mathematics

MDPI

Article

The Rescaled Pólya Urn and the Wright—Fisher Process with Mutation

Giacomo Aletti [1,†] **and Irene Crimaldi** [2,*,†]

[1] Environmental Science and Policy Department, Università degli Studi di Milano, 20133 Milan, Italy; giacomo.aletti@unimi.it
[2] IMT School for Advanced Studies Lucca, 55100 Lucca, Italy
* Correspondence: irene.crimaldi@imtlucca.it
† These authors contributed equally to this work.

Abstract: In recent papers the authors introduce, study and apply a variant of the Eggenberger—Pólya urn, called the "rescaled" Pólya urn, which, for a suitable choice of the model parameters, exhibits a reinforcement mechanism mainly based on the last observations, a random persistent fluctuation of the predictive mean and the almost sure convergence of the empirical mean to a deterministic limit. In this work, motivated by some empirical evidence, we show that the multidimensional Wright—Fisher diffusion with mutation can be obtained as a suitable limit of the predictive means associated to a family of rescaled Pólya urns.

Keywords: Pólya urn; predictive mean; urn model; Wright—Fisher diffusion

check for updates

Citation: Aletti, G.; Crimaldi, I. The Rescaled Pólya Urn and the Wright—Fisher Process with Mutation. *Mathematics* **2021**, *9*, 2909. https://doi.org/10.3390/math9222909

Academic Editors: Emanuele Dolera and Federico Bassetti

Received: 8 October 2021
Accepted: 13 November 2021
Published: 15 November 2021

1. Introduction

The well-known standard Eggenberger—Pólya urn [1,2] works as follows. An urn initially contains $N_{0,i}$ balls of color i, for $i = 1, \dots, k$, and at each time-step, a ball is drawn from the urn and then it is returned into the urn together with $\alpha > 0$ additional balls of the same color (here and in the following, the expression "number of balls" is not to be understood literally, but all the quantities are real numbers, not necessarily integers). Hence, denoting by $N_{n,i}$ the number of balls of color i inside the urn at time-step n, we have

$$N_{n,i} = N_{n-1,i} + \alpha\xi_{n,i} \qquad \text{for } n \geq 1,$$

where $\xi_{n,i} = 1$ if the drawn ball at time-step n is of color i, and $\xi_{n,i} = 0$ otherwise. The parameter α tunes the reinforcement mechanism: the greater the α, the greater the dependence of $N_{n,i}$ on $\sum_{h=1}^{n} \xi_{h,i}$.

In [3–5], the rescaled Pólya (RP) urn has been introduced, studied, generalized and applied. This model differs from the original one by the introduction of a parameter β such that

$$\begin{aligned} N_{n,i} &= b_i + B_{n,i} && \text{with} \\ B_{n+1,i} &= \beta B_{n,i} + \alpha\xi_{n+1,i} && n \geq 0. \end{aligned}$$

Therefore, at time-step 0, the urn contains $b_i + B_{0,i} > 0$ balls of color i and the parameters $\alpha > 0$ and $\beta \geq 0$ regulate the reinforcement mechanism. More precisely, the term $\beta B_{n,i}$ connects $N_{n+1,i}$ to the "configuration" at time-step n by means of the "scaling" parameter β, and the term $\alpha\xi_{n+1,i}$ connects $N_{n+1,i}$ to the outcome of the drawing at time-step $n+1$ by means of the parameter α. The case $\beta = 1$ corresponds to the standard Eggenberger—Pólya urn with an initial number $N_{0,i} = b_i + B_{0,i}$ of balls of color i. When $\beta < 1$, the RP urn model shows the following three characteristics:

(i) A reinforcement mechanism mainly based on the last observations;
(ii) A random persistent fluctuation of the predictive mean $\psi_{n,i} = E[\xi_{n+1,i} = 1|, \xi_{h,j}, 0 \leq h \leq n, 1 \leq j \leq k]$;

(iii) The almost sure convergence of the empirical mean $\sum_{n=1}^{N} \xi_{n,i}/N$ to the deterministic limit $p_i = b_i / \sum_{i=1}^{n} b_i$, and a chi-squared goodness of fit result for the long-term probability distribution $\{p_1, \ldots, p_k\}$.

Regarding point (iii), we specifically have that the chi-squared statistics

$$\chi^2 = N \sum_{i=1}^{k} \frac{(O_i/N - p_i)^2}{p_i},$$

where N is the sample size and $O_i = \sum_{n=1}^{N} \xi_{n,i}$ the number of sampled observations equal to i, is asymptotically distributed as $\chi^2(k-1)\lambda$, with $\lambda > 1$. Therefore, the presence of correlation among observations attenuates the effect of N, which multiplies the chi-squared distance between the observed frequencies and the expected probabilities. This is a key feature for statistical applications in the framework of a "big sample", where a small value of the chi-squared distance might be significant, and hence a correction related to the correlation between observations is required. In [3,5], a possible application in the context of clustered data was described, with independence between clusters and correlation due to a reinforcement mechanism inside each cluster.

In [4], the RP urn was applied as a good model for the evolution of the sentiment associated with Twitter posts. Precisely, we analyzed three data sets: (i) the "COVID-19 epidemic" data set covers the period from 21 February to 20 April to 2020 and includes tweets in Italian about the COVID-19 epidemic; (ii) the "Migration debate" data set refers to the period from 23 January to 22 February 2019 and the collected posts are related to the Italian debate on migration; (iii) the "10 days of traffic" data set collects the entire traffic of posts in Italian in the period from 1 September to 10 September 2019. For every post, the relative sentiment, that is, the positive or negative connotation of the text, was computed using the polyglot python module developed in [6], which provides a numerical value $v \in [-1, 1]$ for the sentiment of a post (for a survey on sentiment analysis, also known as opinion mining, we refer to [7] and references therein). We fixed a threshold T so that a tweet with $v > T$ was classified as a tweet with a positive sentiment and one with $v < -T$ was classified as a tweet with a negative sentiment. Tweets with a value $v \in [-T, T]$ were discarded. We took the following different values for T: $T = 0$, $T = 0.35$ and $T = 0.5$. We applied the RP urn model, ordering the tweets according to their creation time and taking each tweet with a positive/negative classification as an extraction in the urn model. More specifically, we applied the RP model with $k = 2$: the time series of the tweets represents the time series of the extractions from the urn, that is, the random variables $\xi_{n,1}$. The event $\{\xi_{n,1} = 1\}$ means that tweet n exhibits a positive sentiment, while $\{\xi_{n,1} = 0\}$ means that tweet n exhibits a negative sentiment. For all the considered data sets, the estimated values of β were strictly smaller than 1, but very near to 1 (details about the parameters estimation can be found in [4]). Note that the RP urn dynamics with such a value for β cannot be approximated by the standard Pólya urn ($\beta = 1$), because one would lose the fluctuations of the predictive means and the possibility of touching the barriers $\{0, 1\}$. In this work, we show that the law of such an RP urn process can be approximated by a Wright—Fisher diffusion with mutation. More precisely, we prove that the multidimensional Wright—Fisher diffusion with mutation can be obtained as a suitable limit of the predictive means associated with a family of RP urns with $\beta \in$[0,1), $\beta \to 1$. As an example, in Figure 1, for the data set "COVID-19 epidemic", we show the plot of the process $(\psi_{n,1})_n$, reconstructed from the data (details about the reconstruction process can be found in [4]) and rescaled in time as $t = n(1-\beta)^2$, the plot of a simulated (by the Euler–Maruyama method) trajectory of the Wright—Fisher process, the plot of the approximation of this trajectory by means of the RP urn and the approximation of the data process by means of the standard Pólya urn.

