Smoothing distributions for conditional Fleming–Viot and Dawson–Watanabe diffusions

Filippo Ascolani, Antonio Lijoi, Matteo Ruggiero

Carlo Alberto Notebooks

Smoothing distributions for conditional Fleming–Viot and Dawson–Watanabe diffusions

FILIPPO ASCOLANI

Department of Decision Sciences and BIDSA, Bocconi University

ANTONIO LIJOI

Department of Decision Sciences and BIDSA, Bocconi University

MATTEO RUGGIERO*

University of Torino and Collegio Carlo Alberto

April 27, 2022

Abstract

We study the distribution of the unobserved states of two measure-valued diffusions of Fleming–Viot and Dawson–Watanabe type, conditional on observations from the underlying populations collected at past, present and future times. If seen as nonparametric hidden Markov models, this amounts to finding the smoothing distributions of these processes, which we show can be explicitly described in recursive form as finite mixtures of laws of Dirichlet and gamma random measures respectively. We characterize the time-dependent weights of these mixtures, accounting for potentially different time intervals between data collection times, and fully describe the implications of assuming a discrete or a nonatomic distribution for the underlying process that drives mutations. In particular, we show that with a non atomic mutation offspring distribution, the inference automatically upweights mixture components that carry, as atoms, observed types shared at different collection times. The predictive distributions for further samples from the population conditional on the data are also identified and shown to be mixtures of generalized Pólya urns, conditionally on a latent variable in the Dawson–Watanabe case.

*matteo.ruggiero@unito.it

1 Introduction and summary of results

1.1 Problem statement

We consider the problem of evaluating the unknown state of a Fleming–Viot (FV) or a Dawson–Watanabe (DW) diffusion (Watanabe, 1968; Dawson, 1975; Fleming and Viot, 1979), two of the most popular measure-valued models for stochastic population dynamics, conditional on partial information on the underlying population collected at past, present and future times. See Ethier and Kurtz (1986; 1993); Dawson (2010); Feng (2010) for reviews on these processes. This task relates to the literature on hidden Markov models (Cappé et al., 2005) and consists in determining the so-called *smoothing distributions* of the marginal states X_{t_i} of an unobserved Markov process, often called the hidden or latent *signal*, evaluated at time t_i given samples $Y_{t_0}, \ldots, Y_{t_N}$ from the observation model, which is parameterised by the signal state, collected before and after t_i. Here $0 = t_0 < \cdots < t_N$ and $0 < i < N$ (for ease of exposition we denote here single observations at different times but later we will relax this constraint). These conditional distributions are typically used to improve previous estimates obtained at a certain time once additional observations become available at later times, often resulting in a smoother estimated trajectory for the unobserved signal, and they also constitute the starting point for performing Bayesian inference on the model parameters (see, e.g., Kon Kam King et al. (2020)). When all data are collected at the same time t_i, the problem reduces to a simple Bayesian update of the law of X_{t_i} given the observations. In the case of the FV and DW diffusions we consider here, taken at stationarity, this simplifies to updating the law of a Dirichlet process and a gamma random measure, results which date back to the seminal works of Ferguson (1973); Antoniak (1974); Lo (1982); Lo and Weng (1989). When only past and present data are considered, the smoothing problem reduces to a *filtering* problem, which evaluates the law of X_{t_i} given $Y_{t_0}, \ldots, Y_{t_i}$.

Closed form expressions for filtering and smoothing distributions can typically only be obtained for a handful of simple models even when the signal is finite-dimensional, e.g., by assuming linear Gaussian systems for both signal and observations or by taking a finite state space for the signal. In these cases the associated computations have complexity that grows linearly with the number of observations. Other modelling formulations, instead, typically require some form of approximation (see, e.g., Del Moral (1998); Doucet et al. (2011)). Intermediate results can sometimes be obtained and entail computations whose complexity grows polynomially with the number of observations, yielding so-called *computable filters* Chaleyat-Maurel and Genon-Catalot (2006). Along these lines, Chaleyat-Maurel and Genon-Catalot (2009) obtained a computable filter for a two-type Wright–Fisher (WF) diffusion and Papaspiliopoulos and Ruggiero (2014) for a K-type version, while Papaspiliopoulos et al. (2016) extended the latter to measure-valued signals driven by a FV or a DW diffusion, whereby

the associated hidden Markov model can be regarded as nonparametric. Recently, Kon Kam King et al. (2020) solved the smoothing problem for a K-type WF diffusion.

In this paper we characterize the laws of the marginal states of FV and DW diffusions, given samples from the respective underlying populations collected before and after the state temporal index, thus solving the smoothing problem for nonparametric hidden Markov models with these measure-valued unobserved signals. We show that these distributions can be written as finite mixtures of laws of Dirichlet and gamma random measures respectively, whose time-dependent mixture weights are fully described and can account for different time intervals between data collection times. As a byproduct of the above results, we describe the predictive distribution for further samples from the population given the entire dataset, which are shown to be mixtures of generalized Pólya urns. Our results prove that computable smoothing and conditional sampling from the population are feasible with signals given by these FV and DW processes, bringing forward these models as possible canonical choices in a nonparametric framework for hidden Markov models.

To introduce our setting in more detail, let $\mathcal{Y}$ be a Polish space and α a finite non null Borel measure on $\mathcal{Y}$. The class of FV processes we consider here is the so-called infinitely-many-neutral alleles model (Ethier and Kurtz, 1986) with parent-independent mutation, whose transition function can be written (Ethier and Griffiths, 1993a; Walker et al., 2007)

$$
P_t^{\mathrm{FV}}(x, \mathrm{d}x') = \sum_{m=0}^{\infty} d_m(t) \int_{\mathcal{Y}^m} \Pi_{\alpha + \sum_{i=1}^m \delta_{y_i}}(\mathrm{d}x') x^m(\mathrm{d}y_1, \ldots, \mathrm{d}y_m). \tag{1}
$$

Here x^m denotes an m-fold product measure, $d_m(t)$ is the probability that a $\mathbb{Z}_+$-valued death process starting at infinity with infinitesimal rates $n(\theta + n - 1)/2$ from n to $n-1$ is in state m at time t (see Griffiths, 1980; Tavaré, 1984), and Π_α is the law of a Dirichlet random measure (Ferguson, 1973), i.e., such that the projection over any Borel partition $A_1, \ldots, A_K$ of $\mathcal{Y}$ has the Dirichlet distribution with parameters $\alpha(A_1), \ldots, \alpha(A_K)$. Alternatively,

$$
\Pi_\alpha(\cdot) = \mathbb{P}\left\{ \sum_{i \geq 1} p_i \delta_{V_i} \in \cdot \right\}, \qquad (p_1, p_2, \ldots) \sim \mathrm{PD}(\theta), \qquad V_i \overset{\mathrm{iid}}{\sim} P_0, \tag{2}
$$

where $\mathrm{PD}(\theta)$ denotes a Poisson–Dirichlet distribution (Kingman, 1975), with parameter $\theta := \alpha(\mathcal{Y})$, and $P_0 := \alpha/\theta$. Note that Π_α is the reversible measure of (1), while P_0, often called the mutant *offspring distribution*, is the stationary distribution of the $\mathcal{Y}$-valued autonomous Markov process that drives mutations in the population. In this paper we will consider both cases of P_0 being discrete and nonatomic, the former reducing the above formulation to the so-called *unlabeled* model (Ethier and Kurtz, 1981), and we will show the different statistical implications that this choice determines in terms of smoothing.

Next we will consider the class of DW processes given by the branching measure-valued diffusions studied in Ethier and Griffiths (1993b), which takes values in the space of finite positive measures and has transition function

$$P_t^{\mathrm{DW}}(z, \mathrm{d}z') = \sum_{m=0}^{\infty} d_m^{|z|,\beta}(t) \int_{\mathcal{Y}^m} \Gamma_{\alpha+\sum_{i=1}^m \delta_{y_i}}^{\beta+S_t}(\mathrm{d}z')(z/|z|)^m(\mathrm{d}y_1, \ldots, \mathrm{d}y_m), \qquad (3)$$

where $|z| := z(\mathcal{Y})$ denotes, here and throughout, the total mass of z,

$$d_m^{|z|,\beta}(t) = \mathrm{Po}(m \mid S_t|z|), \quad S_t = \beta/(e^{\beta t/2} - 1),$$

with $\mathrm{Po}(\cdot|\lambda)$ a Poisson pmf with mean λ, and Γ_α^β is the law of a gamma random measure with shape α and rate $\beta > 0$, i.e., a positive atomic measure z such that for any finite collection of pairwise disjoint Borel subsets $A_1, \ldots, A_K$ of $\mathcal{Y}$, the random variables $z(A_1), \ldots, z(A_K)$ are independent and $z(A_i) \sim \mathrm{Gamma}(\alpha(A_i), \beta)$.

FV processes have well-known countable representations in terms of limits of particle systems when the population size goes to infinity (Ethier and Kurtz, 1993; Donnelly and Kurtz, 1996), and it is therefore a well established interpretation that they describe the evolving frequencies of alleles observed at a single locus in an ideally infinite population. On the other hand, DW processes arise as scaling limits of branching particle systems (?) and have a close relationship with FV processes, as, roughly speaking, the latter can be constructed as a normalised, unit-mass version of the former (Etheridge and March, 1991; Perkins, 1991). It is therefore natural to see these models and the associated sampling from the underlying particle representations as dynamic extensions of classical statistical settings that are widely used in Bayesian nonparametric statistics. More precisely, denoting by $Y_t^i \in \mathcal{Y}$ the i-th observation collected at time t, for the first model we let

$$Y_t^i|X_t = x \overset{\mathrm{iid}}{\sim} x, \quad X \sim \mathrm{FV}_\alpha. \qquad (4)$$

Here $X \sim \mathrm{FV}_\alpha$ denotes that X is a FV process with transition (1), and given $X_t = x$, the observations are conditionally *iid* random samples from the current underlying population summarised by the atomic probability measure x (recall that (2) is supported by discrete probability measures). For the second model, we let

$$\mathbf{Y}_t^i|Z_t = z \overset{\mathrm{iid}}{\sim} \mathrm{PP}(z), \quad Z \sim \mathrm{DW}_{\alpha,\beta}. \qquad (5)$$

Here $Z \sim \mathrm{DW}_{\alpha,\beta}$ denotes that Z is a DW process with transition (3), and given $Z_t = z$, the i-th draw consists in a vector of observations from a Poisson point process with intensity z. That is, a number $m \sim \mathrm{Po}(|z|)$ of random samples are collected from the normalised measure $\bar{z} := z/|z|$, i.e.

$$Y_{t,j}^i|Z = z, m \overset{\mathrm{iid}}{\sim} \bar{z}, \quad j = 1, \ldots, m, \qquad m|Z = z \sim \mathrm{Po}(|z|). \qquad (6)$$

4

Both settings can be seen as nonparametric hidden Markov models, since the signals are measure-valued and the observations are conditionally independent given the signal state. In this framework, we are interested in determining the distributions of X_t and Z_t given observations from the respective models collected at times before and after t.

1.2 Summary of results

We informally summarize our main results on the smoothing distributions in the following theorem, and then briefly discuss its interpretation and implications. We will denote throughout by $\mathcal{L}(X_t)$ and $\mathcal{L}(Z_t)$ the laws of X_t and Z_t, and assume observations are taken at discrete times $0 = t_0 < t_1 < \cdots < t_N = T$ where equidistance between collection times is not assumed. We also adopt here the compact notation $\mathbf{Y}_{t_{0:N}}$ for the entire dataset, postponing for the later sections the specification of its precise structure.

Theorem 1.1. *Assume model* (4) *with* X *being a FV diffusion with transition function* (1), *and that the data* $\mathbf{Y}_{t_{0:N}}$ *feature* K *distinct values* $y_1, \ldots, y_K$. *Then for every* $0 \leq i \leq N$ *there exists a set* $\mathbf{M} \subset \mathbb{Z}_+^K$ *of finite cardinality and weights* $\{w_{\mathbf{m}}, \mathbf{m} \in \mathbf{M}\}$ *summing up to one such that*

$$\mathcal{L}(X_{t_i} | \mathbf{Y}_{t_{0:N}}) = \sum_{\mathbf{m} \in \mathbf{M}} w_{\mathbf{m}} \Pi_{\alpha + \sum_{j=1}^K m_j \delta_{y_j}}. \tag{7}$$

Similarly, assume model (5) *with* Z *being a DW diffusion with transition function* (3), *and that the data* $\mathbf{Y}'_{t_{0:N}}$ *feature* K' *distinct values* $y'_1, \ldots, y'_{K'}$. *Then for every* $0 \leq i \leq N$ *there exists a set* $\mathbf{M}' \subset \mathbb{Z}_+^{K'}$ *of finite cardinality, a constant* $b \in \mathbb{R}_+$ *and weights* $\{w_{\mathbf{m}}^b, \mathbf{m} \in \mathbf{M}'\}$ *summing up to one, such that*

$$\mathcal{L}(Z_{t_i} | \mathbf{Y}'_{t_{0:N}}) = \sum_{\mathbf{m} \in \mathbf{M}'} w_{\mathbf{m}}^b \Gamma_{\alpha + \sum_{j=1}^{K'} m_j \delta_{y'_j}}^{\beta + b}.$$

The previous results states that all smoothing distributions associated to conditional FV and DW diffusions, given samples from the respective past, present and future populations, are finite mixtures of laws of Dirichlet and gamma random measures respectively. The later sections will present all the details that characterize the elements displayed above, among which the precise form of the mixture weights, which have a recursive nature, and the fact that the mixture components carry only part of the information available from the dataset, through the atomic components $\sum_{j=1}^K m_j \delta_{y_j}$ and $\sum_{j=1}^{K'} m_j \delta_{y'_j}$ of the parameter measures. An interesting implication of our results emerges when α is nonatomic, i.e., when the mutant offsprings in the population underlying the measure-valued diffusions are new types with probability one. In this case, the smoothing distributions become highly informative from an inferential perspective, and provide an automatic upweighting of the information which

5

emerged at more than one collection time. Specifically, mixture components which carry types observed at multiple times (as atoms in the parameter measure) end up having weights of a greater order of magnitude with respect to the others. This provides valuable information on the population lineage structure and is consistent with the expected behaviour of a smoothing distribution, supposed to remove some of the noise incorporated in the previous estimates.

Another byproduct of our results is the availability of the conditional laws for further samples from the population at a certain time, e.g., in the context of (7), the laws of $Y'_{t_i}|\mathbf{Y}_{t_{0:N}}$, of $Y''_{t_i}|\mathbf{Y}_{t_{0:N}}, Y'_{t_i}$, and so on. We identify these distributions, which extend results obtained in Ascolani et al. (2021) and are shown for both models to take the form of mixtures of generalized Pólya urn schemes with time-dependent weights, in the DW case conditional on a latent variable sampled from a mixture of negative binomial densities.

