# On a Variance-Reduction Correction for the Temporal-Difference Learning in the Stochastic Continuous Setting

Ziad Kobeissi, Francis Bach

# On a Variance-Reduction Correction for Temporal-Difference Learning in the Stochastic Continuous Setting

**Ziad Kobeissi**
Institut Louis Bachelier
Inria Paris
ziad.kobeissi@inria.fr

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

## Abstract

We consider the problem of policy evaluation for continuous-time processes using the temporal-difference learning algorithm. More precisely, from the time discretization of a stochastic differential equation, we intend to learn the continuous value function using TD(0). First, we show that the standard TD(0) algorithm is doomed to fail when the time step tends to zero because of the stochastic part of the dynamics. Then, we propose an additive zero-mean correction to the temporal difference making it robust with respect to vanishing time steps. We propose two algorithms: the first one being model-based since it requires to know the drift function of the dynamics; the second one being model-free. We prove the convergence of the model-based algorithm to the continuous-time solution under a linear-parametrization assumption in two different regimes: one with a convex regularization of the problem; and the second using the Polyak-Juditsy averaging method with constant step size and without regularization. The convergence rate obtained in the latter regime is comparable with the state of the art for the simpler problem of linear regression using stochastic gradient descent methods. From a totally different perspective, our method may be applied to solve second-order elliptic equations in non-divergent form using machine learning.

## 1 Introduction

Policy evaluation is one of the main building blocks of modern reinforcement learning (RL). One of the most basic and regarded algorithms for policy evaluation is known as temporal-difference learning (TD), [35]. While TD was originally proposed in the tabular case, its large-scale applicability has been greatly improved by its combination with parametric function approximation [10]. In this case, its theoretical analysis is challenging.

Moreover, TD (and more generally most of RL algorithms) is naturally designed for only handling discrete time processes. In practice, numerous applications in the literature are obtained as time discretizations of continuous processes. In this case, an overwhelming majority of research papers proceed as follows: first, they consider a time discretization of the continuous problem; second, they use RL algorithms on the discretized problem as if it was a simple discrete problem. This approach has been proven to converge to the right solution when the time step tends to zero, only in the deterministic setting, see [12, 14, 28]. However, we argue in the present work that this result does not hold in the stochastic setting, i.e., learning becomes impossible without additional care as the time step vanishes. According to our knowledge, the discretization error has never been regarded in the stochastic set-up before. Yet, it seems absolutely crucial to consider randomness for policy evaluation, since exploration is impossible in the deterministic setting. The present work aims at

Preprint. Under review.

filling this gap by proposing a robust extension of the TD(0) algorithm adapted to vanishing time steps, and proving convergence results under a linear parametrization assumption.

Even if most of the arguments in this work concern the limit when the time step tends to zero, we argue in Section 4.4 that our method may be used for a small fixed non-vanishing time step, making the constants in the convergent rates independent of the time step. This could be particularly efficient for improving accuracy and computation speed of reinforcement learning algorithms on stochastic high-frequency models, even when they do not derive from the discretization of a stochastic differential equation.

From a different perspective, the limit process when the time discretization tends to zero can be characterized using the theory of partial differential equations (PDE). More precisely, it consists in solving a discounted second-order elliptic equation in non-divergent form, see [15]. Therefore, the present analysis may be seen as a new method for numerically solving such class of PDE from observations with potentially unknown drift and diffusion functions.

### 1.1 Related literature

**Temporal-difference learning.** The TD algorithm was introduced in the tabular case by [35], with later convergence results for linearly dependent features [11]. Asymptotic stochastic approximation results were derived by [19] for the tabular case, and by [33] when using linear approximations, with a non-asymptotic analysis in the *i.i.d.* sampling case [27].

**Stochastic iterative methods.** The analysis of TD requires tools from stochastic approximation [7], and many of the required tools have been derived for stochastic gradient descent (SGD) [8] and reused here. The convergence results presented in the present paper may be compared to standard results on RL algorithms, see [6, 21] for TD(0). The techniques in the proof (especially concerning the fast-convergence results in Section 4.3) are adapted from the literature on SGD methods [1, 31] to the non symmetric setting. In particular, [1] consists in 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 [2], which proposed a continuous-time counterpart to $Q$-learning; it was later extended by [37]. From a different perspective, [9] 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, [14] derived algorithms for both policy evaluation and policy improvement. Similar deterministic approaches of continuous-time RL have recently been explored by [28, 40]. In order to balance between exploration and exploitation, [39] added an entropy-regularization term to a similar 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. Some mesh-dependent learning algorithms have been developed, see [25, 26, 29], 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 [20], or [34] for the *Deep Galerkin Method*, or [5, 17, 18] 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 [30, 4] 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.

### 1.2 Contributions

We start by giving theoretical evidences that learning methods based on the temporal difference, are not adapted to high-frequency optimization. More precisely, those methods are doomed to fail when

applied to the discretization of a continuous stochastic problem, when we let the discretization step tends to zero. This claim is made clear by the first equality in Lemma 3.3 in Section 3.2, where it is shown that the variance of the standard rescaled temporal difference tend to infinity, making learning impossible.

Then, we propose a correction to the temporal difference, based on the Taylor expansion at a neighborhood of the continuous problem when the time step tends to zero. Namely, the variance of the corrected rescaled temporal difference stays bounded at the limit, see the second equality in Lemma 3.3.

The rest of the paper focuses on TD(0). However, we would like to insist on the fact that the conclusions of the latter two paragraphs hold not only for TD(0), but for any RL algorithm implying the temporal difference, such as SARSA or Q-learning [36]. Moreover, this holds for nonlinear parametrizations as well as linear ones.

From the corrected TD(0) method, we propose two algorithms. The first one is model-based since it requires to know the drift function. Such an assumption may seem affordable in some models coming for instance from physics, robotics, or finance. Note that we never assume that the diffusion matrix is known and that our method may be applied for arbitrary distributions of noise. The second algorithm is model-free as it uses regression to learn the drift function (and the noise) in the same time as the value function.

Finally, under a linear parametrization assumption, we prove the convergence of the model-based algorithms. First, using usual decreasing learning steps and regularization methods, we obtain standard convergence rates for the regularized problem, see Theorem 4.1; we also prove non-asymptotic bounds on the approximation error to the solution of the unregularized problem, see Corollary 4.3. Second, coming back to the original unregularized problem, we prove convergence results with constant learning step and a Polyak-Juditsky averaging method, see Theorem 4.4 in Section 4.3. Our rate of convergence, of order $1/k$, is analogous to the optimal rate of convergence for the simpler problem of regression with SGD methods and without strong-convexity assumptions.

