



**HAL**  
open science

# On the irreducibility and convergence of a class of nonsmooth nonlinear state-space models on manifolds

Armand Gissler, Alain Durmus, Anne Auger

## ► To cite this version:

Armand Gissler, Alain Durmus, Anne Auger. On the irreducibility and convergence of a class of nonsmooth nonlinear state-space models on manifolds. 2024. hal-04713677

**HAL Id: hal-04713677**

**<https://inria.hal.science/hal-04713677v1>**

Preprint submitted on 30 Sep 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# On the irreducibility and convergence of a class of nonsmooth nonlinear state-space models on manifolds

Armand Gissler<sup>1,2</sup>, Alain Durmus<sup>2</sup>, and Anne Auger<sup>1,2</sup>

<sup>1</sup>Inria Saclay Île-de-France, Palaiseau, France

<sup>2</sup>CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, Palaiseau, France

February 6, 2024

## Abstract

In this paper, we analyze a large class of general nonlinear state-space models on a state-space  $\mathsf{X}$ , defined by the recursion  $\phi_{k+1} = F(\phi_k, \alpha(\phi_k, U_{k+1}))$ ,  $k \in \mathbb{N}$ , where  $F, \alpha$  are some functions and  $\{U_{k+1}\}_{k \in \mathbb{N}}$  is a sequence of i.i.d. random variables. More precisely, we extend conditions under which this class of Markov chains is irreducible, aperiodic and satisfies important continuity properties, relaxing two key assumptions from prior works. First, the state-space  $\mathsf{X}$  is supposed to be a smooth manifold instead of an open subset of a Euclidean space. Second, we only suppose that  $F$  is locally Lipschitz continuous.

We demonstrate the significance of our results through their application to Markov chains underlying optimization algorithms. These schemes belong to the class of evolution strategies with covariance matrix adaptation and step-size adaptation.

**Keywords:** Markov chains, irreducibility, aperiodicity, T-chain, deterministic control model, CMA-ES.

## 1 Introduction

Consider a nonlinear state-space model defined by the recursion:

$$\phi_{k+1} = G(\phi_k, \xi_{k+1}), \tag{1.1}$$

where the sequence  $\{\xi_{k+1}\}_{k \in \mathbb{N}}$  consists of independent and identically distributed (i.i.d.) random variables,  $G : \mathsf{X} \times \mathsf{W} \rightarrow \mathsf{X}$  is a continuous function, and  $\mathsf{X}, \mathsf{W}$  are two measurable spaces. Nonlinear state-space models (1.1) form a class of Markov chains that have been first popularized in stochastic control theory [27, 23, 24, 25]. This has spurred extensive analysis and has a well-established historical context. In particular, for nonlinear autoregressive models, i.e., where  $G$  can be written as  $G(x, u) = \tilde{G}(x) + u$ , ergodicity has been widely investigated [8, 5, 34, 28, 13]. Moreover, connections have been established between the stability of (1.1) and the one of some Ordinary Differential Equation (ODE) [19]. The idea of analyzing (1.1) from the perspective of control theory, where  $u$  is regarded as a control parameter, was initially proposed in [31] within the context of diffusion processes. This approach was subsequently employed with success in [20] and [21]. It has been then applied in [25, 26] when  $G$  is infinity differentiable and  $\mathsf{X}, \mathsf{W}$  are open sets of Euclidean spaces, to establish the irreducibility, aperiodicity and topological properties of the Markov kernel associated to (1.1).

Besides stochastic control models, (1.1) also encompasses many algorithms in optimization and Markov chain Monte Carlo algorithms. The variable  $\phi_k$  corresponds to the state of an algorithm at iteration  $k$  and  $U_{k+1}$  represents the random components used to update this state. However, for certain classes of algorithms, especially those arising from zeroth-order optimization, the function  $G$  may not be continuous. Nevertheless, it can be written using the Markov chain model introduced in [10] as

$$\phi_{k+1} = F(\phi_k, \alpha(\phi_k, U_{k+1})), \tag{1.2}$$

for some continuous function  $F : \mathsf{X} \times \mathsf{W} \rightarrow \mathsf{X}$  and a potentially discontinuous function  $\alpha : \mathsf{X} \times \mathsf{U} \rightarrow \mathsf{W}$  for some measurable space  $\mathsf{U}$  and  $\{U_{k+1}\}_{k \in \mathbb{N}}$  a sequence of i.i.d. random variables valued in  $\mathsf{U}$ . Extensions of the connection

between the  $\varphi$ -irreducibility and the stability of the associated deterministic control model [26, Chapter 7] to Markov chains following (2.1) have been established in [10].

A particularly relevant algorithm of the form (1.2) in Evolution Strategies (ES) is ES with Covariance Matrix Adaptation (CMA-ES) [18, 17] often regarded as the state-of-the-art algorithm for numerical derivative-free optimization of difficult problems with tremendous applications in many domains (e.g., in biology [9, 30], medicine [29], machine learning [1, 15])<sup>1</sup>. Yet, while we have ample empirical evidences of its linear convergence on wide classes on functions, a convergence proof together with a convergence rate is still an open question. In order to extend linear convergence results from step-size adaptive ES [7, 32] to CMA-ES, a first step is to show the irreducibility and topological properties of the kernel associated to the underlying Markov chain. However, previous works [25, 10] cannot be applied since (i) the state space  $X$  of this chain is a smooth manifold whereas previous analysis supposed that they were open subsets of a Euclidean space (ii) the function  $F$  is supposed to be continuously differentiable in existing results while certain step-size updates on which CMA-ES relies are only locally Lipschitz. The purpose of the present paper is to resolve these two limitations and pave the way to a complete convergence analysis of CMA-ES.

In this context, the objective of this paper is to extend these results [10] to the case where the sets  $X$  and  $W$  are smooth manifolds and where  $F$  is locally Lipschitz instead of continuously differentiable.

The paper is organized as follows. In Section 2.1, we provide a precise definition of the class of nonlinear state-space models under investigation. In Sections 2.1.1 and 2.1.2, we illustrate and motivate our analysis through two ES algorithms—one employing covariance matrix adaptation and the other utilizing step-size adaptation. Next, in Section 2.1.3, we outline the assumptions necessary for establishing our main results and deriving the irreducibility and aperiodicity of our model. Our main results are presented in Section 2.2 and are subsequently applied in Section 3 to the two algorithms introduced in Sections 2.1.1 and 2.1.2. Finally, proofs are gathered in Section 4. Note that some of the proofs and useful definitions are given in the appendix.

## 2 Main results

### 2.1 The model and assumptions

Let  $X, W$  be two (smooth, connected) manifolds (see Definition A.1) of dimensions  $n$  and  $p$  respectively, endowed with their Borel  $\sigma$ -fields denoted by  $\mathcal{B}(X)$  and  $\mathcal{B}(W)$  respectively. We let  $\text{dist}_X$  and  $\text{dist}_W$  be two distance functions on  $X$  and  $W$  which induce the topology of  $X$  and  $W$  respectively. As a consequence of [22, Proposition 13.2, Theorem 13.29], such distance functions always exist.

We consider in this paper Markov chains taking values in  $X$  and associated with the general recursion

$$\phi_{k+1} = F(\phi_k, \alpha(\phi_k, U_{k+1})) \quad (2.1)$$

where  $F : X \times W \rightarrow X$  and  $\alpha : X \times U \rightarrow W$  are measurable functions, and  $\{U_{k+1}\}_{k \in \mathbb{N}}$  is a sequence of i.i.d. random variables valued in a measurable space  $(U, \mathcal{U})$ , chosen independently of the initial state  $\phi_0$ . Throughout the paper, we denote by  $P$  the Markov kernel associated to (2.1). As emphasized in the introduction, this class of models is a natural extension of nonlinear state-space models defined on manifolds. To illustrate the interest of such processes, we provide the following examples.

#### 2.1.1 An instructive example: CMA-ES

We introduce here a simplified version of the numerical optimization algorithm called evolution strategy with covariance matrix adaptation (CMA-ES) [18, 17], which, for an objective function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , aims to solve:

$$\text{find } x^* \in \underset{x \in \mathbb{R}^d}{\text{Argmin}} f(x). \quad (\text{P})$$

To this end, it approximates the optimum  $x^*$  of the objective function  $f$  by a multivariate normal distribution  $\mathcal{N}(m_k, C_k)$  for a mean  $m_k \in \mathbb{R}^d$  and a covariance matrix  $C_k \in \mathcal{S}_{++}^d$  that are updated iteratively. More precisely, for each  $k \in \mathbb{N}$ , given  $m_k \in \mathbb{R}^d$  and  $C_k \in \mathcal{S}_{++}^d$ , the algorithm can be described as follows. First, a population of  $\lambda \geq 2$  offspring is sampled using

$$U_{k+1}^1, \dots, U_{k+1}^\lambda \sim \mathcal{N}(0, I_d) \quad \text{i.i.d. and independently of } (m_k, C_k), \quad (2.2)$$

---

<sup>1</sup>As of September 2023, the two main Python implementations of the CMA-ES algorithm `cma` and `cmaes` have more than 5 millions and 45 millions downloads respectively.

so that, conditionally to  $(m_k, C_k)$ , the offspring satisfy  $m_k + \sqrt{C_k}U_{k+1}^i \sim \mathcal{N}(m_k, C_k)$ , for  $i = 1, \dots, \lambda$ . Next, we rank the offspring so that we define a permutation  $s_{k+1} \in \mathfrak{S}_\lambda$  of  $\{1, \dots, \lambda\}$  satisfying

$$f\left(m_k + \sqrt{C_k}U_{k+1}^{s_{k+1}(1)}\right) \leq \dots \leq f\left(m_k + \sqrt{C_k}U_{k+1}^{s_{k+1}(\lambda)}\right). \quad (2.3)$$

Then, given the  $\mu \in \{1, \dots, \lambda\}$  best offspring, the mean is moved towards the best solutions with the following update

$$m_{k+1} = m_k + \sqrt{C_k} \sum_{i=1}^{\mu} w_i U_{k+1}^{s_{k+1}(i)} \quad (2.4)$$

and the covariance matrix update reads

$$C_{k+1} = (1-c)C_k + c\sqrt{C_k} \left( \sum_{i=1}^{\mu} w_i \left( U_{k+1}^{s_{k+1}(i)} \right) \left( U_{k+1}^{s_{k+1}(i)} \right)^\top \right) \sqrt{C_k}. \quad (2.5)$$

It increases the likelihood to sample in the directions where good solutions were found. In the above equations, the weights  $w_1 \geq \dots \geq w_\mu > 0$  satisfy  $\sum_{i=1}^{\mu} w_i = 1$ , and we call  $c \in (0, 1)$  the learning rate for the covariance matrix. In ES, the function values are not used explicitly to update the state variables. It influences the update only through the ranking of candidate solutions via the permutation  $s_{k+1}$ . Consequently, the algorithms are invariant with respect to strictly increasing transformations of the objective function (that preserve the ranking). In this context, a natural class of functions to analyze the convergence of ES are scaling-invariant functions [7, 33]. A function  $f$  is said to be scaling-invariant w.r.t.  $x^*$  if, for every  $x, y \in \mathbb{R}^d$  and  $\rho > 0$ , we have

$$f(x + x^*) \leq f(y + x^*) \Leftrightarrow f(\rho x + x^*) \leq f(\rho y + x^*). \quad (2.6)$$

Convergence of step-size adaptive ES on scaling-invariant functions with smooth level sets was established –for specific assumptions on the step-size update– in previous work [32]. Assuming that the objective function  $f$  satisfies (2.6), we define then the following quantities

$$z_k = \frac{m_k - x^*}{\sqrt{R(C_k)}} \quad ; \quad \Sigma_k = \frac{C_k}{R(C_k)} \quad (2.7)$$

where  $R = \det(\cdot)^{1/d} : \mathcal{S}_{++}^d \rightarrow \mathbb{R}_+$ . We assume w.l.o.g. that  $x^* = 0$ . Then the sequence  $\{(z_k, \Sigma_k)\}_{k \in \mathbb{N}}$  defines a time-homogeneous Markov chain which obeys to the model (2.1), see Proposition C.1, with  $\mathbf{X} = \mathbb{R}^d \times R^{-1}(\{1\})$ ,  $\mathbf{U} = \mathbb{R}^{d \times \lambda}$ ,  $\mathbf{W} = \mathbb{R}^{d \times \mu}$ , and

$$F: \mathbf{X} \times \mathbf{W} \rightarrow \mathbf{X} \\ ((z, \Sigma), (v_1, \dots, v_\mu)) \mapsto \left( \frac{z + \sqrt{\Sigma} \sum_{i=1}^{\mu} w_i v_i}{R^{1/2}(K(\Sigma, v_1, \dots, v_\mu))}, \frac{K(\Sigma, v_1, \dots, v_\mu)}{R(K(\Sigma, v_1, \dots, v_\mu))} \right) \quad (2.8)$$

where

$$K(\Sigma, v_1, \dots, v_\mu) = (1-c)\Sigma + c\sqrt{\Sigma} \left( \sum_{i=1}^{\mu} w_i v_i v_i^\top \right) \sqrt{\Sigma},$$

and with

$$\alpha: \mathbf{X} \times \mathbf{U} \rightarrow \mathbf{W} \\ ((z, \Sigma), (u_1, \dots, u_\lambda)) \mapsto (u_{s(1; z, \Sigma, u_{1:\lambda})}, \dots, u_{s(\mu; z, \Sigma, u_{1:\lambda})}) \quad (2.9)$$

where given  $u_{1:\lambda} = (u_1, \dots, u_\lambda) \in (\mathbb{R}^d)^\lambda$ ,  $z \in \mathbb{R}^d$  and  $\Sigma \in \mathcal{S}_{++}^d$ , we denote by  $s(\cdot; z, \Sigma, u_{1:\lambda})$  a permutation that sorts the  $f(z + \sqrt{\Sigma}u_i)$ ,  $i = 1, \dots, \lambda$ . To ensure uniqueness of this permutation, we impose a tie-break, e.g., if  $i < j$  are such that  $f(z + \sqrt{\Sigma}u_i) = f(z + \sqrt{\Sigma}u_j)$ , then  $s(\cdot; z, \Sigma, u_{1:\lambda})^{-1}(i) < s(\cdot; z, \Sigma, u_{1:\lambda})^{-1}(j)$ . Note that  $\mathbf{X}$  is not an open subset of a Euclidean space (but it is a smooth manifold by the preimage theorem, see e.g. [14, Chapter 1, Section 4]), hence the results in [10] do not apply. We show in Section 3 that our results apply and we prove that  $\{(z_k, \Sigma_k)\}_{k \in \mathbb{N}}$  defines a  $\varphi$ -irreducible aperiodic T-chain and that all compact subsets of  $\mathbf{X}$  are small.

If we establish moreover that the chain  $\{(z_k, \Sigma_k)\}_{k \in \mathbb{N}}$  is positive recurrent, then we obtain that CMA-ES behaves linearly, as stated below.

**Theorem 2.1.** Consider a scaling-invariant function with respect to  $x^*$  and the Markov chain  $\{(z_k, \Sigma_k)\}_{k \in \mathbb{N}}$  defined in (2.7) ensuing from CMA-ES minimizing  $f$ . Suppose that  $\{(z_k, \Sigma_k)\}_{k \in \mathbb{N}}$  is a  $\varphi$ -irreducible aperiodic positive recurrent chain with invariant probability measure  $\pi$ . If the function  $(z, \Sigma) \mapsto \log \|z\|$  is  $\pi$ -integrable on  $\mathbb{R}^d \times \mathcal{S}_{++}^d$ , then almost surely we have

$$\lim_{k \rightarrow \infty} \frac{1}{k} \log \frac{\|m_k - x^*\|}{\|m_0 - x^*\|} = -\text{CR} \in \mathbb{R}. \quad (2.10)$$

When moreover  $\text{CR} > 0$ , we say that CMA-ES converges linearly to  $x^*$ .

*Proof.* Assume that  $x^* = 0$ . Since CMA-ES is invariant by translation [6], (2.10) would generalize to any value of  $x^*$ . Since  $\{(z_k, \Sigma_k)\}_{k \in \mathbb{N}}$  is supposed to be  $\varphi$ -irreducible, aperiodic and positive recurrent, by [26, Theorem 17.0.1], we know that for all  $\pi$ -integrable function  $g$ , we have that

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{k=0}^{T-1} g(z_k, \Sigma_k) = \int g(z, \Sigma) d\pi(z, \Sigma). \quad (2.11)$$

However, we have

$$\begin{aligned} \frac{1}{T} \log \frac{\|m_T\|}{\|m_0\|} &= \frac{1}{T} \sum_{k=0}^{T-1} (\log \|m_{k+1}\| - \log \|m_k\|) \\ &= \frac{1}{T} \sum_{k=0}^{T-1} (\log \|z_{k+1}\| - \log \|z_k\|) + \end{aligned} \quad (2.12)$$

$$\frac{1}{2dT} \sum_{k=0}^{T-1} \log \det \left( \underbrace{(1-c)\Sigma_k + \sqrt{\Sigma_k} \left( \sum_{i=1}^{\mu} w_i \left( U_{k+1}^{s_{k+1}(i)} \right) \left( U_{k+1}^{s_{k+1}(i)} \right)^\top \right) \sqrt{\Sigma_k}}_{=: \tilde{\Sigma}_{k+1}} \right). \quad (2.13)$$

But, by assumption,  $(z, \Sigma) \mapsto \log \|z\|$  is  $\pi$ -integrable. Moreover,  $\det(\Sigma_k) = 1$ , hence

$$\det(\tilde{\Sigma}_{k+1}) = \det \left( (1-c)I_d + c \sum_{i=1}^{\mu} w_i \left( U_{k+1}^{s_{k+1}(i)} \right) \left( U_{k+1}^{s_{k+1}(i)} \right)^\top \right).$$

Moreover,

$$1 - c \leq \det \left( (1-c)I_d + c \sum_{i=1}^{\mu} w_i \left( U_{k+1}^{s_{k+1}(i)} \right) \left( U_{k+1}^{s_{k+1}(i)} \right)^\top \right)^{1/d} \leq 1 - c + c \max_{i=1, \dots, \mu} \|U_{k+1}^i\|^2$$

which defines an integrable quantity, since the vectors  $U_{k+1}^i$ ,  $k \in \mathbb{N}$ ,  $i = 1, \dots, \lambda$ , are standard Gaussian vectors of  $\mathbb{R}^d$ . Applying (2.11) to (2.12) and (2.13), we find the stated result with

$$\text{CR} = -\frac{1}{2d} \mathbb{E}_{(z, \Sigma) \sim \pi} \left[ \det \left( (1-c)I_d + c \sum_{i=1}^{\mu} w_i \left( U_1^{s(i; z, \Sigma, U_1^{1:\lambda})} \right) \left( U_1^{s(i; z, \Sigma, U_1^{1:\lambda})} \right)^\top \right) \right]. \quad (2.14)$$

□

The previous theorem illustrates how the  $\varphi$ -irreducibility and aperiodicity of  $\{(z_k, \Sigma_k)\}_{k \in \mathbb{N}}$  is instrumental to obtain linear convergence of CMA-ES.<sup>2</sup>

<sup>2</sup>The variant of CMA-ES presented here differs significantly from the default CMA-ES (used in applications) where both step-size adaptation and covariance matrix adaptation are used. In addition, the covariance matrix update presents an additional mechanism (rank-one update). The combination of all the mechanisms is important to obtain fast convergence in many situations. This variant with however a learning rate on the mean update has been analyzed in previous theoretical works [3], and it has been proven to be a discretized version of a natural gradient update on the manifold of probability distributions [2].

