1 Proofs

We now give proofs for all the results presented in the paper. Most proofs follow standard patterns from calculus and numerical schemes for differential equations, except for Theorem 3, which uses an argument specific to reinforcement learning to prove that the continuous-time advantage function contains all the necessary information for policy improvement.

The first result presented is a proof of convergence for discretized trajectories.

Lemma 1. *Let $F\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n$ and $\pi\colon \mathcal{S} \to \mathcal{A}$ be the dynamic and policy functions. Assume that, for any a, $s \to F(s,a)$ and $s \to F(s, \pi(s))$ are $\mathcal{C}^1$, bounded and K-lipschitz. For a given s_0, define the trajectory $(s_t)_{t\geq 0}$ as the unique solution of the differential equation*

$$\frac{ds_t}{dt} = F(s_t, \pi(s_t)). \tag{1}$$

For any $\delta t > 0$, define the discretized trajectory $(s^k_{\delta t})_k$ which amounts to maintaining each action for a time interval δt; it is defined by induction as $s^0_{\delta t} = s_0$, $s^{k+1}_{\delta t}$ is the value at time δt of the unique solution of

$$\frac{d\tilde{s}_t}{dt} = F(\tilde{s}_t, \pi(s^k_{\delta t})) \tag{2}$$

with initial point $s^k_{\delta t}$. Then, there exists $C > 0$ such that, for every $t \geq 0$

$$\|s_t - s^{\lfloor \frac{t}{\delta t} \rfloor}_{\delta t}\| \leq \delta t \frac{C}{K} e^{Kt}. \tag{3}$$

Therefore, discretized trajectories converge pointwise to continuous trajectories.

Proof. The proof mostly follos the classical argument for convergence of the Euler scheme for differential equations. For any k, define

$$e^k_{\delta t} = \|s^k_{\delta t} - s_{\delta t k}\|. \tag{4}$$

Let $\tilde{s}_t$ be the solution of Eq. (62) with initial state s^k_δ. This $\tilde{s}_t$ is $\mathcal{C}^2$ on $[0, \delta t]$. Consequently, the Taylor integral formula gives

$$s^{k+1}_{\delta t} = s^k_{\delta t} + F(s^k_{\delta t}, \pi(s^k_{\delta t}))\delta t + \int_0^{\delta t} (\delta t - t) \frac{d^2 \tilde{s}_t}{dt^2} dt \tag{5}$$

$$s_{\delta t(k+1)} = s_{\delta t k} + F(s_{\delta t k}, \pi(s_{\delta t k}))\delta t + \int_0^{\delta t} (\delta t - t) \frac{d^2 s_{t+\delta t k}}{dt^2} dt. \tag{6}$$

Now, both $d^2 s_t/dt^2$ and $d^2 \tilde{s}_t/dt^2$ are uniformly bounded, by boundedness and Lipschitzness of $s \to F(s, \pi(s))$ and $s \to F(s, \pi(s^k_{\delta t}))$. Consequently, there exists C such that

$$e^{k+1}_{\delta t} \leq \|s^k_{\delta t} - s_{\delta t k}\| + \|F(s^k_{\delta t}, \pi(s^k_{\delta t})) - F(s_{\delta t k}, \pi(s_{\delta t k}))\|\delta t + C\delta t^2 \tag{7}$$

$$\leq (1 + K\delta t)e^k_{\delta t} + C\delta t^2. \tag{8}$$

Now, it is easy to prove by induction that

$$e^k_{\delta t} \leq (1 + K\delta t)^k (e^0_{\delta t} + \frac{C}{K}\delta t) - \frac{C}{K}\delta t. \tag{9}$$

As $e^0_{\delta t} = 0$, this translates to

$$e^k_{\delta t} \leq ((1 + K\delta t)^k - 1)\delta t \frac{C}{K} \tag{10}$$

$$\leq (e^{K\delta t k} - 1)\delta t \frac{C}{K}. \tag{11}$$

Consequently,

$$e^{\lfloor t/\delta t \rfloor}_{\delta t} \leq (e^{K(t+\delta t)} - 1)\delta t \frac{C}{K}. \tag{12}$$

Finally, by boundedness, of $s \to F(s, \pi(s))$, there exists C' such that

$$\|s_{\delta t \lfloor t/\delta t \rfloor} - s_t\| \leq \delta t C'. \tag{13}$$

Combining Eq. (13) with Eq. (12), one can find C'' such that

$$\|s_t - s^{\lfloor t/\delta t \rfloor}_{\delta t}\| \leq \delta t \frac{C''}{K} e^{Kt}. \tag{14}$$

□

In what follows, we assume that the continuous-time reward function $r \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is bounded, to ensure existence of V^π and $V^\pi_{\delta t}$ for all δt .

Theorem 1. *Assume that $r \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is bounded, and that $s \to r(s, \pi(s))$ is L_r-Lipschitz continuous, then for all $s \in \mathcal{S}$, one has $V^\pi_{\delta t}(s) = V^\pi(s) + o(1)$ when $\delta t \to 0$.*

Proof. Let $\tilde{s}^t_{\delta t} := s^{\lfloor t/\delta t \rfloor}_{\delta t}$. We have:

$$V^\pi_{\delta t}(s) = \int_t \gamma^t r(\tilde{s}^t_{\delta t}, \pi(\tilde{s}^t_{\delta t}))dt + O(\delta t) \tag{15}$$

