DETERMINISTIC ACTOR–CRITIC WITHOUT AN ACTION-GRADIENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Deterministic policy gradient methods such as DDPG and TD3 improve the actor by ascending the action-gradient of a learned critic, and their analysis assumes that the action value is differentiable in the action. That assumption fails on agents trained on commonly used continuous control benchmarks, and whether it fails is decided by the trained agent and not by the task alone. Differentiability in the action is governed by a competition between the discount factor and the rate at which the closed loop separates nearby states. We give the threshold below which the action-gradient exists, together with the stricter one below which the action-curvature does, and show that a loop that stretches fast enough on a hyperbolic attractor leaves no gradient at almost every state of its basin. The actor ascends the gradient of an action value smoothed by the noise TD3 adds to its targets and by the resolution of the critic. The width of the smoothing sets the size of that gradient: it is at most an inverse power of the width, and in the mean square near the attractor it grows as an inverse power when the noise is reduced. Averaged over the states the actor visits, this slope is the gradient of the return, up to a positive factor and the square of the noise, which is why the method still works beyond the threshold. On agents trained on standard benchmarks we measure the separation rate and the exponent it predicts, test the prediction against direct measurements of the action value, and give a diagnostic that detects, from the simulator alone, when a trained agent lies beyond the threshold.

1 INTRODUCTION

Deterministic actor–critic methods are among the most widely used algorithms for continuous control. A deterministic actor $\mu_{\boldsymbol{\theta}}$ maps each state $\mathbf{s}$ to an action, a critic $Q_{\boldsymbol{\phi}}(\mathbf{s}, \mathbf{a})$ is fitted to the discounted return that follows action $\mathbf{a}$ in state $\mathbf{s}$, and the actor is improved by moving its output in the direction in which the critic increases,

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta \, \nabla_{\boldsymbol{\theta}} \mu_{\boldsymbol{\theta}}(\mathbf{s})^{\top} \, \nabla_{\mathbf{a}} Q_{\boldsymbol{\phi}}(\mathbf{s}, \mathbf{a})\big|_{\mathbf{a}=\mu_{\boldsymbol{\theta}}(\mathbf{s})}, \tag{1}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are the parameters of the actor and the critic, and η is a step size. DDPG [1] and TD3 [2] are the standard instances. The update in Eq. 1 is justified by the deterministic policy gradient theorem [3], with $\nabla_{\mathbf{a}} Q_{\boldsymbol{\phi}}$ standing in for the action-gradient of the true action value $Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})$, the return obtained by playing $\mathbf{a}$ in $\mathbf{s}$ and following $\mu_{\boldsymbol{\theta}}$ thereafter. The theorem and the convergence analyses built on it [4] require $Q^{\mu_{\boldsymbol{\theta}}}$ to be differentiable in the action. This is usually taken for granted: the reward is smooth, the simulator is smooth, and so, it is assumed, is the discounted sum of rewards along the trajectory that the action sets in motion.

We show that this assumption fails on agents trained in the environments on which these methods are routinely evaluated. Take a trained actor and two copies of the simulator in the same state. In the first copy the actor plays its own action; in the second it plays the same action displaced by a small amount h; from the next step on, the actor controls both copies. The displacement moves the next state, and from then on the closed loop—the environment under the control of $\mu_{\boldsymbol{\theta}}$—acts on the difference between the two copies. Over the steps that the discount still counts, the difference either stays as small as it began or grows by a factor at every step. In the first case each reward changes in proportion to that difference, since the reward is a smooth function of the state, and the return changes in proportion to h: the action value has a slope. In the second the difference becomes as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

large as the state itself, the two trajectories have nothing in common, and the rewards they collect differ by an amount unrelated to h. When that happens, making h smaller only postpones the moment of separation, by a fixed number of steps for each halving of h, while the discount removes a fixed fraction of the return for each of those steps. When the separation outruns the discount, the change in return no longer shrinks in proportion to h, and the action value can lose its slope altogether (Section 3).

Which of the two happens depends on two rates, the rate at which the closed loop separates nearby states and the rate at which the discount forgets the future, and there is a threshold between them: below it the effect of the displacement summed over the future is finite and the action-gradient exists; beyond it the change in return need shrink only as a power of the displacement, fixed by the two rates. Which side a trained agent falls on belongs to the closed loop that training has produced, and not to the task alone, and it can be read off two copies of the simulator without ever evaluating the action value (Section 3.3).

Parts of this picture are known. How regular a value function can be, given the discount and how fast the dynamics separate nearby states, has been studied for Lipschitz decision processes [5, 6, 7]. In chaotic systems the objective of policy optimisation is known to become a fractal function of the policy parameters [8]. The same competition between discount and expansion makes gradients taken through a differentiable simulator explode over long horizons [9, 10, 11]. Noise in the policy optimises a smoothed objective [12], and TD3 smooths in the same way by perturbing the action at which it reads its targets [2]. Where the gradient is lost, zeroth-order estimates [13] and learning the deterministic policy through a stochastic one [14] do without it.

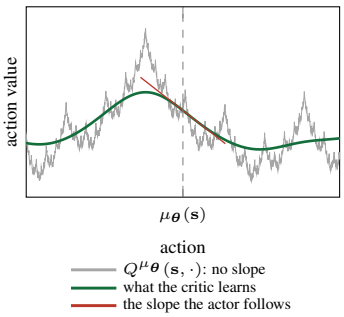

Figure 1: An action value with no slope, the smoothed value the critic learns, and the slope it hands the actor.

We ask what the standard algorithms meet and what they compute. Neither DDPG nor TD3 ever fits its critic to the action value of the deployed actor: the critic is a network of finite flexibility, trained only at the actions the exploration noise places in the buffer, and TD3 reads its targets at actions displaced by a noise of its own. What such a critic represents is that action value smoothed over those widths, and a smoothed value has a slope where the rough one has none (Figure 1). The narrower the smoothing, the larger that slope, and its average over the states the actor visits points along the gradient of the return. The three devices that supply it are the ones usually taken for implementation detail, the width of the exploration, the noise TD3 adds to its targets and the size of the network. Where the separation outruns the discount, the slope they supply is the one the actor climbs (Section 4).

We give the threshold below which the action-gradient exists and the stricter one below which the action-curvature does, show that a loop stretching fast enough on a hyperbolic attractor leaves no gradient at almost every state of its basin, and that neither the thresholds nor the exponent can be improved (Section 3). We characterise the gradient the critic supplies in its place, and the noise and the resolution that set its size (Section 4). On 42 agents trained with TD3 on 7 tasks we read from two copies of the simulator which side of the threshold each agent lies on, cross the threshold by varying the discount at fixed dynamics, test the predicted exponent against direct measurements of the action value, and measure the resolution of the critic and the slope the smoothed action value hands the actor (Section 5).

2 Setting

We consider a Markov decision process with states $\mathbf{s} \in \mathcal{S} \subseteq \mathbb{R}^{d_S}$, actions $\mathbf{a}$ in a box $\mathcal{A} \subset \mathbb{R}^{d_A}$, a reward $r(\mathbf{s}, \mathbf{a})$ with $|r| \leq r_{\max}$, and a discount factor $\gamma \in (0, 1)$, with deterministic dynamics $\mathbf{s}' = F(\mathbf{s}, \mathbf{a})$. This is the setting of the physics simulators on which deterministic actor–critic methods are evaluated, and in it whatever roughness the action value has comes from the dynamics and the actor alone. A deterministic actor $\mu_{\boldsymbol{\theta}}$ maps states to actions, and when it acts without noise the state moves through the map $f(\mathbf{s}) = F(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}))$, the closed loop. We assume that F, r

and $\mu_{\boldsymbol{\theta}}$ are continuously differentiable with bounded derivatives, to the order each theorem names (Assumption 1), the regularity assumed for Lipschitz decision processes [5, 15, 7].

Since the dynamics are deterministic, $\mathbf{s}$ and $\mathbf{a}$ determine the whole trajectory that follows, and the action value of the actor is a sum along it,

$$Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s},\mathbf{a}) = r(\mathbf{s},\mathbf{a}) + \sum_{t\geq 1} \gamma^t \, r\big(\mathbf{s}_t, \mu_{\boldsymbol{\theta}}(\mathbf{s}_t)\big), \qquad \mathbf{s}_1 = F(\mathbf{s},\mathbf{a}), \quad \mathbf{s}_{t+1} = f(\mathbf{s}_t). \tag{2}$$

The action enters this sum through the reward at $(\mathbf{s},\mathbf{a})$ and through the next state $\mathbf{s}_1$, from which the closed loop takes over and fixes every later state, so the tail depends on $\mathbf{a}$ only through $\mathbf{s}_1$. With $V^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}) = Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s},\mu_{\boldsymbol{\theta}}(\mathbf{s}))$ the value of the actor's own action, Eq. 2 says that $Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s},\mathbf{a}) = r(\mathbf{s},\mathbf{a}) + \gamma V^{\mu_{\boldsymbol{\theta}}}(F(\mathbf{s},\mathbf{a}))$, and the action-gradient, wherever it exists, is

$$\nabla_{\mathbf{a}} Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s},\mathbf{a}) = \nabla_{\mathbf{a}} r(\mathbf{s},\mathbf{a}) + \gamma \, \partial_{\mathbf{a}} F(\mathbf{s},\mathbf{a})^{\top} \, \nabla_{\mathbf{s}} V^{\mu_{\boldsymbol{\theta}}}\big(F(\mathbf{s},\mathbf{a})\big). \tag{3}$$

The first term exists under Assumption 1, and the second is there as soon as $V^{\mu_{\boldsymbol{\theta}}}$ has a gradient in the state at $\mathbf{s}_1$, so the action-gradient at $(\mathbf{s},\mathbf{a})$ rests on the regularity of the value one step ahead.

Training maximises the return from a random initial state $\mathbf{s}_0$ drawn from the reset distribution p_0,

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}_0 \sim p_0}\big[V^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}_0)\big]. \tag{4}$$

Differentiating $V^{\mu_{\boldsymbol{\theta}}}$ along the trajectory, with Eq. 3 at every step, and collecting the terms gives the formula of the deterministic policy gradient theorem [3],

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{1-\gamma} \, \mathbb{E}_{\mathbf{s}\sim d^{\mu_{\boldsymbol{\theta}}}}\Big[\nabla_{\boldsymbol{\theta}} \mu_{\boldsymbol{\theta}}(\mathbf{s})^{\top} \, \nabla_{\mathbf{a}} Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s},\mathbf{a})\big|_{\mathbf{a}=\mu_{\boldsymbol{\theta}}(\mathbf{s})}\Big], \tag{5}$$

where $d^{\mu_{\boldsymbol{\theta}}} = (1-\gamma)\sum_{t\geq 0}\gamma^t p_t$ and p_t is the distribution of the state at step t when $\mu_{\boldsymbol{\theta}}$ is followed from $\mathbf{s}_0 \sim p_0$: the states the actor visits, weighted by the discount. The derivation is written out in Appendix A, and every one of its steps differentiates the value at the next state, so it asks that the action-gradient exist and stay bounded along the trajectories, an assumption the convergence analyses of DDPG and TD3 make [4].

Training. DDPG [1] and TD3 [2] learn $\mu_{\boldsymbol{\theta}}$ and $Q_{\boldsymbol{\phi}}$ together, following the loop of Figure 7 of Appendix F, and the actor moves along $\nabla_{\mathbf{a}} Q_{\boldsymbol{\phi}}$, with $Q_{\boldsymbol{\phi}}$ trained by regression on bootstrapped targets. Noise enters that loop when the agent acts, playing $\mathbf{a} = \mu_{\boldsymbol{\theta}}(\mathbf{s}) + \sigma\boldsymbol{\xi}$, with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and the sum clipped to the action box, and the critic is trained only at these actions. Noise enters again when the targets are formed, and TD3 reads the critic at a perturbed action,

$$y = r + \gamma \, Q_{\bar{\boldsymbol{\phi}}}\big(\mathbf{s}', \mu_{\bar{\boldsymbol{\theta}}}(\mathbf{s}') + \tilde{\sigma}\boldsymbol{\zeta}\big), \qquad \boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \tag{6}$$

where $\bar{\boldsymbol{\phi}}$ and $\bar{\boldsymbol{\theta}}$ are slowly updated copies of the critic and the actor. DDPG is the case $\tilde{\sigma} = 0$, and TD3 also clips $\boldsymbol{\zeta}$ and takes the minimum of two critics. At the end of training the actor is deployed without noise, and the two widths σ and $\tilde{\sigma}$ each smooth the value that the critic learns; the return of an actor that plays with noise is studied by Montenegro et al. [14] for stochastic policy gradients.

Assumption 1. *F, r and $\mu_{\boldsymbol{\theta}}$ are k times continuously differentiable with bounded derivatives, in the state, the action and the parameters, with $k \geq 1$ the order each theorem names.*

3 When the action-gradient exists

By Eq. 2 the value at $\mathbf{s}_1$ is a sum along the trajectory $\mathbf{s}_1$, $\mathbf{s}_2 = f(\mathbf{s}_1)$, $\mathbf{s}_3 = f(\mathbf{s}_2), \ldots$ of the closed loop. Moving $\mathbf{s}_1$ by a small amount moves every later state, and $\nabla_{\mathbf{s}} V^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}_1)$ adds up the effect of the move on each reward, weighted by the discount. Whether the sum of those effects is finite is settled by a competition between the discount, which shrinks the later terms, and the closed loop, which can enlarge the move it passes from one state to the next.

3.1 The displacement and how it grows

Suppose the action is displaced by h along a unit direction $\mathbf{e} \in \mathbb{R}^{d_A}$. Two Jacobians describe what happens next: $\mathbf{B}(\mathbf{s},\mathbf{a}) = \partial_{\mathbf{a}} F(\mathbf{s},\mathbf{a})$ says how a small change of the action moves the next state, and

the Jacobian $\mathbf{A}(\mathbf{s}) = Df(\mathbf{s})$ of the closed loop says how a change of the state at one step moves the state at the next. After t steps the state $\mathbf{s}_t$ has moved by approximately $h\,\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e})$, with

$$\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e}) = \frac{\partial \mathbf{s}_t}{\partial \mathbf{a}}\,\mathbf{e} = \mathbf{A}(\mathbf{s}_{t-1}) \cdots \mathbf{A}(\mathbf{s}_1)\,\mathbf{B}(\mathbf{s}, \mathbf{a})\,\mathbf{e} \;\in \mathbb{R}^{d_S}. \tag{7}$$

The growth of $\boldsymbol{\delta}_t$ with t is set by the Jacobians of the closed loop, and we summarise it by a single number: Λ is a separation rate of the closed loop if $\|\mathbf{A}(\mathbf{s}_n) \cdots \mathbf{A}(\mathbf{s}_1)\| \le C\Lambda^n$ along every trajectory, for some constant C. Since $\mathbf{B}$ is bounded (Assumption 1), Eq. 7 gives, with C enlarged to carry that bound, for every state, action, direction and step,

$$\|\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e})\| \le C\,\Lambda^t. \tag{8}$$

A closed loop that reabsorbs disturbances admits a $\Lambda < 1$, and one in which nearby states drift apart admits only a $\Lambda > 1$; Λ belongs to the closed loop, and not to the task alone.

Differentiating Eq. 2 term by term and applying the chain rule to each $\mathbf{s}_t$ gives the action-gradient as a sum over future steps,

$$\nabla_{\mathbf{a}} Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a}) = \nabla_{\mathbf{a}} r(\mathbf{s}, \mathbf{a}) + \sum_{t \ge 1} \gamma^t \Big(\frac{\partial \mathbf{s}_t}{\partial \mathbf{a}}\Big)^{\top} \nabla \bar{r}(\mathbf{s}_t), \qquad \bar{r}(\mathbf{s}) = r\big(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s})\big), \tag{9}$$

where $\bar{r}$ is the reward collected along the closed loop and $\nabla\bar{r}$ its gradient in the state, the dependence of the action on the state included, with $\|\nabla\bar{r}\| \le G$ because the reward and the actor have bounded derivatives (Assumption 1). Along $\mathbf{e}$ the term of step t is $\gamma^t \boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e})^{\top} \nabla\bar{r}(\mathbf{s}_t)$, of size at most $CG(\gamma\Lambda)^t$ by Eq. 8. When $\gamma\Lambda < 1$ these bounds are a geometric sequence whose sum is finite and independent of the state and the action, so the series converges uniformly to a sum continuous in $\mathbf{a}$ and Eq. 9 holds: the action-gradient exists and stays bounded along the trajectories, as Eq. 5 asks. The j-th derivative of a term contains the displacement j times, which replaces $\gamma\Lambda$ by $\gamma\Lambda^j$.

Theorem 1 (Derivatives in the action)**.** *Under Assumption 1, if $j \le k$ and $\gamma\Lambda^j < 1$, then for every state $\mathbf{s}$ the action value $Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \cdot)$ is j times continuously differentiable, and its j-th derivative is bounded uniformly in the state and the action (proof in Appendix B.1).*

The condition $\gamma\Lambda^j < 1$ reads $\Lambda < \gamma^{-1/j}$, so that at $\gamma = 0.99$ the action value keeps a gradient in a closed loop that enlarges displacements by one percent per step and a curvature in one that enlarges them by half a percent, and past one percent neither is guaranteed.

3.2 Beyond the threshold

When $\gamma\Lambda > 1$ the bounds $CG(\gamma\Lambda)^t$ on the terms of Eq. 9 grow with t and their sum diverges, and what can be bounded is the change in $Q^{\mu_{\boldsymbol{\theta}}}$ produced by a finite move $h > 0$ along $\mathbf{e}$. Writing $\mathbf{s}'_t$ for the states of the trajectory the moved action $\mathbf{a} + h\mathbf{e}$ starts, Eq. 2 gives

$$Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a} + h\mathbf{e}) - Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a}) = \big[r(\mathbf{s}, \mathbf{a} + h\mathbf{e}) - r(\mathbf{s}, \mathbf{a})\big] + \sum_{t \ge 1} \gamma^t \Big[r\big(\mathbf{s}'_t, \mu_{\boldsymbol{\theta}}(\mathbf{s}'_t)\big) - r\big(\mathbf{s}_t, \mu_{\boldsymbol{\theta}}(\mathbf{s}_t)\big)\Big]. \tag{10}$$

The first bracket is bounded by $\|\nabla_{\mathbf{a}} r\|\,h$ (Assumption 1), and each bracket in the sum has two bounds, since while the two trajectories are close their states differ by at most $C\,h\,\Lambda^t$ and their rewards by at most G times that, whereas at any step the reward is bounded by $r_{\max}$, so that

$$\Big|r\big(\mathbf{s}'_t, \mu_{\boldsymbol{\theta}}(\mathbf{s}'_t)\big) - r\big(\mathbf{s}_t, \mu_{\boldsymbol{\theta}}(\mathbf{s}_t)\big)\Big| \le \min\big\{\, C\,G\,h\,\Lambda^t,\; 2r_{\max} \,\big\}. \tag{11}$$

The first bound is smaller for early steps and the second for late ones, and they are equal at $t = \log(2r_{\max}/CG)/\log\Lambda + \log(1/h)/\log\Lambda$, whose first term does not depend on h and multiplies the bound by a constant (Appendix B.2 keeps it), leaving

$$t^* = \frac{\log(1/h)}{\log\Lambda}. \tag{12}$$

We call t^* the separation time, the step at which the two trajectories stop being close. Splitting the sum there leaves early terms $C\,G\,h\,(\gamma\Lambda)^t$ that grow geometrically, with a sum dominated by the

last one, $C\,G\,\gamma^{t^*}$ because $h\Lambda^{t^*} = 1$, and late terms $2r_{\max}\gamma^t$ that shrink geometrically, with a sum dominated by the first one, $2r_{\max}\gamma^{t^*}$, so that both parts are bounded by a multiple of γ^{t^*}, and

$$\gamma^{t^*} = \exp\Big(-\log\frac{1}{h}\cdot\frac{\log(1/\gamma)}{\log\Lambda}\Big) = h^{\alpha}, \qquad \alpha = \frac{\log(1/\gamma)}{\log\Lambda}, \tag{13}$$

an exponent below one exactly when $\gamma\Lambda > 1$.

Theorem 2 (Hölder exponent)**.** *Under Assumption 1 with $k \geq 1$, if $\gamma\Lambda > 1$ then for every state, every action, every unit direction $\mathbf{e}$ and every h, for a constant c that does not depend on them,*

$$\big|Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a} + h\mathbf{e}) - Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})\big| \leq c\,|h|^{\alpha}, \qquad \alpha = \frac{\log(1/\gamma)}{\log\Lambda} < 1 \tag{14}$$

(proof in Appendix B.2).

A function that satisfies Eq. 14 is called Hölder continuous with exponent α, and for $\alpha < 1$ the bound allows the difference quotient to grow like $|h|^{\alpha-1}$ as h shrinks but also allows it to stay bounded, so the bound does not decide whether the gradient is lost. Bounds of this kind are known [5, 7, 8]; all of them bound the change of the value from above.

3.3 Reading the threshold off two copies of the simulator

The action value $Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})$ of Eq. 2 sums rewards along the single trajectory that $(\mathbf{s}, \mathbf{a})$ starts, and a change of $\mathbf{a}$ moves that trajectory alone, so Theorems 1 and 2 hold at $\mathbf{s}$ with Λ replaced by a separation rate $\Lambda(\mathbf{s})$ of the trajectories that start there and exponent $\alpha(\mathbf{s}) = \log(1/\gamma)/\log\Lambda(\mathbf{s})$. The same closed loop can be below the threshold at one state and beyond it at another.

Which side a state is on can be read off two copies of the simulator, without computing a derivative. Both are started in $\mathbf{s}$, the second plays a first action displaced by $h\mathbf{e}$, and the actor controls both from then on. Since the reward and the actor have bounded derivatives (Assumption 1), the rewards they collect at step t differ by at most $G\,d_t(h)$, with $d_t(h)$ the distance between their states at step t, so the distances bound the change of the action value,

$$\big|Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}) + h\mathbf{e}) - Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}))\big| \leq \|\nabla_{\mathbf{a}} r\|\,h + G\,S(h), \qquad S(h) = \sum_{t\geq 1}\gamma^t\,d_t(h). \tag{15}$$

Running the copies along a ladder of displacements $h_1 < \cdots < h_n$ gives S as a function of h, and by Eq. 15 the action value is at least as regular as S is, so that the slope of $\log S(h)$ against $\log h$ is the exponent that $Q^{\mu_{\boldsymbol{\theta}}}$ inherits. While the two trajectories stay close, their distance is $\|\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e})\|\,h + o(h)$ by Eq. 7, so that the terms of $S(h)$ before the separation time $t^*(h)$ of Eq. 12 are proportional to h; at that time their distance stops following the displacement, the sum is carried by that step, and $S(h)$ scales as $\gamma^{t^*(h)} = h^{\alpha(\mathbf{s})}$.

The same exponent can be read off two separation times, since halving h postpones the separation by the time in which the loop doubles a displacement, while the discount halves the weight of a reward every $\log 2/\log(1/\gamma)$ steps, which is 69 at $\gamma = 0.99$, and the exponent is the delay divided by that number,

$$D(\mathbf{s}) = t^*(h/2) - t^*(h) = \frac{\log 2}{\log\Lambda(\mathbf{s})}, \qquad \alpha(\mathbf{s}) \simeq \frac{D(\mathbf{s})}{69} \ \text{ at } \gamma = 0.99. \tag{16}$$

Appendix B.6 gives the rule that reads t^* off the distances, and what it returns when the copies never part.

3.4 On a hyperbolic attractor the gradient is absent

Whether the gradient survives beyond the threshold depends on where the trajectories of the loop go and on how it stretches displacements there. A loop that carries a bounded region of states into itself draws the trajectories that start there onto the set the region shrinks to, the attractor of the loop. The attractor is hyperbolic when the state space splits at each point into directions the loop contracts and directions it stretches [16]. One of those directions is stretched the most, at a rate that, averaged along a trajectory, is the same at all starting states outside a set of zero volume [17], and we write λ_1 for that rate, with $\alpha_1 = \log(1/\gamma)/\lambda_1$, so that λ_1 plays for the fastest direction the role that $\log\Lambda$ plays in Theorem 2.

Assumption 2. *The loop carries a bounded region of states into itself and the attractor inside it is hyperbolic. The fastest of its stretched directions is a line along which the loop stretches by at least $\Lambda_{\min} > 1$ at every point, the action can move the next state along that line, and the value is not Lipschitz along it (full statement in Appendix B.4).*

The exponent of the action value at a state $\mathbf{x}$ is read by comparing its change there with a power of the displacement, the ratio of the two,

$$R_\beta(h) = \frac{\left|Q^{\mu_{\boldsymbol{\theta}}}\left(\mathbf{x}, \mu_{\boldsymbol{\theta}}(\mathbf{x}) + h\mathbf{e}\right) - Q^{\mu_{\boldsymbol{\theta}}}\left(\mathbf{x}, \mu_{\boldsymbol{\theta}}(\mathbf{x})\right)\right|}{|h|^\beta}, \tag{17}$$

which vanishes as $h \to 0$ when the action value changes more slowly than $|h|^\beta$ and is unbounded when it changes faster, so that the power separating the two behaviours is the exponent at $\mathbf{x}$.

Theorem 3 (No gradient at almost every state)**.** *Let Assumption 1 with $k \geq 1$ and Assumption 2 hold, let $\gamma\Lambda_{\min} > 1$, and let $\mathbf{e}$ be a direction of the action that moves the next state along the fastest direction. Then, outside a set of zero volume, at every state $\mathbf{x}$ whose trajectory enters the region, the ratio of Eq. 17 tends to zero as $h \to 0$ for every $\beta < \alpha_1$, and $\limsup_{h\to 0} R_\beta(h) = +\infty$ for every $\beta > \alpha_1$ (proof in Appendix B.5).*

The exponent at those states is therefore α_1 exactly, and at $\beta = 1$, where the ratio is the difference quotient at the action the actor plays, that quotient is unbounded, so that the derivative is not there. Since λ_1 is at most $\log \Lambda$, α_1 is at least the α of Theorem 2, with equality when the loop stretches the fastest direction at exactly the rate Λ. Neither the thresholds of Theorem 1 nor the exponent of Theorem 2 can be improved, since Appendix B.3 gives a system with one state variable whose action value at one state is a Weierstrass function plus a smooth term, and all of them are attained there [18].

4 What the actor follows

Where the loop stretches fast enough on a hyperbolic attractor, $Q^{\mu_{\boldsymbol{\theta}}}$ has no gradient in the action at almost every state of its basin, but the actor never sees $Q^{\mu_{\boldsymbol{\theta}}}$: it moves along $\nabla_{\mathbf{a}} Q_{\boldsymbol{\phi}}$, where $Q_{\boldsymbol{\phi}}$ is a network of finite flexibility, trained only at the actions the agent executed, which the exploration noise scatters by σ around $\mu_{\boldsymbol{\theta}}(\mathbf{s})$, on targets that TD3 reads at actions perturbed by $\tilde{\sigma}\boldsymbol{\zeta}$. Each of these three facts smooths the action value that the critic represents, so that what the actor follows is the slope of a smoothed action value, at a width that the noise of the targets and the resolution of the network set together.

Assumption 3. *The target noise $\boldsymbol{\zeta}$ has zero mean, identity covariance and a density with bounded derivatives on a bounded support. A few actions between them move the next state in every direction of the state space, uniformly in the state and in the noise those actions carry. The loops whose actions carry a noise of that size, and the loops of the actors near $\boldsymbol{\theta}$, keep the region of states of Assumption 2, are hyperbolic there as f is, and forget their initial state on that region at one rate common to all of them. Training starts from states whose density is smooth and supported in that region (full statement in Appendices C.1 and C.2).*

4.1 The critic represents a smoothed action value

A critic with finitely many parameters cannot follow structure finer than a scale of its own, which for a network trained by gradient descent comes from the bias towards the coarse features of its target [19], documented for value approximation in continuous control by Yang et al. [20]. That scale is its resolution, one value in each variable, ℓ in the action and ℓ_s in the state. Fitted to a function that is Hölder with exponent α, a critic that resolves the state at ℓ_s returns a slope of at most a constant times $\ell_s^{\alpha-1}$ in the mean over the pairs it is trained on, as if the function had been smoothed over ℓ_s, which Appendix C.3 proves for a class linear in its parameters at the projected fixed point of temporal-difference regression [21].

A second smoothing comes from the targets of Eq. 6. Write $\pi_{\tilde{\sigma}}$ for the actor that plays $\mu_{\boldsymbol{\theta}}(\mathbf{s}) + \tilde{\sigma}\boldsymbol{\zeta}$ at every step, with a fresh noise each time, and $V^{\pi_{\tilde{\sigma}}}$ and $Q^{\pi_{\tilde{\sigma}}}$ for its value and action value,

$$V^{\pi_{\tilde{\sigma}}}(\mathbf{s}) = \mathbb{E}_{\boldsymbol{\zeta}}\Big[Q^{\pi_{\tilde{\sigma}}}\left(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}) + \tilde{\sigma}\boldsymbol{\zeta}\right)\Big], \qquad Q^{\pi_{\tilde{\sigma}}}(\mathbf{s}, \mathbf{a}) = r(\mathbf{s}, \mathbf{a}) + \gamma\, V^{\pi_{\tilde{\sigma}}}\left(F(\mathbf{s}, \mathbf{a})\right), \tag{18}$$

the first of the two holding because that actor plays $\mu_{\boldsymbol{\theta}}(\mathbf{s}) + \tilde{\sigma}\boldsymbol{\zeta}$ at its first step and follows itself afterwards. Reading the first at the next state and putting it in the second gives the map that carries a critic to the target of Eq. 6 in the mean over $\boldsymbol{\zeta}$, ignoring the minimum over the twin critics. That map shrinks the distance between two bounded functions by the factor γ, as the Bellman operator does, so that it has one bounded fixed point, $Q^{\pi_{\tilde{\sigma}}}$, which a critic with capacity to spare represents. In DDPG, where $\tilde{\sigma} = 0$, that fixed point is $Q^{\mu_{\boldsymbol{\theta}}}$ itself, and beyond the threshold the critic represents it only as finely as its resolution allows.

The slope of $Q^{\pi_{\tilde{\sigma}}}$ comes from the noise of the targets, which turns the value at a state into an average: by Eq. 18 the value at $\mathbf{s}$ is the mean of $Q^{\pi_{\tilde{\sigma}}}$ over the actions the first noise can take, and the noises of the steps that follow carry that mean over a window around the state those actions reach. That window opens in every direction of the state, as Assumption 3 asks. When the action moves, the window moves with it, so the derivative can be taken on the density of the window, of width $\tilde{\sigma}$, whose derivatives are of size $1/\tilde{\sigma}$ and multiply the variation of $V^{\pi_{\tilde{\sigma}}}$ across the window; that variation is at most a constant times $\tilde{\sigma}^{\alpha}$ by the argument of Theorem 2 applied to two trajectories driven by the same noise, so that the slope is of size at most $\tilde{\sigma}^{\alpha-1}$.

Theorem 4 (Size of the slope)**.** *Let Assumptions 1 and 3 hold with $k \geq 3$, and let $\gamma\Lambda > 1$ hold along the closed loop and along the trajectories whose actions carry the noise. Then for every state, every unit direction $\mathbf{e}$ of the action and every $\tilde{\sigma} \in (0, 1]$ the slope and the curvature along $\mathbf{e}$ exist and, for a constant C that depends on none of those three and with $\partial_{\mathbf{e}}$ the derivative in the direction $\mathbf{e}$,*

$$\left|\partial_{\mathbf{e}} Q^{\pi_{\tilde{\sigma}}}\big(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s})\big)\right| \leq C\,\tilde{\sigma}^{\,\alpha-1}, \qquad \left|\partial^2_{\mathbf{e}} Q^{\pi_{\tilde{\sigma}}}\big(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s})\big)\right| \leq C\,\tilde{\sigma}^{\,\alpha-2}. \tag{19}$$

If moreover the hypotheses of Theorem 3 hold, with $\mathbf{e}$ the direction they name, a power of $\tilde{\sigma}$ bounds the slope from below in the root mean square: on a neighbourhood $\mathcal{U}$ of the attractor, and for every $\tilde{\sigma}$ small enough, $\big(\mathrm{vol}(\mathcal{U})^{-1} \int_{\mathcal{U}} \partial_{\mathbf{e}} Q^{\pi_{\tilde{\sigma}}}(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}))^2\, d\mathbf{s}\big)^{1/2} \geq c\,\tilde{\sigma}^{\,\alpha_1-1}$, for a constant c, with $\mathrm{vol}(\mathcal{U})$ the volume of $\mathcal{U}$ and α_1 the exponent of Theorem 3 (proof in Appendix C.1).

The target noise smooths the action value over $\tilde{\sigma}$, which the algorithm fixes, and the critic represents what it is given down to ℓ_s, so that the slope it hands the actor is bounded in the mean by a constant times $\max\{L_F\tilde{\sigma}, \ell_s\}^{\alpha-1}$, with $L_F = \|\partial_{\mathbf{a}} F\|_\infty$ carrying a width in the action into a width in the state; that bound diverges as the coarser of the two scales is refined, since the exponent is below one. The noise and the resolution decide a second quantity, how closely the trained critic's slope follows that of the value it is fitted to: the difference is at most a constant times $\ell_s\,\tilde{\sigma}^{\alpha-2}$ for a critic linear in its parameters (Proposition 6), small against the slope itself only when the class resolves the value well below the noise. For a model-based actor, D'Oro and Jaśkowski [22] fit the critic so that its action-gradient is right. For a neural critic ℓ_s has no formula, but it can be measured (Section 5).

4.2 The average over states is the right gradient

The update of Eq. 1 averages the slope over the states the actor visits, and with a converged critic and those states in place of the buffer it is in the mean

$$\mathbf{g}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{s}\sim d^{\mu_{\boldsymbol{\theta}}}}\Big[\nabla_{\boldsymbol{\theta}}\mu_{\boldsymbol{\theta}}(\mathbf{s})^{\top}\,\nabla_{\mathbf{a}} Q^{\pi_{\tilde{\sigma}}}\big(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s})\big)\Big], \tag{20}$$

with $d^{\mu_{\boldsymbol{\theta}}}$ the discounted distribution of states of Section 2, which has a density. The derivative of J in $\boldsymbol{\theta}$ can be taken on that density, so that how rough $V^{\mu_{\boldsymbol{\theta}}}$ is has no bearing on it, and the same holds for Eq. 20, which pairs $\nabla_{\mathbf{a}} Q^{\pi_{\tilde{\sigma}}}$ with the same density.

