MA52128 Report (Full Lenght): Feynman-Kac formulae and Sequential Monte Carlo

Jan Olucha Fuentes

January 8, 2026

Notation

- For a process $X_0, X_1, X_2, \cdots$, we write $X_{0:T}$ to refer to the vector $(X_0, \cdots, X_T)$. Likewise, if $A_0, \cdots, A_T$ is a collection of sets, we denote their Cartesian product $A_0 \times \cdots \times A_T$ by $A_{0:T}$.
- We often write measures in *integral form*. For example, when working with a probability kernel $M : E_1 \times \mathcal{B}(E_2) \to [0, 1]$, we often write expressions like

  $$\nu(dy) = M(x, dy),$$

  as shorthand to indicate that the measure ν on $(E_2, \mathcal{B}(E_2))$ is defined as

  $$\nu(A) = \int_A M(x, dy), \qquad A \in \mathcal{B}(E_2)$$

- $\mathbf{N}, \mathbf{Z}, \mathbf{R}$: the natural numbers, the integers and the real numbers respectively.
- For a measurable space $(E, \mathcal{E})$, we write $\mathcal{P}(E)$ for the set of all probability measures on $(E, \mathcal{E})$.
- For a measurable space $(E, \mathcal{E})$, we write $\mathcal{B}_b(E)$ as the set of measurable bounded functions on E.
- We say that $f(n) \asymp g(n)$ for two quantities that depend on some n if there are constants $c < C$ independent of n such that

  $$cf(n) \leq g(n) \leq Cg(n).$$

Abstract

In this report, we describe the general framework of Feynman–Kac formulae. We motivate their introduction through two statistical applications that give rise to the same Feynman–Kac structure. We present the sequence of Feynman–Kac formulae as a flow on the space of measures, along with three different particle interpretations of these formulae. In the spirit of seeking variance reduction methods for the simulation from these measures, we motivate and study the Particle Filter algorithm. Finally, a second approach for variance reduction, through a change of reference measure is discussed.

Statement of Authorship

The contents and ideas of this report have been obtained mainly from the books [1] and [3], although the author has rephrased and motivated, whenever possible, the explanations in his own words. All diagrams have been created using the tikzcd package or with Python. Generative AI (ChatGPT-5, OpenAI, https://chatgpt.com) has been used to aid in the production of the Python-created diagrams, as well as for grammatical and spelling checks.

Contents

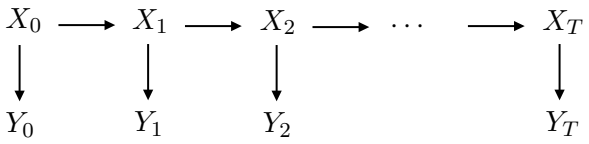


Figure 1: Graphical representation of a State-Space model.

1 Motivation

In this section we describe two examples from different application areas and see that after some thought they all share a common structure. These discussions correspond to Chapters 2 and 3 from [1].

1.1 State-Space Models and the Filtering Problem

A state-space model is in essence a setup in which we have two processes: a Markov chain $(X_n)_{n\geq 0}$ which we can't directly observe, and some noisy observations $(Y_n)_{n\geq 0}$ of $(X_n)_{n\geq 0}$. Formally:

Definition 1.1 (State-Space Model)**.** *A State-Space Model consists of a pair of processes:*

1. *A discrete-time Markov chain $X = (X_n)_{n\geq 0}$ on a measurable state space $(E, \mathcal{E})$ with transition kernels $(M_n)_{n\geq 0}$ and initial law μ_0.*
2. *A process (Y_n) taking values in some measurable state space $(H, \mathcal{H})$ which conditional on $X_n = x_n$, is sampled according to some densities $f_n(y_n|x_n)$.*

Remark 1.2. *According to the sampling scheme above, the joint law is given by*

$$\mathbf{P}(X_{0:T} \in dx_{0:T}, Y_{0:T} \in dy_{0:T}) = \mu_0(dx_0) \prod_{t=1}^{T} M_t(x_{t-1}, dx_t) \prod_{t=0}^{T} f_n(y_t|x_t) dy_t.$$

Let $\{(X_t), (Y_t), (P_t), (f_t)\}$ be a State-Space model as we described it above. Imagine we wish to obtain information about the process X (we may refer to this as the latent variables) given observations of the process Y. This is referred to as the *filtering problem*.

To this end, let $F : E^T \to \mathbf{R}$ be a measurable function, and suppose we want to compute the expected value of a function of the *latent variables* $X_{0:T}$ given the observations $Y_{0:T}$. We can write this out fully as

$$\mathbf{E}\left[F(X_{0:T}) \mid \{Y_{0:T} = y_{0:T}\}\right] = \frac{\int_{E^{T+1}} F(x_{0:T}) \mu_0(dx_0) \prod_{t=1}^{T} M_t(x_{t-1}, dx_t) \prod_{t=0}^{T} f_t(y_t|x_t)}{\int_{E^{T+1}} \mu_0(dx_0) \prod_{t=1}^{T} M_t(x_{t-1}, dx_t) \prod_{t=0}^{T} f_t(y_t|x_t)}.$$

The key observation is that we can rewrite this as $\mathbf{E}_{\mathbf{Q}_T}[F(X_{0:T})]$, where $\mathbf{Q}_T$ is the measure on $(E^T, \mathcal{E}^T)$ given by:

$$\mathbf{Q}_T(dx_{0:T}) \propto \left(\prod_{t=0}^{T} f_t(y_t|x_t) \right) \underbrace{\mu_0(dx_0) \prod_{t=1}^{T} M_t(x_{t-1}, dx_t)}_{\mathbf{P}_T(dx_{0:T})},$$

where $\mathbf{P}_T$ is just the measure on $(E^T, \mathcal{E}^T)$ induced by the Markov chain X. Thus we have seen that we have expressed our problem in a neat way by considering a tilted measure $\mathbf{Q}_T$, where the **weights** where the likelihoods. $f_t(y_t|x_t)$.

1.2 Rare Event Simulation

Suppose we have a Markov chain $X = (X_n)_{n\geq 0}$ taking values in a state space E with some initial law $\mu_0 \in \mathcal{P}(E)$. For a fixed time $T > 0$ and a measurable $A \in \mathcal{E}$, define

$$A_T = \{X_t \in A : t \leq T\}.$$

We may treat A_T as a *rare event*.

Example 1.3. Let X be a Simple Symmetric Random Walk (SSRW) on the integers. A rare event of interest, to which we will come back later in this report, would be the event

$$A_T = \{|X_t| < k : t \leq T\}$$

as we will show later, the probability of this even decays exponentially.

We may be interested in computing

$$\mathbf{E}[F(X_{0:T})|\{X \text{ is alive at time } T\}],$$

which is of course nothing but

$$\frac{1}{Z}\mathbf{E}\left[F(X_{0:T})\prod_{t\leq T}\mathbf{1}(|X_t| < k)\right],$$

where $Z = \mathbf{E}\left[\prod_{t\leq T}\mathbf{1}(|X_t| < k)\right] = \mathbf{P}(\{X \text{ is alive at time } T\})$. We now notice that just like in the previous discussion of State-Space Models, we have expressed our quantity of interest as an expectation $\mathbf{E}_{\mathbf{Q}_T}[F(X_0, \cdots, X_T)]$ under a tilted measure

$$\mathbf{Q}_T(dx_0, \cdots, dx_T) \propto \left(\prod_{t\leq T}\mathbf{1}(|x_t| < k)\right)\mathbf{P}\left((X_0, \cdots, X_T) \in (dx_0, \cdots, dx_T)\right).$$

In both of these examples, we have seen appear measures $\mathbf{Q}_T$ of the form

$$\mathbf{Q}_T(dx_0, \cdots, dx_T) \propto \left(\prod_{t\leq T} G_t(x_t)\right)\mathbf{P}\left((X_0, \cdots, X_T) \in (dx_0, \cdots, dx_T)\right). \tag{1.1}$$

The commonality between these two examples was that we had some Markov measure

$$\mathbf{P}\left((X_0, \cdots, X_T) \in (dx_0, \cdots, dx_T)\right)$$

(meaning, the law of a Markov chain up to some time) weighed at each time step by some potentials G_t. These types of measures are the so-called **Feynman-Kac measures**, to which we devote the next chapter of the report to present in slightly greater generality.

2 Feynman-Kac Models

To describe measures of the type 1.1 in full generality, it is useful to introduce a very general framework for Markov chains. We allow the chain to take values in possibly different state spaces at each time step. Denoting the state spaces by $(E_0, \mathcal{E}_0), (E_1, \mathcal{E}_1), \ldots$, we will see that *any* such Markov chain $(X_n)_{n\geq 0}$ can be described canonically[1] as the coordinate projections on the product space

$$\left(\prod_{n\geq 0} E_n,\ \prod_{n\geq 0} \mathcal{E}_n\right),$$

that is,

$$X_m : (\omega_n)_{n\geq 0} \mapsto \omega_m.$$

Different probability measures on $\left(\prod_{n\geq 0} E_n,\ \prod_{n\geq 0} \mathcal{E}_n\right)$ then correspond to different dynamics of the Markov chain. This discussion corresponds to [3, Section 2.2].

2.1 Canonical Probability Spaces

We now describe a rigorous construction of a Markov chain as described above. That is, given an initial law μ_0 on a measurable space $(E_0, \mathcal{E}_0)$ and a collection of transition kernels $\{M_n\}_n$, where

$$M_n : E_n \times \mathcal{E}_{n+1} \to [0, 1]$$

(i.e. the kernels map the n^{th} measurable space to the $(n+1)^{\text{st}}$ one), we seek the following:

1. A probability space $(\Omega, \mathcal{F}, \mathbf{P})$ on which
2. A sequence of random variables $(X_n)_{n\geq 0}$ is defined such that

$$\mathbf{P}((X_0, \ldots, X_n) \in d(x_0, \ldots, x_n)) = \mu_0(dx_0)M_1(x_0, dx_1)\cdots M_n(x_{n-1}, dx_n).$$

We begin by defining (in the integral sense) a measure $\mathbf{P}_{\mu,n}$ on

$$\Omega_n = \prod_{k=0}^{n} E_k$$

equipped with the product σ-algebra

$$\mathcal{F}_n = \prod_{k=0}^{n} \mathcal{E}_k$$

(recall that this is the smallest σ-algebra for which the projection maps $\pi_j : \Omega_n \to E_j$ are measurable), by setting

$$\mathbf{P}_{\mu,n}(d(x_0, \ldots, x_n)) = \mu(dx_0)M_1(x_0, dx_1)\cdots M_n(x_{n-1}, dx_n).$$

By the Ionescu–Tulcea Theorem (cf. [4, Page 249]), there exists a probability measure $\mathbf{P}_\mu$ on

$$\Omega = \prod_{k\geq 0} E_k$$

[1] The word *canonical* is mathematicians' way of saying: "we didn't have to choose anything, so you can't complain about our choices."

equipped with the product σ-algebra $\prod_{k\geq 0}\mathcal{E}_k$, such that $\mathbf{P}_\mu$ coincides with $\mathbf{P}_{\mu,n}$ on all cylinder sets C_n of the form

$$C_n(A_0,\dots,A_n) = \{(\omega_k)_{k\geq 0} : \omega_k \in A_k \text{ for } 0\leq k\leq n\} = A_0\times\cdots\times A_n\times\prod_{k\geq n+1}E_k.$$

That is,

$$\mathbf{P}_\mu(C_n(A_0,\dots,A_n)) = \mathbf{P}_{\mu,n}(A_0\times\cdots\times A_n)\,.$$

On this probability space

$$\left(\prod_{k\geq 0}E_k,\ \prod_{k\geq 0}\mathcal{E}_k,\ \mathbf{P}_\mu\right)$$

we define the process $(X_n)_{n\geq 0}$ by

$$X_n\big((\omega_k)_{k\geq 0}\big) = \omega_n,$$

that is, as the canonical coordinate projections. Observe that under $\mathbf{P}_\mu$, the law of $(X_n)_{n\geq 0}$ is precisely the Markov law with initial distribution μ and transition kernels $(M_n)_n$. Indeed,

$$\mathbf{P}_\mu((X_0,\dots,X_n)\in A_0\times\cdots\times A_n) = \mathbf{P}_\mu(C_n(A_0,\dots,A_n)) = \int_{A_0\times\cdots\times A_n}\mu(dx_0)M_1(x_0,dx_1)\cdots M_n(x_{n-1},dx_n),$$

where the first equality follows from the fact that (X_n) is defined as the coordinate projection, and the second equality follows from the construction of $\mathbf{P}_\mu$ via the Ionescu–Tulcea theorem.
In this way, we can treat Markov chains $(X_n)_{n\geq 0}$ with $X_n\in E_n$ in a "universal" manner by considering a single canonical process (the coordinate projections), whose dynamics are entirely determined by the choice of the measure $\mathbf{P}_\mu$ as described above. We refer to the measures $\mathbf{P}_\mu$ as **reference measures**.

2.2 Feynman-Kac Formulae

In this section, which corresponds to [3, Section 2.3], we describe what a Feynman–Kac model is. As discussed in the motivation section, the main ingredients are a Markov chain and a sequence of potential functions, which are used to weight the probabilities of the paths taken by the chain. We present these models in their most general form, in light of Section 2.1.

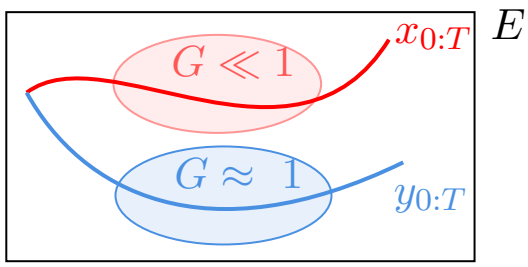
To this end, let $(E_n,\mathcal{E}_n)$ be a sequence of measurable spaces and let $M=(M_n)_n$ be a sequence of Markov kernels from E_n to E_{n+1} describing the transitions of a Markov chain. Let furthermore $\mu\in\mathcal{P}(E_0)$ be a probability measure on the initial space, and construct the probability space

$$\left(\prod_n E_n,\ \prod_n\mathcal{E}_n,\ X,\ \mathbf{P}_\mu\right),$$

where X_n denotes the projection onto the n^{th} coordinate and $\mathbf{P}_\mu$ is the reference measure associated with the Markov chain started from the initial law μ and transition kernels M.
Let $G=(G_n)_n$ be a collection of measurable functions, where each G_n is $\mathcal{E}_n$-measurable, bounded, and non-negative, and such that

$$\mathbf{E}_\mu\left[\prod_{k\leq n}G_k(X_k)\right] > 0 \quad \text{for all } n.$$

We may now define the Feynman–Kac path measures.



$$\mathbf{Q}_T(x_{0:T}) \ll \mathbf{Q}_T(y_{0:T})$$

Figure 2: Visual depiction of the FK path model: in this figure we see two typical paths $x_{0:T}, y_{0:T}$ by time T of the chain with dynamics $\mathbf{P}_\mu$. The FK path model makes some paths more likely than others, with path passing through the red region of low potential being assigned a much lower probability than the blue path which passes through a region of high potential.

Definition 2.1 (Path models)**.** *With X, $\mathbf{P}_\mu$, and G as given above, we define the* prediction *and* updated *Feynman–Kac path models* associated with (G, M) *as the sequences of probability measures on the path space $E_{0:n}$ given by:*

1. **(Prediction):**

$$\mathbf{Q}_{\mu,n}(d(x_0, \ldots, x_n)) = \frac{1}{Z_n}\left(\prod_{k \le n-1} G_k(x_k)\right)\mathbf{P}_{\mu,n}(d(x_0, \ldots, x_n)).$$

2. **(Update):**

$$\widehat{\mathbf{Q}}_{\mu,n}(d(x_0, \ldots, x_n)) = \frac{1}{\widehat{Z}_n}\left(\prod_{k \le n} G_k(x_k)\right)\mathbf{P}_{\mu,n}(d(x_0, \ldots, x_n)).$$

When convenient we may write $\{\mu_0, M, G\}$ to indicate an FK model with initial law μ_0, transition kernels $M = (M_n)_n$, and potentials $G = (G_n)_n$, or simply $\{G, \mathbf{P}_\mu\}$ if its clear that $\mathbf{P}_\mu$ is the law constructed from the initial law μ_0 and the kernels M.