### 2.1.2 A nonsmooth example: a step-size adaptive ES

We present here an other simplification of CMA-ES where instead of adapting a full covariance matrix, a scaling factor called step-size is adapted such that the covariance matrix reads  $\sigma_k^2 I_d$ . In this step-size adaptive algorithm, the optimum  $x^* \in \mathbb{R}^d$  of the problem (P) is approximated by a multivariate normal distribution  $\mathcal{N}(m_k, \sigma_k^2 I_d)$ , where the mean  $m_k \in \mathbb{R}^d$  and the step-size  $\sigma_k > 0$  are updated as follows. For  $k \in \mathbb{N}$ , given a mean  $m_k \in \mathbb{R}^d$  and a step-size  $\sigma_k > 0$ , we sample  $U_{k+1}^1, \dots, U_{k+1}^\lambda$ , rank them by defining the permutation  $s_{k+1} \in \mathfrak{S}_\lambda$  and update the mean  $m_{k+1}$  according to (2.2), (2.3), (2.4), respectively, where we replace  $C_k$  by  $\sigma_k^2 I_d$ . The step-size update obeys

$$\sigma_{k+1} = \sigma_k \times \exp \left( \frac{1}{d_\sigma} \left( \frac{\sqrt{\mu_{\text{eff}}} \left\| \sum_{i=1}^\mu w_i U_{k+1}^{s_{k+1}(i)} \right\|}{\mathbb{E} \|\mathcal{N}(0, I_d)\|} - 1 \right) \right) \quad (2.15)$$

where we define  $\mu_{\text{eff}} = \sum_{i=1}^\mu w_i^2$  and fix  $d_\sigma > 0$  (usually  $d_\sigma \approx 1$ ). Moreover, as in Section 2.1.1, we assume  $f$  to be scaling-invariant, see (2.6). W.l.o.g. we suppose that  $f$  is scaling-invariant w.r.t.  $x^* = 0$ . Then, by defining

$$z_k = \frac{m_k - x^*}{\sigma_k}, \quad (2.16)$$

we get that the sequence  $\{z_k\}_{k \in \mathbb{N}}$  is a time-homogeneous Markov chain which obeys to the model (2.1) (see [32, Proposition 4]) with  $\mathbf{X} = \mathbb{R}^d$ ,  $\mathbf{U} = \mathbb{R}^{d \times \lambda}$ ,  $\mathbf{W} = \mathbb{R}^{d \times \mu}$ ,

$$F: \mathbf{X} \times \mathbf{W} \rightarrow \mathbf{X} \\ (z, (v_1, \dots, v_\mu)) \mapsto \left( z + \sum_{i=1}^\mu w_i v_i \right) \times \exp \left( -\frac{1}{d_\sigma} \left( \frac{\sqrt{\mu_{\text{eff}}} \left\| \sum w_i v_i \right\|}{\mathbb{E} \|\mathcal{N}(0, I_d)\|} - 1 \right) \right) \quad (2.17)$$

and

$$\alpha: \mathbf{X} \times \mathbf{U} \rightarrow \mathbf{W} \\ (z, (u_1, \dots, u_\lambda)) \mapsto (u_{s(1; z, I_d, u_{1:\lambda})}, \dots, u_{s(\lambda; z, I_d, u_{1:\lambda})}) \quad (2.18)$$

where we define the permutation  $s(\cdot; z, I_d, u_{1:\lambda})$  as in Section 2.1.1. Here,  $F$  is not continuously differentiable, and we cannot use the results of [10] to analyze this chain. However the stability of an alternative strategy where (2.15) is replaced by a smooth update of the step-size has already been analyzed [32].

### 2.1.3 Assumptions

We consider the following assumptions on the functions  $F$  and  $\alpha$  to establish ergodicity of the Markov kernel defined via (2.1):

**H1.** For any  $x \in \mathbf{X}$ , the distribution  $\mu_x$  of the random variable  $\alpha(x, U_1)$  admits a density, denoted by  $p_x$ , with respect to a  $\sigma$ -finite measure  $\zeta_{\mathbf{W}}$ , such that:

- (i) The function  $(x, w) \mapsto p_x(w)$  is lower semicontinuous, i.e., for any  $(\bar{x}, \bar{w}) \in \mathbf{X} \times \mathbf{W}$ , we have  $\liminf p_x(w) \geq p_{\bar{x}}(\bar{w})$  when  $(x, w) \rightarrow (\bar{x}, \bar{w})$ .
- (ii) For  $A \subset \mathcal{B}(\mathbf{W})$ ,  $\zeta_{\mathbf{W}}(A) = 0$  if and only if  $A$  is negligible, i.e.,  $\text{Leb}(\varphi(A \cap U)) = 0$  for any chart  $(\varphi, U)$  of  $\mathbf{W}$ , where  $\text{Leb}$  stands for the Lebesgue measure.

The condition **H1** is a generalization of [10, A4], where  $W \subset \mathbb{R}^p$  was instead an open subset of an Euclidean space and  $\zeta_{\mathbf{W}}$  a Lebesgue measure. Besides, in that context, **H1**(i) has already been considered in [10]. If  $\mathbf{W}$  is equipped with a smooth Riemannian metric which makes  $W$  a Riemannian manifold, a  $\sigma$ -finite measure satisfying **H1**(ii) would be the Lebesgue-Riemann volume measure [4, Chapter XII and Proposition XII.1.6].

We assume moreover the following on the map  $F: \mathbf{X} \times \mathbf{W} \rightarrow \mathbf{X}$ .

**H2.**  $F: \mathbf{X} \times \mathbf{W} \rightarrow \mathbf{X}$  is locally Lipschitz, see Definition B.3, on  $\mathbf{X} \times \mathbf{W}$  with respect to the distance  $\text{dist}_{\mathbf{X}} \oplus \text{dist}_{\mathbf{W}}$ , defined by  $\text{dist}_{\mathbf{X}} \oplus \text{dist}_{\mathbf{W}}((x, w), (x', w')) = \text{dist}_{\mathbf{X}}(x, x') + \text{dist}_{\mathbf{W}}(w, w')$  for every  $((x, w), (x', w')) \in (\mathbf{X} \times \mathbf{W})^2$ .

This assumption encompasses the continuous differentiability assumption made in [10], in the case where  $\mathbf{X}$  and  $\mathbf{W}$  are open subsets of some Euclidean spaces. Indeed, any continuously differentiable function is in particular locally Lipschitz. For example, the update (2.17), introduced in Section 2.1.2, satisfies the assumption **H2**.

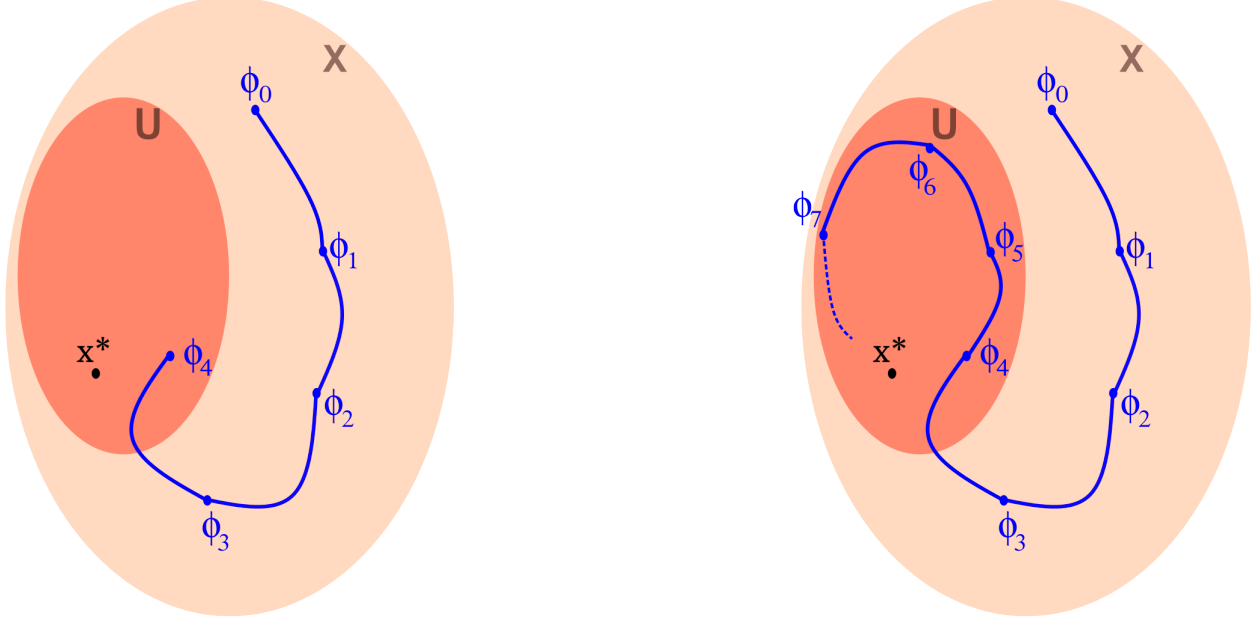


Figure 1: Left: Illustration of a globally attracting state  $x^*$ , for any neighborhood  $U$  of  $x^*$  and any starting state  $\phi_0$ , there exists a  $k$ -steps path from  $\phi_0$  to  $U$ . Right: Illustration of a steadily attracting state  $x^*$ , for any neighborhood  $U$  of  $x^*$  and any starting state  $\phi_0$ , there exist  $T > 0$  and  $k$ -steps paths from  $\phi_0$  to  $U$  for every  $k \geq T$ .

For our last assumption regarding the functions  $F$  and  $\alpha$ , we need to introduce further notations and notions introduced in [26, 10]. The *extended transition map*  $S_x^k: \mathbb{W}^k \rightarrow \mathbb{X}$  can be defined inductively via

$$S_x^{k+1}(w_{1:k+1}) := F(S_x^k(w_{1:k}), w_{k+1}), \quad S_x^0 := x, \quad (2.19)$$

for  $k \in \mathbb{N}$ ,  $x \in \mathbb{X}$  and  $w_{1:k+1} = (w_1, \dots, w_{k+1}) \in \mathbb{W}^{k+1}$ . The value  $S_x^k(w_{1:k})$  corresponds to the  $k^{\text{th}}$  iterate of the chain  $\phi_k$  defined via (2.1), conditionally to  $\phi_0 = x$  and  $\alpha(\phi_t, U_{t+1}) = w_t$  for  $t = 0, \dots, k-1$ . Remark that, by composition, if  $F$  is continuous (ly locally Lipschitz), then so is  $(x, w_{1:k}) \mapsto S_x^k(w_{1:k})$ . Similarly, we define the extended probability density  $p_x^k$  via

$$p_x^{k+1}(w_{1:k+1}) := p_x^k(w_{1:k}) p_{S_x^k(w_{1:k})}(w_{k+1}), \quad p_x^1(w_1) := p_x(w_1). \quad (2.20)$$

The function  $p_x^k$  is then the density of the random variable  $(\alpha(\phi_0, U_1), \dots, \alpha(\phi_{k-1}, U_k))$ , with  $\phi_0 = x$ , w.r.t. the product measure  $\zeta_{\mathbb{W}}^{\otimes k}$ . If  $(x, w) \mapsto p_x(w)$  is l.s.c., then  $(x, w_{1:k}) \mapsto p_x^k(w_{1:k})$  is l.s.c. as well. In this case, the control sets

$$\mathcal{O}_x^k := \{w_{1:k} \in \mathbb{W}^k \mid p_x^k(w_{1:k}) > 0\} \quad (2.21)$$

are nonempty open subsets of  $\mathbb{W}^k$ . The control set  $\mathcal{O}_x^k$  corresponds to the set of paths  $w_{1:k}$  starting at  $x$  which have positive density  $p_x^k(w_{1:k})$ .

Moreover, for  $x \in \mathbb{X}$ ,  $A$  a measurable subset of  $\mathbb{X}$ , and  $k > 0$ , we say that  $w_{1:k} \in \mathbb{W}^k$  is a  *$k$ -steps path* from  $x$  to  $A$  if  $w_{1:k} \in \mathcal{O}_x^k$  and  $S_x^k(w_{1:k}) \in A$ , implying that  $A$  is then reachable by  $P$  from  $x$ . A point  $x^* \in \mathbb{X}$  is said to be a *globally attracting state* if for any  $y \in \mathbb{X}$  and any neighborhood  $U$  of  $x^*$ , there exist  $k > 0$  and a  $k$ -steps path between  $y$  and  $U$  (the original definition of a globally attracting state is actually given in (4.3) and we show in Proposition 4.1 the equivalence with this latter statement). It is said to be *steadily attracting* if for any  $y \in \mathbb{X}$  and any neighborhood  $U$  of  $x^*$ , there exists  $T > 0$  such that for every  $k \geq T$ , we can find a  $k$ -steps path between  $y$  and  $U$ . Note that any steadily attracting state is in particular globally attracting. These two notions are illustrated in Figure 1. Assuming **H1** and **H2**, as emphasized in Theorem 4.15, we show that the kernel  $P$  defined via (2.1) is  $\varphi$ -irreducible exhibiting the existence of a globally attracting state (it is in fact an equivalence). On a related note, we deduce in Theorem 4.17 that the existence of a steadily attracting state is equivalent to the  $\varphi$ -irreducibility and aperiodicity of  $P$ .

We introduce now the notation  $\partial f$  for the *Clarke's generalized Jacobian* of a locally Lipschitz function  $f: X \rightarrow Y$  between two manifolds  $X$  and  $Y$ . These Jacobians have been defined in [11], and we recall the definition in the Euclidean case in Definition B.6. For the sake of completeness, we define here and give basic properties in Appendix B of the Clarke's Jacobian for functions defined on manifolds.

*Proposition and Definition 2.2* (Clarke's generalized Jacobian on manifolds). Let  $X$  and  $Y$  be two manifolds and  $f: X \rightarrow Y$  be locally Lipschitz at  $x_0 \in X$ . Let  $(\varphi, U)$  be a local chart of  $X$  around  $x_0$  and  $(\psi, V)$  be a local chart of  $Y$  around  $f(x_0)$ . Define  $g = \psi \circ f \circ \varphi^{-1}$ . Then  $g: \varphi(U) \rightarrow \psi(V)$  is locally Lipschitz at  $\varphi(x_0)$ , and we can define

$$\partial f(x_0) = \{ \mathcal{D}\psi^{-1}(g \circ \varphi(x_0)) \circ h \circ \mathcal{D}\varphi(x_0) \mid h \in \partial g(\varphi(x_0)) \}, \quad (2.22)$$

where  $\mathcal{D}$  denotes the usual differential operator, and  $\partial$  the Clarke differential operator. This definition does not depend on the choice of the charts  $(\varphi, U)$  and  $(\psi, V)$ .

*Proof.* See Appendix B. □

In the case of a differentiable function  $f$ , the definition of Clarke's generalized Jacobian corresponds to the definition of the Jacobian, i.e.,  $\partial f(x) = \{ \mathcal{D}f(x) \}$ . The notion of Clarke's generalized Jacobian is used to formulate the controllability condition for an element  $x \in X$ :

$$\text{there exists } w_{1:k} \in \overline{\mathcal{O}_x^k} \text{ such that } \partial S_x^k(w_{1:k}) \text{ is of maximal rank.} \quad (C_x)$$

Note that here,  $\partial S_x^k(w_{1:k})$  is of maximal rank, is understood as any element of  $\partial S_x^k(w_{1:k})$  is of rank  $n$ , the dimension of  $X$ . In Corollary 4.13, we show that  $P$  is a T-chain assuming that condition  $(C_x)$  holds for every state  $x \in X$ . In comparison to **H3** below, we do not assume that states  $x$  for which  $(C_x)$  holds are globally attracting. However, we show that if  $(C_{x^*})$  holds for  $x^*$  a globally attracting state, then it holds for every state in  $X$ , see Proposition 4.8.

**H3.** *The controllability condition  $(C_{x^*})$  is satisfied for a globally attracting state  $x^*$ .*

Alternatively, if we want to prove aperiodicity on top of  $\varphi$ -irreducibility, we assume instead the following.

**H4.** *The controllability condition  $(C_{x^*})$  is satisfied for a steadily attracting state  $x^*$ .*

Remark that **H4** implies **H3**. Assumptions **H3-H4** also appear in [10] but with the additional condition that the functions  $S_x^k$  are continuously differentiable for  $x \in X$  and  $k > 0$ , condition that we relax here. Globally and steadily attracting states are characterized by Proposition 4.1 and Proposition 4.4(ii) respectively below. In Section 3, we give one example of a smooth model on manifolds, and one example of a nonsmooth model on a Euclidean space, for which we show that **H4** holds.

## 2.2 Main results

Before stating our main results, we introduce concepts that are needed for their statements. Given  $P$  a Markov kernel on  $(X, \mathcal{B}(X))$ , we define  $P^1 = P$  and for  $k \geq 1$ ,  $x \in X$  and  $A \in \mathcal{B}(X)$ ,  $P^{k+1}(x, A) = \int P(y, A) P^k(x, dy)$ . We say that  $P$  is  $\varphi$ -irreducible when there exists a nontrivial measure  $\varphi$  on  $\mathcal{B}(X)$  such that for any  $A \in \mathcal{B}(X)$  with  $\varphi(A) > 0$ , we have  $\sum_{k \geq 1} P^k(x, A) > 0$  for every  $x \in X$ . Let  $b$  be a probability distribution on  $\mathbb{N}$ , and let  $K_b$  be the transition kernel defined by  $K_b(x, A) := \sum_{k \geq 0} b(k) P^k(x, A)$ . A substochastic transition kernel  $T$  with  $K_b \geq T$  such that  $x \mapsto T(x, A)$  is lower semicontinuous for every  $A \in \mathcal{B}(X)$  is called a *continuous component* of  $K_b$ . If  $P$  admits a distribution  $b$  such that there exists a continuous component  $T$  of  $K_b$  with  $T(\cdot, X) > 0$ , then  $P$  is called a *T-chain*.

A set  $C \in \mathcal{B}(X)$  is called *petite* if there exist a probability distribution  $b$  on  $\mathbb{N}$  and a nontrivial measure  $\nu_b$  on  $\mathcal{B}(X)$  such that  $K_b(x, A) \geq \nu_b(A)$  for every  $x \in X$  and  $A \in \mathcal{B}(X)$ . If moreover  $b = \delta_a$  the Dirac distribution at some  $a \in \mathbb{N}$ , then  $C$  is called *a-small*.

If  $P$  is  $\varphi$ -irreducible, then the family  $(D_i)_{i=1, \dots, d} \in \mathcal{B}(X)^d$  is called a *d-cycle* when

$$\begin{cases} P(x, D_{i+1}) = 1 \text{ for } x \in D_i \text{ and } i = 0, \dots, d-1 \text{ mod } d \\ \varphi((\cup_{1 \leq i \leq d} D_i)^c) = 0 \text{ for any irreducibility measure } \varphi. \end{cases} \quad (2.23)$$

By [26, Theorem 5.4.4 and Proposition 5.2.4], if  $P$  is  $\varphi$ -irreducible, then there exist  $d \geq 1$  and a *d-cycle*. The *period* of  $P$  is the largest integer  $d$  for which there exists a *d-cycle*. If the period of  $P$  is equal to 1, then  $P$  is said to be *aperiodic*.

If  $P$  is  $\varphi$ -irreducible, then  $P$  is said to be recurrent when for every  $A \in \mathcal{B}(X)$  such that  $\varphi(A) > 0$  and for every  $x \in A$ , we have  $\sum_{k=1}^{\infty} P^k(x, A) = +\infty$ . We say that  $P$  is positive when it admits an invariant probability measure.



In practice, in order to show that  $P$  is positive recurrent, it is sufficient to establish a Foster-Lyapunov condition, see e.g., [26, Theorem 15.0.1], i.e., the existence of a function  $V: \mathsf{X} \rightarrow [1, +\infty]$  finite at least at one point of  $\mathsf{X}$ , of a petite set  $C$  and of constants  $b < \infty$  and  $\rho \in (0, 1)$ , such that, for any  $x \in \mathsf{X}$ , we have

$$\int V(y)P(x, dy) \leq \rho V(x) + b\mathbb{1}\{x \in C\}. \quad (2.24)$$

We have now all the tools to state our main contribution.

**Theorem 2.3.** *Assume **H1-H2** and **H3**. Then, the Markov kernel  $P$  defined via (2.1) is a  $\varphi$ -irreducible T-chain, and any compact set is petite. If moreover **H4** holds, then, the Markov kernel  $P$  is aperiodic, and any compact set is small. In addition, if  $P$  is positive recurrent, then  $\{\phi_k\}_{k \in \mathbb{N}}$  is ergodic, i.e.,  $P$  admits a unique stationary distribution  $\pi$  and for  $\pi$ -almost every  $x \in \mathsf{X}$ ,*

$$\lim_{k \rightarrow +\infty} \|\delta_x P^k - \pi\|_{\text{TV}} = 0. \quad (2.25)$$

In addition to the assumptions of Theorem 2.3, if we suppose that  $P$  is Harris recurrent, then a Law of Large Numbers holds, see [26, Theorem 17.0.1]. For any  $\pi$ -integrable function  $g$ , a Markov chain  $\{\phi_k\}_{k \in \mathbb{N}}$  associated to the kernel  $P$  satisfies

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{k=0}^{T-1} g(\phi_k) = \int g d\pi. \quad (2.26)$$

In particular, a Foster-Lyapunov condition (2.24) implies Harris recurrence.

