Biometrika (2022), **xx**, x, *pp*. 1–12
Printed in Great Britain

Uniqueness and global optimality of the maximum likelihood estimator for the generalized extreme value distribution

BY LIKUN ZHANG
Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, U.S.A.
likunz@lbl.gov

BENJAMIN A. SHABY
Department of Statistics, Colorado State University, Fort Collins, Colorado 80523, U.S.A.
bshaby@colostate.edu

SUMMARY

Some key words: Block maximum; Convergence rate; Global maximum; Law of large numbers; Profile likelihood; Support.

1. INTRODUCTION

Classical extreme value theory was introduced almost a century ago (Fisher & Tippett, 1928) and is in wide practical use, yet a basic theoretical elucidation of likelihood-based inference under its central distributional construct remains incomplete. Here, we fill in some important gaps. The generalized extreme value (GEV) distribution arises as the only limit of suitably renormalized maxima over independent and identically distributed random variables, and has therefore routinely been used in modelling the tail behaviour of observed phenomena. However, as the support of the density depends on its parameters, standard regularity conditions of classic asymptotic theory are not satisfied. It is only recently that consistency and asymptotic normality of the maximum likelihood estimator (MLE), found locally on a restricted compact set, have been established. In this paper, we show that the local MLE uniquely and globally maximizes the GEV log-likelihood function, provided that the shape parameter is between -1 and the number of samples. In addition, we establish a number of convergence properties related to the GEV, including uniform consistency of a class of limit relations, revealing a much richer understanding of the likelihood than has previously appeared.

The family of GEV distributions forms a continuous parametric family with respect to $\theta = (\tau, \mu, \xi)$ on some measurable space $(\mathcal{X}, \mathcal{A})$:

$$P_\theta(y) = \begin{cases} \exp\left[-\left\{1+\xi\left(\frac{y-\mu}{\tau}\right)\right\}^{-1/\xi}\right], & \xi \neq 0, \\ \exp\left\{-\exp\left(-\frac{y-\mu}{\tau}\right)\right\}, & \xi = 0, \end{cases}$$

where $1 + \xi(y-\mu)/\tau > 0$ for $\xi \neq 0$, and the scale parameter $\tau > 0$, location parameter $\mu \in \mathbb{R}$, and shape parameter $\xi \in \mathbb{R}$. The GEV distribution unites the Gumbel, Fréchet and Weibull distributions into a single family to allow various shapes.

The estimation of GEV parameters, especially the shape parameter ξ, is pivotal in studying tail behaviour. The Pickands (Pickands, 1975), probability weighted moments (Hosking et al., 1985) and method of moments quantile estimators (Madsen et al., 1997) are among many estimators available (Beirlant et al., 2004). In this paper, we focus on the asymptotic properties of maximum likelihood estimators. Let p_θ denote the density function of P_θ with respect to some dominating measure $\mathcal{P}$. Since the support of the GEV density function depends on θ, the regularity conditions for standard likelihood inference do not hold, which gives rise to fundamental difficulties when studying the existence, consistency and asymptotic normality of the MLE.

Suppose $\theta_0 = (\tau_0, \mu_0, \xi_0)$ is the true parameter, and $Y_1, \ldots, Y_n$ are independent observations from P_{θ_0}. Cohen (1986, 1988) assumed $\xi_0 = 0$ and considered samples drawn from the Gumbel distribution. He proved the consistency and asymptotic normality of the MLE based either on a fitted Gumbel distribution or on a fitted GEV distribution. The support of a Gumbel distribution is independent of its parameters, which makes it easier to examine the asymptotic behaviour of the MLE. Smith (1985) was the first to consider the MLE of a large class of irregular parametric families, and his formulation includes the GEV distribution when $-1 < \xi_0 < 0$. Treating the samples as coming from a distribution in the domain of attraction of a GEV, Dombry (2015) derived the existence of a *local* MLE, implicitly defined as a solution of the score function, under the setting of triangular arrays of block maxima when $\xi_0 > -1$. He proved that for any fixed compact set $K \subset \{\theta : \tau > 0, \mu \in \mathbb{R}, \xi > -1\}$ that contains θ_0, the maximum of the likelihood function in K is confined in an arbitrarily smaller neighbourhood $\tilde{K}$ of θ_0 for all n large enough. The corresponding local MLE

$$\hat{\theta}_n = \arg\max_{\theta \in K} L_n(\theta)$$

solves the score functions and converges almost surely to θ_0. We denote the entries of $\hat{\theta}_n$ as $(\hat{\tau}_n, \hat{\mu}_n, \hat{\xi}_n)$ throughout the remainder of the paper.

Bücher & Segers (2017) extended the result of Dombry (2015) in the simpler setting where $Y_1, \ldots, Y_n$ are independent observations from a GEV distribution, establishing a $O_p(n^{-1/2})$ rate of convergence for the local MLE, and refining the incomplete proof of Smith (1985) to establish the asymptotic normality of $\hat{\theta}_n$ for $\xi_0 > -1/2$ and a pre-specified set K. Subsequently, Dombry & Ferreira (2019) proved the asymptotic normality of the MLE using a different approach. Their results are again based upon local MLE for a likelihood function of block maxima that are approximately GEV distributed. Thus the limiting distribution has a non-trivial bias whose exact expression depends on the asymptotic growth of block size compared to the number of blocks.