Note that a standard approach for obtaining the above smoothing distributions would in principle entail computing the marginal law of the signal at time t_n via $\int \xi_{t_{n-1}}(dx)P_{t_n-t_{n-1}}(x, dx')$ (cf. (9) below) where P_t is the transition function of the signal and $\xi_{t_{n-1}}$ here stands for the law of the signal at time t_{n-1} given some past data, the output of which must then be conditioned to future data points. These operations, carried out directly, are particularly troublesome for various reasons, the most prominent being the fact that the transition functions (1) and (3) are intractable (in fact $d_m(t)$ itself has an infinite series expansion; see Jenkins and Spanò (2017)). It is thus unfeasible to apply Bayes' Theorem directly in this general framework for updating with new observations since the signals are measure-valued. Theorem 1.1 instead shows that with the two models we consider, by carefully combining arguments that exploit the duality properties of the signals and the projective properties of the laws involved, the intractable measure given by conditioning the above integral to future observations can in fact be reduced to a finite mixture whose computation can therefore be carried out exactly.

1.3   General notation and organisation of the paper

We will take as sampling (or *type*) space a complete and separable metric space $\mathcal{Y}$, so that the DW process takes values on the space of finite non null atomic measures on $\mathcal{Y}$ and the FV on its subspace of probability measures. We will denote the FV and DW states by X_t and Z_t respectively, while the corresponding bold capitals $\mathbf{X}_t$ and $\mathbf{Z}_t$ will later denote the K-dimensional respective projections onto measurable partitions of $\mathcal{Y}$, at each fixed t. The main parameter for both (1) and (3) is a finite non null measure α on $\mathcal{Y}$, with total mass $\theta = \alpha(\mathcal{Y})$ and normalized version $P_0 = \alpha/\alpha(\mathcal{Y})$. Integer vectors of multiplicities will be denoted by $\mathbf{m}, \mathbf{n}, \ldots$, with $\mathbf{m} = (m_1, \ldots, m_K) \in \mathbb{Z}_+^K$ and $|\mathbf{m}| = \sum_{j=1}^K m_j$, such that $\mathbf{m} \le \mathbf{n}$

if and only if $m_j \leq n_j$ for every j. The symbol $\mathbf{0}$ will denote the vector containing all zeros. These multiplicities will sometimes carry a subscript to indicate to what collection time they refer to, e.g., $\mathbf{n}_i$ will be a shorthand notation for $(n_{t_i,1}, \ldots, n_{t_i,K})$ referred to time t_i. Similarly, observations from (4) collected at time t_i will be denoted $\mathbf{Y}_i$ for m samples $Y_{t_i}^1, \ldots, Y_{t_i}^m$, with m given. We will not index $\mathbf{Y}_i$ by the sample size m as the information available will be encoded in the distinct types $y_1, \ldots, y_K$ observed and their respective multiplicities $\mathbf{n}_i$. The notation for a collection of observations from model (5) is a bit more involved and will be described in detail in Section 4.1.

The rest of the paper is organised as follows. In Section 2 we define certain operators on measures which are useful for computing several quantities of interest for finite-dimensional signals. These will then be exploited in the subsequent sections for obtaining our results on the measure-valued signals. In Sections 3 and 4 we study the FV and DW cases respectively, where each section starts with preliminary results particularly focussed on the finite-dimensional model counterparts, to then derive the smoothing distributions in full detail and conclude by discussing the implications on conditional sampling of further observations from the population.

2 Some operators on measures

To obtain the smoothing distributions for the two measure-valued signals discussed in the introduction, we are going to exploit the projective properties of the FV and the DW process. To this end, we need to set a few tools that ease notation and computation for the respective finite-dimensional counterparts.

Consider a Markov process $\mathbf{X}$ on $\mathbb{R}^K$, $K < \infty$, with transition function P_t and initial distribution ν. We assume P_t is reversible with reversible measure ν_0. In this section we regard $\mathbf{X}$ as generic, but as anticipated above this setting will be used to model finite-dimensional projections of the measure-valued diffusions X_t and Z_t. The dimension K can therefore be thought of as representing the number of cells in which types in the population have been grouped or binned. Accordingly, *iid* observations collected at time t given $\mathbf{X}_t = \mathbf{x}_t$ generate multiplicities associated to the K groups which can be encoded into a vector $\mathbf{n} \in \mathbb{Z}_+^K$, whose associated density is $p(\mathbf{n}|\mathbf{x}_t)$. We can then define the following operators acting on measures ξ:

- *Update*:
$$\mathcal{U}_{\mathbf{n}}(\xi)(\mathrm{d}\mathbf{x}) := \frac{p(\mathbf{n}|\mathbf{x})\xi(\mathrm{d}\mathbf{x})}{p_\xi(\mathbf{n})}, \qquad p_\xi(\mathbf{n}) := \int p(\mathbf{n}|\mathbf{x})\xi(\mathrm{d}\mathbf{x}). \tag{8}$$

  This provides the conditional measure ξ given observations with associated multiplici-

ties $\mathbf{n}$. It is analogous to Bayes' Theorem with $\mathbf{x}$ acting as the random parameter and ξ as its prior: hence, $\mathcal{U}_{\mathbf{n}}(\xi)$ yields the posterior, whereas the denominator $p_\xi(\mathbf{n})$ is the marginal likelihood of $\mathbf{n}$ obtained by integrating out the random parameter $\mathbf{x}$. Here the multiplicities $\mathbf{n}$ are observed at the same time $\mathbf{x}$ refers to, as in (4) and (5).

- *Forward propagation*:

$$\mathcal{F}_t(\xi)(\mathrm{d}\mathbf{x}') := \xi P_t(\mathrm{d}\mathbf{x}') = \int \xi(\mathrm{d}\mathbf{x}) P_t(\mathbf{x}, \mathrm{d}\mathbf{x}'). \tag{9}$$

This yields the unconditional measure of X_{s+t} if ξ is that of X_s, once the initial state is integrated out.

- *Backward propagation*:

$$\mathcal{B}_t(\xi)(\mathrm{d}\mathbf{x}') := \xi Q_t(\mathrm{d}\mathbf{x}') = \int \xi(\mathrm{d}\mathbf{x}) Q_t(\mathbf{x}, \mathrm{d}\mathbf{x}'). \tag{10}$$

It is the forward propagation obtained by using the transition function of the time reversal of the signal, denoted here Q_t.

With a slight abuse of notation, when ξ is dominated by a sigma-finite measure μ on $\mathbb{R}^K$, we specialize the previous operators as acting on densities, e.g., if $\xi(\mathrm{d}\mathbf{x}) = f(\mathbf{x})\mu(\mathrm{d}\mathbf{x})$, then $\mathcal{U}_{\mathbf{n}}(f)(\mathbf{x}) := p(\mathbf{n}|\mathbf{x})f(\mathbf{x})/p_f(\mathbf{n})$ and $p_f(\mathbf{n}) := \int p(\mathbf{n}|\mathbf{x})f(\mathbf{x})\mu(\mathrm{d}\mathbf{x})$, and similarly for (9) and (10).

Note that the specific form of the backward transition is not necessary for our treatment, as we will leverage on Bayes' Theorem. See, e.g., Lemma 3.2 below.

Remark 2.1. Expanding on the above, and assuming all probability distributions of interest are dominated by μ, one could define a *smoothing operator* acting on two densities f_s, f_u, indexed by $s < u$, by letting

$$\mathcal{S}_{s,t,u}^{\mathbf{n}}(f_s, f_u)(\mathbf{x}) := C\, \mathcal{F}_{t-s}(f_s)(\mathbf{x}) \mathcal{B}_{u-t}(f_u)(\mathbf{x}) \mathcal{U}_{\mathbf{n}}(f_0)(\mathbf{x})/f_0(\mathbf{x})^2,$$

for every $\mathbf{x}$ such that $f_0(\mathbf{x}) > 0$, where $s < t < u$, f_0 is the density of ν_0 with respect to μ and C is a normalising constant that makes the left hand side a density. This yields the distribution of $\mathbf{X}_t$, given observations at time t, if $X_s \sim f_s$ and $X_u \sim f_u$, obtained by jointly propagating $\mathbf{X}_s$ forward of a $t - s$ interval, $\mathbf{X}_u$ backward of a $u - t$ interval, and then conditioning on observations collected at time t. The rationale of this operator can be outlined by considering that, for $t_0 < t_1 < t_2$, if $y_{t_1}|\mathbf{x}_{t_1} \sim p(y_{t_1}|\mathbf{x}_{t_1})$, we have

$$p(\mathbf{x}_{t_1}|\mathbf{x}_{t_0}, y_{t_1}, \mathbf{x}_{t_2}) \propto p(\mathbf{x}_{t_1}|\mathbf{x}_{t_0})p(y_{t_1}|\mathbf{x}_{t_1}, \mathbf{x}_{t_0})p(\mathbf{x}_{t_2}|\mathbf{x}_{t_1}, \mathbf{x}_{t_0}, y_{t_1})$$
$$= p(\mathbf{x}_{t_1}|\mathbf{x}_{t_0})p(y_{t_1}|\mathbf{x}_{t_1})p(\mathbf{x}_{t_2}|\mathbf{x}_{t_1})$$

where we have used the fact that conditionally on $\mathbf{x}_{t_1}$, y_{t_1} is independent on everything else, together with the Markov property. By virtue of Bayes' Theorem, we now have $p(\mathbf{x}_{t_2}|\mathbf{x}_{t_1}) = p(\mathbf{x}_{t_2})p(\mathbf{x}_{t_1}|\mathbf{x}_{t_2})/p(\mathbf{x}_{t_1})$ and $p(\mathbf{x}_{t_1}|y_{t_1}) = p(\mathbf{x}_{t_1})p(y_{t_1}|\mathbf{x}_{t_1})/p(y_{t_1})$ whereby the previous expression is proportional to $p(\mathbf{x}_{t_1}|\mathbf{x}_{t_0})p(\mathbf{x}_{t_1}|\mathbf{x}_{t_2})p(\mathbf{x}_{t_1}|y_{t_1})/p(\mathbf{x}_{t_1})^2$. This operator will not be essential for our calculations of the next sections, but it can provide a unified treatment of the previous operators applied at stationarity. In fact, we have the following special cases:

$$\mathcal{S}^{\mathbf{0}}_{s,t,u}(f_s, f_0) = \mathcal{F}_{t-s}(f_s), \qquad \mathcal{S}^{\mathbf{0}}_{s,t,u}(f_0, f_u) = \mathcal{B}_{u-t}(f_u), \qquad \mathcal{S}^{\mathbf{n}}_{s,t,u}(f_0, f_0) = \mathcal{U}_{\mathbf{n}}(f_0).$$

$\square$

Appropriate compositions of the above operators allow to represent all quantities of interest in this framework. For example $p(\mathbf{x}_{t_i}|\mathbf{n}_{i-1}, \mathbf{n}_i, \mathbf{n}_{i+1}) = \mathcal{S}^{\mathbf{n}_i}_{t_{i-1},t_i,t_{i+1}}(\mathcal{U}_{\mathbf{n}_{i-1}}(f_0), \mathcal{U}_{\mathbf{n}_{i+1}}(f_0))$ identifies the distribution of $\mathbf{X}_{t_i}$ given observations at times t_{i-1}, t_i, t_{i+1}, obtained by first updating the stationary measure at times t_{i-1}, t_{i+1} given observations with multiplicities $\mathbf{n}_{t-1}, \mathbf{n}_{t_{i+1}}$ respectively, then propagating both distributions to the intermediate time t_i, and finally updating the output of the last operation given the multiplicities observed at time t_i. This type of distributions will be instrumental for the proof of Theorem 1.1.

3 Fleming–Viot signals

3.1 Preliminary results

In this section we assume the signal is a FV process and the observation model (4) holds. A projection of the Dirichlet law in (2) onto a measurable partition $(A_1, \ldots, A_K)$ of the sampling space $\mathcal{Y}$ yields a Dirichlet distribution with parameters $(\alpha(A_1), \ldots, \alpha(A_K))$, whose density with respect to the Lebesgue measure on the $(K-1)$-dimensional simplex, denoted $\pi_{\boldsymbol{\alpha}}(\mathbf{x})$, $\boldsymbol{\alpha} = (\alpha(A_1), \ldots, \alpha(A_K))$, is proportional to $\mathbf{x}^{\boldsymbol{\alpha}-\mathbf{1}} := x_1^{\alpha(A_1)-1} \ldots x_K^{\alpha(A_K)-1}$. With a little abuse of notation, we will use the symbol $\pi_{\boldsymbol{\alpha}}$ to denote both the Dirichlet density and the corresponding measure. Similarly, a projection of a FV process with transition (1) onto the same partition yields a Wright–Fisher diffusion, denoted $\mathrm{WF}_{\boldsymbol{\alpha}}$, with transition function

$$P_t(\mathbf{x}, d\mathbf{x}') = \sum_{m=0}^{\infty} d_m(t) \sum_{\mathbf{m} \in \mathbb{Z}_+^K : |\mathbf{m}|=m} \mathbf{x}^{\mathbf{m}} \binom{m}{\mathbf{m}} \pi_{\boldsymbol{\alpha}+\mathbf{m}}(d\mathbf{x}'),$$

and $d_m(t)$ as in (1), which has reversible distribution $\pi_{\boldsymbol{\alpha}}$. We will denote by $p_t(\cdot|\mathbf{x})$ the corresponding density function. In this scenario, (4) reduces to

$$Y_t^i|\mathbf{X}_t = \mathbf{x} \overset{\text{iid}}{\sim} \mathrm{Categorical}(\mathbf{x}), \quad \mathbf{X} \sim \mathrm{WF}_{\boldsymbol{\alpha}}, \tag{11}$$

whereby for each i, $Y_t^i = j$ with probability x_j, for $j = 1, \ldots, K$, and the update operator (8) yields the familiar Bayesian update for Dirichlet distributions $\mathcal{U}_{\mathbf{n}}(\pi_{\boldsymbol{\alpha}}) = \pi_{\boldsymbol{\alpha}+\mathbf{n}}$. It is useful to note for later reference that in (8) the marginal likelihood is

$$m(\mathbf{n}) := p_{\pi_\alpha}(\mathbf{n}) = \frac{B(\boldsymbol{\alpha}+\mathbf{n})}{B(\boldsymbol{\alpha})}, \qquad B(\boldsymbol{\alpha}) := \frac{\prod_{j=1}^K \Gamma(\alpha_j)}{\Gamma(|\boldsymbol{\alpha}|)} \tag{12}$$

often called Dirichlet-Categorical distribution. Define now

$$h(\mathbf{x}, \mathbf{n}) := \frac{p(\mathbf{n}|\mathbf{x})}{m(\mathbf{n})} = \frac{B(\boldsymbol{\alpha})}{B(\boldsymbol{\alpha}+\mathbf{n})}\mathbf{x}^{\mathbf{n}}, \qquad \mathbf{n} \in \mathbb{Z}_+^K, \tag{13}$$

where $p(\mathbf{n}|\mathbf{x})$ is the categorical likelihood in (11) expressed in terms of multiplicities of types. It will also be useful to note that for $\mathbf{n}, \mathbf{m} \in \mathbb{Z}_+^K$, we have

$$h(\mathbf{x}, \mathbf{n})h(\mathbf{x}, \mathbf{m}) = c(\mathbf{n}, \mathbf{m})h(\mathbf{x}, \mathbf{n}+\mathbf{m}) \tag{14}$$

where

$$c(\mathbf{n}, \mathbf{m}) = \frac{m(\mathbf{n}+\mathbf{m})}{m(\mathbf{n})m(\mathbf{m})} = \frac{B(\boldsymbol{\alpha})B(\boldsymbol{\alpha}+\mathbf{m}+\mathbf{n})}{B(\boldsymbol{\alpha}+\mathbf{n})B(\boldsymbol{\alpha}+\mathbf{m})}.$$

Recall now that the WF diffusion is known to have moment-dual given by Kingman's *typed* coalescent. More specifically, let $\mathbf{M}_t$ be a death process on $\mathbb{Z}_+^K$ with rates $\lambda_{\mathbf{m}} = m_j(\theta+|\mathbf{m}|-1)/2$ from $\mathbf{m}$ to $\mathbf{m} - \mathbf{e}_j$, where $\mathbf{e}_j$ is the canonical vector in direction j. Then the following duality identity

$$\mathbb{E}[h(\mathbf{X}_t, \mathbf{m})|\mathbf{X}_0 = \mathbf{x}] = \mathbb{E}[h(\mathbf{x}, \mathbf{M}_t)|\mathbf{M}_0 = \mathbf{m}] \tag{15}$$

holds with h as in (13). We will denote by $p_{\mathbf{n},\mathbf{m}}(t)$ the transition probabilities of $\mathbf{M}_t$ (cf. Papaspiliopoulos et al. (2016), Section 4.2). The above duality is used to prove the following Lemma, needed later.

