J. Appl. Probab. **61,** 244–254 (2024)
*doi:*10.1017/jpr.2023.37

INFORMATION-THEORETIC CONVERGENCE OF EXTREME VALUES TO THE GUMBEL DISTRIBUTION

OLIVER JOHNSON ,* *University of Bristol*

Abstract

We show how convergence to the Gumbel distribution in an extreme value setting can be understood in an information-theoretic sense. We introduce a new type of score function which behaves well under the maximum operation, and which implies simple expressions for entropy and relative entropy. We show that, assuming certain properties of the von Mises representation, convergence to the Gumbel distribution can be proved in the strong sense of relative entropy.

Keywords: entropy; extreme value theory; information theory; von Mises representation

2020 Mathematics Subject Classification: Primary 60F99
Secondary 62B10; 94A15

1. Introduction and notation

It is well known that convergence to the Gaussian distribution in the central limit theorem regime can be understood in an information-theoretic sense, following the work in [6, 8, 27], and in particular [3], which proved convergence in relative entropy (see [15] for an overview of this work). While a traditional characteristic function proof of the central limit theorem may not give a particular insight into why the Gaussian is the limit, this information-theoretic argument (which can be understood to relate to Stein's method [28]) offers an insight on this.

To be specific, we can understand this convergence through the (Fisher) score function with respect to location parameter $\rho_X(x) = f_X'(x)/f_X(x) = (\log f_X(x))'$ of a random variable X with density f_X, where $'$ represents the spatial derivative. Two key observations are (i) that a standard Gaussian random variable Z is characterized by having linear score $\rho_Z(x) = -x$, and (ii) there is a closed-form expression for the score of the sum of independent random variables as a conditional expectation (projection) of the scores of the individual summands (see, e.g., [6]). As a result of this, the score function becomes 'more linear' in the central limit theorem regime (see [16, 17]). Similar arguments can be used to understand 'law of small numbers' convergence to the Poisson distribution [18].

However, there exist other kinds of probabilistic limit theorems which we would like to understand in a similar framework. In this paper we will consider a standard extreme value theory setup [25]: we take independent and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots \sim X$ and define $M_n = \max(X_1, \ldots, X_n)$ and $N_n = (M_n - b_n)/a_n$ for some normalizing sequences a_n and b_n. We want to consider whether N_n converges in relative entropy

Received 3 August 2022; revision received 21 April 2023.
* Postal address: School of Mathematics, University of Bristol, Fry Building, Woodland Road, Bristol, BS8 1UG, UK.
Email: O.Johnson@bristol.ac.uk

to a standard extreme value distribution. This type of extreme value analysis naturally arises in a variety of contexts, including the modelling of natural hazards, world record sporting performances, and applications in finance and insurance.

In this paper we show how to prove convergence in relative entropy for the case of a Gumbel (type I extreme value) limit, by introducing a different type of score function, which we refer to as the max-score Θ_X, and which is designed for this problem. Corresponding properties to those described above hold for this new quantity: (i) a Gumbel random variable X can be characterized by having linear max-score Θ_X (see Example 1.1), and (ii) there is a closed-form expression (Lemma 1.1) for the max-score of the maximum of independent random variables.

In Section 2 we show that the entropy and relative entropy can be expressed in terms of the max-score, in Section 3 we show how to calculate the expected value of the max-score in the maximum regime, and in Section 4 we relate this to the standard von Mises representation (see [25, Chapter 1]) to deduce convergence in relative entropy in Theorem 4.1

Our aim is not to provide a larger class of random variables than papers such as [11, 12, 21] for which convergence to the Gumbel takes place, but rather to use ideas from information theory to understand why this convergence may be seen as natural, and to prove convergence in a strong (relative entropy) sense. So while, for example, it is known that the standardized maximum converges in total variation (see, e.g., [24, p. 159]) convergence in relative entropy is stronger. This follows, for example, by Pinsker's inequality (see, e.g., [20]) for densities f and g,

$$\|f-g\|_{\mathrm{TV}}^2 \le 2D(f\,\|\,g) = 2\int f(x)\log\left(\frac{f(x)}{g(x)}\right)\mathrm{d}x,$$

where here and throughout we write log for the natural logarithm. Further, relative entropy (also known as Kullback–Leibler or KL divergence) is a valuable object of study in its own right, since the logarithm term can be thought of as log-likelihood for comparing two densities. As a result, D provides many fundamental limits in statistical estimation, classification, and hypothesis testing problems; see, for example, [10, Chapter 12] for a general survey, or [2] for an application in machine learning.

We briefly remark that entropy was studied in the Gumbel convergence regime in [23, 26] using direct computation based on the density. For example, [23, Theorems 2.1, 2.3] are proved by writing the entropy as an integral, decomposing that integral into regions, and using a variety of techniques to bound the resulting terms, using formulas arising from the density being in the domain of attraction, and dealing with tail terms appropriately. In contrast with their work, the aim of this paper is to give an elementary and direct proof under relatively simple conditions which hopefully gives some insight into why convergence to the Gumbel takes place, rather than to necessarily provide the strongest possible result.

