Understanding frequency distributions of path-dependent processes with non-multinomial maximum entropy approaches

2017 New J. Phys. 19 033008

(http://iopscience.iop.org/1367-2630/19/3/033008)

New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG
IOP Institute of Physics

Published in partnership with: Deutsche Physikalische Gesellschaft and the Institute of Physics

OPEN ACCESS

RECEIVED
1 September 2016

REVISED
8 February 2017

ACCEPTED FOR PUBLICATION
17 February 2017

PUBLISHED
6 March 2017

PAPER

Understanding frequency distributions of path-dependent processes with non-multinomial maximum entropy approaches

Rudolf Hanel[1], **Bernat Corominas-Murtra**[1] **and Stefan Thurner**[1,2,3,4]

[1] Section for Science of Complex Systems, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria
[2] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, United States of America
[3] IIASA, Schlossplatz 1, A-2361 Laxenburg, Austria
[4] Complexity Science Hub Vienna, Josfstätterstrasse 39, A-1090 Vienna, Austria

E-mail: stefan.thurner@meduniwien.ac.at

Keywords: Pólya urn processes, statistical mechanics, maximum entropy principle, relative entropy, information divergence

Abstract

Path-dependent stochastic processes are often non-ergodic and observables can no longer be computed within the ensemble picture. The resulting mathematical difficulties pose severe limits to the analytical understanding of path-dependent processes. Their statistics is typically non-multinomial in the sense that the multiplicities of the occurrence of states is not a multinomial factor. The maximum entropy principle is tightly related to multinomial processes, non-interacting systems, and to the ensemble picture; it loses its meaning for path-dependent processes. Here we show that an equivalent to the ensemble picture exists for path-dependent processes, such that the non-multinomial statistics of the underlying dynamical process, by construction, is captured correctly in a functional that plays the role of a relative entropy. We demonstrate this for self-reinforcing Pólya urn processes, which explicitly generalize multinomial statistics. We demonstrate the adequacy of this constructive approach towards non-multinomial entropies by computing frequency and rank distributions of Pólya urn processes. We show how microscopic update rules of a path-dependent process allow us to explicitly construct a non-multinomial entropy functional, that, when maximized, predicts the time-dependent distribution function.

1. Introduction

'It seems questionable whether the Boltzmann principle alone, meaning without a complete [...] mechanical description or some other complementary description of the process, can be given any meaning'. Einstein's famous critical comment on the completeness of Boltzmann entropy [1], is still thought provoking. For ergodic systems, e.g. [2], over a well defined set of states, this critique has turned out to be of minor relevance. Here we demonstrate how Einstein's observation becomes relevant again when dealing with non-ergodic, path-dependent systems or processes, i.e. processes where ensemble and time averages cease to yield identical results and the ensemble descriptions of a processes fails to describe the dynamics of a particular process (e.g. compare [3]).

Moreover, for path dependent systems we have to specify what we mean with 'entropy', since no unique generalization of entropy from equilibrium to non-equilibrium systems exists. However, Boltzmann's principle is grounded in the idea that in large systems the most likely samples we may draw from a process, i.e. the so called *maximum-configuration*, also characterize the typical samples, while it becomes very unlikely to draw atypical samples. In fact we will demonstrate the possibility to directly construct 'entropic functionals' from the microscopic properties determining the dynamics of a large class of non-ergodic processes using maximum-configuration framework. In this approach we identify relative entropy (up to a multiplicative constant) with the logarithm of the probability to observe a particular macro state (which typically is represented by a histogram over a set of observables states), compare e.g. [4]. By construction, maximization of the resulting entropy

functionals leads to adequate predictions of statistical properties of non-ergodic processes, in maximum configuration.

For ergodic processes it is possible to replace time-averages of observables by their ensemble-averages, which leads to a tremendous simplification of computations. In particular, this is true for systems composed of independent particles or for Bernoulli processes, i.e. processes where samples are drawn independently, and the states of the independent components or observations collectively follow a multinomial statistics. The multinomial statistics of such a system with W observable states $i = 1, \ldots, W$ is captured by a functional that coincides with Shannon entropy [5], $H(p) = -\sum_{i=1}^{W} p_i \log p_i$. In this context $p = (p_1, \ldots, p_W)$ is the empirical *relative frequency* distribution of observing states i in an experiment of drawing from the process for N times, i.e. $p = k/N$ is the normalized *histogram* of the experiment where state i has been drawn k_i times. Clearly, $\sum_i k_i = N$. In this context $H(p)$ can be understood as the logarithm of the multinomial factor, i.e. $-\sum_{i=1}^{W} p_i \log p_i \sim \frac{1}{N} \log \binom{N}{k}$, where and $\binom{N}{k} = N! \Big/ \prod_{i=1}^{W} k_i!$ (e.g. compare [6]).

Maximization of Shannon entropy under constraints therefore is a way of finding the most likely relative frequency distribution function (normalized histogram of sampled events) one will observe when measuring a system, provided that it follows a multinomial statistics. Constraints represent knowledge about the system. Bernoulli processes with multinomial statistics are characterized by the prior probabilities, $q = (q_1, \ldots, q_W)$. In general, the set of parameters characterizing a process, we denote by θ. In the multinomial case $\theta \equiv q$.

