EXACT PHYLODYNAMICS VIA STRUCTURED MARKOV GENEALOGY PROCESSES

AARON A. KING, QIANYING LIN, AND EDWARD L. IONIDES

ABSTRACT.

1. Introduction.

Problem of phylodynamics. Factorization of problem into two subproblems.

Relation to previous work. Existing methods (Volz et al., 2009; Stadler, 2010). Large-population, small sample-size approximations.

Extension of previous results (King et al., 2022). Broader class of state-spaces. Accommodating discrete structure.

Classes of Markov processes. Utility and flexibility of Markov assumptions.

Population process induces Markov history and genealogy processes. Using these, we derive equations for the likelihood of a genealogy conditional on the history. We then integrate out the history to obtain nonlinear filtering equations, the solution of which yields the likelihood. These readily lend themselves to a family of sequential Monte Carlo algorithms for computing the likelihood. We demonstrate with several examples.

In the following, we show a Markov population process of the kind that is a staple in epidemiology induces a Markov process on the space of genealogies. We then show how one can comput the likelihood of a given genealogy.

2. From population processes to genealogy processes.

2.1. Population processes.

Motivating examples: compartmental models. Wide variety of models. Linear chain trick. Migration, superspreading, competition between strains.

Another perspective on the Markov processes is to be had from its Markov state transition diagram (Fig. 2).

Mathematical notation. Denote the underlying probability space by $(\Omega, \mathcal{B}, \mathrm{Prob})$. We will assume that our population process is a time-inhomogeneous Markov jump process, $\mathbf{X}_t$, parameterized by time $t \in \mathbb{R}_+ := \{t \in \mathbb{R} \mid t \geqslant 0\}$ and taking values in some space $\mathbb{X}$. In earlier work (King et al., 2022), we limited ourselves to the case $\mathbb{X} = \mathbb{Z}^d$, but here we assume only that $\mathbb{X}$ is a complete metric space with a countable dense subset, *i.e.,* a Polish space. The population process is completely specified by its

Date: July 13, 2023.

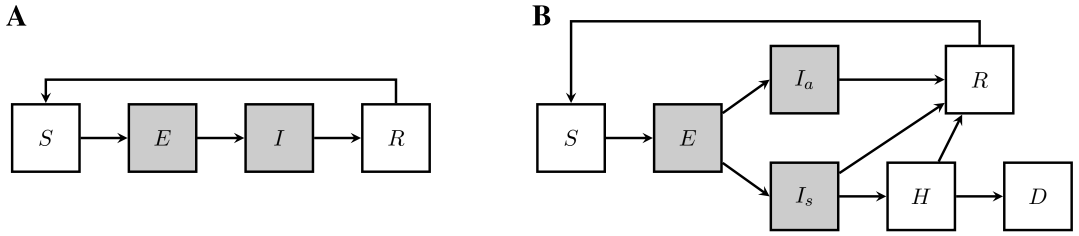

FIGURE 1. Examples of compartmental models. Demes are shaded. [Perhaps another one or two examples here?] [We could add dots to the deme compartments to signify individuals. . . .]

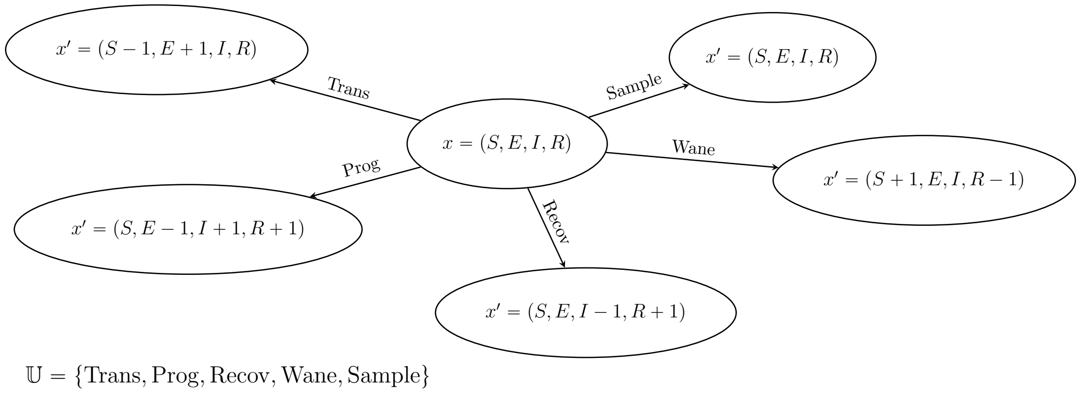

FIGURE 2. Markov state transition diagram for an SEIR model. The state is characterized by four numbers, S, E, I, and R. From a given state x, there are five possible kinds of events $x \mapsto x'$ as shown. From the point of view of the induced genealogy process, Trans (transmission) is of birth type, Prog (progression) is of migration type, and Recov (recovery) is of death type, while Wane (loss or waning of immunity) is of neutral type. Note that, in this formulation, when a Sample (sampling) event occurs, the state does not change.

initial-state distribution, p_0, and its transition rates α. In particular, we suppose that

$$\mathrm{Prob}\left[\mathbf{X}_0 \in \mathcal{E}\right] = \int_{\mathcal{E}} p_0(x)\,\mathrm{d}x \tag{1}$$

for all measurable sets $\mathcal{E} \subseteq \mathbb{X}$. For any $t \in \mathbb{R}_+$, $x, x' \in \mathbb{X}$, we think of the quantity $\alpha(t, x, x')$ as the instantaneous hazard of a jump from x to x'. More precisely, the transition rates have the following properties:

$$\alpha(t, x, x') \geqslant 0, \qquad \int_{\mathbb{X}} \alpha(t, x, x')\,\mathrm{d}x' < \infty,$$

for all $t \in \mathbb{R}_+$ and $x, x' \in \mathbb{X}$. Henceforth, we understand that integrals are taken over all of $\mathbb{X}$ unless otherwise specified. Let $\mathbf{N}_t$ be the number of jumps that $\mathbf{X}$ has taken by time t. We assume that $\mathbf{N}_t$ is a simple counting process so that

$$\mathrm{Prob}\left[\mathbf{N}_{t+\Delta} = n+1 \mid \mathbf{N}_t = n\right] = \Delta \int \alpha(t, x, x')\,\mathrm{d}x' + o(\Delta), \qquad \mathrm{Prob}\left[\mathbf{N}_{t+\Delta} > n+1 \mid \mathbf{N}_t = n\right] = o(\Delta),$$

$$\mathrm{Prob}\left[\mathbf{X}_{t+\Delta} \in \mathcal{E} \mid \mathbf{X}_t = x, \mathbf{N}_{t+\Delta} - \mathbf{N}_t = 1\right] = \frac{\int_{\mathcal{E}} \alpha(t, x, x')\,\mathrm{d}x'}{\int \alpha(t, x, x')\,\mathrm{d}x'} + \frac{o(\Delta)}{\Delta}.$$

[Do we need the last term?] We will further assume that $\mathbf{X}_t$ is non-explosive, *i.e.*, , that $\mathrm{Prob}\left[\mathbf{N}_t < \infty\right] = 1$ for all t. [Is this equivalent to non-explosivity? Or merely an implication?]

