Statistics of Superior Records

Eli Ben-Naim

Los Alamos National Laboratory

with: Pearson Miller (Yale), Paul Krapivsky (Boston)

E. Ben-Naim and P.L. Krapivsky, arXiv:1305:4227

Talk, paper available from: http://cnls.lanl.gov/~ebn

Statistical dynamics of complex systems, arrabida, Portugal, July 3, 2013

Plan

I. Records: background & basic properties

II. Superior records

III. Inferior records

IV. Incremental records

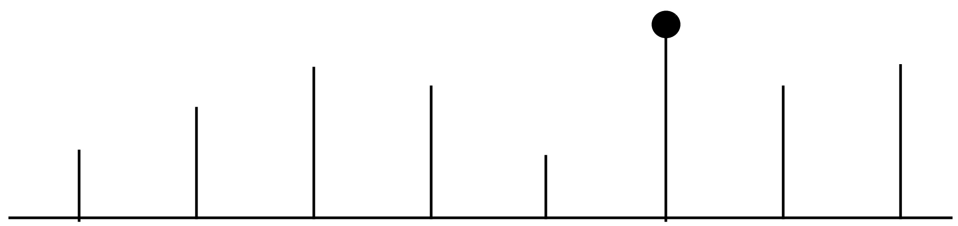
V. General distribution functions

Motivation

- Weather: record high & low temperatures — Havlin 03
- Finance: stock prices — Bouchaud 03
- Insurance: extreme/catastrophic events — Embrechts 97
- Evolution: growth rate of species — Krug 05
- Sports
- Data analysis: record high & low define span

Records and extreme values are ubiquitous

Records



- Record = largest variable in a series

$$X_N = \max(x_1, x_2, \ldots, x_N)$$

- <u>Independent and identically distributed variables</u>

$$\int_0^\infty dx\, \rho(x) = 1$$

- Canonical case: uniform distribution

$$\rho(x) = 1 \quad \text{for} \quad 0 \leq x \leq 1$$

- What is the average record?
- What is the distribution of the record?

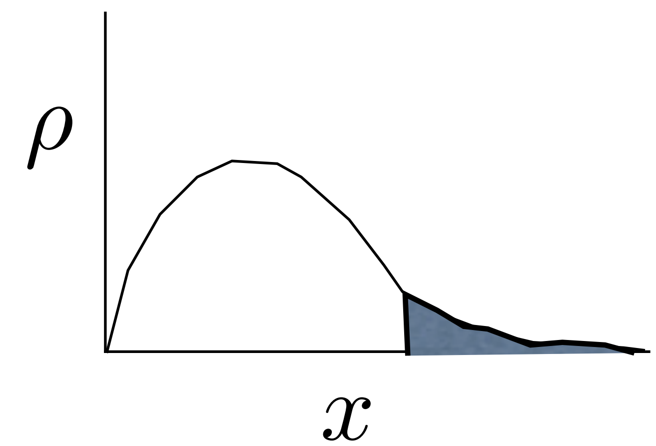
Statistics of extreme values

Feller 68
Gumble 04
Ellis 05

Distribution of the record

- Probability that one variable is $> x$

$$R(x) = \int_x^\infty dy\, \rho(y)$$



- Probability that record is $> x$

$$R_N(x) = 1 - [1 - R(x)]^N$$

- Self-similar distribution

$$R_N(x) \simeq \Psi(s) \quad \text{with} \quad s = RN$$

$$N \to \infty$$
$$R \to 0$$

- Exponential similarity function

$$\Psi(s) = 1 - e^{-s}$$

1. Distribution of extreme values is universal
2. Tail of the distribution function dominates

Fisher-Tippett-Gumble

The average record

- Cumulative distribution function

$$R_N(x) = 1 - [1 - R(x)]^N$$

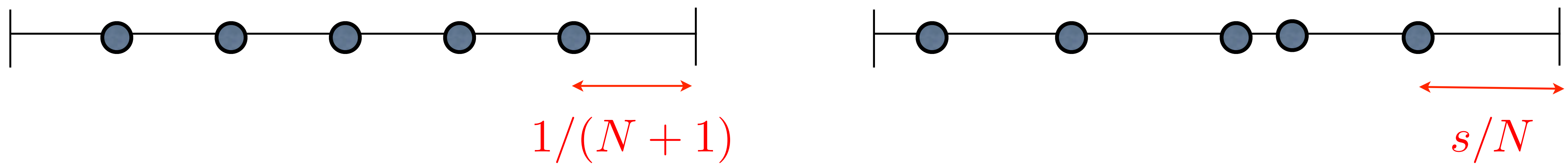
- Probability distribution function is its derivative
- Average record

$$A_N = -\int_0^\infty dx\, x\, \frac{dR_N}{dx} = N \int_0^\infty dx\, x\, \rho\, (1 - R)^{N-1}$$

- Change of variable $x = x(R)$

$$A_N = N \int_0^1 dR\, (1 - R)^N x(R)$$

Example: uniform distribution



- The variable x is randomly distributed in [0:1]

$$\rho(x) = 1 \quad \text{for} \quad 0 \le x \le 1$$

- Cumulative distribution function is linear

$$R(x) = 1 - x$$

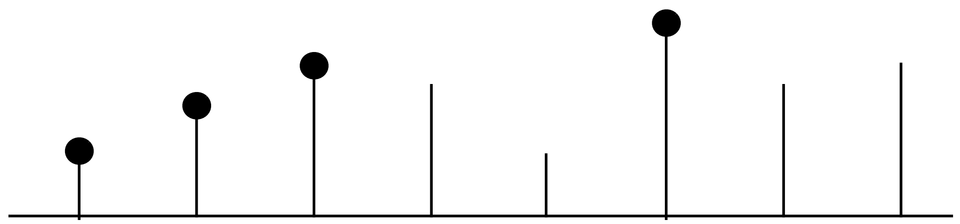
- Average record $A_N = N\int_0^1 dR\,(1-R)^N$

$$A_N = \frac{N}{N+1} \qquad \Longrightarrow \qquad 1 - A_N \simeq N^{-1}$$

- Scaling behavior

$$1 - [1-(1-x)]^N \to 1 - e^{-s} \qquad s = (1-x)N$$

Average number of records



- Probability that Nth variable is a record

$$P_N = \frac{1}{N}$$

- Average number of records = harmonic number

$$M_N = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{N}$$

- Grows logarithmically with number of variables

$$M_N \simeq \ln N + \gamma \qquad \gamma = 0.577215$$

Behavior is independent of distribution function
Number of records is quite small