## 2 Context

### 2.1 From continuous problem to discretization

In the continuous-time stochastic setting, the state, denoted by $(X_t)_{t\in[0,\infty)}$, satisfies the following stochastic differential equation (SDE),

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \tag{1}$$

where $W$ is a $d$-dimensional Brownian motion and $\sigma$ is a matrix-valued function. The value function is defined by (with $\rho > 0$ the continuous discount factor):

$$V(x) = \mathbb{E}\Big[\int_0^\infty e^{-\rho t} r(X_t)dt \Big| X_0 = x\Big]. \tag{2}$$

The simplest idea to approximate the solution of such an infinite dimensional optimization problem is to discretize the continuous dynamics and rewards with respect to time with a sufficiently small time step $\Delta t > 0$. The simplest discretization scheme is the Euler-Maruyama scheme [22], that we adopt here for simplicity (however the arguments of this work extend straightforwardly to higher-order discretization schemes),

$$X_{t_{i+1}} = S_{\Delta t}(X_{t_i}, \xi_i) := X_{t_i} + \Delta t\, b(X_{t_i}) + \sqrt{\Delta t}\sigma(X_{t_i})\xi_i, \tag{3}$$

where $S_{\Delta t}$ is the step operator, $t_i = i\Delta t$ for $i \in \mathbb{N}$, and $(\xi_i)_{i\geq 0}$ are independent distributed random vectors with zero mean and identity covariance matrix. The discrete value function is then defined by

$$V_{\Delta t}(x) = \mathbb{E}\Big[\sum_{i=0}^{\infty} e^{-i\rho\Delta t} r(X_{t_i}) \Big| X_0 = x\Big]. \tag{4}$$

Here, we insist on the fact that we do not assume that the $\xi$ are normally distributed. Indeed, Donsker's Theorem [13] implies a *Gaussianization* phenomenon of the noise when $\Delta t$ tends to zero. Therefore, the methods proposed here hold for any distribution of noise with finite second-order moments. In particular they may be implemented when the distribution of the noise in unknown.

### 2.2 Assumption on the observations and discussions

Let $(\Delta t_k)_{k\geq 0}$ be a sequence of vanishing time steps. We access the observations $(X_k, X'_k)_{k\geq 0}$, a sequence of independent coupled random variables such that $X_k$ is distributed according to $m_k$ the stationary distribution of the discrete dynamics (3), and $X'_k = S_{\Delta t_k}(X_k, \xi_k)$ where $(\xi_k)_{k\geq}$ are i.i.d. latent random variables with zero means and identity covariance matrices.

Let $m$ be the stationary distribution of the continuous dynamics (1), we introduce $X$ a random variable distributed according to $m$ and independent of the observations. The sequence $(m_k)_{k\geq 0}$ is weakly convergent to $m$ and a convergence rate is given in the following theorem.

**Theorem 2.1** (Theorem 14.5.1 from [22]). *For $f \in C^4(\Omega;\mathbb{R})$, there exists $C > 0$ depending only on the $C^4$-norm of $f$ such that $|\mathbb{E}\left[f(X_k) - f(X)\right]| \leq C\Delta t_k$.*

In the rest of this section, we discuss the above assumptions on the distributions of the observations, and explain under which alternative settings we believe our analysis holds.

First, according to us, the most restrictive part of the latter assumptions, which is still largely regarded in the literature, is that the observations are independent. The most regarded alternative is to assume that the observations are sampled from the dynamics, i.e., $X'_k = X_{k+1}$. Such an assumption is out of the scope of the present work. We still believe that the results here hold in this case under a strong mixing assumption on the Markov Chains induced by the discrete dynamics. However, the convergence rates and the details in the proofs would be affected. Moreover, the obtained results are in general not suited as well for practice, since the proofs depend on the Markov Chains to be computed until very large times (corresponding to the mixing time multiplied by a big constant).

Concerning the assumption on $X_k$ being distributed according to $m_k$, it may be relaxed straightforwardly in multiple cases. The simplest case is when the sequence of laws satisfy a similar convergence property as in Theorem 2.1 (the exponent on $\Delta t$ might be lower than one).

Another interesting setting is when $X_k$ is distributed according to $m$, and $X'_k = X_{\Delta t_k}$ with $(X_t)_{0\leq t\leq \Delta t_k}$ satisfying the continuous dynamics (1) and $X_0 = X$. This corresponds to having observations which come directly from the continuous dynamics. In this case, the results hold and the proofs may be repeated with Taylor expansions replaced by Itô calculus. To illustrate this claim and give more insights on how to proceed, in Appendix C.1, we adapted the proof of one of the main results of this paper to this case, namely Lemma 3.3.

Considering offline observations is out of the scope of the present work but it will be considered in the case of Q-learning for a future work consisting in a similar analysis of the continuous time limit in stochastic setting.

### 2.3 Temporal difference with function approximation

Let us consider the problem of approximating $V$ defined in (2) using a parametrized function $v(\cdot,\theta)$, where $\theta \in \Theta$ and $\Theta$ is the parameter set. For $k \geq 0$, we denote $\theta_k$ the learned parameter at iteration $k$. We define the rescaled temporal difference as,

$$\widetilde{\delta}_k = \widetilde{\delta}_{\Delta t_k}(X_k, X'_k, \theta) := (\Delta t_k)^{-1}\big(v(X_k,\theta_k) - \gamma v(X'_k,\theta_k) - r(X_k)\Delta t\big), \tag{5}$$

where the normalization constant $(\Delta t_k)^{-1}$ is chosen in the latter definition because $\mathbb{E}[\widetilde{\delta}_{\Delta t_k}]$ admits a nontrivial limit when $\Delta t_k$ tends to 0. Choosing a different order of magnitude in the normalization constant would be pointless as it leads to a convergence in average to zero or infinity. In the following, the term *rescaled* will be omitted since no non-rescaled temporal difference will be considered.

At least heuristically, we get the following Taylor expansion of $\widetilde{\delta}_k$ when $\Delta t \to 0$,

$$\begin{aligned}\widetilde{\delta}_k &= (\Delta t_k)^{-1}\big[v - (1-\rho\Delta t_k)\big(v + \nabla_x v^\top(\Delta t_k\, b + \sqrt{\Delta t_k}\sigma\xi) + \Delta t_k \xi^\top\sigma^\top(D_x^2 v)\sigma\xi\big) - \Delta t_k\, r\big] + o(1)\\ &= \rho v - \nabla_x v^\top b - (\Delta t_k)^{-\frac{1}{2}}\nabla_x v^\top\sigma\xi - \xi^\top\sigma^\top(D_x^2 v)\sigma\xi - r + o(1),\end{aligned}$$

where we omitted the arguments in the functions $\widetilde{\delta}_{\Delta t}$, $v$, $b$, $\sigma$ and $r$, to simplify the notations. One may notice that the third term in the right-hand side of the latter equality has zero mean when $\xi$ is a centered random variable, but its variance is of order $(\Delta t_k)^{-1}$. We thus obtain that any RL

algorithm based on such a temporal difference is doomed to fail when $\Delta t_k \to 0$ because of the latter variance tending to infinity.

However, removing this term with a divergent variance is possible by simply adding a term in the temporal difference. We define the *corrected* temporal difference as,

$$\begin{aligned} &\delta_k = \delta_{\Delta t_k}(X_k, X'_k, \theta_k) := (\Delta t_k)^{-1}\big(v(X_k,\theta_k) - \gamma v(X'_k,\theta_k) - r(X_k)\Delta t_k + Z_k\big), \\ &\text{where } Z_k = (X'_k - X_k - b(X_k)\Delta t_k)\cdot \nabla_x v(X_k,\theta_k). \end{aligned} \tag{6}$$

One may notice that $Z_k$ satisfy $Z_k = \sqrt{\Delta t_k}\nabla_x v(X_k,\theta_k)^\top \sigma(X_k)\xi_k$, but we prefer the formulation from (6) to insist on the fact that $Z_k$ can be computed without observing the noise $\sigma(X_k)\xi_k$.

