PROBABILISTIC STRUCTURES
IN EVOLUTION

DFG SPP 1590

COLLABORATIVE RESEARCH CENTER | SFB 680
Molecular Basis of
Evolutionary Innovations

Adaptive walks and record processes

Joachim Krug
Institute for Theoretical Physics, University of Cologne

joint work with S.-C. Park, J. Neidhart, S. Nowak and I.G. Szendro

Nonequilibrium Statistical Physics of Complex Systems, KIAS, Seoul, 6.7.2016

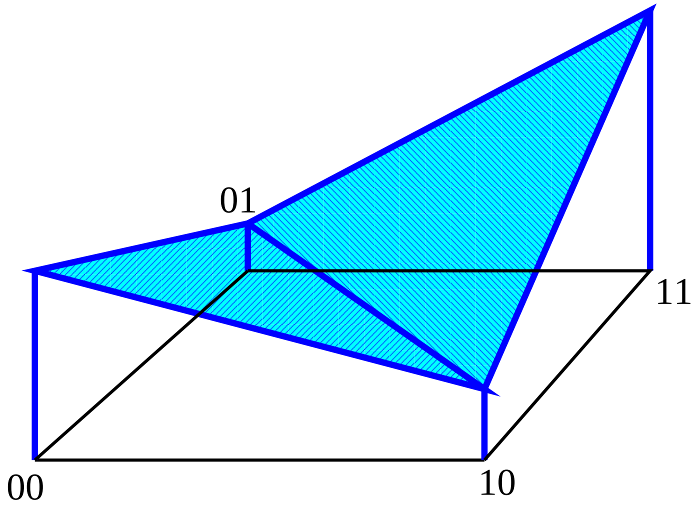
Fitness landscapes

- Genotypes are binary sequences $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_L)$ with $\sigma_i \in \{0, 1\}$ (presence/absence of mutation).

- Together with the Hamming distance $d(\sigma, \sigma') = \sum_{i=1}^{L} 1 - \delta_{\sigma_i, \sigma_i'}$ this defines the Hamming space $\mathbb{H}_2^L$ which is the L-dimensional hypercube

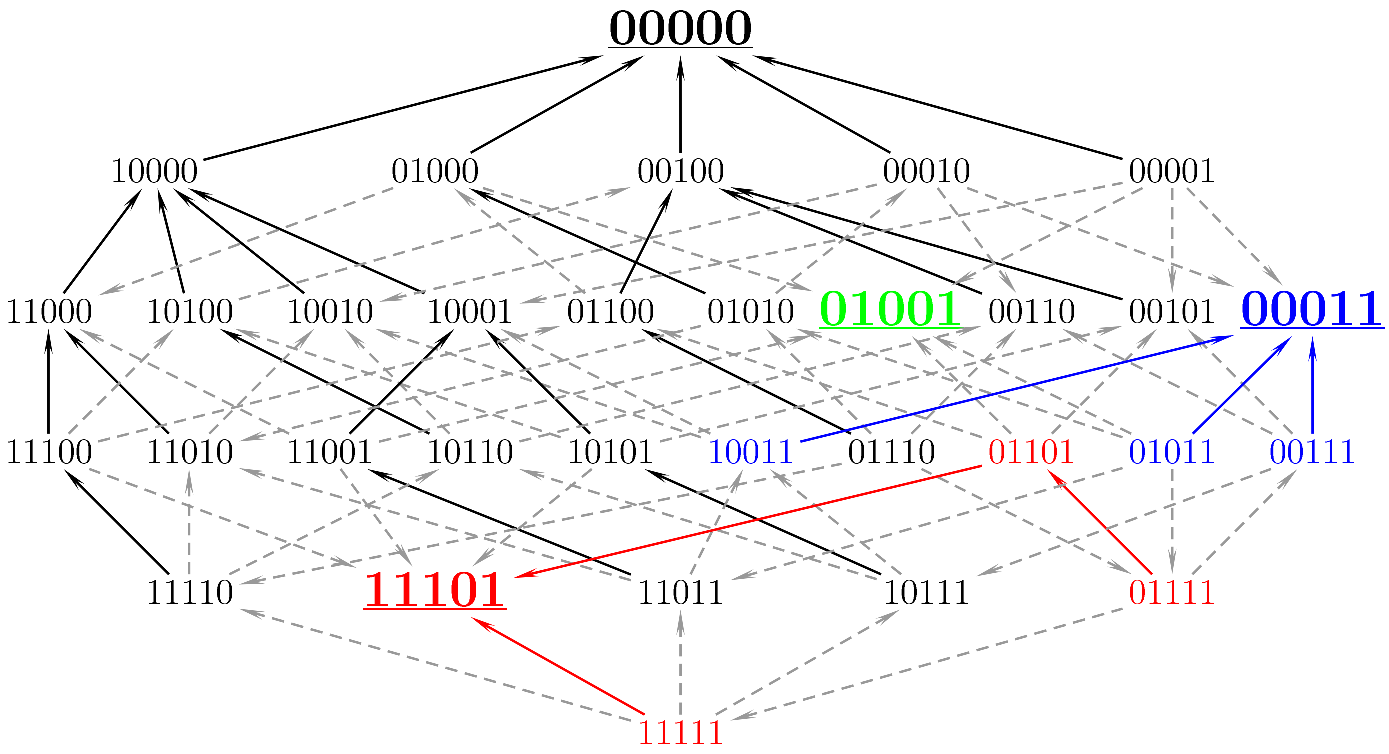
- A fitness landscape is a real-valued function $f(\sigma)$ on $\mathbb{H}_2^L$

- Interactions between the fitness effects of different mutations may induce multiple adaptive peaks:



Example: The *Aspergillus niger* fitness landscape

J.A.G.M. de Visser, S.C. Park, JK, American Naturalist **174**, S15 (2009)

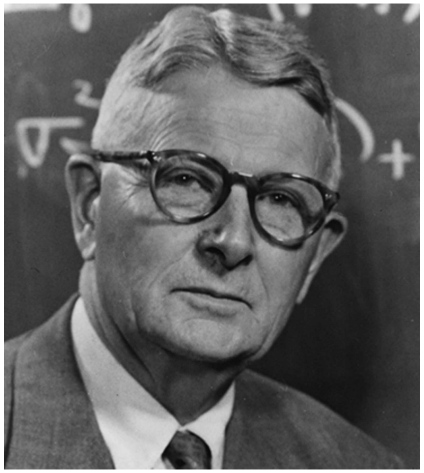


- Combinations of 5 individually deleterious marker mutations

- Arrows point towards higher fitness

- For a survey of other examples see J.A.G.M. de Visser, JK, Nat. Rev. Gen. 2014

Evolutionary accessibility



"In a rugged field of this character, selection will easily carry the species to the nearest peak, but there will be innumerable other peaks that will be higher but which are separated by valleys..."

Sewall Wright, 1932

- Accessibility of fitness landscapes can be quantified by the number of local fitness peaks or the number of fitness-monotonic pathways

  Franke et al. 2011, Hegarty & Martinsson 2014, Berestycki, Brunet, Shi 2016...

