Reinforcement Learning in near-continuous time for continuous state-action spaces

Lorenzo Croissant[1, 2] **Marc Abeille**[2] **Bruno Bouchard**[1]
[1]CEREMADE, Université Paris Dauphine-PSL, CNRS [2]Criteo AI Lab
{croissant,bouchard}@ceremade.dauphine.fr
m.abeille@criteo.com

Abstract

We consider the Reinforcement Learning problem of controlling an unknown dynamical system to maximise the long-term average reward along a single trajectory. Most of the literature considers system interactions that occur in discrete time and discrete state-action spaces. Although this standpoint is suitable for games, it is often inadequate for mechanical or digital systems in which interactions occur at a high frequency, if not in continuous time, and whose state spaces are large if not inherently continuous. Perhaps the only exception is the Linear Quadratic framework for which results exist both in discrete and continuous time. However, its ability to handle continuous states comes with the drawback of a rigid dynamic and reward structure. This work aims to overcome these shortcomings by modelling interaction times with a Poisson clock of frequency ε^{-1}, which captures arbitrary time scales: from discrete ($\varepsilon = 1$) to continuous time ($\varepsilon \downarrow 0$). In addition, we consider a generic reward function and model the state dynamics according to a jump process with an arbitrary transition kernel on $\mathbb{R}^d$. We show that the celebrated optimism protocol applies when the sub-tasks (learning and planning) can be performed effectively. We tackle learning within the eluder dimension framework and propose an approximate planning method based on a diffusive limit approximation of the jump process. Overall, our algorithm enjoys a regret of order $\tilde{\mathcal{O}}(\varepsilon^{1/2}T + \sqrt{T})$. As the frequency of interactions blows up, the approximation error $\varepsilon^{1/2}T$ vanishes, showing that $\tilde{\mathcal{O}}(\sqrt{T})$ is attainable in near-continuous time.

1 Introduction

Controlling a dynamical system to drive it to optimal long-term average behaviour is a key challenge in many applications, ranging from mechanical engineering to econometrics. Reinforcement Learning (RL) aims to do so when the system is a priori unknown by tackling jointly both the control and the statistical inference of the system. This joint objective is even more important in the online version of the problem, in which one interacts with the system along a single trajectory (no resets or episodes). In the last decades, the insights of Bandit Theory (see e.g. [28]) have been leveraged to tackle the RL problem, while addressing the inherent exploration-exploitation dilemma that naturally arises in sequential decision-making (see e.g. [35, § 4.2]). However, most literature considers interactions that occur in discrete time, which is not always applicable when events are triggered by a digital system. Such systems are pervasive in finance and advertising, for instance, and typically have interactions occurring at a very high frequency, with each interaction having only a marginal impact on the state of the system.

Near-continuous time, continuous state-space. A natural approach to plan in such systems is to directly model the problem in continuous time. This is the common approach in finance, see for

16th European Workshop on Reinforcement Learning (EWRL 2023).

instance [15, 17, 31]. However, the continuous time approach conflicts with the sample-based nature of statistical learning theory that fundamentally takes place in discrete time. As such, learning requires careful modelling of the data-generating process and its arrival times. We consider interactions governed by a Poisson clock, setting the expected inter-arrival time of the clock to a parameter $\varepsilon \in (0,1)$. This allows us to model a continuum of situations: from discrete time $\varepsilon = 1$, to continuous time $\varepsilon \downarrow 0$. We are interested in the regime in which $\varepsilon \ll 1$.

Concurrently, a prerequisite for real-world applicability is the ability to model complex dynamics and rich reward signals for continuous state variables. With this in mind, we focus on the model-based approach where the transition and the reward function belong to a parameterised class of functions operating on a continuous state-action space. This level of generality poses challenges regarding all three key sub-tasks of RL: which are planning, learning, and the explore-exploit trade-off.

Discrete and continuous control. For discrete-time dynamics on finite state-action spaces, the planning problem falls under the umbrella of Markov Decision Processes (MDPs) which have been extensively reviewed in [33]. The finite nature of MDPs is at the heart of their theoretical and computational success. Their extension to countable or even continuous state spaces is, however, non-trivial; see e.g. [11, § 4.6, p.245] for a review of the challenges. Perhaps the only exception which retains those nice theoretical and computational properties is the celebrated Linear Quadratic (LQ) framework [21]. However, both frameworks are limited in their expressive power. In contrast, the continuous-time theory of Stochastic Control has demonstrated how to effectively solve the control problem for arbitrary regular dynamics on continuous state-spaces. It enjoys a rich and mature literature [5, 6, 30], both on the theoretical aspects as well as numerical solvers based on Partial Differential Equations (PDEs), another storied field [9, 12, 25]. The near-continuous time framework lies between the two theories, and recent results of [4] show how to navigate between them and approximately solve the planning problem in the high-frequency interactions regime by solving its diffusive counterpart.

Learning non-linear systems. Similar to the planning problem, the natural way to move beyond finite Markov chain models and towards continuous state dynamics is through linear models. The least-squares estimator enjoys strong theoretical guarantees including adaptive confidence sets that can be efficiently maintained online, see e.g. [2]. Extensions [32, 34] showed how to extend this approach to richer model classes through the use of Non-Linear Least Squares (NLLS). This framework subsumes standard least squares and has been successful in many dynamics by retaining its key properties regarding confidence sets. While providing a protocol for learning with NLLS, Russo and Van Roy characterised, in [34], the trade-off between the richness of the model and the hardness of its learning through two quantities of the model class: the log-covering number, and the eluder dimension which summarises the difficulty of turning the information from data into predictive power.

Optimistic exploration. Optimism in the Face of Uncertainty (OFU) has proven highly successful in sequential decision-making from bandits to RL. The works of [7, 10, 20] showed how to extend the celebrated UCB [8] algorithm from bandits to finite MDPs; later, extensions were made to continuous state in the LQ setting, see e.g. [1, 3, 16] and references therein. Extension from bandit to MDP and then to LQ raised new challenges that persist in our setting. First, the agent should not revise its behaviour too often to prevent dithering, which requires the design of a lazy update-scheme. Second, generic continuous states-spaces models come with inherent unboundedness, and one must carefully address stability issues.

In this work, we consider the near-continuous time system interaction model and propose an optimistic algorithm for online reinforcement learning in the average reward setting[1]. Our approach builds on the work of [4] and the connection to the diffusive regime to address the planning sub-task, yielding $\varepsilon^{1/2}$-optimal policies. Furthermore, we perform the learning with NLLS extending the work of [34] to our near-continuous time and unbounded state setting. Underlying the extension of both these two approaches is a careful treatment of the state boundedness which we do with Lyapunov stability arguments. Overall, our algorithm enjoys near-optimal performance as its regret scales with $\tilde{\mathcal{O}}(\varepsilon^{1/2}T + \sqrt{T})$. As the frequency of interactions increases ($\varepsilon \downarrow 0$) the approximation error vanishes, showing that $\tilde{\mathcal{O}}(\sqrt{T})$ is attainable in near-continuous time.

[1] Also known as, *average cost per stage*, *long-run average*, or *ergodic* setting.

2 Setting

We consider an agent interacting with its environment to maximise a long-term average reward. At each interaction, it observes the current state of the system $x \in \mathbb{R}^d$, takes action $a \in \mathbb{A} \subset \mathbb{R}^{d_\mathbb{A}}$, and receives reward $r(x, a)$, for $r : \mathbb{R}^d \times \mathbb{A} \to \mathbb{R}$. The system then transitions to the state x' according to

$$x' = x + \mu_{\theta^*}(x, a) + \Sigma\xi \quad \text{with} \quad \xi \sim \mathcal{N}(0, \mathrm{I}_d),$$

$\Sigma \in \mathbb{R}^{d\times d}$, and in which $\mu_{\theta^*} : \mathbb{R}^d \times \mathbb{A} \to \mathbb{R}^d$ is the deterministic motion of the system[2]. Contrasting with the standard setting, we consider here the interactions to occur in a random fashion, which we model by an independent Poisson process of intensity ε^{-1}. As such, ε parameterises the mean wait time between events and gives us a direct control on the frequency of interactions.

State dynamics. Let $\Omega := \mathbb{D}$ be the space of *càdlàg* functions from $[0, +\infty)$ to $\mathbb{R}^d$, and let $\mathbb{P}$ be a probability measure on Ω. We formalise the interaction time and the noise process as a marked $\mathbb{P}$-compound Poisson process $(N_t)_{t\in\mathbb{R}_+}$ of intensity $\varepsilon^{-1} \geq 1$. We denote by $(\tau_n)_{n\in\mathbb{N}}$ its arrival (interaction) times, with $\tau_0 := 0$, and by $(\xi_n)_{n\in\mathbb{N}}$ its marks, which are independent of everything else and drawn i.i.d. according to the centred standard Gaussian measure ν on $\mathbb{R}^d$. We encode the information available at time $t \in \mathbb{R}_+$ in the σ-algebra $\mathcal{F}_t := \sigma((\tau_n, \xi_n)_{\tau_n \leq t})$ and with the filtration $\mathbb{F}$ defined as the completion of $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$. Let $\mathcal{A}$ be the set of $\mathbb{F}$-adapted $\mathbb{A}$-valued processes, referred to as *controls*. For any initial state $x_0 \in \mathbb{R}^d$ and $\alpha \in \mathcal{A}$, we let X^{α,θ^*} denote the pathwise-unique solution of

$$\begin{cases} X^{\alpha,\theta^*}_{\tau_n} = X^{\alpha,\theta^*}_{\tau_{n-1}} + \mu_{\theta^*}(X^{\alpha,\theta^*}_{\tau_{n-1}}, \alpha_{\tau_{n-1}}) + \Sigma\xi_n \\ X^{\alpha,\theta^*}_{\tau_0} = x_0 \end{cases} . \tag{1}$$

In (1), we model the dynamic according to a jump process and X^{α,θ^*} is then defined at any time $t \in \mathbb{R}_+$ by considering that it is piece-wise constant on each interval $[\tau_{n-1}, \tau_n)$, $n \in \mathbb{N}^*$. Although involved, this definition allows us to define the state process at any time and feature the interplay of the Poisson and wall-time clocks.

Reinforcement learning problem. In our model based paradigm, ignorance about the system is condensed to a single parameter set $\Theta \subset \mathbb{R}^{d_\Theta}$ containing the unknown nominal parameter θ^*. To single out the RL challenges, we further assume that θ^* only affects the drift assuming other quantities (i.e. Σ, ε, and r) are known to the agent. For any $x_0 \in \mathbb{R}^d$, we evaluate the performance of any strategy $\alpha \in \mathcal{A}$ with the long-term average reward criterion defined by

$$\rho^\alpha_{\theta^*}(x_0) := \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{n=1}^{N_T} r(X^{\alpha,\theta^*}_{\tau_n}, \alpha_{\tau_n})\right] . \tag{2}$$

The goal of the agent is to accumulate as much reward as possible, i.e. to compete with the best an omniscient agent can achieve: $\rho^*_{\theta^*}(x_0) := \sup_{\alpha\in\mathcal{A}} \rho^\alpha_{\theta^*}(x_0)$. We evaluate the quality of a learning algorithm generating α according to its regret.

Definition 2.1. For any $T \in \mathbb{R}_+$, $x_0 \in \mathbb{R}^d$, and $\alpha \in \mathcal{A}$, the regret of α is

$$\mathcal{R}_T(\alpha) := T\rho^*_{\theta^*}(x_0) - \sum_{n=1}^{N_T} r(X^{\alpha,\theta^*}_{\tau_n}, \alpha_{\tau_n}) . \tag{3}$$

Noticing that N_T is the number of events up to time T, the definitions of the optimal performance (2) and the regret (3) highlight the interplay between the wall-clock (T) and Poisson clock (N_T). The agent's realised trajectory uses the Poisson clock, which governs interactions, while the ideal performance is understood per unit of wall-clock time.

2.1 Working Assumptions

Of particular interest in our approach is the high-frequency regime in which $\varepsilon \downarrow 0$. In this framework, many interactions occur per unit of time, each of which is of negligible impact both in terms of

[2] While the additive noise structure is a design choice that simplifies the analysis, the choice of parameterising the drift as $x + \mu_{\theta^*}(x, a)$ instead of $\mu_{\theta^*}(x, a)$ does not affect its generality and is made only for convenience.

dynamics and reward. This regime can be encoded by introducing, for any parameter $\theta \in \Theta$, rescaled coefficients $(\bar{\mu}_\theta, \bar{\Sigma}, \bar{r})$ connected to the original parametrisation by

$$\mu_\theta = \varepsilon \bar{\mu}_\theta \,, \quad \Sigma = \varepsilon^{\frac{1}{2}} \bar{\Sigma} \,, \quad \text{and } r = \varepsilon \bar{r} \,.$$

In this rescaled parametrisation, $\bar{\mu}_\theta$, $\bar{\Sigma}$, and $\bar{r}$ are understood as independent of ε. To improve legibility, we will use both representations (μ_θ, Σ, r) and $(\bar{\mu}_\theta, \bar{\Sigma}, \bar{r})$. While the scaling of μ_θ and r in ε arises naturally, the one of Σ is a design choice: we consider the covariance $\Sigma\Sigma^\top$ to be linear in ε. Known as the diffusive regime, this preserves stochasticity[3] as $\varepsilon \downarrow 0$. We now impose regularity assumptions on the drift and reward signal, uniformly over the possible parametrisations and controls $(\alpha, \theta) \in \mathcal{A} \times \Theta$. We take $\|\cdot\|$ to be the Euclidian norm on $\mathbb{R}^d$ and $\|\cdot\|_{\mathrm{op}}$ for the operator norm on $\mathbb{R}^{d\times d}$ associated to $\|\cdot\|$.

Assumption 1. The map $(\bar{\mu}, \bar{r})$ is continuous, and there is $L_0 > 0$ such that for all $(\theta, a) \in \Theta \times \mathbb{A}$

$$L_0 > \sup_{x\in\mathbb{R}^d} \frac{\|\bar{\mu}_\theta(x,a)\|}{1+\|x\|} + \sup_{x\neq x'} \frac{\|\bar{\mu}_\theta(x,a) - \bar{\mu}_\theta(x',a)\|}{\|x-x'\|} + \sup_{x\in\mathbb{R}^d} \|\bar{r}(x,a)\| + \sup_{x\neq x'} \frac{\|\bar{r}(x,a)-\bar{r}(x',a)\|}{\|x-x'\|} .$$

Furthermore, $L_0 > \|\bar{\Sigma}\|_{\mathrm{op}}$ and $\bar{\Sigma}\bar{\Sigma}^\top \succeq \varsigma \, \mathrm{I}_d$ for some $\varsigma > 0$, where $\succeq$ denotes the Loewner order.

Assumption 1 mainly imposes regularity on both $\bar{\mu}_\theta$ and $\bar{r}$ through a Lipschitz condition. We also assume rewards to be bounded, which may be relaxed, but doing so is highly technical and involves trading-off the growth of r with the stability of the process (see Assumption 2). Note that we do not assume boundedness of $\bar{\mu}_\theta$. Finally, we assume non-degeneracy of the noise by requiring $\bar{\Sigma}$ to be full rank.

We conclude with Assumption 2 to ensure stability of the state process. Let $\mathbb{R}^d_* := \mathbb{R}^d \setminus \{0\}$ and $\mathbb{R}_+ := (0, +\infty)$. For $k \in \mathbb{N}$, let $\mathcal{C}^k(\mathbb{R}^d_*; \mathbb{R}_+)$ denote the set of k-times continuously differentiable functions from $\mathbb{R}^d_*$ to $\mathbb{R}_+$. Let ∇ and ∇^2 denote the gradient and Hessian operator respectively.

Assumption 2. There is $(\ell_{\mathscr{V}}, L_{\mathscr{V}}, \mathfrak{c}_{\mathscr{V}}, M_{\mathscr{V}}, M'_{\mathscr{V}}) \in \mathbb{R}^5_+$ and a Lyapunov function $\mathscr{V} \in \mathcal{C}^2(\mathbb{R}^d_*; \mathbb{R}_+)$ satisfying, for any $(x, x', a, \theta) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{A} \times \Theta$, $x \neq x'$, and $\varepsilon \in (0,1)$:

$$\text{(i.)} \quad \ell_{\mathscr{V}} \|x - x'\| \leq \mathscr{V}(x - x') \leq L_{\mathscr{V}} \|x - x'\| \,,$$

$$\text{(ii.)} \quad \sup_{x\in\mathbb{R}^d_*} \|\nabla \mathscr{V}(x)\| \leq M_{\mathscr{V}} \text{ and } \sup_{x\in\mathbb{R}^d_*} \left\|\nabla^2 \mathscr{V}(x)\right\|_{\mathrm{op}} \leq M'_{\mathscr{V}} \,,$$

$$\text{(iii.)} \quad \mathscr{V}(x + \varepsilon\bar{\mu}(x,a) - x' - \varepsilon\bar{\mu}(x',a)) \leq (1 - \varepsilon\mathfrak{c}_{\mathscr{V}})\mathscr{V}(x - x') \,. \tag{4}$$

Assumption 2 is a Lyapunov-like condition through the function $\mathscr{V}$. The condition (i.) requires that $\mathscr{V}$ behaves similarly to a norm, while (ii.) asks that $\mathscr{V}$ be smoothly differentiable everywhere but at 0 and (iii.) imposes a contraction condition on the drifts.

Connection to linear stability. Stability theory has been extensively studied the special case of linear dynamics. In this case, we recover Assumption 2 from the Continuous Algebraic Riccati Equation (CARE; see e.g. [27, § 4.4]). Considering linear dynamics $\bar{\mu}_\theta(x,a) = \bar{A}x + \bar{B}a$ (given matrices $(\bar{A}, \bar{B})$ of appropriate dimensions), continuous stability is guaranteed when the eigenvalues of $\bar{A}$ have negative real-part or, equivalently, by the existence of a positive semi-definite matrix P solving the CARE $\bar{A}^\top P + P\bar{A} = -\,\mathrm{I}_d$. For this P, its associated norm $\mathscr{V} = \|\cdot\|_P$ is the appropriate Lyapunov function for Assumption 2. Indeed, conditions (i.) and (ii.) follow as $\mathscr{V}$ is a norm and, for $\varepsilon \leq 1/2\lambda_{\max}(P)$, we have

$$\begin{aligned}
\mathscr{V}(x + \varepsilon\bar{\mu}(x,a) - x' - \varepsilon\bar{\mu}(x',a))^2 &= (x-x')^\top (P + \varepsilon\bar{A}^\top P + \varepsilon P\bar{A} + \varepsilon^2 P)(x - x') \\
&= (x-x')^\top (P - \varepsilon\,\mathrm{I}_d + \varepsilon^2 P)(x-x') \\
&\leq (x-x')^\top (P - \varepsilon P/\lambda_{\max}(P) + \varepsilon^2 P)(x-x') \\
&\leq (1 - \varepsilon/2\lambda_{\max}(P))\mathscr{V}(x-x')^2 \,.
\end{aligned}$$

Taking the square-root and using $\sqrt{1-\varepsilon/2\lambda_{\max}(P)} \leq 1 - \varepsilon/4\lambda_{\max}(P)$ leads to (iii.) with $\mathfrak{c}_{\mathscr{V}} = 1/4\lambda_{\max}(P)$.

[3]Another common, but more rigid, regime is to consider $\Sigma = \varepsilon\bar{\Sigma}$, whose limit regime is deterministic and known as the fluid limit, see [18].

3 Main results

Our main contribution is a demonstration of the OFU protocol in the near-continuous time continuous state-action RL problem. The ingredients of OFU are: learning from accumulated data to design confidence sets; lazy updates to trade off policy revision and learning guarantees; and planning amongst plausible parameterisations. We summarise this protocol in Algorithm 1.

Algorithm 1 `OFU-Diffusion`

```
Input: confidence level δ, initial state x_0, initial control ϖ_0
for n ∈ ℕ* do
  At time τ_n, receive r(X^{ϖ,θ*}_{τ_{n-1}}, ϖ_{τ_{n-1}}) and X^{ϖ,θ*}_{τ_n}.
  if n satisfies (7) then
    n_k ← n, k ← k + 1,
    Compute θ̂_{n_k} using (5) and C_{n_k}(δ/3) with (6).
    θ̃_k ← argmax_{θ∈C_{n_k}(δ/3)} ρ̄*_θ
    π_k ← π̄*_{θ̃_k} using (9)
  end if
  Play ϖ_{τ_n} := π_k(X^{ϖ,θ*}_{τ_n}).
end for
```

Learning. Our algorithm proceeds by episodes, indexed by $k \in \mathbb{N}$ with n_k denoting the start of the k^{th} episode. At each n_k, Algorithm 1 revises its knowledge using the Non-Linear Least-Square fit and the associated confidence set $\mathcal{C}_{n_k}(\delta)$, defined (for $\beta_n(\delta)$ given in (13) for all $n \in \mathbb{N}$) by

$$\hat{\theta}_{n_k} \in \operatorname*{argmin}_{\theta \in \Theta} \sum_{n=0}^{n_k-1} \left\| X^{\varpi,\theta^*}_{\tau_{n+1}} - X^{\varpi,\theta^*}_{\tau_n} - \mu_\theta(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) \right\|^2 , \tag{5}$$

$$\mathcal{C}_{n_k}(\delta) := \left\{ \theta \in \Theta : \sqrt{\sum_{n=0}^{n_k-1} \left\| \mu_\theta(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) - \mu_{\hat{\theta}_{n_k}}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) \right\|^2} \le \beta_{n_k}(\delta) \right\} . \tag{6}$$

Lazy Updates. Our episodic scheme follows the same rationale as in [1, 20], and triggers updates as soon as enough information is collected. Formally, it constructs a sequence of episodes $\{S_k\}_{k\in\mathbb{N}}$ whose starting times are defined by $n_0 := 0$ and, for any $k \in \mathbb{N}$, n_{k+1} is the first time $n > n_k$ satisfying (7)

$$\sqrt{\sup_{\theta \in \mathcal{C}_{n_k}(\delta)} \sum_{i=0}^{n} \left\| \mu_\theta(X^{\varpi,\theta^*}_{\tau_i}, \varpi_{\tau_i}) - \mu_{\hat{\theta}_{n_k}}(X^{\varpi,\theta^*}_{\tau_i}, \varpi_{\tau_i}) \right\|^2} > 2\beta_n(\delta) . \tag{7}$$

Planning. At the heart of our proposal is the way in which we address the optimistic planning, detailed in Section 4.3. For a given parameter $\theta \in \mathcal{C}_{n_k}(\delta)$, we leverage the connection between our setting and its continuous-time counterpart. We consider continuous-time controls $\bar{\alpha} \in \bar{\mathcal{A}}$ with diffusive average reward given by

$$\bar{\rho}^{\bar{\alpha}}_\theta(x_0) := \liminf_{T\to+\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T \bar{r}(\bar{X}^{\bar{\alpha},\theta}_t, \bar{\alpha}_t)\mathrm{d}t\right] \text{ in which } \begin{cases} \mathrm{d}\bar{X}^{\bar{\alpha},\theta}_t = \bar{\mu}_\theta(\bar{X}^{\bar{\alpha},\theta}_t, \bar{\alpha}_t)\mathrm{d}t + \bar{\Sigma}\mathrm{d}W_t \\ \bar{X}^{\bar{\alpha},\theta}_0 = x_0 \end{cases} \tag{8}$$

in which W denotes a $\mathbb{P}$-Brownian motion, $\bar{\mathbb{F}}$ its filtration, and $\bar{\mathcal{A}}$ the set $\mathbb{A}$-valued $\bar{\mathbb{F}}$-predictable processes. This diffusive problem gives us an optimality criterion and associated optimal control[4]:

$$\bar{\rho}^*_\theta(x_0) := \sup_{\alpha \in \mathcal{A}} \bar{\rho}^{\bar{\alpha}}_\theta(x_0) \text{ and } \bar{\pi}^*_\theta \circ \bar{X}^{\bar{\alpha},\theta} \in \operatorname*{argmax}_{\bar{\alpha}\in\bar{\mathcal{A}}} \bar{\rho}^{\bar{\alpha}}_\theta(x_0) \tag{9}$$

which approximates the original jump-process problem $\rho^*_\theta(x_0)$. This problem admits a Hamilton-Jacobi-Bellman (HJB) equation (given in (18) below) characterising an optimal policy $\bar{\pi}^*_\theta : \mathbb{R}^d \to \mathbb{A}$ which yields a computable optimal Markov control for (9).

[4]We will use the obvious notational confusion between the policy $\bar{\pi}^*_\theta$ and the control process it generates.

Theorem 3.1. *Under Assumptions 1 and 2, for any $\delta \in (0,1)$, $x_0 \in \mathbb{R}^d$, and $\gamma \in (0,1)$, there is a pair $(C_\gamma, C) \in \mathbb{R}_+^2$ of constants independent of ε such that Algorithm 1 achieves*

$$R_T(\varpi) \leq 2C_\gamma \varepsilon^{\frac{\gamma}{2}} T + C\sqrt{\mathrm{d}_{E,T\varepsilon^{-1}} \log(\mathscr{N}^{\varepsilon}_{T\varepsilon^{-1}}) T \log(T\delta^{-1})} \tag{10}$$

with probability at least $1-\delta$, in which $\mathrm{d}_{E,T\varepsilon^{-1}}$ is the $2\varepsilon/\sqrt{T}$-eluder dimension (see [34, Def. 4.] and (56) in Appendix C.2) of the class $\{\mu_\theta\}_{\theta\in\Theta}$ restricted to a ball of radius $\mathcal{O}(\sqrt{\log(T/\varepsilon)})$, and $\log(\mathscr{N}^{\varepsilon}_{T\varepsilon^{-1}})$ is the $\varepsilon^2\|\bar{\Sigma}\|^2_{\mathrm{op}}/T$-log-covering number of this same restricted class.

Theorem 3.1 contains two terms of different nature. The linear term is inherited from the diffusive approximation planning method and scales with $C_\gamma \varepsilon^{\gamma/2}$. The dependency of the constant in γ is inherited from the analysis of [4] and $C_\gamma < +\infty$ holds for $\gamma < 1$. Quantifying the behaviour of C_γ as $\gamma \uparrow 1$ is technically intricate. Nevertheless, our bound indicates that the long run approximation error vanishes as $\varepsilon \downarrow 0$ almost as a fast as $\sqrt{\varepsilon}$. The second term quantifies all other sources of error, and exhibits the expected scaling in the complexity measures of [34], in terms of both eluder dimension and log-covering numbers, as well as the $\sqrt{T}$ horizon dependency.