The standard Fisher score was used in an extreme value context in [4] in a version of Stein's method. Extreme value distributions were considered in the context of Tsallis entropy in [5]. However, this particular max-score framework is new, to the best of our knowledge.

Definition 1.1. For an absolutely continuous random variable $Z \in \mathbb{R}$ we write the distribution function as $F_Z(z) = \mathbb{P}(Z \le z)$, the tail distribution function as $\overline{F}_Z(z) = \mathbb{P}(Z > z) = 1 - F_Z(z)$, the density as $f_Z(z) = F_Z'(z)$, and the hazard function as $h_Z(z) = f_Z(z)/\overline{F}_Z(z)$. We define the max-score function as

$$\Theta_Z(z) := \log(f_Z(z)/F_Z(z)) = \log h_Z(z) + \log(1 - F_Z(z)) - \log F_Z(z), \tag{1.1}$$

where the second result follows on rearranging.

https://doi.org/10.1017/jpr.2023.37 Published online by Cambridge University Press

We can express the max-score as $\Theta_Z(z) = \log(\tau_Z(z))$, where $\tau_Z(z) = f_Z(z)/F_Z(z) = (\mathrm{d}/\mathrm{d}z) \log F_Z(z)$ is the reversed hazard rate from [7]. Since $F_Z(\infty) = 1$, we can write

$$F_Z(x) = \exp\left(-\int_x^\infty \tau_Z(z)\,\mathrm{d}z\right) = \exp\left(-\int_x^\infty \mathrm{e}^{\Theta_Z(z)}\,\mathrm{d}z\right), \tag{1.2}$$

so the max-score function defines the distribution function.

We now remark that the Gumbel distribution has a linear max-score function.

Example 1.1. A Gumbel random variable Y with parameters μ and β has distribution function $F_Y(y) = \exp(-\mathrm{e}^{-(y-\mu)/\beta})$, so, in the notation above, $\tau_Y(y) = (\mathrm{d}/\mathrm{d}y) \log F_Y(y) = \exp(-(y-\mu)/\beta)/\beta$ and a Gumbel random variable has max-score

$$\Theta_Y(y) = \log \tau_Y(y) = -\log \beta - \frac{(y-\mu)}{\beta}.$$

Indeed, using (1.2), we can see the property of having a linear max-score Θ_Y characterizes the Gumbel.

For future reference in this paper, note that (see [19, (1.25)]) $\mathbb{E}Y = \mu + \beta\gamma$, where γ is the Euler–Mascheroni constant, and the moment-generating function is $\mathcal{M}_Y(t) = \mathrm{e}^{\mu t}\Gamma(1-\beta t)$ (see [19, (1.23)]).

We can further state how the max-score function behaves under the maximum and rescaling operations.

Lemma 1.1. *If we write $M_n = \max(X_1, \ldots, X_n)$ and $N_n = (M_n - b_n)/a_n$, then*

$$\Theta_{N_n}(z) = \log(na_n) + \Theta_X(a_n z + b_n), \tag{1.3}$$

$$\Theta_{N_n}(N_n) = \log(na_n) + \Theta_X(M_n). \tag{1.4}$$

Proof. As usual (see, e.g., [25, Section 0.3]), we know that, by independence,

$$F_{M_n}(x) = \mathbb{P}(\max(X_1, \ldots, X_n) \le x) = \mathbb{P}\left(\bigcap_{i=1}^n \{X_i \le x\}\right) = F_X(x)^n, \tag{1.5}$$

so that $F_{N_n}(x) = F_{M_n}(a_n x + b_n) = F_X(a_n x + b_n)^n$. This means that $f_{N_n}(x) = na_n F_X(a_n x + b_n)^{n-1} f_X(a_n x + b_n)$, so $f_{N_n}(x)/F_{N_n}(x) = na_n f_X(a_n x + b_n)/F_X(a_n x + b_n)$, and (1.3) follows on taking logarithms. The second result, (1.4), follows by direct substitution using the fact that $M_n = a_n N_n + b_n$. □

Example 1.2. In particular, if X is exponential with parameter λ, so, with $a_n = 1/\lambda$ and $b_n = \log n/\lambda$,

$$\Theta_X(z) = \log\left(\frac{\lambda \mathrm{e}^{-\lambda z}}{1 - \mathrm{e}^{-\lambda z}}\right),$$

this gives

$$\Theta_{N_n}(z) = \log(n/\lambda) + \log\left(\frac{\lambda \mathrm{e}^{-z}/n}{1 - \mathrm{e}^{-z}/n}\right) = -z - \log(1 - \mathrm{e}^{-z}/n).$$

Hence, letting $n \to \infty$, we know that $\Theta_{N_n}(z)$ converges pointwise to $-z$, which is the max-score of the standard Gumbel (with parameters $\mu = 0$ and $\beta = 1$); see Example 1.1.