- However, even if uphill pathways exist it is not clear if populations can find them

- Here we take a dynamic viewpoint and consider populations navigating a rugged fitness landscape through adaptive walks with local rules

SSWM dynamics

- SSWM = Strong Selection/Weak Mutation — Gillespie 1983, Orr 2002

- Weak mutation: Each new mutation goes to fixation or is lost before the next one arrives

- Strong selection: The fixation probability of a mutation of selective advantage s in a population of size N is

$$\pi(s,N) \approx \frac{1-\exp[-2s]}{1-\exp[-2Ns]} \approx 1-\exp[-2s]$$

for $s > 0$ and $\pi = 0$ for $s \leq 0$, provided $N|s| \gg 1$

- Under these conditions the population performs an uphill adaptive walk in sequence space that terminates at a local fitness maximum

- Formally, an adaptive walk is a Markov chain on $\mathbb{H}_2^L$ with absorption at local maxima

Adaptive walks

- Four flavors of adaptive walks differing in their transition probabilities:

  True Adaptive Walk (TAW)
  Transition rate is proportional to the fitness difference between the resident and mutant genotype $(s \ll 1)$ Gillespie 1983, Orr 2002

  Random Adaptive Walk (RAW) Macken & Perelson 1989
  All fitter genotypes are chosen with equal probability $(s \longrightarrow \infty)$

  Greedy Adaptive Walk (GAW) Orr 2003
  The most fit genotype is chosen deterministically

  Reluctant Adaptive Walk (RELAW)
  The least fit among the fitter genotypes is chosen deterministically
  Bussolari et al. 2003

- Of interest is the length ℓ (= mean number of steps) and height f^* (= mean achieved fitness) of such walks

Walk length in uncorrelated landscapes

In the uncorrelated House-of-Cards/Mutational Landscape model fitness values are i.i.d. random variables. The following results refer to walks starting at a low fitness genotype:

- RAW: $\ell \approx \ln(L) + 0.099$ for large L Flyvbjerg & Lautrup 1992
- GAW: $\ell \to \sum_{k=1}^{\infty} \frac{1}{k!} = e - 1 \approx 1.71828...$ Orr 2003
- RELAW: $\ell \to L + \mathcal{O}(1)$ S. Nowak & JK 2015
- TAW length asymptotics depends on the extreme value index κ of the fitness distribution according to J. Neidhart & JK 2011, Jain 2011

$$\ell \approx \frac{1-\kappa}{2-\kappa} \ln(L) + c_\kappa \quad \text{for} \quad \kappa < 1.$$

- For relative initial fitness $f_0 \in [0,1]$ let $L \to (1 - f_0)L$

Adaptive walks and record processes: i.i.d. case

- For $L \longrightarrow \infty$ the RAW never stops but remains well defined as a stochastic process: The $k+1$'th fitness value f_{k+1} along the walk is a random draw from the fitness distribution $P(f)$ conditioned on $f > f_k$, hence a record

- For finite L the RAW stops when $1 - P(f_k) \approx \frac{1}{L}$, which implies that the time elapsed in the record process is $\sim L$

- At this point the number of records $\approx$ number of steps in the walk is

$$\ell = \ln L + \mathscr{O}(1)$$

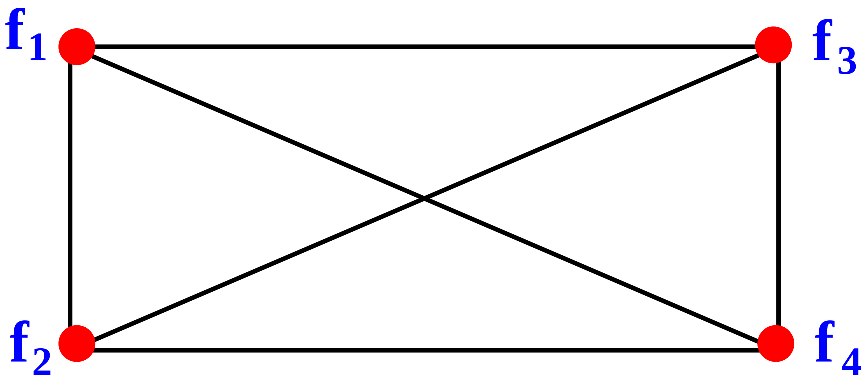
- Like the distribution of record numbers, the distribution of walk lengths is Poisson with mean $\ln L$ Flyvbjerg & Lautrup 1992

- For the GAW with $L \longrightarrow \infty$, the probability that the walk takes at least k steps is equal to the probability $\frac{1}{k!}$ that k i.i.d. random numbers are increasingly ordered S.-C. Park, JK, JTB 2016

The Gillespie approximation

Gillespie 1983

- A precise relation between adaptive walks and record processes holds when the genotype space is a complete graph:



- The order in which genotypes are probed by mutations defines a permutation of the fitness values $f_1, f_2, \ldots, f_L$ and the number of walk steps is equal to the number of records - 1

- The expected number of steps is $\sum_{k=2}^{L} \frac{1}{k} \approx \ln L + \gamma - 1 \approx \ln L - 0.42$

Orr 2002

- The approximation by a complete graph is correct to leading order also for the other walk types

J. Neidhart & JK 2011

Adaptive walks on correlated fitness landscapes

The Rough Mount Fuji model

- Linear (“Mt. Fuji”) landscape with a random component Aita et al. 2000

$$f(\sigma) = cd(\sigma, \sigma^{(0)}) + \eta(\sigma), \qquad c > 0$$

η: i.i.d. random variables $\sigma^{(0)}$: reference sequence

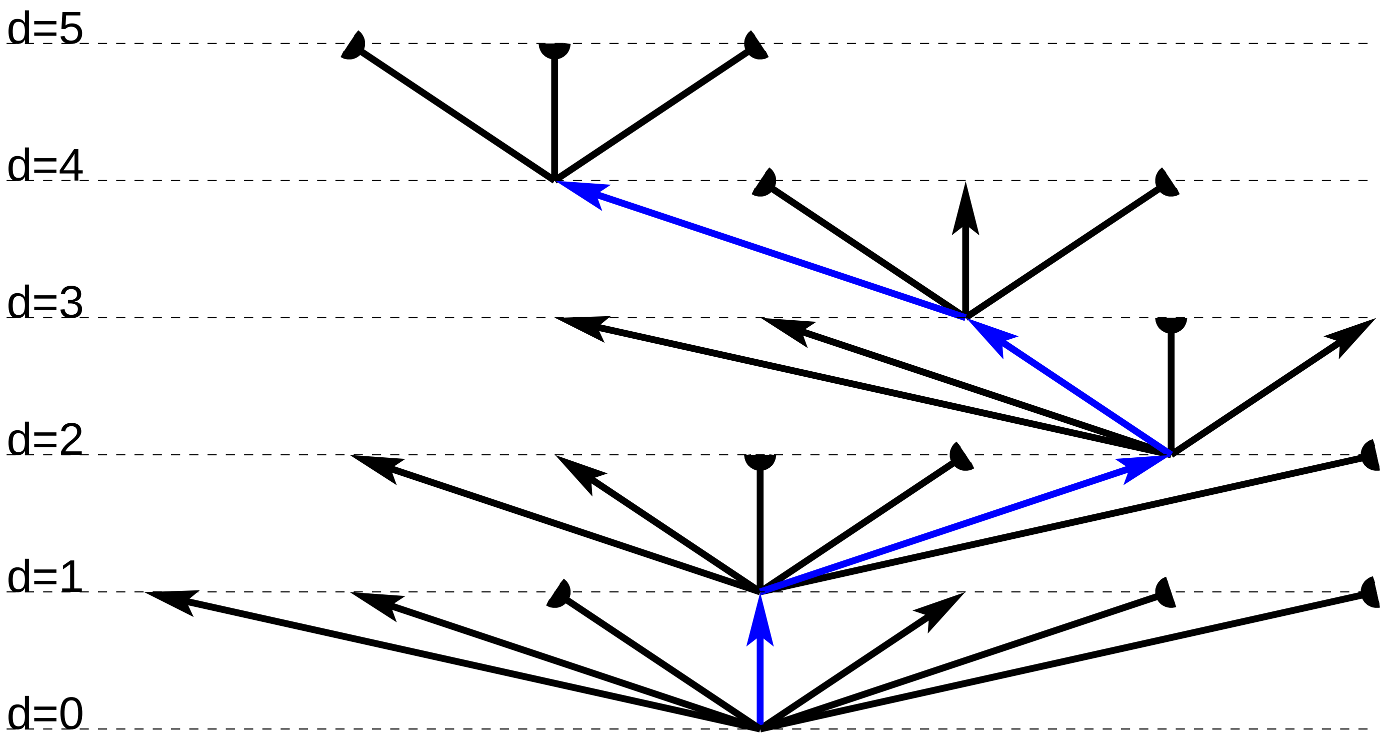
- Fitness-monotonic paths from the reference sequence to the global maximum are certain to exist for any $c > 0$ Hegarty & Martinsson 2014

- How large does the fitness gradient c have to be to allow an adaptive walk to traverse the entire landscape?

