Improving extreme value statistics

Ashivni Shekhawat

Phys. Rev. E **90**, 052148 — Published 24 November 2014

DOI: 10.1103/PhysRevE.90.052148

Improving extreme value statistics

Ashivni Shekhawat[1,2]

[1] *Miller Institute for Basic Research in Science, UC Berkeley, Berkeley, CA USA*
[2] *Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA*

The rate of convergence in extreme value statistics is non-universal and can be arbitrarily slow. Further, the relative error can be unbounded in the tail of the approximation, leading to difficulty in extrapolating the extreme value fit beyond the available data. We introduce the T method, and show that by using simple nonlinear transformations the extreme value approximation can be rendered rapidly convergent in the bulk, and asymptotic in the tail, thus fixing both issues. The transformations are often parameterized by just one parameter which can be estimated numerically. The classical extreme value method is shown to be a special case of the proposed method. We demonstrate that vastly improved results can be obtained with almost no extra cost.

PACS numbers: 02.50.-r

Extreme value statistics provides a universal statistical description of rare events. Such events dictate the fate of a vast variety of phenomena spanning science, engineering, and humanities alike. Examples include hydrology and one-hundred-year floods [1–3], climatology [4, 5], ground states of random energy model [6] and directed polymers [7], catastrophic fracture [8–11], novelty detection [12], stock price fluctuations [13, 14], risk management [15, 16], large insurance claims [17, 18] and so on. The field of extreme value statistics is based on a mathematical theorem [19–22] which states that (with mild restrictions) the distribution of maximum or minimum of a large set of uncorrelated random variables approaches a universal form. Thus, the same universal theory can be used to model the statistics of large floods, large insurance claims, low energy states of disordered systems, or large cracks in a bridge. Unfortunately, this universality breaks down in two related ways. Firstly, the rate of convergence to the universal extreme value form is non-universal, and can be arbitrarily slow [23–26]. Secondly, the mathematical theorem guarantees only uniform convergence, thus the relative error in the tail of the distribution is also non-universal and can be unbounded [27]. In spite of the vast applicability of extreme value statistics and extensive research on the subject, no method is available to overcome these issues. In this paper we propose the first method that gives a universal guarantee on the rate of convergence as well as on the relative error in the tail.

The non-universal behavior in convergence has important practical consequences that are seldom discussed outside of specialized mathematical literature. In several applications such as hydrology or reliability engineering, the amount of data available is rather small. For instance, in hydrology even a century of data collection leads to a mere hundred observations of the flood level at a given geographical location. Similarly in reliability engineering, conducting long and/or repeated experiments can be extremely time and resource intensive. In such cases statistical inference about the extremes needs to be made on the basis of a small dataset. Clearly, slow convergence to the limit form can lead to large statistical errors. Further, by the very nature of the subject, rare events are often outside the range of the available data (think predicting a so-called 'thousand year flood' based on a century worth of data). Thus, extrapolation needs to made in the tail of the distribution, where the relative error is non-universal, and can be extremely large. It is clear that the issue of non-universal rate of convergence and error of extrapolation must be examined carefully.

We demonstrate the non-universal behavior in extreme value statistics and the associated issues with a simple example. Consider an (admittedly contrived) industrial process that grinds out metallic disks whose area, A, is distributed exponentially, so that $P(A < a) = 1 - e^{-a}$. Two analysts are given 100 boxes, each containing 10 such disks. They are asked to approximate the probability distribution of the radii of largest disk in each of the boxes. The first analyst (say, Bob) simply measures the radius of the largest disk in each box, and fits these 100 observation to an extreme value form, perhaps using maximum likelihood estimation. The second analyst (say, Alice) decides to take a different route. She measures the areas of the largest disk in each box, and fits this data to an extreme value distribution. Then she can predict the probability of the radii by using a simple transformation. Both Bob and Alice report their findings (figure 1), and the employers who know the exact distribution fire Bob. What went so wrong for Bob?