The above may be compactly summarized by stating that if $w(t,x)$ satisfies the Kolmogorov forward equation (KFE),

$$\frac{\partial w}{\partial t}(t,x) = \int w(t,x')\,\alpha(t,x',x)\,\mathrm{d}x' - \int w(t,x)\,\alpha(t,x,x')\,\mathrm{d}x', \tag{2}$$

and if, moreover, $w(0,x) = p_0(x)$, then $\int_{\mathcal{E}} w(t,x)\,\mathrm{d}x = \mathrm{Prob}\left[\mathbf{X}_t \in \mathcal{E}\right]$ for every measurable $\mathcal{E} \subseteq \mathbb{X}$. Eq. 2 is sometimes called the *master equation* for $\mathbf{X}_t$.

Without loss of generality, one can assume, as we do here, that the sample paths $t \mapsto \mathbf{X}_t(\omega)$ for $\omega \in \Omega$ are right-continuous with left limits (càdlàg). In fact, all of the processes we will describe in this paper will be taken to be càdlàg, and we will frequently need to refer to the left-limit. Accordingly, if $\mathbf{Z}_t$ is any càdlàg random process, we define

$$\widetilde{\mathbf{Z}}_t := \begin{cases} \lim_{t'\uparrow t} \mathbf{Z}_t, & t > 0, \\ \mathbf{Z}_0, & t = 0. \end{cases}$$

Note that $\widetilde{\mathbf{Z}}_t$ is thus left-continuous with right limits.

Structured populations, demes. In an *unstructured* Markov population process, every lineage is exactly like every other. King et al. (2022) showed how every such process induces an unstructured Markov genealogy process. Here, our aim is to expand the theory considerably by allowing our population of lineages to have discrete structure. In particular, we suppose that there are a countable set of subpopulations that differ in their vital rates, but within each of which, individual lineages are statistically identical. We call these subpopulations *demes*, and use the symbol $\mathbb{D}$ to denote an index set for them.

For any $i \in \mathbb{D}$, we let $n_i(\mathbf{X}_t)$ denote the number of lineages present in deme i at time t, *i.e.,* the *occupancy* of deme i. Thus $n(\mathbf{X}_t) \in \mathbb{Z}_+^{\mathbb{D}}$ is the vector of deme occupancies.

Jump marks. In the following, it will be useful to break the jumps into distinct categories. For this purpose, we let $\mathbb{U}$ be a countable set of jump *marks* such that

$$\alpha(t,x,x') = \sum_u \alpha_u(t,x,x').$$

In Fig. 2, we use the marks to distinguish biologically distinct events. Here and in the following, sums over u are taken over the whole of $\mathbb{U}$ unless otherwise indicated.

Let us define the *jump mark* process, $\mathbf{U}_t$, to be the mark of the latest jump as of time t. As usual, we take the sample paths, $t \mapsto \mathbf{U}_t(\omega)$ for $\omega \in \Omega$, to be càdlàg. Observe that $\mathbf{U}_t$ is a random, but not a Markov, process.

2.2. Examples.

SEIRS model.

SIIR model.

Linear birth-death model.

Moran model and the Kingman coalescent.

2.3. The history, inventory, and genealogy processes.

History process. For $\omega \in \Omega$, $t \mapsto (\mathbf{X}_t(\omega), \mathbf{U}_t(\omega))$ is a càdlàg function of time. Let the *history process*, $\mathcal{H}_t$, be the restriction of this random function to the interval $[0, t]$. Note that $\mathcal{H}_t$ is a Markov process.

The non-explosiveness assumption implies that a.s., for every t, there is a finite, increasing sequence of random event times, which we denote by $\mathsf{event}(\mathcal{H}_t) := (\mathbf{T}_k)_{k=1}^{\mathbf{K}_t}$; the length, $\mathbf{K}_t$, of this sequence is itself random. However, conditional on $\mathcal{H}_t$, $\mathbf{T}_k$ and $\mathbf{K}_t$, together with the mark process $\mathbf{U}_t$, and the population process $\mathbf{X}_t$ are deterministic. One can write down an explicit probability measure for $\mathcal{H}_t$ in terms of these:

$$\begin{aligned}\pi^{\mathcal{H}}(\mathrm{d}\mathcal{H}_t) =& p_0(X_0)\,\mathrm{d}X_0 \prod_{k=1}^{K_t}\left(\alpha_{U_{T_k}}\left(T_k, \widetilde{X}_{T_k}, X_{T_k}\right)\mathrm{d}X_{T_k}\,\mathrm{d}T_k\right)\\ &\times \exp\left(-\sum_u \int_0^t \int \alpha_u(t', \widetilde{X}_{t'}, x')\,\mathrm{d}x'\,\mathrm{d}t'\right).\end{aligned} \tag{3}$$

Inventories. Our goal in this paper is to probabilistically characterize how the genealogical relationships among lineages evolve through time. Accordingly, we develop some notation for this purpose. To begin with, we assign to each lineage a unique number $j \in \mathbb{Z}_+$. This can be done in any fashion, so long as no two lineages ever receive the same number. For example, when a new lineage arises, we can assign it the smallest integer that has not yet been assigned. We will define the *inventory process*, $\mathbf{I}_t$, so that, for every lineage $j \in \mathbb{Z}_+$, $\mathbf{I}_t(j)$ is the deme in which j is found. However, when t is before the birth or after the death of lineage j, then clearly $\mathbf{I}_t(j) \notin \mathbb{D}$. We say in this case that lineage j is in the *underdeme*, which we denote using the symbol $\eth$, so that we can write $\mathbf{I}_t(j) = \eth$. We define $\overline{\mathbb{D}} := \mathbb{D} \cup \{\eth\}$ so that $\mathbf{I} : \mathbb{R}_+ \times \mathbb{Z}_+ \to \overline{\mathbb{D}}$.