Random adaptive walks on the RMF landscape

S.-C. Park, I.G. Szendro, J. Neidhart, JK, Physical Review E **91**, 042707 (2015)

- RAW starts from the reference sequence $\sigma^{(0)}$ and takes only 'uphill' steps that increase $d(\sigma,\sigma^{(0)})$, which is a good approximation if $\ell \ll L$



Random adaptive walks on the RMF landscape

S.-C. Park, I.G. Szendro, J. Neidhart, JK, Physical Review E **91**, 042707 (2015)

- RAW starts from the reference sequence $\sigma^{(0)}$ and takes only 'uphill' steps that increase $d(\sigma, \sigma^{(0)})$, which is a good approximation if $\ell \ll L$

- Then the joint probability $Q_l(y,L)$ that the walk takes at least l steps and reaches a genotype with random fitness component y satisfies

$$Q_{l+1}(y,L) = p(y) \int_{-\infty}^{y+c} dx\, Q_l(x,L) \frac{1-P(x-c)^{L-l}}{1-P(x-c)}$$

- For $L \to \infty$ this reduces to a recursion relation for a modified record process, where the condition for the $k+1$'th record reads $Y_{k+1} > Y_k - c$

- This is known as the δ-exceedance record process with $\delta = -c$

Balakrishnan, Balasubramanian, Panchapakesan 1996

δ-exceedance records and δ-records

- Various modified record processes have been introduced to account for effects of measurement error and noise Edery et al. 2013

- For records from i.i.d. sequences, the most studied model are δ-records defined by the condition

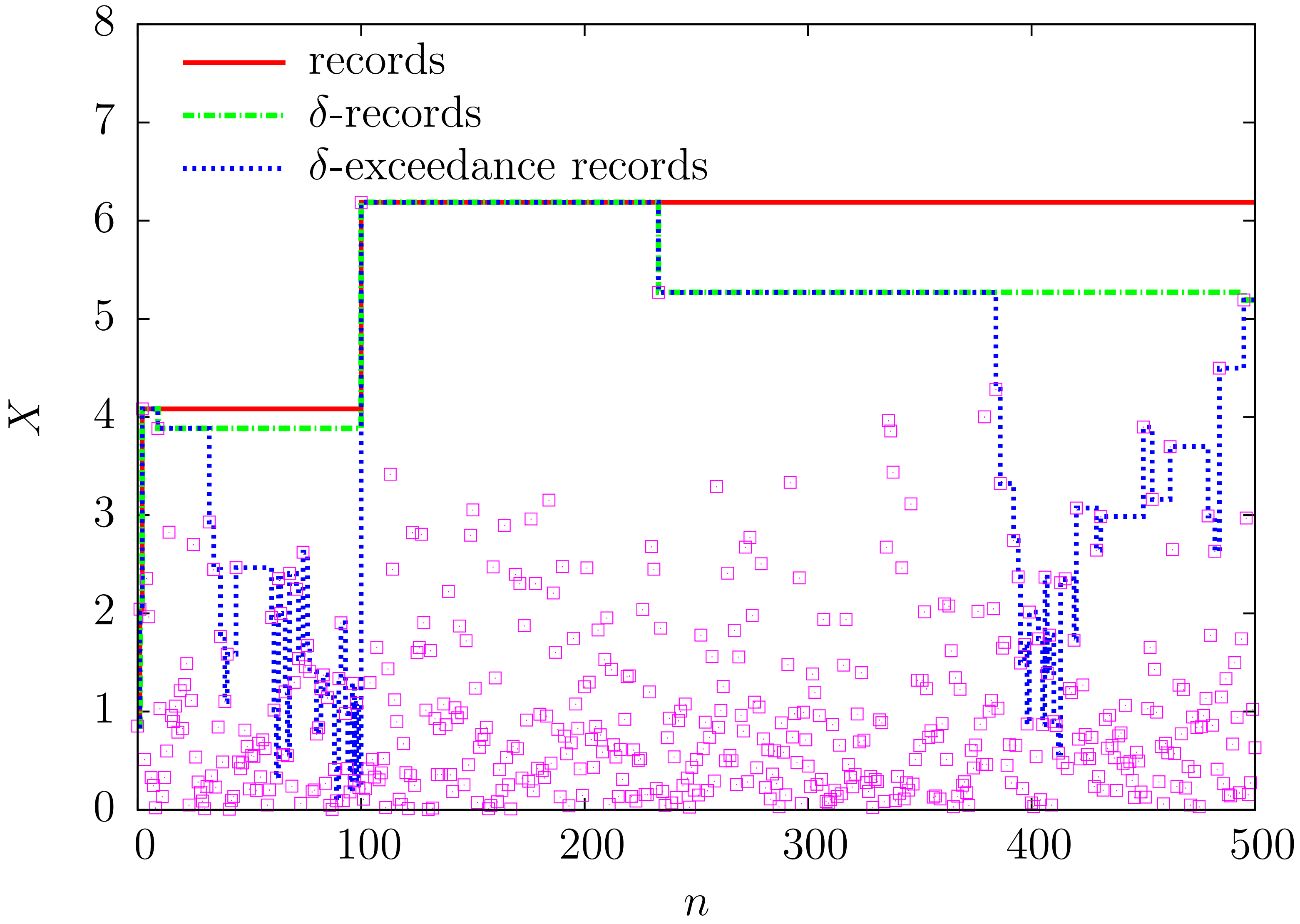
$$X_n > \max\{X_1, X_2, \ldots, X_{n-1}\} + \delta$$

for the occurrence of a δ-record at time n Gouet et al. 2007

- For δ-records the threshold for record occurrence is defined in terms of the true record sequence, which is non-stationary and unbounded whenever the underlying distribution has unbounded support

- In contrast, for the δ-exceedance record process with $\delta < 0$ the threshold can decrease and the process can enter a stationary phase even for unbounded RV's

δ-exceedance records and δ-records



- Sample paths for exp(1) random variables and $\delta = -c = -1$

Phase transition for exponential RV's

- For exponential random variables with unit mean the distribution $Q_l(y)$ of the l'th record value is

$$Q_l(y) = -\frac{d}{dy}\left[\sum_{n=0}^{l} y\frac{(y+cn)^{n-1}}{n!}e^{-y-cn}\right]$$