However, the local MLE $\hat{\theta}_n$ studied by Dombry (2015), Bücher & Segers (2017) and Dombry & Ferreira (2019) may not attain a unique, global maximum of the log-likelihood

$$L_n(\theta) = \sum_{i=1}^{n} l_\theta(Y_i),$$

in which $l_\theta : \theta \mapsto \log p_\theta(y)$, and $\theta \in \Omega_n = \{\theta : p_\theta(Y_i) > 0, i = 1, \ldots, n\}$. Amongst other things, the uniform and global properties of L_n in Ω_n are needed in Bayesian theory to develop optimal decision rules and perform posterior-based inference (Hartigan, 1983), to establish asymptotic posterior normality (von Mises, 1931; Chen, 1985), and to construct rule-based noninformative priors (Bernardo, 2005).

In this paper, we consider $\theta_0 \in \Theta = (0, \infty) \times \mathbb{R} \times (-1/2, \infty)$. We will prove that the local MLE gives a unique, global maximum point for the log-likelihood function by following a two-step strategy:

(I) We first construct a small compact set $\tilde{K}$ containing θ_0 in its interior, and prove that for all large n, L_n in $\tilde{K}$ is strictly concave and attains a unique maximum;

(II) We then specify a larger compact set K, explicitly defined in terms of θ_0, such that $\tilde{K} \subset K$. We prove for all large n, the global maximum must be attained in K; that is, $\arg\max_{\theta \in \Theta} L_n(\theta) = \hat{\theta}_n$.

Specialising Proposition 2 in Dombry (2015) to the exact GEV setting, we have $\hat{\theta}_n \in \tilde{K}$ for all large n. One can therefore conclude that $L_n(\hat{\theta}_n)$ is indeed the unique and global maximum L_n; the global optimality is ensured by (II), while the uniqueness is ensured by (I). This main result is stated in the following theorem.

THEOREM 1 (GLOBAL OPTIMALITY AND UNIQUENESS). *Suppose $Y_1, Y_2, \ldots$ are independently sampled from P_{θ_0} and $\hat{\theta}_n$ is the sequence of local maxima of L_n that is found on a fixed compact neighbourhood of θ_0. Define $\Theta_n = \{\theta \in \Theta : -1/2 < \xi < n-1\}$. Then there almost surely exists $N > 0$ such that for all $n > N$, L_n is uniquely maximized in Θ_n and*

$$\arg\max_{\theta \in \Theta_n} L_n(\theta) = \hat{\theta}_n.$$

Remark 1. One may object that the optimality result is not truly global because of the restriction $\xi < n-1$. As the shape parameters are less than 1 for most observed data-generating processes, the ever-expanding Θ_n is hardly a restriction and does not interfere with the derivation of asymptotic posterior properties.

2. PRELIMINARIES

2.1. *The joint likelihood function and its support*

First we define the finite endpoint of the support when $\xi \neq 0$ as

$$\beta = \beta(\theta) = \mu - \frac{\tau}{\xi}. \quad (1)$$

This one-to-one mapping from (τ, μ, ξ) to (τ, β, ξ) will be used to simplify notation. In addition, define

$$W_i(\theta) = 1 + \xi\left(\frac{Y_i - \mu}{\tau}\right) = \frac{\xi}{\tau}(Y_i - \beta),$$

which helps simplify the log-likelihood function:

$$L_n(\theta) = -n\log\tau - \frac{\xi+1}{\xi}\sum_{i=1}^{n}\log W_i(\theta) - \sum_{i=1}^{n} W_i^{-1/\xi}(\theta) \quad (\xi \neq 0). \quad (2)$$

When $\xi \to 0$, $W_i^{-1/\xi}(\theta) \to \exp\{-(Y_i-\mu)/\tau\}$, so $L_n(\theta)$ with $\xi = 0$ is included in this formulation.

It can be easily verified that the domain of the log-likelihood function,

$$\Omega_n = \{\theta \in \Theta : \xi(Y_i - \beta) > 0, i = 1, \ldots, n\}, \quad (3)$$

is not a convex set, so Taylor expansion will not be helpful for studying $L_n(\theta)$. This precludes the use of routine tools such as the mean-value theorem and makes it difficult to approximate the difference of the function on a certain intervals. Nonetheless, if we slice Ω_n at different levels of ξ, every cross-section is convex; see Fig. 1 for illustration. On a cross-section at a fixed ξ, the

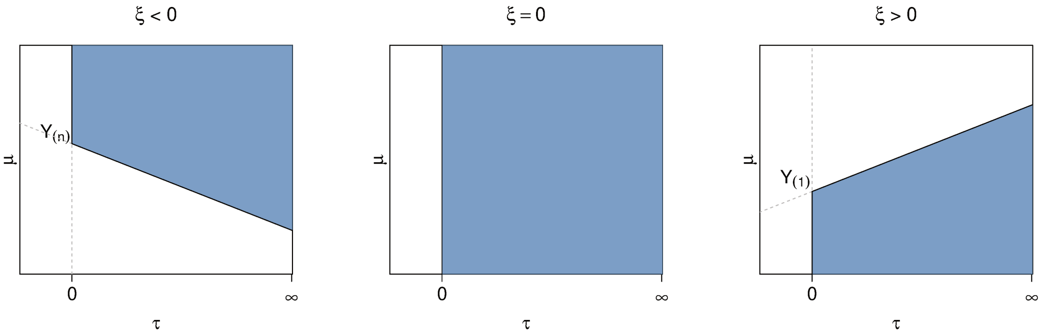

Fig. 1. Slicing the support Ω_n at different levels of $\xi \in (-1/2, \infty)$. A cross-section at any ξ, shown in the shaded area, is convex with respect to (τ, μ). When $\xi \neq 0$, the linear boundary of the cross-section has a slope of $1/\xi$.

value of $\beta = \mu - \tau/\xi$ can be construed as the intercept of the line which has a slope of $1/\xi$ and passes through (τ, μ). When $\xi > 0$, the condition in (3) imposes $\beta < Y_{(1)}$, and when $\xi < 0$, the intercept $\beta > Y_{(n)}$, where $Y_{(1)}$ and $Y_{(n)}$ are the sample minimum and maximum. Therefore, for any $\theta \in \Theta$, we can immediately tell whether $\theta \in \Omega_n$ using only $Y_{(1)}$ and $Y_{(n)}$.