A great deal of insight about convergence issues in extreme value statistics can be gained by analyzing this simple example. However, in order to appreciate the issues involved we need to briefly recollect the standard results of extreme value theory (details in Refs. [20–22], and [28]). Let $\tilde{X}$ be a random variable with a cumulative distribution function (cdf) $F(\cdot)$, thus $P(\tilde{X} < x) = F(x)$. Let $X_n = \max(\tilde{X}_1, \ldots, \tilde{X}_n)$ be the maximum of n independent and identically distributed (iid) random variables drawn from distribution $F(\cdot)$. The distribution of X_n is then given by $P(X_n < x) = F(x)^n$. Theorems in extreme value statistics state that under mild restrictions $F(a_n x + b_n)^n \to G_\gamma(x)$, where $G_\gamma(x) = \exp\{-(1+\gamma x)^{-1/\gamma}\}$ is the universal extreme value distri-

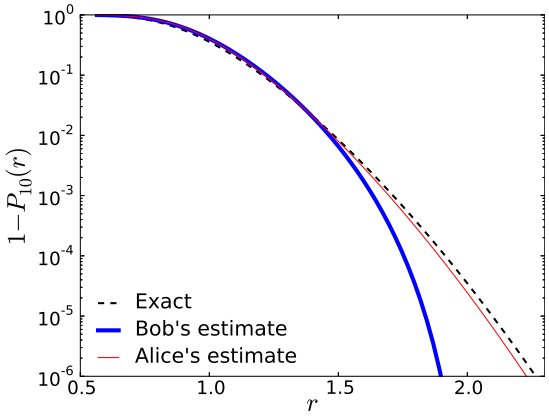

FIG. 1. (Color online) Comparison of the exact and estimated probability of the radius of the largest of 10 disks being greater than r. Bob's estimate (thick blue line), based on the extreme value theory, largely underestimate the probability of observing disks with large radii, while Alice's estimate (thin red line), based on treating area as the primary random variable, and transforming back the results to get probability of radii, works much better.

bution [19, 21, 22]. The exact value of $\gamma,\ a_n,\ b_n$ depends on the details of $F(\cdot)$, but the form of $G_\gamma(\cdot)$ is universal. The cases $\gamma\ =,\ >,\ <\ 0$ correspond to the Gumbel (or type I), the Fréchet (or type II), and the Weibull (or type III) distributions, respectively. Thus, $G_\gamma((x-b_n)/a_n)$ can be used as an approximation for $F(x)^n$. In practical applications $\gamma,\ a_n,\ b_n$ are treated as parameters to be fit to data, and are typically estimated by using maximum likelihood estimation (MLE). Predictions are made by extrapolating the obtained fit. The success of the extreme value theory is due to its stark simplicity. In any application where the statistics are dominated by rare events, one can use a simple analytical form for the distribution function that is parameterized by just three free parameters.

Let us now analyze the Alice-Bob example. Let $P_{10}(A<a),\ P_{10}(R<r)$ be the probabilities that the area, radius of the largest disk in the box of 10 disks are lesser than a, r, respectively. Clearly, $P_{10}(A<a)=(1-e^{-a})^{10}$ and $P_{10}(R<r)=(1-e^{-\pi r^2})^{10}$. It is also clear that the tail of radius distribution, $P_{10}(R<r)$, decays faster than exponentially for large r, thus any extreme value distribution, $G_\gamma(r)$, will not be able to model it accurately. On the other hand, the tail of the area distribution, $P_{10}(A<a)\sim 1-10e^{-a}$, decays exponentially, and can be modeled accurately by $G_0(a-\log 10)=\exp(-10e^{-a})\sim 1-10e^{-a}$. Thus, there is an inherent advantage to working with A as the random variable being fit to extreme value distributions, even though A and R are simply related by $A=\pi R^2$. This is an example of non-universal behavior in extreme value statistics. After a fit has been obtained for A, the probability for R can be obtained easily by transforming back via $P_{10}(R<r)=P_{10}(A<\pi r^2)$. Figure 1 shows a comparison of Bob's and Alice's estimates and the exact result. Since there were 100 boxes, the empirical data was available at a probability level of $1-P=10^{-2}$. Up to this level both estimates agree reasonably with the exact result. However, at $r=1.91$, the exact result is $1-P_{10}(R<r)=10^{-4}$, Bob's estimate is $1-P_B(r)=4.5\times10^{-7}$, and Alice's estimate is $1-P_A(r)=7.3\times10^{-5}$. Thus, Alice's estimate is off by about 25%, while Bob's is off by more than two orders of magnitude. Indeed, considerable advantage can be gained by working with a suitable nonlinear transformation of the data. The challenge is to identify the correct transformation, and this is precisely what the method to be proposed in this letter does. We call the proposed method the T method ('T' for Transformation based method).

