Diffusive Limit Control and Reinforcement Learning

Lorenzo Croissant

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Université Paris Dauphine-PSL

Diffusive Limit Control and Reinforcement Learning

with applications to online auctions

Soutenue par
Lorenzo CROISSANT
Le 08 Décembre 2023

École doctorale n°543
École doctorale SDOSE

Spécialité
Mathématiques

Composition du jury :

Mathieu ROSENBAUM Professeur des universités, École Polytechnique	*Président*
Jean-François CHASSAGNEUX Professeur des universités, Université Paris Cité	*Rapporteur*
Vianney PERCHET Professeur des universités, ENSAE	*Rapporteur*
Pierre CARDALIAGUET Professeur des universités, Université Paris Dauphine-PSL	*Examinateur*
Marc HOFFMANN Professeur des universités, Université Paris Dauphine-PSL	*Examinateur*
Charles-Albert LEHALLE Global head - quantitative research & development Abu Dhabi investment authority	*Examinateur*
Athena PICARELLI Associate professor, Università di Verona	*Examinatrice*
Bruno Bouchard Professeur des universités, Université Paris Dauphine-PSL	*Directeur de thèse*
Marc ABEILLE Staff researcher, Criteo	*Invité*

Dauphine | PSL
UNIVERSITÉ PARIS

To Rachel.

Diffusive Limit Control and Reinforcement Learning

with applications to online auctions

LORENZO CROISSANT

Université Paris Dauphine-PSL

Acknowledgments

A Ph.D. is a long and arduous pilgrimage and, as I now look back, I would like to pay homage to those who were a part of it.

First, I wish to thank those that this manuscript and myself had the honour to have as reviewers, Jean-François Chassagneux and Vianney Perchet, for the care they took in reading my manuscript and their insightful comments. Likewise, I would like to thank my examiners Pierre Cardaliaguet, Marc Hoffmann, Charles-Albert Lehalle, Athena Picarelli, and Mathieu Rosenbaum for dedicating their time to my work and coming to the defence.

For guiding me down the whole path, I am immensely grateful to my supervisors, Bruno Bouchard at Dauphine and Marc Abeille at Criteo. Their rigour, insight, and above all their high expectations pushed me to grow to where I am today, both as a mathematician and to a large extent as a human being. Their constant support and encouraging hand never failed me, and this investment of time and care has touched me profoundly. For all of this, I must remain ever in their debt. I hope to live up to their example.

I was immensely lucky to participate in multiple research groups during my thesis, which was an incredibly enriching experience.

First, at Criteo I had the joy of working within the EEL team. I would like to thank Jeremie, Louis, and Thibaut for the good times, Morgane for many funny tea breaks about literature, and finally Clément for fostering this great environment through thick and thin. Thanks also to the members I had the joy of working alongside: Othman, Ilana, Skander, Ludovic, Ugo, Jean-Yves et alia.

Second, the inseparable FairPlay team, whose weekly reading groups and group meetings allowed me to discover so much great mathematics. Thanks to all the students for their unending enthusiasm: Mathieu, Côme, Sasila, Ziyad, Hafedh, Mike, et alia. Watching you grow is a daily pleasure. Thanks to the postdocs for the many great and leisurely discussions: Felipe, Dorian, Nadav, and Hugo. For it all, the smiles and the laughter, my thanks to Vianney and Patrick.

The CEREMADE was a haven for me during the COVID times and beyond. A great thank you to the B227 office, its OGs Theo, Charly, and João, and the new kids Antoine, Antoine, and Ugo, for the best of times. Many thanks also to the lifelong friends I made. The older crop: Giovanni, Kathi, Rodrigue, Adechola, Ruihua, Lukasz et alia. But also the younger generations: Daniele (and Eugi), Umberto, et alia, and

all those I don't have the room to list. Many thanks also to Zhenjie, Yating, Idriss, and Emeric for their advice, mentoring, and for the good times.

Finally, I would like to thank the members of the FiME group and EDF R&D for the many fruitful discussions: Adrien, Nadia, Olivier, and Damien who organised so many great events which I had the chance to attend.