Lemma 3.1. *Let $\mathbf{n}_{i+1}$ be the multiplicities observed at time t_{i+1}. Then*

$$p(\mathbf{n}_{i+1}|\mathbf{x}_{t_i}) = m(\mathbf{n}_{i+1}) \sum_{\mathbf{0}\leq\mathbf{m}\leq\mathbf{n}_{i+1}} p_{\mathbf{n}_{i+1},\mathbf{m}}(t_{i+1} - t_i)h(\mathbf{x}_{t_i}, \mathbf{m}). \tag{16}$$

Proof. Using (13), we have

$$p(\mathbf{n}_{i+1}|\mathbf{x}_{t_i}) = \int p(\mathbf{n}_{i+1}|\mathbf{x}_{t_{i+1}})p_{t_{i+1}-t_i}(\mathbf{x}_{t_{i+1}}|\mathbf{x}_{t_i})\mathrm{d}\mathbf{x}_{t_{i+1}}$$

$$= m(\mathbf{n}_{i+1}) \int h(\mathbf{x}_{t_{i+1}}, \mathbf{n}_{i+1})p_{t_{i+1}-t_i}(\mathbf{x}_{t_{i+1}}|\mathbf{x}_{t_i})\mathrm{d}\mathbf{x}_{t_{i+1}}$$

$$= m(\mathbf{n}_{i+1})\mathbb{E}[h(\mathbf{X}_{t_{i+1}}, \mathbf{n}_{i+1})|\mathbf{X}_{t_i} = \mathbf{x}_{t_i}],$$

where the integral is over the $(K-1)$-dimensional simplex, from which (15) leads to the result. $\qquad\square$

The next Lemma formalizes the fact that a backward propagation, after a change of measure with respect to the stationary distribution, yields an analogous distributional result to a forward propagation, somehow carrying over the reversibility of FV processes to their conditional versions. In the following statement, the forward and backward operators $\mathcal{F}_t, \mathcal{B}_t$ are applied to laws of FV states by extension of (9)-(10), with P_t as in (1).

Lemma 3.2. *Assume* (4), *let* $y_1, \ldots, y_K$ *be distinct values and* $\mathbf{n} \in \mathbb{Z}_+^K$. *Then* $\mathcal{F}_t(\Pi_{\alpha + \sum_{j=1}^K n_j \delta_{y_j}}) = \mathcal{B}_t(\Pi_{\alpha + \sum_{j=1}^K n_j \delta_{y_j}})$, *with* $\mathcal{F}_t, \mathcal{B}_t$ *as in* (9)-(10), *and in particular*

$$\mathcal{B}_t\left(\Pi_{\alpha + \sum_{j=1}^K n_j \delta_{y_j}}\right) = \sum_{\mathbf{0} \leq \mathbf{k} \leq \mathbf{n}} p_{\mathbf{n},\mathbf{k}}(t) \Pi_{\alpha + \sum_{j=1}^K k_j \delta_{y_j}}. \tag{17}$$

Proof. Denote by $\mathbf{n}_{i-1}$ and $\mathbf{n}_{i+1}$ the multiplicities of types for observations sampled at times t_{i-1} and t_{i+1} respectively in the setting of (11). Then Papaspiliopoulos and Ruggiero (2014) showed that

$$\mathcal{F}_{t_i - t_{i-1}}(\pi_{\boldsymbol{\alpha} + \mathbf{n}_{i-1}})(\mathbf{x}_{t_i}) = \sum_{\mathbf{0} \leq \mathbf{k} \leq \mathbf{n}_{i-1}} p_{\mathbf{n}_{i-1},\mathbf{k}}(t_{i+1} - t_i) \pi_{\boldsymbol{\alpha} + \mathbf{k}}(\mathbf{x}_{t_i}).$$

Furthermore, using (8) first and then Lemma 3.1 we have

$$\begin{aligned}
\mathcal{B}_{t_{i+1} - t_i}(\pi_{\boldsymbol{\alpha} + \mathbf{n}_{i+1}})(\mathbf{x}_{t_i}) &= p(\mathbf{x}_{t_i} | \mathbf{n}_{i+1}) = \frac{p(\mathbf{x}_{t_i}) p(\mathbf{n}_{i+1} | \mathbf{x}_{t_i})}{m(\mathbf{n}_{i+1})} \\
&= p(\mathbf{x}_{t_i}) \sum_{\mathbf{0} \leq \mathbf{k} \leq \mathbf{n}_{i+1}} p_{\mathbf{n}_{i+1},\mathbf{k}}(t_{i+1} - t_i) h(\mathbf{x}_{t_i}, \mathbf{k}) \\
&= \sum_{\mathbf{0} \leq \mathbf{k} \leq \mathbf{n}_{i+1}} p_{\mathbf{n}_{i+1},\mathbf{k}}(t_{i+1} - t_i) h(\mathbf{x}_{t_i}, \mathbf{k}) p(\mathbf{x}_{t_i}) \\
&= \sum_{\mathbf{0} \leq \mathbf{k} \leq \mathbf{n}_{i+1}} p_{\mathbf{n}_{i+1},\mathbf{k}}(t_{i+1} - t_i) \pi_{\boldsymbol{\alpha} + \mathbf{k}}(\mathbf{x}_{t_i})
\end{aligned}$$

where the last identity follows from from (13) and (8). By equating $\mathbf{n}_{i-1}$ with $\mathbf{n}_{i+1}$ and $t_i - t_{i-1}$ with $t_{i+1} - t_i$, one can now see that $\mathcal{B}_t(\pi_{\boldsymbol{\alpha} + \mathbf{n}}) = \mathcal{F}_t(\pi_{\boldsymbol{\alpha} + \mathbf{n}})$, with $\mathcal{B}_t$ as in (10). The fact that $\mathcal{F}_t(\Pi_{\alpha + \sum_{j=1}^K n_j \delta_{y_j}})$ equals the right hand side of (17) now follows from Theorem 3.1 in Papaspiliopoulos et al. (2016), and the same proof can be used to show (17), by seeing $\mathcal{B}_t(\pi_{\boldsymbol{\alpha} + \mathbf{n}})$ as the projection of $\mathcal{B}_t(\Pi_{\alpha + \sum_{j=1}^K n_j \delta_{y_j}})$ onto an arbitrary partition, from which the first statement also follows. $\square$

Thanks to this equivalence between backward and forward propagation, we have that the same result of Lemma 3.1 holds with $\mathbf{n}_{i+1}$ replaced by $\mathbf{n}_{i-1}$, i.e., referred to time t_{i-1}, leading to the expression obtained by replacing $t_{i+1} - t_i$ with $t_i - t_{i-1}$ in the right hand side of (16).

To conclude the section, we can interpret (17) as providing a sort of *distributional bridge* which may be of independent interest. The heart of the intuition can be easily appreciated by considering a single observation $y \in \mathcal{Y}$ collected at time s, whereby conditioning (2) on y adds a point mass at y to the parameter measure α and $\mathcal{B}_t(\Pi_{\alpha+\delta_y}) = p_{1,0}(t)\Pi_\alpha + p_{1,1}(t)\Pi_{\alpha+\delta_y}$. Hence, an interval t prior to observing y, the distribution of X_{s-t} has law as in (2) which carries a point mass at y with probability $p_{1,1}(t)$. This is the probability that a single lineage in the underlying coalescent is not lost by mutation in an interval t. Furthermore, since $p_{1,1}(t) = e^{-\theta t/2} \to 1$ as $t \to 0$, we have $\mathcal{B}_t(\Pi_{\alpha+\delta_{y_0}}) \to \Pi_{\alpha+\delta_{y_0}}$ in total variation, i.e., the probability mass progressively concentrates on the component which carries the observed atom. More generally, we can state the following implication of Lemma 3.2, whose proof, omitted here, is a simple extension of the previous argument.

Corollary 3.3. *In the context of Lemma 3.2, we have $\mathcal{B}_t(\Pi_{\alpha+\sum_{i=1}^K n_i\delta_{y_i}}) \to \Pi_{\alpha+\sum_{i=1}^K n_i\delta_{y_i}}$ in total variation as $t \to 0$.*

Hence, conditioning on the knowledge of future observations from the underlying population yields a law for the current FV state that assigns gradually increasing probability to point masses associated to the observed values.

3.2   Smoothing distributions

Using the results of the previous section, the characterization of the smoothing distribution for conditional FV processes will be provided in three steps. First, in Theorem 3.4, we show that conditioning on observations collected at adjacent times yields a finite mixture of laws of Dirichlet random measures; then, in Proposition 3.6, we give a full description of the mixture weights for different choices of the offspring distribution P_0 (cf. (2)); finally, in Proposition 3.7, we show how the general expression of Theorem 1.1 can be obtained by recursive computation based on the previous results.

We denote by $\mathbf{Y}_{i-1}, \mathbf{Y}_i, \mathbf{Y}_{i+1}$ vectors of observations collected as in (4) at times t_{i-1}, t_i, t_{i+1} respectively, with associated multiplicities $\mathbf{n}_{i-1}, \mathbf{n}_i, \mathbf{n}_{i+1}$ for the distinct values $(y_1, \ldots, y_K)$ observed overall.

Theorem 3.4. *Under model (4), let $\mathbf{Y}_{i-1}, \mathbf{Y}_i, \mathbf{Y}_{i+1}$ be as above. Then there exist weights summing up to one, denoted $w_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}(\Delta_i, \Delta_{i+1})$, for $\mathbf{k}_{i-1} \leq \mathbf{n}_{i-1}, \mathbf{k}_{i+1} \leq \mathbf{n}_{i+1}$, such that*

$$\mathcal{L}(X_{t_i} \mid \mathbf{Y}_{i-1}, \mathbf{Y}_i, \mathbf{Y}_{i+1}) =$$
$$= \sum_{\mathbf{0} \leq \mathbf{k}_{i-1} \leq \mathbf{n}_{i-1}} \sum_{\mathbf{0} \leq \mathbf{k}_{i+1} \leq \mathbf{n}_{i+1}} w_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}(\Delta_i, \Delta_{i+1})\Pi_{\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}}, \tag{18}$$

where $\Delta_i = t_i - t_{i-1}$, $\Delta_{i+1} = t_{i+1} - t_i$, *and*

$$\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}} = \alpha + \sum_{j=1}^{K}(k_{i-1,j} + n_{i,j} + k_{i+1,j})\delta_{y_j}. \tag{19}$$

Remark 3.5. In the previous result, we have used notation $\mathbf{k}_{i-1}$ and $\mathbf{k}_{i+1}$ for the integrating variables, whose indices should help the intuition by indicating the time point they refer to. Note however that in principle these quantities are determined at time t_i, being the number of lineages in a time-reversed genealogy. Instead of using generic integrating variables $\mathbf{i}, \mathbf{j}$, we choose to adopt this notational convention here and later for the sake of readability. $\square$

Proof. Without loss of generality, let $i = 1$ and denote $X_1 = X_{t_1}$. Given a measurable partition $\mathcal{A} = (A_1, \ldots, A_m)$ of $\mathcal{Y}$, let $X_1(\mathcal{A}) := (X_1(A_1), \ldots, X_1(A_m))$ and denote by $\mathbf{Y}(\mathcal{A})$ the list of labels derived from binning $\mathbf{Y}$ into $\mathcal{A}$, i.e., whose i-th element is j if $Y_i \in A_j$. Further, let $\{\mathcal{B}_n, n \geq 1\} = \{(B_1^n, \ldots, B_n^n), n \geq 1\}$ be a sequence of increasingly finer partitions of $\mathcal{Y}$ such that $\mathcal{B}_n$ is finer than $\mathcal{A}$, for every n, and such that $\max_j \operatorname{diam}(B_j^n) \to 0$ as n diverges. Since $\mathcal{B}_n$ is increasingly finer, we have that $\left(\mathbb{E}\left[f(X_1(\mathcal{A}))|\mathbf{Y}_0(\mathcal{B}_n), \mathbf{Y}_1(\mathcal{B}_n), \mathbf{Y}_2(\mathcal{B}_n)\right]\right)_n$ is a martingale for every bounded and continuous function f (see Proposition V.2.7 in Cinlar (2011)). Thus, by the martingale convergence theorem we have that $X_1(\mathcal{A})|\mathbf{Y}_0(\mathcal{B}_n), \mathbf{Y}_1(\mathcal{B}_n), \mathbf{Y}_2(\mathcal{B}_n)$ converges weakly to $X_1(\mathcal{A})|\mathbf{Y}_0, \mathbf{Y}_1, \mathbf{Y}_2$ as $n \to \infty$. The left hand side of the previous expression can be characterized, by virtue of de Finetti's Theorem, in terms of the predictive distributions of $Y_1^{1:k}(\mathcal{A})|\mathbf{Y}_0(\mathcal{B}_n), \mathbf{Y}_1(\mathcal{B}_n), \mathbf{Y}_2(\mathcal{B}_n)$ for arbitrary k, where $Y_1^{1:k}(\mathcal{A})$ denotes k samples from $X_1(\mathcal{A})$. Without loss of generality, let now n be large enough that different observations lie in different sets of $\mathcal{B}_n$, and write, for brevity, $X_{1,n} := X_1(\mathcal{B}_n)$ and $\mathbf{Y}_{i,n} := \mathbf{Y}_i(\mathcal{B}_n)$. Let also $p(y^{1:k}|\mathbf{Y}_{0,n}, \mathbf{Y}_{1,n}, \mathbf{Y}_{2,n})$ be the density of the vector $Y_1^{1:k}(\mathcal{A})$ evaluated at $y^{1:k}$, conditional on the binned observation $\mathbf{Y}_{0,n}, \mathbf{Y}_{1,n}, \mathbf{Y}_{2,n}$. Then we have