Remark 2.2. *These measures are well defined, since the normalisation constants (often referred to as partition functions) are non-zero by the assumption that*

$$\mathbf{E}_\mu\left[\prod_{k \le n} G_k(X_k)\right] > 0 \quad \textit{for all } n.$$

A useful way to interpret these measures is as follows. A Markov chain assigns probabilities to paths of length n (that is, to elements of $E_{0:n}$) via its reference measure restricted to the σ-algebra up to time n, namely $\mathbf{P}_{\mu,n}$. By introducing the potential functions G_k, we may reweight these paths—making some more likely than others—by multiplying $\mathbf{P}_{\mu,n}(d(x_0, \ldots, x_n))$ by (possibly time-dependent) factors G_k evaluated at each step x_k along the path.
We now introduce the flow of time marginals:

Definition 2.3. *The sequence of measures $(\gamma_n)_n$ (resp. $(\widehat{\gamma}_n)_n$) on $(E_n, \mathcal{E}_n)$ defined by setting for a function $f \in \mathcal{B}_b(E_n)$*

$$\gamma_n(f) = \mathbf{E}_\mu \left[f(X_n) \prod_{k \leq n-1} G_k(X_k) \right],$$

and

$$\widehat{\gamma}_n(f) = \mathbf{E}_\mu \left[f(X_n) \prod_{k \leq n} G_k(X_k) \right],$$

are called the unnormalised prediction (resp. updated) Feynman-Kac models *associated to (G, M). If we divide these measures by $\gamma_n(1)$ (resp. $\widehat{\gamma}_n(1)$), we obtain the probability measures η_n (resp. $\widehat{\eta}_n$) which are called the normalised Feynman-Kac models associated to (G, M).*

A recursion formula and the motivation for Particle Filtering.

There is a nice way of relating the model η_{n+1} from η_n through two step procedure which we describe below. This recursion will motivate a particle approximation for the measures η_n which we will study in depth in Section 3. We first define the following transformation:

Definition 2.4 (Gibbs-Boltzmann transformation)**.** *Let (G_n) be a sequence of potentials as described above, the mappings $\psi_n : \mathcal{P}_n(E_n) \to \mathcal{P}_n(E_n)$ given by*

$$\psi_n(\eta)(dx) = \frac{1}{\eta(G_n)} G_n(x) \eta(dx),$$

are called the Gibbs-Boltzmann transformations. Here $\mathcal{P}_n(E_n)$ is the subset of $\mathcal{P}(E_n)$ consisting of measures η with the property that $\eta(G_n) > 0$.

This transformation is merely tilting the measure η with a potential G_n, and normalising to obtain a probability measure again. We now have the following proposition [3, Proposition 2.3.1]:

Proposition 2.5 (Flow of prediction models,)**.** *Let $(\eta_n)_n$ and $(\widehat{\eta}_n)_n$ be the prediction and updated models described above. Then*

$$\eta_{n+1} = \psi_n(\eta_n) M_{n+1}, \quad \textit{and} \quad \widehat{\eta}_{n+1} = \psi_{n+1}\left(\widehat{\eta}_n M_{n+1}\right).$$

Proof. First we show the fact that $\widehat{\eta}_n = \psi_n(\eta_n)$. Indeed, lets start by noting that:

$$\widehat{\gamma}_n(f) = \mathbf{E}_\mu \left[f(X_n) G_n(X_n) \prod_{k \leq n-1} G_k(X_k) \right] = \gamma_n(G_n f).$$

With this in mind we can observe that

$$\widehat{\eta}_n(f) = \frac{\widehat{\gamma}_n(f)}{\widehat{\gamma}(1)} = \frac{\gamma_n(G_n f)/\gamma_n(1)}{\gamma_n(G_n)/\gamma_n(1)} = \frac{\eta_n(G_n f)}{\eta_n(G_n)} = \psi_n(\eta_n)(f).$$

Now that we have established this, we can finish off by noting that:

$$\begin{aligned}\gamma_n(f) &= \mathbf{E}_\mu\left[f(X_n)\prod_{k\leq n-1}G_k(X_k)\right]\\ &= \mathbf{E}_\mu\left[(M_nf)(X_{n-1})\prod_{k\leq n-1}G_k(X_k)\right] = \widehat{\gamma}_{n-1}(M_nf).\end{aligned}$$

In other words, $\gamma_n = \widehat{\gamma}_{n-1}M_n$. As such, we see that

$$\eta_n(f) = (\widehat{\eta}_{n-1}M_n)(f) = (\psi_{n-1}(\eta_{n-1})M_n)(f).$$

Similarly we obtain the recursion for the updated models. □

2.3 A Particle Interpretation

2.3.1 Markov Chain with Killing

We now describe an interpretation of the Feynman–Kac model as the conditional distribution of a Markov process that has not been killed ([3, Section 2.5.1]). We will interpret potential functions as killing rates. For this purpose, we assume that the potential functions are strictly positive and moreover bounded above by 1.

Definition 2.6 (Boltzmann Multiplicative Operator)**.** *For a collection of potentials $G = (G_n)_n$, we define the map $\mathcal{G}_n : \mathcal{B}_b(E_n) \to \mathcal{B}_b(E_n)$ by*

$$\mathcal{G}_n(f) : x \mapsto f(x)G_n(x).$$

Remark 2.7. *We may view $\mathcal{G}_n$ as an integral operator with kernel $\mathcal{G}_n(x, dy) = G_n(x)\delta_x(dy)$. In this way,*

$$\mathcal{G}_n(f)(x) = \int G_n(x)\delta_x(dy)f(y) = G_n(x)f(x).$$

Notice that $\int \mathcal{G}_n(x, dy) \leq 1$, so we may interpret $\mathcal{G}_n$ as a sub-Markov kernel.

We can turn these kernels into genuine Markov kernels by adjoining a common cemetery state Δ to all state spaces E_n, which we denote by E_n^Δ. We then extend the remaining objects as follows:

1. For a function $f \in \mathcal{B}_b(E_n)$, we extend it to a function on E_n^Δ by setting $f(\Delta) = 0$.
2. For a kernel $M_n : E_n \times \mathcal{E}_{n+1} \to [0, 1]$, we extend it to a kernel $M_n^\Delta : E_n^\Delta \times \mathcal{E}_{n+1} \to [0, 1]$ by setting $M_n^\Delta(\Delta, \cdot) = \delta_\Delta$, and for $x \in E_n$, we define $M_n^\Delta(x, \cdot) = M_n(x, \cdot)$.
3. Finally, the extension of $\mathcal{G}_n$ is given by the kernel $\mathcal{G}_n^\Delta$ defined as

$$\mathcal{G}_n^\Delta(x_n, dy_n) = G_n(x_n)\delta_{x_n}(dy_n) + (1 - G_n(x_n))\delta_\Delta(dy_n).$$

Remark 2.8. *This extension has the following effects:*

1. *For M_n^Δ: if the process is in the cemetery state, the only possible transition is to remain in the cemetery state. Otherwise, M_n^Δ behaves exactly like M_n. This step is referred to as* exploration.

2. *For $\mathcal{G}_n^\Delta$: if X_n is not in the state Δ (i.e. it is alive), then after one step under $\mathcal{G}_n^\Delta$ it remains alive and equal to X_n with probability $G_n(X_n)$, and is killed (moves to the cemetery state) with probability $1 - G_n(X_n)$. Moreover,*

$$\int_{E_n^\Delta} \mathcal{G}_n^\Delta(x, dy) = \int_{E_n} \mathcal{G}_n^\Delta(x, dy) + \int_{\{\Delta\}} \mathcal{G}_n^\Delta(x, dy) = G_n(x) + (1 - G_n(x)) = 1.$$

Thus, $\mathcal{G}_n^\Delta$ is a genuine Markov kernel. This step is referred to as killing.

With this in mind, we define the following transition kernels:

$$Q_{n+1}^\Delta = \mathcal{G}_n^\Delta M_{n+1}^\Delta.$$

Let $\mu \in \mathcal{P}(E_0)$ be given, and define $\mathbf{P}_\mu^\Delta$ to be the probability measure on the canonical path space whose finite-dimensional distributions correspond to those of a Markov chain with initial law μ and transition kernels Q^Δ up to time n.

Remark 2.9. *We may think of the chain $(X_n)_{n\geq 0}$ under the law $\mathbf{P}_\mu^\Delta$ as evolving in two steps:*

$$X_n \xrightarrow{\text{killing}} \widehat{X}_n \xrightarrow{\text{exploration}} X_{n+1}.$$

This can be interpreted as a Markov process with transitions M evolving in an absorbing medium, where the potentials G determine the absorption rate *at each point.*

Proposition 2.10. *The updated Feynman–Kac model $\widehat{\mathbf{Q}}_{\mu,n}$ associated with (G, M) represents the conditional law of the killed process X given that it is alive at time n.*

Proof. Let T denote the killing time of the process $(X_n)_{n\geq 0}$. Then

$$\begin{aligned}
\mathbf{P}_\mu^\Delta(T > n) &= \mathbf{P}_\mu^\Delta(X_0 \in E_0, X_1 \in E_1, \ldots, X_n \in E_n) \\
&= \int_{E_0 \times \cdots \times E_n} \mu(dx_0) Q_1^\Delta(x_0, dx_1) \cdots Q_n^\Delta(x_{n-1}, dx_n) \\
&= \int_{E_0 \times \cdots \times E_n} \mu(dx_0) G_0(x_0) M_1(x_0, dx_1) G_1(x_1) \cdots \\
&= \mathbf{E}_\mu \left[\prod_{k \leq n} G_k(X_k) \right].
\end{aligned}$$

This shows that $\widehat{\gamma}_n(1) = \mathbf{P}_\mu^\Delta(T > n)$. By a similar argument, we also obtain

$$\widehat{\gamma}_n(f) = \mathbf{E}_\mu^\Delta \left[f(X_n) \mathbf{1}_{\{T>n\}} \right].$$

□

Remark 2.11. *We refer to the sets $G_n^{-1}(\{0\})$ and $G_n^{-1}((0,1))$ as the* hard *and* soft *obstacles at time n, respectively. Note also that*

$$\eta_n(f) = \mathbf{E}_\mu^\Delta[f(X_n) \mid T \geq n], \qquad \widehat{\eta}_n(f) = \mathbf{E}_\mu^\Delta[f(X_n) \mid T > n],$$

where $\mathbf{E}_\mu^\Delta$ denotes expectation with respect to the law under which $(X_n)_n$ is the killed Markov chain described above.

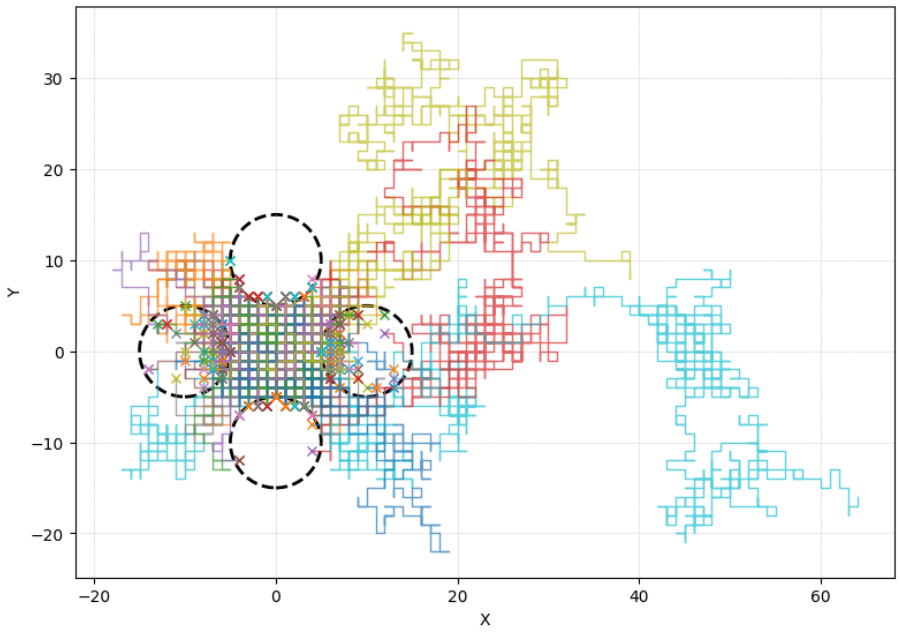


Figure 3: Simulation of 500 simple random walks over 1000 steps with two hard obstacles and two soft obstacles.

2.3.2 Interacting Process Interpretation

In the previous section we say how the distribution flows $(\eta_t)_{t\geq 0}$ could be seen as the law of a Markov chain at time t conditioned to survive some killing procedure. If this were to be simulated on a computer, this interpretation would have an advantage of the simulated particles being independent from each other (we will come to the bad downsides later). In this section we present an alternative particle interpretation for the distribution flows. This discussion corresponds to [3, Section 2.5.2], since its mostly for completeness, the reader may choose to skip directly to Section 3.

Proposition 2.12. *Let $S_{n,\eta}(x, dy)$ be defined via*

$$S_{n,\eta}(x, dy) = G_n(x)\delta_x(dy) + (1 - G_n(x))\psi_n(\eta)(dy),$$

where, recall, that ψ_n is the n^{th} Boltzmann-Gibbs transformation associated to potentials G. Then, if we set $K_{n+1,\eta} = S_{n,\eta}M_{n+1}$, we have that:

1. *$K_{n+1,\eta}$ are probability kernels $E_n \times \mathcal{E}_{n+1} \to [0, 1]$.*
2. *The distribution flow $(\eta_n)_{n\geq 0}$ satisfies the relation*

$$\eta_{n+1} = \eta_n K_{n+1,\eta_n}.$$

Proof. The proof that for any $x \in E_n$ we have that $K_{n+1,\eta}(x, E_{n+1}) = 1$ is a simple calculation which we skip. We will however, show that $\eta_{n+1} = \eta_n K_{n+1,\eta_n}$. Recall from previous discussions that we have already proved that η_{n+1} is equal to $\psi_n(\eta_n)M_{n+1}$. It thus suffices to show that $\eta_n S_{n,\eta_n}$ is indeed equal to $\psi_n(\eta_n)$.