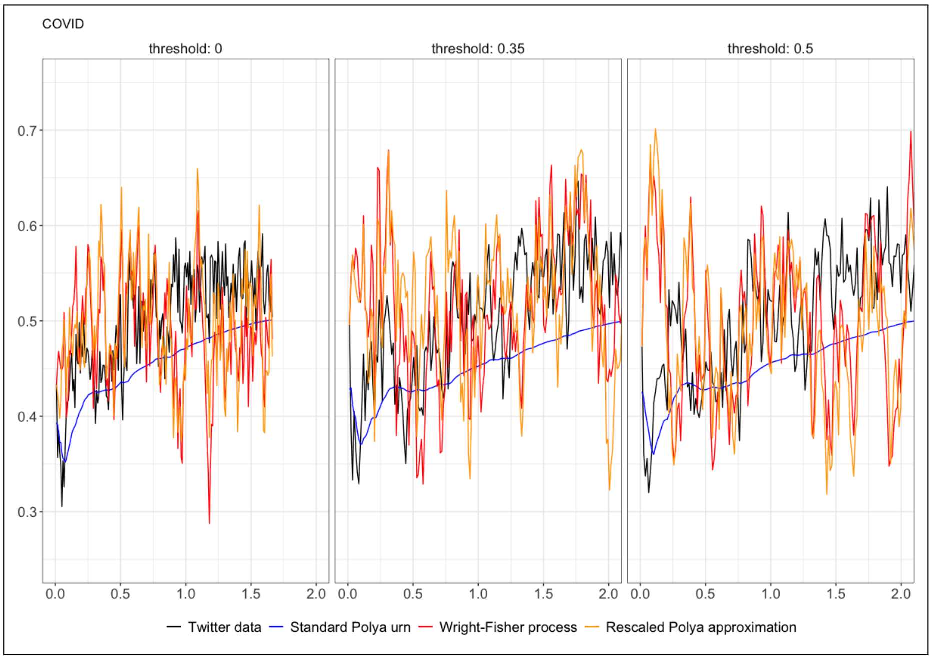

Figure 1. "COVID-19 epidemic" Twitter data set: the black line is the process $(\psi_{n,1})_n$, reconstructed from the data and rescaled in time as $t = n(1-\beta)^2$; the red line is a simulated trajectory of the Wright—Fisher process; the orange line is the approximation of this trajectory by means of the RP urn and the blue line is the approximation of the data process by means of the standard Pólya urn. The numbers 0, 0.35 and 0.5 refer to the values chosen for the threshold T. The corresponding estimated values for $1-\beta$ are: 0.000776 (8×10^{-4}), 0.00115 (11×10^{-4}) and 0.00130 (13×10^{-4}).

The Wright–Fisher (WF) class of diffusion processes models the evolution of the relative frequency of a genetic variant, or allele, in a large randomly mating population with a finite number k of genetic variants. When $k = 2$, the WF diffusion obeys the one-dimensional stochastic differential equation

$$dX_t = F(X_t)dt + \sqrt{X_t(1-X_t)}dW_t, \qquad X_0 = x_0, t \in [0,T]. \tag{1}$$

The drift coefficient, $F : [0,1] \to R$, can include a variety of evolutionary forces such as mutation and selection. For example, $F(x) = p_1 - (p_1 + p_2)x = p_1(1-x) - p_2 x$ describes a process with recurrent mutation between the two alleles, governed by the mutation rates $p_1 > 0$ and $p_2 > 0$. The drift vanishes when $x = p_1/(p_1 + p_2)$ which is an attracting point for the dynamics. Equation (1) can be generalized to the case $k > 2$. The WF diffusion processes are widely employed in Bayesian statistics, as models for time-evolving priors [8–11] and as a discrete-time finite-population construction method of the two-parameter Poisson–Dirichlet diffusion [12]. They have been applied in genetics [13–18], in biophysics [19,20], in filtering theory [21,22] and in finance [23,24].

The benefit coming from the proven limit result is twofold. First, the known properties of the WF process can give a description of the RP urn when the parameter β is strictly smaller than one, but very near to one. Second, the given result might furnish the theoretical base for a new simulation method of the WF process. Indeed, the simulation from Equation (1) is highly nontrivial because there is no known closed form expression for the transition function of the diffusion, even in the simple case with null drift [25].

The rest of the paper is organized as follows. In Section 2, we set up our notation and we formally define the RP urn model. Section 3 provides the main result of this work, that is, the convergence result of a suitable family of predictive means associated with RP urns with $\beta \to 1$. In Section 4, employing the boundary classification of the WF diffusion with mutation and connecting it to the parameters of the RP urn model, we introduce an RP urn with a value of β very near to 1 the notion of recessive subsets of colors and the notion of dominant color. These two concepts are related to the possibility of reaching the barriers 0 and 1 by the predictive means of the urn process. Finally, Section 5 summarizes the work and concludes it.

