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

Ziad Kobeissi [1 2] **Francis Bach** [1]

Abstract

We consider the problem of learning through observations the value function of an uncontrolled continuous-time process and a reward function. Using vanishing time steps, we propose two adaptations of the well-known TD(0) algorithm: the first one is similar to the one in discrete time and is model-free; the second is model-based and is obtained by adding a zero-expectation term, resulting in a reduction of its variance. In the linear setting, we prove multiple convergence results for the two algorithms, the model-based one is more flexible and enjoys better convergence rates. In particular, using the Polyak-Juditsky averaging method and a constant learning step, we obtain a convergence rate similar to the state of the art on the simpler problem of linear regression using SGD. Finally, we present simulations showing numerical evidence of our theoretical analysis.

1. Introduction

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

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

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 purpose of the present work is to analyse methods for approximating the continuous-time value function V using only observations of the dynamics and the reward obtained with discrete time steps (which will be vanishing to zero, or at least relatively small). For the discrete-time counterpart of (1), this consists in the field of reinforcement learning (RL), and particularly the well-known problem of *policy evaluation*. While adaptations of the theory of RL to continuous time have already been investigated in the deterministic set-up (i.e., when $\sigma = 0$) (Dayan & Singh, 1996; Doya, 2000; Lutter et al., 2021), we are not aware of theoretical research work concerning the stochastic setting. Yet, observe that the deterministic assumption is almost never satisfied in practice in RL: even when the model is deterministic, stochastic perturbations are added in order to favor exploration (see Appendix C.3 for more details). The present work is therefore the first, up to our knowledge, to investigate this direction from a theoretical perspective.

The present analysis will focus on a standard quantity in RL, named the *temporal difference* (TD). In Section 4, two kinds of TDs will be introduced: the standard TD and the stochastic TD. While the former is model-free since it only relies on observations of the dynamics with known time steps, the latter is model-based in the sense that it also requires the knowledge of the drift function b at any observation, i.e., the averaged direction of the dynamics. One of the main contributions of the present work consists in proving convergence results of the most basic RL method for policy evaluation, namely the TD(0) algorithm (Sutton, 1988), with the two above-mentioned TDs, in the high-frequency regime, i.e., when the time steps of the observations converge to zero. Naturally, the model-based method enjoys better convergence rate properties than the model-free method, theoretically and numerically. However, the former cannot be implemented in a model-free environment, while the latter can.

The results proved in the present work emphasize 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 a theoretical analysis is only possible through the restrictive assumption of considering linear parametrization.

[1]Inria - Departement d'informatique de l'ENS, PSL Research University, Paris, France [2]Institut Louis Bachelier, Paris, France. Correspondence to: Ziad Kobeissi <ziad.kobeissi@inria.fr>, Francis bach <francis.bach@inria.fr>.

Links with optimal control. The principal goal of RL is to solve optimal control problems through learning using observations. RL numerical methods pursuing this goal enter the class of *generalized policy iteration* (GPI) methods (Sutton & Barto, 2018). Those methods generally involve the value function or the Q-function. They consist in alternating two interactive updates: first, the *policy evaluation* process that is studied here with the specific example of the TD(0) algorithm; second, the *policy improvement* process whose analysis in the high-frequency regime is out of the scope of the present work. Therefore, the present work consists in a first step in the theoretical analysis of RL algorithms to solve optimal control problems in continuous time. For a more detailed discussion on a way to extend the numerical methods presented hereafter to include the policy improvement process as well, we refer to Appendix C.

Links with partial differential equations (PDEs). The Feynman-Kac formula states that the solution of

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

is given by (1). We refer to Appendix A for a summary of some standard notations from PDE literature that we are using here. Recall that classical methods for solving PDEs (like finite differences or finite elements) cannot be implemented in dimension higher than three, because of the number of grid points growing exponentially with respect to the dimension. Here, we analyse alternative mesh-free methods based on reinforcement learning allowing to consider higher dimensions. In this sense, these methods enter the class of *stochastic RL-based discretisation schemes* for linear PDEs. The literature already includes other PDE methods entering the latter class, the ones we present here have the advantage to only require observations.

Recall that the analysis of the convergence properties of discretisation schemes, when the discretisation steps tend to zero, is an important part of the PDE literature. Concerning RL methods in the high-frequency regime, we are not aware of any work dealing with analogous convergence properties in the stochastic setting. Therefore, as already pointed out, our work is motivated by the fact that stochastic dynamics are important in practice even for deterministic models since exploration noises are generally added.

Fields of application. Generally, we believe that applications of the present work encompass a lot of control problems in the high-frequency regime where only observations are available (and the deterministic part of the dynamics is known when using the stochastic TD). In particular, it encompasses stochastic models and deterministic models where stochastic perturbations are added to favor exploration (see Appendix C.3 for more details). To give a non-exhaustive list of actual practical domains where we believe our work can be of some use, let us cite: first, the robotics in real time, e.g., Le Lidec et al. (2022), where the physics 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, some financial models like the high-frequency trading, e.g., Germain et al. (2021), whose models are based on stochastic differential equations of the form of (1), where the stochastic part of the dynamics generally consists in idiosyncratic noises; third, models of nuclear fusion, e.g., Degrave et al. (2022), in which high-frequency controls of the magnetic fields are necessary, which incorporate measurement noises as well, and where the physics is too complicated for the models to be complete, so additional noises to mimic unconsidered physical phenomena can lead to improving the robustness.

The learning model. Given a set of parametrized functions $(v(\cdot,\theta))_{\theta\in\Theta}$ (where Θ is the space of parameters, usually given by $\mathbb{R}^{d_\theta}$ for some integer d_θ), a learning method, like TD(0) here, computes a sequence of parameters $(\theta_k)_{k\geq 0}$ which will eventually converge to a limit parameter θ^*. In fact, such convergence properties only exist in the literature under the restrictive assumption of a linear parametrization. In this case $v(\cdot,\theta^*)$ is expected to be a good approximation of V *in some sense*. For TD(0), the wording *in some sense* generally means that $v(\cdot,\theta^*) = V$ when V belongs to the space of parametrized functions; and more generally that $v(\cdot,\theta^*)$ and V are near to each other in some functional norm, when the space of parametrized functions is large enough.

The specific case of Langevin dynamics. Let us recall that the dynamic described in (1) is said to be of Langevin type, if the drift function b is given by the gradient of some potential function $U : \Omega \to \mathbb{R}$, i.e., $b = \nabla_x U$. Such dynamics are particularly useful in optimization or optimal control problems (we refer to Appendix C.1 for a practical example). In this case, we show in the present work that θ^*, the eventual limit in the high-frequency regime of TD(0), does in fact satisfy

$$\theta^* \in \text{argmin}_\theta \, \mathbb{E}_{X\sim m}\left[\ell(v(X,\theta), V(X))\right], \quad (3)$$

where m is the invariant measure of the process $(X_t)_{t\geq 0}$ given in (1), for some loss function ℓ given by

$$\ell(v,w) = \rho(v-w)^2 + \tfrac{1}{2}|\sigma^\top \nabla_x (v-w)|^2. \quad (4)$$

Recall that, in general, either in discrete or continuous time, TD(0) cannot be formulated as a minimization method; here, this is a consequence of $\mathcal{L}$ being a symmetric operator on $L^2(m)$, only when the dynamic is of Langevin type.

Observe that, if $\underline{c} I_d \leq \sigma\sigma^\top \leq \overline{c} I_d$ in the common sense of symmetric matrices for some $\underline{c}, \overline{c} > 0$, the latter loss

function is a norm which is topologically equivalent to the H^1-norm, where H^1 is the Sobolev space of once weakly differentiable functions. Therefore, we expect $\nabla_x v(\cdot, \theta^*)$ to be a good approximation of $\nabla_x V$ as well. This is particularly interesting when considering optimal control problems since, in this case, the optimal control admits a closed form given by a function of $\nabla_x V$ (once again, we refer to Appendix C for more details). Another consequence is that, when considering linear parametrizations, considering a H^1-normalisation of the basis function seems to be more convenient than other standard choice (like the common L^2-normalisation). This choice will be used in the simulations of Section 6.

Main contributions. Recall that the purpose of the present work is to adapt methods from RL, which are naturally designed for discrete time, to continuous time. We are particularly interested in giving theoretical guarantees of convergence under an assumption of small time steps (alternatively called the *high-frequency* regime). The main contributions of the present work are:

- When the dynamic is of Langevin type, we show that TD(0) can be interpreted as a minimization method. In particular, its limit is a minimizer of (3).
- In Section 4, for general parametrization of the learned value function (i.e., linear or non-linear), we introduce the stochastic temporal difference and show that it has better asymptotic behaviour in the high-frequency regime than the standard temporal difference (see Lemma 4.1).
- In Section 5, we state two convergence results of the algorithm TD(0) under the additional assumption that the parametrization is linear. First, under a strong convexity assumption achieved by a regularization of the algorithm, using standard decreasing learning rates, Theorem 5.2 gives a standard rate of convergence for stochastic TD(0). It also states a slower convergence rate for standard TD(0), requiring additional care in the choice of the learning rate (see Remark 5.3). Second, only with stochastic TD(0), with a constant learning rate and without strong-convex assumption, Theorem 5.5 yields a fast rate of convergence using an averaging method. This fast rate of convergence is similar to the state of the art for the simpler linear regression problem with SGD (see Section 4.3 for more details on the links between TD(0) and SGD).
- In Section 6, we present numerical simulations, showing empirically that reinforcement learning algorithms, in the high-frequency regime and in presence of stochastic noises, perform better with the stochastic temporal difference than with its standard counterpart.

