Temporal Difference Learning with Continuous Time and State in the Stochastic Setting

Ziad Kobeissi, Francis Bach

Temporal Difference Learning with Continuous Time and State in the Stochastic Setting

Ziad Kobeissi
Inria & École Normale Supérieure
Institut Louis Bachelier, Paris, France
ziad.kobeissi@inria.fr

Francis Bach
Inria & École Normale Supérieure
PSL Research University, Paris, France
francis.bach@inria.fr

Abstract

We consider the problem of continuous-time policy evaluation. This consists in learning through observations the value function associated to an uncontrolled continuous-time stochastic dynamic and a reward function. We propose two original variants of the well-known TD(0) method using vanishing time steps. One is model-free and the other is model-based. For both methods, we prove theoretical convergence rates that we subsequently verify through numerical simulations. Alternatively, those methods can be interpreted as novel reinforcement learning approaches for approximating solutions of linear PDEs (partial differential equations) or linear BSDEs (backward stochastic differential equations).

1 Introduction

Consider the value function V obtained from a continuous time and state stochastic process $(X_t)_{t\geq 0}$,

$$V(x) = \mathbb{E}\left[\int_0^\infty e^{-\rho t} r(X_t)\, dt \middle| X_0 = x\right] \quad \text{with} \quad dX_t = b(X_t)dt + \sigma(X_t)dW_t, \tag{1}$$

where $r, b, \sigma : \Omega \to \mathbb{R}, \mathbb{R}^d, \mathbb{R}^{d\times d_W}$ are respectively the reward, drift and diffusion functions, $\rho > 0$ is the exponential discount rate and W is a d_W-dimensional Brownian motion. The state $(X_t)_{t\geq 0}$ is a continuous stochastic process valued in a continuous state space Ω and satisfies a stochastic differential equation (SDE). For readers who are not familiar with SDEs, a heuristic interpretation of *"X is a solution to the right-hand side of* (1)*"* could be: given $X_t \in \Omega$ at time $t \geq 0$, between t and $t + dt$ the state is updated by adding a deterministic term given by $b(X_t)\, dt$ and a stochastic term of the form $\sqrt{dt}\, \sigma(X_t)\xi$ with $\xi \approx \mathcal{N}(0, I_d)$ being independent of all previous states. This is the natural extension to continuous time of Markov Chains with Gaussian transitions (which are widely used in literature on stochastic optimisation). For an introduction to the theory of SDE, we refer to [31].

The purpose of the present work is to analyse methods for learning the continuous-time value function V from discrete-time observations of the state and the reward, using vanishing time steps. Therefore, this enters the field of reinforcement learning (RL) and more precisely the well-known problem of *policy evaluation*. However, most of the theory of RL is fundamentally stated with fixed discrete time. Few extensions to continuous-time problems exist in the literature, see [14, 17, 37] for instance, but we are not aware of any theoretical research work dealing with the stochastic setting (i.e., when $\sigma \not\equiv 0$). Such studies are yet of the utmost importance given that: first, a large part of the modern applications of RL come from time discretisations of continuous-time problems; second, the constant improvement of the available computational power makes it possible to consider finer and finer time discretisations (leading to the need of new numerical methods which are robust to decreasing the time steps); third, working in the deterministic set-up is not sufficient since deterministic processes are never considered in practice in RL (even when the model is deterministic, stochastic perturbations are always added in practice to favor exploration, see Appendix C.4 for more details). The present work is the first, up to our knowledge, to investigate this direction from a theoretical perspective.

The results proved in the present work emphasise the fact that the presence of a stochastic part in the dynamics leads to new technical difficulties when the time steps are small. Therefore, standard RL algorithms have to be properly adapted. This also leads to new theoretical analyses. Like usually, most of such an analysis is only possible through the restrictive assumption of linear parametrisations.

1.1 Definitions of the main objects

We approximate V defined in (1) by a parametrised function $x \mapsto v(x, \theta)$, where $\theta \in \Theta$ is the learnt parameter and Θ is the space of parameters (usually given by $\mathbb{R}^{d_\theta}$ for some positive integer d_θ).

Definitions of the temporal differences. Here, an observation is a quadruple $(\Delta t, X, X', R)$, where $\Delta t > 0$ is deterministic, X and X' are two states arising at some times t and $t + \Delta t$, and R is the running reward between those times. See Section 2 for the precise assumptions on the observations. From that, one may usually compute δ the *standard temporal difference* defined by

$$\delta_{\Delta t} = \tfrac{1}{\Delta t} \left(v(X, \theta) - \gamma_{\Delta t} v(X', \theta) - \Delta t \, R \right), \tag{2}$$

where $\gamma_{\Delta t} = e^{-\rho \Delta t}$, and the scaling $\frac{1}{\Delta t}$ is chosen so that the expectation of δ is of order $O(1)$. This quantity was initially designed for discrete-time dynamics and is derived from Bellman's programming principle [7]. In the present work, one of the main contributions is to introduce and analyse $\widetilde{\delta}$ the *stochastic temporal difference*,

$$\widetilde{\delta}_{\Delta t} = \delta_{\Delta t} + \tfrac{1}{\Delta t} Z \quad \text{with} \quad Z = (X' - X - \Delta t \, b(X)) \cdot \nabla_x v(X, \theta). \tag{3}$$

The two temporal differences δ and $\widetilde{\delta}$ only differ by the additional term Z, that here is called *the (stochastic) correction term* or *the variance-reduction term*. Observe that computing the stochastic TD requires to know the drift function b at any observation, but not σ. This means that the expected direction of the dynamics has to be known, but not the law of the noises (which does not have to be gaussian, see Section 5.2). Such an assumption is standard when the physic of the model is known but some uncertainties may generate noises, like measurement noises, model approximations (e.g., statistical or discretisation errors), unpredictable external factors (e.g., imperfections of the ground or wind for robotic) or idiosyncratic noises (e.g., in financial models). In practice, stochastic TD is more meaningful for small time steps, but standard TD can be used in model-free environments.

Definitions of the Algorithms. We denote δ_k (resp. $\widetilde{\delta}_k$) as the standard (resp. stochastic) TD at iteration k, i.e., computed with $(\Delta t_k, X_k, X'_k, R_k)$ the k^{th} observation. For a sequence of learning rates $(\alpha_k)_{k \geq 1}$, the standard and stochastic TD(0) methods are defined by

$$\begin{aligned} \theta_{k+1} &= \theta_k - \alpha_k \delta_k \nabla_\theta v(X_k, \theta_k), \\ \widetilde{\theta}_{k+1} &= \widetilde{\theta}_k - \alpha_k \widetilde{\delta}_k \nabla_\theta v(X_k, \widetilde{\theta}_k). \end{aligned} \tag{TD0}$$

We also introduce the regularised standard and stochastic TD(0) methods as

$$\begin{aligned} \theta_{k+1} &= \Pi_{B_M} \big(\theta_k - \alpha_k (\delta_k \nabla_\theta v(X_k, \theta_k) + \mu \theta_k) \big), \\ \widetilde{\theta}_{k+1} &= \Pi_{B_M} \big(\widetilde{\theta}_k - \alpha_k (\widetilde{\delta}_k \nabla_\theta v(X_k, \widetilde{\theta}_k) + \mu \widetilde{\theta}_k) \big), \end{aligned} \tag{μ-TD0}$$

where $\mu \geq 0$ and Π_{B_M} is the projection on B_M the Euclidean ball of $\mathbb{R}^d$ centered at 0 with radius M.

1.2 Motivations

In the second paragraph of this text, we already motivated the need for theoretical studies of RL methods which are robust with respect to the decrease of time steps in the stochastic setting. In this section, we give more details on the precise types of problems that can be treated using our methods.

Optimal control. Observe that there is no control in (1) and that controlled dynamical systems will not be considered in the main text of the present work (only in Appendix D). Still, our main motivation is to solve optimal control problems in high-frequency regimes. Let us point out that a large part of the modern RL numerical methods are using TD(0) iterations (or basic extensions of it) as a subroutine for doing *policy evaluation* (PE). Those methods are then completed by adding

a *policy improvement* (PI) process which uses the current approximation of the value function (or the Q-function) to improve the control function. For more details we refer to [47] and the class of *generalised policy iterations*. We believe that the main difficulty for extending most of the RL numerical methods to the high-frequency regime precisely lies in the extension of the PE process, since PI often boils down to approximating an *argmax* operator (see [37] for instance) which is generally less related to the dynamics. This and the fact that the assumptions and analysis for PI are different from the ones of PE explain why we only consider PE in the present work. We refer to Appendix D.2 for an informal discussion on one way to extend our numerical methods to PI as well.

Solving partial or stochastic differential equations (PDEs or SDEs). See Appendix A for some standard notations from the PDE literature. The Feynman-Kac formula states that V from (1) satisfies

$$r = \mathcal{L}V := \rho V - \text{tr}\left(\tfrac{\sigma\sigma^\top}{2} D^2_{x,x} V\right) - b \cdot \nabla_x V. \tag{4}$$

Alternatively, the couple $(Y_t, Z_t) := (V(X_t), \sigma(X_t)^\top \nabla_x V(X_t))$ solves the backward SDE (BSDE)

$$dY_t = -(r(X_t) - \rho Y_t)\, dt + Z_t \cdot dW_t, \tag{5}$$

where we recall that solutions of BSDEs are couples of stochastic processes (unlike standard forward SDEs whose solutions are scalar processes), see [39] for an introduction to BSDEs and their application to optimal control. Consequently, another motivation is that the methods analysed in the present work can be viewed as original numerical methods for solving linear PDEs and BSDEs of the forms given above using only discrete observations of the dynamics. Using similar arguments as in the previous paragraph on optimal control, our methods can be extended to solve nonlinear PDEs and BSDEs but a theoretical analysis of such systems is out of the scope of the present work. For instance, adding a PI process allows to solve Hamilton-Jacobi equations and forward-backward SDE systems (FBSDE) arising in stochastic control problems (see Appendix D). More general nonlinear problems can be treated as well using other additional processes, like Picard iterations for instance, see [12].

Some real-world examples. Giving an exhaustive list of the applications for approximating quantities from (1), (4), (5) and their nonlinear counterparts is out of reach since it covers a very large number of stochastic continuous-time problems (with and without control). Moreover, we believe that this number is only going to grow faster in the future as the available computational power is still growing, the discretisation steps become finer and finer in practice and the interest on learning methods is growing inside communities working with continuous-time problems (using PDEs and SDEs for instance). Here, we give only three examples in which high-frequency RL are already studied in the literature. First, the robotics in real time, e.g., [35], where the physic of the models is well known (so is the drift function b) but some external phenomenon can only be implemented through additional noise (e.g., wind, imperfections of the ground or measurement noises). Second, modern finance models are often based on SDEs, see [18] for a thorough introduction, we refer to the surveys [20, 25] for machine learning methods to solve BSDEs arising in finance ; for instance, an application is the high-frequency trading, where the stochastic part comes from idiosyncratic noises. Third, models of nuclear fusion in tokamaks, e.g., [15], requiring high-frequency controls of the magnetic field, while noises may come from measurement errors or simplifications of the physical model which is too complicated to be accurately implemented with limited computational power.

1.3 Main contributions and originalities.

The main contributions of the present work are:

- For general parametrisation of the learnt value function (i.e., linear or non-linear), we show that the standard TD introduced in (2) is not suited to high-frequency regimes (since its variance blows up), while its stochastic counterpart from (3) is. We refer to Proposition 3.1.
- We introduce two original variants of the TD(0) method for estimating continuous-time value functions or for numerically approximating the solutions of linear PDEs or BSDEs. One is model-free and the other is model-based.
- Recall that convergence rates only exist in the literature under the restrictive assumption of a linear parametrisation. Under such an assumption, we prove that standard TD(0) converges for some decreasing sequence of time steps. This is surprising when compared with the

literature on BSDEs since it may be interpreted as: the solution of (5) can be learnt via an iterative method which totally omits the stochastic term "$Z_t \cdot dW_t$", see Remark 4.5.

- Under a linear parametrisation, we prove that stochastic TD(0) converges, with a faster convergence rate and is more robust with respect to the choice of the time steps than standard TD(0). More precisely, the speed of convergence is similar to the one of the simpler algorithm SGD for linear regression (up to the state of the art for Theorem 4.6).
- We show numerical simulations with similar convergence rates as predicted by the theory.
- We give an original interpretation of TD(0) as a minimising method under a particular structure of the dynamics (for a constant σ and a drift of the form $b = \nabla_x U$), see Proposition 4.2.

2 Assumptions

Assumption on the informations. Let us explain more the assumptions on the observations used for computing (TD0) and (μ-TD0). A sequence $(\Delta t_k, X_k, X'_k, R_k)_{k\geq 1}$ of observations is such that $\Delta t_k > 0$ is convergent to zero and (X_k, X'_k, R_k) are independent random variables which can be:

- **Real-world observations:** X_k, X'_k and R_k are the real states and reward obtained using the continuous-time dynamics in (1) on a time interval of length Δt, i.e.,

$$(X_k, X'_k, R_k) = \left(\widetilde{X}_0, \widetilde{X}_{\Delta t_k}, \frac{1}{\Delta t_k} \int_0^{\Delta t_k} r(\widetilde{X}_t)\, dt \right) \text{ where } d\widetilde{X}_t = b(\widetilde{X}_t)dt + \sigma(\widetilde{X}_t)dW_t.$$

- **Observations from a simulator:** X_k, X'_k and R_k are obtained using the Euler-Maruyama discretisation scheme of the SDE in (1). The step operator is $\mathcal{S}_{\Delta t} : (x, z) \mapsto x + \Delta t\, b(x) + \sqrt{\Delta t}\, \sigma(x) z$, then we assume $X'_k = \mathcal{S}_{\Delta t_k}(X_k, \xi_k)$ and $R_k = r(X_k)$.

Assumption on the law of the observations. We denote the law of X_k by m_k, recall that m denotes the stationary measure of the dynamics (1). We assume that m_k is convergent to m in the sense of distributions and that there exists a nonnegative integer p, such that

A1 for any $f \in C^p(\Omega; \mathbb{R})$, there exists $C_f > 0$ such that $|\mathbb{E}[f(X_k) - f(X)]| \leq C_f \Delta t_k$, for $k \geq 0$, where X is distributed according to m.

In practice, such a condition is obtained using ergodic arguments while following the Markov Chain (which can be continuous or discrete depending whether the observations are from the real world or from a simulator); see Appendix C.3 and Theorem C.1 for such a result with $p = 4$.

Choice of the boundary conditions. In discrete state space, boundary conditions are in general missing and unnecessary since the Markov transition probability is naturally designed such that the trajectories stay inside the state space, or may only leave through specific terminal states. In continuous time and state, and especially in stochastic settings, things become much more complicated. Indeed, boundary conditions of different natures appear naturally when establishing the models, each involving different theoretical and numerical difficulties. For instance, homogeneous Neumann conditions correspond to reflexive walls, Dirichlet boundary conditions correspond to exits, periodic boundary conditions are used for modeling some standard non-euclidean geometries (like spherical or cylindrical coordinates) and others like mixed Robin conditions or state constraints may correspond to other physical considerations; we refer to [19] for more details. Those conditions may even differ on different parts of the boundary or different dimensions. For those reasons, we have to make a choice; if our model does not cover all physical aspects of continuous dynamics, it is complex enough to capture the main ideas for extending RL methods to continuous models. We decide to only consider periodic boundary conditions and a state space given by the d-dimensional torus, i.e., $\Omega = \mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$. Nevertheless, we argue that adapting our arguments to different boundary conditions is totally feasible but out of the scope of the present paper (since it would lead to unnecessary technical difficulties that we prefer to avoid for this work to stay as simple as possible).

3 Preliminary results in the general case

The quantity $\mathbb{E}[\delta_{\Delta t}|X]$ (resp. $\mathbb{E}[\widetilde{\delta}_{\Delta t}|X]$ for stochastic TD) is usually dubbed the Bellman error. This quantity is equal to zero if $v(\cdot,\theta)$ is exactly the value function (of the discrete-time dynamics). Therefore, a standard way to check that the learnt function $v(\cdot,\theta)$ is a good approximation of the value function V, is to check if the Bellman error is near zero in some sense for small time steps. However, the Bellman error is not convenient to compute from observations, nor is useful for learning (since for any fixed value of X, a lot of observations are needed). A common alternative which is easier to compute is the average TD squared, it admits the following decomposition,

$$\underbrace{\mathbb{E}_{(X,X')}\left[|\delta_{\Delta t}|^2\right]}_{\text{averaged TD squared}} = \mathbb{E}_X\Big[\underbrace{\mathbb{E}_{X'}\left[\delta_{\Delta t}\middle|X\right]}_{\text{Bellman error}}^2\Big] + \underbrace{\mathbb{E}_X\left[\operatorname{Var}_{X'}\left(\delta_{\Delta t}\middle|X\right)\right]}_{\text{perturbating term}}. \tag{6}$$

The following proposition allows a simple comparison of the orders of magnitude in the latter equality and its counterpart with $\widetilde{\delta}$, showing that $\widetilde{\delta}$ fits better than δ in the high-frequency regime.