For granting me the opportunity to participate in all of this I must also thank Criteo for financing my Ph.D. thesis.

I must also recognise those who helped me start the journey: my master's comrades Louis, Erwan, Baptiste, and Bastien, and especially Solange and Flore who stepped on the path at the same time as me and were always a great source of joy, I hope I was as positive a force for them as they were for me. I am very grateful to Azadeh for tipping the scales at perhaps the most crucial moment of this story and encouraging me to take the very first step in this long journey over 5 years ago.

Outside the ivory towers of universities, I also wish to thank those friends who are still here after a good fifteen years. Thomas, Alexis, Maximillien, Edwan and Galaad, thanks for all the nonsense. And also Antoine, Victor, Leo, Thomas, Raphaël, Charles, Alexandre, Louis Marie not to be undone in this regard. My only wish is for us to share fifteen more! Finally, for fostering curiosity from a young age, a great thanks to my parents and family.

And above all, for always putting the spring in my every step: to Rachel.

This manuscript was produced using the memoir class, and the PSL thesis cover code by Pierre Guillou. It could not have taken this shape without the work of numerous LaTeX package developers, whom I collectively thank for their impressive work.

Abstract

We consider the diffusive limit of a generic pure-jump Markov control problem as the intensity of the driving Poisson process tends to infinity. We quantify the convergence speed in terms of the Hölder exponent of the Hessian of the limit problem. This approximation provides an efficient alternative method for the numerical resolution of these problems when the intensity of jumps is large. Considering the control of unknown systems, we extend this approach to the context of online Reinforcement Learning. Under the Optimism in the Face of Uncertainty paradigm, we leverage the eluder dimension framework for learning and the diffusive limit for approximate resolution of the planning subproblem. Our algorithm extends existing theory from discrete processes to continuous states and actions. Our study of diffusion limit systems is motivated and illustrated by the bidding problem in a high-frequency online auction against a revenue-maximising seller.

Résumé

On considère la limite diffusive d'un problème de contrôle Markovien à sauts purs quelconque lorsque l'intensité de son processus de Poisson tend vers l'infini. On quantifie la vitesse de convergence en fonction de l'exposant de Hölder de l'Hessienne du problème limite. Cette approximation fournit une méthode alternative efficace pour la résolution numérique de ces problèmes lorsque l'intensité des sauts est grande. On s'attache ensuite au problème de l'incertitude dans les systèmes de contrôle, et on étend notre étude au contexte de l'apprentissage par renforcement en ligne. Dans le paradigme de l'optimisme devant l'incertain, on exploite le carcan de la dimension d'eluder pour gérer l'apprentissage et la limite diffusive pour résoudre approximativement le sous-problème de planification. Notre algorithme étend la théorie existante des problèmes discrets aux problèmes avec états et actions continus. Notre étude des systèmes à limite diffusive est motivée et illustrée par le problème d'enchérir dans une enchère séquentielle à haute fréquence contre un vendeur qui maximise son revenu sous contrainte d'utiliser une règle de mise à jour en temps réel.

Contents

Chapter 1

Introduction

The development of automatics in engineering has placed dynamical systems at the heart of most industrial technologies. Early successes were exemplified by manufacturing processes, rocket technology, and other heavy industrial applications. The development of information technologies and the internet has broadened the reach of this trend, with automated dynamical systems becoming ubiquitous in most, if not all, aspects of our everyday lives.

In many of these real-world cases, the problems related to system noise and uncertainty are inevitable. In the context of automation in dynamical systems, accounting for noisy dynamics mathematically has led to the development of Stochastic Control Theory, which has blossomed and branched out since the second half of the twentieth century.

The branch of continuous (in time and space) stochastic control has developed a rich and successful theory, based on Partial Differential Equation (PDE), starting with the works of Pontryagin's research group in the 1960s; see [59] for a bibliographical overview. This PDE-based approach was significantly enriched in the 1980s by the development of the theory of viscosity solutions [48, 49], allowing for many complex and irregular problems to be solved. These new tools led to significant success in applications, most notably within the domain of Financial Mathematics, e.g. [56].