The proof of Theorem 2.3 is postponed to Section 4.2. It relies on intermediary results that to a great extent are generalizations of results in [10]. In particular, Propositions 4.1, 4.2 and 4.4 characterize globally attracting states, reachable states and steadily attracting states respectively. Propositions 4.7, 4.8, 4.10 and 4.11 provide consequences of the assumption of controllability ( $C_x$ ). Lemma 4.1 is a generalization of [25, Lemma 3.0], which turns out to be useful to prove that the controllability condition ( $C_x$ ) implies that the Markov kernel  $P$  is a T-chain, as stated in Proposition 4.12 and Corollary 4.13. Proposition 4.14 characterizes the support of the irreducibility measures of  $P$ , while Theorems 4.15 to 4.18 end the proof of Theorem 2.3.

In contrast to [10], we assume here  $F$  to be locally Lipschitz instead of continuously differentiable. This changes the assumption of controllability ( $C_x$ ), which consists here of a maximal rank condition for every element in the Clarke's generalized Jacobian of the extended transition map at some point, instead of a maximal rank condition of the Jacobian. Furthermore, we assume the sets  $\mathsf{X}$  and  $W$  to be manifolds instead of open subsets of Euclidean spaces. Hence, while in [10] the Jacobian matrix of the extended transition map can be identified to a rectangular matrix, here, the Clarke's generalized Jacobian consists in a set of linear applications between tangent spaces.

## 3 Applications

### 3.1 CMA-ES

We consider in this section the process  $\{(z_k, \Sigma_k)\}_{k \in \mathbb{N}}$  defined in Section 2.1.1. It defines a time-homogeneous Markov chain when the objective function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  is scaling-invariant, see (2.6). As observed in Section 2.1.1, the Markov chain  $\{(z_k, \Sigma_k)\}_{k \in \mathbb{N}}$  obeys (2.1), with  $F$  and  $\alpha$  defined by (2.8) and (2.9) respectively.

Besides, we assume in this section that  $f$  has Lebesgue-negligible level sets, i.e.,  $\text{Leb}(\mathcal{L}_t) = 0$ , with

$$\mathcal{L}_t := \{x \in \mathbb{R}^d \mid f(x) = t\} \quad \text{for } t \in \mathbb{R}. \quad (3.1)$$

Stability of Markov chains defined in the context of ES with step-size adaptation has been proven [7, 32], yielding to linear convergence. We complement these results applying now Theorem 2.3 to show the stability of  $\{(z_k, \Sigma_k)\}_{k \in \mathbb{N}}$ . First, observe that the assumption **H2** is automatically satisfied, since  $F$  is continuously differentiable. As for **H1**, we use the following result.

**Proposition 3.1.** *Suppose that  $f$  has Lebesgue-negligible level sets. Define for any  $\theta = (z, \Sigma) \in \mathsf{X}$  and  $v = (v_1, \dots, v_\mu) \in W$ ,*

$$p_\theta(v) = \frac{\lambda!}{(\lambda - \mu)!} \mathbb{1}\left\{f\left(z + \sqrt{\Sigma}v_1\right) < \dots < f\left(z + \sqrt{\Sigma}v_\mu\right)\right\}$$

$$\times \left(1 - Q_\theta^f(v_\mu)\right)^{\lambda-\mu} \gamma^d(v_1) \dots \gamma^d(v_\mu) \quad (3.2)$$

with  $Q_\theta^f(u) = \int \mathbb{1}\{f(z + \sqrt{\Sigma}\xi) < f(z + \sqrt{\Sigma}u)\} \gamma^d(\xi) d\xi$  and where  $\gamma^d$  is the density of the  $d$ -dimensional standard normal distribution w.r.t. Lebesgue. Then,  $p_\theta$  defines a density (w.r.t. Lebesgue in  $\mathbb{R}^{d\mu}$ ) of the random variable  $\alpha(\theta, U_1)$ .

If  $f$  has Lebesgue-negligible level sets and is continuous, it follows that **H1** holds.

The proof of Proposition 3.1 mimics the one of [10, Proposition 5.2], but is given for completeness in Appendix C. Then, it remains to prove **H4** and in particular to find a steadily attracting state  $\theta^* = (z^*, \Sigma^*)$  for which there exist  $k > 0$  and  $v_{1:k}^* \in \overline{\mathcal{O}_{\theta^*}^k}$  such that  $\partial S_{\theta^*}^k(v_{1:k}^*)$  is of maximal rank. This is achieved in the following proposition proven in Section 4.3.

**Proposition 3.2.** *Suppose that  $f$  is continuous, scaling-invariant with Lebesgue-negligible level sets. Then,*

(i) *the state  $\theta^* = (0, I_d)$  is steadily attracting ;*

(ii) *there exists  $k > 0$  and  $v_{1:k}^* \in \overline{\mathcal{O}_{\theta^*}^k}$  such that  $\mathcal{D}S_{\theta^*}^k(v_{1:k}^*) : \mathbb{W}^k \rightarrow T_{S_{\theta^*}^k(v_{1:k}^*)}\mathbb{X}$  is surjective, hence is full rank, where*

$$T_{S_{\theta^*}^k(v_{1:k}^*)}\mathbb{X} = \mathbb{R}^d \times \ker(\mathcal{D} \det(I_d)),$$

*where  $\det$  is the determinant map on the set of symmetric matrices  $\mathcal{S}^d$ , and  $\ker$  denotes the kernel of a linear application.*

Then, by applying Theorem 2.3, the  $\varphi$ -irreducibility and aperiodicity of the chain  $\{\theta_k\}_{k \in \mathbb{N}}$  follow.

**Theorem 3.3.** *Suppose that  $f$  is continuous, scaling-invariant with Lebesgue-negligible level sets. Then the Markov chain  $\{\theta_k\}_{k \in \mathbb{N}}$  defines a time-homogeneous  $\varphi$ -irreducible aperiodic  $T$ -chain, for which any compact subset of  $\mathbb{X}$  is small.*

## 3.2 The step-size adaptive ES with nonsmooth update

Here, we consider the process  $\{z_k\}_{k \in \mathbb{N}}$  defined in Section 2.1.2. As for CMA-ES, if  $f$  is supposed to be scaling-invariant, then this sequence defines a time-homogeneous Markov chain. In Section 2.1.2, we have established that this chain follows the model (2.1). Like CMA-ES, the following proposition gives a sufficient condition for assumption **H1** to hold. The proof goes as for Proposition 3.1, which can be found in Appendix C.

**Proposition 3.4.** *Suppose that  $f$  has Lebesgue-negligible level sets. Define for all  $z \in \mathbb{X}$  and  $v = (v_1, \dots, v_\mu) \in \mathbb{W}$*

$$p_z(v) = \frac{\lambda!}{(\lambda - \mu)!} \mathbb{1}\{f(z + v_1) < \dots < f(z + v_\mu)\} (1 - Q_z^f(v_\mu))^{\lambda-\mu} \gamma^d(v_1) \dots \gamma^d(v_\mu) \quad (3.3)$$

with  $Q_z^f(u) = \int \mathbb{1}\{f(z + \xi) < f(z + u)\} \gamma^d(\xi) d\xi$  and where  $\gamma^d$  is the density of the  $d$ -dimensional standard normal distribution w.r.t. Lebesgue. Then,  $p_z$  defines a density (w.r.t. Lebesgue in  $\mathbb{R}^{d\mu}$ ) of the random variable  $\alpha((z, K), U_1)$ . Moreover, if  $f$  is (a monotone transformation of) a continuous function, then  $(z, v) \mapsto p_z(v)$  is l.s.c.

As for CMA-ES, assumption **H2** holds since  $F$ , given in (2.17), is the composition of a continuously differentiable function with the Lipschitz function  $x \mapsto \|x\|$ . Regarding **H4**, the next proposition states the existence of a steadily attracting state. The proof follows the same lines as [10, Proposition 5.3], but is given for completeness.

**Proposition 3.5.** *Suppose that  $f$  is continuous, scaling-invariant with Lebesgue-negligible level sets. Then,  $0$  is a steadily attracting state.*

*Proof.* For  $z_0 \in \mathbb{R}^d$ , we set  $v_1 = -[z_0, \dots, z_0] \in \mathbb{R}^{d\mu}$ , and  $v_k = [0, \dots, 0] \in \mathbb{R}^{d\mu}$ . Note that, by Proposition 3.4, since  $f$  has Lebesgue-negligible level sets,  $v_{1:k} \in \overline{\mathcal{O}_{z_0}^k}$ . Moreover, we have  $S_{z_0}^k(v_{1:k}) = 0$  for every  $k \geq 1$ , where  $S_{z_0}^k$  is defined in (2.19). We conclude the proof by using Corollary 4.5.  $\square$

To complete the verification of **H4**, we show in the next proposition that there exists  $v_1 \in \overline{\mathcal{O}_0^1}$  such that  $S_0^1$  is differentiable in  $v_1$  and  $\mathcal{D}S_0^1(v_1)$  is of maximal rank.

**Proposition 3.6.** *Suppose that  $f$  is continuous, scaling-invariant with Lebesgue-negligible level sets. Then,  $S_0^1$  is differentiable in  $v_1 = (0, \dots, 0) \in \overline{\mathcal{O}_0^1}$  and  $\mathcal{D}S_0^1(v_1)$  is of maximal rank.*

*Proof.* Note that, by Proposition 3.4,  $v_1$  belongs to  $\overline{\mathcal{O}_0^1}$ . Moreover, for  $h = (h_1, \dots, h_\mu) \in \mathbb{W}$ , we have by definition of  $F$  and of  $S_0^1$ , see (2.17) and (2.19) respectively, that

$$S_0^1(v_1 + h) = F(0, h) = \exp\left(\frac{1}{d_\sigma} \left(1 - \frac{\sqrt{\mu_{\text{eff}} \|\sum_{i=1}^\mu w_i h_i\|}}{\mathbb{E}\|\mathcal{N}(0, I_d)\|}\right)\right) \times \sum_{i=1}^\mu w_i h_i.$$

A simple Taylor expansion shows that

$$\lim_{h \rightarrow 0} \frac{\left\| S_0^1(v_1 + h) - S_0^1(v_1) - \exp\left(\frac{1}{d_\sigma}\right) \times \sum_{i=1}^\mu w_i h_i \right\|}{\|h\|} = 0, \quad (3.4)$$

ending the proof.  $\square$

Using Theorem 2.3, we deduce the  $\varphi$ -irreducibility and aperiodicity of the chain  $\{z_k\}_{k \in \mathbb{N}}$ .

**Theorem 3.7.** *Suppose that  $f$  is continuous, scaling-invariant with Lebesgue-negligible level sets. Then, the Markov chain  $\{z_k\}_{k \in \mathbb{N}}$  defines a time-homogeneous  $\varphi$ -irreducible aperiodic  $T$ -chain, for which compact subsets of  $\mathbb{X}$  are small.*

Note that in [32], it has been proven that the chain  $\{z_k\}_{k \in \mathbb{N}}$  is  $\varphi$ -irreducible, aperiodic and positive recurrent, on the condition that the step-size obeys to a smooth update instead of (2.15). However, a smooth step-size update was required only to prove the  $\varphi$ -irreducibility and aperiodicity of the chain, since the derivation of these two results rely in [32] on results in [10]. Now that we have proven that the chain  $\{z_k\}_{k \in \mathbb{N}}$  is  $\varphi$ -irreducible and aperiodic even when the step-size update is nonsmooth, we can prove that it is positive recurrent following the proofs of [32].

## 4 Proofs

### 4.1 Preliminary results

#### 4.1.1 Accessibility, attracting and attainable states

In this section, we generalize characterizations of globally and steadily attracting states developed in [10]. In contrast to this reference, we relax assumptions on the sets  $\mathbb{X}$ ,  $\mathbb{U}$  and  $\mathbb{W}$ . Indeed, [10] supposed that these sets were open subsets of Euclidean spaces. Here, we only suppose that they are smooth connected manifolds, as formalized in Section 2.1. The proofs in this section are almost identical to those of [10], but are given for completeness in Appendix C.

For the rest of the paper, let us define  $A_+^0(x) := \{x\}$  and

$$A_+^k(x) := \{S_x^k(w_{1:k}) \mid w_{1:k} \in \mathcal{O}_x^k\} \quad \text{for } k \geq 1. \quad (4.1)$$

The set  $A_+^k(x)$  is the set of states that can be reached by  $\phi_k$  conditionally to  $\phi_0 = x$ . We also define the set of states that can be reached by  $\{\phi_k\}_{k \in \mathbb{N}}$  (in finite time) conditionally to  $\phi_0 = x$  as

$$A_+(x) := \bigcup_{k \in \mathbb{N}} A_+^k(x). \quad (4.2)$$

Then, we say that the control model associated to (2.1) is *forward accessible* if for every  $x \in \mathbb{X}$ ,  $A_+(x)$  has a nonempty interior in  $\mathbb{X}$ . Moreover, with these notations, a point  $x^* \in \mathbb{X}$  is a *globally attracting state*, if for every  $y \in \mathbb{X}$  we have

$$x^* \in \bigcap_{T \geq 1} \overline{\bigcup_{k \geq T} A_+^k(y)}. \quad (4.3)$$

As shown in the next proposition, this definition is equivalent to the statement we used in Section 2.1 to introduce a globally attracting state that for any  $y \in \mathbb{X}$  and any neighborhood  $U$  of  $x^*$ , there exists  $k > 0$  and a  $k$ -steps path between  $y$  and  $U$ .

**Proposition 4.1** (Characterization of globally attracting states). *Suppose **H1**. A point  $x^* \in \mathbb{X}$  is globally attracting if and only if one of the following equivalent conditions holds.*

- (i) For any  $y \in \mathbb{X}$ ,  $x^* \in \overline{A_+(y)}$ .

- (ii) For any  $y \in \mathsf{X}$  and any open subset  $U$  of  $\mathsf{X}$  containing  $x^*$ , there exist  $k > 0$  and a  $k$ -steps path from  $y$  to  $U$ .
- (iii) For any  $y \in \mathsf{X}$ , there exists a sequence  $\{y_k\}_{k>0}$  with  $y_k \in A_+^k(y)$ , from which we can extract a subsequence converging to  $x^*$ .

A point  $x \in \mathsf{X}$  is said to be *reachable* by  $P$  [26, Section 6.1.2] if for any measurable neighborhood  $U$  of  $x$  in  $\mathsf{X}$ , we have

$$\forall y \in \mathsf{X}, \quad \sum_{k \geq 1} P^k(y, U) > 0. \quad (4.4)$$

The equivalence between globally attracting states and reachable states relies on the following proposition.

**Proposition 4.2** (Characterization of reachable states). *Consider the Markov kernel  $P$  defined via Equation (2.1), and suppose **H1** and that  $F$  is continuous. Then for any open subset  $U$  of  $\mathsf{X}$ , any  $x \in \mathsf{X}$  and  $k > 0$ , the following statements are equivalent.*

- (i) There exists a  $k$ -steps path from  $x$  to  $U$ .
- (ii)  $P^k(x, U) > 0$ .

As an immediate consequence of Propositions 4.1 and 4.2, we get the following equivalence between states that are globally attracting by the control model associated to (2.1) and states that are reachable by  $P$ .

**Corollary 4.3.** *Consider the Markov kernel  $P$  defined via (2.1), and suppose **H1** and that  $F$  is continuous. Then  $x \in \mathsf{X}$  is globally attracting if and only if it is reachable by  $P$ .*

Recall that a state  $x^* \in \mathsf{X}$  is *steadily attracting* [10] if for all  $y \in \mathsf{X}$  and all open neighborhood  $U$  of  $x^*$  in  $\mathsf{X}$ , there exists  $T > 0$  such that for all  $k \geq T$  there exists a  $k$ -steps path from  $y$  to  $U$ .

We now state two technical results related to steadily attracting states, which will be instrumental in the proofs of our main results. The first one is the equivalent of [10, Proposition 3.3].

**Proposition 4.4.** *Suppose **H1**. The following statements hold.*

- (i) If  $x^* \in \mathsf{X}$  is steadily attracting, then it is globally attracting.
- (ii) A state  $x^* \in \mathsf{X}$  is steadily attracting if and only if for every  $y \in \mathsf{X}$  we can find a sequence  $\{y_k\}_{k>0}$  with  $y_k \in \overline{A_+^k(y)}$ , which converges to  $x^*$ .
- (iii) Assume  $F$  is continuous. If there exists a steadily attracting state, then every globally attracting state is steadily attracting.

Note however that the statement of [10, Proposition 3.3 (ii)] is slightly different as the element  $y_k$  belongs to  $A_+^k(y)$  while in (ii) above  $y_k$  belongs to  $\overline{A_+^k(y)}$ . It is easy to see that both statements are equivalent.

In addition, we give the following corollary of Proposition 4.4 when  $F$  is assumed continuous.

**Corollary 4.5.** *Suppose **H1** and that  $F$  is continuous. Then, for any  $x \in \mathsf{X}$  and  $k \in \mathbb{N}$  we have the inclusion  $\{S_x^k(w_{1:k}) \mid w_{1:k} \in \overline{\mathcal{O}_x^k}\} \subset \overline{A_+^k(x)}$ . Consequently the following statements are equivalent.*

- (i) The point  $x^*$  is steadily attracting.
- (ii) For every  $x \in \mathsf{X}$  we can find a sequence  $\{y_k\}_{k>0}$  satisfying  $y_k \in \{S_x^k(w_{1:k}) \mid w_{1:k} \in \overline{\mathcal{O}_x^k}\}$  for any  $k \geq 1$ , and which converges to  $x^*$ .
- (iii) For every  $x \in \mathsf{X}$ , for every neighborhood  $U$  of  $x^*$ , there exists  $T > 0$  such that for any  $k \geq T$  we can find  $w_{1:k} \in \overline{\mathcal{O}_x^k}$  satisfying  $S_x^k(w_{1:k}) \in U$ .

The next result turn out to be useful later in order to prove the aperiodicity of the Markov kernel  $P$ , given that it is  $\varphi$ -irreducible. To this end, we need to introduce the notion of attainability, as considered in [26]. We say that a state  $x^* \in \mathsf{X}$  is *attainable* if

$$\forall y \in \mathsf{X}, \quad x^* \in A_+(y). \quad (4.5)$$

**Proposition 4.6.** *Consider the Markov kernel  $P$  defined via (2.1), and suppose **H1**. Let  $x^* \in \mathsf{X}$  be attainable, and set*

$$E := \{a \in \mathbb{N}^* \mid \exists T \in \mathbb{N}, \forall k \geq T, x^* \in A_+^{ak}(x^*)\}. \quad (4.6)$$

*Then, the following statements hold.*

- (i)  $E$  is nonempty and for every  $a, b \in E$ , the greatest common divider of  $a$  and  $b$  satisfies  $\gcd(a, b) \in E$ .
- (ii) If  $\gcd(E) = \max\{c \in \mathbb{N} \mid c \text{ divides } a, \forall a \in E\} = 1$ , then  $x^*$  is steadily attracting.
- (iii) If  $P$  is  $\varphi$ -irreducible, then there exists a  $d$ -cycle (as defined in (2.23)) with  $d = \gcd(E)$ .

#### 4.1.2 Controllability condition

In this section, we generalize results of [10], more precisely [10, Propositions 3.5, 3.6 and 3.7], on the consequences of condition  $(C_x)$ . The main challenge here is to deal with the condition that  $F$  is supposed to be locally Lipschitz only.

First, we generalize [10, Proposition 3.5] and prove that if the controllability condition  $(C_{x^*})$  is satisfied for some  $x^*$  a globally attracting state, then is satisfied for every  $y \in X$ .