Indeed:

$$V_{\delta t}^{\pi}(s) = \sum_{k=0}^{\infty} \gamma^{k\delta t} r(s_{\delta t}^{k}, \pi(s_{\delta t}^{k}))\delta t \tag{16}$$

$$= \sum_{k=0}^{\infty} \gamma^{k\delta t} \int_{u=k}^{k+1} r(\tilde{s}_{\delta t}^{u\delta t}, \pi(\tilde{s}_{\delta t}^{u\delta t}))du \tag{17}$$

$$= \sum_{k=0}^{\infty} \frac{\delta t \log \gamma}{\gamma^{\delta t} - 1} \int_{u=k}^{k+1} \gamma^{u\delta t} r(\tilde{s}_{\delta t}^{u\delta t}, \pi(\tilde{s}_{\delta t}^{u\delta t}))du \tag{18}$$

$$= \frac{\delta t \log \gamma}{\gamma^{\delta t} - 1} \int_{t=0}^{\infty} \gamma^{t} r(\tilde{s}_{\delta t}^{t}, \pi(\tilde{s}_{\delta t}^{t}))dt \tag{19}$$

$$\tag{20}$$

But:

$$\frac{\delta t \log \gamma}{\gamma^{\delta t} - 1} = \frac{\delta t \log \gamma}{\delta t \log \gamma + O(\delta t^2)} \tag{21}$$

$$= 1 + O(\delta t) \tag{22}$$

$$\tag{23}$$

Therefore:

$$V_{\delta t}^{\pi}(s) = \int_{t} \gamma^{t} r(\tilde{s}_{\delta t}^{t}, \pi(\tilde{s}_{\delta t}^{t}))dt + O(\delta t) \tag{24}$$

We now have, for any $T > 0$,

$$|V_{\delta t}^{\pi}(s) - V^{\pi}(s)| = |\int_{t=0}^{\infty} \gamma^{t} \left(r(\tilde{s}_{\delta t}^{t}, \pi(\tilde{s}_{\delta t}^{t})) - r(s_t, \pi(s_t)) \right) dt| + O(\delta t) \tag{25}$$

$$= |\int_{t=0}^{T} \gamma^{t} \left(r(\tilde{s}_{\delta t}^{t}, \pi(\tilde{s}_{\delta t}^{t})) - r(s_t, \pi(s_t)) \right) dt| \tag{26}$$

$$+ |\int_{t=T}^{\infty} \gamma^{t} \left(r(\tilde{s}_{\delta t}^{t}, \pi(\tilde{s}_{\delta t}^{t})) - r(s_t, \pi(s_t)) \right) dt| + O(\delta t) \tag{27}$$

The second term can be bounded by the supremum of the reward:

$$|\int_{t=T}^{\infty} \gamma^{t} \left(r(\tilde{s}_{\delta t}^{t}, \pi(\tilde{s}_{\delta t}^{t})) - r(\tilde{s}_t, \pi(s_t)) \right) dt| \leq 2 \frac{\|r\|_{\infty}}{\log(\frac{1}{\gamma})} \gamma^{T} \tag{28}$$

The first term can be bounded by using Lemma. 1:

$$|\int_{t=0}^{T} \gamma^t \left(r(\tilde{s}^t_{\delta t}, \pi(\tilde{s}^t_{\delta t})) - r(s_t, \pi(s_t))\right) dt| \leq \int_{t=0}^{T} \gamma^t L_r \|s_t - \tilde{s}^t_{\delta t}\| dt \tag{29}$$

$$\leq \int_{t=0}^{T} L_r \frac{C\delta t}{K} \exp((K + \log\gamma)t) dt \tag{30}$$

$$\leq \frac{L_r C}{K(K + \log\gamma)} \exp((K + \log\gamma)T)\delta t \tag{31}$$

Let us set $T := -\frac{1}{K}\log(\delta t)$. By plugging into Eq. (28), we have:

$$|\int_{t=T}^{\infty} \gamma^t \left(r(\tilde{s}^t_{\delta t}, \pi(\tilde{s}^t_{\delta t})) - r(s_t, \pi(s_t))\right) dt| = O(\delta t^{-\log\gamma}) = o(1). \tag{32}$$

By plugging T into equation (31), we have:

$$|\int_{t=0}^{T} \gamma^t \left(r(\tilde{s}^t_{\delta t}, \pi(\tilde{s}^t_{\delta t})) - r(s_t, \pi(s_t))\right) dt| = O(\delta t^{-\frac{\log\gamma}{K}}) = o(1), \tag{33}$$

yielding our result.

□

For the following proof, we further assume that both V^π and $V^\pi_{\delta t}$ are continuously differentiable, and that the gradient and Hessian of $V^\pi_{\delta t}$ w.r.t. s are uniformly bounded in both s and δt. We also assume convergence of $\partial_s V^\pi_{\delta t}(s)$ to $\partial_s V^\pi(s)$ for all s.

Theorem 2. *Under the hypothesis above, there exists $A^\pi : \mathcal{S} \to \mathbb{R}$ such that $A^\pi_{\delta t}$ converges pointwise to A^π as δt goes to 0. Besides,*

$$A^\pi(s, a) = r(s, a) + \partial_s V^\pi(s) F(s, a) + \log\gamma V^\pi(s). \tag{34}$$

Proof. Denote $\tilde{s}^t_{\delta t}(s_0)$ the evaluation at instant t of the solution of $d\tilde{s}_t/dt = F(\tilde{s}_t, \pi(s_0))$ with starting point s_0.

The Bellman equation on $Q^\pi_{\delta t}$ yields

$$Q^\pi_{\delta t}(s, a) = r(s, a)\delta t + \gamma^{\delta t} V^\pi_{\delta t}(\tilde{s}^{\delta t}_{\delta t}(s)). \tag{35}$$

For all s, a first-order Taylor expansion yields

$$\tilde{s}^{\delta t}_{\delta t}(s) = s + F(s, a)\delta t + O(\delta t^2) \tag{36}$$

where the constant in $O()$ is uniformly bounded thanks to the assumptions on the Hessian. Thus, by uniform boundedness of the Hessian of $V^\pi_{\delta t}$,[1]

$$Q^\pi_{\delta t}(s,a) = r(s,a)\delta t + (1+\ln(\gamma)\delta t + O(\delta t^2))(V^\pi_{\delta t}(s) + \delta t \partial_s V^\pi_{\delta t}(s) F(s,a) + O(\delta t^2). \tag{37}$$

Now, this yields

$$A^\pi_{\delta t}(s,a) = r(s,a) + \ln(\gamma) V^\pi_{\delta t}(s) + \partial_s V^\pi_{\delta t}(s) F(s,a) + O(\delta t), \tag{38}$$

and using the convergence of $V^\pi_{\delta t}(s)$ to $V^\pi(s)$ (Thm. 1) and $\partial_s V^\pi_{\delta t}(s)$ to $\partial_s V^\pi(s)$ (hypothesis) yields the result with

$$A^\pi(s,a) = r(s,a) + \ln(\gamma) V^\pi(s) + \partial_s V^\pi(s) F(s,a). \tag{39}$$

□

We now show that policy improvement works with the continous time advantage function, i.e.