We refer to the additional term $Z_k$ as the variance-reduction term. This terminology is due to the facts that the conditional expectation with respect to $X_k$ of $Z_k$ is equal to zero, and that the variance of $\widetilde{\delta}_k$ is of order $(\Delta t_k)^{-1}$ while the variance of $\delta_k$ is uniformly bounded. The use of the letter $Z$ for the notation comes from its continuous counterpart appearing in the backward stochastic differential equation (BSDE) point of view of the optimal control problem (see [30] for instance).

### 2.4 Model-based and model-free alternatives

We define the corrected TD(0) iterations by,

$$\theta_{k+1} = \theta_k - \gamma_k \delta_k \nabla_\theta v(X_k,\theta_k), \tag{TD0}$$

This algorithm is model-based since the drift function $b$ has to be known to compute $Z_k$ and thus $\delta_k$.

However, we may figure out a model-free alternative to the latter algorithm. For instance, we may learn $x \mapsto \widetilde{b}(x,\theta^b)$ and $x, x' \mapsto \zeta(x,x',\theta^\zeta)$ as approximations of $b$ and $\sigma\xi$ respectively, for two parameters $\theta^b, \theta^\zeta \in \Theta$, in the same time as we are learning the value function. one may compute $(\theta^b_k, \theta^\zeta_k)_{k\geq 0}$ as,

$$\begin{aligned} \theta^b_{k+1} &= \theta^b_k - \gamma_k(\Delta t_k)^{-1}\nabla_\theta \widetilde{b}(X_k,\theta^b_k)^\top\big(X'_k - X_k - \sqrt{\Delta t_k}\,\zeta(X_k,X'_k,\theta^\zeta_k)\big), \\ \theta^\zeta_{k+1} &= \theta^\zeta_k - \gamma_k(\Delta t_k)^{-\frac{1}{2}}\nabla_\theta \zeta(X_k,X'_k,\theta^\zeta_k)^\top\big(X'_k - X_k - \Delta t_k\,\widetilde{b}(X_k,\theta^b_k)\big). \end{aligned}$$

This consists in a stochastic gradient descent method on the regression problem of minimizing $\mathbb{E}\big[\big|\Delta t\,\big(\widetilde{b}(X,X',\theta^b) - b(X)\big) + \sqrt{\Delta t}\big(\zeta(X,X',\theta^\zeta) - \sigma(X)\xi\big)\big|^2\big]$ over $(\theta^b,\theta^\zeta)$, with different learning steps for $\widetilde{b}$ and $\zeta$. The drawback with the latter iterations is that $(\widetilde{b},\zeta)$ will eventually approximate $(0, \sigma\xi + \sqrt{\Delta t_k}b)$ instead of $(b,\sigma\xi)$ (especially when $\Delta t_k$ is slowly converging to 0). Therefore, we should add a penalization in the SGD step made on $\theta^\zeta$, in order to make the artificial noises be independent for different observations, and their mean equal to zero. This additional cost might be of the form $\frac{1}{N}\mathbb{E}\big[\sum_{1\leq i<j\leq N}\zeta(X_i,X'_i,\theta^\zeta)\zeta(X_j,X'_j,\theta^\zeta)\big]$, where $N \geq 2$ is the batch size.

## 3 Asymptotic behavior of the temporal difference

From now on, for simplicity of the notations, we assume that $\sigma$ is a constant positive real number. It is straightforward to extend all the results proved here to the case where we take $\widetilde{\sigma}$ a matrix-valued function such that $\widetilde{\sigma}^\top\widetilde{\sigma} \geq \sigma^2 I_d$.

### 3.1 Short review of the continuous problem and boundary conditions

The value function $V$, defined in (2), satisfies the following partial differential equation [6]

$$\mathcal{L}V(x) = \rho V(x) - \tfrac{\sigma^2}{2}V(x) - \nabla_x V(x)\cdot b(x) - r(x) = 0, \tag{7}$$

where $a = \sigma\sigma^\top/2$. Up to now we intentionally omitted to mention the boundary conditions on the state space. In the following, we will make the simplifying assumption that the state space is the $d$-dimensional torus, defined as the hypercube $[0,1)^d$ with periodic boundary conditions, i.e., $\Omega = \mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$. Indeed, the choice of the boundary conditions is in general a difficult problem for continuous state settings, especially when considering PDE-based model, like here with (7). More precisely, here, we need them to allow the existence of stationary measures for the continuous and

discrete dynamics, and we need a priori estimates on the continuous stationary measure $m$ (more precisely, we need a uniform bound on $\nabla_x \ln m$, see Lemma B.2 in the Appendix). In an attempt to separate difficulties, and to stay focused on the main ideas, we prefer to only consider the torus, where these requirements are easily met, rather than overwhelming the readers with unnecessary technical difficulties due to boundary conditions. See Remark 3.2 below, for some insights on how to extend our results to other domains.

Classical results on second-order elliptic equations imply that if we find a candidate function $v$ such that $\mathcal{L}v$ is small, then $v$ and $V$ are close to each other. For instance, see the following lemma.

**Lemma 3.1.** *Assume that $r$ and $b$ are uniformly bounded on the graph of $u$. There exists a constant $C > 0$ such that if $V \in H^2(\Omega; \mathbb{R})$, then we have*

$$\|V - V^u\|_{H^2} \leq C \|\mathcal{L}V\|_\infty.$$

The proof consists in: first, applying the maximum principle (see [15], Section $6.4$, for instance) to get $\|V - V^u\|_\infty \leq \rho^{-1} \|\mathcal{L}V\|_\infty$; second, using Theorem $8.12$ from [16] which yields the desired inequality. More elliptic results, either on weak or strong derivatives, are presented by [24, 23].

Regarding Lemma 3.1, an alternative, that will not be investigated here, is to try to minimize the residual of $\mathcal{L}(\cdot, \theta)$ through learning [25, 32, 34]. Here, we prefer to use the temporal difference instead, in order to make links with usual reinforcement learning algorithms, and to avoid computing second-order derivatives which may be large for high-frequency parametrizations.

Let us mention for later use, that the stationary law $m \in \mathcal{P}(\mathbb{R}^d)$ of the continuous dynamics (1) satisfies the following PDE,

$$-\tfrac{\sigma^2}{2} \Delta_x m - \operatorname{div}(b(x, u(x)) m(x)) = 0. \tag{8}$$

This may be seen as the dual equation to the homogeneous counterpart of (7).

