F_x(z)}{\log F_0(z)} - (\sigma_{\beta_0}(x))^{\alpha_0} \Big| \le (\sigma_-)^{\alpha_0}/2 \qquad \forall z \ge c,$$

where $\sigma_- = \inf_{x \in \mathbb{X}} \sigma_{\beta_0}(x) \in (0, \infty)$. Since $t \le b_r^-$ is equivalent to $\tilde{a}_r t^{-1/\alpha_0} \ge c$, we obtain that

$$\frac{\log F_x(\tilde{a}_r t^{-1/\alpha_0})}{\log F_0(\tilde{a}_r t^{-1/\alpha_0})} \ge (\sigma_{\beta_0}(x))^{\alpha_0} - \Big| \frac{\log F_x(\tilde{a}_r t^{-1/\alpha_0})}{\log F_0(\tilde{a}_r t^{-1/\alpha_0})} - (\sigma_{\beta_0}(x))^{\alpha_0} \Big| \ge (\sigma_-)^{\alpha_0}/2$$

for all $x \in \mathbb{X}$ and all (n, t) such that $t \in [1, b_r^-]$. Finally, fix $\delta \in (0, 1 \wedge \alpha_0)$. After possibly increasing the constant c, the Potter bounds on the regularly varying function $-\log F_0$ (Bingham et al., 1987, Theorem 1.5.6) imply that, for all $u, v \ge c$,

$$\frac{-\log F_0(u)}{-\log F_0(v)} \le (1 + \delta)\Big(\frac{u}{v}\Big)^{-\alpha_0} \max\Big[\Big(\frac{u}{v}\Big)^{\delta}, \Big(\frac{u}{v}\Big)^{-\delta}\Big]. \tag{6.6}$$

Since $a_r \to \infty$, we may choose n_2 such that $\tilde{a}_r \geq c$ for all $n \geq n_2$. We may apply the previous inequality with $u = \tilde{a}_r$ and $v = \tilde{a}_r y$ for $y \in [c/\tilde{a}_r, 1]$ to deduce that

$$\frac{-\log F_0(\tilde{a}_r)}{-\log F_0(\tilde{a}_r y)} \leq (1+\delta) y^{\alpha_0 - \delta} \qquad \forall n \geq n_2, y \in [c/\tilde{a}_r, 1].$$

This in turn implies

$$\frac{-\log F_0(\tilde{a}_r y)}{-\log F_0(\tilde{a}_r)} \geq \frac{1}{1+\delta} y^{-\alpha_0 + \delta} \qquad \forall n \geq n_2, y \in [c/\tilde{a}_r, 1],$$

and after setting $y = t^{-1/\alpha_0}$ as required in (6.5) and recalling $b_r^- = (\tilde{a}_r/c)^{\alpha_0}$, we obtain that

$$\frac{-\log F_0(\tilde{a}_r t^{-1/\alpha_0})}{-\log F_0(\tilde{a}_r)} \geq \frac{1}{1+\delta} t^{1-\delta/\alpha_0} \qquad \forall n \geq n_2, t \in [1, b_r^-].$$

Overall, defining $n_0 = \max(n_1, n_2)$, we have shown that

$$S_{n3}^+(x,t) \leq \exp\big(-k_\delta t^{1-\delta/\alpha_0} \big) \qquad \forall x \in \mathbb{X}, n \geq n_0, t \in [1, b_r^-],$$

where $k_\delta = (\sigma_-/\sigma_+)^{\alpha_0}/\{4(1+\delta)\}$.