2. Related works

Temporal-difference learning. The TD algorithm was introduced in the tabular case by Sutton (1988), with later convergence results for linearly dependent features (Dayan, 1992). Asymptotic stochastic approximation results were derived by Jaakkola et al. (1993) for the tabular case, and by Schapire & Warmuth (1996) when using linear approximations, with a non-asymptotic analysis in the *i.i.d.* sampling case (Lakshminarayanan & Szepesvari, 2018).

Stochastic iterative methods. The analysis of TD requires tools from stochastic approximation (Benveniste et al., 1990), which have mainly been derived for stochastic gradient descent (SGD) (Bottou et al., 2018) and reused here. The convergence results presented in the present paper may be compared to standard results on RL algorithms, see Bellman (1966); Kirk (1970) for TD(0). The techniques in the proof (especially concerning the fast-convergence results in Section 5.3) are adapted from the literature on SGD methods (Polyak & Juditsky, 1992; Bach & Moulines, 2013) to the non symmetric setting. In particular, Bach & Moulines (2013) present 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 optimization problem raised by TD(0).

Continuous time RL. Continuous-time reinforcement learning started with Baird (1993), who proposed a continuous-time counterpart to Q-learning; it was later extended by Tallec et al. (2019). From a different perspective, Bradtke & Duff (1994) extended classical RL algorithms to continuous-time discrete-state Markov decision processes. Then, using deterministic dynamics given by ordinary differential equations (ODE), and based on the Hamilton-Jacobi-Bellman (HJB) equation, Doya (2000) derived algorithms for both policy evaluation and policy improvement. Similar deterministic approaches of continuous-time RL have recently been explored by Lutter et al. (2021); Yildiz et al. (2021). In order to balance between exploration and exploitation, Wang et al. (2020) added an entropy-regularization term to a continuous optimization 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. Solving partial differential equations using learning algorithms is a natural idea. Indeed, in general, 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 Lagaris et al. (1998; 2000); Malek & Beidokhti (2006), 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 using deep learning, let us cite Khoo et al. (2021), or Sirignano & Spiliopoulos (2018) for the *Deep Galerkin Method*, or Beck et al. (2019); Han & Jentzen (2017); Han et al. (2018) where the PDEs are reformulated into a backward stochastic differential equations (BSDE) or extensions to forward-backward stochastic differential equations (FBSDE); we refer to the surveys of Germain et al. (2021); Beck et al. (2020) and the references therein for more results on deep learning methods for PDEs. Our method is also inspired from FBSDE, but we investigate the stationary formulation from a theoretical viewpoint, and use a stochastic semi-gradient method such as TD(0) instead of SGD methods as in most of the references above.

3. Assumptions and Limitations

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 Evans (2010) 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 encompass all physical aspects of continuous dynamics, it is complex enough to capture the main ideas for adapting RL methods to continuous models. Consequently, 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).

Accessible informations. We assume that we access a sequence $(\Delta t_k, X_k, X'_k, R_k)_{k\geq 1}$ of observations where $\Delta t_k > 0$ is convergent to zero and (X_k, X'_k, R_k) are independent random variables which can be:

- **Real-world observations:** X'_k and R_k are the real state and reward obtained from following the continuous-time dynamics in (1) on a time interval of length Δt; they are given by $X'_k = \widetilde{X}_{\Delta t_k}$ and $R_k = \frac{1}{\Delta t_k}\int_0^{\Delta t_k} r(\widetilde{X}_t)\,dt$, where $(\widetilde{X}_t)_{t\geq 0}$ satisfies (1) with $\widetilde{X}_0 = X_k$.
- **Observations from a simulator:** X'_k and R_k are obtained using the Euler-Maruyama discretisation scheme of the SDE in (1); X'_k is given by $X'_k = \mathcal{S}_{\Delta t_k}(X_k, \xi_k)$, with $\xi_k \sim \mathcal{N}(0, I_d)$ and $\mathcal{S}_{\Delta t} : (x, z) \mapsto x + \Delta t\, b(x) + \sqrt{\Delta t}\,\sigma(x)z$; R_k satisfies $R_k = r(X_k)$.

Let us define m_k as the law of X_k and 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 for any $f \in C^p(\Omega;\mathbb{R})$, there exists $C_f > 0$ such that, for all $k \geq 0$,

$$|\mathbb{E}\left[f(X_k) - f(X)\right]| \leq C_f \Delta t_k, \tag{5}$$

where X is distributed according to m. In practice, such condition is obtained by following the Markov Chain (which can be continuous or discrete depending on the nature of the observations), using ergodic arguments. We refer to Appendix G.1 for more details on some ergodic results implying that (5) holds (in particular, Theorem G.1 implies (5) for $p = 4$).

In any case, we assume that we can compute b at any state, but not σ. This means that we know the expected direction of the dynamics, but we ignore the exact intensity of the noise. Such assumption is standard when we know the physics of the models but some uncertainties may generate noise, like measurement noises, model approximations (e.g., statistical errors from particle systems), unpredictable external factors (e.g., imperfections of the ground or wind for robotic) or idiosyncratic noise (e.g., in financial models).

The independence assumption on the observations is, for sure, a limitation. When using observations from a simulator, independent samples are easy to obtain. For real-world observations, almost independent samples are generally obtained by shuffling a large dataset. Alternatively, one might assume that the observations are not independent but are obtained from the dynamics following the Markov Chain (which can be continuous or discrete depending on the nature of the observations). We believe that comparable results can be derived from this other assumption but it is out of the scope of the present paper. Moreover, theoretically this alternative assumption is generally more restrictive since it relies on ergodic limits and very small learning rates, making the number of necessary samples increases dramatically in practice.

Few words about the Gaussian noises. One limitation of our analysis is the fact that, throughout the paper, we consider only Gaussian noises. In fact, this naturally comes from the continuous time and state assumption and our willing to consider continuous Markovian noises. Indeed, Donsker's theorem (Donsker, 1951) implies a *Gaussianization* phenomenon of the noise when Δt tends to zero. However, we believe that non-gaussian noises can be observed in practice even for very short time steps. In this case, we believe that our strategy will still improve numerical results even if we have no clue on how to adapt the present theoretical analysis.

4. The Two Temporal Differences

4.1. Comparing Standard and Stochastic TDs

For a given observation $(\Delta t, X, X', R)$ and a parameter θ, one usually computes δ the *standard temporal difference*,

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

where $\gamma_{\Delta t} = e^{-\rho \Delta t}$, and the scaling $\frac{1}{\Delta t}$ is chosen accordingly for the expectation of δ to be of order $O(1)$. This quantity is initially designed for discrete-time dynamics and is derived from Bellman's programming principle (Bellman, 1966). In the present work, our first main contribution is to introduce and analyse $\widetilde{\delta}$ the *stochastic temporal difference*,

$$\begin{cases} \widetilde{\delta}_{\Delta t} = \delta_{\Delta t} + \frac{1}{\Delta t} Z \\ Z = (X' - X - \Delta t\, b(X)) \cdot \nabla_x v(X,\theta). \end{cases} \tag{6}$$

The two temporal differences δ and $\widetilde{\delta}$ only differ by the additional term Z, that will be called *(stochastic) correction term* in the following. Observe that, given X, Z has a zero expectation, which implies that

$$\mathbb{E}_{X'}\left[\widetilde{\delta}_{\Delta t} \middle| X\right] = \mathbb{E}_{X'}\left[\delta_{\Delta t} \middle| X\right]. \tag{7}$$

The latter quantity is known as the *Bellman error*. A standard way to check that the learned 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. Recall that the Bellman error is exactly equal to zero when $v(\cdot,\theta) = V$. However, the Bellman error is in general not convenient to compute from observations, neither it is useful for learning (since a lot of observations should be computed for any fixed different value of X). Therefore, we generally consider other quantities, easier to compute, like the average square of the TD, which admits the following decomposition,

$$\begin{aligned} \mathbb{E}_{(X,X')}\left[|\delta_{\Delta t}|^2\right] =& \mathbb{E}_X\Big[\underbrace{\mathbb{E}_{X'}\left[\delta_{\Delta t}\middle|X\right]^2}_{\text{Bellman error}}\Big] \\ &+ \underbrace{\mathbb{E}_X\left[\operatorname{Var}_{X'}\left(\delta_{\Delta t}\middle|X\right)\right]}_{\text{perturbating term}}. \end{aligned} \tag{8}$$

A similar formula holds for $\widetilde{\delta}$. The following lemma is a first evidence of how $\widetilde{\delta}$ should be more adapted to the present situation than δ, showing that the perturbating term has a different order of magnitude.