Theorem 3. *Let π and π' be two policies such that both $s \to r(s,\pi(s))$ and $s \to r(s,\pi'(s))$ are continuous. Assume that both V^π and $V^{\pi'}$ are continuously differentiable. Define the advantage function for policies π and π' as in Eq.* (39).

If for all s, $A^\pi(s,\pi'(s)) \geq 0$, then for all s, $V^\pi(s) \leq V^{\pi'}(s)$. Moreover, if for all s, $V^{\pi'}(s) > V^\pi(s)$, then there exists s' such that $A^\pi(s',\pi'(s')) > 0$.

Proof. Let $(s_t)_{t\geq 0}$ be a trajectory sampled from π' i.e. solution of the equation

$$ds_t/dt = F(s_t, \pi'(s_t)) \tag{40}$$

with initial condition $s_0 = s$.

Define

$$B(T) = \int_{t=0}^{T} \gamma^t r(s_t, \pi'(s_t)) dt + \gamma^T V^\pi(s_T). \tag{41}$$

This function if continuously differentiable, and its derivative is

$$\dot{B}(T) = \gamma^T r(s_T, \pi'(s_T)) + \gamma^T \partial_s V^\pi(s) F(s, \pi'(s)) + \gamma^T \ln(\gamma) V^\pi(s_T) \tag{42}$$

$$= \gamma^T A^\pi(s_T, \pi'(s_T)) \geq 0. \tag{43}$$

[1] Without boundedness of the Hessian, we cannot write the second order Taylor expansion of $V^\pi_{\delta t}(\tilde{s}^{\delta t}_{\delta t}(s))$ in term of δt.

Thus B is increasing, and $B(0) = V^{\pi}(s)$, $\lim_{T\to\infty} B(t) = V^{\pi'}(s)$. Consequently, $V^{\pi}(s) \leq V^{\pi'}(s)$. Furthermore, if $V^{\pi}(s) < V^{\pi'}(s)$, then there exists T_0 such that $\dot{B}(T_0) > 0$ (otherwise B is constant), and $A^{\pi}(s_{T_0}, \pi'(s_{T_0})) > 0$.

□

Theorem 4. *Let $\mathcal{A} = \mathbb{R}^{\mathcal{A}}$ be the action space, and let $\mathcal{P}_1 = \mathbb{R}^{p_1}$ and $\mathcal{P}_2 = \mathbb{R}^{p_2}$ be parameter spaces. Let $A\colon \mathcal{P}_1 \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $V\colon \mathcal{P}_2 \times \mathcal{S} \to \mathbb{R}$ be $\mathcal{C}^2$ function approximators with bounded gradients and Hessians. Let $(a_t)_{t\geq 0}$ be a $\mathcal{C}^1$ exploratory action trajectory and $(s_t)_{t\geq 0}$ the resulting state trajectory, when starting from s_0 and following $ds_t/dt = F(s_t, a_t)$. Let $\theta^k_{\delta t}$ and $\psi^k_{\delta t}$ be the discrete parameter trajectories resulting from the gradient descent steps in the main text, with learning rates $\eta^V = \alpha^V \delta t^\beta$ and $\eta^A = \alpha^A \delta t^\beta$ for some $\beta \geq 0$. Then,*

- *If $\beta = 1$ the discrete parameter trajectories converge to continuous parameter trajectories as δt goes to 0.*
- *If $\beta > 1$, parameter trajectories become stationary as δt goes to 0.*
- *If $\beta < 1$, parameters can grow arbitrarily large after an arbitrarily small physical time when δt goes to 0.*

Proof. Let $(s_t, a_t)_{t\geq 0}$ be the trajectory on which parameters are learnt. To simplify notations, define

$$A_\psi(s, a) = \bar{A}_\psi(s, a) - \bar{A}_\psi(s, \pi(s)). \tag{44}$$

Define F as

$$F^\theta(\theta, \psi, s, a) = \alpha^V (r(s, a) + \ln(\gamma) V_\theta(s) + \partial_s V_\theta(s) F(s, a) - A_\psi(s, a)) \partial_\theta V_\theta(s) \tag{45}$$

$$F^\psi(\theta, \psi, s, a) = \alpha^A (r(s, a) + \ln(\gamma) V_\theta(s) + \partial_s V_\theta(s) F(s, a) - A_\psi(s, a)) \partial_\psi A_\psi(s, a). \tag{46}$$

From the bounded Hessians and Gradients hypothesis, V, A, $\partial_s V$, $\partial_\theta V$ and $\partial_\psi A$ are uniformly Lipschitz continuous in θ and ψ, thus F is Lipschitz continuous.

The discrete equations for parameters updates with learning rates $\alpha^V \delta t^\beta$

and $\alpha^A \delta t^\beta$ are

$$\delta Q = r(s_{k\delta t}, a_{k\delta t})\delta t + \gamma^{\delta t} V_{\theta^k_{\delta t}}(s_{(k+1)\delta t}) - V_{\theta^k_{\delta t}}(s_{k\delta t}) - A_\psi(s_{k\delta t}, a_{k\delta t}) \tag{47}$$

$$\theta^{k+1}_{\delta t} = \theta^k_{\delta t} + \alpha^V \delta t^\beta \frac{\delta Q}{\delta t} \partial_\theta V_{\theta^k_{\delta t}}(s_{k\delta t}) \tag{48}$$

$$\psi^{k+1}_{\delta t} = \psi^k_{\delta t} + \alpha^A \delta t^\beta \frac{\delta Q}{\delta t} \partial_\psi A_{\theta^k_{\delta t}}(s_{k\delta t}, a_{k\delta t}) \tag{49}$$

