Boosting reinforcement learning with sparse and rare rewards using Fleming-Viot particle systems

Daniel Mastropietro, Szymon Majewski, Urtzi Ayesta, Matthieu Jonckheere

Boosting reinforcement learning with sparse and rare rewards using Fleming-Viot particle systems

Daniel Mastropietro Daniel.Mastropietro@irit.fr
IRIT
Université de Toulouse
CNRS
31071 Toulouse, France

Szymon Majewski sjm.majewski@gmail.com
École Polytechnique
Paris, France

Urtzi Ayesta urtzi.ayesta@irit.fr
CNRS-IRIT
31071 Toulouse, France
Ikerbasque - Basque Foundation for Science
48009 Bilbao, Spain
UPV/EHU
University of the Basque Country
20018 Donostia, Spain

Matthieu Jonckheere mjonckheer@laas.fr
CNRS-LAAS
31071 Toulouse, France

Abstract

We consider reinforcement learning control problems under the average reward criterion in which non-zero rewards are both sparse and rare, that is, they occur in very few states and have a very small steady-state probability. Using Renewal Theory and Fleming-Viot particle systems, we propose a novel approach that exploits prior knowledge on the sparse structure of the environment to boost exploration of the non-zero rewards. We also demonstrate how to combine the methodology with a policy gradient algorithm to construct the FVRL method that is able to efficiently solve structured control problems under these scenarios. We provide theoretical guarantees of the convergence of both the steady-state probability estimator and the policy gradient learner. Finally, we illustrate the method on an $M/M/1/K$ queue control problem where the objective is to determine the optimum blocking threshold K. Our results show that FVRL learns the optimum blocking threshold much more efficiently than vanilla Monte-Carlo reinforcement learning.

Keywords: Reinforcement Learning, particle systems, queues, Fleming-Viot, average reward, structured learning

1. Introduction

Deep Reinforcement learning methods, by being able to take advantage of large amounts of computational resources, have been extremely successful at solving very complex problems with large state dimensions and sparse rewards, obtaining super-human performance particularly in games Silver et al. (2017). Many of these successes have been obtained in an episodic setting in which, even under sparse rewards conditions, there is a certainty that the episode will eventually finish, at which moment a reward will be observed.

In several application domains, in particular in networking or robotics, the environment is not episodic (i.e. there is no notion of progression as in games) and the rewards are very rarely observed both in space and time, which we refer by sparse and rare, respectively. For example in networking, a fundamental problem is how to dimension the system in order to optimise the performance bearing in mind that the blocking probability (i.e., the probability that the system cannot accept a new data packet, call, or computation task) can be extremely small. As pointed out in Dulac-Arnold

et al. (2019), efficiently managing the exploration task when non-zero rewards are very rarely observed remains a challenge, and this provides the main motivation for the present work. Our starting point is the structural knowledge on the underlying Markov Decision Process (MDP) that in many cases can be leveraged to drastically improve exploration.

In this paper, we focus on model-free approaches (i.e. without previous learning of a specific model), and take as performance criterion the long-run average reward. We assume that we have access to a simulator of the system, or in its absence, to a large amount of data to construct experience replay of the system.

Our approach relies on the identification of large sets $\mathcal{A}$ of states with zero reward, which can be obtained through previous knowledge on the underlying MDP, or from information gained from exploration already performed. Using the notion of implicit conditioning, we first show that the average reward of the original problem can be expressed in terms of modified trajectories that are absorbed in $\mathcal{A}$. We then show that the average reward of trajectories outside $\mathcal{A}$ can be efficiently computed via the so-called Fleming-Viot particle system (FV). We finally introduce FVRL, a reinforcement learning algorithm that uses Fleming-Viot to solve problems with sparse and rare rewards and gradients.

To illustrate the main ideas, we shall consider the case of an $M/M/1/K$ queue, which is one of the fundamental models used in queuing theory to estimate the number of busy lines in a telephone circuit, or the number of packets in an Internet router. In these applications, the objective is to minimise the blocking probability, i.e., the probability that a new packet of voice call is blocked. The stationary probability of being in state K decreases exponentially with K at rate ρ^K, where ρ is the load of the system. The rewards are thus sparse, since blocking only occurs when the system is full, and rare since this happens with very low probability. Theory shows, Darroch and Seneta (1967), that in the case of the $M/M/1/K$ and taking the state 0 as absorbing, the limiting probability of the process of visiting state K, conditioned to not being absorbed (which is approximated by the Fleming-Viot particle system), decays as $\sqrt{K}\rho^{K/2}$. This represents a significant increase over the probability of the original process: for example, for $K = 40$ and $\rho = 0.7$, this probability is about 8,000 times larger, and increases to 6 million times when $\rho = 0.5$. Hence, by shifting the estimation toward a conditioned evolution, we expect both a lower sample complexity for estimation and improvements in control. Our numerical results show that the Fleming-Viot approach overcomes the difficulty of vanilla Monte-Carlo to properly estimate key quantities for reinforcement learning algorithms, such as value functions and gradients, and thus, in turn, FVRL is able to learn the optimum blocking size in the policy learning context for this example.

In summary, we propose a methodology that can be used to boost learning in reinforcement learning environments with the following characteristics: (i) they present sparse non-zero rewards occurring very rarely at unknown states; (ii) thanks to prior knowledge, it is possible to define a set of states presenting zero rewards; (iii) the policy gradients are non-zero in very few and rare sates. The learning goal considered here is to maximise the long-run average reward under stationary regime operation but this could be easily generalized to contexts with discounts. In environments where the occurrence of non-zero rewards has an extremely low probability, this methodology enables learning at admissible learning times, whereas vanilla Monte-Carlo methods fail.

Related work and contribution

Initially, one might think that the problem of estimating rare events could be tackled using Importance Sampling (IS), an area in which there exists a large literature, but we note that the problem at hand impedes the application of IS methods. Firstly, we here look at scenarios where **there is no policy** allowing to explore rare states. In other words, states with reward are very rarely explored under all policies, so IS in this sense is not an option. Secondly, it is also not possible to invoke the original IS principle when looking purely at the evaluation task for a fixed policy, as this requires a change of measure for Monte Carlo, an infeasible task in our case as the transition rates of the Markov chains are unknown.

On the other hand, the problem of sparse rewards (in space) has been of central interest for a long time and the most intuitive solution to sparse (but not necessarily rare) reward problems is reward shaping. Mataric formulated the idea back in 1994 Mataric (1994). However, these methods have a few drawbacks: expertise is needed to shape the rewards, and only very few and quite arbitrary policies might be reached as a consequence of shaping. Our proposal fits more into the idea of curiosity-driven methods as explained for instance in Pathak et al. (2017). The idea is basically to encourage the exploration of unvisited states in the environment. While existing mechanisms are based on including a bonus term in the loss function to favour exploration (Pathak et al. (2017); Burda et al. (2018)), we define a new (more radical) curiosity mechanism adapted to sparse and rare reward environments that forces exploration outside a set of states that have already been explored or are known to be uninformative.

In our main contribution we develop FVRL, a method that combines FV and RL in problems with sparse and rare rewards that are ubiquitous in the control of stochastic networks. To the best of our knowledge, there is no adequate solution to this problem in the state of the art. Our initial results on a simple queuing model illustrate the potential benefits of the approach, and we will investigate in further research its applicability to a wider variety of examples. Even though we do not focus here on deep learning tools, our proposal could definitely be used in combination with them. This is left for future work.

The rest of the paper is organized as follows. Section 2 describes the mathematical setting of the problem, Section 3 the general methodology, in Section 4 we show its applicability in the case of an $M/M/1/K$ queue, and Section 5 presents the numerical results.

2. Problem description

We consider a continuous-time MDP $(\mathcal{S}, A, q, R)$ with a finite state space $\mathcal{S}$, action space A, jump rates q, and rewards R, under the average cost criterion. Throughout the paper we assume that for each policy π the continuous-time Markov Process X_t^π obtained by following the policy π is aperiodic and irreducible. We denote by p^π the stationary distribution of X_t^π, and by $\mathbb{E}^\pi(\eta)$ the expectation of a function $\eta : \mathcal{S} \to \mathbb{R}$ with respect to p^π. We will be interested in computing $\mathbb{E}^\pi(\eta)$ under the assumption that the function η is zero outside of a small set of states $\mathcal{C} \subset \mathcal{S}$. For example, if the rewards are assumed to be sparse, the reward function $r : \mathcal{S} \to \mathbb{R}$ is zero outside a small set $\mathcal{C}$ and its expected value v^π is computed as $v^\pi \doteq \mathbb{E}^\pi(r) = \sum_{y\in\mathcal{C}} r(y)p^\pi(y)$.

The objective of computing $\mathbb{E}^\pi(\eta)$ will be achieved as follows: we will first choose a set $\mathcal{A} \subset \mathcal{S}$, such that $\mathcal{C} \cap \mathcal{A} = \emptyset$. Then, we will use Fleming-Viot particle systems to compute the truncated expectation $\sum_{x\in\mathcal{A}^c} p^\pi(x)\eta(x)$ where η is a function of interest (e.g. the average reward v^π or its gradient). We will choose $\mathcal{A}$ so that η is zero on that set, thus obtaining an estimator of the desired expectation $\mathbb{E}^\pi(\eta)$.