2.2. *Profile likelihood*

Denote the cross-section of Ω_n at a certain ξ by $\Omega_n(\xi)$. The convexity of $\Omega_n(\xi)$ suggests examining the log-likelihood via profiling out (τ, μ):

$$\mathrm{PL}_n(\xi) = \sup_{(\tau,\mu)\in\Omega_n(\xi)} L_n(\theta).$$

The following proposition, whose proof can be found in the Supplementary Material, ensures that $L_n(\theta)$ is uniquely maximized on each cross-section $\Omega_n(\xi)$.

PROPOSITION 1. *Suppose $L_n(\theta)$ is applied to real numbers $y_1, \ldots, y_n$ that are not all equal. For $\xi \in [-1, n-1] \setminus \{0\}$, there exists a unique and global maximizer $(\tau_n(\xi), \mu_n(\xi))$ of L_n on the cross-section $\Omega_n(\xi)$, which can be found by solving*

$$\begin{cases} \tau = \left[n^{-1} \sum_{i=1}^n \{\xi(y_i - \beta)\}^{-1/\xi}\right]^{-\xi}, \\ (\xi + 1) \sum_{i=1}^n \{\xi(y_i - \beta)\}^{-1} = n \sum_{i=1}^n \{\xi(y_i - \beta)\}^{-1-1/\xi} \big/ \sum_{i=1}^n \{\xi(y_i - \beta)\}^{-1/\xi}. \end{cases} \tag{4}$$

For $\xi = 0$, the unique and global maximizer $(\tau_n(0), \mu_n(0))$ on $\Omega_n(0)$ can be found by solving

$$\begin{cases} n\tau = \sum_{i=1}^n \left\{1 - \exp\left(-\frac{y_i - \mu}{\tau}\right)\right\} y_i, \\ n = \sum_{i=1}^n \exp\left(-\frac{y_i - \mu}{\tau}\right). \end{cases}$$

For $\xi \notin [-1, n-1]$, $\mathrm{PL}_n(\xi) = \infty$. Meanwhile, $(\tau_n(\xi), \mu_n(\xi)) = (0, y_{(1)})$ when $\xi > n - 1$ and $(\tau_n(\xi), \mu_n(\xi)) = (0, y_{(n)})$ when $\xi < -1$, in which $y_{(1)}$ and $y_{(n)}$ denote the minimum and maximum values.

Remark 2. The system in Proposition 1 is defined in terms of (τ, β) for convenience, but its solution can be easily transformed into $(\tau_n(\xi), \mu_n(\xi))$ using (1).

Remark 3. By definition, $\mathrm{PL}_n(\xi) = L_n\{\tau_n(\xi), \mu_n(\xi), \xi\}$. Inserting (4) into (2),

$$\mathrm{PL}_n(\xi) = -n\log\left(\frac{1}{n}\sum_{i=1}^{n}[\xi\{y_i-\beta_n(\xi)\}]^{-1/\xi}\right) - \frac{\xi+1}{\xi}\sum_{i=1}^{n}\log[\xi\{y_i-\beta_n(\xi)\}] - n \quad (5)$$

when $\xi \in [-1, n-1] \setminus \{0\}$. By the continuity of L_n at $\xi = 0$, we know that

$$\lim_{\xi\to 0}\mu_n(\xi) = \mu_n(0), \quad \lim_{\xi\to 0}\tau_n(\xi) = \tau_n(0), \quad \lim_{\xi\to 0}\mathrm{PL}_n(\xi) = \mathrm{PL}_n(0).$$

To find the global maximum, we now need only compare the maxima from each cross-section. If the profile likelihood PL_n as a function of ξ were strictly concave in $[-1, n-1]$, it would have a unique maximum at ξ such that $\mathrm{PL}'_n(\xi) = 0$, and then $(\tau_n(\xi), \mu_n(\xi), \xi)$ would be the unique global maximizer for L_n. Unfortunately, PL_n is not a strictly concave function of ξ. The following proposition, whose proof can be found in the Supplementary Material, demonstrates that the first derivative PL'_n is not monotonically decreasing, and it behaves irregularly when ξ approaches the bounds of the interval $(-1, n-1)$.

PROPOSITION 2. *Under the assumptions of Proposition* 1*, the first derivative* PL'_n *is well-defined and continuous in* $\xi \in (-1, n-1)$*. When* $\xi \neq 0$*,* PL'_n *is taken using* (5)*:*

$$\begin{aligned}\mathrm{PL}'_n(\xi) = &-\frac{n}{\xi} - \frac{n\sum_{i=1}^{n}[\xi\{y_i-\beta_n(\xi)\}]^{-1/\xi}\log[\xi\{y_i-\beta_n(\xi)\}]}{\xi^2\sum_{i=1}^{n}[\xi\{y_i-\beta_n(\xi)\}]^{-1/\xi}}\\ &+\frac{1}{\xi^2}\sum_{i=1}^{n}\log[\xi\{y_i-\beta_n(\xi)\}].\end{aligned} \quad (6)$$

For $\xi = 0$*, the first derivative coincides with the limit:*

$$\lim_{\xi\to 0}\mathrm{PL}'_n(\xi) = \frac{n\mu'_n(0) - \sum_{i=1}^{n}\{y_i-\mu_n(0)+\tau'_n(0)\}}{\tau_n(0)} + \frac{\sum_{i=1}^{n}\{y_i-\mu_n(0)+\tau'_n(0)\}^2 - n\tau'_n(0)^2}{2\tau_n(0)^2}.$$