Under uniform boundedness of the Hessian of $s \mapsto V_\theta(s)$, one can show

$$\begin{pmatrix} \theta^{k+1}_{\delta t} \\ \psi^{k+1}_{\delta t} \end{pmatrix} = \begin{pmatrix} \theta^k_{\delta t} \\ \psi^k_{\delta t} \end{pmatrix} + \delta t^\beta F(\theta^k_{\delta t}, \psi^k_{\delta t}, s_{k\delta t}, a_{k\delta t}) + O(\delta t^\beta \delta t), \tag{50}$$

with a O independent of k. With the additional hypothesis that the gradient of $(s, a) \to \bar{A}_\psi(s, a)$ is uniformly bounded, we have

- For $\beta = 1$, a proof scheme identical to that of Thm. 1 shows that discrete trajectories converge pointwise to continuous trajectories defined by the differential equation

  $$\frac{d}{dt} \begin{pmatrix} \theta_t \\ \psi_t \end{pmatrix} = F(\theta_t, \psi_t, s_t, a_t), \tag{51}$$

  which admits unique solutions for all initial parameters, since F is uniformly lipschitz continuous.

- Similarly, for $\beta > 1$, the proof scheme of Thm. 1 shows that discrete trajectories converge pointwise to continuous trajectories defined by the differential equation

  $$\frac{d}{dt} \begin{pmatrix} \theta_t \\ \psi_t \end{pmatrix} = 0 \tag{52}$$

  and thus that trajectories shrink to a single point as δt goes to 0.

We now turn to proving that when $\beta < 1$, trajectories can diverge instantly in physical time. Consider the following continuous MDP,

$$s_t = \sin(t) \tag{53}$$

whatever the actions, with reward 0 everywhere and $0 < \gamma < 1$. The resulting value function is $V(s) = 0$ (since there are no actions, V is independent

of a policy), and the advantage function is 0. We consider the function approximator $V_\theta(s) = \theta s$ (which can represent the true value function). The update rule for θ is

$$\delta Q_{\delta t}^k = \gamma^{\delta t}\theta_{\delta t}^k \sin((k+1)\delta t) - \theta_{\delta t}^k \sin(k\delta t) \tag{54}$$

$$\theta_{\delta t}^{k+1} = \theta_{\delta t}^k + \alpha \delta t^\beta \frac{\gamma^{\delta t}\theta_{\delta t}^k \sin((k+1)\delta t) - \theta_{\delta t}^k \sin(k\delta t)}{\delta t} \sin(k\delta t) \tag{55}$$

Set $K_{\delta t} := \lfloor \delta t^{-\frac{\beta+3}{4}} \rfloor$, then for all $k \leq K_{\delta t}$, $o(k\delta t) = o(1)$ and

$$\theta_{\delta t}^{k+1} = \theta_{\delta t}^k (1 + \alpha \delta t^\beta (1 + o(1)) \sin(k\delta t)) \tag{56}$$

$$\tag{57}$$

Let $\rho_{\delta t}^k := \log \theta_{\delta t}^k$. Then

$$\rho_{\delta t}^k = \rho_{\delta t}^k + \alpha k \delta t^{\beta+1} + o(k\delta t^{\beta+1}). \tag{58}$$

Finally,

$$\rho_{\delta t}^{K_{\delta t}} = \rho_{\delta t}^0 + \alpha \frac{K_{\delta t}(K_{\delta t}+1)}{2} \delta t^\beta + o(K_{\delta t}^2 \delta t^{\beta+1}) \tag{59}$$

$$= \rho_\delta^0 + \alpha \delta t^{\frac{\beta-1}{3}} + o(\delta t^{\frac{\beta-1}{3}}) \tag{60}$$

$$\xrightarrow[\delta t \to 0]{} +\infty. \tag{61}$$

Thus parameters diverge in an infinitesimal physical time when δt goes to 0. □

Theorem 5. *Let $F \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n$ be the dynamic, and $\pi \colon \mathcal{S} \times \mathcal{A} \to [0,1]$ be the policy, such that $\pi(s,\cdot)$ is a probability distribution over $\mathcal{A}$. Assume that F is C^1 with bounded derivatives, and that π is C^1 and bounded. For any $\delta t > 0$, define the discretized trajectory $(s_{\delta t}^k)_k$ which amounts to sample an action from $\pi(s,\cdot)$ and maintaining each action for a time interval δt; it is defined by induction as $s_{\delta t}^0 = s_0$, $s_{\delta t}^{k+1}$ is the value at time δt of the unique solution of*

$$\frac{d\tilde{s}_t}{dt} = F(\tilde{s}_t, a_k) \tag{62}$$

with $a_k \sim \pi(s_{\delta t}^k, \cdot)$ and initial point $s_{\delta t}^k$.

Then the agent's trajectories converge when $\delta t \to 0$ to the solutions of the deterministic *equation:*

$$\frac{ds_t}{dt} = \mathbb{E}_{a \sim \pi(s_t,\cdot)} F(s_t, a). \tag{63}$$

Notably, if π is an epsilon greedy strategy that mixes a deterministic exploitation policy $\pi^{deterministic}$ with an action taken from a noise policy π^{noise} with probability ε at, the trajectory converge to the solutions of the equation:

$$ds_t/dt = (1-\varepsilon)F(s_t, \pi^{deterministic}(s_t)) + \varepsilon \mathbb{E}_{a\sim\pi^{noise}(a|s)} F(s_t, a) \tag{64}$$

Proof. Consider $(s_{\delta t^2})$ the random trajectory of a near-continuous MDP with time-discretization δt^2 obtained by taking at each step k an action a_k along $a_k \sim \pi(a|s^k_{\delta t^2})$ independantly. We have:

$$s^{\lfloor 1/\delta t \rfloor}_{\delta t^2} = s^0_{\delta t^2} + \sum_{k=1}^{\lfloor 1/\delta t \rfloor} s^k_{\delta t^2} - s^{k-1}_{\delta t^2} + O(\delta t^2) \tag{65}$$

$$= s^0_{\delta t^2} + \sum_{k=1}^{\lfloor 1/\delta t \rfloor} F(s^{k-1}_{\delta t^2}, a_{k-1})\delta t^2 + O(\delta t^2) \tag{66}$$