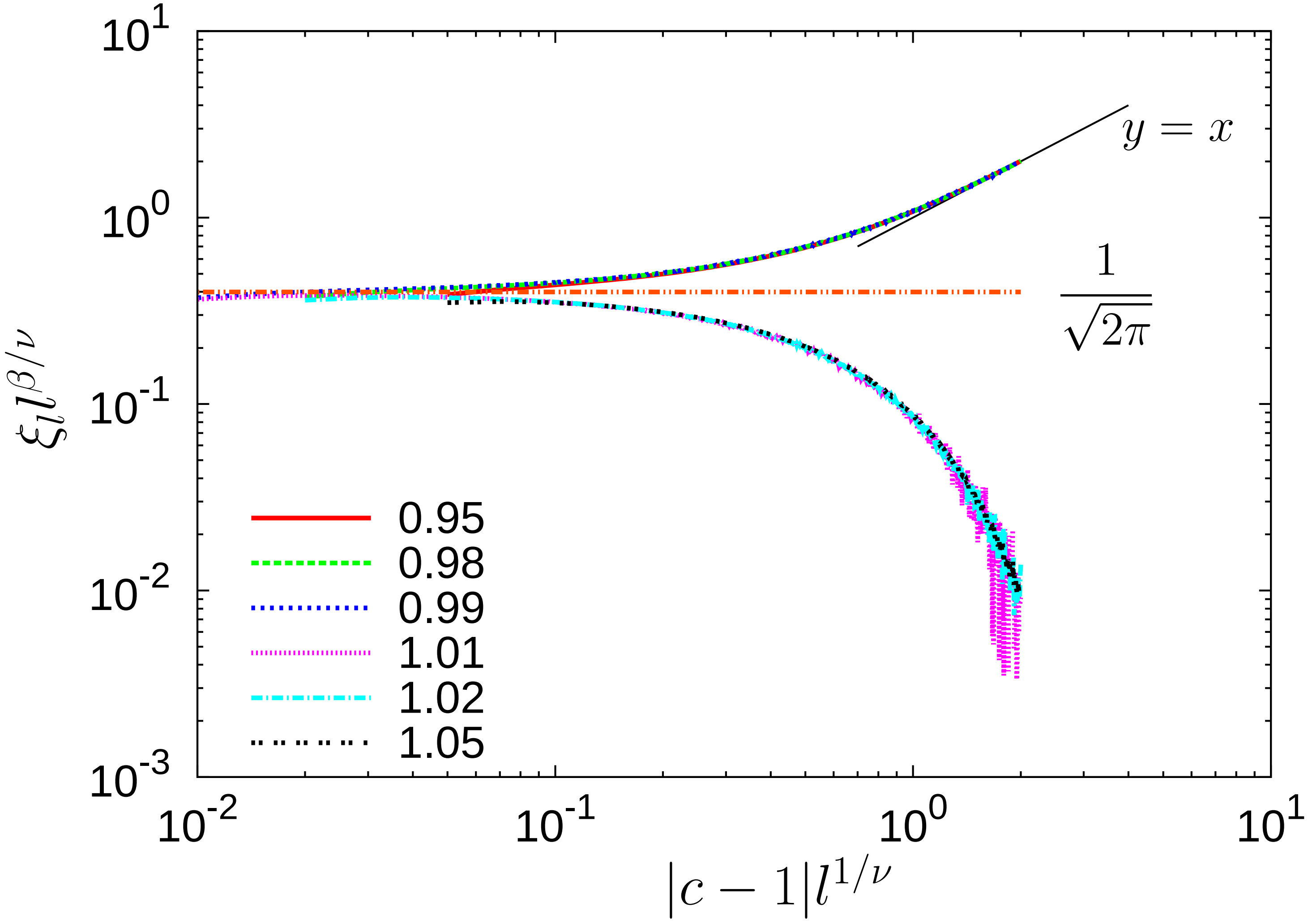
- Expected l'th record value displays a phase transition at $c=1$:

$$z_l \equiv \langle y\rangle_l \approx \begin{cases} (1-c)l, & c<1 \\ \sqrt{2l/\pi}, & c=1, \\ \text{const.}, & c>1. \end{cases}$$

and the mean adaptive walk length behaves as

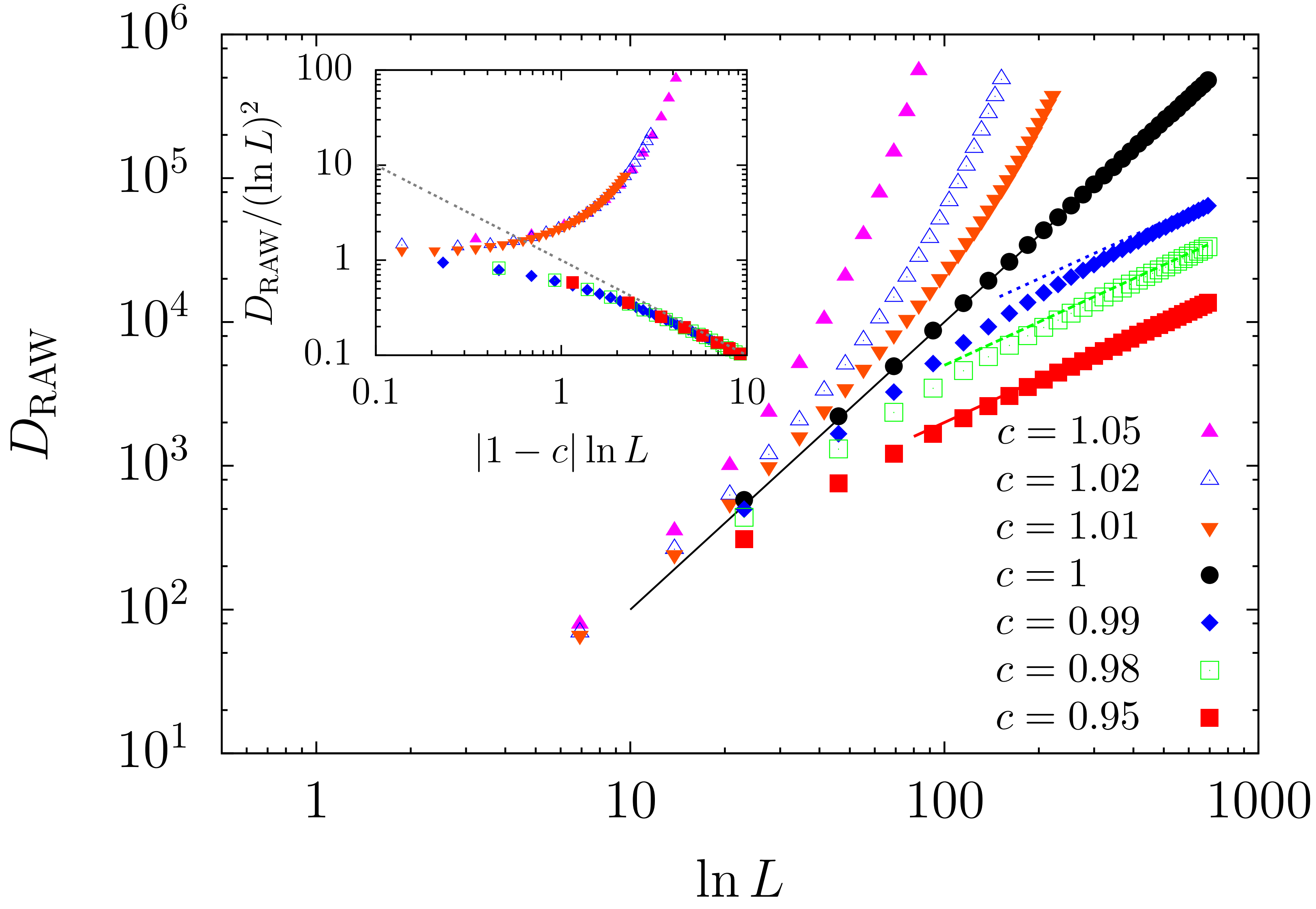
$$\ell \propto \begin{cases} \ln L/(1-c), & c<1 \\ (\ln L)^2, & c=1, \\ O(L), & c>1. \end{cases}$$

Critical behavior



- Scaling plot of order parameter $\xi_l = z_l - z_{l-1}$ with $\beta = 1$ and $\nu = 2$

Transition in the adaptive walk length



Other distributions

- For general distributions with unbounded support, the mean record value z_l satisfies the recursion relation

$$z_{l+1} - z_l = \int_{-\infty}^{\infty} \frac{Q_{l+1}(y)}{h(y)} dy - c,$$

where $h(x) = p(x)/[1 - P(x)]$ is the hazard function.