Denoting the probability to measure a specific histogram by $P(k|\theta, N)$, the most likely histogram $\hat{k}$, that maximizes $P(k|\theta, N)$, is the optimal predictor or the so-called *maximum configuration*. For a multinomial distribution function, $P(k|\theta, N) = \binom{N}{k} \prod_{i=1}^{N} q_i^{k_i}$, where q_i are the prior probabilities (or biases), the functional that is maximized is $\psi(p|\theta) = H(p) + \sum_i p_i \log q_i$, which is (up to a sign) called the *relative entropy* or Kullback–Leibler divergence [7]. The term $H(p)$ coincides with Shannon entropy, the term that depends on q is called *cross-entropy* and is a linear functional in p. By re-parametrizing $q_i = \exp(-\beta\varepsilon_i)$, where $\beta > 0$ is a constant, one gets the standard max-ent functional

$$\psi(p|\theta, N) = H(p) - \beta \sum_i p_i \varepsilon_i \qquad \left[\sum_i p_i = 1\right]. \tag{1}$$

In statistical physics, the constants ε_i typically correspond to energies and β to the so called *inverse temperature* of a system. Maximization of this functional with respect to p yields the most likely empirical distribution function; this is sometimes called the *maximum entropy principle*.

Clearly, systems composed of independent components follow a multinomial statistics. Note that a multinomial statistics is also a direct consequence of working with ensembles of statistically independent systems. In this case the multinomial distribution function reflects the ensemble property and is not necessarily a property of the system itself. Therefore $H(p)$ only has physical relevance for systems that consist of sufficiently independent elements. For path-dependent processes, where ensemble- and time-averages typically yield different results, $H(p)$ remains the entropy of the ensemble picture, but ceases to be the 'physical' entropy that captures the time evolution of a path-dependent process. Obviously, assuming that the entropy functional H, which is consistent with an underlying multinomial statistics, in general also is adequate for characterizing path-dependent processes that are inherently non-multinomial (break multinomial symmetry), is nonsensical.

Surprisingly, the possibility that non-multinomial max-ent functionals can be constructed for path-dependent processes seems to have caught only little attention. In [4] it was noticed that a particular class of non-Markovian random walks with strongly correlated increments can be constructed, where the multiplicity of event sequences is no longer given by the multinomial factor, and the max-ent entropy functional of the process class exactly violates the composition axiom of Khinchin [8]. The general method of constructing a relative entropy principle for a particular process class does not inherently depend on the validity of particular information theoretic axioms, which opens a way for a general treatment of path-dependent, and non-equilibrium processes. We demonstrate this by constructing the max-ent entropy of multi-state *Pólya urn processes* [9, 10].

In multi-state Pólya processes, once a ball of a given color is drawn from an urn, it is replaced by a number of δ balls with the same color- see figure 1. They represent self-reinforcing, path-dependent processes that display the *the rich get richer* and *the winner takes all* phenomenon. Pólya urns are related to the beta-binomial distribution, Dirichlet processes, the Chinese restaurant problem, and models of population genetics. Their mathematical properties were studied in [11, 12], extensions and generalizations of the concept are found in [13, 14], applications to limit theorems in [15–17]. Pólya urns have been used in a wide range of practical applications including response-adaptive clinical trials [18], tissue growth models [19], institutional development [20], computer data structures [21], resistance to reform in EU politics [22], aging of alleles and Ewens's sampling formula [23, 24], image segmentation and labeling [25], and the emergence of novelties in evolutionary scenarios [26, 27]. A notion of Pólya-divergence was recently defined in [28] in the context of

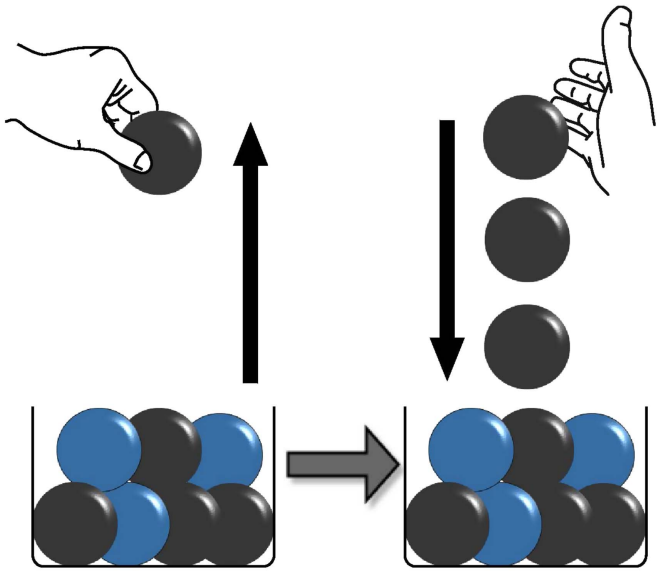

Figure 1. Schematic illustration of a Pólya process. When a ball of a certain color is drawn, it is replaced by $1 + \delta$ balls of the same color. Then the next ball is drawn and the process is repeated for N iterations. Here $\delta = 2$. This reinforcement process creates a history-dependent dynamics. The configurations obtained after successive iterations have non-multinomial structure.