We define $f(s) := \mathbb{E}_{a\sim\pi(s)}[F(s,a)] = \int_{a\in\mathcal{A}} F(s,a)\pi(s,a)$. Since π and F are bounded and C^1, we know that f is C^1. We have:

$$s^{\lfloor 1/\delta t \rfloor}_{\delta t^2} = s^0_{\delta t^2} + \sum_{k=1}^{\lfloor 1/\delta t \rfloor} f(s^{k-1}_{\delta t^2})\delta t^2 + \sum_{k=1}^{\lfloor 1/\delta t \rfloor} (F(s^{k-1}_{\delta t^2}, a_{k-1}) - f(s^{k-1}_{\delta t^2}))\delta t^2 + O(\delta t^2) \tag{67}$$

$$s^{\lfloor 1/\delta t \rfloor}_{\delta t^2} = s^0_{\delta t^2} + \sum_{k=1}^{\lfloor 1/\delta t \rfloor} f(s^{k-1}_{\delta t^2})\delta t^2 + \xi + O(\delta t^2) \tag{68}$$

with $\xi := \delta t^2 \sum_{k=1}^{\lfloor 1/\delta t \rfloor} \left(F(s^{k-1}_{\delta t^2}, a_{k-1}) - f(s^{k-1}_{\delta t^2}) \right)$. By definition, we have $\mathbb{E}[\xi] = 0$. Moreover, by using the independance of actions and the boundness of F, there is $\sigma > 0$ such that:

$$\mathbb{E}[\|\xi\|^2] \leq \sigma^2 \delta t^3 \tag{69}$$

We know that f is C^1 on a compact space. Therefore, there is L_f such that f is L_f Lipschitz, and we have:

$$\| \left(\sum_{k=1}^{\lfloor 1/\delta t \rfloor} f(s^{k-1}_{\delta t^2})\delta t \right) - f(s^0_{\delta t^2}) \| \leq \delta t L_f \sum_{k=1}^{\lfloor 1/\delta t \rfloor} \| s^{k-1}_{\delta t^2} - s^0_{\delta t^2} \| \tag{70}$$

Since F is bounded, we know that $\|s^k_{\delta t^2} - s^{k-1}_{\delta t^2}\| \leq C\delta t$. Therefore:

$$\| \left(\sum_{k=1}^{\lfloor 1/\delta t \rfloor} f(s^{k-1}_{\delta t^2})\delta t \right) - f(s^0_{\delta t^2})\| \leq \delta t L_f C \sum_{k=1}^{\lfloor 1/\delta t \rfloor} k\delta t \tag{71}$$

$$= O(\delta t^2) \tag{72}$$

Therefore:

$$s^{\lfloor 1/\delta t \rfloor}_{\delta t^2} = s^0_{\delta t^2} + f(s^0_{\delta t^2})\delta t + \xi + O(\delta t^2) \tag{73}$$

Therefore, we have $a > 0$ such that $\|s^{\lfloor 1/\delta t \rfloor}_{\delta t^2} - s^0_{\delta t^2} - f(s^0_{\delta t^2})\delta t\| \leq \|\xi\| + a\delta t^2$

We define $(\tilde{s}_{\delta t})$ the deterministic near-continuous process with time discretization δt defined by $\tilde{s}^{k+1}_{\delta t} := s^k_{\delta t} + f(s^k_{\delta t})\delta t$. We have:

$$\|s^{(k+1)\lfloor 1/\delta t \rfloor}_{\delta t^2} - \tilde{s}^{k+1}_{\delta t}\| \leq \|s^{(k+1)\lfloor 1/\delta t \rfloor}_{\delta t^2} - s^{k\lfloor 1/\delta t \rfloor}_{\delta t^2} - f(s^{k\lfloor 1/\delta t \rfloor}_{\delta t^2})\delta t\| + \|s^{k\lfloor 1/\delta t \rfloor}_{\delta t^2} + f(s^{k\lfloor 1/\delta t \rfloor}_{\delta t^2})\delta t - \tilde{s}^{k+1}_{\delta t}\| \tag{74}$$

We know that $\|s^{(k+1)\lfloor 1/\delta t \rfloor}_{\delta t^2} - s^{k\lfloor 1/\delta t \rfloor}_{\delta t^2} - f(s^{k\lfloor 1/\delta t \rfloor}_{\delta t^2})\delta t\| \leq \|\xi_k\| + a\delta t^2$. Moreover:

$$\|s^{k\lfloor 1/\delta t \rfloor}_{\delta t^2} + f(s^{k\lfloor 1/\delta t \rfloor}_{\delta t^2})\delta t - \tilde{s}^{k+1}_{\delta t}\| \leq \|s^{k\lfloor 1/\delta t \rfloor}_{\delta t^2} - \tilde{s}^k_{\delta t}\| + \delta t \|f(s^{k\lfloor 1/\delta t \rfloor}_{\delta t^2}) - f(\tilde{s}^k_{\delta t})\| \tag{75}$$

$$\leq (1 + L_f\delta t)\|s^{k\lfloor 1/\delta t \rfloor}_{\delta t^2} - \tilde{s}^k_{\delta t}\| \tag{76}$$

Therefore, we have:

$$\|s^{(k+1)\lfloor 1/\delta t \rfloor}_{\delta t^2} - \tilde{s}^{k+1}_{\delta t}\| \leq \|\xi_k\| + a\delta t^2 + (1 + L_f\delta t)\|s^{k\lfloor 1/\delta t \rfloor}_{\delta t^2} - \tilde{s}^k_{\delta t}\| \tag{77}$$

By induction, and by taking $k = \lfloor t/\delta t \rfloor$:

$$\|s^{k\lfloor 1/\delta t \rfloor}_{\delta t^2} - \tilde{s}^k_{\delta t}\| \leq \frac{a\delta t}{L_f} \exp(L_f t) + \sum_{j=0}^{\lfloor t/\delta t \rfloor} (1 + \delta t L_f)^j \|\xi_j\| \tag{78}$$