- Assuming that Q_l is well concentrated, the integral can be replaced by $1/h(z_{l+1})$ which is evaluated asymptotically for large z_l

- This analysis shows that the δ-exceedance record process becomes stationary for any $c > 0$ if the tail of $p(y)$ is thinner than exponential, but never for tails fatter than exponential.

- Special role of the exponential distribution reflects that the spacing between subsequent i.i.d. record values is asymptotically constant in this case.

Generalized δ-exceedance record process

S.-C. Park, JK, J. Phys. A (2016)

- Generalize the condition for the $k+1$'th record to $Y_{k+1} > Y_k - \delta_k$ where $\delta_k > 0$ is a deterministic sequence called the handicap

- If $\delta_k = c(k+1)^{b-1}$ with $b > 0$, the sequence of handicaps matches the spacing between subsequent i.i.d. records for distributions of the form

$$P(x) = 1 - \exp[-x^{\alpha}]$$

  with $\alpha = 1/b$, and the exponential case is $\alpha = b = 1$

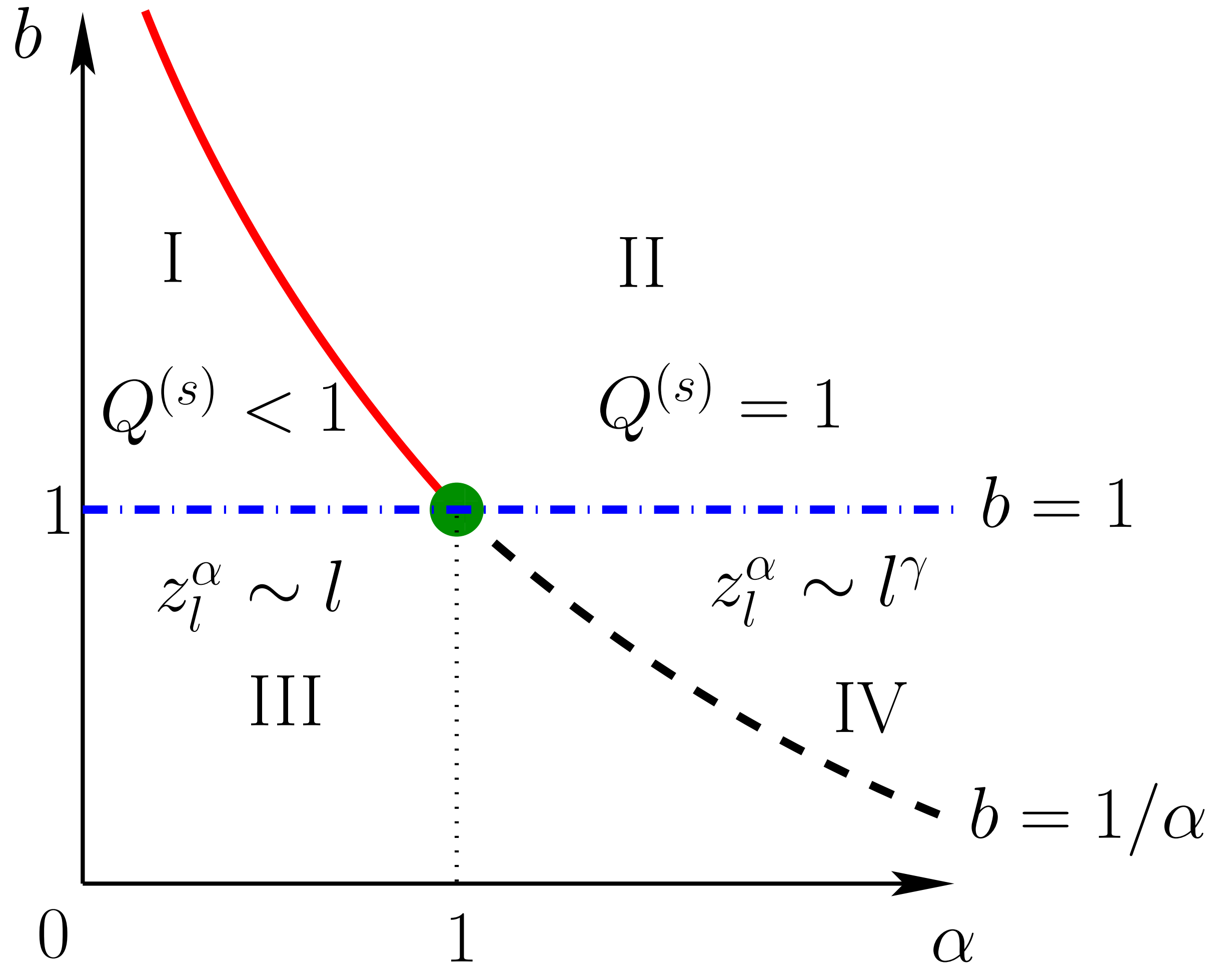
- In biological terms this corresponds to replacing the linear "Mt. Fuji" landscape by a nonlinear (epistatic) deterministic fitness profile

Wiehe 1997

- Epistasis is synergistic/positive (antagonistic/negative) for $b > 1$ ($b < 1$)