Theorem 5 (The update is unbiased up to $\tilde{\sigma}^2$)**.** *Let Assumptions 1 and 3 hold with $k \geq 7$ on the family $\boldsymbol{\theta} \mapsto (F, \mu_{\boldsymbol{\theta}}, r)$. Then J is differentiable at $\boldsymbol{\theta}$ and there is $\sigma_K > 0$, independent of γ, such that*

$$\mathbf{g}(\boldsymbol{\theta}) = (1-\gamma)\,\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) + O(\tilde{\sigma}^2) \tag{21}$$

for $\ell_s \leq L_F\tilde{\sigma} \leq L_F\sigma_K$, with the constant in $O(\tilde{\sigma}^2)$ independent of γ (proof in Appendix C.2).

What Eq. 20 puts in place of the action-gradient of Eq. 5 exists at every state, at a size the noise sets. The actor never computes that average over all states: it draws a minibatch of B states, and the average of the terms over it has mean $\mathbf{g}(\boldsymbol{\theta})$ and, since each term is bounded by a constant times $\tilde{\sigma}^{\alpha-1}$, a variance of at most a constant times $\tilde{\sigma}^{2\alpha-2}/B$, so that refining the smoothing, as long as it stays coarser than the resolution of the critic, leaves the direction of the update where it is while widening the spread around that direction.

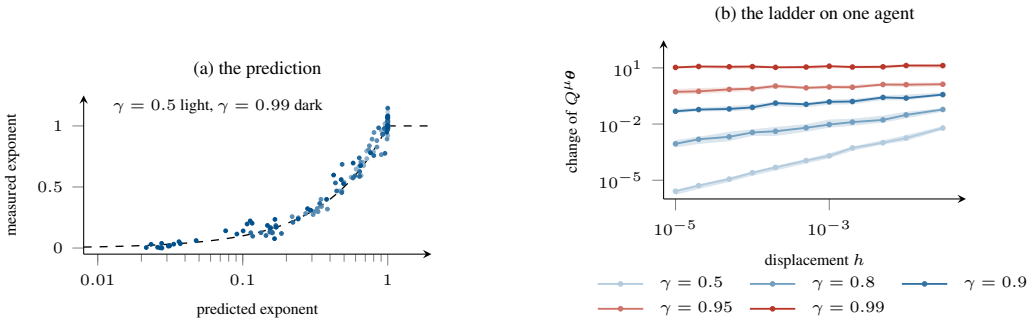

Figure 2: (a) Measured against predicted exponent, one point per agent at each of five discounts; the line is equality. (b) The change of the action value along the ladder of displacements on $HalfCheetah, relu, seed0$, median over its states with the interval resampling them leaves, where the slope of each line is the exponent at that discount: 0.96, 0.50, 0.25, 0.12 and 0.02.

The same mechanisms bound what the exploration noise costs from a single initial state, and where the target noise is the wider of the two, an estimate of that cost from the curvature of the critic falls short of it by a factor $(\tilde{\sigma}/\sigma)^{2-\alpha_1}$ in the mean square (Appendix D).

5 Experiments

A trained agent carries its own separation rate, so the threshold of Section 3 and the exponent of Theorem 2 can be measured on it. We measure them on 42 agents trained with TD3 on 7 tasks, in two ways. One runs two copies of the simulator and never evaluates the action value; the other sums Eq. 2 along the single trajectory the action starts, which a deterministic simulator allows, and reads the exponent as the slope of $\log \left|Q^{\mu_\theta}(\mathbf{s}, \mathbf{a} + h\mathbf{e}) - Q^{\mu_\theta}(\mathbf{s}, \mathbf{a})\right|$ against $\log h$ along the ladder. Both are read at states along the trajectories of the trained actor and reported as the median over those states, so that the exponent of Section 3.3, which belongs to a state, becomes a number that belongs to an agent.

Agents. The tasks are five of the MuJoCo locomotion environments [23] of Gymnasium [24], HalfCheetah, Ant, Swimmer, Hopper and Walker2d, with InvertedDoublePendulum and the classic-control Pendulum, on both of which the actor holds the state near a fixed point. Six actors to a task are trained by the reference implementation of TD3 [2] at the hyperparameters of its authors, among them $\gamma = 0.99$ and $\tilde{\sigma} = 0.2$, half with a relu critic and half with a softplus one. Every measurement restores the simulator to a state the trained actor visits and runs it in double precision, so that two copies given the same actions stay identical to the last digit (Appendix F).

The threshold. The same trajectories give the action value at every discount, since summing the same rewards with a different γ only weights the future differently, so the threshold can be crossed without retraining anything. At $\gamma = 0.5$ no closed loop here doubles a displacement faster than the discount halves a reward, and the exponent measured on the action value is one on every task. Theorem 1 fixes that value for a loop below the threshold, so the reading there calibrates the estimator. As the discount grows the agents cross the threshold, and the change of the action value along the ladder flattens with them (Figure 2b, Table 1): those of HalfCheetah and Ant, whose copies part fastest, at $\gamma = 0.8$; those of Hopper and Swimmer, the slowest, only at $\gamma = 0.99$, where their exponents are 0.40 and 0.19; and five of the six actors of Pendulum not at all. Those six were trained on the same task with the same algorithm and reach returns that differ by less than the spread between their own episodes, yet the sixth has copies that part at every displacement we read: which side an agent falls on belongs to the closed loop that training produced.

The prediction. The two copies give the exponent of an agent as the power of the displacement that the discounted sum of their distances follows (Eq. 15). On the 33 agents whose copies part, at the discount they were trained at, that reading and the exponent measured on the action value differ by a median of 0.020, and on 21 of them the prediction falls inside the interval that resampling the

	two copies		$\hat{\alpha}$ at five discounts				
task	D	α_S	0.5	0.8	0.9	0.95	0.99
HalfCheetah	1.8	0.03	0.98	0.58	0.29	0.13	**0.01**
Ant	2.1	0.03	1.04	0.72	0.32	0.15	**0.03**
Walker2d	9.1	0.21	1.00	0.98	1.00	0.66	**0.23**
Hopper	19.9	0.38	1.00	1.00	1.00	0.98	**0.40**
Swimmer	22.5	0.17	1.00	1.00	1.00	0.93	**0.19**
InvDoublePendulum	4.8	0.11	1.03	0.99	0.86	0.55	**0.15**
Pendulum	2.3	0.04	0.99	0.79	0.40	0.18	**0.04**

Table 1: Medians over the agents of each task whose copies part: the delay D of Eq. 16 and the exponent α_S of the sum of Eq. 15, read one independently of the other and not as $D/69$, and the measured exponent at five discounts (Appendix E has the agents whose copies never part).

trajectories leaves around the measurement (Figure 2a); read through the delay in steps of Eq. 16, the same comparison gives a median gap of 0.034 and 13 agents inside the interval.

The curvature. Theorem 1 gives a second threshold, stricter than the first, and between the two a slope is guaranteed but no curvature. It is read from the same rollouts, by playing the undisplaced action and two displacements in the ratio two, h and $2h$, and following the combination $Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a} + 2h\mathbf{e}) - 2Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a} + h\mathbf{e}) + Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})$, whose exponent in h is two where the curvature exists. The exponent the two copies give an agent puts it in one of three regimes, with both the slope and the curvature, with the slope alone, or with neither, since α fixes the separation rate and with it $\gamma\Lambda^2$ (Table 3 of Appendix E). The exponent of that combination is 2.05 where both the slope and the curvature exist, 1.50 where only the slope does, and 0.28 where neither does, the exponent of the action value itself in that group; the first two groups have the same exponent in the action, both being below the first threshold, so that the curvature alone tells them apart, with intervals that do not meet.

The slope. The two scales that decide the slope the update averages can both be read on the trained agents. The resolution of the critic in the state is read by blurring the critic there and measuring the width at which the slope it hands the actor begins to change; in standard deviations of the states the actor visits, that width has median 0.018. It separates by activation, the relu critics resolving 0.010 against 0.118 for the softplus ones, and the same ordering holds in the action, where the softplus critics of 13 agents are coarser than the target noise (Table 4 of Appendix E). The slope of the action value smoothed in the action is read at widths we set, with no critic in the way. On the agents beyond the threshold, except a few on InvertedDoublePendulum, the spread between the states of an agent grows as the smoothing is refined, while on the Pendulum agents whose copies do not part it stays where it is (Figure 6).

6 Conclusion

Whether the action value of a deterministic actor has a derivative in the action is decided by a race between the discount and the closed loop, and on 33 of the 42 agents we trained the loop wins, so that the value the deterministic policy gradient differentiates is Hölder with an exponent below one, and on a hyperbolic attractor whose loop stretches fast enough its derivative is absent at almost every state, while the update the algorithm performs is defined everywhere: what the critic represents is smoothed, at a width the noise of the targets and the resolution of the network set together, and where the critic resolves the value below that width the slope it hands the actor averages over the states to the gradient of the return, up to the factor $1 - \gamma$ and a remainder of order $\tilde{\sigma}^2$. Refining that width widens the spread around the average to at most a constant times $\tilde{\sigma}^{\alpha-1}$, which a minibatch of B states buys back as $1/\sqrt{B}$, so that the width and the size of the minibatch enter one bound once α is known. Reading the threshold off two copies carries to a dynamics smooth away from the contacts, and to a neural critic, for which the resolutions measured here say what a bound should give.

Reproducibility Statement

The agents are trained by the reference implementation of TD3 at the hyperparameters of its authors, named in Section 5. Appendix F gives the measurement protocol in full: the states at which each quantity is read, the ladder of displacements, the horizon, the resampling behind every interval, and the double precision that keeps two copies of the simulator identical to the last digit. Appendix E reports the readings of every one of the 42 agents, in Tables 5 and 6, so that each number in the body can be traced to the agents behind it. The hypotheses of every theorem are stated in full in Appendix B.4 and the proofs in the appendices that follow it.

Use of Large Language Models

Large language models were used in the preparation of this paper for four tasks. They polished the English of passages written by the authors, without changing what those passages assert. They searched the literature and summarised candidate references, each of which was then checked against its primary source by the authors. They wrote the plotting code that turns the measurement artifacts into the figures, whose output was checked against the numbers the tables report. And they were used as adversarial readers of the proofs, to look for gaps and for hypotheses left unstated; every objection they raised was answered by the authors. The statements of the theorems, the hypotheses they carry and the arguments that establish them are the authors' own, as are the experiments and the claims made about them, and the authors take responsibility for the final content of this work.

References

[1] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

[2] Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, 2018.

[3] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, 2014.

[4] Huaqing Xiong, Tengyu Xu, Lin Zhao, Yingbin Liang, and Wei Zhang. Deterministic policy gradient: Convergence analysis. In *Conference on Uncertainty in Artificial Intelligence*, 2022.

[5] Emmanuel Rachelson and Michail G. Lagoudakis. On the locality of action domination in sequential decision making. In *International Symposium on Artificial Intelligence and Mathematics*, 2010.

[6] Hans Harder and Sebastian Peitz. On the continuity and smoothness of the value function in reinforcement learning and optimal control. In *IEEE Conference on Decision and Control*, 2024.

[7] Davide Maran, Alberto Maria Metelli, and Marcello Restelli. Tight performance guarantees of imitator policies with continuous actions. In *AAAI Conference on Artificial Intelligence*, 2023.

[8] Tao Wang, Sylvia Herbert, and Sicun Gao. Fractal landscapes in policy optimization. In *Advances in Neural Information Processing Systems*, 2023.

[9] Paavo Parmas, Carl Edward Rasmussen, Jan Peters, and Kenji Doya. PIPPS: Flexible model-based policy search robust to the curse of chaos. In *Proceedings of the International Conference on Machine Learning*, 2018.

[10] Luke Metz, C. Daniel Freeman, Samuel S. Schoenholz, and Tal Kachman. Gradients are not all you need. *arXiv preprint arXiv:2111.05803*, 2021.

[11] Hyung Ju Suh, Max Simchowitz, Kaiqing Zhang, and Russ Tedrake. Do differentiable simulators give better policy gradients? In *Proceedings of the International Conference on Machine Learning*, 2022.

[12] Tao Wang, Sylvia Herbert, and Sicun Gao. Mollification effects of policy gradient methods. In *International Conference on Machine Learning*, 2024.

[13] Harshat Kumar, Dionysios S. Kalogerias, George J. Pappas, and Alejandro Ribeiro. Zeroth-order deterministic policy gradient. *arXiv preprint arXiv:2006.07314*, 2020.

[14] Alessandro Montenegro, Marco Mussi, Alberto Maria Metelli, and Matteo Papini. Learning optimal deterministic policies with stochastic policy gradients. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[15] Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Policy gradient in Lipschitz Markov decision processes. *Machine Learning*, 100(2–3):255–283, 2015.

[16] Anatole Katok and Boris Hasselblatt. *Introduction to the Modern Theory of Dynamical Systems*. Cambridge University Press, Cambridge, 1995.

[17] David Ruelle. A measure associated with Axiom-A attractors. *American Journal of Mathematics*, 98(3):619–654, 1976. Theorem, remark and equation numbers refer to the IHES preprint P/73/75.

[18] G. H. Hardy. Weierstrass's non-differentiable function. *Transactions of the American Mathematical Society*, 17(3):301–325, 1916.

[19] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 2019.

[20] Ge Yang, Anurag Ajay, and Pulkit Agrawal. Overcoming the spectral bias of neural value approximation. In *International Conference on Learning Representations*, 2022.

[21] John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

[22] Pierluca D'Oro and Wojciech Jaśkowski. How to learn a useful critic? model-based action-gradient-estimator policy optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 313–324, 2020.

[23] Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[24] Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In *Advances in Neural Information Processing Systems 38, Datasets and Benchmarks Track*, 2025.

[25] Christian Bonatti, Lorenzo J. Díaz, and Marcelo Viana. *Dynamics Beyond Uniform Hyperbolicity*, volume 102 of *Encyclopaedia of Mathematical Sciences*. Springer, 2005.

[26] Sylvain Crovisier and Rafael Potrie. Introduction to partially hyperbolic dynamics. Lecture notes, School on Dynamical Systems, ICTP, Trieste, 2015.

[27] Rufus Bowen and David Ruelle. The ergodic theory of Axiom A flows. *Inventiones Mathematicae*, 29:181–202, 1975.

[28] Rufus Bowen. *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, volume 470 of *Lecture Notes in Mathematics*. Springer, 1975. Second revised edition, J.-R. Chazottes (ed.), 2008; theorem numbers as in that edition.

[29] Lai-Sang Young. What are SRB measures, and which dynamical systems have them? *Journal of Statistical Physics*, 108:733–754, 2002.

[30] Luis Barreira and Yakov Pesin. *Nonuniform Hyperbolicity: Dynamics of Systems with Nonzero Lyapunov Exponents*, volume 115 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 2007.

[31] Edward James McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40(12):837–842, 1934.

[32] Eduardo D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, volume 6 of *Texts in Applied Mathematics*. Springer, 2 edition, 1998.

[33] William Parry and Mark Pollicott. *Zeta Functions and the Periodic Orbit Structure of Hyperbolic Dynamics*, volume 187–188 of *Astérisque*. Société Mathématique de France, 1990.

[34] Roland Zweimüller. Mixing limit theorems for ergodic transformations. *Journal of Theoretical Probability*, 20(4):1059–1071, 2007.

[35] Sébastien Gouëzel and Carlangelo Liverani. Banach spaces adapted to Anosov systems. *Ergodic Theory and Dynamical Systems*, 26(1):189–217, 2006.

[36] Sébastien Gouëzel and Carlangelo Liverani. Compact locally maximal hyperbolic sets for smooth maps: fine statistical properties. *Journal of Differential Geometry*, 79(3):433–477, 2008.

[37] Ronald A. DeVore and George G. Lorentz. *Constructive Approximation*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 1993.

[38] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

A Proofs of Section 2

For a parameter $\boldsymbol{\theta}'$ near $\boldsymbol{\theta}$, the actor $\mu_{\boldsymbol{\theta}'}$ moves the state through the closed loop $f_{\boldsymbol{\theta}'}(\mathbf{s}) = F(\mathbf{s}, \mu_{\boldsymbol{\theta}'}(\mathbf{s}))$ and collects the reward $\bar{r}_{\boldsymbol{\theta}'}(\mathbf{s}) = r(\mathbf{s}, \mu_{\boldsymbol{\theta}'}(\mathbf{s}))$, so that its value $V_{\boldsymbol{\theta}'} = V^{\mu_{\boldsymbol{\theta}'}}$, read at the actor $\mu_{\boldsymbol{\theta}'}$ as in Section 2, satisfies

$$V_{\boldsymbol{\theta}'}(\mathbf{s}) = \sum_{t \geq 0} \gamma^t \, \bar{r}_{\boldsymbol{\theta}'}\big(f^t_{\boldsymbol{\theta}'}(\mathbf{s})\big) = \bar{r}_{\boldsymbol{\theta}'}(\mathbf{s}) + \gamma \, V_{\boldsymbol{\theta}'}\big(f_{\boldsymbol{\theta}'}(\mathbf{s})\big). \tag{22}$$

At $\boldsymbol{\theta}' = \boldsymbol{\theta}$ the loop and the value are those of Section 2, f and $V = V^{\mu_{\boldsymbol{\theta}}}$, and $\bar{r}$ is the reward collected along that loop. The Jacobian of the t-th iterate is the product of the Jacobians along the trajectory,

$$Df^t_{\boldsymbol{\theta}'}(\mathbf{s}) = Df_{\boldsymbol{\theta}'}\big(f^{t-1}_{\boldsymbol{\theta}'}\mathbf{s}\big) \cdots Df_{\boldsymbol{\theta}'}(\mathbf{s}), \qquad Df^0_{\boldsymbol{\theta}'} = \mathbf{I}, \tag{23}$$

and $\nabla_{\boldsymbol{\theta}} \mu_{\boldsymbol{\theta}'} \in \mathbb{R}^{d_A \times d_\theta}$, $\partial_{\mathbf{s}} \mu_{\boldsymbol{\theta}'} \in \mathbb{R}^{d_A \times d_S}$ are the Jacobians of the action in the parameter and in the state. The bounds of Assumption 1 on $\mathbf{B} = \partial_{\mathbf{a}} F$ and on the derivatives of r are B_1 and L_r, and matrix norms are the operator norms induced by the Euclidean norm, which a matrix and its transpose share. Throughout the appendix F and r are read on all of $\mathbb{R}^{d_S} \times \mathbb{R}^{d_A}$ and the actors on all of $\mathbb{R}^{d_S}$, with Assumption 1 and $|r| \leq r_{\max}$ there, so that a displaced action and a segment between two states are admissible; an environment given on $\mathcal{S} \times \mathcal{A}$ is read through such an extension, which leaves the closed loop and its value unchanged on the states and actions the actor visits.

Every appendix writes C for a constant that may differ from one line to the next, with the quantities it depends on named where it first appears in a proof.

Assumption 4 (Actors near $\boldsymbol{\theta}$)**.** *There are an open neighbourhood Θ_0 of $\boldsymbol{\theta}$ and a set $S_0 \subseteq \mathcal{S}$, open in $\mathbb{R}^{d_S}$, with $p_0(S_0) = 1$ and $f_{\boldsymbol{\theta}'}(S_0) \subseteq S_0$ for every $\boldsymbol{\theta}' \in \Theta_0$, such that*

(i) $(\boldsymbol{\theta}', \mathbf{s}) \mapsto \mu_{\boldsymbol{\theta}'}(\mathbf{s})$ is continuously differentiable on $\Theta_0 \times S_0$, with $\|\nabla_{\boldsymbol{\theta}} \mu_{\boldsymbol{\theta}'}\| \leq L_\mu$ and $\|\partial_{\mathbf{s}} \mu_{\boldsymbol{\theta}'}\| \leq L_\mu$, and $\mu_{\boldsymbol{\theta}'}(\mathbf{s})$ lies in the interior of $\mathcal{A}$;

(ii) $\|Df^n_{\boldsymbol{\theta}'}(\mathbf{s})\| \le C_A\Lambda^n$ *for every* $\boldsymbol{\theta}' \in \Theta_0$*, every* $\mathbf{s} \in S_0$ *and every* $n \ge 0$*, with* $\Lambda > 0$*,* $C_A \ge 1$ *and* $\gamma\Lambda < 1$*.*

Assumption 4(ii) gives every actor near $\boldsymbol{\theta}$ a separation rate Λ with $\gamma\Lambda < 1$, the condition of Theorem 1 at $j = 1$, and asks it only on a set of states that the closed loop keeps and that the reset distribution fills. Throughout this section the gradient of $V_{\boldsymbol{\theta}'}$ in the parameter and the integrand of the policy gradient are set to zero outside S_0, a set that every law p_t of the state gives full measure, since $p_0(S_0) = 1$ and $f_{\boldsymbol{\theta}'}(S_0) \subseteq S_0$; both are continuous on the open S_0 and bounded there, which leaves every expectation defined.

The assumption holds, for instance, on a bounded open S_0 that carries p_0, with $\overline{S_0} \subset \mathcal{S}$ and $\overline{f(S_0)} \subset S_0$, when the actor is continuously differentiable in $(\boldsymbol{\theta}', \mathbf{s})$ on a neighbourhood of $\{\boldsymbol{\theta}\} \times \overline{S_0}$ with values in the interior of $\mathcal{A}$, and the closed loop has $\sup_{\overline{S_0}} \|Df\| < \Lambda < 1/\gamma$. The compact set $f(\overline{S_0})$ then lies at a distance $\varepsilon > 0$ from the complement of S_0, and $\mu_{\boldsymbol{\theta}}(\overline{S_0})$ at a distance $\eta > 0$ from the boundary of $\mathcal{A}$; this is the room a small change of the parameter may use. On $\overline{S_0}$ the maps $\mu_{\boldsymbol{\theta}'}$, $f_{\boldsymbol{\theta}'}$ and $Df_{\boldsymbol{\theta}'}$ converge uniformly in the state as $\boldsymbol{\theta}' \to \boldsymbol{\theta}$. A bounded Θ_0 on which they move by less than $\min(\varepsilon, \eta, \Lambda - \sup_{\overline{S_0}} \|Df\|)$ therefore keeps $f_{\boldsymbol{\theta}'}(S_0) \subseteq S_0$ and the played actions inside $\mathcal{A}$, and leaves $\|Df_{\boldsymbol{\theta}'}\| \le \Lambda$ on S_0, so that a product of n such Jacobians has norm at most Λ^n, which is (ii) with $C_A = 1$. The derivatives of the actor, continuous on the compact $\overline{\Theta_0} \times \overline{S_0}$, are bounded there, which is (i).

Lemma 1 (Derivatives of the value)**.** *Under Assumptions 1 and 4,* $(\boldsymbol{\theta}', \mathbf{s}) \mapsto V_{\boldsymbol{\theta}'}(\mathbf{s})$ *is continuously differentiable on* $\Theta_0 \times S_0$*, and for every* $\boldsymbol{\theta}' \in \Theta_0$ *and* $\mathbf{s} \in S_0$

$$\|\nabla_{\mathbf{s}} V_{\boldsymbol{\theta}'}(\mathbf{s})\| \le M_{\mathrm{s}} = \frac{C_A G}{1 - \gamma\Lambda}, \qquad \|\nabla_{\boldsymbol{\theta}} V_{\boldsymbol{\theta}'}(\mathbf{s})\| \le M_{\mathrm{p}} = \frac{C_A B_1 L_\mu G\,\gamma}{(1 - \gamma\bar{\Lambda})^2} + \frac{L_\mu L_r}{1 - \gamma}, \tag{24}$$

where $\bar{\Lambda} = \max(1, \Lambda)$*, and* G*, the bound of Section 3.1 on the gradient* $\nabla\bar{r}$ *of Eq. 9, may be taken equal to* $L_r(1 + L_\mu)$ *on* $\Theta_0 \times S_0$ *by Assumption 4(i).*

Proof. **Differentiating term by term.** We bound the two gradients of the value by differentiating Eq. 22 term by term: the state at step t moves with the parameter through every step before it, which costs a factor $t\bar{\Lambda}^{\,t}$, and the discount pays for it.

The trajectory and its derivatives. Fix $\boldsymbol{\theta}' \in \Theta_0$ and $\mathbf{s} \in S_0$, and let $\mathbf{y}_t = f^t_{\boldsymbol{\theta}'}(\mathbf{s})$, a trajectory that stays in S_0. Differentiating the reward through both of its arguments, and bounding the result by G with Assumption 1 and Assumption 4(i), gives the gradient of the reward along the loop,

$$\nabla_{\mathbf{s}}\bar{r}_{\boldsymbol{\theta}'}(\mathbf{s}) = \nabla_{\mathbf{s}} r\big(\mathbf{s}, \mu_{\boldsymbol{\theta}'}(\mathbf{s})\big) + \partial_{\mathbf{s}}\mu_{\boldsymbol{\theta}'}(\mathbf{s})^\top \nabla_{\mathbf{a}} r\big(\mathbf{s}, \mu_{\boldsymbol{\theta}'}(\mathbf{s})\big), \qquad \|\nabla_{\mathbf{s}}\bar{r}_{\boldsymbol{\theta}'}\| \le G. \tag{25}$$

The parameter enters $\mathbf{y}_{t+1} = F(\mathbf{y}_t, \mu_{\boldsymbol{\theta}'}(\mathbf{y}_t))$ through the state and through the action, so the chain rule gives a recursion for the derivative of the state in the parameter, $\mathbf{Y}_t = \partial\mathbf{y}_t/\partial\boldsymbol{\theta}'$, whose solution is a sum over the steps already taken and whose bound follows from Assumption 4(ii) with $\sum_{j<t}\Lambda^j \le t\bar{\Lambda}^{\,t-1}$:

$$\mathbf{Y}_0 = \mathbf{0}, \qquad \mathbf{Y}_{t+1} = Df_{\boldsymbol{\theta}'}(\mathbf{y}_t)\,\mathbf{Y}_t + \mathbf{B}\big(\mathbf{y}_t, \mu_{\boldsymbol{\theta}'}(\mathbf{y}_t)\big)\nabla_{\boldsymbol{\theta}}\mu_{\boldsymbol{\theta}'}(\mathbf{y}_t), \tag{26}$$

$$\mathbf{Y}_t = \sum_{i=0}^{t-1} Df^{\,t-1-i}_{\boldsymbol{\theta}'}(\mathbf{y}_{i+1})\,\mathbf{B}\big(\mathbf{y}_i, \mu_{\boldsymbol{\theta}'}(\mathbf{y}_i)\big)\nabla_{\boldsymbol{\theta}}\mu_{\boldsymbol{\theta}'}(\mathbf{y}_i), \qquad \|\mathbf{Y}_t\| \le C_A B_1 L_\mu\, t\,\bar{\Lambda}^{\,t-1}. \tag{27}$$

The t-th term of Eq. 22 is continuously differentiable, by induction on t and as a composition of continuously differentiable maps, and differentiating it through $\mathbf{y}_t$ gives

$$\big\|\partial_{\mathbf{s}}\big[\gamma^t\bar{r}_{\boldsymbol{\theta}'}(\mathbf{y}_t)\big]\big\| = \gamma^t\big\|Df^t_{\boldsymbol{\theta}'}(\mathbf{s})^\top\nabla_{\mathbf{s}}\bar{r}_{\boldsymbol{\theta}'}(\mathbf{y}_t)\big\| \le C_A G\,(\gamma\Lambda)^t, \tag{28}$$

$$\begin{aligned}\big\|\partial_{\boldsymbol{\theta}}\big[\gamma^t\bar{r}_{\boldsymbol{\theta}'}(\mathbf{y}_t)\big]\big\| &= \gamma^t\big\|\mathbf{Y}_t^\top\nabla_{\mathbf{s}}\bar{r}_{\boldsymbol{\theta}'}(\mathbf{y}_t) + \nabla_{\boldsymbol{\theta}}\mu_{\boldsymbol{\theta}'}(\mathbf{y}_t)^\top\nabla_{\mathbf{a}} r\big(\mathbf{y}_t, \mu_{\boldsymbol{\theta}'}(\mathbf{y}_t)\big)\big\| \\ &\le C_A B_1 L_\mu G\, t\,\gamma^t\bar{\Lambda}^{\,t-1} + L_\mu L_r\gamma^t,\end{aligned} \tag{29}$$

whose sums over t are a geometric and a differentiated geometric series, finite because $\gamma\Lambda < 1$ and hence $\gamma\bar{\Lambda} = \max(\gamma, \gamma\Lambda) < 1$, and equal to M_{s} and M_{p}.

Uniform convergence of the series. These bounds do not depend on $(\boldsymbol{\theta}', \mathbf{s})$, so both series converge uniformly on $\Theta_0 \times S_0$, as does the series of Eq. 22 itself, whose terms are continuous and at

most $\gamma^t r_{\max}$. On the segments of a ball contained in the open set $\Theta_0 \times S_0$, the Weierstrass test and termwise differentiation therefore give the partial derivatives of $V_{\boldsymbol{\theta}'}(\mathbf{s})$, continuous as uniform limits of continuous functions, so that the value is continuously differentiable there with the stated gradients. At $\boldsymbol{\theta}$ the terms of Eq. 22 are continuous at every state by Assumption 1 and at most $\gamma^t r_{\max}$, so V is continuous everywhere. □

Proposition 1 (Deterministic policy gradient)**.** Under Assumptions 1 and 4, J is continuously differentiable on Θ_0, and Eq. 5 of Section 2 holds:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{1-\gamma}\, \mathbb{E}_{\mathbf{s}\sim d^{\mu_{\boldsymbol{\theta}}}} \Big[\nabla_{\boldsymbol{\theta}} \mu_{\boldsymbol{\theta}}(\mathbf{s})^\top \, \nabla_{\mathbf{a}} Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s},\mathbf{a})\big|_{\mathbf{a}=\mu_{\boldsymbol{\theta}}(\mathbf{s})} \Big]. \tag{30}$$

Proof. **One term per step.** The parameter reaches the return through the action played at every state along the trajectory, so differentiating the value once and unrolling the recursion collects one term per step; those terms are the integrand of Eq. 30, which we write

$$\mathbf{j}(\mathbf{s}) = \nabla_{\boldsymbol{\theta}} \mu_{\boldsymbol{\theta}}(\mathbf{s})^\top \, \nabla_{\mathbf{a}} Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s},\mathbf{a})\big|_{\mathbf{a}=\mu_{\boldsymbol{\theta}}(\mathbf{s})} \quad (\mathbf{s} \in S_0), \qquad \mathbf{j}(\mathbf{s}) = \mathbf{0} \quad (\mathbf{s} \notin S_0). \tag{31}$$

The set of actions $\mathbf{a}$ with $F(\mathbf{s},\mathbf{a}) \in S_0$ is open, being the preimage of the open set S_0 under the continuous $F(\mathbf{s},\cdot)$, and it contains $\mu_{\boldsymbol{\theta}}(\mathbf{s})$, since $f(\mathbf{s}) \in S_0$. That action lies in the interior of $\mathcal{A}$ by Assumption 4(i), so Eq. 3 holds in a neighbourhood of it.

Differentiating the value in the parameter. By the chain rule, differentiating Eq. 22 in the parameter at $\boldsymbol{\theta}' = \boldsymbol{\theta}$ spreads the derivative over the reward, the next state and the value there, the first two in the bracket of Eq. 32 and the third in the term that follows it. That bracket is the action-gradient of Eq. 3 at $\mathbf{a} = \mu_{\boldsymbol{\theta}}(\mathbf{s})$, continuous because $\nabla_{\mathbf{a}} r$, $\mathbf{B}$ and F are continuous by Assumption 1 and $\nabla_{\mathbf{s}} V$ is continuous on S_0, and bounded by $L_r + \gamma B_1 M_{\mathrm{s}}$ through Lemma 1, so that $\mathbf{j}$ of Eq. 31 is continuous on S_0 and bounded by $M_{\mathrm{j}} = L_\mu(L_r + \gamma B_1 M_{\mathrm{s}})$, and Eq. 33 reads the first term as $\mathbf{j}(\mathbf{s})$. Unrolling that recursion along the trajectory $\mathbf{s}_t = f^{\,t}(\mathbf{s}_0)$ of a state $\mathbf{s}_0 \in S_0$ leaves, by the same lemma, a remainder at most $\gamma^T M_{\mathrm{p}}$, which vanishes with T, and gives

$$\nabla_{\boldsymbol{\theta}} V(\mathbf{s}) = \nabla_{\boldsymbol{\theta}} \mu_{\boldsymbol{\theta}}(\mathbf{s})^\top \Big[\nabla_{\mathbf{a}} r(\mathbf{s},\mathbf{a}) + \gamma\, \mathbf{B}(\mathbf{s},\mathbf{a})^\top \nabla_{\mathbf{s}} V\big(f(\mathbf{s})\big) \Big]_{\mathbf{a}=\mu_{\boldsymbol{\theta}}(\mathbf{s})} + \gamma\, \nabla_{\boldsymbol{\theta}} V\big(f(\mathbf{s})\big) \tag{32}$$

$$= \mathbf{j}(\mathbf{s}) + \gamma\, \nabla_{\boldsymbol{\theta}} V\big(f(\mathbf{s})\big), \tag{33}$$

$$\nabla_{\boldsymbol{\theta}} V(\mathbf{s}_0) = \sum_{t<T} \gamma^t\, \mathbf{j}(\mathbf{s}_t) + \gamma^T\, \nabla_{\boldsymbol{\theta}} V(\mathbf{s}_T) \;\xrightarrow[T\to\infty]{}\; \sum_{t\geq 0} \gamma^t\, \mathbf{j}(\mathbf{s}_t). \tag{34}$$

The derivative passes under the expectation, since on a segment of Θ_0 along a unit vector $\mathbf{w}$ the mean value theorem makes the difference quotient of $V_{\boldsymbol{\theta}'}$ equal to $\mathbf{w}^\top \nabla_{\boldsymbol{\theta}} V_{\boldsymbol{\theta}''}(\mathbf{s}_0)$ at an intermediate $\boldsymbol{\theta}''$, hence at most M_{p}, and dominated convergence applies; the sum passes through it as well, dominated by $M_{\mathrm{j}}/(1-\gamma)$. The law of the state at step t is the image p_t of p_0 under the continuous $f^{\,t}$, and $f^{\,t}(S_0) \subseteq S_0$ with $p_0(S_0) = 1$, so that every term reads $\mathbf{j}$ on S_0 alone. Averaging over the initial state, with $d^{\mu_{\boldsymbol{\theta}}} = (1-\gamma)\sum_{t\geq 0} \gamma^t p_t$, gives

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{s}_0\sim p_0}\big[\nabla_{\boldsymbol{\theta}} V(\mathbf{s}_0)\big] = \sum_{t\geq 0} \gamma^t\, \mathbb{E}_{\mathbf{s}_0\sim p_0}\big[\mathbf{j}(f^{\,t}(\mathbf{s}_0))\big] \\ &= \sum_{t\geq 0} \gamma^t\, \mathbb{E}_{\mathbf{s}\sim p_t}\big[\mathbf{j}(\mathbf{s})\big] = \frac{1}{1-\gamma}\, \mathbb{E}_{\mathbf{s}\sim d^{\mu_{\boldsymbol{\theta}}}}\big[\mathbf{j}(\mathbf{s})\big], \end{aligned} \tag{35}$$

which is Eq. 30, since $d^{\mu_{\boldsymbol{\theta}}}$ charges S_0 alone.