Additionally, $\mathrm{PL}'_n(\xi) \to \infty$ *when* $\xi \nearrow n-1$ *and* $\mathrm{PL}'_n(\xi) \to -\infty$ *when* $\xi \searrow -1$*. By the intermediate zero theorem, there must exist a* $\xi \in (-1, n-1)$ *such that* $\mathrm{PL}'_n(\xi) = 0$*.*

If a value of ξ satisfies $\mathrm{PL}'_n(\xi) = 0$, (4) and (6) together ensure that $(\tau_n(\xi), \mu_n(\xi), \xi)$ solves the score equations. Hence this result provides an alternative approach to proving the existence of the local MLE for L_n. However, proving the strong consistency of the local MLE requires n independently P_{θ_0}-distributed random variables.

Figure 2 illustrates some key features of the profile likelihood function. We simulate $Y_1, \ldots, Y_n$ from P_{θ_0} and calculate the log-likelihood PL_n at a grid of ξ values ranging from -1 to $n-1$. For all cases, including $\xi_0 = -0.2$, $\xi_0 = 0$ and $\xi_0 = 0.2$, PL_n appears to be uniquely maximized by the local MLE, which is close to ξ_0. Although it is not a concave function globally, we observe local concavity around ξ_0, which suggests adoption of the two-step strategy introduced in § 1. Roughly speaking, these two steps are established in § 4 via proving (I) PL_n is strictly concave in a small neighbourhood of $\hat{\xi}_n$ and (II) $\mathrm{PL}_n(\xi) < \mathrm{PL}_n(\hat{\xi}_n)$ for ξ far from $\hat{\xi}_n$.

3. CONVERGENCE RATE OF THE SUPPORT BOUNDARY

To prove (I) and (II), we will need to study the distance between the true parameter θ_0 and the boundary of the support Ω_n. It is true from the definition of Ω_n that if $Y_1, \ldots, Y_n$ are drawn from P_{θ_0}, then $\theta_0 \in \Omega_n$ for any $n \geq 1$. It is clear that Ω_n is an open set for any n, so the true parameter θ_0 is always interior to Ω_n. This raises the question: can we always find a neighbourhood of θ_0

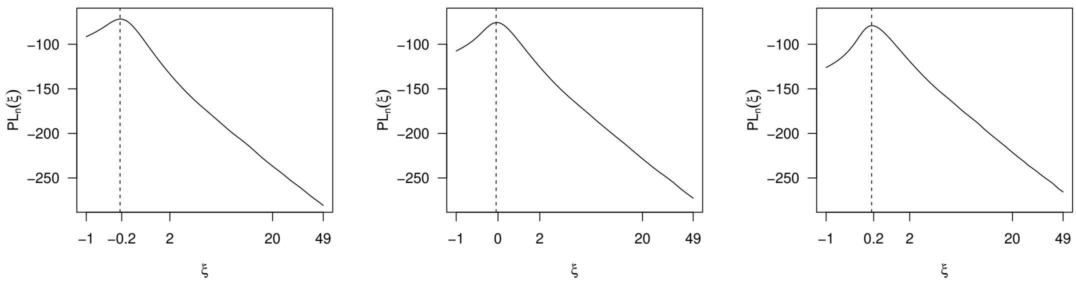

Fig. 2. $\mathrm{PL}_n(\xi)$ under $Y_1, \ldots, Y_n$ sampled from true $\xi_0 = -0.2$ (left), $\xi_0 = 0$ (middle) and $\xi_0 = 0.2$ (right), with dashed lines marking the local MLE $\hat{\xi}_n$. For all scenarios, $(\tau_0, \mu_0) = (1, 0)$ and sample size $n = 50$. We see that $\mathrm{PL}_n(\xi)$ is not concave.

which is contained by Ω_n that is large enough to allow us to examine the log-likelihood in the vicinity of θ_0? Unfortunately, this is not possible because θ_0 becomes arbitrarily close to the boundary as n approaches infinity when $\xi_0 \neq 0$.

To quantify the distance between θ_0 and the boundary of Ω_n, we first assume $\xi_0 > 0$ and examine the cross-section $\Omega_n(\xi_0)$. This is illustrated in Fig. 3, where $\theta_0 = (\tau_0, \mu_0, \xi_0)$ is shown as a red point, and $\beta_0 = \mu_0 - \tau_0/\xi_0$ is the intercept of the line that passes through (τ_0, μ_0) with a slope of $1/\xi_0$. Figure 3 illustrates that the difference of intercepts, $Y_{(1)} - \beta_0$, is a good measure of the distance. By analogy, if true shape parameter $\xi_0 < 0$, the distance can be well-measured by $\beta_0 - Y_{(n)}$.

Since the support of the distribution of P_{θ_0} is bounded below by β_0 when $\xi_0 > 0$, $\lim_{n\to\infty} Y_{(1)} = \beta_0$ almost surely. When $\xi_0 < 0$, the support of the distribution of P_{θ_0} is bounded above by β_0, so $\lim_{n\to\infty} Y_{(n)} = \beta_0$ almost surely. Thus in both cases, the distance between θ_0 and the boundary of Ω_n converges almost surely to zero. Also, Bücher & Segers (2017) showed that $\hat{\theta}_n - \theta_0 = O_p(n^{-1/2})$, so $\hat{\theta}_n$ is also arbitrarily close to θ_0 as n grows, and thus close to the boundary of Ω_n. This is concerning for the purpose of proving global optimality of $\hat{\theta}_n$ because it would be rather challenging to handle the log-likelihood near the boundary of the support.