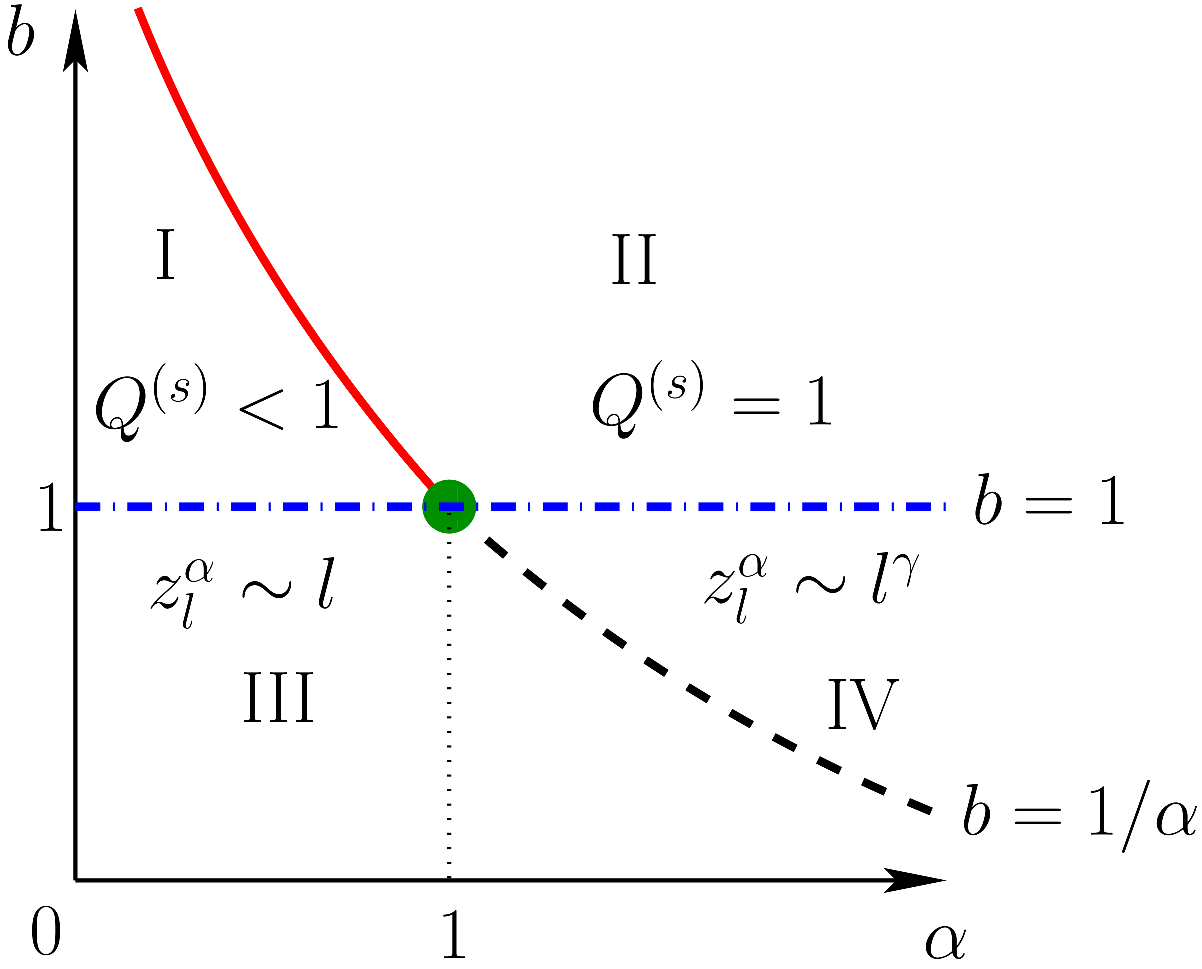
Phase diagram

S.-C. Park, JK, J. Phys. A (2016)



Phase diagram

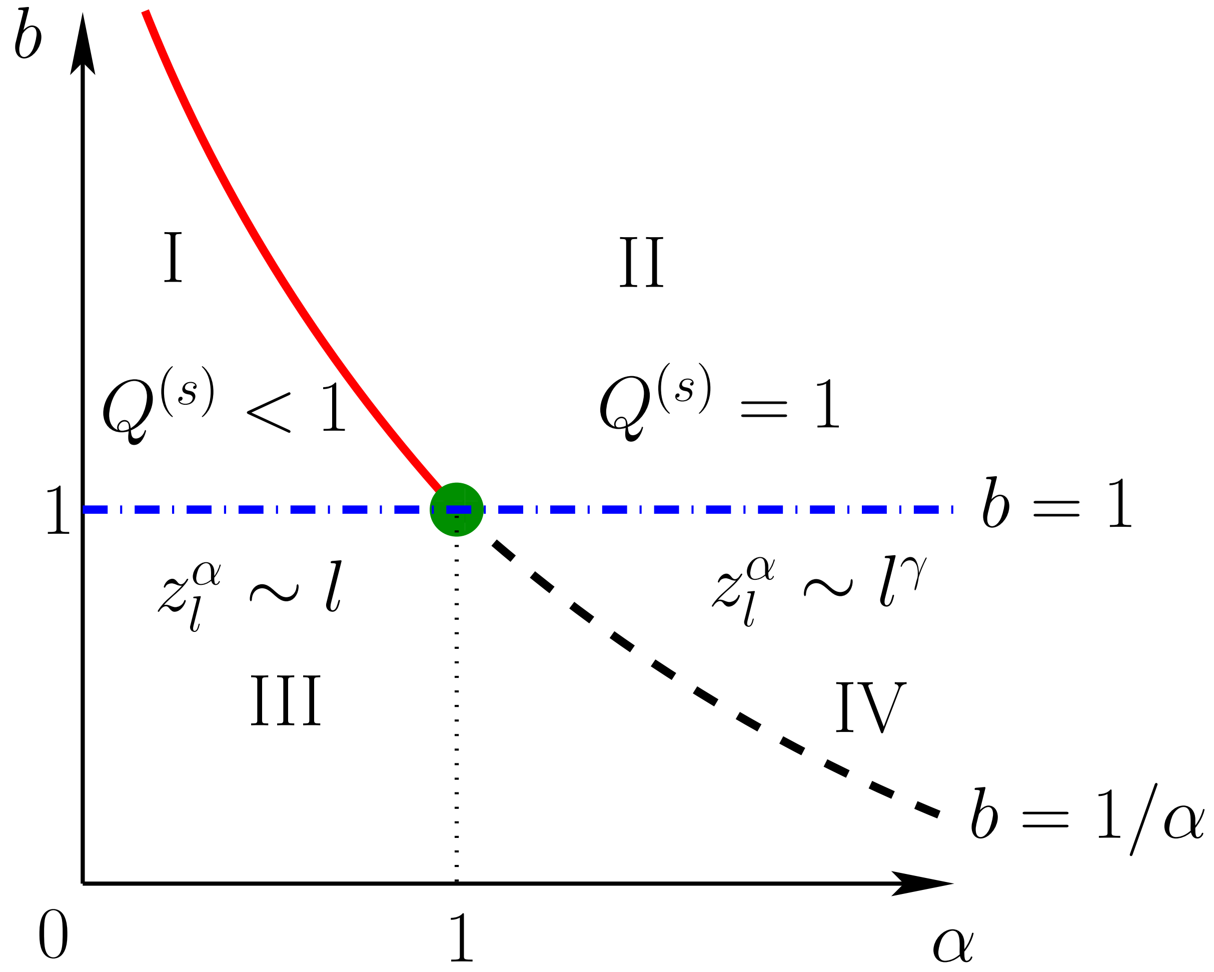
S.-C. Park, JK, J. Phys. A (2016)



- $b < 1$: Mean record value diverges, but the behavior changes from $z_l^\alpha \sim l$ to $z_l^\alpha \sim l^\gamma$ with $\gamma = (1-b)/(1-1/\alpha) < 1$ in region IV.

Phase diagram

S.-C. Park, JK, J. Phys. A (2016)



- $b > 1$: Fraction $Q^{(s)}$ of sample paths become stationary; $Q^{(s)}$ displays a first order phase transition along the red line $b = 1/\alpha > 1$

Stochastic bistability for $b > 1$

- Suppose that the k'th record value $Y_k < \delta_k$ by a fluctuation

- Then the next event satisfies $Y_{k+1} > Y_k - \delta_k$ with probability = 1

- Since Y_{k+1} is an unconstrained draw from $P(x)$, the probability that $Y_{k+1} > \delta_{k+1}$ is

  $$P^{>}_{l+1} = 1 - P(\delta_{l+1}) = \exp[-c^{\alpha}(l+2)^{\alpha(b-1)}] \ll 1$$