The main objective of our paper is two-fold: (i) to propose an efficient algorithm to estimate $\mathbb{E}^\pi(\eta)$ in the case when η is zero outside of a small set of rare states, and (ii) to use this algorithm to improve the estimations of gradients in combination with the policy gradient theorem in order to solve optimal control problems with sparse rewards and/or a specific policy structure. As described in the introduction, we will use the $M/M/1/K$ queue as a guiding example to illustrate the applicability of our approach.

3. Methodology

In this section we present our proposed method to estimate $\mathbb{E}^\pi(\eta)$ for a given policy π and a function $\eta : \mathcal{S} \to \mathbb{R}$ which is zero outside a set $\mathcal{C} \subset \mathcal{S}$. In many cases we only have access to this function through a stochastic oracle, for example when η is the mean reward function r. For simplicity of exposition, we assume that we have direct access to the function η. We also assume throughout this section that the policy π and the chosen set $\mathcal{A} \subset \mathcal{S}$, whose intersection with $\mathcal{C}$ is empty, are fixed. We denote by $\vec{\partial}\mathcal{A}$ the entrance boundary of $\mathcal{A}$, i.e., the set of states $x \in \mathcal{A}$ for which there exists at least a state $y \in \mathcal{A}^c$ with positive jump rate to x, i.e., $q(y, x) > 0$. The set $\vec{\partial}\mathcal{A}^c$ is defined analogously.

3.1 Implicit conditioning

We start by presenting a formula for the expectation $\mathbb{E}^\pi(\eta)$ derived using renewal theory. We define stopping times $T_{\mathcal{A}^c} = \inf\{t > 0 : X_t^\pi \in \mathcal{A}^c\}$ and $T_{\mathcal{A}} = \inf\{t > T_{\mathcal{A}^c} : X_t^\pi \in \mathcal{A}\}$, that is, by $T_{\mathcal{A}^c}$ we denote the first time of entry into $\mathcal{A}^c$ and by $T_{\mathcal{A}}$ the first time of entry into $\mathcal{A}$ after visiting $\mathcal{A}^c$. We also denote by $T_{\mathcal{K}} = \inf\{t > 0 : X_t^\pi \in \mathcal{A}\}$ the first time the process X_t^π leaves the set $\mathcal{A}^c$. Note that $T_{\mathcal{K}} = T_{\mathcal{A}} - T_{\mathcal{A}^c} > 0$.

Finally, we define $p^\pi_{\vec{\partial}\mathcal{A}}(x) \doteq \mathbb{P}(X^\pi_{T_{\mathcal{A}}} = x | X_0^\pi \sim p^\pi), \forall x \in \vec{\partial}\mathcal{A}$, the state distribution of entrance to $\mathcal{A}$ under stationarity, and $\mathbb{P}^{\vec{\partial}\mathcal{A}}(B) \doteq \mathbb{P}(B|X_0^\pi \sim p^\pi_{\vec{\partial}\mathcal{A}})$, the probability of any event B when the Markov process X_t^π starts at a state in $\vec{\partial}\mathcal{A}$ chosen with probability $p^\pi_{\vec{\partial}\mathcal{A}}$. The respective measures for the complement set $\mathcal{A}^c$, $p^\pi_{\vec{\partial}\mathcal{A}^c}$ and $\mathbb{P}^{\vec{\partial}\mathcal{A}^c}$, are defined analogously.

Using renewal theory, the stationary average of an arbitrary function η of the process can be characterised as (see chapter VI in Asmussen (2003))

$$\mathbb{E}^{\pi}(\eta) = \frac{\mathbb{E}^{\vec{\partial}\mathcal{A}}\big(\int_0^{T_{\mathcal{A}}} \eta(X_t^{\pi})dt\big)}{\mathbb{E}^{\vec{\partial}\mathcal{A}} T_{\mathcal{A}}}. \tag{1}$$

We propose a method to estimate the above quantity that penalizes trajectories that enter $\mathcal{A}$ in order to boost exploration of the relevant part of the state space. The penalisation consists in immediately replacing trajectories that enter $\mathcal{A}$ by trajectories outside $\mathcal{A}$. To this end, we use the dynamics of a particle system known as Fleming-Viot (FV) Burdzy et al. (2000) which has been used in the literature to simulate quasi-stationary distributions Burdzy et al. (2000); Asselah et al. (2011). The Fleming-Viot N-particle system with driving process X_t^{π} and absorption set $\mathcal{A}$ is a continuous Markov Process $(\xi_t^{\nu})_{t\geq 0}$ on state space $(\mathcal{A}^c)^N$, constructed as follows. We first choose some probability distribution on $\mathcal{A}^c$ denoted by ν. We sample an N-dimensional vector ξ_0^{ν}, such that $\xi_0^{\nu}(k)$ are i.i.d random variables distributed according to ν. Each particle $\xi_t^{\nu}(k)$ then evolves independently according to the dynamics of X_t^{π}, but whenever it hits a state in $\mathcal{A}$, it immediately jumps to the position of one of the other particles chosen uniformly at random. This mechanism allows us to only explore trajectories outside $\mathcal{A}$, which is where the informative rewards are located.

In order to exploit the Fleming-Viot particle system for the estimation of $\mathbb{E}^{\pi}(\eta)$, the following simple proposition is instrumental.

Proposition 1 *Given a set $\mathcal{A} \subset \mathcal{S}$ and a function $\eta : \mathcal{S} \to \mathbb{R}$ that is zero on $\mathcal{A}$, the following holds:*

$$\mathbb{E}^{\pi}(\eta) = \frac{\int_0^{\infty} \mathbb{E}^{\vec{\partial}\mathcal{A}^c}\Big(\eta(X_t^{\pi}) 1_{T_{\mathcal{K}}>t}\Big)dt}{\mathbb{E}^{\vec{\partial}\mathcal{A}} T_{\mathcal{A}}}. \tag{2}$$

Proof The proof is postponed to appendix A. ■

It readily follows that (2) can be rewritten as:

$$\mathbb{E}^{\pi}(\eta) = \int_0^{\infty} f_{\eta}(t) g(t) dt, \tag{3}$$

where

$$\begin{aligned} f_{\eta}(t) &\doteq \sum_{x\in\mathcal{A}^c} \eta(x) \phi_t^{\vec{\partial}\mathcal{A}^c}(x), \\ g(t) &\doteq \frac{\mathbb{P}^{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{K}} > t)}{\mathbb{E}^{\vec{\partial}\mathcal{A}} T_{\mathcal{A}}}, \end{aligned}$$

and $\phi_t^{\vec{\partial}\mathcal{A}^c}(x) \doteq \mathbb{P}^{\vec{\partial}\mathcal{A}^c}(X_t^{\pi} = x | T_{\mathcal{K}} > t)$ is the probability that the process X_t^{π}, started at a state in $\vec{\partial}\mathcal{A}^c$ chosen with probability $p^{\pi}_{\vec{\partial}\mathcal{A}^c}$, is in x provided it has not been absorbed.

3.2 Estimation with Fleming-Viot particle systems

An estimator of (3) is constructed by estimating each function inside the integral, as follows: g is estimated using regular Monte-Carlo from observations of the stopping times $T_{\mathcal{K}}$ and $T_{\mathcal{A}}$ coming from the simulation of X_t^{π}, starting at a state in $\vec{\partial}\mathcal{A}$ selected uniformly at random; f_{η} is estimated from the simulation of the FV N-particle system driven by X_t^{π} with absorption set $\mathcal{A}$, starting at states in $\vec{\partial}\mathcal{A}^c$ chosen with the probability distribution $p^{\pi}_{\vec{\partial}\mathcal{A}^c}$ estimated by the previous simulation. The estimation details are given in appendix B.

3.3 Bound on the estimation error

It has been proved in a series of papers that, for finite state spaces Cloez and Corujo (2021); Cloez and Thai (2016), uniform in time propagation of chaos holds for Fleming-Viot particle systems. In particular, if ν is a probability measure

on $\mathcal{S}$, then, under assumption of proper initialization, Cloez and Corujo (2021)[Theorem 1.4] shows the following bound on the speed of convergence w.r.t. the number of particles N of the empirical mean $m(\cdot, \xi_t^\nu)$ towards $\phi_t^\nu(x)$:

$$\sup_{\|\varphi\|_\infty \leq 1} \sup_{t \geq 0} \mathbb{E}\left| [m(\cdot, \xi_t^\nu)(\varphi)] - \phi_t^\nu(\varphi) \right| \leq \frac{C_{\text{FV}}}{\sqrt{N}}, \tag{4}$$

where $\|\varphi\|_\infty = \sup_{x \in \mathcal{A}^c} |\varphi(x)|$ and C_{FV} is a positive constant depending on the characteristics of the driving process. Using this result, we can show the following error bound for the estimator of (3) in an idealized case when the simulation for the estimation of g is started according to $p^\pi_{\vec{\partial}\mathcal{A}}$ and the starting positions of the Fleming-Viot particles are i.i.d samples from $p^\pi_{\vec{\partial}\mathcal{A}^c}$.