Therefore, it is imperative that we compare the convergence rate of the distance between θ_0 and the boundary with $n^{-1/2}$ to get a clearer picture of $L_n(\theta)$ near the boundary.

PROPOSITION 3. *Suppose $Y_1, \ldots, Y_n$ are independently sampled from P_{θ_0} and $\epsilon > 0$ is an arbitrary constant.*

(A) If $\xi_0 > 0$, $Y_{(1)} \to \beta_0$ and $Y_{(n)} \to \infty$ almost surely. It also holds almost surely that

$$\lim_{n\to\infty} (\log n)^{(1+\epsilon)\xi_0}(Y_{(1)} - \beta_0) = \infty, \quad \lim_{n\to\infty} (\log n)^{(1-\epsilon)\xi_0}(Y_{(1)} - \beta_0) = 0,$$
$$\lim_{n\to\infty} n^{-(1+\epsilon)\xi_0} Y_{(n)} = 0, \quad \lim_{n\to\infty} n^{-(1-\epsilon)\xi_0} Y_{(n)} = \infty.$$

(B) If $\xi_0 < 0$, $Y_{(1)} \to -\infty$ and $Y_{(n)} \to \beta_0$ almost surely. It also holds almost surely that

$$\lim_{n\to\infty} (\log n)^{(1+\epsilon)\xi_0} Y_{(1)} = 0, \quad \lim_{n\to\infty} (\log n)^{(1-\epsilon)\xi_0} Y_{(1)} = -\infty,$$
$$\lim_{n\to\infty} n^{-(1+\epsilon)\xi_0}(\beta_0 - Y_{(n)}) = \infty, \quad \lim_{n\to\infty} n^{-(1-\epsilon)\xi_0}(\beta_0 - Y_{(n)}) = 0.$$

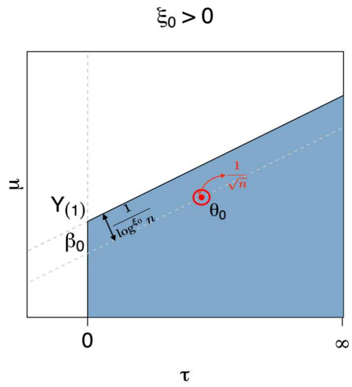

Fig. 3. The cross-section $\Omega_n(\xi_0)$ if true $\xi_0 > 0$. The two parallel dashed lines have a slope of $1/\xi_0$. The bullet point is $\theta_0 = (\tau_0, \mu_0, \xi_0)$. Here we also compare the convergence rates of $\hat{\theta}_n$ and $Y_{(1)}$, which are $n^{-1/2}$ and $1/\log^{\xi_0} n$. The red circle marks the neighbourhood of θ_0 with radius $n^{-1/2}$.

(C) If $\xi_0 = 0$, $Y_{(1)} \to -\infty$ and $Y_{(n)} \to \infty$ almost surely. It also holds almost surely that

$$\lim_{n\to\infty} (\log\log n)^{-1-\epsilon} Y_{(1)} = 0, \quad \lim_{n\to\infty} (\log\log n)^{-1+\epsilon} Y_{(1)} = -\infty,$$
$$\lim_{n\to\infty} (\log n)^{-1-\epsilon} Y_{(n)} = 0, \quad \lim_{n\to\infty} (\log n)^{-1+\epsilon} Y_{(n)} = \infty.$$

Remark 4. When $\xi_0 > 0$, it demonstrates that the convergence rate of $Y_{(1)}$ to β_0 is roughly $1/\log^{\xi_0} n$. The convergence rate of $\hat{\theta}_n$ to θ_0, $n^{-1/2}$, is much faster than the rate of $Y_{(1)}$ to β_0. These two rates are compared schematically in Fig. 3. If $\xi_0 < 0$, the convergence rate of $Y_{(n)}$ to β_0 is n^{ξ_0}, which is still slower than $n^{-1/2}$ because of the restriction $\xi_0 > -1/2$. Thus for a ball neighbourhood of $\hat{\theta}_n$ to be contained in Ω_n, its radius can be up to $1/n^{\epsilon}$ for some $\epsilon \in (0, 1/2)$. This is of vital importance in the proof of (I) and (II).

4. Proof of Theorem 1

4.1. *Smoothness of Hessian matrix*

When $\xi_0 \neq 0$, construct the compact set

$$\tilde{K} = \{\theta \in \Theta : |\tau - \tau_0| \leq r, |\beta - \beta_0| \leq r, |\xi - \xi_0| \leq r\},$$

where r is a small constant to be determined by θ_0 such that the log-likelihood function is locally concave in $\tilde{K}$. Slicing $\tilde{K}$ at different levels of ξ produces parallelograms; see Fig. 4. When $\xi_0 = 0$, $\tilde{K}$ is defined using $|\mu - \mu_0| \leq r$ instead of $|\beta - \beta_0| \leq r$. In this section, we will prove that for all large n, the Hessian matrix of L_n is negative definite in $\tilde{K} \cap \Omega_n$, and hence L_n is strictly concave.