The birth and death times of lineage j are therefore

$$t_j^b = \min\{t | \mathbf{I}_t(j) \neq \eth\} \qquad \text{and} \qquad t_j^d = \sup\{t | \mathbf{I}_t(j) \neq \eth\},$$

respectively. Observe that $n_i(\mathbf{X}_t) = |\{j \mid \mathbf{I}_t(j) = i\}|$ for all $t \in \mathbb{R}_+$ and $i \in \mathbb{D}$. Note also that n does not count the inhabitants of the underdeme.

Jump types. Different kinds of events that occur for the population process can have different kinds of effects on the inventory process, and indeed not every jump affects $\mathbf{I}_t$ at all. From the point of view of the inventory process, there are five distinct *pure types* of jumps, which we enumerate here.

(a) Birth-type events result in the branching of one or more new lineages, each from some existing lineage. If j is one of the new lineages, we use the expression $\mathsf{Anc}(j)$ to refer to its ancestor. Examples of birth-type events include transmission events, speciations, and actual births. It is not assumed that all new lineages arising from a birth event share the same ancestor.
(b) Death-type events result in the extinction of one or more lineages. Examples include recovery, death of a host, and species extinctions.
(c) Migration-type events result in the movement of a lineage from one deme to another. Spatial movements, changes in host age or behavior, and progression of an infection can all be represented as migration-type events.
(d) Sample-type events result in the collection of a sample from a lineage but do not in themselves affect the inventory process.
(e) Neutral-type events result in no change to any of the lineages.

Fig. 2 depicts an example with all five of the pure types. It is not necessary that a jump fall into just one of these types. It is allowable, for instance to have compound jumps that fall into more than one category. For example, sample/death-type events, in which a lineage is simultaneously sampled and removed, have been used, as have birth/death events in which one lineage reproduces at the same moment that another dies. The theory presented here places no restrictions on the complexity of the events that can occur.

However, we do impose the restriction that the *production*, *i.e.,* the deme-specific number of lineages emerging from the event, be constant for all jumps of a given mark. To be precise, the *production* is defined to be a function $r : \mathbb{U} \times \mathbb{D} \to \mathbb{Z}_+$, such that r_i^u lineages of deme i emerge from each event of mark u. We write $r^u = (r_i^u)_{i \in \mathbb{D}}$. Note that we lineages that die as a result of an event do not count in the production. Also, it is important to note that the parent lineage or lineages, if they survive the event, are always counted in the production.

Because different kinds of events may differ not only in the number of offspring they engender, but also in the number of parent lineages, and the distribution of offspring among parents and demes, there is implicitly a deterministic indicator function Q_u, for $u \in \mathbb{U}$, (described below) that captures these properties.

Inventory process. The structure of the state space for the inventory process, $\mathbf{I}_t$, has already been described. It remains to define its stochastic dynamics. The $\mathbf{I}_t$ process is driven by the population process $\mathbf{X}_t$ in that jumps in $\mathbf{I}_t$ do not occur except when jumps in $\mathbf{X}_t$ occur: $\widetilde{\mathbf{I}}_t \neq \mathbf{I}_t$ implies $\widetilde{\mathbf{N}}_t \neq \mathbf{N}_t$.

At jumps of birth type and mark u, the appropriate number of random parents are selected from the appropriate deme(s) and each one sires the appropriate number of offspring in each deme. At jumps of death type, the appropriate number of lineages are selected from the appropriate demes and moved to the underdeme. At jumps of migration type, randomly selected lineages are moved between demes as appropriate. At sample-type and neutral-type jumps, no change occurs. Jumps of compound type are handled in the obvious way.

Crucially, the assumption that the population process $\mathbf{X}_t$ is Markov empowers us to assume that the individual lineages within each deme are exchangeable with respect to the inventory process. Since the law governing $\mathbf{X}_t$ is independent of the identities of the individual lineages, and since the individual lineages are exchangeable by assumption, it follows that, conditional on $\mathcal{H}_t$, the jumps of the inventory are independent of one another.

Genealogies. King et al. (2022), showed how an unstructured population process induces a process on the space of genealogies. Although we now treat a more general case, the construction is much the same, so we abbreviate the presentation. Readers wishing for more detail should consult the earlier paper.

For our purposes, a *genealogy* is a labeled, time-calibrated tree. Its edges represent relationships of ancestry and descent among its nodes. There are three distinct kinds of nodes: (i) *tip nodes*, which represent labeled extant lineages; (ii) *internal nodes*, which represent ancestral events; (iii) *sample nodes*, which represent labeled samples. More formally, we can take a genealogy, G, to be a finite sequence of internal nodes, together with a time. Given G, the time is denoted $T(G)$; this is the time corresponding to the extant lineages. The number of nodes is $K(G)$. We order the nodes temporally, and denote the k-th node $p_k(G)$ and we write $p \in G$ if p is one of the nodes of G. Each $p \in \mathbf{G}$ has a creation-time, $T(p)$, a parent, $\mathsf{Anc}(p)$, and a deme. We let $\mathsf{event}(G)$ denote the sequence of node times: $\mathsf{event}(G) = (p_k(G))_{k=1}^{K(G)}$, where $0 = T(p_1(G)) \leqslant T(p_2(G)) \leqslant \ldots \leqslant T(p_{K(G)}(G)) \leqslant T(G)$. Root nodes

FIGURE 3. Illustration of genealogy processes. [Similar to that of King et al. (2022) but with multiple demes represented.]

are distinguished by being their own parents: p is a root if and only if $\mathsf{Anc}(p) = p$. Every node also has one or more descendant nodes called *children*.

Genealogy processes. The population and inventory processes together induce a stochastic process, $\mathbf{G}_t$, on the space of genealogies. In particular, at each event in the population process, one or more of the following changes happen to the genealogy, according to the type of the event:

(a) A birth-type event at time t results in the creation of one new internal node for each parent lineage. In particular, if j is one of the parent lineages, and p is the new node, then $T(p) = t$, $\mathsf{Anc}(p)$ is the node that was parent to j prior to the event. The children of p include one new tip node for each new lineage sired by j, as well as j itself.
(b) In a death-type event, all the lineages j that die are removed. Internal nodes without children are then recursively removed. Sample nodes are never removed.
(c) In a migration-type event, one node is added for each migrating lineage; each one takes the migrating lineage as child. The ancestor of the new node is that which was parent to the migrating lineage prior to the event. The deme of the lineage changes accordingly.
(d) At a sample-type event, one new sample node is introduced for each sampled lineage. Each one takes the sampled lineage as child. The ancestor of the sample node is that which was parent to the sampled lineage before the vent.
(e) At a neutral-type event, no change is made to the genealogy.
(f) Finally, events of compound type are accommodated by combining the foregoing rules.