Sanov's theorem [29]. This work characterizes Pólya urns in a regime of weak reinforcement. More precisely the Pólya divergence is derived for situations where the ratio between N, the number of samples drawn from the Pólya urn, and the number A_0 of balls initially contained in the urn, are asymptotically fixed by the parameter $\beta \equiv N/A_0$ such that $\infty > \beta > 0$. As a consequence, in the limit $N \to \infty$ the reinforcement parameter $\gamma = \delta/A_0$ asymptotically approaches zero ($\gamma \sim \delta\beta/N \to 0$). So even if the number δ of balls added to the urn at each trial is large, the number of balls initially contained in the urn is much larger. In this regime of weak reinforcement Pólya urns behave similarly to Bernoulli processes. Our constructive approach allows us to access strong reinforcement parameters $\gamma > 0$ and the transition of Pólya urn dynamics from Bernoulli-process like behavior to a winner-takes-all type of dynamics can be studied.

2. Non-multinomial max-ent functionals

The general aim is to construct a max-ent functional for a path-dependent process, which allows us to infer the maximum configuration, i.e. the most likely sample we may draw from a process of interest. From a given class of processes X we select a particular process $X(\theta)$, specified by a set of parameters, θ. Running the processes $X(\theta)$ for N consecutive iterations produces a sequence of *observed states* $x(\theta, N) = [x_1, \ldots, x_N]$, where each x_n takes a value from W possible states. As before, we assume the existence of a most likely histogram $\hat{k}$, that maximizes $P(k|\theta, N)$. To construct a max-ent functional for X, one has to conveniently rescale $P(k|\theta, N)$, which happens in two steps. First, we define $\Psi(p|\theta, N) \equiv \log P(Np|\theta, N)$. Note, if $\hat{k}$ maximizes $P(k|\theta, N)$, then $\hat{p} = \hat{k}/N$ maximizes $\Psi(p|\theta, N)$. Second, a scaling factor $\phi(N)$ can be used to scale out the leading term of the N dependence of Ψ. Typically $\phi(N) = N^c$, for some constant $1 \geqslant c > 0$, compare [4]. $\phi(N)$ corresponds to the *effective* number of degrees of freedom of samples of size N. We *identify* the max-ent functional with $\psi(p|\theta, N) \equiv \Psi(p|\theta, N)/\phi(N)$. Again, if $\hat{k}$ maximizes $P(k|\theta, N)$ with $\sum_i k_i = N$, then $\hat{p} = \hat{k}/N$ maximizes $\psi(p|\theta, N)$, with $\sum_i p_i = 1$. In other words, $\psi(p|\theta, N)$ represents (up to a sign) a functional providing us with a notion of *relative entropy* (information divergence) for the process-class X. If this process-class X is the class of Bernoulli-processes, such that $P(k|q, N)$ is the multinomial distribution, then asymptotically $-\psi(p|q, N) \sim \sum_i p_i(\log p_i - \log q_i)$, is the Kullback–Leibler divergence, and $\phi(N) = N$. In the following we compute $\psi(p|\theta, N)$ for Pólya urn processes.

3. Max-ent functional for Pólya urns

In urn models observable states i are represented by the colors balls contained in the urn can have. The likelihood of drawing a ball of color i is determined by the number of balls contained in the urn. Initially the urn contains a_i balls of color $i = 1, \ldots, W$. The *initial prior probability* to draw a ball of color i is given by $q_i = a_i/A_0$, where $A_0 = \sum_i a_i$ is the total number of balls initially in the urn. Balls are drawn sequentially from the urn. Whenever a

ball of color i is drawn, it is put back into the urn and another δ balls of the same color are added. This defines the multi-state Pólya process [9]. A particular Pólya process is fully characterized by the parameters, $\theta = (q_1, \ldots, q_W; A_0, \delta)$. Drawing without replacement is the *hypergeometric* process, drawing with replacement ($\delta = 0$), is the *multinomial* process.

If $\delta > 0$, after N trials there are $a_i(N) = a_i + \delta k_i$ balls of color i in the urn ($a_i = a_i(0)$). The total number of balls is $A(N) = \sum a_i(N) = A_0 + N\delta$, and the probability to draw a ball of color i in the $(N+1)$ th step is

$$p(i|k, \theta) = \frac{a_i(N)}{A(N)} = \frac{a_i + k_i\delta}{A_0 + N\delta}, \tag{2}$$

which depends on the history of the process in terms of the histogram k. With $x(0) = [\,]$ the *empty sequence*, the probability of sampling sequence x can be computed

$$p(x|\theta) = \prod_{n=1}^{N} p(x_n|k(x(n-1)), \theta) = \frac{\prod_{i=1}^{W} a_i^{(\delta,k_i)}}{A_0^{(\delta,N)}}, \tag{3}$$

where the function $m^{(\delta,r)}$ is defined as

$$m^{(\delta,r)} \equiv m(m+\delta)(m+2\delta)\cdots(m+(r-1)\delta). \tag{4}$$

Note that $m^{(\delta,r)}$ generalizes the multinomial law,

$$\left(\sum_i a_i\right)^{(\delta,N)} = \sum_{\{k|N=\sum_i k_i\}} \binom{N}{k} \prod_{i=1}^{W} a_i^{(\delta,k_i)}, \tag{5}$$

and forms a one-parameter generalization of powers m^r. For $\delta = 0$, $m^{(0,r)} = m^r$ and for $\delta = 1$, $m^{(1,r)} = (m+r-1)!/(m-1)!$.