The insight gained from the above example can be formalized. However, in order to facilitate the discussion we need to introduce two widely used notions of convergence in extreme value statistics. The absolute error of approximation is often characterized by the following asymptotic expansion for large n,

$$F(a_nx+b_n)^n-G_\gamma(x)\sim W(n)G'_\gamma(x)\hat{H}_\gamma[G_\gamma(x)], \quad (1)$$

where the exact form of the function $\hat{H}(\cdot)$ is somewhat complicated [23]. Such expansions are called Edgeworth-type expansions, and are used widely in mathematical statistics [23, 29]. In the above expansion, the rate of convergence is governed by the $F(\cdot)$-dependent non-universal function $W(\cdot)$ [23]. The decay of $W(n)$ for large n can be arbitrarily slow (or arbitrarily fast) depending on $F(\cdot)$. For example, $W(n)=0$ identically if $F(x)=G_\gamma(x)$, $W(n)\sim -1/2n$ if $F(x)=1-e^{-x}$ (the standard exponential distribution), and $W(n)\sim -1/2\log(n/\sqrt{2\pi})$ if $F(x)=\int_{-\infty}^{x}e^{-t^2/2}/\sqrt{2\pi}dt$ (the standard normal distribution) (supplemental section 4). Thus, the rate of convergence can range from infinitely fast to logarithmically slow (or worse). The logarithmic rate of convergence is obviously a cause of concern in practice. Refs. [24, 25] show that the convergence can sometimes be improved by considering penultimate approximations, but the rate still remains logarithmic in several cases of interest. We will show that by using the proposed T method the rate of convergence can be improved considerably. The T method provides a guaranteed $1/n$ convergence as opposed to $1/\log n$ or slower in worse case for classical extreme value statistics.

A second measure of convergence is needed because the Edgeworth expansion characterizes the absolute error, and is not a good measure of the relative error between $F(a_nx+b_n)^n$ and $G_\gamma(x)$ in the upper (or lower) tail. The relative error can be dominated by the higher order terms that are ignored in the Edgeworth expansion. The relative error in the tail is important for extrapolating the results beyond the available data. In such cases, the quality of the upper tail of the approximation is mea-

sured by the ratio [27]

$$L(x) \equiv (1 - F(a_n x + b_n)^n)/(1 - G_\gamma(x)). \qquad (2)$$

Ideally $L(x)$ should stay close to 1. However, practically it can differ significantly from its ideal value of 1 for x close to $x_+ \equiv \sup_x\{x : F(x) < 1\}$ at fixed n. This behavior is characterized by studying the range of x at a given n for which $L(x)$ is close to 1 [27]; the bigger the range, the better the approximation. As before, the behavior is non-universal and a variety of outcomes are possible depending on the details of $F(x)$. $L(x)$ approaches its ideal value of 1 for the exponential distribution, while it decays to 0 rather quickly for the normal distribution (supplemental section 5). This behavior can lead to particularly severe errors and uncertainty when the fit to the extreme value approximation is extrapolated beyond the available data (as in the Alice-Bob example). As mentioned earlier, such extrapolation is carried out routinely in extreme value analysis. We will show how this difficulty can be overcome by using the T method which guarantees that $L(x) \to 1$.