4 Ideas of the Proof

4.1 Stability

Working with unbounded processes and generic drift requires us to prevent state blow-up, which could degrade regret regardless of learning. In Proposition 4.1 we combine the Lyapunov stability of (4) with concentration arguments to show that unstable trajectories can only happen with low probability. A detailed proof is given in Appendix B.

Proposition 4.1. *Under Assumptions 1 and 2, there is a function $H_\delta(n) = \mathcal{O}\big(\sqrt{\log(n\delta^{-1})}\big)$ such that for any $\delta \in (0,1)$, $\alpha \in \mathcal{A}$, $x_0 \in \mathbb{R}^d$, and $\theta \in \Theta$ we have*

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}_+} \frac{\|X_t^{\alpha,\theta}\|}{H_\delta(N_t)} \geq 1\right) \leq \delta\,. \tag{11}$$

Working on the high-probability event of Proposition 4.1 allows us to handle the unbounded state in the learning, planning, and optimism.

4.2 Learning

Confidence Sets. The crux of our analysis is incorporating Proposition 4.1 into the NLLS method of [34] by refining it to be adaptive to the norm of the state process. For $R > 0$, let $\mathcal{B}_2(R) \subset \mathbb{R}^d$ denotes the Euclidean ball of radius R at 0. To adapt the log-covering number, we can work with H_δ by formally defining $\mathscr{N}^{\varepsilon}_n$ as the size of the smallest cover $\mathscr{C}^{\varepsilon}_n$ of $\mathscr{F}_\Theta := (\mu_\theta)_{\theta\in\Theta}$ such that

$$\sup_{\mu_1\in\mathscr{F}_\Theta} \min_{\mu_2\in\mathscr{C}^{\varepsilon}_n} \sup_{x\in\mathcal{B}_2(H_\delta(n))} \|\mu_1(x) - \mu_2(x)\| \leq \frac{\varepsilon\|\bar{\Sigma}\|^2_{\mathrm{op}}}{n}\,. \tag{12}$$

Restricting the domain of $\mathscr{F}_\Theta$ allows us to handle the richness of unbounded models and states while following [34] to define confidence sets. Let $\delta \in (0,1)$, set $\beta_0 := \varepsilon^{\frac{1}{2}}$, and let

$$\beta_n(\delta) := \beta_0 \vee 2\varepsilon^{\frac{1}{2}}\|\bar{\Sigma}\|_{\mathrm{op}} \left(\sqrt{1 + 2\left(\sqrt{2\log\left(\frac{4\pi^2 n^3}{3\delta}\right)} + \sqrt{2\varepsilon^{\frac{1}{2}}\|\bar{\Sigma}\|^{-1}_{\mathrm{op}}\kappa_n(\delta)}\right)} + \sqrt{\kappa_n(\delta)}\right) \tag{13}$$

in which

$$\kappa_n(\delta) := \log\left(\frac{2\pi^2 n^2 \varepsilon \mathscr{N}^{\varepsilon}_n}{3\delta}\left(\|\bar{\Sigma}\|^2_{\mathrm{op}} + 8L_0^2(1 + H_\delta(n))\right)\right)\,.$$

Using this choice $(\beta_n)_{n\in\mathbb{N}}$ and replacing n_k by n in (6) formally defines the confidence sets $(\mathcal{C}_n(\delta))_{n\in\mathbb{N}}$. For any $\alpha \in \mathcal{A}$, the probability that the state process X_t^{α,θ^*} outgrows $H_\delta(N_t)$ is small and, thus, this confidence set will hold with high probability as shown by Proposition 4.2.

Proposition 4.2 (Adapted from [32, Prop. 5]). *Under Assumptions 1 and 2, for any $x_0 \in \mathbb{R}^d$, and $\delta > 0$,*

$$\mathbb{P}\left(\left\{\theta^* \in \bigcap_{n=1}^{\infty} \mathcal{C}_n(\delta)\right\} \cap \left\{\sup_{n\in\mathbb{N}^*} \frac{\left\|X_{\tau_n}^{\varpi,\theta^*}\right\|}{H_\delta(n)} \leq 1\right\}\right) \geq 1-\delta\,, \tag{14}$$

Well-posed confidence sets are insufficient for low-regret approaches in the OFU paradigm. This high confidence (low fit error) of the NLLS estimator must be translated as low online prediction error.

Prediction error. To adapt the ϵ-eluder dimension (defined for $\epsilon > 0$ in [32, Def. 3.]), which we denote $\dim_{\mathrm{E}}$, to our unbounded state we proceed on the trajectory. The relevant extension for us is given for $n \in \mathbb{N}^*$ by the $2\sqrt{\varepsilon/n}$-eluder dimension of the class $\{f|_B\}_{f\in\mathscr{F}_\Theta}$ of elements of $\mathscr{F}_\Theta$ restricted to the set $B_n := \mathcal{B}_2(\sup_{t\leq\tau_n}\|X_t^{\varpi,\theta^*}\|)$, denoted by $\mathrm{d}_{\mathrm{E},n} := \dim_{\mathrm{E}}(\{f|_{B_n}\}_{f\in\mathscr{F}_\Theta}, 2\sqrt{\varepsilon/n})$. In Proposition 4.3 we obtain first and second order prediction error bounds from this eluder dimension. In Proposition 4.3 the order notation $\tilde{\mathcal{O}}$ hides terms that are poly-logarithmic in N_t and $\mathrm{d}_{\mathrm{E},N_t}$ whose the full details are given in Appendix C.2.

Proposition 4.3. *Under Assumptions 1 and 2, for any $\delta \in (0,1)$, $\alpha \in \mathcal{A}$, $x_0 \in \mathbb{R}^d$, and $t \in \mathbb{R}_+$, we have with probability at least $1-\delta$*

$$\sum_{n=1}^{N_t}\left\|\mu_{\hat{\theta}_n}(X_{\tau_n}^{\alpha,\theta^*},\alpha_{\tau_n}) - \mu_{\theta^*}(X_{\tau_n}^{\alpha,\theta^*},\alpha_{\tau_n})\right\| \leq \tilde{\mathcal{O}}\left(\sqrt{\varepsilon \mathrm{d}_{\mathrm{E},N_t}\log(\mathscr{N}_{N_t}^{\varepsilon})N_t} + \mathrm{d}_{\mathrm{E},N_t}\right)\,, \tag{15}$$

and

$$\sum_{n=1}^{N_t}\left\|\mu_{\hat{\theta}_n}(X_{\tau_n}^{\alpha,\theta^*},\alpha_{\tau_n}) - \mu_{\theta^*}(X_{\tau_n}^{\alpha,\theta^*},\alpha_{\tau_n})\right\|^2 \leq \tilde{\mathcal{O}}\left(\mathrm{d}_{\mathrm{E},N_t}\log(\mathscr{N}_{N_t}^{\varepsilon})\right)\,. \tag{16}$$

Lazy updates. We leverage the second order bound (16) of Proposition 4.3 to define our lazy-update scheme (7). We show in Appendix E that this scheme does not degrade the speed at which Algorithm 1 learns by more than a constant factor, while also ensuring that the policy is only updated logarithmically in the number of interactions up to any horizon.

4.3 Planning

Algorithm 1 requires us to be able to plan using any $\theta \in \Theta$, and as such we will extend the definitions of $X^{\alpha,\theta}$, $\rho_\theta^\alpha(x_0)$, $\rho_\theta^*(x_0)$ to any $(\alpha,\theta) \in \mathcal{A}\times\Theta$ by replacing θ^* by θ in (1) and (2). Let $\mathscr{A}$ be the set of measurable maps from $\mathbb{R}^d$ to $\mathbb{A}$. For a given $\theta \in \Theta$, the well-posedness of the control problem $\rho_\theta^*(x_0)$ and its resolution are non-trivial.

Proposition 4.4 (Adapted from [4, Thm. 2.3, Rem. 2.4.]). *Under Assumptions 1 and 2, there is $L_W \in \mathbb{R}_+$, independent of ε, such that for any $\theta \in \Theta$*

(i.) The map $x \mapsto \rho_\theta^*(x)$ is constant, taking only one value which we denote by $\rho_\theta^* \in \mathbb{R}$;

(ii.) There is an L_W-Lipschitz function W_θ^* such that

$$\varepsilon\rho_\theta^* = \max_{a\in\mathbb{A}}\left\{\mathbb{E}[W_\theta^*(x+\mu_\theta(x,a)+\Sigma\xi)] - W_\theta^*(x) + r(x,a)\right\}\ \forall x \in \mathbb{R}^d\,; \tag{17}$$

(iii.) There is $\pi_\theta^* \in \mathscr{A}$, such that for all $x \in \mathbb{R}^d$, $\pi_\theta^*(x)$ maximises the right hand side in (17), and $\pi_\theta^* \circ X^{\pi_\theta^*,\theta}$ is an optimal Markov control, i.e. $\rho_\theta^{\pi_\theta^*}(\cdot) \equiv \rho_\theta^*$.

Proposition 4.4.(i.) shows that the control problem ρ_θ^* is independent of the initial conditions and meaningfully ergodic, which follows from stability analysis of the process using (4). Points (ii.) and (iii.) show that there is an optimal policy, which can be computed by solving the HJB equation (17). As before, confusing policies in $\mathscr{A}$ and controls in $\mathcal{A}$, we will write ρ_θ^π and $X^{\pi,\theta}$ to simplify notation. Unfortunately (17) is an integral equation with low regularity, owing to the non-local jumps of the system, which complicates its analysis and the construction of numerical solvers.

Diffusion limit. In the limit regime of interest, i.e. as $\varepsilon \downarrow 0$, the non-local behaviour of (17) vanishes and it becomes a diffusive HJB equation. The associated diffusive control problem $\bar{\rho}_\theta^*(x_0)$ has been extensively studied, see e.g. [5, 6].

Proposition 4.5 (Adapted from [4, Thm. 3.4.]). *Under Assumptions 1 and 2, for any $\theta \in \Theta$,*

(i.) The map $x \mapsto \bar{\rho}^*_\theta(x)$ is constant, taking only one value which we denote by $\bar{\rho}^*_\theta \in \mathbb{R}$.

(ii.) There is an L_W-Lipschitz function $\bar{W}^*_\theta \in \mathcal{C}^2(\mathbb{R}^d;\mathbb{R})$ such that

$$\bar{\rho}^*_\theta = \max_{a\in\mathbb{A}} \left\{ \bar{\mu}_\theta(x,a)^\top \nabla \bar{W}^*_\theta(x) + \bar{r}(x,a) \right\} + \frac{1}{2}\operatorname{Tr}[\bar{\Sigma}\bar{\Sigma}^\top \nabla^2 \bar{W}^*_\theta(x)],\ \forall x \in \mathbb{R}^d . \tag{18}$$

(iii.) There is $\bar{\pi}^*_\theta \in \mathscr{A}$ such that, for all $x \in \mathbb{R}^d$, $\bar{\pi}^*_\theta(x)$ maximises the right hand side in (18)*, and $\bar{\pi}^*_\theta \circ \bar{X}^{\bar{\pi}^*_\theta,\theta}$ is an optimal Markov control, i.e. $\bar{\rho}^{\bar{\pi}^*_\theta}_\theta(\cdot) \equiv \bar{\rho}^*_\theta$.*

Proposition 4.5 ensures that the diffusive problem satisfies all the properties of Proposition 4.4 (ergodicity, optimal policy, and HJB equation). However, the HJB (18) is now a second-order local PDE instead of a non-local integral equation. This local equation does not have cross-dependencies between points: the solution at x depends only on its derivatives at x, which is fundamentally simpler than the non-local behaviour of (17). Moreover, this diffusive PDE belongs to a well-studied family, both from the points of view of theory [19, 26] and of numerics [23, 24]. These facts motivate the use of these tools to construct approximate planning methods for (17) in the near-continuous time regime as $\varepsilon \downarrow 0$.

Proposition 4.6 (Adapted from [4, Thm. 3.6.]). *Under Assumptions 1 and 2, for any $\gamma \in (0,1)$, there is a constant $C_\gamma > 0$, independent of ε, such that, for any $\theta \in \Theta$,*

$$|\bar{\rho}^*_\theta - \rho^*_\theta| \le C_\gamma \varepsilon^{\frac{\gamma}{2}} \text{ and } \rho^*_\theta - \rho^{\bar{\pi}^*_\theta}_\theta(0) \le C_\gamma \varepsilon^{\frac{\gamma}{2}} . \tag{19}$$

Moreover, there is a function $e_\theta : \mathbb{R}^d \to \mathbb{R}$ such that,

$$\varepsilon \rho^{\bar{\pi}^*_\theta}_\theta(0) = \mathbb{E}[\bar{W}^*_\theta(x + \mu_\theta(x,a) + \Sigma\xi)] - \bar{W}^*_\theta(x) + r(x, \bar{\pi}^*_\theta(x)) + e_\theta(x) ,\ \forall x \in \mathbb{R}^d \tag{20}$$

and there is $C'_\gamma > 0$, independent of ε, such that $|e_\theta(x)| \le C'_\gamma \varepsilon^{1+\frac{\gamma}{2}}(1 + \|x\|^3)$ for all $x \in \mathbb{R}^d$.

Proposition 4.6, combined with (18) provides a certifiable approximation for solving the control problem (2) with off-the-shelf diffusive HJB solvers, at a cost independent of ε. An example of this methodology is seen in [4, § 4], in which [4, Fig. 1, p. 30] shows the reduction in computational effort. Proposition 4.6 also provides in (20) an HJB-like representation of the approximation, which provides a key with which to analyse the regret incurred when using this approximation.

4.4 Regret Decomposition

To sketch the proof of Theorem 3.1, we work on the high-probability event of Proposition 4.2, and omit martingale measurability issues this could cause. We will also ignore the randomness of jump times and consider $T \lesssim \varepsilon N_T$, with $\lesssim$ denoting inequality up to a constant. Appendix E is dedicated to a complete proof.

Proof sketch of Theorem 3.1. Let $k : \mathbb{N} \to \mathbb{N}$ map an event n to the episode $k(n)$ to which it belongs and let $\theta_n := \tilde{\theta}_{k(n)}$. We begin the regret decomposition by applying the HJB-like equation (20) of Proposition 4.6.(iii.) to the rewards collected along the trajectory $r(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n})$ in the definition of the regret. Conditioning as appropriate, this yields

$$\mathcal{R}_T(\varpi) = T\rho^*_{\theta^*} - \varepsilon \sum_{n=1}^{N_T} \rho^{\bar{\pi}^*_{\theta_n}}_{\theta_n}(0) \tag{R_1}$$

$$+ \sum_{n=1}^{N_T} \mathbb{E}[\bar{W}^*_{\theta_n}(\tilde{X}^{\varpi,\theta_n}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] - \bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}) \tag{R_2}$$

$$+ \sum_{n=1}^{N_T} e_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}) \tag{R_3}$$

in which $\tilde{X}^{\varpi,\theta}_{\tau_{n+1}} := X^{\varpi,\theta^*}_{\tau_n} + \mu_\theta(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) + \Sigma\xi_{n+1}$, for $(n,\theta) \in \mathbb{N} \times \Theta$, is a counterfactual one-step transition assuming parameter $\theta \in \Theta$.

On the event of Proposition 4.2, θ^* is in $\cap_{n\in\mathbb{N}}\mathcal{C}_n(\delta)$ and the optimism of Algorithm 1 ensures that $\bar{\rho}^*_{\theta^*} \leq \bar{\rho}^*_{\theta_n} = \bar{\rho}^{\bar{\pi}^*_{\theta_n}}_{\theta_n}$ for all $n \in \mathbb{N}$. Combining this with Proposition 4.6, show that (R_1) decomposes into

$$R_1 \lesssim \varepsilon \left(\sum_{n=1}^{N_T} (\rho^*_{\theta^*} - \bar{\rho}^*_{\theta^*}) + \sum_{n=1}^{N_T} \left(\bar{\rho}^*_{\theta_n} - \rho^{\bar{\pi}^*_{\theta_n}}_{\theta_n} \right) \right) \leq 4 N_T C_\gamma \varepsilon^{1+\frac{\gamma}{2}} .$$

Also by Proposition 4.6, $R_3 \leq \varepsilon^{1+\frac{\gamma}{2}} N_T (1 + H_\delta(N_T)^3)$. Thus $R_1 + R_3 \lesssim C_\gamma \varepsilon^{\frac{\gamma}{2}} T$.

For (R_2), the identity

$$\tilde{X}^{\varpi,\theta}_{\tau_{n+1}} = \tilde{X}^{\varpi,\theta^*}_{\tau_{n+1}} - \mu_{\theta^*}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) + \mu_\theta(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n})$$

combined with the Lipschitzness of $\bar{W}^*_\theta$ from Proposition 4.5, yields

$$\begin{aligned} R_2 \leq L_{\bar{W}} & \sum_{n=1}^{N_T} \left\| \mu_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) - \mu_{\theta^*}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) \right\| && (R_4) \\ + & \sum_{n=1}^{N_T} \mathbb{E}[\bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_{n+1}}) - W^*_{\theta_{n+1}}(X^{\varpi,\theta^*}_{\tau_{n+1}}) | \mathcal{F}_{\tau_n}] && (R_5) \\ + & \sum_{n=1}^{N_T} \mathbb{E}[W^*_{\theta_{n+1}}(X^{\varpi,\theta^*}_{\tau_{n+1}}) | \mathcal{F}_{\tau_n}] - W^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}) , && (R_6) \end{aligned}$$

by adding and subtracting $\mathbb{E}[\bar{W}^*_{\theta_{n+1}}(\tilde{X}^{\varpi,\theta^*}_{\tau_{n+1}}) | \mathcal{F}_{\tau_n}] = \mathbb{E}[\bar{W}^*_{\theta_{n+1}}(X^{\varpi,\theta^*}_{\tau_{n+1}}) | \mathcal{F}_{\tau_n}]$. (R_6) is a martingale term, which we can bound using concentration theory. Our lazy update-scheme ensures that $\theta_n \neq \theta_{n+1}$ only $\mathcal{O}(\log(N_T))$ times by time T, keeping (R_5) small.

It remains to show that the lazy update-scheme, does not degrade the learning of (R_4), which is controlled by improvements to Proposition 4.3 in Appendix C which yield

$$\sum_{n=1}^{N_T} \sup_{(\theta_1,\theta_2)\in\mathcal{C}_{k(n)}(\delta)^2} \left\| \mu_{\theta_1}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) - \mu_{\theta_2}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) \right\| \lesssim \tilde{\mathcal{O}}(\sqrt{\mathrm{d_E}(T\varepsilon^{-1}) \log(\mathscr{N}^{\varepsilon}_{T\varepsilon^{-1}}) T}) .$$

□

5 Conclusion

In this work we proposed a general framework for the Reinforcement Learning problem of controlling an unknown dynamical system, on a continuous state-action space, to maximise the long-term average reward along a single trajectory. In particular, we focused on the understudied high-frequency systems driven by many small movements. Modelling such systems as controlled jump processes, we provided an optimistic algorithm which leverages Non-Linear Least Squares for learning and the diffusive limit regime for approximate planning. This proof of concept calls for several further refinements to be implementable in practice.

Optimism. The optimistic step of Algorithm 1 chooses $\tilde{\theta}_n$ in an inefficient manner. Like in UCRL2 [20], optimistic exploration can be performed at the same time as planning by solving an expanded HJB equation, i.e. (18) with the maximum now taken over $(a, \theta) \in \mathbb{A} \times \Theta$. Since our assumptions are uniform in θ, this is possible up to a modified regret decomposition, as in [20].

Lazy updates. The way we quantify learning progress to design the lazy update-scheme (7) remains fundamentally discrete. Computationally cheaper lazy update-schemes might be obtained through simpler heuristics. For instance, the scaling of the drift with ε suggests it could be possible to update periodically, directly in terms of the wall-clock time T.

Case-by-case. As a proof of concept, we endeavoured to study the RL problem in high generality. However, practical applications must use all available model information to refine the method ad-hoc. This is true for the learning method (replace NLLS with a fit specialised to the model at hand and bound the eluder dimension and log-covering numbers), and for numerical schemes on the PDE (18) which are built on a case-by-case basis for $d > 1$, see [25].

References

[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 1–26, Budapest, Hungary, 09–11 Jun 2011. PMLR. URL `https://proceedings.mlr.press/v19/abbasi-yadkori11a.html`.

[2] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL `https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf`.

[3] Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in linear-quadratic regulators via Lagrangian relaxation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 23–31. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/abeille20a.html`.

[4] Marc Abeille, Bruno Bouchard, and Lorenzo Croissant. Diffusive limit approximation of pure jump optimal ergodic control problems, September 2022. URL `http://arxiv.org/abs/2209.15284`. arXiv:2209.15284 [math].

[5] Ari Arapostathis, Vivek S. Borkar, and Mrinal K. Ghosh. *Ergodic control of diffusion processes*. Number 143 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, 2012. ISBN 9781139003605.

[6] Mariko Arisawa and Pierre-Louis Lions. On ergodic stochastic control. *Communications in partial differential equations*, 23(11-12):2187–2217, 1998.

[7] Peter Auer and Ronald Ortner. Logarithmic Online Regret Bounds for Undiscounted Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL `https://proceedings.neurips.cc/paper/2006/hash/c1b70d965ca504aa751ddb62ad69c63f-Abstract.html`.

[8] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

[9] Guy Barles and Panagiotis E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic analysis*, 4(3):271–283, 1991.

[10] Peter L. Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 35–42, Arlington, VA, 2009. AUAI Press. ISBN 9780974903958.

[11] Dimitri P Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, Belmont, MA, 3rd edition, 2011. ISBN 1886529442.

[12] J. Frédéric Bonnans and Housnaa Zidani. Consistency of Generalized Finite Difference Schemes for the Stochastic HJB Equation. *SIAM Journal on Numerical Analysis*, 41(3):1008–1021, January 2003. URL `https://epubs.siam.org/doi/abs/10.1137/S0036142901387336`. Publisher: Society for Industrial and Applied Mathematics.

[13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press, Oxford, 1st edition, 2013. ISBN 9780199535255.

[14] V. V. Buldygin and I. V. Kozachenko. *Metric characterization of random variables and random processes*, volume 188 of *Translations of mathematical monographs*. American Mathematical Society, Providence, RI, 2000. ISBN 9780821805336.

[15] Hong Chen and David D Yao. *Fundamentals of queueing networks: Performance, asymptotics, and optimization*, volume 46 of *Stochastic Modelling and Applied Probability*. Springer, New York, NY, 2001. ISBN 9781441928962.

[16] Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1300–1309. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/cohen19b.html`.

[17] Rama Cont and Peter Tankov. *Financial modelling with jump processes*. Chapman & Hall/CRC financial mathematics series. Chapman & Hall/CRC, Boca Raton, FL, 2004. ISBN 978-1-58488-413-2.

[18] Joaquin Fernandez-Tapia, Olivier Guéant, and Jean-Michel Lasry. Optimal Real-Time Bidding Strategies. *Applied Mathematics Research eXpress*, September 2016. ISSN 1687-1200, 1687-1197. URL `https://academic.oup.com/amrx/article-lookup/doi/10.1093/amrx/abw007`.

[19] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*, volume 224 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2nd edition, 1983. ISBN 354013025-X.

[20] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010. ISSN 1533-7928.

[21] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[22] Rafail Khasminskii. *Stochastic Stability of Differential Equations*, volume 66 of *Stochastic Modelling and Applied Probability*. Springer, Berlin, Heidelberg, 2012. ISBN 9783642232794.

[23] Peter Knabner and Lutz Angermann. *Numerical methods for elliptic and parabolic partial differential equations*. Number 44 in Texts in applied mathematics. Springer, New York, NY, 2003. ISBN 9780387954493.

[24] Harold J. Kushner. *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, volume 129 of *Mathematics in Science and Engineering*. Elsevier, 1977. ISBN 9780124301405.

[25] Harold J. Kushner and Paul Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time*, volume 24 of *Stochastic Modelling and Applied Probability*. Springer New York, New York, NY, 2001. ISBN 978-1-4612-6531-3 978-1-4613-0007-6.

[26] Olga A. Ladyzhenskaya and Nina N. Ural'tseva. *Linear and quasilinear elliptic equations*, volume 46 of *Mathematics in Science and Engineering*. Elsevier, 1968. ISBN 9780124328501.

[27] Peter Lancaster and Leiba Rodman. *Algebraic riccati equations*. Clarendon press, 1995. ISBN 9780198537953.

[28] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. ISBN 9781108571401.

[29] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*, volume 23 of *Classics in Mathematics*. Springer, Berlin, Heidelberg, 1991. ISBN 9783642202117.

[30] Pierre-Louis Lions. Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations part 2: viscosity solutions and uniqueness. *Communications in partial differential equations*, 8(11):1229–1276, 1983.

[31] Anna A. Obizhaeva and Jiang Wang. Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*, 16(1):1–32, 2013.

[32] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1*, pages 1466–1474, 2014. URL `https://proceedings.neurips.cc/paper/2014/hash/1141938ba2c2b13f5505d7c424ebae5f-Abstract.html`.

[33] Martin L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, NJ, 2005. ISBN 9780471727828.

[34] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 2256–2264, 2013. URL `https://proceedings.neurips.cc/paper/2013/hash/41bfd20a38bb1b0bec75acf0845530a7-Abstract.html`.

[35] Csaba Szepesvári. *Algorithms for reinforcement learning*. Number 4 in Synthesis lectures on artificial intelligence and machine learning. Morgan & Claypool Publishers, 2010.

Appendices

A Preliminaries

A.1 Organisation of Appendices

We prove the results one by one, starting with stability, then learning, planning, and finally concluding with the regret proof of Theorem 3.1.

In Appendix B, we go over the probabilistic properties of our problem and show several bounds on the stability of the process, in the sense of high-probability and moment boundedness. In particular the main objective of this appendix is to prove Proposition 4.1.

In Appendix C, we show a generalisation of the existing theory of learning with NLLS to the case of unbounded functions on unbounded domains. The key results are Propositions 4.2 and 4.3

In Appendix D, we provide a characterisation of the control part of the RL problem we analyse, including the diffusion limit approximation, namely Propositions 4.4 to 4.6.

In Appendix E, we perform regret analysis and collect the last few results used to prove the regret bound of Theorem 3.1. This includes treatement of the lazy update-scheme.

The remainder of Appendix A is devoted to notations and short-hands used throughout, but each appendix is meant to be as notationally stand-alone as possible.