**Remark 3.2.** *Let us give some hints to extend the present results to other state spaces or boundary conditions. The case $\Omega = \mathbb{R}^d$ is more involved than the torus, but one may get some insights on the simple case consisting in taking a drift function deriving from a potential, i.e., $b(x) = \nabla_x U(x)$ for some $U$. In this case, there exists a stationary probability measure of the continuous dynamics if $Z = \int_\Omega e^{-2U(x)/\sigma^2} dx$ is finite, given by $m = e^{-2U/\sigma^2}/Z$. Moreover, here, $\nabla_x \ln(m)$ is bounded if and only if $\nabla_x U$ is bounded. This simple example emphasizes the fact that, for $\mathbb{R}^d$, some restrictions should be satisfied for the results to hold, for instance that the drift function should be bounded and pointing out in the direction of a compact subset of $\mathbb{R}^d$, with sufficient magnitude.*

*Alternatively, one may be interested in considering $\Omega$ as a smooth bounded subset of $\mathbb{R}^d$ with, for instance, Dirichlet or Neumann conditions [15]. In this case, other restrictions appear, but we also believe that our results may be adapted up to making additional assumptions.*

### 3.2 Bellman error and asymptotic analysis of the residual gradient algorithm

In this section and the next one, we assume that $\theta$ is a fixed parameter, $\Delta t$ is the step size, $(X, X')$ are random variable such that $X' = S_{\Delta t}(X, \xi)$ for some latent random variable $\xi$ with zero mean and identity covariance matrix.

Instead of directly making the asymptotic analysis of TD(0), we start by considering another method, namely the residual gradient algorithm (RG, [3]). Indeed, similar non-robustness phenomena happen for both algorithms when the time steps vanish, but RG allows a more quantitative analysis since it is directly linked with the Bellman error. More precisely, RG is a SGD method to reduce the mean-square error of the Bellman error, which here, can be decomposed in two terms as follows,

$$\mathbb{E}_{(X,X')}\big[\big|\widetilde{\delta}_{\Delta t}(X, X', \theta)\big|^2\big] = \underbrace{\mathbb{E}_X\big[\mathbb{E}_{X'}\big[\widetilde{\delta}_{\Delta t}(X, X', \theta)\big|X\big]^2\big]}_{\text{Bellman error}} + \underbrace{\mathbb{E}_X\big[\operatorname{Var}_{X'}\big(\widetilde{\delta}_{\Delta t}(X, X', \theta)\Big|X\big)\big]}_{\text{perturbating term}}. \tag{9}$$

A similar formula holds when $\widetilde{\delta}$ is replaced by $\delta$. The asymptotic behavior of the terms inside the expectation $\mathbb{E}_X$, at fixed $x$, is characterized in the following lemma.

**Lemma 3.3.** *Assume that $r$ and $b$ are uniformly bounded, and that $v$ admits bounded continuous derivatives in $x$ everywhere up to order two. For $X = \delta_x$, the means and variances of $\widetilde{\delta}$ and $\delta$*

*satisfy,*

$$\lim_{\Delta t\to 0}\mathbb{E}_{X'}[\widetilde{\delta}_{\Delta t}(x,X',\theta)]=\mathcal{L}v(x,\theta),\quad \textit{and}\quad \lim_{\Delta t\to 0}\Delta t\,\mathrm{Var}_{X'}(\widetilde{\delta}_{\Delta t}(x,X',\theta))=\sigma^2\left|\nabla_x v(x,\theta)\right|^2,$$

$$\lim_{\Delta t\to 0}\mathbb{E}_{X'}[\delta_{\Delta t}(x,X',\theta)]=\mathcal{L}v(x,\theta),\quad \textit{and}\quad \lim_{\Delta t\to 0}\mathrm{Var}_{X'}\left(\delta_{\Delta t}(x,X',\theta)\right)=\frac{\sigma^4}{2}\mathrm{tr}\left(D_x^2 v(x,\theta)^2\right).$$

The proof is straightforward, using a similar expansion as the one proved in Lemma B.1, below in the Appendix, but only up to order two instead of order four, and Lemma B.5, also in the Appendix.

Passing to the limit $\Delta t\to 0$ in (9), we obtain

$$\lim_{\Delta t\to 0}\mathbb{E}_{(X,X')}\left[\left|\widetilde{\delta}_{\Delta t}(X,X',\theta)\right|^2\right]=\begin{cases}+\infty \text{ if } v(\cdot,\theta) \text{ is not constant,}\\ \mathbb{E}\left[\left(\rho C+r(X)\right)^2\right] \text{ if } v(\cdot,\theta)=C.\end{cases}\tag{10}$$

Consequently, for the residual gradient method, when $\Delta t$ tends to zero, the perturbating term in the decomposition (9) totally overwhelms the Bellman error and the only approximations that can be learnt are constant functions.

The same arguments used on $\delta$ implies,

$$\lim_{\Delta t\to 0}\mathbb{E}_{(X,X')}\left[\,|\delta_{\Delta t}(X,X',\theta)|^2\,\right]=\mathbb{E}_X\left[\mathcal{L}v(X,\theta)^2\right]+\frac{\sigma^4}{2}\mathbb{E}_X\left[\mathrm{tr}\left(D_x^2 v(X,\theta)^2\right)\right].\tag{11}$$

In this case, learning is possible even if we see an additional term appearing on the right-hand side. In the following, we focus on TD(0), however, one may derive similar results for RG that may be found in Section C.2 in the Appendix.

### 3.3 Asymptotic analysis of TD(0)

Let us recall that the corrected TD(0) algorithm is given in (TD0). Similar decomposition and analysis of the asymptotic behavior as in (9) and Lemma 3.3 respectively, yield

$$\mathrm{Var}\left(\widetilde{\delta}_{\Delta t}\nabla_\theta v\right)\sim\mathrm{Var}\left(\mathcal{L}v\nabla_\theta v\right)+\frac{\sigma^2}{2\Delta t}\mathbb{E}\left[\left|\nabla_x v\right|^2\left|\nabla_\theta v\right|^2\right],$$

where the arguments of $v$ and $\mathcal{L}v$ are $(X,\theta)$ and the ones of $\widetilde{\delta}$ are $(X,\xi,\theta)$. This implies the rest of our first contribution, for the case of TD(0). Indeed, learning is impossible at the limit $\Delta t=0$ because the variance of the update becomes infinite (except if $|\nabla_x v||\nabla_\theta v|$ is uniformly equal to zero which does not seem likely to happen for a large class of functions $v$).

Repeating the same arguments for $\delta$ instead of $\widetilde{\delta}$, we obtain

$$\lim_{\Delta t\to 0}\mathrm{Var}\left(\delta_{\Delta t}\nabla_\theta v\right)=\mathrm{Var}\left(\mathcal{L}v\nabla_\theta v\right)+\frac{\sigma^4}{2}\mathbb{E}\left[\mathrm{tr}\left((D_x^2 v)^2\right)\left|\nabla_\theta v\right|^2\right].$$

Therefore, the latter variance stays uniformly bounded when $\Delta t$ tends to zero and learning remains possible even at the limit.

## 4 Linear setting

### 4.1 Assumptions

We say that a function is $C^\ell$ for $\ell\geq 1$ if it admits continuous and bounded derivatives up to order $\ell$. Let us make the following assumptions:

- **H1** The function $v$ is linear with respect to $\theta$, more precisely the set of parametrized functions is $\mathcal{V}_\theta=\left\{x\mapsto v(x,\theta)=\theta^\top\varphi(x),\theta\in\Theta\right\}$ for $\varphi:\Omega\to\mathbb{R}^{d_\theta}$, and $\Theta=\mathbb{R}^{d_\theta}$ for some $d_\theta\geq 1$.
- **H2** The functions $r$ and $b$ are $C^4$.
- **H3** The feature vector $\varphi$ is $C^6$.