Distribution of number of records

- Probability that N variables have n records satisfies recursion equation

$$Q_n(N) = \left(1 - N^{-1}\right) Q_n(N-1) + N^{-1} Q_{n-1}(N-1)$$

- Given in terms of Stirling numbers Graham, Knuth, Patashnik 89

$$Q_n(N) = \frac{1}{N!} \begin{bmatrix} N \\ n \end{bmatrix}$$

- Variance related to second harmonic numbers

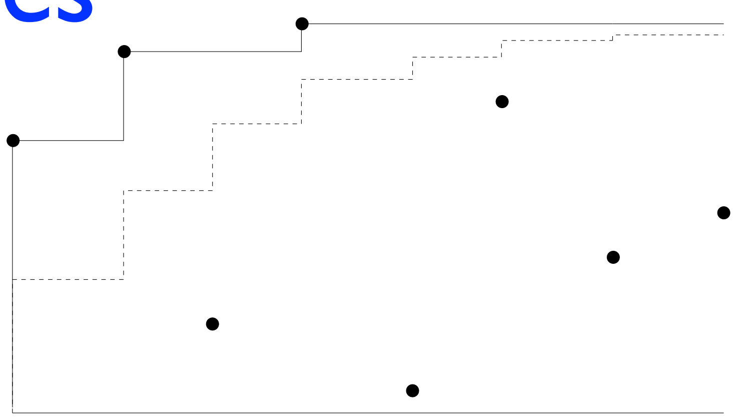
$$V_N = \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{N}\right) - \left(1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{N^2}\right)$$

- Approaches a normal distribution

$$Q_n(N) \to \frac{1}{\sqrt{2\pi \ln N}} \exp\left[-\frac{(n - \ln N)^2}{2 \ln N}\right]$$

Superior Records

- Start with sequence of random variables



$$\{x_1, x_2, x_3, \ldots, x_N\}$$

- Calculate the sequence of records

$$\{X_1, X_2, X_3, \ldots, X_N\} \quad \text{where} \quad X_n = \max(x_1, x_2, \ldots, x_n\}$$

- Compare with the expected average

$$\{A_1, A_2, A_3, \ldots, A_N\} = \{1/2, 2/3, 3/4, \ldots, N/(N+1)\}$$

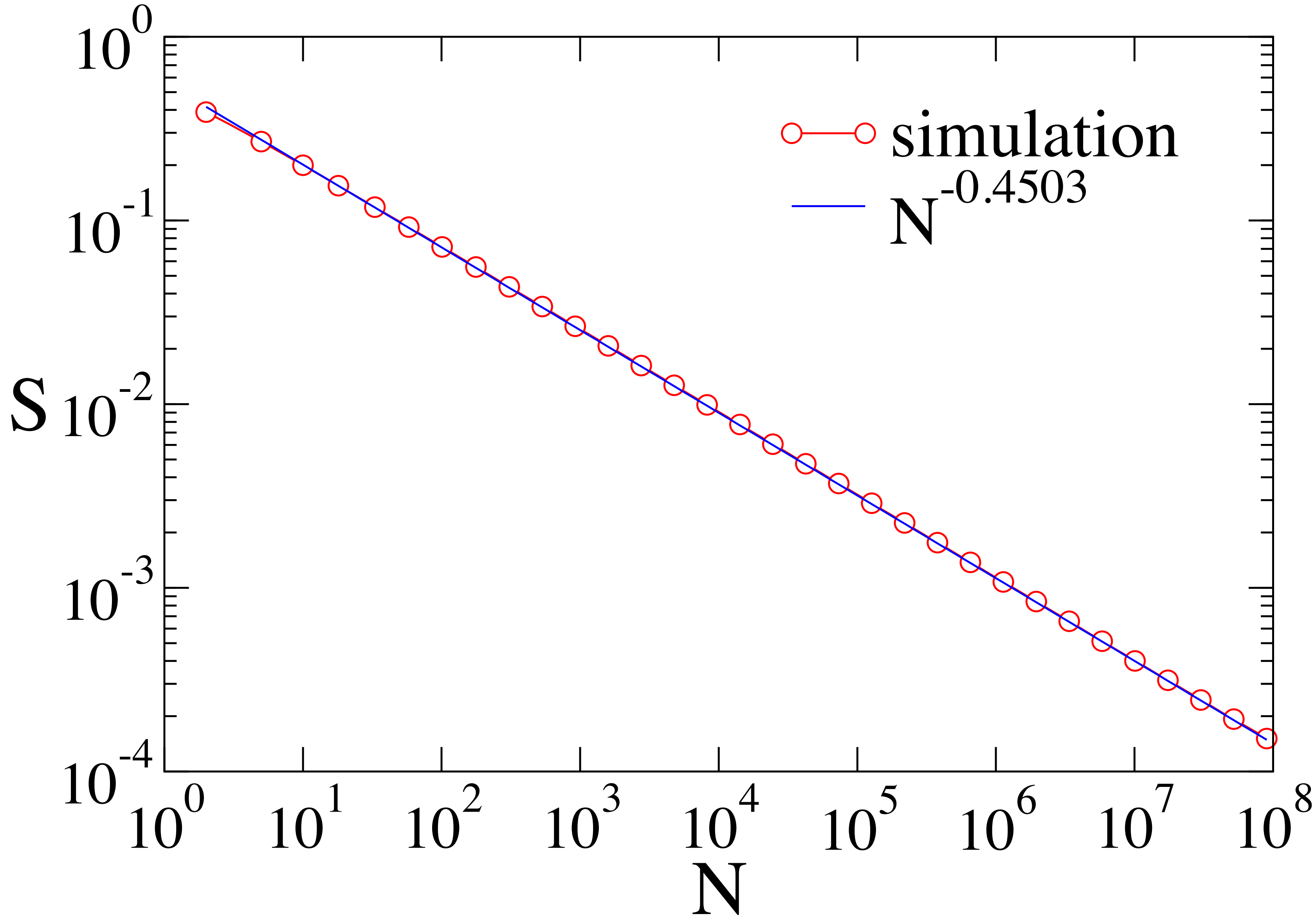
- Superior sequence = records always exceeds average

$$X_n > A_n \quad \text{for all} \quad 1 \leq n \leq N$$

- What fraction S_N of sequences is superior?

measure of "performance"

Numerical simulations



$S_N \sim N^{-\beta}$

$$\boxed{\beta = 0.4503 \pm 0.0002}$$

Power law decay with <u>nontrivial</u> exponent

Distribution of superior records

- Cumulative probability distribution $F_N(x)$ that:
  1. Sequence is superior ($X_n > A_n$ for all n) and
  2. Current record is larger than x ($X_N > x$)
- Gives the desired probability immediately

$$S_N = F_N(A_N)$$

- Recursion equation

$$F_{N+1}(x) = \underbrace{x\,F_N(x)}_{\text{old record holds}} + \underbrace{(1-x)\,F_N(A_N)}_{\text{a new record is set}} \qquad x > A_{N+1}$$