A.2 General notation

The set of natural numbers including 0 is denoted $\mathbb{N}$, while $\mathbb{N}^* := \mathbb{N} \setminus \{0\}$ denotes the set of (strictly) positive integers. For $n \in \mathbb{N}^*$, we use $[n]$ to denote the set of positive integers up to and including n, i.e. $[n] := \{1, \dots, n\}$. Let $\mathbb{R}$ denote the set of real numbers and define $\mathbb{R}_+ := (0, +\infty)$ and $\mathbb{R}^d_* := \mathbb{R}^d \setminus \{0\}$. The space of sequences taking values in S will be denoted by $S^{\mathbb{N}}$. For $S \subset \mathbb{R}^d$, we also denote the complement of S by $S^{\mathrm{c}} := \mathbb{R}^d \setminus S$, we use the same notation for the complement of a probability event.

We denote by $\langle \cdot | \cdot \rangle$ the inner product on $\mathbb{R}^d$, by $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^d$, and by $\|\cdot\|_{\mathrm{op}}$ the associated operator norm on $\mathbb{R}^{d \times d}$. For $R \in \mathbb{R}_+$ and $x \in \mathbb{R}^d$, we denote the Euclidean ball of radius R centred at x by $\mathcal{B}_2(x, R)$, and when $x = 0$ we use the shorthand $\mathcal{B}_2(R)$ for $\mathcal{B}_2(0, R)$.

For $d \geq 1$, $\mathcal{D} \subset \mathbb{R}^d$ and $\mathcal{D}' \subset \mathbb{R}$, we denote the space of continuous functions from $\mathcal{D}$ to $\mathcal{D}'$ by $\mathcal{C}^0(\mathcal{D}; \mathcal{D}')$. For any $k \in \mathbb{N}^*$, we denote $\mathcal{C}^k(\mathcal{D}; \mathcal{D}')$ the subset of $\mathcal{C}^0(\mathcal{D}; \mathcal{D}')$ containing all functions which are continuously differentiable up to order k.

A.3 Problem dependent notation

The space of *càdlàg* (rcll) functions from $[0, +\infty)$ to $\mathbb{R}^d$, for $d \in \mathbb{N}^*$, is denoted $\mathbb{D}$ and $\mathbb{P}$ is a probability measure on $\Omega := \mathbb{D}$. $(N_t)_{t \in \mathbb{R}_+}$ denotes a marked $\mathbb{P}$-compound Poisson process of intensity $\varepsilon^{-1} > 1$, $(\tau_n)_{n \in \mathbb{N}}$ denotes the sequence of its arrival times, with $\tau_0 := 0$, and $(\xi_n)_{n \in \mathbb{N}}$ denotes the sequence of its marks. Namely, the sequences $(\tau_n)_{n \in \mathbb{N}}$ and $(\xi_n)_{n \in \mathbb{N}}$ are independent, $(\tau_{n+1} - \tau_n)_{n \in \mathbb{N}}$ is i.i.d. with exponential distribution of parameter ε and $(\xi_n)_{n \in \mathbb{N}}$ is i.i.d. with standard Gaussian measure on $\mathbb{R}^d$, which we denoted by ν.

For $t \in [0, +\infty)$, $\mathcal{F}_t := \sigma((\tau_n, \xi_n)_{\tau_n \leq t})$ and the filtration $\mathbb{F}$ is the completion of $(\mathcal{F}_t)_{t \in \mathbb{R}_+}$. The set of $\mathbb{F}$-adapted $\mathbb{A}$-valued processes, which we consider as admissible controls, is denoted $\mathcal{A}$. For any $(x_0, \alpha, \theta) \in \mathbb{R}^d \times \mathcal{A} \times \Theta$, $X^{\alpha,\theta}$ is the solution of

$$\begin{cases} X^{\alpha,\theta}_{\tau_n} = X^{\alpha,\theta}_{\tau_{n-1}} + \mu_\theta(X^{\alpha,\theta}_{\tau_{n-1}}, \alpha_{\tau_{n-1}}) + \Sigma \xi_n \\ X^{\alpha,\theta}_{\tau_0} = x_0 \end{cases} . \tag{21}$$

When specifying the dependence on the initial condition $x_0 \in \mathbb{R}^d$ is necessary, we write $X^{x_0,\alpha,\theta}$. This process is defined for any $t \in [0, +\infty)$ by considering its trajectories as piece-wise constant on any interval of the form $[\tau_{n-1}, \tau_n)$ for $n \in \mathbb{N}^*$. For any $(x_0, \alpha, \theta) \in \mathbb{R}^d \times \mathcal{A} \times \Theta$, the control problem is denoted by

$$\rho_\theta^*(x_0) := \sup_{\alpha\in\mathcal{A}} \rho_\theta^\alpha(x_0) \text{ in which } \rho_\theta^\alpha(x_0) := \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{n=1}^{N_T} r(X_{\tau_n}^{x_0,\alpha,\theta}, \alpha_{\tau_n})\right].$$

We denote by W a $\mathbb{P}$-Wiener process (a.k.a Brownian motion), by $\bar{\mathbb{F}}$ the $\mathbb{P}$-augmentation of the filtration it generates, and by $\bar{\mathcal{A}}$ the collection of $\mathbb{A}$-valued and $\bar{\mathbb{F}}$-predictable processes. For any $(x_0, \bar{\alpha}, \theta) \in \mathbb{R}^d \times \bar{\mathcal{A}} \times \Theta$, we denote by $\bar{X}^{\bar{\alpha},\theta}$ (or $\bar{X}^{x_0,\bar{\alpha},\theta}$ if specifying the initial condition) the solution of

$$\begin{cases} \mathrm{d}\bar{X}_t^{\bar{\alpha},\theta} = \bar{\mu}_\theta(\bar{X}_t^{\bar{\alpha},\theta}, \bar{\alpha}_t)\mathrm{d}t + \bar{\Sigma}\mathrm{d}W_t \\ \bar{X}_0^{\bar{\alpha},\theta} = x_0 \end{cases}. \tag{22}$$

The associated control problem is denoted by

$$\bar{\rho}_\theta^*(x_0) := \sup_{\bar{\alpha}\in\bar{\mathcal{A}}} \bar{\rho}_\theta^{\bar{\alpha}}(x_0) \text{ in which } \bar{\rho}_\theta^{\bar{\alpha}}(x_0) := \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T r(\bar{X}_t^{x_0,\bar{\alpha},\theta}, \bar{\alpha}_t)\mathrm{d}t\right].$$

According to Propositions 4.4 and 4.5, we defined the constants $\rho_\theta^* := \rho_\theta^*(0)$ and $\bar{\rho}_\theta^* := \bar{\rho}_\theta^*(0)$. For $\theta \in \Theta$, $\bar{\pi}_\theta^*$ denotes a policy in $\mathscr{A}$ (the set of measurable maps from $\mathbb{R}^d$ to $\mathbb{A}$) which maximises the right-hand side of the HJB equation (17) associated to $\bar{\rho}_\theta^*$ (see Proposition 4.5). Throughout, we use the same notation for policies and the Markov controls they induce, provided there is no ambiguity.

We use ϖ to denote the control process output of Algorithm 1 mathematically. For any $\omega \in \Omega$, the trajectory generated by Algorithm 1 is therefore defined as in (21) by $X^{\varpi,\theta^*}_{\cdot}(\omega)$. By definition of Algorithm 1, in its k^{th} episode (i.e. for $t \in [\tau_{n_k}, \tau_{n_k+1})$), $\varpi_t = \pi_k(X_t^{\varpi,\theta^*})$, with $\pi_k := \bar{\pi}^*_{\tilde{\theta}_k}$.

Throughout these appendices, we will use the shorthand $\psi_\theta^\varepsilon(x,a) := x + \varepsilon\bar{\mu}_\theta(x,a)$, for any $(x,a,\theta) \in \mathbb{R}^d \times \mathbb{A} \times \theta$.

B State Process Stability

A key aspect of our setting is that both the state process $X^{\alpha,\theta}$, for any $(\alpha,\theta)\in\mathcal{A}\times\Theta$, and the drift μ itself are unbounded. This can lead to an exponential blow-up of the state process, which can be harmful to both the learning and control aspects. In order to avoid this difficulty we imposed Assumption 2, which corresponds to a stochastic Lyapunov condition, and ensures that the state will not explode in expectation. We reinforce this result by leveraging concentration theory to obtain the high-probability bound of Proposition 4.1. Appendix B.1 is dedicated to its proof, and it will be used in the proofs of learning results and high-probability regret bounds (Appendices C and E).

Proposition 4.1. *Under Assumptions 1 and 2, there is a function $H_\delta(n)=\mathcal{O}\big(\sqrt{\log(n\delta^{-1})}\big)$ such that for any $\delta\in(0,1)$, $\alpha\in\mathcal{A}$, $x_0\in\mathbb{R}^d$, and $\theta\in\Theta$ we have*

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}_+}\frac{\|X_t^{\alpha,\theta}\|}{H_\delta(N_t)}\geq 1\right)\leq\delta\,. \tag{11}$$

Unlike learning and regret, the analysis of the control task is done in expectation via the HJB equation. Here the unbounded drift will materialise as higher moments of $X^{\alpha,\theta}$. The counterpart of Proposition 4.1 in this case is a moment result, given by Lemma B.5, which is proved in Appendix B.2 and will then be used in Appendix D.

Lemma B.5. *Under Assumptions 1 and 2, for any $p\geq 2$, there is a constant $\mathfrak{c}'_p>0$ independent of ε such that*

$$\mathbb{E}\left[\|X_t^{x_0,\alpha,\theta}\|^p\right]\leq\frac{1}{\ell_{\mathscr{V}}^p}\left(L_{\mathscr{V}}^p e^{-\frac{\mathfrak{c}_{\mathscr{V}}}{4}t}\|x_0\|^p+\frac{4\mathfrak{c}'_p}{\mathfrak{c}_{\mathscr{V}}}\left(1-e^{-\frac{\mathfrak{c}_{\mathscr{V}}}{4}t}\right)\right)\,,$$

for any $(x_0,\alpha,\theta)\in\mathbb{R}^d\times\mathcal{A}\times\Theta$ and $t\in[0,+\infty)$.

B.1 Proof of Proposition 4.1

This appendix is dedicated to the proof of Proposition 4.1 which is a high probability bound on the state process. This proof follows the Chernoff method. Thus, we will derive an exponential moment bound for the state process in Lemma B.2. We will first obtain a stochastic stability condition in expectation in Lemma B.1. In what follows, let $R_\varepsilon:=\sqrt{8d\log(1/\varepsilon)}$ and $\xi\sim\nu$.

Lemma B.1. *Under Assumptions 1 and 2,*

(i.) *for any $(\eta,x,a,\theta)\in\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{A}\times\Theta$, we have*

$$\mathscr{V}(\psi_\theta^\varepsilon(x,a)-\sqrt{\varepsilon}\eta)\leq(1-\varepsilon\mathfrak{c}_{\mathscr{V}})\mathscr{V}(x-\sqrt{\varepsilon}\eta)+\varepsilon M_{\mathscr{V}}L_0(1+\|\eta\|)\,; \tag{23}$$

(ii.) *and, for any $(a,\theta)\in\mathbb{A}\times\Theta$, and any $x\notin\mathcal{B}_2(\varepsilon^{\frac{1}{2}}\|\bar\Sigma\|_{\mathrm{op}}R_\varepsilon)$ we have*

$$\mathbb{E}[\mathscr{V}(\psi_\theta^\varepsilon(x,a)+\Sigma\xi)]\leq(1-\varepsilon\mathfrak{c}_{\mathscr{V}})\mathscr{V}(x)+\varepsilon\mathfrak{c}'_{\mathscr{V}}$$

in which $\mathfrak{c}'_{\mathscr{V}}$ is a constant independent of ε.

Proof.

(i.) By Lipschitzness of $\mathscr{V}$ and (4), for any $(\eta,x,a,\theta)\in\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{A}\times\Theta$, we have

$$\begin{aligned}\mathscr{V}(\psi_\theta^\varepsilon(x,a)-\sqrt{\varepsilon}\eta)&=\mathscr{V}(\psi_\theta^\varepsilon(x,a)-\psi_\theta^\varepsilon(\sqrt{\varepsilon}\eta,a)+\varepsilon\bar\mu(\sqrt{\varepsilon}\eta,a))\\&\leq\mathscr{V}(\psi_\theta^\varepsilon(x,a)-\psi_\theta^\varepsilon(\sqrt{\varepsilon}\eta,a))+M_{\mathscr{V}}\varepsilon\left\|\bar\mu(\sqrt{\varepsilon}\eta,a)\right\|\\&\leq(1-\varepsilon\mathfrak{c}_{\mathscr{V}})\mathscr{V}(x-\sqrt{\varepsilon}\eta)+M_{\mathscr{V}}\varepsilon\left\|\bar\mu(\sqrt{\varepsilon}\eta,a)\right\|\,,\end{aligned}$$

from which (23) follows by using Assumption 1, which implies $\|\bar\mu(\sqrt{\varepsilon}\eta,a)\|\leq L_0(1+\sqrt{\varepsilon}\|\eta\|)\leq L_0(1+\|\eta\|)$ since $\varepsilon\in(0,1)$.

(ii.) For any $x\in\mathbb{R}^d$, by the symmetry of the law of $\bar\Sigma\xi$, by (23) applied for $\eta=\bar\Sigma\xi$, and by taking the expectation, we have

$$\begin{aligned}\mathbb{E}[\mathscr{V}(\psi_\theta^\varepsilon(x,a)+\Sigma\xi)]&=\mathbb{E}[\mathscr{V}(\psi_\theta^\varepsilon(x,a)-\sqrt{\varepsilon}\bar\Sigma\xi)]\\&\leq(1-\varepsilon\mathfrak{c}_{\mathscr{V}})\mathbb{E}[\mathscr{V}(x-\sqrt{\varepsilon}\bar\Sigma\xi)]+\varepsilon M_{\mathscr{V}}L_0(1+\|\bar\Sigma\|_{\mathrm{op}}\mathbb{E}[\|\xi\|])\,.\end{aligned} \tag{24}$$

Since ξ is a standard Gaussian, $\|\xi\|^2$ is a random variable following a χ^2 distribution with d degrees of freedom, thus $\mathbb{E}[\|\xi\|^2] = d$, and by Jensen's inequality $\mathbb{E}[\|\xi\|] \leq \sqrt{d}$. Thus the second term is bounded by $\varepsilon M_{\mathscr{V}} L_0(1 + \|\bar{\Sigma}\|_{\mathrm{op}} \sqrt{d})$.

We now focus on bounding $\mathbb{E}[\mathscr{V}(x - \Sigma\xi)]$. We would like to use a Taylor expansion, but care needs to be taken to handle the non-differentiability of $\mathscr{V}$ at 0. Under the expectation, we distinguish two events: the event on which $\|\xi\| < R_\varepsilon$, which supports the main mass of ν, and the event on which $\|\xi\| \geq R_\varepsilon$, corresponding to the tails.

(a) For the first event we consider (on which $\|\xi\| < R_\varepsilon$), for any $x \notin \mathcal{B}_2(\|\Sigma\|_{\mathrm{op}} R_\varepsilon)$, we must have $0 \notin \mathcal{B}_2(x, \|\Sigma\xi\|)$, and thus $0 \notin (x + \Delta\Sigma\xi)_{\Delta\in[0,1]}$. Since this line segment doesn't contain 0 (the only point at which $\mathscr{V}$ is not continuously differentiable), we can perform a second-order Taylor expansion of $\mathscr{V}$ to obtain

$$\begin{aligned}
&\mathbb{E}[\mathscr{V}(x{+}\Sigma\xi)\mathbb{1}_{\{\|\xi\|<R_\varepsilon\}}] \\
&\qquad \leq \mathbb{E}\left[\left(\mathscr{V}(x) + \xi^\top\Sigma^\top\nabla\mathscr{V}(x) + \frac{1}{2}\operatorname{Tr}[\Sigma\xi\xi^\top\Sigma^\top\nabla^2\mathscr{V}(\hat{x})]\right)\mathbb{1}_{\{\|\xi\|<R_\varepsilon\}}\right]
\end{aligned}$$

for some $\hat{x} \in (x + \Delta\Sigma\xi)_{\Delta\in[0,1]}$. By the Cauchy-Schwartz inequality and the derivative bounds of Assumption 2, we obtain

$$\begin{aligned}
\mathbb{E}[\mathscr{V}(x+\Sigma\xi)\mathbb{1}_{\{\|\xi\|_2<R_\varepsilon\}}] &\leq \mathscr{V}(x) + \mathbb{E}[\xi^\top\mathbb{1}_{\{\|\xi\|<R_\varepsilon\}}]\Sigma^\top\nabla\mathscr{V}(x) + \frac{\varepsilon}{2}M'_{\mathscr{V}}\|\bar{\Sigma}\|_{\mathrm{op}}^2 \\
&\leq \mathscr{V}(x) + \frac{\varepsilon}{2}M'_{\mathscr{V}}\|\bar{\Sigma}\|_{\mathrm{op}}^2\,,
\end{aligned}$$

since $\mathbb{E}[\xi^\top\mathbb{1}_{\{\|\xi\|<R_\varepsilon\}}] = 0$ by the rotational invariance property of a truncated Gaussian.

(b) On the second event (on which $\|\xi\| \geq R_\varepsilon$), we cannot use a Taylor expansion. Instead, we use the Lipschitzness of $\mathscr{V}$ followed by the Cauchy-Schwartz inequality, and then apply a sub-Gaussian concentration inequality (see e.g. [29, (3.5)]):

$$\begin{aligned}
\mathbb{E}[\mathscr{V}(x+\Sigma\xi)\mathbb{1}_{\{\|\xi\|\geq R_\varepsilon\}}] &\leq \mathscr{V}(x) + M_{\mathscr{V}}\|\Sigma\|_{\mathrm{op}}\mathbb{E}[\|\xi\|\,\mathbb{1}_{\{\|\xi\|\geq R_\varepsilon\}}] \\
&\leq \mathscr{V}(x) + M_{\mathscr{V}}\|\Sigma\|_{\mathrm{op}}\sqrt{\mathbb{E}[\|\xi\|^2]\mathbb{P}(\|\xi\|\geq R_\varepsilon)} \\
&\leq \mathscr{V}(x) + M_{\mathscr{V}}\|\Sigma\|_{\mathrm{op}}\sqrt{4de^{-\frac{R_\varepsilon^2}{8d}}} \\
&\leq \mathscr{V}(x) + 2\varepsilon M_{\mathscr{V}}\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{d}\,.
\end{aligned}$$

To complete the proof, we combine both cases in (24), and let

$$\mathfrak{c}'_{\mathscr{V}} := M_{\mathscr{V}}L_0(1+\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{d}) + 2M_{\mathscr{V}}\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{d} + \frac{M'_{\mathscr{V}}}{2}\|\bar{\Sigma}\|_{\mathrm{op}}^2.$$

□

Lemma B.2. *Under Assumptions 1 and 2, for any $(x_0, \alpha, \theta) \in \mathbb{R}^d \times \mathcal{A} \times \Theta$ and any $\lambda \in \mathbb{R}_+$, we have*

$$\mathbb{E}[e^{\lambda\mathscr{V}(X_{\tau_n}^{x_0,\alpha,\theta})}] \leq (n+1)\exp\left(\lambda\left(\frac{\mathfrak{c}'_{\mathscr{V}}}{\mathfrak{c}_{\mathscr{V}}} + L_{\mathscr{V}}(\varepsilon^{\frac{1}{2}}\|\bar{\Sigma}\|_{\mathrm{op}}R_\varepsilon + \|x_0\|)\right) + \frac{\lambda^2 M_{\mathscr{V}}^2\|\bar{\Sigma}\|_{\mathrm{op}}^2}{2\mathfrak{c}_{\mathscr{V}}}\right),$$

for any $n \in \mathbb{N}$.

Proof. For $n \in \mathbb{N}^*$, let us define the following events for $i < n$: $E_{i,n-1} := \{i = \sup\{j \in \{0,\ldots,n-1\} : \|X_{\tau_j}^{\alpha,\theta}\| \leq \|\Sigma\|_{\mathrm{op}}R_\varepsilon\}\}$ and $\bar{E}_{n-1} := \{\min_{j\in\{0,\ldots,n-1\}}\|X_{\tau_j}^{\alpha,\theta}\| > \|\Sigma\|_{\mathrm{op}}R_\varepsilon\}$. Note that both these events are $\mathcal{F}_{\tau_{n-1}}$-measurable and that $\cup_{i\leq n-1}E_{i,n-1} = \bar{E}_{n-1}^c$, so that $\{\bar{E}_{n-1}, E_{0,n-1},\ldots,E_{n-1,n-1}\}$ induces a partition of Ω. We begin by working conditionally to each of these events, and in a second part we will collect them to bound $\mathbb{E}[\exp(\lambda\mathscr{V}(X_{\tau_n}^{\alpha,\theta})]$.

For any $0 \leq i < n$, by the tower rule and by adding and subtracting $\mathbb{E}[\exp\left(\mathbb{E}[\lambda\mathscr{V}(X^{\alpha,\theta}_{\tau_n})|\mathcal{F}_{\tau_{n-1}}]\right)\mathbb{1}_{E_{i,n-1}}]$, we have

$$\begin{aligned}
\mathbb{E}[e^{\lambda\mathscr{V}(X^{\alpha,\theta}_{\tau_n})}\mathbb{1}_{E_{i,n-1}}] &= \mathbb{E}[\mathbb{E}[e^{\lambda\mathscr{V}(X^{\alpha,\theta}_{\tau_n})}|\mathcal{F}_{\tau_{n-1}}]\mathbb{1}_{E_{i,n-1}}] \\
&= \mathbb{E}\Bigg[\exp\left(\mathbb{E}[\lambda\mathscr{V}(X^{\alpha,\theta}_{\tau_n})|\mathcal{F}_{\tau_{n-1}}]\right)\mathbb{1}_{E_{i,n-1}} \\
&\qquad \times \mathbb{E}\left[\exp\left(\lambda\mathscr{V}(X^{\alpha,\theta}_{\tau_n}) - \mathbb{E}[\lambda\mathscr{V}(X^{\alpha,\theta}_{\tau_n})|\mathcal{F}_{\tau_{n-1}}]\right)|\mathcal{F}_{\tau_{n-1}}\right]\Bigg].
\end{aligned}$$

Using a result for Lipschitz functions of Gaussian random variables (see e.g. [13, Thm 5.5]) applied to $\mathscr{V}$ and ξ, we obtain

$$\begin{aligned}
\mathbb{E}[e^{\lambda\mathscr{V}(X^{\alpha,\theta}_{\tau_n})}\mathbb{1}_{E_{i,n-1}}] &\leq e^{\frac{\lambda^2}{2}M^2_{\mathscr{V}}\|\Sigma\|^2_{\mathrm{op}}}\mathbb{E}\Bigg[\exp\left(\mathbb{E}[\lambda\mathscr{V}(X^{\alpha,\theta}_{\tau_n})|\mathcal{F}_{\tau_{n-1}}]\right)\mathbb{1}_{E_{i,n-1}}\Bigg] \\
&= e^{\frac{\lambda^2}{2}M^2_{\mathscr{V}}\|\Sigma\|^2_{\mathrm{op}}}\mathbb{E}\Bigg[\exp\Big(\mathbb{E}[\lambda\mathscr{V}(\psi^{\varepsilon}_{\theta}(X^{\alpha,\theta}_{\tau_{n-1}},\alpha_{\tau_{n-1}})+\Sigma\xi_n)|\mathcal{F}_{\tau_{n-1}}]\Big)\mathbb{1}_{E_{i,n-1}}\Bigg].
\end{aligned} \tag{25}$$

If $i = n-1$, $\|X^{\alpha,\theta}_{\tau_{n-1}}\| \leq \|\Sigma\|_{\mathrm{op}}R_\varepsilon$ on the event $E_{i,n-1}$, and thus we have

$$\begin{aligned}
\mathbb{E}\left[\lambda\mathscr{V}(\psi^{\varepsilon}_{\theta}(X^{\alpha,\theta}_{\tau_{n-1}},\alpha_{\tau_{n-1}})+\Sigma\xi_n)\big|\mathcal{F}_{\tau_{n-1}}\right] &\leq \mathbb{E}\left[\lambda L_{\mathscr{V}}\left\|X^{\alpha,\theta}_{\tau_{n-1}}+\mu(X^{\alpha,\theta}_{\tau_{n-1}},\alpha_{\tau_{n-1}})+\Sigma\xi\right\|\big|\mathcal{F}_{\tau_{n-1}}\right] \\
&\leq \lambda L_{\mathscr{V}}\left((1+L_0)\|\Sigma\|_{\mathrm{op}}R_\varepsilon + 1 + \|\Sigma\|_{\mathrm{op}}\sqrt{d}\right)
\end{aligned}$$

by using the fact that $\mathbb{E}[\|\xi\|] \leq \sqrt{\mathbb{E}[\|\xi\|^2]} = \sqrt{d}$, as $\xi \sim \nu$. Noticing that $\sup_{\varepsilon\in(0,1)}\varepsilon^{\frac{1}{2}}R_\varepsilon = \sqrt{8de^{-1}}$, let us introduce

$$C_H := L_{\mathscr{V}}\left((1+L_0)\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{8de^{-1}} + 1 + \|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{d}\right). \tag{26}$$

Combining this with (25) yields

$$\mathbb{E}[e^{\lambda\mathscr{V}(X^{\alpha,\theta}_{\tau_n})}\mathbb{1}_{E_{i,n-1}}] \leq \exp\left(\frac{\lambda^2}{2}M^2_{\mathscr{V}}\|\Sigma\|^2_{\mathrm{op}} + \lambda C_H\right), \tag{27}$$

in the case $i = n-1$.