The first assumption, **H1**, is common in the theoretical literature [1, 31, 38], since very little is known about theoretical guarantees in the non-linear case. It is also common to assume some regularity the functions of the model, i.e., $b$ and $r$, and the feature vector $\varphi$. The precise regularity needed here comes from applying Theorem 2.1 to $b$, $r$ and its derivatives of $v$ up to order two.

### 4.2 Non-asymptotic analysis using regularization and decreasing learning steps

This section may be thought of as a warm-up for the next one. First, for a regularization constant $\mu > 0$, let us identify the potential limit $\theta^*$ of TD(0), it should satisfy:

$$\mathbb{E}_{X\sim m}\left[\varphi(X)\mathcal{L}_{\rm lin}\varphi(X)\right]\theta^* = \mathbb{E}_{x\sim m}\left[r(X)\varphi(X)\right], \tag{12}$$

where $\mathcal{L}_{\rm lin}$ is the linear part of $\mathcal{L}$, satisfying $\mathcal{L} = \mathcal{L}_{\rm lin} - r$. From Lemma B.2 in the appendix, the symmetric part of $\mathbb{E}\left[\varphi(X)\mathcal{L}_{\rm lin}\varphi(X)\right]$ is positive, and it has full rank if either $\mathbb{E}\left[\varphi(X)\varphi(X)^\top\right]$ or $\mathbb{E}\left[\nabla_x\varphi(X)\nabla_x\varphi(X)^\top\right]$ has full rank. In this case, $\theta^*$ exists, is uniquely defined and satisfies $|\theta^*| \leq M$ for some $M > 0$. We define $\Pi_{B(0,M)}$ as the projection on $B(0, M)$ the Euclidean ball of $\mathbb{R}^d$ centered at 0 with radius $M$. Let us consider the projected regularized TD(0) algorithm as,

$$\widetilde{\theta}_{k+1} = \Pi_{B(0,M)}\big(\widetilde{\theta}_k - \gamma_k(\delta_{\Delta t_k}(X_k, X_k', \widetilde{\theta}_k)\varphi(X_k) + \mu\widetilde{\theta}_k)\big),$$

For $\mu > 0$, the proof of convergence of such an algorithm is simple and may easily be compared with the literature. In particular, we use the common decreasing assumption on the learning step, i.e., that it is proportional to $1/(\mu(k+1))$, and we obtain the usual convergence rate in $1/k$.

**Theorem 4.1.** *Assume* **H1**, **H2**, **H3**, $\mu > 0$, $\gamma_k = \frac{2}{\mu(k+1)}$ *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* $k \geq 1$,

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

*where* $\theta_\mu^*$ *satisfies* $\left(\mu I_d + \mathbb{E}_{X\sim m}\left[\varphi(X)\mathcal{L}_{\rm lin}\varphi(X)\right]\right)\theta_\mu^* = \mathbb{E}_{x\sim m}\left[r(X)\varphi(X)\right]$.

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 of averaging method in the next section.

Moreover, the distance between $\theta^*$ and $\theta_\mu^*$ might be bounded as follows.

**Lemma 4.2.** *Under the same assumption as in Theorem 4.1, there exists* $C > 0$ *such that, for any* $\mu > 0$, $|\theta^* - \theta_\mu^*| \leq C\mu$.

The latter two results directly implies the following non-asymptotic error bound.

**Corollary 4.3.** *Under the same assumption as in Theorem 4.1, after* $K \geq 2$ *iterations with* $\mu = K^{-\frac{1}{4}}$, *we obtain*

$$\left|\widetilde{\theta}_K - \theta^*\right|^2 \leq \frac{C}{\sqrt{K}}.$$

### 4.3 Averaging method with constant learning step and no relaxation

In this section, we use the Polyak-Juditsky averaging method, see [31], to accelerate the convergence of the TD(0) algorithm. In the same spirit as the results from [1], we get the convergence of the algorithm with constant learning step and without regularization assumption and without projection map. Moreover, the convergence rate is competitive with the state of the art for the simpler problem of linear regression using SGD methods. Therefore, following (12), the potential limit $\theta^*$ satisfies

$$\Pi\mathcal{L}v(\cdot, \theta^*) = 0, \tag{13}$$

where $\Pi$ is the $L^2(m)$-orthogonal projector onto $\mathcal{V}_\Theta$.

**Theorem 4.4.** *Assume* **H1**, **H2** *and* **H3**. *and that* $\theta^*$ *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* $\gamma < R^{-2}$, $k \geq 1$,

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

*where* $\bar{\theta}_k = \frac{1}{k}\sum_{i=0}^{k-1}\theta_i$, *for* $k \geq 1$, *and* $H = \mathbb{E}\left[\varphi(X)\mathcal{L}_{\rm lin}\varphi(X)^\top\right]$.

*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* $\gamma < R^{-2}$, $k \geq 0$, *the latter inequalities are replaced with the following ones respectively*

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

The proof is adapted from [1] with the extra difficulties that the linear operators applied to $\theta_k$ in (TD0) are different for any $k \geq 0$, and they are not symmetric, even there 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 assumptions on $\sum_i \Delta t_i^2$.

### 4.4 Discussion on the case of non-vanishing time steps

In practice, we never iterate until convergence, and the algorithms are stopped after a certain number of iterations are computed, or after some threshold is reached. Therefore, we never really consider the case $\Delta t \to 0$.

Consequently, an interesting different setting consists in considering non-vanishing time steps. In this section, we consider $(\Delta t_k)_{k\geq 0}$ to be constant and equal to some $\Delta t > 0$. Classical results of convergence of TD(0) apply here to state similar results as the one proved in the present work. However, the constants in the rates of convergence of the latter classical results will depend on $\Delta t$. Typically, those constants will tend to infinity when $\Delta t$ tends to zero.

Here, we can do better by extending the results of the vanishing setting to the non-vanishing setting. Namely, if $\Delta t$ is small enough, i.e., $\Delta t \leq \Delta t_0$ for some $\Delta t_0 > 0$, the results of Sections 4.2 and 4.3 hold with constants independent of $\Delta t$. This extension is justified because all the inequalities and arguments needed to prove the original results, hold for small $\Delta t$ up to changing a little bit the associated constants, using Theorem 2.1.

## 5 Conclusion

In the present work, we proved that standard reinforcement learning method based on the temporal difference are not adapted to solve continuous stochastic optimization, nor their discretizations using small time-steps. We proposed a correction to the temporal difference, in order to overcome the latter problem and obtain robust algorithms with respect to vanishing time steps.

This allows us to introduce two algorithms base on TD(0) using the corrected temporal difference. The first one is model-based since it requires to compute the drift function (but not the diffusion coefficient), and the second is model-free as it learns the drift function on the fly.