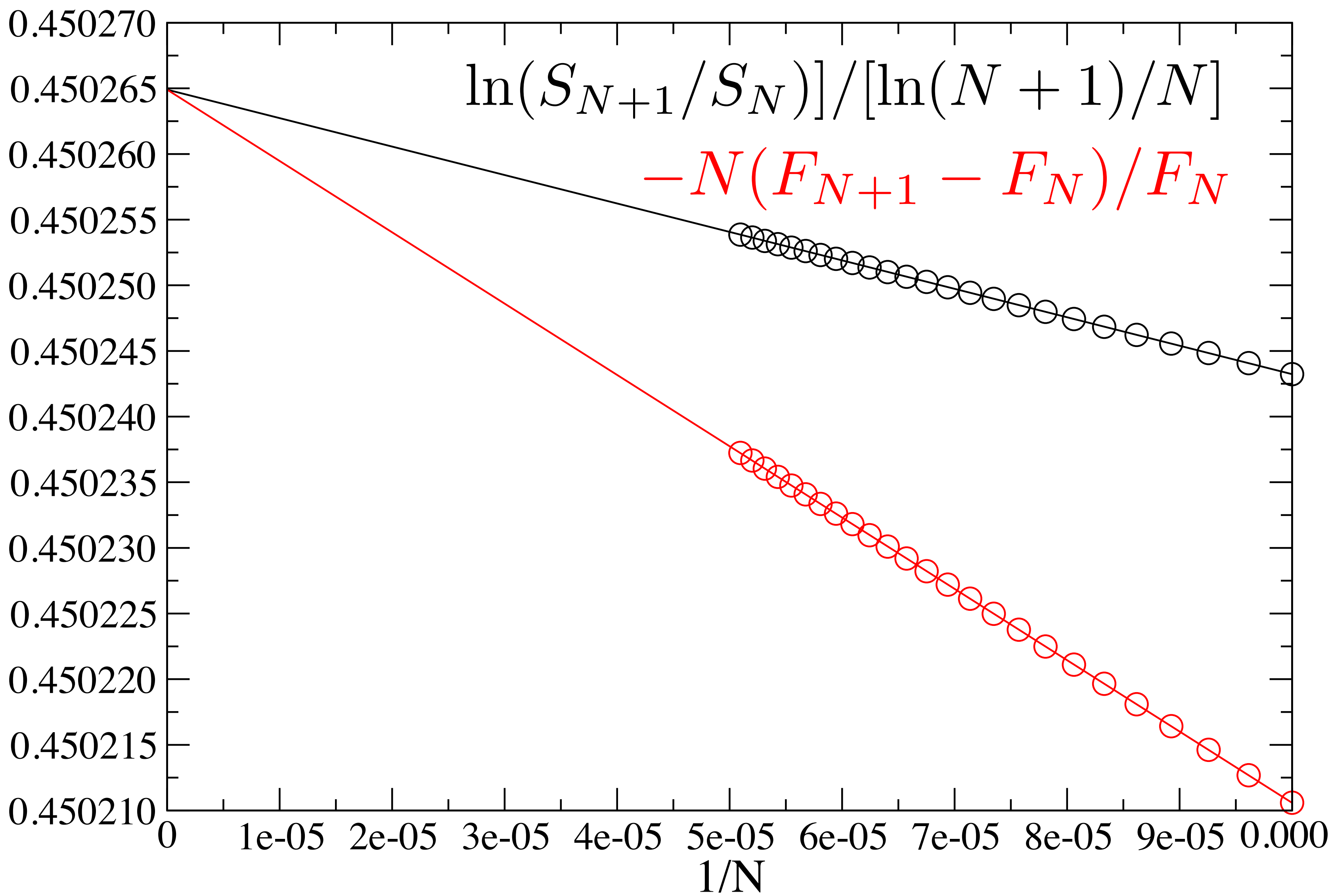
- Recursive solution

$$F_1(x) = 1 - x$$
$$F_2(x) = \tfrac{1}{2}\left(1 + x - 2x^2\right)$$
$$F_3(x) = \tfrac{1}{18}\left(7 + 2x + 9x^2 - 18x^3\right)$$
$$F_4(x) = \tfrac{1}{576}\left(191 + 33x + 64x^2 + 288x^3 - 576x^4\right)$$

$$\xRightarrow{S_N = F_N(A_N)}$$

$$S_1 = \tfrac{1}{2}$$
$$S_2 = \tfrac{7}{18}$$
$$S_3 = \tfrac{191}{576}$$
$$S_4 = \tfrac{35393}{120000}$$

Enumeration



$$\beta = 0.450265$$

Exponent obtained with improved precision

Still, what about the distribution of superior records?

Can the exponent be obtained analytically?

Similarity transformation

- Convert recursion equation

$$F_{N+1}(x) = x\, F_N(x) + (1-x)\, F_N(A_N)$$

into a differential equation (N plays role of time!)

$$\frac{\partial F_N(x)}{\partial N} = (1-x)\,\left[F_N(A_N) - F_N(x)\right]$$

- Seek a similarity solution ($N \to \infty$ limit)

$$F_N(x) \simeq S_N \Phi(s) \quad \text{with} \quad s = (1-x)N$$

boundary conditions $\Phi(0) = 0 \quad \text{and} \quad \Phi(1) = 1 \quad \left(1 - \frac{N}{N+1}\right) N \to 1$