Lemma 4.1. *Assume that r, b and σ are uniformly 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,*

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

In the one hand, because the variance explodes as $\frac{1}{\Delta t}$, the latter lemma directly implies that the perturbating term in (8) 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

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

Before properly introducing TD(0) in the next section, let us mention that TD(0) is in fact a *stochastic semi-gradient descent* method on (8). Therefore, using the above asymptotic results, at least formally, one may easily think that using $\widetilde{\delta}$ instead of δ should lead to a significant improvement for algorithms like TD(0). This thinking will be made mathematically rigorous in the remaining of the paper.

Since TD(0) is a stochastic semi-gradient method, it admits a stochastic gradient counterpart, called *residual gradient*. The latter method will be analysed in Appendix D. In particular, the residual gradient method is more sensitive to the fact that the perturbating term converges to infinity for δ. This implies that the algorithm will fail to learn the value function in the high-frequency regime when using δ: $v(\cdot,\theta)$ will converge to a constant function. On the other hand, using $\widetilde{\delta}$ instead leads the residual gradient method to learn a regularized approximation of V.

4.2. The TD(0) 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. The standard and stochastic TD(0) methods

write as

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

where α_k is the learning rate. We also introduce the regularized standard and stochastic TD(0) methods as

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

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. At least in the case of a Langevin dynamics, introducing μ in the latter algorithm consist in adding a ℓ_2 penalizations of the form $\rho\mu|\theta|^2$ in the optimization problem (3). In practice, this is used to favor regular solutions and to reduce overparametrization. Here, it will help by making the problem μ-strongly convex. The projection step, consisting in applying Π_{B_M}, is also common in the literature.

4.3. The limit $\rho \to \infty$ and the link with SGD

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 and $\rho\Delta t$ tends to zero. Indeed, from (1) we easily obtain that,

$$\lim_{\rho \to 0} V(x) = r(x),$$

for $x \in \Omega$, and that the loss function (4) converges to the square loss. Moreover, for a fixed starting state $X = x \in \Omega$, in the latter asymptotic regime, we get

$$\widetilde{\delta}_{\Delta t} = v(x, \theta) - r(x) + o(1). \tag{9}$$

At least formally, we can thus express the SGD algorithm, applied to the least-square regression of r, as a limit 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 above) equal to zero. In this situation, the fact that SGD is a particular and strictly simpler case than TD(0) (some difficult terms are removed, such that the one making TD(0) be a bootstrapping algorithm) appears more clearly in this case.

This explains why we do not expect the convergence rates proved in the present work to be better than the state of the art of the literature on SGD. Let us recall that Theorem 5.5 does yield a similar convergence rate as the state-of-the-art results on mere convex problem with averaging method (Bach & Moulines, 2013).

5. Convergence Results in the Linear Setting

Throughout this section, we assume (5) and the following assumptions:

A1 The function v is linear with respect to $\theta \in \mathbb{R}^{d_\theta}$, i.e., $v(x, \theta) = \theta^\top \varphi(x)$ where $\varphi : \Omega \to \mathbb{R}^{d_\theta}$ is called the feature vector.

A2 The functions r, b, σ are C^p, where p comes from (5).

A3 The feature vector φ is C^{p+2}, its coordinate functions are linearly independent.

5.1. Identification of the limits

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

$$\mathbb{E}_m\left[\varphi(X)\mathcal{L}\varphi(X)\right]\theta^* = \mathbb{E}_m\left[r(X)\varphi(X)\right]. \tag{10}$$

The linear independence assumption on the coordinate functions of φ in Assumption **A3** implies that $\mathbb{E}\left[\varphi(X)\varphi(X)^\top\right] \in \mathbb{R}^{d_\theta \times d_\theta}$ is positive definite. Therefore, Lemma F.2 in the Appendix implies that the symmetric part of $\mathbb{E}\left[\varphi(X)\mathcal{L}\varphi(X)\right]$ is positive definite as well. This implies that θ^* is well defined and there exists $M_0 > 0$ such that $|\theta^*| \leq M_0$. The latter quantity consists in a convenient choice for M in (μ-TD0), especially when μ is small. When μ is not small, we will prefer the simpler quantity $M_\mu = \frac{\|r\|_\infty}{\mu}$. In general, for ($\mu$-TD0), we will always assume that

$$M \geq \min(M_0, M_\mu). \tag{11}$$

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

$$\left(\mu I_d + \mathbb{E}_m[\varphi(X)\mathcal{L}\varphi(X)]\right)\theta^*_\mu = \mathbb{E}_m\left[r(X)\varphi(X)\right]. \tag{12}$$

Moreover, the distance between θ^* and θ^*_μ might be bounded.

Lemma 5.1. *There exists $C > 0$ such that, for any $\mu > 0$,*

$$|\theta^* - \theta^*_\mu| \leq C\mu.$$

Finally, recall that if the drift function admits the form $b = \nabla_x U$ for some potential function U, θ^* can alternatively be defined by (3); and θ^*_μ by (3) plus an additional ℓ_2-penalization of the form $\rho\mu|\theta|^2$.

5.2. Regularized TD(0)

This section may be thought of as a warm-up for the next one. We will only consider the algorithms (μ-TD0). The proof of the following theorem is less technical than the one of Theorem 5.2 below, and the result may easily be compared

to classical results in the literature, e.g., Sutton (1988). In particular, we use the common decreasing assumption on the learning rate, i.e., that it is proportional to $1/(\mu(k+1))$. Then, using different value for Δt_k, we obtain the usual convergence rate in $1/k$.

Theorem 5.2. *Take $(\theta_k)_{k\geq 0}$ and $(\widetilde{\theta}_k)_{k\geq 0}$ defined by (μ-TD0) with $\mu > 0$, M satisfying (11), and $\alpha_k = \frac{2}{\mu(k+1)}$. For $c > 0$ and $\Delta t_k = \frac{c}{(k+1)^{\frac{1}{3}}}$ for any $k \geq 0$, there exists $C > 0$ such that, for $k \geq 1$,*

$$\mathbb{E}\left[\left|\theta_k - \theta^*_\mu\right|^2\right] \leq \frac{C}{\mu^2 k^{\frac{2}{3}}}.$$

For $c > 0$ and $\Delta t_k \leq \frac{c}{\sqrt{k+1}}$ for any $k \geq 0$, there exists $C > 0$ such that, for $k \geq 1$,

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

Using an averaging method, we might reduce the factor $1/\mu^2$ into $1/\mu$ with the same assumptions, this is a first motivation to introduce an averaging method in the next section.

Remark 5.3. From Theorem 5.2, it looks like using the standard TD δ, instead of the stochastic one $\widetilde{\delta}$, only leads to a loss in the exponent of the convergence rate. In fact, this is not the only drawback: indeed, when using δ, one has to make sure that $\frac{\alpha_k}{\Delta t_k}$ tends to zeros sufficiently fast, which is not necessary when using $\widetilde{\delta}$. To illustrate the latter claim, consider the case $\alpha_k = \Delta t_k$, with a constant real diffusion $\sigma > 0$ and a Langevin dynamic $b = \nabla_x U$. This boils down to having θ_k approximating a continuous time stochastic process $(\theta_t)_{t\geq 0}$ (where we use the same notation by a slight abuse) which satisfies the SDE

$$d\theta_t = -\tfrac{1}{\rho}\nabla_\theta \ell(V, v(\cdot, \theta_t)) + \sigma dW_t,$$

which is of Langevin type as well, where ℓ is the loss function in (4). In this case, θ_k will converge to a random variable with values in Θ, whose law admits a density proportional to $e^{-\frac{\sigma^2}{2\rho}\ell(V,v(\cdot,\theta))}d\theta$, instead of begin a Dirac mass at θ^* as we would expect (and as stated by Theorem 5.2 under different assumptions). We conclude that another advantage of using $\widetilde{\delta}$ over δ, which does not directly transcript from Theorem 5.2, is that it leads to a more flexible method in which one does not have to precisely investigate the relative size of α with respect to Δt.

Finally, combining the results of Theorem 5.2 and Lemma 5.1, we obtain the following non-asymptotic error bound.

Corollary 5.4. *Under the same assumption as in Theorem 5.2, after $K \geq 2$ iterations with $\mu = K^{-\frac{1}{4}}$, we obtain*

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

5.3. Averaging Stochastic TD(0)

In the same spirit as the results from Bach & Moulines (2013), in this section, we get the convergence of the algorithm TD(0) (instead of SGD for Bach & Moulines (2013)) with constant learning step, without a strong convexity assumption, without a regularization assumption and without a projection map. Recall that Remark 5.3 stated that $\frac{\alpha_k}{\Delta t_k}$ should converge to zero in order to use the standard TD δ. Since here α is constant, we only consider the stochastic TD $\widetilde{\delta}$. In particular, the proof of Theorem 5.5 does not hold for δ.