Proposition 3.1. *Assume that r, b and σ are bounded, and that v admits bounded continuous derivatives in x everywhere up to order two. The means and variances of δ and $\widetilde{\delta}$ given X satisfy,*

$$\begin{aligned}
&\lim_{\Delta t\to 0} \mathbb{E}_{X'}[\delta_{\Delta t}|X] = \lim_{\Delta t\to 0} \mathbb{E}[\widetilde{\delta}_{\Delta t}|X] = \mathcal{L}v(X,\theta) - r(X),\\
&\lim_{\Delta t\to 0} \Delta t \operatorname{Var}_{X'}(\delta_{\Delta t}|X) = \left|\sigma(X)\nabla_x v(X,\theta)\right|^2,\\
&\lim_{\Delta t\to 0} \operatorname{Var}_{X'}\left(\widetilde{\delta}_{\Delta t}|X\right) = 2\text{tr}\left((\sigma\sigma^\top D_x^2 v(X,\theta))^2\right).
\end{aligned}$$

In the one hand, because the variance diverges as $\frac{1}{\Delta t}$, the latter proposition directly implies that the perturbating term in (6) should converge to infinity when the time step is small; thus it would totally overwhelm the interesting term. On the other hand, this does not happen when $\widetilde{\delta}$ replaces δ since, in this case, the perturbating term remains bounded. More precisely, at the limit $\Delta t \to 0$, we get

$$\begin{aligned}
&\lim_{\Delta t\to 0} \mathbb{E}\left[|\delta_{\Delta t}|^2\right] = \begin{cases} +\infty \text{ if } v(\cdot,\theta) \text{ is not constant,} \\ \mathbb{E}\left[\left(\rho C - r(X)\right)^2\right] \text{ if } v(\cdot,\theta) = C. \end{cases}\\
&\lim_{\Delta t\to 0} \mathbb{E}\left[|\widetilde{\delta}_{\Delta t}|^2\right] = \mathbb{E}_X\left[(\mathcal{L}v(X,\theta) - r(X))^2\right] + 2\mathbb{E}_X\left[\text{tr}\left((\sigma\sigma^\top D_x^2 v(X,\theta))^2\right)\right].
\end{aligned}$$

An interesting consequence of the former equality is that, at the limit $\Delta t \to 0$, SGD (or any other gradient descent method) applied to (6) can only converge to a constant. Instead for stochastic TD, the latter equality only implies that SGD would converge to a biased limit, where the bias can be small in practice. In Appendix E we give more details on this method, named *residual gradient* [4], we prove convergence results and some alternatives to reduce the bias are discussed in Section E.2.

The latter paragraph suggests that TD(0) is not a stochastic gradient method (otherwise we would not be able to prove convergence results as we do in the next section). Indeed, it is only a stochastic *semi-gradient* method because the term $\delta\nabla_\theta v$ in (2) is not a gradient in general.

4 Convergence Results in the Linear Setting

From here on, we assume (**A1**) and the following assumptions:

- **A2** The function v is linear with respect to $\theta \in \mathbb{R}^{d_\theta}$, i.e., $v(x,\theta) = \theta^\top\varphi(x)$ where $\varphi : \Omega \to \mathbb{R}^{d_\theta}$.
- **A3** The functions r, b, σ are C^p, where p comes from (**A1**) and b is at least Lipschitz continuous.
- **A4** The feature vector φ is C^{p+2}, its coordinate functions are linearly independent.

4.1 Identification of the limits

The eventual limit of (TD0), named θ^*, is given by

$$\mathbb{E}_m\left[\varphi(X)\mathcal{L}\varphi(X)^\top\right]\theta^* = \mathbb{E}_m\left[r(X)\varphi(X)\right]. \tag{7}$$

The linear independence assumption on the coordinate functions of φ in Assumption **A4** implies that $\mathbb{E}\left[\varphi(X)\varphi(X)^\top\right] \in \mathbb{R}^{d_\theta \times d_\theta}$ is positive definite. Therefore, Lemma G.2 in the Appendix implies that the symmetric part of $\mathbb{E}\left[\varphi(X)\mathcal{L}\varphi(X)^\top\right]$ is positive definite as well. This implies that θ^* is well defined and that there exists $M_0 > 0$ such that $|\theta^*| \leq M_0$. Note that the latter quantity is a convenient choice for M in (μ-TD0), especially when μ is small. When μ is not small, we prefer the simpler quantity $M_\mu = \mu^{-1}\|r\|_\infty$. For ($\mu$-TD0), we always assume that

$$M \geq \min(M_0, M_\mu). \tag{8}$$

In this case, θ^*_μ, the eventual limit of (μ-TD0), satisfies $|\theta^*_\mu| \leq M$ so it is independent of M because it remains unchanged by the projection step Π_{B_M}. It is then given by

$$(\mu I_d + \mathbb{E}_m[\varphi(X)\mathcal{L}\varphi(X)])\,\theta^*_\mu = \mathbb{E}_m\left[r(X)\varphi(X)\right]. \tag{9}$$

The bias induced by the regularisation is the distance between θ^* and θ^*_μ, it is bounded as follows.

Proposition 4.1. *There exists $C > 0$ such that, for any $\mu > 0$, we get $|\theta^* - \theta^*_\mu| \leq C\mu$.*

If V belongs to the set of parametrised functions, i.e., there exists θ_V such that $V = v(\cdot, \theta_V)$, then (7) implies that $\theta^* = \theta_V$, so (TD0) would converge to V eventually. Otherwise, $v(\cdot, \theta^*)$ is known to be a good estimator of V in practice but cannot be expressed as the minimiser of a meaningful functional. Yet, we succeed to do so below under particular structure assumptions on the dynamics.

Proposition 4.2. *Assume that σ is constant and b is of the form $b = \nabla_x U$ for some continuously differentiable function $U : \Omega \to \mathbb{R}$, then θ^* is the solution of the following minimisation problem,*

$$\theta^* \in \operatorname{argmin}_\theta \mathbb{E}_{X\sim m}\left[\ell(v(X,\theta), V(X))\right] \quad \textit{where} \quad \ell(v,w) = \rho(v-w)^2 + \tfrac{1}{2}|\sigma^\top \nabla_x (v-w)|^2.$$

4.2 The regularised TD(0)

In this section, we only consider the regularised algorithms (μ-TD0). Under a common decreasing assumption on the learning rate (that it is proportional to $1/(\mu(k+1))$) and convenient choices on Δt_k, Theorem 4.3 below states a usual convergence rate in $1/k$ for the stochastic TD(0) method. This rate can easily be compared to the literature e.g., [46]. For the standard TD(0), we obtain a slower convergence rate because of δ being not adapted to small time steps (as explained in Section 3).

Theorem 4.3. *Take $(\theta_k)_{k\geq 0}$ and $(\widetilde{\theta}_k)_{k\geq 0}$ defined by (μ-TD0) with $\mu > 0$, M satisfying (8), and $\alpha_k = \frac{2}{\mu(k+1)}$. There exists $C > 0$ such that, for $k \geq 1$,*

$$\mathbb{E}\left[\left|\theta_k - \theta^*_\mu\right|^2\right] \leq \frac{C}{\mu^2 k^{\frac{2}{3}}} \text{ for } \Delta t_k = \frac{1}{(k+1)^{\frac{1}{3}}}, \quad \Big| \quad \mathbb{E}\left[\left|\widetilde{\theta}_k - \theta^*_\mu\right|^2\right] \leq \frac{C}{\mu^2 k} \text{ for } \Delta t_k \leq \frac{1}{\sqrt{k+1}}.$$

Recall that θ^*_μ is a biased limit whose bias is bounded from Proposition 4.1. Combined with Theorem 4.3, we obtain the following approximation results of θ^* using (μ-TD0).

Corollary 4.4. *Under the same assumption as in Theorem 4.3, after $K \geq 2$ iterations, we have*

$$|\theta_K - \theta^*|^2 \leq \frac{C}{K^{\frac{1}{3}}} \text{ for } \mu = K^{\frac{1}{6}}, \qquad \Big| \qquad |\widetilde{\theta}_K - \theta^*|^2 \leq \frac{C}{\sqrt{K}}, \text{ for } \mu = K^{\frac{1}{4}}.$$

Remark 4.5. *The fact that standard TD(0) may converge is surprising at first sight since it may be interpreted as: the solution of the BSDE (5) can be learnt without considering the stochastic term $Z_t \cdot dW_t$. However, there are restrictions on the sequence of time steps to obtain such a convergence results: $\alpha_k/\Delta t_k$ has to tend to zero. Such a restriction is totally unnecessary for stochastic TD(0) for which having both Δt_k and α_k tend to zero is sufficient (and $\sum \alpha_k$ being divergent). We refer to the informal discussion on Appendix C.1 for some insights on the convergence of standard TD(0) and the assumption "$\alpha_k/\Delta t_k \to 0$". In particular, this assumption is sharp since $v(\cdot, \theta_k)$ cannot converge to a good estimator of V if "$\alpha_k/\Delta t_k$" does not tend to zero, see Appendix C.1 as well.*

4.3 Averaging Stochastic TD(0)

In this section, in the same spirit as the results in [1] for SGD, we get the convergence of the stochastic TD(0) algorithm with an averaging method, with a constant learning step, without a strong convexity assumption, without a regularisation assumption and without a projection map. This result cannot be extended to standard TD(0) because $\alpha/\Delta t_k$ cannot converge to zero with $\alpha > 0$ independent

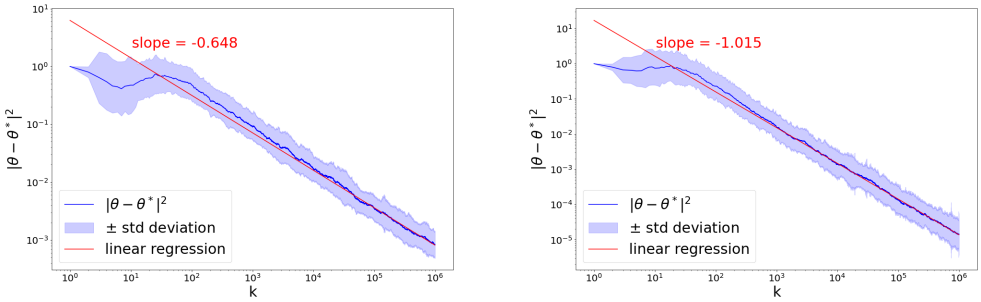

Figure 1: Empirical errors (averaged over 100 runs) for standard TD(0) (left) and stochastic TD(0) (right) under the assumptions of Theorem 4.3. The differences with the rates predicted by the theory are lower than 3%. See Appendix B.1 for details on the model used for the simulations.

of k (see Remark 4.5 for more details). The averaging method we are using here to accelerate the convergence is the Polyak-Juditsky method [40] using $\bar{\theta}_k$ defined by, for $k \geq 1$,

$$\bar{\theta}_k = \frac{1}{k} \sum_{i=0}^{k-1} \widetilde{\theta}_i, \tag{10}$$

We obtain a convergence rate which is competitive with the state of the art for the simpler problem of linear regression using SGD methods (we cannot expect to beat SGD convergence rates since SGD can be viewed as a particularly simple subcase of TD(0), see Appendix C.2 for more details).

Theorem 4.6. *If $\sum_{i=0}^{\infty} \Delta t_i^2$ is finite, there exist $C, R > 0$ such that, for $\alpha < R^{-2}$, $k \geq 1$,*

$$\ell(v(\cdot, \bar{\theta}_k), v(\cdot, \theta^*)) \leq \frac{C}{\alpha k} + \frac{C(d + \mathrm{tr}(HH^{-\top}))}{k}.$$

where $H = \mathbb{E}\left[\varphi(X)\mathcal{L}\varphi(X)^\top\right]$ and ℓ is defined in Proposition 4.2. If $\sum_{i=0}^{k-1} \Delta t_i^2 \leq a \ln(1+k)$ for some $a > 0$ and any $k \geq 0$, then for $\varepsilon > 0$, there exists $C, R > 0$ such that for $\alpha < R^{-2}$, $k \geq 1$,

$$\ell(v(\cdot, \bar{\theta}_k), v(\cdot, \theta^*)) \leq \frac{C}{\alpha k} + \frac{C(d + \mathrm{tr}(HH^{-\top}))}{k^{1-\varepsilon}}.$$

The proof is adapted from [1] with the extra difficulties that the linear operators applied to θ_k in (TD0) are different for each $k \geq 0$, that they are not symmetric, and that their symmetric part has no interesting property (only the symmetric part of the expectation of its limit when $k \to \infty$ has useful properties). Moreover, our sequence of stochastic estimators have vanishing biases that introduce new terms in the proof, this leads to the necessity to add a growth assumption on $\sum_i \Delta t_i^2$.

5 Extensions and limitations

5.1 Extensions

The time-dependent case. In the present analysis, we only considered the stationary case, see (1), mainly because this is the more standard setting in RL. However, we argue that all our results straightforwardly extend to the time-dependent case, i.e., when V is given by

$$V(t_0, x) = \mathbb{E}\left[\int_{t_0}^{T} r(t, X_t)\, dt \middle| X_0 = x\right] \quad \text{with} \quad dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t.$$

In this case, an observation is a quintuple $(t_k, \Delta t_k, X_k, X'_k, R_k)$ where t_k is the time at which X_k is observed. It allows to solve BSDEs with a new approach. Indeed, BSDE solvers from the literature generally divide into two groups. The first one consists of methods which are nonlocal with respect to time, i.e., the updates use entire trajectories from 0 to T, e.g., [22, 21, 36, 41]. The methods from the second group solve a sequence of regression problems starting from the terminal time T and using a backward induction on a fixed time grid, e.g., [2, 20]. Our method enter none of these groups since it is local in time, it is mesh-free and it do not rely on backward inductions.

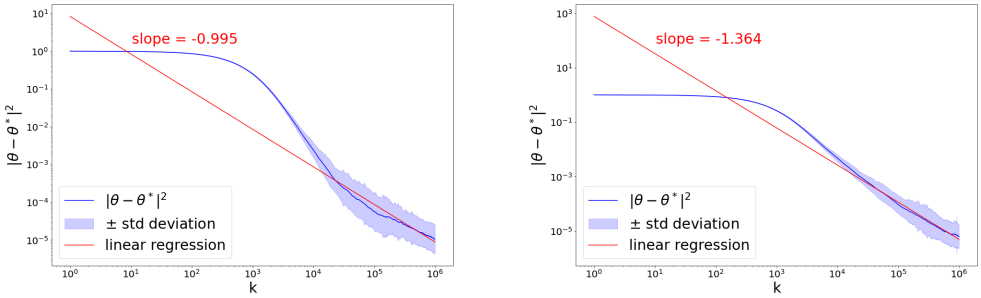

Figure 2: Empirical errors (averaged over 100 runs) for averaged stochastic TD(0) under the assumptions of Theorem 4.6. The learning step is $\alpha = 10^{-3}$. On the left, we use $\Delta t_k = (k+1)^{-\frac{1}{2}}$ and obtain the rate predicted by Theorem 4.6 with an error of 0.5%. On the right, we use $\Delta t_k = (k+1)^{-1}$ and obtain a better rate: we believe that this is due to H being symmetric positive definite and well-conditioned here but not in Theorem 4.6. See Appendix B.1 for details on the simulations.

Nonlinear problems. Recall that our main motivation is precisely to deal with the nonlinear counterparts of (1), (4) and (5). We refer to Section 1.2 for more details.

Other RL methods. We believe that the main ideas of the present work may be largely repeated for most of the RL methods using temporal differences or some of their variants. For instance, we refer to Appendix E for such an extension to residual gradient [4]. We argue that they may be extended to methods based on the Q-function as well (which include the majority of the state-of-the-art RL methods), up to an additional effort since the limit of the Q-function is independent of its second argument when the time steps tend to zero, we refer to [48] for a way to overcome this difficulty.

5.2 Limitations

About the independence assumption. For observations from a simulator, independent samples are easy to obtain. For real-world observations, the independence assumption is false in general but almost independent samples are usually obtained by shuffling a large dataset: we believe that it is sufficient in practice to obtain similar convergence rates as in the independent situation. Alternatively, one might assume that the observations are not independent but are obtained from following the Markov Chain (which is rarely done in practice since datasets are usually shuffled before use). Theoretically, this alternative assumption is generally restrictive since it relies on ergodic limits and very small learning rates, making the number of necessary samples increases dramatically in practice. We believe that comparable results under this assumption can be derived as well, but it is out of the scope of the present paper. We refer to [9] for convergence rates under this assumption and a comparison with the independence setting, in the case of a non-parametric TD(0) method.

Few words about the noises. In the present work, we make the assumption that the noise has independent increments: this is a limitation which is almost always assumed in the RL literature, even in the discrete-time setting. We make a second assumption, that the noise is continuous with respect to the time: this is a limitation since jumps are not allowed, but it still covers a very large number of applications from the real world (e.g., any physical or chemical system). Given these two assumptions, at the limit when Δt tends to zero, noises have to be Gaussian by Donsker's theorem (intuitively, there is a *Gaussianisation* phenomenon of the noise when Δt tends to zero). For observations from a simulator, we assumed the noises to be Gaussian only because it is the simplest way to have consistent noises with respect to changes of time steps (since the sum of independent Gaussians is Gaussian). However, our analysis straightforwardly extends to any family of noises that is (or at least becomes at the limit $\Delta t \to 0$) consistent with respect to changes of time steps.

6 Related works

TD learning and SGD. The TD algorithm was introduced in the tabular case by [46], with later convergence results for linearly dependent features [13]. Asymptotic stochastic approximation results were derived by [26] for the tabular case, and by [44] when using linear approximations, with a non-asymptotic analysis in the *i.i.d.* sampling case [34]. The analysis of TD requires tools from stochastic approximation [8], which have mainly been derived for stochastic gradient descent (SGD) [10] and reused here. The convergence results presented in the present paper may be compared to standard results on RL algorithms, see [7, 29] for TD(0). The proof of Theorem 4.6 is adapted from the literature on SGD [40, 1] to the non-symmetric setting induced by TD(0). In particular, [1] states the state-of-the-art results concerning convergence of SGD methods in the non-strongly convex setting, here we reach similar convergence rates on the more difficult optimisation problem raised by TD(0).