Despite this success when incorporating randomness, control systems in the real world must still account for the problem of uncertainty in the system dynamics. In the continuous-time framework, this problem was addressed by the fields of Adaptive and Robust Control, see for instance [30, 31]. While Robust Control seeks to be resistant to uncertainty, Adaptive Control leverages Statistics in order to learn to control the system. The classical statistical methods used in this domain (e.g. maximum likelihood estimation) and their focus on asymptotic convergence results contrast with more recent Machine

Learning methods.

The branch of discrete (in time and space) Stochastic Control, see e.g. [22, 102], has grown to address the problem of uncertainty in the controlled system from the Machine Learning perspective with Reinforcement Learning, see e.g. [23, 112]. This perspective focuses on task performance, computational efficiency, and finite-time guarantees rather than explanatory modelling.

The field of Reinforcement Learning acquired a great deal of notoriety in the 2010s, with the development of Deep Reinforcement Learning [88]. Some practical applications have been widely publicised, most notably AlphaGo [111] which achieved super-human performance in the game of Go.

These Deep Reinforcement Learning methods heavily leverage statistical methods and thus require intensive exploration efforts to generate data. This implies using control policies that forcibly explore the system. Such policies are inherently in conflict with the goal of the control problem, leading to wasteful behaviour. The study of this trade-off between exploration and exploitation is the focus of the field of Reinforcement Learning Theory.

By using the notion of regret to quantify this trade-off, Reinforcement Learning Theory has developed a fine analysis of the precise challenges of efficient exploration of unknown systems when the end goal is control [16]. By allowing us to merge the conflicting objectives of the statistical learning problem and the control problem, this analysis laid the foundations for the construction of optimal exploration methods for Reinforcement Learning [68].

Unfortunately, owing to its origins in discrete control, the field of Reinforcement Learning Theory has struggled to extend its methods to complex continuous systems, limiting its real-world applicability. The primary hurdle that must be overcome is the analysis and numerical resolution of the control problem on a continuous state-space, both of which are well-studied problems in Stochastic Control.

This observed complementarity between the two fields, stemming from their shared genealogy, suggests that the Machine Learning framework in general, and Reinforcement Learning in particular, holds a lot of promise for the rejuvenation of the problem of learning to control dynamical systems. Conversely, the tools of Stochastic Control Theory have the potential to unlock some of the challenges faced by Reinforcement Learning Theory in the continuous setting.

The main objective of this thesis is to demonstrate this complementarity by solving the continuous-state Reinforcement Learning problem for a category of systems that admit a diffusive limit by leveraging continuous stochastic control both for analytical tools and for computationally efficient approximations. Owing to this dual-field nature, one part of this thesis is primarily control-focused, which we describe in Section 1.1, while another

part focuses more resolutely on Reinforcement Learning, which we describe in Section 1.2, leverages these results. A third part, which is transversal to the two others, is described in Section 1.3 and focuses on a typical use-case based on the problem of high-frequency online auctions in advertising, thus grounding the theoretical approach in a practical setting.

1.1 Diffusive Limit Approximations

Learning Theory is predicated on observing single events (samples) and is thus naturally constructed on discrete-time processes. On the other hand, continuous Control Theory by definition studies processes defined on $[0, T]$, for $T > 0$, or $\mathbb{R}_+$. In order to reconcile these two perspectives, we choose to base our control problem on a pure-jump process whose events arrive according to a Poisson process.

Let N be a random measure on $\mathbb{R}_+ \times \mathbb{R}^{d'}$ with compensator $\eta\nu(\mathrm{d}e)\mathrm{d}t$, for some probability measure ν on $\mathbb{R}^{d'}$, $d' \in \mathbb{N}^*$, $\eta > 0$. Given an initial time $t \in [0, T]$, an initial condition $x \in \mathbb{R}^d$, $d \in \mathbb{N}^*$, and a measurable map $(x, a, e) \in \mathbb{R}^d \times \mathbb{A} \times \mathbb{R}^{d'} \mapsto b(x, a, e) \in \mathbb{R}^d$, we consider the solution to the following Stochastic Differential Equation

$$X^{t,x,\alpha} = x + \int_t^{\cdot} \int_{\mathbb{R}^{d'}} b(X_{s-}^{t,x,\alpha}, \alpha_s, e) N(\mathrm{d}e, \mathrm{d}s)\,, \tag{1.1}$$

in which the control process α belongs to the set $\mathcal{A}$ of $\mathbb{A}$-valued predictable processes, for some $\mathbb{A} \subset \mathbb{R}^{d_{\mathbb{A}}}$ compact, $d_{\mathbb{A}} \in \mathbb{N}^*$.