More precisely, in order to accelerate the convergence of (TD0), We use the Polyak-Juditsky averaging method, see Polyak & Juditsky (1992), for $k \geq 1$,

$$\overline{\theta}_k = \frac{1}{k}\sum_{i=0}^{k-1} \widetilde{\theta}_i, \tag{13}$$

We obtain a convergence rate which is competitive with the state of the art for the simpler problem of linear regression using SGD methods.

Theorem 5.5. *If $\sum_{i=0}^{\infty} \Delta t_i^2$ is finite, there exist $C, R > 0$ such that, for $\alpha < R^{-2}$, $k \geq 1$ and $H = \mathbb{E}\left[\varphi(X)\mathcal{L}\varphi(X)^\top\right]$,*

$$\mathbb{E}\left[v(X, \bar{\theta}_k) - v(X, \theta^*)\right] \leq \frac{C}{\alpha k} + \frac{C(d + \mathrm{tr}(HH^{-\top}))}{k},$$

Assume instead that $\sum_{i=0}^{k-1} \Delta t_i^2 \leq a\ln(1+k)$ for some $a > 0$ and any $k \geq 0$. Then, for any $\varepsilon > 0$, there exists $C, R > 0$ such that for $\alpha < R^{-2}$, $k \geq 0$, we get

$$\mathbb{E}\left[v(X, \bar{\theta}_k) - v(X, \theta^*)\right] \leq \frac{C}{\alpha k} + \frac{C(d + \mathrm{tr}(HH^{-\top}))}{k^{1-\varepsilon}}.$$

The proof is adapted from Bach & Moulines (2013) with the extra difficulties that the linear operators applied to θ_k in (TD0) are different for any $k \geq 0$, and they are not symmetric, even their symmetric part has no interesting properties (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 $\sum_i \Delta t_i^2$.

6. Numerical Simulations

The simulations in this section are made in dimension $d = 1$, in the one-dimensional torus represented as $[-0.5, 0.5]$ with periodic boundary conditions. We consider a linear parametrization with feature vectors given as the first Fourier functions. The state satisfies the dynamic in (1) with a constant diffusion satisfying $\sigma^2/2 = 0.1$, and the drift and reward functions are given by, for $\sigma_g = 0.1$,

$$b = -\frac{\sigma^2}{2}\nabla_x U, \quad r(x) = 20x^3 - 5x,$$
$$\text{where } U = -\ln\Big(e^{-\frac{(x-0.25)^2}{2\sigma_g^2}} + e^{-\frac{(x+0.25)^2}{2\sigma_g^2}}\Big).$$

For more details on the model we refer to Appendix B.1, see Figure 2 for plots of b, U, r and m. In Figure 1, under the assumption of real-world observations described in Section 3, we compare the results obtained by running the two algorithms (TD0) with and without the Polyak-Juditsky averaging (13). The time step is equal to 10^{-6}. In Figure 1, we observe that the algorithms perform better with the stochastic correction term. We also show the convergence rate of the gradient or the learned function.

In Appendix B.2, we experimentally show that the proposed algorithms are robust to: using off-policy drawing (i.e., the samples are not generated according to the invariant measure anymore), using offline drawing (i.e., the samples are not independent anymore), using non-linear parametrizations. In any case, the algorithms using the correction term efficiently learn good approximations of the solution; while the uncorrected algorithms fail. The CPU time for the 20000 iterations of one run of Figure 1 is approximately 30 seconds on a standard laptop. Running all the experiments of this paper took about one hour. The Python code is provided in the supplementary material.

7. Conclusion

In the present work, we proved that standard reinforcement learning methods based on the temporal difference are not adapted to solve continuous stochastic optimization, nor their discretisations using small time-steps. On the one hand, the standard temporal difference can be used to design converging numerical methods (it has the advantage of being model-free), but the convergence rate becomes slower than in the discrete-time situation. Moreover, an additional care has to be taken concerning the choice of the learning rate (see Remark 5.3). On the other hand, we introduce a more adapted object, namely the stochastic temporal difference. Its drawback is that it is model-based. Its advantages are that we obtain the similar rate of convergence than in discrete time, and that it is more flexible (especially concerning the choice of the learning rate).

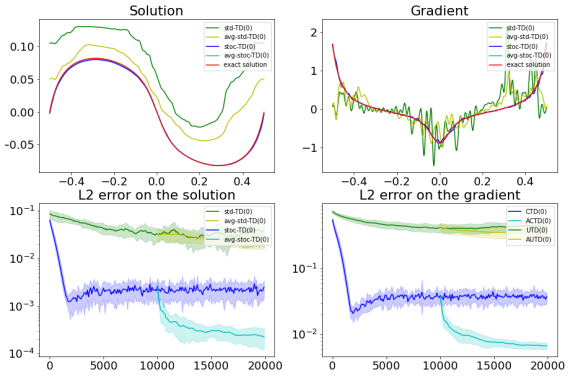

Figure 1. We compare four algorithms: standard TD(0) in green (std-TD(0)), averaged standard TD(0) in yellow (avg-std-TD(0)), stochastic TD(0) in blue (stoc-TD(0)), averaged stochastic TD(0) in cyan (avg-stoc-TD(0)). The learned function and its gradient for one run are shown in top left and right respectively. The convergence speed in L^2-norm are displayed for the learned function and its gradient in the bottom left and right respectively. The exact solution is computed using a PDE solver. The algorithm is online and on-policy, the parameters are: $\Delta t = 10^{-6}$, $\rho = 10$ (i.e., the discount factor is $\gamma = e^{-\rho\Delta t} \approx 1 - 10^{-5}$), the dimension of the feature vector is 101, the batch size is 100, the learning rate is 10^{-3} and a ℓ_2 weight penalization with a coefficient 10^{-5} is added. The averaged algorithms start after 10000 iterations. Each experiment is repeated twenty times with different random seeds, the standard deviation is shown on the two bottom figures.

More precisely, in order to show theoretical convergence guarantees and rates, like it is usual in the literature, we have to adopt the restrictive linear parametrization assumption. We then show two types of convergence results. First, Theorem 5.2 shows convergence First, under a strong convexity assumption achieved by a regularization of the algorithm, using standard decreasing learning rates, Theorem 5.2 shows convergence of TD(0) with both temporal differences. Second, only with stochastic TD(0), with a constant learning rate and without strong-convex assumption, Theorem 5.5 yields a fast rate of convergence using an averaging method (which is comparable to the state of the art for the simpler linear regression problem with SGD).

From a numerical point of view, we show on simulations that adding the stochastic correction term results in better convergence properties. Moreover, this conclusion holds when removing the assumptions that the samples are drawn from the invariant measure, that they are independent, and that the parametrization is linear.

As stated in the introduction, we recall that the original problem of the present work (approximating the continuous time value function (1)) has numerous important consequences in optimal control and numerical resolution of partial differential equations, and might have applications in domains like robotics, finance and nuclear fusion for instance.

References

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

Baird, L. Advantage updating. Technical report, Wright Lab Wright-Patterson AFB OH, 1993.

Beck, C., Jentzen, A., 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.

Beck, C., Hutzenthaler, M., Jentzen, A., and Kuckuck, B. An overview on deep learning-based approximation methods for partial differential equations. *arXiv preprint arXiv:2012.12348*, 2020.

Bellman, R. Dynamic programming. *Science*, 153(3731): 34–37, 1966.

Benveniste, A., Métivier, M., and Priouret, P. *Adaptive Algorithms and Stochastic Approximations*, volume 22. Springer Science & Business Media, 1990.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Bradtke, S. and Duff, M. Reinforcement learning methods for continuous-time Markov decision problems. *Advances in Neural Information Processing Systems*, 7, 1994.

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

Dayan, P. and Singh, S. P. Improving policies without measuring merits. *Advances in Neural Information Processing Systems*, pp. 1059–1065, 1996.

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de Las Casas, D., et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

Donsker, M. D. An invariance principle for certain probability linit theorems. AMS, 1951.

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

Evans, L. C. *Partial differential equations*, volume 19. American Mathematical Soc., 2010.

Germain, M., Pham, H., and Warin, X. Neural networks-based algorithms for stochastic control and pdes in finance. *arXiv preprint arXiv:2101.08068*, 2021.

Han, J. and Jentzen, A. 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.

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

Jaakkola, T., Jordan, M., and Singh, S. Convergence of stochastic iterative dynamic programming algorithms. *advances in Neural Information Processing Systems*, 6, 1993.

Kerimkulov, B., Siska, D., and Szpruch, L. Exponential convergence and stability of howard's policy improvement algorithm for controlled diffusions. *SIAM Journal on Control and Optimization*, 58(3):1314–1340, 2020.

Khoo, Y., Lu, J., and Ying, L. Solving parametric pde problems with artificial neural networks. *European Journal of Applied Mathematics*, 32(3):421–435, 2021.

Kirk, D. E. Optimal control theory: An introduction. 1970.