When the parametrized function is linear with respect to the parameters, we proved two types of convergence results for the model-based method. In Section 4.2, we introduce a regularized algorithm and prove similar convergence rate as standard results (in particular, that make it easy to compare with the literature). Then we deduce a non-asymptotic upper bound of the error between the regularized iterations and the solution of the unregularized problem. In Section 4.3, we consider the original unregularized problem and use a Polyak-Juditsky averaging method, we recover convergence which corresponds to the state-of-the-art rate for the simpler problem of linear regression with SGD methods without strong-convexity assumption.

From another viewpoint, this work consists in an original learning method for solving discounted second-order elliptic equation in non-divergent form, with potentially unknown diffusion (and unknown drift function in the case of the model-free algorithm).

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

## References

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

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

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

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

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

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

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

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

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

[10] S. J. Bradtke and A. G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(1):33–57, 1996.

[11] P. Dayan. The convergence of TD($\lambda$) for general $\lambda$. *Machine Learning*, 8(3):341–362, 1992.

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

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

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

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

[16] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

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

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

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

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

[21] D. E. Kirk. Optimal control theory: An introduction. 1970.

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

[23] N. V. Krylov. *Lectures on Elliptic and Parabolic Equations in Holder spaces*. Number 12. American Mathematical Soc., 1996.

[24] O. A. Ladyzhenskaia and N. N. Uraltseva. *Linear and quasilinear elliptic equations [by] Olga A. Ladyzhenskaya and Nina N. Uraltseva. Translated by Scripta Technica. Translation editor: Leon Ehrenpreis*. Academic Press New York, 1968.

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

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

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

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

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

[30] H. Pham, X. Warin, and M. Germain. Neural networks-based backward scheme for fully nonlinear PDEs. *SN Partial Differential Equations and Applications*, 2(1):1–24, 2021.

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

[32] K. Rudd. *Solving Partial Differential Equations using Artificial Neural Networks*. PhD thesis, Duke University, 2013.

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

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

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

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

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

[38] J. Tsitsiklis and B. Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in Neural information Processing Systems*, 9, 1996.

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

[40] C. Yildiz, M. Heinonen, and H. Lähdesmäki. Continuous-time model-based reinforcement learning. In *International Conference on Machine Learning*, pages 12009–12018. PMLR, 2021.

# A Proof of the main results

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

## A.1 Proof of Theorem 4.1

**Theorem A.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 $\theta_0 \in \Theta$, the sequence $(\theta_k)_{k\geq 0}$ is defined by induction by,*

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

*for $k \geq 0$, where $\gamma_k > 0$ is convergent to zero and $\sum_{k\geq 0}\gamma_k = \infty$, $|\mathbb{E}\left[g_k|\theta_k\right] - A\theta_k - b| \leq (1 + |\theta_k|)\varepsilon_k$, $\varepsilon_k \in \mathbb{R}_+$ 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 $\theta^*$ and*

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

*Proof.* Up to starting the iterative algorithm from $\theta_1$ instead of $\theta_0$, 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 $\theta_k$, imply

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

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

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

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

*Proof of Theorem 4.1.* The proof only consists in checking that we can apply Theorem A.1. Using the same notation as in Theorem A.1, we define,

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

Then, we get

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

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 B.1. From Theorem 2.1, we get

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

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

### A.2 Proof of Theorem 4.4

We start with the following definitions,

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

*Proof of Theorem 4.4.* 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 $\gamma$.

Using Lemma B.1, we get

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

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 B.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,

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

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

$$\begin{aligned}
\eta_{k+1}^r &= (I_d - \gamma H)\eta_k^r + \chi_k^r + \Delta t_k\psi_k^r,\\
\eta_0^0 &= \eta_0, \qquad \eta_0^r = 0 \text{ if } r \geq 1,
\end{aligned} \tag{14}$$

where $\chi_k^r$ and $\psi_k^r$ are defined by

$$\begin{aligned}
\chi_k^0 &= \gamma(H - H_k(X_k))\theta^* + \gamma\left(\mathbb{E}\left[\varphi(X_k)r(X_k)\right] - \varphi(X_k)r(X_k)\right),\\
\psi_k^0 &= \gamma\Delta t_k^{-1}\left(\mathbb{E}[\varphi(X_k)\mathcal{L}v(X_k,\theta^*)] - \mathbb{E}[\varphi(X)\mathcal{L}v(X,\theta^*)]\right) - \gamma\varphi(X_k)R_{2,k}^\top\theta^*,\\
\chi_k^{r+1} &= \gamma(H - H_k(X_k))\eta_k^r,\\
\psi_k^{r+1} &= \gamma\left(\Delta t_k^{-1}(\mathbb{E}[H(X_k)] - H) - \varphi(X_k)R_{2,k}^\top\right)\eta_k^r,
\end{aligned} \tag{15}$$

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 $\chi_r^k$ and $\psi_r^k$.* Here, we prove by induction on $r$ and $k$ that

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

where $R^2 = 3\widetilde{C}\left(\left\|\mathcal{L}_{\rm lin}\varphi + \mathbb{E}[R_0(\cdot,\xi_0)]\right\|_\infty + \Delta t_0^{\frac{1}{2}}\left\|R_1(\cdot,\xi)\right]\right\|_\infty + 2\varepsilon^{-1}\sup_{k\geq 0}\left\|R_2(\Delta t_k,\cdot,\xi)\right]\right\|_\infty + 2\varepsilon^{-1}\Big)$, $0 < \varepsilon < \Delta_0^{-2}$ is a constant that will be defined later, $\widetilde{C}$ is the constant from Lemma B.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 (14) and (20),

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

Then, concerning $\chi_{k+1}^r$, using Lemma B.4, we get

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

Finally, using Lemma B.4 once again for $\psi_{k+1}^r$, we get,

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

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 $\theta^*$ 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

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

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

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

which implies that

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

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

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

where we used (23) to obtain the last line. Then let us pass to $I^r_{k,1}$,

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

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

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

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

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

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 $\sum_{i=0}^{k-1}\Delta t_i^2\le a\ln(1+k)$, we obtain that $C_k\le(1+k)^{a\varepsilon}$ and $\sum_{i=0}^{k-1}C_i$ is of order $k^{1+a\varepsilon}$ leading to the desired inequality up to changing $\varepsilon$ into $a^{-1}\varepsilon$. □

# B Technical lemmas

## B.1 Expansions of the temporal differences

**Lemma B.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*

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

*for some $R_2(\Delta t, x, \xi)$ such that, if $\xi$ 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 $\Delta t$ 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 $\varphi$ 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). □

### B.2 Some lemmas used in the proof of Theorem 4.4

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

$$S^2 \leq \mathrm{tr}(S)S \tag{16}$$

$$A^\top A = -A^2 \leq \frac{2}{\rho\sigma^2}\left\|b + \frac{\sigma^2}{2}\nabla_x \ln(m)\right\|_\infty^2 S^2 \tag{17}$$

$$(SA - AS) \leq 2\sqrt{\frac{2}{\rho\sigma^2}}\left\|b + \frac{\sigma^2}{2}\nabla_x \ln(m)\right\|_\infty S^2, \tag{18}$$

$$\mathbb{E}\left[H(X)H(X)^\top\right] \leq \rho^{-1}\|\mathcal{L}_{\mathrm{lin}}\varphi(X)\|_\infty^2 S. \tag{19}$$