If $i < n-1$, we can apply the same methodology, and continuing from (25) apply Lemma B.1 to obtain

$$\begin{aligned}
\mathbb{E}[e^{\lambda\mathscr{V}(X^{\alpha,\theta}_{\tau_n})}\mathbb{1}_{E_{i,n-1}}] &\leq e^{\frac{\lambda^2}{2}M^2_{\mathscr{V}}\|\Sigma\|^2_{\mathrm{op}}}\mathbb{E}\Bigg[\exp\left(\mathbb{E}\left[\lambda\mathscr{V}(\psi^{\varepsilon}_{\theta}(X^{\alpha,\theta}_{\tau_{n-1}},\alpha_{\tau_{n-1}})+\Sigma\xi_n)|\mathcal{F}_{\tau_{n-1}}\right]\right) \\
&\qquad \times \mathbb{1}_{\{X^{\alpha,\theta}_{\tau_{n-1}} > \|\Sigma\|_{\mathrm{op}}R_\varepsilon\}}\mathbb{1}_{E_{i,n-2}}\Bigg], \\
&\leq e^{\frac{\lambda^2}{2}M^2_{\mathscr{V}}\|\Sigma\|^2_{\mathrm{op}}+\lambda\varepsilon\mathfrak{c}'_{\mathscr{V}}}\mathbb{E}[\exp((1-\varepsilon\mathfrak{c}_{\mathscr{V}})\lambda\mathscr{V}(X^{\alpha,\theta}_{\tau_{n-1}}))\mathbb{1}_{E_{i,n-2}}].
\end{aligned} \tag{28}$$

It remains to use an induction argument in n down to $n = i+1$ and use the fact that $\|X^{\alpha,\theta}_{\tau_i}\| \leq \|\Sigma\|_{\mathrm{op}}R_\varepsilon$ on $E_{i,i}$, to obtain

$$\begin{aligned}
&\mathbb{E}[e^{\lambda\mathscr{V}(X^{\alpha,\theta}_{\tau_n})}\mathbb{1}_{E_{i,n-1}}] \\
&\quad \leq \exp\left(\lambda C_H + \lambda\varepsilon\mathfrak{c}'_{\mathscr{V}}\sum_{k=0}^{n-1-i}(1-\varepsilon\mathfrak{c}_{\mathscr{V}})^k + \frac{\lambda^2 M^2_{\mathscr{V}}\|\Sigma\|^2_{\mathrm{op}}}{2}\sum_{k=0}^{n-1-i}(1-\varepsilon\mathfrak{c}_{\mathscr{V}})^{2k}\right) \\
&\quad \leq \exp\left(\lambda C_H + \lambda\frac{\mathfrak{c}'_{\mathscr{V}}}{\mathfrak{c}_{\mathscr{V}}} + \frac{\lambda^2 M^2_{\mathscr{V}}\|\bar{\Sigma}\|^2_{\mathrm{op}}}{2\mathfrak{c}_{\mathscr{V}}}\right).
\end{aligned} \tag{29}$$

On the event $\bar{E}_{n-1}$, that is if the process is never in the ball $\mathcal{B}_2(\|\Sigma\|_{\mathrm{op}}R_\varepsilon)$ before time τ_n, we use the fact that (28) is valid with $\bar{E}_{n-1}$ and $\bar{E}_{n-2}$ in place of $E_{i,n-1}$ and $E_{i,n-2}$. Applying the induction, we obtain

$$\mathbb{E}[e^{\lambda\mathscr{V}(X^{\alpha,\theta}_{\tau_n})}\mathbb{1}_{\bar{E}_{n-1}}] \leq \exp\left(\lambda L_{\mathscr{V}}\|x_0\| + \lambda\frac{\mathfrak{c}'_{\mathscr{V}}}{\mathfrak{c}_{\mathscr{V}}} + \frac{\lambda^2 M^2_{\mathscr{V}}\|\bar{\Sigma}\|^2_{\mathrm{op}}}{2\mathfrak{c}_{\mathscr{V}}}\right). \tag{30}$$

Using our partition and combining (27), (29), and (30) we can thus write, for any $n \in \mathbb{N}$

$$\mathbb{E}\left[e^{\lambda \mathscr{V}(X^{\alpha,\theta}_{\tau_n})}\right] \leq \mathbb{E}\left[e^{\lambda \mathscr{V}(X^{\alpha,\theta}_{\tau_n})}\left(\mathbb{1}_{\bar{E}_{n-1}} + \sum_{i=0}^{n-1} \mathbb{1}_{E_{i,n-1}}\right)\right]$$
$$\leq (n+1)\exp\left(\lambda\left(\frac{\mathfrak{c}'_{\mathscr{V}}}{\mathfrak{c}_{\mathscr{V}}} + C_H + L_{\mathscr{V}}\|x_0\|\right) + \frac{\lambda^2 M_{\mathscr{V}}^2 \|\bar{\Sigma}\|_{\mathrm{op}}^2}{2\mathfrak{c}_{\mathscr{V}}}\right)$$

which concludes the proof. □

With these two lemmas, we can now prove Proposition 4.1, the main result of this section. First, let us give the exact definition of $H_\delta(n)$:

$$H_\delta(n) := \frac{1}{\ell_{\mathscr{V}}}\left(C_H + L_{\mathscr{V}}\|x_0\|\right) + \frac{\mathfrak{c}'_{\mathscr{V}}}{\ell_{\mathscr{V}}\mathfrak{c}_{\mathscr{V}}} + \frac{M_{\mathscr{V}}}{\ell_{\mathscr{V}}}\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{\frac{2}{\mathfrak{c}_{\mathscr{V}}}\log\left(\frac{\pi^2(n+1)^3}{6\delta}\right)} \tag{31}$$

in which C_H is defined in (26), so that $H_\delta(n) = \mathcal{O}\big(\sqrt{\log(n\delta^{-1})}\big)$.

Proposition 4.1. *Under Assumptions 1 and 2, there is a function $H_\delta(n) = \mathcal{O}\big(\sqrt{\log(n\delta^{-1})}\big)$ such that for any $\delta \in (0,1)$, $\alpha \in \mathcal{A}$, $x_0 \in \mathbb{R}^d$, and $\theta \in \Theta$ we have*

$$\mathbb{P}\left(\sup_{t \in \mathbb{R}_+} \frac{\|X_t^{\alpha,\theta}\|}{H_\delta(N_t)} \geq 1\right) \leq \delta\,. \tag{11}$$

Proof. Fix $n \in \mathbb{N}$, by Markov's inequality and Assumption 2, for any $u > 0$, we have

$$\mathbb{P}\left(\|X^{\alpha,\theta}_{\tau_n}\| > u\right) \leq \mathbb{E}\left[e^{\lambda \ell_{\mathscr{V}}\|X^{\alpha,\theta}_{\tau_n}\|}\right] e^{-\lambda \ell_{\mathscr{V}} u} \leq \mathbb{E}\left[e^{\lambda \mathscr{V}(X^{\alpha,\theta}_{\tau_n})}\right] e^{-\lambda \ell_{\mathscr{V}} u}\,,$$

which implies that

$$\mathbb{P}\left(\|X^{\alpha,\theta}_{\tau_n}\| - \frac{\mathfrak{c}'_{\mathscr{V}}}{\ell_{\mathscr{V}}\mathfrak{c}_{\mathscr{V}}} - \frac{C_H}{\ell_{\mathscr{V}}} - \frac{L_{\mathscr{V}}}{\ell_{\mathscr{V}}}\|x_0\| > u\right)$$
$$\leq \mathbb{E}\left[e^{\lambda \mathscr{V}(X^{\alpha,\theta}_{\tau_n})}\right]\exp\left(-\lambda \ell_{\mathscr{V}}\left(u + \frac{\mathfrak{c}'_{\mathscr{V}}}{\ell_{\mathscr{V}}\mathfrak{c}_{\mathscr{V}}} + \frac{C_H}{\ell_{\mathscr{V}}} + \frac{L_{\mathscr{V}}}{\ell_{\mathscr{V}}}\|x_0\|\right)\right).$$

Applying Lemma B.2, and taking $\lambda = \mathfrak{c}_{\mathscr{V}}\ell_{\mathscr{V}}u/(M_{\mathscr{V}}^2\|\bar{\Sigma}\|_{op}^2)$, we obtain

$$\mathbb{P}\left(\|X^{\alpha,\theta}_{\tau_n}\| > u + \frac{\mathfrak{c}'_{\mathscr{V}}}{\ell_{\mathscr{V}}\mathfrak{c}_{\mathscr{V}}} + \varepsilon^{\frac{1}{2}}\frac{L_{\mathscr{V}}}{\ell_{\mathscr{V}}}\|\bar{\Sigma}\|_{\mathrm{op}}R_\varepsilon + \frac{L_{\mathscr{V}}}{\ell_{\mathscr{V}}}\|x_0\|\right)$$
$$\leq (n+1)\exp\left(-\lambda \ell_{\mathscr{V}} u + \lambda^2\frac{M_{\mathscr{V}}^2\|\bar{\Sigma}\|_{\mathrm{op}}^2}{2\mathfrak{c}_{\mathscr{V}}}\right)$$
$$= (n+1)\exp\left(-\frac{\mathfrak{c}_{\mathscr{V}}\ell_{\mathscr{V}}^2}{2M_{\mathscr{V}}^2\|\bar{\Sigma}\|_{\mathrm{op}}^2}u^2\right).$$

Letting $u = M_{\mathscr{V}}\|\bar{\Sigma}\|_{\mathrm{op}}\ell_{\mathscr{V}}^{-1}\sqrt{2\mathfrak{c}_{\mathscr{V}}^{-1}\log((n+1)/\delta')}$, yields

$$\mathbb{P}\left(\|X^{\alpha,\theta}_{\tau_n}\| \geq \frac{C_H}{\ell_{\mathscr{V}}} + \frac{L_{\mathscr{V}}}{\ell_{\mathscr{V}}}\|x_0\| + \frac{\mathfrak{c}'_{\mathscr{V}}}{\ell_{\mathscr{V}}\mathfrak{c}_{\mathscr{V}}} + \frac{M_{\mathscr{V}}}{\ell_{\mathscr{V}}}\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{\frac{2}{\mathfrak{c}_{\mathscr{V}}}\log\left(\frac{n+1}{\delta'}\right)}\right) \leq \delta'.$$

Setting $\delta' = 6\delta/\pi^2(n+1)^2$, and taking a union bound over $n \in \mathbb{N}$ yields

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}_+}\frac{X_t^{\alpha,\theta}}{H_\delta(N_t)} \geq 1\right) = \mathbb{P}\left(\bigcup\nolimits_{n\in\mathbb{N}}\{\|X^{\alpha,\theta}_{\tau_n}\| \geq H_\delta(n)\}\right) \leq \delta\,,$$

which implies the result since $\delta \in (0,1)$ implies $\log(n^3/\delta) \leq \log(n^3/\delta^3) = 3\log(n/\delta)$. □

B.2 Expectation Bounds of Higher Orders

In this appendix, we will focus on higher moment conditions of the state process, which will be used in the control results of Appendix D. In Lemma B.3 and Corollary B.4 we work to raise the stochastic stability condition from Lemma B.1 to a power $p \geq 2$. Lemma B.5, the main result of this section, will follow from this by arguments of [4].

Lemma B.3. *Under Assumptions 1 and 2, for $p \geq 2$, there is a function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ and a constant $C_p > 0$ independent of ε satisfying*

$$g(x,\eta) \leq \varepsilon C_p \left(1 + \mathscr{V}(x - \sqrt{\varepsilon}\eta)^{p-1}\right)\left(1 + \|\eta\|^p\right) ,$$

for any $(\eta, x) \in \mathbb{R}^d \times \mathbb{R}^d$, such that

$$\mathscr{V}(\psi^{\varepsilon}_{\theta}(x,a) - \sqrt{\varepsilon}\eta)^p \leq (1 - \varepsilon\mathfrak{c}_{\mathscr{V}})\mathscr{V}(x - \sqrt{\varepsilon}\eta)^p + g(x,\eta) \,. \tag{32}$$

for any $(\eta, x, a, \theta) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{A} \times \Theta$.

Proof. We first raise both sides of (23) to the power p

$$\mathscr{V}(\psi^{\varepsilon}_{\theta}(x,a) - \sqrt{\varepsilon}\eta)^p \leq \Big((1 - \varepsilon\mathfrak{c}_{\mathscr{V}})\mathscr{V}(x - \sqrt{\varepsilon}\eta) + \varepsilon M_{\mathscr{V}} L_0(1 + \|\eta\|)\Big)^p .$$

We will now expand the right hand side. Let $a = (1 - \varepsilon\mathfrak{c}_{\mathscr{V}})\mathscr{V}(x - \sqrt{\varepsilon}\eta)$ and $b = \varepsilon M_{\mathscr{V}} L_0(1 + \|\eta\|)$, by the binomial theorem we have

$$\begin{aligned}(a+b)^p &= \sum_{k=0}^{p} \binom{p}{k} a^k b^{p-k} = a^p + b\sum_{k=0}^{p-1}\binom{p}{k} a^k b^{p-1-k} \\ &\leq a^p + b(1+b)^{p-1}(1+a)^{p-1}\sum_{k=0}^{p-1}\binom{p}{k} .\end{aligned}$$

Since $(1 - \varepsilon\mathfrak{c}_{\mathscr{V}}) \in (0,1)$, $\varepsilon \leq 1$, $b \leq 1 + b$, and $\sum_{k=0}^{p-1}\binom{p}{k} \leq 2^p$, by using the binomial identity $(1+a)^q \leq 2^{q-1}(1 + a^q)$ for $(a,q) \in [0,+\infty) \times [1,+\infty)$, we have

$$\begin{aligned}\mathscr{V}(\psi^{\varepsilon}_{\theta}(x,a) - \sqrt{\varepsilon}\eta)^p &\leq (1 - \varepsilon\mathfrak{c}_{\mathscr{V}})\mathscr{V}(x - \sqrt{\varepsilon}\eta)^p \\ &\quad + \varepsilon(1 + M_{\mathscr{V}} L_0(1 + \|\eta\|))^p(1 + \mathscr{V}(x - \sqrt{\varepsilon}\eta)^{p-1})2^{p-2+p} .\end{aligned} \tag{33}$$

Finally, we have

$$\begin{aligned}(1 + M_{\mathscr{V}} L_0(1 + \|\eta\|))^p &= (1 + M_{\mathscr{V}} L_0 + M_{\mathscr{V}} L_0 \|\eta\|)^p \\ &\leq (1 + M_{\mathscr{V}} L_0 + (1 + M_{\mathscr{V}} L_0)\|\eta\|)^p \\ &= (1 + M_{\mathscr{V}} L_0)^p(1 + \|\eta\|)^p \\ &\leq (1 + M_{\mathscr{V}} L_0)^p(1 + \|\eta\|^p)2^{p-1} .\end{aligned} \tag{34}$$

Combining (33) and (34), leads to the required result. □

Recall that $\xi \sim \nu$ is a centred standard Gaussian random variable.

Corollary B.4. *Under Assumptions 1 and 2, for any $p \geq 2$, there is a constant $\mathfrak{c}_p > 0$ independent of ε such that*

$$\mathbb{E}\left[\mathscr{V}(\psi^{\varepsilon}_{\theta}(x,a) + \Sigma\xi)^p\right] \leq \left(1 - \varepsilon\frac{\mathfrak{c}_{\mathscr{V}}}{2}\right)\mathbb{E}[\mathscr{V}(x - \sqrt{\varepsilon}\xi)^p] + \varepsilon\mathfrak{c}_p$$

for any $(x, a, \theta) \in \mathbb{R}^d \times \mathbb{A} \times \Theta$.

Proof.

i. Taking the expectation of the bound on g from Lemma B.3 and applying Hölder's inequality yields

$$\begin{aligned}\mathbb{E}[g(x,\xi)] &\le \varepsilon C_p \mathbb{E}\left[(1+\mathscr{V}(x-\sqrt{\varepsilon}\xi)^{p-1})(1+\|\xi\|^p)\right] \\ &\le \varepsilon C_p \mathbb{E}\left[(1+\mathscr{V}(x-\sqrt{\varepsilon}\xi)^{p-1})^{\frac{(p+1)}{p}}\right]^{\frac{p}{p+1}} \mathbb{E}\left[(1+\|\xi\|^p)^{p+1}\right]^{\frac{1}{p+1}} \\ &\le 4\varepsilon C_p \mathbb{E}\left[1+\mathscr{V}(x-\sqrt{\varepsilon}\xi)^{\frac{(p-1)(p+1)}{p}}\right] \mathbb{E}\left[(1+\|\xi\|^p)^{p+1}\right]^{\frac{1}{p+1}}\ ,\end{aligned}$$

by using the identities: for $(u,v)\in\mathbb{R}_+^2$, $(1+u)^{(p+1)/p}\le 4(1+u^{(p+1)/p})$ and $(1+v)^{p/(p+1)}\le 1+v$. Since ξ has bounded moments of any order,

$$C'_p := 4C_p\mathbb{E}\left[(1+\|\xi\|^p)^{p+1}\right]^{\frac{1}{p+1}}$$

is a finite constant and we have

$$\mathbb{E}\left[g(x,\xi)\right] \le \varepsilon C'_p \mathbb{E}\left[1+\mathscr{V}(x-\sqrt{\varepsilon}\xi)^{p-\frac{1}{p}}\right]\ .$$

ii. Recalling Lemma B.3, we have

$$\begin{aligned}\mathbb{E}\left[\mathscr{V}(\psi_\theta^\varepsilon(x,a)+\Sigma\xi)^p\right] &\le (1-\varepsilon\mathfrak{c}_{\mathscr{V}})\,\mathbb{E}[\mathscr{V}(x-\sqrt{\varepsilon}\xi)^p] + \mathbb{E}[g(x,\xi)] \\ &\le \left(1-\varepsilon\frac{\mathfrak{c}_{\mathscr{V}}}{2}\right)\mathbb{E}[\mathscr{V}(x-\sqrt{\varepsilon}\xi)^p] \\ &\quad + \varepsilon\mathbb{E}\left[C'_p(1+\mathscr{V}(x-\sqrt{\varepsilon}\xi)^{p-\frac{1}{p}}) - \frac{\mathfrak{c}_{\mathscr{V}}}{2}\mathscr{V}(x-\sqrt{\varepsilon}\xi)^p\right]\ . \qquad (35)\end{aligned}$$

iii. Note that, for any $p\ge 2$, the function

$$z\in\mathbb{R}^d \mapsto \frac{\|z\|^{p-\frac{1}{p}}}{1+\|z\|^p}\in\mathbb{R}_+$$

is bounded, so there exists a constant $C''_p>0$ such that, for any $z\in\mathbb{R}^d$,

$$C'_p\mathscr{V}(z)^{p-\frac{1}{p}} - \frac{\mathfrak{c}_{\mathscr{V}}}{2}\mathscr{V}(z)^p \le C''_p\,.$$

Applying this to the expectation in (35), we have

$$\mathbb{E}\left[\mathscr{V}(\psi_\theta^\varepsilon(x,a)+\Sigma\xi)^p\right] \le \left(1-\varepsilon\frac{\mathfrak{c}_{\mathscr{V}}}{2}\right)\mathbb{E}[\mathscr{V}(x+\sqrt{\varepsilon}\xi)^p] + \varepsilon(C''_p+C'_p)\,.$$

Letting $\mathfrak{c}_p := C'_p + C''_p$ completes the proof.

□

Lemma B.5. *Under Assumptions 1 and 2, for any $p\ge 2$, there is a constant $\mathfrak{c}'_p>0$ independent of ε such that*

$$\mathbb{E}\left[\|X_t^{x_0,\alpha,\theta}\|^p\right] \le \frac{1}{\ell_{\mathscr{V}}^p}\left(L_{\mathscr{V}}^p e^{-\frac{\mathfrak{c}_{\mathscr{V}}}{4}t}\|x_0\|^p + \frac{4\mathfrak{c}'_p}{\mathfrak{c}_{\mathscr{V}}}\left(1-e^{-\frac{\mathfrak{c}_{\mathscr{V}}}{4}t}\right)\right)\ ,$$

for any $(x_0,\alpha,\theta)\in\mathbb{R}^d\times\mathcal{A}\times\Theta$ and $t\in[0,+\infty)$.

Proof. Recall from Corollary B.4 that we have

$$\mathbb{E}[\mathscr{V}(\psi_\theta^\varepsilon(x,a)+\Sigma\xi)^p] \le \left(1-\varepsilon\frac{\mathfrak{c}_{\mathscr{V}}}{2}\right)\mathbb{E}[\mathscr{V}(x+\Sigma\xi)^p] + \varepsilon\mathfrak{c}_p \qquad (36)$$

for any $(x,a,\theta)\in\mathbb{R}^d\times\mathbb{A}\times\Theta$. We begin by eliminating the $\Sigma\xi$ from the right-hand side so that we have a proper Lyapunov contraction property on the generator. We expand $\mathscr{V}^p\in\mathcal{C}^2(\mathbb{R}^d;[0,+\infty))$ and use the fact that $\mathbb{E}[\xi]=0$ to obtain

$$\begin{aligned}\mathbb{E}[\mathscr{V}(x+\Sigma\xi)^p] &= \mathscr{V}(x)^p + \varepsilon p\mathbb{E}[\mathscr{V}(x+\Delta\Sigma\xi)^{p-1}\operatorname{Tr}[\xi\bar\Sigma\bar\Sigma^\top\xi^\top\nabla^2\mathscr{V}(x+\Delta\Sigma\xi)]] \\ &\quad + \varepsilon p(p-1)\mathbb{E}[\mathscr{V}(x+\Delta\Sigma\xi)^{p-2}\operatorname{Tr}[\xi\bar\Sigma\bar\Sigma^\top\xi^\top\nabla\mathscr{V}(x+\Delta\Sigma\xi)\nabla\mathscr{V}^\top(x+\Delta\Sigma\xi)]]\end{aligned}$$

for some random variable Δ taking value in $[0,1]$. This is now upper-bounded by using the Lipschitzness of $\mathscr{V}$ and the Cauchy-Schwartz inequality

$$\mathbb{E}[\mathscr{V}(x+\Sigma\xi)^p] \leq \mathscr{V}(x)^p + \varepsilon p M'_{\mathscr{V}} \|\bar{\Sigma}\|^2_{\mathrm{op}} \mathbb{E}[(\mathscr{V}(x) + M_{\mathscr{V}}\Delta\Sigma\,\|\xi\|)^{p-1}\,\|\xi\|^2] \\ + \varepsilon p(p-1)(M_{\mathscr{V}})^2 \|\bar{\Sigma}\|^2_{\mathrm{op}} \mathbb{E}[(\mathscr{V}(x) + M_{\mathscr{V}}\Delta\Sigma\,\|\xi\|)^{p-2}\,\|\xi\|^2]\,.$$

By the binomial theorem as in the proof of Lemma B.3, and as $|\Delta| \leq 1$, we have

$$\mathbb{E}[\mathscr{V}(x+\Sigma\xi)^p] \leq \mathscr{V}(x)^p + \varepsilon\Bigg(pM'_{\mathscr{V}}\|\bar{\Sigma}\|^2_{\mathrm{op}}\mathbb{E}\left[\|\xi\|^2 \sum_{k=0}^{p-1}\binom{p-1}{k}\mathscr{V}(x)^k (M_{\mathscr{V}}\,\|\Sigma\|_{\mathrm{op}}\,\|\xi\|)^{p-1-k}\right] \\ + p(p-1)(M_{\mathscr{V}}\|\bar{\Sigma}\|_{\mathrm{op}})^2\mathbb{E}\left[\sum_{k=0}^{p-2}\binom{p-2}{k}\mathscr{V}(x)^k(M_{\mathscr{V}}\,\|\Sigma\|_{\mathrm{op}}\,\|\xi\|)^{p-2-k}\right]\Bigg).$$

Since $\|\xi\|$ is a sub-Gaussian random variable it has moments of all orders, and we can express the interior of the bracket above as a polynomial in $\mathscr{V}(x)$ of order $p-1$ with finite coefficients $\{a_k\}_{k=0}^{p-1} \subset \mathbb{R}_+$. Recalling (36), we thus have

$$\mathbb{E}[\mathscr{V}(\psi^\varepsilon_\theta(x,a)+\Sigma\xi)^p] \leq (1-\varepsilon\mathfrak{c}_{\mathscr{V}})\left(\mathscr{V}(x)^p + \varepsilon\sum_{k=0}^{p-1}a_k\mathscr{V}(x)^k\right) + \varepsilon\mathfrak{c}_p \\ \leq \left(1-\varepsilon\frac{\mathfrak{c}_{\mathscr{V}}}{4}\right)\mathscr{V}(x)^p + \varepsilon\left(\mathfrak{c}_p - \frac{\mathfrak{c}_{\mathscr{V}}}{4}\mathscr{V}(x)^p + \sum_{k=0}^{p-1}a_k\mathscr{V}(x)^k\right)$$

As in part iii. of the proof of Corollary B.4, the interior of the second bracket is a continuous function which goes to $-\infty$ as $\|x\| \to +\infty$, so there must be a constant $\mathfrak{c}'_p \in \mathbb{R}_+$ (independent of ε) such that

$$\mathfrak{c}_p + \sup_{x\in\mathbb{R}^d}\left(-\frac{\mathfrak{c}_{\mathscr{V}}}{4}\mathscr{V}(x)^p + \sum_{k=0}^{p-1}a_k\mathscr{V}(x)^k\right) \leq \mathfrak{c}'_p < +\infty\,.$$

Therefore, we have the desired Lyapunov generator condition

$$\mathbb{E}[\mathscr{V}(\psi^\varepsilon_\theta(x,a)+\Sigma\xi)^p] \leq \left(1-\varepsilon\frac{\mathfrak{c}_{\mathscr{V}}}{4}\right)\mathscr{V}(x)^p + \varepsilon\mathfrak{c}'_p\,,$$

which is equivalently written for any $(x,a) \in \mathbb{R}^d \times \mathbb{A}$ as

$$\frac{1}{\varepsilon}\int(\mathscr{V}(\psi^\varepsilon_\theta(x,a)+\Sigma e)^p - \mathscr{V}(x)^p)\nu(\mathrm{d}e) \leq -\frac{\mathfrak{c}_{\mathscr{V}}}{4}\mathscr{V}(x)^p + \mathfrak{c}'_p\,. \tag{37}$$

By Itô's Lemma, (37), and a localisation argument, we have, for any $t \geq t_0 \geq 0$, that

$$\mathbb{E}[\mathscr{V}(X^{x_0,\alpha,\theta}_t)^p] = \mathbb{E}[\mathscr{V}(X^{x_0,\alpha,\theta}_{t_0})^p] \\ + \mathbb{E}\left[\int_{t_0}^t \frac{1}{\varepsilon}\int(\mathscr{V}(\psi^\varepsilon_\theta(X^{x_0,\alpha,\theta}_s,\alpha_s)+\Sigma e)^p - \mathscr{V}(X^{x_0,\alpha,\theta}_s)^p)\nu(\mathrm{d}e)\mathrm{d}s\right] \\ \leq \mathbb{E}[\mathscr{V}(X^{x_0,\alpha,\theta}_{t_0})^p] - \frac{\mathfrak{c}_{\mathscr{V}}}{4}\int_{t_0}^t \mathbb{E}\left[\mathscr{V}(X^{x_0,\alpha,\theta}_s)^p\right]\mathrm{d}s + (t-t_0)\mathfrak{c}'_p\,.$$

By a simple comparison argument for ODEs, we then obtain

$$\mathbb{E}[\mathscr{V}(X^{x_0,\alpha,\theta}_t)^p] \leq e^{-\frac{\mathfrak{c}_{\mathscr{V}}}{4}t}\mathscr{V}(x_0)^p + \frac{4\mathfrak{c}'_p}{\mathfrak{c}_{\mathscr{V}}}\left(1-e^{-\frac{\mathfrak{c}_{\mathscr{V}}}{4}t}\right).$$

Using now Assumption 2, we obtain

$$\mathbb{E}[\|X^{x_0,\alpha,\theta}_t\|^p] \leq \frac{1}{\ell^p_{\mathscr{V}}}\left(L^p_{\mathscr{V}}e^{-\frac{\mathfrak{c}_{\mathscr{V}}}{4}t}\|x_0\|^p + \frac{4\mathfrak{c}'_p}{\mathfrak{c}_{\mathscr{V}}}\left(1-e^{-\frac{\mathfrak{c}_{\mathscr{V}}}{4}t}\right)\right)\,.$$

□

C Concentration Inequality and Online Prediction Error

The key result of this section, Proposition 4.2, builds heavily on [34, Prop. 5]. Proposition 4.2 differs from this existing result in three ways. First, it is *any-time* i.e. does not require *a priori* knowledge of a time horizon. This is a minor technical refinement, but it is of practical importance. Second, it applies to a pure-jump process defined on $\mathbb{R}_+$. This apparent complexity vanishes when the filtration of the pure-jump process is chosen correctly, as the state process is piece-wise constant. Third, and most important, it applies to learning in a function class ($\mathscr{F}_\Theta$) of unbounded drifts for an unbounded process $X^{\alpha,\theta}$, which is an inherent difficulty in handling continuous state RL problems.