Continuous-time RL. Continuous-time RL started with [3], who proposed a continuous-time counterpart to Q-learning; it was later extended by [48]. From a different perspective, [11] extended classical RL algorithms to continuous-time discrete-state Markov decision processes. Then, using deterministic dynamics given by ordinary differential equations, and based on the Hamilton-Jacobi-Bellman equation, [17] derived algorithms for both policy evaluation and policy improvement. Similar deterministic approaches of continuous-time RL have recently been explored by [37, 50]. In order to balance between exploration and exploitation, [49] added an entropy-regularisation term to a continuous optimisation problem, the authors concluded that Gaussian controls are optimal for their relaxed problems, leading to a similar SDE system as the one studied in the present work.

Learning methods for solving PDEs and SDEs. Solving partial differential equations using learning algorithms is a natural idea. Indeed, classical methods such as finite differences, finite elements, or Galerkin methods cannot be computed for dimensions higher than three because of the size of the grid becoming too large. Some mesh-dependent learning algorithms have been developed, see [32, 33, 38], but they suffer from the same computational difficulties in high dimensions as the classical methods. There has been a surge of works during the last five years for solving high-dimensional PDEs or SDEs using deep learning, let us cite [28, 45] for the *Deep Galerkin Method*, or [6, 23, 24] for methods based on BSDEs or FBSDEs; we refer to the surveys [5, 20, 25] and the references therein for more results on deep learning methods for PDEs, BSDEs or FBSDEs. The methods presented in the present work are different to all the above-mentioned methods since they are mesh-free, local in time and do not rely on backward inductions.

7 Conclusion

In the present work, we prove that standard reinforcement learning methods based on the temporal difference are not adapted to solve stochastic continuous-time optimisation problems (see Proposition 3.1), nor their discretisations using small time steps. We then propose two original numerical methods based on the well-known TD(0) algorithm using vanishing time steps. We prove theoretical convergence rates (Theorem 4.3 and 4.6) under the assumption of linear parametrisations (which is always needed in the literature for proving explicit rates). We subsequently verify those rates through numerical simulations (Figures 1 and 2). The first method is model-free, its convergence rate is slower than the standard rates for discrete-time settings because it uses a quantity which is not suited to small time steps (see Section 3). Moreover, additional care has to be taken concerning the choice of the sequence of time steps (see Remark 4.5). The second method is model-based (it only requires to know the drift function b), it admits better rates of convergence and is more robust to changes of the time steps. More precisely, Theorem 4.3 shows convergence rates for both algorithms using standard decreasing learning rates, a strong convexity assumption (induced by the regularisation parameter μ) and a projection step. Theorem 4.6 shows a fast-convergence rate for the model-based method with a constant learning rate, without strong-convexity assumption and using an averaging method. This rate is similar to the state of the art for the simpler linear regression problem with SGD [1].

Alternatively, the two methods analysed in the present work are novel approaches to numerically solve linear PDEs and BSDEs using observations; their main advantage on the existing methods are that they are local in time, mesh-free and that they do not rely on any backward induction. Those problems (linear or nonlinear) are central in the vast domain of mathematical modelling and have uncountable applications, we refer to Section 1.2 for a more thorough discussion and some examples.

Acknowledgements. We thank Justin Carpentier for fruitful discussions related to this work. This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001(PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grant SEQUOIA 724063).

References

[1] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. *Advances in Neural Information Processing Systems*, 26:773–781, 2013.

[2] Achref Bachouch, Côme Huré, Nicolas Langrené, and Huyen Pham. Deep neural networks algorithms for stochastic control problems on finite horizon: numerical applications. *Methodology and Computing in Applied Probability*, 24(1):143–178, 2022.

[3] Leemon Baird. Advantage updating. Technical report, Wright Lab Wright-Patterson AFB OH, 1993.

[4] Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the International Conference on Machine Learning*, pages 30–37, 1995.

[5] Christian Beck, Martin Hutzenthaler, Arnulf Jentzen, and Benno Kuckuck. An overview on deep learning-based approximation methods for partial differential equations. *arXiv preprint arXiv:2012.12348*, 2020.

[6] Christian Beck, Arnulf Jentzen, et al. Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. *Journal of Nonlinear Science*, 29(4):1563–1619, 2019.

[7] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[8] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*, volume 22. Springer Science & Business Media, 1990.

[9] Éloïse Berthier, Ziad Kobeissi, and Francis Bach. A non-asymptotic analysis of non-parametric temporal-difference learning. *Advances in Neural Information Processing Systems*, 2022.

[10] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[11] Steven Bradtke and Michael Duff. Reinforcement learning methods for continuous-time Markov decision problems. *Advances in Neural Information Processing Systems*, 7, 1994.

[12] Jean-François Chassagneux, Junchao Chen, Noufel Frikha, Chao Zhou, et al. A learning scheme by sparse grids and picard approximations for semilinear parabolic pdes. Technical report, 2021.

[13] Peter Dayan. The convergence of TD(λ) for general λ. *Machine Learning*, 8(3):341–362, 1992.

[14] Peter Dayan and Satinder P. Singh. Improving policies without measuring merits. *Advances in Neural Information Processing Systems*, pages 1059–1065, 1996.

[15] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

[16] Monroe D. Donsker. An invariance principle for certain probability linit theorems. AMS, 1951.

[17] Kenji Doya. Reinforcement learning in continuous time and space. *Neural Computation*, 12(1):219–245, 2000.

[18] Nicole El Karoui, Shige Peng, and Marie Claire Quenez. Backward stochastic differential equations in finance. *Mathematical finance*, 7(1):1–71, 1997.

[19] Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Soc., 2010.

[20] Maximilien Germain, Huyên Pham, and Xavier Warin. Neural networks-based algorithms for stochastic control and pdes in finance. *arXiv preprint arXiv:2101.08068*, 2021.

[21] Emmanuel Gobet and Rémi Munos. Sensitivity analysis using itô–malliavin calculus and martingales, and application to stochastic optimal control. *SIAM Journal on control and optimization*, 43(5):1676–1713, 2005.

[22] Jiequn Han et al. Deep learning approximation for stochastic control problems. *arXiv preprint arXiv:1611.07422*, 2016.

[23] Jiequn Han and Arnulf Jentzen. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics*, 5(4):349–380, 2017.

[24] Jiequn Han, Arnulf Jentzen, and E Weinan. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

[25] Ruimeng Hu and Mathieu Lauriere. Recent developments in machine learning methods for stochastic control and games. *arXiv preprint arXiv:2303.10257*, 2023.

[26] Tommi Jaakkola, Michael Jordan, and Satinder Singh. Convergence of stochastic iterative dynamic programming algorithms. *advances in Neural Information Processing Systems*, 6, 1993.

[27] Bekzhan Kerimkulov, David Siska, and Lukasz Szpruch. Exponential convergence and stability of howard's policy improvement algorithm for controlled diffusions. *SIAM Journal on Control and Optimization*, 58(3):1314–1340, 2020.

[28] Yuehaw Khoo, Jianfeng Lu, and Lexing Ying. Solving parametric pde problems with artificial neural networks. *European Journal of Applied Mathematics*, 32(3):421–435, 2021.

[29] Donald E. Kirk. Optimal control theory: An introduction. 1970.

[30] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*, volume 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1992.

[31] Peter E Kloeden and Eckhard Platen. *Stochastic differential equations*. Springer, 1992.

[32] Isaac E. Lagaris, Aristidis Likas, and Dimitrios I. Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Transactions on Neural Networks*, 9(5):987–1000, 1998.

[33] Isaac E Lagaris, Aristidis C Likas, and Dimitris G Papageorgiou. Neural-network methods for boundary value problems with irregular boundaries. *IEEE Transactions on Neural Networks*, 11(5):1041–1049, 2000.

[34] Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355, 2018.

[35] Quentin Le Lidec, Louis Montaut, Cordelia Schmid, Ivan Laptev, and Justin Carpentier. Leveraging randomized smoothing for optimal control of nonsmooth dynamical systems. *arXiv preprint arXiv:2203.03986*, 2022.

[36] Charles-Albert Lehalle and Robert Azencott. Piecewise affine neural networks and nonlinear control. In *ICANN 98: Proceedings of the 8th International Conference on Artificial Neural Networks, Skövde, Sweden, 2–4 September 1998 8*, pages 633–638. Springer, 1998.

[37] Michael Lutter, Shie Mannor, Jan Peters, Dieter Fox, and Animesh Garg. Value iteration in continuous actions, states and time. *arXiv preprint arXiv:2105.04682*, 2021.

[38] Alaeddin Malek and R Shekari Beidokhti. Numerical solution for high order differential equations using a hybrid neural network—optimization method. *Applied Mathematics and Computation*, 183(1):260–271, 2006.

[39] Shige Peng. Backward stochastic differential equations and applications to optimal control. *Applied Mathematics and Optimization*, 27(2):125–144, 1993.

[40] Boris Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855, 1992.

[41] Demetri Psaltis, Athanasios Sideris, and Alan A Yamamura. A multilayered neural network controller. *IEEE control systems magazine*, 8(2):17–21, 1988.

[42] Martin L Puterman and Shelby L Brumelle. On the convergence of policy iteration in stationary dynamic programming. *Mathematics of Operations Research*, 4(1):60–69, 1979.

[43] ML Puterman. On the convergence of policy iteration for controlled diffusions. *Journal of Optimization Theory and Applications*, 33:137–144, 1981.

[44] Robert E. Schapire and Manfred K. Warmuth. On the worst-case analysis of temporal-difference learning algorithms. *Machine Learning*, 22(1):95–121, 1996.

[45] Justin Sirignano and Konstantinos Spiliopoulos. DGM: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, 375:1339–1364, 2018.

[46] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

[47] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: an Introduction*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2018.

[48] Corentin Tallec, Léonard Blier, and Yann Ollivier. Making deep Q-learning methods robust to time discretization. In *International Conference on Machine Learning*, pages 6096–6104, 2019.

[49] Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21:198–1, 2020.

[50] Cagatay Yildiz, Markus Heinonen, and Harri Lähdesmäki. Continuous-time model-based reinforcement learning. In *International Conference on Machine Learning*, pages 12009–12018, 2021.

A Some standard notations from PDE literature

In this section, we recall the definition of some standard differential operators. For $n, m \geq 1$, let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a function which admits partial derivatives in any direction up to order two, for $x \in \mathbb{R}^n$ we define:

- the first-order derivative (or Jacobian) of f at x as $D_x f(x) \in \mathbb{R}^{m \times n}$ such that $D_x f(x)_{i,j} = \partial_{x_j} f_i(x)$;
- if $m = 1$, the gradient of f at x as $\nabla_x f(x) \in \mathbb{R}^n$ such that $\nabla_x f(x)_j = \partial_{x_j} f(x)$;
- if $n = m$, the divergence of f at x as $\mathrm{div}(f)(x) = \sum_{j=1}^n \partial_{x_j} f_j(x)$;
- if $m = 1$, the second order derivative (or Hessian) of f at x as $D_x^2 f(x) \in \mathbb{R}^{n \times n}$ such that $D_x^2 f(x)_{i,j} = \partial_{x_i} \partial_{x_j} f(x)$;
- if $m = 1$, the Laplacian of f at x, as $\Delta_x f(x) = \sum_{i=1}^n \partial_{x_i} \partial_{x_i} f(x)$.

Occasionally, the Hessian and the Laplacian might be used even if $m > 1$. For the Laplacian, it only consists in applying the Laplacian coordinate-wise. For the Hessian, it outputs a tensor in dimension three, such that $D_x^2 f(x)_{i,j,k} = \partial_{x_j} \partial_{x_k} f_i(x)$.

B Numerical simulations

B.1 Description of the model

Let us describe all the details of the numerical simulations leading to Figure 1 and 2. In dimension $d = 1$, the one-dimensional torus is thought as the set $[-0.5, 0.5]$ completed with periodic boundary conditions. We take σ and b as in Proposition 4.2, and U and r given by

$$U = -\tfrac{\sigma^2}{2} \ln(2 - \cos(2\pi x)) \text{ and } r(x) = \left(\rho + \tfrac{4\pi^2\sigma^2}{2-\cos(2\pi x)}\right) \sin(2\pi x).$$

We refer to the top of Figure 3 for a graphical representations of r and $-U$. The function $-U$ has to be thought of as a potential function, i.e., the dynamic heads toward the minimum of $-U$.

One can easily check that V is then given by

$$V(x) = \sin(2\pi x),$$

by checking that (4) is satisfied. The invariant measure m of the dynamics in (1) satisfies

$$m(x) = \frac{e^{\frac{2U(x)}{\sigma^2}}}{\int_\Omega e^{\frac{2U(y)}{\sigma^2}} dy} = \frac{\sqrt{3}}{2 - \cos(2\pi x)}.$$

The function V and m are represented at the bottom of Figure 3.

The repartition function F of the distribution m satisfies

$$F(x) = m\big([-0.5, x]\big) = \pi \arctan(\sqrt{3} \tan(\pi x)).$$

Observe that, if u is a uniform random variable on $[0, 1]$ then $F^{-1}(u)$ is distributed according to m. Therefore, it becomes easy to sample i.i.d. observations from m given that, for $z \in [0, 1]$,

$$F^{-1}(z) = \pi^{-1} \arctan\left(\tfrac{1}{\sqrt{3}} \tan(\pi(z - \tfrac{1}{2}))\right).$$

The feature vector φ for the learning is

$$\varphi(x) = \big(1, \sin(2\pi x), \cos(2\pi x)\big)^\top,$$

so that V belongs to the set of parametrised functions. The parameters for the learning are $\alpha_k = \frac{2}{k+1}$, $\Delta t_k = \sqrt[3]{\alpha_k}$ for standard TD(0) and $\Delta t_k = \sqrt{\alpha_k}$ for stochastic TD(0). We take $\mu = 0$ (i.e., no regularisation) but we obtain the convergence rates of Theorem (4.3) like if μ was equal to 0.5 because the matrix $H = \mathbb{E}\left[\varphi(X)\mathcal{L}\varphi(X)^\top\right]$ is positive symmetric definite with $H \geq 0.5 I_d$.

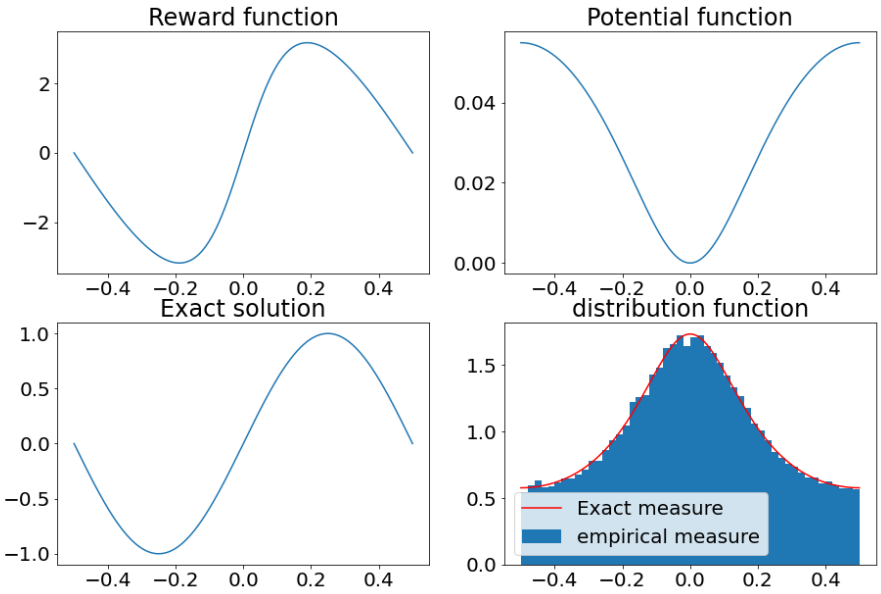

Figure 3: Here are some graphical representations of the model described in Section B.1. Namely, the reward function is at the top left; the potential is at the top right; the exact solution V is at the bottom left; the measure m and an empirical approximation of it with 10^5 samples divided in 100 subintervals are at the bottom right. We used $\rho = 1$ and $\sigma^2 = 0.1$.

C Insights and informal discussions

C.1 Insights on the convergence of standard TD(0)

As we already pointed out in Section 1.3 and Remark 4.5, the very fact that the model-free TD(0) method may converge is already surprising as it may be interpreted as: the solution of (5) can be learnt via an iterative method which totally omits the stochastic term "$Z_t \cdot dW_t$".

The present section aims at giving insights on why this method should converge in a particular setting, i.e., when $\alpha_k/\Delta t_k$ converges to zero. Conversely, we will see that it cannot converge outside of the latter setting. All the arguments here are very informal and should not be considered as rigorous (we refer to Appendix F for the rigorous proof).

Let us replace the discrete iteration variable k by a continuous variable $s \in [0, \infty)$ with the standard analogous

$$s \approx \sum_{i=0}^{k-1} \alpha_i \quad \text{and} \quad \Delta s \approx \alpha_k.$$

We assume that all indexes k may be changed into the continuous variable s and that we can give a sense, at least informally, to continuous-iteration observations $(\Delta t_s, X_s, X'_s, r_s)_{s \geq 0}$. We get,

$$\begin{aligned} d\theta_s \approx \Delta\theta_s &\approx \theta_{s+\Delta s} - \theta_s \\ &\approx -\Delta s\, \delta_s \nabla_\theta v \\ &\approx -\Delta s\, (\mathcal{L}v - r_s + (\Delta t_s)^{-\frac{1}{2}} (\sigma^\top \nabla_x v) \cdot \xi_s) \nabla_\theta v \\ &\approx -\Delta s\, (\mathcal{L}v - r_s) \nabla_\theta v + \sqrt{\tfrac{\Delta s}{\Delta t_s}} (\sigma^\top \nabla_x v) \cdot (\sqrt{\Delta s}\, \xi_s) \nabla_\theta v \end{aligned}$$

where the third line is obtained using Lemma G.1, with ξ_s a normally distributed random variable.