It is sometimes indicated in the literature that the slow rate of convergence is limited to the cases where $\gamma = 0$, i.e. the Gumbel (type I) distribution. This is incorrect (supplemental section 14). However, it is true that out of the most commonly used distributions, the cases with $\gamma = 0$ are more prone to such issues. We shall restrict our discussion to such case from here onwards. However, our method is equally applicable to other cases.

The essential insight gained from the Alice-Bob example was that it can be advantageous if we apply a suitable nonlinear transformation to the data, instead of working with the raw data itself. Suppose that we have data vector $\mathbf{X}$, where each $\mathbf{X}[i] = \max(\tilde{X}_1, \ldots, \tilde{X}_n)$, and $\tilde{X}_i$ are iid random variables drawn from an underlying distribution $F(\cdot)$. What is the optimal choice for a transformed variable $\tilde{Y} = \hat{T}(\tilde{X})$? Clearly, if we choose $\hat{T}(\cdot) = G_0^{-1}(F(\cdot))$, then the distribution of $\tilde{Y}$ becomes $P(\tilde{Y} < y) = G_0(y)$, and the transformed data $\mathbf{Y}$ is described perfectly by an extreme value distribution with linear rescaling, $P(\mathbf{Y}[i] < y) = G_0((y - b_n)/a_n)$, where $a_n = 1,\ b_n = \log n$. The distribution of $\mathbf{X}$ can now be recovered *exactly* as $P(\mathbf{X}[i] < x) = G_0((\hat{T}(x) - b_n)/a_n)$. Thus, working with a suitably transformed variable completely suppresses the systematic errors of the extreme value approximation in the sense of Eqs. 1, 2. However, there is a slight problem with this scheme: it demands that to construct $\hat{T}(\cdot)$ we know $F(\cdot)$, which if we knew, we could calculate the distribution of $\mathbf{X}$ $(= F(\cdot)^n)$ exactly without this elaborate scheme anyway!

The problem can be made more tractable by observing that since the large (or small) observations that dominate the extreme value statistics are described by the tail of the distribution $F(\cdot)$, it is not necessary to know $F(\cdot)$ in full detail to construct a reasonable transformation. It is intuitive that the first few terms in the asymptotic expansion of $F(\cdot)$ must suffice. We will show that just the leading term in the expansion is sufficient to arrive at an excellent approximation. Let $F(x) \sim 1 - \sum_{i=0}^{\infty} f_i(x)$ be an asymptotic expansion for large x (or for x near x_+ if the distribution $F(\cdot)$ has bounded support), where the gauge functions $f_i(\cdot)$ are monotonic. Then the transformation $\tilde{Y} = T(\tilde{X}) = -\log f_0(\tilde{X})$ has the following properties

$$F(T^{-1}(a_n x + b_n))^n - G_0(x) \sim \frac{1}{2n} G_0'(x) \hat{H}_0[G_0(x)], \quad (3)$$

$$(1 - F(T^{-1}(a_n x + b_n))^n)/(1 - G_0(x)) \to 1, \qquad (4)$$

where, $a_n = 1,\ b_n = \log n$, $\hat{H}_0[G_0(x)] = e^{-x} + x - 1$ (the proofs are simple, see supplemental section 7). Thus, irrespective of the details of $F(\cdot)$, we have guaranteed that the absolute error converges as $1/n$, and the relative error in the tail vanishes. This is a significant improvement over the classical extreme value statistics, where there is no guarantee on either of these behaviors. Clearly, $G_0((T(x) - b_n)/a_n)$ is a good approximation for $F(x)^n$.