Showing this is just a calculation too. Let $f \in \mathcal{B}_b(E_n)$, then

$$\begin{aligned}(\eta_n S_{n,\eta_n})(f) &= \int_{E_n \times E_n} \eta_n(dx)\left[G_n(x)\delta_x(dy) + (1-G_n(x))\psi_n(\eta_n)(dy)\right] f(y) \\ &= \int_{E_n \times E_n} \eta_n(dx) G_n(x)\delta_x(dy) f(y) + \int_{E_n \times E_n} \eta_n(dx)(1-G_n(x))\psi_n(\eta_n)(dy) f(y) \\ &= \int_{E_n} \eta_n(dx) G_n(x) f(x) + \psi_n(\eta_n)(f)\left(1-\eta_n(G_n)\right) \\ &= \eta_n(G_n f) + \frac{\eta_n(G_n f)}{\eta_n(G_n)}(1-\eta_n(G_n)) = \psi_n(f)\end{aligned}$$

□

What is the particle interpretation of this result? It all boils down to a choice of a wise probability measure $\mathbf{K}_{\eta_0}$ on the canonical probability space:

Definition 2.13. *Let $\left(\prod_{n\geq 0} E_n, \prod_{n\geq 0} \mathcal{E}_n, (X_n)_{n\geq 0}\right)$, be the canonical space, and $(K_{n+1,\mu_n})_{\mu_n \in \mathcal{P}(E_n), n\geq 0}$ be the family of probability kernels described above. For an initial measure $\eta_0 \in \mathcal{P}(E_n)$, let $(\eta_n)_{n\geq 0}$, satisfy the recurrence relation $\eta_{n+1} = \eta_n K_{n+1,\eta_n}$ (i.e: the FK distribution flows). Construct with these a measure $\mathbf{K}_0$ on the canonical space under which $(X_n)_{n\geq 0}$ is a Markov Chain with transition kernels $(K_{n+1,\eta+n})$, that is to say:*

$$\mathbf{K}_{\eta_0}((X_0, \cdots, X_n) \in (dx_0, \cdots, dx_n)) = \eta_0(dx_0) K_{1,\eta_0}(x_0, dx_1) \cdots K_{n,\eta_{n-1}}(x_{n-1}, dx_n).$$

The measure $\mathbf{K}_{\eta_0}$ is called the McKean measure *associated to the kernels $(K_{n+1,\mu_n})_{\mu_n \in \mathcal{P}(E_n), n\geq 0}$.*

Remark 2.14. *Markov Chains for which the transition kernels not only depend on current position but also on the current distribution are called* non-linear *Markov Chains. The reason for this is the following. Say we have a transition kernel M_{n+1}. Then we know that $\mathcal{L}(X_{n+1}) = \mathcal{L}(X_n)M_{n+1}$, so we can think of M_{n+1} as an operator. This operator is in fact linear! For if μ and ν are two measures on $(E_n, \mathcal{E}_n)$, by the linearity of the integral we have that $(\mu+\nu)M_{n+1} = \mu M_{n+1} + \nu M_{n+1}$. Now, for the case where the kernels M depend on the current distribution, then we do not in general have $(\mu+\nu)M_{\nu+\mu} = \mu M_\mu + \nu M_\nu$. This is where the term* non-linear *originates from.*

Proposition 2.15. *Let $(X_n)_{n\geq 0}$ have law $\mathbf{K}_{\eta_0}$ as described above, then $\mathcal{L}(X_n) = \eta_n$.*

Proof. Let $f \in \mathcal{B}_b(E_n)$, and denote by $\bar{\mathbf{E}}_{\eta_0}$ the expectation with respect to the measure $\mathbf{K}_{\eta_0}$. Then

$$\begin{aligned}\mathcal{L}(X_n)(f) &= \bar{\mathbf{E}}_{\eta_0}[f(X_n)] \\ &= \int_{E_0 \times \cdots \times E_n} f(x_n)\eta_0(dx_0)K_{1,\eta_0}(x_0,dx_1)\cdots K_{n,\eta_{n-1}}(x_{n-1},dx_n) \\ &= \int_{E_1 \times \cdots \times E_n} f(x_n) K_{2,\eta_1}(x_1,dx_2)\cdots K_{n,\eta_{n-1}}(x_{n-1},dx_n) \underbrace{\int_{E_0} \eta_0(dx_0)K_{1,\eta_0}(x_0,dx_1)}_{\eta_1(dx_1)} \\ &= \cdots = \int_{E_n} f(x_n)\eta_n(dx_n)\end{aligned}$$

□

Remark 2.16. *We now have established that the non-linear Markov Chain with dynamics prescribed by $\mathbf{K}_{\eta_0}$ has law at time n equal to η_n. But what is the particle interpretation behind this? Well, once again, we can look at the transition kernel K_{n+1,η_n} and notice that it involves a two-step transition. First we apply S_{n,η_n} and then we explore according to the original dynamics M_{n+1}, so let's think of what S is doing. Recall that*

$$S_{n,\eta}(x, dy) = G_n(x)\delta_x(dy) + (1 - G_n(x))\psi_n(\eta)(dy).$$

In the particle interpretation, we can call applying S to the measure as performing an "interacting jump" on the particle side: if the particle is currently at $x \in E_n$, then with probability $G_n(x)$, remain at x (this is given by the $\delta_x(dy)$ term); and with probability $1 - G_n(x)$, resample the particle's location according to the measure $\psi_n(\eta_n)$, which depends on the "cloud of particles" η_n. This is where the name interacting particle interpretation comes from. In summary: the dynamics under $\mathbf{K}_{\eta_0}$ may be thought of as:

$$X_n \overset{\text{interacting jump}}{\longrightarrow} \widehat{X}_n \overset{\text{exploring}}{\longrightarrow} X_{n+1}.$$

Where:

1. *$\widehat{X}_n = X_n$ with probability $G_n(X_n)$, and $\widehat{X}_n \sim \psi_n(\eta_n)$ with probability $1 - G_n(X_n)$.*
2. *$X_{n+1} \sim M_{n+1}(\widehat{X}_n, \cdot)$*

3 Particle Filtering

After providing a couple of alternative particle interpretations for the meaning of the FK flows, let's focus on the question of simulating from these measures, or computing integrals with respect to these measures. The killing interpretation gives us a straightforward (but naive) way of simulating from $\widehat{\eta}_n$. Simply: for a large fixed N, simulate N particle evolutions: $(X_t^i)_{t\leq n}$, for $i = 1, \cdots, N$ that evolve under the killing procedure described before. Our naive estimator for $\widehat{\eta}_n(f)$ would then be:

$$\frac{1}{\#\text{Surviving particles}} \sum_{i=1}^{N} f(X_n^i)\mathbf{1}\{X_n^i \text{ survived}\}. \tag{3.1}$$

The issue with this is that in many examples, the event of survival may be a *rare event*, so that for example its probability decays exponentially fast in n. What this means is that even if we start with a large number N of particles, the sum above may be taken with respect to an almost zero number of surviving particles. We will make this argument more precise in Section 5, but this will cause the estimator 3.1 to have a relative variance that grows exponentially, which in effect makes the approximation terrible.

However, we can be a bit more clever and exploit the structure of the measures $\widehat{\eta}_n$ to come up with a clever way of using an approximation to $\widehat{\eta}_n$ to give an approximation to $\widehat{\eta}_{n+1}$. This will be the idea of the Particle Filter, and the story begins by recalling a key result, namely that the measures satisfy a recursion

$$\widehat{\eta}_{n+1} = \psi_{n+1}\big(\widehat{\eta}_n M_{n+1}\big).$$

We will see how this recursion naturally motivates the so-called *Particle Filter*, an algorithm that constructs an approximation to $\widehat{\eta}_n$ through particle simulation. We will then study the convergence and stability properties of this algorithm.

3.1 Motivation of the Particle Filter

Suppose we have a particle approximation to $\widehat{\eta}_{n-1}$, defined by

$$\widehat{\eta}_{n-1}^N(dx) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_{n-1}^i}(dx),$$

Where N is some presumably large number of particles. Note that we are not assuming the particles X_{n-1}^i are independent. For the purposes of this motivation section, we simply assume that the empirical measure they generate approximates $\widehat{\eta}_{n-1}$ sufficiently well. It is then natural to assert that

$$\widehat{\eta}_n(dx) \approx \psi_n\big(\widehat{\eta}_{n-1}^N M_n\big). \tag{3.2}$$

We can push this approximation one step further. Observe that

$$\widehat{\eta}_{n-1}^N M_n(dx) = \frac{1}{N}\sum_{i=1}^{N} M_n(X_{n-1}^i, dx).$$

An unbiased way to approximate this mixture of measures is to sample $X_n^i \sim M_n(X_{n-1}^i, dx)$ and write

$$\widehat{\eta}_{n-1}^N M_n(dx) \approx \frac{1}{N}\sum_{i=1}^{N} \delta_{X_n^i}(dx).$$

Substituting this approximation into (3.2), we obtain

$$\begin{aligned}\widehat{\eta}_n &\approx \psi_n\left(\frac{1}{N}\sum_{i=1}^{N}\delta_{X_n^i}\right)\\ &= \sum_{i=1}^{N}\left(\frac{G_n(X_n^i)}{\sum_j G_n(X_n^j)}\right)\delta_{X_n^i} := \sum_{i=1}^{N} W_n^i \delta_{X_n^i}.\end{aligned}$$

Finally, we continue this approximation scheme as follows. Since this approximation to $\widehat{\eta}_n$ is itself a weighted mixture of point masses, we can produce an *unweighted* particle approximation by resampling: for each i, select one of the particles $X_n^1, \ldots, X_n^N$—say particle j—with probability W_n^j, and define $\widehat{X}_n^i$ equal to X_n^j. This yields the empirical measure

$$\frac{1}{N}\sum_{i=1}^{N}\delta_{\widehat{X}_n^i}.$$

Which we claim approximates $\widehat{\eta}_n$. In summary, starting with an empirical measure that approximates $\widehat{\eta}_{n-1}$, we have produced another empirical measure which approximates $\widehat{\eta}_n$. At this point, the procedure can be iterated indefinitely by repeating the steps described in this motivation section. In fact, this is precisely the *Particle Filtering* algorithm. What remains is to state it formally and to make the above approximations rigorous by showing that they are valid in the particle limit $N \to \infty$.

3.2 The Particle Filter

We now state the definition of the Particle Filter (PF) algorithm:

Definition 3.1 (Generic Particle Filter)**.** *A Particle Filter is given by the following algorithm:*

Require: *A Feynman-Kac model $\{\mu_0, (M_t), (G_t)\}$, a fixed number N of particles to simulate, a resampling scheme, and a finite time horizon T.*

1: **for** $n = 1$ **to** N **do**
2: *Sample* $X_0^n \sim \mu_0$.
3: *Compute* W_0^n
4: **end for**
5: **for** $t = 1$ **to** T **do**
6: **for** $n = 1$ **to** N **do**
7: *Choose ancestor:* $A_t^n \sim$ *Resample*$(W_{t-1}^{1:N})$
8: *Sample* $X_t^n \sim M_t(X_{t-1}^{A_t^n}, dx_t)$
9: **end for**
10: *Compute* $W_t^{1:N}$
11: **end for**

Remark 3.2. *There are several ways that the resampling step in line 7 could be implemented. The one we will focus on here is the so-called multinomial resampling scheme, which simply chooses $A_t^n = j$ with probability W_{t-1}^j, i.e: we choose the ancestor of particle n to be particle j with probability equal to the weight of particle X_{t-1}^j.*

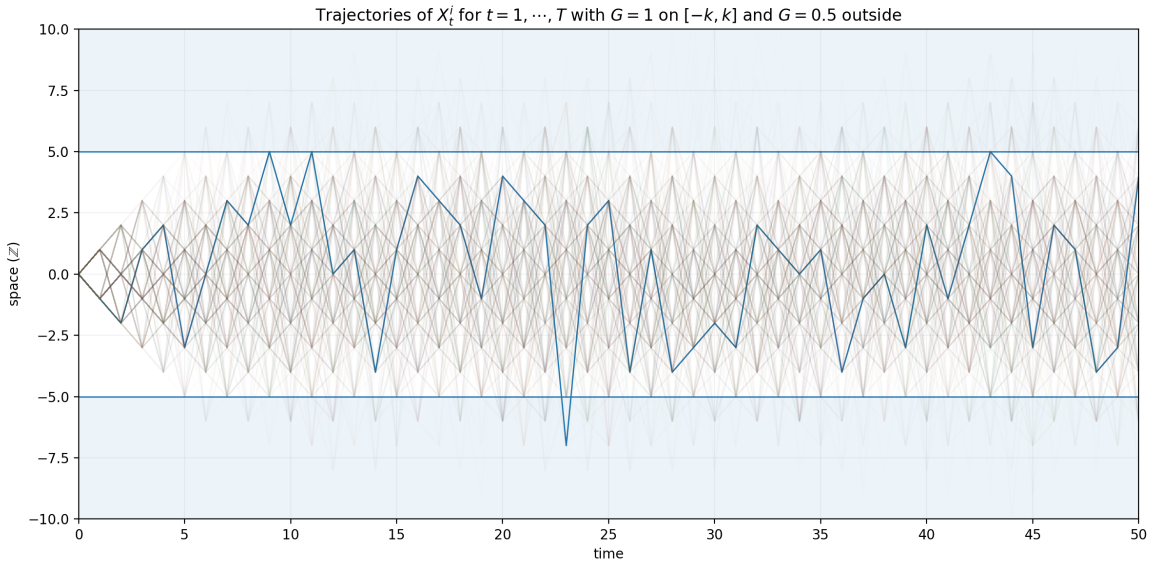


Figure 4: Trajectories of the Particle Filter algorithm described above (with multinomial resampling) with 500 particles. The reference measure for the Markov chain is that of a Simple Symmetric Random Walk on $\mathbf{Z}$, and the potentials G_t are all equal to $G(x) = \frac{1}{2}\left(1 + \mathbf{1}_{\{|x| \leq k\}}\right)$. One trajectory shown in full opacity, and the remaining trajectories are shown in reduced opacity. We see see that particles don't tend to spend much time in the regions of low potential, since the resampling step pushes particles towards areas of high potential.

In light of the motivation section, it is natural to interpret the final collection of particles $(X_T^n)_{n \leq N}$ as the output of the algorithm, from which we can consider objects such as

1. An approximation of $\widehat{\eta}_{T-1} M_T(dx)$ given by

$$\frac{1}{N} \sum_{n=1}^{N} \delta_{X_T^n}(dx). \tag{3.3}$$

2. An approximation of $\widehat{\eta}_T(dx)$ given by

$$\widehat{\eta}_T^N(dx) = \sum_{n=1}^{N} W_T^n \delta_{X_T^n}(dx). \tag{3.4}$$

Several questions of importance arise when looking at these estimates:

1. What kind of convergence results can we derive for $\widehat{\eta}_T^N(f)$ where $f \in \mathcal{B}_b(E_T)$? As we will see, it will turn out that $\widehat{\eta}_T^N(f) \to \widehat{\eta}_T(f)$ as $N \to \infty$ in $\mathcal{L}^2$ norm as well as in the almost sure sense.

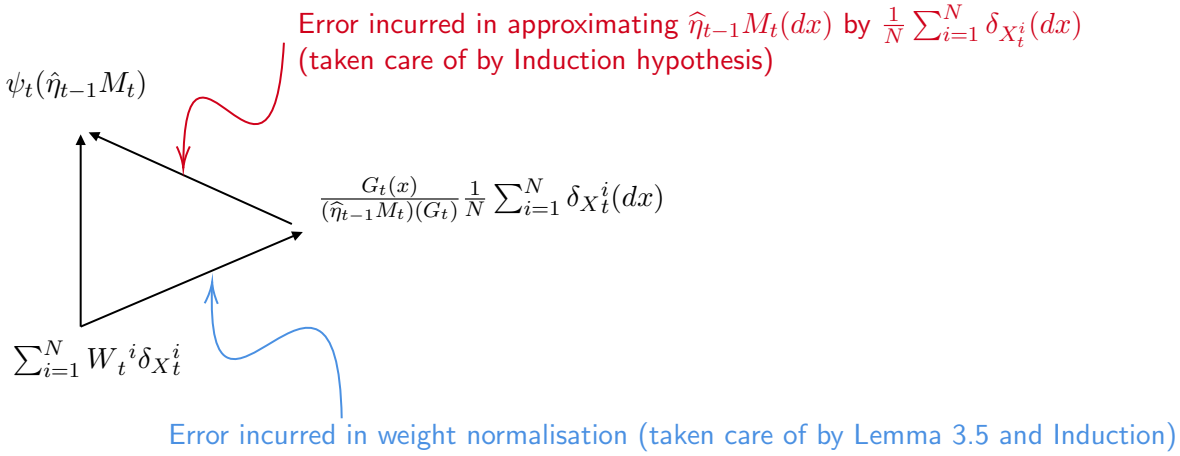
2. Can we say anything about how the error $\widehat{\eta}_T^N(f) - \widehat{\eta}(f)$ is distributed? As we will see, it will turn out to be Gaussian under appropriate rescaling.

3. What is the asymptotic distribution of the error of our approximation?

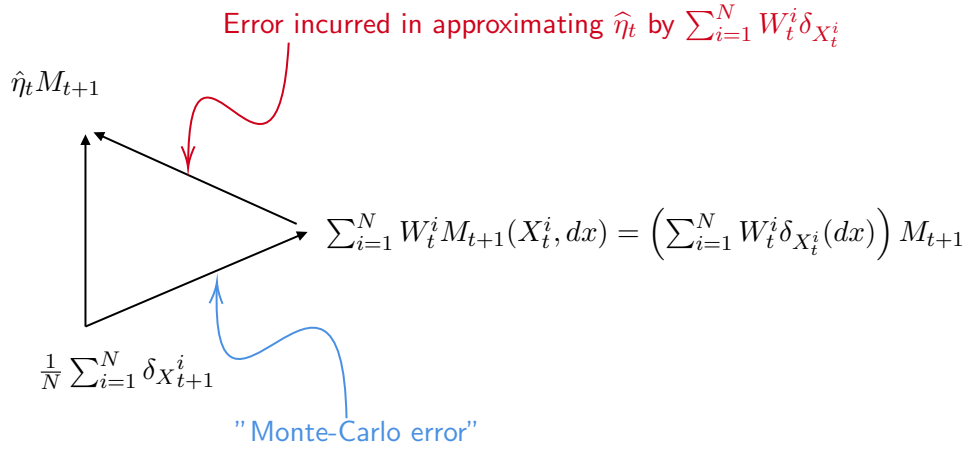
3.3 $\mathcal{L}^2$ convergence

While proving that $\widehat{\eta}_t^N(f) \to \widehat{\eta}_t(f)$ in $\mathcal{L}^2$ norm is not *technically difficult*, it is easy to get lost in the calculations without a clear understanding of why each step is performed. The argument is inductive in nature. Specifically, we first show that if Approximation 3.3 is accurate at time t, then Approximation 3.4 is also accurate at time t. We then show that if Approximation 3.4 is accurate at time t, then Approximation 3.3 is accurate at time $t+1$.