**Proposition 4.7.** *Suppose **H1** and **H2**. Let  $x^* \in X$  be a globally attracting state. If  $(C_{x^*})$  holds, then for every  $y \in X$ ,  $(C_y)$  holds.*

*Proof.* By  $(C_{x^*})$ , there exist  $k > 0$  and  $w_{1:k}^* \in \overline{\mathcal{O}_{x^*}^k}$  such that  $\partial S_{x^*}^k(w_{1:k}^*)$  is of rank  $n$ , the dimension of  $X$ . See that, by Proposition B.10, we can assume that  $w_{1:k}^* \in \mathcal{O}_{x^*}^k$ . Moreover, the function  $S^k : (z, w_{1:k}) \mapsto S_z^k(w_{1:k})$  is locally Lipschitz (since  $F$  is locally Lipschitz), hence according to Proposition B.9,  $\lim_{z \rightarrow x^*} \partial S^k(z, w_{1:k}) \subset \partial S^k(x^*, w_{1:k})$  and since  $\partial S^k(z, w_{1:k}) = \partial_z S^k(z, w_{1:k}) \times \partial S_z^k(w_{1:k})$ , we obtain  $\lim_{z \rightarrow x^*} \partial S_z^k(w_{1:k}^*) \subset \partial S_{x^*}^k(w_{1:k}^*)$ . Since rank is lower semicontinuous, we deduce that there exists an open neighborhood  $U$  of  $x^*$  such that for any  $z \in U$ ,  $\partial S_z^k(w_{1:k}^*)$  is of rank  $n$ . Moreover,  $z \mapsto p_z^k(w_{1:k}^*)$  is lower semicontinuous, so, up to taking  $U$  smaller, we can suppose that for any  $z \in U$ ,  $p_z^k(w_{1:k}^*) > 0$ , i.e.,  $w_{1:k}^* \in \mathcal{O}_z^k$ .

Let  $y \in X$ . Since  $x^*$  is a globally attracting state, then by Proposition 4.1, there exist  $t_0 > 0$  and  $u_{1:t_0}$  a  $t_0$ -steps path from  $y$  to  $x \in U$ , i.e.,  $u_{1:t_0} \in \mathcal{O}_y^{t_0}$  and  $x = S_y^{t_0}(u_{1:t_0}) \in U$ . Since  $\mathcal{O}_y^{t_0}$  is open and  $S_y^{t_0}$  is continuously locally Lipschitz, by Corollary B.5 we can assume w.l.o.g. that  $S_y^{t_0}$  is differentiable at  $u_{1:t_0}$ .

Since  $x \in U$ , then  $\partial S_x^k(w_{1:t_0}^*)$  is of maximal rank, using the chain rule, see Proposition B.11, we deduce that, for  $T = t_0 + k$  and  $u_{t_0+1:T+k} = w_{1:k}^*$ , we have that  $\partial S_y^T(u_{1:T})$  is of maximal rank.  $\square$

The next proposition states that if we find a point  $x^*$ ,  $k > 0$  and  $w_{1:k}^* \in \mathcal{O}_{x^*}^k$  which satisfy the forementioned controllability condition  $(C_{x^*})$ , that is,  $\partial S_{x^*}^k(w_{1:k}^*)$  is of maximal rank, then, using Proposition 4.9, we can find  $u_{1:k}^* \in \mathcal{O}_{x^*}^k$  as closed as we want from  $w_{1:k}^*$  such that  $S_{x^*}^k$  is differentiable in  $u_{1:k}^*$  and  $\mathcal{D}S_{x^*}^k(u_{1:k}^*)$  is of maximal rank. In other words, our controllability condition  $(C_{x^*})$  implies a full rank condition as used in [10].

**Proposition 4.8.** *Suppose **H1** and **H2**. Let  $x^* \in X$  and suppose that  $(C_{x^*})$  holds. Then, condition  $(R_{x^*})$  stated below holds.*

*Proof.* By  $(C_{x^*})$  and by Proposition B.10, there exist  $k > 0$  and  $w_{1:k}^* \in \mathcal{O}_{x^*}^k$  such that  $\partial S_{x^*}^k(w_{1:k}^*)$  is of maximal rank. By Proposition 4.9 below, for any neighborhood  $W \subset W$  of  $w_{1:k}^*$ , there exists  $u_{1:k}^* \in W$ , such that  $S_{x^*}^k$  is differentiable in  $u_{1:k}^*$ , with  $\text{rank } \mathcal{D}S_{x^*}^k(u_{1:k}^*) = n$ . However  $\mathcal{O}_{x^*}^k$  is open, so we can take  $W = \mathcal{O}_{x^*}^k$  and complete the proof.  $\square$

**Proposition 4.9.** *Suppose that  $f: X \rightarrow Y$  is locally Lipschitz at  $x_0 \in X$ , and that  $\partial f(x_0)$  is of maximal rank, i.e., any  $h \in \partial f(x_0)$  is of maximal rank. Then, there exists a neighborhood  $U$  of  $x_0$  such that for any  $y \in U$ ,  $\partial f(y)$  is of maximal rank. Moreover, for every neighborhood  $V \subset U$  of  $x_0$ , there exists  $y_0 \in V$  such that  $f$  is differentiable at  $y_0$  and  $\mathcal{D}f(y_0)$  is of maximal rank.*

*Proof.* Let  $A = \{h \in \mathcal{L}(T_{x_0}X, T_{f(x_0)}Y) \mid h \text{ is not of maximal rank}\}$ . Since the application rank is l.s.c., then  $A$  is a closed set. By Proposition B.8,  $\partial f(x_0)$  is compact, and disjoint from  $A$  since it is assumed to be of maximal rank. Thus  $\text{dist}(\partial f(x_0), A) > 0$ , where  $\text{dist}$  is a metric induced by a norm on the finitely dimensioned affine space  $\mathcal{L}(T_{x_0}X, T_{f(x_0)}Y)$ . Moreover, there exists  $h^* \in \partial f(x_0)$  such that for every  $h \in \partial f(x_0)$  we have

$$\text{dist}(h, A) \geq \text{dist}(h^*, A) = \text{dist}(\partial f(x_0), A) > 0.$$

By [11, Proposition 2.6.2(c)], there exists a neighborhood  $U$  of  $x_0$  such that for all  $y \in U$ ,  $\text{dist}(\partial f(y), A) \geq \text{dist}(h^*, A)/2 > 0$ , thus  $\partial f(y)$  is of maximal rank. The second part follows from Rademacher's theorem, see Corollary B.5.  $\square$

From now on, we can assume a full rank condition, i.e.,

$$\text{there exists } w_{1:k} \in \mathcal{O}_x^k \text{ such that } \mathcal{D}S_x^k(w_{1:k}) \text{ exists and is of maximal rank,} \quad (\mathbf{R}_x)$$

instead of the controllability condition  $(\mathbf{C}_x)$ . We can then use Proposition 4.8 to extend our results. The next proposition states that if we can find a globally attracting state  $x^*$  satisfying the maximal rank condition  $(\mathbf{R}_{x^*})$ , then we can find an attainable state. It generalizes [10, Proposition 3.6].

**Proposition 4.10.** *Suppose **H1** and **H2**. Let  $x^* \in \mathbf{X}$  and suppose that there exist  $k > 0$  and  $w_{1:k}^* \in \mathcal{O}_{x^*}^k$  such that  $(\mathbf{R}_{x^*})$  is satisfied with  $w_{1:k}^*$ .*

(i) *There exists  $U$  a neighborhood of  $x^*$  such that for any  $x \in U$ , there exists  $w_{1:k} \in \mathcal{O}_x^k$  for which  $S_x^k(w_{1:k}) = S_{x^*}^k(w_{1:k}^*)$ .*

(ii) *If  $x^*$  is globally attracting, then  $S_{x^*}^k(w_{1:k}^*)$  is attainable, see (4.5).*

*Proof.* (i) Let  $(U, \varphi)$  be a local chart of  $\mathbf{X}$  around  $x^*$ ,  $(V, \theta)$  a local chart of  $\mathbf{X}$  around  $S_{x^*}^k(w_{1:k}^*)$ , and  $(W, \psi)$  a local chart of  $\mathbf{W}^k$  around  $w_{1:k}^*$ , such that the following differentiable function is well-defined

$$\begin{aligned} \tilde{S}^k: \quad & \varphi(U) \times \psi(W) \subset \mathbb{R}^n \times \mathbb{R}^{kp} \rightarrow V \subset \mathbb{R}^n \\ & (x, w) = ((x_1, \dots, x_n), (w_1, \dots, w_{kp})) \mapsto \tilde{S}_x^k(w) := \theta \circ S_{\varphi^{-1}(x)}^k \circ \psi^{-1}(w). \end{aligned}$$

We recall that the positive integers  $n$  and  $p$  are the dimensions of  $\mathbf{X}$  and  $\mathbf{W}$ , respectively. By composition, observe that  $\mathcal{D}_w \tilde{S}^k(\varphi(x^*), \psi(w_{1:k}^*))$  is surjective. Hence, we can find coordinates  $i_1, \dots, i_n$  of  $\mathbb{R}^{kp}$  such that

$$\det \left[ \mathcal{D}_{w_{i_1}} \tilde{S}^k \mid \dots \mid \mathcal{D}_{w_{i_n}} \tilde{S}^k \right] (\varphi(x^*), \psi(w_{1:k}^*)) = n.$$

Note that, up to a permutation of indices in the chart  $\psi$ , we can assume w.l.o.g. that  $i_1, \dots, i_n$  equal respectively  $kp - n + 1, \dots, kp$ . To ease the presentation, we use the following abuse of notation  $(w_1^*, \dots, w_{kp}^*) = \psi(w_{1:k}^*)$ . Then, by the implicit function theorem, see Theorem B.13, there exist neighborhoods  $M$  of  $(\varphi(x^*), w_1^*, \dots, w_{kp-n}^*)$  and  $N$  of  $(w_{kp-n+1}^*, \dots, w_{kp}^*)$ , and a  $\mathcal{C}^1$  function  $g: M \rightarrow N$  such that, for every  $(x_1, \dots, x_n, w_1, \dots, w_{kp-n}) \in M$ , we have

$$\tilde{S}_{(x_1, \dots, x_n)}^k(w_1, \dots, w_{kp-n}, g(x_1, \dots, x_n, w_1, \dots, w_{kp-n})) = \tilde{S}_{\varphi(x^*)}^k(w_1^*, \dots, w_{kp}^*).$$

This proves (i).

(ii) Suppose that  $x^*$  is globally attracting. Let  $U \subset \mathbf{X}$  be a neighborhood of  $x^*$  satisfying (i), and let  $y \in \mathbf{X}$ . Then, by Proposition 4.1(ii), there exist  $k_1 > 0$  and  $w_{1:k_1} \in \mathcal{O}_y^{k_1}$  such that  $S_y^{k_1}(w_{1:k_1}) \in U$ . Since  $U$  satisfies (i), there exists  $w_{k_1+1:k_1+k} \in \mathcal{O}_{S_y^{k_1}(w_{1:k_1})}^k$  with  $S_y^{k_1+k}(w_{1:k_1+k}) = S_{x^*}^k(w_{1:k}^*)$ .  $\square$

We discuss in the next proposition the forward accessibility of the control model (2.1). We recall that it is said to be forward accessible if for every  $x \in \mathbf{X}$ , the subset  $A_+(x) \subset \mathbf{X}$  defined in (4.2) of states that can be reached in finite time starting from  $x$ , has a nonempty interior. The next proposition generalizes [10, Proposition 3.7].

**Proposition 4.11.** *Suppose **H1** and **H2**. If for every  $x \in \mathbf{X}$ ,  $(\mathbf{R}_x)$  holds, then the control model associated to (2.1) is forward accessible.*

*Furthermore, if  $F$  is smooth (infinitely differentiable), the control model is forward accessible if and only if for every  $x \in \mathbf{X}$ ,  $(\mathbf{R}_x)$  holds.*

*Proof.* We apply the Local Submersion Theorem [14, Chapter 1.4]. Since  $S_x^k$  is a submersion at  $w_{1:k}$ , there exist local charts  $(W, \psi)$  of  $\mathbf{W}^k$  around  $w_{1:k}$  and  $(V, \varphi)$  of  $\mathbf{X}$  around  $S_x^k(w_{1:k})$  such that

$$\varphi \circ S_x^k \circ \psi(u_1, \dots, u_{kp}) = (u_1, \dots, u_n) \quad \text{for all } (u_1, \dots, u_{kp}) \in \psi(W).$$

Therefore, since  $\varphi$  is a continuous bijection (by definition of a local chart), then there exists a neighborhood  $U$  of  $S_x^k(w_{1:k})$  such that  $S_x^k(W) = U$ . Moreover,  $\mathcal{O}_{x^*}^k$  is an open subset of  $\mathbf{W}^k$ , so we can assume  $W \subset \mathcal{O}_{x^*}^k$ . Therefore,  $U \subset A_+^k(x)$ , which hence has a nonempty interior.

Suppose now that  $F$  is smooth and that the control model is forward accessible. Then, for every  $x \in \mathbf{X}$ ,  $\text{int}(A_+(x)) \neq \emptyset$ . Since  $A_+(x) = \cup_{k \geq 0} A_+^k(x)$ , we deduce that there exists  $k \in \mathbb{N}$  such that  $\text{int}(A_+^k(x)) \neq \emptyset$ . Since  $\text{int}(A_+^0(x)) = \text{int}(\{x\}) = \emptyset$ , we find that necessarily  $k > 0$ . By Sard's theorem [14, Appendix 1], we have that the set  $N := \{\mathbf{w} \in \mathcal{O}_x^k \mid \text{rank } \mathcal{D}S_x^k(\mathbf{w}) < n\}$  is of measure zero, that is, for all charts  $(\varphi, U)$  of  $\mathbf{X}$ , we have  $\text{Leb } \varphi(N \cap U) = 0$ , hence  $\text{int}(N) = \emptyset$ . We deduce that there exists  $w_{1:k} \in \mathcal{O}_x^k \setminus N$ , i.e., such that  $\text{rank } \mathcal{D}S_x^k(w_{1:k}) = n$ .  $\square$

## 4.2 Proofs of the main results: verifiable conditions for irreducibility and aperiodicity

### 4.2.1 T-chain and irreducibility

Here, we generalize the main results of [10] on  $\varphi$ -irreducibility of Markov kernels  $P$  defined via (2.1), as well as the T-chain property.

First, we slightly generalize [25, Lemma 3.0] to our context, that is, for a locally Lipschitz function between manifolds instead of a smooth function between open subsets of Euclidean spaces.

**Lemma 4.1.** *Let  $X_1$  be a  $n$ -dimensional manifold,  $\tilde{W}_1$  a  $m$ -dimensional manifold,  $\hat{W}_1$  a  $n$ -dimensional manifold, equipped with their respective Borelian  $\sigma$ -fields and with a measure  $\zeta_X$  (resp.  $\zeta_{\tilde{W}}$ ,  $\zeta_{\hat{W}}$ ), which satisfies that for any  $A \in \mathcal{B}(X_1)$  (resp. of  $\mathcal{B}(\tilde{W}_1)$ ,  $\mathcal{B}(\hat{W}_1)$ ),  $\zeta_X(A) = 0$  (resp.  $\zeta_{\tilde{W}}(A) = 0$ ,  $\zeta_{\hat{W}}(A) = 0$ ) if and only if  $\varphi(A \cap U)$  is Lebesgue-negligible for every chart  $(\varphi, U)$ .*

*Let  $G: (x, \tilde{w}, \hat{w}) \in X_1 \times \tilde{W}_1 \times \hat{W}_1 \mapsto z \in X_1$  be a locally Lipschitz map differentiable in  $(x_0, \tilde{w}_0, \hat{w}_0)$  such that  $\text{rank } \mathcal{D}_{\tilde{w}} G(x_0, \tilde{w}_0, \hat{w}_0) = n$ . Then,*

- (i) *There exists an open subset  $X \times \tilde{W} \times \hat{W} \subset X_1 \times \tilde{W}_1 \times \hat{W}_1$  containing  $(x_0, \tilde{w}_0, \hat{w}_0)$  such that for any  $x \in X$ , the measure defined by*

$$\nu(x, \cdot): A \subset X_1 \mapsto \int_{\tilde{W}} \int_{\hat{W}} \mathbb{1}_A \{G(x, \tilde{w}, \hat{w})\} d\zeta_{\tilde{W}}(\tilde{w}) d\zeta_{\hat{W}}(\hat{w}) \quad (4.7)$$

*is equivalent to the measure  $\zeta_X$  on an open subset  $\mathcal{R}_x$  of  $X_1$ .*

- (ii) *There exist  $c > 0$ ,  $U_{x_0}$  an open subset of  $X_1$  containing  $x_0$ ,  $V_{x_0}^{\tilde{w}_0, \hat{w}_0}$  an open subset of  $\tilde{W}_1$  containing  $G(x_0, \tilde{w}_0, \hat{w}_0)$  such that for every  $x \in X$  and every measurable subset  $A$  of  $X_1$ , we have  $\nu(x, A) \geq c \mathbb{1}_{U_{x_0}}(x) \times \zeta_{X_1}(A \cap V_{x_0}^{\tilde{w}_0, \hat{w}_0})$ .*

*Proof.* First we prove the lemma when  $X_1, \tilde{W}_1, \hat{W}_1$  are open subsets respectively of  $\mathbb{R}^n, \mathbb{R}^m, \mathbb{R}^n$ , and  $\zeta_X, \zeta_{\tilde{W}}, \zeta_{\hat{W}}$  are assumed to be the Lebesgue measures on  $\mathbb{R}^n, \mathbb{R}^m, \mathbb{R}^n$  respectively.

Define the function

$$G^*: (x, \tilde{w}, \hat{w}) \in X_1 \times \tilde{W}_1 \times \hat{W}_1 \mapsto (x, \tilde{w}, G(x, \tilde{w}, \hat{w})) \in \mathbb{R}^{n+m+n}$$

Then, since  $\mathcal{D}_{\tilde{w}} G(x_0, \tilde{w}_0, \hat{w}_0)$  is of rank  $n$ , then  $\mathcal{D}G^*(x_0, \tilde{w}_0, \hat{w}_0)$  exists and is a full-rank squared matrix. Therefore, the inverse function theorem –as stated in Theorem B.12– applies and we find a neighborhood  $X \times \tilde{W} \times \hat{W}$  of  $(x_0, \tilde{w}_0, \hat{w}_0)$ , a neighborhood  $\mathcal{R}$  of  $(x_0, \tilde{w}_0, z_0)$  (where  $z_0 := G(x_0, \tilde{w}_0, \hat{w}_0)$ ), and a locally Lipschitz function  $H^*: \mathcal{R} \rightarrow X \times \tilde{W} \times \hat{W}$  such that

$$H^*(G^*(x, \tilde{w}, \hat{w})) = (x, \tilde{w}, \hat{w}) \quad \text{for every } (x, \tilde{w}, \hat{w}) \in X \times \tilde{W} \times \hat{W}.$$

Thus, there exists a locally Lipschitz function  $H: \mathcal{R} \rightarrow \hat{W}$  such that

$$H(x, \tilde{w}, G(x, \tilde{w}, \hat{w})) = \hat{w} \quad \text{for every } (x, \tilde{w}, \hat{w}) \in X \times \tilde{W} \times \hat{W}.$$

Then, by the chain rule, see [11, Theorem 2.6.6], for every  $(x, \tilde{w}, \hat{w}) \in X \times \tilde{W} \times \hat{W}$  at which  $G$  admits a partial derivative w.r.t.  $\hat{w}$ , we have that

$$\mathcal{D}_z H(x, \tilde{w}, G(x, \tilde{w}, \hat{w})) = [\mathcal{D}_{\hat{w}} G(x, \tilde{w}, \hat{w})]^{-1} \quad (4.8)$$

which is thus invertible. Moreover, by [11, Proposition 2.6.2(c)],  $\mathcal{D}_z H$  is continuous at points on which it is defined (which is dense by Rademacher's theorem [12, Theorem 3.2]). Therefore, there exists  $h_0 > 0$  such that in each of these points, by (4.8) we have

$$|\det \mathcal{D}_z H| \geq h_0. \quad (4.9)$$

Then, applying [16, Theorem 3] and Fubini's theorem, we get

$$\nu(x, A) = \int_A \left( \int \mathbb{1}_{\mathcal{R}}(x, \tilde{w}, z) |\det \mathcal{D}_z H| d\tilde{w} \right) dz, \quad (4.10)$$

so that

$$p(x, z) := \int \mathbb{1}_{\mathcal{R}}(x, \tilde{w}, z) |\det \mathcal{D}_z H| d\tilde{w} \quad (4.11)$$

defines a density w.r.t. Lebesgue for  $\nu(x, \cdot)$ . The rest of proof goes as in [25, Lemma 3.0], that we recall here for completeness.