The same bounds near $\boldsymbol{\theta}$. Since the bounds of Lemma 1 hold at every $\boldsymbol{\theta}' \in \Theta_0$, the same argument gives $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}') = \mathbb{E}_{\mathbf{s}_0\sim p_0}[\nabla_{\boldsymbol{\theta}} V_{\boldsymbol{\theta}'}(\mathbf{s}_0)]$ there, and the continuity of those gradients makes $\nabla_{\boldsymbol{\theta}} J$ continuous on Θ_0. □

Proposition 1 is the deterministic reading of the policy gradient theorem of Silver et al. [3], proved for transitions with a density on a compact state space under their conditions A.1: the transition density and its gradient in the action, the reward and its gradient in the action, the actor and its Jacobian in the parameters, and the initial density, are continuous in all their variables. Their matrix $\nabla_{\boldsymbol{\theta}} \mu_{\boldsymbol{\theta}}(\mathbf{s})$ is the transpose of ours, and their time starts at 1, so that with the time counted from zero

their discounted state distribution is $\sum_{t\geq 0}\gamma^t p_t = d^{\mu_{\boldsymbol{\theta}}}/(1-\gamma)$, which is not normalised. Their reward integrated against it is $\sum_{t\geq 0}\gamma^t \mathbb{E}_{\mathbf{s}\sim p_t}[\bar{r}(\mathbf{s})]$, which is $J(\boldsymbol{\theta})$ by Eq. 22 and the change of variables of Eq. 35, so the two formulas differentiate the same function, and the factor $1/(1-\gamma)$ of Eq. 30 is carried there by the distribution. A deterministic transition has no density, because the next state is a single point of zero volume that carries probability one. The bounds that their proof draws from the compact state space, on $\nabla_{\boldsymbol{\theta}} V$, on the action-gradient at the played action and on $\nabla_{\boldsymbol{\theta}}\mu_{\boldsymbol{\theta}}$, are here M_{p}, $L_r + \gamma B_1 M_{\mathrm{s}}$ and L_μ; bounds of this kind are assumed outright in their conditions A.2, on which their Theorem 2 rests.

B Proofs of Section 3

B.1 Proof of Theorem 1

Theorem 1 (Derivatives in the action)**.** *Under Assumption 1, if $j \leq k$ and $\gamma\Lambda^j < 1$, then for every state $\mathbf{s}$ the action value $Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s},\cdot)$ is j times continuously differentiable, and its j-th derivative is bounded uniformly in the state and the action.*

We prove the theorem by differentiating Eq. 2 term by term. A displaced action moves the whole trajectory, a derivative of order p of the state at step t carries p copies of that displacement and so grows like $\bar{\Lambda}^{p\,t}$, and the discount beats that growth when $\gamma\bar{\Lambda}^{\,j} < 1$.

Throughout, H_p bounds the derivatives of order p of F, $\mu_{\boldsymbol{\theta}}$ and r, those of F and r taken jointly in the state and the action, so that $\|\mathbf{B}\| \leq H_1$ and $\|\nabla_{\mathbf{a}} r\| \leq H_1$ and the constants B_1 and L_r of Appendix A may both be taken equal to H_1; $C_A \geq 1$ is the constant of the separation rate Λ of Section 3.1, so that $\|Df^{\,n}(\mathbf{x})\| \leq C_A\Lambda^n$ at every state $\mathbf{x}$ of a trajectory of the closed loop and every $n \geq 0$; and $\bar{r}(\mathbf{s}) = r(\mathbf{s},\mu_{\boldsymbol{\theta}}(\mathbf{s}))$ is the reward collected along that loop. To carry one rate whether Λ is above or below one we write $\bar{\Lambda} = \max(1,\Lambda)$, again a separation rate with the same constant and with $\gamma\bar{\Lambda}^{\,j} < 1$, since either $\bar{\Lambda} = \Lambda$, or $\bar{\Lambda} = 1$ and $\gamma\bar{\Lambda}^{\,j} = \gamma < 1$. A derivative of order p along directions $\mathbf{w}_1,\dots,\mathbf{w}_p$ is multilinear and symmetric in them, and its norm is the largest norm it takes on unit directions. In Appendices B.1 and B.2, C depends only on C_A and on $H_1,\dots,H_p$, where p is the order of the derivative it accompanies.

Lemma 2 (Derivatives of the trajectory)**.** *Let $\mathbf{s}$ be a state, $\mathbf{a}$ an action, and let $\mathbf{s}_1(\mathbf{u}) = F(\mathbf{s},\mathbf{a}+\mathbf{u})$ and $\mathbf{s}_{t+1}(\mathbf{u}) = f(\mathbf{s}_t(\mathbf{u}))$ be the trajectory opened by the displaced action, k times continuously differentiable in $\mathbf{u}$ by Assumption 1 and induction on t. Then, for every $\mathbf{u}$, every m with $m+1 \leq k$ and every $t \geq 1$,*

$$\left\|D^{m+1}\mathbf{s}_t(\mathbf{u})\right\| \leq M_m\, t^m\, \bar{\Lambda}^{(m+1)t}, \tag{36}$$

with constants M_m that depend only on m, C_A and $H_1,\dots,H_{m+1}$.

Proof. Fix unit directions $\boldsymbol{v}_0,\dots,\boldsymbol{v}_m$ of the action space and write $\partial_P\mathbf{s}_t$ for the derivative of $\mathbf{s}_t$ along the directions $\boldsymbol{v}_l$ with $l \in P$, so that $\|D^{m+1}\mathbf{s}_t\|$ is the largest norm of $\partial_{\{0,\dots,m\}}\mathbf{s}_t$ over such directions, and the bound at $|P|$ directions serves for $\partial_P\mathbf{s}_t$. At $m = 0$ the derivative is the displacement of Eq. 7 at the action $\mathbf{a}+\mathbf{u}$:

$$\partial_{\{0\}}\mathbf{s}_t = Df^{\,t-1}(\mathbf{s}_1)\,\mathbf{B}(\mathbf{s},\mathbf{a}+\mathbf{u})\,\boldsymbol{v}_0, \qquad \|\partial_{\{0\}}\mathbf{s}_t\| \leq C_A B_1 \bar{\Lambda}^{\,t-1} \leq M_0\bar{\Lambda}^{\,t}, \quad M_0 = C_A B_1. \tag{37}$$

Let Eq. 36 hold at every order at most m, with $m+1 \leq k$. Differentiating $\mathbf{s}_t = f(\mathbf{s}_{t-1})$ along the $m+1$ directions spreads the derivative over the partitions of the positions, each of which arises once, since a new position either opens a block of its own or joins one of the blocks, and removing it from its block recovers the partition it came from:

$$\partial_{\{0,\dots,m\}}\mathbf{s}_t = \mathbf{A}(\mathbf{s}_{t-1})\,\partial_{\{0,\dots,m\}}\mathbf{s}_{t-1} + \boldsymbol{\rho}_{t-1}, \tag{38}$$

$$\boldsymbol{\rho}_{t-1} = \sum_{\{P_1,\dots,P_{p+1}\}} D^{p+1}f(\mathbf{s}_{t-1})\left[\partial_{P_1}\mathbf{s}_{t-1},\dots,\partial_{P_{p+1}}\mathbf{s}_{t-1}\right], \tag{39}$$

the sum running over the partitions of $\{0,\dots,m\}$ into $p+1 \geq 2$ non-empty blocks, one factor per block. The blocks of a term have sizes $n_1+1,\dots,n_{p+1}+1$ with $n_1+\dots+n_{p+1} = m-p$, so each has size at most m and the induction hypothesis bounds every factor, while the chain rule bounds

$\|D^{p+1} f\|$ by C, the closed loop being F composed with $\mathbf{s} \mapsto (\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}))$ and that derivative existing because $p+1 \leq m+1 \leq k$:

$$\left\| D^{p+1} f(\mathbf{s}_{t-1}) \left[\partial_{P_1} \mathbf{s}_{t-1}, \ldots, \partial_{P_{p+1}} \mathbf{s}_{t-1} \right] \right\| \leq C \prod_{l=1}^{p+1} M_{n_l} (t-1)^{n_l} \bar{\Lambda}^{(n_l+1)(t-1)} \tag{40}$$

$$= C\,(t-1)^{m-p}\,\bar{\Lambda}^{(m+1)(t-1)} \tag{41}$$

$$\leq C\,t^{m-1}\,\bar{\Lambda}^{(m+1)(t-1)}, \tag{42}$$

the last step using $p \geq 1$. Adding over the partitions, whose number depends only on m, gives $\|\boldsymbol{\rho}_{t-1}\| \leq C\,t^{m-1} \bar{\Lambda}^{(m+1)(t-1)}$. Unrolling Eq. 38 down to $t=1$ reaches $\partial_{\{0,\ldots,m\}} \mathbf{s}_1$, a derivative of order $m+1$ of F in the action of norm at most H_{m+1}, and a sum over the earlier steps. The separation rate bounds that sum, with $(i+1)^{m-1} \leq t^{m-1}$ and $\bar{\Lambda}^{\,t-1-i} \bar{\Lambda}^{(m+1)i} \leq \bar{\Lambda}^{(m+1)t}$ at its index $i \leq t-1$; collecting the two terms through $\bar{\Lambda} \geq 1$ and $(t-1)t^{m-1} \leq t^m$ gives

$$\|\partial_{\{0,\ldots,m\}} \mathbf{s}_t\| \leq \left\| Df^{\,t-1}(\mathbf{s}_1)\, \partial_{\{0,\ldots,m\}} \mathbf{s}_1 \right\| + \sum_{i=1}^{t-1} \left\| Df^{\,t-1-i}(\mathbf{s}_{i+1})\, \boldsymbol{\rho}_i \right\| \tag{43}$$

$$\leq C_A H_{m+1} \bar{\Lambda}^{\,t-1} + (t-1)\,C\,t^{m-1} \bar{\Lambda}^{(m+1)t} \tag{44}$$

$$\leq M_m\, t^m\, \bar{\Lambda}^{(m+1)t}, \tag{45}$$

which is Eq. 36 at order $m+1$ and closes the induction. □

Proof of Theorem 1. **The terms of the displaced sum.** Fix a state $\mathbf{s}$ and an action $\mathbf{a}$. By Eq. 2, the displaced action collects

$$Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a}+\mathbf{u}) = \sum_{t \geq 0} \gamma^t r_t(\mathbf{u}), \qquad r_0(\mathbf{u}) = r(\mathbf{s}, \mathbf{a}+\mathbf{u}), \quad r_t(\mathbf{u}) = \bar{r}\big(\mathbf{s}_t(\mathbf{u})\big) \;\; (t \geq 1), \tag{46}$$

whose terms are at most $\gamma^t r_{\max}$ in absolute value, so that the series converges at every $\mathbf{u}$ and $|Q^{\mu_{\boldsymbol{\theta}}}| \leq r_{\max}/(1-\gamma)$. The reward along the closed loop is r composed with the same map as f, so it is k times continuously differentiable, as is each r_t, and the chain rule bounds $\|D^q \bar{r}\|$ by C as it bounds $\|D^q f\|$; the G of Section 3.1 is one of these constants. Differentiating $r_t = \bar{r} \circ \mathbf{s}_t$ spreads the derivative over the partitions of the directions as in Eq. 39, one factor per block, and Lemma 2 bounds every factor, the blocks of a term having sizes $m_1+1, \ldots, m_q+1$ with $m_1 + \cdots + m_q = p-q$, each at most $p \leq j \leq k$ and adding to p in the exponent of $\bar{\Lambda}$:

$$D^p r_t(\mathbf{u}) \big[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p \big] = \sum_{\{P_1, \ldots, P_q\}} D^q \bar{r}(\mathbf{s}_t) \big[\partial_{P_1} \mathbf{s}_t, \ldots, \partial_{P_q} \mathbf{s}_t \big], \tag{47}$$

$$\left\| D^p r_t(\mathbf{u}) \right\| \leq C \prod_{l=1}^{q} M_{m_l}\, t^{m_l}\, \bar{\Lambda}^{(m_l+1)t} = C\,t^{p-q}\,\bar{\Lambda}^{p\,t} \tag{48}$$

$$\leq C\,t^{p-1}\,\bar{\Lambda}^{p\,t} \qquad (t \geq 1,\; 1 \leq p \leq j), \tag{49}$$

the last step using $q \geq 1$, while the derivatives of order p of r_0 are at most H_p.

Summing the derivatives. For $p \leq j$ the ratio $\gamma \bar{\Lambda}^{\,p}$ is at most $\gamma \bar{\Lambda}^{\,j} < 1$, so the derivatives of order p of the terms of Eq. 46 are dominated, uniformly in $\mathbf{u}$, by a summable sequence:

$$\left\| D^p r_0(\mathbf{u}) \right\| \leq H_p, \qquad \gamma^t \left\| D^p r_t(\mathbf{u}) \right\| \leq C\,t^{p-1} \big(\gamma \bar{\Lambda}^{\,p} \big)^t \quad (t \geq 1), \tag{50}$$

$$\Sigma_p = H_p + C \sum_{t \geq 1} t^{p-1} \big(\gamma \bar{\Lambda}^{\,p} \big)^t < \infty. \tag{51}$$

Each of these series therefore converges uniformly, by the Weierstrass test, and may be differentiated term by term along the coordinates, where Eq. 49 holds since it holds along arbitrary directions, so that, by an induction on the order,

$$D^p \big[\mathbf{a} \mapsto Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a}) \big] (\mathbf{a}+\mathbf{u}) = \sum_{t \geq 0} \gamma^t D^p r_t(\mathbf{u}), \qquad 1 \leq p \leq j. \tag{52}$$

Each derivative of Eq. 52 is continuous as a uniform limit of continuous functions and has norm at most Σ_p, a bound uniform in the state and the action that depends only on j, γ, $\bar{\Lambda}$, C_A and $H_1, \ldots, H_j$. □

At $p = 2$ the two partitions of two directions bound the term of step t of the series of second derivatives by $C\,t\,(\gamma\bar{\Lambda}^2)^t$, the rate $(\gamma\Lambda^2)^t$ of Section 3.1 with a factor t, when $\Lambda \geq 1$.

B.2 Proof of Theorem 2

Theorem 2 (Hölder exponent)**.** *Under Assumption 1 with $k \geq 1$, if $\gamma\Lambda > 1$ then for every state, every action, every unit direction $\mathbf{e}$ and every h, for a constant c that does not depend on them,*

$$\big|Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a} + h\mathbf{e}) - Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})\big| \leq c\,|h|^{\alpha}, \qquad \alpha = \frac{\log(1/\gamma)}{\log \Lambda} < 1. \tag{53}$$

The constant may be taken to be

$$c = L_r + (2r_{\max})^{1-\alpha}\,(C_\Gamma G)^{\alpha}\Big(\frac{\gamma\Lambda}{\gamma\Lambda - 1} + \frac{1}{1-\gamma}\Big) + \frac{2r_{\max}}{1-\gamma}, \qquad C_\Gamma = \frac{C_A B_1}{\Lambda}. \tag{54}$$

A move of the action changes the action value through the reward collected now and through the distance between the two trajectories that follow, weighted by the discount.

Proposition 2 (Distance bound)**.** Let Assumption 1 hold with $k \geq 1$. For every state $\mathbf{s}$, action $\mathbf{a}$, unit direction $\mathbf{e}$ and $h > 0$, let $\mathbf{s}_t$ and $\mathbf{s}'_t$ be the states at step t of the trajectories started by $\mathbf{a}$ and $\mathbf{a} + h\mathbf{e}$. Then

$$\big|Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a} + h\mathbf{e}) - Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})\big| \leq L_r\,h + G\,S(h), \qquad S(h) = \sum_{t \geq 1} \gamma^t\,\|\mathbf{s}'_t - \mathbf{s}_t\|, \tag{55}$$

a bound that holds trivially when $S(h)$ is infinite.

Proof. By Eq. 2 the two action values are $r(\mathbf{s}, \mathbf{a}) + \sum_{t \geq 1} \gamma^t \bar{r}(\mathbf{s}_t)$ and the same with $\mathbf{a} + h\mathbf{e}$ and $\mathbf{s}'_t$, both series converging because every reward is at most $r_{\max}$ in absolute value. Subtracting them term by term gives Eq. 10,

$$Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a} + h\mathbf{e}) - Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a}) = \big[r(\mathbf{s}, \mathbf{a} + h\mathbf{e}) - r(\mathbf{s}, \mathbf{a})\big] + \sum_{t \geq 1} \gamma^t \big[\bar{r}(\mathbf{s}'_t) - \bar{r}(\mathbf{s}_t)\big]. \tag{56}$$

On the segment $u \mapsto \mathbf{s}_t + u(\mathbf{s}'_t - \mathbf{s}_t)$, $u \in [0, 1]$, the derivative of $\bar{r}$ is $\nabla\bar{r}^{\top}(\mathbf{s}'_t - \mathbf{s}_t)$, at most $G\|\mathbf{s}'_t - \mathbf{s}_t\|$ by the Cauchy–Schwarz inequality, and on the segment $u \mapsto \mathbf{a} + uh\mathbf{e}$ the derivative of $r(\mathbf{s}, \cdot)$ is at most $L_r h$, so the fundamental theorem of calculus gives

$$\big|\bar{r}(\mathbf{s}'_t) - \bar{r}(\mathbf{s}_t)\big| \leq G\,\|\mathbf{s}'_t - \mathbf{s}_t\|, \qquad \big|r(\mathbf{s}, \mathbf{a} + h\mathbf{e}) - r(\mathbf{s}, \mathbf{a})\big| \leq L_r\,h. \tag{57}$$

Bounding each term of Eq. 56 in this way and adding gives Eq. 55. □

Proof of Theorem 2. **Large displacements.** A negative h is a positive move along $-\mathbf{e}$ and $h = 0$ changes nothing, so let $h > 0$. Since $\gamma < 1 < \gamma\Lambda$, the rate Λ is above one, hence $\bar{\Lambda} = \Lambda$, and $0 < \alpha < 1$, so that $h^{\alpha} \geq 1$ for $h \geq 1$ and $h \leq h^{\alpha}$ for h below one. For $h \geq 1$ each action value is at most $r_{\max}/(1-\gamma)$ in absolute value, so their difference is at most $2r_{\max}/(1-\gamma) \leq 2r_{\max}h^{\alpha}/(1-\gamma) \leq c\,h^{\alpha}$; let then $0 < h < 1$.

The distance between the two trajectories. Let $\mathbf{s}_t(u)$ be the trajectory opened by the action $\mathbf{a} + u\mathbf{e}$, which Lemma 2 makes continuously differentiable in u, with $\mathbf{s}_t(0) = \mathbf{s}_t$, $\mathbf{s}_t(h) = \mathbf{s}'_t$ and derivative the displacement of Eq. 7 at the action $\mathbf{a} + u\mathbf{e}$, written $\boldsymbol{\delta}_t(u)$, of norm at most $C_A B_1 \Lambda^{t-1} = C_\Gamma \Lambda^t$ by Eq. 37. The fundamental theorem of calculus applied to each coordinate, and the bound of the norm of an integral by the integral of the norm, give

$$\mathbf{s}'_t - \mathbf{s}_t = \int_0^h \boldsymbol{\delta}_t(u)\,du, \qquad \|\mathbf{s}'_t - \mathbf{s}_t\| \leq C_\Gamma\,h\,\Lambda^t. \tag{58}$$

Each reward is at most $r_{\max}$ in absolute value, so each bracket of Eq. 56 carries the two bounds of Eq. 11,

$$\big|\bar{r}(\mathbf{s}'_t) - \bar{r}(\mathbf{s}_t)\big| \leq \min\big\{C_\Gamma G\,h\,\Lambda^t,\ 2r_{\max}\big\}. \tag{59}$$

If $C_\Gamma G = 0$ or $r_{\max} = 0$ every bracket vanishes and the difference is at most $L_r h \leq c\,h^{\alpha}$, so let both $C_\Gamma G$ and $r_{\max}$ be positive.

Splitting at the crossing time. The two bounds of Eq. 59 cross where $C_\Gamma G\, h\, \Lambda^t = 2r_{\max}$, at the real number

$$T(h) = t_0 + t^*(h), \qquad t_0 = \frac{\log\big(2r_{\max}/(C_\Gamma G)\big)}{\log\Lambda}, \qquad t^*(h) = \frac{\log(1/h)}{\log\Lambda}, \tag{60}$$

where $t^*(h)$ is the separation time of Eq. 12 and t_0, the term that Section 3.2 drops, does not depend on h. Since $\alpha\log\Lambda = \log(1/\gamma)$, every real τ has $\gamma^\tau = \Lambda^{-\alpha\tau}$, so that

$$C_\Gamma G\, h\, \Lambda^{T(h)} = 2r_{\max}, \qquad \gamma^{T(h)} = \Big(\frac{C_\Gamma G}{2r_{\max}}\Big)^\alpha h^\alpha. \tag{61}$$

Split the sum of Eq. 56 at $T = T(h)$, with the first bound of Eq. 59 before T and the second from T on. The early terms sum a geometric progression of ratio $\gamma\Lambda > 1$ over the n integers of $[1, T)$, at most $(\gamma\Lambda)^{n+1}/(\gamma\Lambda - 1)$ with $n < T$, which is then read through Eq. 61, and the late ones a geometric progression of ratio $\gamma < 1$ from an integer not below T,

$$\sum_{1\le t<T} \gamma^t\, C_\Gamma G\, h\, \Lambda^t \le \frac{C_\Gamma G\, h\, (\gamma\Lambda)^{T+1}}{\gamma\Lambda - 1} = \frac{2r_{\max}\,\gamma\Lambda}{\gamma\Lambda - 1}\,\gamma^T, \tag{62}$$

$$\sum_{t\ge\max(T,1)} \gamma^t\, 2r_{\max} \le \frac{2r_{\max}}{1-\gamma}\,\gamma^T. \tag{63}$$

When $T \le 1$ the first sum is empty and its bound holds because the right side is positive. Both bounds are multiples of $2r_{\max}\gamma^T = (2r_{\max})^{1-\alpha}(C_\Gamma G)^\alpha h^\alpha$, and adding the change of the reward collected at $\mathbf{s}$, at most $L_r h \le L_r h^\alpha$, gives Eq. 53 with the constant of Eq. 54, which depends on neither $\mathbf{s}$, nor $\mathbf{a}$, nor $\mathbf{e}$. □

B.3 The thresholds cannot be improved

Take one state variable and one action variable, both real numbers, with

$$F(x,a) = \Lambda\,(x+a), \qquad \mu_{\boldsymbol\theta}(x) = 0, \qquad r(x,a) = \cos x, \tag{64}$$

for a number $\Lambda > 1$.

Proposition 3 (The thresholds are attained)**.** For the system of Eq. 64, Assumption 1 holds for every k and Λ is the smallest separation rate of the closed loop. At the state $x = 0$, if $\gamma\Lambda > 1$, no bound $|Q^{\mu_{\boldsymbol\theta}}(0, a+h) - Q^{\mu_{\boldsymbol\theta}}(0,a)| \le c\,|h|^\beta$ with $\beta > \alpha$ holds at any action, α being the exponent of Eq. 13. At the same state, for every j, the action value is j times differentiable exactly when $\gamma\Lambda^j < 1$.

At $x = 0$ the action value is a Weierstrass function plus a smooth term.

Theorem 6 (18, Theorem 1.31 and §1.33)**.** *Let $0 < \gamma < 1$, let $\Lambda > 1$, let $\alpha = \log(1/\gamma)/\log\Lambda$ and let $W(u) = \sum_{t\ge0}\gamma^t\varphi(\Lambda^t u)$ with φ the cosine or the sine. If $\gamma\Lambda \ge 1$, then W has a finite derivative at no point. If $\gamma\Lambda > 1$, then $W(u+h) - W(u) = O(|h|^\alpha)$ at every u, and $(W(u+h) - W(u))/|h|^\alpha$ tends to zero at no u.*

Hardy writes a, b and πx for γ, Λ and u; the change of variable $u = \pi x$ is linear and affects neither statement.

Proof of Proposition 3. **The system satisfies the assumptions.** The three functions of Eq. 64 are infinitely differentiable with bounded derivatives of every order from the first, so Assumption 1 holds for every k, and $r_{\max} = 1$. The closed loop $f(x) = \Lambda x$ multiplies every displacement by Λ at every step, so Λ is a separation rate with constant $C_A = 1$; a separation rate Λ' with constant C' has

$$\Lambda^n \le C'\Lambda'^n \quad (n \ge 0), \qquad \Lambda \le C'^{1/n}\Lambda', \tag{65}$$

and $C'^{1/n} \to 1$, so that $\Lambda' \ge \Lambda$. From $x = 0$ the action a leads to $x_t = \Lambda^t a$, and Eq. 2 gives

$$Q^{\mu_{\boldsymbol\theta}}(0,a) = 1 + \sum_{t\ge1}\gamma^t\cos\big(\Lambda^t a\big) = 1 - \cos a + W(a), \qquad W(u) = \sum_{t\ge0}\gamma^t\cos\big(\Lambda^t u\big). \tag{66}$$

The second equality removes from $W(a)$ its term at $t = 0$, which is $\cos a$.

Above the first threshold. Let $\gamma\Lambda > 1$. Theorem 2 applies, since Assumption 1 holds and Λ is a separation rate, so the action value changes by at most a multiple of $|h|^\alpha$. Write

$$\rho(a,h) = \frac{Q^{\mu_\theta}(0,a+h) - Q^{\mu_\theta}(0,a)}{|h|^\alpha}. \tag{67}$$

The term $1-\cos a$ of Eq. 66 changes by at most $|h|$, so $\rho(a,h)$ differs from $(W(a+h)-W(a))/|h|^\alpha$ by at most $|h|^{1-\alpha}$, which tends to zero with h because $\gamma\Lambda > 1$ puts α below one; by Theorem 6, $\rho(a,h)$ therefore tends to zero at no action. A bound $c\,|h|^\beta$ with $\beta > \alpha$ at one action would keep $\rho(a,h)$ below $c\,|h|^{\beta-\alpha}$ and would send it to zero, which happens at no action: the exponent of Theorem 2 cannot be raised at this state, at any action.

The higher thresholds. Let now $\gamma\Lambda^j \geq 1$, and let $p \leq j$ be the smallest integer with $\gamma\Lambda^p \geq 1$, which is at least one since $\gamma < 1$, so that $\gamma\Lambda^{p-1} < 1$ by minimality. The derivative of order $m \leq p-1$ of the t-th term of W is bounded by $(\gamma\Lambda^m)^t \leq (\gamma\Lambda^{p-1})^t$, so W can be differentiated term by term $p-1$ times, and

$$W^{(p-1)}(u) = \sum_{t\geq 0} \left(\gamma\Lambda^{p-1}\right)^t \varphi_p\left(\Lambda^t u\right), \qquad \varphi_p \in \{\cos, -\sin, -\cos, \sin\}, \tag{68}$$

a series of the same kind as W, with $\gamma\Lambda^{p-1} < 1$ in the place of γ, a sine or a cosine, and a sign that does not affect differentiability. Since $\gamma\Lambda^{p-1}\cdot\Lambda = \gamma\Lambda^p \geq 1$, Theorem 6 leaves $W^{(p-1)}$ without a finite derivative at any point, so $Q^{\mu_\theta}(0,\cdot)$, which differs from W by the smooth term $1-\cos a$, has no derivative of order p or higher. With Theorem 1, which gives the derivative when $\gamma\Lambda^j < 1$, the j-th derivative at $x = 0$ exists for those j and for no other. □

B.4 Hyperbolic attractors and Assumption 2

The closed loops of Theorem 3 carry a bounded region of states into itself and, inside it, contract some directions and stretch the others. A *trapping region* is a bounded open set U with $\overline{f(U)} \subset U$, so that a trajectory that enters U stays there; the sets $f^t(U)$ shrink onto the *attractor* $\Omega = \bigcap_{t\geq 0} f^t(U)$, a compact set with $f(\Omega) = \Omega$ on which these trajectories accumulate, and the *basin* $\mathcal{B} = \bigcup_{t\geq 0} f^{-t}(U)$ is the open set of states whose trajectory enters U. The attractor is *hyperbolic* when at each of its points the state space splits as $E^s(\mathbf{x}) \oplus E^u(\mathbf{x})$, with $Df(\mathbf{x})E^{s,u}(\mathbf{x}) = E^{s,u}(f\mathbf{x})$ and constants $C_h \geq 1$, $\varrho < 1 < \Lambda_-$ such that, for $t \geq 0$,

$$\|Df^t(\mathbf{x})\mathbf{w}\| \leq C_h\varrho^t\|\mathbf{w}\| \quad (\mathbf{w}\in E^s(\mathbf{x})), \qquad \|Df^t(\mathbf{x})\mathbf{w}\| \geq C_h^{-1}\Lambda_-^t\|\mathbf{w}\| \quad (\mathbf{w}\in E^u(\mathbf{x})) \tag{69}$$

[16]. Among the unstable directions we single out the fastest: the unstable subspace splits as $E^{u,2} \oplus E^{uu}$, both carried along the trajectory, with E^{uu} a line whose stretch $\Lambda_{uu}(\mathbf{x}) = \|Df(\mathbf{x})|_{E^{uu}(\mathbf{x})}\|$ dominates that of $E^{u,2}$,

$$\|Df^n(\mathbf{x})|_{E^{u,2}(\mathbf{x})}\| \leq C_D\vartheta^n\Lambda_{uu}^{(n)}(\mathbf{x}), \qquad \Lambda_{uu}^{(n)}(\mathbf{x}) = \prod_{i<n}\Lambda_{uu}(f^i\mathbf{x}) = \|Df^n(\mathbf{x})|_{E^{uu}(\mathbf{x})}\|,$$
$$\Lambda_{uu}^{(t,T)}(\mathbf{x}) = \frac{\Lambda_{uu}^{(T)}(\mathbf{x})}{\Lambda_{uu}^{(t)}(\mathbf{x})}, \tag{70}$$

with $C_D \geq 1$ and $\vartheta < 1$, the dominated splitting of Bonatti et al. [25, Appendix B] and Crovisier and Potrie [26]. The splitting $E^s \oplus E^{u,2} \oplus E^{uu}$ is dominated in their sense: Eq. 69 gives the domination of E^s by the other two, with constant C_h^2 and ratio ϱ/Λ_-, and Eq. 70 that of $E^{u,2}$ by E^{uu}, with C_D and ϑ.

The strong rate is the average of $\log\Lambda_{uu}$ against the SRB measure ν of Ω, the invariant probability measure that the trajectories of every state of U outside a set of zero volume trace out [17], and it sets the exponent of Theorem 3:

$$\lambda_1 = \int \log\Lambda_{uu}\, d\nu, \qquad \alpha_1 = \frac{\log(1/\gamma)}{\lambda_1}. \tag{71}$$

Almost every state means every state outside a set of zero volume in the state space, never outside a set of ν-measure zero.

Assumption 2 (full statement). (H1) *There is a trapping region U on which f is twice continuously differentiable and one-to-one, with an invertible Jacobian; Ω is hyperbolic with $E^u \neq \{0\}$, and f is transitive on Ω, some trajectory in Ω being dense in it; the Jacobian of f is invertible at almost every state of $\mathcal{B}$.*

(D) *On Ω, $E^u = E^{u,2} \oplus E^{uu}$ with $\dim E^{uu} = 1$, $E^{u,2} = \{0\}$ being allowed, and Eq. 70 holds.*

(H2$_\mathcal{B}$) *The conditional measures of ν on the local unstable leaves are equivalent to the volume of the leaves, and the holonomy along the local stable leaves between two transversals is absolutely continuous.*

(E) *The loop stretches the strong direction by at least $\Lambda_{\min} > 1$ at every point of Ω, that is $\Lambda_{uu} \geq \Lambda_{\min}$, so that $\lambda_1 \geq \log \Lambda_{\min}$.*

A uniformly hyperbolic attractor with its SRB measure is the setting of Bowen and Ruelle [27], Bowen [28] and Ruelle [17], and the dominated splitting (D) is the standard notion of partially hyperbolic dynamics [25]. On an Axiom-A attractor the conditional measures of the SRB measure carry a positive Hölder density on the unstable leaves, by the construction of Bowen and Ruelle [27] and Ruelle [17], and the stable holonomy is absolutely continuous [29, 30]; (H2$_\mathcal{B}$) asks these two properties of the map of (H1). The results quoted here are proved for such diffeomorphisms and use the map only on a neighbourhood of Ω, and under (H1) f is a twice differentiable embedding of U onto the open set $f(U)$.

Theorem 7 (Hyperbolic attractors)**.** *Under (H1) and (D) the following hold.*

(i) For ϵ small and $\mathbf{x} \in \Omega$, the local stable leaf $W^s_\epsilon(\mathbf{x}) = \{\mathbf{y} : \ d(f^t\mathbf{y}, f^t\mathbf{x}) \leq \epsilon \ \forall t \geq 0\}$ and the local unstable leaf $W^u_\epsilon(\mathbf{x})$, the states with a backward trajectory within ϵ of that of $\mathbf{x}$, are embedded discs tangent to $E^s(\mathbf{x})$ and $E^u(\mathbf{x})$ that depend continuously on $\mathbf{x}$; there are $C_1 \geq 1$ and $\varrho_1 < 1$ with $d(f^t\mathbf{y}, f^t\mathbf{x}) \leq C_1\varrho_1^t d(\mathbf{y}, \mathbf{x})$ for $\mathbf{y} \in W^s_\epsilon(\mathbf{x})$, and $W^u_\epsilon(\mathbf{x}) \subset \Omega$ [16, Theorems 6.2.8 and 6.4.9], [17, Theorem 1.1 and Proposition 1.2(a)].