This third extension is non-trivial and leads us to significantly reshuffle the proof structure of [34], and to incorporate some self-normalised inequality arguments as well as high-probability bounds on the state from Appendix B. While many of the original ideas are still used, the way they link together has changed and thus we will include, in Appendix C.1, a complete derivation for the sake of clarity. In this spirit, we will prove a generic result (Theorem C.3), which itself implies Proposition 4.2.

Proposition 4.2 (Adapted from [32, Prop. 5]). *Under Assumptions 1 and 2, for any $x_0 \in \mathbb{R}^d$, and $\delta > 0$,*

$$\mathbb{P}\left(\left\{\theta^* \in \bigcap_{n=1}^{\infty} \mathcal{C}_n(\delta)\right\} \cap \left\{\sup_{n\in\mathbb{N}^*} \frac{\left\|X^{\varpi,\theta^*}_{\tau_n}\right\|}{H_\delta(n)} \leq 1\right\}\right) \geq 1-\delta\,, \tag{14}$$

Proposition 4.2 ensures that the sets $(\mathcal{C}_n(\delta))_{n\in\mathbb{N}}$ defined in (6) are valid confidence sets. In order to bound the regret, we need to go further and to bound the online prediction error of functions within these confidence sets along the trajectory (see. (57)).

For any $n \in \mathbb{R}$, let $\mathrm{d}_{\mathrm{E},n}$ denotes the $2\sqrt{\varepsilon/n}$-eluder dimension of the model class restricted to the set $B_n := \mathcal{B}_2(\sup_{s\leq\tau_n}\|X^{\varpi,\theta^*}_s\|)$, i.e. $\mathrm{d}_{\mathrm{E},n} := \dim_{\mathrm{E}}(\{f|_{B_n}\}_{f\in\mathscr{F}_\Theta}, 2\sqrt{\varepsilon/n})$. In Appendix C.2, we derive a general result (Proposition C.7) from which Proposition 4.2 follows.

Proposition 4.3. *Under Assumptions 1 and 2, for any $\delta \in (0,1)$, $\alpha \in \mathcal{A}$, $x_0 \in \mathbb{R}^d$, and $t \in \mathbb{R}_+$, we have with probability at least $1-\delta$*

$$\sum_{n=1}^{N_t}\left\|\mu_{\hat{\theta}_n}(X^{\alpha,\theta^*}_{\tau_n},\alpha_{\tau_n}) - \mu_{\theta^*}(X^{\alpha,\theta^*}_{\tau_n},\alpha_{\tau_n})\right\| \leq \tilde{\mathcal{O}}\left(\sqrt{\varepsilon \mathrm{d}_{\mathrm{E},N_t}\log(\mathscr{N}^{\varepsilon}_{N_t})N_t} + \mathrm{d}_{\mathrm{E},N_t}\right)\,, \tag{15}$$

and

$$\sum_{n=1}^{N_t}\left\|\mu_{\hat{\theta}_n}(X^{\alpha,\theta^*}_{\tau_n},\alpha_{\tau_n}) - \mu_{\theta^*}(X^{\alpha,\theta^*}_{\tau_n},\alpha_{\tau_n})\right\|^2 \leq \tilde{\mathcal{O}}\left(\mathrm{d}_{\mathrm{E},N_t}\log(\mathscr{N}^{\varepsilon}_{N_t})\right)\,. \tag{16}$$

C.1 Confidence sets

In this section, we work in a generic online learning framework, so that our results can be more easily compared and contrasted with [32, 34] and others. We, therefore, introduce some dedicated notation and a stand-alone assumption for this section.

Consider a set of functions $\mathscr{F}$ from $\mathbb{R}^d \to \mathbb{R}^d$, and fix $f^* \in \mathscr{F}$. We will study pairs of (random) $\mathbb{R}^d$-valued sequences $((X_i)_{i\in\mathbb{N}}, (Y_i)_{i\in\mathbb{N}})$ generated as

$$Y_i = f^*(X_i) + \xi_i$$

for $(\xi_i)_{i\in\mathbb{N}}$ a stochastic process in some filtered probability space $(\Omega', \mathcal{H}_\infty, \mathbb{H}, \mathbb{P})$, with each ξ_i independent of everything else up to time i. We take $\mathcal{H}_i$ as the completion of $\sigma(\{\xi_j\}_{j\leq i})$, for $i \in \mathbb{N}$, and we let $\mathbb{H} = (\mathcal{H}_i)_{i\geq 0}$.

Given some $\mathbb{R}^d$-valued and $\mathbb{H}$-adapted sequences $(Z_i)_{i\in\mathbb{N}}$ and $(Z'_i)_{i\in\mathbb{N}}$, and some $n \in \mathbb{N}^*$, let us define

$$\langle Z|Z'\rangle_n := \sum_{i=0}^{n-1}\langle Z_i|Z'_i\rangle \text{ and } \|Z\|_n := \sqrt{\langle Z|Z\rangle_n}\,.$$

While $\|\cdot\|_n$ is not a norm, it plays this role and we follow here the notational convention of [34]. We will extend the definitions of $\langle\cdot|\cdot\rangle_n$ and $\|\cdot\|_n$ to $n = 0$ by simply taking the empty sum to be 0, i.e. $\langle Z, Z'\rangle_0 := 0$.

To simplify notation, we will drop the sequence $(X_i)_{i\in\mathbb{N}}$ when it is an argument to a function inside $\|\cdot\|_n$ or $\langle\cdot|\cdot\rangle_n$: i.e. $\|f\|_n$ stands for $\|(f(X_i))_{i\in\mathbb{N}}\|_n$. With this notation in mind, for any $n \in \mathbb{N}$, we define $\hat{f}_n$ as an arbitrary element of

$$\operatorname*{argmin}_{f\in\mathscr{F}} \|Y - f\|_n^2 \,.$$

In other words $\hat{f}_n$ is a non-linear least-square fit in $\mathscr{F}$ using the first n points of $(X_i, Y_i)_{i\in\mathbb{N}}$. In this generic setting, we introduce Assumption 3, which in our end-goal application subsumes Assumptions 1 and 2 and Proposition 4.1.

Assumption 3. There is $(L, \Gamma) \in \mathbb{R}_+^2$ and a function $H_\delta : \mathbb{N} \to \mathbb{R}_+$ such that

$$\sup_{f\in\mathscr{F}} \sup_{x\in\mathbb{R}^d} \frac{\|f(x)\|}{1+\|x\|} \leq L\,,$$

and for all $i \in \mathbb{N}^*$, ξ_i is an $\mathcal{H}_{i-1}$-conditionally Γ^2-sub-Gaussian random variable, ξ_0 is Γ^2-sub-Gaussian, and the sequence $(X_i)_{i\in\mathbb{N}}$ satisfies

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} \frac{\|X_n\|}{H_\delta(n)} > 1\right) < \delta$$

for all $\delta \in (0, 1)$.

Let $(\mathscr{C}_n^\Gamma)_{n\in\mathbb{N}^*}$ denote a deterministic sequence of finite covers of $\mathscr{F}$, whose cardinalities are respectively given by $(\mathscr{N}_n^\Gamma)_{n\in\mathbb{N}^*}$, such that for all $n \in \mathbb{N}^*$

$$\sup_{f\in\mathscr{F}} \min_{g\in\mathscr{C}_n^\Gamma} \sup_{x\in\mathcal{B}(H_\delta(n))} \|f(x) - g(x)\| \leq \frac{\Gamma^2}{n}\,.$$

The definition of this cover corresponds to one used in [34] with a domain restricted to lie in the high-probability region of the state process instead of the whole domain. This ensures the cover remains finite for all $n \in \mathbb{N}^*$.

For any $\delta \in (0, 1)$, $n \in \mathbb{N}^*$, and $f \in \mathscr{F}$ let us define the quantities

$$\begin{aligned}
L_n^1(\delta) &:= \log((\Gamma^2 + 8L^2(1 + \sup_{i\leq n}\|X_i\|_2^2))\mathscr{N}_n^\Gamma \delta^{-1})\,,\\
L_n^0(\delta) &:= L_n^1(6\delta\pi^{-2}n^{-2})\,,\\
C_n^1(f) &:= \Gamma^2 + \|f - f^*\|_n^2\\
C_n^2(f) &:= \sup_{i\leq n}\|f(X_i) - \hat{f}_n(X_i)\|\,,
\end{aligned}$$

and the event

$$\begin{aligned}
\mathscr{E}_n^0(\delta) := \Bigg\{ \left\|\hat{f}_n - f^*\right\|_n &\leq 2\Gamma\sqrt{L_n^1\left(\frac{3\delta}{\pi^2 n^2}\right)}\\
&+ 2\sqrt{\Gamma^2 + 2\Gamma\left(n \sup_{g\in\mathscr{C}_n^\Gamma} C_n^2(g)\sqrt{2\log\left(\frac{4\pi^2 n^3}{3\delta}\right)} + \sqrt{2n \sup_{g\in\mathscr{C}_n^\Gamma} C_n^2(g) L_n^1\left(\frac{3\delta}{\pi^2 n^2}\right)}\right)}\Bigg\}.
\end{aligned} \tag{38}$$

Building upon the proof method of [34], the cornerstone of this section is Lemma C.2, which shows that, with high-probability, f^* is contained in all the elements of a sequence of confidence sets, each centred at $\hat{f}_n$ in the $\|\cdot\|_n$ norm.

Lemma C.2. *Under Assumption 3, for $n \in \mathbb{N}^*$ and $\delta \in (0, 1)$, we have*

$$\mathbb{P}\left(\bigcap_{n\in\mathbb{N}^*} \mathscr{E}_n^0(\delta)\right) \geq 1 - \delta\,.$$

We begin the proof of Lemma C.2 by giving the concentration inequality of Lemma C.1.

Lemma C.1. *Under Assumption 3, for all* $n \in \mathbb{N}^*$, $\delta \in (0,1)$, *and* $f \in \mathscr{F}$

$$\mathbb{P}\left(|\langle \xi | f - f^* \rangle_n| \geq \Gamma \sqrt{2(\Gamma^2 + \|f - f^*\|_n)\log\left(\frac{\Gamma^2 + \|f - f^*\|_n}{\delta}\right)}\right) \leq \delta\,.$$

Proof. This proof relies on extensively studied arguments for self-normalised inequalities, but we include it for completeness because it uses non standard constants. Let us begin by fixing $f \in \mathscr{F}$. For all $n \in \mathbb{N}$, let

$$Z_n(f) := \langle \xi | f - f^* \rangle_n\,.$$

For any $\lambda \in \mathbb{R}$, let us define the process $(M_n^\lambda(f))_{n\in\mathbb{N}}$ defined by

$$M_n^\lambda(f) := \exp\left(\lambda Z_n(f) - \frac{\lambda^2\Gamma^2}{2}\|f - f^*\|_n^2\right)\,.$$

Let us show that $M_n^\lambda(f)$ is a conditional supermartingale. For any $n \in \mathbb{N}$, we have

$$\mathbb{E}\left[M_{n+1}^\lambda(f)|\mathcal{H}_n\right] = M_n^\lambda(f)\mathbb{E}\left[\exp\left(\lambda\langle \xi_{n+1}|f(X_n) - f^*(X_n)\rangle_n\right)\middle|\mathcal{H}_n\right]e^{-\frac{\lambda^2\Gamma^2}{2}\|f(X_n) - f^*(X_n)\|_n^2}. \tag{39}$$

By the Cauchy-Schwartz inequality

$$|\langle \xi_n | f(X_n) - f^*(X_n)\rangle_n| \leq \|\xi_n\|_n \|f(X_n) - f^*(X_n)\|_n$$

and thus, since ξ_n is conditionally Γ^2-subgaussian with variance Γ^2, $\|\xi_n\|$ is Γ^2-subgaussian. Therefore

$$\mathbb{E}\left[\exp\left(\lambda\langle \xi_n|f(X_n) - f^*(X_n)\rangle_n - \frac{\lambda^2\Gamma^2}{2}\|f(X_n) - f^*(X_n)\|_n^2\right)|\mathcal{H}_n\right] \leq 1$$

and thus, by (39), $M_n^\lambda(f)$ is a supermartingale. By definition of $\langle\cdot|\cdot\rangle_0$ and $\|\cdot\|_0$, $M_0^\lambda(f) = 1$, so that $\mathbb{E}[M_n^\lambda(f)] \leq 1$ for all $n \in \mathbb{N}$.

We now perform a Laplace trick. Let Φ be the Gaussian measure of mean 0 and variance Γ^{-4} on $\mathbb{R}$, and let us define the process $(M_n(f))_{n\in\mathbb{N}}$ by

$$\begin{aligned}
M_n(f) :&= \int M_n^\lambda(f)\Phi(\mathrm{d}\lambda)\\
&= \int \exp\left(\lambda Z_n(f) - \frac{\lambda^2\Gamma^2}{2}\|f - f^*\|_n^2\right)\Phi(\mathrm{d}\lambda)\\
&= \frac{1}{\Gamma^2 + \|f - f^*\|_n^2}\exp\left\{\frac{Z_n^2(f)}{2\Gamma^2(\Gamma^2 + \|f - f^*\|_n^2)}\right\}\,.
\end{aligned}$$

By Markov's inequality, $\mathbb{P}(M_n(f) \geq \delta^{-1}) \leq \delta$, and thus

$$\mathbb{P}\left(Z_n(f) \geq \Gamma\sqrt{2(\Gamma^2 + \|f - f^*\|_n^2)\log\left(\frac{\Gamma^2 + \|f - f^*\|_n^2}{\delta}\right)}\right) \leq \delta\,.$$

□

We will turn to the proof of Lemma C.2. Recall (38), which defined for $\delta \in (0,1)$ and $n \in \mathbb{N}^*$, the event

$$\begin{aligned}
\mathscr{E}_n^0(\delta) := \Bigg\{ \left\|\hat{f}_n - f^*\right\|_n &\leq 2\Gamma\sqrt{L_n^1\left(\frac{3\delta}{\pi^2 n^2}\right)}\\
&+ 2\sqrt{\Gamma^2 + 2\Gamma\left(n\sup_{g\in\mathscr{C}_n^\Gamma} C_n^2(g)\sqrt{2\log\left(\frac{4\pi^2 n^3}{3\delta}\right)} + \sqrt{2n\sup_{g\in\mathscr{C}_n^\Gamma} C_n^2(g) L_n^1\left(\frac{3\delta}{\pi^2 n^2}\right)}\right)}\Bigg\}.
\end{aligned}$$

Lemma C.2. *Under Assumption 3, for $n \in \mathbb{N}^*$ and $\delta \in (0,1)$, we have*

$$\mathbb{P}\left(\bigcap_{n\in\mathbb{N}^*} \mathcal{E}_n^0(\delta)\right) \geq 1-\delta\,.$$

Proof. The proof builds on elements of [34]. We begin by giving two small auxiliary results which we will use.

i. Let $n \in \mathbb{N}^*$, and $\delta \in (0,1)$, by a union bound over the family of conditionally sub-Gaussian random variables $(\|\xi_i\|)_{i\in[n]}$, we have

$$\mathbb{P}\left(\sup_{i\leq n}\|\xi_i\| \geq \Gamma\sqrt{2\log\left(\frac{2n}{\delta}\right)}\right) \leq \delta \tag{40}$$

ii. For any $f \in \mathscr{F}$, and $n \in \mathbb{N}^*$ we have

$$\begin{aligned}\|f^*-Y\|_n^2 - \|f-Y\|_n^2 &= \langle f^*-Y|f^*-Y\rangle_n - \langle f-f^*+f^*-Y|f-f^*+f^*-Y\rangle_n\\ &= \langle f^*-Y|f^*-Y\rangle_n - \langle f-f^*|f-f^*\rangle_n\\ &\quad + 2\langle Y-f^*|f-f^*\rangle_n - \langle Y-f^*|Y-f^*\rangle_n\\ &= -\|f-f^*\|_n^2 + 2\langle \xi|f-f^*\rangle_n\,. \end{aligned} \tag{41}$$

Applying (41) with $f := \hat{f}_n$, the n-point non-linear least-square fit, leads to a non positive left hand side and thus

$$\left\|\hat{f}_n - f^*\right\|_n^2 \leq 2\left|\langle \xi|f-f^*\rangle_n\right|\,.$$

At the same time, for all $n \in \mathbb{N}^*$, by definition of $\mathscr{C}_n^\Gamma$, it holds that for all $g \in \mathscr{C}_n^\Gamma$

$$\begin{aligned}\left\|\hat{f}_n - f^*\right\|_n^2 &\leq 2\left|\langle \xi|g-f^*\rangle_n\right| + 2\left|\langle \xi|\hat{f}_n-g\rangle_n\right|\\ &\leq 2\left|\langle \xi|g-f^*\rangle_n\right| + 2n\sup_{i\leq n}\|\xi_i\|_2\, C_n^2(g)\,. \end{aligned}\tag{42}$$

Combining (40) and (42), we obtain, for all $\delta \in (0,1)$, $n \in \mathbb{N}^*$, and $g \in \mathscr{C}_n^\Gamma$, that

$$\mathbb{P}\left(\left\|\hat{f}_n - f^*\right\|_n^2 \geq 2\left|\langle \xi|g-f^*\rangle_n\right| + 2nC_n^2(g)\Gamma\sqrt{2\log\left(\frac{2n}{\delta}\right)}\right) \leq \delta \tag{43}$$

Let us now provide two bounds on $C_n^1(g)$ we will use. For all $n \in \mathbb{N}^*$, $\delta \in (0,1)$ and $g \in \mathscr{C}_n^\Gamma$, let

$$C_n^1(g) \leq \Gamma^2 + 8L^2(1+\sup_{i\leq n}\|X_i\|^2)\,. \tag{44}$$

$$C_n^1(g) \leq \Gamma^2 + \left\|\hat{f}_n - f^*\right\|_n^2 + \left\|g-\hat{f}_n\right\|_n^2 \leq C_n^1(\hat{f}_n) + nC_n^2(g)\,, \tag{45}$$

Applying Lemma C.1 for each $g \in \mathscr{C}_n^\Gamma$, by a union bound over $g \in \mathscr{C}_n^\Gamma$, we have for any $\delta_0(n) \in (0,1)$ (to be fixed at the end), that

$$\delta_0(n) \geq \mathbb{P}\left(\sup_{g\in\mathscr{C}_n^\Gamma}\left|\langle \xi|g-f^*\rangle_n\right| \geq \Gamma\sqrt{2\sup_{g\in\mathscr{C}_n^\Gamma} C_n^1(g)\log\left(\frac{\sup_{g\in\mathscr{C}_n^\Gamma} C_n^1(g)\mathscr{N}_n^\Gamma}{\delta_0(n)}\right)}\right)\,.$$

Applying (44) and (45) this becomes

$$\delta_0(n) \geq \mathbb{P}\left(\sup_{g\in\mathscr{C}_n^\Gamma}\left|\langle \xi|g-f^*\rangle_n\right| \right.$$

$$\left. \geq \Gamma\sqrt{2(C_n^1(\hat{f}_n) + n\sup_{g\in\mathscr{C}_n^\Gamma} C_n^2(g))\log\left(\frac{(\Gamma^2+8L^2(1+\sup_{i\leq n}\|X_i\|^2))\mathscr{N}_n^\Gamma}{\delta_0(n)}\right)}\right)$$

and thus

$$\delta_0(n) \geq \mathbb{P}\left(\sup_{g\in\mathscr{C}_n^\Gamma} |\langle \xi | g - f^* \rangle_n| \geq \Gamma\sqrt{2L_n^1(\delta_0(n))}\left(\sqrt{C_n^1(\hat{f}_n)} + \sqrt{n \sup_{g\in\mathscr{C}_n^\Gamma} C_n^2(g)}\right)\right). \tag{46}$$

Combining (43) and (46) by a union bound gives us

$$\delta_0(n) \geq \mathbb{P}\Bigg(\left\|\hat{f}_n - f^*\right\|_n^2 \geq 2\Gamma\sqrt{2L_n^1\left(\frac{\delta_0(n)}{2}\right)}\left(\sqrt{C_n^1(\hat{f}_n)} + \sqrt{n \sup_{g\in\mathscr{C}_n^\Gamma} C_n^2(g)}\right) + 2nC_n^2(g)\Gamma\sqrt{2\log\left(\frac{4n}{\delta_0(n)}\right)}\Bigg).$$

For all $n \in \mathbb{N}^*$, on the complement of this event (whose probability is at least $1 - \delta_0(n)$) we have

$$C_n^1(\hat{f}_n) \leq \Gamma^2 + \Gamma\sqrt{2C_n^1(\hat{f}_n)L_n^1(\delta_0(n)/2)} + h_n^\Gamma, \tag{47}$$

in which

$$h_n^\Gamma := 2\Gamma\left(n \sup_{g\in\mathscr{C}_n^\Gamma} C_n^2(g)\sqrt{2\log\left(\frac{4n}{\delta_0(n)}\right)} + \sqrt{2n \sup_{g\in\mathscr{C}_n^\Gamma} C_n^2(g) L_n^1\left(\frac{\delta_0(n)}{2}\right)}\right).$$

Viewing (47) as a second order polynomial in $\sqrt{C_n^1(\hat{f}_n)}$, we obtain via its roots that

$$\begin{aligned}\sqrt{C_n^1(\hat{f}_n)} &\leq \Gamma\sqrt{L_n^1(\delta_0(n)/2)} + \sqrt{\left(\Gamma\sqrt{L_n^1(\delta_0(n)/2)}\right)^2 + 4(\Gamma^2 + h_n^\Gamma)} \\ &\leq 2\Gamma\sqrt{L_n^1(\delta_0(n)/2)} + 2\sqrt{\Gamma^2 + h_n^\Gamma}.\end{aligned}$$

Since $\left\|\hat{f}_n - f^*\right\|_n \leq \sqrt{C_n^1(\hat{f}_n)}$ by definition of $C_n^1(\hat{f}_n)$, we have

$$\begin{aligned}\left\|\hat{f}_n - f^*\right\|_n \leq 2\sqrt{\Gamma^2 + 2\Gamma\left(n \sup_{g\in\mathscr{C}_n^\Gamma} C_n^2(g)\sqrt{2\log\left(\frac{4n}{\delta_0(n)}\right)} + \sqrt{2n \sup_{g\in\mathscr{C}_n^\Gamma} C_n^2(g) L_n^1\left(\frac{\delta_0(n)}{2}\right)}\right)} \\ + 2\Gamma\sqrt{L_n^1(\delta_0(n)/2)}.\end{aligned}$$

Therefore, letting

$$\begin{aligned}\mathcal{E}_n^1(\delta) := \Bigg\{ \left\|\hat{f}_n - f^*\right\|_n \leq 2\Gamma\sqrt{L_n^1\left(\frac{\delta}{2}\right)} \\ + 2\sqrt{\Gamma^2 + 2\Gamma\left(n \sup_{g\in\mathscr{C}_n^\Gamma} C_n^2(g)\sqrt{2\log\left(\frac{4n}{\delta}\right)} + \sqrt{2n \sup_{g\in\mathscr{C}_n^\Gamma} C_n^2(g) L_n^1\left(\frac{\delta}{2}\right)}\right)}\Bigg\},\end{aligned}$$

we have, for all $n \in \mathbb{N}^*$, that $\mathbb{P}(\mathcal{E}_n^1(\delta_0(n))) \geq \delta_0(n)$. Letting $\delta_0(n) = \frac{6}{\pi^2 n^2}\delta$, by a union bound we obtain

$$\mathbb{P}\left(\bigcap_{n\in\mathbb{N}^*} \mathcal{E}_n^1(\delta_0(n))\right) \geq 1 - \delta\frac{6}{\pi^2}\sum_{n=1}^\infty \frac{1}{n^2} = 1 - \delta.$$

Noting that $\mathcal{E}_n^0(\delta) = \mathcal{E}_n^1(\delta_0(n))$ for all $\delta \in (0,1)$ and $n \in \mathbb{N}^*$ completes the proof. □

In the proof of Lemma C.2, we used self-normalised inequalities to generalise the results of [34] to unbounded states. We now incorporate the high probability bound of Assumption 3 and formalise confidence sets, which will prove Theorem C.3. Theorem C.3 can then be specified for our setting by merging it with the results of Appendix B in Proposition 4.2.