This choice of state dynamics makes the motion of the system evident: the agent observes the current state of the system and chooses an action based on past information. After waiting an exponentially distributed time, an event happens and the state of the system jumps to a random point dependent on the current state and the action chosen. This Markov nature of the system is crucial.

At the same time, this type of random waiting time corresponds naturally to many digital systems, whose state changes are driven by exogenous events. Typical examples of this kind of system are given by queueing problems and by online advertising auctions. Low values of η cause the system to behave more like a discrete-time system (with a random number of events N_T close to ηT), while $\eta \to +\infty$ corresponds to a fully continuous-time system.

The scale of some internet-based digital systems has grown to regimes under which η is large enough that the system can be realistically approximated by a continuous-time diffusive system. This is notably the case for online advertising auctions, which we use as an example throughout and give an overview of in Section 1.3.

1.1.1 Diffusive limits in stochastic control

Consider the finite-horizon control problem, which is defined for $(t,x) \in [0,T] \times \mathbb{R}^d$ by

$$V_T(t,x) := \sup_{\alpha \in \mathcal{A}} \left[\int_t^T r(X_{s-}^{t,x,\alpha}, \alpha_s) \mathrm{d}N_s \right] \tag{1.2}$$

in which $N_t := N(\mathbb{R}^{d'}, [0,t])$ and $(x,a) \in \mathbb{R}^d \times \mathbb{A} \mapsto r(x,a) \in \mathbb{R}$ is a reward function. Under some regularity conditions on the coefficients b and r (say, bounded and Lipschitz in x uniformly in other arguments), we can show that V_T is the unique bounded viscosity solution to the Hamilton-Jacobi-Bellman (HJB) equation

$$0 = \partial_t V_T + \sup_{a \in \mathbb{A}} \left\{ \mathcal{L}^a V_T + \eta r(\cdot, a) \right\} \text{ on } [0,T) \times \mathbb{R}^d, \tag{1.3}$$

with the boundary condition $V_T(T,\cdot) = 0$ on $\mathbb{R}^d$, in which, given a function $\phi \in \mathcal{C}_b^0([0,T) \times \mathbb{R}^d; \mathbb{R})$, the operator

$$\mathcal{L}^a : \phi \mapsto \eta \int_{\mathbb{R}^{d'}} [\phi(\cdot, \cdot + b(\cdot, a, e)) - \phi] \nu(\mathrm{d}e) \in \mathcal{C}_b^0([0,T) \times \mathbb{R}^d; \mathbb{R}) \tag{1.4}$$

is the infinitesimal generator (see e.g. [72, § 9.7]) associated to (1.1) with $\alpha \equiv a$.

Because (1.4), and thus (1.3), is non-local, analysis of (1.3) is difficult: both in regards to regularity estimates and numerical resolution. Furthermore, this effect is compounded when η grows large: numerical integration error and $\|\partial_t V_T\|$ both scale with η.

Many practical systems have large values of η, but maintain some degree of structure at the macroscopic scale, due to having proportionally small increments. Perhaps the most well-known example of this type of phenomenon is in financial markets, e.g. [52, 67]. In the case of (1.1), there are several possible scaling regimes, such as the fluid limit regime of [54], but we choose to focus on the diffusive limit regime because it preserves the stochasticity of the process as $\eta \to +\infty$.