When an event results in the addition of one or more new nodes to a genealogy, the lineages which are children of that node are said to *emerge* from the event. Thus, after a birth-type event, the emerging lineages include all the new offspring as well as their parents. Likewise, at pure migration- or sample-type events, each migrating or sampled lineage emerges from the event. At pure death-type events, no lineages emerge. In general, at an event of mark u, there are r_i^u emergent lineages in deme i.

Pruning. Although the process just described yields a genealogy that relates all extant members of the population, and all samples, the data we ultimately wish to analyze will contain only the samples. We therefore describe how the genealogy process is *pruned* to yield the sample-only genealogy. Given a genealogy $\mathbf{G}_t$, one obtains the *pruned genealogy*, $\mathbf{P}_t = \mathsf{prune}(\mathbf{G}_t)$ by first dropping every tip node and then recursively dropping every internal node without children. In a pruned genealogy only internal and sample nodes remain, and sample nodes are found at all of the leaves and some of the interior nodes of the genealogy. Observe that the pruned genealogy retains information not only about how much ancestry is shared by any pair of sample lineages, but also about where among the demes each lineage was through time. Note also that $T(\mathbf{P}_t) = T(\mathbf{G}_t) = t$.

Alternative representation of a pruned genealogy. Consider the finite set of all samples represented in pruned genealogy P. Beginning with 1, assign natural numbers to these in such a way that if sample j is ancestral to sample j', then $j < j'$. For example, we can order the samples by their times, resolving any ties arbitrarily. Let $\mathsf{lin}(P)$ denote the set of lineage numbers. Using this ordering, we can uniquely associate each point on a genealogical tree with the least of those lineages that descend from that point. In particular, any lineage $j \in \mathsf{lin}(P)$, corresponding to a sample taken at time t_j^s, can be traced backward from node to node until either it coalesces with some lesser (*i.e.,* older) lineage at some time $t_j^o > 0$ or a root is reached (in which case, we define $t_j^o = 0$). Each node encountered along the way represents a genealogical event from which j emerges. Moreover, at each time $t \in [t_j^o, t_j^s)$, lineage j is in precisely one of the demes $\mathbb{D}$. However, for $t \notin [t_j^o, t_j^s)$, lineage j does not exist. To express this, we again say that lineage j is in the underdeme, $\eth$.

It will be useful to distill the elements that characterize a pruned genealogy. Accordingly, given a pruned genealogy, P, we make the following definitions.

(a) Let $\mathsf{ct}^P : \mathbb{Z}_+ \times \mathbb{R}_+ \to \mathbb{Z}_+$ be such that, for $j \in \mathbb{Z}_+$, $\mathsf{ct}_j^P(t)$ is the counting process which increases by 1 at each event along lineage j.

(b) Let $\mathsf{anc}^P : \mathbb{Z}_+ \times \mathbb{R}_+ \to \mathbb{Z}_+$ be such that $\mathsf{anc}_j^P(t)$ indicates the unique lineage, ancestral of lineage j, and alive at time t. In particular, we posit that $\mathsf{anc}_j^P(t) = j$ for $t \in [t_j^o, t_j^s)$ and $\mathsf{anc}_j^P(t) = 0$ for $t > t_j^s$, so that the function is well defined for all j and t.

(c) Let $\mathsf{deme}^P : \mathbb{Z}_+ \times \mathbb{R}_+ \to \overline{\mathbb{D}}$ be such that, for $j \in \mathbb{Z}_+$, $\mathsf{deme}_j^P(t)$ indicates in which deme lineage j lies at time t. In particular $\mathsf{deme}_j^P(t) = \eth$ if $t \notin [t_j^o, t_j^s)$.

(d) Let $\mathbb{Y} = (\mathbb{Z}_+ \times \mathbb{Z}_+ \times \overline{\mathbb{D}})^{\mathbb{Z}_+}$ and define $Y^P : \mathbb{R}_+ \to \mathbb{Y}$ such that, for $j \in \mathbb{Z}_+$ and $t \in \mathbb{R}_+$, $Y_j^P(t) = \Big(\mathsf{ct}_j^P(t), \mathsf{anc}_j^P(t), \mathsf{deme}_j^P(t)\Big)$.

For $j > |\mathsf{lin}(P)|$, we adopt the convention that $t_j^o = t_j^s = \infty$, so that $\mathsf{ct}_j^P(t) = 0$, $\mathsf{anc}_j^P(t) = 0$, and $\mathsf{deme}_j^P(t) = \eth$ for all $t \in \mathbb{R}_+$. It is easy to see that Y^P is well defined, piecewise constant, and càdlàg, and that the map $P \mapsto Y^P$ is one-to-one. However, not every piecewise constant, càdlàg map $Y : \mathbb{R}_+ \to \mathbb{Y}$ defines a pruned genealogy.

Given $\eta \in \mathbb{Y}$, we will use the notation $\mathsf{ct}(\eta)$, $\mathsf{anc}(\eta)$, and $\mathsf{deme}(\eta)$ to refer to the coordinates of η.

Since $\mathbf{G}_t$ is a stochastic process, both $\mathbf{P}_t := \mathsf{prune}(G_t)$ and $Y^{\mathbf{P}_t}$ are stochastic processes as well. In fact, the latter two are Markovian, since each contains within itself all of its past history. The process $\mathbf{G}_t$ is not Markovian, though $(\mathbf{X}_t, \mathbf{G}_t)$ is.

To visualize $\mathbf{P}_t$, one can make a correspondence between demes and colors. Then a pruned genealogy is visualized as a tree with colored branches. Knowing the function $\mathsf{deme}^{\mathbf{P}_t}$ is equivalent to knowing the coloring, while $\mathsf{ct}^{\mathbf{P}_t}$ determines the locations of events in the genealogy and $\mathsf{anc}^{\mathbf{P}_t}$ determines the topology. Note in particular that, if $y = Y^{\mathbf{P}_t}$, then $y_j(t') \neq \widetilde{y}_j(t')$ if and only if t' is the time of an event from which lineage j emerges. Note also that, there can be events on a pruned genealogy where the color does not change. That is, where ct_j increments, but neither anc_j nor deme_j change.