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

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

where the last line is obtained using (8) and the following integration by parts,

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

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

*Second step: proving the four inequalities.* The first inequality (16) is straightforward, it only relies on the fact that $S$ is symmetric and positive. The fourth inequality (19) is straightforward using the definitions of $H(X)$ and $S$. The third inequality (18) is a consequence of (17). Therefore, there is only (17) left to prove. Let us take $\lambda \in \mathbb{C}$ a complex eigenvalue of $H$, and $\theta$ 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 B.3.** *For $\gamma \leq R^{-2}$, the following two inequalities hold for any $k \geq 0$,*

$$(I_d - \gamma H^\top)(I_d - \gamma H) \leq I_d - \gamma S \tag{20}$$

$$\left(I_d - (I - \gamma H^\top)^k\right)\left(I_d - (I - \gamma H)^k\right) \leq \gamma^2 k^2 H^\top H, \tag{21}$$

$$\left(I_d - (I - \gamma H^\top)^k\right)\left(I_d - (I - \gamma H)^k\right) \leq 4\left(1 + \frac{2}{\rho\sigma^2}\|b + \nabla_x \ln(m)\|_\infty^2\right) I_d, \tag{22}$$

$$\left(I_d - (I - \gamma H^\top)^k\right)\left(H^{-1} + H^{-\top}\right)\left(I_d - (I - \gamma H)^k\right) \leq 2\gamma k\left(1 + \sqrt{\frac{2}{\rho\sigma^2}}\|b + \nabla_x \ln(m)\|_\infty\right) I_d. \tag{23}$$

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 (23) is a straightforward consequence of (21) and (22). Then, concerning (20), it is sufficient to write $(I_d - \gamma H^\top)(I_d - \gamma H) = I_d - 2\gamma S + \gamma^2\left(S^2 + SA - AS - A^2\right)$, and use the definition of $R$, (16), (17) and (18). Therefore, it only remains to prove (21) and (22).

*First step: proving* (21). Let us proceed by induction, the case $k = 0$ is straightforward. Let us denote $y_k = \left(I_d - (I_d - \gamma H)^k\right)$ and assume that the inequality holds for $k$. One may notice that for $\theta \in \mathbb{R}^d$, using (20), we obtain

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

which implies $y_k^\top \left(I_d - \gamma H\right)^\top H + H^\top \left(I_d - \gamma H\right) y_k \leq 2\gamma k H^\top H$. Using the latter inequality, the relation $y_{k+1} = (I_d - \gamma H)y_k + \gamma H$, and (20) again, we get

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

This concludes the induction.

*Second step: proving* (22). In this step, we will only work with the complex eigenvalues of $H$: let $\lambda \in \mathbb{C}$ be one of them, we get

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

This implies

$$\begin{aligned}
\left|1-(1-\gamma\lambda)^k\right| &\leq \gamma|\lambda| \sum_{j=0}^{k-1} |1-\gamma\lambda|^j \\
&\leq \frac{\gamma|\lambda|}{1-|1-\gamma\lambda|} \\
&\leq \frac{\gamma|\lambda|}{1-(1-\gamma\Re(\lambda))^{\frac{1}{2}}} \text{ using (20),} \\
&\leq \frac{\gamma|\lambda|}{1-(1-\frac{\gamma}{2}\Re(\lambda))} \text{ because } \gamma\Re(\lambda) \leq 1, \\
&\leq 2\sqrt{1+\frac{\Im(\lambda)^2}{\Re(\lambda)^2}} \\
&\leq 2\left(1+\frac{2}{\rho\sigma^2}\|b+\nabla_x \ln(m)\|_\infty^2\right)^{\frac{1}{2}},
\end{aligned}$$

where the last inequality comes from a similar argument as in the proof of (17). This concludes the proof.

□

**Lemma B.4.** *Assume* **H3**. *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 2.1. 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 Theorem 2.1, 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. □

### B.3 Calculus of variances and covariances

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

$$\operatorname{Var}\left(\xi \cdot \nabla_x v(x)\right) = \left|\nabla_x v(x)\right|^2, \tag{24}$$

$$\operatorname{Var}\left(\xi^\top D^2 v(x)\xi - \Delta_x v(x)\right) = \operatorname{tr}\left(D_x^2 v(x)^2\right). \tag{25}$$

*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 $(X, \xi)$ and $(X, P^\top \xi)$ have the same law and $\xi$ is independent of $X$ and $D$, this implies

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

This concludes the proof. □

# C Complementary results

## C.1 Counterpart of Lemma 3.3 with discrete observations of the continuous dynamic

In this section, we prove the counterpart of Lemma 3.3 to the case when the samples are directly observed from the continuous dynamics (1), instead of the discrete one (3). Basically, the proof consists in replacing the Taylor expansions used in the other proofs by Itô Calculus. We believe that all results in the present paper may be adapted to this case, using similar arguments.

**Lemma C.1.** *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 $\delta_{\Delta t}^{\mathrm{cont}}$ the continuous temporal difference by,*

$$\begin{aligned}
&\delta_{\Delta t}^{\mathrm{cont}} = (\Delta t)^{-1}\left[v(x,\theta_k) - \gamma v(X_{\Delta t},\theta) - r(x)\Delta t + \sigma W_{\Delta t}\cdot\nabla_x v(x,\theta)\right], \\
&\textit{where } dX_t = b(X_t) + \sigma dW_t, \ \textit{and} \ X_0 = x.
\end{aligned}$$

*The mean and variance of $\delta_{\Delta t}^{\mathrm{cont}}$ satisfy,*

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

*Proof.* In this proof, the dependence of $v$ in $\theta$ is omitted. From Itô calculus, we have,

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

Therefore, the continuous temporal difference satisfies,

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

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

Take $g : \Omega \to \mathbb{R}$ a bounded continuous function (we take $g = \nabla_x v\cdot b$ for the proof of the convergence of the third term, and $g = \Delta_x v$ for the proof concerning the fourth term). We define $A$ as a set of measure zero such that $(X_t(\omega))_{0\le t\le 1}$ is continuous for any $\omega \in \Omega_X\backslash A$ (where $\Omega_X$ 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 $\omega$), 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 $\lim_{\Delta t\to 0}\mathbb{E}[\delta_{\Delta t}^{\mathrm{cont}}]=\mathcal{L}v(x,\theta)$.

Similar arguments imply that

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

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

$$\lim_{\Delta t\to 0}\mathrm{Var}(\delta_{\Delta t}^{\mathrm{cont}})=\lim_{\Delta t\to 0}\frac{\sigma^2}{\Delta t^2}\mathbb{E}\left[\left|\int_0^{\Delta t}(\nabla_x v(X_t)-\nabla_x v(x))\cdot dW_t\right|^2\right]. \tag{26}$$

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

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

Let us prove that the first integrable in the latter equality leads to a vanishing term only in the limit (26). This time, we take $g=D_x^2 v\, b+\nabla_x\Delta_x v$, let us consider the following sequence of inequalities

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

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

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

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

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

This concludes the proof. □

### C.2 Similar analysis of the residual gradient method

In this section, we state similar results as Theorems 4.1 and 4.4 while replacing the used algorithm by the residual gradient method, instead of TD(0). We want to insists that some notations in this section and the followings will differ from the rest of the article, more precisely, it consists in $\theta^*$, $\theta_\mu^*$, $(\theta_k)_{k\ge 0}$, $(\widetilde{\theta}_k)_{k\ge 0}$. Indeed, those quantities are defined using TD(0), while here they come from RG. Let us start by defining those quantities here.