To clarify how these approximations are related, and how one leads to the other, it is helpful to think in terms of "triangles" in the space of measures. In the triangle below, we illustrate how to pass from Approximation 3.4 to the object it aims to approximate via two intermediate steps. The first step incurs error due to the approximation of the complicated weights (in particular the denominator). The second step incurs error from approximating $\widehat{\eta}_{t-1}M_t(dx)$ by the empirical measure $\frac{1}{N}\sum_{i=1}^N \delta_{X_t^i}(dx)$.



And on the triangle below, we illustrate how to pass from Approximation 3.3 to the object it tries to approximate: $\widehat{\eta}_t M_{t+1}$: the first error is incurred due to a "Monte-Carlo" error (called Monte-Carlo error since $\frac{1}{N}\sum_{i=1}^N \delta_{X_{t+1}^i}$ is quite literally a Monte-Carlo estimate of the measure $\left(\sum_{i=1}^N W_t^i \delta_{X_t^i}(dx)\right) M_{t+1}$) This will be taken care of by Lemma 3.4. The second error is due to approximating $\widehat{\eta}_t$ by $\sum_{i=1}^N W_t^i \delta_{X_t^i}$. This will be taken care of by the induction hypothesis.



Proof of $\mathcal{L}^2$ convergence

First of all, a quick remark: as hinted in the description of the algorithm (cf. Definition 3.1), there are several ways we could go about resampling the particles, but to go in line with the discussion in the motivation section, we will work in the case of multinomial resampling, i.e: $\mathbf{P}(\{A_t^n = m\}|\mathcal{F}_{t-1}) = W_{t-1}^m$, where $(\mathcal{F}_t)_{t\geq 0}$ is the filtration generated by the particle process $(X_t^i)_{i\leq N, t\geq 0}$. A key observation is that by conditioning on the ancestor:

$$\mathbf{P}\left(X_t^n \in dx|\mathcal{F}_{t-1}\right) = \sum_{m=1}^{N} \mathbf{P}\left(A_t^n = m|\mathcal{F}_{t-1}\right)\mathbf{P}\left(X_t^n \in dx|A_t^n = m\right) \tag{3.5}$$

$$= \sum_{m=1}^{N} W_t^m M_t(X_{t-1}^m, dx) \tag{3.6}$$

we obtain an expression for the conditional distribution of X_t^n given the history of the process up to time $t-1$. In particular, we see that conditional on $\mathcal{F}_{t-1}$, the particles at time t are i.i.d, with distribution equal to the one above. We will now start working towards proving the following: [1, Proposition 11.3].

Theorem 3.3 ($\mathcal{L}^2$ convergence)**.** *Suppose that the potential functions (G_t) of the Feynman-Kac model are bounded and strictly positive. Then, for any time t there are constants c_t and c_t' such that for any $f \in \mathcal{B}_b(E_t)$, we have*

$$\left\| \frac{1}{N}\sum_{n=1}^{N} f(X_t^n) - \widehat{\eta}_{t-1}M_t(f) \right\|_2^2 \leq c_t \frac{\|f\|_\infty^2}{N}, \tag{3.7}$$

(replacing $\widehat{\eta}_{-1}M_0$ by μ_0) and

$$\left\| \sum_{n=1}^{N} W_t^n f(X_t^n) - \widehat{\eta}_t(f) \right\|_2^2 \leq c_t' \frac{\|f\|_\infty^2}{N}, \tag{3.8}$$

As explained in the discussion above, to prove these statements we will need a pair of Lemmas:

Lemma 3.4 (Monte-Carlo approximation)**.** *Let f as above. Then*

$$\left\| \frac{1}{N}\sum_{n=1}^{N} f(X_t^n) - \sum_{n=1}^{N} W_{t-1}^n (M_t f)(X_{t-1}^n) \right\|_2^2 \leq \frac{1}{N}\|f\|_\infty^2 .$$

Proof. The proof simply relies on the observation we made in equation 3.6. Then, we observe that

$$\mathbf{E}\left[\frac{1}{N}\sum_{n=1}^{N} f(X_t^n)\middle|\mathcal{F}_{t-1}\right] = \mathbf{E}[f(X_t^n)|\mathcal{F}_{t-1}] = \int_{E_n} \sum_{m=1}^{N} W_{t-1}^m M_t(X_{t-1}^m, dx) f(x)$$

$$= \sum_{n=1}^{N} W_{t-1}^n (M_t f)(X_{t-1}^n)$$

Thus, we see that:

$$\left\| \frac{1}{N}\sum_{n=1}^{N} f(X_t^n) - \sum_{n=1}^{N} W_{t-1}^n (M_t f)(X_{t-1}^n) \right\|_2^2 = \mathbf{E}\left[\mathbf{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N} f(X_t^n) - \sum_{n=1}^{N} W_{t-1}^n (M_t f)(X_{t-1}^n) \right)^2 \middle| \mathcal{F}_{t-1} \right] \right] \tag{3.9}$$

$$= \mathbf{E}\left[\frac{1}{N} \mathsf{Var}(f(X_t^n)|\mathcal{F}_{t-1}) \right] \tag{3.10}$$

$$\leq \frac{1}{N} \|f\|_\infty . \tag{3.11}$$

where in this last inequality we have used the that $\mathsf{Var}(X) \leq \mathbf{E}X^2$. □

Now we need a second Lemma, which quantifies the error made by normalisation of the weights:

Lemma 3.5. *Define the normalised potentials $\bar{G}_t = \frac{G_t}{\ell_t}$, where $\ell_t = \widehat{\eta}_{t-1} M_t(G_t)$. Then*

$$\left\| \sum_{n=1}^{N} W_t^n f(X_t^n) - \frac{1}{N}\sum_{n=1}^{N} \bar{G}_t(X_t^n) f(X_t^n) \right\|_2^2 \leq \|f\|_\infty^2 \left\| \frac{1}{N}\sum_{n=1}^{N} \bar{G}_t(X_t^n) - 1 \right\|_2^2$$

Proof. The proof relies on the following simple observation. If we multiply and divide by ℓ_t we obtain:

$$\sum_{n=1}^{N} W_t^n f(X_t^n) = \frac{\sum_{n=1}^{N} \bar{G}_t(X_t^n) f(X_t^n)}{\sum_{n=1}^{N} \bar{G}_t(X_t^n)}, \tag{3.12}$$

and so we see that

$$\sum_{n=1}^{N} W_t^n f(X_t^n) - \frac{1}{N}\sum_{n=1}^{N} \bar{G}_t(X_t^n) f(X_t^n) = \sum_{n=1}^{N} W_t^n f(X_t^n) \left(1 - \frac{1}{N}\sum_{n=1}^{N} \bar{G}_t(X_t^n) \right)$$

$$\leq \|f\|_\infty \left(1 - \frac{1}{N}\sum_{n=1}^{N} \bar{G}_t(X_t^n) \right)$$

squaring and taking expectations gives the claim. □

Now we are ready to prove Theorem 3.3.

Proof of Theorem 3.3. The idea is as follows: we will show that the bounds hold by induction on t. We do so by showing that 3.7 at time t implies 3.8 at time t, and that 3.8 at time t implies 3.7 at time $t+1$. Notice that the base case holds due to a standard Monte-Carlo estimate for i.i.d random variables. Let's start.

We assume 3.7 holds at time t. Then

$$\sum_{n=1}^{N} W_t^n f(X_t^n) - \widehat{\eta}_t(f) = \sum_{n=1}^{N} W_t^n f(X_t^n) - \frac{1}{N}\sum_{n=1}^{N} \bar{G}_t(X_t^n) f(X_t^n) \tag{3.13}$$

$$+ \frac{1}{N}\sum_{n=1}^{N} \bar{G}_t(X_t^n) f(X_t^n) - \widehat{\eta}_t(f) \tag{3.14}$$

Since $\mathbf{E}[(X+Y)^2] \leq 2\left(\mathbf{E}X^2 + \mathbf{E}Y^2\right)$, we can take each of these terms separately. Note that term 3.13 can be taken care of by using Lemma 3.5. Then we note that

$$\|f\|_\infty^2 \left\| \frac{1}{N}\sum_{n=1}^N \bar{G}_t(X_t^n) - 1 \right\|_2^2 = \|f\|_\infty^2 \left\| \frac{1}{N}\sum_{n=1}^N \bar{G}_t(X_t^n) - \widehat{\eta}_{t-1}M_t(\bar{G}_t) \right\|_2^2,$$

and this can be taken care of by our assumption that inequality 3.7 holds at time t, with $f = \bar{G}_t$. In other words, we get that

$$\left\| \sum_{n=1}^N W_t^n f(X_t^n) - \frac{1}{N}\sum_{n=1}^N \bar{G}_t(X_t^n) f(X_t^n) \right\|_2^2 \leq \|f\|_\infty^2 \, c_t \frac{\left\|\bar{G}_t\right\|_\infty^2}{N} \tag{3.15}$$

Continuing, we now look at the second term 3.14. The trick is that $\widehat{\eta}_t(f) = \widehat{\eta}_{t-1}M_t(\bar{G}_t \times f)$, where $\bar{G}_t \times f$ is usual multiplication. Then we can apply the assumption that 3.7 holds at time t to the function $\bar{G}_t \times f$, and get that

$$\left\| \frac{1}{N}\sum_{n=1}^N \bar{G}_t(X_t^n) f(X_t^n) - \widehat{\eta}_t(f) \right\|_2^2 \leq c_t \frac{\left\|\bar{G}_t \times f\right\|_\infty^2}{N} \leq c_t \left\|\bar{G}_t\right\|_\infty^2 \frac{\|f\|_\infty^2}{N} \tag{3.16}$$

Thus combining $\mathbf{E}[(X+Y)^2] \leq 2\left(\mathbf{E}X^2 + \mathbf{E}Y^2\right)$ with bounds 3.15 and 3.16 gives that the bound 3.8 holds with $c_t' = 4c_t \left\|\bar{G}_t\right\|_\infty^2$. Now that we have proven that bound 3.7 at time t implies bound 3.8 at time t, we are going to show that bound 3.8 at time t implies 3.7 at time $t+1$. Then we will be done. In a similar way, we simply add and subtract an intermediate quantity to our target quantity so that we can apply a triangle-like inequality and estimate each term individually:

$$\frac{1}{N}\sum_{n=1}^N f(X_{t+1}^n) - \widehat{\eta}_t M_{t+1}(f) = \frac{1}{N}\sum_{n=1}^N f(X_{t+1}^n) - \sum_{n=1}^N W_t^n (M_{t+1}f)(X_t^n) \tag{3.17}$$

$$+ \sum_{n=1}^N W_t^n (M_{t+1}f)(X_t^n) - \widehat{\eta}_t M_{t+1}(f) \tag{3.18}$$

The term 3.17 we can bound using Lemma 3.4, which gives that

$$\left\| \frac{1}{N}\sum_{n=1}^N f(X_{t+1}^n) - \sum_{n=1}^N W_t^n (M_{t+1}f)(X_t^n) \right\|_2^2 \leq \frac{1}{N}\|f\|_\infty^2 \tag{3.19}$$

Then, term 3.18 can be bounded using bound 3.8 at time t with the function $M_{t+1}f$. Noting that $(\widehat{\eta}_t M_{t+1})(f) = \widehat{\eta}_t\left((M_{t+1}f)\right)$. Finally, its easy to see that $\|M_{t+1}f\|_\infty^2 \leq \|f\|_\infty^2$, and so we get

$$\left\| \sum_{n=1}^N W_t^n (M_{t+1}f)(X_t^n) - \widehat{\eta}_t M_{t+1}(f) \right\|_2^2 \leq c_t' \frac{\|f\|_\infty^2}{N}. \tag{3.20}$$

Finally, combining once again $\mathbf{E}[(X+Y)^2] \leq 2(\mathbf{E}X^2 + \mathbf{E}Y^2)$, with bounds 3.19 and 3.20, we get that

$$\left\| \frac{1}{N}\sum_{n=1}^N f(X_{t+1}^n) - \widehat{\eta}_t M_{t+1}(f) \right\|_2^2 \leq 2(1+c_t')\frac{\|f\|_\infty^2}{N},$$

which is bound 3.7 at time $t+1$, hence finishing the proof. □

3.4 Almost Sure Convergence

The idea for this proof is once again to perform a two-step induction proof, as well as the derivation of a fourth moment bound to replicate a step in the proof of the Strong Law of Large Numbers.

Theorem 3.6 (Almost Sure Convergence of Particle Filter)**.** *Assume the potentials (G_t) satisfy the same assumptions as before. Then:*

1. *For any $f \in \mathcal{B}_b(E_t)$, we have that for all t, almost surely as $N \to \infty$:*

$$\frac{1}{N}\sum_{n=1}^{N} f(X_t^n) \to \widehat{\eta}_{t-1}M_t(f) \tag{3.21}$$

   writing $\widehat{\eta}_{-1}M_0 := \mu_0$.

2. *For any f such that $G_t \times f \in \mathcal{B}_b(E_t)$, we have that for all t, almost surely as $N \to \infty$:*

$$\sum_{n=1}^{N} W_t^n f(X_t^n) \to \widehat{\eta}_t(f) \tag{3.22}$$

Remark 3.7. *Note that if $f \in \mathcal{B}_b(E_t)$, then $G_t \times f \in \mathcal{B}_b(E_t)$ because the potentials are bounded by assumption. Thus the second statement holds for a larger class of functions.*

Proof. It is clear that convergence statement 3.21 holds at $t = 0$, since X_0^n are i.i.d from μ_0, and so we apply the Strong Law of Large Numbers. Now assume that convergence statement 3.21 holds at time t, let us show that 3.22 holds at time t. This is simply due to the fact that we can write

$$\sum_{n=1}^{N} W_t^n f(X_t^n) = \frac{\frac{1}{N}\sum_{n=1}^{N} \bar{G}_t(X_t^n)f(X_t^n)}{\frac{1}{N}\sum_{n=1}^{N} \bar{G}_t(X_t^n)},$$

and both numerator and denominator converge: to $\widehat{\eta}_{t-1}M_t(\bar{G}_t \times f) = \widehat{\eta}_t(f)$ and $\widehat{\eta}_{t-1}M_t(\bar{G}_t) = 1$ respectively, using our assumption of 3.21 with the functions $\bar{G}_t \times f$ and $\bar{G}_t$ respectively. This shows that statement 3.22 holds at time t.

Now let's assume that statement 3.22 holds at time $t-1$, let's show that statement 3.21 holds at time t. The idea is to consider for $n = 1, \cdots, N$, the random variables

$$Z_t^n = f(X_t^n) - \sum_{n=1}^{N} W_{t-1}^n (M_t f)(X_{t-1}^n).$$

Recalling from equation 3.6 the shape of the distribution of the X_t^n's conditional on $\mathcal{F}_{t-1}$, it is easy to see that conditional on $\mathcal{F}_{t-1}$, the Z_t^n's are centred i.i.d random variables. Now, consider:

$$\mathbf{E}\left[\left(\sum_{n=1}^{N} Z_t^n\right)^4 \middle| \mathcal{F}_{t-1}\right] = \sum_{i,j,k,l} \mathbf{E}[Z_t^i Z_t^j Z_t^k Z_t^l \mid \mathcal{F}_{t-1}], \tag{3.23}$$

Sine conditional on $\mathcal{F}_{t-1}$, the Z_t^n's are i.i.d and centred, there won't be many surviving terms, only:

1. Those where $i = j = k = l$, of which there are N terms, and each of them will contribute an $\mathbf{E}[(Z_t^1)^4 \mid \mathcal{F}_{t-1}]$ to the sum.