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

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

Lagaris, I. E., Likas, A. C., and Papageorgiou, D. G. Neural-network methods for boundary value problems with irregular boundaries. *IEEE Transactions on Neural Networks*, 11(5):1041–1049, 2000.

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

Le Lidec, Q., Montaut, L., Schmid, C., Laptev, I., and Carpentier, J. Leveraging randomized smoothing for optimal control of nonsmooth dynamical systems. *arXiv preprint arXiv:2203.03986*, 2022.

Lutter, M., Mannor, S., Peters, J., Fox, D., and Garg, A. Value iteration in continuous actions, states and time. *arXiv preprint arXiv:2105.04682*, 2021.

Malek, A. and Beidokhti, R. S. Numerical solution for high order differential equations using a hybrid neural networkoptimization method. *Applied Mathematics and Computation*, 183(1):260–271, 2006.

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

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

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

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

Sirignano, J. and Spiliopoulos, K. DGM: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, 375:1339–1364, 2018.

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

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: an Introduction*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2018.

Tallec, C., Blier, L., and Ollivier, Y. Making deep Q-learning methods robust to time discretization. In *International Conference on Machine Learning*, pp. 6096–6104, 2019.

Wang, H., Zariphopoulou, T., and Zhou, X. Y. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21:198–1, 2020.

Yildiz, C., Heinonen, M., and Lähdesmäki, H. Continuous-time model-based reinforcement learning. In *International Conference on Machine Learning*, pp. 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. More details about the model

The state satisfies the dynamics in (1) with a constant diffusion $\sigma > 0$ and a drift function b given by:

$$b = -\frac{\sigma^2}{2} \nabla_x U, \quad \text{where } U = \frac{1}{2\sigma_g^2} x^2 - \ln(\cosh(\tfrac{0.25x}{\sigma_g^2})).$$

One may notice that the function U has not the same form as the one given in Section 6. The present function U is derived from the one in Section 6 by adding an additive constant, this does not change the drift, neither the invariant measure (it is only more convenient for calculation). The function U is called the potential and $\sigma_g = 0.1$ may be different from σ, the index g stands for *generator*. We refer to the top left and bottom left figures in 2 for a graphical representation of b and U respectively. Here, m the invariant measure of the SDE admits a simple closed form, as the Gibbs measure of the potential U, i.e.,

$$m = N^{-1} e^{-U} = N^{-1} \Big(e^{-\frac{(x-0.25)^2}{2\sigma_g^2}} + e^{-\frac{(x+0.25)^2}{2\sigma_g^2}} \Big).$$

where $N > 0$ is a normalizing constant given by $N = \int_{[-0.5,0.5]} e^{-U(x)} dx$ such that m is a probability density function. One may recognize a sum of two Gaussian conditioned to belong to $[-0.5, 0.5]$, see the bottom left figure in 2 for a graphical representation. Consequently, it is easy to sample from the invariant density. The reward function is given by the following polynomial function:

$$r(x) = 20x^3 - 5x,$$

see the top left figure in 2 for a graphical representation.

In order to learn in the torus, the Fourier basis seems particularly adapted, so the linear feature vectors φ will be taken as

$$\varphi_1(x) = 1, \quad \varphi_{2k}(x) = \frac{\sin(2k\pi x)}{\sqrt{\frac{1}{2}(1 + 4k^2\pi^2)}}, \quad \text{and} \quad \varphi_{2k+1}(x) = \frac{\cos(2k\pi x)}{\sqrt{\frac{1}{2}(1 + 4k^2\pi^2)}},$$

for $1 \leq k \leq (d_\theta - 1)/2$ where d_θ is assumed to be odd. Let us recall that the derivative of φ is involved in the computation of the variance-reduction term. Therefore we believe (and we saw experimentally) that normalizing the basis in H^1-norm leads to better learning performances. For another argument to favor a H^1-normalisation, we refer to the discussion on the case of Langevin dynamics in Section 1.

B.2. Changing the learning assumptions

In this section, we show more numerical simulations in which we remove some of the main assumptions that was necessary to obtain the theoretical convergence rates stated in Section 5.

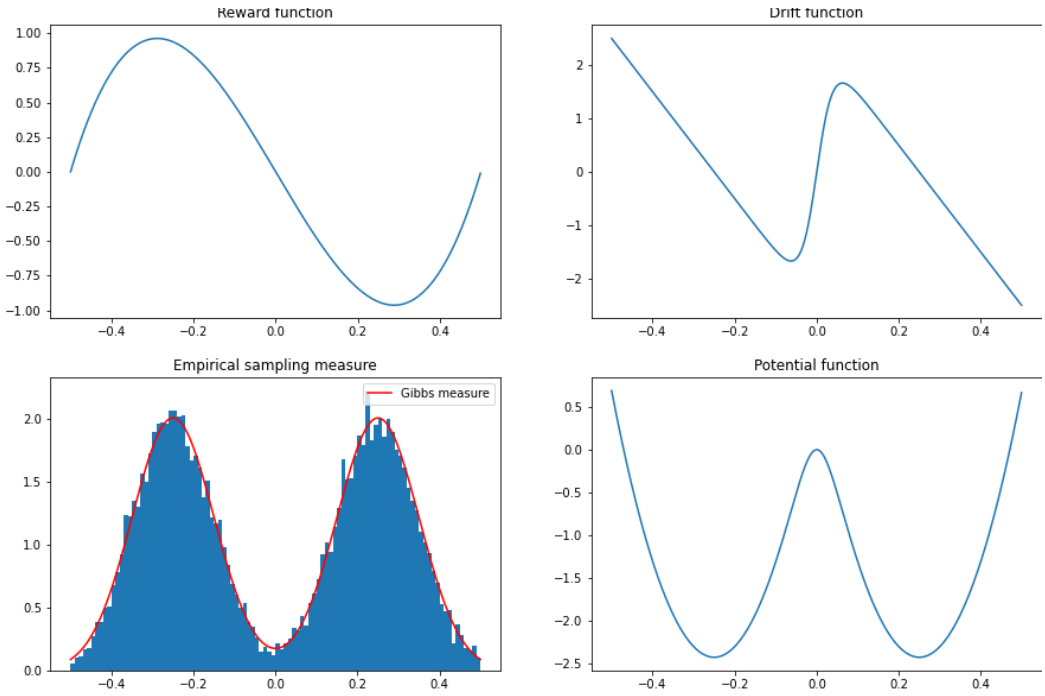

Figure 2. Here are some graphical representations from the Markov Chain. Namely, the reward function is at the top left; the drift is at the top right; on bottom left, is a histogram of presence of 10000 samples divided in 100 subintervals; the potential U is at bottom right. We took $\sigma^2/2 = 0.1$.

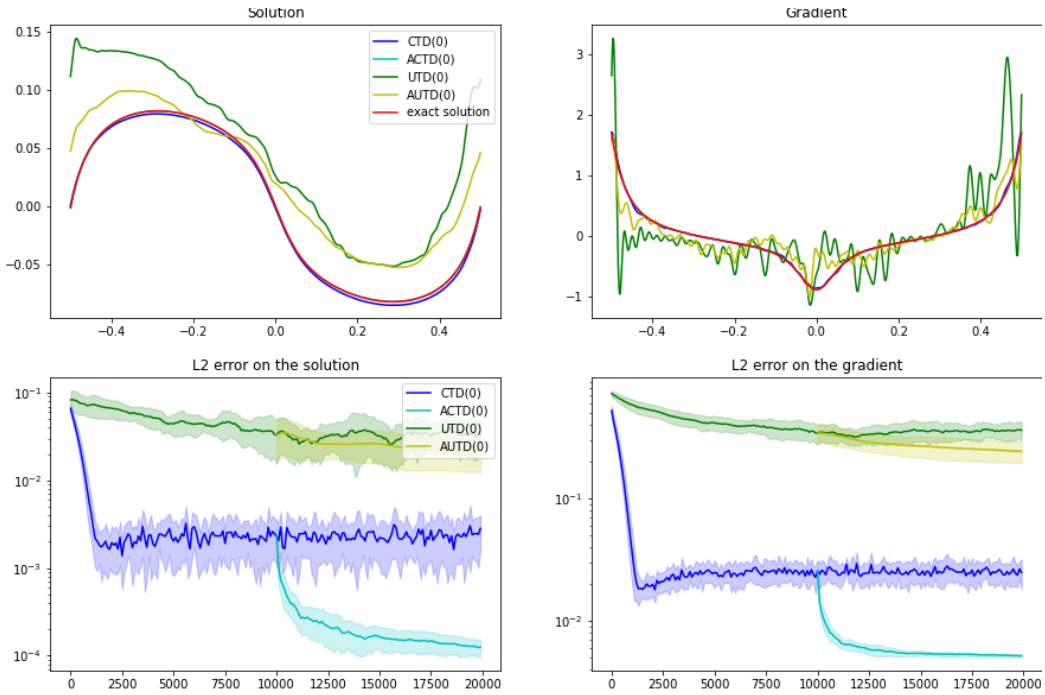

Figure 3. Same parameters as in Figure 1 but the only thing changing is the fact that the samples are now drawn using the uniform distribution.