Lineage count, saturation. In the following, we will find that we need to count the deme-specific numbers of lineages present at a given time. Accordingly, for any $\eta \in \mathbb{Y}$, and $i \in \mathbb{D}$, let us define

$$\ell_i(\eta) := |\{j \in \mathbb{Z}_+ \mid \mathsf{deme}(\eta_j) = i\}| \in \mathbb{Z}_+, \qquad \ell(\eta) := (\ell_i(\eta))_{i \in \mathbb{D}} \in \mathbb{Z}_+^{\mathbb{D}}.$$

Note that lineages j for which $\mathsf{deme}(\eta_j) = \eth$ are not counted. With this definition, it follows that, for any pruned genealogy P, $\ell_i(Y^P(t))$ is the number of lineages in deme i at time t and $\ell(Y^P(t)) \in \mathbb{Z}_+^{\mathbb{D}}$ is the non-negative integer vector telling how many lineages lie within each of the demes at time t.

We will also have occasion to refer to the deme-specific number of lineages emerging from a given event. Therefore, for $\eta, \eta' \in \mathbb{Y}$, and $i \in \mathbb{D}$, let us define

$$s_i(\eta, \eta') := \big|\{j \in \mathbb{Z}_+ \mid \mathsf{deme}(\eta_j) = i \ \ \& \ \ \mathsf{ct}(\eta'_j) \neq \mathsf{ct}(\eta_j)\}\big|, \qquad s(\eta, \eta') := (s_i(\eta, \eta'))_{i \in \mathbb{D}} \in \mathbb{Z}_+^{\mathbb{D}}.$$

With this definition, for any P, if $y = Y^P$, then $s_i(\widetilde{y}(t), y(t))$ is the number of lineages in deme i that emerge from an event at time t and $s(\widetilde{y}(t), y(t))$ is the non-negative integer vector telling how many lineages in each deme emerge at time t. In particular, $s(\eta,\eta) = 0$ so that if $t \notin \mathsf{event}(P)$, then $s(\widetilde{y}(t), y(t)) = 0$. We refer to $s\left(\widetilde{y}(t), y(t)\right)$ as the *saturation* at time t.

2.4. Obscured genealogy.

As we have just seen, a pruned genealogy contains information about the full history of each sample lineage, including the times at which it entered or exited any deme, sired offspring, or was sampled. The data we seek to analyze will typically lack much of this information. Accordingly, we define the *obscured genealogy* to be that obtained by discarding all information about demes and events not visible from the topology of the tree alone. In particular, if $\mathbf{P}$ is a pruned genealogy and $\mathbf{V} = \mathsf{obs}(\mathbf{P})$ is the corresponding obscured genealogy, $\mathbf{V}$ is uniquely determined by the function $t \mapsto \mathsf{anc}(Y^{\mathbf{P}}(t))$.

Binomial ratio. For $n, r, \ell, n \in \mathbb{Z}_+^{\mathbb{D}}$, define the *binomial ratio*

$$\begin{pmatrix} n & \ell \\ r & s \end{pmatrix} := \begin{cases} \dfrac{\prod_{i\in\mathbb{D}} \dbinom{n_i - \ell_i}{r_i - s_i}}{\prod_{i\in\mathbb{D}} \dbinom{n_i}{r_i}}, & \forall i\ n_i \geqslant \{\ell_i, r_i\} \geqslant s_i \geqslant 0 \\ 0, & \text{otherwise} \end{cases}$$

Observe that $\begin{pmatrix} n & \ell \\ r & s \end{pmatrix} \in [0,1]$. Moreover, in consequence of the Chu-Vandermonde identity, we have

$$\sum_{s\in\mathbb{Z}_+^{\mathbb{D}}} \begin{pmatrix} n & \ell \\ r & s \end{pmatrix} \binom{\ell}{s} = 1,$$

whenever $n_i \geqslant \{\ell_i, r_i\} \geqslant 0$ for all i.

3. Results.

Likelihood conditional on history. Our first result will be an expression for the likelihood of a given pruned genealogy P^* given the history $\mathcal{H}_t$ of the population process up to time $t = T(P^*)$. Of course, not every pruned genealogy is compatible with $\mathcal{H}_t$. In particular, P^* is only compatible with $\mathcal{H}_t$ if every event in P^* coincides with an event in $\mathcal{H}_t$. That is, only if $\mathsf{event}(P^*) \subseteq \mathsf{event}(\mathcal{H}_t)$. Moreover, certain events in P^* may be incompatible with all possible histories. For example, if P^* has an event in which a lineage moves from deme i to deme i' but there are no $u \in \mathbb{U}$ for which this is possible, then P^* is incompatible with the population process itself. Let us define the function $Q : \mathbb{U} \times \mathbb{Y} \times \mathbb{Y} \to \{0,1\}$ so that $Q_u(\eta,\eta') = 1$ if and only if a change $\eta \to \eta'$ in a pruned genealogy is compatible with an event of mark u at that time. More formally, $Q_u(\eta,\eta') = 1$ if and only if there are $\omega \in \Omega$ and $t' < t \in \mathbb{R}_+$ such that $\mathbf{U}_{t'}(\omega) = u$, $\widetilde{Y}^{\mathsf{prune}(\mathbf{G}_t(\omega))}(t') = \eta$, and $Y^{\mathsf{prune}(\mathbf{G}_t(\omega))}(t') = \eta'$. Using this, we define

$$\phi_u(\xi,\eta,\eta') := \begin{pmatrix} n(\xi) & \ell(\eta') \\ r^u & s(\eta,\eta') \end{pmatrix} Q_u(\eta,\eta').$$

for $\xi \in \mathbb{X}$, $\eta, \eta' \in \mathbb{Y}$, and $u \in \mathbb{U}$. Recall that $n : \mathbb{X} \to \mathbb{Z}_+$ is the deme-occupancy function, r^u is the production of a jump of mark u, ℓ is the lineage-count function, and s is the saturation.