The unregularized corrected RG algorithm consists in computing $(\theta_k)_{k\geq 0}$ as,

$$\theta_{k+1} = \theta_k - \gamma_k \delta_k \nabla_\theta \delta_k, \tag{RG}$$

and the $\mu$-regularized corrected RG algorithm consists in computing $(\widetilde{\theta}_k)_{k\geq 0}$ as,

$$\widetilde{\theta}_{k+1} = \Pi_{B(0,M)}\big(\widetilde{\theta}_k - \nabla_\theta \frac{\gamma_k}{2}(\delta_{\Delta t_k}(X_k, X'_k, \widetilde{\theta}_k)^2 - \mu\widetilde{\theta}_k)\big),$$

where $M$ is assumed to be a known upper bound of $\theta^*_\mu$ which is defined later as the limit of the $\mu$-regularized.

Let us define $F_\mu$ the RG cost,

$$F_\mu(\theta) = \mathbb{E}_X\left[\mathcal{L}v(X,\theta)^2\right] + \frac{\sigma^4}{2}\mathbb{E}_X\left[\mathrm{tr}\left(D_x^2 v(X,\theta)^2\right)\right] + \frac{\mu}{2}|\theta|^2.$$

Then, $\theta^*_\mu$ is defined by,

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

Similarly, we define $F = F_0$ and $\theta^* = \theta^*_0$. The following theorem is the counterpart to RG of Theorem 4.1, it concerns the convergence rate of the regularized RG method.

**Theorem C.2.** *Assume* **H1**, **H2**, **H3**, $\mu > 0$, $\gamma_k = \frac{2}{\mu(k+1)}$ *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 the proof in the following section C.3.

Then, we state below the counterpart to RG of Theorem 4.4, it concerns the convergence rate of the unregularized RG method with constant learning step and an averaging method.

**Theorem C.3.** *Assume* **H1**, **H2** *and* **H3**. *and that* $\theta^*$ *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* $\gamma < R^{-2}$, $k \geq 1$,

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

*where* $\bar{\theta}_k = \frac{1}{k}\sum_{i=0}^{k-1}\theta_i$, *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* $\gamma < R^{-2}$, $k \geq 0$, *the latter inequalities are replaced with the following ones respectively*

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

The proof consists in repeating the proof of Theorem 4.4, it is even simpler since here all the linear operators are symmetric.

### C.3 Proof of Theorem C.2

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

**Theorem C.4.** *Let* $f : \Theta \to \mathbb{R}$ *be* $\mu$*-convex, L-semi-concave, and such that* $\theta^* = \mathrm{argmin}_\theta f(\theta)$ *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,*

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

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

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

*Proof.* Up to starting the iterative algorithm from $\theta_1$ instead of $\theta_0$, 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 $\theta_k$, imply

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

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

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

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

*Proof of Theorem C.2.* This proof consists in checking that we can apply Theorem C.4, using the following notations,

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

Thus, we get,

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

Then, from Theorem 2.1, 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 C.4 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. □

### C.4 Possible extensions of RG

One important weakness of RG algorithm is the presence of the term $\frac{\sigma^4}{2}\mathbb{E}_X\left[\mathrm{tr}\left(D_x^2 v(X,\theta)^2\right)\right]$ in the definition of $F_\mu$. In this section, we propose four alternatives to remove it. More precisely, we replace $F_\mu$ 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 $\theta_\mu^* = \mathrm{argmin}\,\widetilde{F}_\mu^*$. Similarly, we take $\widetilde{F} = \widetilde{F}_0$ and $\theta^* = \theta_0^*$.

In particular, Theorems C.2 and C.3 hold with this four alternatives.

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

$$\theta_{k+1} = \theta_k - \gamma_k \left( \nabla_\theta |\delta_k|^2 + \mu \theta_k \right), \quad \text{with } \delta_k = \frac{1}{n_k} \sum_{i=0}^{n_k - 1} \delta_{\Delta t_k}(X_{k,t_i}, X_{k,t_{i+1}}, \theta_k), \tag{MS-RG}$$

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 C.2 then hold with $\gamma_k = \frac{4}{\mu(k+1)}$, $n_k \geq c^{-1}\sqrt{k+1}$, $n_k \Delta t_k \leq c/\sqrt{k+1}$, and $\sigma_k \leq c k^{-\frac{1}{8}}$, for some $c > 0$, for any $k \geq 0$. In particular the proofs or the counterparts to the multistep setting of Theorems C.2 and C.3, are similar to the originals but using Lemma C.5 below instead of Lemma B.1.

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

$$\theta_{k+1} = \theta_k - \gamma_k \left( \nabla_\theta |\delta_k|^2 + \mu \theta_k \right), \quad \text{with } \delta_k = \delta_{\sigma_k, \Delta t_k}(X_k, \xi_k, \theta_k), \tag{$\sigma$RG}$$

where $\delta_{\sigma_k, \Delta t_k}$ is $\delta_{\Delta t_k}$ where $\sigma$ has been replaced by $\sigma_k$ in (3) 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 4.1 then hold with $\gamma_k = \frac{4}{\mu(k+1)}$, $\Delta t_k \leq c/\sqrt{k+1}$, and $\sigma_k \leq c k^{-\frac{1}{8}}$, for $k \geq 0$.

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

$$\theta_{k+1} = \theta_k - \gamma_k \left( \nabla_\theta |\delta_k|^2 + \mu \theta_k \right), \quad \text{with } \delta_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \delta_{\Delta t_k}(X_k, \xi_{k,i}, \theta_k), \tag{MB-RG}$$

where $(N_k)_{k \geq 0}$ are the size of the mini-batches. The conclusions of Theorem 4.1 then hold with $\gamma_k = \frac{4}{\mu(k+1)}$, $\Delta t_k \leq c/\sqrt{k+1}$, and $N_k \geq c^{-1}\sqrt{k}$, for $k \geq 0$.

**Changing the law of the noise.** Note that the perturbating term from (11) 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 $\xi_k$ 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 $\zeta_k$ is a random vector, each of its coordinate being an independent Rademacher random variable.

Using Donsker's theorem [13], 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: $\Delta t_k$ is replaced by $\Delta t_k^{\frac{1}{2}}$ (this is a consequence of the central limit theorem). The conclusions of Theorem 4.1 then hold with $\gamma_k = \frac{4}{\mu(k+1)}$ and $\Delta t_k \leq c/(k+1)$, for $k \geq 0$.

### C.5 Counterpart to Lemma B.1 in the multi-step setting

**Lemma C.5.** *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*

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

*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

$$\delta_{\Delta t}^n(x, \xi, \theta) = \frac{1}{n} \sum_{i=0}^{n-1} \delta_{\Delta t}(X_{t_i}, X_{t_{i+1}}, \theta). \tag{27}$$

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 B.1 imply

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

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