Then, let us use a new continuous-iteration Brownian motion $(\widetilde{W}_s)_{s\geq 0}$ (which is naturally different from $(W_t)_{t\geq 0}$ the continuous-time Brownian motion from the dynamics in (1)) and its standard discretised approximation $d\widetilde{W}_s \approx \sqrt{\Delta s}\,\xi_s$. We get that $(\theta_s)_{s\geq 0}$ is a solution to the SDE

$$d\theta_s = -(\mathcal{L}v - r_s)\nabla_\theta v\, ds + «\sqrt{\tfrac{\alpha_k}{\Delta t_k}}»\nabla_\theta v \nabla_x v^\top \sigma\, d\widetilde{W}_s, \tag{11}$$

The first consequence of the latter calculus is that θ cannot converge to a deterministic limit if the diffusion term $«\sqrt{\tfrac{\alpha_k}{\Delta t_k}}»\nabla_\theta v \nabla_x v^\top \sigma$ does not converge to zero. This should only happen if $\sqrt{\tfrac{\alpha_k}{\Delta t_k}}$ tends to zero. Observe that we did not use any assumption on the parametrisation $v(\cdot,\theta)$. In particular, the informal derivation of (11) holds for nonlinear parametrisation, so does the latter conclusion.

From now on, we explicitly assume that v is linear in θ, i.e., $v(\cdot,\theta) = \theta^\top \varphi$ and that $\sqrt{\tfrac{\alpha_k}{\Delta t_k}}$ does tend to zero. Consequently, the dynamics of θ_s tends to be deterministic at the limit and we get

$$\begin{aligned}\frac{d}{ds}\mathbb{E}\left[|\theta_s - \theta^*|^2\right] &= -2(\theta-\theta^*)^\top \mathbb{E}_{X\sim m}\left[\varphi(\mathcal{L}v(X,\theta) - r(X))\right] \\ &= -2(\theta-\theta^*)^\top \mathbb{E}\left[\varphi\mathcal{L}\varphi^\top\right](\theta-\theta^*),\end{aligned}$$

where the definition of θ^* in (7) is used to get the last line. Recall that the symmetric part of $\mathbb{E}\left[\varphi\mathcal{L}\varphi^\top\right]$ is positive definite by Lemma G.2. We just proved that $\mathbb{E}\left[|\theta_s - \theta^*|^2\right]$ is a Lyapunov function, which implies that θ_s should converge to θ^*.

C.2 Comparison between TD(0) and SGD: the limit $\rho \to \infty$

We already mentioned that TD(0) may be seen as an extension of SGD. A way to make the latter statement rigorous here is to consider the limit when ρ tends to infinity. Indeed, from (1) we easily obtain that, for $x \in \Omega$,

$$\lim_{\rho\to\infty} \rho V(x) = r(x),$$

and that ℓ/ρ converges to the square loss, where ℓ is defined in Proposition 4.2. In addition, if $\rho\Delta t$ tends to zero, from a fixed starting state $X = x \in \Omega$, we get

$$\widetilde{\delta}_{\Delta t} = \rho v(x,\theta) - r(x) + o(1).$$

Formally, we can thus view the least-square regression of r using SGD as a particularly simple limit case of TD(0).

In the discrete-time setting, the latter analysis is even easier and only consists in taking the discrete discount factor (named γ in (2)) equal to zero. In this situation, the fact that SGD is an instance of TD(0) in a particularly simple setting (some difficult terms are removed, such that the one making TD(0) be a bootstrapping algorithm) appears more clearly.

This explains why we do not expect the convergence rates proved in the present work to beat the state of the art of the literature on SGD. Let us recall that Theorem 4.6 does yield a similar convergence rate as the state-of-the-art results on mere convex problem with averaging method [1].

C.3 Convergence of the Discrete Markov Chain

In this section, we give a condition under which the convergence results (**A1**) holds. Here, we consider the case of observations obtained from a simulator, like stated in Section 2. Moreover, we assume that m_k, the law of X_k, is in fact the stationary measure of the discretisation of the dynamic in (1) by the Euler-Maruyama scheme. In this case, the sequence $(m_k)_{k\geq 0}$ is weakly convergent to m and a convergence rate is given in the following theorem.

Theorem C.1 (Theorem 14.5.1 from [30])**.** *For $f \in C^4(\Omega;\mathbb{R})$, there exists $C > 0$ depending only on the C^4-norm of f such that $|\mathbb{E}\left[f(X_k) - f(X)\right]| \leq C\Delta t_k$.*

C.4 Formal discussion on the origins of the noises

In standard discrete-time reinforcement learning, two kinds of noises are often considered. The first one comes from the transition probability of the MDP, this is the intrinsic noise of the model. The

second kind of noise is an artificial noise which is generally added to favor exploration. Exploration is used in order to exit non-interesting local minima and converge to more robust solutions.

Here, the noise in (1) is represented by a Brownian motion, but we never explained if this noise was intrinsic to the model or artificially added for exploration. In this section, we show that it can be any or both of the two propositions. More precisely, we introduce the following three classes of models:

- **M1** stochastic models of the form of (1) with an intrinsic noise,
- **M2** deterministic models with linear dynamics with respect to the controls, and an artificial noise added for exploration, regularisation or for smoothing the control (see [35]).
- **M3** stochastic models of the form of (1), with linear dynamics with respect to the controls, and an artificial noise.

For the first class of models **M1**, the noise is part of the model and cannot be tuned. The last two classes seem more interesting in the framework of RL, and more specifically in the theoretical study of the exploration/exploitation trade-off. In class **M2**, the dynamics has the following form,

$$\frac{d}{dt}x_t = A(x_t)u(x_t) + B(x_t),$$

where u is the control function, A and B are respectively matrix-valued and vector-valued functions. Then, in order to encourage exploration, instead of choosing a deterministic control function (e.g., being greedy with respect to some criterion), one generally adds noises in the choice of u. Gaussian noises are often considered in discrete dynamics because of their simplicity to sample, or because they are the minimisers of some entropy-relaxations of the optimisation problems (see [49] for instance). Therefore, at least at the discrete level, it is natural to change u into its noisy counterpart $u + \sigma(x, u)\xi_i/\sqrt{\Delta t}$. This leads to the following dynamics,

$$X_{t_{i+1}} = X_{t_i} + \Delta t \left(A(X_{t_i})u(X_{t_i}) + B(X_{t_i})\right) + \sqrt{\Delta t}A(X_{t_i})\sigma(X_{t_i}, u(X_{t_i}))\xi_i,$$

which admits a similar form as the observations from a simulator in Section 2, with $A\sigma$ replacing σ. This time, the noise is tunable and a particular interesting regime consists in letting σ tends to zero. The class **M3** consists in a mix between the two other classes, with $A\sigma_{\text{art}} + \sigma_{\text{int}}$ replacing σ this time, where σ_{art} and σ_{int} are the artificial and intrinsic noises respectively. The noise is tunable in some measure but the regime $\sigma \to 0$ is in general prohibited.

D Links optimal control problems

D.1 A short review of the optimal control problem in continuous time

Let us consider the controlled counterpart of (1),

$$dX_t = b(X_t, u(X_t))dt + \sigma(X_t, u(X_t))dW_t, \tag{12}$$

where $u : \Omega \to \mathcal{A}$ is a control function and $\mathcal{A}$ is set of admissible controls. The controller aims at maximising the following quantity over the set of admissible functions u,

$$J(u) = \mathbb{E}\Big[\int_0^\infty e^{-\rho t} r(X_t, u(X_t))dt\Big], \tag{13}$$

where X_0 is distributed according to some probability measure $\mu_0 \in \mathcal{P}(\Omega)$. Contrary to the ones in the main text, the prototypes of the drift, diffusion and reward functions are given by $b : \Omega \times \mathcal{A} \to \mathbb{R}^d$, $\sigma : \Omega \times \mathcal{A} \to \mathbb{R}^d$ and $r : \Omega \times \mathcal{A} \to \mathbb{R}$, i.e., they may depend on the control.

A natural approach from optimal control theory is to introduce the value V^u associated to a specific control function u, defined by:

$$V^u(x) = \mathbb{E}\Big[\int_0^\infty e^{-\rho t} r(X_t, u(X_t))dt \Big| X_0 = x\Big].$$

Then, solving the optimisation problem (13) boils down to compute V^* and u^*, respectively the optimal value function and the optimal control, satisfying

$$V^*(x) = \max_u V^u(x) \qquad \text{and} \qquad u^*(x) \in \operatorname{argmax}_u V^u(x).$$

Moreover, under mild regularity assumption, V^* can be characterised as the solution of a partial differential equation, named Hamilton-Jacobi-Bellman (HJB) equation (see [7] for more details), given by

$$\rho V^* - \max_{u \in \mathcal{A}} H(x, \nabla_x V^*(x), D_x^2 V^*(x), u) = 0, \tag{14}$$

where the Hamiltonian H is defined by, for $p \in \mathbb{R}^d$, $z \in \mathbb{R}^{d \times d}$ and $u \in \mathcal{A}$,

$$H(x, p, z, u) = r(x, u) + p \cdot b(x, u) + \frac{1}{2} \text{tr}\left(\left(\sigma\sigma^\top\right)(x, u) z\right). \tag{15}$$

Moreover, the optimal control belongs to the argmax the Hamiltonian, i.e.,

$$u^*(x) \in \text{argmax}_u \, H(x, \nabla_x V^*(x), D_x^2 V^*(x), u).$$

Uncontrolled diffusion. When the diffusion σ does not depend on the control, The HJB equation (14) can be rewritten as

$$\rho V^* - \text{tr}\left(\tfrac{\sigma\sigma^\top}{2} D_{x,x}^2 V\right) - \widetilde{H}(x, \nabla_x V^*) = 0,$$

where, here, for $p \in \mathbb{R}^d$, the reduced Hamiltonian is defined by,

$$\widetilde{H}(x, p) = \max_{u \in \mathbb{R}^d} \left\{p \cdot b(x, u) + r(x, u)\right\}.$$

In this case and under standard assumptions on b and L, the BSDE (5) is replaced by the FBSDE system

$$\begin{cases} dX_t = \nabla_p \widetilde{H}(X_t, Z_t)\, dt + \sigma(X_t) dW_t, \\ dY_t = -(r(X_t) - \rho Y_t)\, dt + Z_t^\top \sigma(X_t)\, dW_t. \end{cases}$$

Unlike the couple of SDEs from (1) and (5), the latter system is coupled through its both equations, making it much more complex to solve. Observe that it cannot be solved neither forward nor backward, it generally requires a fix point argument on the whole system.

A simple example. Let us consider the simpler case where σ is a constant positive real number, b is given by $b(x, u) = u$, r is concave with respect to u, and the control space is $\mathcal{A} = \mathbb{R}^d$. In this case, the HJB equation (14) can be rewritten as

$$\rho V^* - \tfrac{\sigma^2}{2} \Delta_x V^* - \widetilde{H}(x, \nabla_x V^*) = 0,$$

where, here, for $p \in \mathbb{R}^d$, the reduced Hamiltonian is defined by,

$$\widetilde{H}(x, p) = \max_{u \in \mathbb{R}^d} \left\{p \cdot u + r(x, u)\right\}.$$

In particular, $\widetilde{H}$ is the Legendre's transform (or convex conjugate) of $-r$ with respect to its second argument, let us recall that $-r$ is assumed to be convex with respect to u. In this case, the optimal control u^* admits the following closed form,

$$u^*(x) = \nabla_p \widetilde{H}(x, \nabla_x V^*(x)). \tag{16}$$

Remark D.1. *If $\widetilde{H}$ and $\nabla_p \widetilde{H}$ admit a known closed form, the latter example is particularly simple. This conclusions easily extend to the case of b being a more general affine function with respect to u, i.e., of the following form,*

$$b(x, u) = b_0(x) + b_1(x) u,$$

where $b_0 : \Omega \to \mathbb{R}^d$ and $b_1 : \Omega \to \mathbb{R}^{d \times d}$ are vector-valued and matrix-valued functions respectively.

Hopefully, this class of control problems is in fact of high interest since it actually contains a lot of models from modern control theory and reinforcement learning. One may for instance consider the cases where the dependence of r with respect to u is: either a characteristic function of a compact subset of $\mathbb{R}^d$; or a power function of $|u|$ (in the next paragraph, we present the quadratic case).

The quadratic case. Under the same assumptions as in the latter example, let us consider the particular case when r is separated with a quadratic part in u. More precisely, let us consider r to be given by,

$$r(x,u) = -\frac{|u|^2}{2} + r_0(x),$$

where $r_0 : \Omega \to \mathbb{R}$ is the state reward function. In this case, the conditions on V^* and u^* may be written as follows,

$$\begin{aligned} &\rho V^*(x) - \tfrac{\sigma^2}{2}\Delta_x V^*(x) - \tfrac{1}{2}|\nabla_x V^*(x)|^2 = r_0(x), \\ &u^*(x) = \nabla_x V^*(x). \end{aligned}$$

Finally, let us mention that the latter problem is not strictly speaking a linear-quadratic problem even if b is linear and r is quadratic with respect to u. Indeed, linear-quadratic control problems requires r to be quadratic with respect to the couple (x,u) (and b linear with respect to (x,u)), which is not the case here.

D.2 Informal extension to solve control problems

Let us assume that the observations are from a simulation, as defined in Section 2. The analysis in the present section can then easily be repeated for real-world observations with different notations. Using the Euler-Maruyama discretisation scheme on (12) we obtain the controlled discrete step operator

$$X_{t_{i+1}} = S_{\Delta t}(X_{t_i}, u(X_{t_i}), \xi_i) := X_{t_i} + \Delta t\, b(X_{t_i}, u(X_{t_i})) + \sqrt{\Delta t}\sigma(X_{t_i}, u(X_{t_i}))\xi_i.$$

This time, the latter step operator does not characterise a Markov chain, but a Markov decision process associated with the reward function r.

In reinforcement learning, such MDP are usually solved using iterative algorithms in the class of *generalised policy iteration* (GPI) methods [47]. Those algorithms are generally based on the value function or the Q-function. They consist in alternating two interactive updates: the *policy evaluation* and the *policy improvement*. Here, using the framework introduced in the present work, the policy evaluation consists in computing an approximation of the value function V^u associated to some control function u. Conversely, the policy improvement consists in updating the control function u in order to maximise its associated value function. These two processes are therefore antagonist in the sense that updating u makes our current approximation of V^u being less accurate; and vice-versa, when V^u is updated, the optimal response to it changes as well.

In the present work, outside of the present section, we focus on the policy evaluation process. However, one may easily figure out how it may be extended to GPI methods in order to solve MDPs with vanishing time steps. The simplest example consist in the *policy iteration* method which is described below.

Theoretical policy iteration. It consists in computing an approximation of the value function between any update of the control function. Namely, starting from an initial arbitrary control function u^0, we compute the approximating sequences $(V^\ell)_{\ell\geq 0}$ and $(u^\ell)_{\ell\geq 0}$ as follows:

$$V^\ell = V^{u^\ell} \qquad \text{and} \qquad u^{\ell+1} \in \operatorname{argmax}_u H(x, \nabla_x V^\ell(x), D_x^2 V^\ell(x), u),$$

where V^{u^ℓ} is the value function associated to the control function u^ℓ, and $u^{\ell+1}$ is the best response given the value function V^ℓ. Let us recall that H is defined in (15) as the Hamiltonian.

This algorithm is convergent with a super-linear convergence rate, see [42, 43, 27]. In particular, this method may be seen as a Newton algorithm applied to some infinite-dimensional fixed-point operator.

As the terminology *theoretical* suggests, the latter method is not implementable in practice with finite computational power, because of our assumption on continuous state and control spaces. This is not the case of the following iterative method.

Approximate policy iteration. This method is inspired by the latter one, theoretical policy iteration. However, this time the policy-evaluation step is only made using functional approximation using one of the two TD(0) methods (model-based or model-free) presented in the present work.

We assume that the policy improvement step can be done (resp. approximatively solved), for any value function $V : \Omega \to \mathbb{R}$. More precisely, there exists an operator $\mathcal{U}$ taking V as an argument and outputting a control function $u = \mathcal{U}(V)$, such that $u(x)$ is a maximiser (resp. almost a maximiser) of $u' \mapsto H(x, \nabla_x V(x), D_x^2 V(x), u')$. For instance, $\mathcal{U}$ may admit a closed form if the system reduces to (16) as in the example of the previous section. Otherwise, one may use an approximating iterative method to construct $\mathcal{U}$, for instance with an actor-critic method.

Like in the main text, we consider a parametrised value function $x \mapsto v(x, \theta)$ for some parameter $\theta \in \Theta$. Starting from an arbitrary initial parameter θ^0, let $(\theta^\ell)_{\ell \geq 0}$ be a sequence of parameters such that $x \mapsto v(x, \theta^\ell)$ is approximating the above sequence $(V^\ell)_{\ell \geq 0}$ in the theoretical policy iteration method. The sequence of control functions is defined by,

$$u^{\ell+1} = \mathcal{U}(v(\cdot, \theta^\ell)).$$

Then, at iteration $\ell \geq 1$, we compute θ^ℓ using the stochastic TD(0) method from (TD0) (alternatively we could have chosen the standard TD(0) method), i.e.,

$$\theta^\ell = \lim_{k \to \infty} \theta_k^\ell, \text{ where } \theta_{k+1}^\ell = \theta_k^\ell - \alpha_k^\ell \widetilde{\delta}_k^\ell \nabla_\theta v(X_k^\ell, \theta_k^\ell),$$

using the following definitions of the counterpart of (3),

$$\widetilde{\delta}_k^\ell = \widetilde{\delta}_{\Delta t_k}(X_k^\ell, \widetilde{X}_k^\ell, \theta_k^\ell) := (\Delta t_k^\ell)^{-1}\big(v(X_k^\ell, \theta_k^\ell) - \gamma_{\Delta t_k} v(\widetilde{X}_k^\ell, \theta_k^\ell) - r(X_k^\ell, u(X_k^\ell))\Delta t_k^\ell + Z_k\big),$$
$$\text{where } Z_k^\ell = \Big(\widetilde{X}_k^\ell - X_k^\ell - b(X_k^\ell, u^k(X_k^\ell))\Delta t_k^\ell\Big) \cdot \nabla_x v(X_k^\ell, \theta_k^\ell),$$
$$\text{and } \widetilde{X}_k^\ell = S_{\Delta_k^\ell}(X_k^\ell, u^\ell(X_k^\ell), \xi_k^\ell).$$

For a fixed $\ell \geq 1$, the sequences $(\Delta t_k^\ell)_{k \geq 0}$, $(X_k^\ell)_{k \geq 0}$, $(\xi_k^\ell)_{k \geq 0}$ and $(\alpha_k^\ell)_{k \geq 0}$ satisfy similar assumptions as their counterparts in the main text.