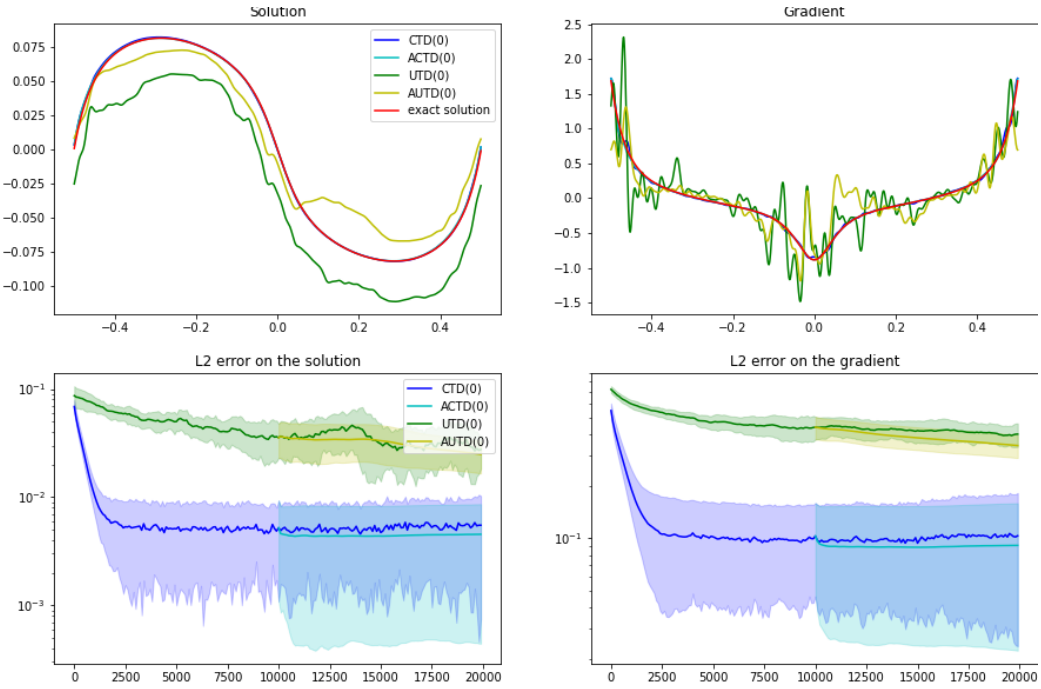

Figure 4. Same parameters as in Figure 1 but the only thing changing is the fact that the total number of samples is 1000. Each sample is used 2000 times through the learning.

First, let us remove the assumption that the samples are generated according to the invariant measures. In figure 1, the samples are now generated from the uniform distribution. We observe that the conclusions of Figure 1 hold and that the same improvement happen when adding the variance-reduction term.

Then, let us remove the assumption that the samples are independent. We consider offline learning, meaning that a limited set of samples is sampled, then they are used multiple times in the learning routine. In Figure 4 we take only 1000 samples, each will be used 2000 times during the learning. We would like to insists that this represents a small number of samples, therefore it should have a negative impact on the convergence rate of all algorithms. We refer to Figure 5 for an example of a histogram of presence with only 1000 samples. In particular, one may notice that some array are over-represented, some are sub-represented, and worst, some are not represented at all. We would like to insist on the fact that the goal here is to show experimentally that the variance-reduction correction improves the convergence properties of the algorithms even in this extreme situation.

In Figure 6, we consider a nonlinear and non-regular parametrization function. More precisely, the value function admits the following form:

$$x \mapsto v(\cos(x), \sin(x), \theta),$$

where v is the function generated by a fully-connected neural network with four hidden layers and twenty neurons in each layer. The value function v is not regular since we are using ReLU activation functions. We see in Figure 6, that the conclusions of the main text seem to hold in this situation.

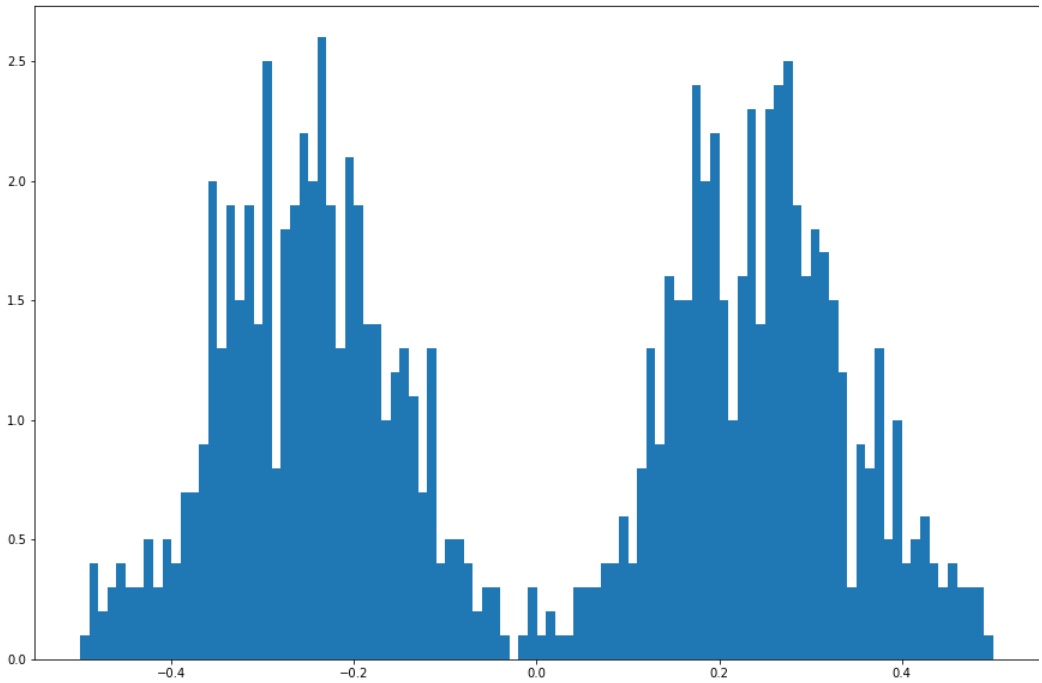

Figure 5. The histogram of presence when only 1000 samples are drawn from the invariant measure. The histogram is divided in 100 subintervals.

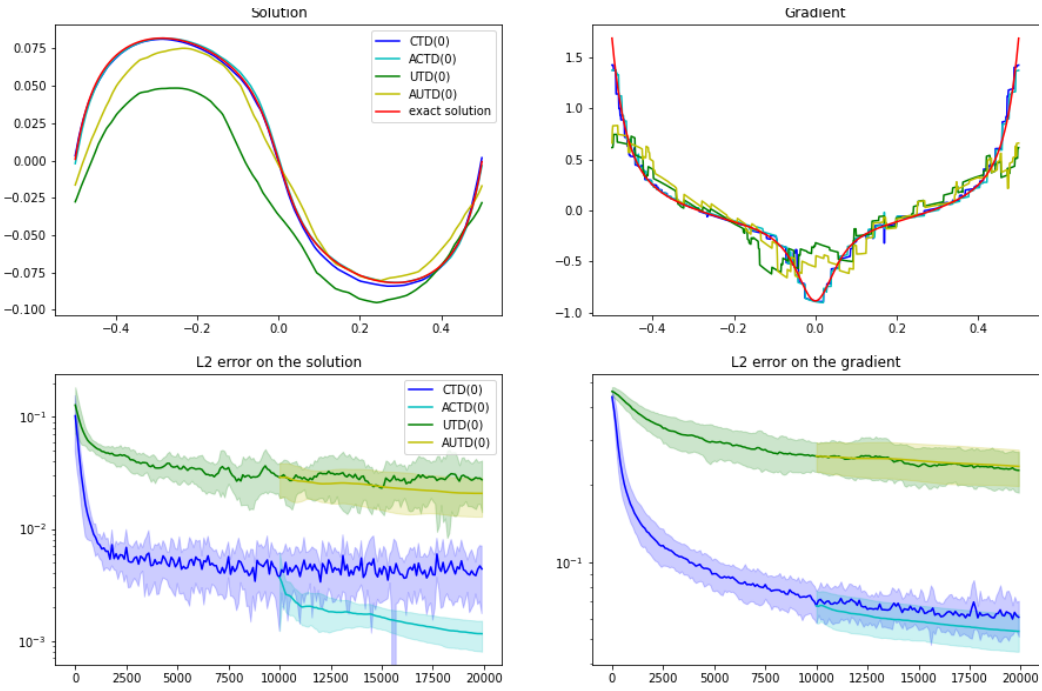

Figure 6. Same parameters as in Figure 1 but the only thing changing is the parametrization, here the value function is given by $x \mapsto v(\cos(x), \sin(x), \theta)$ where θ are the weights of a fully-connected neural network with four hidden layers and twenty neurons per layer.