Fix  $x \in X$  and let  $\mathcal{R}_x$  be the open subset of  $\mathbb{R}^n$  defined by

$$\mathcal{R}_x = \left\{ z \in X_1 \mid \exists \tilde{w} \in \tilde{W}, (x, \tilde{w}, z) \in \mathcal{R} \right\}.$$

Then, note that  $p(x, z)$  is positive if and only if  $z \in \mathcal{R}_x$ , and zero otherwise. This proves (i). For (ii), observe that, since  $\mathcal{R}$  is a neighborhood of  $(x_0, \tilde{w}_0, z_0)$ , then it contains a nonempty open subset  $X_0 \times \tilde{W}_0 \times Z_0$  containing  $(x_0, \tilde{w}_0, z_0)$ . We get then that  $p(x, z) \geq h_0 \times \text{Leb}(\tilde{W}_0)$  for every  $(x, z) \in X_0 \times Z_0$ . Then,

$$\nu(x, A) \geq h_0 \text{Leb}(\tilde{W}_0) \mathbb{1}\{x \in X_0\} \times \text{Leb}(A \cap Z_0)$$

which proves (ii).

Now suppose that  $X_1, \tilde{W}_1, \hat{W}_1$  are manifolds.

Let  $(\varphi, X_2)$  be a local chart of  $X_1$  around  $x_0$ ,  $(\tilde{\psi}, \tilde{W}_2)$  be a local chart of  $\tilde{W}_1$  around  $\tilde{w}_0$ ,  $(\hat{\psi}, \hat{W}_2)$  be a local chart of  $\hat{W}_1$  around  $\hat{w}_0$ , and  $(\eta, X_3)$  be a local chart of  $X_1$  around  $z_0 = G(x_0, \tilde{w}_0, \hat{w}_0)$ .

Then, define the locally Lipschitz map

$$G^{\text{loc}}: (x, \tilde{w}, \hat{w}) \in \varphi(X_2) \times \tilde{\psi}(\tilde{W}_2) \times \hat{\psi}(\hat{W}_2) \mapsto z = \eta \circ G(\varphi^{-1}(x), \tilde{\psi}^{-1}(\tilde{w}), \hat{\psi}^{-1}(\hat{w})) \in \mathbb{R}^n.$$

Thus, (i) and (ii) hold with  $G^{\text{loc}}$ , and

$$\nu^{\text{loc}}(x, A) = \int_{\tilde{W}_0} \int_{\hat{W}_0} \mathbb{1}_A(x, \tilde{w}, \hat{w}) \eta \circ G(\varphi^{-1}(x), \tilde{\psi}^{-1}(\tilde{w}), \hat{\psi}^{-1}(\hat{w})) d\tilde{w} d\hat{w}$$

is equivalent to the Lebesgue measure, for all  $x \in X_0$ , and  $X_0 \times \tilde{W}_0 \times \hat{W}_0$  being a neighborhood of  $(\varphi(x_0), \tilde{\psi}(\tilde{w}_0), \hat{\psi}(\hat{w}_0))$ . But, by assumption on the measures  $\zeta_X, \zeta_{\tilde{W}}$  and  $\zeta_{\hat{W}}$ ,  $\nu(x, \cdot)$  is locally equivalent to  $\rho^{-1} \circ \nu^{\text{loc}}(\varphi(x), \cdot)$  for all local chart  $(\rho, A)$  of  $X_1$ , thus is locally equivalent to  $\eta^{-1} \circ \text{Leb}_n$  where  $\text{Leb}_n$  is the Lebesgue measure of  $\mathbb{R}^n$ . Thus,  $\nu(x, \cdot)$  is equivalent to  $\zeta_X$ . This proves (i).

Now apply (ii) to  $G^{\text{loc}}$ , and find  $c > 0$ ,  $U_{x_0}$  an open of  $\varphi(X_2)$  containing  $\varphi(x_0)$ ,  $V_{x_0}^{\tilde{w}_0, \hat{w}_0}$  an open of  $X_3$  containing  $\eta(z_0)$ , such that

$$\nu^{\text{loc}}(x, A) \geq c \mathbb{1}_{U_{x_0}}(x) \times \text{Leb}_n(A \cap V_{x_0}^{\tilde{w}_0, \hat{w}_0}) \quad \text{for every } x \in \varphi(X), A \subset \mathbb{R}^n.$$

But, by assumption on  $\zeta_X$ , we find  $L_1^\eta, L_2^\eta > 0$  such that

$$\begin{aligned} \nu(x, A) &\geq L_1^\eta \times \nu^{\text{loc}}(\varphi(x), \eta(A \cap X_3)) \\ &\geq L_1^\eta \times c \mathbb{1}_{U_{x_0}}(\varphi(x)) \times \text{Leb}_n(\eta(A \cap X_3) \cap V_{x_0}^{\tilde{w}_0, \hat{w}_0}) \\ &\geq L_1^\eta \times c \mathbb{1}_{\varphi^{-1}(U_{x_0})}(x) \times L_2^\eta \times \zeta_X(A \cap \eta^{-1}(V_{x_0}^{\tilde{w}_0, \hat{w}_0})) \end{aligned}$$

for all  $x \in X$  and  $A \subset X_1$ , which proves (ii). □

We can now state the following result.

**Proposition 4.12.** *Consider the Markov kernel  $P$  defined via (2.1), and suppose **H1-H2**. Let  $x \in X$ .*

(i) *If  $(R_x)$  holds for some  $k > 0$  and  $w_{1:k} \in \mathcal{O}_x^k$ , then there exist  $c > 0$ , and open subsets  $U_x$  and  $V_x^{w_{1:k}}$  of  $X$  containing  $x$  and  $S_x^k(w_{1:k})$  respectively, such that*

$$P^k(y, A) \geq c \zeta_X(A) \quad \text{for every } y \in U_x \text{ and } A \in \mathcal{B}(X), \quad (4.12)$$

*for some nontrivial measure  $\zeta_X$  on  $V_x^{w_{1:k}}$ . That is,  $U_x$  is a  $k$ -small set.*

(ii) *If furthermore  $F$  is smooth (infinitely differentiable), and if there exist  $k > 0$ ,  $c > 0$  and  $(\varphi, V)$  a local chart of  $X$  such that*

$$P^k(x, A) \geq c \text{Leb} \circ \varphi(A \cap V) \quad \text{for every } A \in \mathcal{B}(X), \quad (4.13)$$

*then  $(R_x)$  holds.*



*Proof.* Condition  $(R_x)$  implies that  $\text{rank } \mathcal{D}S_x^k(w_{1:k}) = n$  for some  $k > 0$  and  $w_{1:k} \in \mathcal{O}_x^k$ . Since  $(\bar{x}, \bar{w}_{1:k}) \mapsto p_{\bar{x}}^k(\bar{w}_{1:k})$  is l.s.c., and  $p_x^k(w_{1:k}) > 0$ , then there exist  $p_0 > 0$  and a neighborhood  $X_1 \times W_1$  of  $(x, w_{1:k})$  such that  $p_{\bar{x}}^k(\bar{w}_{1:k}) \geq p_0$  for every  $\bar{x} \in X_1$  and every  $\bar{w}_{1:k} \in W_1$ . Then, for every  $y \in X_1$ , we have

$$P^k(y, A) = \int_{\mathcal{O}_y^k} \mathbb{1}_A(S_y^k(\bar{w}_{1:k})) p_y^k(\bar{w}_{1:k}) d\zeta_W^{\otimes k}(\bar{w}_{1:k}) \geq p_0 \int_{W_1} \mathbb{1}_A(S_y^k(\bar{w}_{1:k})) d\zeta_W^{\otimes k}(\bar{w}_{1:k}). \quad (4.14)$$

Since  $S_x^k$  is a submersion at  $w_{1:k}$ , by the Local Submersion theorem, there exists a local chart  $(V, \psi)$  of  $W^k$  around  $w_{1:k}$  and a local chart  $(U, \varphi)$  of  $X$  around  $S_x^k(w_{1:k})$ , such that for every  $(\bar{w}_1, \dots, \bar{w}_{kp}) \in \psi(V)$ , we have

$$\varphi \circ S_x^k \circ \psi^{-1}(\bar{w}_1, \dots, \bar{w}_{kp}) = (\bar{w}_1, \dots, \bar{w}_n). \quad (4.15)$$

Note that, up to taking  $V$  and  $W_1$  smaller, we can assume  $V = W_1$ , and that  $\psi(W_1) = \hat{W}_1 \times \tilde{W}_1$  is a rectangle of  $\mathbb{R}^{kp}$ , with  $\hat{W}_1 = \{(\bar{w}_1, \dots, \bar{w}_n) \in \mathbb{R}^n \mid \exists (\bar{w}_{n+1}, \dots, \bar{w}_{kp}) \in \mathbb{R}^{kp-n}, (\bar{w}_1, \dots, \bar{w}_{kp}) \in \psi(W_1)\} \subset \mathbb{R}^n$  and likewise  $\tilde{W}_1 \subset \mathbb{R}^{kp-n}$ . Hence, the function

$$G: X \times \hat{W}_1 \times \tilde{W}_1 \rightarrow X \\ (\bar{x}, (\bar{w}_1, \dots, \bar{w}_n), (\bar{w}_{n+1}, \dots, \bar{w}_{kp})) \mapsto S_{\bar{x}}^k \circ \psi^{-1}(\bar{w}_1, \dots, \bar{w}_{kp}) \quad (4.16)$$

satisfies, by (4.15), that  $\text{rank } \mathcal{D}_{(w_1, \dots, w_n)} G(x, w_1, \dots, w_{kp}) = n$ . Then, by Lemma 4.1 and **H1(ii)**, there exist  $c > 0$ ,  $U_x$  an open subset of  $X$  containing  $x$ ,  $(\varphi, V_x^{w_{1:k}})$  a local chart of  $X$  around  $S_x^k(w_{1:k})$  such that, for every  $y \in X$ , we have

$$\int_{W_1} \mathbb{1}_A(S_y^k(\bar{w}_{1:k})) d\zeta_W^k(\bar{w}_{1:k}) \geq c \mathbb{1}_{U_x}(y) \zeta_X(A \cap V_x^{w_{1:k}}), \quad (4.17)$$

with  $\zeta_X = \text{Leb} \circ \varphi(\cdot \cap V_x^{w_{1:k}})$  is a measure which satisfies the assumption required in Lemma 4.1 on  $V_x^{w_{1:k}}$ . Combining (4.14) and (4.17) gives  $P^k(y, A) \geq cp_0 \zeta_X(A \cap V_x^{w_{1:k}})$  for every  $y \in U_x \cap X_1$ , proving (i).

Suppose now that  $F$  is smooth, then  $(\bar{x}, \bar{w}_{1:k}) \mapsto S_{\bar{x}}^k(\bar{w}_{1:k})$  is smooth for all  $k > 0$ . Take  $k > 0$ ,  $c > 0$  and  $(\varphi, V)$  a local chart such that (4.13) holds. Let  $N = \{S_x^k(\bar{w}_{1:k}) \in X \mid \bar{w}_{1:k} \in \mathcal{O}_x^k, \text{rank } \mathcal{D}S_x^k(\bar{w}_{1:k}) < n\}$ . By Sard's theorem, we know that  $\text{Leb} \circ \varphi(N \cap V) = 0$ , implying that  $P^k(x, V \setminus N) \geq c \text{Leb} \circ \varphi(V \setminus N) = c \text{Leb} \circ \varphi(V) > 0$ . Hence there exists  $w_{1:k} \in \mathcal{O}_x^k$  such that  $S_x^k(w_{1:k}) \in V \setminus N$ , i.e.,  $\text{rank } \mathcal{D}S_x^k(w_{1:k}) = n$ .  $\square$

Following [10, Corollary 4.1], we now deduce sufficient conditions for the Markov kernel  $P$  to define a T-chain.

**Corollary 4.13.** *Consider the Markov kernel  $P$  defined via (2.1), and suppose **H1** and **H2**. Suppose that for any  $x \in X$ ,  $(C_x)$  holds. Then  $X$  can be written as the union of open small sets and thus  $P$  is a T-chain.*

*Proof.* First, using Proposition 4.8, for all  $x \in X$ ,  $(R_x)$  holds, i.e., there exist  $k > 0$  and  $u_{1:k} \in \mathcal{O}_x^k$  such that  $S_x^k$  is differentiable in  $u_{1:k}$  and  $\text{rank } \mathcal{D}S_x^k(u_{1:k}) = n$ .

Proposition 4.12 implies that for every  $x \in X$ , there exists an open neighborhood  $U_x$  of  $x$  in  $X$  which is a  $k$ -small set. Denoting  $a$  the Dirac distribution in  $k$ , we find that  $U_x$  is  $\nu_a$ -petite, hence, by [26, Proposition 6.2.3],  $K_a$  possesses a continuous component  $T$  which is nontrivial on  $U_x$  and in particular at  $x$ . Thus, by [26, Proposition 6.2.4],  $P$  is a T-chain.  $\square$

We now characterize the support of the maximal irreducibility measure of  $P$ . We recall that, by [26, Proposition 4.2.2], any  $\varphi$ -irreducible Markov kernel  $P$  admits a maximal irreducibility measure  $\psi$ , that is,  $P$  is  $\psi$ -irreducible and for every irreducibility measure  $\varphi$  of  $P$ , we have that  $\text{supp } \varphi \subset \text{supp } \psi$ . The proof mimics the one of [10, Proposition 4.2], and is given for completeness in Appendix C.

**Proposition 4.14.** *Suppose that  $P$  is a  $\psi$ -irreducible Markov kernel, defined via (2.1), with  $\psi$  a maximal irreducibility measure, that **H1** holds and that  $F$  is continuous. Then*

$$\text{supp } \psi = \{x^* \in X \mid x^* \text{ is globally attracting}\}. \quad (4.18)$$

Furthermore, if  $x^* \in X$  is globally attracting, then

$$\text{supp } \psi = \overline{A_+(x^*)}. \quad (4.19)$$

We now state our core results, from which we deduce Theorem 2.3. Assuming the controllability condition is satisfied at every  $x$ , there is equivalence between the irreducibility of  $P$  and the existence of a globally attracting state.

**Theorem 4.15.** *Consider the Markov kernel  $P$  defined via (2.1), and suppose **H1** and **H2**. Suppose  $(C_x)$  is satisfied for every  $x \in \mathsf{X}$ . Then  $P$  is  $\varphi$ -irreducible if and only if a globally attracting state exists.*

*Proof.* By Proposition 4.8, we know that  $(R_x)$  holds for any  $x \in \mathsf{X}$ . If  $P$  is  $\varphi$ -irreducible, then by Proposition 4.14, any point of the support of the nontrivial measure  $\varphi$  is globally attracting, hence there exists a globally attracting state. Conversely, if  $x^* \in \mathsf{X}$  is globally attracting, then, by Corollary 4.3,  $x^*$  is reachable by  $P$ , and by Corollary 4.13,  $P$  is a T-chain. As a result, by [26, Proposition 6.2.1],  $P$  is  $\varphi$ -irreducible.  $\square$

We deduce from this theorem our first practical result in order to prove the irreducibility, the T-chain property of a Markov kernel following the model investigated. If assumptions **H1** and **H2** are satisfied for a Markov kernel defined via (2.1), the theorem below implies that one needs to find a globally attracting state  $x^*$  where the controllability condition  $(C_{x^*})$  is satisfied to obtain the  $\varphi$ -irreducible and T-chain property of the Markov kernel.

**Theorem 4.16** (Practical condition for  $\varphi$ -irreducibility and T-chain property). *Consider the Markov kernel  $P$  defined via (2.1), and suppose **H1-H3**. Then  $P$  is a  $\varphi$ -irreducible T-chain, and thus every compact set of  $\mathsf{X}$  is petite.*

*Proof.* By Proposition 4.7, for any  $x \in \mathsf{X}$ ,  $(C_x)$  holds. Then, by Corollary 4.13,  $P$  is a T-chain and by Theorem 4.15,  $P$  is  $\varphi$ -irreducible, and by [26, Theorem 6.2.5] all compact sets of  $\mathsf{X}$  are petite.  $\square$

This latter theorem constitutes the first part of our main result stated in Theorem 2.3 while the second part relates to the aperiodicity of the kernel which is developed in the next section.

## 4.2.2 Aperiodicity

In this section, we provide conditions for  $P$  to be aperiodic. We start with the following characterization which is the counterpart to Theorem 4.15 for a kernel to be  $\varphi$ -irreducible aperiodic.

**Theorem 4.17.** *Consider the Markov kernel  $P$  defined via (2.1), and suppose **H1** and **H2**. If for every  $x \in \mathsf{X}$ ,  $(C_x)$  holds, then  $P$  is a  $\varphi$ -irreducible aperiodic Markov kernel if and only if there exists a steadily attracting state.*

*Proof.* First suppose that  $P$  is  $\varphi$ -irreducible and aperiodic. By Theorem 4.15, there exists a globally attracting state  $x^* \in \mathsf{X}$ . Besides, by Proposition 4.10, there exists an attainable state  $y^*$ , to which we apply Proposition 4.6 (iii), so that there exists a  $d$ -cycle. However,  $P$  is aperiodic, so  $d = 1$ . Thus, by Proposition 4.6 (ii),  $y^*$  is steadily attracting.

Conversely, suppose that there exists a steadily attracting state  $x^*$ . By Proposition 4.4 (i),  $x^*$  is globally attracting, so that by Theorem 4.15,  $P$  is  $\varphi$ -irreducible. It remains to prove that it is aperiodic. By Proposition 4.8,  $(R_{x^*})$  holds for some  $k > 0$  and  $w_{1:k}^* \in \mathcal{O}_{x^*}^k$ . Therefore, we apply Proposition 4.10, hence  $y^* := S_{x^*}^k(w_{1:k}^*)$  is attainable. Let  $U$  be a neighborhood of  $x^*$  which satisfies Proposition 4.10 (i). Since  $x^*$  is steadily attracting, there exists  $T > 0$ , such that for every  $t \geq T$ , there exists  $u_{1:t} \in \mathcal{O}_{y^*}^t$  such that  $z := S_{y^*}^t(u_{1:t}) \in U$ . As  $U$  satisfies Proposition 4.10 (i), then there exists  $w_{1:k} \in \mathcal{O}_z^k$  such that  $S_z^k(w_{1:k}) = S_{x^*}^k(w_{1:k}^*) = y^*$ . All in all, we have that for every  $t \geq T$ , there exists  $w_{1:k+t} \in \mathcal{O}_{y^*}^{k+t}$  such that  $y^* = S_{y^*}^{k+t}(w_{1:k+t})$ , hence  $y^* \in A_+^{k+t}(y^*)$ . By Proposition 4.6 (iii), there exists a 1-cycle, i.e.,  $P$  is aperiodic.  $\square$

We now state our main practical condition to ensure that  $P$  is aperiodic.

**Theorem 4.18** (Practical condition for  $\varphi$ -irreducibility and aperiodicity). *Consider the Markov kernel  $P$  defined via (2.1), and suppose **H1-H2** and **H4**. Then  $P$  is a  $\varphi$ -irreducible aperiodic T-chain, and every compact set of  $\mathsf{X}$  is small.*

*Proof.* By Proposition 4.7,  $(C_x)$  holds for any  $x \in \mathsf{X}$ . Thus, by Theorems 4.16 and 4.17, we have that  $P$  is a  $\varphi$ -irreducible aperiodic T-chain for which compact sets of  $\mathsf{X}$  are petite. Note that, by [26, Theorem 5.5.7], any petite set is small.  $\square$

## 4.3 Proofs for the application to CMA-ES

*Proof of Proposition 3.2 (i).* By Corollary 4.5, it is sufficient to find, for any  $\theta_0 = (z_0, \Sigma_0) \in \mathsf{X}$ , a sequence  $\{v_k\}_{k \geq 1}$  such that  $v_{1:k} \in \overline{\mathcal{O}_{\theta_0}^k}$  for every  $k \geq 1$ , and  $\lim_{k \rightarrow \infty} S_{\theta_0}^k(v_{1:k}) = (0, I_d)$ . Since  $f$  has Lebesgue negligible level sets, we have, by Proposition 3.1, that for every  $(z, \Sigma) \in \mathsf{X}$ , and for every  $u \in \mathbb{R}^d$ , the element  $v = (u, \dots, u)$  of  $\mathsf{W} = \mathbb{R}^{d\mu}$  belongs to  $\overline{\mathcal{O}_{(z, \Sigma)}^1}$ .

Then, set  $v_1^1 = \dots = v_1^\mu = -\Sigma_0^{-1/2} z_0$  so that  $v_1 = (v_1^1, \dots, v_1^\mu) \in \overline{\mathcal{O}_{\theta_0}^1}$  and  $z_1 = 0$ . Note that  $S^1(v_1) = \theta_1 := (0, \Sigma_1)$ , for some  $\Sigma_1 \in \mathcal{S}_{++}^d$ . Next, consider  $(e_1, \dots, e_d)$  an orthogonal basis of eigenvectors of the positive definite matrix  $\Sigma_1$ , with  $\Sigma_1 e_i = \lambda_i(\Sigma_1) e_i$ , where  $\lambda_i(\cdot)$  denotes the function that maps a symmetric matrix to its  $i$ -th largest eigenvalue (counted with multiplicity).