Other RL methods for solving MDPs. Like the latter adaptation to continuous time of the approximate policy iteration method, most of the RL algorithms using temporal difference may be adapted using the current framework in order to be more robust to vanishing time steps. This includes in particular approximate value iteration, Q-learning, SARSA, actor-critic methods and others.... The changes only consists in replacing any temporal difference in an algorithm by one of the two TD(0) algorithms presented here.

Remark D.2. *For the model-based algorithm, adding the variance-reduction correction can only benefit to the policy evaluation process. This explains why we chose to focus only on policy evaluation algorithms like TD learning in the present work. Another reason for not considering the policy improvement process is that the necessary assumptions for making its analysis are different from the ones considered here. Therefore, we believe that dealing with the policy improvement process in continuous time in a separate future contribution will allow a better understanding of each work and more flexibility to extend our results.*

E The residual gradient method

E.1 Extensions of the main results to the residual gradient method

In this section, we state similar results as Theorems 4.3 and 4.6 while replacing the TD(0) methods by residual gradient (RG) methods. We want to insist that, in Section E, the notations θ^*, θ_μ^*, $(\theta_k)_{k \geq 0}$ and $(\widetilde{\theta}_k)_{k \geq 0}$ stand for different quantities than in the rest of the article. This is due to the iterations and limits from RG methods being different from the ones from TD(0). Let us start by defining those quantities here. The standard and stochastic RG methods write as

$$\begin{aligned} \theta_{k+1} &= \theta_k - \alpha_k \delta_k \nabla_\theta \delta_k, \\ \widetilde{\theta}_{k+1} &= \widetilde{\theta}_k - \alpha_k \widetilde{\delta}_k \nabla_\theta \widetilde{\delta}_k, \end{aligned} \tag{RG}$$

where α_k is the learning rate. We also introduce the regularised standard and stochastic RG methods as

$$\begin{aligned} \theta_{k+1} &= \Pi_{B_M}\left(\theta_k - \tfrac{\alpha_k}{2} \nabla_\theta (|\delta_k|^2 + \mu |\theta_k|^2)\right), \\ \widetilde{\theta}_{k+1} &= \Pi_{B_M}\left(\widetilde{\theta}_k - \tfrac{\alpha_k}{2} \nabla_\theta (|\widetilde{\delta}_k|^2 + \mu |\widetilde{\theta}_k|^2)\right), \end{aligned} \tag{μ-RG}$$

where $\mu \geq 0$ and M is assumed to be a known upper bound of θ^*_μ which is defined later as the limit of the μ-regularised method. Let us define F_μ the RG cost,

$$F_\mu(\theta) = \underbrace{\mathbb{E}_X\left[(\mathcal{L}v(X,\theta) - r(X))^2\right]}_{\text{interesting term}} + \underbrace{\tfrac{1}{2}\mathbb{E}_X\left[\mathrm{tr}\left((\sigma\sigma^\top D_x^2 v(X,\theta))^2\right)\right]}_{\text{bias from RG being a gradient method}} + \underbrace{\tfrac{\mu}{2}|\theta|^2}_{\text{bias from the regularisation}},$$

Observe that the third term in the definition of F_μ is a similar bias as the one we have for TD(0), it can be tuned in a similar fashion as in Corollary 4.4. However, the second term is a bias coming from RG being a gradient method (more details in Section 3) and is more problematic since it cannot be reduced, but if σ can be tuned.

We then define θ^*_μ as the eventual limit of RG (which admits a double bias),

$$\theta^*_\mu = \mathrm{argmin}_{\theta\in\Theta} F_\mu(\theta).$$

We define $F = F_0$ and $\theta^* = \theta^*_0$. The following theorem is the counterpart to RG of Theorem 4.3, it concerns the convergence rate of the regularised stochastic RG method.

Theorem E.1. *Assume* **A2***,* **A3***,* **A4***, $\mu > 0$, $\alpha_k = \frac{2}{\mu(k+1)}$ and $\Delta t_k \leq c/\sqrt{k+1}$, for some $c > 0$ and for any $k \geq 0$. The sequence $(\widetilde{\theta}_k)_{k\geq 0}$ is convergent, and there exists $C > 0$ such that, for $k \geq 1$,*

$$\mathbb{E}\left[\left|\widetilde{\theta}_k - \theta^*_\mu\right|^2\right] \leq \frac{C}{\mu^2 k}.$$

We refer to Section F.3 for the proof.

Unlike Theorem 4.3, the latter theorem does not stand any convergence result for the standard RG method. This is because standard RG method may only converge to a constant function and not to a good estimator of V, more details are given in Section 3.

Then, we state below the counterpart to RG of Theorem 4.6, it concerns the convergence rate of the unregularised RG method with constant learning step and an averaging method.

Theorem E.2. *Assume* **A2***,* **A3** *and* **A4***. and that θ^* is bounded. If $\sum_{i=0}^\infty \Delta t_i^2$ is finite, there exist $C, R > 0$ such that, the following inequality holds for $\alpha < R^{-2}$, $k \geq 1$,*

$$\mathbb{E}\left[|\mathcal{L}v(X,\bar{\theta}_k) - \mathcal{L}v(X,\theta^*)|^2\right] \leq \frac{C}{\alpha k} + \frac{Cd}{k},$$

where $\bar{\theta}_k = \frac{1}{k}\sum_{i=0}^{k-1}\widetilde{\theta}_i$, for $k \geq 1$.

If instead we assume that $\sum_{i=0}^{k-1}\Delta t_i^2 \leq a\ln(1+k)$ for some $a > 0$ for any $k \geq 0$, then for any $\varepsilon > 0$ there exists $C, R > 0$ such that for $\alpha < R^{-2}$, $k \geq 0$, the latter inequalities are replaced with the following ones respectively

$$\mathbb{E}\left[|\mathcal{L}v(X,\bar{\theta}_k) - \mathcal{L}v(X,\theta^*)|^2\right] \leq \frac{C}{\alpha k} + \frac{Cd}{k^{1-\varepsilon}}.$$

Observe that the left-hand sides in the latter theorem are less convenient to work with than the ones in Theorem 4.6; this is another advantage in favor of the semi-gradient method TD(0) over real gradient methods like RG.

The proof of Theorem E.2 is a simple adaptation of the ones of Theorem 4.6, it is even simpler since here all the linear operators are symmetric.

E.2 Possible extensions of RG

One important weakness of RG algorithm is the presence of the term $\mathbb{E}_X\left[\mathrm{tr}\left((\sigma\sigma^\top D_x^2 v(X,\theta))^2\right)\right]$ in the definition of F_μ. In this section, we propose four alternatives to remove it. More precisely, we will try to minimise $\widetilde{F}_\mu$ instead of F_μ, where $\widetilde{F}_\mu$ is defined by

$$\widetilde{F}_\mu(\theta) = \mathbb{E}_X\left[(\mathcal{L}v(X,\theta) - r(X))^2\right] + \frac{\mu}{2}|\theta|^2,$$

and θ^*_μ is now defined by $\theta^*_\mu = \mathrm{argmin}\,\widetilde{F}^*_\mu$. Similarly, we take $\widetilde{F} = \widetilde{F}_0$ and $\theta^* = \theta^*_0$.

In particular, Theorems E.1 and E.2 hold with the following four alternatives.

Multi-step RG. This algorithm consists in the following induction relation,

$$\widetilde{\theta}_{k+1} = \widetilde{\theta}_k - \tfrac{\alpha_k}{2}\nabla_\theta \left(|\overline{\delta}_k|^2 + \mu |\widetilde{\theta}_k|^2 \right), \text{ with } \overline{\delta}_k = \tfrac{1}{n_k}\sum_{i=0}^{n_k-1} \widetilde{\delta}_{\Delta t_k}(X_{k,t_i}, X_{k,t_{i+1}}, \theta_k), \quad \text{(MS-RG)}$$

where $X_{k,0} = X_k$ and $X_{k,t_{i+1}} = S_{\Delta t_k}(X_{k,t_i}, \xi_{k,i})$, for $0 \leq i < n_k$ and $n_k \geq 1$ a sequence converging to infinity. The conclusions of Theorem E.1 then hold with $\alpha_k = \frac{4}{\mu(k+1)}$, $n_k \geq c^{-1}\sqrt{k+1}$, $n_k \Delta t_k \leq c/\sqrt{k+1}$, and $\sigma_k \leq c k^{-\frac{1}{8}}$, for some $c > 0$, for any $k \geq 0$. In particular the proofs or the counterparts to the multistep setting of Theorems E.1 and E.2, are similar to the originals but using Lemma G.6 below instead of Lemma G.1.

Vanishing viscosities. This algorithm consists in the following induction relation,

$$\widetilde{\theta}_{k+1} = \widetilde{\theta}_k - \tfrac{\alpha_k}{2}\nabla_\theta \left(|\widetilde{\delta}_k|^2 + \mu |\widetilde{\theta}_k|^2 \right), \text{ with } \widetilde{\delta}_k = \widetilde{\delta}_{\sigma_k, \Delta t_k}(X_k, \xi_k, \theta_k), \quad (\sigma\text{RG})$$

where $\widetilde{\delta}_{\sigma_k,\Delta t_k}$ is $\widetilde{\delta}_{\Delta t_k}$ where σ has been replaced by σ_k in the dynamics and (3). Here, we assume that we may choose the intensity of the noise, which is only possible when the noise have been added artificially to a deterministic problem (see class **M2**), which is in general interesting for the three following reasons: allowing exploration, having regular continuous-time solutions and having full-supported invariant measures of the dynamics.

The conclusions of Theorem E.1 then hold with $\alpha_k = \frac{4}{\mu(k+1)}$, $\Delta t_k \leq c/\sqrt{k+1}$, and $\sigma_k \leq c k^{-\frac{1}{8}}$, for $k \geq 0$.

Using mini-batches. Another alternative consists in using mini-batches, i.e.,

$$\theta_{k+1} = \theta_k - \tfrac{\alpha_k}{2}\nabla_\theta \left(|\overline{\delta}_k|^2 + \mu |\widetilde{\theta}_k|^2 \right), \text{ with } \overline{\delta}_k = \tfrac{1}{N_k}\sum_{i=1}^{N_k} \widetilde{\delta}_{\Delta t_k}(X_k, \xi_{k,i}, \theta_k), \quad \text{(MB-RG)}$$

where $(N_k)_{k\geq 0}$ are the size of the mini-batches. The conclusions of Theorem E.1 then hold with $\alpha_k = \frac{4}{\mu(k+1)}$, $\Delta t_k \leq c/\sqrt{k+1}$, and $N_k \geq c^{-1}\sqrt{k}$, for $k \geq 0$.

Changing the law of the noise. Note that the perturbating term from (6) comes from the variance of a term involving $\xi_k^\top D_x^2 v \xi_k - \Delta_x v$. Let us make the simple observation that, in dimension $d = 1$, the latter expression is null if ξ_k is a Rademacher random variable. This argument can be generalised to dimension $d > 1$. Since $D^2 v(X_k, \theta_k)$ is symmetric, we can find P an orthogonal matrix and D a diagonal matrix such that $D_x^2 v(X_k, \theta_k) = P^\top D P$. Therefore, it we can take $\xi_k = P^\top \zeta_k$ where ζ_k is a random vector, each of its coordinate being an independent Rademacher random variable.

Using Donsker's theorem [16], the random process at the limit is still a Brownian motion even if the increments before convergence are not Gaussian anymore. However, the weak convergence of the sequence $(m_k)_{k\geq 0}$ is slower here: Δt_k is replaced by $\Delta t_k^{\frac{1}{2}}$ (this is a consequence of the central limit theorem). The conclusions of Theorem E.1 then hold with $\alpha_k = \frac{4}{\mu(k+1)}$ and $\Delta t_k \leq c/(k+1)$, for $k \geq 0$.

F Proof of the main results

Here, C is a constant that can change from line to line and is independent from $(\alpha_k)_{k\geq 0}$, $(\theta_k)_{k\geq 0}$ and μ.

We will also make the simplification assumption that σ is always a constant positive number, only to simplifies the notations. No additional difficulties (like regularity) come from such a simplification since the learnt function v is assumed to be regular.

F.1 Proof of Theorem 4.3

In order to prove Theorem 4.3, we will need the following theorem first.

Theorem F.1. *Let $A \in \mathbb{R}^{d\times d}$ be a square matrix such that $A + A^\top \geq 2\mu I_d$ for some $\mu > 0$, and $b, \theta^* \in \mathbb{R}^d$ such that $A\theta^* = b$ and $|\theta^*| \leq M$ for some $M \geq 0$. For $\theta_0 \in \Theta$, define the sequence $(\theta_k)_{k\geq 0}$ by induction as*

$$\theta_{k+1} = \Pi_{B(0,M)}\left(\theta_k - \alpha_k g_k\right),$$

for $k \geq 0$, where $\alpha_k > 0$ is convergent to zero and $\sum_{k\geq 0}\alpha_k = \infty$, $|\mathbb{E}\left[g_k|\theta_k\right] - A\theta_k + b| \leq (1 + |\theta_k|)\varepsilon_k$, $\varepsilon_k \geq 0$ is convergent to zero, and $\mathbb{E}\left[|g_k|^2\middle|\theta_k\right] \leq c_k(1 + |\theta_k|^2)$ for some $c_k \geq 0$. Then $(\theta_k)_{k\geq 0}$ is convergent in expectation to θ^* and

$$\mathbb{E}\left[|\theta_k - \theta^*|^2\right] \leq 4M^2 e^{-\mu\sum_{i=0}^{k-1}\alpha_i} + C(1+M^2)\sum_{i=0}^{k-1}\alpha_i\left(\alpha_i + \mu^{-1}\varepsilon_i^2\right)e^{-\mu\sum_{j=i+1}^{k-1}\alpha_j}.$$

Proof. Up to starting the iterative algorithm from θ_1 instead of θ_0, we may assume that $|\theta_k| \leq M$ for every $k \geq 0$. For $k \geq 0$, let us denote $y_k = \mathbb{E}\left[|\theta_k - \theta^*|^2\right]$. We recall that $|\Pi_{B(0,M)}(\theta) - \theta^*| \leq |\theta - \theta^*|$ for any $\theta \in \Theta$, since $\theta^* \in B(0, M)$. This and the induction relation satisfied by θ_k, imply

$$\begin{aligned}
y_{k+1} &= \mathbb{E}\left[\left|\Pi_B\left(\theta_k - \alpha_k g_k\right) - \theta^*\right|^2\right] \\
&\leq \mathbb{E}\left[\left|\theta_k - \theta^* - \alpha_k g_k\right|^2\right] \\
&\leq y_k - 2\alpha_k\mathbb{E}\left[(\theta_k - \theta^*)^\top g_k\right] + \alpha_k^2\mathbb{E}\left[|g_k|^2\right] \\
&\leq y_k - 2\alpha_k\mathbb{E}\left[(\theta_k - \theta^*)^\top\mathbb{E}\left[g_k|\theta_k\right]\right] + \alpha_k^2\mathbb{E}\left[\mathbb{E}\left[|g_k|^2|\theta_k\right]\right] \\
&\leq y_k - 2\alpha_k\mathbb{E}\left[(\theta_k - \theta^*)^\top(A\theta_k - b)\right] + 2\alpha_k\varepsilon_k\mathbb{E}\left[|\theta_k - \theta^*|(1 + |\theta_k|)\right] + c_k\alpha_k^2\mathbb{E}\left[(1 + |\theta_k|^2)\right] \\
&\leq y_k - \alpha_k\mathbb{E}\left[(\theta_k - \theta^*)^\top(A + A^\top)(\theta_k - \theta^*)\right] + \mu\alpha_k\mathbb{E}\left[|\theta_k - \theta^*|^2\right] + 2(1+M^2)\mu^{-1}\alpha_k\varepsilon_k^2 + c_k(1+M^2)\alpha_k^2 \\
&\leq (1 - \mu\alpha_k)y_k + (1+M^2)\alpha_k\left(C\mu^{-1}\varepsilon_k^2 + c_k\alpha_k\right) \\
&\leq e^{-\mu\alpha_k}y_k + (1+M^2)\alpha_k\left(C\mu^{-1}\varepsilon_k^2 + c_k\alpha_k\right).
\end{aligned}$$

where we used a Young inequality to get to the fifth line. Therefore, we obtain,

$$y_k \leq e^{-\mu\sum_{i=0}^{k-1}\alpha_i}y_0 + C(1+M^2)\sum_{i=0}^{k-1}\alpha_i\left(c_i\alpha_i + \mu^{-1}\varepsilon_i^2\right)e^{-\mu\sum_{j=i+1}^{k-1}\alpha_j},$$

which leads to the desired inequality using $y_0 \leq (|\theta_0| + |\theta^*|)^2 \leq 4M^2$. □

Proof of Theorem 4.3. The proof only consists in checking that we can apply Theorem F.1. Let us start by the proof for stochastic TD(0). Using the same notation as in Theorem F.1, we define,

$$A = \mathbb{E}\left[\varphi(X)\mathcal{L}(X)\right] + \mu I_d, \quad b = \mathbb{E}\left[r(X)\varphi(X)\right], \text{ and } g_k = \widetilde{\delta}_k\varphi(X_k) + \mu\theta_k.$$

Then, we get

$$\begin{aligned}
\mathbb{E}\left[g_k|\theta_k\right] &= \mathbb{E}\left[\varphi(X_k)\left(\mathcal{L}\varphi(X_k) + R_{0,k} + \Delta t_k^{\frac{1}{2}}R_{1,k} + \Delta t_k R_{2,k}\right)\right]\theta_k + \mu\theta_k + \mathbb{E}\left[\varphi(X_k)r(X_k)\right] \\
&= \mathbb{E}\left[\varphi(X_k)\mathcal{L}\varphi(X_k)\right]\theta_k + \mu\theta_k + \mathbb{E}\left[\varphi(X_k)r(X_k)\right] + \Delta t_k\mathbb{E}\left[\varphi(X_k)R_{2,k}^\top\right]\theta_k,
\end{aligned}$$

where $R_{0,k} = R_0(X_k, \xi_k)$, $R_{1,k} = R_1(X_k, \xi_k)$ and $R_{2,k} = R_2(\Delta t_k, X_k, \xi_k)$ are given in Lemma G.1. From (**A1**), we get

$$\left|\mathbb{E}\left[\varphi(X_k)\mathcal{L}\varphi(X_k)\right] - A\right| \leq C, \text{ and } \left|\mathbb{E}\left[\varphi(X_k)r(X_k)\right] - b\right| \leq C.$$

Therefore, we obtain $|\mathbb{E}\left[g_k|\theta_k\right] - A\theta_k - b| \leq C(1 + |\theta_k|)\varepsilon_k$ with $\varepsilon_k = \Delta t_k$. The fact that $\mathbb{E}\left[|g_k|^2\middle|\theta_k\right] \leq c_k(1 + |\theta_k|^2)$, with $c_k = C$ being independent of k, is straightforward. Finally, $A + A^\top \geq 2\mu I_d$ comes from Lemma G.2. Theorem F.1 and the inequalities $|\theta^*| \leq C\mu^{-1}$ and $\exp\left(-\sum_{j=i+1}^{k}\frac{1}{j}\right) \leq i/k$ for $k > i \geq 0$, conclude the proof for stochastic TD(0).