The probability of observing a particular histogram k after N trials becomes

$$P(k|\theta, N) = \binom{N}{k} \frac{\prod_{i=1}^{W} a_i^{(\delta,k_i)}}{A_0^{(\delta,N)}}, \tag{6}$$

with $\sum_{\{k_i \geqslant 0 | \sum_i k_i = N\}} P(k|\theta, N) = 1$. Note that $P(k|\theta, N)$ is almost of multinomial form, it is a multinomial factor times a term depending on θ. One might conclude that the max-ent functional for Pólya processes is Shannon entropy in combination with a generalized cross-entropy term that depends on θ. However, this turns out to be wrong, since contributions from the generalized powers $m^{(\delta,r)}$ in equation (6) cancel the multinomial factor almost completely. To see this we first rewrite

$$\begin{aligned} a_i^{(\delta,k_i)} &= a_i(a_i+\delta)\cdots(a_i+(k_i-1)\delta) \\ &= (a_i+\delta)\cdots(a_i+k_i\delta)\tfrac{a_i}{a_i+k_i\delta} \\ &= k_i!\delta^{k_i}(1+\tfrac{a_i}{\delta})\cdots\left(1+\tfrac{a_i}{k_i\delta}\right)\tfrac{a_i}{a_i+k_i\delta} \\ &\sim k_i!\delta^{k_i}(k_i+1)^{\frac{a_i}{\delta}}\tfrac{a_i}{a_i+k_i\delta}, \end{aligned} \tag{7}$$

where we use $\sum_{r=1}^{s} \frac{1}{r} \sim \log(s+1)$ and $1 + y \sim \exp(y)$, which is valid for sufficiently small $y = a_i/\delta$, i.e. for sufficiently large δ. With the notation $\gamma \equiv \delta/A_0$ we obtain

$$P(k|\theta, N) = \frac{1+\gamma N}{\left(1+\frac{1}{N}\right)^{\frac{1}{\gamma}}} \prod_{i=1}^{W} \frac{\left(p_i+\frac{1}{N}\right)^{\frac{q_i}{\gamma}}}{1+N\gamma\frac{p_i}{q_i}}, \tag{8}$$

where $k = pN$. Following the construction discussed above, we identify $\Psi(p|\theta, N) = \log P(pN|\theta, N)$, which no longer scales explicitly with N, but $\phi(N) = 1$ ($c = 0$), so that $\psi = \Psi$. Inserting equation (7) into (6), leads to the expression

$$\begin{aligned} \psi(p|\theta) = &-\sum_{i=1}^{W}\left(1-\tfrac{q_i}{\gamma}\right)\log\left(p_i+\tfrac{1}{N}\right) \\ &-\sum_{i=1}^{W}\log\left(1+\tfrac{1}{N\gamma}\frac{q_i-\gamma}{p_i+\frac{1}{N}}\right)+\sum_{i=1}^{W}\log q_i \\ &-\tfrac{1}{\gamma}\log\left(1+\tfrac{1}{N}\right)+\log(1+N\gamma)-W\log(N\gamma). \end{aligned} \tag{9}$$

More precisely, the *finite size* Pólya 'entropy' can be conveniently identified with the terms in $\psi(p|\theta)$ that do not depend on q,

$$H^{\text{Pólya}}(p) = -\sum_{i=1}^{W} \log\left(p_i + \frac{\lambda}{N}\right), \tag{10}$$

where $\lambda > 0$ can in principle be chosen freely. Up to a constant depending only on γ and N, the finite size cross-entropy can be identified with

$$H^{\text{Pólya}}_{\text{cross}}(p|q) = -\sum_i \left[\frac{q_i}{\gamma}\log\left(p_i + \frac{1}{N}\right) - \log\left(1 + \frac{1}{N\gamma}\frac{q_i - \lambda\gamma}{p_i + \frac{\lambda}{N}}\right) + \log q_i\right]. \tag{11}$$

Convenient choices for λ are the following. $\lambda = 1$, represent $\log\psi$ as in equation (9). Alternatively, one may choose $\lambda = 1/(W\gamma)$, which is a convenient choice if one considers a uniform initial distribution, $q_i = 1/W$, of balls in the urn. The finite size Pólya entropy equation (10), yields a well defined entropy even if some states i have vanishing probability $p_i = 0$.