However, while this gives us some intuition as to why the Gumbel is the limit in this case, pointwise convergence of the score function does not seem a particularly strong sense of convergence. We now discuss the question of convergence in relative entropy.

https://doi.org/10.1017/jpr.2023.37 Published online by Cambridge University Press

2. Max-score function and entropy

We next show that we can use the max-score function to give an alternative formulation for the entropy of a random variable, which allows us to quickly find the entropy of a Gumbel distribution. We first state a simple lemma, which follows directly from the fact that both $F_X(X)$ and $1 - F_X(X)$ are uniformly distributed.

Lemma 2.1. *For any continuous random variable X with distribution function F_X, $\mathbb{E}\log F_X(X) = \mathbb{E}\log(1 - F_X(X)) = -1$.*

Proposition 2.1. *For an absolutely continuous random variable X with max-score function Θ_X, the entropy $H(X)$ satisfies $H(X) = 1 - \mathbb{E}\Theta_X(X)$.*

Proof. The key observation is that $\log f_X(x) = \log F_X(x) + \Theta_X(x)$, so

$$\begin{aligned} H(X) &= -\int_{-\infty}^{\infty} f_X(x)\log f_X(x)\,\mathrm{d}x \\ &= -\int_{-\infty}^{\infty} f_X(x)\log F_X(x)\,\mathrm{d}x - \int_{-\infty}^{\infty} f_X(x)\Theta_X(x)\,\mathrm{d}x \qquad (2.1) \\ &= 1 - \mathbb{E}\Theta_X(X), \end{aligned}$$

where we apply Lemma 2.1 to find the first term of (2.1). □

In particular, we recover the entropy of Y, a Gumbel distribution (see, e.g., [22, Theorem 1.6(iii)]).

Example 2.1. For Y, a Gumbel distribution with parameters μ and β, using Example 1.1,

$$H(Y) = 1 - \mathbb{E}\left(-\log\beta - \frac{Y-\mu}{\beta}\right) = 1 + \log\beta + \gamma,$$

since $\mathbb{E}Y = \mu + \beta\gamma$.

We can use similar arguments to give an expression for the relative entropy $D(X \parallel Y)$ where Y is Gumbel.

Proposition 2.2. *Given absolutely continuous random variables X and Y, we can write the relative entropy from X to Y as*

$$D(X \parallel Y) = \mathbb{E}(\Theta_X(X) - \Theta_Y(X)) + \mathbb{E}(-\log F_Y(X) - 1). \qquad (2.2)$$

In particular, if Y is a Gumbel random variable with parameters μ and β, then

$$D(X \parallel Y) = \left(\mathbb{E}\Theta_X(X) + \log\beta + \frac{\mathbb{E}X - \mu}{\beta}\right) + (\mathbb{E}\mathrm{e}^{-(X-\mu)/\beta} - 1), \qquad (2.3)$$

assuming both sides of the expression are finite.

Proof. We can write $D(X \parallel Y)$ as

$$\begin{aligned} \int f_X(x)\log\left(\frac{f_X(x)}{f_Y(x)}\right)\mathrm{d}x &= -H(X) - \int f_X(x)\log f_Y(x)\,\mathrm{d}x \\ &= (\mathbb{E}\Theta_X(X) - 1) - \int f_X(x)\log F_Y(x)\,\mathrm{d}x - \int f_X(x)\Theta_Y(x)\,\mathrm{d}x, \end{aligned}$$

https://doi.org/10.1017/jpr.2023.37 Published online by Cambridge University Press

using Proposition 2.1, which implies (2.2). We deduce (2.3) using the values of F_Y and Θ_Y from Example 1.1. □

Observe that in the case of X itself Gumbel with the same parameters as Y, both bracketed terms in (2.3) vanish. We can rewrite the first term as $\mathbb{E}(\Theta_X(X) - \Theta_Y(X))$, using the value of the max-score in the Gumbel case (Example 1.1). This suggests that (as in [15]) we may wish to consider this term as a standardized score function with the relevant linear term subtracted off. We can rewrite the second term in (2.3) as $e^{\mu/\beta}\mathcal{M}_X(-1/\beta) - 1$, where $\mathcal{M}_X(t)$ is the moment-generating function. Since (see Example 1.1) the moment-generating function of a Gumbel random variable is $\mathcal{M}_Y(t) = e^{\mu t}\Gamma(1-\beta t)$, we know that in the Gumbel case $e^{\mu/\beta}\mathcal{M}_X(-1/\beta) - 1 = e^{\mu/\beta}e^{-\mu/\beta}\Gamma(2) - 1 = 0$.

3. Expected max-score of the standardized maximum

We now consider the behaviour of the expected max-score of the standardized maximum $N_n = (M_n - b_n)/a_n$, using the representation (1.1). We first state a technical lemma which holds for all continuous random variables X.