2. The Rescaled Pólya Urn

For a vector $\boldsymbol{x} = (x_1, \dots, x_k)^\top \in \mathbb{R}^k$, we set $|\boldsymbol{x}| = \sum_{i=1}^k |x_i|$ and $\|\boldsymbol{x}\|^2 = \boldsymbol{x}^\top \boldsymbol{x} = \sum_{i=1}^k |x_i|^2$. Moreover we denote by $\mathbf{1}$ and $\mathbf{0}$ the vectors with all the components equal to 1 and equal to 0, respectively.

Let $\alpha > 0$ and $\beta \geq 0$. At time-step 0, the urn contains $b_i + B_{0,i} > 0$ distinct balls of color i, with $i = 1, \dots, k$. We set $\boldsymbol{b} = (b_1, \dots, b_k)^\top$ and $\boldsymbol{B_0} = (B_{0,1}, \dots, B_{0,k})^\top$. We suppose $b = |\boldsymbol{b}| > 0$ and we set $\boldsymbol{p} = \frac{\boldsymbol{b}}{b}$. At each time-step $(n+1) \geq 1$, a ball is drawn at random from the urn and we define the random vector $\boldsymbol{\xi_{n+1}} = (\xi_{n+1,1}, \dots, \xi_{n+1,k})^\top$ as

$$\xi_{n+1,i} = \begin{cases} 1 & \text{when the drawn ball at time-step } n+1 \text{ is of color } i \\ 0 & \text{otherwise.} \end{cases}$$

The number of balls inside the urn is updated as follows:

$$\boldsymbol{N_{n+1}} = \boldsymbol{b} + \boldsymbol{B_{n+1}} \qquad \text{with} \qquad \boldsymbol{B_{n+1}} = \beta \boldsymbol{B_n} + \alpha \boldsymbol{\xi_{n+1}}\,, \tag{2}$$

which gives

$$\boldsymbol{B_n} = \beta^n \boldsymbol{B_0} + \alpha \beta^n \sum_{h=1}^{n} \beta^{-h} \boldsymbol{\xi_h}\,. \tag{3}$$

Similarly, from the equality

$$|\boldsymbol{B_{n+1}}| = \beta |\boldsymbol{B_n}| + \alpha\,,$$

we get, using $\sum_{h=0}^{n-1} x^h = (1 - x^n)/(1 - x)$,

$$|\boldsymbol{B_n}| = \beta^n |\boldsymbol{B_0}| + \alpha \sum_{h=1}^{n} \beta^{n-h} = \beta^n \left(|\boldsymbol{B_0}| - \frac{\alpha}{1-\beta} \right) + \frac{\alpha}{1-\beta}\,. \tag{4}$$

Setting $r_n^* = |\boldsymbol{N_n}| = b + |\boldsymbol{B_n}|$, that is the total number of balls inside the urn at time-step n, we get the relations

$$r_{n+1}^* = r_n^* + (\beta - 1)|\boldsymbol{B_n}| + \alpha \tag{5}$$

and

$$r_n^* = b + \frac{\alpha}{1-\beta} + \beta^n \left(|\boldsymbol{B_0}| - \frac{\alpha}{1-\beta} \right). \tag{6}$$

Denoting by $\mathcal{F}_0$ the trivial σ-field and setting $\mathcal{F}_n = \sigma(\boldsymbol{\xi_1}, \dots, \boldsymbol{\xi_n})$ for $n \geq 1$, the conditional probabilities $\boldsymbol{\psi_n} = (\psi_{n,1}, \dots, \psi_{n,k})^\top$ of the extraction process, also called *predictive means*, are

$$\boldsymbol{\psi_n} = E[\boldsymbol{\xi_{n+1}} | \mathcal{F}_n] = \frac{\boldsymbol{N_n}}{|\boldsymbol{N_n}|} = \frac{\boldsymbol{b} + \boldsymbol{B_n}}{r_n^*} \qquad n \geq 0 \tag{7}$$

and, from (3) and (4), we have

$$\boldsymbol{\psi_n} = \frac{\boldsymbol{b} + \beta^n \boldsymbol{B_0} + \alpha \sum_{h=1}^{n} \beta^{n-h} \boldsymbol{\xi_h}}{b + \frac{\alpha}{1-\beta} + \beta^n \left(|\boldsymbol{B_0}| - \frac{\alpha}{1-\beta} \right)}\,. \tag{8}$$

The dependence of $\boldsymbol{\psi_n}$ on $\boldsymbol{\xi_h}$ is regulated by the factor $f(h,n) = \alpha \beta^{n-h}$, with $1 \leq h \leq n$, $n \geq 0$. In the case of the standard Eggenberger—Pólya urn (i.e., the case $\beta = 1$), each observation $\boldsymbol{\xi_h}$ has the same "weight" $f(h,n) = \alpha$. Instead, when $\beta < 1$ the factor $f(h,n)$ increases with h, and the main contribution is given by the most recent drawings. The case $\beta = 0$ is an extreme case, for which $\boldsymbol{\psi_n}$ depends only on the last drawing $\boldsymbol{\xi_n}$.

By means of (7), together with (2) and (5), we get

$$\boldsymbol{\psi_{n+1}} - \boldsymbol{\psi_n} = -\frac{(1-\beta)}{r^*_{n+1}} b\big(\boldsymbol{\psi_n} - \boldsymbol{p}\big) + \frac{\alpha}{r^*_{n+1}}\big(\boldsymbol{\xi_{n+1}} - \boldsymbol{\psi_n}\big). \tag{9}$$

Setting $\Delta\boldsymbol{M_{n+1}} = \boldsymbol{\xi_{n+1}} - \boldsymbol{\psi_n}$ and letting $\epsilon_n = b(1-\beta)/r^*_{n+1}$ and $\delta_n = \alpha/r^*_{n+1}$, from (9) we obtain

$$\boldsymbol{\psi_{n+1}} - \boldsymbol{\psi_n} = -\epsilon_n(\boldsymbol{\psi_n} - \boldsymbol{p}) + \delta_n \Delta\boldsymbol{M_{n+1}}\,. \tag{10}$$