For standard TD(0), the proof is similar but we have to change c_k into $c_k = C(\Delta t_k)^{-1}$. □

F.2 Proof of Theorem 4.6

We start with the following definitions,

$$\begin{aligned}
S &= \rho\mathbb{E}\left[\varphi(X)\varphi(X)^\top\right] + \frac{\sigma^2}{2}\mathbb{E}\left[D_x\varphi(X)D_x\varphi(X)^\top\right] \\
A &= \mathbb{E}\left[\varphi(X)\left(\frac{\sigma^2}{2}\nabla_x \ln(m) + b\right)D_x\varphi(X)^\top\right] \\
H(x) &= \varphi(x)\mathcal{L}\varphi(x)^\top \\
H_k(x) &= H(x) + \mathbb{E}\left[H(X) - H(X_k)\right] \\
H &= \mathbb{E}\left[H(X)\right].
\end{aligned}$$

Proof of Theorem 4.6. Here, $C > 0$ stands for a generic constant which value may change from line to line, it depends on the constants in the assumptions and is independent of k, of the smallest eigenvalue of S and of α.

Using Lemma G.1, we get

$$\theta_{k+1} = \theta_k - \alpha\varphi(X_k)\left(\mathcal{L}\varphi(X_k) + R_0(X_k,\xi_k) + \Delta t_k^{\frac{1}{2}}R_1(X_k,\xi_k) + \Delta t_k R_2(\Delta t_k, X_k, \xi_k)\right)^\top \theta_k + \alpha\varphi(X_k)r(X_k),$$

where $R_0(x,\xi)^\top\theta = \frac{\sigma^2}{2}\left(\xi^\top D_x^2 v(x,\theta)\xi - \Delta_x v(x,\theta)\right)$, and R_1 and R_2 can be red in Lemma G.1, and we get $\mathbb{E}_\xi[R_0(x,\xi)] = \mathbb{E}[R_1(x,\xi)] = 0$. Take $\eta_k = \theta_k - \theta^*$, it satisfies the following induction relation,

$$\eta_{k+1} = \left(I_d - \alpha H_k(X_k)\right)\eta_k - \alpha\left(H_k(X_k)\theta^* + \varphi(X_k)r(X_k)\right) - \alpha\left(H - \mathbb{E}[H(X_k)] + \Delta t_k\varphi(X_k)R_{2,k}^\top\right)(\eta_k + \theta^*),$$

where $H_k(x) = \varphi(x)(\mathcal{L}\varphi(x) + R_{0,k} + \Delta t_k^{\frac{1}{2}}R_{1,k})^\top + H - \mathbb{E}[H(X_k)]$, in particular $\mathbb{E}[H_k(X_k)] = H$. One may easily check that η_k can be rewritten as $\eta_k = \sum_{r=0}^{k-1}\eta_k^r$, where η_k^r is defined by,

$$\begin{aligned}
\eta_{k+1}^r &= (I_d - \alpha H)\eta_k^r + \chi_k^r + \Delta t_k\psi_k^r, \\
\eta_0^0 &= \eta_0, \qquad \eta_0^r = 0 \text{ if } r \geq 1,
\end{aligned} \tag{17}$$

where χ_k^r and ψ_k^r are defined by

$$\begin{aligned}
\chi_k^0 &= \alpha(H - H_k(X_k))\theta^* + \alpha\left(\varphi(X_k)r(X_k) - \mathbb{E}\left[\varphi(X_k)r(X_k)\right]\right), \\
\psi_k^0 &= \alpha\Delta t_k^{-1}\left(\mathbb{E}[H(X_k)] - H\right)\theta^* + \alpha\Delta t_k^{-1}\mathbb{E}[\varphi(X_k)r(X_k) - \varphi(X)r(X)] - \alpha\varphi(X_k)R_{2,k}^\top\theta^*, \\
\chi_k^{r+1} &= \alpha(H - H_k(X_k))\eta_k^r, \\
\psi_k^{r+1} &= \alpha\left(\Delta t_k^{-1}(\mathbb{E}[H(X_k)] - H) - \varphi(X_k)R_{2,k}^\top\right)\eta_k^r,
\end{aligned} \tag{18}$$

where we used that $\mathbb{E}\left[\varphi(X)\mathcal{L}v(X,\theta^*)\right] = 0$ to get the second line. One may notice that $\eta_k^r = 0$ if $r \geq k$.

First step: getting bounds on the covariance matrices of χ_r^k and ψ_r^k. Here, we prove by induction on r and k that

$$\begin{aligned}
\mathbb{E}\left[\eta_k^r \otimes \eta_k^r\right] &\leq 3C_k\alpha^r R^{2r} I_d, \\
\mathbb{E}\left[\chi_k^r \otimes \chi_k^r\right] &\leq C_k\alpha^{\max(r+1,2)}R^{2r}S, \\
\mathbb{E}\left[\psi_k^r \otimes \psi_k^r\right] &\leq \varepsilon C_k\alpha^{\max(r+1,2)}R^{2r}S,
\end{aligned}$$

where $R^2 = 3\widetilde{C}\left(\left\|\mathcal{L}\varphi + \mathbb{E}[R_0(\cdot,\xi_0)]\right\|_\infty + \Delta t_0^{\frac{1}{2}}\left\|R_1(\cdot,\xi)\right]\right\|_\infty + 2\varepsilon^{-1}\sup_{k\geq 0}\left\|R_2(\Delta t_k,\cdot,\xi)]\right\|_\infty + 2\varepsilon^{-1}\right)$, $0 < \varepsilon < \Delta t_0^{-2}$ is a constant that will be defined later, $\widetilde{C}$ is the constant from Lemma G.4 and $C_k = \left(|\theta^*|^2 + \eta_0^\top S\eta_0\right)\exp(\varepsilon\sum_{i=0}^{k-1}\Delta t_i^2)$.

For $k \geq 0$, and $r \geq 1$, let us prove the results for $(k+1, r)$ while assuming that it holds for (k, r), $(k, r-1)$ and $(k+1, r-1)$. For $b_k = \varepsilon \Delta t_k^2$, we get from (17) and (25),

$$\begin{aligned}
\mathbb{E}\left[\eta_{k+1}^r \otimes \eta_{k+1}^r\right] &\leq (1+b_k)\,\mathbb{E}\left[(I_d - \alpha H)\eta_k^r \otimes \eta_k^r (I_d - \alpha H^\top)\right] + \mathbb{E}\left[\chi_k^r \otimes \chi_k^r\right] + \Delta t_k^2 (1 + b_k^{-1})\mathbb{E}\left[\psi_k^r \otimes \psi_k^r\right] \\
&\leq 3C_k \alpha^r R^{2r}\,(1+b_k)\,(I_d - \alpha H)(I_d - \alpha H^\top) + C_k \alpha^{r+1} R^{2r} S + \varepsilon C_k \Delta t_k^2 \alpha^{r+1} R^{2r}(1 + b_k^{-1}) S \\
&\leq 3C_k \alpha^r R^{2r}(1 + \varepsilon \Delta t_k^2)(I_d - \alpha S) + \alpha^{r+1} R^{2r} C_k (2 + \varepsilon \Delta t_k^2) S \\
&\leq 3C_k \alpha^r R^{2r}(1 + \varepsilon \Delta t_k^2) I_d \leq 3C_k e^{\varepsilon \Delta t_k^2} \alpha^r R^{2r} I_d = 3C_{k+1} \alpha^r R^{2r} I_d.
\end{aligned}$$

Then, concerning χ_{k+1}^r, using Lemma G.4, we get

$$\begin{aligned}
\mathbb{E}\left[\chi_{k+1}^r \otimes \chi_{k+1}^r\right] &\leq 3C_{k+1} \alpha^{r+1} R^{2r-2} \mathbb{E}\left[(H - H_k(X_k))(H - H_k(X_k))^\top\right] \\
&\leq 3C_{k+1} \alpha^{r+1} R^{2r-2} \mathbb{E}\left[H_k(X_k) H_k(X_k)^\top\right] \\
&\leq 3C_{k+1} \alpha^{r+1} R^{2r-2} \left\| \mathcal{L}\varphi + \mathbb{E}[R_0(\cdot, \xi_k) + \Delta t_k^{\frac{1}{2}} R_1(\cdot, \xi^k)] \right\|_\infty \mathbb{E}\left[\varphi(X_k) \otimes \varphi(X_k)^\top\right] \\
&\leq C_{k+1} \alpha^{r+1} R^{2r} S.
\end{aligned}$$

Finally, using Lemma G.4 once again for ψ_{k+1}^r, we get,

$$\begin{aligned}
\mathbb{E}\left[\psi_{k+1}^r \otimes \psi_{k+1}^r\right] &\leq 6C_{k+1} \alpha^{r+1} R^{2r-2} \left(\Delta t_k^{-2} (\mathbb{E}[H(X_k)] - H)(\mathbb{E}[H(X_k)] - H)^\top + \mathbb{E}\left[|R_{2,k}|^2 \varphi(X_k) \otimes \varphi(X_k)\right]\right) \\
&\leq \varepsilon C_{k+1} \alpha^{r+1} R^{2r} S.
\end{aligned}$$

It remains to prove the inequalities for $k = 0$ and $r = 0$. Concerning $r = 0$, the proof is similar but we use the boundedness of θ^* and r instead of the induction assumption. Then $k = 0$ and $r \geq 1$ is straightforward since $\eta_0^r = \chi_0^r = \psi_0^r = 0$.

Second step: getting a bound on $\mathbb{E}\left[(\bar{\eta}_k^r)^\top S \bar{\eta}_k^r\right]$. Namely, we will prove that

$$\mathbb{E}\left[(\bar{\eta}_k^r)^\top S \bar{\eta}_k^r\right] \leq \frac{C \alpha^{\max(r-1,0)} R^{2r}}{k} \mathrm{tr}(I_d + H^{-\top} H) \left(\frac{1}{k} \sum_{i=0}^{k-1} C_i + \frac{1}{k} \left(\sum_{i=0}^{k-1} \Delta t_i C_i^{\frac{1}{2}}\right) + \widetilde{\delta}_{r=0} \alpha^{-1}\right),$$

for some constant $C > 0$. First, we notice that

$$\begin{aligned}
\eta_k^r &= (I_d - \alpha H)^{k-1} \eta_0^r + \sum_{i=0}^{k-1} (I_d - \alpha H)^{k-1-i} (\chi_i^r + \Delta t_i \psi_i^r) \\
\bar{\eta}_k^r &= \frac{1}{\alpha k} H^{-1} \left(I_d - (I_d - \alpha H)^k\right) \eta_0^r + \frac{1}{\alpha k} \sum_{i=0}^{k-1} \left(I_d - (I_d - \alpha H)^{k-i}\right) H^{-1} (\chi_i^r + \Delta t_i \psi_i^r),
\end{aligned}$$

this and the fact that $\mathbb{E}[\chi_i^r] = 0$ imply

$$\begin{aligned}
\mathbb{E}\left[(\bar{\eta}_k^r)^\top S \bar{\eta}_k^r\right] &\leq \frac{3}{\alpha^2 k^2} (\eta_0^r)^\top \left(I_d - (I_d - \alpha H)^k\right)^\top H^{-\top} S H^{-1} \left(I_d - (I_d - \alpha H)^k\right) \eta_0^r \\
&\quad + \frac{3}{\alpha^2 k^2} \sum_{i=0}^{k-1} \mathbb{E}\left[(\chi_i^r)^\top \left(I_d - (I_d - \alpha H^\top)^{k-i}\right) H^{-\top} S H^{-1} \left(I_d - (I_d - \alpha H)^{k-i}\right) (\chi_i)^r\right] \\
&+ \frac{3}{\alpha^2 k^2} \sum_{0 \leq i,j \leq k-1} \Delta t_i \Delta t_j \mathbb{E}\left[(\psi_i^r)^\top \left(I_d - (I_d - \alpha H^\top)^{k-i}\right) H^{-\top} S H^{-1} \left(I_d - (I_d - \alpha H)^{k-j}\right) \psi_j^r\right].
\end{aligned}$$

Let us define $I_{k,0}^r$, $I_{k,1}^r$ and $I_{k,2}^r$ as the first, second and third term, respectively, in the right-hand side of the latter inequality. One may notice that $I_{k,0}^r = 0$ if $r \geq 1$. Then concerning, $I_{k,0}^0$, we get

$$\begin{aligned}
I_{k,0}^0 &= \frac{3}{2\alpha^2 k^2} \eta_0^\top \left(I_d - (I_d - \alpha H^\top)^k\right) \left(H^{-\top} + H^{-1}\right) \left(I_d - (I_d - \alpha H)^k\right) \eta_0 \\
&\leq \frac{C}{\alpha^2 k} \eta_0^\top \eta_0 \leq \frac{C}{\alpha^2 k},
\end{aligned}$$

where we used (28) to obtain the last line. Then let us pass to $I^r_{k,1}$,

$$
\begin{aligned}
I^r_{k,1} &= \frac{3}{2\alpha^2 k^2}\sum_{i=0}^{k-1}\mathbb{E}\left[(\chi^r_i)^\top\left(I_d-(I_d-\alpha H^\top)^{k-i}\right)\left(H^{-\top}+H^{-1}\right)\left(I_d-(I_d-\alpha H)^{k-i}\right)\chi^r_i\right]\\
&= \frac{3}{2\alpha^2 k^2}\mathrm{tr}\sum_{i=0}^{k-1}\left(I_d-(I_d-\alpha H)^{k-i}\right)\mathbb{E}\left[\chi^r_i\otimes\chi^r_i\right]\left(I_d-(I_d-\alpha H^\top)^{k-i}\right)\left(H^{-\top}+H^{-1}\right)\\
&\le \frac{3\alpha^{\max(r-1,0)}R^{2r}}{2k^2}\mathrm{tr}\sum_{i=0}^{k-1}C_i\left(I_d-(I_d-\alpha H)^{k-i}\right)S\left(I_d-(I_d-\alpha H^\top)^{k-i}\right)\left(H^{-\top}+H^{-1}\right)\\
&= \frac{3\alpha^{\max(r-1,0)}R^{2r}}{4k^2}\mathrm{tr}\sum_{i=0}^{k-1}C_i\left(I_d-(I_d-\alpha H^\top)^{k-i}\right)\left(I_d-(I_d-\alpha H)^{k-i}\right)\left(2I_d+HH^{-\top}+H^{-1}H^\top\right)\\
&\le \frac{C\alpha^{\max(r-1,0)}R^{2r}}{k^2}\mathrm{tr}(I_d+HH^{-\top})\sum_{i=0}^{k-1}C_i.
\end{aligned}
$$

Then, concerning $I^r_{k,2}$, using the triangular inequality, we get

$$
\begin{aligned}
I^r_{k,2} &\le \frac{3}{2\alpha^2 k^2}\left(\sum_{i=0}^{k-1}\Delta t_i\mathbb{E}\left[(\psi^r_i)^\top\left(I_d-(I_d-\alpha H^\top)^{k-i}\right)\left(H^{-\top}+H^{-1}\right)\left(I_d-(I_d-\alpha H)^{k-i}\right)\psi^r_i\right]^{\frac12}\right)^2\\
&\le \frac{C\alpha^{\max(r-1,0)}R^{2r}}{2k^2}\left(\sum_{i=0}^{k-1}\Delta t_i\left[C_i\mathrm{tr}\left(I_d+H^{-\top}H\right)\right]^{\frac12}\right)^2\\
&= \frac{C\alpha^{\max(r-1,0)}R^{2r}}{2k^2}\mathrm{tr}\left(I_d+H^{-\top}H\right)\left(\sum_{i=0}^{k-1}\Delta t_i C_i^{\frac12}\right)^2,
\end{aligned}
$$

where we obtained the second line with similar arguments as in the calculus of the bound of $I^r_{k,1}$ above.