It remains to consider the case $t \in (b_r^-, b_r^+]$. Since $t \mapsto S_{n3}^+(x,t)$ is decreasing, the previous display implies that

$$\begin{aligned}
S_{n3}^+(x,t) \leq S_{n3}^+(x, b_r^-) &\leq \exp\Big(-k_\delta (b_r^-)^{1-\delta/\alpha_0} \Big) \\
&= \exp\Big(-k_\delta \Big(\frac{b_r^-}{t}\Big)^{1-\delta/\alpha_0} t^{1-\delta/\alpha_0} \Big) \\
&\leq \exp\Big(-k_\delta \Big(\frac{b_r^-}{b_r^+}\Big)^{1-\delta/\alpha_0} t^{1-\delta/\alpha_0} \Big) \\
&= \exp\Big(-k_\delta \Big(\frac{1}{c}\Big)^{\alpha_0-\delta} t^{1-\delta/\alpha_0} \Big) \qquad \forall x \in \mathbb{X}, n \geq n_0, t \in (b_r^-, b_r^+].
\end{aligned}$$

This bound is also an upper bound for the bounds that have been derived for the cases $t \in [1, b_r^-]$ and $t > b_r^+$, whence we have overall shown that

$$S_{n3}^+(x,t) \leq \exp\Big(-k_\delta \Big(\frac{1}{c}\Big)^{\alpha_0-\delta} t^{1-\delta/\alpha_0} \Big) \qquad \forall x \in \mathbb{X}, n \geq n_0, t \geq 1.$$

The right-hand side defines a survival function associated with a random variable that is easily square integrable (recall that $\delta < \alpha_0$).

Next, consider the case $j = 2$. Since $\log_-(y) \leq y^{-1}$ for all $y > 0$, we have $S_{n2}(x,t) \leq S_{n3}(x, t^{\alpha_0})$. We may hence reuse the bound obtained for S_{n3}.

Finally, consider the case $j = 1$. Observing that $\log_+(y) > t$ is equivalent to $y > e^t$ (for $t \geq 0$), we have

$$S_{n1}(x,t) = Q_n\Big(x, \big\{y \geq 1/a_r : y > \sigma_{\beta_0}(x) e^t\big\}\Big) \leq Q_n\Big(x, \big\{y : y > \sigma_{\beta_0}(x) e^t\big\}\Big).$$

By definition of Q_n, we may rewrite the right-hand side of the previous display as

$$\mathsf{P}\left(\frac{\max(\xi_1^x,\ldots,\xi_r^x)\vee 1}{a_r} > \sigma_{\beta_0}(x)e^t\right),$$

where $\xi_1^x,\xi_2^x,\ldots$ are iid random variables with cdf F_x. Choose n_3 sufficiently large such that $1/a_r \le \sigma_-$ for all $n \ge n_3$. Then $1/a_r \le \sigma_{\beta_0}(x)e^t$ for all $t\ge 0, x\in\mathbb{X}$ and $n\ge n_3$, whence the previous probability is bounded by

$$\begin{aligned}\mathsf{P}\left(\frac{\max(\xi_1^x,\ldots,\xi_r^x)}{a_r} > \sigma_{\beta_0}(x)e^t\right) \le r\,\mathsf{P}\left(\xi_1^x > a_r\sigma_{\beta_0}(x)e^t\right) &\le r\,\mathsf{P}\left(\xi_1^x > a_r\sigma_-e^t\right)\\ &= r\{1-F_x(a_r\sigma_-e^t)\}\\ &\le -r\log F_x(a_r\sigma_-e^t),\end{aligned}$$

where the last inequality follows from $1-z\le -\log z$ for all $z\in[0,1]$. Next, rewrite the right-hand side of the previous display as

$$-r\log F_0(a_r\sigma_-)\cdot\frac{\log F_0(a_r\sigma_-e^t)}{\log F_0(a_r\sigma_-)}\cdot\frac{\log F_x(a_r\sigma_-e^t)}{\log F_0(a_r\sigma_-e^t)}.$$