Lemma 3.1. *For M_n the maximum of n independent copies of absolutely continuous random variable X:*

(i) The expected value $\mathbb{E}F_X(M_n) = -1/n$ is the same for all F_X.

(ii) The expected value $\mathbb{E}\log(1 - F_X(M_n)) = -H_n$ is the same for all F_X, where we write $H_n := 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$ to be the nth harmonic number.

Proof. Part (i) is a simple corollary of Lemma 2.1 Recalling from (1.5) that $F_{M_n}(x) = F_X(x)^n$, we know from Lemma 2.1 that $-1 = \mathbb{E}\log F_{M_n}(M_n) = n\mathbb{E}\log F_X(M_n)$, and the result follows on rearranging.

Part (ii) requires a slightly more involved calculation. By standard manipulations, we know that $-\log(1 - F_X(X))$ is exponential with parameter 1. Now, since $-\log(1 - F_X(t))$ is increasing in t, we can write

$$\begin{aligned} -\log(1 - F_X(M_n)) &= -\log\Big(1 - F_X\Big(\max_{1\le i\le n} X_i\Big)\Big) \\ &= \max_{1\le i\le n} -\log(1 - F_X(X_i)) \sim \max_{1\le i\le n} E_i, \end{aligned}$$

where the E_i are independent exponentials with parameter 1. It is well known that the expected value of $\max_{1\le i\le n} E_i = H_n$, the nth harmonic number. The simplest proof of this is to write $\max_{1\le i\le n} E_i = \sum_{i=1}^n U_i$, where the U_i are independent exponentials with parameter $n - i + 1$. (This follows from the memoryless property of E_i, by thinking of U_1 as the time for the first exponential event to happen, U_2 as the time for the second, and so on.) Since $\mathbb{E}U_i = 1/(n - i + 1)$, the result follows. □

We can put this together to deduce the following result.

Lemma 3.2. *For any absolutely continuous X, writing $N_n = (M_n - b_n)/a_n$ for any sequence of norming constants a_n, b_n, we deduce that*

$$\mathbb{E}\Theta_{N_n}(N_n) = (\log a_n + \mathbb{E}\log h_X(M_n)) + (\log n - H_n) + \frac{1}{n}.$$

https://doi.org/10.1017/jpr.2023.37 Published online by Cambridge University Press

Proof. Using the representation in (1.4) of the score in Lemma 1.1 and the expression in (1.1), we know that

$$\begin{aligned}\mathbb{E}\Theta_{N_n}(N_n) &= \log(na_n) + \mathbb{E}\Theta_X(M_n)\\ &= \log(na_n) + \mathbb{E}\log h_X(M_n) + \mathbb{E}\log(1 - F_X(M_n)) - \mathbb{E}\log F_X(M_n)\\ &= (\log a_n + \mathbb{E}\log h_X(M_n)) + (\log n - H_n) + \frac{1}{n},\end{aligned}$$

using the two parts of Lemma 3.1. □

Note that only the first bracketed term of Lemma 3.2 depends on the particular choice of X.

4. Von Mises representation and convergence in relative entropy

We will demonstrate convergence of relative entropy in a restricted version of the domain of maximum attraction. In order to work in terms of relative entropy, we need to assume that X is absolutely continuous. Additionally, we recall the definition of a distribution function F_X having a representation of von Mises type [25, (1.5)].

Definition 4.1. Assume that the upper limit of the support of X is $x_0 := \sup\{x : F_X(x) < 1\}$ (which may be finite or infinite) and

$$F_X(x) = 1 - c(x)\exp\left(-\int_{z_0}^{x} \frac{1}{g(u)}\,\mathrm{d}u\right) = 1 - c(x)\exp(-G(x)), \tag{4.1}$$

for some auxiliary function g such that $g'(x) \to 0$ as $x \to x_0$, and $\lim_{x\to x_0} c(x) = c > 0$.

Assuming the von Mises representation (4.1) holds, we can write the density as $f_X(x) = (c(x)G'(x) - c'(x))\exp(-G(x))$, or on dividing by $1 - F_X(x) = c(x)\exp(-G(x))$ we can deduce that the hazard function satisfies

$$h_X(x) = \frac{1}{g(x)} - \frac{c'(x)}{c(x)}. \tag{4.2}$$

The canonical choice of norming constants is given in [25, Proposition 1.1(a)] (see also [13, Table 3.4.4]) as (a_n, b_n) satisfying $1/n = \overline{F}(b_n)$ and $a_n = g(b_n)$. Note that (see [25, Proposition 1.4]) the normalized maximum $N_n = (M_n - b_n)/a_n$ converges in distribution to the Gumbel if and only if the representation (4.1) holds. See [13, Table 3.4.4] for a list of eight types of distributions whose standardized maximum converges to the Gumbel, some of which we give as examples below.