Theorem 2 *Assume that we start the simulation of X_t^π for the estimator $\hat{g}$ according to the distribution $p^\pi_{\vec{\partial}\mathcal{A}}$, that M return cycles to $\mathcal{A}$ under stationarity are observed during that simulation, and that we compute the estimator $\hat{f}_\eta$ using the FV particle system started at the positions of N i.i.d samples from $p^\pi_{\vec{\partial}\mathcal{A}^c}$. Let η be a state function such that $\eta(x) = 0$ for $x \in \mathcal{A}$ and $\sup_{x \in \mathcal{A}^c} |\eta(x)| \leq 1$. Then the following bound holds:*

$$\mathbb{E}\left| \hat{\mathbb{E}}^\pi(\eta) - \mathbb{E}^\pi(\eta) \right| \leq \frac{C_{FV} \mathbb{E}^{\vec{\partial}\mathcal{A}^c} T_{\mathcal{K}}}{\sqrt{N} \mathbb{E}^{\vec{\partial}\mathcal{A}} T_{\mathcal{A}}} + \frac{C_g}{\sqrt{M}} + \mathcal{O}\left(\frac{1}{M}\right),$$

where C_g is given in terms of the distributions of $T_{\mathcal{K}}, T_{\mathcal{A}}$, and is approximately equal to $3 \frac{\mathbb{E}^{\vec{\partial}\mathcal{A}^c} T_{\mathcal{K}}}{\mathbb{E}^{\vec{\partial}\mathcal{A}} T_{\mathcal{A}}}$.

A more precise statement of Theorem 2 together with the proof are provided in Appendix C.

3.4 Performance measures

Theorem 2 ensures that the estimation error is of the same order as Monte-Carlo, which gives minimal guarantees for FV. However, the estimation error should not be our only focus to evaluate the difference between FV and Monte-Carlo, especially when the ultimate goal is the convergence of a reinforcement learning algorithm. Indeed, in the control problem, a noisy but still informative signal might be very useful compared to no signal at all. Our main idea when replacing MC by FV is to trade the observation of a very rare event in the original problem by the observation of a more common event for FV particles. Although a fully rigorous analysis of the probability to observe a non-zero reward is out of the scope of this paper, we can give rough estimations using the results of Groisman and Jonckheere (2013); Cloez and Corujo (2021). The dynamics of a tagged particle of the FV process converges in N to a one dimensional Markov process with stationary distribution ν_Q^π (see Groisman and Jonckheere (2013)), where

$$\nu_Q^\pi(B) = \lim_{t \to \infty} \mathbb{P}^{\vec{\partial}\mathcal{A}^c}(X_t^\pi \in B | T_{\mathcal{K}} > t).$$

If the state space is finite, this one dimensional process in turn converges in distribution exponentially fast to its stationary distribution. Hence, the probability of finding a non-zero signal for the FV process in a finite time interval is of the order of $\nu_Q^\pi(\mathcal{C})$. Provided that a complete cycle can be observed with a similar probability (which commands the estimation of $g(t)$ in (3)), the result is that we have replaced the MC probability of an informative signal $\mathbb{E}^\pi(\eta)$ by $\mathbb{E}^{\nu_Q^\pi}(\eta)$ which can be significantly higher, depending on the Markov chain dynamics and on the set $\mathcal{A}$.

3.5 FVRL: Policy gradient approach with FV

In this subsection we show how the Fleming-Viot estimation introduced in Section 3.2 can be combined with the policy gradient theorem to solve optimal control problems in environments with sparse and rare rewards under the average reward criterion. As noted above, the FV estimation can be instrumental in two ways: to evaluate value functions and to evaluate gradients. Indeed, there are many MDPs (see for instance Ross and Tsang (1989); Ross (1995); Koole (1998); Bonald et al. (2004); Koole (2007)) where the optimal policies are known to be of threshold type (because of underlying monotonicity properties). When the rewards structure are sparse and rare, this leads to natural parameterisations of

the policies where their gradients are non-zero only in very few states. We provide here a general illustration of the methodology, and we will apply it to the concrete example of an $M/M/1/K$ queue in Section 4.

Let us consider the case in which an agent interacts with an environment (available in practice either through a simulator or through experience replay of historical data), with the aim of minimising the long-run average cost. As customary in the literature, we let S_t, R_t and A_t denote the state, reward and action at the t-th time step, respectively. We denote by π_θ the policy function parameterised by θ. We recall that we assume that a positive reward is accrued only in states in $\mathcal{C}$. It follows from classical MDP theory that the optimal policy for the average reward criterion is also bias optimal, i.e., when the one step reward (thought of here as a one step cost) at time t is $R_t - v^{\pi_\theta}$, see Puterman (2005). We thus define the state-action value function as

$$Q_\theta(x,a) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty}(R_t - v^{\pi_\theta}) \mid S_0 = x, A_0 = a\right], \tag{5}$$

and we seek to find the θ that minimises the global (undiscounted) cost, i.e. $v^{\pi_\theta} \doteq \sum_x p^{\pi_\theta}(x)\sum_a \pi_\theta(x,a)Q_\theta(x,a)$.

We propose to use a gradient descent algorithm to learn the optimum parameter θ that minimises v^{π_θ}. Denoting by X the random variable associated to the steady-state distribution under π, it follows from the Policy Gradient Theorem for the average reward criterion Sutton et al. (2000) that

$$\nabla_\theta v^{\pi_\theta} \quad = \quad \mathbb{E}^{\pi_\theta}[Q_\theta(X,a)\nabla_\theta \pi_\theta(a|X)] = \sum_{x\in\mathcal{S}} p^{\pi_\theta}(x)\sum_a Q_\theta(x,a)\nabla_\theta\pi_\theta(a|x) \tag{6}$$

Under this assumption, the FV system can also be leveraged to estimate the gradients through the estimation of $p^{\pi_\theta}(x)$, leading to the FVRL algorithm. The details of the estimation process are given in appendix D.

Proposition 3 *Given a continuously differentiable parameterisation v^{π_θ}, the policy gradient FVRL algorithm converges with probability 1 to the optimum parameter θ^*.*

The proof is a consequence of H. J. Kushner (2003) (see Theorem 2.1 in section 5.2) and Bottou et al. (2018), that consider the convergence of stochastic approximation algorithms with bias, as is the case with the FV estimator used in the average reward gradient estimation.

Of particular interest are cases where the gradients in (6) are non-zero (or significant) only for rare states (because of e.g. structural properties of the underlying policies as explained above). In those cases our method can find informative gradients where vanilla methods would not[1].

4. Application to queuing systems

In this section we apply the methodology outlined in Section 3.1 to determine the optimum blocking threshold in a toy-model queuing system. To simplify the exposition and be able to do theoretical computations, we shall assume that the underlying MDP is a simple $M/M/1/K$ queue with fixed rates, but the method could also be applicable to unknown state-dependent rates. The $M/M/1/K$ queue-length occupancy, denoted by X_t^π, with a policy π that blocks an incoming job when $X_t^\pi = K$, measures the number x of jobs waiting to be served in the buffer at a given time. It is a continuous-time discrete space stochastic process living on $\{0,1,\dots,K\}$ with upward rate from any state x (except $x = K$) to $x+1$ given by λ, and downward rate from x (except $x = 0$) to $x-1$ equal to μ. We denote by $\rho \doteq \lambda/\mu$ the load of the system which is typically smaller than 1 in real applications. We suppose that the only state with non-zero reward is $x = K$, i.e., when a new arrival is blocked. Simple computations show that if $\rho < 1$, $p^\pi(K) \doteq \mathbb{P}(X_t^\pi = K) = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}}$. Even for moderate values of K, this probability can be extremely small. For

1. We also believe that, even in cases where the parameterisation leads to some policy gradients being non-zero inside $\mathcal{A}$, the assumption of sparsity of rewards actually implies that the gradients of the value function inside $\mathcal{A}$ are very flat and uninformative. In those scenarios, FV could again be leveraged as a good exploration mechanism. This falls outside the scope of this paper and will be investigated in future research.

example with $\rho = 0.5$ and $K = 20$, $p^\pi(K)$ becomes of order 10^{-6}. Regarding the rewards, we consider that the cost of acceptance $r(x, a = 1) = 0$, $\forall x$, and the cost of rejection $r(x, a = 0) = B(1 + b^{x-x_{ref}})$, where B, b and x_{ref} are positive constants. The choice of these parameters ensures that the optimal policy is of threshold type, i.e., there exists a state $0 < K < \infty$ such that $a = 1$ is optimal for all $x < K$, and $a = 0$ is optimal for $x = K$ Koole (1998). The value of x_{ref} is the reference queue size that is instrumental in defining the optimum threshold, K^* (which is close to x_{ref}). We will first deploy FV to estimate efficiently the steady-state probabilities $p^\pi(x)$ as outlined in Section 3.2, and then use a policy gradient approach, see Section 3.5, to derive the optimal job-acceptance policy, when blocking generates costs.

4.1 Estimating the blocking probability in an $M/M/1/K$ queue with Fleming-Viot

In this section we assume there is no blocking until state K. From the description above, clearly $\mathcal{C} = \{K\}$ in (2), and any set $\mathcal{A} = \{0, 1, \ldots, J-1\}$ with $J \leq K$ is a valid absorption set for the FV estimator. From equation (2), we can write the stationary blocking probability as

$$p^\pi(K) = \frac{\int_0^\infty \phi_t^J(K)\mathbb{P}^J(T_\mathcal{K} > t)dt}{\mathbb{E}^{J-1}T_\mathcal{A}}. \tag{7}$$

As outlined in Section 3.2, the empirical distribution of the states $m(x, \xi_t^J)$ converges, for fixed time, to $\phi_t^J(K)$ as the number of particles N of the FV system increases. Therefore, we can achieve a precise estimate of (7) by estimating $\phi_t^J(K)$ with the empirical distribution of the FV particles' positions. The three different quantities contributing to the FV estimator are computed as follows: (i) $P^J(T_\mathcal{K} > t)$ and $E^{J-1}(T_\mathcal{A})$ are estimated from the absorption cycles observed during the simulation of a single queue system started at $x = J - 1$ and allowed to run until a sufficiently large number of arrival events T are observed. Each cycle contributes once to the estimation of both quantities. In the case of $P^J(T_\mathcal{K} > t)$, the survival time $T_\mathcal{K}$ is measured from the time the queue visits $x = J$ from below until it visits $x = J - 1$ from above. The tail of the trajectory that is not absorbed by the end of the simulation is discarded; (ii) $\phi_t^J(K)$ is estimated as the empirical distribution of the blocking queue occupation size K, i.e., $m(K, \xi_t^J)$, from the simulation of an N-particle FV system, each starting at $x = J$ and allowed to run until the maximum observed survival time from step (i), as explained in appendix B.