- Similarity function obeys first-order ODE

$$\Phi'(s) + (1 - \beta\, s^{-1})\Phi(s) = 1$$

Similarity solution gives distribution of scaled record

Similarity Solution

- Equation with yet unknown exponent

$$\Phi'(s) + (1 - \beta\, s^{-1})\Phi(s) = 1$$

- General solution

$$\Phi(s) = s \int_0^1 dz\, z^{-\beta} e^{s(z-1)}$$

- Boundary condition dictates the exponent

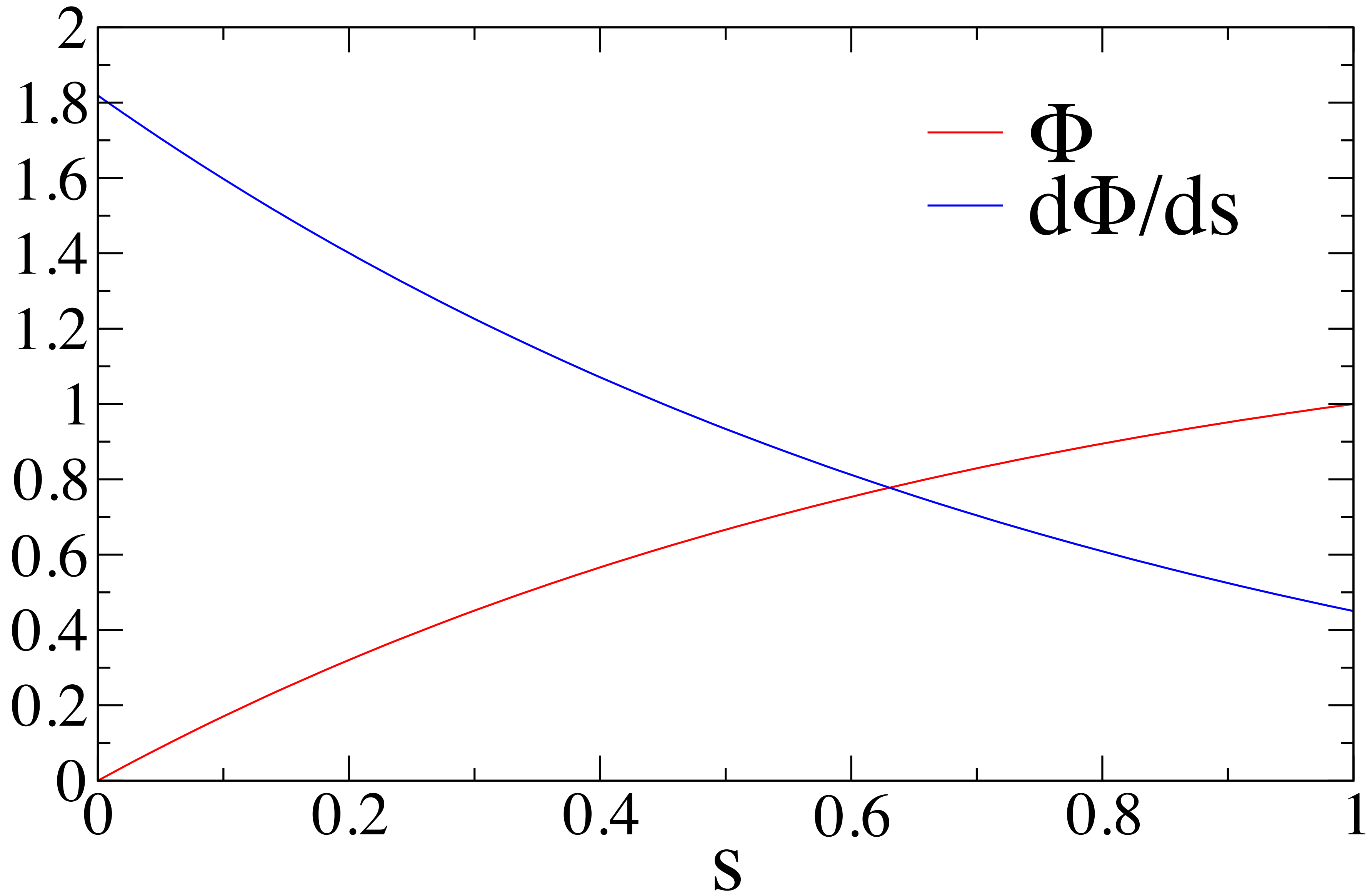
$$\int_0^1 dz\, z^{-\beta} e^{(z-1)} = 1$$

- Root is a transcendental number

$$\beta = 0.450265027495$$

Analytic solution for distribution and exponent

Distribution of records (for superior sequences)



scaling variable $s = (1 - x)N$

The average record

- Similarity function immediately gives average

$$\langle s \rangle = -\int_0^1 ds\, s\, \Phi'(s)$$

- Average record

$$1 - \langle x \rangle \simeq C\, N^{-1}$$

- Constant follows from the similarity function

$$C = \int_0^1 ds\, [1 - \Phi(s)]$$

- Constant is nontrivial

$$C = 0.388476$$

Similarity function characterizes all records statistics

Summary I

- Compare record with expected average
- Superior sequence consistently "outperforms" average
- Probability a sequence is superior decays as power law

$$S_N \sim N^{-\beta}$$

- Exponent is nontrivial, can be obtained analytically

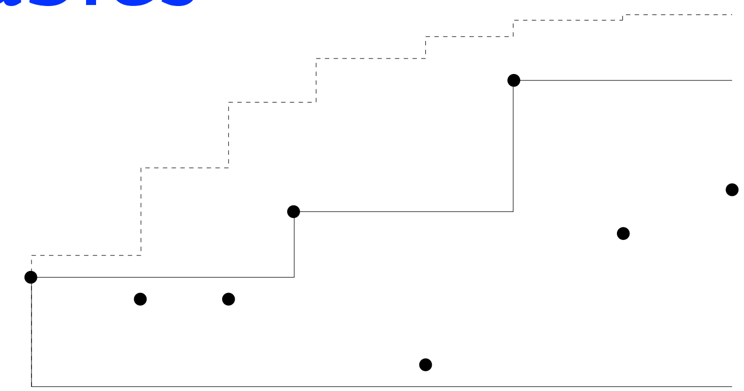
$$\beta = 0.450265$$

- Distribution function can be obtained as well

Inferior records

- Start with sequence of random variables

$$\{x_1, x_2, x_3, \ldots, x_N\}$$



- Calculate the sequence of records

$$\{X_1, X_2, X_3, \ldots, X_N\} \quad \text{where} \quad X_n = \max(x_1, x_2, \ldots, x_n\}$$

- Compare with the expected average

$$\{A_1, A_2, A_3, \ldots, A_N\} = \{1/2, 2/3, 3/4, \ldots, N/(N+1)\}$$

- Inferior sequence = records always below average

$$X_n > A_n \quad \text{for all} \quad 1 \leq n \leq N$$

- What fraction of sequences are inferior?

$$I_N \sim N^{-\alpha}$$

expect power law decay, different exponent

Probability sequence is inferior

- Start with sequence of random variables

$$\{A_1, A_2, A_3, \ldots, A_N\} = \{1/2, 2/3, 3/4, \ldots, N/(N+1)\}$$

- One variable

$$x_1 < \frac{1}{2} \quad \Longrightarrow \quad I_1 = \frac{1}{2}$$

- Two variables

$$x_1 < \frac{1}{2} \quad \text{and} \quad x_2 < \frac{2}{3} \quad \Longrightarrow \quad I_2 = \frac{1}{2} \times \frac{2}{3} = \frac{1}{3}$$

- Recursion equation (no interactions between variables)