2. Those where two indices are the same, and the other two indices are the same, but different. E.g: $i = j = 1$, and $k = l = 2$. Each of these terms will contribute an $\mathbf{E}[(Z_t^1)^2 \mid \mathcal{F}_{t-1}]^2$, and there will be

$$\binom{4}{2} \sum_{i=1}^{N} \sum_{j \neq i} 1 = 6N(N-1)$$

Effectively, 3.23 becomes:

$$\mathbf{E}\left[\left(\sum_{n=1}^{N} Z_t^n \right)^4 \middle| \mathcal{F}_{t-1} \right] = N\mathbf{E}[(Z_t^1)^4 \mid \mathcal{F}_{t-1}] + 6N(N-1)\mathbf{E}[(Z_t^1)^2 \mid \mathcal{F}_{t-1}]^2.$$

Now, since we are trying to prove statement 3.21, we can assume $f \in \mathcal{B}_b(E_t)$, and so each of these expectations is bounded by some absolute constant c, which means that

$$\mathbf{E}\left[\left(\sum_{n=1}^{N} Z_t^n \right)^4 \middle| \mathcal{F}_{t-1} \right] \leq cN^2,$$

so by taking expectations, we see that the usual fourth moment of $\sum_{n=1}^{N} Z_t^n$ also shares the same upper bound. Now we can replicate the proof strategy of the Strong Law of Large Numbers under assumption of bounded fourth moment. Namely: let $\epsilon > 0$ be given, then

$$\mathbf{P}\left(\left| \frac{1}{N} \sum_{n=1}^{N} Z_t^n \right| > \epsilon \right) \leq \frac{\mathbf{E}\left[\left(\sum_{n=1}^{N} Z_t^n \right)^4 \right]}{\epsilon^4 N^4} \leq \frac{cN^2}{\epsilon^4 N^4} = c' \frac{1}{N^2},$$

which is summable in N, and so by the Borel-Cantelli Lemma, almost surely:

$$\frac{1}{N} \sum_{n=1}^{N} Z_t^n \to 0.$$

However,

$$\begin{aligned} \frac{1}{N} \sum_{n=1}^{N} Z_t^n &= \frac{1}{N} \sum_{n=1}^{N} f(X_t^n) - \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{N} W_{t-1}^m (M_t f)(X_{t-1}^m) \\ &= \frac{1}{N} \sum_{n=1}^{N} f(X_t^n) - \sum_{n=1}^{N} W_{t-1}^n (M_t f)(X_{t-1}^n). \end{aligned}$$

And due to our assumption of 3.22 holding, we can choose the function to be $M_t f$ and we get that the second term on the last equality above, converges almost surely to $(\widehat{\eta}_{t-1} M_t)(f)$. In other words, we have shown that almost surely:

$$\frac{1}{N} \sum_{n=1}^{N} f(X_t^n) \to (\widehat{\eta}_{t-1} M_t)(f),$$

completing the proof. □

3.5 A Central Limit Theorem

In the previous sections, we studied how the unweighted (resp. weighted) empirical measures given by the particles $(X_t^i)_{i\leq n}$ are good approximations to $\widehat{\eta}_{t-1}M_t$ (resp. $\widehat{\eta}_t$) as the number of particles became large. Since these approximations are random, the errors themselves will also be random. We would now like to understand the asymptotic distribution of these errors. This will be given by a Central Limit Theorem for Particle Filters.

Theorem 3.8 (CLT for Particle Filters)**.** *Under the same assumptions as before (potentials are positive and upper bounded), we have that for any $f \in \mathcal{B}_b(E_t)$:*

$$\sqrt{N}\left(\frac{1}{N}\sum_{n\leq N} f(X_t^n) - \widehat{\eta}_{t-1}M_t(f)\right) \Rightarrow \mathcal{N}(0, \tilde{V}_t(f)) \tag{3.24}$$

and for any f such that $G_t \times f \in \mathcal{B}_b(E_t)$:

$$\sqrt{N}\left(\sum_{n\leq N} W_t^n f(X_t^n) - \widehat{\eta}_t(f)\right) \Rightarrow \mathcal{N}(0, V_t(f)) \tag{3.25}$$

where the variances are defined recursively by:

1. $\tilde{V}_0(f) = \mathit{Var}_{\mu_0}(f)$.
2. $V_t(f) = \tilde{V}_t\left(\bar{G}_t \times (f - \widehat{\eta}_t(f))\right)$ *for* $t \geq 0$.
3. $\tilde{V}_t(f) = V_{t-1}(M_t f) + \mathit{Var}_{\widehat{\eta}_{t-1}M_t}(f)$ *for* $t \geq 1$.

The proof strategy, as usual, will consist on this "two-step induction".

Proof. Note that at $t = 0$, using the convention $\widehat{\eta}_{t-1}M_t = \mu_0$, we get that statement 3.24 holds by the normal CLT. Now we will show that 3.24 at time t implies 3.25 at time t. This will just be an application of Slutsky's Theorem. Note that

$$\begin{aligned}\sqrt{N}\left(\sum_{n\leq N} W_t^n f(X_t^n) - \widehat{\eta}_t(f)\right) &= \frac{\sqrt{N}\left(\frac{1}{N}\sum_{n\leq N}\bar{G}_t(X_t^n)f(X_t^n) - \bar{G}_t(X_t^n)\widehat{\eta}_t(f)\right)}{\frac{1}{N}\sum_{n\leq N}\bar{G}_t(X_t^n)} \\ &= \frac{\sqrt{N}\left(\frac{1}{N}\sum_{n\leq N}\bar{G}_t(X_t^n)\bar{f}(X_t^n)\right)}{\frac{1}{N}\sum_{n\leq N}\bar{G}_t(X_t^n)}\end{aligned}$$

Then, using the assumption that $\bar{G}_t \times \bar{f} \in \mathcal{B}_b(E_t)$, 3.24 implies that the numerator converges to a $\mathcal{N}\left(0, \tilde{V}(\bar{G}_t \times (f - \widehat{\eta}_t(f))\right)$. The denominator converges almost surely to 1 by the almost sure convergence Theorem, and so the ratio converges to a $\mathcal{N}(0, V_t(f))$, showing that 3.25 holds at time t. Now we will show how 3.25 at time $t-1$ implies 3.24 at time t.

Let's first state the ingredients we'll need and then prove them individually.

If we define the following two quantities:

$$\Delta_1 = \sqrt{N}\left(\frac{1}{N}\sum_{n=1}^{N} f(X_t^n) - \sum_{n=1}^{N} W_{t-1}^n (M_t f)(X_{t-1}^n)\right),$$

and

$$\Delta_2 = \sqrt{N}\left(\sum_{n=1}^{N} W_{t-1}^n (M_t f)(X_{t-1}^n) - \widehat{\eta}_{t-1} M_t(f)\right),$$

then we quickly see that the characteristic function of our object of interest, i.e:

$$\sqrt{N}\left(\frac{1}{N}\sum_{n\leq N} f(X_t^n) - \widehat{\eta}_{t-1} M_t(f)\right),$$

is simply $\mathbf{E}[\exp\{iu(\Delta_1 + \Delta_2)\}]$. Since Δ_2 is $\mathcal{F}_{t-1}$ measurable, we can take it out of the conditional expectation and see that

$$\mathbf{E}[\exp\{iu(\Delta_1 + \Delta_2)\}] = \mathbf{E}\left[\exp\{iu\Delta_2\}\mathbf{E}[\exp\{iu\Delta_1\} \mid \mathcal{F}_{t-1}]\right].$$

Now, observe that Δ_2 is of the shape of statement 3.25 with the choice of function $M_t \times f$ but at $t-1$, and so we know that $\Delta_2 \Rightarrow Y$, where $Y \sim \mathcal{N}(0, V_{t-1}(M_t f))$. Now we state the main ingredient of the proof. Suppose we have shown that

$$\mathbf{E}[\exp\{iu\Delta_1\} \mid \mathcal{F}_{t-1}] \xrightarrow{\mathbf{P}} \exp\left(-\frac{\sigma^2 u^2}{2}\right) \tag{3.26}$$

where $\sigma^2 = \mathsf{Var}_{\widehat{\eta}_{t-1}M_t}(f)$. Then by Slutsky's Theorem, and the Continuous Mapping Theorem:

$$\exp\left(iu\Delta_2\right)\mathbf{E}[\exp\{iu\Delta_1\} \mid \mathcal{F}_{t-1}] \Rightarrow \exp\left(iuY - \frac{\sigma^2 u^2}{2}\right).$$

But now, since everything above is bounded, (in particular less than 1), we can take a continuous function that is equal to 1 on the unit disk and then decreases to zero, in conjunction to Portmanteaus Theorem, to assert that

$$\mathbf{E}\left[\exp\left(iu\Delta_2\right)\mathbf{E}[\exp\{iu\Delta_1\} \mid \mathcal{F}_{t-1}]\right] \to \mathbf{E}\left[\exp\left(iuY - \frac{\sigma^2 u^2}{2}\right)\right].$$

But now since Y was a Normally distributed random variable, we know how its characteristic function will look like, and so we see that the right hand side above will actually be $\exp\left(\frac{u^2(\tilde{V}_t(f))}{2}\right)$. In other words, the characteristic function of

$$\sqrt{N}\left(\frac{1}{N}\sum_{n\leq N} f(X_t^n) - \widehat{\eta}_{t-1} M_t(f)\right)$$

converges to the characteristic function of a $\mathcal{N}(0, \tilde{V}_t(f))$, so by Levy's Theorem we are done. Therefore all that's remaining to do is to show 3.26. We can start by noticing that Δ_1 is nothing but $\sqrt{N}\frac{1}{N}\sum_{n=1}^{N} Z_t^n = \frac{1}{\sqrt{N}}\sum_{n=1}^{N} Z_t^n$, where Z_t^n is as we defined previously:

$$Z_t^n = f(X_t^n) - \sum_{m=1}^{N} W_{t-1}^m (M_t f)(X_{t-1}^m),$$

and that these where conditionally i.i.d given $\mathcal{F}_{t-1}$. From this we notice that $\mathbf{E}[\exp(iu\Delta_1) \mid \mathcal{F}_{t-1}] = (\mathbf{E}[\exp\{iuN^{-\frac{1}{2}}Z_t^1\} \mid \mathcal{F}_{t-1}])^N$. Recall that if two complex numbers u, v have norm at most 1, we have the inequality $|u^N - v^N| \leq N|u-v|$. From this we gather that almost surely:

$$\left|\mathbf{E}\left[\exp\{iu\Delta_1\} \mid \mathcal{F}_{t-1}\right] - \exp\left(-\frac{\sigma^2u^2}{2}\right)\right| \leq N\left|\mathbf{E}\left[\exp\left\{iuN^{-\frac{1}{2}}Z_t^1\right\} \mid \mathcal{F}_{t-1}\right] - \exp\left(-\frac{\sigma^2u^2}{2N}\right)\right| \quad (3.27)$$

So now we'll look at the first term inside this absolute value. Recall from the Taylor expansion of e^x that

$$|\exp(ix) - 1 - ix + x^2/2| \leq |x|^3/6.$$

We may use this to see that

$$\mathbf{E}[\exp\{iuN^{-\frac{1}{2}}Z_t^1\} \mid \mathcal{F}_{t-1}] = 1 + iuN^{-\frac{1}{2}}\underbrace{\mathbf{E}[Z_t^1 \mid \mathcal{F}_{t-1}]}_{0} - \frac{u^2}{2N}\underbrace{\mathbf{E}[(Z_t^1)^2 \mid \mathcal{F}_{t-1}]}_{:=\sigma_N^2} + R_N \quad (3.28)$$

where $|R_N| \leq \frac{|u|^3}{6N^{3/2}}\mathbf{E}[|Z_t^1|^3 \mid \mathcal{F}_{t-1}]$. In other words:

$$N\left|\mathbf{E}\left[\exp\left\{iuN^{-\frac{1}{2}}Z_t^1\right\} \mid \mathcal{F}_{t-1}\right] - 1 + \frac{u^2}{2N}\sigma_N^2\right| \overset{a.s}{\to} 0 \quad (3.29)$$

To put this together with 3.27, we just need to bound

$$N\left|1 - \frac{u^2}{2N}\sigma_N^2 - \exp\left(-\frac{\sigma^2u^2}{2N}\right)\right| = N\left|\frac{u^2}{2N}\left(\sigma_N^2 - \sigma^2\right) + \mathcal{O}(N^{-2})\right| \quad (3.30)$$

So if we can show that $\sigma_N^2 \overset{\mathbf{P}}{\to} \sigma^2$, then combining 3.30 with 3.29 will give 3.26. Which will finish the proof. To show this, we write

$$\begin{aligned}
\sigma_N^2 &:= \mathbf{E}\left[(Z_t^1)^2 \mid \mathcal{F}_{t-1}\right] && (3.31)\\
&= \mathbf{E}\left[\left(f(X_t^n) - \sum_{n=1}^N W_{t-1}^n(M_tf)(X_{t-1}^n)\right)^2\middle|\mathcal{F}_{t-1}\right] && (3.32)\\
&= \mathbf{E}\left[f(X_t^n)^2 \mid \mathcal{F}_{t-1}\right] - \left(\sum_{n=1}^N W_{t-1}^n(M_tf)(X_{t-1}^n)\right)^2 && (3.33)\\
&= \underbrace{\sum_{n=1}^N W_{t-1}(M_tf^2)(X_{t-1}^n)}_{A_N} - \left(\underbrace{\sum_{n=1}^N W_{t-1}^n(M_tf)(X_{t-1}^n)}_{B_N}\right)^2 && (3.34)
\end{aligned}$$

Here step 3.33 comes from the fact that the second term is $\mathcal{F}_{t-1}$ measurable. The last step comes from the fact that we know the distribution of X_t^n conditional on $\mathcal{F}_{t-1}$. We are almost done now. From the assumption 3.25 at time $t-1$, we know that

$$\sqrt{N}(A_N - \widehat{\eta}_{t-1}(M_tf^2)) \Rightarrow \mathcal{N}(0, \cdots).$$

From Slutsky's Theorem, this means that in fact $A_N - \widehat{\eta}_{t-1}(M_tf^2) \Rightarrow 0$, but convergence in distribution to a constant implies convergence in probability, and so $A_N \overset{\mathbf{P}}{\to} \widehat{\eta}_{t-1}(M_t^2f) = \widehat{\eta}_{t-1}M_t(f^2)$. In a similar way, $B_N \overset{\mathbf{P}}{\to} \widehat{\eta}_{t-1}M_t(f)$. From this it follows that $\sigma_N^2 \overset{\mathbf{P}}{\to} \widehat{\eta}_{t-1}M_t(f^2) - (\widehat{\eta}_{t-1}M_t(f))^2 =: \sigma^2$. As required. $\square$

4 Stability of Particle Filters

In the previous sections we saw some Convergence Theorems related to Particle Filters (cf. Theorems 3.6, 3.3). The downside of these Theorems is that they work in the particle limit $N \to \infty$ and for a fixed time $t \geq 0$. In practice however, a Particle Filter will be implemented by fixing some $N \geq 1$ to be your number of particles, and then iterate the algorithm until some target time t. The CLT 3.8 told us that

$$\sqrt{N}\left(\widehat{\eta}_t^N(f) - \widehat{\eta}_t(f)\right) \Rightarrow \mathcal{N}(0, V_t(f)),$$

which informally says that $\widehat{\eta}_t^N(f) - \widehat{\eta}_t(f) \approx \mathcal{N}\left(0, \frac{V_t(f)}{N}\right)$. What is the issue here? Well, from the statement of the CLT we know that the variances $V_t(f)$ are defined recursively, and a priori it is not clear whether they behave "well" as $t \to \infty$. Indeed, if $V_t(f)$ is unbounded as $t \to \infty$, then our approximation actually keeps getting worse and worse as time progresses, given that we have a fixed number of particles. It is crucial to understand the growth of $V_t(f)$. Let's start with the following Lemma, which will give us an (albeit ugly) explicit representation of the variance.