3. Main Result

Consider the RP urn with parameters $\alpha > 0$, $\beta \in [0,1)$, $b > 0$ and $\boldsymbol{B_0}$ such that $|\boldsymbol{B_0}| = r(\beta) = \alpha/(1-\beta)$. Consequently, the total number of balls in the urn along the time-steps is constantly equal to $r^*(\beta) = b + r(\beta)$ and if we denote by $\boldsymbol{\psi}^{(\beta)} = (\boldsymbol{\psi_n^{(\beta)}})_n$ the predictive means corresponding to the fixed value β, we have the dynamics

$$\boldsymbol{\psi_n^{(\beta)}} - \boldsymbol{\psi_{n-1}^{(\beta)}} = -\epsilon(\beta)\big(\boldsymbol{\psi_{n-1}^{(\beta)}} - \boldsymbol{p}\big) + \delta(\beta)\Delta\boldsymbol{M_n^{(\beta)}}\,, \tag{11}$$

where

$$\epsilon(\beta) = \frac{b(1-\beta)^2}{\alpha + b(1-\beta)}, \qquad \delta(\beta) = \frac{\alpha(1-\beta)}{\alpha + b(1-\beta)} \tag{12}$$

and $\boldsymbol{\Delta M_n^{(\beta)}} = \boldsymbol{\xi_n^{(\beta)}} - \boldsymbol{\psi_{n-1}^{(\beta)}}$. Note that we have $\epsilon(\beta) \sim c\delta(\beta)^2$ for $\beta \to 1$, with $c = b/\alpha > 0$. Finally, we define $\boldsymbol{X}^{(\beta)} = (\boldsymbol{X_t^{(\beta)}})_{t\geq 0}$, where

$$\boldsymbol{X_t^{(\beta)}} = \boldsymbol{\psi^{(\beta)}_{\lfloor t/(1-\beta)^2 \rfloor}} \qquad \Longleftrightarrow \qquad \boldsymbol{X_t^{(\beta)}} = \boldsymbol{\psi_{n-1}^{(\beta)}},\ t \in [\,(n-1)(1-\beta)^2, n(1-\beta)^2\,). \tag{13}$$

The following result holds true:

Theorem 1. *Suppose that $\boldsymbol{X_0^{(\beta)}}$ weakly converges towards some process $\boldsymbol{X_0}$ when $\beta \to 1$. Then, for $\beta \to 1$, the family of stochastic processes $\{\boldsymbol{X^{(\beta)}},\ \beta \in [0,1)\}$ weakly converges towards the k-alleles Wright—Fisher diffusion $\boldsymbol{X} = (\boldsymbol{X_t})_{t\geq 0}$, with type-independent mutation kernel given by $\boldsymbol{p}$ and with dynamics*

$$d\boldsymbol{X_t} = -b\frac{\boldsymbol{X_t} - \boldsymbol{p}}{\alpha}dt + \Sigma(\boldsymbol{X_t})d\boldsymbol{W_t}, \tag{14}$$

with $\Sigma(\boldsymbol{X_t})\Sigma(\boldsymbol{X_t})^\top = \Big(\mathrm{diag}(\boldsymbol{X_t}) - \boldsymbol{X_t}\boldsymbol{X_t}^\top\Big)$ and $\mathbf{1}^\top\Sigma(\boldsymbol{X_t}) = \mathbf{0}^\top$, that is,

$$\Sigma(\boldsymbol{X_t})_{ij} = \begin{cases} 0 & \textit{if } X_{t,i}X_{t,j} = 0 \textit{ or } i < j \\ \sqrt{X_{t,i}\frac{\sum_{l=i+1}^k X_{t,l}}{\sum_{l=i}^k X_{t,l}}} & \textit{if } i = j \textit{ and } X_{t,i}X_{t,j} \neq 0 \\ -X_{t,i}\sqrt{\frac{X_{t,j}}{\sum_{l=j}^k X_{t,l}\sum_{l=j+1}^k X_{t,l}}} & \textit{if } i > j \textit{ and } X_{t,i}X_{t,j} \neq 0. \end{cases} \tag{15}$$

Proof. Fix a sequence (β_n), with $\beta_n \in [0,1)$ and $\beta_n \to 1$. The sequence of processes $\{\boldsymbol{X}^{(\beta_n)},\ n \in \mathbb{N}\}$ is bounded, hence we have to prove the tightness of the sequence in the space $D^k[0,\infty)$ of right-continuous functions with the usual Skorohod topology, and the characterization of the law of the unique limit process.

For any $f \in C_b^2$, define

$$
\begin{aligned}
\gamma_n^{(\beta,f)}(\boldsymbol{x}) &= \widehat{A}^{(\beta)} f((n-1)(1-\beta)^2)(\boldsymbol{x}) \\
&= E\Big[\frac{f(\boldsymbol{X}^{(\beta)}_{\boldsymbol{n(1-\beta)^2}}) - f(\boldsymbol{X}^{(\beta)}_{\boldsymbol{(n-1)(1-\beta)^2}})}{(1-\beta)^2}\Big|\boldsymbol{X}^{(\beta)}_{\boldsymbol{(n-1)(1-\beta)^2}} = \boldsymbol{x}\Big] \\
&= E\Big[\frac{f(\boldsymbol{\psi}_{\boldsymbol{n}}^{(\beta)}) - f(\boldsymbol{\psi}_{\boldsymbol{n-1}}^{(\beta)})}{(1-\beta)^2}\Big|\boldsymbol{\psi}_{\boldsymbol{n-1}}^{(\beta)} = \boldsymbol{x}\Big] \\
&\overset{\text{by } \boldsymbol{\psi}_n^{(\beta)} - \boldsymbol{\psi}_{n-1}^{(\beta)} = -\epsilon(\beta)\left(\boldsymbol{\psi}_{n-1}^{(\beta)} - \boldsymbol{p}\right) + \delta(\beta)\Delta \boldsymbol{M}_n^{(\beta)}}{=} \frac{1}{(1-\beta)^2}\Bigg(E\Big[f(\boldsymbol{x}) + \sum_i \frac{\partial f}{\partial x_i}(\boldsymbol{x})(-\epsilon(\beta)(x_i - p_i) + \delta(\beta)\Delta M_{n,i}{}^{(\beta)}) \\
&\quad + \tfrac{1}{2}\delta(\beta)^2 \sum_{ij} \frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x})\Delta M_{n,i}^{(\beta)} \Delta M_{n,j}^{(\beta)} + O((1-\beta)^3)\Big|\mathcal{F}_{n-1}\Big] - f(\boldsymbol{x})\Bigg) \\
&= -\tfrac{b}{\alpha + b(1-\beta)} \sum_i \frac{\partial f}{\partial x_i}(\boldsymbol{x})(x_i - p_i) + \frac{1}{2}\tfrac{\alpha^2}{(\alpha + b(1-\beta))^2} \sum_{ij} \frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x})(x_i \mathbb{1}_{i=j} - x_i x_j) \\
&\quad + O(1-\beta)\,.
\end{aligned}
\tag{16}
$$