The proposed method, which we call the T method, is now clear. We parameterize the transformation $T(x) = -\log f_0(x)$ by a parameter β. For several common distributions, the forms $T(x) = x^\beta, (\log(x))^\beta$ are suitable (see supplemental section 13 for a discussion of how to chose $T(x)$ based on the data). Then we have a parameter vector $\theta = (\beta, a_n, b_n)$, and a model $G_0((T(\mathbf{X}|\beta) - b_n)/a_n)$. Given data vector $\mathbf{X}$ the parameter vector θ can be estimated via the maximum likelihood method by maximizing the following likelihood function

$$\mathcal{L}(\theta|\mathbf{X}) = \sum_i \log \partial_{\mathbf{X}_i} G_0((T(\mathbf{X}_i|\beta) - b_n)/a_n). \qquad (5)$$

Thus, the T method is no more complicated or computationally taxing than the classical extreme value statistics where one maximizes the likelihood function $\sum_i \log \partial_{\mathbf{X}_i} G_\gamma((\mathbf{X}_i - b_n)/a_n)$ to estimate the parameter vector (γ, a_n, b_n).

We test the proposed method on data generated from the normal and the lognormal distribution, as well as from simulations of height of Brownian excursions [30], and strengths of local load sharing fiber bundles [9, 31] (details of these models can be found in the cited references and supplemental section 11). The fiber bundle is a widely used model of fracture, while the Brownian excursion finds applications in stochastic processes, graph theory [32], and theory of fluctuating interfaces [33]. In the supplemental section 9 we demonstrate applications to a meteorological data set (the Venice sea level data [34, 35]), and in section 9 to data drawn from exponential distribution. The applicability of extreme value statistics to such a wide variety of phenomena again shows its universal nature. In all the cases we have studied so far, the T method presents an improvement over the classical method.

We generate random samples $\mathbf{X}$ of size 100. For the normal and lognormal case each $\mathbf{X}[i] =$

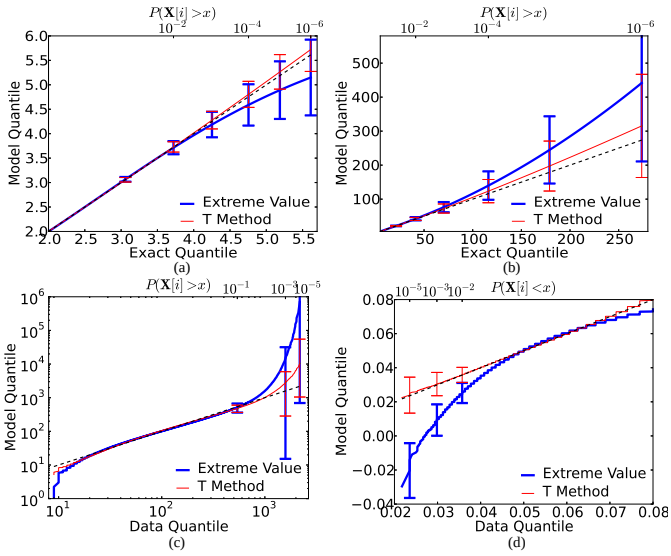

FIG. 2. (Color online) Comparison of the classical extreme value approximation with the T method for (a) normal distribution, (b) lognormal distribution, (c) height of a Brownian excursion [30], and (d) strength of a fiber bundle [9, 31]. The main graph shows the QQ plot, i.e. model versus empirical (or exact, in the normal and lognormal case) inverse cdf, while the upper x-axis shows the appropriate exceedance corresponding to the quantile on the main x-axis. The dashed black line is a guide to the eye and shows the ideal result. The solid lines show the quantiles averaged over 1000 Monte Carlo runs, where the thick blue line is used for the extreme value theory, while the thin red line is used for the T method. The errorbars show the 2-standard deviation range. In each Monte Carlo run the model fit to a sample of size $m = 100$, and the fit is extrapolated to a probability level of 10^{-5} or lower. It is clear that the T method yields better predictions and less variance even when extrapolated well beyond the range of the available data. Note the differences in the fiber bundle (case (d)) because bundle strengths are modeled as minima, as opposed to maxima in the other three cases.