Notation. *To avoid cluttering the document with parenthesis, for a measure μ and a function f, we will write μf instead of $\mu(f)$.*

Lemma 4.1 (Representation of the variances)**.** *The variances $V_t(f)$ as defined in the statement of Theorem 3.8 satisfy the following equation:*

$$V_t(f) = \sum_{s=0}^{t} (\widehat{\eta}_{s-1} M_s) \left[\left\{ \bar{G}_s R_{s+1:t}(f - \widehat{\eta}_t f) \right\}^2 \right]$$

where $R_{s+1:t} := R_{s+1} \circ R_{s+2} \circ \cdots \circ R_t(f)$, and $R_t(f) := M_t(\bar{G}_t \times f)$

Proof. The proof is by induction. The base case $t = 0$ holds by looking at the CLT (Theorem 3.8), and noting that $V_0(f) = \mathsf{Var}_{\mu_0}(\bar{G}_0(f - \widehat{\eta}_t f))$. Recall that we use the convention $\widehat{\eta}_{-1} M_0 := \mu_0$. Then note that $\mu_0[\bar{G}_0(f - \widehat{\eta}_0 f)] = 0$ since $\widehat{\eta}_0 := \bar{G}_0 \mu_0$, and so $\mathsf{Var}_{\mu_0}(\bar{G}_0(f - \widehat{\eta}_t f)) = \mu_0 \left[\left\{ \bar{G}_0(f - \widehat{\eta}_0 f) \right\}^2 \right]$, which is exactly what the claim entails for $t = 0$.

Assume now that the claim holds at time $t - 1$, let's show using the recursion from the CLT that the claim holds at time t. From the recursion, we have that

$$V_t(f) = V_{t-1}(\underbrace{M_t(\bar{G}_t \times (f - \widehat{\eta}_t f))}_{R_t(f - \widehat{\eta}_t f)}) + \mathsf{Var}_{\widehat{\eta}_{t-1} M_t}(\bar{G}_t \times (f - \widehat{\eta}_t f)) \tag{4.1}$$

Now we examine each term individually. From our Inductive Hypothesis, we see that:

$$V_{t-1}(R_t(f - \widehat{\eta}_t f)) = \sum_{s=0}^{t-1} (\widehat{\eta}_{s-1} M_s) \left[\left\{ \bar{G}_s R_{s+1:t-1}\left(R_t(f - \widehat{\eta}_t f) - \widehat{\eta}_{t-1}\left(R_t(f - \widehat{\eta}_t f)\right)\right) \right\}^2 \right] \tag{4.2}$$

Now observe something important: $\widehat{\eta}_{t-1}(R_t f) = (\widehat{\eta}_{t-1}M_t)(\bar{G}_t f) = \widehat{\eta}_t f$, therefore, the term $\widehat{\eta}_{t-1}\left(R_t(f - \widehat{\eta}_t f)\right)$ in 4.2 is zero, and so 4.2 actually becomes

$$V_{t-1}(R_t(f - \widehat{\eta}_t f)) = \sum_{s=0}^{t-1} (\widehat{\eta}_{s-1}M_s)\left[\left\{\bar{G}_s R_{s+1:t-1}\left(R_t(f - \widehat{\eta}_t f)\right)\right\}^2\right] \tag{4.3}$$

$$= \sum_{s=0}^{t-1} (\widehat{\eta}_{s-1}M_s)\left[\left\{\bar{G}_s R_{s+1:t}\left(f - \widehat{\eta}_t f\right)\right\}^2\right] \tag{4.4}$$

This is almost what we need. Let's look at the second term of 4.1. Since

$$\widehat{\eta}_{t-1}M_t(\bar{G}_t \times (f - \widehat{\eta}_t f)) = 0,$$

(this is how $\bar{G}_t$ is defined), we see that

$$\mathsf{Var}_{\widehat{\eta}_{t-1}M_t}(\bar{G}_t \times (f - \widehat{\eta}_t f)) = \widehat{\eta}_{t-1}M_t\left[\left\{\bar{G}_t(f - \widehat{\eta}_t)\right\}^2\right],$$

which combining with 4.4 finishes the claim. □

4.1 Strongly Mixing Kernels

Let's recall some important definitions.

Definition 4.2 (Total Variation distance)**.** *Let $\mathbf{P}$ and $\mathbf{Q}$ be two probability measures on some measurable space $(E, \mathcal{E})$ we define their total variation distance as*

$$\|\mathbf{P} - \mathbf{Q}\|_{TV} := \sup_{A \in \mathcal{E}} |\mathbf{P}(A) - \mathbf{Q}(A)|.$$

Remark 4.3. *There are several equivalent definitions of total variation distance. One that will be useful to us is the following:*

$$\|\mathbf{P} - \mathbf{Q}\|_{TV} = \sup_{\Delta f \leq 1} |\mathbf{P}(f) - \mathbf{Q}(f)|,$$

where $\Delta f = \sup_{x,y} |f(x) - f(y)|$ is the maximum variation of a function. If $\mathbf{P}$ and $\mathbf{Q}$ both admit densities $p(x)$ and $q(x)$ with respect to some common measure dx, it is also true that

$$\|\mathbf{P} - \mathbf{Q}\|_{TV} = \frac{1}{2}\int |p(x) - q(x)|dx.$$

Definition 4.4 (Contractivity Coefficient)**.** *Let $M : E_n \times \mathcal{E}_{n+1} \to [0, 1]$ be a Markov kernel. We define its Contractivity Coefficient ρ_M as*

$$\rho_M := \sup_{x,y} \|M(x, \cdot) - M(y, \cdot)\|_{TV}.$$

We call M strongly mixing if $\rho_M < 1$.

The contraction coefficient of a Markov kernel has the following functional interpretation

Lemma 4.5. *Let M be a Markov kernel with contraction coefficient ρ. Let $\varphi \in \mathcal{B}_b(E)$. Then*

$$\Delta(M\varphi) \leq \rho\Delta\varphi$$

Remark 4.6. *In a probabilistic sense, we can interpret this as follows: if a Markov chain $(X_n)_{n\geq 1}$ has transition kernels M, then the maximal oscillation of the function $\mathbf{E}_x[\varphi(X_1)]$, i.e: φ evaluated after one step of the chain, is smaller than the original maximal oscillation times a factor ρ.*

Proof. First we note that for constants $c \geq 0$, $\Delta(cf) = c\Delta(f)$. Thus

$$\Delta(M\varphi) = \Delta\varphi \times \Delta\left(M\left(\frac{\varphi}{\Delta\varphi}\right)\right).$$

Now note that

$$\begin{aligned}
\Delta\left(M\left(\frac{\varphi}{\Delta\varphi}\right)\right) &= \sup_{x,y}\left|M\left(x, \frac{\varphi}{\Delta\varphi}\right) - M\left(y, \frac{\varphi}{\Delta\varphi}\right)\right| \\
&\leq \sup_{x,y}\sup_{\Delta\phi\leq 1}|M(x,\phi) - M(y,\phi)| \\
&= \sup_{x,y}\|M(x,\cdot) - M(y,\cdot)\|_{\mathsf{TV}} \leq \rho
\end{aligned}$$

□

4.2 Asymptotic Stability of Variances

We will work under the following two assumptions:

Assumption 4.7 (Kernels)**.** *We assume that the Kernels (M_t) of our Markov Process X_t all admit a density $m_t(x_t|x_{t-1})$ with respect to some fixed measure dx and that there exists a constant $c_M \geq 1$ such that for all $t, x_t, x_{t-1}, x'_{t-1}$:*

$$\frac{m_t(x_t|x_{t-1})}{m_t(x_t|x'_{t-1})} \leq c_m$$

It turns out this assumption is sufficient to guarantee that the kernel is strongly mixing:

Proposition 4.8. *If Assumption 4.7 is fulfilled, then the kernels M_t are strongly mixing.*

Proof. We use the equivalent expression for TV distance

$$\left\|M_t(x_{t-1},\cdot) - M_t(x'_{t-1},\cdot)\right\|_{\mathsf{TV}} = \sup_A M_t(x_{t-1}, A) - M_t(x'_{t-1}, A)$$

and observe that

$$\begin{aligned}
M_t(x_{t-1}, A) - M_t(x'_{t-1}, A) &= \int_A \left(m_t(x_{t-1}, x) - m_t(x'_{t-1}, x)\right) dx \\
&\leq (1 - c_M^{-1}) \int_A m_t(x_{t-1}, x) dx \\
&\leq 1 - c_M^{-1}
\end{aligned}$$

where we've used the fact that the ratio of the densities is bounded below by $1/c_M$ □

Assumption 4.9 (Potentials)**.** *The potential functions (G_t) are uniformly bounded, i.e: there exists constants c_l and c_u such that for all t:*

$$0 < c_l \leq G_t \leq c_u.$$

And the result we will now work towards proving is the following:

Theorem 4.10 (Asymptotic Stability of Variance)**.** *Under assumptions 4.7 and 4.9 we have that for any $f \in \mathcal{B}_b(E_t)$, the variances $V_t(f)$ as defined in the CLT 3.8 are bounded uniformly in time.*

In order to progress with the proof of Theorem 4.10, we first need to make a detour and discuss how the Feynman-Kac measure $\mathbf{Q}_t$ can actually be viewed as a the law of a Markov process $(Z_0, \cdots, Z_t)$ up to the given time.

4.2.1 Feynman-Kac measure as a Markov measure

Proposition 4.11. *Let $\mathbf{Q}_t$ be a Feynman-Kac measure as defined earlier. $\mathbf{Q}_t$ is a Markov measure with initial law*

$$\mathbf{Q}_{0|t}(dx_0) = \frac{1}{L_t} H_{0:t}(x_0) G_0(x_0) \mathbf{M}_0(dx_0)$$

and transition kernels

$$Q_{s|t}(x_{s-1}, dx_s) = \frac{H_{s:t}(x_s)}{H_{s-1:t}(x_{s-1})} G_s(x_{s-1}, x_s) M_s(x_{s-1}, dx_s).$$

Where $H_{s:t}(x_s)$ is defined by

$$H_{s:t}(x_s) = \int_{\mathcal{X}^{t-s}} \prod_{i=s+1}^{t} G_i(x_{i-1}, x_i) M_i(x_{i-1}, dx_i), \qquad s < t$$

and $H_{t:t} = 1$.

Proof. We verify that $\mathbf{Q}_0|t(dx_0)$ is a probability measure. Indeed:

$$\int_{\mathcal{X}} \mathbf{Q}_{0|t}(dx_0) = \frac{1}{L_t} \int_{\mathcal{X}} \int_{\mathcal{X}^t} \left(\prod_{i=1}^{t} G_i(x_{i-1}, x_i) M_i(x_{i-1}, dx_i) \right) G_0(x_0) \mathbf{M}_0(dx_0) = \int_{\mathcal{X}^{t+1}} \mathbf{Q}_t(dx_{0:t}) = 1.$$

Let's verify that $Q_{s|t}(x_{s-1}, \cdot)$ is a probability measure. For this we need to note the following recursive fact:

$$\int_{\mathcal{X}} H_{s:t}(x_s) G_s(x_{s-1}, x_s) M_s(x_{s-1}, dx_s) =$$

$$\int_{\mathcal{X}} \int_{\mathcal{X}^{t-s}} \prod_{i=s+1}^{t} G_i(x_{i-1}, x_i) M_i(x_{i-1}, dx_i) G_s(x_{s-1}, x_s) M_s(x_{s-1}, dx_s) =$$

$$\int_{\mathcal{X}^{t-(s-1)}} \prod_{i=s}^{t} G_i(x_{i-1}, x_i) M_i(x_{i-1}, dx_i) = H_{s-1:t}(x_{s-1})$$

from this calculation we see that $Q_{s|t}(x_{s-1}, \cdot)$ integrates to 1. Finally, the fact that $\mathbf{Q}_t$ is the product of the initial measure and the Markov kernels follows from the telescoping of the product of all the $H_{s:t}$. □

Remark 4.12. *The kernels $Q_{s|t}$* **depend** *on t, hence why the notation emphasizes this dependence. Moreover, we can interpret the quantity $H_{s:t}(x_s)$ as the expected value of the product of all potentials from time $s+1$ up to t of the Markov process X started at time s at position x_s, which evolves with kernels $\{M_t\}$.*

The connection between this Markovian view of $\mathbf{Q}_t$ and the whole story of the variances is due to the following proposition. Recall from Lemma 4.1 that we were considering a sum of integrals of objects of the form $R_{s+1:t}(\varphi)$, where $R_{s+1:t} := R_{s+1} \circ \cdots \circ R_t$, and $R_t(\varphi) = M_t(\bar{G}_t \times \varphi)$

Proposition 4.13. *With $R_t, Q_{s|t}$ as defined before, and $\bar{H}_{s:t}$ defined as the $H_{s:t}$ above but with $\bar{G}_s$ instead of G_s, we have that*

$$R_{s+1:t}(\varphi) = \bar{H}_{s:t} \times Q_{s+1:t|t}(\varphi),$$

where $Q_{s+1:t|t}(\varphi) = Q_{s+1|t} \cdots Q_{t|t}(\varphi)$

Proof. Recall from the definition of the kernels $Q_{s|t}$ that

$$Q_{s|t}(x_{s-1}, dx_s) = \frac{H_{s:t}(x_s)}{H_{s-1:t}(x_{s-1})} G_s(x_{s-1}, x_s) M_s(x_{s-1}, dx_s).$$

We can replace the H by the $\bar{H}$, as well as the G by the $\bar{G}$, so long as we also replace the definition of the initial law to drop the L_t. Then we see that

$$Q_{s:t|t}(\varphi)(x_{s-1}) = \frac{1}{\bar{H}_{s-1:t}(x_{s-1})} \int \prod_{i=s}^{t} \bar{G}_i(x_{i-1}, x_i) M_i(x_{i-1}, dx_i) \varphi(x_t)$$

rearranging gives that $\bar{H}_{s-1:t}(x_{s-1}) Q_{s:t|t}(\varphi)(x_{s-1}) = R_{s:t}(\varphi)(x_{s-1})$ as required. □

We need one last fact about the Markov process that gives rise to $\mathbf{Q}_t$ before moving on with all the ingredients of the proof.

Proposition 4.14. *Let assumptions 4.7 and 4.9 hold, then the Markov Process defined by having law $\mathbf{Q}_t$ is strongly mixing.*

Proof. Since we are assuming that the kernels M_t admit densities $m_t(x_t|x_{t-1})$, we see that the kernels $Q_{s|t}$ admit a density $q_s(x_s|x_{s-1})$ given by

$$q_s(x_s|x_{s-1}) = \frac{H_{s:t}(x_s)}{H_{s-1:t}(x_{s-1})} G_s(x_{s-1}, x_s) m_s(x_s|x_{s-1}).$$

Observe how

$$\begin{aligned} H_{s:t}(x_s) &= \int G_{s+1}(x_s, x_{s+1}) m_{s+1}(x_{s+1}|x_s) H_{s+1}(x_{s+1}) dx_{s+1} \\ &\leq \frac{c_u}{c_l} c_M \int G_{s+1}(x'_s, x_{s+1}) m_{s+1}(x_{s+1}|x'_s) H_{s+1}(x_{s+1}) dx_{s+1} \\ &= \frac{c_u}{c_l} c_M H_{s:t}(x'_s). \end{aligned}$$

and so

$$\frac{q_s(x_s|x_{s-1})}{q_s(x_s|x'_{s-1})} = \frac{H_{s-1}(x'_{s-1})}{H_{s-1}(x_{s-1})} \frac{m_s(x_s|x_{s-1})}{m_s(x_s|x'_{s-1})} \leq c_G c_M^2$$

where $c_G = \frac{c_u}{c_l}$. Therefore, by proposition 4.8, we are done. □

We may now progress with the ingredients of the proof. First, a couple technical lemmas:

Lemma 4.15. *Let ψ and φ be two continuous and bounded functions, with $\psi \geq 0$ and φ satisfy $\sup \varphi \geq 0$ and $\inf \varphi \leq 0$. Then*

$$\Delta(\psi\varphi) \leq \|\psi\|_\infty \Delta(\varphi)$$

Proof. Recalling from the definition that $\Delta(f)$ is $\sup_{x,y} |f(x) - f(y)| = \sup_x f(x) - \inf_y f(y)$. It is clear that $\sup_x \psi(x)\varphi(x) \leq \|\psi\|_\infty \sup_x \varphi(x)$. Now notice, that since $\psi \geq 0$, $\inf \psi\varphi \geq \|\psi\|_\infty \inf \varphi$. Indeed: If $\inf \varphi = 0$, then since $\psi \geq 0$, $\inf \varphi\psi = 0$. Otherwise, once again using non-negativity of ψ and the assumption that $\inf \varphi$ is now strictly less than zero, it will follow that $\inf \psi\varphi \geq \|\psi\|_\infty \inf \varphi$. This finishes the claim. □

The next Lemma just says that any value of a bounded function φ lies within $\pm\Delta\varphi$ of its average.