(ii) For nearby $\mathbf{x}, \mathbf{y} \in \Omega$, $W^s_\epsilon(\mathbf{x})$ and $W^u_\epsilon(\mathbf{y})$ meet in exactly one point, which lies in Ω and depends continuously on the pair [16, Propositions 6.4.13 and 6.4.21], [17, Proposition 1.2(d)]; for $\mathbf{z} \in \Omega$ the box $[W^u_\epsilon(\mathbf{z}), W^s_\epsilon(\mathbf{z})]$ of these intersections therefore contains a neighbourhood of $\mathbf{z}$ in the state space, covered by the local stable leaves of the points of $W^u_\epsilon(\mathbf{z})$.

(iii) Every state of $\mathcal{B}$ lies on the stable leaf $W^s(\mathbf{z}) = \bigcup_{n\geq 0} f^{-n}(W^s_\epsilon(f^n\mathbf{z}))$ of some $\mathbf{z} \in \Omega$ [28, Proposition 3.10].

(iv) Through each $\boldsymbol{q} \in \Omega$ passes a continuously differentiable curve $W^{uu}(\boldsymbol{q})$, tangent to E^{uu} at each of its points and inside the unstable leaf $\bigcup_{n\geq 0} f^n(W^u_\epsilon(f^{-n}\boldsymbol{q}))$ of $\boldsymbol{q}$, with $f(W^{uu}(\boldsymbol{q})) = W^{uu}(f\boldsymbol{q})$, whose local pieces depend continuously on $\boldsymbol{q}$ in the C^1 topology; two such curves are disjoint or equal [26, Theorem 4.3 and Remarks 4.2 and 4.4].

(v) The fields E^s and E^u are Hölder on Ω [16, Theorem 19.1.6]; $E^s \oplus E^{u,2}$ and E^{uu} are Hölder by Crovisier and Potrie [26, Theorem 4.11] applied to the splittings of Eq. 70 of f and of f^{-1}, and so is $\log \Lambda_{uu}$, a Lipschitz function of Df and E^{uu} bounded below by $\log(C_h^{-1}\Lambda_-)$.

(vi) The SRB measure ν exists, is ergodic and has full support on Ω; Ω is the union of finitely many disjoint closed pieces permuted cyclically by f, each containing the unstable leaves of its points; on each piece a power of f with the normalised restriction of ν is isomorphic to a Bernoulli shift, hence mixing [17, Theorem 1.5 and Remark 1.6(a′)]. That remark asks the periodic points to be dense as well, which (H1) gives: since $f(U) \subset U$, the attractor is locally maximal, the periodic points of such a set are dense among the non-wandering points of the map restricted to it [16, Corollary 6.4.20], and every point of Ω is non-wandering.

Under (E) with $\gamma\Lambda_{\min} > 1$ the strong rate satisfies $\lambda_1 \geq \log \Lambda_{\min} > \log(1/\gamma)$, so that the exponent α_1 of Theorem 3 is below one. In the other direction $\Lambda^{(n)}_{uu}(\boldsymbol{q}) \leq \|Df^n(\boldsymbol{q})\| \leq C_A\Lambda^n$, so that taking logarithms, dividing by n and integrating against ν gives $\lambda_1 \leq \log \Lambda$ and $\alpha_1 \geq \alpha$, with equality when the loop stretches the strong direction at exactly the rate Λ at every point of Ω. Along a strong leaf distances are measured by arc length, dist_{uu}; we write $W^{uu}_\delta(\boldsymbol{q})$ for the points of $W^{uu}(\boldsymbol{q})$ within arc length δ of $\boldsymbol{q}$, and

$$\mathrm{osc}^{uu}_\delta(\boldsymbol{q}) = \sup\left\{|V(\mathbf{y}) - V(\boldsymbol{q})| : \ \mathbf{y} \in W^{uu}_\delta(\boldsymbol{q})\right\} \tag{72}$$

for the oscillation of the value along it.

Definition 1 (Non-degenerate reward)**.** The reward is non-degenerate along the strong direction, condition (ND), when V is not Lipschitz along the strong leaves of Ω: for every K and every $\delta_* > 0$ there are $\boldsymbol{q} \in \Omega$ and $\delta \le \delta_*$ with $\operatorname{osc}_\delta^{uu}(\boldsymbol{q}) > K\delta$.

A degenerate reward is one whose discounted sum telescopes, with a remainder that is Lipschitz along the strong leaves. If $\bar r - c = \psi - \gamma\,\psi \circ f$ on Ω for a constant c and a bounded ψ, the sum of Eq. 2 along a trajectory of Ω telescopes to $V = c/(1-\gamma) + \psi$, and V is Lipschitz along the strong leaves when ψ is; conversely, when V is, the Bellman equation $\bar r = V - \gamma V \circ f$ is of this form with $\psi = V$ and $c = 0$.

At a state $\mathbf{x} \in \mathcal{B}$ a move of the action along a unit direction $\mathbf{e}$ moves the next state $\mathbf{x}_0 = f(\mathbf{x})$ along $\mathbf{v}(\mathbf{x}) = \mathbf{B}(\mathbf{x})\mathbf{e}$ with $\mathbf{B}(\mathbf{x}) = \mathbf{B}(\mathbf{x}, \mu_{\boldsymbol{\theta}}(\mathbf{x}))$, and $\mathbf{x}_0$ lies on the stable leaf of some $\mathbf{z} \in \Omega$ (Theorem 7(iii)). Condition (A) is that there is a unit direction $\mathbf{e}$, the same at every state, for which the loop carries this move at the strong rate of the trajectory of $\mathbf{z}$,

$$\liminf_{t\to\infty} \frac{\|Df^t(\mathbf{x}_0)\,\mathbf{v}(\mathbf{x})\|}{\Lambda_{uu}^{(t)}(\mathbf{z})} > 0 \qquad \text{for almost every } \mathbf{x} \in \mathcal{B}. \tag{73}$$

The condition does not depend on the choice of $\mathbf{z}$: two such points have trajectories that approach each other exponentially, and since $\log \Lambda_{uu}$ is Hölder the ratio of their products $\Lambda_{uu}^{(t)}$ stays between two positive constants. At a state of Ω, where $\mathbf{z} = \mathbf{x}_0$, (A) asks that $\mathbf{v}(\mathbf{x})$ have a component along $E^{uu}(\mathbf{x}_0)$, since the components along $E^{u,2}$ and E^s grow more slowly by Eqs. 69 and 70.

Assumption 2 is (H1), (D), (H2$_{\mathcal{B}}$) and (E) together with (ND) and (A).

B.5 Proof of Theorem 3

Theorem 3 (No gradient at almost every state)**.** *Let Assumption 1 hold with $k \ge 1$ and Assumption 2 hold, with $\gamma\Lambda_{\min} > 1$ and $\mathbf{e}$ the direction of (A). Then at almost every state $\mathbf{x}$ of the basin the ratio*

$$R_\beta(h) = \frac{\left|Q^{\mu_{\boldsymbol{\theta}}}\left(\mathbf{x}, \mu_{\boldsymbol{\theta}}(\mathbf{x}) + h\mathbf{e}\right) - Q^{\mu_{\boldsymbol{\theta}}}\left(\mathbf{x}, \mu_{\boldsymbol{\theta}}(\mathbf{x})\right)\right|}{|h|^\beta}$$

tends to zero as $h \to 0$ for every $\beta < \alpha_1$, and $\limsup_{h\to 0} R_\beta(h) = +\infty$ for every $\beta > \alpha_1$.

The proof follows two nearby states backwards along the loop: their predecessors are closer by the whole stretch the fastest direction has accumulated, while the difference between their values is only discounted, so that when the stretch outruns the discount the difference quotient is unbounded.

Replacing f by an iterate f^n, γ by γ^n and $\bar r$ by the reward $\sum_{i<n} \gamma^i\, \bar r \circ f^i$ collected over n steps leaves the value V, the leaves of Theorem 7 and the exponent α_1 unchanged, since $\log(1/\gamma)$ and λ_1 are both multiplied by n, and Eq. 73 read at the multiples of n still holds. We take n large and a multiple of the number of pieces of Theorem 7(vi), which the loop permutes cyclically, so that the iterate carries each piece into itself. For such an n the iterate satisfies, at every point of Ω, by Eqs. 69 and 70,

$$\|Df|_{E^s}\| \le \rho < 1, \qquad \|Df^{-1}|_{E^u}\| \le \Lambda_u^{-1} < 1, \qquad \|Df|_{E^{u,2}}\| \le \theta_D\,\Lambda_{uu}, \quad \theta_D < 1, \tag{74}$$

and $\Lambda_{uu} \ge \Lambda_{\min}^n$ with $\gamma^n \Lambda_{\min}^n > 1$. We work with the iterate and keep the names f, γ, $\bar r$ and $\Lambda_{\min}$, the last of them now the rate of the iterate, that is the $\Lambda_{\min}$ of (E) raised to the power n, so that in these names $\Lambda_{uu} \ge \Lambda_{\min}$ and $\gamma\Lambda_{\min} > 1$. The action value keeps the original discount γ_0, $Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{a}) = r(\mathbf{x}, \mathbf{a}) + \gamma_0 V(F(\mathbf{x}, \mathbf{a}))$, and $\mathbf{x}_0 = F(\mathbf{x}, \mu_{\boldsymbol{\theta}}(\mathbf{x}))$ is the next state of the original loop. We write L for a Lipschitz constant of $\bar r$ on U, Λ^+ for a bound on $\|Df\|$ there, $\lambda_{uu} = n\lambda_1$ for the strong rate of the iterate, and $\Theta = \sum_{j\ge 1} (\gamma\Lambda_{\min})^{-j}$.

Let $\eta_V \in (0, 1]$ satisfy $\gamma(\Lambda^+)^{\eta_V} < 1$. For $\boldsymbol{q}, \boldsymbol{q}' \in U$, and for $\boldsymbol{q}$ on the local stable leaf of $\boldsymbol{p} \in \Omega$,

$$|V(\boldsymbol{q}) - V(\boldsymbol{q}')| \le \sum_{t\ge 0} \gamma^t L \min\left(2\operatorname{diam} U, (\Lambda^+)^t d(\boldsymbol{q}, \boldsymbol{q}')\right) \le C_V\, d(\boldsymbol{q}, \boldsymbol{q}')^{\eta_V}, \tag{75}$$

$$|V(\boldsymbol{q}) - V(\boldsymbol{p})| \le \sum_{t\ge 0} \gamma^t L C_1 \varrho_1^t\, d(\boldsymbol{q}, \boldsymbol{p}) = L_s\, d(\boldsymbol{q}, \boldsymbol{p}). \tag{76}$$

Eq. 75 splits each term with $\min(a,b) \le a^{1-\eta_V} b^{\eta_V}$; Eq. 76 is Theorem 7(i).

By Theorem 7(vi), ν is the average of finitely many measures ν_i carried by closed pieces Ω_i that contain the unstable leaves of their points; the original loop carries each piece onto the next, and the mixing of the power of f on each piece makes each ν_i ergodic for the iterate. Along the original loop the value at two points of a strong leaf obeys

$$V(\mathbf{y}) - V(\mathbf{w}) = \bar{r}_0(\mathbf{y}) - \bar{r}_0(\mathbf{w}) + \gamma_0\big[V(f_0\mathbf{y}) - V(f_0\mathbf{w})\big], \tag{77}$$

with f_0 and $\bar{r}_0$ the original loop and reward. So if V were Lipschitz along the strong leaves of one piece it would be so on the preceding one, and hence on all of Ω; under Definition 1 it is Lipschitz along the strong leaves of no piece. The mean of $\log \Lambda_{uu}$ against ν_i is λ_{uu} for every i, since the original loop carries ν_i to ν_{i+1}.

Since $\log \Lambda_{uu}$ and the fields of Theorem 7(v) are Hölder with a common exponent η', the first with constant H, and the iterate contracts the unstable leaves backwards by Λ_u, two points $\mathbf{y}, \boldsymbol{q}$ of a local unstable leaf at leaf distance at most δ_0 have

$$C_d^{-1} \le \frac{\Lambda_{uu}^{(m)}(f^{-m}\mathbf{y})}{\Lambda_{uu}^{(m)}(f^{-m}\boldsymbol{q})} \le C_d, \qquad \log C_d = \frac{H\delta_0^{\eta'}}{\Lambda_u^{\eta'} - 1} \qquad (m \ge 0), \tag{78}$$

and when $\mathbf{y}$ lies on the strong leaf of $\boldsymbol{q}$, $\mathrm{dist}_{uu}(f^{-m}\mathbf{y}, f^{-m}\boldsymbol{q}) \le C_d\, \mathrm{dist}_{uu}(\mathbf{y},\boldsymbol{q})/\Lambda_{uu}^{(m)}(f^{-m}\boldsymbol{q})$, since f^{-m} shrinks the tangent of the strong leaf by $1/\Lambda_{uu}^{(m)}$ at each of its points.

Lemma 3 (The strong leaf is determined by the rate)**.** *There are $\epsilon_2 > 0$ and $\delta_1 > 0$ such that, for every $C_0 > 0$, if $\boldsymbol{q}, \mathbf{y} \in \Omega$ and*

$$d\big(f^{-m}\mathbf{y}, f^{-m}\boldsymbol{q}\big) \le \min\Big(\epsilon_2,\ \frac{C_0}{\Lambda_{uu}^{(m)}(f^{-m}\boldsymbol{q})}\Big) \qquad (m \ge 0), \tag{79}$$

then $\mathbf{y} \in W_{\delta_1}^{uu}(\boldsymbol{q})$, on the strong leaf of $\boldsymbol{q}$ within arc length δ_1.

Proof. By the first bound of Eq. 79, $\mathbf{y}$ lies on the local unstable leaf of $\boldsymbol{q}$, and each backward image $\mathbf{y}_j = f^{-j}\mathbf{y}$ on that of $\boldsymbol{q}_j = f^{-j}\boldsymbol{q}$, at leaf distance at most a multiple of $\epsilon_2\Lambda_u^{-j}$. Read in coordinates on this leaf along $E^{u,2}(\boldsymbol{q}_j) \oplus E^{uu}(\boldsymbol{q}_j)$, the local strong leaf of $\boldsymbol{q}_j$ is the graph of a continuously differentiable map from the strong axis to $E^{u,2}(\boldsymbol{q}_j)$ of slope at most $\frac{1}{2}$, by Theorem 7(iv); let T_j be the distance of $\mathbf{y}_j$ from that graph along $E^{u,2}$. One step of f then gives

$$T_{j-1} \le \theta_D\, \Lambda_{uu}(\boldsymbol{q}_j)\,(1+\omega_j)\,T_j, \tag{80}$$

$$T_0 \le \theta_D^m\, \Lambda_{uu}^{(m)}(\boldsymbol{q}_m) \prod_{j=1}^{m}(1+\omega_j)\,T_m \tag{81}$$

$$\le C\,C_0\,\theta_D^m \prod_{j=1}^{m}(1+\omega_j) \xrightarrow[m\to\infty]{} 0. \tag{82}$$

In Eq. 80 the map f, read in these coordinates, carries the strong leaf of $\boldsymbol{q}_j$ into that of $\boldsymbol{q}_{j-1}$, and its derivative at the origin respects the splitting and multiplies $E^{u,2}$ by at most $\theta_D\Lambda_{uu}(\boldsymbol{q}_j)$, by Eq. 74; that derivative is equicontinuous in the point, so the mean value theorem gives the factor $1+\omega_j$, with $\omega_j \to 0$ as the leaf distance of $\mathbf{y}_j$ from $\boldsymbol{q}_j$ tends to zero. Eq. 81 iterates from $j = m$ and uses $\prod_{j=1}^{m}\Lambda_{uu}(\boldsymbol{q}_j) = \Lambda_{uu}^{(m)}(\boldsymbol{q}_m)$. Eq. 82 bounds T_m by a multiple of $d(\mathbf{y}_m, \boldsymbol{q}_m) \le C_0/\Lambda_{uu}^{(m)}(\boldsymbol{q}_m)$, and the limit holds because $\frac{1}{m}\sum_{j\le m}\log(1+\omega_j) \to 0$. Hence $T_0 = 0$, and $\mathbf{y}$ lies on the graph, that is on the strong leaf of $\boldsymbol{q}$. □

Near Ω we extend the three fields to Hölder fields $\hat{E}^s$, $\hat{E}^{u,2}$, $\hat{E}^{uu}$ on a neighbourhood U' of Ω, with the same exponent, by McShane's extension theorem [31] in local charts and a partition of unity, with projections $\hat{\Pi}^s$, $\hat{\Pi}^{u,2}$, $\hat{\Pi}^{uu}$ of bounded norm, and for $a > 0$ we write $\hat{\mathcal{C}}_a(\boldsymbol{q}) = \{\mathbf{w} : |\hat{\Pi}^s\mathbf{w}| + |\hat{\Pi}^{u,2}\mathbf{w}| \le a\,|\hat{\Pi}^{uu}\mathbf{w}|\}$ for the strong cone of opening a. On Ω, Df respects the splitting and shrinks the other components against the strong one by $\max(\theta_D, \rho/\Lambda_{\min}) < 1$; at a point $\boldsymbol{q} \in U'$, $Df(\boldsymbol{q})$ and the fields differ from their values at the nearest point of Ω by at most a multiple of $d(\boldsymbol{q},\Omega)^{\eta'}$. Hence there are $\theta < 1$, $a_1 > 0$ and C_6 such that for $\boldsymbol{q} \in U'$ close to Ω and $a \le a_1$

$$Df(\boldsymbol{q})\,\hat{\mathcal{C}}_a(\boldsymbol{q}) \subset \hat{\mathcal{C}}_{\theta a + C_6 d(\boldsymbol{q},\Omega)^{\eta'}}(f\boldsymbol{q}). \tag{83}$$

Lemma 4 (A curve in the strong cone sweeps the strong leaf)**.** *There is $\delta_3 > 0$ such that for every $\delta \le \delta_3$ and $\epsilon > 0$ there is $a_* > 0$ with the following property. Let $\boldsymbol{q} \in \Omega$ and let $\Gamma : [-2\delta, 2\delta] \to U'$ be a curve parametrised by arc length with $d(\Gamma(0), \boldsymbol{q}) \le a_*$, $\Gamma'(\sigma) \in \hat{\mathcal{C}}_{a_*}(\Gamma(\sigma))$ and $d(\Gamma(\sigma), \Omega) \le a_*$ for every σ. Then every point of $W^{uu}_\delta(\boldsymbol{q})$ lies within ϵ of Γ.*

Proof. If not, there are $a_n \to 0$, points $\boldsymbol{q}_n \in \Omega$, curves Γ_n as in the statement and $\mathbf{y}_n \in W^{uu}_\delta(\boldsymbol{q}_n)$ at distance at least ϵ from Γ_n. Along a subsequence $\boldsymbol{q}_n \to \boldsymbol{q}$ and $\mathbf{y}_n \to \mathbf{y} \in W^{uu}_\delta(\boldsymbol{q})$, by the continuity of the strong leaves, and the curves, being 1-Lipschitz, converge uniformly to a curve Γ in Ω with $\Gamma(0) = \boldsymbol{q}$, by the Arzelà–Ascoli theorem. Oriented so that their tangents have a positive component along a continuous unit field of $\hat{E}^{uu}$ near $\boldsymbol{q}$, the curves have tangents converging weakly-$*$ in L^∞ to Γ', and since that component tends to one while $\|\Gamma'\| \le 1$, $\Gamma'(\sigma)$ is the unit vector of E^{uu} at $\Gamma(\sigma)$ at all but a set of parameters of zero length. The backward images of such a curve shrink at the strong rate, as in Eq. 78, so that $d(f^{-m}\Gamma(\sigma), f^{-m}\boldsymbol{q}) \le C_d|\sigma|/\Lambda^{(m)}_{uu}(f^{-m}\boldsymbol{q})$ for every m; with δ_3 small enough that $2\delta_3 C_d \le \epsilon_2$ and $2\delta_3 \le \delta_1$, Lemma 3 places $\Gamma(\sigma)$ on the strong leaf of $\boldsymbol{q}$, and Γ is the arc-length parametrisation of $W^{uu}(\boldsymbol{q})$ on $[-2\delta, 2\delta]$. So $\mathbf{y} = \Gamma(\sigma_\infty)$ for some σ_∞, and $d(\mathbf{y}_n, \Gamma_n(\sigma_\infty)) \to 0$, a contradiction. □

A move of the action at a state $\mathbf{x} \in \mathcal{B}$ moves the next state along the curve $c(h) = F(\mathbf{x}, \mu_{\boldsymbol{\theta}}(\mathbf{x}) + h\mathbf{e})$, whose images $\mathcal{C}_t = f^t \circ c$ we follow along the trajectory of a point $\mathbf{z} \in \Omega$ with $\mathbf{x}_0 \in W^s(\mathbf{z})$; $\ell_t(h)$ is the length of $\mathcal{C}_t$ between the parameters 0 and h, and $\mathbf{z}_t = f^t\mathbf{z}$.

Lemma 5 (The actuated curve grows at the strong rate)**.** *Let $\epsilon \le \epsilon_{\max}$, a radius fixed once with the local leaves of Theorem 7(i), and let $\mathbf{x}_0 \in W^s_\epsilon(\mathbf{z})$. There are $\delta_4 > 0$ and C_5, which depend on neither $\mathbf{x}$ nor ϵ, and n_1, A'_x, $h_* > 0$, which depend on $\mathbf{x}$, such that:*

- *(a) for $|h| \le h_*$ and $t \ge n_1$ with $A'_x|h|\Lambda^{(t)}_{uu}(\mathbf{z}) \le \delta_4$, $\ell_t(h) \le A'_x|h|\Lambda^{(t)}_{uu}(\mathbf{z})$;*
- *(b) if Eq. 73 holds at $\mathbf{x}$, there are n_2, T_0 and $a_x \le A_x \le C_5 a_x$ such that for $T \ge T_0$ and $|h| \le h_*$ with $A_x|h|\Lambda^{(T)}_{uu}(\mathbf{z}) \le \delta_4$ one has $a_x|h|\Lambda^{(T)}_{uu}(\mathbf{z}) \le \ell_T(h)$ and $\ell_t(h) \le A_x|h|\Lambda^{(t)}_{uu}(\mathbf{z})$ for $n_2 \le t \le T$; the tangent of $\mathcal{C}_T$ lies in $\hat{\mathcal{C}}_{a_T}$ and $d(\mathcal{C}_T(\sigma), \Omega) \le d_T$ for $|\sigma| \le h_*$, with a_T and d_T independent of h and tending to zero.*

Proof. **Block domination in the adapted chart.** In the chart at $\mathbf{z}_t$ adapted to $E^{uu}(\mathbf{z}_t) \oplus E^{u,2}(\mathbf{z}_t) \oplus E^s(\mathbf{z}_t)$, with the maximum of the three norms, the matrix of $Df(\mathbf{z}_t)$ is block diagonal, $\mathrm{diag}(J_t, X_t, Y_t)$ with J_t a number comparable to $\Lambda_{uu}(\mathbf{z}_t)$; $J^{(t,t')} = J_{t'-1}\cdots J_t$, and likewise $X^{(t,t')}$ and $Y^{(t,t')}$. Once N is large, Eq. 74 makes the strong block dominate over a block of N steps:

$$\max\big(\|X^{(t,t+N)}\|, \|Y^{(t,t+N)}\|\big) \le \tfrac{1}{8}\big|J^{(t,t+N)}\big|. \tag{84}$$

By (H1) Df is Lipschitz near Ω, so at a point within r of $\mathbf{z}_t$ its matrix differs from that of $Df(\mathbf{z}_t)$ by at most a multiple of r, and along a trajectory that stays within r of that of $\mathbf{z}$ the product over a block is within $C\,r$ of the block diagonal one. For $\mathbf{u} = (\xi, \boldsymbol{\eta}, \boldsymbol{\zeta})$ of slope $\mathrm{sl}(\mathbf{u}) = \max(\|\boldsymbol{\eta}\|, \|\boldsymbol{\zeta}\|)/|\xi| \le 1$ and its image $\mathbf{u}' = (\xi', \boldsymbol{\eta}', \boldsymbol{\zeta}')$ under the product over the block, this gives, for r below a uniform radius r^*,

$$\mathrm{sl}(\mathbf{u}') \le \tfrac{1}{4}\mathrm{sl}(\mathbf{u}) + C\,r, \qquad |\xi'| = |J^{(t,t+N)}|\,|\xi|\,(1 \pm C\,r), \tag{85}$$

while a vector with slope above 1 has its non-strong part multiplied by at most $\frac{1}{4}|J^{(t,t+N)}|$.

The curve stays near the orbit. The distance of a point of the curve from the trajectory of $\mathbf{z}$ is at most the distance the trajectory of $\mathbf{x}_0$ keeps from that trajectory, plus the length of the curve, and while the bound of (a) holds that length decays backwards from the last admissible time T:

$$d\big(\mathcal{C}_t(\sigma), \mathbf{z}_t\big) \le C_1\varrho_1^t\epsilon + \ell_t(h) \quad (\sigma \text{ between 0 and } h), \qquad \ell_t(h) \le \delta_4\,\Lambda_{\min}^{-(T-t)}. \tag{86}$$

The radii of the blocks are therefore summable, forward from n_1 and backward from T, and the factors $1 \pm Cr$ multiply to numbers between two uniform constants D_- and D_+. The block diagonal product has norm $|J^{(t,t+N)}|$ in the maximum norm, so the tangent of the curve grows over the blocks by at most $D_+|J^{(n_1,t)}|$, which is (a), by induction on t with n_1 the time after which the trajectory of $\mathbf{x}_0$ is within $r^*/2$ of that of $\mathbf{z}$.

The slope enters the strong cone. For (b), suppose the slope of $\mathbf{u}_t = Df^t(\mathbf{x}_0)\mathbf{v}$ stayed above 1 at every block start. Then

$$\|\mathbf{u}_{t+N}\| \le \tfrac{1}{4}\big|J^{(t,t+N)}\big|\,\|\mathbf{u}_t\|, \tag{87}$$

$$\|\mathbf{u}_t\|/\Lambda_{uu}^{(t)}(\mathbf{z}) \xrightarrow[t\to\infty]{} 0 \quad \text{along the block starts,} \tag{88}$$

against Eq. 73. So the slope is at most 1 at some block start and, by Eq. 85, at most $\frac{1}{2}$ at the next one, a time n_2 that depends on $\mathbf{x}$. By continuity in h, for $|h| \le h_*$ the vector $Df^{n_2}(c(h))c'(h)$ has slope at most 1 and strong coordinate $\xi_{n_2}(h)$ between c_v and $4c_v$, with $c_v > 0$, and from n_2 on

$$\big|\xi_{t+N}(h)\big| = \big|J^{(t,t+N)}\big|\,\big|\xi_t(h)\big|\,(1 \pm C\,r), \tag{89}$$

$$a_x\,\Lambda_{uu}^{(t)}(\mathbf{z}) \;\le\; \big\|\partial_h \mathcal{C}_t(h)\big\| \le\; A_x\,\Lambda_{uu}^{(t)}(\mathbf{z}), \qquad \frac{A_x}{a_x} \le C_5. \tag{90}$$

In Eq. 87 a vector of slope above one has its norm equal to its non-strong part, which Eq. 84 multiplies by at most $\frac{1}{4}|J^{(t,t+N)}|$, and Eq. 88 follows because $|J^{(t,t+N)}|$ is comparable to $\Lambda_{uu}^{(t,t+N)}(\mathbf{z})$. Eq. 89 is Eq. 85 read on the strong coordinate, the slope staying at most 1 from n_2 on; Eq. 90 collects it over the blocks with the radii of Eq. 86, the ratio A_x/a_x staying below a uniform C_5 because c_v cancels. Integrating Eq. 90 over the parameter gives the bounds on the lengths. After a uniform time the curve lies on the stable leaves of points of Ω, by Theorem 7(iii) and the product structure, so $d(\mathcal{C}_t(\sigma), \Omega) \le d_t \to 0$ geometrically.

The opening of the cone. The slope $\frac{1}{2}$ in the chart at $\mathbf{z}_t$ puts the tangent in a strong cone of uniform opening a_{T_0} once the curve is close to Ω, and Eq. 83 iterated from there gives

$$a_T \le \theta^{\,T-T_0} a_{T_0} + C_6 \sum_{T_0 \le t < T} \theta^{\,T-1-t} d_t^{\eta'} \;\xrightarrow[T\to\infty]{}\; 0. \tag{91}$$

On Ω the cone is contracted at every opening, so Eq. 83 may be read with a_1 prescribed, at the price of a smaller neighbourhood and a C_6 that depends on it; we take $a_1 = 2$. The opening a_{T_0} is at most twice the slope, hence at most 1, and T_0 is taken so large that $C_6 \sum_{t \ge T_0} d_t^{\eta'} \le 1$, which the geometric decay of d_t allows. Every a_T is then at most a_1. □

Lemma 6 (Returns to a rough scale)**.** *Under Definition 1, for every K, every $\delta_* > 0$ and every piece Ω_i there is $\delta \le \delta_*$ such that $B_{K,\delta} = \{\boldsymbol{q} \in \Omega_i : \operatorname{osc}_\delta^{uu}(\boldsymbol{q}) > K\delta\}$ is open in Ω_i and $\nu_i(B_{K,\delta}) > 0$. Hence there is an f-invariant Borel set $\mathcal{G} \subset \Omega$ with $\nu(\mathcal{G}) = 1$ whose points $\mathbf{z}$ satisfy $\frac{1}{T}\log\Lambda_{uu}^{(T)}(\mathbf{z}) \to \lambda_{uu}$ and visit each of these sets, for rational K and δ_*, infinitely often.*

Proof. The oscillation is lower semicontinuous, by the continuity of the strong leaves, so $B_{K,\delta}$ is open. If $\nu_i(B_{K,\delta}) = 0$ for every $\delta \le \delta_*$, then on a set Z of full ν_i-measure $\operatorname{osc}_\delta^{uu} \le K\delta$ for every rational $\delta \le \delta_*$, hence $|V(\mathbf{y}) - V(\boldsymbol{q})| \le K\,\mathrm{dist}_{uu}(\mathbf{y}, \boldsymbol{q})$ for $\boldsymbol{q} \in Z$ and $\mathbf{y} \in W_{\delta_*}^{uu}(\boldsymbol{q})$; Z is dense in Ω_i, the support of ν_i, and by the continuity of the strong leaves and of V the bound extends to every point of Ω_i, so that V is Lipschitz along the strong leaves of Ω_i, which is excluded. The second part is Birkhoff's theorem [16] on each piece, for $\log\Lambda_{uu}$ and for the indicators of the countably many sets $B_{K,\delta}$. □

Lemma 7 (From the attractor to the basin)**.** *If $\mathcal{G} \subset \Omega$ is Borel, f-invariant and $\nu(\mathcal{G}) = 1$, then at almost every $\mathbf{x} \in \mathcal{B}$ the next state $\mathbf{x}_0$ lies on the stable leaf of a point of $\mathcal{G}$.*

Proof. For ν-almost every $\mathbf{z}$, $\mathcal{G}$ fills the local unstable leaf of $\mathbf{z}$ up to a set of zero volume for the leaf volume of (H2$_\mathcal{B}$), and by the product structure a neighbourhood of $\mathbf{z}$ is the union of the local stable leaves of the points of that leaf; by the absolute continuity of the stable holonomy and Fubini's theorem in these product coordinates, the stable leaves of the points of $\mathcal{G}$ fill that neighbourhood up to a set of zero volume. Since ν has full support, finitely many such neighbourhoods cover a neighbourhood of Ω, which every trajectory from $\mathcal{B}$ enters; $\mathcal{G}$ is invariant, and the preimage under f of a subset of $\mathcal{B}$ of zero volume has zero volume, since the Jacobian of f is invertible almost everywhere on $\mathcal{B}$ by (H1), so that f is a local diffeomorphism there, whose local inverses carry sets of zero volume to sets of zero volume. □

Proof of Theorem 3. **Choosing the constants.** Let C_5 and δ_4 be the constants of Lemma 5, and set

$$K = \lceil 32 L C_5 \Theta \rceil, \qquad \delta_* = \min\big(\delta_3,\ \delta_4/(4C_5)\big). \tag{92}$$

Lemma 6 with this δ_* gives δ and the set $B_{K,\delta}$ on each piece, and the invariant set $\mathcal{G}$; we then put $\epsilon = \min\big\{\epsilon_{\max}, (K\delta/(4C_V))^{1/\eta_V}\big\}$ and take a_* from Lemma 4. By Lemma 7 and (A), at almost every $\mathbf{x} \in \mathcal{B}$ the next state lies on the stable leaf of some $\mathbf{z} \in \mathcal{G}$ and Eq. 73 holds; we fix such an $\mathbf{x}$. Replacing $\mathbf{x}_0$ and $\mathbf{z}$ by $f^m\mathbf{x}_0$ and $f^m\mathbf{z}$, with $f^m\mathbf{x}_0 \in W^s_\epsilon(f^m\mathbf{z})$, changes $V(c(h)) - V(\mathbf{x}_0)$ by a sum over m steps that is at most a multiple of $|h|$, and Eq. 73 by a positive factor, so we take $\mathbf{x}_0 \in W^s_\epsilon(\mathbf{z})$, with $\mathbf{z}$ in a piece Ω_i and δ, ϵ, a_* and $B = B_{K,\delta}$ the constants fixed for that piece.