Remark 4 *There is trade-off between a small and a large J, the state that defines the size of the absorption set $\mathcal{A}$: for smaller J, the return times to $\mathcal{A}$ will be smaller, requiring fewer arrival events T for the estimation of (i), but at the same time visiting the rare blocking state K will be rarer, requiring a larger number of particles N for the estimation of (ii). The opposite is true for larger values of J.*

4.2 Learning the optimum threshold K

In this section we explain how the optimum integer-valued K is learned using the policy gradient methodology presented in Section 3.5. Following Massaro et al. (2019), we propose a parameterised acceptance policy $\pi(a = 1|x)$ that is a linear step function of the state x, that is deterministic for x outside the interval $(\theta, \theta + 1)$ and decreases linearly from 1 to 0 in such interval. That is, the acceptance policy parameterised by the positive-real-valued θ, is defined as:

$$\pi_\theta(a = 1|x) = \begin{cases} 1 & \text{if } x \leq \theta, \\ x - \theta + 1 & \text{if } \theta < x < \theta + 1, \\ 0 & \text{if } x \geq \theta + 1. \end{cases}$$

Note that the policy is deterministic for integer-valued θ, in which case the blocking size is $K = \theta + 1$.

We use a gradient descent algorithm to learn the optimum parameter θ that minimises the long-run expected cost v^{π_θ}. Using expression (6), the gradient of v^{π_θ} becomes

$$\frac{\partial v^{\pi_\theta}}{\partial \theta} = p^{\pi_\theta}(K-1)\left[Q_\theta(K-1, 1) - Q_\theta(K-1, 0)\right], \tag{8}$$

where $K - 1$ is the smallest integer that is larger than or equal to θ. Observe that this parameterisation leads, as expected, to gradients being 0 for $x < K - 1$. Note also that the gradient is discontinuous at θ and $\theta + 1$, making the assumptions

of Proposition 3 not fully satisfied. However, these two points have measure zero and therefore, with probability 1, no discontinuity is observed[2].

5. Results

In this section we present the results of using the FV approach for the estimation of the stationary probability as described in Section 4.1 on the one hand, and in the estimation of the optimum threshold of the control problem described in Section 4.2 on the other hand. In each context, the method's performance is compared with vanilla Monte-Carlo (MC). An $M/M/1/K$ queue system is used as test bench as it provides a well-known environment to verify the correctness and accuracy of estimators. The system serves, at rate $\mu = 1$, single-class jobs arriving at rate $\lambda = 0.7$, hence it has load $\rho = 0.7$.

5.1 Estimates of the blocking probability

In this section we study the convergence of the FV estimator as both N and T increase. To simplify the analysis, we make T increase proportional to N so that plots are produced as a function of N. The constant of proportionality is 100.

The FV estimator of the blocking probability follows from the methodology described in Section 4.1. The benchmark MC estimator, on the other hand, is computed by a direct application of expression (1), i.e., as the fraction of the time spent at state K and the total time of cycles returning to the initial state $x = J - 1$, observed during the simulation of the queue. To guarantee a fair comparison between the two methods, we allow the queue to run until it reaches the same number of events observed in the two-phase FV estimator. In addition, the queue simulation starts at $x = J - 1$, the boundary of the absorption set $\mathcal{A}$ used in FV, so that both methods start at the same distance from the blocking state.

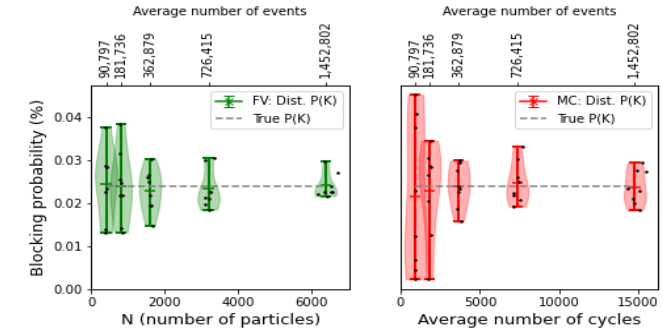
Figure 1 compares the accuracy of the estimation of the blocking probability between FV and MC. We considered the cases $K = 20$ and $K = 40$, which are regarded to represent moderate and large capacities based on their blocking probabilities of order 10^{-4} and 10^{-7}, respectively. The absorption set size J is chosen as the closest integer to $K/2$.

With $K = 20$, both methods behave similarly, giving unbiased estimates of the probability, although FV presents slightly smaller variance. We note that as N and T grow, the variance of the estimator reduces, achieving less than 50% error. With $K = 40$, FV provides an accurate estimate, while MC fails completely as it estimates zero probability, even for simulations with a large number of events, such as one million.

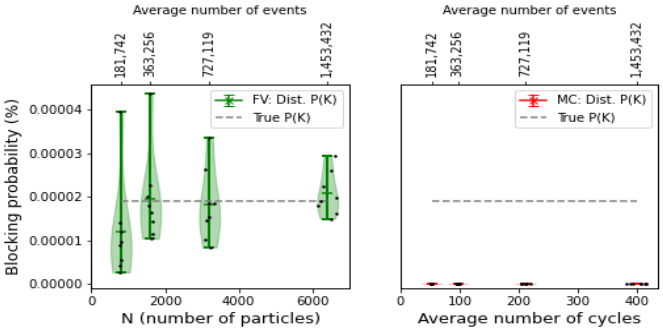
5.2 Policy gradient learner of the optimum threshold K

In this section we use the FVRL method to learn (in a model free context) the optimal blocking size, K^*, in an $M/M/1/K$ queue that minimises the expected blocking cost. The modeling parameters, in particular the transitions rates, and the reward structure are as defined in Section 4. Given $\rho = 0.7$, we chose $b = 3$ $(> 1/\rho)$ so that $0 < K^* < \infty$, and thus the problem of determining the optimum blocking size is non-trivial. Parameter B simply defines a scale and was chosen equal to 5. The setup of the experiments was done as follows: we chose fairly large values of the reference queue size x_{ref} to tune the optimum blocking size K^* on which to experiment; an even larger value $(x_{ref} + 10)$ was chosen for the initial blocking size guess, so that, already at the onset, blocking occurs rarely. Two different J/K fractions were considered in order to experiment with different sizes of the absorption set $\mathcal{A}$, 0.3 and 0.5. The values of the number of particles N and of the number of arrivals T were chosen in accordance to the trade-off described in Remark 4 in Section 4.1, namely larger N and smaller T for the smaller $J/K = 0.3$ value, and smaller N and larger T for the larger $J/K = 0.5$ value. For each setup we ran the FVRL policy learner on 800 learning steps using the chosen J as the size of the absorption set $\mathcal{A}$[3]. The value of J is updated at the start of each learning step to the integer that

2. A special case occurs when θ is integer, in which case the discontinuities would be observed with non-zero probability in the gradient descent algorithm under the following scenario: an integer value is chosen for the initial guess of θ, and integer-valued clipping (e.g. to ± 1) is used for the next θ estimated by the algorithm. This problem is solved by simply not choosing an integer-valued initial guess of θ.
3. After each learning step, the gradient of the average reward in (6) is estimated as described in appendix D, allowing up to 250 arrival events until mixing is observed. The estimate $\hat{\eta}(x)$ in appendix D is based on the average of 100 replications of its estimation procedure (or less if some replications do not reach mixing, which occurs very rarely).



(a) Capacity $K = 20$, on 8 replications per experiment



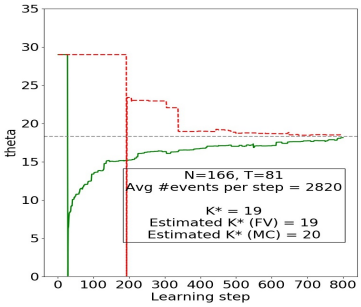
(b) Capacity $K = 40$, on 8 replications per experiment

Figure 1: Convergence analysis of the Fleming-Viot (left) and Monte-Carlo (right) estimators of the blocking probability. The number of particles considered in each FV experiment are $N = 400, 800, 1600, 3200, 6400$, but (b) excludes 400. In the MC experiments, the average number of observed return cycles to the initial state $J - 1$ is indicated on the bottom horizontal axis.

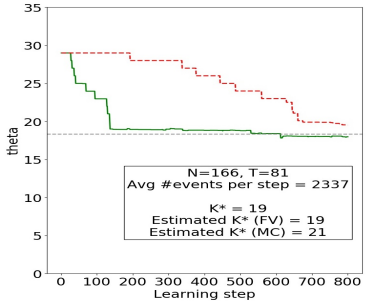
is closest to the J/K fraction of K considered. At the end of each learning step, parameter θ is updated by gradient descent using a constant learning rate of $\alpha = 10$, and the number of observed events at the learning step is recorded. This number was used as the number of events at which the MC learner, which follows the same policy gradient approach but estimates the probability $p^{\pi_\theta}(K-1)$ in expression (8) using Monte-Carlo instead of Fleming-Viot, is stopped at the corresponding learning step. This, together with the fact that each learning step in the Monte-Carlo learner is started at $J - 1$, where J is defined as in FVRL, allows a fair comparison between the two methods.