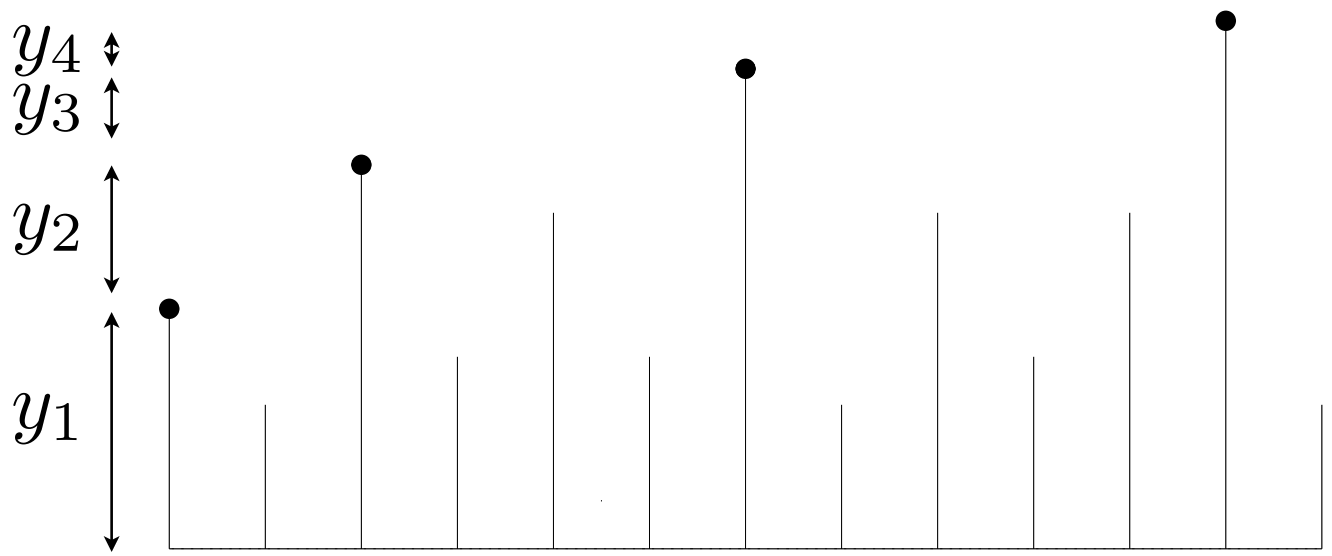
$$I_{N+1} = I_N \frac{N}{N+1}$$

- Simple solution

$$I_N = \frac{1}{N+1} \qquad I_N \sim N^{-1}$$

power law decay with trivial exponent

Incremental Records



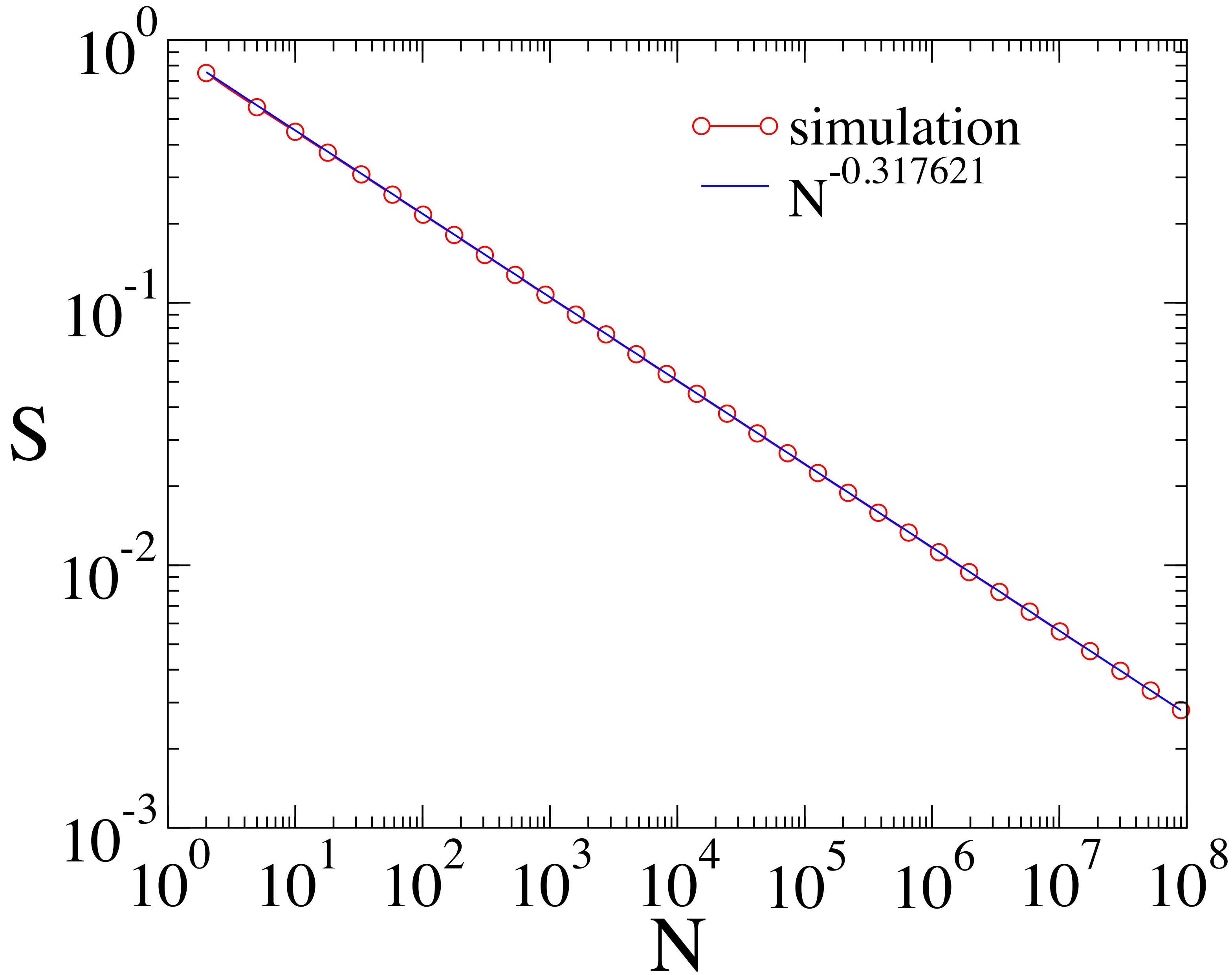
Incremental sequence of records

every record improves upon previous record by yet smaller amount

- No a priori knowledge of distribution, no parameters
- Definition does involve memory
- What fraction S_N of sequences is incremental?

$$S_N \sim N^{-\nu} \qquad \nu = 0.31762101$$

Numerical Simulations



$S_N \sim N^{-\nu}$ $\qquad \nu = 0.3176 \pm 0.0002$

Power law decay with nontrivial exponent

Distribution of records

- Probability a sequence is inferior and record $< x$

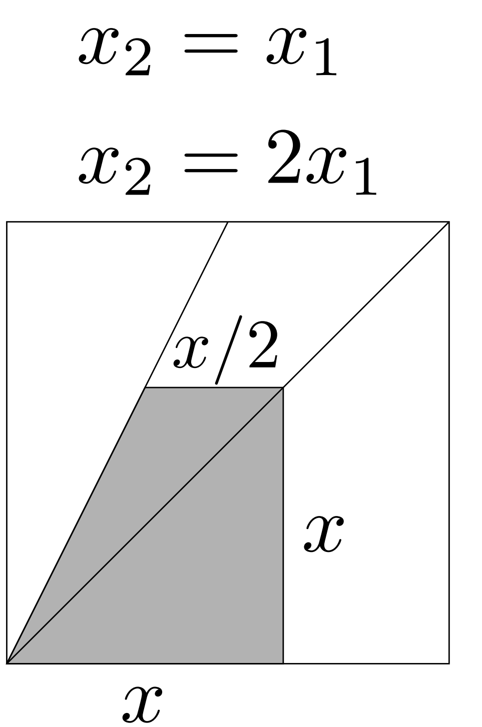
$$G_N(x) \quad \Longrightarrow \quad S_N = G_N(1)$$