Eq. (4.7) implies $-r\log F_0(a_r\sigma_-) = (\sigma_-)^{-\alpha_0}+\mathrm{o}(1)$, whence there exists $n_3\in\mathbb{N}$ such that $0\le -r\log F_0(a_r\sigma_-)\le 2(\sigma_-)^{-\alpha_0}$ for all $n\ge n_3$. Next, Eq. (6.6) applied with $u=a_r\sigma_-e^t$ and $v=a_r\sigma_-$ yields

$$\frac{\log F_0(a_r\sigma_-e^t)}{\log F_0(a_r\sigma_-)}\le(1+\delta)(e^t)^{-\alpha_0+\delta} = (1+\delta)e^{-t(\alpha_0-\delta)}$$

for all $n\ge n_4$ and $t\ge 0$; here, n_4 is chosen such that $a_r\sigma_-\ge c$ for all $n\ge n_4$. Finally, by Eq. (4.8), there exists $n_5\in\mathbb{N}$ such that for all $n\ge n_5$, all $x\in\mathbb{X}$ and all $t\ge 0$,

$$0\le\frac{\log F_x(a_r\sigma_-e^t)}{\log F_0(a_r\sigma_-e^t)}\le 2(\sigma_+)^{\alpha_0}.$$

Overall, we have shown that

$$S_{n1}(x,t)\le 4(1+\delta)(\sigma_+/\sigma_-)^{\alpha_0}e^{-t(\alpha_0-\delta)}\qquad\forall x\in\mathbb{X}, n\ge n_3\vee n_4\vee n_5, t\ge 0,$$

and since $\delta<\alpha_0$, the proof is finished. □

6.4 Proofs for Section 5

Proof of Lemma 5.1. Fix $\delta_0>0$. By continuity of measure, it is sufficient to show that, with probability one, '$d_H(\hat\eta_n,\eta_0)<\delta_0$ for all sufficiently large n'. Equivalently, defining $\Delta := \{\eta\in H\colon d_H(\eta,\eta_0)\ge\delta_0\}$, we need to show that '$\hat\eta_n\notin\Delta$ for all sufficiently large n' with probability one.

For that purpose, note that, for any $\eta\ne\eta_0$, we have

$$\lim_{\delta\downarrow 0} Pm_{B(\eta,\delta)} = M(\eta) < M(\eta_0) < \infty$$

by Lemma 6.10 and condition (A1), where Lemma 6.10 is applicable by (A2) and the first part of (A3). Hence, after possibly decreasing $\delta = \delta(\eta) > 0$ from (A3), we can assume that

$$Pm_{\overline{B}(\eta,\delta)} < M(\eta_0). \tag{6.7}$$

We may cover the set Δ by the collection of open balls $\{B(\eta, \delta(\eta)) : \eta \in \Delta\}$. Since Δ is compact by compactness of H, we can choose a finite subcover, say $\Delta \subseteq \bigcup_{i=1}^{K} B_i$ where $B_i = B(\eta_i, \delta(\eta_i))$.

Now, for each $i = 1, \ldots, K$,

$$\limsup_{n\to\infty} \mathbb{P}_n m_{B_i} \le \limsup_{n\to\infty} \mathbb{P}_n m_{\mathrm{cl}(B_i)} \le Pm_{\mathrm{cl}(B_i)} < M(\eta_0)$$

with probability one, where we used condition (A3) and Eq. (6.7) at the second and third inequality, respectively. As a consequence,

$$\limsup_{n\to\infty} \sup_{\eta\in\Delta} \mathbb{P}_n m_\eta \le \limsup_{n\to\infty} \max_{i=1,\ldots,K} \mathbb{P}_n m_{B_i} < M(\eta_0)$$

with probability one.

Also, by condition (A4) and since $M_n(\hat\eta_n) \ge M_n(\eta_0) - \mathrm{o}_{a.s.}(1)$, we further have

$$M(\eta_0) \le \liminf_{n\to\infty} M_n(\eta_0) \le \liminf_{n\to\infty} M_n(\hat\eta_n) = \liminf_{n\to\infty} \mathbb{P}_n m_{\hat\eta_n}$$

with probability one. Let Ω_0 denote the intersection of the two previous probability-one events, and define $N = \{\hat\eta_n \in \Delta \text{ infinitely often}\}$. The event $N \cap \Omega_0$ is then empty: indeed, on $N \cap \Omega_0$ we would have

$$\liminf_{n\to\infty} \mathbb{P}_n m_{\hat\eta_n} \le \limsup_{n\to\infty} \sup_{\eta\in\Delta} \mathbb{P}_n m_\eta < M(\eta_0) \le \liminf_{n\to\infty} \mathbb{P}_n m_{\hat\eta_n};$$

a contradiction. Hence, $\mathsf{P}(N) = \mathsf{P}(N \cap \Omega_0^c) \le \mathsf{P}(\Omega_0^c) = 0$, which was to be shown. $\square$