In this regime, we let $\eta = \eta_\varepsilon := \varepsilon^{-1}$ for $\varepsilon > 0$, and assume the coefficient scales as $b = \varepsilon b_1 + \varepsilon^{1/2} b_2$, for some locally bounded maps b_1 and b_2 such that $\int b_2(\cdot, e) \nu(\mathrm{d}e) = 0$ and $\inf \int b_2(\cdot, e)^2 \nu(\mathrm{d}e) > 0$. As $\varepsilon \downarrow 0$, the process $X^{t,x,\alpha}$ of (1.1) converges to a controlled diffusion process, see [66]. For $(t,x) \in [0,T) \times \mathbb{R}^d$ and $\bar{\alpha} \in \bar{\mathcal{A}}$, with $\bar{\mathcal{A}}$ the set of $\mathbb{A}$-valued processes predictable relative to the filtration of a d-dimensional Wiener process W, this diffusion is given by the (unique) strong solution to the Stochastic Differential Equation (SDE)

$$\bar{X}^{t,x,\alpha} = \int_t^\cdot \mu(\bar{X}_s^{t,x,\alpha}, \alpha_s) \mathrm{d}s + \int_t^\cdot \sigma(\bar{X}_s^{t,x,\alpha}, \alpha_s) \mathrm{d}W_s \,, \tag{1.5}$$

whose coefficients are given by

$$\mu := \int_{\mathbb{R}^{d'}} b_1(\cdot, e)\nu(\mathrm{d}e)\,,\ \sigma := \left(\int_{\mathbb{R}^{d'}} b_2(\cdot, e)b_2(\cdot, e)^\top \nu(\mathrm{d}e)\right)^{\frac{1}{2}},$$

which are maps from $\mathbb{R}^d \times \mathbb{A}$ to $\mathbb{R}^d$ and $\mathbb{R}^{d\times d}$, respectively.

In terms of the control problem, this is reflected in the PDE (1.3) (using a second order Taylor expansion in $\mathcal{L}^a$) by the fact that $\varepsilon V_T \to \bar{V}_T$ (see e.g. [57] for a similar analysis) $\bar{V}_T$ being the solution of

$$0 = \partial_t \bar{V}_T + \sup_{a\in\mathbb{A}} \left\{\bar{\mathcal{L}}^a + r(\cdot, a)\right\} \text{ on } [0, T) \times \mathbb{R}^d, \tag{1.6}$$

with the boundary condition $\bar{V}_T(T, \cdot) = 0$ on $\mathbb{R}^d$, in which, given any function $\phi \in \mathcal{C}^{(1,2)}([0, T) \times \mathbb{R}^d; \mathbb{R})$, the operator

$$\bar{\mathcal{L}}^a : \phi \mapsto \mu(\cdot, a)^\top \mathrm{D}_x\phi + \frac{1}{2}\mathrm{Tr}\left[\sigma(\cdot, a)\sigma(\cdot, a)^\top \mathrm{D}^2_{xx}\phi\right] \in \mathcal{C}^0([0, T) \times \mathbb{R}^d; \mathbb{R})$$

is the infinitesimal generator of (1.5) with $\alpha \equiv a$. Equation (1.6) is a standard diffusive HJB whose analysis and numerical resolution are both well documented, see e.g. [78, 86] and references therein.

This method of approximation via the diffusion limit has a storied history. It has been studied extensively in the context of queuing networks, see the books [43, 77] as well as [62], but has also seen use in the context of Actuarial Science [21, 45] and Financial Mathematics [95], amongst others. However, it appears that, in the context of a general problem, the convergence rate of this approximation has not been studied.

Chapter 3 focuses on the study of this convergence rate to the diffusive limit, with $d = d' = 1$ for simplicity. We consider the renormalised family (indexed by $\varepsilon \in \mathbb{R}_+$) of pre-limit problems

$$V^\varepsilon_T(t, x) := \sup_{\alpha\in\mathbb{A}} \mathbb{E}\left[\int_t^T \varepsilon r(X^{t,x,\alpha}_{s-}, \alpha_s)\mathrm{d}N^\varepsilon_s\right] \tag{1.7}$$

in which N^ε is the analogue of N when $\eta = \eta_\varepsilon$. To determine the convergence of V^ε_T to $\bar{V}_T$, we use the HJB-based methodology described above.