Lemma 4.16. *Let $\varphi \in C_b(\mathcal{X})$ and assume there is a measure $\mathbf{P}$ for which $\mathbf{P}(\varphi) = 1$. Then*

$$\|\varphi\|_\infty \leq 1 + \Delta\varphi.$$

Proof. For all x, x', we have that $|\varphi(x) - \varphi(x')| \leq \Delta\varphi$, and so

$$\varphi(x') - \Delta\varphi \leq \varphi(x) \leq \varphi(x') + \Delta\varphi.$$

Now take expectation with respect to $X' \sim \mathbf{P}$ and we see that $1 - \Delta\varphi \leq \varphi(x) \leq 1 + \Delta\varphi$. This gives the claim since $-(1 + \Delta\varphi) \leq 1 - \Delta\varphi$. □

Remark 4.17. *If there is a measure for which the integral is actually zero, then the lemma really does say that the infinity norm is bounded above by the total variation of the function.*

The final technical Lemma of the proof is the following (cf. [1, Page 183])

Lemma 4.18 (Bound on $\left\|\bar{H}_{s:t}\right\|_\infty$)**.** *We have that*

$$\left\|\bar{H}_{s:t}\right\|_\infty \leq \prod_{i=1}^{t-s} \left(1 + \rho_M \rho_Q^{i-1} c_G\right)$$

Now we are ready to state and prove the main Theorem of the section.

Proof of Theorem 4.10. We have the following:

$$\begin{aligned}
\Delta(R_{s+1:t}(\varphi)) &= \Delta(\bar{H}_{s:t} \times Q_{s+1:t|t}(\varphi)) \\
&\leq \left\|\bar{H}_{s:t}\right\|_\infty \Delta(Q_{s+1:t|t}(\varphi)) \\
&\leq \left\|\bar{H}_{s:t}\right\|_\infty \rho_Q^{t-s} \Delta(\varphi)
\end{aligned}$$

where the first equality is due to Proposition 4.13, the middle inequality is due to Lemma 4.15, and the final inequality is due to iterated application of Lemma 4.5. Now we will use this as well as the bound from Lemma 4.18 in the result we obtained in Lemma 4.1 to show the Theorem. Recall that

$$V_t(f) = \sum_{s=0}^{t} (\widehat{\eta}_{s-1} M_s) \left[\left\{ \bar{G}_s R_{s+1:t}(f - \widehat{\eta}_t f) \right\}^2 \right], \tag{4.5}$$

and notice that since $\widehat{\eta}_s(R_{s+1}(f)) := \widehat{\eta}_s(M_{s+1}\bar{G}_{s+1}f) = \widehat{\eta}_{s+1}(f)$, we have that $\widehat{\eta}_s\left(R_{s+1:t}(f - \widehat{\eta}_t f)\right) = \widehat{\eta}_t(f - \widehat{\eta}_t f) = 0$, and so there is a probability measure (i.e: $\widehat{\eta}_s$) under which $R_{s+1:t}(f - \widehat{\eta}_t f)$ has zero mean, whence it follows (replicating the proof of Lemma 4.16 with a zero instead of a one) that $\|R_{s+1:t}(f - \widehat{\eta}_t f)\|_\infty \leq \Delta R_{s+1:t}(f - \widehat{\eta}_t f)$. Now using the calculation above, it follows that in fact

$$\|R_{s+1:t}(f - \widehat{\eta}_t f)\|_\infty \leq \prod_{i=1}^{t-s}\left(1 + \rho_M \rho_Q^{i-1} c_G\right)\rho_Q^{t-s}\Delta f$$

Plugging this into 4.5, by taking out $\|R_{s+1:t}(f - \widehat{\eta}_t f)\|_\infty^2$ from the integral, we get that

$$\begin{aligned} V_t(f) &\leq \sum_{s=0}^{t}\left(\prod_{i=1}^{t-s}\left(1 + \rho_M \rho_Q^{i-1} c_G\right)\rho_Q^{t-s}\Delta f\right)^2 (\widehat{\eta}_{s-1}M_s)(\bar{G}_s^2) \\ &\leq (\Delta f)^2 c_G^2 \sum_{s=0}^{t}\prod_{i=1}^{t-s}\left(1 + \rho_M \rho_Q^{i-1} c_G\right)^2 \rho_Q^{2(t-s)} \end{aligned}$$

where the last inequality follows from the fact that $\bar{G}_t = \frac{G_t}{\ell_t} \leq \frac{c_u}{\ell_t}$ and $\ell_t = \widehat{\eta}_{t-1}M_t(G_t) \geq c_l$. Now we can just complete the proof by noticing the following trick:

$$\begin{aligned} V_t(f) =& (\Delta f)^2 c_G^2 \sum_{s=0}^{t}\exp\left(2\sum_{i=1}^{t-s}\log(1 + \rho_M \rho_Q^{i-1} c_G)\right)\rho_Q^{2(t-s)} \\ &\leq (\Delta f)^2 c_G^2 \sum_{s=0}^{t}\exp\left(2\rho_M c_G \sum_{i=1}^{t-s}\rho_Q^{i-1}\right)\rho_Q^{2(t-s)} \\ &\leq (\Delta f)^2 c_G^2 \exp\left(\frac{2\rho_M c_G}{1 - \rho_Q}\right) \times \frac{1}{1 - \rho_Q^2}. \end{aligned}$$

which is uniformly bounded in time. □

Summary ☕. *Let us summarise the previous two chapters on Particle Filtering:*

1. *We were interested in computing an integral of the form $\widehat{\eta}_t(f)$ for some $f \in \mathcal{B}_b(E_t)$, where $\widehat{\eta}_t$ is the updated Feynman-Kac model.*

2. *From Proposition 2.5, we know that the sequence of models $(\widehat{\eta}_t)_{t\geq 0}$ satisfy the recursion given by $\widehat{\eta}_{t+1} = \psi_{n+1}(\widehat{\eta}_t M_{t+1})$. As discussed in Section 3.1, this recursion motivated us to successively use approximations, so that a particle approximation of $\widehat{\eta}_t$ yields a particle approximation to $\widehat{\eta}_{t+1}$.*

3. *This intuition was formalised in the Particle Filter Algorithm (cf. Definition 3.1), whose inputs are the ingredients of an FK model, i.e: an initial law μ_0, some potentials $G = (G_t)_{t\geq 0}$ and some transition kernels $M = (M_t)_{t\geq 0}$, as well as a number N of particles to simulate. This algorithm outputs at time t a collection of particles $(X_t^i)_{i\leq N}$.*

4. *In the next two sections, we saw how these particles could be used to construct approximations to $\widehat{\eta}_t M_{t+1}$ and $\widehat{\eta}_{t+1}$. In particular, we saw that their integrals against test functions converges in $\mathcal{L}^2$ and the almost sure sense to the "correct integrals". We saw how the proofs had a "two-step recursion" nature, and in the proof of the $\mathcal{L}^2$ case, we saw how this proof structure reflected our intuition of how the errors propagate in the successive approximations.*

5. *We then saw how the error $\eta_t^N(f) - \eta_t(f)$ was approximately Gaussian for large N, with variance $\frac{V_t(f)}{N}$, where $V_t(f)$ was introduced through a recursive relation. This formalises our intuition, that for a larger number of particles and a fixed time horizon, the precision of our estimates will be better. However, the question remained whether as $t \to \infty$, $V_t(f)$ would remain bounded, for if it didn't, it would mean that as t grows, the approximations created by the particle filter would get worse and worse.*

6. *This last question was answered in the positive in Section 4, where we saw that under some mixing conditions on the kernels and some further regularity on the potentials, $V_t(f)$ was uniformly bounded in time.*

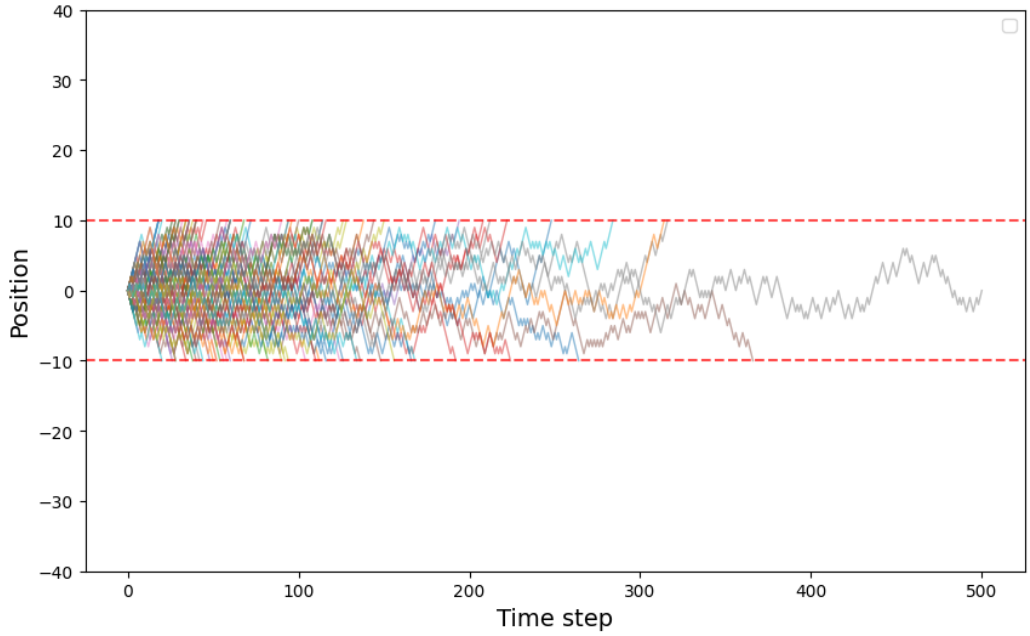


Figure 5: Simulation of 100 i.i.d copies of a SSRW killed when exiting the interval $[-10, 10]$ after 500 steps. From this realisation, only one particle has survived.

5 Variance Reduction by Changing Reference Measure

Notice that the Particle Filter could be thought of as a kind of mechanism for reducing variance in the simulation $\widehat{\eta}_n$. Indeed, we saw in Section 4 that under some mixing conditions on the kernels and some boundedness conditions on the potentials, we had that the variance of the error remained stable as time went through. In this last chapter, we will discuss an alternative way of reducing the variance of such approximations. The key idea will be to change the dynamics of our Markov chain to make the event of survival less rare. Throughout this chapter, we will continually use the example that we began in Section 1.2.

Example 5.1 (Continuation of Example 1.3)**.** Let $(X_t)_{t\geq 0}$ be a simple symmetric random walk (SSRW for short) on the integers, which one could think of as a (very simplistic model of a) radioactive particle moving in some medium. Let's imagine that at the levels $\pm k$ we have two "absorbing barriers", which cause a particle to die upon hitting them. The natural *rare event* to consider here is

$$A_t = \{|X_n| < k, \text{ for all } n \leq t\},$$

and we may be interested in quantities such as $\mathbf{P}_x(A_t)$, or $\mathbf{E}_x[f(X_t)|A_t]$, etc. Let's make the following observation however:

Lemma 5.2. *The probability of the event A_t decays exponentially in t.*

Sketch. Since the probability of jumping upwards is positive, we have that there exists constants $m \in \mathbf{N}$ and $\delta > 0$ such that for any $x \in \{-k+1, \cdots, k-1\}$,

$$\mathbf{P}_x(X_m > 2k) \geq \delta,$$

from which it follows that for any starting point x in "allowed interval", we have that $\mathbf{P}_x(A_m) < 1 - \delta$. Now we can repeatedly use the Markov Property and apply this same reasoning to conclude that

$$\mathbf{P}_x(A_{jm}) < (1-\delta)^j.$$

Finally, by writing any t as $jm + r$ for some j, r, we get the desired claim. □

Let's assume we try to estimate this through the naive approach described at the start of Section 3, i.e: by simulating n i.i.d paths $\left(X_0^{(i)}, X_1^{(i)}, \cdots, X_t^{(i)}\right)$ up to time t from a SSRW on $\mathbf{Z}$, and simply counting how many of them stay in the interval $I_k = \{-k+1, \cdots, k-1\}$. By the Strong Law of Large Numbers, we know that as $k \to \infty$, we will have almost sure convergence of our estimator

$$\widehat{p} = \frac{1}{n} \sum_{i \leq n} \mathbf{1}\left\{\left(X_0^{(i)}, X_1^{(i)}, \cdots, X_t^{(i)}\right) \in I_k\right\}$$

to our desired probability. However, let's have a closer look at what happens to the variance of our estimator at a fixed sample size n:

$$\begin{aligned}
\mathsf{Var}_0\left(\frac{1}{n}\sum_{i\leq n} \mathbf{1}\{(X_0^{(i)}, X_1^{(i)}, \cdots, X_t^{(i)}) \in I_k\}\right) &= \frac{1}{n^2}\sum_{i\leq n} \mathsf{Var}_0\left(\mathbf{1}\{(X_0^{(i)}, X_1^{(i)}, \cdots, X_t^{(i)})\right) \\
&= \frac{1}{n^2}\sum_{i\leq n} \mathbf{P}_0(\tau > t)(1 - \mathbf{P}_0(\tau > t)) \\
&= \frac{1}{n} p_0(t)\left(1 - p_0(t)\right)
\end{aligned}$$

where $p_0(t) = \mathbf{P}_0(\tau > t)$ for convenience. Naturally, this variance will be very small since $p_0(t)$ is also really small, but the key observation is that we should compare it to the real value we are trying to estimate (Indeed, if the variance of an estimator is of order 10^{-10} but the real value is of order 10^{-20}, we actually have an incredibly bad estimator!). Since for large t, we said that $p_0(t) \sim \exp(-ct)$, this relative variance actually blows up if t is large relative to n. Indeed, the relative variance is:

$$\frac{\sqrt{\frac{p_0(t)(1-p_0(t))}{n}}}{p_0(t)} = \frac{1}{\sqrt{np_0(t)}}\sqrt{(1-p_0(t))} \sim \frac{\exp(c't)}{\sqrt{n}}$$

In more precise words, as we increase t, we need exponentially many samples to maintain our error "stable". An intuitive explanation for this phenomenon is as follows: let's consider the problem of estimating $\widehat{\eta}_n(f) = \mathbf{E}[f(X_n)|A_n]$. The naive estimator would then be

$$\frac{1}{\#\text{Surviving Particles}} \sum_{\text{survivors}} f(X_n^{(i)}).$$

However, since the vast majority of paths are killed before time n, the estimator above effectively uses a vanishing number of samples. In other words, under the law of the SSRW, most trajectories contribute nothing to the expectation of interest, the law is "putting its mass in the wrong place". An intuitive way to fix this would be to change the reference measure so that a more meaningful fraction of these paths "reach the rare event". This is precisely what we will discuss in the next two sections. First, we will discuss some general technicalities, and then present a specific example.

5.1 Change of Reference Measure

Let us return to the general framework introduced earlier. We consider a canonical probability space

$$\left(\prod_{n\geq 0} E_n,\ \prod_{n\geq 0} \mathcal{E}_n,\ X,\ \mathbf{P}_\mu \right),$$

where $\mathbf{P}_\mu$ is a reference measure under which the canonical process X is a Markov chain with initial distribution μ and transition kernels $M_n : E_n \times \mathcal{E}_{n+1} \to [0,1]$.