Proof of Lemma 5.2. The first two paragraphs of the proof are the same as those of the proof of Theorem 5.1. On the one hand, we have

$$M_n(\hat\eta_n) \ge M_n(\eta_0) - \mathsf{o}_\mathsf{P}(1) \ge M(\eta_0) - \mathsf{o}_\mathsf{P}(1), \qquad n \to \infty, \tag{6.8}$$

by the assumption on the estimator and by Assumption (A4'). On the other hand, with $\delta(\eta_i)$ and B_i for $i = 1, \ldots, K$ defined as in the proof of Lemma 5.1, we have

$$\mathbb{P}_n m_{B_i} \le \mathbb{P}_n m_{\mathrm{cl}(B_i)} \le Pm_{\mathrm{cl}(B_i)} + \mathsf{o}_\mathsf{P}(1) = Pm_{\overline{B}(\eta_i,\delta(\eta_i))} + \mathsf{o}_\mathsf{P}(1), \qquad n \to \infty,$$

by Assumption (A3') and thus

$$\begin{aligned}\sup_{\eta\in\Delta} M_n(\eta) &\le \max_{i=1,\ldots,K} \mathbb{P}_n m_{B_i} \\ &\le \max_{i=1,\ldots,K} Pm_{\overline{B}(\eta_i,\delta(\eta_i))} + \mathsf{o}_\mathsf{P}(1) \le M(\eta_0) - \varepsilon + \mathsf{o}_\mathsf{P}(1) \end{aligned} \tag{6.9}$$

for some $\varepsilon > 0$, since $Pm_{B(\eta_i,\delta(\eta_i))} < M(\eta_0)$ for all $i = 1, \ldots, K$. Together, Eqs. (6.8) and (6.9) imply that the event $\{\hat\eta_n \in \Delta\}$ is included in the event $\{M(\eta_0) - \mathsf{o}_\mathsf{P}(1) \le M(\eta_0) - \varepsilon + \mathsf{o}_\mathsf{P}(1)\}$, and the probability of that event goes to zero as $n \to \infty$. Hence, $\mathsf{P}(\hat\eta_n \in \Delta) \to 0$, as required. $\square$

Lemma 6.10 (Convergence of suprema over shrinking balls)**.** *Suppose that Assumption (A2) is met and let $\eta \in H$ be arbitrary. If $Pm_{\overline{B}(\eta,\delta)} < \infty$ for some $\delta > 0$, then*

$$\lim_{\delta \downarrow 0} Pm_{B(\eta,\delta)} = Pm_\eta.$$

Proof. Fix $\eta \in H$, and let $\delta_n \downarrow 0$ as $n \to \infty$. For $z \in \mathbb{E}$, define $f_n(z) = m_{B(\eta,\delta_n)}(z)$, and note that there exists n_0 such that $Pf_n \le Pf_{n_0} < \infty$ for all $n \ge n_0$ by monotonicity and the assumption that $Pm_{\overline{B}(\eta,\delta)} < \infty$ for some $\delta > 0$. If $Pf_{n_0} = -\infty$, there is nothing to show, so assume $Pf_{n_0} > -\infty$. By the monotone convergence theorem applied to the non-decreasing sequence $(f_{n_0} - f_n)\mathbb{1}_{\{|f_{n_0}|<+\infty\}}$,

$$\lim_{n\to\infty} Pf_n = P\Big(\lim_{n\to\infty} f_n\Big) = \int_{\mathbb{E}} \lim_{n\to\infty} \sup_{\eta' \in B(\eta,\delta_n)} m_{\eta'}(z)\, P(\mathrm{d}z).$$

By Assumption (A2), the integration domain on the right-hand side can be replaced by $\mathbb{E} \setminus V_\eta$, and on that domain, we have

$$\lim_{n\to\infty} \sup_{\eta' \in B(\eta,\delta_n)} m_{\eta'}(z) = m_\eta(z)$$

for all $z \in \mathbb{E} \setminus V_\eta$ by upper semicontinuity. This implies the assertion. $\square$

References

Aliprantis, C. D. and K. C. Border (2006). *Infinite dimensional analysis: a hitchhiker's guide* (Third ed.). Springer.