Two different learning strategies of θ were used and compared: one where the increment of θ after each learning step is left unbounded (as long as it doesn't go below 0.1), and one where the increment is clipped to ± 1. The results are shown in figure 2, where we clearly see how FVRL outperforms vanilla Monte-Carlo in the large $K^* = 24$ case, where MC fails to learn. In the moderate $K^* = 19$ case, both methods are able to learn, but only when clipping is used does FV learn visibly faster than MC. We also make the interesting observation that even large expected relative estimation errors of $\hat{\phi}_t^J(K)$ and $\hat{\mathbb{E}}(T_{\mathcal{A}})$ (100% and 150%) [4] allow FVRL to learn the optimum blocking size after a reasonable number of steps, with almost no difference in convergence speed. On the other hand, a much smaller convergence rate is observed in plots (a), (c) and (e) (compared to the clipping case of (b), (d) and (f)) where the estimated θ jumps abruptly to near the allowed minimum of 0.1, due to a large estimated gradient at the very beginning, while the gradient thereafter is relatively small for θ values smaller than the optimum[5]. Clipping helps avoid these big jumps and makes FVRL learn faster compared to the strategy of unbounded θ updates.

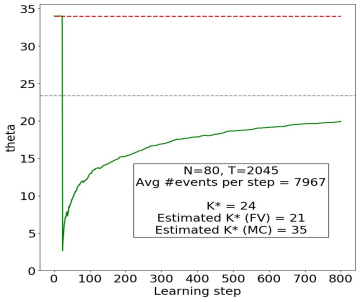
4. Estimation errors of $\hat{\phi}_t^J(K)$ and $\hat{\mathbb{E}}(T_{\mathcal{A}})$ are estimated using simple calculations based on the dynamics of the $M/M/1/K$ queue and of the corresponding absorbed process.
5. These small gradient values are due to an asymmetry in the cost function being optimised which has smaller gradients for θ values smaller than the optimum than for θ values larger than the optimum.



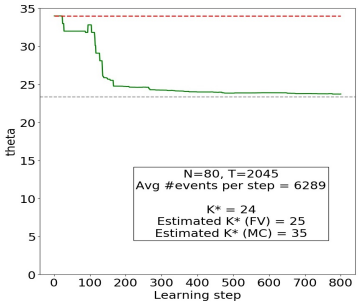
(a) (19, 30, 0.3, 100%)



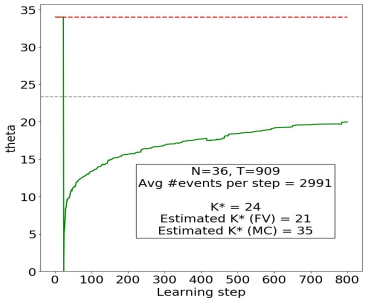
(b) (19, 30, 0.3, 100%) clip



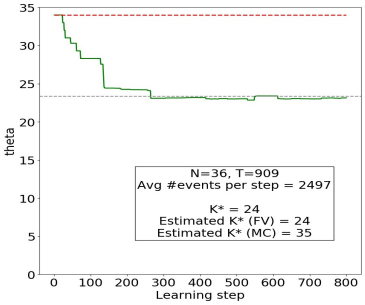
(c) (24, 35, 0.5, 100%)



(d) (24, 35, 0.5, 100%) clip



(e) (24, 35, 0.5, 150%)



(f) (24, 35, 0.5, 150%) clip

Figure 2: Convergence of FVRL (solid green) to the optimum blocking size K^* (dashed gray line) compared to vanilla Monte-Carlo (dashed red) for the $M/M/1/K$ problem. Each plot caption indicates (the optimum K^* value, the initial K_0 guess, the fraction J/K, the relative error (%) expected for $\hat{\phi}_t^J(K)$ and $\hat{\mathbb{E}}(T_{\mathcal{A}})$). Although plots show the result of a single simulation, a total of three replications were carried out in each case, leading to similar results.

6. Conclusions and future work

The Fleming-Viot particle system was presented as an efficient alternative to Monte-Carlo for the exploration of environments where rewards are sparse and their occurrence is rare. Its application to the estimation of the blocking probability in an $M/M/1/K$ queuing system served as a test bench, where the method proved to be much more efficient than Monte-Carlo for large capacities K, where the latter completely fails. Results on optimal control of the queue using policy gradient were also presented, where the proposed FVRL algorithm is able to find the optimum parameter in situations where Monte-Carlo fails. In this case, the accuracy of the blocking probability estimator is not as crucial as in the estimation problem, because the algorithm is able to learn as long as it receives a signal from the rare states.

In future work, we plan to combine the FVRL algorithm with approximation methods like neural networks to deal with higher dimensional scenarios, such as queues receiving multi-class jobs. We also intend to extend it to environments other than queues, as well as explore the choice of the absorption set $\mathcal{A}$ in an adaptive fashion, i.e., based on the information gathered during exploration about the states that give no rewards.

Acknowledgements

This work was partially supported by the French National Research Agency under the program "Investments for the Future" with reference code ANR-11-LABX-0040.

References

S. Asmussen. Applied Probability and Queues. Applications of mathematics : stochastic modelling and applied probability. Springer, 2003. ISBN 9780387002118. URL https://books.google.fr/books?id=BeYaTxesKy0C.

A. Asselah, P.A. Ferrari, and P. Groisman. Quasistationary distributions and Fleming-Viot processes in finite spaces. J. Appl. Probab., 48(2):322–332, 2011. ISSN 0021-9002. doi: 10.1239/jap. URL http://dx.doi.org/10.1239/jap/1308662630.

Thomas Bonald, Matthieu Jonckheere, and Alexandre Proutière. Insensitive load balancing. ACM Sigmetrics Performance Evaluation Review, 32(1):367–377, 2004.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. SIAM Review, 60(2):223–311, jan 2018. doi: 10.1137/16m1080173.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. arXiv preprint arXiv:1808.04355, 2018.

K. Burdzy, R. Holyst, and P. March. A Fleming-Viot particle representation of the Dirichlet Laplacian. Comm. Math. Phys., 214(3):679–703, 2000. ISSN 0010-3616. doi: 10.1007/s002200000294. URL http://dx.doi.org/10.1007/s002200000294.

Bertrand Cloez and Josué Corujo. Uniform in time propagation of chaos for a moran model. arXiv preprint arXiv:2107.10794, 2021.

Bertrand Cloez and Marie-Noémie Thai. Quantitative results for the fleming–viot particle system and quasi-stationary distributions in discrete space. Stochastic Processes and their Applications, 126(3):680–702, 2016.

J. N. Darroch and E. Seneta. On quasi-stationary distributions in absorbing continuous-time finite Markov chains. J. Appl. Probability, 4:192–196, 1967. ISSN 0021-9002.

Gabriel Dulac-Arnold, Daniel J. Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. CoRR, abs/1904.12901, 2019. URL http://arxiv.org/abs/1904.12901.

P. Groisman and M. Jonckheere. Simulation of quasi-stationary distributions on countable spaces. Markov Process. Related Fields, 19(3):521–542, 2013. ISSN 1024-2953.

G. Yin H. J. Kushner. Stochastic Approximation and Recursive Algorithms and Applications. Springer-Verlag, 2003. doi: 10.1007/b97441.

Ger Koole. Structural results for the control of queueing systems using event-based dynamic programming. Queueing systems, 30(3):323–339, 1998.

Ger Koole. Monotonicity in Markov reward and decision chains: Theory and applications, volume 1. Now Publishers Inc, 2007.

Antonio Massaro, Francesco De Pellegrini, and Lorenzo Maggi. Optimal trunk-reservation by policy learning. In IEEE INFOCOM 2019, apr 2019. doi: 10.1109/infocom.2019.8737552.

Maja J Mataric. Reward functions for accelerated learning. In Machine learning proceedings 1994, pages 181–189. Elsevier, 1994.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 2778–2787. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/pathak17a.html.

M. L. Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, 2005.

Keith W. Ross. Multiservice Loss Models for Broadband Telecommunication Networks. Springer, 1995. ISBN 978-3-540-19918-2.

Keith W. Ross and Danny H.K. Tsang. The stochastic knapsack problem. IEEE Transactions on Communications, 37 (7):740–747, July 1989. ISSN 1558-0857. doi: 10.1109/26.31166.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. nature, 550 (7676):354–359, 2017.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In Advances in neural information processing systems, pages 1057–1063, 2000.

Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge University Press, 2019.

Appendix A. Proof of Proposition 1

To reach any state in $\mathcal{C} \in \mathcal{A}^c$ the process needs to go through the entrance boundary of $\mathcal{A}^c$, denoted $\vec{\partial}\mathcal{A}^c$.

Thus, from the Renewal Reward Theorem Asmussen (2003), using the definition of the hitting times $T_{\mathcal{A}^c}$ and $T_{\mathcal{A}}$, and noting that $T_{\mathcal{A}} > T_{\mathcal{A}^c}$, it follows that

$$\mathbb{E}^{\pi}(\eta) = \frac{\mathbb{E}^{\vec{\partial}\mathcal{A}}\big(\int_0^{T_{\mathcal{A}^c}} \eta(X_t)dt + \int_{T_{\mathcal{A}^c}}^{\infty} \eta(X_t)\mathbf{1}_{t<T_{\mathcal{A}}}(t)dt\big)}{\mathbb{E}^{\vec{\partial}\mathcal{A}}T_{\mathcal{A}}},$$

Clearly the first integral is 0, as η is 0 in $\mathcal{A}$ by assumption.