For $\delta \in (0,1)$, let $\beta_0 \in \mathbb{R}_+$ and let us define the sequence $(\mathcal{C}_n(\delta))_{n\in\mathbb{N}}$ in which

$$\mathcal{C}_n(\delta) := \left\{ f \in \mathscr{F} : \|f - \hat{f}_n\|_n \leq \beta_n \right\} \tag{48}$$

with

$$\beta_n(\delta) := \beta_0 \vee 2\Gamma \left(\sqrt{1 + 2\left(\sqrt{2\Gamma \log\left(\frac{8n}{\delta}\right)} + \sqrt{2L_n^0\left(\frac{\delta}{4}\right)} \right)} + \sqrt{L_n^0\left(\frac{\delta}{4}\right)} \right) . \tag{49}$$

Theorem C.3. *Under Assumption 3, we have for all $\delta \in (0,1)$*

$$\mathbb{P}\left(\left\{ \bigcap_{n\in\mathbb{N}^*} \{f^* \in \mathcal{C}_n(\delta)\} \right\} \bigcap \left\{ \sup_{n\in\mathbb{N}^*} \frac{\|X_n\|}{H_\delta(n)} \leq 1 \right\} \right) \leq \delta$$

Proof. Fix $\delta \in (0,1)$, and assume $\omega \in \{\omega' \in \Omega : \sup_{n\in\mathbb{N}^*} \|X_n(\omega')\|_2 / H_\delta(n) \leq 1\}$. In this case we have the following bound, for all $n \in \mathbb{N}^*$

$$2n \min_{g\in\mathscr{C}_n^\Gamma} C_n^2(g) \leq 2\Gamma^2$$

by definition of $\mathscr{C}_n^\Gamma$ as a $\Gamma^2 n^{-1}$ cover on $\mathcal{B}_2(H_\delta(n))$. Therefore, the event

$$\left\{ \bigcap_{n\in\mathbb{N}^*} \mathcal{E}_n^0(\delta) \right\} \bigcap \left\{ \sup_{n\in\mathbb{N}^*} \frac{\|X_n\|_2}{H_\delta(n)} \leq 1 \right\}$$

is contained in the event

$$\mathcal{E}^0(\delta) := \left\{ \bigcap_{n\in\mathbb{N}^*} \left\{ \left\| f^* - \hat{f}_n \right\|_n \leq \beta_n(2\delta) \right\} \right\} \bigcap \left\{ \sup_{n\in\mathbb{N}^*} \frac{\|X_n\|_2}{H_\delta(n)} \leq 1 \right\} .$$

By Lemma C.2, Assumption 3, and a union bound, $\mathbb{P}\left(\mathcal{E}^0(\delta)\right) \geq 1 - 2\delta$, and we obtain the result by (48) and (49), i.e. by definition of $\mathcal{C}_n(\delta)$. ☐

Proposition 4.2 (Adapted from [32, Prop. 5]). *Under Assumptions 1 and 2, for any $x_0 \in \mathbb{R}^d$, and $\delta > 0$,*

$$\mathbb{P}\left(\left\{ \theta^* \in \bigcap_{n=1}^{\infty} \mathcal{C}_n(\delta) \right\} \cap \left\{ \sup_{n\in\mathbb{N}^*} \frac{\|X_{\tau_n}^{\varpi,\theta^*}\|}{H_\delta(n)} \leq 1 \right\} \right) \geq 1 - \delta \,, \tag{14}$$

Proof. The proof follows by applying Theorem C.3 to this setting. Where $(X_i)_{i\in\mathbb{N}} := ((X_{\tau_i}^{\varpi,\theta^*}, \varpi_{\tau_i}))_{i\in\mathbb{N}}$, $(Y_i)_{i\in\mathbb{N}} := (X_{\tau_{i+1}}^{\varpi,\theta^*} - X_{\tau_i}^{\varpi,\theta^*})_{i\in\mathbb{N}}$, $\mathscr{F} := \mathscr{F}_\Theta$ and with $(\xi_{n+1})_{n\in\mathbb{N}}$ and $(\beta_n(\delta))_{n\in\mathbb{N}^*}$ as defined in Section 2 and (13) respectively. This sets $\Gamma = \|\Sigma\|_{\mathrm{op}} = \varepsilon^{\frac{1}{2}} \|\bar{\Sigma}\|_{\mathrm{op}}$. The only subtlety is that the process X^{ϖ,θ^*} is measured at random times, but since these times are independent of anything else, and the process is almost surely constant between them, they do not affect the proof. ☐

C.2 Widths of confidence sets

In Appendix C.1, we showed how to design confidence sets along a trajectory of $X^{\alpha,\theta}$ for learning μ by using NLLS to minimise a fit error of the form

$$\sum_{n=1}^{N} \left\| \mu_1(X_{\tau_n}^{\alpha,\theta^*}, \alpha_{\tau_n}) - \mu_2(X_{\tau_n}^{\alpha,\theta^*}, \alpha_{\tau_n}) \right\| ,$$

for $(\mu_1, \mu_2) \in \mathcal{C}_N(\delta)$ and $N \in \mathbb{N}^*$. When analysing the regret of such a learning algorithm this is not sufficient: instead of the fit error, we need to control a prediction error of the form

$$\sum_{n=1}^{N} \left\| \mu_{\theta_n}(X_{\tau_n}^{\alpha,\theta^*}, \alpha_{\tau_n}) - \mu_{\theta^*}(X_{\tau_n}^{\alpha,\theta^*}, \alpha_{\tau_n}) \right\| ,$$

for $(\mu_{\theta_n})_{n\in\mathbb{N}} \subset \mathscr{F}_\theta$ such that $\mu_{\theta_n} \in \mathcal{C}_n(\delta)$ for all $n \in \mathbb{N}$. The difference is that μ_{θ_n} changes over time, so that the sum counts the errors in predicting the next state made by the sequence $(\mu_{\theta_n})_{n\in\mathbb{N}}$.

In fact, since we will want to implement lazy-updates, we will need a more general result where the μ_{θ_n} are not all in their respective $\mathcal{C}_n(\delta)$ but rather are from a piece-wise constant sequence with $\mu_{\theta_n} := \mu_{\theta_{k(n)}} \in \mathcal{C}_{k(n)}(\delta)$, where $k(n) \le n$ for all $n \in \mathbb{N}$. Therefore, as in Appendix C.1, we begin by showing a general result in the learning framework of [34] (Proposition C.4), then apply it to our setting to prove Proposition 4.3. Using the notation of Appendix C.1, let $\mathscr{F}$ be a function class of functions from $\mathbb{R}^d \to \mathbb{R}^d$, and recall the arbitrary sequence $(X_n)_{n\in\mathbb{N}} \subset \mathbb{R}^d$.

The ϵ-eluder dimension of a function class $\mathscr{F}$, for $\epsilon \in \mathbb{R}_+$, introduced in [34] is a notion of dimension which is perfectly tailored to converting fit errors into prediction errors. We defer to [34] for its technical definition. Unlike [34], we must adapt our eluder dimension to work with unbounded functions on unbounded processes. Failing to do so would lead our results to be largely vacuous since the eluder dimension of $\mathscr{F}$ might be infinite for any ϵ.

We work with a modified eluder dimension, which takes three arguments: a function class $\mathscr{F}$ whose elements have for domain a set $\mathcal{X} \subset \mathbb{R}^d$; a set $S \subset \mathcal{X}$; and $\epsilon \in \mathbb{R}_+$. Our modified eluder dimension is the ϵ-eluder dimension of $\{f|_S : f \in \mathscr{F}\}$, the class containing the restrictions to S of elements of $\mathscr{F}$, which we denote by $\dim_{\mathrm{E}}^S(\mathscr{F}, \epsilon)$. In this way, the eluder dimension of [34] is $\dim_{\mathrm{E}}^{\mathcal{X}}(\mathscr{F}, \epsilon)$. For $n \in \mathbb{N}^*$, let $B_n := \mathcal{B}_2(\sup_{i\in[n]} \|X_i\|)$ and, for any $u \in \mathbb{R}_+$, let us define the sequence $(\mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(u))_{n\in\mathbb{N}^*}$, in which

$$\mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(u) := \dim_{\mathrm{E}}^{B_n}\left(\mathscr{F}, \frac{2u}{\sqrt{n}}\right)$$

for all $n \in \mathbb{N}^*$ and $u \in \mathbb{R}_+$.

Proposition C.4. *Let $(\tilde{\beta}_i)_{i\in\mathbb{N}}$ be a non-decreasing positive real-valued sequence, $(\tilde{f}_i)_{i\in\mathbb{N}}$, and $(\mathscr{F}_i)_{i\in\mathbb{N}}$ be a sequence of subsets of $\mathscr{F}$ of the form $\mathscr{F}_i := \{f \in \mathscr{F} : \|f - \tilde{f}_i\|_i \le \tilde{\beta}_i\}$. Then, for any $n \in \mathbb{N}$, we have*

$$\sum_{i=1}^{n} \sup_{(f,f')\in\mathscr{F}_n^2} \|f(X_i) - f'(X_i)\| \le 2\tilde{\beta}_n \sqrt{\mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde{\beta}_0) n} + \mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde{\beta}_0) \sup_{i\in[n]} \|X_i\| \,, \tag{50}$$

and

$$\sum_{i=1}^{n} \sup_{(f,f')\in\mathscr{F}_n^2} \|f(X_i) - f'(X_i)\|^2 \le 4\tilde{\beta}_n^2 \mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde{\beta}_0)\left(3 + \log\left(\frac{n \sup_{i\in[n]} \|X_i\|}{16\tilde{\beta}_n^4 (\mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde{\beta}_0))^2}\right)\right)$$
$$+ 2\mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(1 + 2\tilde{\beta}_n^2 \mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde{\beta}_0))(1 + \sup_{i\in[n]} \|X_i\|) \,. \tag{51}$$

To prove Proposition C.4, the key result of [34] we leverage is Lemma C.5 which we combine with two functional inequalities given in Lemma C.6.

For a function class $\mathscr{F}$ with domain $\mathcal{X} \subset \mathbb{R}^d$, and any $x \in \mathcal{X}$, let us define

$$\Lambda(\mathscr{F}; x) = \sup_{(f_1,f_2)\in\mathscr{F}^2} \|f_1(x) - f_2(x)\| \,.$$

The quantity $\Lambda(\mathscr{F}, x)$ is the maximal prediction gap at x between two functions in $\mathscr{F}$. Bounding the prediction error along $(X_i)_{i\in\mathbb{N}}$ of a sequence of function classes $(\mathscr{F}_i)_{i\in\mathbb{N}} \subset \mathscr{F}$ means bounding $\sum_{i=1}^n \Lambda(\mathscr{F}_i, X_i)$ in terms of $n \in \mathbb{N}$.

Lemma C.5. *[[34, Prop.3]] Let $(\tilde{f}_i)_{i\in\mathbb{N}}$ be a sequence of elements of $\mathscr{F}$, $(\mathscr{F}_i)_{i\in\mathbb{N}}$ be a sequence of subsets of $\mathscr{F}$ of the form $\mathscr{F}_i := \{f \in \mathscr{F} : \|f - \tilde{f}_i\|_i \le \tilde{\beta}_i\}$. For any $\epsilon \in (0, 1)$ and $n \in \mathbb{N}$, one has*

$$\sum_{i=1}^{n} \mathbb{1}_{\{\Lambda(\mathscr{F}_i; X_i) > \epsilon\}} \le \left(\frac{4\tilde{\beta}_n^2}{\epsilon^2} + 1\right) \dim_{\mathrm{E}}^{B_n}(\mathscr{F}, \epsilon) \,.$$

Proof. Following the proof of [34, Prop.3], the only modification involves the bound $\|\overline{f} - \underline{f}\|_n \le \tilde{\beta}_n$, for any $(\overline{f}, \underline{f}) \in \mathscr{F}_n^2$, which holds by assumption. □

Lemma C.6. *Let $(x_i)_{i\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}^*}$. Assume there is a family of positive sequences $((\zeta_n^\epsilon)_{n\in\mathbb{N}})_{\epsilon\in\mathbb{R}_+}$ and a family of positive constants $(\chi^\epsilon)_{\epsilon\in\mathbb{R}_+}$ such that, for any $n \in \mathbb{N}^*$ and $\epsilon > 0$,*

$$\sum_{i=1}^{n} \mathbb{1}_{\{x_i>\epsilon\}} \le \frac{\zeta_n^\epsilon}{\epsilon^2} + \chi^\epsilon \tag{52}$$

then the following two inequalities hold

$$\sum_{i=1}^{n} x_i \le 2\sqrt{n\zeta_n^\epsilon} + \chi^\epsilon \sup_{i\in[n]} x_i \tag{53}$$

$$\sum_{i=1}^{n} x_i^2 \le \zeta_n^\epsilon \left(3 + \log\left(\frac{n\sup_{i\in[n]} x_i^2}{(\zeta_n^\epsilon)^2}\right)\right) + \chi^\epsilon(2+\zeta_n^\epsilon)(1+\sup_{i\in[n]} x_i^2)\,. \tag{54}$$

Proof.

i. For $\epsilon > 0$, we have by (52)

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\epsilon)\mathbb{1}_{\{x_i>\epsilon\}} &= \sum_{i=1}^{n}\int_\epsilon^{x_i} \mathbb{1}_{\{x_i>u\}}\mathrm{d}u \\
&\le \int_\epsilon^{\sup_{i\in[n]} x_i} \sum_{i=1}^{n} \mathbb{1}_{\{x_i>u\}}\mathrm{d}u \\
&\le \int_\epsilon^{\sup_{i\in[n]} x_i} \frac{\zeta_n^\epsilon}{u^2} + \chi^\epsilon \mathrm{d}u \\
&= \chi \sup_{i\in[n]} x_i - \frac{\zeta_n^\epsilon}{\sup_{i\in[n]} x_i} - \chi^\epsilon\epsilon + \frac{\zeta_n^\epsilon}{\epsilon}\,,
\end{aligned}$$

and thus

$$\sum_{i=1}^{n}(x_i-\epsilon)\mathbb{1}_{\{x_i>\epsilon\}} \le \frac{\zeta_n^\epsilon}{\epsilon} + \chi^\epsilon \sup_{i\in[n]} x_i\,. \tag{55}$$

Combining (55) with

$$\sum_{i=1}^{n}(x_i-\epsilon) \le \sum_{i=1}^{n}(x_i-\epsilon)\mathbb{1}_{\{x_i>\epsilon\}}$$

yields

$$\sum_{i=1}^{n} x_i \le n\epsilon + \frac{\zeta_n^\epsilon}{\epsilon} + \chi^\epsilon \sup_{i\in[n]} x_i\,.$$

Setting $\epsilon = \sqrt{\zeta_n^\epsilon/n}$ yields (53).

ii. To prove (54), we iterate the bound (55)

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\epsilon)^2\mathbb{1}_{\{x_i>\epsilon\}} &= 2\sum_{i=1}^{n}\int_\varepsilon^{x_i}(x_i-u)\mathbb{1}_{\{x_i>u\}}\mathrm{d}u \\
&\le 2\sum_{i=1}^{n}\int_\epsilon^{\sup_{i\in[n]} x_i}(x_i-u)\mathbb{1}_{\{x_i>u\}}\mathrm{d}u \\
&\le 2\int_\epsilon^{\sup_{i\in[n]} x_i} \frac{\zeta_n^\epsilon}{\epsilon} + \chi^\epsilon \sup_{i\in[n]} x_i \mathrm{d}u \\
&\le 2\left(\chi(\sup_{i\in[n]} x_i^2 - \sup_{i\in[n]} x_i\epsilon) + \zeta_n^\epsilon \log\left(\frac{\sup_{i\in[n]} x_i}{\epsilon}\right)\right) \\
&\le 2\zeta_n^\epsilon \log\left(\frac{\sup_{i\in[n]} x_i}{\epsilon}\right) + 2\chi^\epsilon \sup_{i\in[n]} x_i^2\,.
\end{aligned}$$

Now, by some algebraic manipulations of $\sum_{i=1}^n x_i^2$, completing the square, discarding negative terms, and using (55) in the third step, we get

$$\begin{aligned}\sum_{i=1}^n x_i^2 &\le \sum_{i=1}^n x_i^2 \mathbb{1}_{\{x_i>\epsilon\}} + \epsilon^2 \sum_{i=1}^n \mathbb{1}_{\{x_i>\epsilon\}} \\ &\le \sum_{i=1}^n (x_i-\epsilon)^2 \mathbb{1}_{\{x_i>\epsilon\}} + 2\epsilon \sum_{i=1}^n x_i \mathbb{1}_{\{x_i>\epsilon\}} + n\epsilon^2 \\ &\le 2\zeta_n^\epsilon \log\left(\frac{\sup_{i\in[n]} x_i}{\epsilon}\right) + 2\chi^\epsilon \sup_{i\in[n]} x_i^2 + \epsilon\left(\frac{\zeta_n^\epsilon}{\epsilon} + \chi^\epsilon \sup_{i\in[n]} x_i + \epsilon n\right) + n\epsilon^2 .\end{aligned}$$

Taking $\epsilon = \zeta_n^\epsilon/\sqrt{n}$ and factoring, using also $u \le 1+u^2$ for $u \in \mathbb{R}$, yields

$$\sum_{i=1}^n x_i^2 \le \zeta_n^\epsilon \left(3 + \log\left(\frac{n \sup_{i\in[n]} x_i^2}{(\zeta_n^\epsilon)^2}\right)\right) + \chi^\epsilon(2+\zeta_n^\epsilon)(1+\sup_{i\in[n]} x_i^2) .$$

□

Proposition C.4. *Let $(\tilde\beta_i)_{i\in\mathbb{N}}$ be a non-decreasing positive real-valued sequence, $(\tilde f_i)_{i\in\mathbb{N}}$, and $(\mathscr{F}_i)_{i\in\mathbb{N}}$ be a sequence of subsets of $\mathscr{F}$ of the form $\mathscr{F}_i := \{f \in \mathscr{F} : \|f - \tilde f_i\|_i \le \tilde\beta_i\}$. Then, for any $n \in \mathbb{N}$, we have*

$$\sum_{i=1}^n \sup_{(f,f')\in\mathscr{F}_n^2} \|f(X_i) - f'(X_i)\| \le 2\tilde\beta_n \sqrt{\mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde\beta_0) n} + \mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde\beta_0) \sup_{i\in[n]} \|X_i\| , \tag{50}$$

and

$$\begin{aligned}\sum_{i=1}^n \sup_{(f,f')\in\mathscr{F}_n^2} \|f(X_i) - f'(X_i)\|^2 \le 4\tilde\beta_n^2 \mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde\beta_0)\left(3 + \log\left(\frac{n \sup_{i\in[n]}\|X_i\|}{16\tilde\beta_n^4(\mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde\beta_0))^2}\right)\right) \\ + 2\mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(1 + 2\tilde\beta_n^2 \mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde\beta_0))(1 + \sup_{i\in[n]}\|X_i\|) .\end{aligned} \tag{51}$$

Proof. The proof consists in applying Lemma C.6 to Lemma C.5, with $x_i = \Lambda(\mathscr{F}_i, X_i)$, $\zeta_n^\epsilon = 4\tilde\beta^2 \dim_{\mathrm{E}}^{B_n}(\mathscr{F}, \epsilon)$ ($B_n := \mathcal{B}_2(\sup_{i\in[n]}\|X_i\|)$), and $\chi^\epsilon = \dim_{\mathrm{E}}^{B_n}(\mathscr{F}, \epsilon)$. When we set the value of ϵ in the proof of Lemma C.6, χ^ϵ becomes

$$\dim_{\mathrm{E}}^{B_n}\left(\mathscr{F}, \sqrt{frac4\tilde\beta_n^2 n}\right) \le \dim_{\mathrm{E}}^{B_n}\left(\mathscr{F}, \sqrt{\frac{4\tilde\beta_0^2}{n}}\right)$$

as $(\tilde\beta_n)_{n\in\mathbb{N}}$ is non-decreasing and the eluder dimension is decreasing in its third argument. An analogue remark holds for ζ_n^ϵ. We can thus substitute $\zeta_n^\epsilon = 4\tilde\beta_n^2 \mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde\beta_0)$ and $\chi^\epsilon = \mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde\beta_0)$ in (53) and (54), which gives the result. □

We now apply Proposition C.4 to our setting. For $n \in \mathbb{N}^*$, let us recall the shorthand notation

$$\mathrm{d}_{\mathrm{E},n} := \dim_{\mathrm{E}}^{B_n}\left(\mathscr{F}_\Theta, 2\sqrt{\frac{\varepsilon}{n}}\right) \tag{56}$$

in which we extended the notation from $(X_i)_{i\in\mathbb{N}}$ to $X^{\alpha,\theta}$ in the evident manner.

Proposition C.7. *Under Assumptions 1 and 2, for any $(\alpha, \theta) \in \mathcal{A} \times \Theta$ and $t \in \mathbb{R}_+$, any non-decreasing positive real-valued sequence $(\tilde\beta_n)_{n\in\mathbb{N}}$, any $(\tilde\mu_n)_{n\in\mathbb{N}} \subset \mathscr{F}_\Theta$, and any sequence $(\mathscr{F}_n)_{n\in\mathbb{N}}$ of subsets of $\mathscr{F}_\theta$ of the form*

$$\mathscr{F}_n = \left\{\mu \in \mathscr{F}_\Theta : \sqrt{\sum_{i=0}^{n-1} \left\|\mu_n(X_{\tau_i}^{\alpha,\theta}, \alpha_{\tau_i}) - \tilde\mu_n(X_{\tau_i}^{\alpha,\theta}, \alpha_{\tau_i})\right\|_2^2} \le \tilde\beta_n\right\} ,$$

we have

$$\sum_{n=1}^{N_t} \sup_{(\mu_1,\mu_2)\in\mathscr{F}_n} \left\|\mu_1(X^{\alpha,\theta}_{\tau_n},\alpha_{\tau_n}) - \mu_2(X^{\alpha,\theta}_{\tau_n},\alpha_{\tau_n})\right\| \le 2\beta_{N_t}\sqrt{\mathrm{d}_{\mathrm{E},N_t}} + \mathrm{d}_{\mathrm{E},N_t}\sup_{s\le t}\left\|X^{\alpha,\theta}_s\right\| \,, \tag{57}$$

and

$$\sum_{n=1}^{N_t} \sup_{(\mu_1,\mu_2)\in\mathscr{F}_n} \left\|\mu_1(X^{\alpha,\theta}_{\tau_n},\alpha_{\tau_n}) - \mu_2(X^{\alpha,\theta}_{\tau_n},\alpha_{\tau_n})\right\|^2$$
$$\le 4\beta^2_{N_T}\mathrm{d}_{\mathrm{E},N_t}\left(3+\log\left(\frac{N_t\sup_{s\le t}\left\|X^{\alpha,\theta}_s\right\|}{16\beta^4_{N_t}\mathrm{d}^2_{\mathrm{E},N_t}}\right)\right) + 2\mathrm{d}_{\mathrm{E},N_t}(1+2\beta^2_{N_t}\mathrm{d}_{\mathrm{E},N_t})(1+\sup_{s\le t}\left\|X^{\alpha,\theta}_s\right\|^2). \tag{58}$$

Proof. Immediate by applying Proposition C.4 to our setting, as we did in the proof of Proposition 4.2. □

Under the event of Proposition 4.2, which ensures that $\theta^* \in \cap_{n\in\mathbb{N}}\mathcal{C}_n(\delta)$, we can derive from Proposition C.7 a bound on the prediction error relative to the true dynamics X^{α,θ^*} generated by the control $\alpha\in\mathcal{A}$, in particular we are interested in $\alpha=\varpi$.

Proposition 4.3. *Under Assumptions 1 and 2, for any* $\delta\in(0,1)$, $\alpha\in\mathcal{A}$, $x_0\in\mathbb{R}^d$, *and* $t\in\mathbb{R}_+$, *we have with probability at least* $1-\delta$

$$\sum_{n=1}^{N_t}\left\|\mu_{\hat\theta_n}(X^{\alpha,\theta^*}_{\tau_n},\alpha_{\tau_n}) - \mu_{\theta^*}(X^{\alpha,\theta^*}_{\tau_n},\alpha_{\tau_n})\right\| \le \tilde{\mathcal{O}}\left(\sqrt{\varepsilon\mathrm{d}_{\mathrm{E},N_t}\log(\mathscr{N}^{\varepsilon}_{N_t})N_t} + \mathrm{d}_{\mathrm{E},N_t}\right)\,, \tag{15}$$

and

$$\sum_{n=1}^{N_t}\left\|\mu_{\hat\theta_n}(X^{\alpha,\theta^*}_{\tau_n},\alpha_{\tau_n}) - \mu_{\theta^*}(X^{\alpha,\theta^*}_{\tau_n},\alpha_{\tau_n})\right\|^2 \le \tilde{\mathcal{O}}\left(\mathrm{d}_{\mathrm{E},N_t}\log(\mathscr{N}^{\varepsilon}_{N_t})\right)\,. \tag{16}$$

Proof. This follows from Proposition C.7 by choosing $(\tilde\beta_n)_{n\in\mathbb{N}} = (\beta_n(\delta))_{n\in\mathbb{N}}$ and $(\mathscr{F}_n)_{n\in\mathbb{N}} = (\mathcal{C}_n(\delta))_{n\in\mathbb{N}}$, i.e. choosing $(\tilde\mu_n)_{n\in\mathbb{N}} = (\mu_{\hat\theta_n})_{n\in\mathbb{N}}$, the NLLS fit on n points. It is key to notice that these choices of $(\tilde\beta_n)_{n\in\mathbb{N}}$, $(\mathscr{F}_n)_{n\in\mathbb{N}}$, and $(\tilde\mu_n)_{n\in\mathbb{N}}$ are adapted to $\mathbb{F}$, and therefore we can apply Proposition C.7 on the event of Proposition 4.2 without issues. This yields

$$\sum_{n=1}^{N_t}\left\|\mu_{\hat\theta_n}(X^{\alpha,\theta^*}_{\tau_n},\alpha_{\tau_n}) - \mu_{\theta^*}(X^{\alpha,\theta^*}_{\tau_n},\alpha_{\tau_n})\right\| \le 2\beta_{N_t}(\delta)\sqrt{\mathrm{d}_{\mathrm{E},N_t}} + \mathrm{d}_{\mathrm{E},N_t}H_\delta(N_T)\,,$$

and

$$\sum_{n=1}^{N_t}\left\|\mu_{\hat\theta_n}(X^{\alpha,\theta^*}_{\tau_n},\alpha_{\tau_n}) - \mu_{\theta^*}(X^{\alpha,\theta^*}_{\tau_n},\alpha_{\tau_n})\right\|^2 \le 4\beta_{N_T}(\delta)^2\mathrm{d}_{\mathrm{E},N_t}\left(3+\log\left(\frac{N_tH_\delta(N_t)}{16\beta_{N_t}(\delta)^4\mathrm{d}^2_{\mathrm{E},N_t}}\right)\right)$$
$$+ 2\mathrm{d}_{\mathrm{E},N_t}(1+2\beta_{N_t}(\delta)^2\mathrm{d}_{\mathrm{E},N_t})(1+H^2_\delta(N_t))\,.$$

To obtain the estimates of (15)–(16), it suffices to recall the definitions of $\beta_n(\delta)$ (i.e. (13)) and $H_\delta(n)$ (i.e. (31)). □

D Planning and Diffusive Limit Approximation

Our work builds upon [4], but with specialised results for our setting. This paper recovers the key results of this section (Propositions 4.4 to 4.6) under a stronger and more abstract set of assumptions. For the comfort of the reader we thus present the necessary steps to extend their results to our assumptions. Since our assumptions do not directly subsume theirs, we exhibit in each case from Assumptions 1 and 2 how to recover the keystone results which underpin the technical arguments of [4].