Berlinet, A., F. Liese, and I. Vajda (2000). Necessary and sufficient conditions for consistency of M-estimates in regression models with general errors. *J. Statist. Plann. Inference 89*(1-2), 243–267.

Bingham, N. H., C. M. Goldie, and J. L. Teugels (1987). *Regular variation.* Cambridge: Cambridge University Press.

Bücher, A. and J. Segers (2018). Maximum likelihood estimation for the Fréchet distribution based on block maxima extracted from a time series. *Bernoulli 24*(2), 1427–1462.

Chandra, T. K. and A. Goswami (1992). Cesàro uniform integrability and the strong law of large numbers. *Sankhyā Ser. A 54*(2), 215–231.

Dawid, A. P. (2007). The geometry of proper scoring rules. *Ann. Inst. Statist. Math. 59*(1), 77–93.

Dawid, A. P., M. Musio, and L. Ventura (2016). Minimum scoring rule inference. *Scand. J. Stat. 43*(1), 123–138.

Dombry, C. (2015). Existence and consistency of the maximum likelihood estimators for the extreme value index within the block maxima framework. *Bernoulli 21*(1), 420–436.

Dudley, R. M. (2002). *Real analysis and probability*, Volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge. Revised reprint of the 1989 original.

Einmahl, J. H. J., L. de Haan, and C. Zhou (2016). Statistics of heteroscedastic extremes. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 78*(1), 31–51.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc. 102*(477), 359–378.

Gut, A. (1992). The weak law of large numbers for arrays. *Statist. Probab. Lett. 14*(1), 49–52.

Hu, T. C., F. Móricz, and R. L. Taylor (1989). Strong laws of large numbers for arrays of rowwise independent random variables. *Acta Math. Hungar. 54*(1-2), 153–162.

Klein, N. (2024). Distributional regression for data analysis. *Annual Review of Statistics and Its Application 11*, 321–346.

Kneib, T., A. Silbersdorff, and B. Säfken (2023). "rage against the mean" – a review of distributional regression approaches. *Econometrics and Statistics 26*, 99–123.

Lachout, P., E. Liebscher, and S. Vogel (2005). Strong convergence of estimators as ϵ_n-minimisers of optimisation problems. *Ann. Inst. Statist. Math. 57*(2), 291–313.

Liese, F. and I. Vajda (1994). Consistency of m-estimates in general regression models. *Journal of Multivariate Analysis 50*(1), 93–114.

Padoan, S. A., M. Ribatet, and S. A. Sisson (2010). Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc. 105*(489), 263–277.

Philip, S., S. Kew, G. J. van Oldenborgh, F. Otto, R. Vautard, K. van der Wiel, A. King, F. Lott, J. Arrighi, R. Singh, and M. van Aalst (2020). A protocol for probabilistic extreme event attribution analyses. *Advances in Statistical Climatology, Meteorology and Oceanography 6*(2), 177–203.

Salibián-Barrera, M. (2006). The asymptotics of mm-estimators for linear regression with fixed designs. *Metrika 63*, 283–294.

Székely, G. J. (2003). E-statistics: The energy of statistical samples. Technical Report 2003-16, Bowling Green State University, Dept. of Mathematics and Statistics.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (2023). *Weak convergence and empirical processes—with applications to statistics* (Second ed.). Springer Series in Statistics. Springer, Cham.

Varadarajan, V. S. (1958a). On the convergence of sample probability distributions. *Sankhyā 19*, 23–26.

Varadarajan, V. S. (1958b). Weak convergence of measures on separable metric spaces. *Sankhyā 19*, 15–22.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statistics 20*, 595–601.