We first study the regularity of the solution $\bar{V}_T$ of (1.6) and show that $\bar{V}_T \in \mathcal{C}^{(1,2)}_b([0, T) \times \mathbb{R}; \mathbb{R}) \cap \mathcal{C}^0([0, T] \times \mathbb{R}; \mathbb{R})$ with $\partial_{xx}\bar{V}_T$ being γ-Hölder in space for some $\gamma \in (0, 1)$. We then show that the order of convergence of V_T to $\bar{V}_T$ is $\varepsilon^{\gamma/2}$ in the sense that

$$\left|V^\varepsilon_T(t, x) - \bar{V}_T(t, x)\right| \le C(T - t)\varepsilon^{\frac{\gamma}{2}}. \tag{1.8}$$

We further show how to use (1.6) to construct a control which is $\varepsilon^{\gamma/2}$-optimal for (1.7). Subsequently, we show how to construct error correction terms for this approximation, in the first order and higher orders, at the expense of increased complexity of approximation. This is predicated on the construction of a further PDE, or system of PDEs, and the extensive use of viscosity solutions arguments.

Using an example inspired by the online auction problem (see also Section 1.3), we demonstrate this approximation method and show the computational gains it can achieve relative to directly solving (1.3).

1.1.2 Ergodic control in the diffusive limit

Chapter 3 demonstrates that the diffusive limit approximation can lead to significant computational improvements as compared to the direct resolution of (1.3). However, in many practical problems, the finite horizon fails to capture the true nature of the control problem. When T is very large, the key question becomes whether the system enters a steady-state behaviour for most of the duration, or is dominated by transient behaviour.

These long-term control problems, absent a natural form of discount, are best represented by the ergodic control problem on (1.1) defined by

$$\rho^*(x) := \sup_{\alpha \in \mathcal{A}} \liminf_{T \to +\infty} \frac{1}{T} \mathbb{E}\left[\int_0^T r(X_{s-}^{0,x,\alpha}, \alpha_s) \mathrm{d}N_s \right] \tag{1.9}$$

for any $x \in \mathbb{R}^d$. This criterion is also the one we will use in Reinforcement Learning, see Section 1.2. Due to the limit inferior, the feasibility of (1.9) is inherently tied to the properties of the process $X^{0,x,\alpha}$ itself, and in particular to the possibility of steady-state behaviour (ergodicity).

For example, if the system cannot be stabilised, i.e. if no control can prevent the mass of the process from escaping to $+\infty$, then ρ^* will be ill-defined. See also [102, § 8.3.1] for a discussion in the discrete control case. This property demonstrates how the ergodic problem captures some fundamental complexities of the control problem which other settings may miss, and highlights why it is of particular interest from a theoretical perspective.

Solving the ergodicity difficulty of the limit inferior, and thus showing that (1.9) is well-posed, is achieved through Lyapunov conditions, see e.g. [87]. We make Lyapunov assumptions on the process in which the Lyapunov functions behave as $\|\cdot\|^p$ for some $p \geq 3$.

In Chapter 4, we first show the ergodicity of (1.1) from its Lyapunov assumptions using Ordinary Differential Equation (ODE) methods. This analysis shows uniform Lipschitz estimates for the value functions of the discounted control problem on the pure-jump processes (1.1), allowing us to

show that the ergodic problem is well-posed by using the vanishing discount method, see e.g. [14, 25].

This method shows that the control problem is meaningfully ergodic, in that it forgets initial conditions: there is $\rho^* \in \mathbb{R}$ such that $\rho^*(x) = \rho^*$ for every $x \in \mathbb{R}^d$. Moreover, the value ρ^* accompanied by an auxiliary decision function $w : \mathbb{R}^d \to \mathbb{R}$, which is Lipschitz, form a viscosity solution couple to the ergodic HJB equation

$$0 = -\rho^* + \sup_{a \in \mathbb{A}} \left\{ \mathcal{L}^a w + \eta r(\cdot, a) \right\} \text{ on } \mathbb{R}^d. \tag{1.10}$$

which is the ergodic analog of (1.3). A measurable pointwise maximiser of the maximum in (1.10) yields an optimal stationary decision policy and thus an optimal Markov control.