We note that, for any $f \in C_b^2$, the partial derivatives in (16) are uniformly bounded, as $\boldsymbol{x}$ belongs to the compact simplex $S = \{x_i \geq 0, \sum_i x_i = 1\}$. The family $\{\gamma_n^{(\beta,f)}(\boldsymbol{x}), n \in \mathbb{N}, \beta < 1, \boldsymbol{x} \in S\}$ is then uniformly integrable. Thus, as a consequence of [26] (Theorem 4) (or [27] (ch. 7.4.3, Theorem 4.3, p. 236)), we have that the sequence of processes $\{\boldsymbol{X}^{(\beta_n)}, n \in \mathbb{N}\}$ is tight in the space of right-continuous functions with the usual Skorohod topology. Since, for any n and t, $\boldsymbol{X}_{\boldsymbol{t}}^{(\beta_n)} \in S$, then $\boldsymbol{1}^\top \Sigma(\boldsymbol{X}_{\boldsymbol{t}}) = \boldsymbol{0}^\top$. Moreover, the generator of the limit process is determined by the limit

$$
\begin{aligned}
Af(t)(\boldsymbol{x}) &= \lim_{n\to\infty} \gamma^{(\beta_n,f)}_{\lfloor t/(1-\beta)^2 \rfloor}(\boldsymbol{x}) \\
&= -\tfrac{b}{\alpha} \sum_i \frac{\partial f}{\partial x_i}(\boldsymbol{x})(x_i - p_i) + \frac{1}{2}\sum_{ij} \frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x})(x_i \mathbb{1}_{i=j} - x_i x_j).
\end{aligned}
$$

Hence, the weak limit of the sequence of the bounded processes $\boldsymbol{X}^{(\beta_n)}$ is the diffusion process

$$
d\boldsymbol{X_t} = -b\frac{\boldsymbol{X_t} - \boldsymbol{p}}{\alpha}dt + \Sigma(\boldsymbol{X_t})d\boldsymbol{W_t}, \qquad \Sigma(\boldsymbol{X_t})\Sigma(\boldsymbol{X_t})^\top = \Big(\mathrm{diag}(\boldsymbol{X_t}) - \boldsymbol{X_t}\boldsymbol{X_t}^\top\Big).
$$

The expression (15) follows from [28] (Corollary 3). □

Remark 1 (Limiting ergodic distribution)**.** *Since the simplex has dimension $k-1$ with respect to the Lebesgue measure, it is convenient to change the notations. Let T^{k-1} be the $k-1$-dimensional simplex defined by*

$$
T^{k-1} := \{\boldsymbol{y} \in \mathbb{R}^{k-1} : y_1 \geq 0, \ldots, y_{k-1} \geq 0, 1 - y_1 - y_2 - \cdots - y_{k-1} \geq 0\},
$$

where, with the old definition, we have $x_i = y_i, i < k$ and $x_k := 1 - y_1 - y_2 - \cdots - y_{k-1}$. Obviously, there is a one-to-one natural correspondence between T^{k-1} and the simplex $\{\boldsymbol{x} \in \mathbb{R}^k : x_1 \geq 0, \ldots, x_k \geq 0, \sum_i x_i = 1\}$ defined by

$$
\boldsymbol{y} = (y_1, \ldots, y_{k-1}) \quad \longleftrightarrow \quad (y_1, \ldots, y_{k-1}, 1 - y_1 - y_2 - \cdots - y_{k-1}) = (x_1, \ldots, x_{k-1}, x_k) = \boldsymbol{x}.
$$

The Markov diffusion process $\boldsymbol{X_t}$ in (14) may be redefined as $\boldsymbol{Y_t} = (X_{t,1}, \ldots, X_{t,k-1})$ on $\boldsymbol{y} \in T^{k-1}$ with the corresponding generator

$$
Lf(\boldsymbol{y}) = -\tfrac{b}{\alpha}\sum_{i=1}^{k-1} \frac{\partial f}{\partial y_i}(\boldsymbol{y})(y_i - p_i) + \frac{1}{2}\sum_{i,j=1}^{k-1} \frac{\partial^2 f}{\partial y_i \partial y_j}(\boldsymbol{y})(y_i \mathbb{1}_{i=j} - y_i y_j). \tag{17}
$$

The Kolmogorov forward equation for the density $p(\boldsymbol{y},t)$ of the limiting process $\boldsymbol{Y}_t$ is

$$\frac{\partial}{\partial t}p(\boldsymbol{y},t)=\frac{1}{2}\Bigg(\tfrac{b}{\alpha}\sum_{i=1}^{k-1}\frac{\partial}{\partial y_i}\Big(p(\boldsymbol{y},t)(y_i-p_i)\Big)\\ +\sum_{i=1}^{k-1}\frac{\partial^2}{\partial y_i^2}\Big(y_i(1-y_i)p(\boldsymbol{y},t)\Big)-2\sum_{1\le i<j\le k-1}\frac{\partial^2}{\partial y_i\partial y_j}\Big(y_iy_jp(\boldsymbol{y},t)\Big)\Bigg). \quad (18)$$

Therefore, it is not hard to show that the limit invariant ergodic distribution is

$$p(\boldsymbol{y})=\frac{1}{B(2\frac{b}{\alpha}\boldsymbol{p})}(1-y_1-\cdots-y_{k-1})^{\frac{2b(1-p_1-\cdots-p_{k-1})}{\alpha}-1}\prod_{i=1}^{k-1}y_i^{\frac{2bp_i}{\alpha}-1}, \quad (19)$$

because it satisfies (18) *(see also [29]). The above distribution is the Dirichlet distribution Dir*$\big(2\frac{b}{\alpha}\boldsymbol{p}\big)$ *as a function of* $\boldsymbol{x}=(\boldsymbol{y},1-y_1-\cdots-y_{k-1})$.