Third step: getting the desired bound. Using the triangular inequality on the norm induced by S, we obtain

$$
\begin{aligned}
\mathbb{E}\left[(\bar\eta_k)^\top S\bar\eta_k\right] &\le \left(\sum_{r=0}^{k-1}\mathbb{E}\left[(\bar\eta^r_k)^\top S\bar\eta^r_k\right]^{\frac12}\right)^2\\
&\le 2\mathbb{E}\left[(\bar\eta^r_0)^\top S\bar\eta^r_0\right]+2\left(\sum_{r=1}^{k-1}\mathbb{E}\left[(\bar\eta^r_k)^\top S\bar\eta^r_k\right]^{\frac12}\right)^2\\
&\le \frac{C}{\alpha k}+\frac{C}{k^2(1-\alpha^{\frac12}R)}\mathrm{tr}(I_d+HH^{-\top})\left(\sum_{i=0}^{k-1}C_i+\left(\sum_{i=0}^{k-1}\Delta t_i C_i^{\frac12}\right)^2\right).
\end{aligned}
$$

Therefore, if $\sum_{k=0}^\infty \Delta t_k^2$ is finite, then C_k is uniformly bounded and we can conclude by taking $\varepsilon=\Delta t_0^{-2}$. If instead $\sum_{i=0}^{k-1}\Delta t_i^2\le a\ln(1+k)$, we obtain that $C_k\le(1+k)^{a\varepsilon}$ and $\sum_{i=0}^{k-1}C_i$ is of order $k^{1+a\varepsilon}$ leading to the desired inequality up to changing ε into $a^{-1}\varepsilon$. □

F.3 Proof of Theorem E.1

Let us start by proving the following theorem on stochastic gradient descent methods.

Theorem F.2. *Let $f:\Theta\to\mathbb{R}$ be μ-convex, L-semi-concave, and such that $\theta^*=\mathrm{argmin}_\theta f(\theta)$ satisfies $|\theta^*|\le M$ for some $M>0$. For $\theta_0\in\Theta$, the sequence $(\theta_k)_{k\ge0}$ is defined by induction using the following projected stochastic gradient descent method,*

$$\theta_{k+1}=\Pi_{B(0,M)}\left(\theta_k-\alpha_k g_k\right),$$

for $k \geq 0$, where $\alpha_k > 0$ is convergent to zero, and $\sum_{k\geq 0} \alpha_k = \infty$, $|\mathbb{E}[g_k|\theta_k] - f'(\theta_k)| \leq (1+|\theta_k|)\varepsilon_k$, $\varepsilon_k \in \mathbb{R}_+$ is convergent to zero, and $\mathbb{E}\left[|g_k|^2 \middle| \theta_k\right] \leq C(1+|\theta_k|^2)$. Then $(\theta_k)_{k\geq 0}$ is convergent in expectation to θ^*, and

$$\mathbb{E}\left[|\theta_k - \theta^*|^2\right] \leq 4M^2 e^{-\frac{\mu}{2}\sum_{i=0}^{k-1}\alpha_i} + C(1+M^2)\sum_{i=0}^{k-1}\alpha_i\left(\alpha_i + \mu^{-1}\varepsilon_i^2\right)e^{-\frac{\mu}{2}\sum_{j=i+1}^{k-1}\alpha_j}.$$

Proof. Up to starting the iterative algorithm from θ_1 instead of θ_0, we may assume that $|\theta_k| \leq M$ for every $k \geq 0$. For $k \geq 0$, let us denote $b_k = |\theta_k - \theta^*|^2$. We recall that $|\Pi_{B(0,M)}(\theta) - \theta^*| \leq |\theta - \theta^*|$ for any $\theta \in \Theta$, since $\theta^* \in B(0,M)$. This and the induction relation satisfied by θ_k, imply

$$\begin{aligned}
b_{k+1} &\leq \mathbb{E}\left[|\theta_k - \theta^* - \alpha_k g_k|^2\right] \\
&\leq b_k - 2\alpha_k \mathbb{E}\left[(\theta_k - \theta^*)^\top g_k\right] + \alpha_k^2 \mathbb{E}\left[|g_k|^2\right] \\
&\leq b_k - 2\alpha_k \mathbb{E}\left[(\theta_k - \theta^*)^\top \mathbb{E}\left[g_k|\theta_k\right]\right] + \alpha_k^2 \mathbb{E}\left[\mathbb{E}\left[|g_k|^2|\theta_k\right]\right] \\
&\leq b_k - 2\alpha_k \mathbb{E}\left[(\theta_k - \theta^*)^\top f'(\theta_k)\right] + 2\alpha_k\varepsilon_k \mathbb{E}\left[|\theta_k - \theta^*|(1+|\theta_k|)\right] + C\alpha_k^2\mathbb{E}\left[(1+|\theta_k|^2)\right] \\
&\leq b_k - 2\alpha_k \mathbb{E}\left[f(\theta^*) - f(\theta_k) - \frac{\mu}{2}|\theta_k - \theta|^2\right] + 2(1+M)\alpha_k\varepsilon_k\mathbb{E}\left[|\theta_k - \theta^*|\right] + C(1+M^2)\alpha_k^2 \\
&\leq (1-\mu\alpha_k)b_k + \frac{\mu}{2}\alpha_k \mathbb{E}\left[|\theta_k - \theta^*|^2\right] + 4(1+M^2)\mu^{-1}\alpha_k\varepsilon_k^2 + C(1+M^2)\alpha_k^2 \\
&\leq (1 - \frac{\mu}{2}\alpha_k)b_k + C(1+M^2)\alpha_k(\mu^{-1}\varepsilon_k^2 + \alpha_k) \\
&\leq e^{-\frac{\mu}{2}\alpha_k} b_k + C(1+M^2)\alpha_k(\mu^{-1}\varepsilon_k^2 + \alpha_k),
\end{aligned}$$

where we used the μ-strong convexity of f to get to the fifth line, and a Young inequality to obtain the sixth line. Therefore, we obtain

$$b_k \leq e^{-\frac{\mu}{2}\sum_{i=0}^{k-1}\alpha_i} b_0 + C(1+M^2)\sum_{i=0}^{k-1}\alpha_i\left(\alpha_i + \mu^{-1}\varepsilon_i^2\right)e^{-\frac{\mu}{2}\sum_{j=i+1}^{k-1}\alpha_j},$$

which leads to the desired inequality using $b_0 \leq (|\theta_0| + |\theta^*|)^2 \leq 4M^2$. □

Proof of Theorem E.1. This proof consists in checking that we can apply Theorem F.2, using the following notations,

$$f(\theta) = \mathbb{E}\left[|\mathcal{L}v(X,\theta)|^2\right] + \frac{\sigma^4}{2}\mathbb{E}\left[\mathrm{tr}\left(D_x^2 v(X,\theta)^2\right)\right] + \frac{\mu}{2}|\theta|^2, \text{ and } g_k = \nabla_\theta|\delta_k|^2 + \mu\theta_k.$$

Thus, we get,

$$\begin{aligned}
\mathbb{E}\left[g_k|\theta_k\right] &= \mathbb{E}\left[\nabla_\theta\left|\mathcal{L}(X_k,\theta_k) + R_{0,k}^\top\theta_k + \Delta t_k^{\frac{1}{2}} R_{1,k}^\top\theta_k + \Delta t_k R_{2,k}^\top\theta_k\right|^2\right] + \mu\theta_k \\
&= \mathbb{E}\left[\nabla_\theta\left|\mathcal{L}(X_k,\theta_k)\right|^2\right] + \mathbb{E}\left[\nabla_\theta\left|R_{0,k}^\top\theta_k\right|^2\right] + \mu\theta_k + \Delta t_k\mathbb{E}\left[\nabla_\theta\left|R_{1,k}^\top\theta_k\right|^2\right] + 2\Delta t_k\mathbb{E}\left[\nabla_\theta\left(\widetilde{\delta}_k R_{2,k}^\top\theta_k\right)\right].
\end{aligned}$$

Then, from (**A1**), we obtain

$$\left|\mathbb{E}\left[\nabla_\theta\left|\mathcal{L}(X_k,\theta_k) + R_{0,k}^\top\theta_k + \Delta t_k^{\frac{1}{2}} R_{1,k}^\top\theta_k + \Delta t_k R_{2,k}^\top\theta_k\right|^2\right] + \mu\theta_k - f'(\theta_k)\right| \leq C(1+|\theta_k|).$$

This implies that $|\mathbb{E}[g_k|\theta_k] - f'(\theta_k)| \leq C\Delta t_k(1+|\theta_k|)$. The fact that $\mathbb{E}\left[|g_k|^2|\theta_k\right] \leq C(1+|\theta_k|^2)$ is straightforward. Theorem F.2 and the inequalities $|\theta^*|^2 \leq C\mu^{-1}$ and $\exp\left(-\sum_{j=i+1}^{k}\frac{1}{j}\right) \leq i/k$ for $k > i \geq 0$, conclude the proof. □

G Technical results

G.1 Proof of Proposition 3.1

Proof of Proposition 3.1. The proof differs whether we assume that the observations come from a simulator or from the real world (see Section 3 for precise definitions).

For observations coming from a simulator, it is a direct consequence of Lemma G.1 below.

In the following, we make the proof in the case of real-world observations for $\widetilde{\delta}$ (the proof for δ is similar).

In this proof, the dependence of v in θ is omitted. From Itô calculus, we have,

$$v(X_{\Delta t}) = v(X_0) + \int_0^{\Delta t} \left(\nabla_x v(X_t) \cdot b(X_t) + \tfrac{\sigma^2}{2} \Delta_x v(X_t) \right) dt + \sigma \int_0^{\Delta t} \nabla_x v(X_t) \cdot dW_t.$$

Therefore, the continuous temporal difference satisfies,

$$\begin{aligned} \widetilde{\delta}_{\Delta t}^{\text{cont}} = \mathcal{L}v(x) + \frac{1 - e^{-\rho \Delta t} - \rho \Delta t}{\Delta t} v(x) - \frac{e^{-\rho \Delta t}}{\Delta t} \int_0^{\Delta t} (\nabla_x v(X_t) \cdot b(X_t) - \nabla_x v(x) \cdot b(x)) dt \\ - \frac{\sigma^2 e^{-\rho \Delta t}}{2 \Delta t} \int_0^{\Delta t} (\Delta_x v(X_t) - \Delta_x v(x)) dt - \frac{\sigma e^{-\rho \Delta t}}{\Delta t} \int_0^{\Delta t} (\nabla_x v(X_t) - \nabla_x v(x)) \cdot dW_t. \end{aligned}$$

In the latter equality, the last term has zero mean, the second is convergent to zero and we prove in the following that the third and fourth are also convergent to zero.

Take $g : \Omega \to \mathbb{R}$ a bounded continuous function (we take $g = \nabla_x v \cdot b$ for the proof of the convergence of the third term, and $g = \Delta_x v$ for the proof concerning the fourth term). We define A as a set of measure zero such that $(X_t(\omega))_{0 \leq t \leq 1}$ is continuous for any $\omega \in \Omega_X \backslash A$ (where Ω_X is the sample space of the random process X). For any $\omega \in \Omega_X \backslash A$, Heine's Theorem states that $t \in [0, 1] \to X_t(\omega)$ admits a uniform modulus of continuity (which depends on ω), this implies that

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_0^{\Delta t} (g(X_t(\omega)) - g(x)) dt = 0.$$

We just proved that $\frac{1}{\Delta t} \int_0^{\Delta t} (g(X_t(\omega)) - g(x)) dt$ converges almost surely to zero, moreover it is uniformly bounded because g is bounded, so by the dominated convergence theorem, we obtain:

$$\lim_{\Delta t \to 0} \mathbb{E} \left[\frac{1}{\Delta t} \int_0^{\Delta t} (g(X_t) - g(x)) dt \right] = 0.$$

As a consequence, we obtain $\lim_{\Delta t \to 0} \mathbb{E}[\widetilde{\delta}_{\Delta t}^{\text{cont}}] = \mathcal{L}v(x, \theta)$.

Similar arguments imply that

$$\lim_{\Delta t \to 0} \mathbb{E} \left[\frac{1}{\Delta t} \int_0^{\Delta t} |g(X_t) - g(x)|^2 \, dt \right] = 0,$$

so the only term on the right-hand side of the latter expansion of $\widetilde{\delta}_{\Delta t}^{\text{cont}}$ whose variance does not vanish at the limit is the last, i.e.,

$$\lim_{\Delta t \to 0} \text{Var}(\widetilde{\delta}_{\Delta t}^{\text{cont}}) = \lim_{\Delta t \to 0} \frac{\sigma^2}{\Delta t^2} \mathbb{E} \left[\left| \int_0^{\Delta t} (\nabla_x v(X_t) - \nabla_x v(x)) \cdot dW_t \right|^2 \right]. \tag{19}$$

Using Itô calculus on $\nabla_x v(X_t)$, we obtain

$$\nabla_x v(X_t) - \nabla_x v(x) = \int_0^t \left(D_x^2 v(X_s) b(X_s) + \nabla_x \Delta_x v(X_s) \right) ds + \int_0^t D^2 v(X_s) dW_s.$$

Let us prove that the first integrable in the latter equality leads to a vanishing term only in the limit (19). This time, we take $g = D_x^2 v\, b + \nabla_x \Delta_x v$, let us consider the following sequence of inequalities

$$\mathbb{E}\left[\left|\int_0^{\Delta t}\int_0^t g(X_s)ds\cdot dW_t\right|^2\right] = \int_0^{\Delta t}\mathbb{E}\left[\left|\int_0^t g(X_s)ds\right|^2\right]dt \le \int_0^{\Delta t} t^2\|g\|_\infty^2 dt = \frac{\Delta t^3}{3}\|g\|_\infty^2,$$

Indeed, once we multiply by $\frac{\sigma^2}{\Delta t^2}$, this leads to a term of order Δt which will vanish at the limit $\Delta t \to 0$. Let us consider the only remaining part of the variance,

$$\begin{aligned}
\frac{\sigma^2}{\Delta t^2}\mathbb{E}\left[\left|\int_0^{\Delta t}\int_0^t \sigma D_x^2 v(X_s)dW_s\cdot dW_t\right|^2\right] &= \frac{\sigma^4}{\Delta t^2}\int_0^{\Delta t}\mathbb{E}\left[\left|\int_0^t D_x^2 v(X_s)dW_s\right|^2\right]dt \\
&= \frac{\sigma^4}{\Delta t^2}\int_0^{\Delta t}\int_0^t \mathbb{E}\left[\mathrm{tr}\left(D_x^2 v(X_s)^2\right)\right]ds\,dt \\
&= \frac{\sigma^4}{\Delta t^2}\int_0^{\Delta t}(\Delta t - s)\mathbb{E}\left[\mathrm{tr}\left(D_x^2 v(X_s)^2\right)\right]ds \\
&= \sigma^4\int_0^1 (1-u)\mathbb{E}\left[\mathrm{tr}\left(D_x^2 v(X_{u\Delta t})^2\right)\right]du,
\end{aligned}$$

where the last line is obtained using the change of variable $s = u\,\Delta t$. Using once again the dominated convergence theorem, we obtain:

$$\lim_{\Delta t\to 0}\mathrm{Var}(\widetilde{\delta}_{\Delta t}^{\mathrm{cont}}) = \sigma^4\mathbb{E}\left[\mathrm{tr}\left(D_x^2 v(x)^2\right)\right]\int_0^1(1-u)du = \frac{\sigma^4}{2}\mathbb{E}\left[\mathrm{tr}\left(D_x^2 v(x)^2\right)\right].$$

This concludes the proof. □

G.2 Expansions of the temporal differences

Lemma G.1. *Assume that the observations come from a simulator as defined in Section 3, for $(x,\xi,\theta)\in\Omega\times\mathbb{R}^d\times\Theta$ and $0<\Delta t<1$, we have*

$$\begin{aligned}
\widetilde{\delta}_{\Delta t}(x, S_{\Delta t}(x,\xi),\theta) &= \mathcal{L}v(x) + R_0(x,\xi)^\top\theta + \Delta t^{\frac{1}{2}} R_1(x,\xi)^\top\theta + \Delta t R_2(\Delta t, x,\xi)^\top\theta \\
R_0(x,\xi)^\top\theta &= \frac{\sigma^2}{2}\left(\Delta_x v(x) - \xi^\top D_x^2 v(x)\xi\right), \\
R_1(x,\xi)^\top\theta &= \rho\sigma\nabla_x v(x)\cdot\xi - \frac{\sigma}{2}b(x,u(x))^\top D_x^2 v(x)\xi - \frac{\sigma^3}{6}d_x^3 v(x)(\xi,\xi,\xi),
\end{aligned}$$

for some $R_2(\Delta t, x,\xi)$ such that, if ξ is a random variable normally distributed with zero mean and identity covariance matrix, then for $p\ge 1$, $\mathbb{E}\left[|R_2(\Delta t,x,\xi)|^p\right]$ is bounded uniformly with respect to Δt and x.

A similar result with real-world observations can be derived with additional terms in R_0 and R_1 which do depend on Δt but vanish when Δt tends to zero. Its proofs is a direct consequence of the proof of Proposition 3.1 in the previous section.