Theorem 1. *Let P^* be a given pruned genealogy, $t = T(P^*)$, and $y^* = Y^{P^*}$. Then*

$$\mathrm{Prob}\left[P^* | \mathcal{H}_t\right] = \mathbb{1}\{\mathsf{event}(P^*) \subseteq \mathsf{event}(\mathcal{H}_t)\} \times \prod_{e\in\mathsf{event}(\mathcal{H}_t)} \phi_{U_e}\Big(X_e, \widetilde{y}^*(e), y^*(e)\Big).$$

Proof. As we have already observed, if $\mathsf{event}(P^*) \not\subseteq \mathsf{event}(\mathcal{H}_t)$, then $\mathrm{Prob}\left[P^*\right] = 0$. Similarly, if there is any event of P^* which is incompatible with the population process, $\mathrm{Prob}\left[P^*\right] = 0$. Let us therefore suppose that neither of these conditions hold. Recall that, conditional on $\mathcal{H}_t$, each jump of the inventory process is independent of the others. Moreover, at each event $e \in \mathsf{event}(\mathcal{H}_t)$, a jump of mark $\mathbf{U}_e$ occured, with a production of $r^{U_e} = (r_i)_{i\in\mathbb{D}}$, resulting in a new deme-occupancy of $n(X_e) = (n_i)_{i\in\mathbb{D}}$. In P^*, at time e, there are $\ell_i = \ell_i(y^*(t))$ lineages in deme i, of which $s_i = s_i(\tilde{y}^*(t), y^*(t))$ are emergent. The exchangeability of lineages within demes implies that each lineage present in a deme at time e was equally likely to have been one of the emergent lineages. In particular, at time e, the probability that s_i of the ℓ_i deme-i lineages were among the r_i of n_i lineages emergent in the inventory process is the same as the probability that, drawing ℓ_i balls without replacement from an urn containing r_i black balls and $n_i - r_i$ white balls, one ends up with s_i black balls, namely

$$\frac{\dbinom{n_i - \ell_i}{r_i - s_i}\dbinom{\ell_i}{s_i}}{\dbinom{n_i}{r_i}}.$$

Since our genealogies are labelled, we are interested in the probability that the specific lineages are involved. We therefore divide by $\dbinom{\ell_i}{s_i}$. Since, again conditional on $\mathcal{H}_t$, the inventory process within each deme is independent of the others, we have established that

$$\mathrm{Prob}\left[P^*|\mathcal{H}_t\right] = \prod_{e\in\mathsf{event}(\mathcal{H}_t)} \begin{pmatrix} n(X_e) & \ell(y^*(e)) \\ r^{U_e} & s(\tilde{y}^*(e), y^*(e)) \end{pmatrix}.$$

Returning to the possibility that P^* is incompatible with $\mathcal{H}_t$, since $\mathrm{Prob}\left[P^*\right] = 0$ if either any $Q_u = 0$ or of any event of P^* is not an event of $\mathcal{H}_t$, we obtain the result. □

Our second result concerns the likelihood of a given obscured genealogy conditional on the history.

Theorem 2. *Let V^* be a given obscured genealogy and $t = T(V^*)$. Let $\pi : \mathbb{U} \times \mathbb{R}_+ \times \mathbb{X}^2 \times \mathbb{Y}^2 \to \mathbb{R}_+$ be a probability kernel,* i.e., *$\pi_u(t, \xi, \xi', \eta, \eta') \geqslant 0$ and $\sum_{\eta'} \pi_u(t, \xi, \xi', \eta, \eta') = 1$ for all $u \in \mathbb{U}$, $t \in \mathbb{R}_+$, $\xi, \xi' \in \mathbb{X}$, and $\eta \in \mathbb{Y}$. Suppose moreover that $\pi_u(t, \xi, \xi', \eta, \eta') > 0$ for all t, ξ, ξ' whenever $Q_u(\eta, \eta') = 1$. Let $\mathbf{y}_t$ be the Markov jump process generated by the kernel π, with jump times fixed at* $\mathsf{event}(\mathcal{H}_t)$*. Then*

$$\mathrm{Prob}\left[V^*|\mathcal{H}_t\right] = \mathrm{E}\left[\prod_{e\in\mathsf{event}(\mathcal{H}_t)} \frac{\phi_{U_e}(X_e, \tilde{\mathbf{y}}_t, \mathbf{y}_t)}{\pi_{U_e}(e, \tilde{X}_e, X_e, \tilde{\mathbf{y}}_e, \mathbf{y}_e)}\right],$$

the expectation being taken over $\mathbf{y}_t$.

Proof. First, observe that, since obs is a deterministic operator,

$$\mathrm{Prob}\left[V^*|\mathcal{H}_t\right] = \int \mathbb{1}\{\mathsf{obs}(P) = V^*\}\, \mathrm{Prob}\left[P|\mathcal{H}_t\right] \mathrm{d}P, \tag{4}$$

the integral being taken over all pruned genealogies. Our strategy will be to evaluate Eq. 4 using importance sampling: we will propose pruned genealogies compatible with V^* as sample paths from a Markov process on $\mathbb{Y}$ and treat the integral in Eq. 4 as an expectation over these paths. Conditional on $\mathcal{H}_t$, the probability kernel π generates a Markov chain, $\hat{\mathbf{y}}_k$, on $\mathbb{Y}$ such that

$$\mathrm{Prob}\left[\hat{\mathbf{y}}_k = \hat{y}_k|\hat{\mathbf{y}}_{k-1} = \hat{y}_{k-1}\right] = \pi_{U_{t_k}}(t_k, \tilde{X}_{t_k}, X_{t_k}, \hat{y}_{k-1}, \hat{y}_k).$$

$\hat{\mathbf{y}}_k$ is the embedded chain of a unique, piecewise-constant, càdlàg process $\mathbf{y}_t$, with jump times at t_k. The conditions on π guarantee that for every P such that $\mathsf{obs}(P) = V^*$, then Y^P is a sample path of this process and, conversely, that for every sample path y, there is a pruned genealogy $P = P(y)$ such that $\mathsf{obs}(P) = V^*$ and $Y^P(t) = y_t$.

We can therefore write

$$\mathrm{Prob}\left[V^*|\mathcal{H}_t\right] = \mathrm{E}\left[\frac{\mathrm{Prob}\left[P(\mathbf{y})|\mathcal{H}_t\right]}{\pi(\mathbf{y}|\mathcal{H}_t)}\right],$$

the expectation being taken with respect to the random process $\mathbf{y}_t$. Here, by definition,

$$\pi(\mathbf{y}|\mathcal{H}_t) = \prod_{e \in \mathsf{event}(\mathcal{H}_t)} \pi_{U_e}(e, \widetilde{X}_e, X_e, \widetilde{\mathbf{y}}_e, \mathbf{y}_e).$$

The result follows from Theorem 1. □

Two approaches possible:

1. Use the filter equations developed in the Appendix to compute $\mathrm{E}\left[\mathbb{1}\{\mathsf{obs}(\mathbf{P}_t) = \mathbf{V}_t\}\right]$. In effect, we compute $\mathrm{Prob}\left[\mathbf{P}_t\right]$ and then sum over paintings y.
2. Change the order of the summation: $\mathrm{E}\left[\mathbb{1}\{\mathsf{obs}(\mathbf{P}_t) = \mathbf{V}_t\}\right] = \mathrm{E}\left[\mathrm{E}\left[\mathbb{1}\{\mathsf{obs}(\mathbf{P}_t) = \mathbf{V}_t\} \mid \mathcal{H}_t\right]\right]$. That is, conditional on $\mathcal{H}_t$, the likelihood of $\mathbf{V}_t$ is sum over all $\mathbf{P}_t$ such that $\mathsf{obs}(\mathbf{P}_t) = \mathbf{V}_t$. Therefore, any importance sampling scheme for $\mathbf{P}_t$ will do. In particular, we choose a scheme that paints the tree forward in time, driven by a mimic of $\mathbf{X}_t$. We show that this is equivalent to the specific filter equation.