Therefore, if $\varepsilon > 0$, we have :

$$\mathbb{P}\left(\|s_{\delta t^2}^{k\lfloor 1/\delta t\rfloor} - \tilde{s}_{\delta t}^{k}\| > \varepsilon\right) \leq \mathbb{P}\left(\sum_{j=0}^{\lfloor t/\delta t\rfloor}(1+\delta t L_f)^j \|\xi_j\| > \varepsilon - \frac{a\delta t}{L_f}\exp(L_f t)\right) \tag{79}$$

$$\leq \frac{\mathbb{E}\left[\sum_{j=0}^{\lfloor t/\delta t\rfloor}(1+\delta t L_f)^j \|\xi_j\|\right]}{\varepsilon - \frac{a\delta t}{L_f}\exp(L_f t)} \tag{80}$$

$$\leq \frac{\mathbb{E}\left[\|\xi\|\right]}{\varepsilon - \frac{a\delta t}{L_f}\exp(L_f t)}\frac{\exp(L_f t)}{L_f \delta t} \tag{81}$$

$$\tag{82}$$

But $\mathbb{E}\left[\|\xi\|\right] \leq \sqrt{\mathbb{E}\left[\|\xi\|^2\right]} \leq \sigma\delta t^{3/2}$. Therefore, we have:

$$\mathbb{P}\left(\|s_{\delta t^2}^{\lfloor t/\delta t\rfloor\lfloor 1/\delta t\rfloor} - \tilde{s}_{\delta t}^{k}\| > \varepsilon\right) = O(\sqrt{\delta t}) \tag{83}$$

Therefore, the process $t \mapsto s_{\delta t^2}^{\lfloor t/\delta t\rfloor\lfloor 1/\delta t\rfloor}$ converges in probability to $\tilde{s}$. Furthermore, by a similar argument than in Lemma 1, we know that the discretized process $\tilde{s}$ converge to the continuous process defined by $\frac{ds}{dt} = f(s_t)$. We can conclude ou result.

□

2 Implementation details

All the details specifying our implementation are given in this section. We first give precise pseudo code descriptions for both *Continuous Deep Advantage Updating* (Alg. 1), as well as the variants of DDPG (Alg. 2) and DQN (Alg. 3) used.

For DDPG and DQN, two different settings were experimented with:

- One with time discretization scalings, to keep the comparison fair. In this setting, the discount factor is still scaled as $\gamma^{\delta t}$, rewards are scaled as $r\delta t$, and learning rates are scaled to obtain parameter updates of order δt. As RMSprop is used for all experiments, this amounts to using a learning rate scaling as $\alpha^Q = \tilde{\alpha}^Q \delta t$, $\alpha^\pi = \tilde{\alpha}^\pi \delta t$.

- One without discretization scalings. In that case, only the discount factor is scaled as $\gamma^{\delta t}$, to prevent unfair shortsightedness. All other

Algorithm 1 Discrete DAU

Inputs:
θ and ψ, parameters of V_θ and $\bar{A}_\psi$.
π^{explore} and $\nu_{\delta t}$ defining an exploration policy.
$\mathbf{opt}_V$, $\mathbf{opt}_A$, $\alpha^V \delta t$ and $\alpha^A \delta t$ optimizers and learning rates.
$\mathcal{D}$, buffer of transitions (s, a, r, d, s'), with d the episode termination signal.
δt and γ, time discretization and discount factor.
nb_epochs number of epochs.
nb_steps, number of steps per epoch.
nb_learn, number of learning step per epoch
N, training batch size

Observe initial state s^0
$t \leftarrow 0$
for $e = 0,$ **nb_epochs do**
  for $j = 1,$ **nb_steps do**
    $a^t \leftarrow \pi^{\text{explore}}(s^t, \nu^t_{\delta t})$.
    Perform a^t and observe $(r^{t+1}, d^{t+1}, s^{t+1})$.
    Store $(s^t, a^t, r^{t+1}, d^{t+1}, s^{t+1})$ in $\mathcal{D}$.
    $t \leftarrow t + 1$
  end for
  for $k = 0,$ **nb_learn do**
    Sample a batch of N random transitions from $\mathcal{D}$
    for i=0, N **do**
      $$\delta Q^i \leftarrow \delta t\big(\bar{A}_\psi(s^i, a^i) - \bar{A}_\psi(s^i, \pi_\phi(s^i))\big) - \Big(r^i \delta t + (1 - d^i)\gamma^{\delta t} V_\theta(s'^i) - V_\theta(s^i)\Big)$$
    end for
    $$\Delta\theta \leftarrow \frac{1}{N} \sum_{i=1}^{N} \frac{\delta Q^i \partial_\theta V_\theta(s^i)}{\delta t}$$
    $$\Delta\psi \leftarrow \frac{1}{N} \sum_{i=1}^{N} \frac{\delta Q^i \partial_\psi \left(\bar{A}_\psi(s^i, a^i) - \max_{a'} \bar{A}_\psi(s^i, a')\right)}{\delta t}$$
    Update θ with $\mathbf{opt}_1$, $\Delta\theta$ and learning rate $\alpha^V \delta t$.
    Update ψ with $\mathbf{opt}_2$, $\Delta\psi$ and learning rate $\alpha^A \delta t$.
  end for
end for

Algorithm 2 DDPG

Inputs:

ψ and ϕ, parameters of Q_ψ and π_ϕ.

ψ' and ϕ', parameters of target networks $Q_{\psi'}$ and $\pi_{\phi'}$.

π^{explore} and ν defining an exploration policy.

$\mathbf{opt}_Q$, $\mathbf{opt}_\pi$, α^Q and α^π, optimizers and learning rates.

$\mathcal{D}$, buffer of transitions (s, a, r, d, s'), with d the episode termination signal.

γ discount factor.

τ target network update factor.

nb_epochs number of epochs.

nb_steps, number of steps per epoch.

Observe initial state s^0

$t \leftarrow 0$

for $e = 0$, **nb_epochs do**

  for $j = 1$, **nb_steps do**

    $a^k \leftarrow \pi^{\text{explore}}(s^k, \nu^k)$.

    Perform a^k and observe $(r^{k+1}, d^{k+1}, s^{k+1})$.

    Store $(s^k, a^k, r^{k+1}, d^{k+1}, s^{k+1})$ in $\mathcal{D}$.