Let then  $\kappa \geq 0$  and define  $v_2^1 = \dots = v_2^\mu = -\Sigma_1^{-1/2} \kappa e_2$  and  $\theta_2 = (z_2, \Sigma_2) = S_{\theta_1}^1(v_2)$ . Then let  $v_3^1 = \dots = v_3^\mu = -\Sigma_2^{-1/2} z_2$  and  $\theta_3 = (z_3, \Sigma_3) = S_{\theta_2}^1(v_3)$ . Then, we have  $z_2 = \kappa e_2$  and  $z_3 = 0$ . Besides,  $\Sigma_2 = r_2^{-1} \times ((1-c)\Sigma_1 + c\kappa^2 e_2 e_2^\top)$  with  $r_2 = \det((1-c)\Sigma_1 + c\kappa^2 e_2 e_2^\top)^{1/d}$  depends continuously on the choice of  $\kappa \geq 0$ . Moreover, we have  $1-c \leq r_2 \leq 1-c + c\kappa^2$ . Then,

$$r_3 \Sigma_3 = \Sigma_1 + c(1-c)^{-2} \kappa^2 \times (r_2 + 1 - c) e_2 e_2^\top =: K_3 \quad (4.20)$$

for some  $r_3 > 0$ . However, the eigenvalue of the matrix  $K_3$  associated to the eigenvector  $e_1$  equals  $\lambda_1(\Sigma_1)$  for any value of  $\kappa$ , while the eigenvalue of  $K_3$  associated to the eigenvector  $e_2$  depends continuously on  $\kappa \geq 0$  and tends to  $+\infty$  when  $\kappa \rightarrow \infty$  and to  $\lambda_2(\Sigma_1) \leq \lambda_1(\Sigma_1)$  when  $\kappa \rightarrow 0$ . Hence, there exists a value of  $\kappa \geq 0$  such that the eigenvalues of  $K_3$  associated respectively to the eigenvectors  $e_1$  and  $e_2$  are equal. Setting  $\kappa$  to this value, we get then that  $\lambda_1(\Sigma_3) = \lambda_2(\Sigma_3)$ .

Repeating eventually these steps  $(d-1)$  times, we find  $v_4, \dots, v_{1+2(d-1)} \in \mathbb{W}$  such that, denoting  $\theta_k = (z_k, \Sigma_k) = S_{\theta_0}^k(v_{1:k})$ ,  $z_{1+2(d-1)} = 0$ , and  $\lambda_1(\Sigma_{1+2(d-1)}) = \dots = \lambda_d(\Sigma_{1+2(d-1)})$ , with  $v_k \in \overline{\mathcal{O}_{\theta_{k-1}}^1}$  for each  $k > 0$ . However,  $\det(\Sigma_k) = 1$  for every  $k \in \mathbb{N}$ , thus  $\Sigma_{1+2(d-1)} = I_d$ .

For the next steps  $k \geq 2d$ , we choose  $v_k = 0 \in \overline{\mathcal{O}_{\theta_{k-1}}^1}$ , so that, by induction, we obtain  $\theta_k = (0, I_d)$ . By Corollary 4.5, we find that  $(0, I_d)$  is a steadily attracting state.  $\square$

*Proof of Proposition 3.2 (ii).* Let  $k > 0$  and  $v_{1:k} \in \overline{\mathcal{O}_{\theta_0}^k}$ , with  $\theta_0 = (z_0, \Sigma_0) := (0, I_d)$ . We find here values for  $k > 0$  and  $v_1, \dots, v_k \in \mathbb{W}$  such that the map

$$\mathcal{D}S_{(0, I_d)}^k(v_{1:k}) : T_{(v_{1:k})} \mathbb{W}^k \rightarrow T_{S_{(0, I_d)}^k(v_{1:k})} \mathbb{X}$$

is full-rank, i.e., is surjective. We remind that  $\mathbb{W} = (\mathbb{R}^d)^\mu$ , hence  $T_{(v_{1:k})} \mathbb{W}^k = \mathbb{W}^k = \mathbb{R}^{d \times \mu \times k}$ . Moreover, we have  $\mathbb{X} = \mathbb{R}^d \times \det^{-1}(\{1\})$ , therefore, by [22, Proposition 5.38]

$$T_{S_{(0, I_d)}^k(v_{1:k})} \mathbb{X} = \mathbb{R}^d \times \ker \mathcal{D} \det(\Sigma_k),$$

where  $\theta_t = (z_t, \Sigma_t) = S_{\theta_0}^t(v_{1:t})$  for each  $t = 0, \dots, k$ .

We define then inductively the covariance matrix before normalization as

$$K_{t+1} = (1-c)K_t + c\sqrt{K_t} \sum_{i=1}^{\mu} w_i (v_{t+1}^i) (v_{t+1}^i)^\top \sqrt{K_t}$$

with  $K_0 = \Sigma_0 = I_d$ , so that, by induction, we have for every  $t = 0, \dots, k$ ,  $\Sigma_t = \frac{K_t}{\det(K_t)^{1/d}}$ . Let us introduce (small) perturbations  $h_t = (h_t^1, \dots, h_t^\mu) \in \mathbb{W}$  for  $t = 1, \dots, k$ , and let us denote the perturbed process as

$$\theta_t^h = (z_t^h, \Sigma_t^h) = S_{\theta_0}^t(v_{1:t} + h_{1:t}).$$

Define  $K_t^h \in \mathcal{S}_{++}^d$  similarly. Set  $k_0 = d(d+1)/2$  the dimension of  $\mathcal{S}^d$ , and set  $k = k_0(k_0 - 1) + 1$ . Then, set  $v_1, \dots, v_{k_0}$  as follows. Define  $\psi_1, \dots, \psi_{k_0}$  nonzero vectors of  $\mathbb{R}^d$ , such that  $(\psi_1 \psi_1^\top, \dots, \psi_{k_0} \psi_{k_0}^\top)$  forms a basis of  $\mathcal{S}^d$ .

For  $t = 1, \dots, k_0$ , using Proposition 3.1, we set  $v_t = (K_{t-1}^{-1/2} \psi_t, \dots, K_{t-1}^{-1/2} \psi_t) \in \overline{\mathcal{O}_{\theta_{t-1}}^1}$ , so that  $K_t = (1-c)K_{t-1} + c\psi_t \psi_t^\top$ . Fix then  $\kappa_t^1 \in \mathbb{R}$ , and let  $\varepsilon_1 > 0$  be an arbitrary small positive quantity. Set  $h_t^1 = \dots = h_t^\mu = \frac{1}{2} \kappa_t^1 \varepsilon_1 K_{t-1}^{-1/2} \psi_t$  and then

$$K_t^h = (1-c)K_{t-1}^h + c\psi_t \psi_t^\top + \varepsilon_1 \kappa_t^1 c\psi_t \psi_t^\top + \varepsilon_1 A_t^1(\varepsilon_1),$$

where  $A_t^1(\varepsilon_1) \in \mathcal{S}^d$  tends to 0 when  $\varepsilon_1 \rightarrow 0$ . Then, we get by induction,

$$K_{k_0}^h = K_{k_0} + \varepsilon_1 \sum_{t=1}^{k_0} \kappa_t^1 (1-c)^{k_0-t} c\psi_t \psi_t^\top + \varepsilon_1 A_{k_0}^1(\varepsilon_1).$$

Likewise,  $A_{k_0}^1(\varepsilon_1)$  defines a symmetric matrix which then tends to 0 when  $\varepsilon_1$  tends to 0. Repeat these steps  $k_0 - 1$  times with  $\varepsilon_2, \dots, \varepsilon_{k_0-1} > 0$  instead of  $\varepsilon_1 > 0$  and  $\kappa_t^2, \dots, \kappa_t^{k_0-1} \in \mathbb{R}$  instead of  $\kappa_t^1 \in \mathbb{R}$ . All in all, we have finally, since  $k = k_0(k_0 - 1) + 1$ ,

$$K_{k-1}^h = K_{k-1} + \sum_{s=1}^{k_0-1} \left[ \varepsilon_s \sum_{t=1}^{k_0} \kappa_t^s (1-c)^{k_0(k_0-1)-sk_0+k_0-t} c \psi_t \psi_t^\top + \varepsilon_s A_{k-1}^s(\varepsilon_s) \right].$$

Again, for each  $s = 1, \dots, k_0 - 1$ ,  $A_{k-1}^s(\varepsilon_s)$  defines a symmetric matrix which tends to 0 when  $\varepsilon_s$  tends to 0. Now, consider  $(S_1, \dots, S_{k_0-1})$  a basis of  $\ker \mathcal{D} \det(\Sigma_{k-1})$ . For  $s = 1, \dots, k_0 - 1$ , we set now the real values  $\kappa_t^s$ ,  $t = 1, \dots, k_0$  such that we have

$$\sum_{t=1}^{k_0} \kappa_t^s (1-c)^{k-1-sk_0+k_0-t} c \psi_s \psi_s^\top = S_s.$$

This is possible since  $(\psi_1 \psi_1^\top, \dots, \psi_{k_0} \psi_{k_0}^\top)$  is a basis of  $\mathcal{S}^d$ . Then,

$$K_{k-1}^h = K_{k-1} + \sum_{t=1}^{k_0-1} \varepsilon_t S_t + \varepsilon_t A_{k-1}^t(\varepsilon_t).$$

Yet, since  $S_t \in \ker \mathcal{D} \det(\Sigma_{k-1})$ , we have then

$$\begin{aligned} \Sigma_{k-1}^h &= \frac{K_{k-1}^h}{\det(K_{k-1}^h)^{1/d}} = \frac{K_{k-1} + \sum_{t=1}^{k_0-1} \varepsilon_t S_t + \varepsilon_t A_{k-1}^t(\varepsilon_t)}{\det\left(K_{k-1} + \sum_{t=1}^{k_0-1} \varepsilon_t S_t + \varepsilon_t A_{k-1}^t(\varepsilon_t)\right)^{1/d}} \\ &= \Sigma_{k-1} + r \sum_{t=1}^{k_0-1} \varepsilon_t S_t + \varepsilon_t B^t(\varepsilon_t), \end{aligned}$$

where we set  $r = \det(K_{k-1})^{-1/d}$ , and the symmetric matrices  $B^t(\varepsilon_t)$  tend to 0 when  $\varepsilon_t \rightarrow 0$ , for  $t = 1, \dots, k_0 - 1$ . Lastly, set  $v_k = 0 \in \mathcal{O}_{\theta_{k-1}}^1$ , and let  $h_k^1 = \dots = h_k^\mu = \Sigma_{k-1}^{-1/2} \xi_{k_0}$ , for some arbitrary small vector  $\xi_{k_0} \in \mathbb{R}^d$ . Then,  $\Sigma_k = \Sigma_{k-1}$  and

$$z_k^h = z_k + (1-c)^{-1/2} \xi_{k_0} + l(\varepsilon_1, \dots, \varepsilon_{k_0-1}) + \|(\varepsilon_1, \dots, \varepsilon_{k_0-1}, \xi_{k_0})\| h(\varepsilon_1, \dots, \varepsilon_{k_0-1}, \xi_{k_0}),$$

where  $l: \mathbb{R}^{k_0-1} \rightarrow \mathbb{R}^d$  is linear and  $h(\varepsilon_1, \dots, \varepsilon_{k_0-1}, \xi_{k_0})$  tends to 0 when  $\|(\varepsilon_1, \dots, \varepsilon_{k_0-1}, \xi_{k_0})\| \rightarrow 0$ . Furthermore,

$$\Sigma_k^h = \Sigma_k + r \sum_{t=1}^{k_0-1} \varepsilon_t S_t + \varepsilon_t B^t(\varepsilon_t) + c \xi_{k_0} \xi_{k_0}^\top.$$

Finally,

$$\frac{S_{(0, I_d)}^k(v_{1:k} + h_{1:k}) - S_{(0, I_d)}^k(v_{1:k}) - \left( \frac{(1-c)^{-1/2} \xi_{k_0} + l(\varepsilon_1, \dots, \varepsilon_{k_0-1})}{r \sum_{t=1}^{k_0-1} \varepsilon_t S_t} \right)}{\|(\varepsilon_1, \dots, \varepsilon_{k_0-1}, \xi_{k_0})\|} \quad (4.21)$$

tends to 0 when  $\|(\varepsilon_1, \dots, \varepsilon_{k_0-1}, \xi_{k_0})\| \rightarrow 0$ . Therefore,

$$\mathcal{D}S_{(0, I_d)}^k(v_{1:k}) h_{1:k} = \left( \frac{(1-c)^{-1/2} \xi_{k_0} + l(\varepsilon_1, \dots, \varepsilon_{k_0-1})}{r \sum_{t=1}^{k_0-1} \varepsilon_t S_t} \right) \quad (4.22)$$

defines a surjective map from  $\mathbb{W}^k$  to  $\mathbb{R}^d \times \ker \mathcal{D} \det(\Sigma_k)$ . Indeed, if  $H_\Sigma \in \ker \mathcal{D} \det(\Sigma_k)$  and  $h_z \in \mathbb{R}^d$ , then there exist  $\varepsilon_1, \dots, \varepsilon_{k_0-1} \in \mathbb{R}$  such that  $r \sum_{t=1}^{k_0-1} \varepsilon_t S_t = H_\Sigma$ , and then there exists  $\xi_{k_0} \in \mathbb{R}^d$  such that  $\mathcal{D}S_{(0, I_d)}^k(v_{1:k}) h_{1:k} = (h_z; H_\Sigma)$ .  $\square$

## References

- [1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

- [2] Youhei Akimoto, Yuichi Nagata, Isao Ono, and Shigenobu Kobayashi. Bidirectional Relation between CMA Evolution Strategies and Natural Evolution Strategies. In Robert Schaefer, Carlos Cotta, Joanna Kołodziej, and Günter Rudolph, editors, *Parallel Problem Solving from Nature, PPSN XI*, Lecture Notes in Computer Science, pages 154–163, Berlin, Heidelberg, 2010. Springer.
- [3] Youhei Akimoto, Yuichi Nagata, Isao Ono, and Shigenobu Kobayashi. Theoretical Foundation for CMA-ES from Information Geometry Perspective. *Algorithmica*, 64(4):698–716, December 2012.
- [4] Herbert Amann and Joachim Escher. *Analysis III*. Birkhäuser, Basel, 2009.
- [5] H. Z. An and S. G. Chen. A note on the ergodicity of non-linear autoregressive model. *Statistics & Probability Letters*, 34(4):365–372, June 1997.
- [6] Anne Auger. Thèse d’habilitation à diriger des recherches “Analysis of Comparison-based Stochastic Continuous Black-Box Optimization Algorithms”, May 2016.
- [7] Anne Auger and Nikolaus Hansen. Linear Convergence of Comparison-based Step-size Adaptive Randomized Search via Stability of Markov Chains. *SIAM Journal on Optimization*, 26(3):1589–1624, January 2016.
- [8] Rabi Bhattacharya and Chanho Lee. On geometric ergodicity of nonlinear autoregressive models. *Statistics & Probability Letters*, 22(4):311–315, March 1995.
- [9] Jonathan Bieler, Rosamaria Cannavo, Kyle Gustafson, Cedric Gobet, David Gatfield, and Felix Naef. Robust synchronization of coupled circadian and cell cycle oscillators in single mammalian cells. *Molecular systems biology*, 10(7):739, 2014.
- [10] Alexandre Chotard and Anne Auger. Verifiable conditions for the irreducibility and aperiodicity of Markov chains by analyzing underlying deterministic models. *Bernoulli*, 25(1):112–147, February 2019.
- [11] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, January 1990.
- [12] Lawrence Craig Evans and Ronald F Gariepy. *Measure Theory and Fine Properties of Functions, Revised Edition*. Chapman and Hall/CRC, New York, April 2015.
- [13] Peter W Glynn, John E Glynn, and Sanatan Rai. Recurrence classification for a family of non-linear storage models. *Probability and Mathematical Statistics*, 37, 2017.
- [14] Victor Guillemin and Alan Pollack. *Differential Topology*. American Mathematical Soc., 2010.
- [15] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.
- [16] Piotr Hajlasz. Change of variables formula under minimal assumptions. In *Colloquium Mathematicae*, volume 64, pages 93–101, 1993.
- [17] Nikolaus Hansen, Sibylle D. Müller, and Petros Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation*, 11(1):1–18, March 2003.
- [18] Nikolaus Hansen and Andreas Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, June 2001.
- [19] Jianyi Huang, Ioannis Kontoyiannis, and Sean P. Meyn. The ODE Method and Spectral Theory of Markov Operators. In Bozena Pasik-Duncan, editor, *Stochastic Theory and Control*, Lecture Notes in Control and Information Sciences, pages 205–221, Berlin, Heidelberg, 2002. Springer.
- [20] Kanji Ichihara and Hiroshi Kunita. A classification of the second order degenerate elliptic operators and its probabilistic characterization. *Z. Wahrscheinlichkeitstheorie verw Gebiete*, 30(3):235–254, 1974.
- [21] Wolfgang Kliemann. Recurrence and Invariant Measures for Degenerate Diffusions. *The Annals of Probability*, 15(2):690–707, 1987.
- [22] John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2012.

- [23] S. Meyn and P. Caines. A new approach to stochastic adaptive control. *IEEE Transactions on Automatic Control*, 32(3):220–226, March 1987.
- [24] S. P. Meyn and P. E. Caines. Stochastic controllability and stochastic Lyapunov functions with applications to adaptive and nonlinear systems. In Norbert Christopeit, Kurt Helmes, and Michael Kohlmann, editors, *Stochastic Differential Systems*, Lecture Notes in Control and Information Sciences, pages 235–257, Berlin, Heidelberg, 1989. Springer.
- [25] S. P. Meyn and P. E. Caines. Asymptotic Behavior of Stochastic Systems Possessing Markovian Realizations. *SIAM Journal on Control and Optimization*, 29(3):535–561, May 1991.
- [26] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, December 2012.
- [27] S.P. Meyn and L.J. Brown. Model reference adaptive control of time varying and stochastic systems. *IEEE Transactions on Automatic Control*, 38(12):1738–1753, December 1993.
- [28] Abdelkader Mokkadem. *Critères de mélange pour des processus stationnaires. Estimation sous des hypothèses de mélange. Entropie des processus linéaires*. PhD thesis, Université Paris-Sud, September 1987.
- [29] Cécile Patte, Pierre-Yves Brillet, Catalin Fetita, Jean-François Bernaudin, Thomas Gille, Hilario Nunes, Dominique Chapelle, and Martin Genet. Estimation of regional pulmonary compliance in idiopathic pulmonary fibrosis based on personalized lung poromechanical modeling. *Journal of Biomechanical Engineering*, 144(9):091008, 2022.
- [30] Maria Rodriguez-Fernandez, Pedro Mendes, and Julio R Banga. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems*, 83(2-3):248–265, 2006.
- [31] Daniel W. Stroock and S. R. S. Varadhan. On the Support of Diffusion Processes with Applications to the Strong Maximum Principle. In *Contributions to Probability Theory*, pages 333–360. University of California Press, December 1972.
- [32] Cheikh Toure, Anne Auger, and Nikolaus Hansen. Global linear convergence of evolution strategies with recombination on scaling-invariant functions. *Journal of Global Optimization*, 86(1):163–203, May 2023.
- [33] Cheikh Toure, Armand Gissler, Anne Auger, and Nikolaus Hansen. Scaling-invariant Functions versus Positively Homogeneous Functions. *Journal of Optimization Theory and Applications*, 191(1):363–383, October 2021.
- [34] J.-F. Yao and J.-G. Attali. On stability of nonlinear AR processes with Markov switching. *Advances in Applied Probability*, 32(2):394–407, June 2000.

## A Background on manifolds

We call throughout the paper a manifold a smooth manifold. We recall below the definition of a manifold, and we refer to [4] for more details.

*Definition A.1* (Manifolds). A topological space  $X$  is said to be a topological manifold of dimension  $n$  if it is a second countable Hausdorff space that is locally Euclidean of dimension  $n$ .