Example 4.1. We can illustrate the representation in (4.1) as follows:

- For the exponential we can take $c(x) = 1$, $z_0 = 0$, $x_0 = \infty$, $g(u) = 1/\lambda$, and $b_n = \log n/\lambda$.
- For the gamma distribution with shape parameter α and rate parameter β we can take $c(x) = 1$, $z_0 = 0$, $x_0 = \infty$, and $g(u) = \Gamma(\alpha, \beta u)/(\beta(\beta u)^{\alpha-1}\exp(-\beta u))$, where $\Gamma(\cdot,\cdot)$ is the upper incomplete gamma function. Note that as $u \to \infty$ we know that $g(u) \to 1/\beta$.
- For the standard Gaussian distribution, we can take $g(x) = (1 - \Phi(x))/\phi(x)$ (for ϕ and Φ the standard normal density and distribution functions), and note that the Mills ratio $g(x) \simeq 1/x$ as $x \to \infty$.

https://doi.org/10.1017/jpr.2023.37 Published online by Cambridge University Press

- For the 'Weibull-like' distribution of [13, Table 3.4.4] with $\overline{F} \sim Kx^{\alpha} \exp(-cx^{\tau})$ (with $\tau > 0$ and $\alpha \in \mathbb{R}$), we can take $G(x) = cx^{\tau} - \alpha \log x$, so that $g(x) = x/(\tau cx^{\tau} - 1)$.
- For the Benktander type II distribution of [13, Table 3.4.4], with

$$\overline{F}(x) = x^{\beta-1} \exp(-\alpha(x^{\beta} - 1)/\beta),$$

  for $\alpha > 0$, $0 < \beta < 1$, we can take $c(x) = 1$, $z_0 = 1$, and $g(x) = x/(1 - \beta + \alpha x^{\beta})$.
- For the example of $F_X(x) = 1 - \exp(-x/(1-x))$ given in [14] (see also [25, p. 39]) we can take $c(x) = 1$, $z_0 = 0$, $x_0 = 1$, $g(u) = (1-u)^2$, and $b_n = \log n/(1 + \log n)$. This is an example of 'exponential behaviour at x_0' in the sense of [13, Table 3.4.4], where we can take $g(x) = (x_0 - x)^2/\alpha$.

We now state a restricted technical condition that we can use to give a simple proof of convergence in relative entropy.

Condition 4.1.

(i) *Assume* $\ell(t) := 1 - c'(t)g(t)/c(t) \to 1$ *as* $t \to x_0$.

(ii) *Assume there exists a constant* $\sigma < 1$ *such that* $\log(g(x)/x^{\sigma})$ *is bounded and continuous, and that* $\gamma := \lim_{x \to x_0} g(x)/x^{\sigma}$ *is finite and non-zero.*

(iii) *Assume that* $\int_{-\infty}^{0} |x|^k \, \mathrm{d}F_X(x) < \infty$ *for all* k.

Note that Condition 4.1 (i) holds automatically with equality when $c(x)$ is constant, which is the restricted version of the von Mises condition stated as [25, (1.3)], and which includes all but the Weibull-like part of Example 4.1. Note that Condition 4.1 (ii) is satisfied for the first five examples in Example 4.1 (taking $\sigma = 0$ for the exponential and gamma examples, $\sigma = -1$ for the Gaussian, $\sigma = 1 - \tau$ for the Weibull-like distribution, and $\sigma = 1 - \beta$ for the Benktander type II distribution). We discuss how the analysis can be adapted in the final example in Remark 4.2.

Note that we would ideally like to weaken Condition 4.1 (ii) to allow $g(x)/x^{\sigma}$ to be slowly varying at x_0 (for g to be regularly varying with index $-\sigma$); however, we leave this as future work.

Lemma 4.1. *Under Condition 4.1 (iii):*

(i) *The mean* $\mathbb{E}N_n$ *converges to the Euler–Mascheroni constant* γ.

(ii) *By Taylor's theorem, the moment-generating function converges as follows:*

$$\lim_{n \to \infty} \mathcal{M}_{N_n}(t) = \Gamma(1 - t). \tag{4.3}$$

Proof. Note that (see [25, Proposition 2.1]) under Condition 4.1 (iii) the kth moment of N_n converges:

$$\lim_{n \to \infty} \mathbb{E}(N_n)^k = (-1)^k \Gamma^{(k)}(1), \tag{4.4}$$

where $\Gamma^{(k)}(x)$ is the kth derivative of the Γ function at x.

https://doi.org/10.1017/jpr.2023.37 Published online by Cambridge University Press

We deduce convergence of the moment-generating function by Taylor's theorem:

$$\lim_{n\to\infty} \mathcal{M}_{N_n}(t) = \lim_{n\to\infty} \sum_{r=0}^{\infty} \frac{t^r \mathbb{E}(N_n)^r}{r!} = \lim_{n\to\infty} \sum_{r=0}^{\infty} \frac{t^r \mathbb{E}(N_n)^r}{r!}$$
$$= \sum_{r=0}^{\infty} \frac{(-t)^r \Gamma^{(r)}(1)}{r!} = \Gamma(1-t). \qquad \square$$