Reaching the oscillation along the strong leaf. At a time $T \geq T_0$ with $\mathbf{z}_T \in B$ there is $\mathbf{y} \in W^{uu}_\delta(\mathbf{z}_T)$ with $|V(\mathbf{y}) - V(\mathbf{z}_T)| > K\delta$. By Lemma 5(b), the curve $\mathcal{C}_T$ reparametrised by arc length from $\mathcal{C}_T(0) = f^T\mathbf{x}_0$ has length at least $a_x\delta_4/A_x \geq \delta_4/C_5 \geq 4\delta$ on each side within the parameters $|h| \leq \delta_4/(A_x\Lambda^{(T)}_{uu}(\mathbf{z}))$, and for T large its arc $|\sigma| \leq 2\delta$ satisfies the hypotheses of Lemma 4 with $\boldsymbol{q} = \mathbf{z}_T$, since $d(f^T\mathbf{x}_0, \mathbf{z}_T) \leq C_1\varrho_1^T\epsilon$, $a_T \to 0$ and $d_T \to 0$. So a point of that arc, at arc length $|\sigma_T| \leq 2\delta$ from $\mathcal{C}_T(0)$, lies within ϵ of $\mathbf{y}$; we let h_T be its parameter, for which Lemma 5(b) gives

$$|h_T| \leq \frac{|\sigma_T|}{a_x\Lambda^{(T)}_{uu}(\mathbf{z})} \leq \frac{2\delta}{a_x\Lambda^{(T)}_{uu}(\mathbf{z})} \xrightarrow[T\to\infty]{} 0. \tag{93}$$

The gap at the return time. By Eq. 75, for T large,

$$\big|V(\mathcal{C}_T(h_T)) - V(f^T\mathbf{x}_0)\big| \geq \big|V(\mathbf{y}) - V(\mathbf{z}_T)\big| - C_V\epsilon^{\eta_V} - L_sC_1\varrho_1^T\epsilon \geq \tfrac{5}{8}K\delta, \tag{94}$$

the two subtracted terms being the move from $\mathbf{y}$ to $\mathcal{C}_T(h_T)$, within ϵ, and the move from $\mathbf{z}_T$ to $f^T\mathbf{x}_0$ along the stable leaf.

The discounted head of the telescope. The Bellman equation along the two trajectories gives

$$V(c(h_T)) - V(\mathbf{x}_0) = \sum_{t<T}\gamma^t\big[\bar r(f^t c(h_T)) - \bar r(f^t\mathbf{x}_0)\big] + \gamma^T\big[V(\mathcal{C}_T(h_T)) - V(f^T\mathbf{x}_0)\big], \tag{95}$$

whose head is small against γ^T. Its terms with $t < n_2$ are at most $C(\mathbf{x})|h_T|$, which is below $\frac{1}{16}K\delta\gamma^T$ for T large because $\gamma^T\Lambda^{(T)}_{uu}(\mathbf{z}) \to \infty$; for $n_2 \leq t < T$,

$$\sum_{n_2\leq t<T}\gamma^t L\,\ell_t(h_T) \overset{\text{(i)}}{\leq} LA_x|h_T|\,\gamma^T\Lambda^{(T)}_{uu}(\mathbf{z})\sum_{j\geq 1}(\gamma\Lambda_{\min})^{-j} \overset{\text{(ii)}}{\leq} 2LC_5\Theta\,\delta\,\gamma^T \overset{\text{(iii)}}{\leq} \tfrac{1}{16}K\delta\,\gamma^T,$$

where (i) is Lemma 5(b) with $\Lambda^{(t,T)}_{uu}(\mathbf{z}) \geq \Lambda^{T-t}_{\min}$, (ii) is Eq. 93 with $A_x \leq C_5a_x$, and (iii) is the choice of K. With Eq. 94 this gives $|V(c(h_T)) - V(\mathbf{x}_0)| \geq \frac{1}{2}K\delta\gamma^T$, and since $Q^{\mu_{\boldsymbol\theta}}(\mathbf{x}, \mu_{\boldsymbol\theta}(\mathbf{x}) + h\mathbf{e}) - Q^{\mu_{\boldsymbol\theta}}(\mathbf{x}, \mu_{\boldsymbol\theta}(\mathbf{x})) = [r(\mathbf{x}, \mu_{\boldsymbol\theta}(\mathbf{x}) + h\mathbf{e}) - r(\mathbf{x}, \mu_{\boldsymbol\theta}(\mathbf{x}))] + \gamma_0[V(c(h)) - V(\mathbf{x}_0)]$ with $r(\mathbf{x}, \cdot)$ Lipschitz with constant L_r,

$$R_\beta(h_T) \geq \frac{\gamma_0K\delta}{2}\Big(\frac{a_x}{2\delta}\Big)^\beta\exp\Big(T\Big[\beta\,\tfrac{1}{T}\log\Lambda^{(T)}_{uu}(\mathbf{z}) - \log\tfrac{1}{\gamma}\Big]\Big) - L_r|h_T|^{1-\beta}.$$

The bracket tends to $\beta\lambda_{uu} - \log(1/\gamma)$, which is positive exactly when $\beta > \alpha_1$; the times T with $\mathbf{z}_T \in B$ are infinitely many, and $h_T \to 0$. So $\limsup_{h\to 0} R_\beta(h) = +\infty$ for $\beta \in (\alpha_1, 1]$, and for $\beta > 1$ this follows from $\beta = 1$.

The upper bound. Let $\mathbf{x}$ be such that $\mathbf{x}_0$ lies on the stable leaf of a point $\mathbf{z}$ with $\frac{1}{T}\log\Lambda^{(T)}_{uu}(\mathbf{z}) \to \lambda_{uu}$, which holds at almost every $\mathbf{x}$ by Lemmas 6 and 7, and reduce as in the lower bound to $\mathbf{x}_0 \in W^s_\epsilon(\mathbf{z})$. For $0 < |h| \leq h_*$ let $T(h)$ be the largest T with $A'_x|h|\Lambda^{(T)}_{uu}(\mathbf{z}) \leq \delta_4$, so that, Λ_{uu} being at most Λ^+,

$$\frac{\delta_4}{A'_x\Lambda^+} \leq |h|\,\Lambda^{(T(h))}_{uu}(\mathbf{z}) \leq \frac{\delta_4}{A'_x}. \tag{96}$$

Eq. 95 at $T = T(h)$, with Lemma 5(a) for $n_1 \leq t < T$, a bound $C(\mathbf{x})|h|$ for $t < n_1$ and $|V| \leq r_{\max}/(1-\gamma_0)$ in the last term, gives

$$|V(c(h)) - V(\mathbf{x}_0)| \leq C(\mathbf{x})|h| + LA'_x|h|\,\gamma^T\Lambda^{(T)}_{uu}(\mathbf{z})\,\Theta + \frac{2r_{\max}}{1-\gamma_0}\gamma^T \leq C'(\mathbf{x})\,\gamma^{T(h)},$$

by Eq. 96 and $\gamma^T \Lambda_{uu}^{(T)}(\mathbf{z}) \to \infty$. Eq. 96 and the limit at $\mathbf{z}$ then give

$$\log(1/|h|) = T(h)\big(\lambda_{uu} + o(1)\big), \tag{97}$$

$$\gamma^{T(h)} = |h|^{\alpha_1 + o(1)}, \tag{98}$$

$$R_\beta(h) \le L_r |h|^{1-\beta} + \gamma_0 C'(\mathbf{x})\, |h|^{\alpha_1 - \beta + o(1)} \xrightarrow[h\to 0]{} 0 \qquad (\beta < \alpha_1). \tag{99}$$

Eq. 98 reads Eq. 97 through $\alpha_1 \lambda_{uu} = \log(1/\gamma)$, and the limit in Eq. 99 holds because $\alpha_1 < 1$. □

B.6 The two-copy test

The two-copy test

1. Run the actor from $\mathbf{s}$ once with its own first action $\mu_{\boldsymbol{\theta}}(\mathbf{s})$, and once for each displacement of a ladder $h_1 < \cdots < h_n$ with first action $\mu_{\boldsymbol{\theta}}(\mathbf{s}) + h_i\,\mathbf{e}$, recording at every step the distance of each displaced run from the undisplaced one; $S(h_i)$ is the discounted sum of those distances, and $\alpha(\mathbf{s})$ is the slope of $\log S$ against $\log h$ along the ladder.
2. The copies part at h_i when their distance stops being h_i/h_{i-1} times the distance at h_{i-1}, at the step $t^*(h_i)$, and $D(\mathbf{s})$ is the number of steps by which halving the displacement postpones that step.
3. If $\alpha(\mathbf{s}) < 1$, the agent is beyond the threshold at $\mathbf{s}$ and $Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \cdot)$ inherits the exponent $\alpha(\mathbf{s})$; copies that never part leave the question open, since the test reads only the distances along the runs it makes.

Fix a state $\mathbf{s}$ and a unit direction $\mathbf{e}$ of the action. The undisplaced run is the trajectory $\mathbf{s}_t$ of the actor from $\mathbf{s}$, and the run displaced by $h > 0$ is the trajectory $\mathbf{s}_t(h)$ whose first action is $\mu_{\boldsymbol{\theta}}(\mathbf{s}) + h\mathbf{e}$, the actor being in control from the second step on; the two copies record at every step the distance $d_t(h) = \|\mathbf{s}_t(h) - \mathbf{s}_t\|$. The two runs differ only in the first action, so the separation rate that matters is that of the trajectories that start at $\mathbf{s}$. Let $\mathcal{O}$ be a convex open set of actions that contains the displaced first actions, and let $\Lambda(\mathbf{s}) > 0$ and $C_A(\mathbf{s}) \ge 1$ satisfy

$$\|Df^n(\mathbf{x})\| \le C_A(\mathbf{s})\, \Lambda(\mathbf{s})^n \qquad (n \ge 0) \tag{100}$$

at every state $\mathbf{x}$ that a trajectory from $\mathbf{s}$ with first action in $\mathcal{O}$ reaches at some step $t \ge 1$.

Proposition 4 (Theorems 1 and 2 at one state)**.** Under Assumption 1 with $k \ge 1$, Theorems 1 and 2 hold for $\mathbf{a} \mapsto Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})$ on $\mathcal{O}$, with $\Lambda(\mathbf{s})$ and $C_A(\mathbf{s})$ in place of Λ and C_A in their hypotheses, constants and conclusions, and in particular with the exponent $\alpha(\mathbf{s}) = \log(1/\gamma)/\log \Lambda(\mathbf{s})$.

Proof. The proofs in Appendices B.1 and B.2 use the separation rate only through the norms of products of Jacobians along trajectories that start at $\mathbf{s}$ with a first action in $\mathcal{O}$, taken from their first step on, and Eq. 100 bounds each of them; the actions $\mathbf{a} + u\mathbf{e}$ with u between 0 and h, the only ones Appendix B.2 reads, stay in $\mathcal{O}$ because it is convex. □

Since $h \mapsto \mathbf{s}_t(h)$ is continuously differentiable with derivative the displacement of Eq. 7,

$$\mathbf{s}_t(h) - \mathbf{s}_t = h\,\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e}) + o(h), \qquad d_t(h) = \|\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e})\|\, h + o(h) \qquad (h \to 0), \tag{101}$$

and when $k \ge 2$ the remainder is controlled at every step. Write $\bar{\Lambda} = \max(1, \Lambda(\mathbf{s}))$ and M_1 for the constant of Eq. 36 at $C_A(\mathbf{s})$, fix a ratio $q > 1$ and the smaller displacement $h' = h/q$, and let t be a step at which $\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e}) \ne \mathbf{0}$. Then

$$\omega_t(h) = \frac{h\, M_1\, t\, \bar{\Lambda}^{2t}}{2\, \|\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e})\|}, \tag{102}$$

$$\big|d_t(h) - h\, \|\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e})\|\big| \le \tfrac{1}{2} M_1\, t\, \bar{\Lambda}^{2t} h^2, \tag{103}$$

$$\Big|\frac{d_t(h)}{q\, d_t(h')} - 1\Big| \le 4\, \omega_t(h) \qquad \big(\omega_t(h) \le \tfrac{1}{2}\big). \tag{104}$$

Eq. 103 integrates twice over the moved action the bound $M_1 t \bar{\Lambda}^{2t}$ that Eq. 36 puts on the second derivative of the trajectory, and Eq. 104 divides it by $h\|\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e})\|$ and compares the two runs. While the displacement is small the ratio of the two distances is therefore close to q, and it can leave a band around q only after $\omega_t(h)$ has grown.

The rule at one state. The test reads a ladder of displacements $h_1 < \cdots < h_n$, each h_i against h_{i-1} with $q = h_i/h_{i-1}$, at the readable steps, those where none of the runs at 0, h_{i-1} and h_i has terminated, $d_t(h_{i-1}) > 0$ and the horizon is not reached. The copies at $h = h_i$ part at the first step $\hat{t}(h)$ from which the ratio $d_t(h)/d_t(h_{i-1})$ lies outside $[q(1-\epsilon_{\mathrm{tol}}), q(1+\epsilon_{\mathrm{tol}})]$ at n_{p} consecutive readable steps with both distances at least $d_{\min}$, and at which $d_t(h) > d_1(h)$, so that a distance that has only shrunk does not count as a parting; we use $\epsilon_{\mathrm{tol}} = 0.2$, $n_{\mathrm{p}} = 5$ and $d_{\min} = 10^{-10}$. When there is no such step the state receives at h the first of these readings that applies:

ended together, when both copies terminate at the same step;

no reading, when they terminate at different steps or only one of them does;

no effect, when $d_t(h) = 0$ at every readable step;

not drawing apart, when the distance at the last readable step is at most $d_1(h)$;

still drawing apart, otherwise.

The parting step. At a state where the displacement grows at an exponential rate, $\lambda(\mathbf{s}, \mathbf{e}) = \lim_t \frac{1}{t} \log \|\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e})\| > 0$, the rule is read under the hypothesis that the linear regime ends near the step at which $h\|\boldsymbol{\delta}_t\|$ reaches a fixed size. Let $\hat{t}_\infty(h)$ be the parting step that the rule returns for the pair h/q and h, applied with no horizon and with no termination of the episode.

(L) There are $d_{\mathrm{lin}} > 0$, $h_L \in (0, 1]$ and an integer $K_L \geq 0$, independent of h, such that for $0 < h \leq h_L$ the step $\hat{t}_\infty(h)$ exists and lies within K_L steps of the first step $\tau(h)$ with $h\|\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e})\| \geq d_{\mathrm{lin}}$.

Where the ratio q changes along the ladder, (L) is taken at each of its values, with common constants. It does not follow from $\lambda(\mathbf{s}, \mathbf{e}) > 0$: in the example of Appendix B.3, $d_t(h) = \Lambda^t h$ at every step, the ratio stays at q and the copies never part. The rule uses only the readable steps up to $n_{\mathrm{p}} - 1$ steps later to decide whether a step is the parting step, so the test with horizon T_{hor} returns $\hat{t}(h) = \hat{t}_\infty(h)$ whenever $\hat{t}_\infty(h) + n_{\mathrm{p}} - 1 \leq T_{\mathrm{hor}}$ and none of the runs at 0, h/q and h terminates before that step. When moreover $c_- e^{\lambda(\mathbf{s},\mathbf{e})t} \leq \|\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e})\| \leq c_+ e^{\lambda(\mathbf{s},\mathbf{e})t}$ for $t \geq 1$, $\tau(h)$ differs from $\log(d_{\mathrm{lin}}/h)/\lambda(\mathbf{s}, \mathbf{e})$ by a bounded amount, and in the number of halvings $x = \log_2(1/h)$

$$\hat{t}_\infty(h) = \frac{\log 2}{\lambda(\mathbf{s}, \mathbf{e})}\, x + \varpi(h), \quad |\varpi(h)| \leq K_\varpi, \qquad \frac{\log 2}{\lambda(\mathbf{s}, \mathbf{e})} \geq D(\mathbf{s}), \quad \frac{\log(1/\gamma)}{\lambda(\mathbf{s}, \mathbf{e})} \geq \alpha(\mathbf{s}), \quad (105)$$

the inequalities holding because $\|\boldsymbol{\delta}_t\| \leq C_A(\mathbf{s}) B_1 \Lambda(\mathbf{s})^{t-1}$ gives $\lambda(\mathbf{s}, \mathbf{e}) \leq \log \Lambda(\mathbf{s})$, with equality when the displacement grows by the factor $\Lambda(\mathbf{s})$ per step. The parting step thus grows by $\log 2/\lambda(\mathbf{s}, \mathbf{e})$ steps per halving, the delay $D(\mathbf{s})$ of Eq. 16 when $\lambda(\mathbf{s}, \mathbf{e}) = \log \Lambda(\mathbf{s})$.

Lemma 8. *Under the hypotheses of Theorem 3, with* $\mathbf{e}$ *its direction,* $\lambda(\mathbf{s}, \mathbf{e}) = \lambda_1$ *at almost every state* $\mathbf{s}$ *of the basin.*

Proof. For $t \geq 1$, $\boldsymbol{\delta}_{t+1}(\mathbf{s}, \mathbf{e}) = Df^t(\mathbf{x}_0)\mathbf{v}$ with $\mathbf{x}_0 = f(\mathbf{s})$ and $\mathbf{v} = \mathbf{B}(\mathbf{s})\mathbf{e}$. At almost every $\mathbf{s} \in \mathcal{B}$, $\mathbf{x}_0$ lies on the stable leaf of a point $\mathbf{z}$ at which $\frac{1}{t} \log \Lambda_{uu}^{(t)}(\mathbf{z}) \to \lambda_1$ (Lemma 7 with the Birkhoff points of $\log \Lambda_{uu}$) and Eq. 73 holds. There Lemma 5(a), read along the multiples of the iterate with bounded factors in between, bounds $\|Df^t(\mathbf{x}_0)\mathbf{v}\|$ above by a multiple of $\Lambda_{uu}^{(t)}(\mathbf{z})$, and Eq. 73 bounds it below by a multiple for t large; taking logarithms and dividing by t gives $\lambda(\mathbf{s}, \mathbf{e}) = \lambda_1$. □

Over the states and along the ladder. When several states of an agent are read, the states counted at a displacement are those whose reading is a parting, ended together, no effect, not drawing apart or still drawing apart. The displacement receives the status *parts* when the copies part at more than half of them, with parting step the median of $\hat{t}(h)$ over the counted states, taken to be $+\infty$ where the copies do not part; it receives *never* when the states that part or are still drawing apart are at most half, and *undetermined* otherwise. Going down the ladder from the largest displacement, the test drops a displacement whose status is not parts, and drops one when the displacement below it has status parts and a parting step no later than its own; the first displacement not dropped is kept, together with every displacement below it.

The agent then receives one of four statuses. If h_2, the smallest displacement that is read, does not have status parts, the agent receives that status, never or undetermined. Otherwise the kept

displacements with status parts and a parting step later than the first step give the points (x_i, y_i), with $x_i = \log_2(1/h_i)$ and y_i the median parting step; with fewer than two points the agent is unreadable, and with two or more the measured delay $\hat{D}$ is the slope of the least-squares line through them, a slope $\hat{D} \le 0$ giving the status never and a positive slope the status parts.

The least-squares slope is an average of the difference quotients with nonnegative weights that sum to one,

$$\hat{D} = \sum_{i<j} w_{ij} \, \frac{y_j - y_i}{x_j - x_i}, \qquad w_{ij} = \frac{(x_j - x_i)^2}{\sum_{i'<j'} (x_{j'} - x_{i'})^2}, \tag{106}$$

and if $y_i = a x_i + \varpi_i$ with $|\varpi_i| \le K_\varpi$, then $|\hat{D} - a| \le K_\varpi \sum_i |x_i - \bar{x}| / \sum_i (x_i - \bar{x})^2$, a multiple of K_ϖ fixed by the ladder. When every state follows Eq. 105 with a slope between $a_{\min}$ and $a_{\max}$, the median of the parting steps over the states differs by at most K_ϖ from a function whose difference quotients all lie in that interval, so $\hat{D}$ lies within that multiple of K_ϖ of $[a_{\min}, a_{\max}]$. The uniform stretching condition (U_{uu}) of Appendix D, Eq. 157, turns the limit of Lemma 8 into the two-sided bound $c_- e^{\lambda_1 t} \le \|\boldsymbol{\delta}_t(\mathbf{s}, \mathbf{e})\| \le c_+ e^{\lambda_1 t}$ that Eq. 105 asks for. Under the hypotheses of Theorem 3, (U_{uu}) and (L), the slopes equal $\log 2/\lambda_1$ at almost every state of the basin, and the exponent read from the delay, $\hat{D} \log(1/\gamma)/\log 2$, estimates α_1. The protocol does not pick its states at random over the state space: they are the images $f^t(\mathbf{s}_0)$ at a fixed step of an initial state with a density that vanishes outside $\mathcal{B}$. Such a state falls where Lemma 8 holds with probability one: the set of exceptions has zero volume, and since $\mathcal{B}$ is invariant the preimage of that set under f^t has zero volume as well, by the argument of Lemma 7.

The status never does not place an agent below the threshold: it is given when the slope through the kept points is $\hat{D} \le 0$, or when at h_2 the copies of most counted states end together, have no effect or do not draw apart. The test sees only the distances along the runs from the states it reads, not every product of Jacobians.

The discounted sum of the distances. By Proposition 2, the change of the action value along $\mathbf{e}$ is at most $L_r h + G\, S(h)$ with $S(h) = \sum_{t \ge 1} \gamma^t d_t(h)$, so a bound $S(h) \le K' h^\beta$ gives the exponent $\min(1, \beta)$ to the action value. When $\gamma \Lambda(\mathbf{s}) < 1$, Eq. 58 bounds $d_t(h)$ by $C_\Gamma(\mathbf{s}) h \Lambda(\mathbf{s})^t$, $C_\Gamma(\mathbf{s}) = C_A(\mathbf{s}) B_1 / \Lambda(\mathbf{s})$, at every step for h small, so that

$$S(h) \le \frac{C_\Gamma(\mathbf{s})\, \gamma \Lambda(\mathbf{s})}{1 - \gamma \Lambda(\mathbf{s})}\, h, \tag{107}$$

and the exponent of S is at least one. It is one when $\boldsymbol{\delta}_{\bar{t}}(\mathbf{s}, \mathbf{e}) \ne \mathbf{0}$ at some step $\bar{t}$, since Eq. 101 then gives $d_{\bar{t}}(h) \ge \frac{1}{2} \|\boldsymbol{\delta}_{\bar{t}}(\mathbf{s}, \mathbf{e})\| h$ for h small, hence $S(h) \ge \frac{1}{2} \gamma^{\bar{t}} \|\boldsymbol{\delta}_{\bar{t}}(\mathbf{s}, \mathbf{e})\| h$.

When $\gamma \Lambda(\mathbf{s}) > 1$ and $\mathbf{s}$ lies in a bounded open set U with $\overline{f(U)} \subset U$ whose states have norm at most R_U, both runs stay in U for h small, and splitting $S(h)$ at the separation time $t^*(h)$ of Eq. 12, where $\gamma^{t^*(h)} = h^{\alpha(\mathbf{s})}$, with $d_t(h) \le C_\Gamma(\mathbf{s}) h \Lambda(\mathbf{s})^t$ before it and $d_t(h) \le 2R_U$ after it, gives as in Eqs. 62 and 63

$$S(h) \le \Big(\frac{C_\Gamma(\mathbf{s})\, \gamma \Lambda(\mathbf{s})}{\gamma \Lambda(\mathbf{s}) - 1} + \frac{2R_U}{1 - \gamma} \Big) h^{\alpha(\mathbf{s})}, \tag{108}$$

When an episode ends at step T_{end}, the recorded sum misses at most $2R_U \gamma^{T_{\mathrm{end}}}/(1-\gamma) = 2R_U \gamma^{T_{\mathrm{end}} - t^*(h)} h^{\alpha(\mathbf{s})}/(1-\gamma)$, which is small against $h^{\alpha(\mathbf{s})}$ when the episode lasts many steps beyond the parting.

C Proofs of Section 4

C.1 Proof of Theorem 4

Part (a) turns the noise of the targets into a derivative: the value of $\pi_{\tilde{\sigma}}$ at a state is an average over the actions the noise can take, so that moving the first action moves the window of that average, which the later noises cover and where accessibility enters. The derivative then falls on the density of that window, at a cost of $1/\tilde{\sigma}$, against a variation of the value across it of $\tilde{\sigma}^\alpha$. Part (b) shows that the oscillation of the value along the strong leaves, the one Theorem 3 produces, survives the averaging,

so that the slope is bounded below as well, in mean square on a neighbourhood of the attractor and with the exponent α_1 of that theorem.

The actor $\pi_{\tilde\sigma}$ adds $\tilde\sigma\boldsymbol\zeta$ to its action at every step, and by Eq. 18 its action value and its value satisfy

$$Q^{\pi_{\tilde\sigma}}(\mathbf{s},\mathbf{a}) = r(\mathbf{s},\mathbf{a}) + \gamma\, V^{\pi_{\tilde\sigma}}\big(F(\mathbf{s},\mathbf{a})\big), \qquad V^{\pi_{\tilde\sigma}} = \bar r_{\tilde\sigma} + \gamma\, T_{\tilde\sigma} V^{\pi_{\tilde\sigma}}, \tag{109}$$

with $\bar r_{\tilde\sigma}(\mathbf{x}) = \mathbb{E}[r(\mathbf{x},\mu_{\boldsymbol\theta}(\mathbf{x})+\tilde\sigma\boldsymbol\zeta)]$ and $(T_{\tilde\sigma} g)(\mathbf{x}) = \mathbb{E}[g(f_{\tilde\sigma\boldsymbol\zeta}(\mathbf{x}))]$, where $f_{\mathbf{u}}(\mathbf{x}) = F(\mathbf{x},\mu_{\boldsymbol\theta}(\mathbf{x})+\mathbf{u})$ is the closed loop with its action displaced by $\mathbf{u}$. Conditions on the noise and on the separation rate carry the proof.

(N) The law of $\boldsymbol\zeta$ has zero mean, identity covariance and a density φ supported in $\{\|\boldsymbol\zeta\| \le c_0\}$, with two bounded derivatives in part (a) and one in part (b).

(SR) The separation rate holds along the trajectories whose actions carry offsets: for offsets with $\|\mathbf{u}_j\| \le c_0$,

$$\big\|D(f_{\mathbf{u}_{t-1}} \circ \cdots \circ f_{\mathbf{u}_0})(\mathbf{x})\big\| \le C_A \Lambda^t \qquad (t \ge 1), \tag{110}$$

which for zero offsets is the bound defining Λ in Section 3.1.

The actions of a few steps also have to move the state in every direction. Along the displaced trajectory $\mathbf{y}_0 = f_{\mathbf{u}_{-1}}(\mathbf{s})$, $\mathbf{y}_{j+1} = f_{\mathbf{u}_j}(\mathbf{y}_j)$, with $\mathbf{G}_j = D(f_{\mathbf{u}_{N-1}} \circ \cdots \circ f_{\mathbf{u}_j})(\mathbf{y}_j)$, $\mathbf{G}_N = \mathbf{I}$ and $\mathbf{B}_{\mathbf{u}}(\mathbf{y}) = \partial_{\mathbf{a}} F(\mathbf{y}, \mu_{\boldsymbol\theta}(\mathbf{y}) + \mathbf{u})$, the columns of

$$\mathbf{K}_N(\mathbf{s},\underline{\mathbf{u}}) = \big[\, \mathbf{G}_1\mathbf{B}_{\mathbf{u}_0}(\mathbf{y}_0) \;\big|\; \cdots \;\big|\; \mathbf{G}_N\mathbf{B}_{\mathbf{u}_{N-1}}(\mathbf{y}_{N-1}) \,\big] \in \mathbb{R}^{d_S \times N d_A} \tag{111}$$

are the moves of $\mathbf{y}_N$ produced by the actions of the N steps that follow the first.

Definition 2 (Accessibility in N steps)**.** The action reaches every direction of the state in N steps when there is $c_{\mathrm{acc}} > 0$ such that $\mathbf{K}_N(\mathbf{s},\underline{\mathbf{u}})$ has d_S-th singular value at least c_{acc} at every state $\mathbf{s}$ and for every choice of offsets with $\|\mathbf{u}_j\| \le c_0$.

For constant Jacobians and $N = d_S$ this is Kalman's condition for the controllability of a linear system [32]; it allows $d_A < d_S$, since $\mathbf{K}_N$ collects the actuation of N steps.

Theorem 4 (Size of the slope)**.** *(a) Let Assumption 1 hold with $k \ge 3$, let $\gamma\Lambda > 1$ with (SR), let $\boldsymbol\zeta$ satisfy (N), and let the action reach every direction of the state in N steps. There is C, which depends on Λ, C_A, c_0, N, c_{acc}, the derivatives of φ and those of F, $\mu_{\boldsymbol\theta}$ and r up to order three, such that for every state $\mathbf{s}$, unit direction $\mathbf{e}$ and $\tilde\sigma \in (0,1]$ the two derivatives exist and*

$$\big|\partial_{\mathbf{e}} Q^{\pi_{\tilde\sigma}}\big(\mathbf{s},\mu_{\boldsymbol\theta}(\mathbf{s})\big)\big| \le C\,\tilde\sigma^{\,\alpha-1}, \qquad \big|\partial^2_{\mathbf{e}} Q^{\pi_{\tilde\sigma}}\big(\mathbf{s},\mu_{\boldsymbol\theta}(\mathbf{s})\big)\big| \le C\,\tilde\sigma^{\,\alpha-2}. \tag{112}$$

(b) Let the hypotheses of Theorem 3 hold, with $\mathbf{e}$ the direction of (A) and Assumption 1 for $k \ge 2$ on U, let every displaced loop $f_{\mathbf{u}}$ with $\|\mathbf{u}\| \le c_0$ map $\overline{U}$ into U, let the action reach every direction of the state in N steps from every state $\mathbf{y}$ of U, with $\mathbf{K}_N$ built from $\mathbf{y}_0 = \mathbf{y}$, and let $\boldsymbol\zeta$ satisfy (N). There are an open neighbourhood $\mathcal{U}$ of Ω, $c > 0$ and $\tilde\sigma_0 > 0$ such that for $\tilde\sigma \le \tilde\sigma_0$ the slope exists at every $\mathbf{x} \in \mathcal{U}$ and

$$\Big(\frac{1}{\mathrm{vol}(\mathcal{U})}\int_{\mathcal{U}} \partial_{\mathbf{e}} Q^{\pi_{\tilde\sigma}}\big(\mathbf{x},\mu_{\boldsymbol\theta}(\mathbf{x})\big)^2\, d\mathbf{x}\Big)^{1/2} \ge c\,\tilde\sigma^{\,\alpha_1-1}. \tag{113}$$

The upper bound. The noise of the N steps that follow can reproduce the effect of the first action, as in the argument of Section 4.1.