To simplify the following analysis we consider the limit $N \to \infty$ of this functional, where the notion of 'information divergence' for Pólya processes, essentially reduces to

$$\psi(p|\theta) = -\sum_{i=1}^{W} \log p_i + \frac{1}{\gamma}\sum_{i=1}^{W} q_i \log p_i + \sum_{i=1}^{W} \log q_i, \tag{12}$$

up to terms of order $1/N$ and terms that do not explicitly depend on p_i or q_i. In this limit the asymptotic Pólya 'entropy' is given by,

$$H^{\text{Pólya}}(p) = -\sum_{i=1}^{W} \log p_i. \tag{13}$$

We observe that one cannot derive $H^{\text{Pólya}}(p)$ from the multiplicity of the system, which gets canceled by counter terms, as we have seen above. In addition, we note that the q dependent terms, $\sum_i q_i \log p_i$, in equation (12) play the role of the Pólya 'cross-entropy', which is no longer linear in p.

Maximizing $\psi(p|\theta)$ with respect to p on $\sum p_i = 1$, either leads to the solution

$$p_i = \frac{1}{\zeta}(q_i - \gamma), \tag{14}$$

for $0 < p_i < 1$, or, if this can not be satisfied, to boundary solutions $p_i = 0$. ζ is a normalization constant. There exist three scenarios:

(A) For $\gamma < \min(q)$, equation (14) is the max-ent solution for all i (no boundary-solutions). The limit $\gamma \to 0$ provides the correct multinomial limit $p_i \to q_i$.

(B) If $\max(q) > \gamma > \min(q)$, ψ gets maximal for those i with $q_i > \gamma$ and follows solution equation (14); those i where $q_i < \gamma$ are boundary-solutions, $p_i = 0$.

(C) For $\gamma > \max(q)$ all p_i are boundary-solutions, meaning that one winner i takes it all, $p_i = 1$, while all other states have vanishing probability.

Since $\partial^2\psi_{\text{Pólya}}/\partial p_i^2 < 0$ if $\gamma < q_i$, for all i, case (A) applies. If $q_i < \gamma$, equation (14) becomes negative but also unstable and is replaced by a boundary solution: cases (B) and (C). The Pólya max-ent not only allows us to predict p_i from the initial prior probabilities q_i, it also identifies γ as the crucial parameter that distinguishes between the three regimes of Pólya urn dynamics[5]. For sufficiently large but finite N, the analysis above is more involved but solvable.

Assuming uniformly distributed priors, $q_i = 1/W$ for all i, the max-ent result equation (14) correctly predicts uniformly distributed $\hat{p}_i = 1/W$, while observed distributions p may strongly deviate from this prediction. This result reflects the fact that despite the Pólya urn process being inherently instable (e.g. winner takes all) with little chance of predicting who in particular will win, i.e. which color of balls will dominate the others, repeating the experiment many times every color of balls has the same chance to win (or biased according to the priors q). This discrepancy between ensemble average and time average makes it impossible to predict who in particular will win or loose in the course of time. However, using detailed information about the process one can predict *how* winners win. In particular one can (i) predict the onset of instability, i.e. the emergence of colors i that will effectively never be drawn, at $\gamma_{\text{crit}} = \min(q)$ (compare figure 2), and (ii) construct a maximum entropy

[5] Note that a Pólya urn U_1 that initially contains A_0 balls and has evolved for N steps with $\gamma = \delta/A_0$, can be regarded as another Pólya urn, U_2, in its initial state, containing $A_N = A_0 + \delta N$ balls, that evolves with an effective reinforcement parameter $\gamma(N) = \delta/A_N$, and the initial distribution of balls $q(N) = p(i|k(N), \theta)$, where $k(N)$ is the histogram of colors drawn in the first N steps of the original urn process U_1. Obviously the asymptotic behavior of Pólya urns gets determined early on in the process, where the effective reinforcement parameter $\gamma(N)$ is largest. The probability of a Pólya urn to enter a winner-takes-all dynamics, i.e. to end up in one of the scenarios A, B, or C, depends on the reinforcement parameter γ.

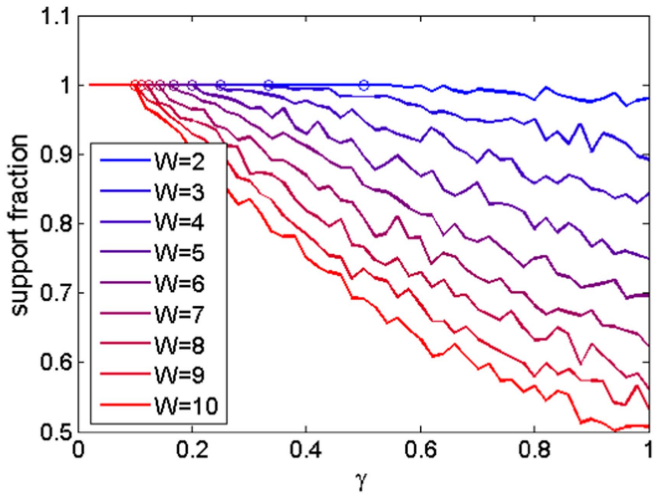

Figure 2. The fraction of distinct colors contained in the Pólya urn, which at least get sampled once within the first $N = 500$ steps of the process, for numbers of colors $W = 2, 3, \ldots, 10$ for uniformly distributed initial conditions $q_i = 1/W$, $i = 1, \ldots, W$, evaluated from 250 runs for each $\gamma = 0.01, 0.02, \ldots, 1$. The onset of instability $\gamma_{\text{crit}}(q) = \min(q) = 1/W$ (circular markers) is very well reproduced experimentally.

functional for predicting the time dependent frequency distribution of a process, i.e. the number of times one observes states *i* for *n* times. As a consequence, one also can derive the rank distributions of the process, i.e. the frequency of observing balls of some color after ranking those frequencies according to their magnitude.