  for large l

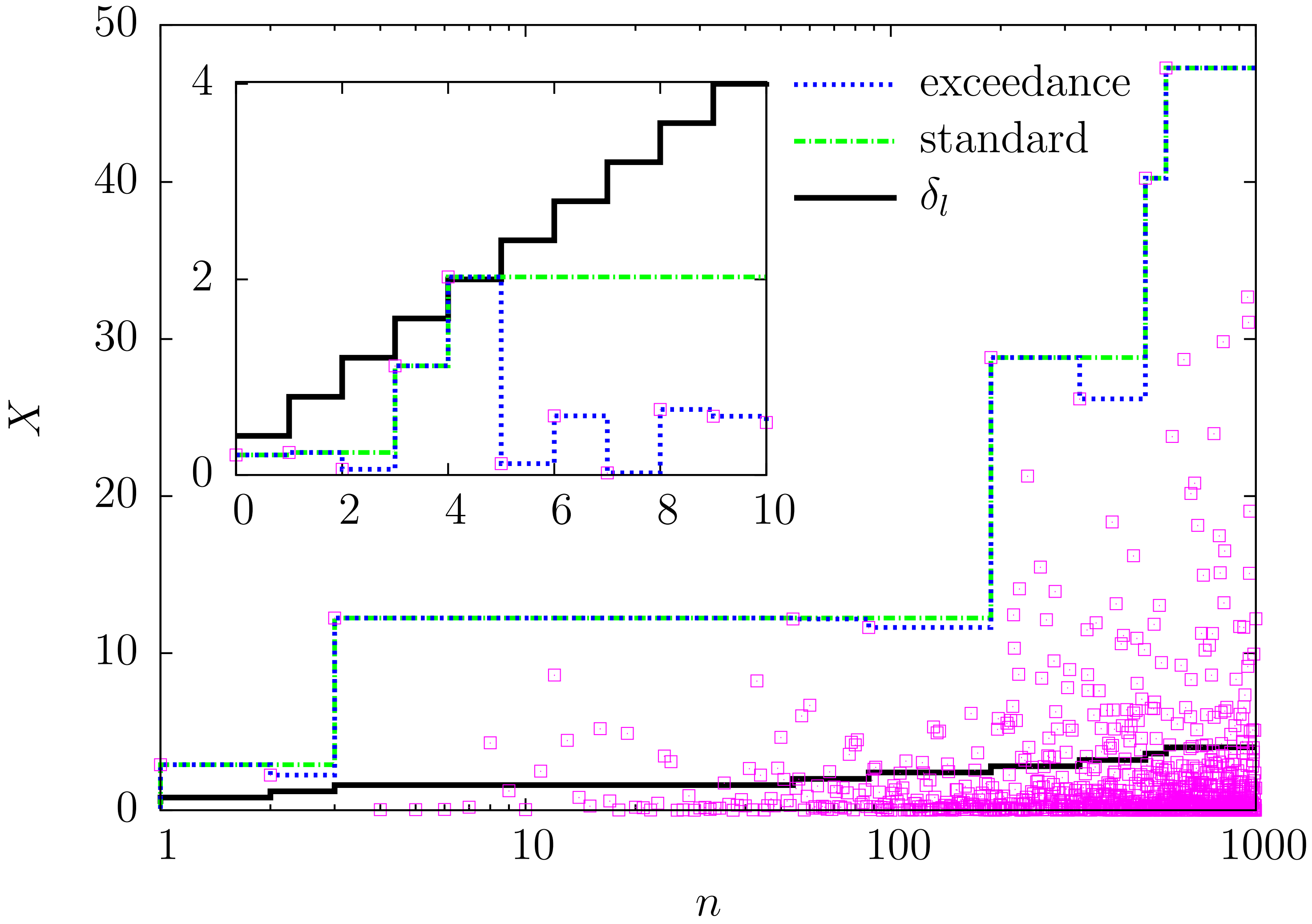
- Thus the process can become trapped in a stationary phase where $Y_k < \delta_k$ and all events are “records”

- This implies a decomposition of the distribution of record values

  $$Q_l(x) = Q^{(s)}\rho(x) + [1 - Q^{(s)}]\tilde{\rho}_l(x - \tilde{z}_l)$$

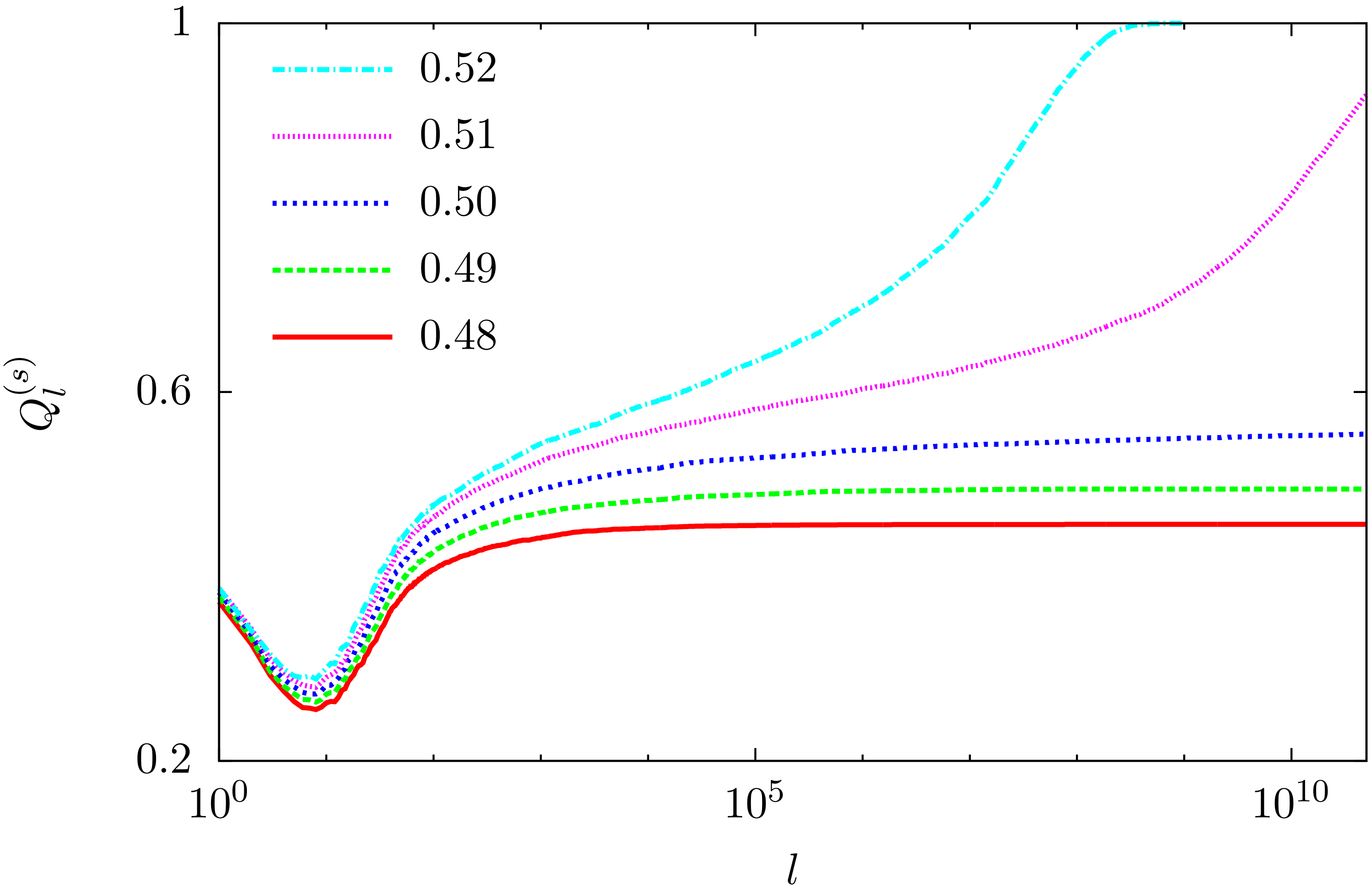
  where $\tilde{z}_l \longrightarrow \infty$ for $l \longrightarrow \infty$ and the “order parameter” $0 < Q^{(s)} \leq 1$