$$
\begin{aligned}
p(y^{1:k}|\mathbf{Y}_{0,n}, \mathbf{Y}_{1,n}, \mathbf{Y}_{2,n}) &\propto p(y^{1:k}, \mathbf{Y}_{0,n}, \mathbf{Y}_{1,n}, \mathbf{Y}_{2,n}) = \mathbb{E}\left[p(y^{1:k}, \mathbf{Y}_{0,n}, \mathbf{Y}_{1,n}, \mathbf{Y}_{2,n}|X_{1,n})\right] \\
&= \mathbb{E}\left[p(y^{1:k}|X_{1,n})p(\mathbf{Y}_{0,n}|X_{1,n})p(\mathbf{Y}_{1,n}|X_{1,n})p(\mathbf{Y}_{2,n}|X_{1,n})\right]
\end{aligned} \tag{20}
$$

where in the last identity we have used the conditional independence of the observations given the signal state (cf. (4)). By Lemma 3.1 and the subsequent comment we get

$$p(\mathbf{Y}_{0,n}|X_{1,n}) \propto \sum_{\mathbf{0} \leq \mathbf{k}_{i-1} \leq \mathbf{n}_{i-1}} p_{\mathbf{n}_{i-1}, \mathbf{k}_{i-1}} h(X_{1,n}, \mathbf{k}_{i-1})$$

$$p(\mathbf{Y}_{2,n}|X_{1,n}) \propto \sum_{\mathbf{0} \leq \mathbf{k}_{i+1} \leq \mathbf{n}_{i+1}} p'_{\mathbf{n}_{i+1}, \mathbf{k}_{i+1}} h(X_{1,n}, \mathbf{k}_{i+1})$$

$$p(\mathbf{Y}_{1,n}|X_{1,n}) \propto h(X_{1,n}, \mathbf{n}_i)$$

where $p_{\mathbf{n}_{i-1},\mathbf{k}_{i-1}} := p_{\mathbf{n}_{i-1},\mathbf{k}_{i-1}}(\Delta_i)$ and $p'_{\mathbf{n}_{i+1},\mathbf{k}_{i+1}} := p_{\mathbf{n}_{i+1},\mathbf{k}_{i+1}}(\Delta_{i+1})$. By linearity and by (14), (20) is thus proportional to

$$\sum_{0 \leq \mathbf{k}_{i-1} \leq \mathbf{n}_{i-1}} \sum_{0 \leq \mathbf{k}_{i+1} \leq \mathbf{n}_{i+1}} p_{\mathbf{n},\mathbf{k}_{i-1}} p'_{\mathbf{m},\mathbf{k}_{i+1}} \frac{m^{(n)}(\mathbf{k}_{i-1} + \mathbf{n}_i + \mathbf{k}_{i+1})}{m^{(n)}(\mathbf{k}_{i-1})m^{(n)}(\mathbf{n}_i)m^{(n)}(\mathbf{k}_{i+1})}$$
$$\times \mathbb{E}\left[p(y^{1:k}|X_{1,n})h(X_{1,n}, \mathbf{k}_{i-1} + \mathbf{n}_i + \mathbf{k}_{i+1})\right],$$

where $m^{(n)}$ denotes the marginal distribution in (12) relative to the model induced by the partition $\mathcal{B}_n$. Moreover, using (13) and (8) it can be seen that

$$\mathbb{E}\left[p(\mathbf{n}'|\mathbf{x})h(\mathbf{x}, \mathbf{n})\right] = \int p(\mathbf{n}'|\mathbf{x})\frac{p(\mathbf{n}|\mathbf{x})}{m(\mathbf{n})}p(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{m(\mathbf{n}, \mathbf{n}')}{m(\mathbf{n})} = m_{\mathbf{n}}(\mathbf{n}'),$$

with $m_{\mathbf{n}}(\mathbf{n}') := p(\mathbf{n}'|\mathbf{n})$, hence we can write

$$\mathbb{E}\left[p(y^{1:k}|X_{1,n})h(X_{1,n}, \mathbf{k}_{i-1} + \mathbf{n}_i + \mathbf{k}_{i+1})\right] = m_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}(y^{1:k}).$$

Note that the above identity holds since $\mathcal{B}_n$ is finer than $\mathcal{A}$. Hence the left hand side of (20) equals

$$\sum_{0 \leq \mathbf{k}_{i-1} \leq \mathbf{n}_{i-1}} \sum_{0 \leq \mathbf{k}_{i+1} \leq \mathbf{n}_{i+1}} p_{\mathbf{n}_{i-1},\mathbf{k}_{i-1}} p'_{\mathbf{n}_{i+1},\mathbf{k}_{i+1}}$$
$$\times \frac{m^{(n)}(\mathbf{k}_{i-1} + \mathbf{n}_i + \mathbf{k}_{i+1})}{C_n m^{(n)}(\mathbf{k}_{i-1})m^{(n)}(\mathbf{n}_i)m^{(n)}(\mathbf{k}_{i+1})} m_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}(\mathbf{n}')$$

where C_n is a normalizing constant and $\mathbf{n}'$ is the vector of multiplicities associated to $y^{1:k}$. Since $m_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}(\mathbf{n}')$ is the distribution induced by the Pólya Urn scheme of the Dirichlet–multinomial model, it follows that the law of $Y_1^{1:k}(\mathcal{A})|\mathbf{Y}_{0,n}, \mathbf{Y}_{1,n}, \mathbf{Y}_{2,n}$ is exchangeable. Note in particular that this marginal distribution does not depend on the partition $\mathcal{B}_n$, since $\mathcal{A}$ is a coarser partition. Given the arbitrariness of k, we can appeal to de Finetti's Theorem to conclude that the law of $X_1(\mathcal{A})|\mathbf{Y}_{0,n}, \mathbf{Y}_{1,n}, \mathbf{Y}_{2,n}$ is given by

$$\sum_{0 \leq \mathbf{k}_{i-1} \leq \mathbf{n}_{i-1}} \sum_{0 \leq \mathbf{k}_{i+1} \leq \mathbf{n}_{i+1}} p_{\mathbf{n}_{i-1},\mathbf{k}_{i-1}} p'_{\mathbf{n}_{i+1},\mathbf{k}_{i+1}}$$
$$\times \frac{m^{(n)}(\mathbf{k}_{i-1} + \mathbf{n}_i + \mathbf{k}_{i+1})}{C_n m^{(n)}(\mathbf{k}_{i-1})m^{(n)}(\mathbf{n}_i)m^{(n)}(\mathbf{k}_{i+1})} \pi_{\alpha(\mathcal{A})+\mathbf{k}_{i-1}(\mathcal{A})+\mathbf{n}_i(\mathcal{A})+\mathbf{k}_{i+1}(\mathcal{A})} \tag{21}$$

where $\alpha(\mathcal{A}) = (\alpha(A_1), \ldots, \alpha(A_m))$ and $\mathbf{k}_{i-1}(\mathcal{A}), \mathbf{n}_i(\mathcal{A}), \mathbf{k}_{i+1}(\mathcal{A})$ denote the multiplicities projected onto $\mathcal{A}$. The limit as $n \to \infty$ can now be computed by virtue of the martingale convergence theorem. The proof is completed by observing that the limiting weights do not depend on the partition $\mathcal{B}_n$ and the previous display coincides with the projection onto the partition $\mathcal{A}$ of a finite mixture of laws of Dirichlet processes. $\square$

The previous result provides an explicit representation of the conditional law of a FV state given observations at adjacent times, but does not investigate in full detail the mixture weights, denoted generically in the Theorem statement, which we do next. To pursue this task, we need to compute

$$\lim_{n\to\infty} \frac{m^{(n)}(\mathbf{k}_{i-1} + \mathbf{n}_i + \mathbf{k}_{i+1})}{C_n m^{(n)}(\mathbf{k}_{i-1}) m^{(n)}(\mathbf{n}_i) m^{(n)}(\mathbf{k}_{i+1})}.$$

since the term $p_{\mathbf{n}_{i-1},\mathbf{k}_{i-1}} p'_{\mathbf{n}_{i+1},\mathbf{k}_{i+1}}$ in (21) does not depend on n. In the setting of Theorem 3.4, denote now by

$$D_{i-1} = \{j \in \{1,\dots,K\} : n_{i-1,j} > 0 \text{ and either } n_{i,j} > 0 \text{ or } n_{i+1,j} > 0\},$$
$$D_{i+1} = \{j \in \{1,\dots,K\} : n_{i+1,j} > 0 \text{ and either } n_{i,j} > 0 \text{ or } n_{i-1,j} > 0\},$$

the set of distinct values in $\mathbf{Y}_{i-1}$ shared with $\mathbf{Y}_i$ or $\mathbf{Y}_{i+1}$, and those in $\mathbf{Y}_{i+1}$ shared with $\mathbf{Y}_{i-1}$ or $\mathbf{Y}_i$, respectively. Then

$$\mathcal{D} = \left\{ (\mathbf{k},\mathbf{k}') \leq (\mathbf{n}_{i-1},\mathbf{n}_{i+1}) : k_j > 0, k'_{j'} > 0, \forall j \in D_{i-1} \text{ and } j' \in D_{i+1} \right\}$$

is the set of multiplicities $(\mathbf{k},\mathbf{k}')$ not greater than $(\mathbf{n}_{i-1},\mathbf{n}_{i+1})$ such that the frequency of distinct values shared between different collection times is strictly positive. For example, if t_i is the current time index, suppose we have $\mathbf{n}_{i-1} = (1,3,0)$, $\mathbf{n}_i = (0,0,1)$ and $\mathbf{n}_{i+1} = (0,2,1)$, whereby of the three types observed overall, at time t_{i-1} we observed multiplicities 1 and 3 for the first two, at time t_i an instance of a third type, and so on. Then

$$\mathcal{D} = \{(\mathbf{i},\mathbf{j}) : \mathbf{i} \leq (1,3,0), \mathbf{j} \leq (0,2,1), i_2 > 0, j_2 > 0, j_3 > 0\}$$

is given by vectors of multiplicities not greater than $(\mathbf{n}_{i-1}, \mathbf{n}_{i+1})$, with positive entries for type two, which is shared by times t_{i-1}, t_{i+1}, and for type three, limited to the second coordinate, since it is shared by time t_{i+1} and the current time. In other words, multiplicities not greater than those observed, with positive entries for types: (i) observed at both times different from the current, or (ii) observed at the current time and at least another time. Notice that $\mathcal{D} = \emptyset$ corresponds to the case in which no values are shared between the three collection times t_{i-1}, t_i and t_{i+1}, which holds, for example, when all the observations are distinct.

Before stating the result, note that when P_0 is supported by a countably infinite set, $\mathbf{m}(\mathbf{n})$ can be defined by extension of (12), where all but a finite number of terms simplify in the ratio. Let also $a^{(b)} = a(a+1)\dots(a+b-1)$ denote the Pochhammer symbol.

Proposition 3.6. *In the setting of Theorem 3.4, let $\tilde{p} = p_{\mathbf{n}_{i-1},\mathbf{k}_{i-1}}(\Delta_i) p_{\mathbf{n}_{i+1},\mathbf{k}_{i+1}}(\Delta_{i+1})$. Then*

A. if P_0 is discrete,

$$w^{\mathbf{n}_{i-1},\mathbf{n}_{i+1}}_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}(\Delta_i,\Delta_{i+1}) \propto \tilde{p} \frac{m(\mathbf{k}_{i-1} + \mathbf{n}_i + \mathbf{k}_{i+1})}{m(\mathbf{k}_{i-1})m(\mathbf{n}_i)m(\mathbf{k}_{i+1})}; \tag{22}$$

15

B. if P_0 is nonatomic and $\mathcal{D} = \emptyset$,

$$w^{\mathbf{n}_{i-1}, \mathbf{n}_{i+1}}_{\mathbf{k}_{i-1}, \mathbf{n}_i, \mathbf{k}_{i+1}}(\Delta_i, \Delta_{i+1}) \propto \tilde{p} \, \frac{\theta^{(|\mathbf{k}_{i-1}|)} \theta^{(|\mathbf{k}_{i+1}|)}}{(\theta + |\mathbf{n}_i|)^{(|\mathbf{k}_{i-1}| + |\mathbf{k}_{i+1}|)}};$$

C. if P_0 is nonatomic and $\mathcal{D} \neq \emptyset$,

$$w^{\mathbf{n}_{i-1}, \mathbf{n}_{i+1}}_{\mathbf{k}_{i-1}, \mathbf{n}_i, \mathbf{k}_{i+1}}(\Delta_i, \Delta_{i+1}) \propto \tilde{p} \, \frac{\theta^{(|\mathbf{k}_{i-1}|)} \theta^{(|\mathbf{k}_{i+1}|)}}{(\theta + |\mathbf{n}_i|)^{(|\mathbf{k}_{i-1}| + |\mathbf{k}_{i+1}|)}} \prod_{j=1}^{K} \frac{(k_{i-1,j} + n_{i,j} + k_{i+1,j} - 1)!}{(k_{i-1,j} - 1)! \, (n_{i,j} - 1)! \, (k_{i+1,j} - 1)!}$$

if $(\mathbf{k}_{i-1}, \mathbf{k}_{i+1}) \in \mathcal{D}$, and zero otherwise.

Proof. Statement A follows from the fact that ultimately the limit partition sets with positive multiplicities will be those coinciding with the support points of P_0.