Theorem 4.1. *If the distribution function of X satisfies the von Mises representation (4.1) with $x_0 = \infty$ and Condition 4.1 holds, there exist norming constants a_n and b_n satisfying $1/n = \overline{F}(b_n)$ and $a_n = g(b_n)$ such that $N_n = (M_n - b_n)/a_n$ satisfies $\lim_{n\to\infty} D(N_n \parallel Y) = 0$, where Y is a standard Gumbel distribution (with $\beta = 1$ and $\mu = 0$).*

Proof. We use the norming constants from [25, Proposition 1.1(a)] (see also [13, Table 3.4.4]). We can write the first term in the relative entropy expression (2.3) in the case $\mu = 0$ and $\beta = 1$ using Lemma 3.2 as

$$\mathbb{E}\Theta_{N_n}(N_n) + \mathbb{E}N_n = (\log a_n + \mathbb{E}\log h_X(M_n)) + (\log n - H_n) + \frac{1}{n} + \mathbb{E}N_n. \tag{4.5}$$

We can consider the behaviour as $n \to \infty$ of the four terms in (4.5) separately:

(i) We can write the first term in (4.5) in terms of the σ of Condition 4.1(ii) as

$$\log a_n + \mathbb{E}\log h_X(M_n) = \log g(b_n) + \mathbb{E}\log h_X(M_n)$$
$$= \log\left(\frac{g(b_n)}{b_n^\sigma}\right) - \mathbb{E}\log\left(\frac{g(M_n)}{M_n^\sigma}\right)$$
$$- \sigma\,\mathbb{E}\log\left(\frac{M_n}{b_n}\right) + \mathbb{E}\log(g(M_n)h_X(M_n)). \tag{4.6}$$

(a) Since $b_n \to x_0$ and $M_n \to x_0$ in distribution, we know that the first two terms of (4.6) tend to $\log\gamma - \log\gamma = 0$, using the portmanteau lemma.

(b) We can control the third term of (4.6) by writing $M_n = b_n + a_n N_n$ (and recalling that $a_n = g(b_n)$) to obtain

$$\mathbb{E}\log\left(\frac{M_n}{b_n}\right) = \mathbb{E}\log\left(1 + \frac{a_n N_n}{b_n}\right) = \sum_{k=1}^{\infty} \frac{(-1)^k}{k}\left(\frac{a_n}{b_n}\right)^k \mathbb{E}N_n^k,$$

and we can use the facts that $a_n/b_n = g(b_n)/b_n \sim \gamma b_n^{\sigma-1} \to 0$ and (by (4.4)) that $\mathbb{E}N_n^k$ converges to a finite constant to deduce that this term tends to zero.

(c) Using the representation of the hazard function in (4.2) we know that the fourth term of (4.6) equals $\mathbb{E}\log(1 - c'(M_n)g(M_n)/c(M_n)) = \mathbb{E}\log\ell(M_n)$, so, since $M_n \to x_0$ in distribution, we know that $\mathbb{E}\log(g(M_n)h_X(M_n)) \to \log\ell(x_0) = 0$, by the portmanteau lemma.

Hence, overall, the first term of (4.5) tends to zero.

(ii) It is a standard result that $\log n - H_n$ is a monotonically increasing sequence that converges to $-\gamma$.

https://doi.org/10.1017/jpr.2023.37 Published online by Cambridge University Press

(iii) Clearly, the third term in (4.5) converges to zero.

(iv) Lemma 4.1 tells us that the final term converges to γ.

Putting this all together, we deduce that (4.5) converges to $0-\gamma+0+\gamma=0$.

In the case $\mu=0$ and $\beta=1$, the second term in the relative entropy expression (2.3) becomes $\mathbb{E}e^{-N_n}-1=\mathcal{M}_{N_n}(-1)-1\to\Gamma(2)-1=0$, by (4.3).

Corollary 4.1. *Assume that the distribution function F_X has a von Mises representation (4.1) whose auxiliary function g satisfies Condition 4.1. Then the entropy of the normalized maximum $N_n=(M_n-b_n)/a_n$ satisfies $\lim_{n\to\infty}H(N_n)=1+\gamma$, which is the entropy of the corresponding Gumbel distribution.*

Proof. By Proposition 2.1, $H(N_n)=1-\mathbb{E}\Theta_{N_n}(N_n)=(1+\mathbb{E}N_n)-\mathbb{E}(\Theta_{N_n}(N_n)+N_n)$. The first term converges to $1+\gamma$ by Lemma 4.1, and the second term is precisely (4.5) and converges to zero as described in the proof of Theorem 4.1. □

Remark 4.1. Note that, for the exponential case of Example 4.1, Condition 4.1 is satisfied, so we can deduce convergence in relative entropy. Indeed, since g is constant in this case, the first term of (4.5) vanishes, meaning that we can deduce that $\mathbb{E}\Theta_{N_n}(N_n)=\log n-H_n+1/n$ and the entropy is exactly $H(N_n)=1+H_n-\log n-1/n$, which value may be of independent interest. Using the fact that $H_n=\log n+\gamma+1/(2n)+O(1/n^2)$, we can deduce that $H(N_n)=1+\gamma-1/(2n)+O(1/n^2)$. In the spirit of [17] and other papers, it may be of interest to ask under what conditions the convergence in Corollary 4.1 is at rate $O(1/n)$ in this way.