- One variable

$$G_1(x) = x \quad \Longrightarrow \quad S_1 = 1$$

- Two variables

$$G_2(x) = \frac{3}{4}\,x^2 \quad \Longrightarrow \quad S_2 = \frac{3}{4}$$



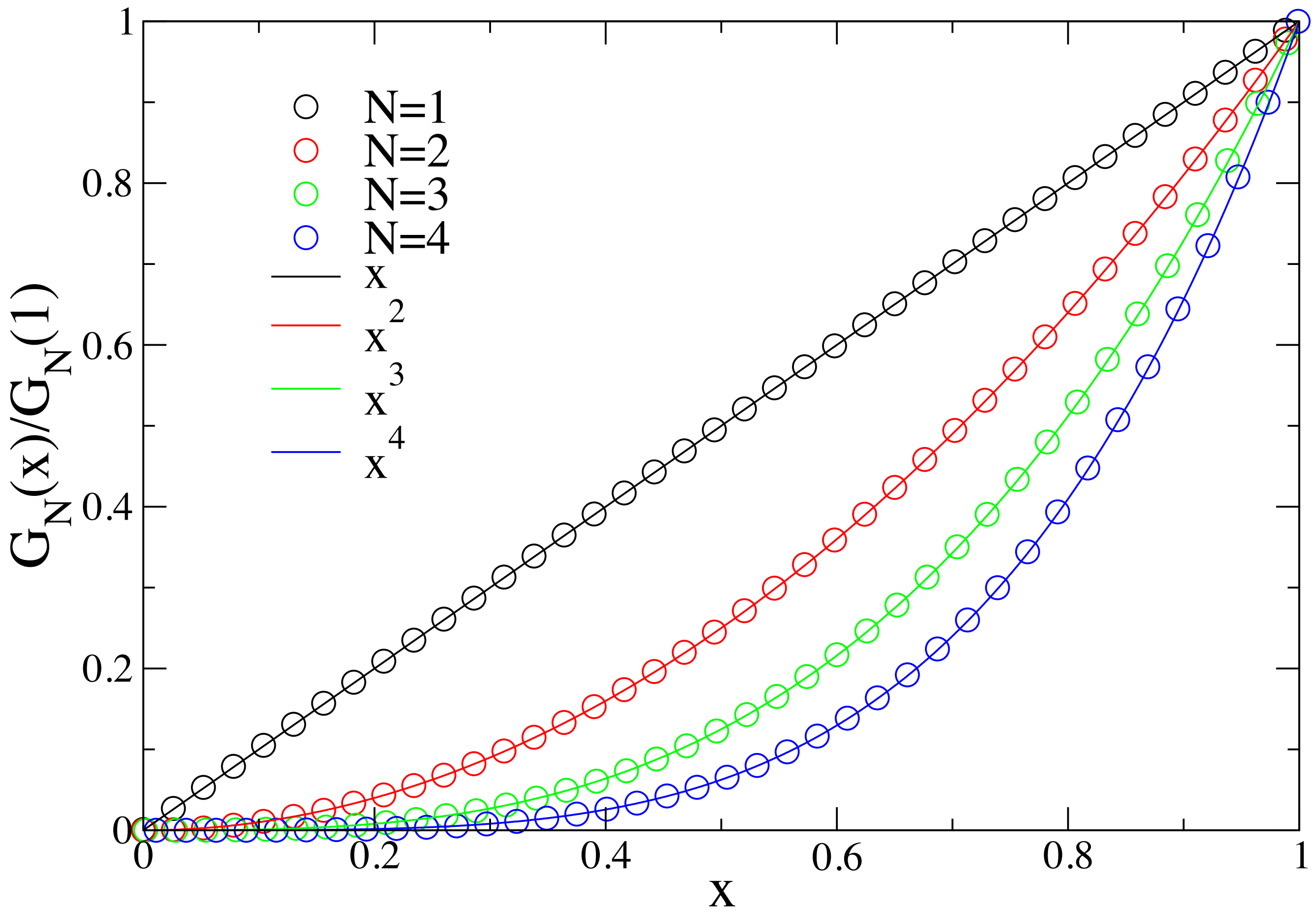
- In general, conditions are scale invariant $\quad x \to a\,x$
- Distribution of records for incremental sequences

$$G_N(x) = S_N\,x^N$$

- Distribution of records for all sequences equals x^N

Statistics of records follow fisher-tippett-gumble!

Scaling behavior



- Distribution of records for incremental sequences

$$G_N(x)/S_N = x^N = [1-(1-x)]^N \to e^{-s}$$

- Same scaling variable

$$s = (1-x)N$$

Exponential similarity function

Distribution of records

- Probability distribution $S_N(x,y)dxdy$ that:
  1. Sequence is incremental
  2. Current record is in range $(x,x+dx)$
  3. Latest increment is in range $(y,y+dy)$ *with* $0<y<x$
- Gives the probability a sequence is incremental

$$S_N = \int_0^1 dx \int_0^x dy\, S_N(x,y)$$

- Recursion equation incorporates memory

$$S_{N+1}(x,y) = \underbrace{x\, S_N(x,y)}_{\text{old record holds}} + \underbrace{\int_y^{x-y} dy'\, S_N(x-y,y')}_{\text{a new record is set}}$$

- Evolution equation includes integral, has memory

$$\frac{\partial S_N(x,y)}{\partial N} = -(1-x)S_N(x,y) + \int_y^{x-y} dy'\, S_N(x-y,y')$$

Similarity transformation

- Assume record and increment scale similarly

$$y \sim 1 - x \sim N^{-1}$$

- Introduce a scaling variable for the increment

$$s = (1-x)N \quad \text{and} \quad z = yN$$

- Seek a similarity solution

$$S_N(x,y) = N^2 S_N \,\Psi(s,z)$$

- Eliminate time out of the master equation

$$\left(2 - \nu + s + s\frac{\partial}{\partial s} + z\frac{\partial}{\partial z}\right)\Psi(s,z) = \int_z^\infty dz' \,\Psi(s+z,z')$$

Factorizing solution

- Assume record and increment decouple

$$\Psi(s,z) = e^{-s}\, f(z)$$

- Substitute into equation for similarity solution

$$\left(2-\nu+s+s\frac{\partial}{\partial s}+z\frac{\partial}{\partial z}\right)\Psi(s,z) = \int_z^\infty dz'\,\Psi(s+z,z')$$