Assume now, without loss of generality, that the partition $\mathcal{B}_n$ is such that the first observation lies in B_1^n, the second in B_2^n and so on. The density of a vector $\mathbf{k}$ of multiplicities is in this case determined by the Blackwell–MacQueen Pólya urn scheme (Blackwell and MacQueen, 1973) to be

$$m^{(n)}(\mathbf{k}) = \frac{\prod_{j=1}^{K} \prod_{h=0}^{k_j-1} \left(\theta P_0(B_j^n) + h \right)}{\theta^{(|\mathbf{k}|)}},$$

with the convention that $\prod_{h=0}^{-1} = 1$. Denoting $m_{\mathbf{n}}(\mathbf{n}') := p(\mathbf{n}'|\mathbf{n})$, as in the proof of Theorem 3.4, it follows that

$$\frac{m^{(n)}(\mathbf{k}_{i-1} + \mathbf{n}_i + \mathbf{k}_{i+1})}{m^{(n)}(\mathbf{k}_{i-1}) m^{(n)}(\mathbf{n}_i) m^{(n)}(\mathbf{k}_{i+1})} = \frac{m^{(n)}_{\mathbf{n}_i}(\mathbf{k}_{i-1} + \mathbf{k}_{i+1})}{m^{(n)}(\mathbf{k}_{i-1}) m^{(n)}(\mathbf{k}_{i+1})}$$

$$= \frac{\theta^{(|\mathbf{k}_{i-1}|)} \theta^{(|\mathbf{k}_{i+1}|)}}{(\theta + |\mathbf{n}_i|)^{(|\mathbf{k}_{i-1}| + |\mathbf{k}_{i+1}|)}} \prod_{j=1}^{K} \frac{\prod_{h=0}^{k_{i-1,j} + k_{i+1,j} - 1} \left(\theta P_0(B_j^n) + n_{i,j} + h \right)}{\prod_{h=0}^{k_{i-1,j} - 1} \left(\theta P_0(B_j^n) + h \right) \prod_{h=0}^{k_{i+1,j} - 1} \left(\theta P_0(B_j^n) + h \right)}.$$

If $\mathcal{D} = \emptyset$, since no values are shared across times, we have that, for every j, at most one between $k_{i-1,j}$ and $k_{i+1,j}$ is non zero, and in such case we have $n_{i,j} = 0$. Then, if $k_{i-1,j} > 0$ we have $k_{i+1,j} = 0$ and $n_{i,j} = 0$, so

$$\frac{\prod_{h=0}^{k_{i-1,j} + k_{i+1,j} - 1} \left(\theta P_0(B_j^n) + n_{i,j} + h \right)}{\prod_{h=0}^{k_{i-1,j} - 1} \left(\theta P_0(B_j^n) + h \right) \prod_{h=0}^{k_{i+1,j} - 1} \left(\theta P_0(B_j^n) + h \right)} = 1,$$

and the same happens when $k_{i+1,j} > 0$, $k_{i-1,j} = 0$ and $n_{i,j} = 0$. This leads to statement B.

16

If $\mathcal{D} \neq \emptyset$, since some values are shared across times, there exists a j such that one of the following is true: (i) $n_{i,j} > 0$ and $k_{i-1,j} > 0$; (ii) $n_{i,j} > 0$ and $k_{i+1,j} > 0$; (iii) $k_{i-1,j} > 0$ and $k_{i+1,j} > 0$. In case (i)

$$\frac{\prod_{h=0}^{k_{i-1,j}+k_{i+1,j}-1} \left(\theta P_0(B_j^n) + n_{i,j} + h\right)}{\prod_{h=0}^{k_{i-1,j}-1} \left(\theta P_0(B_j^n) + h\right) \prod_{h=0}^{k_{i+1,j}-1} \left(\theta P_0(B_j^n) + h\right)} \rightarrow \infty$$

as $n \rightarrow \infty$, since $P_0(B_j^n) \rightarrow 0$ and the denominator vanishes. Case (ii) is obtained similarly. In case (iii), rewrite the weights as

$$
\begin{aligned}
&\frac{\prod_{h=0}^{k_{i-1,j}+k_{i+1,j}-1} \left(\theta P_0(B_j^n) + n_{i,j} + h\right)}{\prod_{h=0}^{k_{i-1,j}-1} \left(\theta P_0(B_j^n) + h\right) \prod_{h=0}^{k_{i+1,j}-1} \left(\theta P_0(B_j^n) + h\right)} \\
&= \frac{\prod_{h=0}^{k_{i-1,j}-1} \left(\theta P_0(B_j^n) + n_{i,j} + h\right)}{\prod_{h=0}^{k_{i-1,j}-1} \left(\theta P_0(B_j^n) + h\right)} \frac{\prod_{h=k_{i-1,j}}^{k_{i-1,j}+k_{i+1,j}-1} \left(\theta P_0(B_j^n) + n_{i,j} + h\right)}{\prod_{h=0}^{k_{i+1,j}-1} \left(\theta P_0(B_j^n) + h\right)}.
\end{aligned}
\tag{23}
$$

Here the left factor is such that

$$\frac{\prod_{h=0}^{k_{i-1,j}-1} \left(\theta P_0(B_j^n) + n_{i,j} + h\right)}{\prod_{h=0}^{k_{i-1,j}-1} \left(\theta P_0(B_j^n) + h\right)} \geq 1$$

and the right factor can be written

$$\frac{\prod_{h=k_{i-1,j}}^{k_{i-1,j}+k_{i+1,j}-1} \left(\theta P_0(B_j^n) + n_{i,j} + h\right)}{\prod_{h=0}^{k_{i+1,j}-1} \left(\theta P_0(B_j^n) + h\right)} = \frac{\prod_{h=0}^{k_{i+1,j}-1} \left(\theta P_0(B_j^n) + k_{i-1,j} + n_{i,j} + h\right)}{\prod_{h=0}^{k_{i+1,j}-1} \left(\theta P_0(B_j^n) + h\right)}.$$

Therefore, the left hand side of (23) is greater than or equal to

$$\frac{\prod_{h=0}^{k_{i+1,j}-1} \left(\theta P_0(B_j^n) + k_{i-1,j} + n_{i,j} + h\right)}{\prod_{h=0}^{k_{i+1,j}-1} \left(\theta P_0(B_j^n) + h\right)}$$

which diverges to infinity as $n \rightarrow \infty$ as well. Thus, nodes with shared observations have divergent unnormalized weights. Let $\mathcal{S} = D_{i-1} \cup D_{i+1}$ be the set of shared values and let

17

$(\mathbf{k}_{i-1}, \mathbf{k}_{i+1}) \in \mathcal{D}$. Then, we can write the associated weight as

$$\frac{\theta^{(|\mathbf{k}_{i-1}|)}\theta^{(|\mathbf{k}_{i+1}|)}}{(\theta + |\mathbf{n}_i|)^{(|\mathbf{k}_{i-1}|+|\mathbf{k}_{i+1}|)}}$$
$$\times \prod_{j=1}^{K} \frac{\prod_{h=0}^{k_{i-1,j}+n_{i,j}+k_{i+1,j}-1}\left(\theta P_0(B_j^n) + h\right)}{\prod_{h=0}^{k_{i-1,j}-1}\left(\theta P_0(B_j^n) + h\right)\prod_{h=0}^{n_{i,j}-1}\left(\theta P_0(B_j^n) + h\right)\prod_{h=0}^{k_{i+1,j}-1}\left(\theta P_0(B_j^n) + h\right)}$$
$$= \frac{\theta^{(|\mathbf{k}_{i-1}|)}\theta^{(|\mathbf{k}_{i+1}|)}}{(\theta + |\mathbf{n}_i|)^{(|\mathbf{k}_{i-1}|+|\mathbf{k}_{i+1}|)}}$$
$$\times \prod_{j \in \mathcal{S}} \frac{\prod_{h=1}^{k_{i-1,j}+n_{i,j}+k_{i+1,j}-1}\left(\theta P_0(B_j^n) + h\right)}{\prod_{h=1}^{k_{i-1,j}-1}\left(\theta P_0(B_j^n) + h\right)\prod_{h=1}^{n_{i,j}-1}\left(\theta P_0(B_j^n) + h\right)\prod_{h=1}^{k_{i+1,j}-1}\left(\theta P_0(B_j^n) + h\right)}$$
$$\times \frac{1}{\prod_{j \in \mathcal{S}} P_0(B_j^n)\prod_{h=0}^{n_{i,j}-1} P_0(B_j^n)}.$$

Here the third factor on the right hand side is common to each node and is cancelled upon normalizing, while the second factor converges to the product in statement C, proving the result. $\qquad\square$

We now have a full description of (18), which is a finite mixture of laws of Dirichlet random measures whose parameter measure $\alpha + \sum_{j=1}^{K}(k_{i-1,j} + n_{i,j} + k_{i+1,j})\delta_{y_j}$ contains, besides the unnormalised offspring measure α, the current observations $\mathbf{Y}_i$ and a subset of the observations $(\mathbf{Y}_{i-1}, \mathbf{Y}_{i+1})$ collected at adjacent times. The mixture weights are in turn determined by the following two elements. The first is given by the transition probabilities of the death process associated to the typed coalescent, which determines the probability that past and future data are atoms in the respective random measures as a function of the distance between time t_i and the adjacent times. As the lags Δ_i and Δ_{i+1} grow, the number of survived lineages is lower with higher probability and the random measures in the mixture will carry, on average, less information in terms of types observed at different times. The second element is the joint marginal likelihood of past, present and future data. For instance, when the offspring distribution is discrete, the ratio in (22) is higher when $m(\mathbf{k}_{i-1} + \mathbf{n}_i + \mathbf{k}_{i+1}) > m(\mathbf{k}_{i-1})m(\mathbf{n}_i)m(\mathbf{k}_{i+1})$, i.e. when sampling jointly $\mathbf{k}_{i-1}$, $\mathbf{n}_i$ and $\mathbf{k}_{i+1}$ has high probability relative to sampling them separately: this provides a smoothing effect by favouring the nodes $(\mathbf{k}_{i-1}, \mathbf{n}_i, \mathbf{k}_{i+1})$ with the same types collected at different times. Such mechanism comes to an extreme when the offspring distribution is nonatomic. In this case, the weights of the mixture components that do not carry atoms observed at multiple times, in the sense of the set $\mathcal{D}$, vanish in the limit.

The above results do not include the case of α having both a continuous and discrete component. The same tools used for proving Proposition 3.6 can in principle be used to deal

with this case as well, where we expect parts A, B and C of the statement to hold for the respective parts of the parameters measure. In essence, values drawn by the discrete part of α are subjected to the probability that lineages survive as controlled by the term $\tilde{p}$ (hence ultimately by the death process), whereas values drawn from the continuous part of α are also, in addition, subjected to whether they are shared across collection times. A full description of such results would require a cumbersome notation and would not add further valuable insight, hence, we will not pursue this task here.

We can now revisit the first statement of Theorem 1.1 in light of the above findings. Let $\mathbf{Y}_{t_{0:N}}$ be the entire dataset sampled in model (4), and let K be the number of distinct values in $\mathbf{Y}_{t_{0:N}}$. Denoting by $\overleftarrow{\mathbf{n}}_{i-1}$ the total multiplicities of the vector $\mathbf{Y}_{t_{0:i-1}}$, we know there exist weights $\{v_{1,\mathbf{m}}\}$ such that

$$\mathcal{L}(X_{t_i}|\mathbf{Y}_{t_{0:i-1}}) = \sum_{\mathbf{k}_{i-1} \leq \overleftarrow{\mathbf{n}}_{i-1}} v_{1,\mathbf{k}_{i-1}} \Pi_{\alpha + \sum_{j=1}^{K} k_{i-1,j}\delta_{y_j}}. \tag{24}$$

This can be obtained recursively starting from $X_{t_0} \sim \Pi_\alpha$, where the reversible measure Π_α acts as prior (or unconditional distribution) for the marginal state of the signal. Upon observing $\mathbf{Y}_{t_0}$, the update yields $\mathcal{L}(X_{t_0}|\mathbf{Y}_{t_0}) = \Pi_{\alpha + \sum_{j=1}^{K_0} n_{0,j}\delta_{y_j}}$, where $\mathbf{n}_0$ is the vector of multiplicities of the K_0 distinct values in $\mathbf{Y}_{t_0}$. Propagating forward the previous through $\mathcal{F}_t$, one obtains

$$\mathcal{L}(X_{t_1}|\mathbf{Y}_{t_0}) = \sum_{\mathbf{k}_0 \leq \mathbf{n}_0} p_{\mathbf{n}_0,\mathbf{k}_0}(t_1 - t_0) \Pi_{\alpha + \sum_{j=1}^{K_0} k_{0,j}\delta_{y_j}},$$

which can then be updated once data in t_1 become available by observing that (8) satisfies

$$\mathcal{U}_{\mathbf{n}}\left(\sum_{i=1}^{H} w_i \xi_i\right) = \sum_{i=1}^{H} \frac{w_i p_{\xi_i}(\mathbf{n})}{\sum_{h=1}^{H} w_h \xi_h(\mathbf{n})} \mathcal{U}_{\mathbf{n}}(\xi_i).$$

Proceeding in this way, alternating updates and forward propagations, leads to (24); see Papaspiliopoulos et al. (2016), Section 3.1, for further details. Denoting now by $\overrightarrow{\mathbf{n}}_{i+1}$ the total multiplicities of the vector $\mathbf{Y}_{t_{i+1:N}}$, by virtue of Lemma 3.2 and of the linearity of (10), there exist weights $\{v_{2,\mathbf{n}}\}$ such that

$$\mathcal{L}(X_{t_i}|\mathbf{Y}_{t_{i+1:N}}) = \sum_{\mathbf{k}_{i+1} \leq \overrightarrow{\mathbf{n}}_{i+1}} v_{2,\mathbf{k}_{i+1}} \Pi_{\alpha + \sum_{j=1}^{K} k_{i+1,j}\delta_{y_j}}. \tag{25}$$

This can also be obtained by working backwards from $X_{t_N} \sim \Pi_\alpha$, then updating given the multiplicities $\mathbf{n}_N$ of the K_N distinct values in $\mathbf{Y}_{t_N}$, which yields $\mathcal{L}(X_{t_N}|\mathbf{Y}_{t_N}) = \Pi_{\alpha + \sum_{j=1}^{K_N} n_{N,j}\delta_{y_j}}$, then, using Lemma 3.2, propagating backwards to get

$$\mathcal{L}(X_{t_{N-1}}|\mathbf{Y}_{t_N}) = \sum_{\mathbf{k}_N \leq \mathbf{n}_N} p_{\mathbf{n}_N,\mathbf{k}_N}(t_N - t_{N-1}) \Pi_{\alpha + \sum_{j=1}^{K_N} k_{N,j}\delta_{y_j}},$$

and so on. The following proposition connects the two above distributions to yield the general representation in Theorem 1.1.