We begin by the well-posedness results for the pure jump case (Proposition 4.4) and the diffusive limit case (Proposition 4.5), and then focus on the approximation result linking the two regimes (Proposition 4.6). In [4], Proposition 4.4 corresponds to Theorem 2.3. and Remark 2.4. In Appendix D.1, we show how it follows from Assumptions 1 and 2 by proving the two intermediary results used in [4] to prove the result.

Proposition 4.4 (Adapted from [4, Thm. 2.3, Rem. 2.4.]). *Under Assumptions 1 and 2, there is $L_W \in \mathbb{R}_+$, independent of ε, such that for any $\theta \in \Theta$*

(i.) The map $x \mapsto \rho^*_\theta(x)$ is constant, taking only one value which we denote by $\rho^*_\theta \in \mathbb{R}$;

(ii.) There is an L_W-Lipschitz function W^*_θ such that

$$\varepsilon\rho^*_\theta = \max_{a\in\mathbb{A}} \{\mathbb{E}[W^*_\theta(x+\mu_\theta(x,a)+\Sigma\xi)] - W^*_\theta(x) + r(x,a)\} \ \forall x \in \mathbb{R}^d\,; \tag{17}$$

(iii.) There is $\pi^*_\theta \in \mathscr{A}$, such that for all $x \in \mathbb{R}^d$, $\pi^*_\theta(x)$ maximises the right hand side in (17), and $\pi^*_\theta \circ X^{\pi^*_\theta,\theta}$ is an optimal Markov control, i.e. $\rho^{\pi^*_\theta}_\theta(\cdot) \equiv \rho^*_\theta$.

In [4], Proposition 4.5 corresponds to Theorem 3.4. In Appendix D.2, we show that it also follows from Assumptions 1 and 2 by proving that [4, Assumption 5] holds under Assumptions 1 and 2.

Proposition 4.5 (Adapted from [4, Thm. 3.4.]). *Under Assumptions 1 and 2, for any $\theta \in \Theta$,*

(i.) The map $x \mapsto \bar\rho^*_\theta(x)$ is constant, taking only one value which we denote by $\bar\rho^*_\theta \in \mathbb{R}$.

(ii.) There is an L_W-Lipschitz function $\bar W^*_\theta \in \mathcal{C}^2(\mathbb{R}^d;\mathbb{R})$ such that

$$\bar\rho^*_\theta = \max_{a\in\mathbb{A}} \left\{\bar\mu_\theta(x,a)^\top \nabla \bar W^*_\theta(x) + \bar r(x,a)\right\} + \frac{1}{2}\operatorname{Tr}[\bar\Sigma\bar\Sigma^\top \nabla^2 \bar W^*_\theta(x)],\ \forall x \in \mathbb{R}^d\,. \tag{18}$$

(iii.) There is $\bar\pi^*_\theta \in \mathscr{A}$ such that, for all $x \in \mathbb{R}^d$, $\bar\pi^*_\theta(x)$ maximises the right hand side in (18), and $\bar\pi^*_\theta \circ \bar X^{\bar\pi^*_\theta,\theta}$ is an optimal Markov control, i.e. $\bar\rho^{\bar\pi^*_\theta}_\theta(\cdot) \equiv \bar\rho^*_\theta$.

Remark D.1. Proposition 4.5.(iii.) is not stated as is in [4, Thm. 3.4], but it follows from it by the same arguments as [4, Remark 2.4].

Propositions 4.4 and 4.5 together ensure that both the prelimit and limit regimes are well posed, while Proposition 4.6 gives the rate of convergence of the control problems along this limit. This result is essentially contained in the proof of [4, Thm. 3.6], but since its statement is different, we include a proof for completeness in Appendix D.3.

Proposition 4.6 (Adapted from [4, Thm. 3.6.]). *Under Assumptions 1 and 2, for any $\gamma \in (0,1)$, there is a constant $C_\gamma > 0$, independent of ε, such that, for any $\theta \in \Theta$,*

$$|\bar\rho^*_\theta - \rho^*_\theta| \le C_\gamma \varepsilon^{\frac{\gamma}{2}} \text{ and } \rho^*_\theta - \rho^{\bar\pi^*_\theta}_\theta(0) \le C_\gamma \varepsilon^{\frac{\gamma}{2}}\,. \tag{19}$$

Moreover, there is a function $e_\theta : \mathbb{R}^d \to \mathbb{R}$ such that,

$$\varepsilon\rho^{\bar\pi^*_\theta}_\theta(0) = \mathbb{E}[\bar W^*_\theta(x+\mu_\theta(x,a)+\Sigma\xi)] - \bar W^*_\theta(x) + r(x,\bar\pi^*_\theta(x)) + e_\theta(x)\,,\ \forall x \in \mathbb{R}^d \tag{20}$$

and there is $C'_\gamma > 0$, independent of ε, such that $|e_\theta(x)| \le C'_\gamma \varepsilon^{1+\frac{\gamma}{2}}(1+\|x\|^3)$ for all $x \in \mathbb{R}^d$.

D.1 Proof of Proposition 4.4

In [4], Theorem 2.3 and Remark 2.4 follow from Lemmas A.1 and A.2, which respectively give a mixing condition and a moment bound for $X^{\alpha,\theta}$. We already proved [4, Lemma A.2] in Lemma B.5. Moreover, Lemma D.2 which reproduced [4, Lemmas A.1] holds with only minor modifications of the proof from [4].

Lemma B.5. *Under Assumptions 1 and 2, for any $p \geq 2$, there is a constant $\mathfrak{c}'_p > 0$ independent of ε such that*

$$\mathbb{E}\left[\|X_t^{x_0,\alpha,\theta}\|^p\right] \leq \frac{1}{\ell_{\mathscr{V}}^p}\left(L_{\mathscr{V}}^p e^{-\frac{\mathfrak{c}_{\mathscr{V}}}{4}t}\|x_0\|^p + \frac{4\mathfrak{c}'_p}{\mathfrak{c}_{\mathscr{V}}}\left(1 - e^{-\frac{\mathfrak{c}_{\mathscr{V}}}{4}t}\right)\right),$$

for any $(x_0, \alpha, \theta) \in \mathbb{R}^d \times \mathcal{A} \times \Theta$ and $t \in [0, +\infty)$.

Lemma D.2. *For any $(x, x') \in \mathbb{R}^d \times \mathbb{R}^d$, $\theta \in \Theta$, and $\alpha \in \mathcal{A}$,*

$$\mathbb{E}\left[\|X_t^{x,\alpha,\theta} - X_t^{x',\alpha,\theta}\|\right] \leq \frac{L_{\mathscr{V}}}{\ell_{\mathscr{V}}}\|x - x'\|\, e^{-\mathfrak{c}_{\mathscr{V}} t}$$

for any $t \in [0, +\infty)$.

Proof. We can follow the proof of [4] using Assumption 2 directly without resorting to the higher order Lyapunov function ζ which they use. □

D.2 Proof of Proposition 4.5

Proposition 4.5, such as it is stated in [4, Thn 3.4.] relies on their Assumption 5. This assumption contains two conditions, which we will show respectively in Lemmas D.3 and D.4.

As detailed in [4, Remark 3.2.(i)], the first condition can be shown by proving an analogue of [4, Lemma A.1] for the diffusive limit process (22). In terms of arguments of the proof, this analogue requires only a change in the stochastic generator used in Itô's Lemma[5]. In the proof of Lemma D.3, we, therefore, show how to adapt [4, Lemma A.1] to the generator of the diffusion under Assumptions 1 and 2.

In the proof of [4, Lemma A.1], there are two key steps. First, study the discounted version of the control problem, and show that it is equi-Lipschitz continuous in the discount, which rests on the result in Lemma D.3. Then one takes the vanishing discount limit in the HJB equation using the theory of viscosity solutions to complete the proof.

Lemma D.3. *For any $(x_0, x'_0) \in \mathbb{R}^d \times \mathbb{R}^d$, $\theta \in \Theta$, $\alpha \in \mathcal{A}$,*

$$\mathbb{E}\left[\left\|\bar{X}_t^{x,\alpha,\theta} - \bar{X}_t^{x',\alpha,\theta}\right\|\right] \leq \frac{L_{\mathscr{V}}}{\ell_{\mathscr{V}}}\|x - x'\|\, e^{-\mathfrak{c}_{\mathscr{V}} t}$$

for any $t \in [0, +\infty)$.

Proof. If $x_0 = x'_0$, this is trivially true by pathwise-uniqueness, so we suppose $x_0 \neq x'_0$. Let us consider $(x_1, x_2) \in \mathbb{R}^d \times \mathbb{R}^d$ with $x_1 \neq x_2$. By a Taylor expansion in (4), we obtain as $\varepsilon \to 0$

$$(\bar{\mu}(x_1, a) - \bar{\mu}(x_2, a))^\top \nabla\mathscr{V}(x_1 - x_2) \leq -\mathfrak{c}_{\mathscr{V}}\mathscr{V}(x_1 - x_2)\,. \tag{59}$$

The Lyapunov function $\mathscr{V}$ is not differentiable at 0, so we will construct an approximating sequence for it. Let erf denote the error function and let $\mathscr{V}_\iota := \mathscr{V}\operatorname{erf}(\iota\mathscr{V})$ for $\iota > 0$. Note that $\mathscr{V}_\iota \in \mathcal{C}^1(\mathbb{R}^d; \mathbb{R}_+)$ and $\mathscr{V}_\iota$ is Lipschitz, let us show that it satisfies (59) everywhere.

Let $z := x_1 - x_2$. Since $z \neq 0$ we have

$$\nabla\mathscr{V}_\iota(z) = \nabla\mathscr{V}(z)\left(\operatorname{erf}(\iota\mathscr{V}(z)) + \frac{2\iota}{\sqrt{\pi}}\mathscr{V}(z)e^{-\iota^2\mathscr{V}^2(z)}\right).$$

[5] For a general overview of this sort of stability results and of Stochastic Lyapunov conditions in the diffusive case, see e.g. [22, § 5.7].

By Assumption 2, this implies that

$$(\bar{\mu}_\theta(x_1,a)-\bar{\mu}_\theta(x_2,a))^\top \nabla \mathscr{V}_\iota(z) \leq -\mathfrak{c}_{\mathscr{V}} \mathscr{V}(z)\mathrm{erf}(\iota \mathscr{V}(z)) - \frac{2\iota}{\sqrt{\pi}} \mathfrak{c}_{\mathscr{V}} \mathscr{V}(z)^2 e^{-\iota^2 \mathscr{V}^2(z)}$$
$$\leq -\mathfrak{c}_{\mathscr{V}} \mathscr{V}_\iota(z)\,. \tag{60}$$

Since $\nabla \mathscr{V}_\iota$ is continuous in z, and so is the right-hand side, we can let $\|z\| \to 0$ and conclude the bound also holds for $x_1 = x_2$.

We now apply Itô's lemma for the process $\bar{X}^{x,\alpha,\theta} - \bar{X}^{x',\alpha,\theta}$ to $\mathscr{V}_\iota$. Using (60), this yields, for $t \geq t_0 \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[\mathscr{V}_\iota\left(\bar{X}_t^{x_0,\alpha,\theta} - \bar{X}_t^{x_0',\alpha,\theta}\right)\right] \\
&\quad \leq \mathbb{E}\left[\mathscr{V}_\iota\left(\bar{X}_{t_0}^{x_0,\alpha,\theta} - \bar{X}_{t_0}^{x_0',\alpha,\theta}\right)\right] \\
&\quad + \mathbb{E}\left[\int_{t_0}^t \left(\bar{\mu}_\theta\left(\bar{X}_s^{x_0,\alpha,\theta}, \alpha_s\right) - \bar{\mu}_\theta\left(\bar{X}_s^{x_0',\alpha,\theta}, \alpha_s\right)\right)^\top \nabla \mathscr{V}_\iota\left(\bar{X}_s^{x_0,\alpha,\theta} - \bar{X}_s^{x_0',\alpha,\theta}\right) \mathrm{d}s\right] \\
&\quad \leq \mathbb{E}\left[\mathscr{V}_\iota\left(\bar{X}_{t_0}^{x_0,\alpha,\theta} - \bar{X}_{t_0}^{x_0',\alpha,\theta}\right)\right] - \int_{t_0}^t \mathfrak{c}_{\mathscr{V}} \mathbb{E}\left[\mathscr{V}_\iota\left(X_s^{x_0,\alpha,\theta} - X_s^{x_0',\alpha,\theta}\right)\right] \mathrm{d}s\,.
\end{aligned}$$

We conclude by the same ODE comparison argument as in the proof of Lemma B.5 and then pass to the limit as $\iota \to 0$ to obtain the claimed result using Assumption 2.(i). ☐

While Lemma D.3 showed that [4, Assumption 5.(i)] is implied by Assumptions 1 and 2. It remains now to verify their Assumption 5.(ii). Note that by [4, Remark 3.2.(ii)], an equation of the form of their (3.3) is sufficient to do so. Lemma D.4 gives exactly this result with (61), by noting that [4, (3.4)] holds by Assumption 2.

Lemma D.4. *Under Assumptions 1 and 2, for any $p \geq 2$ there are $(\bar{\mathfrak{c}}_p, \bar{\mathfrak{c}}_p') \in \mathbb{R}_+^2$ such that*

$$\bar{\mu}_\theta(x,a)^\top \nabla \mathscr{V}(x)^p + \mathrm{Tr}[\bar{\Sigma}\bar{\Sigma}^\top \nabla^2 \mathscr{V}(x)^p] \leq -\bar{\mathfrak{c}}_p \mathscr{V}(x)^p + \bar{\mathfrak{c}}_p' \tag{61}$$

for any $(x,a,\theta) \in \mathbb{R}^d \times \mathbb{A} \times \Theta$.

Proof. Let us take $(x,x') \in \mathbb{R}^d \times \mathbb{R}^d$ such that $\|x-x'\| \geq \varepsilon/(1-\varepsilon L_0)$, which implies $\|x - x' + \Delta(\mu_\theta(x,a) - \mu_\theta(x',a))\| > 0$ for any $\Delta \in [0,1]$ and for all $(a,\theta) \in \mathbb{A} \times \Theta$ and we can expand (4), which gives

$$-\varepsilon \mathfrak{c}_{\mathscr{V}} \mathscr{V}(x-x') \geq (\mu_\theta(x,a) - \mu_\theta(x',a))^\top \nabla \mathscr{V}(x-x')$$
$$+ \frac{1}{2}(\mu_\theta(x,a) - \mu_\theta(x',a))^\top \nabla^2 \mathscr{V}(\hat{x})(\mu_\theta(x,a) - \mu_\theta(x',a))\,,$$

in which $\hat{x} = x + \hat{\Delta}(x'-x)$ for some $\hat{\Delta} \in [0,1]$. Thus

$$(\bar{\mu}_\theta(x,a) - \bar{\mu}_\theta(x',a))^\top \nabla \mathscr{V}(x-x')$$
$$\leq -\mathfrak{c}_{\mathscr{V}} \mathscr{V}(x-x') - \frac{\varepsilon}{2}(\bar{\mu}_\theta(x,a) - \bar{\mu}_\theta(x',a))^\top \nabla^2 \mathscr{V}(\hat{x})(\bar{\mu}_\theta(x,a) - \bar{\mu}_\theta(x',a))\,.$$

Letting $\varepsilon \to 0$, the constraint on (x,x') vanishes as well as the second term (on compact sets), and we recover

$$(\bar{\mu}_\theta(x,a) - \bar{\mu}_\theta(x',a))^\top \nabla \mathscr{V}(x-x') + \frac{1}{2}\mathrm{Tr}[\bar{\Sigma}\bar{\Sigma}^\top \nabla^2 \mathscr{V}(x-x')] \leq -\mathfrak{c}_{\mathscr{V}} \mathscr{V}(x-x') + \frac{d}{2}\|\bar{\Sigma}\|_{\mathrm{op}}^2 M_{\mathscr{V}}'\,.$$

Taking $x' = 0$ implies that

$$\bar{\mu}_\theta(x,a)^\top \nabla \mathscr{V}(x) + \frac{1}{2}\mathrm{Tr}[\bar{\Sigma}\bar{\Sigma}^\top \nabla^2 \mathscr{V}(x)] \leq -\mathfrak{c}_{\mathscr{V}} \mathscr{V}(x) + C$$

for all $(x,a) \in \mathbb{R}_*^d \times \mathbb{A}$, in which $C := d\|\bar{\Sigma}\|_{\mathrm{op}}^2 M_{\mathscr{V}}'/2 + L_0 M_{\mathscr{V}}$.

Notice that, since $\mathscr{V} \in \mathcal{C}^2(\mathbb{R}^d_*; \mathbb{R}_+)$ and vanishes at 0 (see Assumption 1), $\mathscr{V}(\cdot)^p$ can be extended by continuity at 0 so that $\mathscr{V}(\cdot)^p \in \mathcal{C}^2(\mathbb{R}^d; \mathbb{R}_+)$. For any $(x, a, \theta) \in \mathbb{R}^d \times \mathbb{A} \times \Theta$, let

$$\begin{aligned}
k(x,a) :&= \bar{\mu}_\theta(x,a)^\top \nabla \mathscr{V}(x)^p + \frac{1}{2}\operatorname{Tr}\left[\bar{\Sigma}\bar{\Sigma}^\top \nabla^2 \mathscr{V}(x)^p\right] \\
&= p\bar{\mu}_\theta(x,a)^\top \nabla \mathscr{V}(x)\mathscr{V}(x)^{p-1} \\
&\quad + \frac{1}{2}\operatorname{Tr}\left[\bar{\Sigma}\bar{\Sigma}^\top \left(p\mathscr{V}(x)^{p-1}\nabla^2\mathscr{V}(x) + p(p-1)\mathscr{V}(x)^{p-2}\nabla\mathscr{V}(x)\nabla^\top\mathscr{V}(x)\right)\right] \\
&= p\mathscr{V}^{p-1}(x)\left(\bar{\mu}_\theta(x,a)^\top \nabla\mathscr{V}(x) + \frac{1}{2}\operatorname{Tr}[\bar{\Sigma}\bar{\Sigma}^\top\nabla^2\mathscr{V}(x)]\right) \\
&\quad + \frac{p(p-1)}{2}\mathscr{V}(x)^{p-2}\operatorname{Tr}[\bar{\Sigma}\bar{\Sigma}^\top\nabla\mathscr{V}(x)\nabla^\top\mathscr{V}(x)] \\
&\leq -p\mathfrak{c}_{\mathscr{V}}\mathscr{V}(x)^p + Cp\mathscr{V}(x)^{p-1} + \frac{dp(p-1)}{2}(\|\bar{\Sigma}\|_{\mathrm{op}}M_{\mathscr{V}})^2\mathscr{V}(x)^{p-2}
\end{aligned}$$

and we can now choose $\bar{\mathfrak{c}}_p = -p\mathfrak{c}_{\mathscr{V}}/2$, for which there exists a constant $\bar{\mathfrak{c}}'_p$ such that

$$-\bar{\mathfrak{c}}_p\mathscr{V}^p(x) + Cp\mathscr{V}^{p-1}(x) + \frac{dp(p-1)}{2}(\|\bar{\Sigma}\|_{\mathrm{op}}M_{\mathscr{V}})^2\mathscr{V}^{p-2}(x) \leq \bar{\mathfrak{c}}'_p$$

for all $x \in \mathbb{R}^d$. □

D.3 Proof of Proposition 4.6

The rest of this section is dedicated to showing Proposition 4.6 using modifications of the proof of [4, Thm. 3.6.] to which it corresponds. Here we produce a self-contained proof in order to clarify how (20) is derived from the proof.

Proposition 4.6 (Adapted from [4, Thm. 3.6.]). *Under Assumptions 1 and 2, for any $\gamma \in (0,1)$, there is a constant $C_\gamma > 0$, independent of ε, such that, for any $\theta \in \Theta$,*

$$|\bar{\rho}^*_\theta - \rho^*_\theta| \leq C_\gamma\varepsilon^{\frac{\gamma}{2}} \text{ and } \rho^*_\theta - \rho^{\bar{\pi}^*_\theta}_\theta(0) \leq C_\gamma\varepsilon^{\frac{\gamma}{2}}\,. \tag{19}$$

Moreover, there is a function $e_\theta : \mathbb{R}^d \to \mathbb{R}$ such that,

$$\varepsilon\rho^{\bar{\pi}^*_\theta}_\theta(0) = \mathbb{E}[\bar{W}^*_\theta(x + \mu_\theta(x,a) + \Sigma\xi)] - \bar{W}^*_\theta(x) + r(x, \bar{\pi}^*_\theta(x)) + e_\theta(x)\,,\ \forall x \in \mathbb{R}^d \tag{20}$$

and there is $C'_\gamma > 0$, independent of ε, such that $|e_\theta(x)| \leq C'_\gamma\varepsilon^{1+\frac{\gamma}{2}}(1 + \|x\|^3)$ for all $x \in \mathbb{R}^d$.

Proof. The first part of Proposition 4.6, i.e. (19), corresponds to [4, Thm. 3.6.], which we previously showed holds in our setting by verifying its assumptions. We now prove the second claim. Let

$$\delta r^\varepsilon_\theta(x,a) := \bar{\mu}_\theta(x,a)^\top\nabla\bar{W}^*_\theta(x) + \frac{1}{2}\operatorname{Tr}[\bar{\Sigma}\bar{\Sigma}^\top\nabla^2\bar{W}^*_\theta(x)] - \frac{1}{\varepsilon}\left(\mathbb{E}\left[\bar{W}^*_\theta(\psi^\varepsilon_\theta(x,a) + \Sigma\xi)\right] - \bar{W}^*_\theta(x)\right).$$

From (18), and Proposition 4.5.(iii.) we have

$$\begin{aligned}
\bar{\rho}^*_\theta &= \max_{a\in\mathbb{A}}\left\{\bar{\mu}_\theta(x,a)^\top\nabla\bar{W}^*_\theta + \frac{1}{2}\operatorname{Tr}[\bar{\Sigma}\bar{\Sigma}^\top\nabla^2\bar{W}^*_\theta(x)] + \bar{r}(x,a)\right\} \\
&= \bar{\mu}_\theta(x,\bar{\pi}^*_\theta(x))^\top\nabla\bar{W}^*_\theta(x) + \frac{1}{2}\operatorname{Tr}[\bar{\Sigma}\bar{\Sigma}^\top\nabla^2\bar{W}^*_\theta(x)] + \bar{r}(x,\bar{\pi}^*_\theta(x))
\end{aligned}$$

which implies

$$\varepsilon\rho^{\bar{\pi}^*_\theta}_\theta(0) = \mathbb{E}[\bar{W}^*_\theta(\psi^\varepsilon_\theta(x,\bar{\pi}^*_\theta(x)) + \Sigma\xi)] - \bar{W}^*_\theta(x) + r(x,\bar{\pi}^*_\theta(x)) + \varepsilon(\delta r^\varepsilon_\theta(x,\bar{\pi}^*_\theta(x)) + \bar{\rho}^*_\theta - \rho^{\bar{\pi}^*_\theta}_\theta(0)).$$

Note that $|\delta r^\varepsilon_\theta(x,\bar{\pi}^*_\theta(x))| \leq \sup_{a\in\mathbb{A}}|\delta r^\varepsilon_\theta(x,a)|$, which by [4, (3.10)] is bounded by $c_\gamma\varepsilon^{\frac{\gamma}{2}}(1 + \|x\|^3)$ for some constant $c_\gamma > 0$. An application of (19) yields

$$\bar{\rho}^*_\theta - \rho^{\bar{\pi}^*_\theta}_\theta(0) = \bar{\rho}^*_\theta - \rho^*_\theta + \rho^*_\theta - \rho^{\bar{\pi}^*_\theta}_\theta(0) \leq 2C_\gamma\varepsilon^{\frac{\gamma}{2}}$$

and, at the same time, $\bar{\rho}^*_\theta - \rho^{\bar{\pi}^*_\theta}_\theta(0) \geq \bar{\rho}^*_\theta - \rho^*_\theta \geq -C_\gamma\varepsilon^{\frac{\gamma}{2}}$. Therefore, there is a function $e_\theta : \mathbb{R}^d \to \mathbb{R}$ such that (20) holds, which also satisfies

$$|e_\theta(x)| \leq (2C_\gamma + c_\gamma)\varepsilon^{1+\frac{\gamma}{2}}(1 + \|x\|^3)\,.$$

□

E Regret Analysis

In this final appendix, we complete the analysis of the regret of Algorithm 1 and prove Theorem 3.1. First, we will give the regret decomposition, and then in the later sections we will bound terms one by one calling upon the results of the previous appendices.