Remark 2 (Transition density of the limit process)**.** *The transition density* $p(\boldsymbol{y_0},\boldsymbol{y};t)$ *is defined by*

$$P(\boldsymbol{Y_t}\in S|\boldsymbol{Y_0}=\boldsymbol{y_0})=\int_{S\cap T^{k-1}}p(\boldsymbol{y_0},\boldsymbol{y};t)d\boldsymbol{y}$$

and it can be represented in terms of series of orthogonal polynomials [30] as shown in [31]. Moreover, we refer to [9,32,33] for the explicit form of the reproducing kernel orthogonal polynomials.

4. Recessive and Dominant Colors in an RP Urn with β Near to 1

Let $\boldsymbol{J}=\{J_1,\ldots,J_{k_J}\}$ be a partition of $\{1,\ldots,k\}$, in that $J_l\neq\varnothing$, $J_{i_1}\cap J_{i_2}=\varnothing$, and $\cup_{l=1}^{k_J}=\{1,\ldots,k\}$. Here k_j denotes the cardinality of $\boldsymbol{J}$. Define the k_J-dimensional objects $(\boldsymbol{\psi_n^{(\beta,J)}})_n$, $(\boldsymbol{\xi_n^{(\beta,J)}})_n$ and $\boldsymbol{p}^{(J)}$ as

$$\left.\begin{aligned}\psi_{n,i}^{(\beta,\boldsymbol{J})}&=\sum_{l\in J_i}\psi_{n,l}^{(\beta)}\\ \xi_{n,i}^{(\beta,\boldsymbol{J})}&=\sum_{l\in J_l}\xi_{n,l}^{(\varepsilon)}\\ p_i^{(\boldsymbol{J})}&=\sum_{l\in J_i}p_l\end{aligned}\right\}\qquad\text{for } i=1,\ldots,k_J,$$

and $\boldsymbol{X_t^{(\beta,J)}}=\boldsymbol{\psi}^{(\beta,J)}_{\lfloor t/(1-\beta)^2\rfloor}$. With these definitions, from (11), we immediately get that $(\boldsymbol{\psi_n^{(\beta,J)}})_n$ is a k_J-dimensional RP urn following the dynamics

$$\boldsymbol{\psi_n^{(\beta,J)}}-\boldsymbol{\psi_{n-1}^{(\beta,J)}}=-\epsilon(\beta)\big(\boldsymbol{\psi_{n-1}^{(\beta,J)}}-\boldsymbol{p}^{(J)}\big)+\delta(\beta)\big(\boldsymbol{\xi_n^{(\beta,J)}}-\boldsymbol{\psi_{n-1}^{(\beta,J)}}\big) \quad (20)$$

and that Theorem 1 holds for $\boldsymbol{X_t^{(\beta,J)}}$. Consequently, the convergence to the Wright—Fisher diffusion still holds if we group together some components of the process. For instance, when we consider two groups of components, we have the following result:

Corollary 1. *Let* $\boldsymbol{J}=\{J,J^c\}$ *with* $J\neq\varnothing$, $J^c\neq\varnothing$. *Under the hypothesis of Theorem 1, each component of the sequence of processes* $\boldsymbol{X_t^{(\beta,J)}}$ *converges, for* $\beta\to 1$, *to the one-dimensional diffusion process with values in* $[0,1]$ *that satisfies the SDE*

$$dX_{t,i}^{(\boldsymbol{J})}=-b\frac{X_{t,i}^{(\boldsymbol{J})}-p_i}{\alpha}dt+(-1)^{i+1}\sqrt{X_{t,i}^{(\boldsymbol{J})}(1-X_{t,i}^{(\boldsymbol{J})})}dW_t.$$

In addition, $X_{t,1}^{(\boldsymbol{J})}=\sum_{l\in J}X_{t,l}$ *and* $X_{t,2}^{(\boldsymbol{J})}=\sum_{l\in J^c}X_{t,l}$.

Now, if we further specialize the grouping choice to $\boldsymbol{J} = (\{i\}, \{1, \ldots, i-1, i+1, \ldots, k\})$, we get:

Corollary 2. *Under the conditions of Theorem 1 the i-th component of the sequence of processes $\boldsymbol{X}^{(\beta)}$ converges, for $\beta \to 1$, to the one-dimensional diffusion $(X_{t,i})_{t\geq 0}$ with values in $[0,1]$ satisfying the SDE*

$$dX_{t,i} = -b\frac{X_{t,i} - p_i}{\alpha}dt + \sqrt{X_{t,i}(1 - X_{t,i})}dW_t.$$

For instance, the above two results are useful in order to translate the well-known classification of the boundaries of the WF process with mutation [34] (p. 239, Example 8) (see also [35]) to the RP urn model when the parameter β is strictly smaller than 1, but very near to 1. Indeed, Corollary 1 implies that $Z_t = \sum_{l\in J} X_{t,l}$ satisfies the SDE

$$\begin{aligned} dZ_t &= -b\frac{Z_t - \sum_{l\in J} p_l}{\alpha}dt + \sqrt{Z_t(1 - Z_t)}dW_t \\ &= \Big(-\frac{b}{\alpha}\Big(1 - \sum_{l\in J} p_l\Big)Z_t + \frac{b}{\alpha}\sum_{l\in J} p_l(1 - Z_t)\Big)dt + \sqrt{Z_t(1 - Z_t)}dW_t. \end{aligned}$$

Setting $a_0 = \frac{b}{\alpha}\sum_{l\in J} p_l$ and $a_1 = \frac{b}{\alpha} - a_0$ and noting that $\cap_{i\in J}\{X_{t,i} = 0\} = \{Z_t = 0\}$, we obtain:

(1) $a_0 < 1/2$, i.e., $\sum_{l\in J} p_l < \frac{\alpha}{2b}$, if and only if $P(\exists t\colon \cap_{i\in J}\{X_{t,i} = 0\}) = 1$;
(2) $a_0 \geq 1/2$, i.e., $\sum_{l\in J} p_l \geq \frac{\alpha}{2b}$, if and only if $P(\exists t\colon \cap_{i\in J}\{X_{t,i} = 0\}) = 0$.