Although the fixed larger compact set K is yet to be specified, we know from the strong consistency of the local MLE that $\hat{\theta}_n \in \tilde{K}$ for large sample size n. It is of interest to study $L_n''(\hat{\theta}_n)$, the Hessian at $\hat{\theta}_n$. The log-likelihood $L_n(\theta)$ in (2) and elements of its Hessian matrix

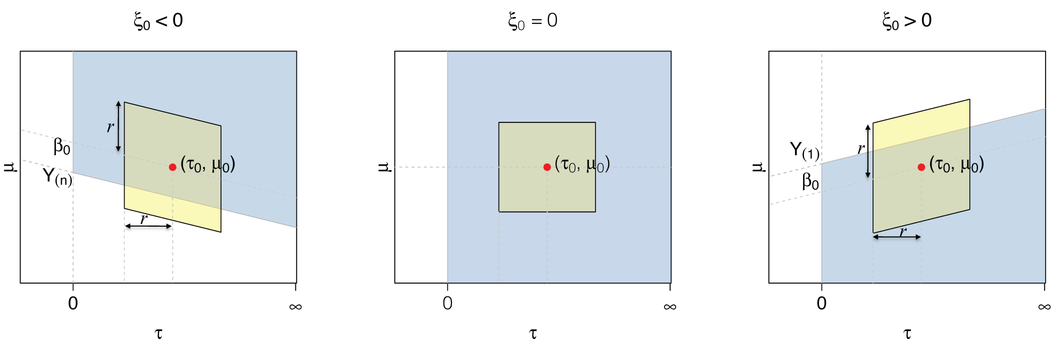

Fig. 4. Illustrating $\tilde{K}$ for $\xi_0 < 0$ (left), $\xi_0 = 0$ (middle) and $\xi_0 > 0$ (right). In all cases, the set $\tilde{K}$ sliced at $\xi = \xi_0$ is shown in yellow, with $\Omega_n(\xi_0)$ shown in blue. For $\xi_0 \neq 0$, the slice is a parallelogram when sliced at any ξ in $(\xi_0 - r, \xi_0 + r)$.

$L_n''(\theta)$ can all be written as linear combinations of sums of the form

$$\sum_{i=1}^{n} W_i^{-k-a/\xi}(\theta) \log^b W_i(\theta),$$

where $k, b = 0, 1, 2$, $a = 0, 1$; see the Supplementary Material for the expressions for the Hessian.

For constants k and a such that $k\xi_0 + a + 1 > 0$, it is straightforward to obtain

$$E_{\theta_0}\left\{W^{-k-a/\xi_0}(\theta_0) \log^b W(\theta_0)\right\} = (-\xi_0)^b \Gamma^{(b)}(k\xi_0 + a + 1),$$

where $W(\theta_0) = \xi_0(Y - \beta_0)/\tau_0$ with $Y \sim P_{\theta_0}$, and $\Gamma^{(b)}$ is the bth-order derivative of the Gamma function. Since $\{W_i^{-k-a/\xi}(\theta_0) \log^b W_i(\theta_0) : i = 1, 2, \ldots\}$ is an independent and identically distributed sequence, the strong law of large numbers gives

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} W_i^{-k-a/\xi_0}(\theta_0) \log^b W_i(\theta_0) = (-\xi_0)^b \Gamma^{(b)}(k\xi_0 + a + 1)$$

almost surely.

To examine $L_n''(\hat{\theta}_n)$, we replace θ_0 with $\hat{\theta}_n$ in the preceding averages. Since $\lim_{n\to\infty} \hat{\theta}_n = \theta_0$ almost surely, the continuity of the sums with respect to θ permits a pseudo large law of numbers for the elements in $L_n''(\hat{\theta}_n)$.

PROPOSITION 4. *Suppose $Y_1, Y_2, \ldots$ are independently sampled from P_{θ_0} and $\hat{\theta}_n$ is the local MLE of $L_n(\theta)$ that is strongly consistent. Then for constants k and a such that $k\xi_0 + a + 1 > 0$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} W_i^{-k-a/\hat{\xi}_n}(\hat{\theta}_n) \log^b W_i(\hat{\theta}_n) = (-\xi_0)^b \Gamma^{(b)}(k\xi_0 + a + 1) \tag{7}$$

almost surely, where b is a non-negative integer.

The proof of this result depends on Proposition 3. For details see the Supplementary Material. Proposition 4 ensures that $L_n''(\hat{\theta}_n)$ behaves like $L_n''(\theta_0)$ for large n. For the next result, we show that if we carefully select r for $\tilde{K}$ based on the value of θ_0, $L_n''(\theta)$ can be approximated by $L_n''(\hat{\theta}_n)$ in $\tilde{K} \cap \Omega_n$, yielding the negative-definiteness of $L_n''(\theta)$ in this neighbourhood.

PROPOSITION 5. *Let $Y_1, Y_2, \ldots$ be independently sampled from P_{θ_0} and let $\hat{\theta}_n$ be the local MLE of $L_n(\theta)$ that is strongly consistent. For a small $r > 0$ chosen by the rule specified in the Supplementary Material, there almost surely exists N such that, for any $n > N$ and $\theta \in \tilde{K} \cap \Omega_n$,*

$$I_3 - A_0(r) \leq L_n''(\theta)\{L_n''(\hat{\theta}_n)\}^{-1} \leq I_3 + A_0(r), \tag{8}$$

where I_3 is the 3×3 identity matrix and $A_0(r)$ is a 3×3 symmetric positive-semidefinite matrix whose elements only depend on θ_0 and the radius r, and whose largest eigenvalue tends to zero as $r \to 0$.

As a side result, we extend the limit relations in (7) to obtain uniform consistency as the powers of the W_i terms change in a closed interval. In Proposition 4, changing the power continuously produces a continuous path of the limit. If we fix the non-negative integer b and regard $\Phi_n(\alpha) = n^{-1} \sum_{i=1}^n W_i^{-\alpha}(\hat{\theta}_n) \log^b W_i(\hat{\theta}_n)$ as a stochastic process, $\Phi_n(\alpha)$ converges pointwise almost surely to $\Phi(\alpha) = (-\xi_0)^b \Gamma^{(b)}(\alpha\xi_0 + 1)$. The following result, which we prove in the Supplementary Material, says that the rate of convergence of sequences of $\Phi_n(\alpha)$ is essentially the same within a closed interval of α. That is, there is uniform consistency, which is a stronger property than stochastic equicontinuity. The uniformity will be crucial to proving step (II).