Changing the integration variable to $u = t - T_{\mathcal{A}}$ and using that $T_{\mathcal{K}} = T_{\mathcal{A}} - T_{\mathcal{A}^c}$ yields

$$\mathbb{E}^{\pi}(\eta) = \frac{\mathbb{E}^{\vec{\partial}\mathcal{A}}\big(\int_0^{\infty} \eta(X_t)\mathbf{1}_{t<T_{\mathcal{K}}}(t)dt\big)}{\mathbb{E}^{\vec{\partial}\mathcal{A}}T_{\mathcal{A}}},$$

Using the fact that $X^{\pi}_{T_{\mathcal{A}^c}}$ is distributed according to $p^{\pi}_{\vec{\partial}\mathcal{A}^c}$ when $X_0^{\pi} \sim p^{\pi}_{\vec{\partial}\mathcal{A}}$, and exchanging integral and expectation, expression (2) of Proposition 1 is obtained.

Appendix B. Estimation of $\mathbb{E}^{\pi}(\eta)$

An estimator of $\mathbb{E}^{\pi}(\eta)$ using (3) is constructed from estimators $\hat{f}_{\eta}, \hat{g}$ of functions f_{η}, g. The estimator is then given by

$$\hat{\mathbb{E}}^{\pi}(\eta) = \int_0^{\infty} \hat{f}_{\eta}(t)\hat{g}(t)dt. \tag{9}$$

First, we explain the construction of the estimator $\hat{g}$ of g. We run the Markov Chain X_t starting from a chosen state $x_0 \in \vec{\partial}\mathcal{A}$. Denote $\tau_{\mathcal{A},0} = 0$. We define a sequence of stopping times $\tau_{\mathcal{A},i}, \tau_{\mathcal{A}^c,i}$ by

$$\begin{aligned} \tau_{\mathcal{A}^c,i} &= \inf_{t>\tau_{\mathcal{A},i-1}} \{X_t^{\pi} \in \mathcal{A}^c\}, \\ \tau_{\mathcal{A},i} &= \inf_{t>\tau_{\mathcal{A}^c,i}} \{X_t^{\pi} \in \mathcal{A}\}, \end{aligned}$$

for $i \geq 1$. We also define $T_{\mathcal{K},i} = \tau_{\mathcal{A},i} - \tau_{\mathcal{A}^c,i}$ and $T_{\mathcal{A},i} = \tau_{\mathcal{A},i} - \tau_{\mathcal{A},i-1}$. We run the chain X_t until we obtain random variables $\{T_{\mathcal{A},i}, \mathcal{T}_{\mathcal{K},i}\}_{i=1}^{M_0+M}$. We consider the first M_0 cycles to be burn-in, and define a Monte-Carlo estimator of $\hat{g}$ based on the remaining M samples in the following way:

$$\hat{g}(t) = \frac{\sum_{i=M_0+1}^{M_0+M} \mathbf{1}_{\mathcal{T}_{\mathcal{K},i}>t}}{\sum_{i=M_0+1}^{M_0+M} T_{\mathcal{A},i}} \tag{10}$$

Next, we explain how f can be estimated using the Fleming-Viot N-particle system driven by X_t^{π} with absorption set $\mathcal{A}$ and denoted $(\xi_t^{\nu})_{t\geq 0}$, where ν is some probability distribution on $\mathcal{A}^c$. Let $m(\cdot, \xi) : \mathcal{S} \to [0,1]$ denote the empirical distribution of the N particles with positions described by vector ξ, defined as the empirical mean $m(x,\xi) \doteq \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\xi(i)=x}$. Since $m(\cdot, \xi_t^{\nu})$ is an estimator of the measure ϕ_t^{ν}, $\sum_{x\in\mathcal{A}^c} \eta(x)m(x,\xi_t^{\nu})$ is an estimator of $\sum_{x\in\mathcal{A}^c} \eta(x)\phi_t^{\nu}$.

Ideally then, to estimate f_{η} we would like to have access to a sample of size N from $p^{\pi}_{\vec{\partial}\mathcal{A}^c}$. This is unlikely to be the case, unless the boundary $\vec{\partial}\mathcal{A}^c$ has just one point. In practice, we propose to approximate the measure $p^{\pi}_{\vec{\partial}\mathcal{A}^c}$ by: $\overline{\nu} = \frac{1}{M}\sum_{i=M_0+1}^{M_0+M} \mathbf{1}_{X_{T_{\mathcal{A}^c,i}}}$, that is, by the empirical distribution of the first entry points into $\mathcal{A}^c$ from the simulation used to compute the estimator $\hat{g}$. We can now define:

$$\hat{f}_{\eta}(t) = \sum_{x\in\mathcal{A}^c} \eta(x)m(x,\xi_t^{\overline{\nu}}). \tag{11}$$

We note that, by construction, $\hat{g}(t) = 0$ for $t > T_{\mathcal{K},\max} = \max\{T_{\mathcal{K},i} : M_0 + 1 \leq i \leq M_0 + M\}$. Therefore, since we wish to compute $\hat{\mathbb{E}}^\pi(\eta) = \int_0^\infty \hat{f}_\eta(t)\hat{g}(t)dt$, we only need to simulate the Fleming-Viot process until the time $T_{\mathcal{K},\max}$ is reached. Also, since both $\hat{f}_\eta(t)$ and $\hat{g}(t)$ are almost surely piecewise constant functions and $\hat{g}(t) = 0$ for $t > T_{\mathcal{K},\max}$, the integral $\int_0^\infty \hat{f}_\eta(t)\hat{g}(t)dt$ is a finite sum that can be easily computed.

Appendix C. Error of Fleming-Viot estimation

In this appendix we prove the following more precise statement of Theorem 2.

Theorem 2 *Assume that we start the simulation of X_t^π for the estimator $\hat{g}$ according to the distribution $p^\pi_{\vec{\partial}\mathcal{A}}$, that M return cycles to $\mathcal{A}$ under stationarity are observed during that simulation, and that we compute the estimator $\hat{f}_\eta$ using the FV particle system started at the position of N i.i.d samples from $p^\pi_{\vec{\partial}\mathcal{A}^c}$. Let η be a function of the state such that $\eta(x) = 0$ for $x \in \mathcal{A}$ and $\sup_{x\in\mathcal{A}^c} |\eta(x)| \leq 1$. Then the following bound holds:*

$$\mathbb{E}\left|\hat{\mathbb{E}}^\pi(\eta) - \mathbb{E}^\pi(\eta)\right| \leq \frac{\mathbb{E}^{\vec{\partial}\mathcal{A}^c}T_{\mathcal{K}}}{\mathbb{E}^{\vec{\partial}A}T_{\mathcal{A}}}\frac{C}{\sqrt{N}} + \frac{\left(Var^{\vec{\partial}\mathcal{A}}T_{\mathcal{A}}\right)^{1/2}\mathbb{E}^{\vec{\partial}\mathcal{A}^c}T_{\mathcal{K}}}{(\mathbb{E}^{\vec{\partial}A}T_{\mathcal{A}})^2\sqrt{M}} + \frac{1}{\sqrt{M}}\int_0^\infty \sqrt{F_{\mathcal{K}}(t)(1-F_{\mathcal{K}}(t))}dt + \mathcal{O}\left(\frac{1}{M}\right),$$

where $F_{\mathcal{K}}$ is the distribution function $F_{\mathcal{K}}(t) = \mathbb{P}^{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{K}} \leq t)$. All the moments and the integral in the above bounds are finite, and we have approximately:

$$\begin{aligned}\left(Var^{\vec{\partial}\mathcal{A}}T_{\mathcal{A}}\right)^{1/2} &\approx \mathbb{E}^{\vec{\partial}A}T_{\mathcal{A}}\\ \int_0^\infty \sqrt{F_{\mathcal{K}}(t)(1-F_{\mathcal{K}}(t))}dt &\approx 2E^{\vec{\partial}A^c}T_{\mathcal{K}},\end{aligned}$$

hence the sum of constants next to the terms of order $\mathcal{O}\left(\frac{1}{\sqrt{M}}\right)$ is approximately equal to $3\frac{\mathbb{E}^{\vec{\partial}\mathcal{A}^c}T_{\mathcal{K}}}{\mathbb{E}^{\vec{\partial}\mathcal{A}}T_{\mathcal{A}}}$.