Suppose now that we wish to work with a different set of dynamics. More precisely, assume we choose an alternative initial distribution $\bar{\mu}$ such that $\mu \ll \bar{\mu}$, and replace the transition kernels M_n by kernels $\bar{M}_n$ satisfying

$$M_n(x,\cdot) \ll \bar{M}_n(x,\cdot), \qquad \text{for all } n \text{ and } x \in E_n.$$

Equivalently, this corresponds to changing the reference measure from $\mathbf{P}_\mu$ to a new measure $\bar{\mathbf{P}}_{\bar{\mu}}$, under which X is a Markov chain with initial distribution $\bar{\mu}$ and transition kernels $\bar{M}_n$.

Remark 5.3. *In the context of the previous discussion, these new dynamics will be chosen so that "rare events become less rare" under this new measure.*

Given a collection of potentials G, a natural question is whether the Feynman–Kac model associated with $(G, \mathbf{P}_\mu)$ can be related to a new model $(\bar{G}, \bar{\mathbf{P}}_{\bar{\mu}})$ for a suitable choice of modified potentials $\bar{G}$. In other words: if we change the dynamics from $\mathbf{P}_\mu$ to $\bar{\mathbf{P}}_\mu$, can we still keep the expectations the same if we alter the potentials slightly? We quickly see from the assumptions above that the answer is positive:

Proposition 5.4. *Let $\mathbf{P}_\mu$ and $\bar{\mathbf{P}}_{\bar{\mu}}$ be the probability measures described above. Let G be a collection of potentials. Then defining new potentials*

$$\bar{G}(x_0,\cdots,x_n) = G(x_0,\cdots,x_n)\frac{dM_n(x_{n-1},\cdot)}{d\bar{M}_n(x_{n-1},\cdot)}(x_n), \quad \textit{where } \frac{dM_0}{d\bar{M}_0}(x_0) := \frac{d\mu}{d\bar{\mu}}(x_0),$$

we have that the FK models associated to $(G, \mathbf{P}_\mu)$ are equivalent to those associated to $(\bar{G}, \bar{\mathbf{P}}_{\bar{\mu}})$.

Proof. Note that we can write for a set $A_0 \times \cdots \times A_n \in \mathcal{E}_0 \times \cdots \times \mathcal{E}_n$:

$$\begin{aligned}
\mathbf{P}_{\mu,n}(A_0 \times \cdots \times A_n) &= \int_{A_0} \mu(dx_0) \int_{A_1} M_1(x_0, dx_1) \cdots \int_{A_n} M_n(x_{n-1}, dx_n) \\
&= \int_{A_0} \mu(dx_0) \int_{A_1} M_1(x_0, dx_1) \cdots \int_{A_n} \frac{dM_n(x_{n-1},\cdot)}{d\bar{M}_n(x_{n-1},\cdot)}(x_n) \bar{M}_n(x_{n-1}, dx_n)
\end{aligned}$$

unravelling all this, we get that

$$\mathbf{P}_{\mu,n}(A_0 \times \cdots \times A_n) = \mathbf{E}_{\bar{\mathbf{P}}_{\mu,n}}\left[\frac{d\mu}{d\bar{\mu}} \prod_{k\leq n} \frac{dM_k(x_{k-1},\cdot)}{d\bar{M}_k(x_{k-1},\cdot)} \mathbf{1}_A \right]$$

so in fact we see that

$$\frac{d\mathbf{P}_{\mu,n}}{d\bar{\mathbf{P}}_{\mu,n}}(x_0,\cdots,x_n) = \frac{d\mu}{d\bar{\mu}}(x_0) \prod_{k\leq n} \frac{dM_k(x_{k-1},\cdot)}{d\bar{M}_k(x_{k-1},\cdot)}(x_k).$$

Now the proof follows immediately. Indeed:

$$\begin{aligned}\widehat{\gamma}_n(f) &:= \mathbf{E}_{\mathbf{P}_\mu}\left[f(X_n)\prod_{k\leq n}G_k(X_k)\right]\\ &= \mathbf{E}_{\mathbf{P}_{\mu,n}}\left[f(X_n)\prod_{k\leq n}G_k(X_k)\right]\\ &= \mathbf{E}_{\bar{\mathbf{P}}_{\bar{\mu},n}}\left[f(X_n)\prod_{k\leq n}G_k(X_k)\frac{d\mathbf{P}_{\mu,n}}{d\bar{\mathbf{P}}_{\bar{\mu},n}}(X_0,\cdots,X_n)\right]\\ &= \mathbf{E}_{\bar{\mathbf{P}}_{\bar{\mu},n}}\left[f(X_n)\prod_{k\leq n}\bar{G}_k(X_k)\right].\end{aligned}$$

□

Remark 5.5. *The proof above shows that the law $\mathbf{P}_\mu$ is* locally *absolutely continuous with respect to $\bar{\mathbf{P}}_\mu$. This means that for any finite time horizon n, the n-dimensional distribution $\mathbf{P}_{\mu,n}$ is absolutely continuous with respect to $\bar{\mathbf{P}}_{\mu,n}$.*

A way of formalising the intuition that "variance can be reduced if the new dynamics make the rare event less rare" is through the following proposition, whose content corresponds to [3, Exercise 12.3.3].

Proposition 5.6. *Let μ be a probability measure on some measurable space $(\mathcal{X},\mathcal{B}(\mathcal{X}))$, f be a non-negative bounded function on $\mathcal{X}$, and $(X_i)_{i\leq N}$ be i.i.d samples from μ. Furthermore, let $\bar{\mu}$ be another probability measure such that $\bar{\mu}\gg\mu$, $\bar{f}=f\frac{d\mu}{d\bar{\mu}}$, and $(\bar{X}_i)_{i\leq N}$ be i.i.d samples from $\bar{\mu}$. Finally, let μ^N and $\bar{\mu}^N$ be the empirical measures obtained from each collection of particles respectively. Then*

$$\mathit{Var}\left(\bar{\mu}^N(\bar{f})\right)=\mathit{Var}\left(\mu^N(f)\right)-\frac{1}{N}\mu\left(f^2\left(1-\frac{d\mu}{d\bar{\mu}}\right)\right)$$

Proof. First of all, it is easy to verify that that $N\mathsf{Var}(\mu^N(f))=\mu((f-\mu(f))^2)=\mu(f^2)-\mu(f)^2$, as well as a similar expression for $\bar{\mu}$ and $\bar{f}$, as well as the fact that $\mathbf{E}[\bar{\mu}^N(\bar{f})]=\mu(f)$. From this we see that,

$$\begin{aligned}N\mathsf{Var}(\bar{\mu}^N(\bar{f})) &= \bar{\mu}((\bar{f}-\mu(f))^2)\\ &= \bar{\mu}(\bar{f}^2)-\mu(f)^2\\ &= \mu\left(f^2\times\frac{d\mu}{d\bar{\mu}}\right)-\mu(f)^2+\mu(f^2)-\mu(f^2)\\ &= N\mathsf{Var}(\mu^N(f))-\mu\left(f^2\left(1-\frac{d\mu}{d\bar{\mu}}\right)\right).\end{aligned}$$

□

Example 5.7. Let's specialise back our motivating example, where $\mathbf{P}_{x_0,n}$ is the law of a SSRW on $\mathbf{Z}$, i.e: $\mathbf{P}_n(x_0,\cdots,x_n)$ is the probability that a SSRW on $\mathbf{Z}$ up to time n takes the path $(x_0,\cdots,x_n)$. In light of Proposition 5.6, we can take $\mu=\mathbf{P}_{x_0,n}$ and $f(x_0,\cdots,x_n)=\prod_{i\leq n}\mathbf{1}\{x_i\in I_k\}$, where $I_k=\{-k+1,\cdots,k-1\}$. In this case, as we have seen, we may choose some different transition probabilities

$\bar{M}(x,y)$ which give rise to a different law $\bar{\mathbf{P}}_{x_0,n}$ on path space, and Proposition 5.6 tells us that by simulating from $\bar{\mathbf{P}}_{x_0,n}$ instead of $\mathbf{P}_{x_0,n}$, we can obtain a variance reduction so long as

$$\mathbf{P}_{x_0,n}\left(\mathbf{1}\{\text{path is alive}\}\left(1-\frac{d\mathbf{P}_{x_0,n}}{d\bar{\mathbf{P}}_{x_0,n}}\right)\right)>0.$$

This condition can be heuristically interpreted as: *on average* surviving paths are more like likely under $\bar{\mathbf{P}}_{x_0,n}$ than under $\mathbf{P}_{x_0,n}$. The corresponding interpretation on the general case where we perform killing with rates given by potentials $G(x)$ is precisely that variance reduction is achieved if paths that spend more time on areas of high potential are more likely under $\bar{\mathbf{P}}$ than $\mathbf{P}$.

5.2 Doob's Transform

To conclude the report, we present one way of constructing the new dynamics $\bar{M}$ which we have been discussing so far. We will then show a specific example and sketch an argument of why it produces a reduction in variance.

Definition 5.8. *Let M be a transition kernel from a space $E \to E$ (we assume homogeneity for simplicity), and let $U : E \to \mathbf{R}$ be a* positive *function, which we call a* weight function *on E. The kernel*

$$Q(x,dy)=\frac{U(y)}{U(x)}M(x,dy)$$

is called the Doob's h-transform of M with weight function U.

Remark 5.9. *The kernel $Q(x,dy)$ as defined above* is not necessarily a probability kernel, *i.e: $Q(x,E)$ might not be equal to 1 in general. It is quickly seen however, that $Q(x,dy)$ is a probability kernel if and only if U is harmonic for M, i.e:*

$$\int_E M(x,dy)U(y)=U(x).$$

The idea in the definition above is that U encodes the regions of space where we want our Markov chain to spend more time in. To account for the case where U is not harmonic, we can then define a new transition kernel,

$$\bar{M}(x,dy)=\frac{Q(x,dy)}{Q(x,S)}=\frac{1}{\int\frac{U(z)}{U(x)}M(x,dz)}\frac{U(y)}{U(x)}M(x,dy)$$

And with these new kernels, build the measure $\bar{\mathbf{P}}$. Let's say for simplicity that we start both chains at the same point x_0 of state-space, then in light of the calculations done in section 5.1, we know that

$$\frac{d\mathbf{P}_{x_0,n}}{d\bar{\mathbf{P}}_{x_0,n}}(x_0,\cdots,x_n)=\left(\prod_{k=0}^{n}Q(x_k,S)\right)\frac{U(x_0)}{U(x_n)} \tag{5.1}$$

In particular, we may now use this to relate expectations with respect to the measure $\mathbf{P}_{x_0}$ to expectations with respect to the measure $\bar{\mathbf{P}}_{x_0}$. Namely:

$$\mathbf{E}_{x_0}\left[f(X_n)\prod_{k\le n}G_k(X_k)\right]=\bar{\mathbf{E}}_{x_0}\left[f(X_n)\left(\prod_{k\le n}G_k(X_k)Q(X_k,S)\right)\frac{U(x_0)}{U(X_n)}\right], \tag{5.2}$$

where $\mathbf{E}_{x_0}$ and $\bar{\mathbf{E}}_{x_0}$ denote expectations with respect to $\mathbf{P}_0$ and $\bar{\mathbf{P}}_{x_0}$ respectively. Let's return to the example of the SSRW with killing. The reason why the Doob transform is useful for variance reduction comes precisely from the following Proposition.

Proposition 5.10. *Let $(X_n)_n$ be a Markov chain with transition kernel P on some state space E. Let $A \subset E$ be a set of "allowed" states. Let P_A be the restriction of P to A, i.e: $P_A : A \times A \to [0,1]$ given by $P_A(x,y) = P(x,y)$. Let (λ, U) be an eigenpair for P_A, i.e: $P_A U = \lambda U$. Then for $x \in A$:*

$$\lim_{n\to\infty} \mathbf{P}_x(X_1 = y | X_1, \cdots, X_n \in A) = \frac{1}{\lambda}\frac{U(y)}{U(x)}P(x,y)$$

For a proof, we refer the reader to [2, Section 3.2]. In other words, the Doob transform with weight function U as in the proposition above will give rise to the "survival" process, i.e: the original chain "conditioned on living forever". Let's now see precisely how much variance reduction is achieved by having performed this change of reference measure:

Example 5.11. Let's show explicitly an example the Doob transform discussed above. In the case of the SSRW on $\mathbf{Z}$, let P be its transition matrix and I_k the interval $\{-k+1, \cdots, k-1\}$. We are then interested in finding a function $U(x)$ and an eigenvalue λ such that

$$(P_{I_k} U)(x) = \frac{U(x+1) - U(x-1)}{2} = \lambda U(x)$$

with the boundary conditions $U(x) = 0$ for $x \in \{-k, k\}$. By recalling that $\cos(\theta) = \frac{1}{2}\left(\exp(i\theta) + \exp(-i\theta)\right)$, one quickly verifies that

$$\begin{cases} U(x) = \cos\left(\frac{\pi x}{2k}\right) & x \in I_k \\ \lambda = \cos\left(\frac{\pi}{2k}\right) \end{cases}$$

solves the system above. As such, the resulting Doob transform looks like

$$\bar{P}(x,y) = \frac{1}{\cos\left(\frac{\pi}{2k}\right)}\frac{\cos\left(\frac{\pi y}{2k}\right)}{\cos\left(\frac{\pi x}{2k}\right)}P(x,y).$$

Using these dynamics and all the theory discussed above, we see that for any $x \in I_k$:

$$\mathbf{E}_x[f(X_n)\mathbf{1}\{\text{particle is alive at time } n\}] = \lambda^n \bar{\mathbf{E}}_x\left[f(X_n)\frac{U(x)}{U(X_n)}\right].$$

With this expression we can now see how by simulating the right hand side expectation we can achieve a drastic reduction in variance. Indeed: say we try to simulate

$$\widehat{\gamma}_n^N(f) = \frac{1}{N}\sum_{i=1}^N f(X_n^i)\mathbf{1}\{X_n^i \text{ is alive}\},$$

where $(X_t^i)_{t\le n}$ were i.i.d simulated from $\mathbf{P}_{x,n}$. Since f is bounded, we can write

$$\mathsf{Var}(\widehat{\gamma}_n^N(f)) \asymp \frac{1}{N}\mathsf{Var}(\mathbf{1}\{X_n^i \text{ is alive}\}) \qquad \widehat{\gamma}_n(f) \asymp \mathbf{P}_{x,n}(X_n^i \text{ is alive})$$

which means, in the same way as we saw earlier, that the relative variance of this approximation:

$$\frac{\sqrt{\mathsf{Var}(\widehat{\gamma}_n^N(f))}}{\widehat{\gamma}_n(f)} \sim_{n\to\infty} \frac{\exp(cn)}{\sqrt{N}} \qquad \text{for some } c > 0,$$

is of order $\exp(cn)$ as $n \to \infty$, which is not too good. On the other hand, we know that to compute $\widehat{\gamma}_n(f)$, it suffices to estimate $\bar{\mathbf{E}}_x\left[f(X_n)/U(X_n)\right]$ and then weigh the result by the deterministic quantity $U(x)\lambda^n$. Now notice that for the estimator

$$m^N(f) = \frac{1}{N}\sum_{i=1}^{N} f(\bar{X}_n^i)/U(\bar{X}_n^i),$$

where $(\bar{X}_t^i)t \leq n$ are i.i.d simulations from $\bar{\mathbf{P}}_{x,n}$, we have that both

$$\mathsf{Var}(m^N(f)) = \frac{1}{N}\mathsf{Var}(f(X_n)/U(X_n)) \qquad \text{and} \qquad \mathbf{E}[m^N(f)] = \bar{\mathbf{E}}_x[f(X_n)/U(X_n)]$$

are order 1 as $n \to \infty$, which means that our relative variance is order 1 too as $n \to \infty$. A drastic reduction compared to the naive method.

References

[1] Nicolas Chopin and Omiros Papaspiliopoulos. *An introduction to sequential Monte Carlo*. Springer Series in Statistics. Springer, 2020.

[2] Pierre Collet, Servet Martínez, and Jaime San Martín. *Quasi-Stationary Distributions: Markov Chains, Diffusions and Dynamical Systems*. Springer, Berlin, Heidelberg, 1 edition, 2013.

[3] Pierre Del Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems With Applications*. 05 2004.

[4] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer New York, 2 edition, 1996.