$\max(\tilde{X}_1, \ldots, \tilde{X}_n)$, $n = 100$, and $\tilde{X}_i$ are drawn from normal and the lognormal distributions, respectively. For the case Brownian excursions, $\tilde{X}_i$ are heights of excursions less than 10^6 steps long (without this condition the excursions can extend indefinitely). For the fiber bundle $\mathbf{X}[i]$ is the strength of a bundle consisting of 1000 fibers. The strength of the bundle is dictated by the strength of its weakest flaw, thus this is a case where we study the minima of random variables (the formulation remains almost unchanged, supplemental section 10). We estimate the parameter vector $\theta = (\beta, a_n, b_n)$ by using MLE (Eq. 5) with $T(x|\beta) = x^\beta$ for normal, $(\log(x))^\beta$ for lognormal and excursion heights, and $x^{-\beta}$ for bundle strengths. The fits to datasets of size 100 are then extrapolated to a probability level (or reliability level, in case of fiber bundle strength) of 10^{-5} or lower. The result of the extrapolation is compared to exact results in normal and lognormal case, and to high precision numerical results for the case of fiber bundles and excursions. Figure 2 shows a favorable comparison of the results obtained by the T method with the classical extreme value approximation. It is clear that the suggested method out-performs the classical extreme value approximation with the same number of parameters.

The classical formulation of the extreme value statistics is limited to iid random variables, and as such, so is the T method. One future direction would be to extend the T method to stationary sequences, where the classical formulation is applicable under mild restrictions [36, 37]. Our understanding of the non-stationary and strongly correlated random variables is limited. The few available analytical solutions seem to indicate that there is no universality (and thus no general method) for the case of non-stationary and strongly correlated variables [7, 38]. One possible direction for the correlated case would be build on the filtering methods [39] that generate independent extremes from dependent observations.

In summary, we have suggested a simple method, which we call the T method, to alleviate the problem of slow convergence of classical extreme value approximations. The method works by estimating a simple nonlinear transformation that defines a new random variable that has better convergence properties in the extreme value sense. Some previous authors have studied rates of convergence in nonlinear scaling in extreme value statistics (see Refs. [40, 41]). Their results are rather remarkable, however, their focus has been on studying $W(n)$ for specific transformations (power transformation, for example) rather than constructing numerical methods of wide applicability. In this sense the proposed T method is complementary to their results. The T method was applied successfully to distributions in the domain of attraction of the Gumbel (type I) distribution. We hope that application of our method will lead to more reliable estimates of probabilities of extremes in a large number of applications.

ACKNOWLEDGMENTS

The author acknowledges support from DOE-BES DE-FG02-07ER46393 while he was at Cornell University, and from the Miller Institute for Basic Research in Science at University of California, Berkeley. The author would like to thank Prof. James P. Sethna and Prof. Stefano Zapperi for insightful discussions and comments, and Prof. Robert O. Ritchie for hosting him at the Lawrence Berkeley National Laboratory.

[1] R. W. Katz, M. B. Parlange, and P. Naveau, Advances in Water Resources **25**, 1287 (2002).