Note that  $X$  is said to be a Hausdorff space if for every pair of distinct points  $x, y \in X$ , there exist neighborhoods  $U$  of  $x$  and  $V$  of  $y$  that are disjoint. Moreover,  $X$  is said to be second countable if there exists a countable basis, that is, a countable collection  $\mathcal{B}$  of open subsets of  $X$  such that any open subset of  $X$  can be written as the union of sets in  $\mathcal{B}$ .

Finally,  $X$  is locally Euclidean when for every  $x \in X$ , there exists a neighborhood  $U$  of  $x$ , an open set  $V$  of  $\mathbb{R}^n$  and a homeomorphism (i.e., a continuous bijection with a continuous reciprocal function)  $\varphi: U \rightarrow V$ . We call  $(\varphi, U)$  a chart around  $x$ .

Besides, a manifold  $X$  is said to be smooth if it is topological, locally Euclidean, and if every charts  $(\varphi, U)$  and  $(\psi, V)$  around any point  $x \in X$  are such that  $\varphi \circ \psi^{-1}$  is continuously differentiable.

Given  $X$  a ( $n$ -dimensional) manifold, and  $x \in X$ , we denote by  $T_x X$  the tangent space of  $X$  in  $x$ . We refer to [4, Chapter XI] or to [14, Chapter 1, Section 2] for a formal definition of tangent spaces.

We introduce the measurability on a smooth manifold via the following definition. We refer to [4, Chapter XII] for further details.

*Definition A.2.* A subset  $A \subset X$  is said to be measurable if for all  $x \in X$ , there exists a chart around  $x$  denoted  $(\varphi, U)$  such that  $\varphi(A \cap U)$  is measurable (in  $\mathbb{R}^n$ ).

## B Clarke's generalized derivative of locally Lipschitz functions on manifolds

Clarke's generalized Jacobian is defined for locally Lipschitz functions  $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$  [11]. We define here the Clarke's derivative for locally Lipschitz functions  $f: X \rightarrow Y$  where  $X$  and  $Y$  are smooth manifolds. First, let us define formally what a locally Lipschitz function between manifolds is.

*Definition B.3.* Let  $X$  and  $Y$  be two manifolds, equipped with their distance functions  $d_X$  and  $d_Y$  respectively, and  $f: X \rightarrow Y$  a function.

- (i)  $f$  is said to be Lipschitz if there exists  $L > 0$  such that for all  $x, y \in X$  we have  $d_Y(f(x), f(y)) \leq L \times d_X(x, y)$ .
- (ii)  $f$  is said to be locally Lipschitz at  $x \in X$  if there exists a neighborhood  $U$  of  $x$  in  $X$  such that the restriction of  $f$  to  $U$  is Lipschitz.

As stated below, a function is locally Lipschitz if and only if it is locally Lipschitz in the charts.

**Proposition B.4.** *If  $f: X \rightarrow Y$  is locally Lipschitz at  $x \in X$ , then for all local charts  $(\varphi, U)$  of  $X$  around  $x$  and  $(\psi, V)$  of  $Y$  around  $f(x)$ , the function  $\psi \circ f \circ \varphi^{-1}$  is locally Lipschitz at  $\varphi(x)$ .*

*Proof.* See that both  $\varphi^{-1}$  and  $\psi$  are  $\mathcal{C}^1$  hence are locally Lipschitz at all points of their domains. By composition we find that  $\psi \circ f \circ \varphi^{-1}$  is locally Lipschitz at  $\varphi(x)$ .  $\square$

Rademacher's theorem [12, Theorem 3.2], states that a locally Lipschitz function is almost everywhere differentiable. This is easily extended to locally Lipschitz functions on manifolds.

**Corollary B.5** (Rademacher's theorem). *Let  $\zeta_X$  be a measure on  $X$ , which is locally equivalent to the Lebesgue measure, that is, for any measurable subset  $A$  of  $X$ , then  $\zeta_X(A) = 0$  if and only if for every charts  $(\varphi, U)$  of  $X$ ,  $\text{Leb} \circ \varphi(A \cap U) = 0$ . Then, any function  $f: X \rightarrow Y$  locally Lipschitz at every  $x \in X$ , is differentiable  $\zeta_X$ -almost everywhere.*

*Proof.* Consider local charts  $(\varphi, U)$  of  $X$  around  $x$  and  $(\psi, V)$  of  $Y$  around  $f(x)$ . Let us prove that for  $\zeta_X$ -almost every point  $y$  of  $U$ ,  $f$  is differentiable at  $y$ . See that by Proposition B.4,  $g := \psi \circ f \circ \varphi^{-1}$  is locally Lipschitz on  $\varphi(U)$ . Thus, by [12, Theorem 3.2], we have that  $g$  is differentiable  $\text{Leb}$ -almost everywhere on  $\varphi(U)$ . Thus, since  $\varphi$  and  $\psi^{-1}$  are  $\mathcal{C}^1$ , and since the measures  $\zeta_X$  and  $\text{Leb} \circ \varphi$  are equivalent on  $U$ , then  $f = \psi^{-1} \circ g \circ \varphi$  is differentiable  $\zeta_X$ -almost everywhere.  $\square$

We give now the definition of Clarke's Jacobian for locally Lipschitz functions on Euclidean spaces.

*Definition B.6* (Clarke's generalized Jacobian). Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be locally Lipschitzian at  $x_0 \in \mathbb{R}^n$ . Define

$$\partial f(x_0) = \text{conv} \left\{ \lim_{t \rightarrow \infty} \mathcal{D}f(x_t) \mid x_t \rightarrow x_0, f \text{ is differentiable in all } x_t \right\} \quad (\text{B.1})$$

where  $\mathcal{D}f(x_t) \in \mathbb{R}^{n \times m}$  is the Jacobian matrix of  $f$  at  $x_t$  (when defined) and  $\text{conv}$  denotes the convex hull.

We generalize now this definition to locally Lipschitz functions on manifolds.

*Proposition and Definition B.7* (Clarke's generalized Jacobian on manifolds). Let  $X$  and  $Y$  be two manifolds. Let  $f: X \rightarrow Y$  be locally Lipschitzian at  $x_0 \in X$ . Let  $(\varphi, U)$  be a local chart of  $X$  around  $x_0$  and  $(\psi, V)$  be a local chart of  $Y$  around  $f(x_0)$ . Define  $g = \psi \circ f \circ \varphi^{-1}$ . Then  $g: \varphi(U) \rightarrow \psi(V)$  is locally Lipschitz at  $\varphi(x_0)$ , and we can define

$$\partial f(x_0) = \left\{ \mathcal{D}\psi^{-1}(g \circ \varphi(x_0)) \circ h \circ \mathcal{D}\varphi(x_0) \mid h \in \partial g(\varphi(x_0)) \right\}. \quad (\text{B.2})$$

*Proof.* The maps  $\psi$  and  $\varphi^{-1}$  are by definition continuously differentiable, hence are locally Lipschitz. Therefore, by composition,  $g$  is locally Lipschitz. Furthermore, note that the expression (2.22) does not depend on the choice of the charts. Indeed, let  $(\varphi_1, U)$  and  $(\varphi_2, U)$  be two charts of  $X$  at  $x_0$  and  $(\psi_1, V)$  and  $(\psi_2, V)$  be two charts of  $Y$  at  $f(x_0)$ , such that  $g_1 = \psi_1 \circ f \circ \varphi_1^{-1}$  and  $g_2 = \psi_2 \circ f \circ \varphi_2^{-1}$  are well defined. Then, note that  $g_2 = \psi_2 \circ \psi_1^{-1} \circ g_1 \circ \varphi_1 \circ \varphi_2^{-1}$ . Apply then the chain rule [11, Corollary of Theorem 2.6.6] to  $g_2$  and get

$$\begin{aligned} \partial g_2(\varphi_2(x_0)) \\ = \{ \mathcal{D}\psi_2(f(x_0))\mathcal{D}\psi_1^{-1}(g_1(\varphi_1(x_0))) \circ h \circ \mathcal{D}\varphi_1(x_0)\mathcal{D}\varphi_2^{-1}(\varphi_2(x_0)) \mid h \in \partial g_1(\varphi_1(x_0)) \}. \end{aligned}$$

Therefore,

$$\begin{aligned} \{ \mathcal{D}\psi_1^{-1}(g_1 \circ \varphi_1(x_0)) \circ h \circ \mathcal{D}\varphi_1(x_0) \mid h \in \partial g_1(\varphi_1(x_0)) \} \\ = \{ \mathcal{D}\psi_2^{-1}(g_2 \circ \varphi_2(x_0)) \circ h \circ \mathcal{D}\varphi_2(x_0) \mid h \in \partial g_2(\varphi_2(x_0)) \}. \end{aligned}$$

□

We also state the next result, which would be useful to prove Proposition 4.9.

**Proposition B.8.** *If  $f: X \rightarrow Y$  is locally Lipschitzian at  $x_0 \in X$ , then  $\partial f(x_0)$  is nonempty, compact and convex.*

*Proof.* This follows from [11, Proposition 2.6.2(a)] and Proposition and Definition 2.2. □

We now transpose the uppercontinuity of Clarke's Jacobians to the context of locally Lipschitz functions between manifolds.

**Proposition B.9.** *Let  $f: X \rightarrow Y$  be locally Lipschitz at  $x_0$ . Then,  $\lim_{x \rightarrow x_0} \partial f(x) \subset \partial f(x_0)$ .*

*Proof.* Let  $(\varphi, U)$  and  $(\psi, V)$  be two local charts respectively of  $X$  and  $Y$  at  $x_0$  and  $f(x_0)$ , such that  $\tilde{f} = \psi \circ f \circ \varphi^{-1}$  is well defined. Then, by Proposition and Definition B.7, we have, for any  $x \in U$

$$\partial f(x) = \left\{ \mathcal{D}\psi^{-1}(\tilde{f} \circ \varphi(x)) \circ h \circ \mathcal{D}\varphi(x) \mid h \in \partial \tilde{f}(\varphi(x)) \right\}.$$

By applying [11, Proposition 2.6.2] to  $\tilde{f}$ , we find that  $\lim_{x \rightarrow x_0} \partial \tilde{f}(\varphi(x)) \subset \partial \tilde{f}(x_0)$ , which ends the proof. □

The next proposition is actually a very important requirement for our analysis. It states that if we can find a point for which the generalized differential of a locally Lipschitz function in this point is of maximal rank, then we can find a point closed to it in which the function is differentiable and the derivative is full rank.

**Proposition B.10.** *Suppose that  $f: X \rightarrow Y$  is locally Lipschitzian at  $x_0 \in X$ , and that  $\partial f(x_0)$  is of maximal rank, i.e., all  $h \in \partial f(x_0)$  is of maximal rank. Then, there exists a neighborhood  $U$  of  $x_0$  such that for all  $y \in U$ ,  $\partial f(y)$  is of maximal rank. Moreover, for all neighborhood  $V \subset U$  of  $x_0$ , there exists  $y_0 \in V$  such that  $f$  is differentiable at  $y_0$  and  $\mathcal{D}f(y_0)$  is of maximal rank.*

*Proof.* Let  $A = \{h \in \mathcal{L}(T_{x_0}X, T_{f(x_0)}Y) \mid h \text{ is not of maximal rank}\}$ . Since the application rank is l.s.c., then  $A$  is a closed set. By Proposition B.8,  $\partial f(x_0)$  is compact, and disjoint from  $A$  since it is assumed to be of maximal rank. Thus  $d(\partial f(x_0), A) > 0$ , where  $d$  is a metric induced by some norm on the affine space  $\mathcal{L}(T_{x_0}X, T_{f(x_0)}Y)$  of finite dimension. Moreover, there exists  $h^* \in \partial f(x_0)$  such that for all  $h \in \partial f(x_0)$  we have

$$d(h, A) \geq d(h^*, A) = d(\partial f(x_0), A) > 0,$$

By [11, Proposition 2.6.2(c)], there exists a neighborhood  $U$  of  $x_0$  such that for all  $y \in U$ ,  $d(\partial f(y), A) \geq d(h^*, A)/2 > 0$ , thus  $\partial f(y)$  is of maximal rank. The second part follows from Corollary B.5. □

Next, we state a chain rule for the generalized Jacobian on manifolds.

**Proposition B.11** (Chain rule). *Let  $X, Y$  and  $Z$  be three manifolds. If  $f: X \rightarrow Y$  is locally Lipschitz at  $x_0 \in X$ , and if  $g: Y \rightarrow Z$  is differentiable at  $f(x_0)$ , then we have*

$$\partial(g \circ f)(x_0) = \{\mathcal{D}g(f(x_0))h \mid h \in \partial f(x_0)\}.$$



*Proof.* Let  $(\varphi, U)$ ,  $(\psi, V)$  and  $(\nu, W)$  be local charts respectively of  $X$ ,  $Y$  and  $Z$  around  $x_0$ ,  $f(x_0)$  and  $g \circ f(x_0)$ . Define  $\tilde{f} = \psi \circ f \circ \varphi^{-1}: \varphi(U) \rightarrow \psi(V)$  and  $\tilde{g} = \nu \circ g \circ \psi^{-1}: \psi(V) \rightarrow \nu(W)$ . Then, by Proposition and Definition B.7, we obtain

$$\partial(g \circ f)(x_0) = \left\{ \mathcal{D}\nu^{-1}(g \circ f(x_0)) \circ H \circ \mathcal{D}\varphi(x_0) \mid H \in \partial(\tilde{g} \circ \tilde{f})(\varphi(x_0)) \right\}.$$

Now we apply the chain rule from [11, Corollary of Theorem 2.6.6]. Since  $\tilde{g}$  is differentiable at  $\tilde{f}(\varphi(x_0))$  and  $\tilde{f}$  is locally Lipschitz at  $\varphi(x_0)$ , we have then

$$\partial(\tilde{g} \circ \tilde{f})(\varphi(x_0)) = \left\{ \mathcal{D}\tilde{g}(\tilde{f}(\varphi(x_0))) \circ H \mid H \in \partial\tilde{f}(\varphi(x_0)) \right\}.$$

Then,

$$\begin{aligned} \partial(g \circ f)(x_0) &= \left\{ \mathcal{D}\nu^{-1}(g \circ f(x_0)) \circ \mathcal{D}\tilde{g}(\tilde{f}(\varphi(x_0))) \circ H \circ \mathcal{D}\varphi(x_0) \mid H \in \partial\tilde{f}(\varphi(x_0)) \right\} \\ &= \left\{ \mathcal{D}g(f(x_0)) \circ \mathcal{D}\psi^{-1}(f(x_0)) \circ H \circ \mathcal{D}\varphi(x_0) \mid H \in \partial\tilde{f}(\varphi(x_0)) \right\} \\ &= \left\{ \mathcal{D}g(f(x_0)) \circ H \mid H \in \partial f(x_0) \right\}, \end{aligned}$$

the last line being obtained by applying Proposition and Definition B.7 to  $f$ .  $\square$

Lastly, the next two theorems are extensions of the inverse function theorem and of the implicit function theorem to our context.

**Theorem B.12** (Inverse function theorem). *Let  $X$  and  $Y$  be two manifolds of dimension  $n$ . Let  $f: X \rightarrow Y$  be locally Lipschitzian at  $x_0 \in X$ . Suppose that  $\partial f(x_0)$  is of maximal rank, i.e., for all  $h \in \partial f(x_0)$ , we have  $\text{rank } h = n$ . Then, there exist a neighborhood of  $x_0$  in  $X$ , a neighborhood  $V$  of  $f(x_0)$  in  $Y$  and a Lipschitzian function  $g: V \rightarrow U$  such that*

- (i)  $g(f(u)) = u$  for all  $u \in U$  ;
- (ii)  $f(g(v)) = v$  for all  $v \in V$ .

*Proof.* Let  $(\varphi, U)$  be a local chart of  $X$  around  $x_0$  and  $(\psi, V)$  a local chart of  $Y$  around  $f(x_0)$ . Define then  $\tilde{f} = \psi \circ f \circ \varphi^{-1}$ . Since  $\partial f(x_0)$  is of maximal rank, by the chain rule, using Proposition and Definition 2.2, then  $\partial\tilde{f}(\varphi(x_0))$  is of maximal rank. Then, up to taking  $U$  and  $V$  smaller, by the Inverse function theorem applied to  $\tilde{f}$  as stated in [11, Theorem 7.1.1], then there exists a Lipschitz function  $\tilde{g}: \psi(V) \rightarrow \varphi(U)$  such that  $\tilde{g}(\tilde{f}(\tilde{u})) = \tilde{u}$  for  $\tilde{u} \in \varphi(U)$  and  $\tilde{f}(\tilde{g}(\tilde{v})) = \tilde{v}$  for  $\tilde{v} \in \psi(V)$ . Define then  $g = \varphi^{-1} \circ \tilde{g} \circ \psi: U \rightarrow V$  to get

$$g(f(u)) = \varphi^{-1} \circ \tilde{g} \circ \psi(f(u)) = \varphi^{-1} \circ \tilde{g}(\tilde{f}(\varphi(u))) = \varphi^{-1} \circ \varphi(u) = u$$

for all  $u \in U$ , and

$$f(g(v)) = f \circ \varphi^{-1} \circ \tilde{g} \circ \psi(v) = \psi^{-1} \circ \tilde{f}(\tilde{g}(\psi(v))) = \psi^{-1} \circ \psi(v) = v$$

for all  $v \in V$ .  $\square$

**Theorem B.13** (Implicit function theorem). *Let  $X$ ,  $Y$  and  $Z$  be manifolds of dimensions respectively  $m$ ,  $k$  and  $k$ . Let  $f: X \times Y \rightarrow Z$  be locally Lipschitzian at  $(x_0, y_0) \in X \times Y$ . Moreover, assume that the partial generalized differential  $\partial_y f(x_0, y_0)$  is of maximal rank. Then there exists a neighborhood  $U$  of  $x_0$  and a Lipschitz function  $g: U \rightarrow Y$  such that  $g(x_0) = y_0$ , and for all  $x \in U$ ,*

$$f(x, g(x)) = f(x_0, y_0). \tag{B.3}$$

*Proof.* Define  $F(x, y) = (x, f(x, y))$  a function  $X \times Y \rightarrow X \times Z$ , which is locally Lipschitz at  $(x_0, y_0)$ . Define  $n = m + k$  and note that the dimensions of  $X \times Y$  and  $X \times Z$  both equal  $n$ . Besides, since  $\partial_y f(x_0, y_0)$  is of maximal rank, we find that  $\partial F(x_0, y_0)$  is of maximal rank. Thus we can apply the inverse function theorem to  $F$  and find neighborhoods  $U$ ,  $V$ , and  $W$  respectively of  $x_0$  in  $X$ ,  $y_0$  in  $Y$  and  $f(x_0, y_0)$  in  $Z$ , as well as a Lipschitz function  $G: U \times W \rightarrow U \times V$  such that for all  $(x, z) \in U \times W$  we have

$$F(G(x, z)) = (x, z).$$

Note that then  $G(x, z) = (x, \tilde{G}(x, z))$  for some  $\tilde{G}(x, z) \in V$ , so that  $f(x, \tilde{G}(x, z)) = z$ . Therefore, define  $g(x) = \tilde{G}(x, f(x_0, y_0))$  to get

$$f(x, g(x)) = f(x_0, y_0). \tag{B.3}$$

$\square$

## C Additional proofs

**Proposition C.1.** Consider an objective function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  which is scaling-invariant w.r.t.  $x^* = 0$ . Then, the sequence  $\{z_k, \Sigma_k\}_{k \in \mathbb{N}}$  defined by (2.7) is a time-homogeneous Markov chain which follows (2.1), with functions  $F$  and  $\alpha$  defined by (2.8) and (2.9) respectively.