PROPOSITION 6 (UNIFORM CONSISTENCY). *Suppose $Y_1, Y_2, \ldots$ are independently sampled from P_{θ_0} and $\hat{\theta}_n$ is the local MLE of $L_n(\theta)$ that is strongly consistent. Let b be a non-negative integer and I be a closed interval on the real line such that $\alpha\xi_0 + 1 > 0$ for $\alpha \in I$. Write $\Phi_n(\alpha) = n^{-1} \sum_{i=1}^n W_i^{-\alpha}(\hat{\theta}_n) \log^b W_i(\hat{\theta}_n)$ and $\Phi(\alpha) = (-\xi_0)^b \Gamma^{(b)}(\alpha\xi_0 + 1)$. Then*

$$\lim_{n \to \infty} \sup_{\alpha \in I} |\Phi_n(\alpha) - \Phi(\alpha)| \to 0$$

almost surely.

4.2. *Step (I) and its proof*

PROPOSITION 7 (STEP (I)). *Let $Y_1, Y_2, \ldots$ be independently sampled from P_{θ_0} and let $\hat{\theta}_n$ be the local MLE of $L_n(\theta)$ that is strongly consistent. Then we can find some $r > 0$ small enough such that $L_n(\theta)$ is a strictly concave function in $\tilde{K} \cap \Omega_n$. Namely, there almost surely exists $N > 0$ such that for all $n > N$,*

$$L_n''(\theta) < 0 \quad (\theta \in \tilde{K} \cap \Omega_n).$$

Therefore, $\hat{\theta}_n$ is an unique maximum point in $\tilde{K}$.

Proof. Proposition 4 ensures that

$$\lim_{n \to \infty} \frac{1}{n} L_n''(\hat{\theta}_n) = -I(\theta_0)$$

almost surely, where $I(\theta_0)$ is the Fisher information of P_{θ_0}, and we know $|I(\theta_0)| > 0$ for all $\xi_0 > -1/2$. Therefore, $I(\theta_0)$ is positive definite, and there almost surely exists $N > 0$ such that for all $n > N$, $L_n''(\hat{\theta}_n) < 0$.

By Proposition 5, $A_0(r)$ only depends on θ_0 and r. We now fix r small enough such that the smallest eigenvalue of $I_3 - A_0(r)$ is positive. By (8),

$$L_n''(\theta) \leq L_n''(\hat{\theta}_n)\{I_3 - A_0(r)\} < 0.$$

The choice of r only depends on θ_0. □

4.3. *Step (II) and its proof*

Step (II) confines the global MLE to a fixed compact set K which is constructed using the values of θ_0 such that $\tilde{K} \subset K$. Since $\hat{\theta}_n = \arg\max_{\theta \in K} L_n(\theta)$ by definition, we can deduce the global optimality of $\hat{\theta}_n$.

PROPOSITION 8 (STEP (II)). *Let $Y_1, Y_2, \ldots$ be independently sampled from P_{θ_0} and $(\mu_n(\xi), \tau_n(\xi))$ be the maximizer of L_n on the cross-section $\Omega_n(\xi)$. Then for large n, the global maximum must be in a cube K whose vertices are only dependent on the value of θ_0; that is, there almost surely exists $N > 0$ such that for all $n > N$,*

$$\arg\max_{\theta \in \Theta_n} L_n(\theta) \in K.$$

Proof. We detail the construction of the cube K in the Supplementary Material. Denote the range of ξ in K by J. Then

$$J = \begin{cases} [0, C_0\xi_0], & \xi_0 > 0, \\ [C_1\xi_0, 0], & \xi_0 < 0, \\ [-C_2/\log n, C_2/\log\log n], & \xi_0 = 0, \end{cases} \tag{9}$$

in which $C_2 = \exp(\gamma)$, where γ is the Euler-Mascheroni constant, and $C_0, C_1 > 1$ are fixed constants such that $(1/x - 1)\log\tau_0 + \xi_0 \log\Gamma(1/x) > 0$ when $x > C_0$, and $-\log x + \gamma + 0.1 < 0$ when $x > C_1$.

Utilising Proposition 1 and 2 from § 2, we show in the Supplementary Material that

$$\mathrm{PL}_n(\xi) < \mathrm{PL}_n(\hat{\xi}_n) \quad (\xi \notin J) \tag{10}$$

and

$$(\mu_n(\xi), \tau_n(\xi), \xi) \in K \quad (\xi \in J). \tag{11}$$

By (9), ξ_0 is in the interior of J. Since $\hat{\xi}_n$ converges almost surely to ξ_0, we have $\hat{\xi}_n \in J$ for sufficiently large n. Denote $K_1 = \{\theta \in \Theta : \xi \in J\}$. Clearly, $K \subset K_1$ and (10) implies $\arg\max_{\theta \in \Theta_n} L_n(\theta) \in K_1$. When $\xi \in J$, (11) encloses the unique maximizer $(\mu_n(\xi), \tau_n(\xi))$ on $\Omega_n(\xi)$ in K. Equivalently, $\arg\max_{\theta \in K_1} L_n(\theta) \in K$. Combining (10) and (11), $\arg\max_{\theta \in \Theta_n} L_n(\theta) \in K$. □

4.4. *Completing the Proof of Theorem* 1

Proposition 2 in Dombry (2015) ascertained that for all large n, the argmax point on the set K defined in Proposition 8 is confined in any smaller neighbourhood $\tilde{K}$. Although his result was developed within the framework of triangular arrays of block maxima, the proof can be adapted to work on independent and identically distributed GEV samples.