Proof Using the notation $f_\eta, g, \hat{f}_\eta, \hat{g}$ introduced in (3) and (11), (10), we have $\mathbb{E}^\pi(\eta) = \int_0^\infty f_\eta(t)g(t)dt$ and $\hat{\mathbb{E}}^\pi(\eta) = \int_0^\infty \hat{f}_\eta(t)g(t)dt$. We are thus interested in bounding:

$$\mathbb{E}\left|\int_0^\infty \hat{f}_\eta\hat{g}dt - \int_0^\infty f_\eta g dt\right|$$

We start by decomposing the problem of upper bounding the above quantity into two subproblems in the following way:

$$\mathbb{E}\left|\int_0^\infty \hat{f}_\eta\hat{g}dt - \int_0^\infty f_\eta g dt\right| \leq \mathbb{E}\left|\int_0^\infty (f_\eta - \hat{f}_\eta)g dt\right| + \mathbb{E}\left|\int_0^\infty \hat{f}_\eta(\hat{g} - g)dt\right| \tag{12}$$

We start with bounding the first term on the right hand side. For this purpose we will need the uniform propagation of chaos bound presented in 4, that is:

$$\sup_{\|\varphi\|_\infty\leq 1}\sup_{t\geq 0}\mathbb{E}\left|[m(\cdot,\xi_t^\nu)(\varphi)] - \phi_t^\nu(\varphi)\right| \leq \frac{C_{\text{FV}}}{\sqrt{N}},$$

form which it follows that:

$$\sup_{t\geq 0}\mathbb{E}\left|[\hat{f}_\eta(t) - f_\eta(t)\right| \leq \frac{C_{\text{FV}}}{\sqrt{N}}. \tag{13}$$

As was mentioned in subsection 3.3, this bound follows directly from Cloez and Corujo (2021)[Theorem 1.4]. The assumptions of Cloez and Corujo (2021)[Theorem 1.4] have a very general form, but it is easy to check that they are

trivially satisfied in our simple case. The assumption (I) on initialization is satisfied by our assumption that the FV particle system is started at the position of N i.i.d samples from $p^{\pi}_{\vec{\partial}\mathcal{A}^c}$. The assumption (C1) has several parts. The uniform bound on selection rates (that is in our case the intensities of jumps out of $\mathcal{A}^c$) follows from the fact that the state space is finite. The rest of assumption (C1) is trivially satisfied when we take $V^d_\mu(x)$ to be the intensity of jump out of $\mathcal{A}^c$ from the state $x \in \mathcal{A}^c$ for any μ, and set function $V^b_\mu(y), V^s_\mu(x,y)$ equal to zero. Finally the assumption (C2) follows from the fact that we are working with an irreducible Markov Chain on a finite state space. Therefore, using the triangle inequality and then the inequality 13, we obtain:

$$\begin{aligned}\mathbb{E}\left|\int_0^\infty (f_\eta - \hat{f}_\eta) g dt\right| &\leq \mathbb{E}\int_0^\infty \left|(f_\eta - \hat{f}_\eta)\right| g dt \\ &\leq \int_0^\infty \mathbb{E}\left|(f_\eta - \hat{f}_\eta)\right| g dt \\ &\leq \frac{C_{\mathrm{FV}}}{\sqrt{N}}\int_0^\infty g dt = \frac{\mathbb{E}^{\vec{\partial}\mathcal{A}^c} T_{\mathcal{K}}}{\mathbb{E}(T_{\mathcal{A}})}\frac{C_{\mathrm{FV}}}{\sqrt{N}},\end{aligned}$$

where in the last line we also use the 'wedding cake decomposition' $\int_0^\infty, \mathbb{P}^{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{K}} > t)dt = \mathbb{E}^{\vec{\partial}\mathcal{A}^c} T_{\mathcal{K}}$.
Since $\sup_{x\in\mathcal{A}^c}|\eta(x)| \leq 1$, we also have $\left|\hat{f}_\eta(t)\right| \leq 1$ for $t \geq 0$, we get:

$$\left|\int_0^\infty \hat{f}_\eta(\hat{g} - g)dt\right| \leq \int_0^\infty |g - \hat{g}|\, dt.$$

We thus wish to estimate $\mathbb{E}\int_0^\infty |\hat{g} - g|\, dt$. For convenience, we define a random variable $\overline{T}_{\mathcal{A}}$ with the distribution of $T_{\mathcal{A}}$ when X^π_t is started with initial distribution $p^{\pi}_{\vec{\partial}\mathcal{A}}$, and a random variable $\overline{T}_{\mathcal{K}}$ with the distribution of $T_{\mathcal{K}}$ when X^π_t is started at $p^{\pi}_{\vec{\partial}\mathcal{A}^c}$. Since we start the simulation of X^π_t for the purpose of estimating g with distribution $p^{\pi}_{\vec{\partial}\mathcal{A}}$, we do not need any burn-in. We therefore take $M_0 = 0$. We also note that since we start the simulation at the distribution $p^{\pi}_{\vec{\partial}\mathcal{A}}$, it follows from renewal theory Asmussen (2003) that the inter-arrival times $\mathcal{T}_{\mathcal{A},i}$ used to construct the estimator $\hat{g}$ are i.i.d. with distribution $\overline{T}_{\mathcal{A}}$.

We also introduce additional shorthand notation for the numerators and denominators of $g(t)$ and $\hat{g}(t)$. We denote $N_t = \mathbb{P}^{\vec{\partial}\mathcal{A}^c}(T_K > t)$ and $D_{\mathcal{A}} = \mathbb{E}^{\vec{\partial}\mathcal{A}} T_{\mathcal{A}}$. We also denote by $\hat{N}_t, \hat{D}_{\mathcal{A}}$ the estimators of $N_t, D_{\mathcal{A}}$, that is $\hat{N}_t = \frac{1}{M}\sum_{i=1}^M \mathbb{1}(T_{\mathcal{K},i} > t)$ and $\hat{D} = \frac{1}{M}\sum_{i=1}^M T_{\mathcal{A},i}$. We thus have $g(t) = \frac{N_t}{D}$ and $\hat{g}_t = \frac{\hat{N}_t}{\hat{D}}$. We also introduce $D_{\mathcal{K}} = \mathbb{E}^{\vec{\partial}\mathcal{A}^c} T_{\mathcal{K}}$ and $\hat{D}_{\mathcal{K}} = \frac{1}{M}\sum_{i=1}^M T_{\mathcal{K},i}$.

We are interested in bounding:

$$\mathbb{E}\int_0^\infty \left|\frac{\hat{N}_t}{\hat{D}_{\mathcal{A}}} - \frac{N_t}{D_{\mathcal{A}}}\right|$$

Using the triangle inequality $\left|\frac{\hat{N}_t}{\hat{D}_{\mathcal{A}}} - \frac{N_t}{D_{\mathcal{A}}}\right| \leq \left|\frac{\hat{N}_t}{\hat{D}_{\mathcal{A}}} - \frac{\hat{N}_t}{D_{\mathcal{A}}}\right| + \left|\frac{\hat{N}_t}{D_{\mathcal{A}}} - \frac{N_t}{D_{\mathcal{A}}}\right|$ we get:

$$\mathbb{E}\int_0^\infty \left|\frac{\hat{N}_t}{\hat{D}_{\mathcal{A}}} - \frac{N_t}{D_{\mathcal{A}}}\right| dt \leq \mathbb{E}\int_0^\infty \hat{N}_t \left|\frac{1}{\hat{D}_{\mathcal{A}}} - \frac{1}{D_{\mathcal{A}}}\right| dt + \mathbb{E}\int_0^\infty \frac{1}{D_{\mathcal{A}}}\left|\hat{N}_t - N_t\right| dt,$$

Using the formula $\int_0^\infty \hat{N}_t dt = \frac{1}{M}\sum_{i=1}^M T_{\mathcal{K},i} = \hat{D}_{\mathcal{K}}$ the first term on the right hand side of the above bound is equal to $\mathbb{E}\left|\frac{\hat{D}_{\mathcal{K}}}{\hat{D}_{\mathcal{A}}} - \frac{\hat{D}_{\mathcal{K}}}{D_{\mathcal{A}}}\right|$. To bound this quantity, we introduce an event $B = \{\hat{D}_{\mathcal{A}} < \frac{1}{2}D_{\mathcal{A}}\}$. We use the decomposition

$$\mathbb{E}\left|\frac{\hat{D}_{\mathcal{K}}}{\hat{D}_{\mathcal{A}}} - \frac{\hat{D}_{\mathcal{K}}}{D_{\mathcal{A}}}\right| = \mathbb{E}\mathbf{1}_B\left|\frac{\hat{D}_{\mathcal{K}}}{\hat{D}_{\mathcal{A}}} - \frac{\hat{D}_{\mathcal{K}}}{D_{\mathcal{A}}}\right| + \mathbb{E}\mathbf{1}_{B^c}\left|\frac{\hat{D}_{\mathcal{K}}}{\hat{D}_{\mathcal{A}}} - \frac{\hat{D}_{\mathcal{K}}}{D_{\mathcal{A}}}\right| \tag{14}$$

and bound each of the terms separately.

Since we always have $0 \leq \frac{\hat{D}_{\mathcal{K}}}{\hat{D}_{\mathcal{A}}} \leq 1$ and on the set B we have $0 \leq \frac{\hat{D}_{\mathcal{K}}}{D_{\mathcal{A}}} \leq \frac{1}{2}$, we have:

$$\mathbb{E}\mathbf{1}_B \left| \frac{\hat{D}_{\mathcal{K}}}{\hat{D}_{\mathcal{A}}} - \frac{\hat{D}_{\mathcal{K}}}{D_{\mathcal{A}}} \right| \leq \mathbb{P}(B).$$

Since the Markov Process X_t^{π} is aperiodic and irreducible and the state space $\mathcal{S}$ is finite, it is geometrically ergodic. It follows then that there exist constants $C, \lambda > 0$, such that $\mathbb{P}^{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{A}} > t) \leq C\exp(-\lambda t)$. Therefore, by Wainwright (2019)[Theorem 2.13], the random variable $\overline{T}_{\mathcal{A}}$ is subexponential. Furthermore $\mathbb{P}(B) \leq \mathbb{P}^{\vec{\partial}\mathcal{A}^c}\left(\left|\hat{D}_{\mathcal{A}} - D_{\mathcal{A}}\right| \geq \frac{1}{2}D_{\mathcal{A}}\right)$. From the concentration bound for the standard estimator of the mean of for subexponential variables Wainwright (2019)[Equation 2.18], it follows that the quantity $\mathbb{P}(B)$ is exponentially decaying with M, and is thus of order $\mathcal{O}\left(\frac{1}{M}\right)$.