3.1. Rank and frequency distributions of Pólya urns

With the presented max-ent approach we now compute frequency distribution functions. Given the histogram $k = (k_1, k_2, \ldots k_W)$ is obtained after N iterations of the process,we define new variables,

$$n_z(k) = \sum_{i=1}^{W} \chi(k_i = z), \tag{15}$$

where χ is the characteristic function that returns 1 if the argument is true and 0 if false. $n_z(k)$ is the number of colors i that occur z times after running the Pólya process for N iterations. n_z is subject to the two constraints,

$$W = \sum_{z=0}^{N} n_z(k) \quad \text{and} \quad N = \sum_{z=0}^{N} n_z(k) z, \tag{16}$$

which can be included in the maximization procedure introducing Lagrange multipliers, α and β. The probability of observing some $n = (n_1, \ldots, n_N)$ is

$$\tilde{P}(n|\theta, N) = \sum_{\{k_i \geqslant 0 | n = n(k)\}} P(k|\theta, N). \tag{17}$$

Defining the relative frequencies $\pi_z = n_z/W$ and $\bar{p}_z = z/N$ we can construct the max-ent functional from $\tilde{P}(n|\theta, N)$. We identify $\tilde{\psi}(\pi|\theta, N) \equiv \log(\tilde{P}(n|\theta, N))/W$.

For the multinomial $P(k|\theta, N) = \binom{N}{k} \prod_i q_i^{k_i}$, and uniform priors $q_i = 1/W$ we find up to an additive constant,

$$\begin{aligned} \tilde{\psi}(\pi|\theta, N) = &- \sum_{z=0}^{N} \pi_z \log \pi_z \\ &- N \sum_{z=0}^{N} \pi_z \bar{p}_z \log \bar{p}_z. \end{aligned} \tag{18}$$

$\tilde{\psi}(\pi|\theta, N)$ has to be maximized subject to equation (16),

$$\sum_{z=0}^{N} \pi_z = 1 \quad \text{and} \quad \sum_{z=0}^{N} \pi_z \bar{p}_z = \frac{1}{W}, \tag{19}$$

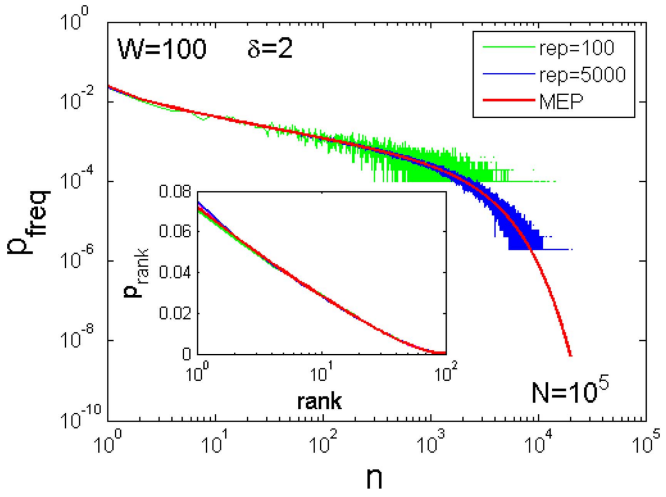

Figure 3. Frequency distribution of a Pólya urn process and uniform initial conditions (red line), for $W = 100$, $\delta = 2, a_i = 1$ for all $i = 1$, and $N = 10^5$ steps. Simulations are shown for 100 (green) and 5000 (blue) repetitions of the process. Inset: rank distributions of the max-ent result and the numerical realizations in semi-log scale.

so that we get the asymptotic solution for large W and large N, $(N \gg W \gg 1)$,

$$\pi_z = \frac{\phi^z}{\zeta\ z!}. \tag{20}$$

This is the Poisson distribution, exactly as expected for multinomial processes. $\phi = N\mathrm{e}^{-(1+\frac{\beta}{N})}$, ζ is a normalization constant, and π_z gets maximal at $\hat{z} = \phi \sim N/W$.