Theorem 3.1. *Under Assumptions 1 and 2, for any $\delta \in (0,1)$, $x_0 \in \mathbb{R}^d$, and $\gamma \in (0,1)$, there is a pair $(C_\gamma, C) \in \mathbb{R}_+^2$ of constants independent of ε such that Algorithm 1 achieves*

$$R_T(\varpi) \leq 2C_\gamma \varepsilon^{\frac{\gamma}{2}} T + C\sqrt{\mathrm{d}_{E,T\varepsilon^{-1}} \log(\mathscr{N}_{T\varepsilon^{-1}}^{\varepsilon}) T \log(T\delta^{-1})} \tag{10}$$

with probability at least $1-\delta$, in which $\mathrm{d}_{E,T\varepsilon^{-1}}$ is the $2\varepsilon/\sqrt{T}$-eluder dimension (see [34, Def. 4.] and (56) *in Appendix C.2) of the class $\{\mu_\theta\}_{\theta\in\Theta}$ restricted to a ball of radius $\mathcal{O}(\sqrt{\log(T/\varepsilon)})$, and $\log(\mathscr{N}_{T\varepsilon^{-1}}^{\varepsilon})$ is the $\varepsilon^2\|\bar{\Sigma}\|_{\mathrm{op}}^2/T$-log-covering number of this same restricted class.*

E.1 Regret Decomposition

Recall that we defined $k : n \in \mathbb{N} \mapsto k(n)$ as the map associating to each event n the episode of Algorithm 1 in which they occur. Like in Section 4.4, let us define $\theta_n = \tilde{\theta}_{k(n)}$ for all $n \in \mathbb{N}$. The regret of Algorithm 1, which generates the control $\varpi \in \mathcal{A}$, is

$$\mathcal{R}_T(\varpi) := T\rho_{\theta^*}^* - \sum_{n=1}^{N_T} r(X_{\tau_n}^{\varpi,\theta^*}, \varpi_{\tau_n})$$

By definition of ϖ in Algorithm 1, $\varpi_{\tau_n} = \bar{\pi}_{\theta_n}^*(X_{\tau_n}^{\varpi,\theta^*})$, so that

$$\mathcal{R}_T(\varpi) := T\rho_{\theta^*}^* - \sum_{n=1}^{N_T} r(X_{\tau_n}^{\varpi,\theta^*}, \bar{\pi}_{\theta_n}^*(X_{\tau_n}^{\varpi,\theta^*}))$$

At the heart of the decomposition is the use of the HJB-type equation (20) applied for each n at the point $X_{\tau_n}^{\varpi,\theta^*}$. For clarity, let us introduce for all $n \in \mathbb{N}$ the random variable $\tilde{X}_{\tau_{n+1}}^{\varpi,\theta_n}$ equal in distribution, conditionally on $\mathcal{F}_{\tau_n}$, to the random variable $\psi_{\theta_n}^{\varepsilon}(X_{\tau_n}^{\varpi,\theta^*}, \varpi_{\tau_n}) + \Sigma\xi_{n+1}$. With this notation (20) becomes

$$\varepsilon\rho_{\theta_n}^{\bar{\pi}_{\theta_n}^*}(0) = \mathbb{E}[\bar{W}_{\theta_n}^*(\tilde{X}_{\tau_{n+1}}^{\varpi,\theta_n})|\mathcal{F}_{\tau_n}] - \bar{W}_{\theta_n}^*(X_{\tau_n}^{\varpi,\theta^*}) + r(X_{\tau_n}^{\varpi,\theta^*}, \bar{\pi}_{\theta_n}^*(X_{\tau_n}^{\varpi,\theta^*})) + e_{\theta_n}(X_{\tau_n}^{\varpi,\theta^*})\,. \tag{62}$$

This *imagined* evolution of the system represents the counterfactual induced by a single step transition at time τ_{n+1}, according to the belief in θ_n. With this notation, applying (62) yields

$$\begin{aligned}
\mathcal{R}_T(\varpi) &= T\rho_{\theta^*}^* - \sum_{n=1}^{N_T} \varepsilon\rho_{\theta_n}^{\bar{\pi}_{\theta_n}^*}(0) + \sum_{n=1}^{N_T} e_{\theta^*}(X_{\tau_n}^{\varpi,\theta^*}) + \sum_{n=1}^{N_T} \mathbb{E}[\bar{W}_{\theta_n}^*(\tilde{X}_{\tau_{n+1}}^{\varpi,\theta_n})|\mathcal{F}_{\tau_n}] - \bar{W}_{\theta_n}^*(X_{\tau_n}^{\varpi,\theta^*})\,. \\
&= (T - \varepsilon N_T)\rho_{\theta^*}^* && (R_1) \\
&\quad + \varepsilon\sum_{n=1}^{N_T}(\rho_{\theta^*}^* - \rho_{\theta_n}^{\bar{\pi}_{\theta_n}^*}(0)) + \sum_{n=1}^{N_T} e_{\theta^*}(X_{\tau_n}^{\varpi,\theta^*}) && (R_2) \\
&\quad + \sum_{n=1}^{N_T} \mathbb{E}[\bar{W}_{\theta_n}^*(\tilde{X}_{\tau_{n+1}}^{\varpi,\theta_n})|\mathcal{F}_{\tau_n}] - \bar{W}_{\theta_n}^*(X_{\tau_n}^{\varpi,\theta^*}). && (63)
\end{aligned}$$

The first term, (R_1), quantifies the deviation of the Poisson clock from its mean. On the other hand, (R_2) quantifies both the optimistic nature of Algorithm 1 and the approximation error of its approximate planning. The third term, (63), resembles a martingale (up to reordering), but it fails to be one on two key counts. First, the element from the family of functions $(\bar{W}_{\theta_n}^*)_{n\in\mathbb{N}}$ used at each step n changes. Second, the expectation terms are with respect to the counterfactual transitions $(\tilde{X}_{\tau_{n+1}}^{\varpi,\theta^*})_{n\in\mathbb{N}}$ while the random terms use the real transitions $(X_{\tau_{n+1}}^{\varpi,\theta^*})_{n\in\mathbb{N}}$.

Note that we can control the difference between the counterfactual and the real trajectory at a one-step time horizon, by using

$$\tilde{X}^{\varpi,\theta}_{\tau_{n+1}} \overset{\mathrm{d}}{=} X^{\varpi,\theta^*}_{\tau_{n+1}} - \mu_{\theta^*}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) + \mu_\theta(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) \,, \tag{64}$$

in which $\overset{\mathrm{d}}{=}$ denotes equality in the same conditionally distributional sense as above. By adding and subtracting relevant terms to exhibit the key quantities we get:

$$\begin{aligned}\sum_{n=1}^{N_T} \mathbb{E}[\bar{W}^*_{\theta_n}(\tilde{X}^{\varpi,\theta_n}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] - \bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}) \le& \sum_{n=1}^{N_T} \mathbb{E}[\bar{W}^*_{\theta_n}(\tilde{X}^{\varpi,\theta_n}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] - \mathbb{E}[\bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] \\ &+ \sum_{n=1}^{N_T} \mathbb{E}[\bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] - \mathbb{E}[\bar{W}^*_{\theta_{n+1}}(X^{\varpi,\theta^*}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] \\ &+ \sum_{n=1}^{N_T} \mathbb{E}[\bar{W}^*_{\theta_{n+1}}(X^{\varpi,\theta^*}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] - \bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}) \,.\end{aligned}$$

Using (64), and the uniform L_W-Lipschitzness of $(\bar{W}^*_{\theta_n})_{n\in\mathbb{N}}$, we get for each $n \in \mathbb{N}$

$$\mathbb{E}[\bar{W}^*_{\theta_n}(\tilde{X}^{\varpi,\theta_n}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] - \mathbb{E}[\bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] \le L_W \left\| \mu_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) - \mu_{\theta^*}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) \right\|$$

and thus the regret term (63) is bounded by

$$\sum_{n=1}^{N_T} \mathbb{E}[\bar{W}^*_{\theta_n}(\tilde{X}^{\varpi,\theta_n}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] - \bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}) \le R_3 + R_4 + R_5$$

in which

$$R_3 := L_W \sum_{n=1}^{N_T} \left\| \mu_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) - \mu_{\theta^*}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) \right\| \tag{R_3}$$

$$R_4 := \sum_{n=1}^{N_T} \mathbb{E}[\bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_{n+1}}) - \bar{W}^*_{\theta_{n+1}}(X^{\varpi,\theta^*}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] \tag{R_4}$$

$$R_5 := \sum_{n=1}^{N_T} \mathbb{E}[\bar{W}^*_{\theta_{n+1}}(X^{\varpi,\theta^*}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] - \bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}) \,. \tag{R_5}$$

At the end of this decomposition, we have constructed a true martingale in (R_5), which we bound in Appendix E.6. The first term (R_3) accumulates the fit error described in Proposition 4.3, up to the lazy updates, which we study in Appendix E.4. The term (R_4) is bounded by the number of effective updates of θ_n (namely, $\sum_{n=1}^{N_T} \mathbb{1}_{\{\theta_{n+1}\neq\theta_n\}}$) in Appendix E.5. Finally, the bounds on (R_1) and (R_2) are given in Appendices E.2 and E.3 respectively.

To combine the high-probability events used to bound (R_1) and (R_5), with the event of Proposition 4.2 used by the other terms, we will perform a union bound. This corresponds to the $\delta/3$ used in the definition of the confidence sets of Algorithm 1.

E.2 Bounding the Poisson clock variation term (R_1)

We bound (R_1) using Lemma E.1 which is a standard sub-exponential concentration result, see e.g. [14, Lemma 4.1]. It implies

$$\mathbb{P}\left(|T - \varepsilon N_T| \ge 2\sqrt{\varepsilon T \log\left(\frac{6}{\delta}\right)} \vee 2\varepsilon \log\left(\frac{6}{\delta}\right)\right) \le \frac{\delta}{3} \,.$$

Lemma E.1. *For any $T \in \mathbb{R}^*_+$ and $\delta \in (0,1)$,*

$$\mathbb{P}\left(|\varepsilon N_T - T| > 2\sqrt{\varepsilon T \log\left(\frac{2}{\delta}\right)} \vee 2\varepsilon \log\left(\frac{2}{\delta}\right)\right) \le \delta \,.$$

Proof. Let $\upsilon := \varepsilon^{-1}T$. For any $\lambda \in [-1,1]$, $\mathbb{E}[e^{\lambda(N_T-\upsilon)}] = \exp(\upsilon(e^{\lambda}-1-\lambda)) \leq e^{\lambda^2\upsilon}$. Therefore, N_T is $(\sqrt{2}\upsilon, 1)$-subexponential (see e.g. [14]) and therefore,

$$\mathbb{P}\left(|N_T - \upsilon| > \epsilon\right) \leq \begin{cases} e^{-\frac{\epsilon^2}{4\upsilon}} \text{ for } \epsilon \in (0, 2\upsilon] \\ e^{-\frac{\epsilon}{2}} \text{ for } \epsilon > 2\upsilon \end{cases},$$

which implies

$$\mathbb{P}\left(|N_T - \upsilon| > 2\sqrt{\upsilon \log\left(\frac{2}{\delta}\right)}\mathbb{1}_{\{\delta \geq e^{-\upsilon}\}} + 2\log\left(\frac{2}{\delta}\right)\mathbb{1}_{\{\delta \leq e^{-\upsilon}\}}\right) \leq \delta\,.$$

□

E.3 Bounding the optimistic approximation term (R_2)

There are two terms in (R_2). The second is the most straightforward as it can be bounded by applying the bound on e_{θ^*} of Proposition 4.6, which yields

$$\sum_{n=1}^{N_T} e_{\theta^*}(X^{\varpi,\theta^*}_{\tau_n}) \leq 2C'_\gamma N_T \varepsilon^{1+\frac{\gamma}{2}}(1 + \sup_{s\leq T}\|X^{\varpi,\theta^*}_s\|^3)\,.$$

We decompose the remaining term of (R_2) into

$$\begin{aligned}\varepsilon\sum_{n=1}^{N_T}(\rho^*_{\theta^*} - \rho^{\bar\pi^*_{\theta_n}}_{\theta_n}) &= \varepsilon\sum_{n=1}^{N_T}\left(\rho^*_{\theta^*} - \bar\rho^*_{\theta^*} + \bar\rho^{\bar\pi^*_{\theta^*}}_{\theta^*} - \bar\rho^{\bar\pi^*_{\theta_n}}_{\theta_n} + \bar\rho^*_{\theta_n} - \rho^*_{\theta_n} + \rho^*_{\theta_n} - \rho^{\bar\pi^*_{\theta_n}}_{\theta_n}\right) \\ &\leq 4N_T C_\gamma \varepsilon^{1+\frac{\gamma}{2}} + \varepsilon\sum_{n=1}^{N_T}\left(\bar\rho^{\bar\pi^*_{\theta^*}}_{\theta^*} - \bar\rho^{\bar\pi^*_{\theta_n}}_{\theta_n}\right)\end{aligned}$$

by applying Proposition 4.6 to all but the second pair of terms.

On the event of Proposition 4.2, with $\delta/3$ in place of δ, we have $\theta^* \in \cap_{n\in\mathbb{N}^*}\mathcal{C}_n(\delta/3)$ and thus, by definition of Algorithm 1, $\bar\rho^{\bar\pi^*_{\theta^*}}_{\theta^*} - \bar\rho^{\bar\pi^*_{\theta_n}}_{\theta_n} \leq 0$ for all $n \in \mathbb{N}^*$. Thus, on this event we have

$$\varepsilon\sum_{n=1}^{N_T}(\rho^*_{\theta^*} - \rho^{\bar\pi^*_{\theta_n}}_{\theta_n}) \leq 4N_T C_\gamma \varepsilon^{1+\frac{\gamma}{2}}\,.$$

E.4 Bounding the prediction error term (R_3)

Because of the lazy updates, $\mu_{\theta_n} = \mu_{\theta_{k(n)}}$ is chosen within $\mathcal{C}_{k(n)}(\delta/3)$ instead of $\mathcal{C}_n(\delta/3)$ preventing us from using directly Proposition C.7. Nevertheless, the lazy update-scheme is designed not to degrade the overall learning performance by more than a constant factor. Leveraging (7),

$$\sum_{i=1}^{n-1}\left\|\mu_{\theta_n}(X^{\varpi,\theta^*}_{\tau_i}, \varpi_{\tau_i}) - \mu_{\theta^*}(X^{\varpi,\theta^*}_{\tau_i}, \varpi_{\tau_i})\right\| \leq \begin{cases}2\beta_n(\delta/3) \text{ if } n < n_k \\ \beta_n(\delta/3) \text{ if } n = n_k\end{cases} \tag{65}$$

As a result, μ_{θ_n} is chosen within an inflated version of $\mathcal{C}_n(\delta/3)$, defined as in (6) but with $\beta_n(\delta/3)$ replaced by $2\beta_n(\delta/3)$. Thus, we can follow the same arguments as in the proof of Proposition 4.3, by applying Proposition C.7 to the inflated confidence sets, up to the constant factor 2 in the bounds. And therefore on the event of Proposition 4.2, we have

$$\begin{aligned}R_3 &= L_W\sum_{n=1}^{N_T}\left\|\mu_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) - \mu_{\theta^*}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n})\right\| \\ &\leq 6L_W\beta_{N_T}(\delta/3)\sqrt{\mathrm{d}_{\mathrm{E},N_t}} + L_W\mathrm{d}_{\mathrm{E},N_t}H_{\delta/3}(N_T)\,.\end{aligned}$$

E.5 Bounding the lazy-update term (R_4)

We observe that (R_4) is bounded by

$$
\begin{aligned}
R_4 &= \sum_{n=1}^{N_T} \mathbb{E}[\bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_{n+1}}) - \bar{W}^*_{\theta_{n+1}}(X^{\varpi,\theta^*}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] \\
&\leq 2L_W \sum_{n=1}^{N_T} \mathbb{E}\left[\left(1 + \|X^{\varpi,\theta^*}_{\tau_{n+1}}\|\right) \mathbb{1}_{\{\theta_n \neq \theta_{n+1}\}} |\mathcal{F}_{\tau_n}\right] \\
&\leq 2L_W \sum_{n=1}^{N_T} \left((1+\varepsilon L_0)(1 + \|X^{\varpi,\theta^*}_{\tau_n}\|) + \varepsilon^{\frac{1}{2}} \|\bar{\Sigma}\|_{\mathrm{op}} \mathbb{E}\left[\|\xi_{n+1}\| \,|\mathcal{F}_{\tau_n}\right]\right) \mathbb{1}_{\{\theta_n \neq \theta_{n+1}\}} \\
&\leq 2L_W(1+\varepsilon L_0)\left(1 + \sup_{s\leq T}\|X^{\varpi,\theta^*}_s\| + \sqrt{d}\varepsilon^{\frac{1}{2}}\|\bar{\Sigma}\|_{\mathrm{op}}\right) \sum_{n=1}^{N_T} \mathbb{1}_{\{\theta_n \neq \theta_{n+1}\}} \,.
\end{aligned}
$$

Thus bounding the number of updates with Lemma E.2 bounds (R_4).

Lemma E.2. *Under Assumptions 1 and 2, Algorithm 1 generates episodes which satisfy for all* $T \in \mathbb{R}_+$ *and* $\delta \in (0,1)$

$$
\begin{aligned}
\sum_{n=1}^{N_T} \mathbb{1}_{\{\theta_n \neq \theta_{n+1}\}} \leq 4\beta_{N_T}(\delta/3)^2 \mathrm{d}_{\mathrm{E},N_t} &\left(3 + \log\left(\frac{N_t \sup_{s\leq t}\|X^{\varpi,\theta^*}_s\|}{16\beta_{N_t}(\delta/3)^4 \mathrm{d}^2_{\mathrm{E},N_t}}\right)\right) \\
&+ 2\mathrm{d}_{\mathrm{E},N_t}(1 + 2\beta_{N_t}(\delta/3)^2 \mathrm{d}_{\mathrm{E},N_t})(1 + \sup_{s\leq t}\|X^{\varpi,\theta^*}_s\|^2) \,.
\end{aligned}
$$

Proof. Consider $k \in \mathbb{N}^*$, by (7), each time we trigger an update we have

$$
\begin{aligned}
2\beta_{n_k}(\delta/3)^2 &< \sup_{\mu_\theta \in \mathcal{C}_{n_{k-1}}(\delta)} \left\|\mu_\theta - \mu_{\hat{\theta}n_{k-1}}\right\|^2_{n_k} \\
&\leq \sup_{\mu_\theta \in \mathcal{C}_{n_{k-1}}(\delta)} \left\|\mu_\theta - \mu_{\hat{\theta}n_{k-1}}\right\|^2_{n_{k-1}} \\
&\quad + \sup_{\mu_\theta \in \mathcal{C}_{n_{k-1}}(\delta)} \sum_{n=n_{k-1}+1}^{n_k} \left\|\mu_\theta(X^{\varpi,\theta}_{\tau_n}, \varpi_{\tau_n}) - \mu_{\hat{\theta}n_{k-1}}(X^{\varpi,\theta}_{\tau_n}, \varpi_{\tau_n})\right\|^2 \\
&\leq \beta_{n_k}(\delta/3)^2 + \sum_{n=n_{k-1}+1}^{n_k} \Lambda(\mathcal{C}_{n_{k-1}}(\delta/3); X^{\varpi,\theta}_{\tau_n}, \varpi_{\tau_n})^2 \,.
\end{aligned}
$$

Summing over all episodes, since the sequence $(\beta_n(\delta/3))_{n\in\mathbb{N}}$ is non-decreasing, we have that for all $T \in \mathbb{R}_+$

$$
\sum_{n=1}^{N_T} \Lambda(\mathcal{C}_{n_k}(\delta/3); (X_{\tau_n}, \varpi_{\tau_n}))^2 \geq \sum_{k=1}^{K_T} \beta_{n_k}(\delta/3)^2 \geq K_T \beta_0(\delta/3)^2 \,,
$$

in which $K_T := k(N_T) \in \mathbb{N}$ is the number of episodes by time T. An application of the second part of Proposition C.7, i.e. (58) now yields the desired result as $\beta_0(\delta/3)^2 = \varepsilon$. □

E.6 Bounding the martingale term (R_5)

Let

$$
Z_n := \mathbb{E}[\bar{W}^*_{\theta_n}(X^{\alpha,\theta^*}_{\tau_n})|\mathcal{F}_{n-1}] - \bar{W}^*_{\theta_n}(X^{\alpha,\theta^*}_{\tau_n}) \,.
$$

By definition

$$
R_5 = \mathbb{E}[\bar{W}^*_{\theta_{N_T+1}}(X^{\varpi,\theta^*}_{\tau_{N_T+1}})|\mathcal{F}_{\tau_{N_T}}] + \bar{W}^*_{\theta_0}(x_0) + \sum_{n=1}^{N_T} Z_n \,.
$$

On the one hand, Z_n is a $L_W\|\Sigma\|_{\mathrm{op}}$-Lipschitz function of ξ_n, which is Gaussian and of mean 0. Therefore, by [13, Thm 5.5], Z_n is $L_W\,\|\Sigma\|_{\mathrm{op}}$-sub-Gaussian and

$$\mathbb{P}\left(\sum_{n=1}^{N_T} Z_n > L_W\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{2\varepsilon N_T\log\left(\frac{1}{\delta}\right)}\right) \le \delta\,. \tag{66}$$

On the other hand, by the uniform Lipschitzness of $(\bar{W}^*_\theta)_{\theta\in\Theta}$, $\bar{W}^*_{\theta_0}(x_0) \le L_W(1+\|x_0\|)$ and

$$\begin{aligned}\mathbb{E}[\bar{W}^*_{\theta_{N_T+1}}(X^{\varpi,\theta^*}_{\tau_{N_T+1}})|\mathcal{F}_{\tau_{N_T}}] &\le L_W(1+\mathbb{E}[\|X^{\varpi,\theta^*}_{\tau_{N_T+1}}\||\mathcal{F}_{\tau_{N_T}}]) \\ &\le L_W(1+\varepsilon L_0+(1+\varepsilon L_0)\|X^{\varpi,\theta^*}_{\tau_{N_T}}\|+\varepsilon^{\frac{1}{2}}\|\bar{\Sigma}\|_{\mathrm{op}}\mathbb{E}[\|\xi_{N_T+1}\|\,|\mathcal{F}_{\tau_{N_T}}]) \\ &\le L_W(1+\varepsilon L_0)\left(1+\sup_{s\le T}\|X^{\varpi,\theta^*}_s\|_2+\varepsilon^{\frac{1}{2}}\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{d}L_W\right). \end{aligned}\tag{67}$$

Combining (66) and (67) yields

$$R_5 \le L_W\left\|\bar{\Sigma}\right\|_{\mathrm{op}}\sqrt{2\varepsilon N_T\log\left(\frac{3}{\delta}\right)}+2L_W(1+\varepsilon L_0)(1+\sup_{s\le T}\|X^{\varpi,\theta^*}_s\|+\varepsilon^{\frac{1}{2}}\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{d}L_W) \tag{68}$$

with probability at least $1-\delta/3$.

E.7 Collecting the bounds

We conclude the proof of Theorem 3.1 by collecting all the terms from Appendices E.2–E.6 and simplifying them. By a union bound over the events listed in steps Appendices E.2, E.4 and E.6, with probability at least $1-\delta$

$$\begin{aligned}\mathcal{R}_T(\varpi) \le\ & 2L_0\left(\sqrt{\varepsilon T\log\left(\frac{6}{\delta}\right)}\vee 2\varepsilon\log\left(\frac{6}{\delta}\right)\right) \\ &+4N_T C_\gamma\varepsilon^{1+\frac{\gamma}{2}}+2C'_\gamma N_T\varepsilon^{1+\frac{\gamma}{2}}(1+H^3_{\delta/3}(N_T)) \\ &+6L_W\beta_{N_T}(\delta/3)\sqrt{\mathrm{d}_{\mathrm{E},N_T}}+L_W\mathrm{d}_{\mathrm{E},N_T}H_{\delta/3}(N_T) \\ &+2L_W(1+\varepsilon L_0)\left((1+H_{\delta/3}(N_T)+d\varepsilon^{\frac{1}{2}}\|\bar{\Sigma}\|_{\mathrm{op}}\right)\left(4\beta_{N_T}(\delta/3)^2\mathrm{d}_{\mathrm{E},N_t}\left(3\right.\right. \\ &\left.\left.+\log\left(\frac{N_t H_{\delta/3}(N_T)}{16\beta_{N_T}(\delta/3)^4\mathrm{d}^2_{\mathrm{E},N_T}}\right)\right)+2\mathrm{d}_{\mathrm{E},N_T}(1+2\beta_{N_t}(\delta/3)^2\mathrm{d}_{\mathrm{E},N_T})(1+H_{\delta/3}(N_T)^2)\right) \\ &+L_W\left\|\bar{\Sigma}\right\|_{\mathrm{op}}\sqrt{2\varepsilon N_T\log\left(\frac{3}{\delta}\right)}+2L_W(1+\varepsilon L_0)(1+H_{\delta/3}(N_T)+\varepsilon^{\frac{1}{2}}\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{d}L_W).\end{aligned}$$

This can be more simply expressed for some constants $C^{(i)}_{\mathcal{R}}\in\mathbb{R}_+$, $i\in[5]$, as

$$\begin{aligned}\mathcal{R}_T(\varpi) \le\ & C^{(1)}_{\mathcal{R}}(C_\gamma+C'_\gamma)\varepsilon^{1+\frac{\gamma}{2}}N_T\log(N_T)^3+C^{(2)}_{\mathcal{R}}\sqrt{\mathrm{d}_{\mathrm{E},N_T}\varepsilon N_T\log\left(\frac{N_T(1+\varepsilon\mathscr{N}^{\varepsilon}_{N_T})}{\delta}\right)} \\ &+C^{(3)}_{\mathcal{R}}\left(1+\varepsilon\mathrm{d}_{\mathrm{E},N_T}\log(N_T)\log(N_T(1+\varepsilon\mathscr{N}^{\varepsilon}_{N_T}))\right)\mathrm{d}_{\mathrm{E},N_T}\log(N_T)^4 \\ &+C^{(4)}_{\mathcal{R}}\sqrt{\varepsilon T\log\left(\frac{1}{\delta}\right)}+C^{(5)}_{\mathcal{R}}\left(1+\log\left(\frac{1}{\delta}\right)\right)\end{aligned}$$

still with probability at least $1-\delta$. On this high-probability event we can write $\mathcal{R}_T(\varpi)$ (up rounding up $T\varepsilon^{-1}$ where necessary and up to a change in the constants) as

$$\mathcal{R}_T(\varpi) \le C_{\mathcal{R}}^{(1)}(C_\gamma + C'_\gamma)\varepsilon^{\frac{\gamma}{2}} T \log\left(\frac{T}{\varepsilon}\right) + C_{\mathcal{R}}^{(2)}\sqrt{\mathrm{d}_{\mathrm{E},T\varepsilon^{-1}} T \log\left(\frac{T\varepsilon^{-1}(1+\varepsilon\mathscr{N}_{T\varepsilon^{-1}})}{\delta}\right)}$$
$$+ C_{\mathcal{R}}^{(3)}\Big(1+\varepsilon \mathrm{d}_{\mathrm{E},T\varepsilon^{-1}} \log(T\varepsilon^{-1})\log(T\varepsilon^{-1}(1+\varepsilon\mathscr{N}^{\varepsilon}_{T\varepsilon^{-1}}))\Big)\mathrm{d}_{\mathrm{E},T\varepsilon^{-1}}\log(T\varepsilon^{-1})^4$$
$$+ C_{\mathcal{R}}^{(4)}\sqrt{\varepsilon T \log\left(\frac{1}{\delta}\right)} + C_{\mathcal{R}}^{(5)}\left(1+\log\left(\frac{1}{\delta}\right)\right) .$$

Considering only the two dominant terms and ignoring logarithmic factors we get the claimed bound.