Lemma 9 (The value of the noisy actor is Hölder, uniformly in the width)**.** *Let $\gamma\Lambda > 1$ and let Eq. 110 hold. For every $\tilde\sigma \in [0,1]$ and all states $\mathbf{x},\mathbf{x}'$, $|V^{\pi_{\tilde\sigma}}(\mathbf{x}') - V^{\pi_{\tilde\sigma}}(\mathbf{x})| \le C_H\|\mathbf{x}'-\mathbf{x}\|^{\alpha}$, with C_H independent of $\tilde\sigma$.*

Proof. Driven by the same noises, the trajectories $\mathbf{y}_t$ from $\mathbf{x}$ and $\mathbf{y}'_t$ from $\mathbf{x}'$ stay within $C_A\Lambda^t\|\mathbf{x}'-\mathbf{x}\|$ of each other at step t by Eq. 110, so that the rewards they collect differ by

$$\big|\bar r_{\tilde\sigma}(\mathbf{y}_t) - \bar r_{\tilde\sigma}(\mathbf{y}'_t)\big| \le \min\big\{G\,C_A\Lambda^t\|\mathbf{x}'-\mathbf{x}\|,\; 2r_{\max}\big\}. \tag{114}$$

Averaging over the noises and splitting the sum at the step where the two bounds cross, as in Eqs. 62 and 63, gives the bound. □

Lemma 10 (The first action is covered by the later noises)**.** *Let the action reach every direction of the state in N steps. For a state $\mathbf{s}$, offsets $\underline{\mathbf{u}} = (\mathbf{u}_{-1}, \dots, \mathbf{u}_{N-1})$ with $\|\mathbf{u}_j\| \le c_0$ and a unit $\mathbf{e}$, let $\mathbf{w} = \mathbf{G}_0 \mathbf{B}_{\mathbf{u}_{-1}}(\mathbf{s})\mathbf{e}$ be the move of $\mathbf{y}_N$ produced by the first action and $\underline{\mathbf{c}} = \mathbf{K}_N^+ \mathbf{w}$ with $\mathbf{K}_N^+ = \mathbf{K}_N^\top (\mathbf{K}_N \mathbf{K}_N^\top)^{-1}$. Then $\mathbf{K}_N \underline{\mathbf{c}} = \mathbf{w}$, $\|\underline{\mathbf{c}}\| \le \Xi_0 = C_A B_1 \Lambda^N / c_{\mathrm{acc}}$, and $\underline{\mathbf{c}}$ is twice continuously differentiable in the offsets, with first and second derivatives bounded by constants Ξ_1 and Ξ_2 that depend on N, Λ^N, c_{acc} and the derivatives of F and $\mu_{\boldsymbol{\theta}}$ up to order three.*

Proof. Since $\|\mathbf{K}_N^\top \mathbf{z}\| \ge c_{\mathrm{acc}} \|\mathbf{z}\|$, $\mathbf{K}_N \mathbf{K}_N^\top$ is invertible with inverse of norm at most c_{acc}^{-2} and $\|\mathbf{K}_N^+\| \le c_{\mathrm{acc}}^{-1}$; $\|\mathbf{w}\| \le C_A B_1 \Lambda^N$ by Eq. 110. The matrix $\mathbf{K}_N$ and the vector $\mathbf{w}$ are products of at most N Jacobians of displaced loops and one derivative in the action, so they are twice continuously differentiable in the offsets with derivatives bounded by polynomials in N times powers of Λ^N, and the derivative of the pseudo-inverse is

$$\partial \mathbf{K}_N^+ = -\mathbf{K}_N^+ (\partial \mathbf{K}_N) \mathbf{K}_N^+ + (\mathbf{I} - \mathbf{K}_N^+ \mathbf{K}_N) (\partial \mathbf{K}_N)^\top \mathbf{K}_N^{+\top} \mathbf{K}_N^+. \tag{115}$$

Differentiating it once more gives Ξ_2. □

Lemma 11 (Integration by parts on the noises)**.** *Under the hypotheses of part (a), fix $\mathbf{s}$, $\mathbf{e}$ and $\tilde{\sigma} \in (0, 1]$, let $\Phi(h, \boldsymbol{\varepsilon})$ be the state $\mathbf{y}_N$ reached with the offsets $(h\mathbf{e}, \tilde{\sigma}\varepsilon_0, \dots, \tilde{\sigma}\varepsilon_{N-1})$, $\varphi_N = \varphi^{\otimes N}$, and for ψ bounded and α-Hölder with constant C_ψ let $A(h) = \int \psi(\Phi(h, \boldsymbol{\varepsilon})) \varphi_N(\boldsymbol{\varepsilon})\, d\boldsymbol{\varepsilon}$. Then A is twice continuously differentiable for $|h| < c_0$, with $|A'(h)| \le C_{A,1} C_\psi \tilde{\sigma}^{\alpha - 1}$ and $|A''(h)| \le C_{A,2} C_\psi \tilde{\sigma}^{\alpha - 2}$, the constants depending on N, Ξ_0, Ξ_1, Ξ_2, c_0 and the derivatives of φ.*

Proof. By the chain rule, with $Z = \tilde{\sigma}^{-1} \underline{\mathbf{c}}$ the field of Lemma 10,

$$\partial_h \Phi = \mathbf{w}, \qquad D_{\boldsymbol{\varepsilon}} \Phi = \tilde{\sigma} \mathbf{K}_N, \qquad D_{\boldsymbol{\varepsilon}} \Phi \, Z = \partial_h \Phi, \tag{116}$$

$$\|Z\| \le \Xi_0 / \tilde{\sigma}, \qquad \left| \operatorname{div}_{\boldsymbol{\varepsilon}} Z \right| \le N d_A \Xi_1. \tag{117}$$

For ψ continuously differentiable, $\partial_h [\psi \circ \Phi] = Z \cdot \nabla_{\boldsymbol{\varepsilon}} [\psi \circ \Phi]$, and integrating by parts against the compactly supported field $\varphi_N Z$,

$$A'(h) = -\int \big(\psi(\Phi(h, \boldsymbol{\varepsilon})) - \psi(\Phi(h, \mathbf{0})) \big) \operatorname{div}_{\boldsymbol{\varepsilon}} (\varphi_N Z)\, d\boldsymbol{\varepsilon}, \tag{118}$$

the constant being subtracted because $\operatorname{div}(\varphi_N Z)$ integrates to zero. The two factors of Eq. 118 are bounded separately:

$$\big| \psi\big(\Phi(h, \boldsymbol{\varepsilon})\big) - \psi\big(\Phi(h, \mathbf{0})\big) \big| \le C_\psi \big(L_N \tilde{\sigma} \big)^\alpha, \qquad L_N = N c_0 C_A B_1 \Lambda^N, \tag{119}$$

$$\big\| \operatorname{div}(\varphi_N Z) \big\|_{L^1} \le \big(N \|\nabla \varphi\|_{L^1} \Xi_0 + N d_A \Xi_1 \big) / \tilde{\sigma}, \tag{120}$$

$$|A'(h)| \le C_{A,1} C_\psi \, \tilde{\sigma}^{\,\alpha - 1}. \tag{121}$$

Eq. 119 holds because on the support of φ_N the state $\Phi(h, \boldsymbol{\varepsilon})$ stays within $L_N \tilde{\sigma}$ of $\Phi(h, \mathbf{0})$, and Eq. 120 collects Eq. 117; their product is Eq. 121. Differentiating Eq. 118 once more and integrating by parts against the field $\operatorname{div}(\varphi_N Z)\, Z$ writes A'' as an integral of the same difference against a function of L^1 norm at most a constant times $\tilde{\sigma}^{-2}$, which is the bound on A''. For a Hölder ψ, its extension to all states with the same constant [31], mollified, gives smooth functions with the same Hölder constant converging uniformly, and the bounds pass to the limit. □

Proof of Theorem 4, part (a). Fix $\mathbf{s}$, $\mathbf{e}$ and $\tilde{\sigma} \in (0, 1]$. By Eq. 109, and by the Bellman equation of $V^{\pi_{\tilde{\sigma}}}$ iterated over N steps,

$$Q^{\pi_{\tilde{\sigma}}} \big(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}) + h\mathbf{e} \big) = r\big(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}) + h\mathbf{e} \big) + \gamma V^{\pi_{\tilde{\sigma}}} \big(f_{h\mathbf{e}}(\mathbf{s}) \big), \tag{122}$$

$$V^{\pi_{\tilde{\sigma}}} \big(f_{h\mathbf{e}}(\mathbf{s}) \big) = B(h) + \gamma^N A(h), \tag{123}$$

$$\big| \partial_{\mathbf{e}} Q^{\pi_{\tilde{\sigma}}} \big(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}) \big) \big| \le H_1 + \gamma C_{B'} + \gamma^{N+1} C_{A,1} C_H \, \tilde{\sigma}^{\,\alpha - 1} \le C \, \tilde{\sigma}^{\,\alpha - 1}. \tag{124}$$

In Eq. 123 the function A is the one of Lemma 11 at $\psi = V^{\pi_{\tilde{\sigma}}}$, and $B(h)$ is the expected reward of the first N steps, whose first two derivatives in h are bounded by a constant $C_{B'}$ independent of $\tilde{\sigma}$, the noise entering B only as offsets of norm at most c_0. Eq. 124 takes $C_\psi = C_H$ from Lemma 9 and the bound on A' from Lemma 11, and collects the three terms through $\tilde{\sigma} \le 1$ and $\alpha < 1$. The second derivative is bounded in the same way, with $C_{A,2}$ in place of $C_{A,1}$. □

Accessibility cannot be dropped from part (a): on the circle, $F(x,a) = 2x + \beta(x)a$ with a smooth β vanishing near 0 and equal to one near $\frac{1}{2}$, $\mu_{\boldsymbol{\theta}} = 0$, $r(x,a) = 1 - \cos 2\pi x$ and $\gamma > \frac{1}{2}$ satisfy every other hypothesis, the actions never move the state from the fixed point 0, and from $x = \frac{1}{2}$ the action value of $\pi_{\tilde{\sigma}}$ changes like $|h|^{\log_2(1/\gamma)}$, an exponent below one, for every $\tilde{\sigma}$, so that the slope does not exist at any width.

The lower bound. By Eq. 109 the slope at $\mathbf{x}$ is

$$\partial_{\mathbf{e}} Q^{\pi_{\tilde{\sigma}}}\big(\mathbf{x}, \mu_{\boldsymbol{\theta}}(\mathbf{x})\big) = \partial_{\mathbf{e}} r\big(\mathbf{x}, \mu_{\boldsymbol{\theta}}(\mathbf{x})\big) + \gamma_0\, g_{\tilde{\sigma}}\big(f_0(\mathbf{x})\big), \quad g_{\tilde{\sigma}} = \nabla V^{\pi_{\tilde{\sigma}}} \cdot W, \quad W(\mathbf{y}) = \mathbf{B}\big(f_0^{-1}\mathbf{y}\big)\mathbf{e}, \tag{125}$$

where f_0 and γ_0 are the loop and the discount before the passage to an iterate of Appendix B.5, and $g_{\tilde{\sigma}}$, the derivative of the value along the field W, is defined and continuous on $f_0(U)$, the field W being smooth there because f_0 is a diffeomorphism onto its image. Since f_0 has Jacobian bounded above and below on U, it is enough to bound $g_{\tilde{\sigma}}$ from below in L^2 on a fixed neighbourhood of Ω. The integral curves ϕ_h of W pass through the next state along the direction $\mathbf{B}(\mathbf{x})\mathbf{e}$ in which the action moves it, as the actuated curve of Appendix B.5 does, and the value along such a curve varies by the integral of $g_{\tilde{\sigma}}$ over it.

Lemma 12 (From the movement along the integral curves to the slope)**.** *Under the hypotheses of part (b), $V^{\pi_{\tilde{\sigma}}}$ is continuously differentiable on U. If $\mathcal{G} \subset f_0(U)$ is a Borel set whose images $\phi_h(\mathcal{G})$, $|h| \le \tau$, lie in a fixed compact subset of $f_0(U)$, and $\sup_{|h|\le\tau} |V^{\pi_{\tilde{\sigma}}}(\phi_h \mathbf{y}) - V^{\pi_{\tilde{\sigma}}}(\mathbf{y})| \ge c_1 \tau^{\alpha_1}$ for every $\mathbf{y} \in \mathcal{G}$, then*

$$\int_{\phi_{[-\tau,\tau]}(\mathcal{G})} g_{\tilde{\sigma}}^2\, d\mathbf{y} \ge \tfrac{1}{4}\, e^{-D_W \tau}\, \mathrm{vol}(\mathcal{G})\, c_1^2\, \tau^{2\alpha_1 - 2}, \tag{126}$$

with D_W a bound on the divergence of W.

Proof. By Eq. 109 iterated N times, $V^{\pi_{\tilde{\sigma}}}$ is a smooth term plus its own average over the state reached after N noisy steps, and the derivative in the state is carried onto the density of the noise by the same integration by parts as in Lemma 11, now with the Kalman matrix built from the state itself, which accessibility on U allows; $V^{\pi_{\tilde{\sigma}}}$ is then continuously differentiable. For $\mathbf{y} \in \mathcal{G}$,

$$c_1 \tau^{\alpha_1} \le \sup_{|h|\le\tau} \big|V^{\pi_{\tilde{\sigma}}}(\phi_h \mathbf{y}) - V^{\pi_{\tilde{\sigma}}}(\mathbf{y})\big| \le (2\tau)^{1/2} \Big(\int_{-\tau}^{\tau} g_{\tilde{\sigma}}(\phi_h \mathbf{y})^2\, dh \Big)^{1/2}, \tag{127}$$

by the fundamental theorem of calculus along $h \mapsto \phi_h \mathbf{y}$ and the Cauchy–Schwarz inequality. Squaring it gives $\int_{-\tau}^{\tau} g_{\tilde{\sigma}}(\phi_h \mathbf{y})^2\, dh \ge \frac{1}{2} c_1^2 \tau^{2\alpha_1 - 1}$, and then

$$\tfrac{1}{2} c_1^2 \tau^{2\alpha_1 - 1}\, \mathrm{vol}(\mathcal{G}) \le \int_{-\tau}^{\tau} \int_{\mathcal{G}} g_{\tilde{\sigma}}\big(\phi_h \mathbf{y}\big)^2\, d\mathbf{y}\, dh \le 2\tau\, e^{D_W \tau} \int_{\phi_{[-\tau,\tau]}(\mathcal{G})} g_{\tilde{\sigma}}^2\, d\mathbf{y}, \tag{128}$$

which is Eq. 126. The first inequality of Eq. 128 integrates the squared form of Eq. 127 over $\mathbf{y} \in \mathcal{G}$ and exchanges the integrals; the second reads each fixed h as the change of variables $\mathbf{y} \mapsto \phi_h \mathbf{y}$, injective onto a subset of the tube with Jacobian at most $e^{D_W \tau}$ by Liouville's formula, and the length 2τ of the interval of h carries $\tau^{2\alpha_1 - 1}$ to $\tau^{2\alpha_1 - 2}$. □

Lemma 13 (The oscillation survives the smoothing)**.** *With K, δ, ϵ and $B_{K,\delta}$ the constants and the set fixed in the proof of Theorem 3, let $\mathbf{y}$ lie on the local stable leaf of $\mathbf{z} \in \Omega$, let Eq. 73 hold at $\mathbf{x} = f_0^{-1}\mathbf{y}$ with the quantities of Lemma 5 bounded by fixed numbers, and let T be a time with $\mathbf{z}_T \in B_{K,\delta}$ and $R(\tau) \le \Lambda_{uu}^{(T)}(\mathbf{z}) \le \Lambda_R R(\tau)$, where $R(\tau) = 2\delta/(c'\tau)$ with c' a lower bound on a_x and $\Lambda_R = (\Lambda^+)^2$, Λ^+ the bound on $\|Df\|$ of Appendix B.5. At the scale $\tau = A_\tau \tilde{\sigma}$, for A_τ large and $\tilde{\sigma}$ small, independently of $\mathbf{y}$,*

$$\sup_{|h|\le\tau} \big|V^{\pi_{\tilde{\sigma}}}(\phi_h \mathbf{y}) - V^{\pi_{\tilde{\sigma}}}(\mathbf{y})\big| \ge \tfrac{1}{8} K \delta\, \gamma^T \ge c_4\, \tau^{\alpha_1} e^{\alpha_1 S_T(\mathbf{z})}, \qquad S_T(\mathbf{z}) = \log \Lambda_{uu}^{(T)}(\mathbf{z}) - T\lambda_{uu}. \tag{129}$$

Proof. The proof of Theorem 3 reads its actuated curve only through the tangent $\mathbf{B}(\mathbf{x})\mathbf{e}$ at $\mathbf{y}$, which is $W(\mathbf{y})$, so its construction runs along $h \mapsto \phi_h \mathbf{y}$ and gives a displacement h_T, a point $\mathbf{z}'$ of the strong leaf of $\mathbf{z}_T$ within ϵ of $f^T \phi_{h_T} \mathbf{y}$ with $|V(\mathbf{z}') - V(\mathbf{z}_T)| > K\delta$, and the bound on the head of

the telescoping. We drive the noisy chains from $\phi_{h_T}\mathbf{y}$ and from $\mathbf{y}$ by the same noise, and write $\mathbf{\Delta}_t$ for the deviation of either chain from the noiseless orbit with the same start. Then

$$|h_T| \le \frac{2\delta}{a_x \Lambda_{uu}^{(T)}(\mathbf{z})} \le \tau, \tag{130}$$

$$\mathbf{\Delta}_{t+1} = M_t \mathbf{\Delta}_t + b_t, \qquad \|b_t\| \le C\tilde{\sigma}, \tag{131}$$

$$\|\mathbf{\Delta}_t\| \le C_3 \tilde{\sigma}\, \Lambda_{uu}^{(t)}(\mathbf{z}) \le C_4 \delta / A_\tau \qquad (t \le T), \tag{132}$$

$$\sum_{t<T} \gamma^t L C \tilde{\sigma}\, \Lambda_{uu}^{(t)}(\mathbf{z}) \le \tfrac{1}{16} K \delta\, \gamma^T, \tag{133}$$

$$\gamma^T = \exp\big(-\alpha_1 \log \Lambda_{uu}^{(T)}(\mathbf{z})\big) \exp\big(\alpha_1 S_T(\mathbf{z})\big). \tag{134}$$

In Eq. 131 the matrix M_t is an average of Jacobians of displaced loops near the orbit, which the strong cones and the distortion estimates bound by the strong rate up to a factor that stays bounded along the orbit, the perturbations being summable; that bound gives the first half of Eq. 132, and the second holds because $\Lambda_{uu}^{(T)}(\mathbf{z}) \le \Lambda_R\, 2\delta/(c' A_\tau \tilde{\sigma})$. For A_τ large, Eq. 132 puts both noisy states at time T within 2ϵ of $\mathbf{z}'$ and of $\mathbf{z}_T$ at every realisation, so that V differs between them by at least $\frac{1}{2}K\delta$, with a sign that does not depend on the noise. The coupling of the noisy and deterministic chains, which gives $|V^{\pi_{\tilde{\sigma}}} - V| \le C\tilde{\sigma}^\eta$ on U for some $\eta > 0$, costs at most $\frac{1}{8}K\delta$ for $\tilde{\sigma}$ small, and Eq. 133, the head of the coupled telescoping added to that of Theorem 3, costs another $\frac{1}{16}K\delta\gamma^T$ for A_τ large; together they give the first inequality of Eq. 129. The second follows from Eq. 134, which reads γ^T through $\alpha_1 \lambda_{uu} = \log(1/\gamma)$, with $\Lambda_{uu}^{(T)}(\mathbf{z}) \le \Lambda_R\, 2\delta/(c'\tau)$. □

The fluctuation S_T of the strong cocycle is of order $\sqrt{T}$, and the exponent in Eq. 129 is α_1 on the event $S_T \ge -C_S$. That this event keeps a probability bounded below at every scale follows from a central limit theorem. The function $\log \Lambda_{uu} - \lambda_{uu}$ is Hölder with zero mean on each piece, so its sums along the orbits of a mixing power of f, divided by the square root of the time, converge in law under ν_i, through the Markov partitions and the Hölder semiconjugacy of Bowen [28], to a centred Gaussian, which is degenerate exactly when the function is a coboundary [33]; and the convergence in law holds as well under the restriction of ν_i to any set of positive ν_i-measure [34].

Proof of Theorem 4, part (b). Let K and δ_* be the constants of Eq. 92, and let δ and $B = B_{K,\delta}$ be the scale and the set Lemma 6 gives for them, with a compact $B' \subset B$ of positive ν_i-measure that has a neighbourhood in Ω inside $\{\mathrm{osc}_\delta^{uu} > K\delta/2\}$, by the lower semicontinuity of the oscillation. Let E_τ be the set of points $\mathbf{z}$ with a time n such that $\Lambda_{uu}^{(n)}(\mathbf{z}) \in [R(\tau), R(\tau)(\Lambda^+)^2)$, $f^n\mathbf{z} \in B'$ and $S_n(\mathbf{z}) \ge -C_S$. The central limit theorem for $\log \Lambda_{uu} - \lambda_{uu}$, applied to the backward sums $S_n^-(\mathbf{w}) = S_n(f^{-n}\mathbf{w})$ under ν_i restricted to B' at the first time n at which the backward cocycle reaches $R(\tau)$, gives $\nu(E_\tau) \ge c_\nu > 0$ for every small τ. Writing $\mathcal{K}$ for a compact set on which the quantities of Lemma 5 are bounded, N' for a neighbourhood of Ω in $\mathcal{B}$ that contains the local stable leaves of every point of Ω, $\mathcal{G}$ for the points $\mathbf{y} \in \mathcal{K} \cap N'$ on the stable leaves of E_τ, and $v_\nu > 0$ for the lower bound that the absolute continuity of (H2$_\mathcal{B}$) puts on the volume of the stable saturation of a set of ν-measure at least c_ν,

$$\mathrm{vol}\big(\text{stable leaves of } E_\tau\big) \ge v_\nu > 0, \tag{135}$$

$$\mathrm{vol}\big(\mathcal{K}^{\mathrm{c}} \cap N'\big) \le v_\nu/2, \qquad \mathrm{vol}(\mathcal{G}) \ge v_\nu/2, \tag{136}$$

$$\sup_{|h|\le\tau} \big|V^{\pi_{\tilde{\sigma}}}(\phi_h \mathbf{y}) - V^{\pi_{\tilde{\sigma}}}(\mathbf{y})\big| \ge c_4 e^{-\alpha_1 C_S} \tau^{\alpha_1} \qquad (\mathbf{y} \in \mathcal{G}), \tag{137}$$

$$\big\|g_{\tilde{\sigma}}\big\|_{L^2(N')} \ge c\, \tilde{\sigma}^{\,\alpha_1 - 1}. \tag{138}$$

Eq. 135 is the first property of (H2$_\mathcal{B}$), the projection of ν along the stable leaves onto a local unstable leaf being absolutely continuous. In Eq. 136 the compact set $\mathcal{K}$ exists because, by Lemma 7, those quantities are finite at almost every point of $\mathcal{B}$. Eq. 137 is Lemma 13 at the scale $\tau = A_\tau \tilde{\sigma}$, applied with the time n at which the backward cocycle reaches $R(\tau)$, the cocycles of two points of one stable leaf differing by a bounded amount, and Eq. 138 is Lemma 12 read with $c_1 = c_4 e^{-\alpha_1 C_S}$ and the volume of Eq. 136. With $\mathcal{U} = f_0^{-1}(N')$, Eqs. 125 and 138 give Eq. 113, the Jacobian of f_0 bounded above and below on U and the term $\partial_{\mathbf{e}} r$ bounded while $\tilde{\sigma}^{\alpha_1 - 1} \to \infty$. □

C.2 Proof of Theorem 5

The update averages the slope against the discounted distribution of states, which has a density, and the derivative passes from the rough value onto that density. We write $\langle g, \varphi \rangle$ for the integral of $g\varphi$ against the volume on U, extended to distributions g. The transfer operator $\mathcal{L}$ of the closed loop, $\langle \mathcal{L} g, \varphi \rangle = \langle g, \varphi \circ f \rangle$, carries the density of the state one step forward, and likewise $\mathcal{L}_{\mathbf{u}}$ for the displaced loop $f_{\mathbf{u}}$ and $\mathcal{L}_{\tilde\sigma} = \mathbb{E}_{\boldsymbol{\zeta}} \mathcal{L}_{\tilde\sigma \boldsymbol{\zeta}}$ for the chain of $\pi_{\tilde\sigma}$; all of them preserve mass. With $T_{\tilde\sigma}$ the operator of Eq. 109, the value of $\pi_{\tilde\sigma}$, its truncation at the horizon T and the unnormalised occupancy are

$$V^{\pi_{\tilde\sigma}} = \sum_{t \geq 0} \gamma^t T_{\tilde\sigma}^t \bar r_{\tilde\sigma}, \qquad V_{\tilde\sigma}^{(T)} = \sum_{t \leq T} \gamma^t T_{\tilde\sigma}^t \bar r_{\tilde\sigma}, \qquad \bar d = \sum_{t \geq 0} \gamma^t \mathcal{L}^t p_0 = \frac{d^{\mu_{\boldsymbol{\theta}}}}{1-\gamma}. \tag{139}$$

The loops near $\boldsymbol{\theta}$ have hyperbolic attractors, and their transfer operators have a spectral gap on anisotropic spaces of distributions. For an integer $p' \geq 1$ and a real $q' \geq 0$ with $p' + q' < k - 1$ let $B^{p',q'}(U)$ be the space of Gouëzel and Liverani [35] localised on the trapping region, the completion of the smooth functions with compact support in U for a norm that measures p' derivatives by integrating on leaves close to the stable ones against test functions of class $C^{q'}$.

Theorem 8 (Anisotropic spaces)**.** *Under (H1), with f mixing on Ω, the following hold.*

(i) $C^{p'}(U)$ embeds continuously in $B^{p',q'}$, and $B^{p',q'}$ embeds compactly in $B^{p'-1,q'+1}$ [35, Remark 4.3], [36, Lemma 4.3].

(ii) $|\langle g, \varphi \rangle| \leq C \|\varphi\|_{C^{q'}} \|g\|_{p',q'}$ [35, Proposition 4.1].

(iii) Derivatives along fields of class $C^{p'+q'}$ are bounded from $B^{p',q'}$ to $B^{p'-1,q'+1}$, and multiplication by a function of class $C^{p'+q'}$ is bounded on $B^{p',q'}$ by its $C^{p'+q'}$ norm; both follow from the definition of the norms [35, Section 9].

(iv) $\mathcal{L}$ is bounded on $B^{p',q'}$, with essential spectral radius at most $\max(\Lambda_-^{-p'}, \varrho^{q'})$ times its spectral radius [36, Proposition 4.4]; that radius is the exponential of the topological pressure, and equals one for the potential of ν [36, Corollary 4.6 and Theorem 6.4]; and it is a simple eigenvalue alone on its circle, on $B^{p',q'}$ when f is orientation mixing and on the orientable subspace $B_+^{p',q'}$ when f is mixing and orientation preserving [36, Theorem 5.1]. Hence $\|\mathcal{L}^t g - \langle g, \mathbf{1} \rangle \nu\|_{p',q'} \leq C \kappa^t \|g\|_{p',q'}$ with $\kappa < 1$.

The spectral gap is quoted on the subspace of distributions that are even under the change of orientation; that subspace contains $\bar d$ and every density and distribution of mass zero the proof of Theorem 5 uses.

The actions enter through the reachability of the expanding directions. For $\mathbf{x} \in \Omega$, an integer $n_0 \geq 1$ and controls $\boldsymbol{c}^{(0)}, \dots, \boldsymbol{c}^{(n_0-1)}$ of the next n_0 actions, let

$$\Xi_{\mathbf{x}}\big(\boldsymbol{c}^{(0)}, \dots, \boldsymbol{c}^{(n_0-1)}\big) = \Big[\sum_{i < n_0} D f^{\,n_0-1-i}(f^{\,i+1}\mathbf{x})\, \mathbf{B}(f^{\,i}\mathbf{x})\, \boldsymbol{c}^{(i)} \Big] \in \mathbb{R}^{d_S} / E^s(f^{\,n_0}\mathbf{x}) \tag{140}$$

be the class, modulo the stable directions, of the move they produce. *Condition (K)* asks that for some such n_0 and some $c_K > 0$ the map $\Xi_{\mathbf{x}}$ have a right inverse of norm at most $1/c_K$ at every $\mathbf{x} \in \Omega$: the actions of n_0 steps reach every expanding direction. It follows from the accessibility of Definition 2 on U, and with $n_0 = 1$ it asks that the moves $\mathbf{B}(\mathbf{x})\boldsymbol{c}$ fill the quotient by $E^s(f\mathbf{x})$.

Theorem 5 (The update is unbiased up to $\tilde\sigma^2$)**.** *Let Assumption 1 hold with $k \geq 7$, with an integer $p \geq 4$ and a real $q \in (0,1)$, $p + q < k - 2$, a reward of class C^{q+3} and $\boldsymbol{\theta} \mapsto \mu_{\boldsymbol{\theta}}$ differentiable into C^{p+q+1}; let every loop $f_{\boldsymbol{\theta}'}$ near $\boldsymbol{\theta}$ satisfy (H1) with the same U and be mixing on its attractor; let p_0 be of class C^p with support in a compact subset of U. Let $\boldsymbol{\zeta}$ satisfy (N), and let there be $\tilde\sigma_1 > 0$ such that the displaced loops $f_{\mathbf{u}}$ with $\|\mathbf{u}\| \leq \tilde\sigma_1 c_0$ satisfy (H1) with the same region and the same constants, with invertible Jacobians on $\overline{U}$, and such that the operators $\mathcal{L}_{\tilde\sigma\boldsymbol{\zeta}}$ with $\tilde\sigma \leq \tilde\sigma_1$ and $\|\boldsymbol{\zeta}\| \leq c_0$ are bounded on each $B^{p',q'}$ uniformly in $\tilde\sigma$ and $\boldsymbol{\zeta}$. Let the operators $\mathcal{L}_{\tilde\sigma}$ and $\mathcal{L}_{\boldsymbol{\theta}'}$ have the spectral gap of Theorem 8 with constants independent of $\tilde\sigma$ and $\boldsymbol{\theta}'$, and $\boldsymbol{\theta}' \mapsto \mathcal{L}_{\boldsymbol{\theta}'}$ be differentiable at $\boldsymbol{\theta}$ from $B^{p',q'}$ to $B^{p'-1,q'+1}$, the stability of the spectral gap and the linear response of this*

setting [35, 36]; and let (K) hold. Then there is $\sigma_K \in (0, \tilde{\sigma}_1]$, independent of γ, such that J is differentiable at $\boldsymbol{\theta}$ and, for $\ell_s \le L_F\tilde{\sigma} \le L_F\sigma_K$,

$$\mathbf{g}(\boldsymbol{\theta}) = (1-\gamma)\,\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) + O(\tilde{\sigma}^2), \tag{141}$$

with a constant independent of γ; at almost every $\mathbf{s} \in U$ the gradient $\nabla_{\mathbf{a}} Q^{\pi_{\tilde{\sigma}}}(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}))$ exists and is at most a constant times $1 + 1/(\tilde{\sigma}(1-\gamma))$.

Lemma 14 (A Lipschitz bound uniform in the horizon)**.** *Under the hypotheses of Theorem 5 there are $\sigma_K > 0$ and C_K, independent of γ, such that for $\tilde{\sigma} \le \sigma_K$, every horizon T, every $\mathbf{x} \in U$ and every $\boldsymbol{v}$, $|\partial_{\boldsymbol{v}} V_{\tilde{\sigma}}^{(T)}(\mathbf{x})| \le C_K(1 + 1/(\tilde{\sigma}(1-\gamma)))\|\boldsymbol{v}\|$.*

Proof. Move the initial state by $\boldsymbol{v}$ and the noise $\boldsymbol{\zeta}_t$ by a control $\boldsymbol{c}_t$ that depends only on the past: the displacement of the state at step t is $\boldsymbol{v}_t$, where $\boldsymbol{v}_0 = \boldsymbol{v}$ and $\boldsymbol{v}_{t+1} = Df_{\tilde{\sigma}\boldsymbol{\zeta}_t}(\mathbf{x}_t)\boldsymbol{v}_t + \tilde{\sigma}\mathbf{B}(\mathbf{x}_t, \mathbf{a}_t)\boldsymbol{c}_t$ along the run $(\mathbf{x}_t, \mathbf{a}_t)$. Integrating by parts in each $\boldsymbol{\zeta}_t$ against its density writes the derivative of $V_{\tilde{\sigma}}^{(T)}$ along $\boldsymbol{v}$ as the expected derivative of the discounted rewards along $(\boldsymbol{v}_t)$ plus $\sum_t \mathbb{E}[G_T\, \boldsymbol{c}_t \cdot \nabla \log p_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}_t)]$, with $p_{\boldsymbol{\zeta}}$ the density of (N), G_T the discounted return over the T steps and $|G_T| \le r_{\max}/(1-\gamma)$. It is therefore enough to choose controls with $\|\boldsymbol{v}_t\| + \tilde{\sigma}\|\boldsymbol{c}_t\| \le C2^{-t/N}\|\boldsymbol{v}\|$ for some C and N. Near Ω the stable directions extend to a continuous invariant field, contracted by Df, on a compact U' with $f(U') \subset \operatorname{int} U'$ and $f^{n_1}(U) \subset U'$, where (K) extends with the constant $c_K/2$; the controls are zero up to the step n_1 at which the orbit enters U'. On blocks of $N \ge n_0$ steps, the last n_0 controls of each block are chosen by the right inverse of Eq. 140 so that the planned move at the end of the block, computed along the noiseless orbit, lies in the stable directions; the stable part is then multiplied by $\frac{1}{4}$ over the block, while the noise moves the orbit and the Jacobians by $O(\tilde{\sigma})$ and the stable field by its modulus of continuity at $O(\tilde{\sigma})$. For $\tilde{\sigma}$ small the quantity $\|\boldsymbol{v}^{\perp}\| + \|\boldsymbol{v}^{s}\|/(4C_N)$, with $\boldsymbol{v}^{\perp}$ and $\boldsymbol{v}^{s}$ the parts off and along the stable directions and $C_N = \sup_{i \le N} \sup_U \|Df^i\|$, halves over every block, which gives the required decay. □

By Lemma 14, $V^{\pi_{\tilde{\sigma}}}$ is Lipschitz with constant $L = C_K(1 + 1/(\tilde{\sigma}(1-\gamma)))$ on every ball contained in U, hence differentiable at almost every state, by Rademacher's theorem, with $\nabla V_{\tilde{\sigma}}^{(T)} \to \nabla V^{\pi_{\tilde{\sigma}}}$ weakly-$*$ in L^∞, as $V_{\tilde{\sigma}}^{(T)} \to V^{\pi_{\tilde{\sigma}}}$ uniformly with gradients bounded by L. Since f is a diffeomorphism onto its image, the chain rule at the states $\mathbf{s}$ where $V^{\pi_{\tilde{\sigma}}}$ is differentiable at $f(\mathbf{s})$, which are almost all, gives the gradient in the action that Theorem 5 bounds. With $\mathbf{w}_j = \mathbf{B}\,\partial\mu_{\boldsymbol{\theta}}/\partial\theta_j$ and $\bar{r}_{\boldsymbol{\theta}}(\mathbf{s}) = r(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}))$ the reward along the loop at the parameter $\boldsymbol{\theta}$, the derivative of the action value at a fixed state, along $\partial\mu_{\boldsymbol{\theta}}(\mathbf{s})/\partial\theta_j$, is

$$D_j Q^{\pi_{\tilde{\sigma}}}(\mathbf{s}) = \frac{\partial \bar{r}_{\boldsymbol{\theta}}}{\partial \theta_j}(\mathbf{s}) + \gamma \left\langle \mathbf{w}_j(\mathbf{s}), \nabla V^{\pi_{\tilde{\sigma}}}(f(\mathbf{s})) \right\rangle \tag{142}$$

and its average against $\bar{d}$ is $\mathcal{F}_j(\tilde{\sigma}) = g_j(\boldsymbol{\theta})/(1-\gamma)$ by Eq. 20 when $\ell_s \le L_F\tilde{\sigma}$.