Proposition 3.7. *Assume* (24) *and* (25) *hold, and let* $\mathbf{n}_i$ *be the vector of multiplicities of* $\mathbf{Y}_{t_i}$ *relative to the* K *distinct values in the whole dataset* $\mathbf{Y}_{t_{0:N}}$. *Then*

$$\mathcal{L}(X_{t_i}|\mathbf{Y}_{t_{0:N}}) = \sum_{\mathbf{k}_{i-1} \leq \overleftarrow{\mathbf{n}}_{i-1}} \sum_{\mathbf{k}_{i+1} \leq \overrightarrow{\mathbf{n}}_{i+1}} p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}} \Pi_{\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}}, \tag{26}$$

where $\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}$ *is as in* (19) *and the weights*

$$p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}} = \sum_{\mathbf{h} \leq \overleftarrow{\mathbf{n}}_{i-1}: \mathbf{h} \geq \mathbf{k}_{i-1}} \sum_{\mathbf{l} \leq \overrightarrow{\mathbf{n}}_{i+1}: \mathbf{l} \geq \mathbf{k}_{i+1}} v_{1,\mathbf{h}} v_{2,\mathbf{l}} w^{\mathbf{h},\mathbf{l}}_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}, \quad \mathbf{k}_{i-1} \leq \overleftarrow{\mathbf{n}}_{i-1}, \mathbf{k}_{i+1} \leq \overrightarrow{\mathbf{n}}_{i+1},$$

with $w^{\mathbf{h},\mathbf{l}}_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}$ *as in Proposition* 3.6, *sum up to one.*

Proof. The first statement follows as in the proof of Theorem 3.4 by noting that conditioning on $\mathbf{Y}_{i-1}, \mathbf{Y}_{i+1}$ in (18) is qualitatively analogous to conditioning to $\mathbf{Y}_{t_{0:i-1}}, \mathbf{Y}_{t_{i+1,N}}$ in (24)-(25), since the main argument (20) is based on the factorization of the likelihoods of the data collected prior, concurrently and after the signal state. The second statement follows by the linearity of the expected value in (20) and by readjusting the weights. $\square$

The proof of Proposition 3.7 clarifies that the smoothing mixture is computed in two steps: first $\mathcal{L}(X_{t_{i-1}}|\mathbf{Y}_{t_{0:i-1}})$ and $\mathcal{L}(X_{t_{i+1}}|\mathbf{Y}_{t_{i+1:N}})$ are computed through backward and forward filtering respectively, then the smoothing operator is applied, as in Theorem 3.4. The first step leads to two mixtures whose number of components is $\prod_{k=1}^{K}(1 + \sum_{j=0}^{i-1} n_{t_j,k})$ and $\prod_{k=1}^{K}(1 + \sum_{j=i+1}^{N} n_{t_j,k})$ respectively, as shown in Section 4 of Kon Kam King et al. (2020). Recall that here K is the number of distinct values observed in the entire dataset, which is considered as given. Since each distinct element of the smoothing distribution is now given by a distinct choice of $\mathbf{k}_{i-1}$ and $\mathbf{k}_{i+1}$, the total number of components in the smoothing distribution is therefore

$$\prod_{k=1}^{K} \left(1 + \sum_{j=0}^{i-1} n_{t_j,k}\right) \left(1 + \sum_{j=i+1}^{N} n_{t_j,k}\right). \tag{27}$$

As expected, smoothing comes at a greater nominal computational cost than filtering, since, roughly speaking, it combines information from both past and future. However, the actual cost of smoothing is expected to be much lower than the nominal, due to two factors. The first, specific to the current modelling assumptions, is that in the scenario of Statement C in Proposition 3.5, with a continuous baseline distribution, the number of components

is automatically pruned by the smoothing operator, which discards values that are not shared across times. Hence (27) represents a crude upper bound. The second factor is that some mixture component weights are typically negligible. This aspect, which had already been noted in Chaleyat-Maurel and Genon-Catalot (2006) and was investigated in detail for Wright–Fisher and Cox–Ingersoll–Ross models in Kon Kam King et al. (2020), suggests various possible pruning strategies that allow to approximate the smoothing distribution, lowering the actual computational cost by some order of magnitudes while keeping a high precision in the approximation.

3.3 Predictive distributions

As a corollary to Proposition 3.7, we can derive the predictive distribution of further samples collected at time t_i, given the original data set $\mathbf{Y}_{t_{0:N}}$. The following statement extends the results in Ascolani et al. (2021).

Corollary 3.8. *In the setting of Proposition 3.7, let* (26) *be the conditional law of X_{t_i} given* $\mathbf{Y}_{t_{0:N}}$. *Then the law of the $(k+1)$th further sample Y^{k+1} from X_{t_i} is*

$$\mathbb{P}(\, Y^{k+1} \in A | \mathbf{Y}_{t_{0:N}}, Y^{1:k}) =$$
$$= \sum_{\mathbf{k}_{i-1} \leq \overleftarrow{\mathbf{n}}_{i-1}} \sum_{\mathbf{k}_{i+1} \leq \overrightarrow{\mathbf{n}}_{i+1}} p_{\mathbf{k}_{i-1}, \mathbf{n}_i, \mathbf{k}_{i+1}} \frac{\alpha_{\mathbf{k}_{i-1} + \mathbf{n}_i + \mathbf{k}_{i+1}}(A) + \sum_{j=1}^{k} \delta_{Y^j}(A)}{\theta + |\mathbf{k}_{i-1}| + |\mathbf{n}_i| + |\mathbf{k}_{i+1}| + k} \tag{28}$$

for every Borel set A of $\mathcal{Y}$, with $\alpha_{\mathbf{k}_{i-1} + \mathbf{n}_i + \mathbf{k}_{i+1}}$ is as in (19) *and $p_{\mathbf{k}_{i-1}, \mathbf{n}_i, \mathbf{k}_{i+1}}$ as in Proposition 3.7.*

Proof. The statement can be easily proved by noting that

$$\mathbb{P}(Y_i \in A | \mathbf{Y}_{t_{0:N}}) = \mathbb{E}\left[\mathbb{P}(Y_i \in A | X_i, \mathbf{Y}_{t_{0:N}}) | \mathbf{Y}_{t_{0:N}}\right] = \mathbb{E}\left[X_i(A) | \mathbf{Y}_{t_{0:N}}\right]$$

and using (26), after recalling that if $W \sim \Pi_\alpha$ with Π_α as in (2) then $\mathbb{E}(W) = \alpha/\alpha(\mathcal{Y})$. $\qquad \square$

Here for brevity we have used the notation Y^{k+1} for the additional $(k+1)$-st sample instead of the correct notation $Y^{|\mathbf{n}_i|+k+1}$, given the original dataset already contains $|\mathbf{n}_i|$ observations sampled at t_i. Recall that the predictive distribution for observations sampled from a Dirichlet random measure, a realization of (2), is described by the Blackwell–MacQueen Pólya urn scheme (Blackwell and MacQueen, 1973), whereby for $\alpha = \theta P_0$,

$$Y_1 \sim P_0, \qquad Y_{k+1} | Y_1, \ldots, Y_k \sim \frac{\alpha + \sum_{j=1}^{k} \delta_{Y_j}}{\theta + k}.$$

It is then clear that (28) is a finite mixture of generalized Pólya urn schemes, whose sampling mechanism can be described as follows. For each $k \geq 1$, given we already observed the further sample $Y^{1:k}$,

- choose a pair $(\mathbf{k}_{i-1}, \mathbf{k}_{i+1})$ with probability $p_{\mathbf{k}_{i-1}, \mathbf{n}_i, \mathbf{k}_{i+1}}$
- draw a categorical random variable $J \in \{1, 2, 3\}$ with probabilities proportional to $\theta, |\mathbf{k}_{i-1}| + |\mathbf{n}_i| + |\mathbf{k}_{i+1}|$ and k respectively
- given $Y^{1:k}$, draw

$$
Y^{k+1} \sim \begin{cases} P_0, & \text{if } J = 1, \\ \frac{\sum_{j=1}^{K} (k_{i-1,j} + n_{i,j} + k_{i+1,j}) \delta_{y_j}}{|\mathbf{k}_{i-1}| + |\mathbf{n}_i| + |\mathbf{k}_{i+1}|}, & \text{if } J = 2, \\ \frac{1}{k} \sum_{j=1}^{k} \delta_{Y^j}, & \text{if } J = 3, \end{cases}
$$

where $(y_1, \ldots, y_K)$ in the second expression are the distinct values in $\mathbf{Y}_{t_{0:N}}$.

We conclude the section with the observation that (26) is the de Finetti measure of the sequence $\{Y^{|\mathbf{n}_i|+k}, k \geq 1\}$, i.e. as $k \to \infty$

$$
\mathbb{P}(Y^{|\mathbf{n}_i|+k+1} \in \cdot \, | \mathbf{Y}_{t_{0:N}}, Y^{|\mathbf{n}_i|+1:|\mathbf{n}_i|+k}) \Rightarrow P^* \qquad \text{a.s.}
$$

where $\Rightarrow$ denotes weak convergence and P^* is the law in (26). This can be proved along the same lines of Proposition 5 in Ascolani et al. (2021), where we also refer the reader for an application of a special case of (28).

4 Dawson–Watanabe signals

4.1 Preliminary results

Assume now the signal is a DW process and the observation model (5) holds. A projection of a gamma random measure Γ_α^β onto a measurable partition $A_1, \ldots, A_K$ of $\mathcal{Y}$ yields independent gamma random variables $\Gamma_\alpha^\beta(A_j) \sim \text{Ga}(\alpha(A_j), \beta)$. Similarly, a projection of a DW process with transition (3) onto the same partition yields independent subcritical continuous-state branching processes (CSBPs), denoted as $\text{CSBP}(\alpha(A_j), \beta)$, with transition function

$$
p_t(z_j, \mathrm{d}z_j') = \sum_{m=0}^{\infty} \text{Po}(m | z_j S_t) \text{Ga}(\mathrm{d}z_j' | \alpha(A_j) + m, \beta + S_t), \qquad S_t = \frac{\beta}{e^{\beta t / 2} - 1};
$$

cf. Papaspiliopoulos et al. (2016). Therefore, the projection of model (5) yields

$$
\mathbf{Y}_{t_n}^i | \mathbf{Z}_t = \mathbf{z} \stackrel{\text{iid}}{\sim} \text{PP}(\mathbf{z}), \quad \mathbf{Z} \sim \text{CSBP}_K(\boldsymbol{\alpha}, \beta), \tag{29}
$$

where $\mathrm{CSBP}_K(\boldsymbol{\alpha}, \beta)$ denotes a vector of K independent CSBPs as above, $\boldsymbol{\alpha} = (\alpha(A_1), \dots, \alpha(A_K))$ and $\mathrm{PP}(\mathbf{z})$ a Poisson point process with intensity $\mathbf{z} = (z(A_1), \dots, z(A_K))$. Cf. (6). In other words, conditionally on $\mathbf{Z}_t = \mathbf{z}$, the number m of observations is sampled from $\mathrm{Po}(|\mathbf{z}|)$ and a vector $\mathbf{Y}_t^1 = (Y_{t,1}^1, \dots, Y_{t,m}^1)$ is collected independently with multinomial distribution with parameters $\mathbf{z}/|\mathbf{z}|$. In this setting we will refer to a single *draw* as the collection of such m-sized vector $\mathbf{Y}_t^1$, with m random. Consider now a number $c \in \mathbb{N}$ of such draws, later referred to as the *cardinality* of the vector, denoted $\mathbf{Y}_t^{1:c} = (\mathbf{Y}_t^1, \dots, \mathbf{Y}_t^c)$ with multiplicities $(\mathbf{n}_{t,1}, \dots, \mathbf{n}_{t,c})$ associated to the distinct values $(y_1, \dots, y_K)$ observed overall. We denote by $\mathbf{N}_t^{1:c} = \sum_{j=1}^c \mathbf{n}_{t,j}$ the vector that collects the overall multiplicities observed for the distinct values. The following Lemma identifies the marginal distribution of the observations.

Lemma 4.1. *Consider a sample* $\mathbf{Y}$ *with multiplicity* $\mathbf{N}$ *and cardinality c from* (29). *Then the marginal likelihood is*

$$r_c(\mathbf{Y}) = \left(\frac{\beta}{\beta + c}\right)^\theta \left(\frac{1}{\beta + c}\right)^{|\mathbf{N}|} \frac{\Gamma(\theta + |\mathbf{N}|)}{\Gamma(\theta) \prod_{j=1}^c \prod_{k=1}^K \Gamma(n_{j,k} + 1)} m(\mathbf{N}) \tag{30}$$

with $m(\mathbf{N})$ *as in* (12) *and* $\theta = \alpha(\mathcal{Y})$.

Proof. Recall that $\mathbf{z}^{\mathbf{m}-\mathbf{1}} = z_1^{m_1-1} \cdots z_K^{m_K-1}$ for $\mathbf{m} = (m_1, \dots, m_K)$. Notice that

$$p\left(\mathbf{Y}_t^j \mid \mathbf{z}\right) = \mathrm{Po}(|\mathbf{n}_j|\, ;\, |\mathbf{z}|) \mathrm{Mult}(\mathbf{n}_j\, ;\, |\mathbf{n}_j|, \mathbf{z}/|\mathbf{z}|) = \frac{e^{-|\mathbf{z}|}|\mathbf{z}|^{|\mathbf{n}_j|}}{|\mathbf{n}_j|!} \binom{|\mathbf{n}_j|}{\mathbf{n}_j} \left(\frac{\mathbf{z}}{|\mathbf{z}|}\right)^{\mathbf{n}_j} = \frac{e^{-|\mathbf{z}|}}{\prod_{k=1}^K n_{j,k}!} \mathbf{z}^{\mathbf{n}_j}$$

Then, writing $\mathbf{N} = \mathbf{N}_t^{1:c}$, we have

$$\begin{aligned}
r_c(\mathbf{Y}) &= \int p(\mathbf{Y}|\mathbf{z}) p(\mathbf{z}) \mathrm{d}\mathbf{z} = \int \frac{e^{-c|\mathbf{z}|}}{\prod_{j=1}^c \prod_{k=1}^K n_{j,k}!} \mathbf{z}^{\mathbf{N}} \frac{\beta^\theta}{\prod_{j=1}^K \Gamma(\alpha_j)} \mathbf{z}^{\boldsymbol{\alpha}-\mathbf{1}} e^{-\beta|\mathbf{z}|} \, \mathrm{d}\mathbf{z} \\
&= \frac{\beta^\theta}{\prod_{j=1}^c \prod_{k=1}^K n_{j,k}! \prod_{j=1}^K \Gamma(\alpha_j)} \int \mathbf{z}^{\boldsymbol{\alpha}+\mathbf{N}-\mathbf{1}} e^{-(\beta+c)\mathbf{z}} \, \mathrm{d}\mathbf{z} \\
&= \left(\frac{\beta}{\beta+c}\right)^\theta \left(\frac{1}{\beta+c}\right)^{|\mathbf{N}|} \frac{1}{\prod_{j=1}^c \prod_{k=1}^K n_{j,k}!} \prod_{j=1}^K \frac{\Gamma(\alpha_j + |\mathbf{N}_j|)}{\Gamma(\alpha_j)}
\end{aligned}$$

whereby using now (12) gives the result. $\qquad\square$

Note that the special case $r_1(\mathbf{Y})$ is given by the product of the marginal distributions associated to Gamma–Poisson and Dirichlet–Multinomial models, the former being a negative

binomial distribution with success probability $1/(\beta + 1)$ and number of failures θ, evaluated at $|\mathbf{N}|$. Thus, denoting

$$\gamma_a(n) = \left(\frac{\beta}{\beta + a}\right)^\theta \left(\frac{1}{\beta + a}\right)^n \frac{\Gamma(\theta + n)}{\Gamma(\theta)}, \quad n \in \mathbb{N}, a > 0, \tag{31}$$

one can rewrite (30) as

$$r_c(\mathbf{Y}) = \frac{\gamma_c(|\mathbf{N}|)}{\prod_{j=1}^{c} \prod_{k=1}^{K} \Gamma(n_{j,k} + 1)} m(\mathbf{N}) \tag{32}$$

with $m(\mathbf{N})$ as in (12). In the following, with an abuse of notation, we will use $\mathbf{N}$ interchangeably with $\mathbf{Y}$, that is $r_c(\mathbf{N})$ and $p(\mathbf{N} \mid \mathbf{z})$ in place of $r_c(\mathbf{Y})$ and $p(\mathbf{Y} \mid \mathbf{z})$: indeed, even if $\mathbf{N}$ does not store the counts $|\mathbf{n}_j|$ associated to each draw, the main quantities to be defined afterwards will depend on $\mathbf{Y}$ only through $\mathbf{N}$ itself.