*Proof.* Let  $i = 1, \dots, \lambda$ . Then, we have

$$f\left(z_k + \sqrt{\Sigma_k} U_{k+1}^i\right) = f\left(R(C_k)^{-1/2} \times \left[m_k + \sqrt{C_k} U_{k+1}^i\right]\right).$$

Since  $f$  is scaling-invariant, this implies that the permutation  $s_{k+1}$  satisfies almost surely that

$$f\left(z_k + \sqrt{\Sigma_k} U_{k+1}^{s_{k+1}(1)}\right) \leq \dots \leq f\left(z_k + \sqrt{\Sigma_k} U_{k+1}^{s_{k+1}(\lambda)}\right).$$

Let  $k \geq 1$ , and observe that

$$K_{k+1} := \Sigma_k + \sqrt{\Sigma_k} \sum_{i=1}^{\mu} w_i \left[U_{k+1}^{s_{k+1}(i)}\right] \left[U_{k+1}^{s_{k+1}(i)}\right]^\top \sqrt{\Sigma_k} = R(\Sigma_k)^{-1} \Sigma_{k+1} = R(C_k)^{-1} C_{k+1}.$$

Since  $R = \det^{1/d}(\cdot)$  is (positively) homogeneous  $R(K_{k+1}) = R(C_k)^{-1} R(C_{k+1})$ . Furthermore, we have that

$$\begin{aligned} z_{k+1} &= (R(C_{k+1}))^{-1/2} \times m_{k+1} \\ &= R(C_k)^{1/2} R(C_{k+1})^{-1/2} R(C_k)^{-1/2} \times \left[m_k + \sqrt{C_k} \sum_{i=1}^{\mu} w_i U_{k+1}^{s_{k+1}(i)}\right] \\ &= R(K_{k+1})^{-1/2} \times \left[m_k + \sqrt{\Sigma_k} \sum_{i=1}^{\mu} w_i^m U_{k+1}^{s_{k+1}(i)}\right]. \end{aligned}$$

Moreover,

$$\Sigma_{k+1} = R(C_{k+1})^{-1} C_{k+1} = R(K_{k+1})^{-1} K_{k+1}$$

All in all, we have

$$(z_{k+1}, \Sigma_{k+1}) = F\left((z_k, \Sigma_k), \alpha\left((z_k, \Sigma_k), (U_{k+1}^1, \dots, U_{k+1}^\lambda)\right)\right),$$

ending the proof.  $\square$

*Proof of Proposition 3.1 and Proposition 3.4.* Let  $\theta = (z, \Sigma) \in \mathbb{X}$ , and  $U^1, \dots, U^\lambda \sim \mathcal{N}(0, I_d)$  i.i.d., and  $U = (U^1, \dots, U^\lambda)$ . Then  $W = \alpha(\theta, U)$  satisfies a.s.

$$W = \sum_{\sigma \in \mathfrak{S}_\lambda} \mathbb{1}\left\{f\left(z + \sqrt{\Sigma} U^{\sigma(1)}\right) < \dots < f\left(z + \sqrt{\Sigma} U^{\sigma(\lambda)}\right)\right\} \times \left(U^{\sigma(1)}, \dots, U^{\sigma(\mu)}\right)$$

where  $\mathfrak{S}_\lambda$  is the set of permutations of  $\{1, \dots, \lambda\}$ .

Hence, by symmetry,

$$\begin{aligned} W &= \frac{1}{(\lambda - \mu)!} \sum_{\sigma \in \mathfrak{S}_\lambda} \mathbb{1}\left\{f\left(z + \sqrt{\Sigma} U^{\sigma(1)}\right) < \dots < f\left(z + \sqrt{\Sigma} U^{\sigma(\mu)}\right)\right\} \\ &\quad \times \prod_{k=\mu+1}^{\lambda} \mathbb{1}\left\{f\left(z + \sqrt{\Sigma} U^{\sigma(k)}\right) < f\left(z + \sqrt{\Sigma} U^{\sigma(k)}\right)\right\} \times \left(U^{\sigma(1)}, \dots, U^{\sigma(\mu)}\right). \end{aligned}$$

Let  $\eta: W \rightarrow \mathbb{R}_+$  be a smooth map with compact support. We obtain

$$\begin{aligned} \mathbb{E}[\eta(W)] &= \frac{1}{(\lambda - \mu)!} \sum_{\sigma \in \mathfrak{S}_\lambda} \int \mathbb{1}\left\{f\left(z + \sqrt{\Sigma} u_{\sigma(1)}\right) < \dots < f\left(z + \sqrt{\Sigma} u_{\sigma(\mu)}\right)\right\} \\ &\quad \times \prod_{k=\mu+1}^{\lambda} \mathbb{1}\left\{f\left(z + \sqrt{\Sigma} u_{\sigma(k)}\right) < f\left(z + \sqrt{\Sigma} u_{\sigma(k)}\right)\right\} \end{aligned}$$

$$\times \eta(u_{\sigma(1)}, \dots, u_{\sigma(\mu)}) \gamma^d(u_1) \dots \gamma^d(u_\lambda) du_1 \dots du_\lambda.$$

However, observe that, for each  $k = \mu + 1, \dots, \lambda$ , we have

$$\int \mathbb{1} \left\{ f \left( z + \sqrt{\Sigma} u_{\sigma(\mu)} \right) < f \left( z + \sqrt{\Sigma} u_{\sigma(k)} \right) \right\} \gamma^d(u_{\sigma(k)}) du_{\sigma(k)} = 1 - Q_\theta^f(u_{\sigma(\mu)}).$$

We deduce then the desired result. Note that Proposition 3.4 is obtained by taking  $\Sigma = I_d$ .  $\square$

*Proof of Proposition 4.1.* First observe that (4.3) is equivalent to (iii). Indeed, if (iii) holds, then for every  $y \in \mathsf{X}$  there exists a sequence  $\{y_k\}_{k \in \mathbb{N}}$  with  $y_k \in A_+^k(y)$ , and with a subsequence converging to  $x^*$ . In that case, for every  $T \geq 1$ , and for every neighborhood  $N$  of  $x^*$ , there exist infinitely many indices  $k \geq T$  such that  $y_k \in N$ . Thus, every neighborhood  $N$  of  $x^*$  intersects  $\bigcap_{T \geq 1} \bigcup_{k \geq T} A_+^k(y)$ , which proves that (4.3) holds. Conversely, assume that (4.3) holds. Then, for every  $T \geq 1$ , there exists  $k \geq T$  such that  $x^* \in \overline{A_+^k(y)}$ . Then, consider  $y_k \in A_+^k(y)$  such that  $\text{dist}_X(x^*, y_k) \leq 1/k$ . This proves (iii).

Now, suppose (iii) and let us prove (i). Let  $y \in \mathsf{X}$ . By (iii), we know that there exists a sequence  $\{y_k\}_{k > 0}$  such that  $y_k \in A_+^k(y)$  and with a subsequence which converge to  $x^*$ . However,  $A_+^k(y) \subset A_+(y)$ , therefore  $\{y_k\}_{k > 0}$  is a sequence with values in  $A_+(y)$  admitting  $x^*$  as an accumulation point, which proves (i).

Next, assume that (i) holds, and let us prove that this implies (ii). Let  $y \in \mathsf{X}$ . By (i),  $x^* \in \overline{A_+(y)}$ . In other words, for any open  $U$  of  $\mathsf{X}$  containing  $x^*$ , we have  $U \cap A_+(y) \neq \emptyset$ . Let  $U$  be such an open subset. Since  $A_+(y) = \bigcup_{k \in \mathbb{N}} A_+^k(y)$ , then there exists  $k \in \mathbb{N}$  such that  $A_+^k(y)$  intersects  $U$ . If  $k \geq 1$ , this proves (ii). Else, if  $k = 0$ , we do the same reasoning with  $z = S_y^1(w_1)$  for some  $w_1 \in \mathcal{O}_y^1$ , which proves (ii).

Last, let us prove that (ii) implies (iii). Suppose (ii), let  $y \in \mathsf{X}$ , and let  $\{k_t\}_{t \geq 1}$  be an increasing sequence of  $\mathbb{N}_{>0}$  which satisfies that, for every  $t \geq 1$ , there exists a  $k_t$ -steps path from  $y$  to  $B(x^*, 1/t)$ . Hence, let  $\{y_k\}_{k > 0}$  be a sequence such that  $y_k \in A_+^k(y)$  for every  $k > 0$ , and with  $y_{k_t} \in B(x^*, 1/t)$ . Then, the subsequence  $\{y_{k_t}\}_{t > 0}$  converges to  $x^*$ , proving (iii).  $\square$

*Proof of Proposition 4.2.* Let  $U$  be an open subset of  $\mathsf{X}$ ,  $x \in \mathsf{X}$  and  $k > 0$ . First let us assume that  $P^k(x, U) > 0$ . However, note that we have

$$P^k(x, U) = \int_{\mathcal{O}_x^k} \mathbb{1}\{S_x^k(w_{1:k}) \in U\} p_x^k(w_{1:k}) d\zeta_W(w_1) \dots d\zeta_W(w_k),$$

which implies that there exists  $w_{1:k} \in \mathcal{O}_x^k$  such that  $S_x^k(w_{1:k}) \in U$ , hence  $w_{1:k}$  is a  $k$ -steps path from  $x$  to  $U$ . Conversely, assume that there exists  $w_{1:k}$  a  $k$ -steps path from  $x$  to  $U$ , i.e., that  $S_x^k(w_{1:k}) \in U$ . Since  $F$  is continuous, then  $S_x^k$  is continuous as well. Therefore there exists an open subset  $V$  of  $W^k$  such that for all  $v_{1:k} \in V$ ,  $S_x^k(v_{1:k}) \in U$ . Then we obtain

$$P^k(x, U) \geq \int_{\mathcal{O}_x^k \cap V} \mathbb{1}\{S_x^k(w_{1:k}) \in U\} p_x^k(w_{1:k}) d\zeta_W(w_1) \dots d\zeta_W(w_k) > 0,$$

since  $\mathcal{O}_x^k \cap V$  is open by intersection, and  $p_x^k$  is l.s.c.  $\square$

*Proof of Proposition 4.4.* The statement (i) is a consequence of Proposition 4.1 (ii).

For (ii), let  $y \in \mathsf{X}$ , and for every integer  $s \geq 1$ , consider the open subset  $U_s = B(x^*, 1/s)$  of  $\mathsf{X}$ . Then, there exists a nondecreasing sequence  $\{T_s\}_{s \geq 1}$ , such that for every  $k \geq T_s$ , there exists a  $k$ -steps path  $w_{1:k}^k \in \mathcal{O}_y^k$  from  $y$  to  $U_s$ . For  $k \in \mathbb{N}$ , define  $y_k = S_y^k(w_{1:k}^k)$ , and observe that  $y_k \in A_+^k(y)$ . Moreover, we have  $y_k \in U_s$  for every  $k \in \{T_s, \dots, T_{s+1} - 1\}$ . Then, the sequence  $\{y_k\}_{k \in \mathbb{N}}$  converges to  $x^*$ . Conversely, suppose that for every  $y \in \mathsf{X}$ , there exists a sequence  $\{y_k\}_{k \in \mathbb{N}}$  converging to  $x^*$  with  $y_k \in \overline{A_+^k(y)}$ . Hence, for every  $k > 0$  there exists  $z_k \in A_+^k(y) \cap B(y_k, 1/k)$ . By definition of  $A_+^k(y)$ , there exists then  $w_{1:k}^k \in \mathcal{O}_y^k$  such that  $z_k = S_y^k(w_{1:k}^k)$ . Besides, since  $y_k$  tends to  $x^*$ , then  $z_k$  tends to  $x^*$  as well. Let  $U$  be a neighborhood of  $x^*$ , so that there exists  $T \in \mathbb{N}$  with  $z_k \in U$  when  $k \geq T$ . Then, for  $k \geq T$ ,  $w_{1:k}^k$  is a  $k$ -steps path from  $y$  to  $U$ . Thus  $x^*$  is steadily attracting, proving (ii).

For (iii), suppose that there exist  $x^*$  a steadily attracting state and  $y^*$  a globally attracting state. Let us prove that  $y^*$  is steadily attracting. Let  $U$  be a neighborhood of  $y^*$  in  $\mathsf{X}$ , and let  $z \in \mathsf{X}$ . Since  $y^*$  is globally attracting, there exist  $k > 0$  and a  $k$ -steps path  $w_{1:k} \in \mathcal{O}_{x^*}^k$  from  $x^*$  to  $U$ , i.e. such that  $S_{x^*}^k(w_{1:k}) \in U$ . Since  $F$  is continuous, then  $S_{x^*}^k$  is continuous, and thus there exists a neighborhood  $V$  of  $x^*$  such that for every  $x \in V$ , we have  $S_x^k(w_{1:k}) \in V$ . Moreover,  $x \mapsto p_x^k(w_{1:k})$  is lower semicontinuous, so up to taking  $V$  a smaller neighborhood of  $x^*$ , we can assume that  $w_{1:k} \in \mathcal{O}_x^k$ . Last,  $x^*$  is steadily attracting, so there exists  $T > 0$  such that for every  $t \geq T$ , there exists a  $t$ -steps path  $v_{1:t}$  from  $z$  to  $V$ , i.e.,  $S_z^t(v_{1:t}) \in V$ . All in all, for every  $t \geq T$ , there exists a  $(t+k)$ -steps path  $[v_{1:t}, w_{1:k}]$  from  $z$  to  $U$ , ending the proof.  $\square$

*Proof of Corollary 4.5.* The inclusion  $\{S_x^k(w_{1:k}) \mid w_{1:k} \in \overline{\mathcal{O}_x^k}\} \subset \overline{A_+^k(y)}$  follows directly from the definition of  $A_+^k(y)$  and the continuity of  $F$ . Let  $x^* \in X$  and assume that (i)  $x^*$  is steadily attracting. Then, by definition, for every  $x \in X$  and every neighborhood  $U$  of  $x^*$ , there exists  $T > 0$  such that for every  $k \geq T$  there is a  $k$ -steps path from  $x$  to  $U$ , hence (iii) holds.

Next, assume that (iii) for every  $x \in X$  and every neighborhood  $U$  of  $x^*$ , there exists  $T > 0$  such that for every  $k \geq T$  we can find  $w_{1:k} \in \overline{\mathcal{O}_x^k}$  with  $S_x^k(w_{1:k}) \in U$ . Then, as in the previous proof, for every integer  $s \geq 1$ , consider the open subset  $U_s = B(x^*, 1/s)$  of  $X$ . Therefore there exists a nondecreasing sequence  $\{T_s\}_{s \geq 1}$ , such that for every  $k \geq T_s$ , there exists  $w_{1:k} \in \overline{\mathcal{O}_x^k}$  with  $S_x^k(w_{1:k}) \in U_s$ . So we find a sequence  $\{y_k\}_{k \in \mathbb{N}}$  such that  $y_k \in \{S_x^k(w_{1:k}) \mid w_{1:k} \in \overline{\mathcal{O}_x^k}\}$  for  $k \in \mathbb{N}$ , and with  $y_k \in U_s$  for  $k \in \{T_s, \dots, T_{s+1} - 1\}$ , which proves (ii).

Last, observe that the implication ‘(ii) implies (i)’ follows directly from Proposition 4.4(ii) and the inclusion  $\{S_x^k(w_{1:k}) \mid w_{1:k} \in \overline{\mathcal{O}_x^k}\} \subset \overline{A_+^k(y)}$ .  $\square$

*Proof of Proposition 4.6.* First, we prove (i). Observe that  $E$  is nonempty. Indeed,  $x^*$  is attainable, so there exist  $a \in \mathbb{N}^*$  and  $w_{1:a} \in \mathcal{O}_{x^*}^a$  with  $x^* = S_{x^*}^a(w_{1:a})$ , and so  $x^* = S_{x^*}^{ak}(w_{1:a}, \dots, w_{1:a})$ . Consider now  $a$  and  $b$  two elements of  $E$  and let us prove that  $d := \gcd(a, b) \in E$ . By definition, there exist  $T_a, T_b > 0$ , such that for every  $k \geq T_a$ ,  $x^* \in A_+^{ak}(x^*)$ , and for every  $k \geq T_b$ ,  $x^* \in A_+^{bk}(x^*)$ . Let  $T \in \mathbb{N}$  be larger than  $a/d$ , so that for every  $k \geq 0$ , the Euclidean division of  $(T+k)d$  by  $a$  provides us  $q_a$  and  $r$  two integers such that  $(T+k)d = q_a a + r$ . Besides, by Bézout’s theorem, we find that  $r = q_b b$  for some  $q_b \in \mathbb{N}$ , hence  $(T+k)d = q_a a + q_b b$ . However, by definition of  $T_a$  and  $T_b$ , we have  $x^* \in A_+^{(q_a+T_a)a}(x^*)$  and  $x^* \in A_+^{(q_b+T_b)b}(x^*)$ . All in all,  $x^* \in A_+^{(T+k)d+T_a a+T_b b}$ , proving (i) since  $d$  divides  $a$  and  $b$ .

To prove (ii), observe that if  $\gcd(E) = 1$ , then, by (i), we have  $1 \in E$ . Then, there exists  $T \in \mathbb{N}$  such that for all  $k \geq T$ ,  $x^* \in A^k(x^*)$ . Let  $y \in X$ . Since  $x^*$  is attainable, there exists  $t \in \mathbb{N}$  such that  $x^* \in A^t(y)$ , so that for all  $k \geq T+t$ ,  $x^* \in A^k(y)$ . Thus,  $x^*$  is steadily attracting.

For (iii), define  $d = \gcd(E)$ , and let  $D_i = \cup_{r \geq 0} A_+^{r(d+i)}(x^*)$  for  $i \in \{0, \dots, d-1\}$ . First observe that the  $D_i$  are disjoint sets. Indeed,  $A_+^i(x^*)$  intersects  $A_+^j(x^*)$  for some integers  $i, j$ , then there exists  $y$  in their intersection. As  $x^*$  is attainable, there exists  $k > 0$  such that  $x^* \in A_+^k(y)$ , hence  $x^* \in A_+^{r(k+i)}(x^*)$  and  $x^* \in A_+^{r(k+j)}(x^*)$  for all  $r \geq 0$ . This implies that  $d$  divides both  $k+i$  and  $k+j$  hence  $d$  divides  $i-j$ . This shows that the sets  $D_i$ ,  $i \in \{0, \dots, d-1\}$ , are disjoint. Moreover, by construction, we have  $P(y, D^{i+1}) = 1$  for all  $i \in \mathbb{Z}/d\mathbb{Z}$ . Finally, observe that the union of the  $D_i$ ,  $i \in \{0, \dots, d-1\}$ , is equal to  $A_+(x^*)$ . Since  $P$  is  $\varphi$ -irreducible, and  $P^k(x^*, A_+(x^*)) = 1$  for all  $k \in \mathbb{N}$ , then the support of  $\varphi$  is included in  $A_+(x^*)$ . All in all, we have that  $\{D_i\}_{0 \leq i \leq d-1}$  is a  $d$ -cycle.  $\square$

*Proof of Proposition 4.14.* First we prove (4.18). Let  $x^* \in \text{supp } \psi$ , and let  $U$  be a neighborhood of  $x^*$ . Then,  $\psi(U) > 0$ , which implies that for every  $y \in X$ ,  $\sum_{k \geq 0} P^k(y, U) > 0$ . This is true for every neighborhood  $U$  of  $x^*$ , hence, by Proposition 4.2 and Proposition 4.1,  $x^*$  is globally attracting.

Conversely, let  $x^* \in X$  be a globally attracting state. Then, by Proposition 4.2 and Proposition 4.1, for every neighborhood  $U$  of  $x^*$  and for every  $y \in X$ , there exists  $k > 0$  such that  $P^k(y, U) > 0$ , hence  $\psi(U) > 0$ . This implies that  $x^* \in \text{supp } \psi$ . All in all, we obtain (4.18).

Now, let us prove (4.19). Consider  $x^* \in X$  a globally attracting state, and let  $y^* \in \text{supp } \psi$ . By (4.18),  $y^*$  is then a globally attracting state. Therefore, by Proposition 4.1,  $y^* \in \overline{A_+(x^*)}$ .

Conversely, let  $y^* \in \overline{A_+(x^*)}$  and let us prove that  $y^*$  is globally attracting. Let  $U$  be a neighborhood of  $y^*$ , so that  $U$  intersects  $A_+(x^*)$ . This implies that there exist  $k \geq 1$  and  $w_{1:k} \in \mathcal{O}_{x^*}^k$  such that  $S_{x^*}^k(w_{1:k}) \in U$ . Since  $F$  is continuous, then  $S_x^k(w_{1:k}) \in U$  for every  $x$  in a neighborhood  $V$  of  $x^*$ . Besides,  $x \mapsto p_x^k(w_{1:k})$  is l.s.c., so, up to taking  $V$  smaller, we can assume that  $w_{1:k} \in \mathcal{O}_x^k$  for every  $x \in V$ . Furthermore,  $x^*$  is globally attracting, so, for every  $z \in X$ , there exist  $t \in \mathbb{N}$  and  $v_{1:t} \in \mathcal{O}_z^t$  such that  $S_z^t(v_{1:t}) \in V$ , hence such that  $[v_{1:t}, w_{1:k}] \in \mathcal{O}_z^{t+k}$  and  $S_z^{t+k}(v_{1:t}, w_{1:k})$ , proving that  $y^*$  is globally attracting. This ends the proof, using (4.18).  $\square$