For the Pólya urn with uniform priors we get from equation (17)

$$\tilde{P}(n|\theta, N) = \frac{1}{Z(\theta, N)} \frac{W!}{\prod_{z=0}^{N} n_z!} \prod_{z=0}^{N} \left[\frac{\left(\tilde{p}_z + \frac{1}{N}\right)^{\frac{1}{W\gamma}}}{\tilde{p}_z + \frac{1}{\gamma WN}} \right]^{n_z}. \tag{21}$$

$Z(\theta, N)$ is the normalization. Up to a constant the max-ent functional $\tilde{\psi}_{\text{Pólya}}(\pi|\theta, N) \equiv \log(P(n|\theta, N))/W$ is

$$\begin{aligned} \tilde{\psi}_{\text{Pólya}}(\pi|\theta, N) = &- \sum_{z=0}^{N} \pi_z \log(\pi_z) \\ &- \sum_{z=0}^{N} \pi_z \log\left(\tilde{p}_z + \tfrac{1}{\gamma WN}\right) \\ &+ \tfrac{1}{W\gamma} \sum_{z=0}^{N} \pi_z \log\left(\tilde{p}_z + \tfrac{1}{N}\right). \end{aligned} \tag{22}$$

maximizing $\tilde{\psi}_{\text{Pólya}}(\pi|\theta, N)$ under the conditions of equation (19) provides the frequency distribution of the Pólya process for uniform priors,

$$\pi_z = \frac{1}{\zeta} \phi^z \frac{(z+1)^{\frac{1}{W\gamma}}}{z + \frac{1}{\gamma W}}, \tag{23}$$

with $\phi = \exp(-\beta)$, and normalization $\zeta = \exp(1 + \alpha) N^{\frac{1}{W\gamma}-1}$.

The rank distribution of states, $f(r)$, can now be obtained as follows. $r = 1$ is the state that occurs most frequently, $r = W$ is the least occupied state. For $r = 1, \ldots, W$ we define intervals $[t_{r+1} t_r]$ with $t_1 = N$ and $t_{W+1} = 0$, such that $\sum_{t_{r+1} \leqslant z < t_r} \pi_z \sim 1/W$. To find t_i we substitute sums by integrals and get

$$\frac{1}{W} = \int_{t_{r+1}}^{t_r} \mathrm{d}z\ \pi_z \quad \text{and} \quad f(r) = \frac{W}{N} \int_{t_{r+1}}^{t_r} \mathrm{d}z\ \pi_z z. \tag{24}$$

Results for the frequency distributions for $a_i = 1$, $W = 100$, and $\delta = 2$ are shown in figure 3, together with a numerical simulation for the same process. The inset shows the rank distribution. The Pólya max-ent predicts frequency and rank distribution extremely well.

The above results were all derived under the assumption that $\gamma > 0$ is sufficiently large. By numerical simulation we find that the solution equation (23) also works remarkably well for very small values of γ, if the value of γ in equation (23) is appropriately renormalised, $\gamma \to \gamma_0$. In particular for $\gamma = 0$ (multinomial process)

we sample the Poisson distribution function, equation (20). The Pólya max-ent solution recovers the Poisson distribution extremely well if $\gamma = 0 \rightarrow \gamma_0(W, N) = 1/(N + 3W)$. In this sense the Pólya max-ent remains adequate in the limit of small γ.

4. Discussion

Pólya urns offer a transparent way to study self-reinforcing systems with explicit path-dependence. They behave similarly to Bernoulli processes if the reinforcement is weak, i.e. if the number of balls initially contained in the urn is large in comparison to the number of balls added to the urn at each trial. This weak reinforcement regime has been studied in [28].

If reinforcement gets stronger, Pólya urns start to behave differently and the Pólya divergence derived in [28] no longer applies. Based on the microscopic rules of the process, we constructively derive the generalized information divergence or relative entropy $-\psi$, for strongly reinforcing Pólya urns. The functional ψ acts as the corresponding non-multinomial max-ent functional. This provides us with an alternative to the *ensemble approach* for path-dependent processes that enables us to predict the statistics of the process. The maximization of the functional leads to an equivalent of the classical *maximum configuration* approach, which by definition predicts the most likely distribution function. In this sense maximum configuration predictions are optimal, and can be used to understand even details of the statistics of path-dependent processes, such as their frequency and rank distributions.

It is interesting to note that the functional playing the role of the entropy in the Pólya processes violates at least two of the four classic information theoretic (Shannon–Khinchin) axioms which determine Shannon entropy [8]. Even more, for the finite size Pólya entropy, three of the four axioms are violated. This indicates that the classes of generalized entropy functionals that are useful for a max-ent approach may be even larger than expected [30, 31]. One might speculate that in this sense the classic information theoretic axioms are too rigorous, when it comes to characterizing information flow and phase space structure in non-stationary, path-dependent, processes. The observation that each particular class of non-multinomial processes requires a matching max-ent functional that can in principle be constructed from the generative rules of a process, opens the applicability of max-ent approaches for a wide range of complex systems in a meaningful way. The generalized max-ent approach in this sense responds to naturally Einsteins comment on Boltzmann's principle.

Finally we note the implications for statistical inference with data from non-multinomial sources, which implicitly involves the estimation of the parameters θ that determine the process that generates the data. In a max-ent approach this is done by fitting classes of curves to the data, that are consistent with the max-ent approach. For doing this, the nature of the process, i.e. its class, needs to be known. For path-dependent processes, which are non-multinomial by nature, entropy will no longer be Shannon entropy H, and the information divergence will no longer be the Kullback–Leibler divergence.