Stochastic bistability for $b = 1/\alpha = 2$



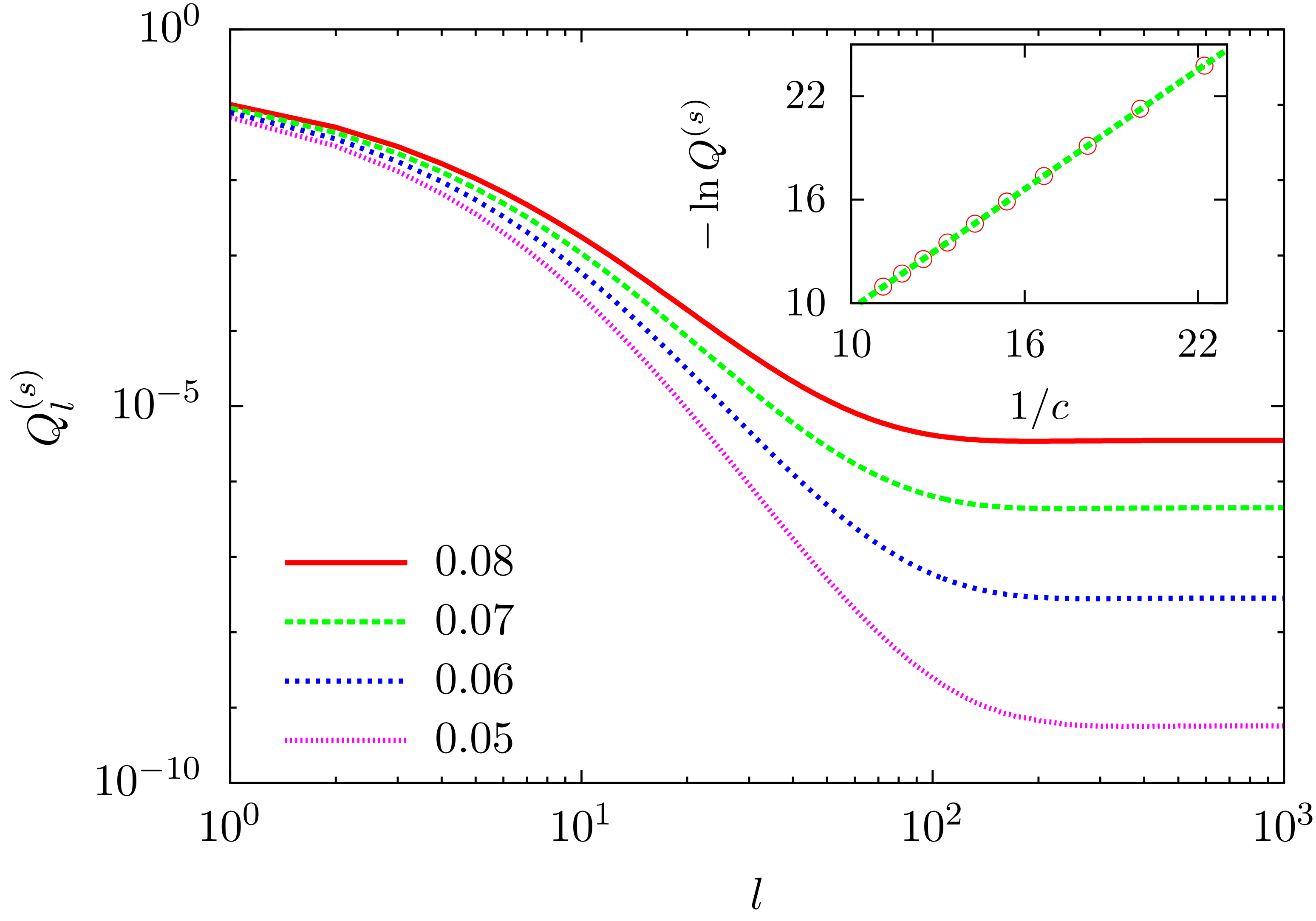
- main plot: $Y_k > \delta_k$ (“normal” phase)
- inset: $Y_k < \delta_k$ (stationary phase)

First order phase transition for $b = 1/\alpha = 2$



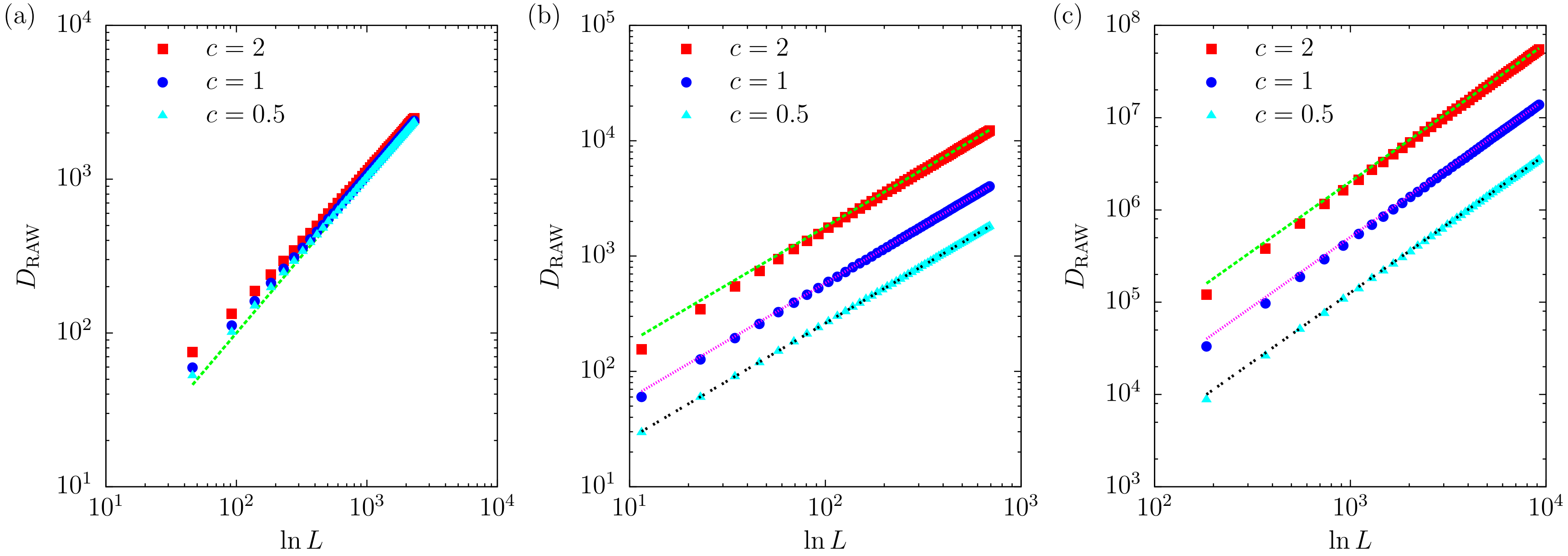
- Asymptotic fraction of stationary paths $Q^{(s)}$ jumps at $c = 1/2$

Behavior of $Q^{(s)}$ for small c



- Inset suggests essential singularity at $c = 0$: $Q^{(s)} \sim \exp[-\chi/c]$

Adaptive walk length for $b = \frac{1}{2} < 1$



(a) $\alpha = 1 < 1/b \Rightarrow \ell \approx \ln L$ independent of c

(b) $\alpha = 2 = 1/b \Rightarrow \ell \approx A(c)\,\ln L$ with $A(c) = [\sqrt{1+c^2} - c]^{-2}$

(c) $\alpha = 4 > 1/b \Rightarrow \ell \approx \sqrt{4c}[\ln L]^{3/2}$

Summary

- Adaptive walks provide a simple yet biologically relevant paradigm of how populations explore complex fitness landscapes

- Random adaptive walks are closely related to record processes

- On correlated fitness landscapes of Rough Mt. Fuji (RMF) type the problem becomes equivalent to δ-exceedance records

- These display a rich phase behavior that arises from the interplay of the tail of the distribution with the deterministic handicap sequence δ_k

- Special role of the exponential distribution appears also in the structural properties of the RMF landscape[1] and for greedy adaptive walks[2]

[1] J. Neidhart, I.G. Szendro, JK, Genetics 2014

[2] S.-C. Park, JK, J. Theor. Biol. 2016