[2] J.-T. Shiau, S. Feng, and S. Nadarajah, Hydrological Pro-

cesses **21**, 2157 (2007).
[3] M. N. Khaliq, T. B. M. J. Ouarda, J.-C. Ondo, P. Gachon, and B. Bobee, Journal of Hydrology **329**, 534 (2006).
[4] T. A. Buishand, Statistica Neerlandica **43**, 1 (1989).
[5] R. L. Smith, Statistical Science **4**, 367 (11 1989).
[6] J.-P. Bouchaud and M. Mezard, Journal of Physics A: Mathematical and General **30**, 7997 (1997).
[7] S. N. Majumdar and P. L. Krapivsky, Physica A: Statistical Mechanics and its Applications **318**, 161 (2003).
[8] W. Weibull, *A statistical theory of the strength of materials* (Stockholm, 1939).
[9] D. G. Harlow and S. L. Phoenix, Journal of Composite Materials **12**, 195 (1978).
[10] P. M. Duxbury, P. L. Leath, and P. D. Beale, Phys. Rev. B **36**, 367 (Jul 1987).
[11] C. Manzato, A. Shekhawat, P. K. V. V. Nukala, M. J. Alava, J. P. Sethna, and S. Zapperi, Phys. Rev. Lett. **108**, 065504 (Feb 2012).
[12] S. J. Roberts, Science, Measurement and Technology, IEE Proceedings - **147**, 363 (Nov 2000).
[13] F. M. Longin, J. Of Business **69** (July 1996).
[14] J. P. Broussard and G. G. Booth, European Journal of Operational Research **104**, 393 (1998).
[15] P. Embrechts, S. I. Resnick, and G. Samorodnitsky, North American Actuarial Journal **3**, 30 (1999).
[16] A. J. McNeil and R. Frey, Journal of Empirical Finance **7**, 271 (2000).
[17] J. Beirlant and J. L. Teugels, Insurance: Mathematics and Economics **11**, 17 (1992).
[18] R. Holger and T. Nader, Scandinavian Actuarial Journal **1997**, 70 (1997).
[19] P. B. Gnedenko, Annals of Mathematics **44**, 423 (1943).
[20] S. I. Resnick, *Extreme values, regular variation and point processes* (Springer Verlag, New York, 2007).
[21] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values*, Vol. 208 (Springer, 2001).
[22] L. De Haan and A. Ferreira, *Extreme value theory: an introduction* (Springer, 2007).
[23] L. de Haan and S. I. Resnick, The Annals of Probability **24**, 97 (1996).
[24] M. I. Gomes, Annals of the Institute of Statistical Mathematics **36**, 71 (1984).
[25] J. P. Cohen, Advances in Applied Probability **14**, pp. 833 (1982), ISSN 00018678.
[26] G. Györgyi, N. R. Moloney, K. Ozogány, Z. Rácz, and M. Droz, Phys. Rev. E **81**, 041135 (Apr 2010).
[27] C. W. Anderson, Journal of the Royal Statistical Society. Series B (Methodological) **40**, 197 (1978).
[28] See Supplemental Material at [URL] for relevant introductory material on extreme value theory, and details of the proofs presented in this paper.
[29] P. Hall, *The bootstrap and Edgeworth expansion* (Springer, 1992).
[30] K. Chung, Arkiv for Matematik **14**, 155 (1976).
[31] P. M. Duxbury and P. L. Leath, Phys. Rev. B **49**, 12676 (May 1994).
[32] D. Aldous, The Annals of Probability **25**, pp. 812 (1997).
[33] S. Majumdar and A. Comtet, Journal of Statistical Physics **119**, 777 (2005).
[34] R. L. Smith, Journal of Hydrology **86**, 27 (1986).
[35] P. Pirazzoli, Acqua Aria **10**, 1023 (1982).
[36] M. Leadbetter, Zeitschrift fr Wahrscheinlichkeitstheorie und Verwandte Gebiete **28**, 289 (1974).
[37] S. Coles, J. Heffernan, and J. Tawn, Extremes **2**, 339 (1999), ISSN 1386-1999.
[38] D. S. Dean and S. N. Majumdar, Phys. Rev. E **64**, 046121 (Sep 2001).
[39] J. A. Tawn, Journal of Hydrology **101**, 227 (1988).
[40] E. Pantcheva, in *Stability Problems for Stochastic Models*, Lecture Notes in Mathematics, Vol. 1155, edited by V. V. Kalashnikov and V. M. Zolotarev (Springer Berlin Heidelberg, 1985) pp. 284–309.
[41] H. M. Barakat, E. M. Nigm., and M. E. El-Adll, Statistical Papers **51**, 149 (2010).