Theorem 4.1 shows that convergence in relative entropy occurs for a range of random variables that are 'well behaved' in some sense. However, observe that the Gnedenko example $g(u)=(1-u)^2$ from Example 4.1 does not satisfy Condition 4.1(ii), so Theorem 4.1 cannot be directly applied in this case. However, it is possible to deduce convergence in relative entropy in this example too, using a relatively simple adaptation of the argument to a class of random variables with finite x_0 such that the following replacement for Condition 4.1(ii) holds.

Condition 4.2. *Assume there exists a constant $\sigma>1$ such that $\log(g(x)/(x_0-x)^\sigma)$ is bounded and continuous, and that $\gamma:=\lim_{x\to x_0}g(x)/(x_0-x)^\sigma$ is finite and non-zero.*

Remark 4.2. The only place where we need to adapt the proof of Theorem 4.1 is in the decomposition of the first term in (4.5), where we can instead use the decomposition

$$\log\left(\frac{g(b_n)}{(x_0-b_n)^\sigma}\right)-\mathbb{E}\log\left(\frac{g(M_n)}{(x_0-M_n)^\sigma}\right)-\sigma\mathbb{E}\log\left(\frac{x_0-M_n}{x_0-b_n}\right).$$

As before, the first two terms tend to $\log\gamma-\log\gamma=0$ by the portmanteau lemma. We can use a similar Taylor expansion,

$$\mathbb{E}\log\left(\frac{x_0-M_n}{x_0-b_n}\right)=-\sum_{k=1}^{\infty}\frac{1}{k}\left(\frac{a_n}{x_0-b_n}\right)^k\mathbb{E}N_n^k,$$

and deduce convergence in relative entropy since $a_n/(x_0-b_n)=g(b_n)/(x_0-b_n)\simeq\gamma(x_0-b_n)^{1-\sigma}\to 0$.

We have shown that there is a natural information-theoretic interpretation of convergence in relative entropy of the standardized maximum to the Gumbel distribution, and provided simple

https://doi.org/10.1017/jpr.2023.37 Published online by Cambridge University Press

conditions under which this occurs. It would be of interest to provide a similar analysis for the other extreme value distributions (as proved for example using different methods in [23])—the Fréchet (Type II) and Weibull (Type III) distributions—which remains an interesting problem for future work, as does the question of the optimal rate of convergence in relative entropy.

We note that a similar function to the max-score can be used to evaluate the entropy of more general order statistics, as studied recently in [9] (see also [29]). That is, given an i.i.d. sample $X_1, X_2, \ldots, X_n$ from F_X, if we write the order statistics $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ then it is well known (see, e.g., [1, (2.2.2)]) that the density of $X_{(r)}$ is $f_{X_{(r)}}(x) = c_{n,r} f_X(x) F_X(x)^{r-1}(1 - F_X(x))^{n-r}$, where $c_{n,r} = n!/(r-1)!/(n-r)!$. Hence, we can provide analysis similar to that based on Proposition 2.1 for the maximum, by writing $H(X_{(r)})$ as

$$
\begin{aligned}
&- \int f_{X_{(r)}}(x)(\log c_{n,r} + \log f_X(x) + (r-1)\log F_X(x) + (n-r)\log(1 - F_X(x)))\,\mathrm{d}x \\
&= -\log c_{n,r} - \mathbb{E}h_X(X_{(r)}) - \int f_{X_{(r)}}(x)((r-1)\log F_X(x) + (n-r-1)\log(1 - F_X(x)))\,\mathrm{d}x \\
&= -\log c_{n,r} - \mathbb{E}h_X(X_{(r)}) - \int c_{n,r} u^{r-1}(1-u)^{n-r}((r-1)\log u + (n-r-1)\log(1-u))\,\mathrm{d}u \\
&= -\log c_{n,r} - \mathbb{E}h_X(X_{(r)}) - ((r-1)(H_{r-1} - H_n) + (n-r-1)(H_{n-r} - H_n))\,\mathrm{d}u,
\end{aligned}
$$

where the final result comes from taking $u = F_X(x)$ and using standard results about beta integrals. Essentially, we recover [9, Lemma 3] in the case of uniform F_X, since $h_X(x) = 1/(1-x)$. Again, most of the terms do not depend on F_X itself, so by bounding the hazard function we can control the behaviour of the entropy.