Proof of Theorem 5. The occupancy $\bar{d}$ has an integrable density and, by Theorem 8, lies in $B^{p,q}$ with $\|\bar{d}\|_{p,q} \le C(1 + \|p_0\|_{C^p})/(1-\gamma)$. With $\mathbf{z}_j = \mathbf{A}^{-1}\mathbf{w}_j$, of class C^{p+q}, the identity $\langle \mathbf{w}_j, (\nabla\varphi)\circ f\rangle = \partial_{\mathbf{z}_j}(\varphi \circ f)$ and an integration by parts, which leaves no boundary term because $\bar{d}\,\mathbf{z}_j$ has compact support in U, give for every φ of class C^1

$$\left\langle \bar{d}, \langle \mathbf{w}_j, (\nabla\varphi)\circ f\rangle \right\rangle = \langle u_j, \varphi\rangle, \qquad u_j = -\mathcal{L}\operatorname{div}(\bar{d}\,\mathbf{z}_j), \tag{143}$$

with $u_j \in B^{p-1,q+1}$ of mass zero. Applied to $V_{\tilde{\sigma}}^{(T)}$, with $\langle u_j, P_{\tilde{\sigma}}^t \bar{r}_{\tilde{\sigma}}\rangle = \langle \mathcal{L}_{\tilde{\sigma}}^t u_j, \bar{r}_{\tilde{\sigma}}\rangle$, Eq. 143 gives

$$\left\langle \bar{d}, \frac{\partial \bar{r}_{\boldsymbol{\theta}}}{\partial \theta_j} \right\rangle + \gamma \left\langle \bar{d}, \langle \mathbf{w}_j, (\nabla V_{\tilde{\sigma}}^{(T)})\circ f\rangle \right\rangle = \left\langle \bar{d}, \frac{\partial \bar{r}_{\boldsymbol{\theta}}}{\partial \theta_j} \right\rangle + \gamma \sum_{t \le T} \gamma^t \left\langle \mathcal{L}_{\tilde{\sigma}}^t u_j, \bar{r}_{\tilde{\sigma}} \right\rangle.$$

As $T \to \infty$ the left side tends to $\mathcal{F}_j(\tilde{\sigma})$, by the weak-$*$ convergence and Eq. 142, and the right side to

$$\mathcal{G}_j(\tilde{\sigma}) = \left\langle \bar{d}, \frac{\partial \bar{r}_{\boldsymbol{\theta}}}{\partial \theta_j} \right\rangle + \gamma \sum_{t \ge 0} \gamma^t \left\langle \mathcal{L}_{\tilde{\sigma}}^t u_j, \bar{r}_{\tilde{\sigma}} \right\rangle, \tag{144}$$

the series converging at the rate of the spectral gap because u_j has mass zero. So $\mathcal{F}_j(\tilde{\sigma}) = \mathcal{G}_j(\tilde{\sigma})$. At $\tilde{\sigma} = 0$ the same series is the derivative of the objective: writing $J(\boldsymbol{\theta}') = \sum_t \gamma^t \langle \mathcal{L}_{\boldsymbol{\theta}'}^t p_0, \bar{r}_{\boldsymbol{\theta}'}\rangle$, the

linear response of the transfer operator, $\partial_{\theta_j}\mathcal{L}_{\boldsymbol{\theta}}\, g = -\operatorname{div}((\mathcal{L}g)(\mathbf{w}_j \circ f^{-1}))$, gives $\partial_{\theta_j}\mathcal{L}_{\boldsymbol{\theta}}\,\bar{d} = u_j$ by Eq. 143, and differentiating term by term, which the uniform spectral gap allows by dominated convergence, and reordering,

$$\frac{\partial J}{\partial \theta_j} = \Big\langle \bar{d}, \frac{\partial \bar{r}_{\boldsymbol{\theta}}}{\partial \theta_j} \Big\rangle + \gamma \sum_{t\geq 0} \gamma^t \big\langle \mathcal{L}^t u_j, \bar{r} \big\rangle = \mathcal{G}_j(0). \tag{145}$$

The functionals $\mathcal{G}_j(\tilde{\sigma})$ and $\mathcal{G}_j(0)$ differ by $O(\tilde{\sigma}^2)$. With $\mathbf{A}_{\mathbf{u}} = Df_{\mathbf{u}}$ the Jacobian of the displaced loop and $\mathbf{B}_{\mathbf{u}}$ of Eq. 111, the fields $\mathbf{X}_{s,\boldsymbol{\zeta}} = \mathbf{A}_{s\boldsymbol{\zeta}}^{-1}\mathbf{B}_{s\boldsymbol{\zeta}}\boldsymbol{\zeta}$ turn $\partial_s(\varphi\circ f_{s\boldsymbol{\zeta}})$ into $\partial_{\mathbf{X}_{s,\boldsymbol{\zeta}}}(\varphi\circ f_{s\boldsymbol{\zeta}})$ at every $s \in [0,\tilde{\sigma}]$. A Taylor expansion in s of $\varphi \circ f_{s\boldsymbol{\zeta}}$, read at $s = \tilde{\sigma}$, has the first-order term $\partial_{\mathbf{X}_{0,\boldsymbol{\zeta}}}(\varphi \circ f)$ linear in $\boldsymbol{\zeta}$, which the zero mean of the noise removes, so that $\|(\mathcal{L}_{\tilde{\sigma}} - \mathcal{L})g\|_{p-3,q+3} \leq C\tilde{\sigma}^2\|g\|_{p-1,q+1}$, the remainder containing two derivatives along fields of class C^{p+q}; likewise $\|\bar{r}_{\tilde{\sigma}} - \bar{r}\|_{C^{q+1}} \leq C\tilde{\sigma}^2$. Telescoping $\mathcal{L}_{\tilde{\sigma}}^t - \mathcal{L}^t = \sum_{i+n=t-1}\mathcal{L}_{\tilde{\sigma}}^i(\mathcal{L}_{\tilde{\sigma}} - \mathcal{L})\mathcal{L}^n$ on vectors of mass zero,

$$\Big|\mathcal{G}_j(\tilde{\sigma}) - \frac{\partial J}{\partial \theta_j}\Big| \leq C\tilde{\sigma}^2 \|\mathbf{z}_j\|_{C^{p+q}} \|\bar{d}\|_{p,q} \Big[\frac{\|\bar{r}\|_{C^{q+3}}}{(1-\kappa_2)^2} + \frac{\|\partial_{\mathbf{a}}^2 r\|_{C^{q+1}}}{1-\kappa_1}\Big],$$

with $\kappa_1 < 1$ the rate of the spectral gap of $\mathcal{L}_{\tilde{\sigma}}$, which Theorem 5 asks for uniformly in $\tilde{\sigma}$, $\kappa_2 = \max(\kappa, \kappa_1)$ with κ the rate of Theorem 8 for $\mathcal{L}$, and the sums over t bounded with the discount dropped. Since $g_j(\boldsymbol{\theta}) = (1-\gamma)\mathcal{F}_j(\tilde{\sigma}) = (1-\gamma)\mathcal{G}_j(\tilde{\sigma})$ and $(1-\gamma)\|\bar{d}\|_{p,q}$ is bounded independently of γ, this is Eq. 141. □

C.3 The resolution of a critic

A critic linear in its parameters is fitted at the projected fixed point of temporal-difference regression [21]. We read the action through its offset $\mathbf{u} = \mathbf{a} - \mu_{\boldsymbol{\theta}}(\mathbf{s})$, writing $\tilde{F}(\mathbf{s},\mathbf{u}) = F(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{u})$ and $\tilde{r}(\mathbf{s},\mathbf{u}) = r(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{u})$, so that in the mean the target of Eq. 6 is given by the operator of Section 4.1,

$$(\mathcal{B}_{\tilde{\sigma}}Q)(\mathbf{s},\mathbf{u}) = \tilde{r}(\mathbf{s},\mathbf{u}) + \gamma\, \mathbb{E}_{\boldsymbol{\zeta}}\big[Q\big(\tilde{F}(\mathbf{s},\mathbf{u}), \tilde{\sigma}\boldsymbol{\zeta}\big)\big], \tag{146}$$

whose fixed point is $Q^{\pi_{\tilde{\sigma}}}$. The data are pairs $(\mathbf{s},\mathbf{u})$ with law $D = \nu \otimes \rho_{\tilde{\sigma}}$, where $\rho_{\tilde{\sigma}}$ is the law of $\tilde{\sigma}\boldsymbol{\zeta}$ and ν a law of states: the offsets carried by the data and the offsets that smooth the targets have the same width, the on-policy case. Three conditions bear on the data and on the class.

(T) The law ν of the states is stationary for the chain the evaluated actor runs, $\nu P_{\tilde{\sigma}} = \nu$ with $(P_{\tilde{\sigma}}v)(\mathbf{s}) = \mathbb{E}[v(\tilde{F}(\mathbf{s},\tilde{\sigma}\boldsymbol{\zeta}))]$, the setting of Tsitsiklis and Van Roy [21].

(R) The class is a product $\mathcal{F} = \mathcal{F}_s \otimes \mathcal{F}_u$ of a closed space $\mathcal{F}_s$ of continuously differentiable functions of the state and a finite-dimensional space $\mathcal{F}_u$ of functions of the offset that contains the constants, and $\mathcal{F}_s$ resolves the state at a scale ℓ_s: in the norm of $L^2(\nu)$, every continuously differentiable g has an approximant $v \in \mathcal{F}_s$ with

$$\|g - v\|_\nu \leq C_J\, \ell_s\, \|\nabla g\|_\infty, \qquad \|\nabla v\|_\nu \leq C'_J\, \|\nabla g\|_\infty, \qquad \|\nabla w\|_\nu \leq \frac{C_B}{\ell_s}\, \|w\|_\nu \quad (w \in \mathcal{F}_s), \tag{147}$$

the first two of Jackson and the third of Bernstein [37].

(F) For $v \in \mathcal{F}_s$ and every $\mathbf{s}$ the map $\mathbf{u} \mapsto v(\tilde{F}(\mathbf{s},\mathbf{u}))$ lies in $\mathcal{F}_u$, as it does on a torus when the action enters additively and $\mathcal{F}_u$ is a Fourier class.

The two scales are read on different variables, the noise on the action and the resolution on the state, and $L_F = \|\partial_{\mathbf{a}}F\|_\infty$ of Section 4.1 carries one into the other, so that $L_F\tilde{\sigma}$ and ℓ_s are lengths on the same axis.

Proposition 5 (Slope of a critic of resolution ℓ_s)**.** Let Assumption 1 hold with $k \geq 1$, let $\gamma\Lambda > 1$, let (T), (R) and (F) hold, let the reward depend on the state only, and let $V^{\pi_{\tilde{\sigma}}}$ be α-Hölder with constant C_H and satisfy $\|\nabla V^{\pi_{\tilde{\sigma}}}\|_\infty \leq C_V\tilde{\sigma}^{\alpha-1}$ for $\tilde{\sigma} \in (0,1]$. Then the projected fixed point $\hat{Q}$ of $\mathcal{B}_{\tilde{\sigma}}$ in $\mathcal{F}$ exists, is unique, and

$$\big\|\partial_{\mathbf{u}}\hat{Q}\big\|_D \leq K \max\{L_F\tilde{\sigma}, \ell_s\}^{\alpha-1},$$
$$K = \gamma L_F \max(1, L_F)^{1-\alpha} \max(C_H, C_V)\Big[C'_J C_\varrho + C_B(1 + C_J C_\varrho)\Big(1 + \frac{1}{\sqrt{1-\gamma^2}}\Big)\Big], \tag{148}$$

with C_ϱ a constant of a fixed mollifier; K depends on neither $\tilde{\sigma}$ nor ℓ_s.

The Hölder bound asked of $V^{\pi_{\tilde\sigma}}$ is that of Lemma 9, where C_H is fixed, and the bound on its gradient follows from the argument of Theorem 4 read for a move of the state.

Proof. **Existence of the projected fixed point.** The transition is non-expansive in $L^2(D)$ by (T) and Jensen's inequality, so $\mathcal{B}_{\tilde\sigma}$ is a γ-contraction, and so is $\Pi\mathcal{B}_{\tilde\sigma}$ on $\mathcal{F}$ with Π the orthogonal projection; the fixed point $\hat Q$ therefore exists and is unique. Since $\hat Q - \Pi Q^{\pi_{\tilde\sigma}}$ lies in $\mathcal{F}$ and is orthogonal to $\Pi Q^{\pi_{\tilde\sigma}} - Q^{\pi_{\tilde\sigma}}$,

$$\big\|\hat Q - Q^{\pi_{\tilde\sigma}}\big\|_D \le \frac{1}{\sqrt{1-\gamma^2}} \inf_{p\in\mathcal{F}} \big\|Q^{\pi_{\tilde\sigma}} - p\big\|_D \tag{149}$$

[21]. With an orthonormal basis of $\mathcal{F}_u$ in $L^2(\rho_{\tilde\sigma})$ that starts with the constant, Π is the product of the projections Π_s and Π_u, and averaging in the offset commutes with it, so the value $\hat V = \mathbb{E}_{\rho_{\tilde\sigma}}\hat Q$ is the projected fixed point in $\mathcal{F}_s$ of the state chain and carries the bound of Eq. 149 in the norm of $L^2(\nu)$.

The exponent comes from the state. The two bounds asked of $V^{\pi_{\tilde\sigma}}$ bound its Hölder seminorm of order $\beta \in [\alpha, 1]$ by $\max(C_H, C_V)\tilde\sigma^{\alpha-\beta}$, and its mollification g_h at scale h satisfies

$$\big\|V^{\pi_{\tilde\sigma}} - g_h\big\|_\infty \le \big[V^{\pi_{\tilde\sigma}}\big]_\beta h^\beta, \qquad \|\nabla g_h\|_\infty \le C_\varrho\big[V^{\pi_{\tilde\sigma}}\big]_\beta h^{\beta-1}. \tag{150}$$

Let $v \in \mathcal{F}_s$ be the Jackson approximant of g_{ℓ_s}. Bernstein applied to $\hat V - v$, with $\|\hat V - v\|_\nu \le \|\hat V - V^{\pi_{\tilde\sigma}}\|_\nu + \|V^{\pi_{\tilde\sigma}} - v\|_\nu$, Eq. 149 and Eq. 150, gives

$$\begin{aligned}\|\nabla\hat V\|_\nu &\le \|\nabla v\|_\nu + \frac{C_B}{\ell_s}\Big(1 + \frac{1}{\sqrt{1-\gamma^2}}\Big)\big\|V^{\pi_{\tilde\sigma}} - v\big\|_\nu \\ &\le \max(C_H, C_V)\,\tilde\sigma^{\alpha-\beta}\ell_s^{\beta-1}\Big[C'_J C_\varrho + C_B(1 + C_J C_\varrho)\Big(1 + \frac{1}{\sqrt{1-\gamma^2}}\Big)\Big],\end{aligned} \tag{151}$$

and the minimum over β of $\tilde\sigma^{\alpha-\beta}\ell_s^{\beta-1}$ is $\max\{\tilde\sigma, \ell_s\}^{\alpha-1}$, at most $\max(1, L_F)^{1-\alpha}\max\{L_F\tilde\sigma, \ell_s\}^{\alpha-1}$, the noise of the targets entering through the displacement of the state it produces. By (F) the target $\gamma\hat V\circ\tilde F$ lies in $L^2(\nu)\otimes\mathcal{F}_u$, so that $\hat Q = \Pi_s\tilde r + \gamma(\Pi_s\otimes I)(\hat V\circ\tilde F)$, whose first term has no derivative in the offset because the reward depends on the state alone, and

$$\big\|\partial_{\mathbf{u}}\hat Q\big\|_D \le \gamma L_F\|\nabla\hat V\|_\nu, \tag{152}$$

which is Eq. 148. In Eq. 151 the two inequalities of the class enter at the same scale, which is why the class constants carry no power of ℓ_s; in Eq. 152 the projection Π_s acts on the state alone and has norm one, so it does not increase the mean square of the derivative in the offset, and the invariance of ν gives $\|(\nabla\hat V)\circ\tilde F\|_D = \|\nabla\hat V\|_\nu$. □

The width the critic imposes on the slope in Eq. 148 is the larger of $L_F\tilde\sigma$, the width of the offsets read on the state, and the scale ℓ_s at which the class resolves the value.

The actor follows the fitted critic, and the slope of Eq. 148 differs from the slope of the value by the error the class makes on that value, divided by ℓ_s; three further conditions turn that reading into a bound.

($\mathbf{T}_C$) The data cover what the target reads: $(\tilde F(\cdot, \mathbf{u}))_*\nu \le C_1\nu$ for every offset of the support and for $\mathbf{u} = \mathbf{0}$, and the law $\rho_{\tilde\sigma}$ of the offsets in the data is dominated by the law ρ at which the targets read them, $\rho_{\tilde\sigma} \le C_2\rho$, with $\gamma^2 C < 1$ for $C = C_1C_2$. Condition (T) asks the second of these with $C_2 = 1$, the two laws being the same there.

($\mathbf{R}_2$) The class of (R) approximates at the second order as well, and does so simultaneously with the gradient: every g of class C^2 has an approximant $Ag \in \mathcal{F}_s$ with $\|g - Ag\|_\nu \le c_2\ell_s^2\|\nabla^2 g\|_\infty$ and $\|\nabla(g - Ag)\|_\nu \le c_1\ell_s\|\nabla^2 g\|_\infty$.

($\mathbf{V}_2$) $\|\nabla V^{\pi_{\tilde\sigma}}\|_\infty \le C_V\tilde\sigma^{\alpha-1}$ and $\|\nabla^2 V^{\pi_{\tilde\sigma}}\|_\infty \le C_{V2}\tilde\sigma^{\alpha-2}$, the two bounds of Eq. 19 read for a displacement of the state.

Proposition 6 (The slope of the fitted critic)**.** Let Assumption 1 hold with $k \ge 2$, let the reward depend on the state alone, and let (T_C), (R), (R_2), (F) and (V_2) hold with C_2 bounded independently

of $\tilde{\sigma}$. Then there is a constant K_C that depends on the model, on the class and on γ through $(1-\gamma^2 C)^{-1/2}$, but on neither ℓ_s nor $\tilde{\sigma}$, such that

$$\left\|\partial_{\mathbf{u}}\hat{Q} - \partial_{\mathbf{u}} Q^{\pi_{\tilde{\sigma}}}\right\|_D \le K_C\, \ell_s\, \tilde{\sigma}^{\,\alpha-2}. \tag{153}$$

Proof. Write $Q = Q^{\pi_{\tilde{\sigma}}}$, $e_{\mathrm{app}} = \inf_{p\in\mathcal{F}} \|Q-p\|_D$, and let p be an approximant. The error on the value, its passage to the derivative and the size of e_{app} obey

$$\left\|\hat{Q} - Q\right\|_D \le \frac{1}{\sqrt{1-\gamma^2 C}}\, e_{\mathrm{app}}, \tag{154}$$

$$\left\|\partial_{\mathbf{u}}(\hat{Q} - p)\right\|_D \le \frac{\gamma L_F \sqrt{C_1 C_2}\, C_B}{\ell_s} \left\|\hat{Q} - p\right\|_D, \tag{155}$$

$$e_{\mathrm{app}} \le c_2\, \ell_s^2 \Big[\left\|\nabla^2 r\right\|_\infty + \gamma\, C'\, \tilde{\sigma}^{\alpha-2}\Big], \tag{156}$$

and adding $\|\partial_{\mathbf{u}}(p-Q)\|_D \le \gamma c_1 \ell_s C' \tilde{\sigma}^{\alpha-2}$, with $C' = M_F C_V + L_F^2 C_{V2}$ and $M_F = \|\partial_{\mathbf{u}}^2 \tilde{F}\|_\infty$, gives Eq. 153.

Eq. 154 is Eq. 149 with the transition non-expansive only up to the constants of (T_C), which replaces γ by $\gamma\sqrt{C}$ in the contraction. Eq. 155 applies the third inequality of Eq. 147 to $\hat{Q} - p$, which lies in the class, and reads the derivative in the offset through $\partial_{\mathbf{u}}\tilde{F}$, bounded by L_F, and the change of measure of (T_C). Eq. 156 takes $p = (\Pi_s r) \otimes 1 + \gamma(\Pi_s \otimes I)((AV^{\pi_{\tilde{\sigma}}}) \circ \tilde{F})$, which (F) keeps in the class, and bounds the three pieces it leaves by the second-order approximation of (R_2), the second derivative of $V^{\pi_{\tilde{\sigma}}} \circ \tilde{F}$ in the offset at a fixed state being at most $C'\tilde{\sigma}^{\alpha-2}$ by (V_2); the remaining term is where the simultaneous approximation is used, on $p - Q$, whose derivative (R_2) controls without differentiating the projection. □

The first order alone would not do: with an approximant that is accurate to $O(\ell_s)$ on the value, Eq. 155 returns a bound of the size of the two slopes it compares, and the proposition would be true and empty. Read against the slope itself, which is of order $\tilde{\sigma}^{\alpha-1}$, Eq. 153 is a relative error of order $\ell_s/\tilde{\sigma}$; read against the $\tilde{\sigma}^2$ of Theorem 5, it is the smaller of the two only when $\ell_s \lesssim \tilde{\sigma}^{\,4-\alpha}$.

D TWO FURTHER CONSEQUENCES

The mechanisms of Section 4 apply as well to the exploration noise σ, with which the actor is trained and without which it is deployed. Write π_σ for the actor that plays $\mu_{\boldsymbol{\theta}}(\mathbf{s}) + \sigma\boldsymbol{\xi}$ at every step, V^{π_σ} for its value, the value of Eq. 18 read at the width σ, and $C_\sigma = V^{\pi_\sigma} - V$ for the cost of exploring from a state. It is built from the displacements of Section 3: the noise injected at step k moves the state, the closed loop multiplies the move by Λ at each later step, and the reward at step t reacts to it. Because the noise has zero mean the first order in σ cancels at each injection, and what is left carries the displacement squared, so that the term of step t is at most σ^2 times a constant times Λ^{2t}, and the series has the ratio $\gamma\Lambda^2$ that decides the curvature in Section 3.1. On an attractor the exponent α_1 is attained under two further conditions: f is topologically mixing on Ω (M), and the strong rate is the same along every orbit,

$$C_{uu}^{-1} e^{\lambda_1 n} \le \Lambda_{uu}^{(n)}(\boldsymbol{q}) \le C_{uu} e^{\lambda_1 n} \qquad (\boldsymbol{q}\in\Omega,\ n\ge 0), \tag{157}$$

which by Livšic's theorem [16] holds exactly when $\log \Lambda_{uu}$ differs from the constant λ_1 by a Hölder coboundary, for instance when every periodic orbit of Ω has mean strong rate λ_1 (U_{uu}).

Theorem 9 (Cost at a fixed state)**.** *Let Assumption 1 hold with $k \ge 2$ and let $\boldsymbol{\xi}$ have zero mean and support in $\{\|\boldsymbol{\xi}\| \le c_0\}$. There is C such that for every state and every $\sigma \in (0,1]$, $|C_\sigma(\mathbf{s})| \le C\sigma^2$ if $\gamma\Lambda^2 < 1$, $C\sigma^2(1+\log(1/\sigma))$ if $\gamma\Lambda^2 = 1$, and $C\sigma^\alpha$ if $\gamma\Lambda^2 > 1$. If moreover the hypotheses of Theorem 3 hold with (M) and (U_{uu}), and $\boldsymbol{\xi}$ has a density with a bounded derivative, then for every open $\mathcal{U}$ with $\overline{\mathcal{U}} \subset U$ there are $c > 0$, C' and $\sigma_0 > 0$ with*

$$\Big(\frac{1}{\mathrm{vol}(\mathcal{U})}\int_{\mathcal{U}} C_\sigma(\mathbf{s})^2\, d\mathbf{s}\Big)^{1/2} \ge c\,\sigma^{\alpha_1} \qquad (\sigma \le \sigma_0), \tag{158}$$

and $\sup_{f(U)} |C_\sigma| \le C'\sigma^{\alpha_1}$ for $\sigma \le \sigma_0$.

When $\gamma\Lambda^2 > 1$ the cost at a fixed state is of order σ^{α} at most, and for a small α the bound hardly falls when the noise is reduced; on the attractor the two bounds close on the power σ^{α_1}. Averaging over the initial state changes the answer, by the mechanism of Section 4.2: a smooth density does not see the roughness of the value. On an attractor the density is carried along the expanding directions, where it keeps its derivatives, and the width w of a reset sets the size of those derivatives. A bound uniform in the centre of the reset carries the largest rate at which a displacement can grow, $\lambda_{\max} = \lim_n \frac{1}{n} \log \sup_U \|Df^n\| \geq \lambda_1$.

Theorem 10 (Cost averaged over a reset of width w)**.** *Let the closed loop satisfy (H1) with F and $\mu_{\boldsymbol{\theta}}$ four times differentiable, r twice differentiable in the action and $\bar{r}$ twice in the state, with bounded derivatives, and $\gamma e^{\lambda_{\max}} > 1$; let $\boldsymbol{\xi}$ have zero mean, covariance the identity and support in $\{\|\boldsymbol{\xi}\| \leq c_0\}$, drawn afresh at every step. Let $p_0(\mathbf{x}) = w^{-d_S}\phi((\mathbf{x}-\mathbf{x}_0)/w)$ for a smooth profile ϕ supported in the unit ball, with centre $\mathbf{x}_0$ in a compact set $\mathcal{K}_0$ of the basin, the Jacobian of f vanishing at no point of the transient of $\mathcal{K}_0$ into U, let $0 < \epsilon_0 < \lambda_u^-$, the slowest rate of expansion along E^u, and let $\alpha_+ = \log(1/\gamma)/(\lambda_{\max} + \epsilon_0)$, which is then below one. There are w_0, σ_1, C and C', independent of w, σ and $\mathbf{x}_0 \in \mathcal{K}_0$, such that for $w \leq w_0$, with J_σ and J the returns averaged over p_0,*

$$|J_\sigma(\boldsymbol{\theta}) - J(\boldsymbol{\theta})| \leq C\sigma^2 w^{\alpha_+ - 2} = C\sigma^{\alpha_+}\Big(\frac{\sigma}{w}\Big)^{2-\alpha_+} \qquad (\sigma \leq \min(w, \sigma_1)), \tag{159}$$

$$|J_\sigma(\boldsymbol{\theta}) - J(\boldsymbol{\theta})| \leq C'\sigma^{\alpha_+} \qquad (w < \sigma \leq \sigma_1). \tag{160}$$

A reset narrower than the noise gives back the order σ^{α_+} of a fixed state, and as the reset widens the bound falls toward σ^2; since the mean of the cost against a smooth density is of order σ^2 at most while the root mean square of the cost is of order σ^{α_1}, the cost takes both signs and cancels in the average.

Write Δ_a for the Laplacian in the action and $(\mathcal{M}_u\psi)(\mathbf{s},\mathbf{a}) = \mathbb{E}[\psi(\mathbf{s}, \mathbf{a}+\sqrt{u}\,\boldsymbol{\xi})]$ for the smoothing at the width $\sqrt{u}$. A natural estimate of the cost, which does not run the noisy actor, reads the curvature Δ_a of the critic, multiplies it by $\sigma^2/2$ and sums it along the deployed trajectory. For a Gaussian noise the cost is an integral of such curvatures over the widths below σ, and the estimate replaces the integral by the curvature at the single width $\tilde{\sigma}$ of the critic, of order $\tilde{\sigma}^{\alpha-2}$ at most by Theorem 4, while the cost reaches σ^{α_1} in the mean square by Eq. 158.

Theorem 11 (The cost is an integral of curvatures)**.** *Let $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, let the reward be bounded and measurable, and let $\mathbf{a} \mapsto Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s},\mathbf{a})$ be continuous at $\mu_{\boldsymbol{\theta}}(\mathbf{s})$ uniformly in $\mathbf{s}$. With $\bar{Q}_u = \mathcal{M}_u Q^{\pi_{\sqrt{u}}}$ the action value of the actor with noise $\sqrt{u}$, smoothed at the same width, and the states $\mathbf{y}_t$ of that actor from $\mathbf{y}_0 = \mathbf{s}$,*

$$C_\sigma(\mathbf{s}) = \frac{1}{2}\int_0^{\sigma^2} K(u)(\mathbf{s})\, du, \qquad K(u)(\mathbf{s}) = \sum_{t\geq 0} \gamma^t\, \mathbb{E}\big[\Delta_a \bar{Q}_u\big(\mathbf{y}_t, \mu_{\boldsymbol{\theta}}(\mathbf{y}_t)\big)\big]. \tag{161}$$

Theorem 12 (The estimate is too small by at least $(\tilde{\sigma}/\sigma)^{2-\alpha_1}$)**.** *Let the hypotheses of the second part of Theorem 9 hold, with Assumption 1 for $k \geq 3$, both noises with densities with two bounded derivatives, and the action reaching every direction of the state in N steps at every state of U. With $P(\mathbf{s}) = \frac{\sigma^2}{2}\sum_{t\geq 0}\gamma^t \Delta_a Q^{\pi_{\tilde{\sigma}}}(\mathbf{s}_t, \mu_{\boldsymbol{\theta}}(\mathbf{s}_t))$ along the deployed trajectory and $\|\cdot\|_{\mathcal{U}}$ the root mean square over an open $\mathcal{U}$ with $\overline{\mathcal{U}} \subset U$, there are $\tilde{\sigma}_*$ and $c > 0$ such that for $0 < \sigma \leq \tilde{\sigma} \leq \tilde{\sigma}_*$*

$$\frac{\|C_\sigma\|_{\mathcal{U}}}{\|P\|_{\mathcal{U}}} \geq c\Big(\frac{\tilde{\sigma}}{\sigma}\Big)^{2-\alpha_1}. \tag{162}$$

Every halving of the exploration noise at a fixed $\tilde{\sigma}$ multiplies the bound of Eq. 162 by $2^{2-\alpha_1}$.

D.1 Proofs

Proof of Theorem 9, upper bound. Write $\rho(\mathbf{s},\mathbf{u}) = r(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{u})$ and $g(\mathbf{s},\mathbf{u}) = F(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{u})$. For a random X with zero mean and $\|X\| \leq c_0\sigma$, and Ψ twice differentiable, Taylor's formula with integral remainder gives

$$\big|\mathbb{E}\Psi(X) - \Psi(0)\big| \leq \tfrac{1}{2}c_0^2\sigma^2 \sup\|D^2\Psi\|, \tag{163}$$

the first order vanishing with the mean. One injection of noise at a state $\mathbf{z}$ is followed through $\chi_j(\mathbf{u}) = f^j(g(\mathbf{z}, \mathbf{u}))$, whose two derivatives obey

$$\|\chi_j'\| \le b_0 \Lambda^j, \qquad \chi_j'' = \mathbf{A}(\chi_{j-1})\chi_{j-1}'' + D^2 f(\chi_{j-1})\big[\chi_{j-1}', \chi_{j-1}'\big], \tag{164}$$
$$\|\chi_j''\| \le K_a \Lambda^j + K_f b_0^2 \sum_{i<j} \Lambda^{j-1-i}\Lambda^{2i},$$

by the chain rule and $\|Df^j\| \le \Lambda^j$, with b_0, K_a and K_f bounds on the first and second derivatives; the sum is $\Lambda^{j-1}(\Lambda^j - 1)/(\Lambda - 1)$ for $\Lambda \neq 1$ and j for $\Lambda = 1$. With $\mathbf{y}_t$ the state the exploring actor reaches at step t and the reward at that step telescoped over the step at which the noise enters, each injection contributes at most σ^2 times the second derivative of $\rho \circ \chi_{t-1-k}$, so that

$$\big|\mathbb{E}\rho(\mathbf{y}_t, \sigma\boldsymbol{\xi}_t) - \rho(\mathbf{s}_t, 0)\big| \le \min\big\{2r_{\max},\ \sigma^2 S_t\big\}, \tag{165}$$
$$|C_\sigma(\mathbf{s})| \le \sum_{t\ge 0} \gamma^t \min\big\{2r_{\max},\ \sigma^2 S_t\big\}, \tag{166}$$

where $S_t = \sum_{k<t} \sup \|D^2(\rho \circ \chi_{t-1-k})\|$ is at most a constant times $(1+t)^2$ when $\Lambda \le 1$ and times Λ^{2t} when $\Lambda > 1$. Summing Eq. 166 as in Eqs. 62 and 63, with ratio $\gamma\Lambda^2$ and the crossing at $\sigma^2\Lambda^{2t} \asymp 1$, where $\gamma^t = \sigma^\alpha$, gives the three cases. □

Proof of Theorem 11. The Gaussian density φ_u of $\sqrt{u}\,\boldsymbol{\xi}$ satisfies the heat equation $\partial_u \varphi_u = \frac{1}{2}\Delta\varphi_u$, so the smoothing $\mathcal{M}_u$ is differentiable in $u > 0$ in operator norm on bounded functions, with derivative $\frac{1}{2}\Delta_a \mathcal{M}_u$ of norm at most $\sqrt{2d_A}/(2u)$. With $\mathcal{E}_u Q(\mathbf{s}) = \mathcal{M}_u Q(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}))$, $\mathcal{P}V(\mathbf{s}, \mathbf{a}) = V(F(\mathbf{s}, \mathbf{a}))$ and $\mathcal{T}_u = \mathcal{E}_u\mathcal{P}$,

$$V_u = V^{\pi_{\sqrt{u}}} = (I - \gamma\mathcal{T}_u)^{-1}\mathcal{E}_u r, \qquad Q^{\pi_{\sqrt{u}}} = r + \gamma\mathcal{P}V_u, \tag{167}$$
$$\partial_u V_u = (I - \gamma\mathcal{T}_u)^{-1}\tfrac{1}{2}\Delta_a \mathcal{M}_u\big(r + \gamma\mathcal{P}V_u\big)\big(\cdot, \mu_{\boldsymbol{\theta}}(\cdot)\big) = \tfrac{1}{2}K(u), \tag{168}$$
$$\|V_u - V\|_\infty \le \big\|\mathcal{E}_u Q^{\mu_{\boldsymbol{\theta}}} - V\big\|_\infty/(1-\gamma). \tag{169}$$

Eq. 168 differentiates the resolvent of Eq. 167. Integrating it from u_1 to σ^2 and letting $u_1 \to 0$ gives Eq. 161: by Eq. 169 the value V_{u_1} tends to V uniformly, since $Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s}, \cdot)$ is continuous at $\mu_{\boldsymbol{\theta}}(\mathbf{s})$ uniformly in $\mathbf{s}$, with a modulus given by Theorem 1 or 2. The identity rests on the heat equation, which is carried by the Gaussian law alone. □

Proof of Theorem 9, lower bound. Under (U_{uu}) the construction of Appendix B.5 runs at the exact rate: the strong products of the orbits near Ω, noisy or not, are within constant factors of $e^{\lambda_1 t}$, by the cones and the distortion bound with the coboundary of Eq. 157, so that V and V^{π_σ} are α_1-Hölder along the unstable directions with constants independent of σ, and the same growth in the proof of the upper bound gives $\sup_{f(U)} |C_\sigma| \le C'\sigma^{\alpha_1}$. For the lower bound we compare the cost at $\mathbf{s}$ with the cost at the state $\Phi_\kappa(\mathbf{s}) = f^{-1}(F(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}) + \kappa\sigma\mathbf{e}))$, which the loop carries to the point of the actuated curve of $\mathbf{s}$ at the parameter $\kappa\sigma$, with $\mathbf{e}$ the direction of (A). By the Bellman equation,

$$C_\sigma(\Phi_\kappa\mathbf{s}) - C_\sigma(\mathbf{s}) = \big[V^{\pi_\sigma}(\Phi_\kappa\mathbf{s}) - V^{\pi_\sigma}(\mathbf{s})\big] - \bar{r}(\Phi_\kappa\mathbf{s}) + \bar{r}(\mathbf{s}) - \gamma\big[V(F(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}) + \kappa\sigma\mathbf{e})) - V(f(\mathbf{s}))\big]. \tag{170}$$

Of the terms of Eq. 170, the reward pair $-\bar{r}(\Phi_\kappa\mathbf{s}) + \bar{r}(\mathbf{s})$ is at most a constant times $|\kappa|\sigma$, since $\bar{r}$ is Lipschitz on U and Φ_κ moves the state by at most a constant times $|\kappa|\sigma$; the other two behave differently below the scale σ:

$$\big|V^{\pi_\sigma}(\Phi_\kappa\mathbf{s}) - V^{\pi_\sigma}(\mathbf{s})\big| \le C\big(|\kappa|\sigma^{\alpha_1} + |\kappa|^{\alpha_1}\sigma^{2\alpha_1} + \sigma\big), \tag{171}$$
$$\big|V\big(F(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}) + \kappa\sigma\mathbf{e})\big) - V\big(f(\mathbf{s})\big)\big| \ge \gamma^n\big(|V(\mathbf{y}) - V(\mathbf{z}_n)| - K_r\, d - 2\omega\big). \tag{172}$$