C. Application to continuous-time reinforcement learning

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

Let us consider the controlled counterpart of (1),

$$dX_t = b(X_t, u(X_t))dt + \sigma(X_t, u(X_t))dW_t, \tag{14}$$

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

$$J(u) = \mathbb{E}\Big[\int_0^\infty e^{-\rho t} r(X_t, u(X_t))dt\Big], \tag{15}$$

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:

$$V^u(x) = \mathbb{E}\Big[\int_0^\infty e^{-\rho t} r(X_t, u(X_t))dt \Big| X_0 = x\Big]. \tag{16}$$

Then, solving the optimization problem (15) 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 characterized as the solution of a partial differential equation, named Hamilton-Jacobi-Bellman (HJB) equation (see (Bellman, 1966) for more details), given by

$$\rho V^* - \max_{u \in \mathcal{A}} H(x, \nabla_x V^*(x), D_x^2 V^*(x), u) = 0, \tag{17}$$

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

$$H(x, p, z, u) = r(x, u) + p \cdot b(x, u) + \frac{1}{2}\text{tr}\left(\left(\sigma\sigma^\top\right)(x, u)z\right). \tag{18}$$

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

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

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 (17) can be rewritten as

$$\rho V^* - \tfrac{\sigma^2}{2}\Delta_x V^* - \widetilde{H}(x, \nabla_x V^*) = 0, \tag{19}$$

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

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

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,

$$u^*(x) = \nabla_p \widetilde{H}(x, \nabla_x V^*(x)). \tag{20}$$

Remark C.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.

C.2. Application of the present work to a discretised MDP

Here, using the Euler-Maruyama discretisation scheme on (14) (like in the main text in Section 3 in the case of observations from a simulator), we obtain the following controlled discrete step operator,

$$X_{t_{i+1}} = S_{\Delta t}(X_{t_i}, u(X_{t_i}), \xi_i) := X_{t_i} + \Delta t\, b(X_{t_i}, u(X_{t_i})) + \sqrt{\Delta t}\sigma(X_{t_i}, u(X_{t_i}))\xi_i. \tag{21}$$

This time, the latter step operator does not characterize 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 *generalized policy iteration* (GPI) methods (Sutton & Barto, 2018). 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 maximize 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 examples consist in the *policy iteration* methods 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 (18) as the Hamiltonian.

This algorithm is convergent with a super-linear convergence rate, see (Puterman & Brumelle, 1979; Puterman, 1981; Kerimkulov et al., 2020). In particular, this method may be seen as an application of a Newton algorithm 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 maximizer (resp. almost a maximizer) 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 (19) 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 parametrized 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 (6),

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

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 other.... The changes only consists in replacing any temporal difference in an algorithm by one of the two TD(0) algorithms presented here.

Remark C.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.

C.3. Intrinsic and artificial noises

In standard discrete-time reinforcement learning, two kinds of noises are often considered. The first one come 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:

- **C1** stochastic models of the form of (1) with an intrinsic noise,
- **C2** deterministic models with linear dynamics with respect to the controls, and an artificial noise added for exploration, regularization or for smoothing the control (see Le Lidec et al. (2022)).
- **C3** 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 **C1**, 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 **C2**, the dynamics has the following form,

$$\frac{d}{dt}x_t = A(x_t)u(x_t) + B(x_t), \tag{22}$$

where 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 minimizers of some entropy-relaxations of the optimization problems (see Wang et al. (2020) 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 3, 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 **C3** 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. The residual gradient method

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

In this section, we state similar results as Theorems 5.2 and 5.5 while replacing the TD(0) methods by residual gradient (RG) methods. We want to insist that, in Section D, 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 regularized standard and stochastic RG methods as

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

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-regularized method. Let us define F_μ the RG cost,

$$F_\mu(\theta) = \mathbb{E}_X\left[(\mathcal{L}v(X,\theta) - r(X))^2\right] + \frac{1}{2}\mathbb{E}_X\left[\text{tr}\left((\sigma\sigma^\top D_x^2 v(X,\theta))^2\right)\right] + \frac{\mu}{2}|\theta|^2.$$

Then, θ^*_μ is defined by,

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

Similarly, we define $F = F_0$ and $\theta^* = \theta^*_0$. The following theorem is the counterpart to RG of Theorem 5.2, it concerns the convergence rate of the regularized RG method.

Theorem D.1. *Assume* **A1**, **A2**, **A3**, $\mu > 0$, $\alpha_k = \frac{2}{\mu(k+1)}$ *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 E.3 for the proof.

Then, we state below the counterpart to RG of Theorem 5.5, it concerns the convergence rate of the unregularized RG method with constant learning step and an averaging method.

Theorem D.2. *Assume A1, A2 and A3. 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 $\alpha < R^{-2}$, $k \geq 1$,*

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

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

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

The proof consists in repeating the proof of Theorem 5.5, it is even simpler since here all the linear operators are symmetric.

D.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 replace F_μ by $\widetilde{F}_\mu$ defined by

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

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

In particular, Theorems D.1 and D.2 hold with this four alternatives.

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

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

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 D.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 $\sigma_k \leq ck^{-\frac{1}{8}}$, for some $c > 0$, for any $k \geq 0$. In particular the proofs or the counterparts to the multistep setting of Theorems D.1 and D.2, are similar to the originals but using Lemma F.6 below instead of Lemma F.1.

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

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

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 (6). 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, 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 5.2 then hold with $\alpha_k = \frac{4}{\mu(k+1)}$, $\Delta t_k \leq c/\sqrt{k+1}$, and $\sigma_k \leq ck^{-\frac{1}{8}}$, for $k \geq 0$.

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

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

where $(N_k)_{k\geq 0}$ are the size of the mini-batches. The conclusions of Theorem 5.2 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 (8) 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 generalized 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 (Donsker, 1951), 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 5.2 then hold with $\alpha_k = \frac{4}{\mu(k+1)}$ and $\Delta t_k \leq c/(k+1)$, for $k \geq 0$.

E. 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 simplification since the learned function v is assumed to be regular.

E.1. Proof of Theorem 5.2

Theorem E.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 $\mathbb{E}\left[|g_k|^2\middle|\theta_k\right] \leq C(1 + |\theta_k|^2)$. Then $(\theta_k)_{k\geq 0}$ is convergent in expectation to θ^* and

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

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 $b_k = \mathbb{E}\left[|\theta_k - \theta^*|^2\right]$. 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

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

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

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

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

Proof of Theorem 5.2. The proof only consists in checking that we can apply Theorem E.1. Using the same notation as in Theorem E.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 F.1. From (5), we get

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

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

E.2. Proof of Theorem 5.5

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 5.5. 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 F.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 F.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 $\mathbb{E}[H_k(X_k)] = H$. On 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,

$$\begin{aligned}\eta_{k+1}^r &= (I_d - \alpha H)\eta_k^r + \chi_k^r + \Delta t_k\psi_k^r, \\ \eta_0^0 &= \eta_0, \qquad \eta_0^r = 0 \text{ if } r \ge 1,\end{aligned} \tag{23}$$

where χ_k^r and ψ_k^r are defined by

$$\begin{aligned}\chi_k^0 &= \alpha(H - H_k(X_k))\theta^* + \alpha\left(\mathbb{E}\left[\varphi(X_k)r(X_k)\right] - \varphi(X_k)r(X_k)\right), \\ \psi_k^0 &= \alpha\Delta t_k^{-1}\left(\mathbb{E}[\varphi(X_k)\mathcal{L}v(X_k,\theta^*)] - \mathbb{E}[\varphi(X)\mathcal{L}v(X,\theta^*)]\right) - \alpha\varphi(X_k)R_{2,k}^\top\theta^*, \\ \chi_k^{r+1} &= \alpha(H - H_k(X_k))\eta_k^r, \\ \psi_k^{r+1} &= \alpha\left(\Delta t_k^{-1}(\mathbb{E}[H(X_k)] - H) - \varphi(X_k)R_{2,k}^\top\right)\eta_k^r,\end{aligned} \tag{24}$$

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 $R^2=3\widetilde{C}\left(\|\mathcal{L}\varphi+\mathbb{E}[R_0(\cdot,\xi_0)]\|_\infty+\Delta t_0^{\frac{1}{2}}\|R_1(\cdot,\xi)]\|_\infty+2\varepsilon^{-1}\sup_{k\geq 0}\|R_2(\Delta t_k,\cdot,\xi)]\|_\infty+2\varepsilon^{-1}\right)$, $0<\varepsilon<\Delta_0^{-2}$ is a constant that will be defined later, $\widetilde{C}$ is the constant from Lemma F.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 (23) and (30),

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

Then, concerning χ_{k+1}^r, using Lemma F.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 F.4 once again for ψ_{k+1}^r, we get,

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

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

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

which implies that

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

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

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

where we used (33) to obtain the last line. Then let us pass to $I^r_{k,1}$,

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

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

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

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

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

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 $\sum_{i=0}^{k-1}\Delta t_i^2 \leq a\ln(1+k)$, we obtain that $C_k \leq (1+k)^{a\varepsilon}$ 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 $a^{-1}\varepsilon$. □