Acknowledgements

We wish to thank the two anonymous reviewers for their helpful comments which helped improve the presentation of this paper.

Funding information

There are no funding bodies to thank relating to the creation of this article.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (2008). *A First Course in Order Statistics*. SIAM, Philadelphia.

[2] Arulkumaran, K., Deisenroth, M. P., Brundage, M. and Bharath, A. A. (2017). A brief survey of deep reinforcement learning. Preprint, arXiv:1708.05866.

[3] Barron, A. R. (1986). Entropy and the central limit theorem. *Ann. Prob.* **14**, 336–342.

[4] Bartholmé, C. and Swan, Y. (2013). Rates of convergence towards the Fréchet distribution. Preprint, arXiv:1311.3896.

[5] Bercher, J.-F. and Vignat, C. (2008). An entropic view of Pickands' theorem. In *Proc. 2008 IEEE Int. Symp. Information Theory (ISIT)*. IEEE, New York, pp. 2625–2628.

[6] Blachman, N. M. (1965). The convolution inequality for entropy powers. *IEEE Trans. Inf. Theory* **11**, 267–271.

[7] Block, H. W., Savits, T. H. and Singh, H. (1998). The reversed hazard rate function. *Prob. Eng. Inf. Sci.* **12**, 69–90.

https://doi.org/10.1017/jpr.2023.37 Published online by Cambridge University Press

[8] BROWN, L. D. (1982). A proof of the central limit theorem motivated by the Cramér–Rao inequality. In *Statistics and Probability: Essays in Honour of C. R. Rao*, eds G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh. North-Holland, New York, pp. 141–148.
[9] CARDONE, M., DYTSO, A. AND RUSH, C. (2023). Entropic central limit theorem for order statistics. *IEEE Trans. Inf. Theory* **69**, 2193–2205.
[10] COVER, T. M. AND THOMAS, J. A. (1991). *Elements of Information Theory*. John Wiley, New York.
[11] DE HAAN, L. AND FERREIRA, A. (2007). *Extreme Value Theory: An Introduction*. Springer, New York.
[12] DE HAAN, L. AND RESNICK, S. I. (1982). Local limit theorems for sample extremes. *Ann. Prob.* 396–413.
[13] EMBRECHTS, P., KLÜPPELBERG, C. AND MIKOSCH, T. (2013). *Modelling Extremal Events: For Insurance and Finance* (Stochastic Modelling and Appl. Prob. **33**). Springer, New York.
[14] GNEDENKO, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.* **44**, 423–453.
[15] JOHNSON, O. T. (2004). *Information Theory and the Central Limit Theorem*. Imperial College Press, London.
[16] JOHNSON, O. T. (2020). Maximal correlation and the rate of Fisher information convergence in the central limit theorem. *IEEE Trans. Inf. Theory* **66**, 4992–5002.
[17] JOHNSON, O. T. AND BARRON, A. R. (2004). Fisher information inequalities and the central limit theorem. *Prob. Theory Relat. Fields* **129**, 391–409.
[18] KONTOYIANNIS, I., HARREMOËS, P. AND JOHNSON, O. T. (2005). Entropy and the law of small numbers. *IEEE Trans. Inf. Theory* **51**, 466–472.
[19] KOTZ, S. AND NADARAJAH, S. (2000). *Extreme Value Distributions: Theory and Applications*. World Scientific, Singapore.
[20] KULLBACK, S. (1967). A lower bound for discrimination information in terms of variation. *IEEE Trans. Inf. Theory* **13**, 126–127.
[21] PICKANDS III, J. (1986). The continuous and differentiable domains of attraction of the extreme value distributions. *Ann. Prob.* **14**, 996–1004.
[22] RAVI, S. AND SAEB, A. (2012). A note on entropies of *l*-max stable, *p*-max stable, generalized Pareto and generalized log-Pareto distributions. *ProbStat Forum* **5**, 62–79.
[23] RAVI, S. AND SAEB, A. (2014). On convergence of entropy of distribution functions in the max domain of attraction of max stable laws. Preprint, arXiv:1402.0277.
[24] REISS, R.-D. (2012). *Approximate Distributions of Order Statistics: With Applications to Nonparametric Statistics*. Springer, New York.
[25] RESNICK, S. I. (2013). *Extreme Values, Regular Variation and Point Processes*. Springer, New York.
[26] SAEB, A. (2014). Rates of convergence for Rényi entropy in extreme value theory. Preprint, arXiv:1402.4316.
[27] STAM, A. J. (1959). Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inf. Control* **2**, 101–112.
[28] STEIN, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. 6th Berkeley Symp. Math. Statist. Prob.*, Vol. 2. University of California Press, Berkeley, CA.
[29] WONG, K. M. AND CHEN, S. (1990). The entropy of ordered sequences and order statistics. *IEEE Trans. Inf. Theory* **36**, 276–284.

https://doi.org/10.1017/jpr.2023.37 Published online by Cambridge University Press