With the same spirit, Corollary 2 states that $Z_t = 1 - X_{t,i}$ satisfies the SDE

$$\begin{aligned} dZ_t &= -b\frac{Z_t - \sum_{l\neq i} p_l}{\alpha}dt + \sqrt{(1 - Z_t)Z_t}dW_t \\ &= \Big(-\frac{b}{\alpha}p_i Z_t + \frac{b}{\alpha}(1 - p_i)(1 - Z_t)\Big)dt + \sqrt{Z_t(1 - Z_t)}dW_t. \end{aligned}$$

Setting $a_0 = \frac{b}{\alpha}(1 - p_i)$ and $a_1 = \frac{b}{\alpha} - a_0$, we get:

(3) $a_0 < 1/2$, i.e., $p_i > 1 - \frac{\alpha}{2b}$, if and only if $P(\exists t\colon \{X_{t,i} = 1\}) = 1$;
(4) $a_0 \geq 1/2$, i.e., $p_i \leq 1 - \frac{\alpha}{2b}$, if and only if $P(\exists t\colon \{X_{t,i} = 1\}) = 0$.

Therefore, for an RP urn with $\beta < 1$, but very near to 1, we can give the following definition:

Definition 1. *We call recessive a non-empty subset $J \subsetneq \{1, \ldots, k\}$ of colors such that $\sum_{l\in J} p_l < \frac{\alpha}{2b}$. We call dominant a color $i \in \{1, \ldots, k\}$ such that $\{1, \ldots, k\} \setminus \{i\}$ is recessive.*

Obviously, every subset of a recessive set is recessive. Moreover, when $\frac{\alpha}{b} > 2(1 - \min_i p_i)$, every set $J \subsetneq \{1, \ldots, k\}$ is recessive. The terms "recessive" and "dominant" are justified by the fact that, recalling properties (1)–(4) of the WF process, if a set of colors is recessive, then we can observe that at some times the corresponding predictive means of the urn process are very near to zero. On the contrary, when a color is dominant, we can observe that at some times the corresponding predictive mean of the urn process is very near to one. In Figure 2, we plot the process $(\psi_{n,1})$ related to the simulation of an RP urn with $k = 2$, $\alpha/b = 1$ and $p = 0.75$, where it is possible to observe the excursions near the barrier 1.

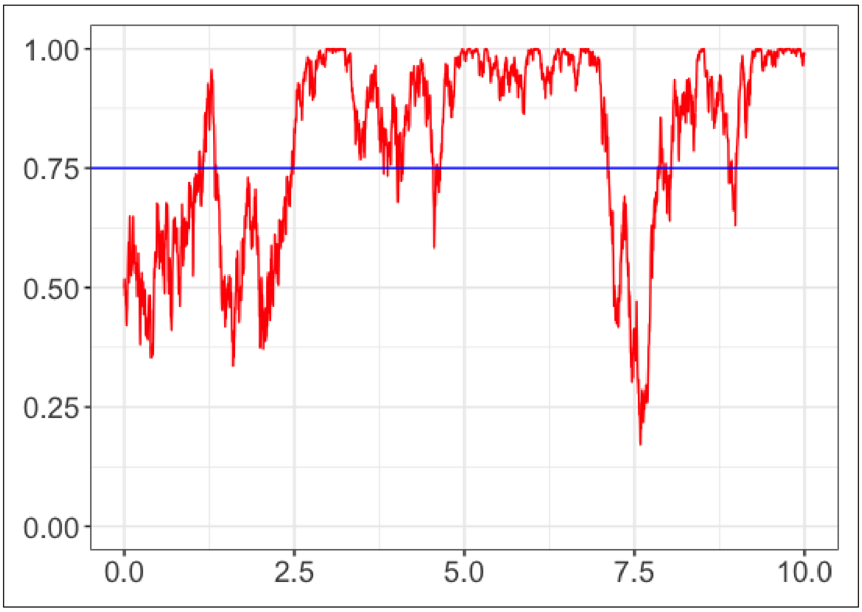

Figure 2. Simulation: plot of the process $(\psi_{n,1})$ related to the simulation of an RP urn with $k = 2$, $\alpha/b = 1$ and $p = 0.75$.

5. Conclusions

We have proven that the multidimensional WF diffusion with mutation can be obtained as the limit of the predictive means associated with a family of RP urns with $\beta < 1$, $\beta \to 1$. As a consequence, the known properties of the WF process can give a description of the RP urn when the parameter β is strictly smaller than 1, but very near to 1. For instance, starting from the known classification of the boundaries for the WF process and connecting it to the model parameters of the RP urn, we have obtained for an RP urn with a value of β very near to one, the notion of recessive subsets of colors and the notion of a dominant color. These two concepts are related to the possibility of reaching the barriers 0 and 1 by the predictive means of the urn process. Other classical problems, together with the corresponding known results for the WF process, can be found in [31]. These results can be used in order to give an approximated answer to the considered problems in the case of an RP urn with a value of β near 1.