Acknowledgments

This work was supported by the Austrian Science Fund FWF under P29252 Generalized information theoretic approaches for history dependent- processes and the FP7 projects LASAGNE no. 318132 and MULTIPLEX no. 318132.

References

[1] Einstein A 1910 Theorie der Opaleszenz von homogenen Flüssigkeiten und Flüssigkeitsgemischen in der Nähe des kritischen Zustandes. *Ann. Phys.* **33** 1275
[2] Landau L D and Lifshitz E M 1980 *Statistical Physics, Course in Theoretical Physics* vol 5 3rd edn (Amsterdam: Elsevier) footnote p 12
[3] Peters O 2011 The time resolution of the St Petersburg paradox *Phil. Trans. R. Soc.* A **396** 4913–31
[4] Hanel R, Thurner S and Gell-Mann M 2014 How multiplicity of random processes determines entropy: derivation of the maximum entropy principle for complex systems *Proc. Natl Acad. Sci. USA* **111** 6905–10
[5] Shannon C E 1948 A mathematical theory of communication *Bell Syst. Tech. J.* **27** 379–423
Shannon C E 1948 A mathematical theory of communication *Bell Syst. Tech. J.* **27** 623–56
[6] Jaynes E T 1968 Prior probabilities *IEEE Trans. Syst. Sci. Cybern.* **4** 227–41
[7] Kullback S and Leibler R A 1951 On information and sufficiency *Ann. Math. Stat.* **22** 79–86
[8] Khinchin A I 1957 *Mathematical Foundations of Information Theory* (New York: Dover)
[9] Eggenberger F and Pólya G 1923 Über die Statistik verketteter Vorgänge *Z. Angew. Math. Mech.* **1** 279–89
[10] Pólya G 1930 Sur quelques points de la théorie des probabilités *Ann. Inst. Henri Poincaré* **1** 117–61
[11] Wallstrom T C 2012 The equalization probability of Pólya urn *Am. Math. Mon.* **119** 516–8
[12] Johnson N L and Kotz S 1977 Urn models and their application: an approach to modern discrete probability theory *Urn Models and their Application: An Approach to Modern Discrete Probability Theory* (New York: Wiley)

[13] Mahmoud H 2008 Pólya urn models *Texts in Texts in Statistical Science* (*Chapman & Hall/CRC Texts in Statistical Science*) (Hoboken, NJ: Taylor and Francis)
[14] Kotz S, Mahmoud H and Robert P 2000 On generalised Pólya urn models *Stat. Prob. Lett.* **49** 163–73
[15] Janson S 2004 Functional limit theorems for multitype branching processes and generalised Pólya urns *Stoch. Process. Appl.* **110** 177–245
[16] Smythe R T 1996 Central limit theorems for urn models *Stoch. Process. Appl.* **65** 115–37
[17] Gouet R 1993 Martingale functional central limit theorems for a generalised Pólya urn *Ann. Prob.* **21** 1624–39
[18] Tolusso D and Wang X 2011 Interval estimation for response adaptive clinical trials *Comput. Stat. Data Anal.* **55** 725–30
[19] Binder B J and Landman K A 2009 Tissue growth and the Pólya distribution *Aust. J. Eng. Edu.* **15** 35–42
[20] Crouch C and Farrell H 2004 Breaking the path of institutional development? alternatives to the new determinism *Ratio. Soc.* **16** 5–43
[21] Bagchi A and Pal A K 1985 Asymptotic normality in the generalised Pólya-eggenberger urn model, with an application to computer data structures *SIAM J. Algebr. Discrete. Methods* **6** 394–405
[22] Geppert T 2012 EU-Agrar-und Regionalpolitik, Wie vergangene Entscheidungen zukünftige Entwicklungen beeinflussen-Pfadabhängigkeit und die Reformfähigkeit von Politikfeldern *PhD Thesis* University of Bamberg Press, Bamberg
[23] Donnelly P 1986 Partition structures, Pólya urns, the Ewens sampling formula, and the ages of alleles *Theor. Popul. Biol.* **30** 271–88
[24] Hoppe F M 1984 Pólya-like urns and the Ewens' sampling formula *J. Math. Biol.* **20** 91–4
[25] Banerjee A, Burlina P and Alajaji F 1999 Image segmentation and labeling using the Pólya urn model *IEEE Trans. Image. Proc.* **8** 1243–53
[26] Alexander J M, Skyrms B and Zabell S 2012 Inventing new signals *Dyn. Games Appl.* **2** 129–45
[27] Tria F, Loreto V, Servedio V D P and Strogatz S H 2014 The dynamics of correlated novelties *Sci. Rep.* **4** 5890
[28] Grendar M and Niven R K 2010 The Pólya information divergence *Info. Sci.* **180** 4189–94
[29] Sanov I N 1957 On the probability of large deviations of random variables *Math. Sb.* **42** 11–44
[30] Hanel R and Thurner S 2011 A comprehensive classification of complex statistical systems and an *ab initio* derivation of their entropy and distribution function *Europhys. Lett.* **93** 20006
[31] Hanel R and Thurner S 2011 When do generalised entropies apply? How phase space volume determines entropy *Europhys. Lett.* **96** 50003