   It may be useful in this to first show that $\mathrm{Prob}\left[\mathbf{P}_t|\mathcal{H}_t\right]$ can be expressed as an expectation over y, where y is constrained to be equal $y^{\mathbf{P}_t}$. Then, show that the sum of the $\mathbf{P}_t$-constraint indicator functions is equal to the $\mathbf{V}_t$-constraint indicator function.

Integrating out the history, filter equations. Suppose we have a pruned genealogy $\mathbf{P}^*$, defined on the time-interval $[0, T]$, with event times $0 = t_0 < t_1 < \cdots < t_n = T$. From the theorem, we have, for $t \notin \mathsf{event}(\mathbf{P}^*)$,

$$\frac{\partial w}{\partial t} = \sum_u \int w(t,x')\,\alpha_u(t,x',x)\,\phi_u\big(x, y(t), y(t)\big)\,\mathrm{d}x' - \sum_u \int w(t,x)\,\alpha_u(t,x,x')\,\mathrm{d}x'. \tag{5}$$

At event times, $t \in \mathsf{event}(\mathbf{P}^*)$, one has

$$w(t,x) = \sum_u \int \widetilde{w}(t,x')\,\frac{\alpha_u(t_k,x',x)}{\mu}\,\phi_u\big(x, \widetilde{y}(t), y(t)\big)\,\mathrm{d}x', \tag{6}$$

In addition, there are the initial and boundary conditions

$$w(0,x) = p_0(x), \qquad w(t,x) = 0 \quad \text{whenever} \quad n(x) < \ell(y(t)). \tag{7}$$

4. Examples.

4.1. SEIRS.

Jumps: $\mathbb{U} = \{\mathrm{Inf}, \mathrm{Prog}, \mathrm{Recov}, \mathrm{Wane}, \mathrm{Birth}, \mathrm{Deaths}_{\mathrm{S}}, \mathrm{Death}_{\mathrm{E}}, \mathrm{Death}_{\mathrm{I}}, \mathrm{Death}_{\mathrm{R}}, \mathrm{Sample}\}$.

Demes: $\mathbb{D} = \{\mathrm{E}, \mathrm{I}\}$.

Jump rates:

- $\alpha_{\mathrm{Inf}}(t,x,x') = \beta(t)\,\frac{x^{\mathrm{S}}x^{\mathrm{I}}}{N(t)}\,\mathbb{1}\{x' = x + (-1,1,0,0)\}$
- $\alpha_{\mathrm{Prog}}(x,x') = \rho\,x^{\mathrm{E}}\,\mathbb{1}\{x' = x + (0,-1,1,0)\}$
- $\alpha_{\mathrm{Recov}}(x,x') = \gamma\,x^{\mathrm{I}}\,\mathbb{1}\{x' = x + (0,0,-1,1)\}$

- $\alpha_{\mathrm{Wane}}(x,x') = \upsilon\, x^{\mathrm{R}}\, \mathbb{1}\{x' = x + (1,0,0,-1)\}$
- $\alpha_{\mathrm{Sample}}(t,x,x') = \psi\, x^{I}\, \mathbb{1}\{x' = x\}$
- $\alpha_{\mathrm{Birth}}(t,x,x') = B(t)\, \mathbb{1}\{x' = x + (1,0,0,0)\}$
- $\alpha_{\mathrm{Death}_k}(x,x') = \mu\, x^{k}\, \mathbb{1}\{x'^{j} = x^{j} - \delta_{jk}\}$, $k \in \{\mathrm{S},\mathrm{E},\mathrm{I},\mathrm{R}\}$

5. Discussion.

References.

King, A. A., Lin, Q., & Ionides, E. L. (2022) Markov genealogy processes. *Theoretical Population Biology* **143**:77–91.

Stadler, T. (2010) Sampling-through-time in birth-death trees. *Journal of Theoretical Biology* **267**:396–404.

Volz, E. M., Kosakovsky Pond, S. L., Ward, M. J., Leigh Brown, A. J., & Frost, S. D. W. (2009) Phylodynamics of infectious disease epidemics. *Genetics* **183**:1421–1430.

A. Filter equations.

Explicit expressions for the probability densities that will arise in the following are not always available. Hence, we will develop some technology for manipulating them that avoids the need for explicit expressions.

Definition 1. Suppose X_t is a continuous-time Markov process with Kolmogorov forward equation (KFE)

$$\frac{\partial w}{\partial t} = \int w(t,x')\,\beta(t,x',x)\,\mathrm{d}x' - \int w(t,x)\,\beta(t,x,x')\,\mathrm{d}x', \tag{8}$$

Suppose that $B(t,x,x') > 0$ and $\lambda(t,x) \in \mathbb{R}$ are given functions. We say that the equation

$$\frac{\partial w}{\partial t} = \int w(t,x')\,\beta(t,x',x)\,B(t,x',x)\,\mathrm{d}x' - \int w(t,x)\,\beta(t,x,x')\,\mathrm{d}x' - \lambda(t,x)\,w(t,x) \tag{9}$$

is the *filter equation* with *driver* X_t, *boost* B, and *decay* λ.