    $k \leftarrow k + 1$

  end for

  for $k = 0$, nb_learn **do**

    Sample a batch of N random transitions from $\mathcal{D}$

    $\tilde{Q}^i \leftarrow r^i + (1 - d^i)\gamma Q_{\psi'}(s'^i, \pi_{\phi'}(s'^i))$

    $\Delta\psi \leftarrow \frac{1}{N} \sum_{i=1}^{N} \left(Q^i - \tilde{Q}^i \right) \partial_\psi Q(s^i, a^i)$

    $\Delta\phi \leftarrow \frac{1}{N} \sum_{i=1}^{N} \partial_a Q_\psi(s^i, \pi_\phi(s^i)) \partial_\phi \pi_\phi(s^i)$

    Update ψ with $\mathbf{opt}_Q$, $\Delta\psi$ and learning rate α^Q.

    Update ϕ with $\mathbf{opt}_\pi$, $\Delta\phi$ and learning rate α^π.

    $\psi' \leftarrow \tau\psi' + (1 - \tau)\psi$

    $\phi' \leftarrow \tau\phi' + (1 - \tau)\phi$

  end for

end for

Algorithm 3 DQN

Inputs:

ψ parameter of Q_ψ.
ψ', parameters of target networks $Q_{\psi'}$.
π^{explore} and ν defining an exploration policy.
$\mathbf{opt}_Q$, α^Q optimizer and learning rate.
$\mathcal{D}$, buffer of transitions (s, a, r, d, s'), with d the episode termination signal.
γ discount factor.
τ target network update factor.
nb_epochs number of epochs.
nb_steps, number of steps per epoch.

Observe initial state s^0
$t \leftarrow 0$
for $e = 0,$ **nb_epochs do**
  for $j = 1,$ **nb_steps do**
    $a^k \leftarrow \pi^{\text{explore}}(s^k, \nu^k)$.
    Perform a^k and observe $(r^{k+1}, d^{k+1}, s^{k+1})$.
    Store $(s^k, a^k, r^{k+1}, d^{k+1}, s^{k+1})$ in $\mathcal{D}$.
    $k \leftarrow k + 1$
  end for
  for $k = 0,$ nb_learn **do**
    Sample a batch of N random transitions from $\mathcal{D}$
    $\tilde{Q}^i \leftarrow r^i + (1 - d^i)\gamma \max_{a'} Q_{\psi'}(s'^i, a')$
    $\Delta\psi \leftarrow \frac{1}{N} \sum_{i=1}^{N} \left(Q^i - \tilde{Q}^i\right) \partial_\psi Q(s^i, a^i)$
    Update ψ with $\mathbf{opt}_Q$, $\Delta\psi$ and learning rate α^Q.
    $\psi' \leftarrow \tau\psi' + (1 - \tau)\psi$
  end for
end for

parameters are set with a reference $\delta t_0 = 1e-2$. For instance, for all δt's, the reward perceived is $r * \delta t_0$, and similarily for learning rates, $\alpha^Q = \tilde{\alpha}^Q \delta t_0$, $\alpha^\pi = \tilde{\alpha}^Q \delta t_0$. These scalings don't depend on the discretization, but perform decently at least for the highest discretization.

2.1 Global hyperparameters

The following hyperparameters are maintained constant throughout all our experiments,

- All networks used are of the form

```
Sequential(
    Linear(nb_inputs, 256),
    LayerNorm(256),
    ReLU(),
    Linear(256, 256),
    LayerNorm(256),
    ReLU(),
    Linear(256, nb_outputs)
).
```

  Policy networks have an additional tanh layer to constraint action range. On certain environments, network inputs are normalized by applying a mean-std normalization, with mean and standard deviations computed on each individual input features, on all previously encountered samples.

- $\mathcal{D}$ is a cyclic buffer of size 1000000.

- **nb_steps** is set to 10, and 256 environments are run in parallel to accelerate the training procedure, totalling 2560 environment interactions between learning steps.

- **nb_learn** is set to 50.

- The physical γ is set to 0.8. It is always scaled as $\gamma^{\delta t}$ (even for unscaled DQN and DDPG).

- N, the batch size is set to 256.

- RMSprop is used as an optimizer without momentum, and with $\alpha = 1 - \delta t$ (or $1 - \delta t_0$ for unscaled DDPG and DQN).
- Exploration is always performed as described in the main text. The OU process used as parameters $\kappa = 7.5$, $\sigma = 1.5$.
- Unless otherwise stated, $\alpha_1 := \tilde{\alpha}^Q = \alpha^V = \alpha^A = 0.1$, $\alpha_2 := \tilde{\alpha}^\pi = \alpha^\pi = 0.03$.
- $\tau = 0.9$

2.2 Environment dependent hyperparameters

We hereby list the hyperparameters used for each environment. Continuous actions environments are marked with a (C), discrete actions environments with a (D).

- **Ant (C)**: State normalization is used. Discretization range: $[0.05, 0.02, 0.01, 0.005, 0.002]$.
- **Cheetah (C)**: State normalization is used. Discretization range: $[0.05, 0.02, 0.01, 0.005, 0.002]$
- **Bipedal Walker (C)**[2]: State normalization is used, $\alpha_2 = 0.02$. Discretization range: $[0.01, 0.005, 0.002, 0.001]$.
- **Cartpole (D)**: $\alpha_2 = 0.02$, $\tau = 0$. Discretization range: $[0.01, 0.005, 0.002, 0.001, 0.0005]$.
- **Pendulum (C)**: $\alpha_2 = 0.02$, $\tau = 0$. Discretization range: $[0.01, 0.005, 0.002, 0.001, 0.0005]$.

3 Additional results

Additional results mentionned in the text are presented in this section.

[2]The reward for Bipedal Walker is modified not to scale with δt. This does not introduce any change for the default setup.