Proof. The proof consists in defining $\varphi : [0,1]\to\mathbb{R}$ by

$$\varphi(s) = e^{-s\rho\Delta t}v\left(x + s\left(b(x,u(x))\Delta t + \sigma\sqrt{\Delta t}\xi\right)\right),$$

and taking the development up to order four,

$$\varphi(1) = \varphi(0) + \varphi'(0) + \frac{\varphi''(0)}{2} + \frac{\varphi'''(0)}{6} + \int_0^1\frac{(1-s)^3}{6}\varphi''''(s)ds.$$

Using $\widetilde{b} \in \mathbb{R}^d$ defined by $\widetilde{b} = b(x, u(x))\Delta t + \sigma\sqrt{\Delta t}\xi$, the latter derivatives of φ are given by

$$\begin{aligned}
\varphi(0) &= v(x) \\
\varphi'(0) &= -\rho\Delta t v(x) + \nabla_x v(x) \cdot \widetilde{b} \\
\varphi''(0) &= \rho^2\Delta t^2 v(x) - 2\rho\Delta t \nabla_x v(x) \cdot \widetilde{b} + d_x^2 v(x)(\widetilde{b}, \widetilde{b}) \\
\varphi'''(0) &= -\rho^3\Delta t^3 v(x) + 3\rho^2\Delta t^2 \nabla_x v(x) \cdot \widetilde{b} - 3\rho\Delta t d_x^2 v(x)(\widetilde{b}, \widetilde{b}) + d_x^3 v(x)(\widetilde{b}, \widetilde{b}, \widetilde{b}) \\
\varphi''''(s) &= e^{-s\rho\Delta t}\left[\rho^4\Delta t^4 v - 4\rho^3\Delta t^3 \nabla_x v \cdot \widetilde{b} + 6\rho^2\Delta t^2 d_x^2 v(\widetilde{b}, \widetilde{b}) - 4\rho\Delta t d_x^3 v(\widetilde{b}, \widetilde{b}, \widetilde{b}) + d^4 v(\widetilde{b}, \widetilde{b}, \widetilde{b}, \widetilde{b})\right].
\end{aligned}$$

We conclude by replacing all the equalities in this proof in (3). □

G.3 Some lemmas used in the proof of Theorem 4.6

Lemma G.2. *The matrices S and A are respectively the symmetric and asymmetric part of H. Moreover, they satisfy*

$$S^2 \leq \mathrm{tr}(S)S \tag{20}$$

$$A^\top A = -A^2 \leq \frac{2}{\rho\sigma^2}\left\|b + \frac{\sigma^2}{2}\nabla_x \ln(m)\right\|_\infty^2 S^2 \tag{21}$$

$$(SA - AS) \leq 2\sqrt{\frac{2}{\rho\sigma^2}}\left\|b + \frac{\sigma^2}{2}\nabla_x \ln(m)\right\|_\infty S^2, \tag{22}$$

$$\mathbb{E}\left[H(X)H(X)^\top\right] \leq \rho^{-1}\|\mathcal{L}\varphi(X)\|_\infty^2 S. \tag{23}$$

Proof. First step: proving that S and A are respectively the symmetric and asymmetric part of H. Take $\theta \in \Theta$, we get:

$$\begin{aligned}
\theta^\top H\theta &= \theta^\top \mathbb{E}\left[\varphi(X)\mathcal{L}\varphi(X)^\top\right]\theta \\
&= \mathbb{E}\left[v(X,\theta)\mathcal{L}v(X,\theta)\right] \\
&= \int_\Omega \left(\rho v - \frac{\sigma^2}{2}\Delta_x v + b(x)\cdot\nabla_x v\right) v(x)m(x)dx \\
&= \rho\mathbb{E}\left[v(X)^2\right] + \frac{\sigma^2}{2}\mathbb{E}\left[|\nabla_x v(X)|^2\right],
\end{aligned}$$

where the last line is obtained by using the fact that m satisfies

$$-D_{x,x}^2 \cdot \left(\frac{\sigma\sigma^\top}{2}m\right) + \mathrm{div}(bm) = 0, \tag{24}$$

and the following integration by parts,

$$\begin{aligned}
\int_\Omega \nabla_x v \cdot b(x)v(x)m(x)dx &= \int_\Omega \frac{1}{2}\nabla_x(v^2)\cdot b(x)m(x)dx \\
&= -\frac{1}{2}\int_\Omega \mathrm{div}(b(x)m(x))v^2(x)dx, \\
-\int_\Omega \Delta_x v(x)v(x)m(x)dx &= \int_\Omega |\nabla_x v|^2 m(x)dx + \int_\Omega \frac{1}{2}\nabla_x(v^2)\cdot\nabla_x m(x)dx \\
&= \int_\Omega |\nabla_x v|^2 m(x)dx - \frac{1}{2}\int_\Omega \Delta_x m(x)v^2(x)dx.
\end{aligned}$$

This implies that S is the symmetric part of H. Then it is straightforward that the asymmetric part of H is equal to A.

Second step: proving the four inequalities. The first inequality (20) is straightforward, it only relies on the fact that S is symmetric and positive. The fourth inequality (23) is straightforward using the definitions of $H(X)$ and S. The third inequality (22) is a consequence of (21). Therefore, there is

only (21) left to prove. Let us take $\lambda \in \mathbb{C}$ a complex eigenvalue of H, and θ an associated normalised eigenvector, it satisfies $\bar{\theta}^\top S\theta = \Re(\lambda)$ and $\bar{\theta}^\top A\theta = i\Im(\lambda)$. Therefore, we get

$$\begin{aligned}
|\Im(\lambda)| &= \left|\bar{\theta}^\top A\theta\right| \\
&= \left|\mathbb{E}\left[\bar{v}(X,\theta)\left(b(X) + \nabla_x \ln m(X)\right)^\top \nabla_x v(X,\theta)\right]\right| \\
&\leq \|b + \nabla_x \ln(m)\|_\infty \mathbb{E}\left[|v(X,\theta)|^2\right]^{\frac{1}{2}} \mathbb{E}\left[|\nabla_x v(X,\theta)|^2\right]^{\frac{1}{2}} \\
&\leq \sqrt{\frac{2}{\rho\sigma^2}} \|b + \nabla_x \ln(m)\|_\infty \bar{\theta}^\top S\theta.
\end{aligned}$$

This concludes the proof. □

Lemma G.3. *For $\alpha \leq R^{-2}$, the following two inequalities hold for any $k \geq 0$,*

$$(I_d - \alpha H^\top)(I_d - \alpha H) \leq I_d - \alpha S \tag{25}$$

$$\left(I_d - (I - \alpha H^\top)^k\right)\left(I_d - (I - \alpha H)^k\right) \leq \alpha^2 k^2 H^\top H, \tag{26}$$

$$\left(I_d - (I - \alpha H^\top)^k\right)\left(I_d - (I - \alpha H)^k\right) \leq 4\left(1 + \frac{2}{\rho\sigma^2}\|b + \nabla_x \ln(m)\|_\infty^2\right) I_d, \tag{27}$$

$$\left(I_d - (I - \alpha H^\top)^k\right)\left(H^{-1} + H^{-\top}\right)\left(I_d - (I - \alpha H)^k\right) \leq 2\alpha k\left(1 + \sqrt{\frac{2}{\rho\sigma^2}}\|b + \nabla_x \ln(m)\|_\infty\right) I_d. \tag{28}$$

The latter lemma would be straight forward if H were symmetric. Conversely, it does not hold if we only assume the eigenvalues of H to be bounded and with positive real part. In fact, we need some bound on the imaginary part of the spectrum of H, depending on its real part.

Proof. One may notice that (28) is a straightforward consequence of (26) and (27). Then, concerning (25), it is sufficient to write $(I_d - \alpha H^\top)(I_d - \alpha H) = I_d - 2\alpha S + \alpha^2\left(S^2 + SA - AS - A^2\right)$, and use the definition of R, (20), (21) and (22). Therefore, it only remains to prove (26) and (27).

First step: proving (26). Let us proceed by induction, the case $k = 0$ is straightforward. Let us denote $y_k = \left(I_d - (I_d - \alpha H)^k\right)$ and assume that the inequality holds for k. One may notice that for $\theta \in \mathbb{R}^d$, using (25), we obtain

$$\begin{aligned}
\theta^\top y_k^\top (I_d - \alpha H)^\top H\theta &\leq \left(\theta^\top y_k^\top (I_d - \alpha H)^\top (I_d - \alpha H) y_k \theta\right)^{\frac{1}{2}} \left(\theta^\top H^\top H\theta\right)^{\frac{1}{2}} \\
&\leq \alpha k \theta^\top H^\top H\theta,
\end{aligned}$$

which implies $y_k^\top \left(I_d - \alpha H\right)^\top H + H^\top \left(I_d - \alpha H\right) y_k \leq 2\alpha k H^\top H$. Using the latter inequality, the relation $y_{k+1} = (I_d - \alpha H)y_k + \alpha H$, and (25) again, we get

$$\begin{aligned}
y_{k+1}^\top y_{k+1} &= y_k^\top (I_d - \alpha H)^\top (I_d - \alpha H) y_k + \alpha y_k^\top (I_d - \alpha H)^\top H + \alpha H^\top (I_d - \alpha H) y_k + \alpha^2 H^\top H \\
&\leq \alpha^2 k^2 H^\top H + 2\alpha^2 k H^\top H + \alpha^2 H^\top H = \alpha^2 (k+1)^2 H^\top H.
\end{aligned}$$

This concludes the induction.

Second step: proving (27). In this step, we will only work with the complex eigenvalues of H: let $\lambda \in \mathbb{C}$ be one of them, we get

$$\begin{aligned}
\left|1 - (1 - \alpha\lambda)^{k+1}\right| &= \left|(1 - \alpha\lambda)\left(1 - (1 - \alpha\lambda)^k\right) + \alpha\lambda\right| \\
&\leq \left(|1 - \alpha\lambda|\left|1 - (1 - \alpha\lambda)^k\right| + \alpha|\lambda|\right).
\end{aligned}$$

This implies

$$\begin{aligned}
\left|1-(1-\alpha\lambda)^k\right| &\leq \alpha|\lambda| \sum_{j=0}^{k-1} |1-\alpha\lambda|^j \\
&\leq \frac{\alpha|\lambda|}{1-|1-\alpha\lambda|} \\
&\leq \frac{\alpha|\lambda|}{1-(1-\alpha\Re(\lambda))^{\frac{1}{2}}} \text{ using (25),} \\
&\leq \frac{\alpha|\lambda|}{1-(1-\frac{\alpha}{2}\Re(\lambda))} \text{ because } \alpha\Re(\lambda) \leq 1, \\
&\leq 2\sqrt{1+\frac{\Im(\lambda)^2}{\Re(\lambda)^2}} \\
&\leq 2\left(1+\frac{2}{\rho\sigma^2}\|b+\nabla_x \ln(m)\|_\infty^2\right)^{\frac{1}{2}},
\end{aligned}$$

where the last inequality comes from a similar argument as in the proof of (21). This concludes the proof. □

Lemma G.4. *Assume* **A4**. *There exists $C>0$ such that the two following inequalities hold for any $k \geq 0$,*

$$\begin{aligned}
\mathbb{E}\left[\varphi(X_k)\otimes\varphi(X_k)\right] &\leq CS, \\
\left(\mathbb{E}\left[H(X_k)\right]-H\right)\left(\mathbb{E}\left[H(X_k)\right]-H\right)^\top &\leq C\Delta t_k^2 S.
\end{aligned}$$

Proof. We recall that the set of admissible functions v is finitely dimensional, therefore the C^4-norm and the $H^1(m)$-norm are equivalent and there exists $C>0$ such that $\|v(\cdot,\theta)\|_{C^4}^2 \leq C\|v(\cdot,\theta)\|_{H^1(m)}^2$. For $\theta\in\Theta$ and $k\geq 0$, this implies

$$\begin{aligned}
\theta^\top \mathbb{E}\left[\varphi(X_k)\otimes\varphi(X_k)\right]\theta &= C\mathbb{E}\left[v(X_k,\theta)^2\right] \\
&\leq C\mathbb{E}\left[v(X,\theta)^2\right] + C\Delta t_k \left\|v(\cdot,\theta)^2\right\|_{C^4} \\
&\leq C\left(1+\Delta t_k\right)\|v(\cdot,\theta)\|_{H^1(m)}^2,
\end{aligned}$$

where the second line is obtained from Theorem (**A1**). Here, C is a constant that can change from line to line. The first inequality is then obtained by recalling that $\|v(\cdot,\theta)\|_{H^1(m)}^2 \leq (\rho^{-1}+2\sigma^{-2})\theta^\top S\theta$.

Concerning the second inequality, we get

$$\begin{aligned}
\left|\left(\mathbb{E}\left[H(X_k)\right]-H\right)\theta\right|^2 &= \left|\mathbb{E}\left[\varphi(X_k)\mathcal{L}v(X_k,\theta)-\varphi(X)\mathcal{L}v(X,\theta)\right]\right|^2 \\
&\leq C\left(\Delta t_k \|v(\cdot,\theta)\|_{C^6}\right)^2 \\
&\leq C\Delta t_k^2 \|v(\cdot,\theta)\|_{H^1(m)}^2,
\end{aligned}$$

where the second line is obtained from (**A1**), and the third line from the fact that the C^6-norm is equivalent to the $H^1(m)$ on the finite dimensional space of functions v. We conclude the same way as we did for the first inequality. □

G.4 Calculus of variances and covariances

Lemma G.5. *Let $(x,\theta)\in\Omega\times\Theta$ and ξ a Gaussian vector with zero mean and identity covariance matrix, the following equalities hold*

$$\operatorname{Var}\left(\xi\cdot\nabla_x v(x)\right) = \left|\nabla_x v(x)\right|^2, \tag{29}$$

$$\operatorname{Var}\left(\xi^\top D^2 v(x)\xi - \Delta_x v(x)\right) = 2\operatorname{tr}\left(D_x^2 v(x)^2\right). \tag{30}$$

Proof. The first equality is straightforward. Since $D^2v(x)$ is symmetric, there exists P an orthogonal matrix and D a diagonal matrix such that $D^2v(x) = P^\top DP$. The couples (X, ξ) and $(X, P^\top \xi)$ have the same law and ξ is independent of X and D, this implies

$$\begin{aligned}
\mathrm{Var}\left(\xi^\top D^2v(x)\xi - \Delta_x v(x)\right) &= \mathbb{E}\left[\left(\xi^\top D^2v(x)\xi - \Delta_x v(x)\right)^2\right] \\
&= \mathbb{E}\left[\left(\left(P^\top\xi\right)^\top D^2v(x)P^\top\xi - \Delta_x v(x)\right)^2\right] \\
&= \mathbb{E}\left[\left(\xi^\top D\xi - \Delta_x v(x)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^d D_i^2\left(\xi_i^2 - 1\right)^2\right] = 2\sum_{i=1}^d D_i^2 = 2\mathrm{tr}\left(D_x^2 v(x)^2\right).
\end{aligned}$$

This concludes the proof. □

G.5 Counterpart to Lemma G.1 in the multi-step setting

Lemma G.6. *There exists $C > 0$ such that, for any $(x, \theta) \in \Omega \times \Theta$, $n \geq 1$, $0 < \Delta t < \frac{1}{n}$ and $\xi = (\xi_i)_{0\leq i<n}$ independent normally distributed random variables with zero mean and identity covariance matrix, we have*

$$\begin{aligned}
\left|\mathbb{E}\left[|\widetilde{\delta}^n_{\Delta t}(x,\xi,\theta)|^2\right] - \mathcal{L}v(x)^2\right| &\leq C\left(1 + |\theta|^2\right)\left(n^{-1} + n\Delta t\right), \\
\left|\mathbb{E}\left[\nabla_\theta|\widetilde{\delta}^n_{\Delta t}(x,\xi,\theta)|^2\right] - \nabla_\theta\mathcal{L}v(x)^2\right| &\leq C\left(1 + |\theta|\right)\left(n^{-1} + n\Delta t\right).
\end{aligned}$$

Proof. Taking $X_0 = x$ and $X_{t_{i+1}} = S_{\Delta t}(X_{t_i}, \xi_i)$ for $0 \leq i < n$, we obtain

$$\widetilde{\delta}^n_{\Delta t}(x,\xi,\theta) = \frac{1}{n}\sum_{i=0}^{n-1}\widetilde{\delta}_{\Delta t}(X_{t_i}, X_{t_{i+1}}, \theta). \tag{31}$$

Let us do the expansion of $\mathcal{L}v(X_{t_i})$ around x up to order two,

$$\mathcal{L}v(X_{t_i}) = \mathcal{L}v(x) + \nabla_x\mathcal{L}v(x)\cdot\widetilde{b} + \int_0^1 (1-s)\widetilde{b}^\top D_x^2\mathcal{L}v\left(x + s\widetilde{b}\right)\widetilde{b}ds,$$

where $\widetilde{b} = \sum_{j=0}^{i-1}\left(b(X_{t_j}, u(X_{t_j}))\Delta t + \sigma\sqrt{\Delta t}\xi_j\right)$. The latter equalities and Lemma G.1 imply

$$\begin{aligned}
\widetilde{\delta}^n(x,\xi,\theta) = \mathcal{L}(x) + \frac{\sigma^2}{2n}\sum_{i=0}^{n-1}\left(\Delta_x v(X_{t_i}) - \xi_i^\top D^2v(X_{t_i})\xi_i\right) + \frac{1}{n\sqrt{\Delta t}}\sum_{i=0}^{n-1}\Big[(n-1-i)\sigma\nabla_x\mathcal{L}v(X_{t_i})\cdot\xi_i \\
+ \rho\sigma\nabla_x v(X_{t_i})\cdot\xi_i - \frac{\sigma}{2}b(X_{t_i}, u(X_{t_i}))^\top D^2v(X_{t_i})\xi_i - \frac{\sigma^3}{6}d_x^3v(X_{t_i})(\xi_i,\xi_i,\xi_i)\Big] + R^n_{\Delta t}(x,\xi,\theta),
\end{aligned}$$

with $\mathbb{E}\left[|R^n_{\Delta t}(x,\xi,\theta)|^2\right] \leq C(1+|\theta|^2)(n^{-1}+n\Delta t)^2$. We conclude by taking the expectation of the square in the latter equality and using the independence of $(\xi_i)_{0\leq i<n}$. □