E.3. Proof of Theorem D.1

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

Theorem E.2. *Let $f : \Theta \to \mathbb{R}$ be μ-convex, L-semi-concave, and such that $\theta^* = \operatorname{argmin}_\theta f(\theta)$ satisfies $|\theta^*| \leq M$ for some $M > 0$. For $\theta_0 \in \Theta$, the sequence $(\theta_k)_{k\geq 0}$ 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 $\sum_{k\geq 0}\alpha_k = \infty$, $|\mathbb{E}\left[g_k|\theta_k\right] - f'(\theta_k)| \leq (1+|\theta_k|)\varepsilon_k$, $\varepsilon_k \in \mathbb{R}_+$ 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

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

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

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

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 (5), 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 $|\mathbb{E}\left[g_k|\theta_k\right] - f'(\theta_k)| \leq C\Delta t_k(1+|\theta_k|)$. The fact that $\mathbb{E}\left[|g_k|^2\middle|\theta_k\right] \leq C(1+|\theta_k|^2)$ is straightforward. Theorem E.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. □

F. Technical lemmas

F.1. Expansions of the temporal differences

Lemma F.1. *For $(x, \xi, \theta) \in \Omega \times \mathbb{R}^d \times \Theta$ and $0 < \Delta t < 1$, there exist $R(x, \xi, \theta)$ such that*

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

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 $p \geq 1$, $\mathbb{E}\left[|R_2(\Delta t, x, \xi)|^p\right]$ is bounded uniformly with respect to Δt and x.

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 (6). □

F.2. Some lemmas used in the proof of Theorem 5.5

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

$$S^2 \leq \mathrm{tr}(S) S \tag{25}$$

$$A^\top A = -A^2 \leq \frac{2}{\rho\sigma^2}\left\| b + \frac{\sigma^2}{2}\nabla_x \ln(m)\right\|_\infty^2 S^2 \tag{26}$$

$$(SA - AS) \leq 2\sqrt{\frac{2}{\rho\sigma^2}}\left\| b + \frac{\sigma^2}{2}\nabla_x \ln(m)\right\|_\infty S^2, \tag{27}$$

$$\mathbb{E}\left[H(X)H(X)^\top\right] \leq \rho^{-1}\|\mathcal{L}\varphi(X)\|_\infty^2 S. \tag{28}$$

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

$$-D^2_{x,x} \cdot \left(\frac{\sigma\sigma^\top}{2} m \right) + \operatorname{div}(bm) = 0, \tag{29}$$

and the following integration by parts,

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

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 (25) is straightforward, it only relies on the fact that S is symmetric and positive. The fourth inequality (28) is straightforward using the definitions of $H(X)$ and S. The third inequality (27) is a consequence of (26). Therefore, there is only (26) left to prove. Let us take $\lambda \in \mathbb{C}$ a complex eigenvalue of H, and θ an associated normalized 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 F.3. *For $\alpha \leq R^{-2}$, the following two inequalities hold for any $k \geq 0$,*

$$(I_d - \alpha H^\top)(I_d - \alpha H) \leq I_d - \alpha S \tag{30}$$

$$\left(I_d - (I - \alpha H^\top)^k\right) \left(I_d - (I - \alpha H)^k\right) \leq \alpha^2 k^2 H^\top H, \tag{31}$$

$$\left(I_d - (I - \alpha H^\top)^k\right) \left(I_d - (I - \alpha H)^k\right) \leq 4 \left(1 + \frac{2}{\rho\sigma^2} \|b + \nabla_x \ln(m)\|^2_\infty\right) I_d, \tag{32}$$

$$\left(I_d - (I - \alpha H^\top)^k\right) \left(H^{-1} + H^{-\top}\right) \left(I_d - (I - \alpha H)^k\right) \leq 2\alpha k \left(1 + \sqrt{\frac{2}{\rho\sigma^2}} \|b + \nabla_x \ln(m)\|_\infty\right) I_d. \tag{33}$$

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 (33) is a straightforward consequence of (31) and (32). Then, concerning (30), 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, (25), (26) and (27). Therefore, it only remains to prove (31) and (32).

First step: proving (31). 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 (30), 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 (30) 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 (32). 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

$$\begin{aligned} \left|1 - (1-\alpha\lambda)^k\right| &\leq \alpha|\lambda| \sum_{j=0}^{k-1} |1-\alpha\lambda|^j \\ &\leq \frac{\alpha|\lambda|}{1 - |1-\alpha\lambda|} \\ &\leq \frac{\alpha|\lambda|}{1 - (1-\alpha\Re(\lambda))^{\frac{1}{2}}} \text{ using (30),} \\ &\leq \frac{\alpha|\lambda|}{1 - (1 - \frac{\alpha}{2}\Re(\lambda))} \text{ because } \alpha\Re(\lambda) \leq 1, \\ &\leq 2\sqrt{1 + \frac{\Im(\lambda)^2}{\Re(\lambda)^2}} \\ &\leq 2\left(1 + \frac{2}{\rho\sigma^2}\|b + \nabla_x \ln(m)\|_\infty^2\right)^{\frac{1}{2}}, \end{aligned}$$

where the last inequality comes from a similar argument as in the proof of (26). This concludes the proof.

□

Lemma F.4. *Assume* **A3**. *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 (5). 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 (5), 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. □

F.3. Calculus of variances and covariances

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

$$\mathrm{Var}\left(\xi \cdot \nabla_x v(x)\right) = \left|\nabla_x v(x)\right|^2, \tag{34}$$

$$\mathrm{Var}\left(\xi^\top D^2 v(x)\xi - \Delta_x v(x)\right) = \mathrm{tr}\left(D_x^2 v(x)^2\right). \tag{35}$$

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

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

This concludes the proof. □

F.4. Counterpart to Lemma F.1 in the multi-step setting

Lemma F.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

$$\widetilde{\delta}^n_{\Delta t}(x,\xi,\theta) = \frac{1}{n}\sum_{i=0}^{n-1}\widetilde{\delta}_{\Delta t}(X_{t_i}, X_{t_{i+1}}, \theta). \tag{36}$$

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 F.1 imply

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

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 $(\xi_i)_{0\leq i<n}$. □

G. Additional results

G.1. Convergence of the Discrete Markov Chain

In this section, we give a condition under which the convergence results (5) holds. Here, we consider the case of observations obtained from a simulator, like stated in Section 3. 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 G.1 (Theorem 14.5.1 from Kloeden & Platen (1992))**.** *For $f \in C^4(\Omega;\mathbb{R})$, there exists $C > 0$ depending only on the C^4-norm of f such that $|\mathbb{E}\left[f(X_k) - f(X)\right]| \leq C\Delta t_k$.*

G.2. Extensions of the proofs in the case of real-world observations

In this section, we prove the counterpart of Lemma 4.1 to the case when the samples are directly observed from the continuous dynamics in (1), this corresponds to the assumption of real-world observations in Section 3. All the other proofs may then be easily repeated and we obtain similar results. Basically, the proof consists in replacing the Taylor expansions used in the other proofs by It Calculus. We believe that all results in the present paper may be adapted to this case, using similar arguments.

Lemma G.2. *Assume that r and b are uniformly continuous, and that v admits bounded continuous derivatives in x everywhere up to order three. For $\Delta t > 0$, $x \in \Omega$, W a Brownian motion and $\theta \in \Theta$, we define $\widetilde{\delta}^{\text{cont}}_{\Delta t}$ the continuous temporal difference by,*

$$\widetilde{\delta}^{\text{cont}}_{\Delta t} = (\Delta t)^{-1}\big[v(x,\theta_k) - \gamma_{\Delta t} v(X_{\Delta t},\theta) - r(x)\Delta t + \sigma W_{\Delta t}\cdot\nabla_x v(x,\theta)\big],$$
$$\text{where } dX_t = b(X_t) + \sigma dW_t, \text{ and } X_0 = x.$$

The mean and variance of $\widetilde{\delta}^{\text{cont}}_{\Delta t}$ satisfy,

$$\lim_{\Delta t\to 0}\mathbb{E}[\widetilde{\delta}^{\text{cont}}_{\Delta t}] = \mathcal{L}v(x,\theta), \text{ and } \lim_{\Delta t\to 0}\text{Var}(\widetilde{\delta}^{\text{cont}}_{\Delta t}) = \frac{\sigma^4}{2}\text{tr}\left(D_x^2 v(x,\theta)^2\right).$$

Proof. In this proof, the dependence of v in θ is omitted. From It 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,

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

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.,

$$\lim_{\Delta t \to 0} \text{Var}(\widetilde{\delta}_{\Delta t}^{\text{cont}}) = \lim_{\Delta t \to 0} \frac{\sigma^2}{\Delta t^2}\mathbb{E}\left[\left|\int_0^{\Delta t} (\nabla_x v(X_t) - \nabla_x v(x)) \cdot dW_t\right|^2\right]. \tag{37}$$

Using It 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 (37). This time, we take $g = D_x^2 v \, b + \nabla_x \Delta_x v$, let us consider the following sequence of inequalities

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

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,

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

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

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

This concludes the proof. □