Filter equations afford a convenient means of computing the likelihood of a given sequence of events. This is facilitated by the following

Lemma 1. *Eq. 9 is satisfied by* $w(t,x) = \int_0^\infty v\,u(t,x,v)\,\mathrm{d}v$*, where* u *satisfies the KFE*

$$\begin{aligned}\frac{\partial u}{\partial t} = &\int u(t,x',v')\,\beta(t,x',x)\,\delta(v,B(t,x',x)\,v')\,\mathrm{d}x'\,\mathrm{d}v' \\ &\qquad - \int u(t,x,v)\,\beta(t,x,x')\,\delta(v',B(t,x,x')\,v)\,\mathrm{d}x'\,\mathrm{d}v' + \tfrac{\partial}{\partial v}\left[\lambda(t,x)\,v\,u(t,x,v)\right].\end{aligned} \tag{10}$$

Here, $\delta(v,v')$ *is the familiar Dirac* δ*.*

Proof.

$$\begin{aligned}\frac{\partial w}{\partial t} &= \int v\,\frac{\partial u}{\partial t}(t,x,v)\,\mathrm{d}v\\ &= \int v\,u(t,x',v')\,\beta(t,x',x)\,\delta(v,B(t,x',x)v')\,\mathrm{d}v\,\mathrm{d}x'\,\mathrm{d}v'\\ &\qquad - \int v\,u(t,x,v)\,\beta(t,x,x')\,\delta(v',B(t,x,x')v)\,\mathrm{d}v\,\mathrm{d}x'\,\mathrm{d}v'\\ &\qquad + \int v\,\tfrac{\partial}{\partial v}\left[\lambda(t,x)\,v\,u(t,x,v)\right]\,\mathrm{d}v.\end{aligned}$$

Evaluating the first integral with respect to v, the second with respect to v', and the third by parts, we obtain

$$\begin{aligned}\frac{\partial w}{\partial t} = \int v'\,u(t,x',v')\,\beta(t,x',x)\,B(t,x',x)\,\mathrm{d}v'\,\mathrm{d}x' - \int v\,u(t,x,v)\,\beta(t,x,x')\,\mathrm{d}v\,\mathrm{d}x'\\ - \lambda(t,x)\int v\,u(t,x,v)\,\mathrm{d}v,\end{aligned}$$

which is simplified to obtain Eq. 9. □

We recognize in Eq. 10 the KFE of a certain process (X_t, V_t). In particular, X_t is the driver with KFE Eq. 8. The V_t process has jumps wherever X_t does, such that when X_t jumps from x to x', V_t jumps by the multiplicative factor $B(t,x,x')$. Between jumps, V_t decays deterministically and exponentially at rate $\lambda(t,x)$. If we view V_t as a weight, then Lemma 1 says that the V_t-weighted average of X_t evolves according to Eq. 9. This motivates the following result, which effectively allows boosts of zero.

Proposition 3. *Suppose X_t is a continuous-time Markov process with state space $\mathbb{X}$ and KFE as in Eq. 8. Let H_t be its history process. That is, for $\omega \in \Omega$ and $t \in \mathbb{R}_+$, $H_t(\omega) : [0,t] \to \mathbb{X}$ such that, for $t' \in [0,t]$, $H_t(\omega)(t') = X_{t'}(\omega)$. Moreover, there are random variables $K \in \mathbb{Z}_+$ and $t_k \in (0,t]$, $k = 1,\dots,K$ such that $\widetilde{X}_t = X_t$ whenever $t \neq t_k$ for all k. Suppose F is a real-valued function of H_t such that*

$$F(H_t) = \prod_k Q\Big(t_k, \widetilde{H}(t_k), H(t_k)\Big)\,B\Big(t_k, \widetilde{H}(t_k), H(t_k)\Big),$$

for some given measurable functions $B > 0$, $Q \in \{0,1\}$. Suppose w satisfies the filter equation

$$\begin{aligned}\frac{\partial w}{\partial t}(t,x) = \int w(t,x')\,\beta(t,x',x)\,Q(t,x',x)\,B(t,x',x)\,\mathrm{d}x' - \int w(t,x)\,\beta(t,x,x')\,Q(t,x,x')\,\mathrm{d}x'\\ - \int w(t,x)\,\beta(t,x,x')\,\Big(1 - Q(t,x,x')\Big)\,\mathrm{d}x'.\end{aligned}$$

Then $\mathrm{E}\left[F(H_t)\right] = \int w(t,x)\mathrm{d}x$.

Proof. Apply Lemma 1 with the driver generated by the rate functions $\beta(t,x,x')\,Q(t,x,x')$. □

An important special case is that of a deterministic driving process. The following result is established by routine calculation.

Proposition 4. *Suppose $X : [0,T] \to \mathbb{X}$ is a deterministic, piecewise constant, càdlàg function. Let E be the set of its jump times. Then the KFE for X_t is Eq. 8 with $\beta(t,x,x') = \sum_{e \in E} \delta(t,e)\delta(x',X_e)$.*

With this driver, the filter equation (Eq. 9) becomes

$$\frac{\partial w}{\partial t} = -\lambda(t,x)\,w(t,x) \quad \textit{for} \quad t \notin E,$$
$$w(e,x) = \delta(x, X_e) \int \widetilde{w}(e,x')\,B(e,x',X_e)\,\mathrm{d}x' \quad \textit{for} \quad e \in E.$$

The results so far allow us to compute expectations over random jumps and jump times. We will have occasion to compute marginal expectations in the situation where some jump times are known. The following result handles this case.

Proposition 5. *Suppose $\mathbf{X}_t$ is a continuous-time Markov process with state space $\mathbb{X}$ and KFE as in Eq. 2. Let $\mathcal{H}_t$ be its history process. Suppose $0 < t_1 < \cdots < t_N \leqslant t$ are fixed times and set $E = \{t_1, \dots, t_N\}$. Let $B > 0$ and $Q \in \{0,1\}$ be given measurable functions. Suppose F_E is an $\mathbb{R}_+$-valued random variable such that*

$$\mathrm{E}\left[F_E \mid \mathcal{H}_t\right] = \prod_{e\in E} B(e, \widetilde{\mathbf{X}}_e, \mathbf{X}_e)\,Q(e, \widetilde{\mathbf{X}}_e, \mathbf{X}_e).$$

Let $\pi(x,x')$ be a given probability kernel and Y_t be the Markov process generated by $\beta(t,x,x') = \left(\alpha(t,x,x') + \sum_{e\in E}\delta(t,e)\,\pi(x,x')\right)\,Q(t,x,x')$. Suppose w satisfies the filter equation with driver Y_t, boost B, and decay $\lambda = \int \alpha(t,x,x')\,\left(1 - Q(t,x,x')\right)\,\mathrm{d}x'$. Then $\mathrm{E}\left[F_E\right] = \int w(t,x)\,\mathrm{d}x$.

Proof. Apply Lemma 1. □

A. A. King, Department of Ecology & Evolutionary Biology, Center for the Study of Complex Systems, and Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 USA

Email address: kingaa@umich.edu

URL: https://kinglab.eeb.lsa.umich.edu/

Q.-Y. Lin, Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM XXXXX USA

E. L. Ionides, Department of Statistics University of Michigan, Ann Arbor, MI 48109 USA