To bound the second term in 14, we observe that the function $h(x) = 1/x$ is Lipschitz continuous on $[a, \infty)$ for any $a > 0$, with the Lipschitz constant $L_a = \sup_{x\in[a,\infty)} |h'(x)| = \frac{1}{a^2}$. Using this fact with $a = D_{\mathcal{A}}/2$, we have:

$$\mathbb{E}\mathbf{1}_{B^c}\hat{D}_{\mathcal{K}} \left| \frac{1}{\hat{D}_{\mathcal{A}}} - \frac{1}{D_{\mathcal{A}}} \right| \leq \frac{4}{D_{\mathcal{A}}^2}\mathbb{E}\hat{D}_{\mathcal{K}} \left| \hat{D}_{\mathcal{A}} - D_{\mathcal{A}} \right|.$$

Using the Cauchy-Schwartz inequality, we get:

$$\begin{aligned}
\mathbb{E}\hat{D}_{\mathcal{K}} \left| \hat{D}_{\mathcal{A}} - D \right| &\leq \left(\mathbb{E}\hat{D}_{\mathcal{K}}^2\right)^{1/2} \left(\mathbb{E}\left| \hat{D}_{\mathcal{A}} - D_{\mathcal{A}} \right|^2\right)^{1/2} \\
&\leq \left(\left(\mathbb{E}\hat{D}_{\mathcal{K}}\right)^2 + \mathrm{Var}(\hat{D}_{\mathcal{K}})\right)^{1/2} \left(\mathrm{Var}(\hat{D}_{\mathcal{A}})\right)^{1/2} \\
&= \frac{\left(\mathrm{Var}^{\vec{\partial}\mathcal{A}} T_{\mathcal{A}}\right)^{1/2}}{\sqrt{M}} \left((D_{\mathcal{K}})^2 + \frac{1}{M}\mathrm{Var}^{\vec{\partial}\mathcal{A}^c}(D_{\mathcal{K}})\right)^{1/2} \\
&\leq \frac{\left(\mathrm{Var}^{\vec{\partial}\mathcal{A}} T_{\mathcal{A}}\right)^{1/2} \mathbb{E}^{\vec{\partial}\mathcal{A}^c} T_{\mathcal{K}}}{\sqrt{M}} + \frac{\left(\mathrm{Var}^{\vec{\partial}\mathcal{A}} T_{\mathcal{A}}\right)^{1/2} \left(\mathrm{Var}^{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{K}})\right)^{1/2}}{M},
\end{aligned}$$

where in the last line we also used $\sqrt{a^2 + b^2} \leq a + b$ for non-negative a, b. We therefore obtain

$$\mathbb{E}\int_0^{\infty} \hat{N}_t \left| \frac{1}{\hat{D}_{\mathcal{A}}} - \frac{1}{D_{\mathcal{A}}} \right| dt \leq \frac{\left(\mathrm{Var}^{\vec{\partial}_{\mathcal{A}}} T_{\mathcal{A}}\right)^{1/2} \mathbb{E}^{\vec{\partial}\mathcal{A}^c} T_{\mathcal{K}}}{(\mathbb{E}^{\vec{\partial}A} T_{\mathcal{A}})^2 \sqrt{M}} + \mathcal{O}(\frac{1}{M})$$

We are left with bounding

$$\frac{1}{D_{\mathcal{A}}}\mathbb{E}\int_0^{\infty} \left| \hat{N}_t - N_t \right| dt.$$

Since $\hat{N}_t$ is an average of M Bernoulli random variables with mean N_t, we have:

$$\begin{aligned}
\frac{1}{D_{\mathcal{A}}}\mathbb{E}\int_0^{\infty} \left| \hat{N}_t - N_t \right| dt &= \frac{1}{D_{\mathcal{A}}}\int_0^{\infty} \mathbb{E}\left| \hat{N}_t - N_t \right| dt \\
&\leq \frac{1}{D_{\mathcal{A}}}\int_0^{\infty} \sqrt{\mathbb{E}\left| \hat{N}_t - N_t \right|^2} dt \\
&= \frac{1}{\sqrt{M} D_{\mathcal{A}}}\mathbb{E}\int_0^{\infty} \sqrt{N_t(1 - N_t)},
\end{aligned}$$

where in the first inequality we use $\mathbb{E}Y \leq \sqrt{EY^2}$ which follows from Cauchy-Schwartz inequality. We note, that $N_t = \mathbb{P}^{\vec{\partial}\mathcal{A}^c}(T_K > t) = 1 - \mathbb{P}^{\vec{\partial}\mathcal{A}^c}(T_K \leq t) = 1 - F_K(t)$. Thus we have

$$\frac{1}{D_{\mathcal{A}}}\mathbb{E}\int_0^{\infty} \left| \hat{N}_t - N_t \right| dt \leq \frac{1}{\sqrt{M}\mathbb{E}^{\vec{\partial}\mathcal{A}} T_{\mathcal{A}}} \int_0^{\infty} \sqrt{F_{\mathcal{K}}(t)(1 - F_{\mathcal{K}}(t))} dt.$$

Combining all of the above inequalities in an obvious manner, we obtain the bound from the thesis.

It follows from exponential ergodicity of X_t^π that $\overline{T}_\mathcal{A}$, $\overline{T}_\mathcal{K}$ have exponential tails, that is, there exist constants $C_\mathcal{A}, \lambda_\mathcal{A}, C_\mathcal{K}, \lambda_\mathcal{K}$ such that $\mathbb{P}^{\vec{\partial}\mathcal{A}}(T_\mathcal{A} > t) \leq C_\mathcal{A} \exp(-\lambda_\mathcal{A} t)$ and $\mathbb{P}^{\vec{\partial}\mathcal{A}^c}(T_\mathcal{K} > t) \leq C_\mathcal{K} \exp(-\lambda_K t)$. Therefore all the moments and the integral in the thesis are finite.

Furthermore, the random variables $\overline{T}_\mathcal{A}, \overline{T}_\mathcal{K}$ are approximately exponential, especially in the tails. When Y is an exponentially distributed variable, we have $\mathrm{Var}(Y) = (\mathbb{E}Y)^2$, and thus we have approximately $\left(\mathrm{Var}^{\vec{\partial}\mathcal{A}} T_\mathcal{A}\right)^{1/2} \approx \mathbb{E}^{\vec{\partial}A} T_\mathcal{A}$.

Also, if Y is an exponential distribution with parameter λ, we have $F_Y(T) = 1 - e^{\lambda t}$, and thus:

$$\begin{aligned}\int_0^\infty \sqrt{F_Y(t)(1-F_Y(t))}dt &\leq \int_0^\infty \sqrt{1-F_Y(t)}dt \\ &= \int_0^\infty e^{-\frac{\lambda t}{2}}dt = \frac{2}{\lambda} = 2\mathbb{E}Y.\end{aligned}$$

This concludes the argument about approximate values of the constant in the bound in the thesis. From those two approximations it follows, that the sum of constants next to the terms of order $\mathcal{O}\left(\frac{1}{\sqrt{M}}\right)$ is approximately equal to $3\frac{\mathbb{E}^{\vec{\partial}\mathcal{A}^c} T_\mathcal{K}}{\mathbb{E}^{\vec{\partial}\mathcal{A}} T_\mathcal{A}}$. ■

Appendix D. Estimation of the gradient of the average reward using FV

In this section we explain how the FV system can be leveraged to estimate the average reward gradients in the policy gradient algorithm, stated in expression (6) in Section 3.5, in two steps:
(1) Estimate for $x \in \mathcal{A}^c$, $\eta(x) = \sum_a Q_\theta(x,a)\nabla_\theta \pi_\theta(a|x)$.
(2) Given an estimate $\hat{\eta}$ of η, use the FV procedure to estimate $\sum_{x\in\mathcal{A}^c} \hat{\eta}(x) p^{\pi_\theta}(x)$.

The estimation of η can be done using properly coupled copies of the trajectories outside $\mathcal{A}$ (used to run the FV system). Although coupling is not strictly necessary, it is a way of accelerating the estimation of the sum in (1) (which, for instance, becomes simply the difference of two Q values in the queue blocking example presented in Section 4.2) when the trajectories meet before the maximum time allowed by the implementation to estimate each Q value separately. It also allows us to obtain an unbiased estimator of an infinite sum, while the naïve estimator might be biased.

More precisely, the estimation method is as follows: we define a Markov chain on an extended state-action space $A \times S$, where A is the set of possible actions over all possible states and S is the set of states. For each state in set C we denote by A_s the set of actions available at state x. We then run $|A_s|$ copies of the Markov chain on the extended state space, each starting at (a_i, x) for different $a_i \in A_s$. When two such chains meet, they continue evolving together forever. Under this procedure, we can see that after all chains meet, the contribution to $\sum_a Q_\theta(x,a)\nabla_\theta\pi_\theta(a|x)$ is zero, because all values contributing to $Q_\theta(x,a)$ are the same for all a for the given x and the derivatives of the policy sum up to zero. Therefore we only need to run the chains until this point. Also, before all the chains meet, the terms $v^{\pi_\theta}(x)$ in (5) cancel out, and therefore we only need to record the observed rewards R_t.