- First order integro-differential equation

$$zf'(z)+(2-\nu)f(z) = e^{-z}\int_z^\infty f(z')dz'$$

- Cumulative distribution of scaled increment

$$g(z) = \int_z^\infty f(z')dz'$$

- Convert into a second order differential equation

$$zg''(z)+(2-\nu)g'(z)+e^{-z}g(z)=0$$

$$g(0)=1$$
$$g'(0)=-1/(2-\nu)$$

Distribution of increment

- Assume record and increment decouple

$$zg''(z) + (2-\nu)g'(z) + e^{-z}g(z) = 0$$

$$g(0) = 1$$
$$g'(0) = -1/(2-\nu)$$

- Two independent solutions

$$g(z) = z^{\nu-1} \quad \text{and} \quad g(z) = \text{const.} \quad \text{as} \quad z \to \infty$$

- The exponent is determined by the tail behavior

$$\nu = 0.317621$$

- The distribution of increment has a broad tail

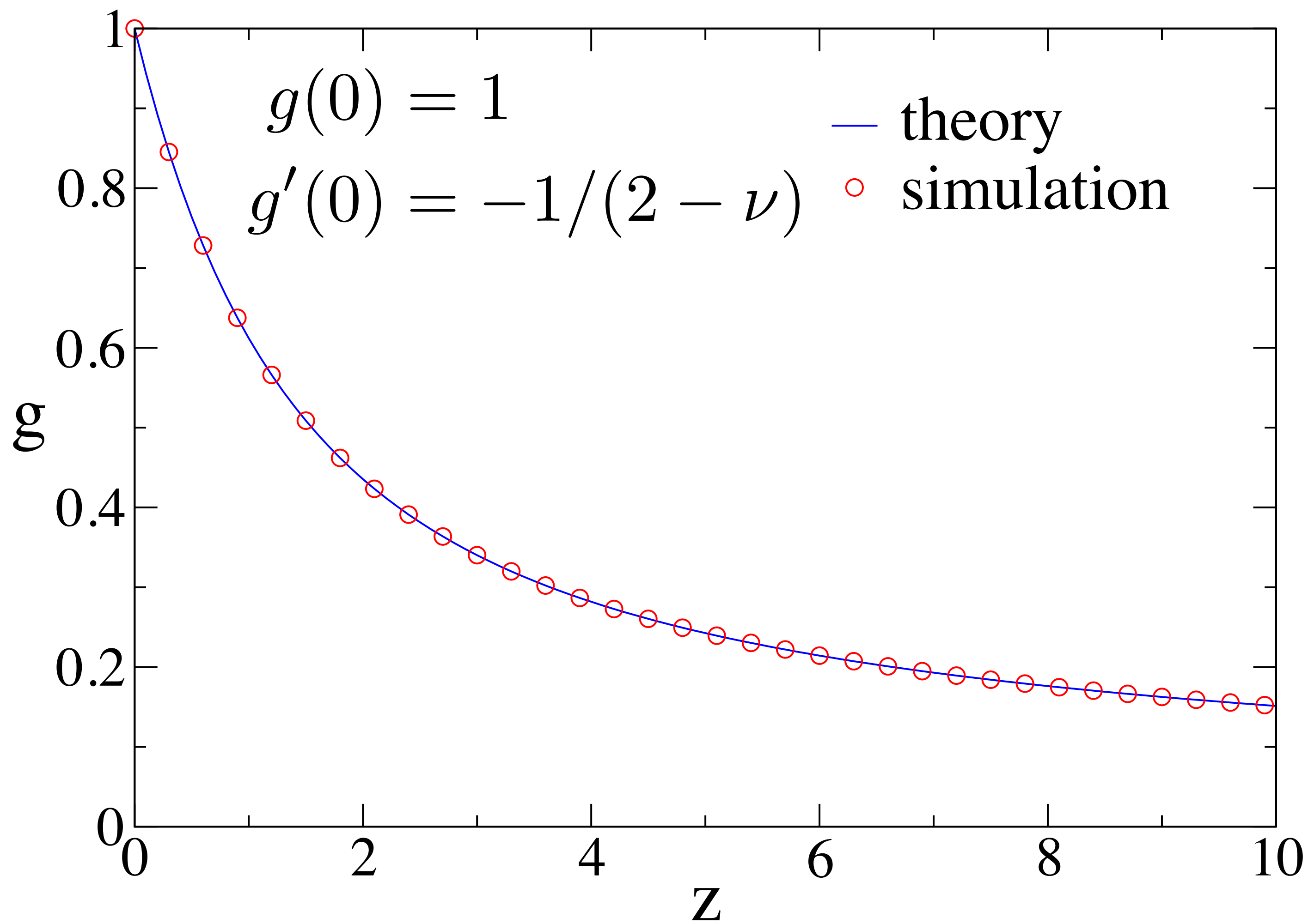
$$P_N(y) \sim N^{-1} y^{\nu-2}$$

Increments can be relatively large

problem reduced to second order ODE

Numerical confirmation

Monte Carlo simulation versus integration of ODE



$$\frac{\langle sz \rangle}{\langle s \rangle \langle z \rangle} \to 1$$

Increment and record become uncorrelated

Summary II

- Incremental sequences: improvement in record diminishes monotonically
- Distribution of record is narrow (exponential)
- Distribution of increment is broad (power law)
- Increment and record become uncorrelated when the sequence becomes very large
- Analytic treatment incorporates memory
- Problem reduces to a second order ODE
- Exponent can be obtained analytically

General distributions

- Arbitrary distribution function
- Single parameter contains information about tail