LEMMA 1 (CONSISTENCY). *Let $K \subset \Theta$ be a compact set that contains θ_0 as an interior point and $Y_1, Y_2, \ldots$ be a sequence of independent and identically distributed random variables with common distribution P_{θ_0}. Then a sequence of estimators $\hat{\theta}_n$ can be found to maximize the log-likelihood L_n over K. For any smaller neighbourhood $\tilde{K}$ of θ_0 such that $\tilde{K} \subset K$, we have $\hat{\theta}_n \in \tilde{K}$ almost surely. Hence $\hat{\theta}_n \to \theta_0$ almost surely as $n \to \infty$.*

Proof. Bücher & Segers (2017) noted that Proposition 2 in Dombry (2015) is applicable for the GEV distributions. Noticing that a GEV distribution is in its own domain of attraction, the block size sequence $m(n)$ is set to be 1 with $a_m = \tau_0$ and $b_m = \mu_0$.

Following the proof in Dombry (2015), $\tilde{K}$ is limited to be a ball neighbourhood of θ_0 with an arbitrarily small radius. It is straightforward to generalize the proof to any small neighbourhood of θ_0 such that $\tilde{K} \subset K$. Because the closure of the set $\Delta = K \setminus \tilde{K}$ is compact, any open cover of Δ has a finite subcover, and the remaining proof applies without modification. □

Combining Proposition 8 and Lemma 1, we obtain

$$\arg\max_{\theta \in \Theta_n} L_n(\theta) \in \tilde{K} \cap \Omega_n,$$

and by the local strict concavity in $\tilde{K} \cap \Omega_n$ ensured by Proposition 7,

$$\hat{\theta}_n = \arg\max_{\tilde{K} \cap \Omega_n} L_n(\theta),$$

whence we conclude that $\hat{\theta}_n$ attains the unique and global maximum of L_n.

5. Discussion

Intermediate results necessary for the proofs of local strict concavity and boundedness of the global MLE unveiled additional characteristics of the GEV likelihood function that may be of independent interest. For example, the profile likelihood attains a unique maximum at each slice of the support, the convergence rate of the support boundary to the local MLE is slower than $n^{-1/2}$, and a class of averages that are the building blocks of the Hessian matrix converge to their limits uniformly. These results enhance our understanding of the GEV likelihood.

In applications, observations are never generated exactly from a GEV distribution; rather, they come from a distribution which we typically assume to be in the domain of attraction of a GEV. Dividing the observations into non-overlapping blocks, we make the approximating assumption that the maxima extracted from each block are GEV distributed. Thus, the asymptotic setup of Dombry (2015) and Dombry & Ferreira (2019) should be viewed as the more realistic, and our work offers theoretical foundations for maximum likelihood estimation using the GEV when the block size is large.

Finally, the number of block maxima in any observational record is limited. For future research, it is important to examine the minimum sample size required for the observations to manifest large-sample behaviour, as had been done for previous asymptotic results in extreme value statistics. Small-sample estimators for the GEV tend to be unstable, so taking advantage of the profile likelihood might provide an effective, and to our knowledge unexplored, approach to estimating the shape parameter. That is, one could first calculate the maximum likelihood on the cross-sections of the support at different levels of ξ, and then find the ξ that maximizes the profile likelihood; see Fig. 2. Doing so is asymptotically guaranteed to find the global MLE, and might improve numerical stability in small samples.

Acknowledgement

The authors thank the referees for carefully reading our manuscript and for giving many constructive comments which greatly improved the quality of the paper.

Supplementary material

The detailed proofs for the aforementioned propositions are shown in the Supplementary Material. There are additional technical results and figures included in this document.

References

Beirlant, J., Goegebeur, Y., Teugels, J. & Segers, J. (2004). *Statistics of Extremes*. John Wiley & Sons, Ltd., Chichester. With contributions from Daniel De Waal and Chris Ferro.

Bernardo, J. M. (2005). Reference analysis. In *Bayesian Thinking: Modeling and Computation*, D. Dey & C. Rao, eds., vol. 25 of *Handbook of Statistics*. Elsevier/North-Holland, Amsterdam, pp. 17–90.

Bücher, A. & Segers, J. (2017). On the maximum likelihood estimator for the generalized extreme-value distribution. *Extremes* **20**, 839–872.

Chen, C. F. (1985). On asymptotic normality of limiting density functions with Bayesian implications. *J. R. Statist. Soc. Ser. B* **47**, 540–546.

Cohen, J. P. (1986). Large sample theory for fitting an approximating Gumbel model to maxima. *Sankhyā Ser. A* **48**, 372–392.

Cohen, J. P. (1988). Fitting extreme value distributions to maxima. *Sankhyā Ser. A* **50**, 74–97.

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

Dombry, C. & Ferreira, A. (2019). Maximum likelihood estimators based on the block maxima method. *Bernoulli* **25**, 1690–1723.

Fisher, R. A. & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society* **24**, 180–190.

Hartigan, J. A. (1983). *Bayes Theory*. Springer, New York.

Hosking, J. R. M., Wallis, J. R. & Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics* **27**, 251–261.

Madsen, H., Rasmussen, P. F. & Rosbjerg, D. (1997). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 1. At-site modeling. *Water Resources Research* **33**, 747–757.

Pickands, III, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119–131.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72**, 67–90.

von Mises, R. (1931). *Wahrscheinlichkeitsrechnung*. Deuticke, Vienna.