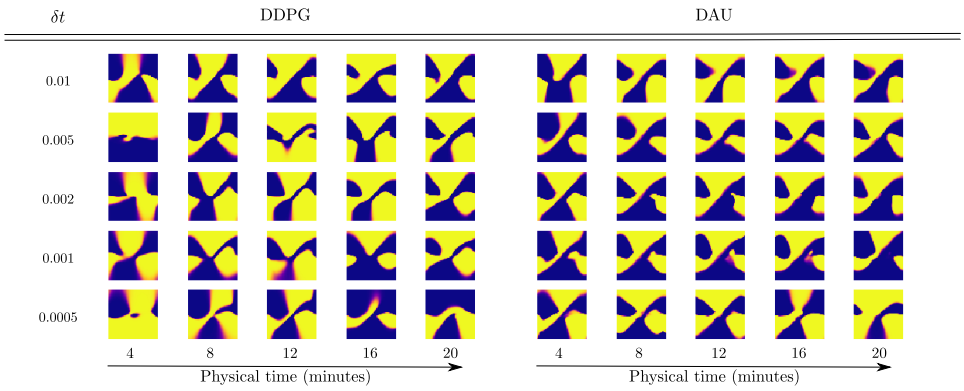

Figure 1: Policies obtained by DDPG (unscaled version) and AU at different instants in physical time of training on the pendulum swing-up environment. Each image represents the policy learnt by the policy network, with x-axis representing angle, and y-axis angular velocity. The lighter the pixel, the closer to 1 the action, the darker, the closer to -1.

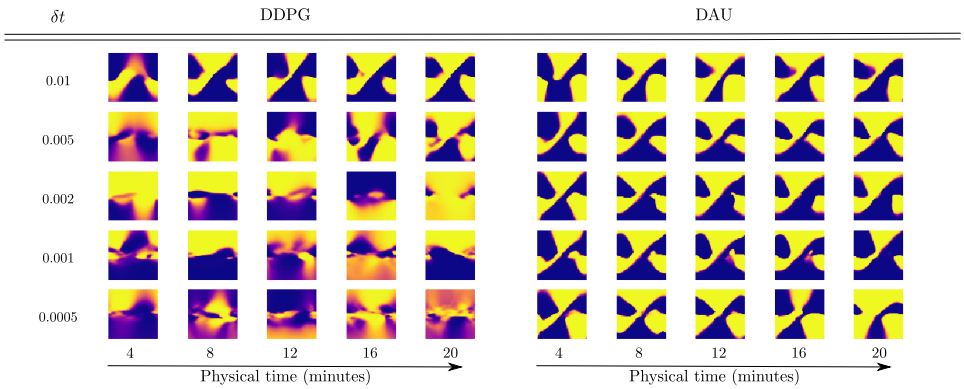

Figure 2: Policies obtained by DDPG (scaled version) and AU at different instants in physical time of training on the pendulum swing-up environment. Each image represents the policy learnt by the policy network, with x-axis representing angle, and y-axis angular velocity. The lighter the pixel, the closer to 1 the action, the darker, the closer to -1.

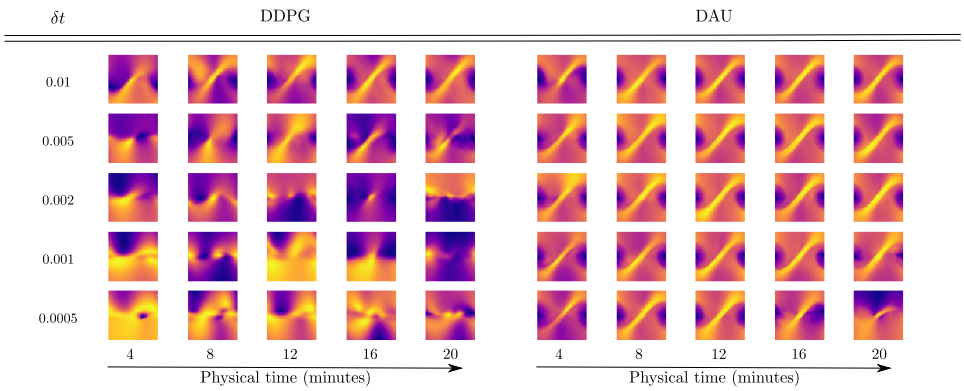

Figure 3: Value functions obtained by DDPG (scaled version) and AU at different instants in physical time of training on the pendulum swing-up environment. Each image represents the value function learnt, with x-axis representing angle, and y-axis angular velocity. The lighter the pixel, the higher the value.

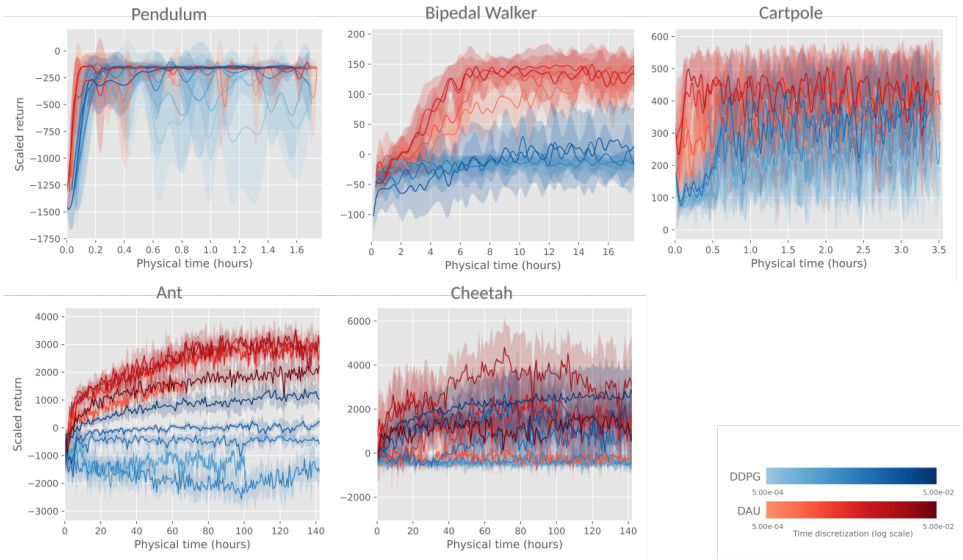

Figure 4: Learning curves for DAU and DDPG (scaled) on classic control benchmarks for various time discretization δt: Scaled return as a function of the physical time spent in the environment.