$$\alpha = \lim_{N\to\infty} N \int_{A_N}^{\infty} dx\, \rho(x)$$

- Equals the exponent for inferior sequences

$$I_N \sim N^{-\alpha}$$

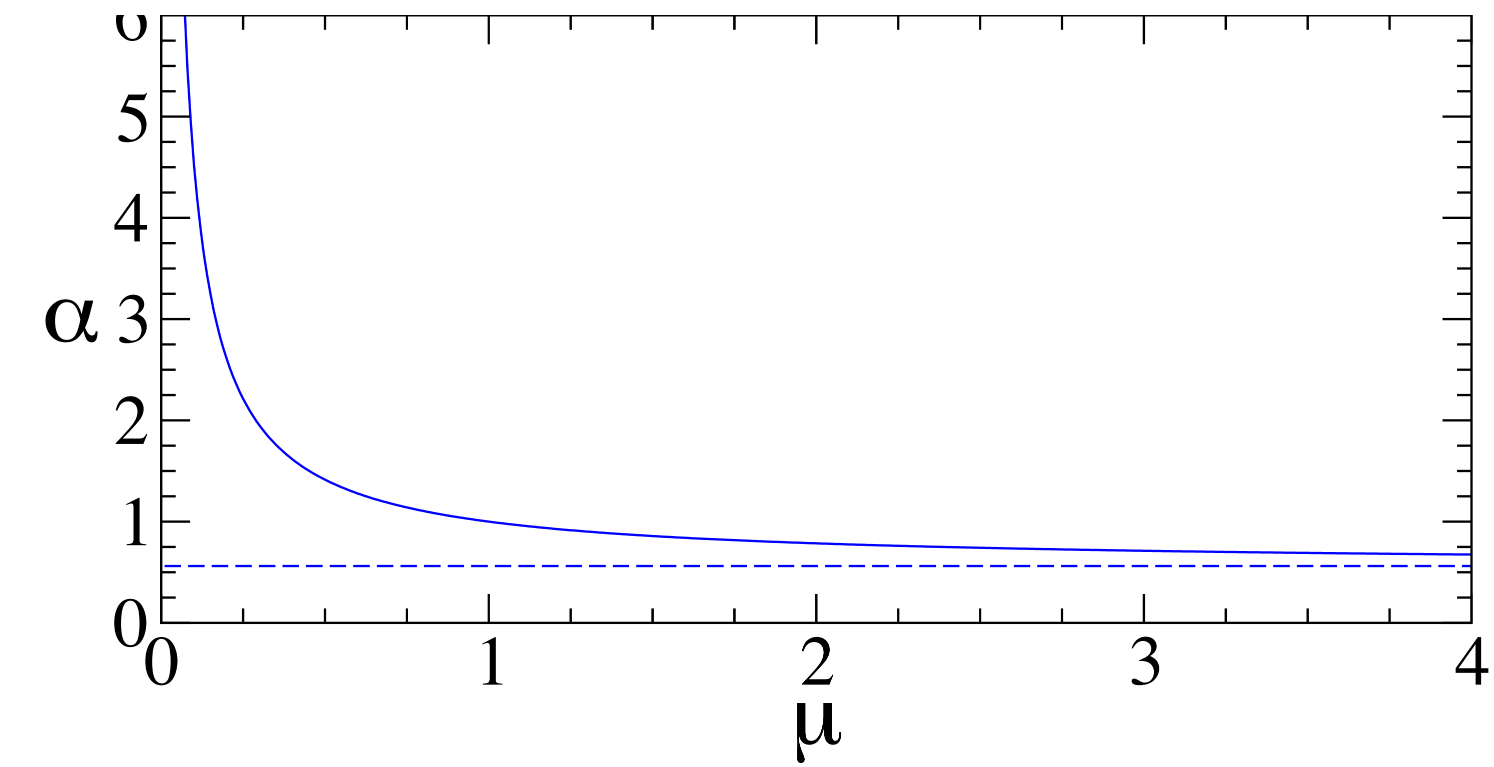
- Exponent for superior sequences

$$\alpha \int_0^1 dz\, z^{-\beta} e^{\alpha(z-1)} = 1$$

- Powerlaw distributions (compact support)

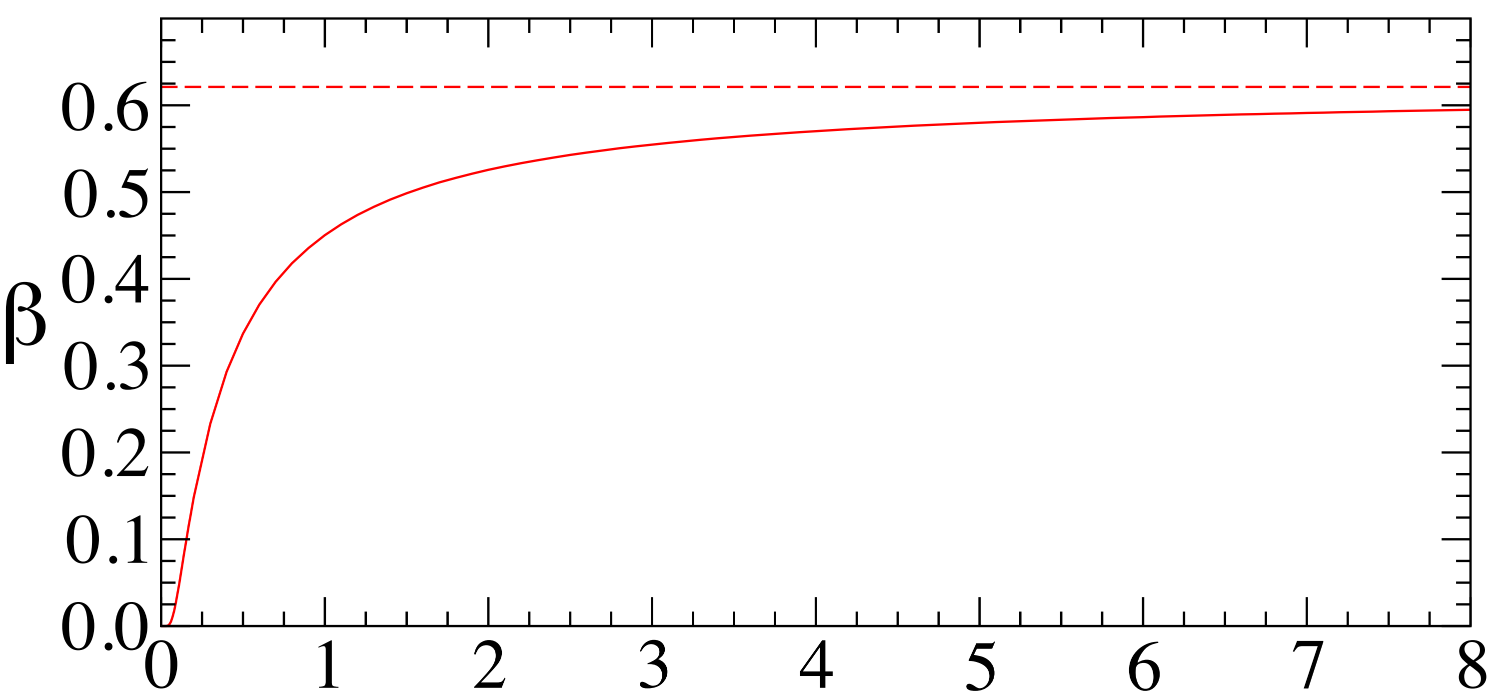
$$R(x) \sim (1-x)^{\mu} \qquad \Longrightarrow \qquad \alpha = \left[\Gamma\left(1+\tfrac{1}{\mu}\right)\right]^{\mu}$$

Continuously varying exponents



$$\alpha_{\mathrm{min}} \le \alpha < \infty$$

$$\alpha_{\mathrm{min}} = e^{-\gamma} = 0.561459$$



$$\beta_{\mathrm{max}} = 0.621127$$

$$0 < \beta \le \beta_{\mathrm{max}}$$

Tail of distribution function controls record statistics

Conclusions

- Studied persistent configuration of record sequences
- Linear evolution equations (but nonlocal/memory)
- Dynamic formulation: treat sequence length as time
- Similarity solutions for distribution of records
- Probability of persistent configuration (superior, inferior, incremental) decays as a power-law
- Power laws exponents are generally nontrivial
- Exponents can be obtained analytically
- Tail of distribution function controls record statistics