In this setting, the action of the update operator (8) is a straightforward extension of the update of a Gamma prior given Poisson observations, yielding here $\mathcal{U}_{\mathbf{N}}(\mathbf{Ga}(\boldsymbol{\alpha}, \beta)) = \mathbf{Ga}(\boldsymbol{\alpha} + \mathbf{N}, \beta + c)$. Proceeding now similarly to Section 3.1, define

$$h(\mathbf{z}, \mathbf{N}, c) = \frac{p(\mathbf{N}|\mathbf{z})}{r_c(\mathbf{N})} = e^{-c|\mathbf{z}|} \left(\frac{\beta + c}{\beta}\right)^\theta (\beta + c)^{|\mathbf{N}|} \prod_{j=1}^{K} \frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j + \mathbf{N}_j)} \mathbf{z}^{\mathbf{N}}, \tag{33}$$

whereby we note that

$$h(\mathbf{z}, \mathbf{N}, c) p(\mathbf{z}) = p(\mathbf{z}|\mathbf{N}, c) \tag{34}$$

and that the product of functions in (33) satisfies the identity

$$h(\mathbf{z}, \mathbf{N}, c) h(\mathbf{z}, \mathbf{M}, d) = \frac{r_{c+d}(\mathbf{N} + \mathbf{M})}{r_c(\mathbf{N}) r_d(\mathbf{M})} h(\mathbf{z}, \mathbf{N} + \mathbf{M}, c + d). \tag{35}$$

Theorem 4.1 in Papaspiliopoulos et al. (2016) showed the following duality relation for h functions as in (33),

$$\mathbb{E}\left[h(\mathbf{Z}_t, \mathbf{m}, c)|\mathbf{Z}_0 = \mathbf{z}\right] = \mathbb{E}\left[h(\mathbf{z}, \mathbf{M}_t, C_t)|\mathbf{M}_0 = \mathbf{m}, C_0 = c\right], \tag{36}$$

where $\mathbf{Z}_t$ is as in (29), $\mathbf{M}_t$ is a death process on $\mathbb{Z}_+^K$ that jumps from $\mathbf{m}$ to $\mathbf{m} - \mathbf{e}_i$ at rate $2\mathbf{m}_i(\beta + C_t)$, and C_t is a deterministic function that decreases to zero as

$$C_t = \frac{\beta c}{(\beta + c)e^{\beta t/2} - c}, \qquad C_0 = c. \tag{37}$$

We will often write C_t^c when it is useful to emphasize the starting point of C_t at time 0. The process C_t can be interpreted as a continuous-path time reversal of the deterministic process

that governs the sample size. See Papaspiliopoulos et al. (2016), Section 1.3 for a discussion. We will denote by $p_{\mathbf{n},\mathbf{m}}^c(t)$ the transition probabilities of $\mathbf{M}_t$, which depend on (37) through c (cf. Papaspiliopoulos et al. (2016), Section 4.3). One implication of the duality relation in (36), given in the following lemma, makes the parallel with Lemma 3.1 in the WF case, and provides the predictive law of observations given a past state of the signal.

Lemma 4.2. *Let* $\mathbf{Y}_{i+1}$ *be the vector of observations collected at time* t_{i+1} *from model* (29), *with multiplicities* $\mathbf{N}_{i+1}$ *and cardinality* c_{i+1}. *Then*

$$p(\mathbf{Y}_{i+1}|\mathbf{z}_{t_i}) = r_{c_{i+1}}(\mathbf{N}_{i+1}) \sum_{\mathbf{n} \leq \mathbf{N}_{i+1}} p_{\mathbf{N}_{i+1},\mathbf{n}}^{c_{i+1}}(t) h(\mathbf{z}_{t_i},\mathbf{n},C_t).$$

Proof. Using (33) we have

$$p(\mathbf{Y}_{i+1}|\mathbf{z}_{t_i}) = \int p(\mathbf{y}_{i+1}|\mathbf{z}_{t_{i+1}}) p_{t_{i+1}-t_i}(\mathbf{z}_{t_{i+1}}|\mathbf{z}_{t_i}) d\mathbf{z}_{t_{i+1}}$$

$$= r_{c_{i+1}}(\mathbf{N}_{i+1}) \int h(\mathbf{z}_{t_{i+1}},\mathbf{N}_{i+1},c_{i+1}) p_{t_{i+1}-t_i}(\mathbf{z}_{t_{i+1}}|\mathbf{z}_{t_i}) d\mathbf{z}_{t_{i+1}}$$

$$= r_{c_{i+1}}(\mathbf{N}_{i+1}) \mathbb{E}[h(\mathbf{Z}_{t_{i+1}},\mathbf{N}_{i+1},c_{i+1})|\mathbf{Z}_{t_i} = \mathbf{z}_{t_i}]$$

from which (36) gives the result. □

The next Proposition corresponds to Lemma 3.2 in the FV case and shows that the backward propagation of a conditional DW state yields the same as a forward propagation.

Proposition 4.3. *Assume* (5), *let* $y_1, \ldots, y_K$ *be distinct values and* $\mathbf{n} \in \mathbb{Z}_+^K$. *Then* $\mathcal{B}_t(\Gamma_{\alpha+\sum_{j=1}^K n_j \delta_{y_j}}^{\beta+c}) = \mathcal{F}_t(\Gamma_{\alpha+\sum_{j=1}^K n_j \delta_{y_j}}^{\beta+c})$ *with* $\mathcal{F}_t, \mathcal{B}_t$ *as in* (9)-(10), *and in particular*

$$\mathcal{B}_t\left(\Gamma_{\alpha+\sum_{j=1}^K n_j \delta_{y_j}}^{\beta+c}\right) = \sum_{\mathbf{0} \leq \mathbf{k} \leq \mathbf{n}} p_{\mathbf{n},\mathbf{k}}^c(t) \Gamma_{\alpha+\sum_{j=1}^K k_j \delta_{y_j}}^{\beta+C_t}.$$

Proof. Let $\mathbf{Y}_{i-1}, \mathbf{Y}_{i+1}$ be vectors sampled at times t_{i-1}, t_{i+1}, with multiplicities $\mathbf{N}_{i-1}, \mathbf{N}_{i+1}$ and cardinalities (c_{i-1}, c_{i+1}), respectively. Then we have

$$p(\mathbf{z}_{t_i}|\mathbf{Y}_{i-1}) = p(\mathbf{z}_{t_i}) \mathbb{E}\left[h(\mathbf{z}_{t_i},\mathbf{M}_{t_i},C_{t_{i+1}})|\mathbf{M}_{t_{i-1}} = \mathbf{N}_{i-1}, C_{t_{i-1}} = c_{i-1}\right]$$

$$= \sum_{\mathbf{0} \leq \mathbf{k} \leq \mathbf{N}_{i-1}} p_{\mathbf{N}_{i-1},\mathbf{k}}^{c_{i-1}}(t_i - t_{i-1}) \mathbf{Ga}(\boldsymbol{\alpha} + \mathbf{k}, \beta + C_{t_i-t_{i-1}})(\mathbf{z}_{t_i})$$

25

where the first equality comes from Proposition 2.2 of Papaspiliopoulos and Ruggiero (2014) and the second from Proposition 4.2 of Papaspiliopoulos et al. (2016) and the time homogeneity of C_t. Since by (8) $p(\mathbf{z}_{t_i}|\mathbf{Y}_{i+1}) = p(\mathbf{Y}_{i+1}|\mathbf{z}_{t_i})p(\mathbf{z}_{t_i})/r(\mathbf{N}_{i+1}, c_{i+1})$, we have

$$p(\mathbf{z}_{t_i}|\mathbf{Y}_{i+1}) = \frac{p(\mathbf{z}_{t_i})}{r(\mathbf{N}_{i+1}, c_{i+1})} r(\mathbf{N}_{i+1}, c_{i+1}) \sum_{0 \leq \mathbf{k} \leq \mathbf{N}_{i+1}} p_{\mathbf{N}_{i+1}, \mathbf{k}}^{c_{i+1}}(t_{i+1} - t_i) h(\mathbf{z}_i, \mathbf{k}, C_t)$$

$$= \sum_{0 \leq \mathbf{k} \leq \mathbf{N}_{i+1}} p_{\mathbf{N}_{i+1}, \mathbf{k}}^{c_{i+1}}(t_{i+1} - t_i) \mathbf{Ga}(\boldsymbol{\alpha} + \mathbf{k}, \beta + C_{t_{i+1}-t_i})(\mathbf{z}_{t_i})$$

where we have used Lemma 4.2 and then (34). The rest of the proof now proceeds along the same lines of Lemma 3.2, with Theorem 3.2 in Papaspiliopoulos et al. (2016) in place of Theorem 3.1. □

4.2 Smoothing distributions

We proceed similarly to Section 3.2, by first providing the smoothing distribution conditional on observations collected at adjacent times, then giving a full description of the mixture weights, and finally showing how the general expression of Theorem 1.1 can be computed. We denote by $\mathbf{Y}_{i-1}, \mathbf{Y}_i, \mathbf{Y}_{i+1}$ vectors of draws collected at times t_{i-1}, t_i, t_{i+1} respectively, each defined as in (5) and the preceding discussion, with cardinalities c_{i-1}, c_i, c_{i+1} and associated multiplicities $\mathbf{N}_{i-1}, \mathbf{N}_i, \mathbf{N}_{i+1}$ for the distinct values $(y_1, \ldots, y_K)$ observed overall.

Theorem 4.4. *Under model* (5), *let* $\mathbf{Y}_{i-1}, \mathbf{Y}_i, \mathbf{Y}_{i+1}$ *be as above. Then there exist weights* $\{w_{\mathbf{k}_{i-1}, \mathbf{N}_i, \mathbf{k}_{i+1}}(\Delta_i, \Delta_{i+1}), \mathbf{k}_{i-1} \leq \mathbf{N}_{i-1}, \mathbf{k}_{i+1} \leq \mathbf{N}_{i+1}\}$ *which sum up to one such that*

$$\mathcal{L}(Z_i | \mathbf{Y}_{i-1}, \mathbf{Y}_i, \mathbf{Y}_{i+1}) =$$
$$= \sum_{0 \leq \mathbf{k}_{i-1} \leq \mathbf{N}_{i-1}} \sum_{0 \leq \mathbf{k}_{i+1} \leq \mathbf{N}_{i+1}} v_{\mathbf{k}_{i-1}, \mathbf{N}_i, \mathbf{k}_{i+1}}^{\mathbf{N}_{i-1}, \mathbf{N}_{i+1}}(\Delta_i, \Delta_{i+1}) \Gamma_{\alpha_{\mathbf{k}_{i-1} + \mathbf{N}_{i-1} + \mathbf{k}_{i+1}}}^{\beta(\Delta_i, \Delta_{i+1})}.$$

where $\alpha_{\mathbf{k}_{i-1} + \mathbf{N}_{i-1} + \mathbf{k}_{i+1}}$ *is as in* (19), $\Delta_i = t_i - t_{i-1}, \Delta_{i+1} = t_{i+1} - t_i$,

$$\beta(\Delta_i, \Delta_{i+1}) = \beta + C_{\Delta_i}^{c_{i-1}} + c_i + C_{\Delta_{i+1}}^{c_{i+1}} \tag{38}$$

and C_t^c *is as in* (37).

Proof. Throughout the proof we use the same notational conventions used in the proof of Theorem 3.4 with the obvious modifications due to the present setting. In particular, considering the projected state $Z_1(\mathcal{A})|\mathbf{Y}_0(\mathcal{B}_n), \mathbf{Y}_1(\mathcal{B}_n), \mathbf{Y}_2(\mathcal{B}_n)$ conditional on binning the observations in a partition $\mathcal{B}_n$ finer than $\mathcal{A}$, we have by the martingale convergence theorem

that $Z_1(\mathcal{A})|\mathbf{Y}_0(\mathcal{B}_n), \mathbf{Y}_1(\mathcal{B}_n), \mathbf{Y}_2(\mathcal{B}_n)$ converges weakly to $Z_1(\mathcal{A})|\mathbf{Y}_0, \mathbf{Y}_1, \mathbf{Y}_2$. An argument similar to that used in the proof of Theorem 3.4, which makes use of Lemma 4.2 and (35), can now be used for characterising the conditional law of the DW states by means of the collection of predictive distributions. Specifically, this yields

$$
\begin{aligned}
\mathcal{L}(\, & \mathbf{Z}_1(\mathcal{A})|\mathbf{Y}_{0,n}, \mathbf{Y}_{1,n}, \mathbf{Y}_{2,n}) = \\
&= \sum_{\mathbf{0} \leq \mathbf{k}_{i-1} \leq \mathbf{N}_{i-1}} \sum_{\mathbf{0} \leq \mathbf{k}_{i+1} \leq \mathbf{N}_{i+1}} p_{\mathbf{N}_{i+1}, \mathbf{k}_{i-1}}^{c_{i-1}} p_{\mathbf{N}_{i+1}, \mathbf{k}_{i+1}}^{c_{i+1}} \\
&\quad \times \frac{r^{(n)}(\mathbf{k}_{i-1} + \mathbf{N}_i + \mathbf{k}_{i+1}, C_{\Delta_i}^{c_{i-1}} + c_i + C_{\Delta_{i+1}}^{c_{i+1}})}{C_n' \, r^{(n)}(\mathbf{k}_{i-1}, C_{\Delta_i}^{c_{i-1}}) r^{(n)}(\mathbf{N}_i, c_i) r^{(n)}(\mathbf{k}_{i+1}, C_{\Delta_{i+1}}^{c_{i+1}})} \\
&\quad \times \mathbf{Ga}(\alpha(\mathcal{A}) + \mathbf{k}_{i-1}(\mathcal{A}) + \mathbf{N}_i + \mathbf{k}_{i+1}(\mathcal{A}), \beta + C_{\Delta_i}^{c_{i+1}} + c_i + C_{\Delta_{i+1}}^{c_{i-1}})
\end{aligned}
$$

where C_n' is normalizing constant and $r^{(n)}$ is the marginal distribution in Lemma 4.1 (with the notational convention here that $r(\cdot, c) := r_c(\cdot)$), relative to the model induced by the partition $\mathcal{B}_n$. Similarly to the FV case, the proof is now completed by observing that the limiting weights do not depend on the partition $\mathcal{B}_n$ and the previous display coincides with the projection onto the partition $\mathcal{A}$ of a finite mixture of laws of Gamma processes. $\qquad \square$

We next provide a full description of the weights appearing in the statement of Theorem 4.4, with the convention $\gamma(\cdot, c) := \gamma_c(\cdot)$, for $\gamma_c(\cdot)$ as in (31).