Author Contributions: Both authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: Irene Crimaldi is partially supported by the Italian "Programma di Attività Integrata" (PAI), project "TOol for Fighting FakEs" (TOFFE) funded by the IMT School for Advanced Studies Lucca. This research received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: Both authors sincerely thank the organizers of the present special issue for their invitation to contribute and Fabio Saracco for having collected and shared with them the analyzed Twitter data sets. Giacomo Aletti is a member of the Italian group "Gruppo Nazionale per il Calcolo Scientifico" of the Italian Institute "Istituto Nazionale di Alta Matematica". Irene Crimaldi is a member of the Italian group "Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni" of the Italian Institute "Istituto Nazionale di Alta Matematica".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Eggenberger, F.; Pólya, G. Über die Statistik verketteter Vorgänge. *ZAMM-J. Appl. Math. Mech./Z. Angew. Math. Mech.* **1923**, *3*, 279–289. [CrossRef]
2. Mahmoud, H.M. *Pólya Urn Models*; Texts in Statistical Science Series; CRC Press: Boca Raton, FL, USA, 2009.
3. Aletti, G.; Crimaldi, I. The Rescaled Pólya Urn: Local reinforcement and chi-squared goodness of fit test. *Adv. Appl. Probab.* Available online: https://iris.imtlucca.it/handle/20.500.11771/19197#.YZNznboRVPZ (accessed on 1 November 2021).
4. Aletti, G.; Crimaldi, I.; Saracco, F. A model for the Twitter sentiment curve. *PLoS ONE* **2021**, *16*, e0249634. [CrossRef]
5. Aletti, G.; Crimaldi, I. Generalized Rescaled Pólya urn and its statistical applications. *arXiv* **2021**, arXiv:2010.06373.
6. Chen, Y.; Skiena, S. Building sentiment lexicons for all major languages. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), Baltimore, MD, USA, 22–27 June 2014; pp. 383–389.
7. Chakraborty, K.; Bhattacharyya, S.; Bag, R. A Survey of Sentiment Analysis from Social Media Data. *IEEE Trans. Comput. Soc. Syst.* **2020**, *7*, 450–464. [CrossRef]
8. Favaro, S.; Ruggiero, M.; Walker, S.G. On a Gibbs sampler based random process in Bayesian nonparametrics. *Electron. J. Stat.* **2009**, *3*, 1556–1566. [CrossRef]
9. Griffiths, R.C.; Spanò, D. Diffusion processes and coalescent trees. In *Probability and Mathematical Genetics, Papers in Honour of Sir John Kingman*; Bingham, N.H., Goldie, C.M., Eds.; LMS Lecture Note Series; Cambridge University Press: Cambridge, UK, 2010; Volume 378, pp. 358–375.
10. Mena, R.; Ruggiero, M. Dynamic density estimation with diffusive Dirichlet mixtures. *Bernoulli* **2016**, *22*, 901–926. [CrossRef]
11. Walker, S.G.; Hatjispyros, S.J.; Nicoleris, T. A Fleming-Viot process and Bayesian nonparametrics. *Ann. Appl. Probab.* **2007**, *17*, 67–80. [CrossRef]
12. Costantini, C.; De Blasi, P.; Ethier, S.; Ruggiero, M.; Spanò, D. Wright-Fisher construction of the two-parameter Poisson-Dirichlet diffusion. *Ann. Appl. Probab.* **2017**, *27*, 1923–1950. [CrossRef]
13. Bollback, J.P.; York, T.L.; Nielsen, R. Estimation of $2N_e s$ from temporal allele frequency data. *Genetics* **2008**, *179*, 497–502. [CrossRef]
14. Gutenkunst, R.N.; Hernandez, R.D.; Williamson, S.H.; Bustamante, C.D. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet.* **2009**, *5*, e1000695. [CrossRef] [PubMed]
15. Malaspinas, A.S.; Malaspinas, O.; Evans, S.N.; Slatkin, M. Estimating allele age and selection coefficient from time-serial data. *Genetics* **2012**, *192*, 599–607. [CrossRef]
16. Schraiber, J.; Griffiths, R.C.; Evans, S.N. Analysis and rejection sampling of Wright-Fisher diffusion bridges. *Theor. Popul. Biol.* **2013**, *89*, 64–74. [CrossRef] [PubMed]
17. Williamson, S.H.; Hernandez, R.; Fledel-Alon, A.; Zhu, L.; Bustamante, C.D. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 7882–7887. [CrossRef]
18. Zhao, L.; Lascoux, M.; Overall, A.D.J.; Waxman, D. The characteristic trajectory of a fixing allele: a consequence of fictitious selection that arises from conditioning. *Genetics* **2013**, *195*, 993–1006. [CrossRef]
19. Dangerfield, C.; Kay, D.; Burrage, K. Stochastic models and simulation of ion channel dynamics. *Procedia Comput. Sci.* **2010**, *1*, 1587–1596. [CrossRef]
20. Dangerfield, C.E.; Kay, D.; MacNamara, S.; Burrage, K. A boundary preserving numerical algorithm for the Wright—Fisher model with mutation. *BIT Numer. Math.* **2012**, *5*, 283–304. [CrossRef]
21. Chaleyat-Maurel, M.; Genon-Catalot, V. Filtering the Wright–Fisher diffusion. *ESAIM Probab. Stat.* **2009**, *13*, 197–217. [CrossRef]
22. Papaspiliopoulos, O.; Ruggiero, M. Optimal filtering and the dual process. *Bernoulli* **2014**, *20*, 1999–2019. [CrossRef]
23. Delbaen, F.; Shirakawa, H. An interest rate model with upper and lower bounds. *Asia-Pac. Financ. Mark.* **2002**, *9*, 191–209. [CrossRef]
24. Gourieroux, C.; Jasiak, J. Multivariate Jacobi process with application to smooth transitions. *J. Econom.* **2006**, *131*, 475–505. [CrossRef]
25. Jenkins, P.A.; Spanò, D. Exact simulation of the Wright-Fisher diffusion. *Ann. Appl. Probab.* **2017**, *27*, 1478–1509. [CrossRef]
26. Kushner, H.J. *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*; MIT Press Series in Signal Processing, Optimization, and Control; MIT Press: Cambridge, MA, USA, 1984; Volume 6.
27. Kushner, H.J.; Yin, G.G. *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed.; Applications of Mathematics; Springer: New York, NY, USA, 2003; Volume 35.
28. Tanabe, K.; Sagae, M. An Exact Cholesky Decomposition and the Generalized Inverse of the Variance-Covariance Matrix of the Multinomial Distribution, with Applications. *J. R. Stat. Soc. Ser. B (Methodol.)* **1992**, *54*, 211–219. [CrossRef]
29. Wright, S. *Evolution and the Genetics of Populations, Volume 2: Theory of Gene Frequencies*; Evolution and the Genetics of Populations; University of Chicago Press: Chicago, IL, USA, 1984.
30. Dunkl, C.F.; Xu, Y. *Orthogonal Polynomials of Several Variables*, 2nd ed.; Encyclopedia of Mathematics and Its Applications; Cambridge University Press: Cambridge, UK, 2014; Volume 155. [CrossRef]
31. Aletti, G.; Crimaldi, I. The rescaled Pólya urn and the Wright—Fisher process with mutation. *arXiv* **2021**, arXiv:2110.01853.
32. Griffiths, R.C.; Spanò, D. Orthogonal polynomial kernels and canonical correlations for Dirichlet measures. *Bernoulli* **2013**, *19*, 548–598. [CrossRef]

33. Griffiths, R.C.; Spanò, D. Multivariate Jacobi and Laguerre polynomials, infinite-dimensional extensions, and their probabilistic connections with multivariate Hahn and Meixner polynomials. *Bernoulli* **2011**, *17*, 1095–1125. [CrossRef]
34. Karlin, S.; Taylor, H.M. *A Second Course in Stochastic Processes*; A Subsidiary of Harcourt Brace Jovanovich; Academic Press, Inc.: New York, NY, USA; London, UK, 1981.
35. Huillet, T. On Wright–Fisher diffusion and its relatives. *J. Stat. Mech. Theory Exp.* **2007**, *2007*, P11006–P11006. [CrossRef]