Eq. 171 holds because the noise of the first step covers the displacement $\kappa\sigma$, so that carrying the derivative onto the density of $\boldsymbol{\xi}$, whose derivative is bounded, makes the exploring return smooth along the actuated direction. Eq. 172 is the telescoping of Appendix B.5, read at the time $n = n_\sigma$ at which $\sigma e^{\lambda_1 n}$ reaches a fixed size M_0, with $\mathbf{z}_n$ the point of Ω within a_* of $f^{n+1}\mathbf{s}$ and $\mathbf{y}$ the point at arc length d along its strong leaf. The head of Eq. 95, the rewards collected over the n steps along the two trajectories, is at most $K_r\gamma^n d$ through the length of the actuated curve, with K_r a multiple of the Lipschitz constant of $\bar{r}$, fixed by $\mathcal{U}$ and independent of $\mathbf{s}$, σ and υ. Each of the two ends of the arc costs $\omega = C_H\epsilon^{\alpha_1}$, with C_H the α_1-Hölder constant of V on $f(U)$: the move from $\mathbf{y}$ to the

pushed forward curve and the move from $\mathbf{z}_n$ to $f^{\,n+1}\mathbf{s}$ stay within the accuracy ϵ of Lemma 4. By that lemma, as $|\kappa|$ runs log-uniformly over $[\upsilon e^{-R}, \upsilon]$ with κ of either sign, for a free $\upsilon \in (0,1]$ and a width R fixed by the distortion constants of Appendix B.5, the actuated curve pushed forward covers a window of scales of the strong leaf of $\mathbf{z}_n$. Under Definition 1 the oscillation along the strong leaves, in mean square over such a window of scales, exceeds a constant on a set $\mathcal{O}_\nu$ of positive ν-measure, and by (M) and the convergence of smooth densities to ν [17] the point $\mathbf{z}_n$ falls near $\mathcal{O}_\nu$ for a set of states $\mathbf{s} \in \mathcal{U}$ whose volume, for every n beyond a fixed step, is at least a constant independent of n and σ. After the average over κ, and with $\gamma^{2n_\sigma} \geq c(\upsilon\sigma)^{2\alpha_1}$, the mean square of Eq. 170 is at least $c\,\upsilon^{2\alpha_1}\sigma^{2\alpha_1} - C(\upsilon^2\sigma^{2\alpha_1} + \upsilon^{2\alpha_1}\sigma^{4\alpha_1} + \sigma^2)$ on that set, and since $\alpha_1 < 1$ the first term wins once υ is small and σ is then small enough. Finally $\mathbf{s} \mapsto \Phi_\kappa(\mathbf{s})$ is a diffeomorphism close to the identity, so the root mean square over $\mathcal{U}$ of the left side of Eq. 170 is at most $(1+\sqrt{2})\|C_\sigma\|_{\mathcal{U}}$, which gives Eq. 158. □

Proof of Theorem 10. With $h_k = \mathcal{L}^k p_0$ the density at step k and $\mathcal{P}_\sigma$ the transfer operator of one noisy step, the difference of the two returns telescopes over the step at which the noise enters,

$$J_\sigma(\boldsymbol{\theta}) - J(\boldsymbol{\theta}) = \sum_{k\geq 0} \gamma^{k+1} \big\langle \widetilde{V}^{\pi_\sigma}, (\mathcal{P}_\sigma - \mathcal{L}) h_k \big\rangle + \Delta_{\mathrm{r}}, \qquad |\Delta_{\mathrm{r}}| \leq \frac{d_A \sigma^2 \sup \|\partial_{\mathbf{a}}^2 r\|}{2(1-\gamma)}, \tag{173}$$

with $\widetilde{V}^{\pi_\sigma}$ the value of the exploring actor started after the injection and Δ_{r} the effect of the noise on the reward itself; each $(\mathcal{P}_\sigma - \mathcal{L})h_k$ has zero mass. On an attractor h_k concentrates on Ω and loses its derivatives in the stable directions, so the pairing is read along discs in the unstable cone, over which the reset disintegrates into standard families of densities that keep two derivatives at the scale w by a Lasota–Yorke inequality. The stable holonomy of the random sequence of maps brings the displaced disc back onto the undisplaced one, and the value, twice differentiable along the random stable leaves, supplies the two derivatives the density lacks. On one disc the pairing is read twice, against the second derivative of the density along the disc and against the oscillation of the value over the support,

$$\big|\big\langle \widetilde{V}^{\pi_\sigma}, (\mathcal{P}_\sigma - \mathcal{L}) h_k \big\rangle\big| \leq C\sigma^2\, w^{-2} e^{-2\lambda_u^- k}, \tag{174}$$

$$\big|\big\langle \widetilde{V}^{\pi_\sigma}, (\mathcal{P}_\sigma - \mathcal{L}) h_k \big\rangle\big| \leq C\sigma^2 \big(w e^{(\lambda_{\max}+\epsilon_0)k}\big)^{\alpha_+}, \tag{175}$$

$$\sum_{k\geq 0} \gamma^{k+1} \big|\big\langle \widetilde{V}^{\pi_\sigma}, (\mathcal{P}_\sigma - \mathcal{L}) h_k \big\rangle\big| \leq C\sigma^2 w^{\alpha_+ - 2}. \tag{176}$$

Eq. 176 splits the sum at the step where $we^{(\lambda_{\max}+\epsilon_0)k}$ is of order one, with Eq. 175 before it and Eq. 174 after. The transient from $\mathcal{K}_0$ into the trapping region takes a fixed number of injective steps. For $w < \sigma$ the adjoint form of the pairing, with the Hölder bound of the value at the exponent α_+, gives $C'\sigma^{\alpha_+}$. □

Proof of Theorem 12. For $\sigma \leq \tilde{\sigma} \leq \tilde{\sigma}_*$,

$$\sup_U \big|\Delta_a Q^{\pi_{\tilde{\sigma}}}\big| \leq C\,\tilde{\sigma}^{\,\alpha_1 - 2}, \tag{177}$$

$$\|P\|_{\mathcal{U}} \leq C\,\sigma^2 \tilde{\sigma}^{\,\alpha_1 - 2}, \tag{178}$$

$$\|C_\sigma\|_{\mathcal{U}} \geq c\,\sigma^{\alpha_1}, \tag{179}$$

$$\frac{\|C_\sigma\|_{\mathcal{U}}}{\|P\|_{\mathcal{U}}} \geq \frac{c}{C}\Big(\frac{\tilde{\sigma}}{\sigma}\Big)^{2-\alpha_1}. \tag{180}$$

Eq. 177 is the integration by parts of Lemma 11, carried out with the oscillation of the value at the exponent α_1 itself, which (U_{uu}) gives along the noisy orbits; Eq. 178 sums it along the deployed trajectory with the weight $\sigma^2/2$; Eq. 179 is Eq. 158. Their quotient, Eq. 180, is Eq. 162. □

E THE MEASUREMENTS IN FULL

Section 5 reports the agreement between the predicted and the measured exponent over the whole grid. Figures 3 to 6 carry the same measurements in more detail. Tables 5 and 6 give the readings of every agent with its bootstrap interval, Table 2 the exponent of every task at the five discounts, with the agents whose copies part and those whose copies never do in separate columns, and Tables 3 to 5 the numbers behind the rest.

task	actors	part	never	undet.	unread.	D	α_S	0.5	0.8	0.9	0.95	0.99
HalfCheetah	6	6	0	0	0	1.8	0.03 / —	0.98 / —	0.58 / —	0.29 / —	0.13 / —	0.01 / —
Ant	6	6	0	0	0	2.1	0.03 / —	1.04 / —	0.72 / —	0.32 / —	0.15 / —	0.03 / —
Walker2d	6	6	0	0	0	9.1	0.21 / —	1.00 / —	0.98 / —	1.00 / —	0.66 / —	0.23 / —
Hopper	6	5	1	0	0	19.9	0.38 / 1.00	1.00 / 0.99	1.00 / 0.98	1.00 / 1.02	0.98 / 1.07	0.40 / 1.08
Swimmer	6	5	1	0	0	22.5	0.17 / 0.98	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	0.93 / 1.00	0.19 / 0.95
InvertedDoublePendulum	6	4	2	0	0	4.8	0.11 / 0.82	1.03 / 1.00	0.99 / 0.98	0.86 / 0.90	0.55 / 0.88	0.15 / 0.83
Pendulum	6	1	5	0	0	2.3	0.04 / 1.00	0.99 / 1.00	0.79 / 1.00	0.40 / 1.00	0.18 / 1.00	0.04 / 1.00

Table 2: The two-copy test task by task: the actors of each task and how many of them the test reads as parting, as never parting, as undetermined and as unreadable, then the columns of Table 1 for the agents whose copies part and, after the slash, for those whose copies never do; the delay D is read on the parting agents alone.

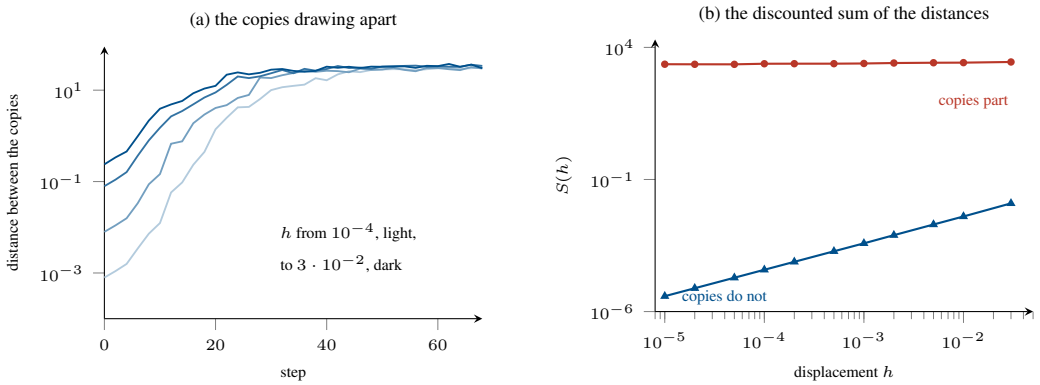

Figure 3: The two readings of the copies. (a) The distance between them on one HalfCheetah actor, at four displacements of the ladder, median over the states. (b) The discounted sum $S(h)$ of those distances (Eq. 15), on that actor, where its slope in these axes is α_S, and on a Pendulum actor whose copies never part.

F Experimental protocol

Every measurement is made after training, in double precision, from internal states that the trained actor visits, and no measurement discards an agent: the 42 actors are every actor we trained.

The agents. The actors are trained by our implementation of TD3, which follows the reference code of its authors [2]: actor and critics have two hidden layers of 256 units, the actor a $\tanh$ output rescaled to the action box, and they are trained with Adam at 3×10^{-4} on minibatches of $B = 256$ transitions, with $\gamma = 0.99$, exploration noise $\sigma = 0.1$, target noise $\tilde{\sigma} = 0.2$ clipped at 0.5, policy delay $d = 2$, $\tau = 0.005$ and 25,000 uniform steps before learning starts; both noises are fractions of the half-width of each action coordinate. Training runs for 300,000 steps, 60,000 on Pendulum, where the uniform steps are the first 1,000, and the measurements use the final checkpoint. The two critic activations are the relu of the reference code and softplus, the actor being a relu network in both cases. Four of the seven tasks, HalfCheetah, Ant, Hopper and Walker2d, have a foot that touches the ground, and three, Pendulum, InvertedDoublePendulum and Swimmer, do not; three seeds and two critic activations give six actors per task, 42 over the seven tasks.

States and copies. The states of a measurement are internal states of the simulator taken one every ten steps along the deterministic trajectories of the trained actor from the initial states of eight seeds: 120 states for the two-copy measurement, fifteen on each trajectory, 40 for the slope of the smoothed value, five on each trajectory, and 120 for the resolution of the critic. Restoring such a state and replaying the same actions reproduces the same trajectory to the last digit, which is what lets two copies be compared.

The two copies. Both copies start in the state $\mathbf{s}$; the first plays $\mu_{\boldsymbol{\theta}}(\mathbf{s})$, the second $\mu_{\boldsymbol{\theta}}(\mathbf{s}) + h\mathbf{e}$ with $\mathbf{e}$ the unit direction of the first action coordinate, the sign of h reversed if the displaced action leaves the box, and both follow $\mu_{\boldsymbol{\theta}}$ for 1500 steps or until the episode ends. The distance between the copies is the Euclidean distance between the simulator states: velocities, and positions with the free

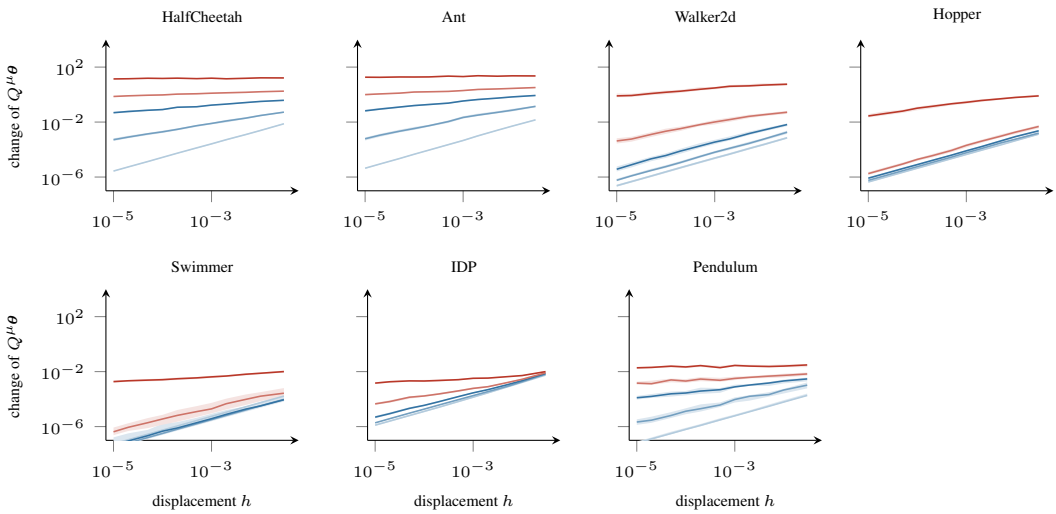

Figure 4: The change of the action value along the ladder of displacements, one panel per task and one line per discount, over the states of the agents of the task whose copies part, with the interval resampling them leaves; the shades run from $\gamma = 0.5$, light, to $\gamma = 0.99$, dark, and the slope of a line is the exponent at that discount.

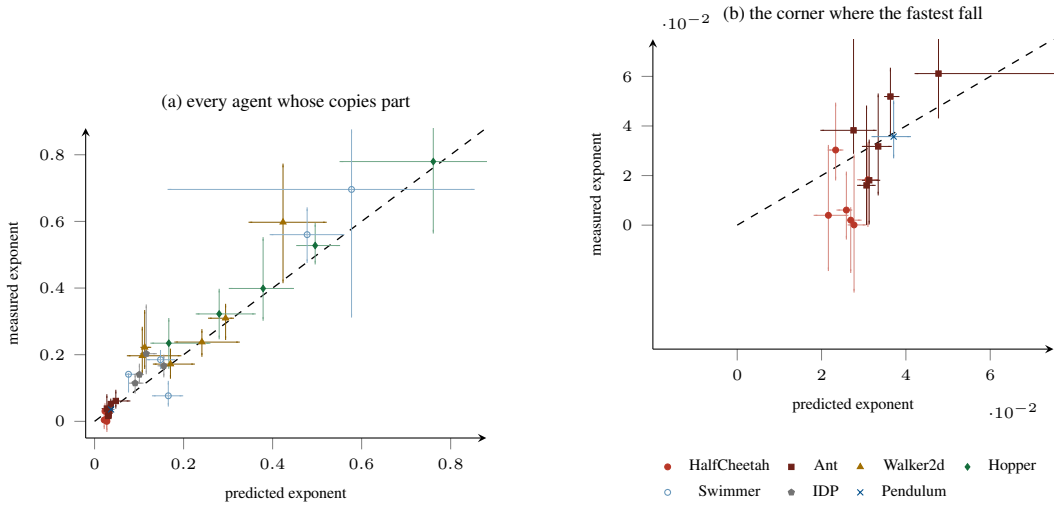

Figure 5: The measured exponent against the predicted one at the discount the agents were trained at, with the interval of both readings, on the 33 agents whose copies part; the line is equality and (b) is the corner of (a).

translations left out. The displacements are

$$h \in \{10^{-5},\ 2\times10^{-5},\ 5\times10^{-5},\ 10^{-4},\ 2\times10^{-4},\ 5\times10^{-4},\\ 10^{-3},\ 2\times10^{-3},\ 5\times10^{-3},\ 10^{-2},\ 3\times10^{-2}\}, \tag{181}$$

the same on every task, with a run at $h = 0$ that checks that the two copies reproduce each other to the last digit. The parting step is read at each state with the rule of Appendix B.6, and the delay D of an agent is the slope $\hat{D}$ of Eq. 106.

The action value and the five discounts. The action value of Eq. 2 is the discounted sum of the rewards of a copy: $\Delta Q(h)$ is the difference of the discounted rewards of the two copies over the 1500 steps, and the same rewards read with $\gamma = 0.5$, 0.8, 0.9, 0.95 and 0.99 give it under five weightings of the future. The exponent $\hat{\alpha}$ is the least-squares slope of the median over the states of $\log|\Delta Q(h)|$ against $\log h$, over the displacements at most 5×10^{-3}, the nine smallest of Eq. 181, where the points still lie on a line; the interval of Table 5 is the 16th to 84th percentile of a bootstrap over the states.

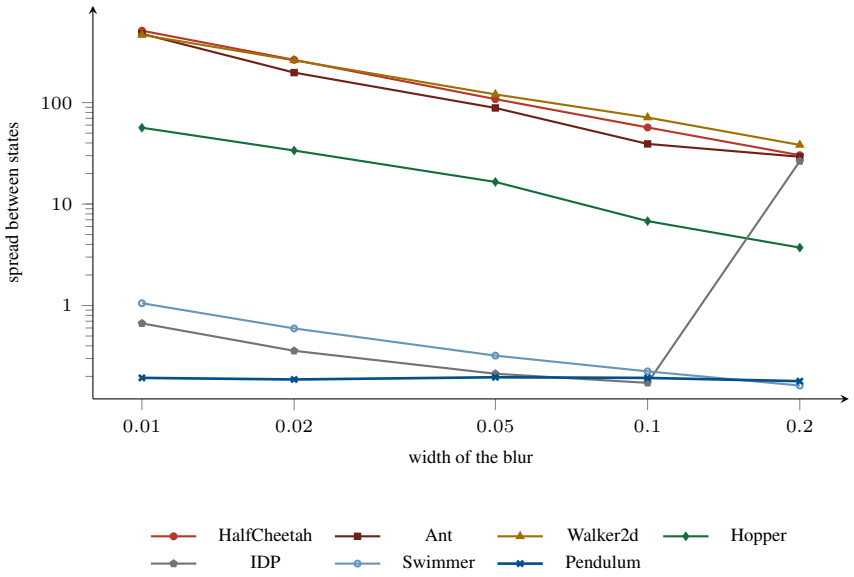

Figure 6: The spread between states of the slope of the smoothed action value against the width of the smoothing, median actor per task, all seven tasks on one axis. The reading at the widest width on InvertedDoublePendulum, IDP in the legend, is not usable: three pairs of states in a thousand carry the whole of that spread.

regime	cases	mean	median	16-84	asserted
curvature	54	2.05	2.05	[1.92, 2.20]	2
slope	21	1.52	1.50	[1.29, 1.72]	(1, 2)
neither	90	0.37	0.28	[0.10, 0.67]	α
—	45	1.97	2.00	[1.91, 2.18]	—

Table 3: The exponent of the second difference of the action value, grouped by the regime the two copies predict: curvature where $\gamma\Lambda^2 < 1$, slope where $\gamma\Lambda < 1 < \gamma\Lambda^2$, neither where $\gamma\Lambda > 1$. One case per agent and discount, the interval the 16th to 84th percentile over the group, and the last row the agents whose copies do not part.

The slope of the smoothed action value. The action value is smoothed in the first coordinate of the action by a Gaussian of width $b\,w_1$, with w_1 the half-width of the action box on that coordinate, and the slope $g_b(\mathbf{s})$ of the smoothed value at $\mathbf{a}_0 = \mu_{\boldsymbol{\theta}}(\mathbf{s})$ is estimated by

$$\frac{1}{K}\sum_{k\le K} z_k\,\frac{Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s},\mathbf{a}_0 + b\,w_1 z_k \mathbf{e}) - Q^{\mu_{\boldsymbol{\theta}}}(\mathbf{s},\mathbf{a}_0 - b\,w_1 z_k \mathbf{e})}{2b\,w_1}, \qquad z_k \sim \mathcal{N}(0,1), \tag{182}$$

the derivative of the Gaussian average written as an average against z [38], in the antithetic form that cancels the even part of the return. Each $Q^{\mu_{\boldsymbol{\theta}}}$ is a rollout of 1500 steps of the trained actor after the displaced first action. The displaced action is clipped to the box before it reaches the simulator, so that beyond the box the rollout reads the value at the boundary while z stays Gaussian. The same z_k serve every width, so that the widths are paired. The widths are $b \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$ half-widths, with $K = 96$ pairs at each of the 40 states. The spread between the states is the root of the variance of g_b over them with the variance of the estimator subtracted, $\sqrt{\max\{\mathrm{var}_{\mathbf{s}}\, g_b(\mathbf{s}) - \overline{\mathrm{se}^2}, 0\}}$, where $\overline{\mathrm{se}^2}$ is the mean squared standard error of the estimates, and the exponent of the spread is the least-squares slope in $\log b$. Over the agents beyond the threshold that exponent is negative on all but three, all of them on InvertedDoublePendulum, where the widest width carries a spread of its own; of the nine agents whose copies do not part, the five on Pendulum have an exponent within 0.05 of zero and the other four, one on Hopper, one on Swimmer and two on InvertedDoublePendulum, have exponents between -0.77 and $+0.97$ with intervals that do not contain zero.

The resolution of the critics. For a trained critic and a state $\mathbf{s}$, $g(\mathbf{s})$ is the slope $\partial_{a_1} Q_{\boldsymbol{\phi}}(\mathbf{s}, \mu_{\boldsymbol{\theta}}(\mathbf{s}))$ that the critic hands the actor, read on the head that the actor follows in the update, and $g_w(\mathbf{s})$ the

task	critic	ℓ	range	$\ell > \tilde{\sigma}$	ℓ_s	range
HalfCheetah	relu	0.072	[0.069, 0.107]	0	0.018	[0.069, 0.107]
HalfCheetah	softplus	0.615	[0.530, 0.700]	3	0.132	[0.530, 0.700]
Hopper	relu	0.056	[0.035, 0.057]	0	0.010	[0.035, 0.057]
Hopper	softplus	0.188	[0.174, 0.401]	1	0.086	[0.174, 0.401]
Walker2d	relu	0.046	[0.046, 0.055]	0	0.012	[0.046, 0.055]
Walker2d	softplus	0.629	[0.350, 0.853]	3	0.114	[0.350, 0.853]
Swimmer	relu	0.024	[0.018, 0.040]	0	0.008	[0.018, 0.040]
Swimmer	softplus	0.573	[0.338, 0.679]	3	0.102	[0.338, 0.679]
Ant	relu	0.035	[0.022, 0.040]	0	0.012	[0.022, 0.040]
Ant	softplus	0.353	[0.324, 0.422]	3	0.155	[0.324, 0.422]
InvertedDoublePendulum	relu	0.004	[0.003, 0.009]	0	0.003	[0.003, 0.009]
InvertedDoublePendulum	softplus	0.014	[0.014, 0.020]	0	—	[0.014, 0.020]
Pendulum	relu	0.006	[0.003, 0.009]	0	0.004	[0.003, 0.009]
Pendulum	softplus	0.081	[0.063, 0.107]	0	0.018	[0.063, 0.107]

Table 4: The resolution of the critic by task and activation: ℓ in half-widths of the action box and ℓ_s in deviations of the states the actor visits, the scale of Proposition 5. Median and range over the readings of the three seeds that fall inside the widths swept, a dash where none does, and $\ell > \tilde{\sigma}$ the count of the seeds coarser in the action than the target noise.

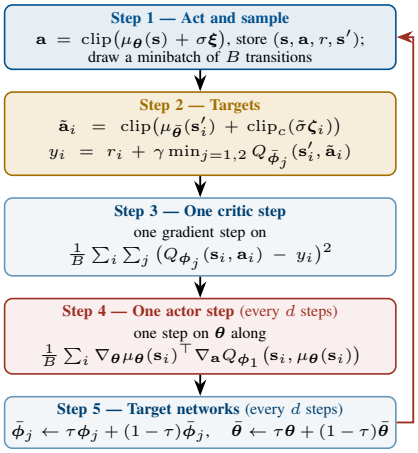

Figure 7: One environment step of TD3 [2], the loop that trains the agents measured here; DDPG is the case $\tilde{\sigma} = 0$ with one critic and $d = 1$.

same slope after smoothing the critic in that coordinate by a Gaussian of width w in half-widths of the box, estimated as in Eq. 182 with a control variate of known mean, the pairs doubled from 512 up to 16,384 until the standard error falls below a fiftieth of $|g_w(\mathbf{s})|$. Readings that never reach that accuracy are left out. The resolution ℓ is the width at which the median over the states of $|g_w(\mathbf{s}) - g(\mathbf{s})|/|g(\mathbf{s})|$ reaches a tenth, interpolated in $\log w$ on a grid of 25 widths from 0.001 to 1; states whose slope is below a hundredth of the median slope are left out. The resolution ℓ_s in the state is read in the same way, with the critic blurred in the state: $g_w(\mathbf{s})$ is then the mean of $\partial_{a_1} Q_{\boldsymbol{\phi}}$ over states drawn around $\mathbf{s}$ with deviation w times the deviation of each state coordinate over the states the actor visits, so that the widths are in deviations of the state, and the slope itself is the exact derivative of the network. Draws are added until the standard error is below a fiftieth of that mean. The smoothing of the critic in the action is Gaussian and unclipped, so at the largest widths it reaches outside the action box, where the critic was never trained; only the first coordinate of the action is smoothed, so on the many-actuator tasks ℓ describes that direction; and the critic is the one at the end of training, read on the states of the final actor.

task	critic	seed	copies	D	from D	α_S	measured	interval
Ant	relu	0	parts	2.8	0.04	0.05	0.06	[0.04, 0.09]
Ant	relu	1	parts	2.2	0.03	0.04	0.05	[0.04, 0.06]
Ant	relu	2	parts	2.0	0.03	0.03	0.02	[0.00, 0.05]
Ant	softplus	0	parts	2.0	0.03	0.03	0.02	[0.00, 0.03]
Ant	softplus	1	parts	1.8	0.03	0.03	0.04	[0.03, 0.08]
Ant	softplus	2	parts	2.2	0.03	0.03	0.03	[0.01, 0.05]
HalfCheetah	relu	0	parts	1.6	0.02	0.02	0.00	[-0.02, 0.03]
HalfCheetah	relu	1	parts	1.7	0.02	0.03	0.01	[-0.01, 0.02]
HalfCheetah	relu	2	parts	2.1	0.03	0.03	0.02	[0.00, 0.03]
HalfCheetah	softplus	0	parts	1.5	0.02	0.02	0.03	[0.02, 0.05]
HalfCheetah	softplus	1	parts	1.9	0.03	0.03	0.00	[-0.02, 0.01]
HalfCheetah	softplus	2	parts	2.2	0.03	0.03	0.00	[-0.03, 0.03]
Hopper	relu	0	never	—	1.00	1.00	1.08	[1.01, 1.17]
Hopper	relu	1	parts	37.3	0.54	0.50	0.53	[0.48, 0.59]
Hopper	relu	2	parts	19.9	0.29	0.28	0.32	[0.25, 0.39]
Hopper	softplus	0	parts	6.5	0.09	0.17	0.23	[0.16, 0.30]
Hopper	softplus	1	parts	27.6	0.40	0.76	0.78	[0.57, 0.98]
Hopper	softplus	2	parts	19.0	0.28	0.38	0.40	[0.31, 0.55]
InvertedDoublePendulum	relu	0	never	—	1.00	0.99	0.98	[0.96, 1.00]
InvertedDoublePendulum	relu	1	never	—	1.00	0.65	0.67	[0.59, 0.74]
InvertedDoublePendulum	relu	2	parts	3.9	0.06	0.10	0.14	[0.12, 0.17]
InvertedDoublePendulum	softplus	0	parts	5.3	0.08	0.16	0.17	[0.14, 0.19]
InvertedDoublePendulum	softplus	1	parts	5.8	0.08	0.12	0.20	[0.15, 0.35]
InvertedDoublePendulum	softplus	2	parts	4.2	0.06	0.09	0.11	[0.09, 0.14]
Pendulum	relu	0	parts	2.3	0.03	0.04	0.04	[0.03, 0.05]
Pendulum	relu	1	never	—	1.00	1.00	1.00	[1.00, 1.00]
Pendulum	relu	2	never	—	1.00	1.00	0.99	[0.98, 1.02]
Pendulum	softplus	0	never	—	1.00	1.00	1.07	[1.07, 1.07]
Pendulum	softplus	1	never	—	1.00	1.00	1.00	[1.00, 1.00]
Pendulum	softplus	2	never	—	1.00	1.00	1.00	[1.00, 1.00]
Swimmer	relu	0	parts	7.4	0.11	0.08	0.14	[0.09, 0.14]
Swimmer	relu	1	parts	22.5	0.33	0.17	0.08	[0.05, 0.12]
Swimmer	relu	2	parts	6.7	0.10	0.15	0.19	[0.16, 0.21]
Swimmer	softplus	0	parts	38.1	0.55	0.48	0.56	[0.48, 0.64]
Swimmer	softplus	1	parts	41.6	0.60	0.58	0.70	[0.32, 0.87]
Swimmer	softplus	2	never	—	1.00	0.98	0.95	[0.88, 1.00]
Walker2d	relu	0	parts	6.0	0.09	0.11	0.22	[0.16, 0.33]
Walker2d	relu	1	parts	15.6	0.23	0.29	0.31	[0.25, 0.35]
Walker2d	relu	2	parts	8.2	0.12	0.24	0.24	[0.20, 0.27]
Walker2d	softplus	0	parts	13.9	0.20	0.42	0.60	[0.42, 0.77]
Walker2d	softplus	1	parts	10.0	0.14	0.17	0.17	[0.13, 0.21]
Walker2d	softplus	2	parts	3.7	0.05	0.11	0.20	[0.15, 0.28]

Table 5: The 42 agents one by one, as the two copies read them, with the exponent measured on the action value at $\gamma = 0.99$ and the interval of a bootstrap over the states. The columns are rounded to two decimals, which on eight agents puts the exponent from D on an endpoint of the interval, so the coverage counted from this table is not the one of Section 5, which is read at full precision.

task	critic	seed	at 1000	curvature	regime	response	together
Ant	relu	0	0.06	0.05	neither	1.1	18
Ant	relu	1	0.05	0.04	neither	1.4	22
Ant	relu	2	0.02	0.04	neither	1.2	5
Ant	softplus	0	0.02	0.01	neither	1.0	0
Ant	softplus	1	0.04	0.06	neither	1.1	13
Ant	softplus	2	0.03	0.05	neither	1.3	0
HalfCheetah	relu	0	0.00	0.03	neither	1.1	120
HalfCheetah	relu	1	0.01	0.01	neither	1.3	120
HalfCheetah	relu	2	0.02	0.02	neither	1.0	120
HalfCheetah	softplus	0	0.03	0.06	neither	1.3	120
HalfCheetah	softplus	1	0.00	-0.00	neither	1.1	120
HalfCheetah	softplus	2	0.00	0.02	neither	1.1	120
Hopper	relu	0	1.08	1.94	—	963.9	82
Hopper	relu	1	0.52	0.56	neither	28.5	108
Hopper	relu	2	0.32	0.30	neither	7.9	34
Hopper	softplus	0	0.23	0.20	neither	4.0	0
Hopper	softplus	1	0.79	0.75	neither	161.7	2
Hopper	softplus	2	0.40	0.42	neither	10.5	15
InvertedDoublePendulum	relu	0	0.98	2.35	—	463.5	120
InvertedDoublePendulum	relu	1	0.67	0.97	—	73.7	120
InvertedDoublePendulum	relu	2	0.14	0.17	neither	2.7	120
InvertedDoublePendulum	softplus	0	0.17	0.17	neither	3.5	120
InvertedDoublePendulum	softplus	1	0.20	0.22	neither	3.5	120
InvertedDoublePendulum	softplus	2	0.11	0.11	neither	1.9	120
Pendulum	relu	0	0.04	0.08	neither	1.2	120
Pendulum	relu	1	1.00	2.23	—	513.4	120
Pendulum	relu	2	0.99	2.32	—	474.7	120
Pendulum	softplus	0	1.07	2.00	—	829.1	120
Pendulum	softplus	1	1.00	1.87	—	495.2	120
Pendulum	softplus	2	1.00	1.99	—	492.5	120
Swimmer	relu	0	0.14	0.08	neither	2.8	120
Swimmer	relu	1	0.08	0.10	neither	1.4	120
Swimmer	relu	2	0.19	0.24	neither	3.6	120
Swimmer	softplus	0	0.56	0.51	neither	35.8	120
Swimmer	softplus	1	0.69	0.73	neither	55.4	120
Swimmer	softplus	2	0.98	1.95	—	429.4	120
Walker2d	relu	0	0.22	0.20	neither	3.5	7
Walker2d	relu	1	0.31	0.29	neither	7.1	0
Walker2d	relu	2	0.24	0.21	neither	4.7	12
Walker2d	softplus	0	0.60	0.47	neither	29.5	12
Walker2d	softplus	1	0.17	0.19	neither	2.8	0
Walker2d	softplus	2	0.20	0.16	neither	3.9	15

Table 6: The same 42 agents, as the action value and the two copies read them. "at 1000" is the measured exponent of Table 5 with the sum cut at a thousand steps, "response" the slope of the smoothed action value, and "together" the number of states at which both copies terminate at the same step.