Proposition 4.5. *In the setting of Theorem 4.4, we have*

$$
v_{\mathbf{k}_{i-1}, \mathbf{N}_i, \mathbf{k}_{i+1}}^{\mathbf{N}_{i-1}, \mathbf{N}_{i+1}}(\Delta_i, \Delta_{i+1}) = \frac{\gamma(|\mathbf{k}_{i-1} + \mathbf{N}_i + \mathbf{k}_{i+1}|, C_{\Delta_i} + c_i + C_{\Delta_{i+1}})}{\gamma(|\mathbf{k}_{i-1}|, C_{\Delta_i}) \gamma(|\mathbf{N}_i|, c_i) \gamma(|\mathbf{k}_{i+1}|, C_{\Delta_{i+1}})} w_{\mathbf{k}_{i-1}, \mathbf{N}_i, \mathbf{k}_{i+1}}^{\mathbf{N}_{i-1}, \mathbf{N}_{i+1}}
$$

where $w_{\mathbf{k}_{i-1}, \mathbf{N}_i, \mathbf{k}_{i+1}}^{\mathbf{N}_{i-1}, \mathbf{N}_{i+1}}$ as in Proposition 3.6.

Proof. The result follows from (32) with an easy calculation. $\qquad \square$

Similarly to what we have done for the FV case, we can now revisit the second statement of Theorem 1.1 in light of the above findings. We leave to the reader the adaptation of the argument preceding Theorem 3.7 to the present setting, which ultimately leads to

$$
\mathcal{L}(Z_{t_i}|\mathbf{Y}_{t_{0:i-1}}) = \sum_{\mathbf{k}_{i-1} \leq \overleftarrow{\mathbf{n}}_{i-1}} \rho_{1, \mathbf{k}_{i-1}} \Gamma_{\alpha + \sum_{j=1}^K k_{i-1,j} \delta_{y_j}}^{\beta + c_1}, \tag{39}
$$

and

$$
\mathcal{L}(Z_{t_i}|\mathbf{Y}_{t_{i+1:N}}) = \sum_{\mathbf{k}_{i+1} \leq \overrightarrow{\mathbf{n}}_{i+1}} \rho_{2, \mathbf{k}_{i+1}} \Gamma_{\alpha + \sum_{j=1}^K k_{i+1,j} \delta_{y_j}}^{\beta + c_2}, \tag{40}
$$

for some sets of weights $\{\rho_{1,\mathbf{k}_{i-1}}\}, \{\rho_{2,\mathbf{k}_{i+1}}\}$, where $\overleftarrow{\mathbf{n}}_{i-1}, \overrightarrow{\mathbf{n}}_{i+1}$ are here multiplicities associated to $\mathbf{Y}_{t_{0:i-1}}, \mathbf{Y}_{t_{i+1:N}}$ respectively, relative to the distinct types $(y_1, \dots, y_K)$ observed overall.

Proposition 4.6. *Assume* (39) *and* (40) *hold, and let* $\mathbf{n}_i$ *be the vector of multiplicities of* $\mathbf{Y}_{t_i}^{1:c_i}$ *relative to the K distinct values in the whole dataset* $\mathbf{Y}_{t_{0:N}}$. *Then*

$$\mathcal{L}(Z_{t_i}|\mathbf{Y}_{t_{0:N}}) = \sum_{\mathbf{k}_{i-1} \leq \overleftarrow{\mathbf{n}}_{i-1}} \sum_{\mathbf{k}_{i+1} \leq \overrightarrow{\mathbf{n}}_{i+1}} p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}} \Gamma_{\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}}^{\beta(\Delta_i,\Delta_{i+1})} \tag{41}$$

with $\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}$ *as in* (19) *and where the weights*

$$p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}} = \sum_{\mathbf{h}\leq\overleftarrow{\mathbf{n}}_{i-1}:\mathbf{h}\geq\mathbf{k}_{i-1}} \sum_{\mathbf{l}\leq\overrightarrow{\mathbf{n}}_{i+1}:\mathbf{l}\geq\mathbf{k}_{i+1}} \rho_{1,\mathbf{h}}\rho_{2,\mathbf{l}} v_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}^{\mathbf{h},\mathbf{l}}, \quad \mathbf{k}_{i-1} \leq \overleftarrow{\mathbf{n}}_{i-1}, \mathbf{k}_{i+1} \leq \overrightarrow{\mathbf{n}}_{i+1}, \tag{42}$$

with $v_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}^{\mathbf{h},\mathbf{l}}$ *as in Proposition* (4.5), *sum up to one.*

Proof. The proof is analogous to that of Proposition 3.7. $\qquad\square$

4.3 Predictive distributions

As a corollary to Proposition 4.6, we can derive the predictive distribution of further samples collected at time t_i, given the original data set $\mathbf{Y}_{t_{0:N}}$. If the latter contains c_i draws at time t_i, i.e., $\mathbf{Y}_{t_i} = (\mathbf{Y}_{t_i}^1, \dots, \mathbf{Y}_{t_i}^{c_i})$, denote for the present purposes by $\mathbf{Y}_{1:m}^{c_i+1} := (Y_1^{c_i+1}, \dots, Y_m^{c_i+1})$ the elements of the (c_i+1)-th draw at t_i, where m is random and we have dropped the subscript t_i.

Corollary 4.7. *In the setting of Proposition 4.6, let* (41) *be the conditional law of* Z_{t_i} *given* $\mathbf{Y}_{t_{0:N}}$, *and let*

$$m|\mathbf{Y}_{t_{0:N}} \sim \sum_{\mathbf{k}_{i-1} \leq \overleftarrow{\mathbf{n}}_{i-1}} \sum_{\mathbf{k}_{i+1} \leq \overrightarrow{\mathbf{n}}_{i+1}} p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}} \text{NegBin}\left(\tilde{\theta}, \{1 + \beta(\Delta_i, \Delta_{i+1})\}^{-1}\right), \tag{43}$$

with $\tilde{\theta} = \theta + |\mathbf{k}_{i-1}| + |\mathbf{n}_i| + |\mathbf{k}_{i+1}|$ *and* $\beta(\Delta_i, \Delta_{i+1})$ *as in* (38). *Then the law of the $(k+1)$-st further element of the (c_i+1)-st draw from* (5) *is*

$$\mathbb{P}(Y_{k+1}^{c_i+1} \in A|\mathbf{Y}_{t_{0:N}}, Y_{1:k}^{c_i+1}, m) =$$

$$= \sum_{\mathbf{k}_{i-1} \leq \overleftarrow{\mathbf{n}}_{i-1}} \sum_{\mathbf{k}_{i+1} \leq \overrightarrow{\mathbf{n}}_{i+1}} \tilde{p}_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}} \frac{\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}(A) + \sum_{j=1}^{k} \delta_{Y_j^{c_i+1}}(A)}{\theta + |\mathbf{k}_{i-1}| + |\mathbf{n}_i| + |\mathbf{k}_{i+1}| + k} \tag{44}$$

for every A Borel set of $\mathcal{Y}$, with $\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}$ as in (19),

$$\tilde{p}_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}} \propto p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}\, p(Y_{1:k}^{c_i+1}|\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1})p(m|\,|\mathbf{k}_{i-1}|,|\mathbf{n}_i|,|\mathbf{k}_{i+1}|)$$

where $p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}$ are as in (42)*, and $p(m|\,|\mathbf{k}_{i-1}|,|\mathbf{n}_i|,|\mathbf{k}_{i+1}|)$, $p(Y_{1:k}^{c_i+1}|\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1})$ are the distribution induced by* (43) *and* (44) *respectively.*

Proof. Following (6), with M the random number of elements of $\mathbf{Y}^{c_i+1}$, we can write

$$\mathbb{P}(\,M=m|\mathbf{Y}_{t_{0:N}}) = \mathbb{E}\Big[\mathbb{P}(M=m|Z_{t_i})|\mathbf{Y}_{t_{0:N}}\Big] = \mathbb{E}\left[\frac{e^{-|Z_{t_i}|}|Z_{t_i}|^m}{m!}\Big|\mathbf{Y}_{t_{0:N}}\right]$$

$$= \sum_{\mathbf{k}_{i-1}\leq\overleftarrow{\mathbf{n}}_{i-1}}\sum_{\mathbf{k}_{i+1}\leq\overrightarrow{\mathbf{n}}_{i+1}} p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}} \int \frac{e^{-z}z^m}{m!}\frac{\beta(\Delta_i,\Delta_{i+1})^{\tilde{\theta}}}{\Gamma(\tilde{\theta})}z^{\tilde{\theta}-1}e^{-\beta(\Delta_i,\Delta_{i+1})\,z}\,\mathrm{d}z$$

$$= \sum_{\mathbf{k}_{i-1}\leq\overleftarrow{\mathbf{n}}_{i-1}}\sum_{\mathbf{k}_{i+1}\leq\overrightarrow{\mathbf{n}}_{i+1}} p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}\mathrm{NegBin}\Big(\tilde{\theta},\{1+\beta(\Delta_i,\Delta_{i+1})\}^{-1}\Big)$$

where we have used Proposition 4.6, with $\tilde{\theta} = |\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}|$ and $\beta(\Delta_i,\Delta_{i+1})$ as in (38), from which the first statement follows (cf. also Lemma 4.1). Notice that, as desired,

$$|\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}| = \alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}(\mathcal{Y}) = \theta + |\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}| = \theta + |\mathbf{k}_{i-1}| + |\mathbf{n}_i| + |\mathbf{k}_{i+1}|.$$

Thus, again from (6), conditional on $M=m$, the m elements of $\mathbf{Y}^{c_i+1}$ are sampled from the marginal distribution induced by the normalized measure $Z_{t_i}/|Z_{t_i}|$. The second statement then follows along the same lines of Corollary 3.8, upon noting that $\Gamma_\alpha^\beta/|\Gamma_\alpha^\beta| \stackrel{\mathrm{d}}{=} \Pi_\alpha$ for every finite measure α. $\qquad\square$

Similar implications to those discussed at the end of Section 3.3 apply here as well.

References

ANTONIAK, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–1174.

ASCOLANI, F., LIJOI, A. and RUGGIERO, M. (2021). Predictive inference with Fleming–Viot-driven dependent Dirichlet processes. *Bayesian Anal.* **16** (2), 371–395.

BLACKWELL, D. and MACQUEEN, J.B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1**, 353–355.

CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in hidden Markov models.* Springer.

CHALEYAT-MAUREL, M. and GENON-CATALOT, V. (2006). Computable infinite-dimensional filters with applications to discretized diffusion processes. *Stoch. Proc. Appl.* **116**, 1447–1467.

CHALEYAT-MAUREL, M. and GENON-CATALOT, V. (2009). Filtering the Wright–Fisher diffusion. *ESAIM Probab. Stat.* **13**, 197–217.

CINLAR, E. (2011). *Probability and stochastics.* Springer.

DAWSON, D.A. (1975). Stochastic evolution equations and related measure processes. *J. Mult. Analysis* **5**, 1–52.

DAWSON, D.A. (2010). *Introductory lectures on stochastic population systems.* Technical Report Series **451**, Laboratory for Research in Statistics and Probability, Carleton University.

DEL MORAL, P. (1998). Measure–valued processes and interacting particle systems. Application to nonlinear filtering problems. *Ann. Appl. Probab.* **8**, 438–495.

DONNELLY, P. and KURTZ, T.G. (1996). A countable representation of the Fleming–Viot measure-valued diffusion. *Ann. Probab.* **24**, 698–742.

DOUCET, R., GARIVIER, A., MOULINES, E. and OLSSON, J. (2011). Sequential Monte Carlo smoothing for general state space hidden Markov models. *Ann. Appl. Probab.* **21**, 2109–2145.

ETHERIDGE, A. and MARCH, P. (1991). A note on superprocesses. *Probab. Theory ReI. Fields* **89**, 141–147.

ETHIER, S.N. and GRIFFITHS, R.C. (1993a). The transition function of a Fleming–Viot process. *Ann. Probab.* **21**, 1571–1590.

ETHIER, S.N. and GRIFFITHS, R.C. (1993b). The transition function of a measure-valued branching diffusion with immigration. In *Stochastic Processes. A Festschrift in Honour of Gopinath Kallianpur* (S. Cambanis, J. Ghosh, R.L. Karandikar and P.K. Sen, eds.), 71-79. Springer, New York.

ETHIER, S.N. and KURTZ, T.G. (1981). The infinitely-many-neutral-alleles diffusion model. *Adv. Appl. Probab.* **13**, 429–452.

ETHIER, S.N. and KURTZ, T.G. (1986). *Markov processes: characterisation and convergence.* Wiley, New York.

ETHIER, S.N. and KURTZ, T.G. (1993). Fleming–Viot processes in population genetics. *SIAM J. Control Optim.* **31**, 345–386.

FENG, S. (2010). *The Poisson–Dirichlet distribution and related topics.* Springer, Heidelberg.

FERGUSON, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*

1, 209–230.

Fleming, W.H. and Viot, M. (1979). Some measure valued Markov processes in population genetics theory. *Indiana Univ. Math. J.* **28**, 817–843.

Griffiths, R.C. (1980). Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theor. Popul. Biol.* **17**, 37–50.

Jenkins, P. A. and Spanò, D. (2017). Exact simulation of the Wright–Fisher diffusion. Ann. App. Probab. **27**, 1478–1509.

Kingman, J.F.C. (1975). Random discrete distributions. *J. Roy. Statist. Soc. Ser. B* **37**, 1–22.

Kon Kam King, G., Papaspiliopoulos, O. and Ruggiero, M. (2020). Exact inference for a class of non-linear hidden Markov models on general state spaces. *Electron. J. Statist.* **15**(1), 2832–2875.

Lo, A.Y. (1982). Bayesian nonparametric statistical inference for Poisson point process. *Z. Wahrsch. Verw. Gebiete* **59**, 55–66.

Lo, A.Y. and Weng, C.-S. (1989). On a class of Bayesian nonparametric estimates: II. Hazard rate estimates. *Ann. Inst. Statist. Math.* **41**, 227–245.

Papaspiliopoulos, O. and Ruggiero, M. (2014). Optimal filtering and the dual process. *Bernoulli* **20**, 1999–2019.

Papaspiliopoulos, O., Ruggiero, M. and Spanó, D. (2016). Conjugacy properties of time-evolving Dirichlet and gamma random measures. *Electronic Journal of Statistics* **10**, 3452–3489.

Perkins, E.A. (1991). Conditional Dawson–Watanabe processes and Fleming–Viot processes. In *Seminar on Stochastic processes*, 143–156, Birkhäuser, Boston.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetic models. *Theor. Popul. Biol.* **26**, 119–164.

Walker, S.G., Hatjispyros, S.J. and Nicoleris, T. (2007). The Fleming–Viot process and Bayesian nonparametrics. *Ann. Appl. Probab.* **17**, 67–80.

Watanabe, S. (1968). A limit theorem of branching processes and continuous state branching processes. *J. Math. Kyoto Univ.* **8**, 141–167.