Using the analogous Lyapunov assumptions and the same ODE method for (1.5), we can show the analogous result for the diffusive limit problem. In particular, that there is a Lipschitz solution $\bar{w} : \mathbb{R}^d \to \mathbb{R}$ to the HJB equation

$$0 = -\bar{\rho}^* + \sup_{a \in \mathbb{A}} \left\{ \bar{\mathcal{L}}^a \bar{w} + r(\cdot, a) \right\} \text{ on } \mathbb{R}^d, \tag{1.11}$$

in which $\bar{\rho}^* \in \mathbb{R}$ is the value of the ergodic control problem on (1.5), defined analogously to (1.9).

Having shown the two problems are well posed and satisfy the requisite HJB equations, we will apply the methodology of Section 1.1.1 to the resolution of ergodic control problem (1.9) and the HJB equation (1.3). To this effect, for $\varepsilon \in \mathbb{R}_+$, we define

$$\rho_\varepsilon^* := \sup_{\alpha \in \mathcal{A}} \liminf_{T \to +\infty} \frac{1}{T} \mathbb{E}\left[\int_0^T \varepsilon r(X_{s-}^{0,0,\alpha}, \alpha_s) \mathrm{d}N_s^\varepsilon \right] \tag{1.12}$$

in which N^ε is defined as in Section 1.1.1.

We first focus on (1.11) and build on the Lipschitzness of $\bar{w}$ to get strong regularity estimates of $\bar{w}$, for $d \in \mathbb{N}$, assuming the volatility σ is independent of the control $a \in \mathbb{A}$, bounded, and satisfies a uniform ellipticity condition. This analysis shows that, for μ Lipschitz (but not-necessarily bounded), $D^2\bar{w}$ is locally γ-Hölder continuous for any $\gamma \in (0, 1)$, with a constant growing at most linearly in x.

This regularity estimate allows us to apply the arguments of Chapter 3, up to the use of moment estimates (derived from Lyapunov conditions) on $X_t^{0,x,\alpha}$, for any $t \in \mathbb{R}_+$, to handle the linear growth of μ and the Hölder constant of $\bar{w}$. This yields the following approximation bound in the ergodic case with unbounded drift, uniform ellipticity, and uncontrolled volatility

$$|\rho_\varepsilon^* - \bar{\rho}^*| \leq C\varepsilon^{\frac{\gamma}{2}}.$$

We also show how to construct a control which is $\varepsilon^{\gamma/2}$-optimal for the pre-limit ergodic control problem (1.12) using the optimal decision policy obtained from (1.11).

Subsequently, we show how to construct error correction terms for this approximation once again, but using a method that differs from the one in Chapter 3. In this instance, the increase in complexity comes from iterating additional PDEs rather than from the construction of an increasingly large system of PDEs. This method is more straightforward than the one in Chapter 3 and relies on a verification argument.

Continuing towards the goal of efficient resolution of (1.9), we delve into the numerical resolution of the ergodic control problem. Because numerical schemes generally need ad-hoc constructions we take $d = 1$ once again, and strengthen the regularity and growth-order conditions in our assumptions. Under these assumptions, we construct a numerical scheme for (1.11), in which we solve an equation of the form

$$0 = -\bar{\rho}_h^{\kappa,*} + \sup_{a\in\mathbb{A}} \left\{ \bar{\mathcal{L}}_h^a \bar{w}_h^\kappa + r(\cdot, a) \right\} \tag{1.13}$$

on a mesh $\bar{M}_h^\kappa$ on the interval $[-\kappa h, \kappa h]$ of fineness $h > 0$ containing $N_\kappa := 2\kappa + 1$ points, $\kappa \in \mathbb{N}^*$, with

$$\bar{\mathcal{L}}_h^a : v \in \mathbb{R}^{N_k} \mapsto P_h v \in \mathbb{R}^{N_k}$$

for a transition matrix $P_h \in \mathbb{R}^{N_\kappa \times N_\kappa}$ which is obtained via finite difference approximation of $\bar{\mathcal{L}}^a$. The solution couple $(\bar{\rho}_h^{\kappa,*}, \bar{w}_h^\kappa)$ of (1.13) corresponds to the value and decision function of an ergodic control problem on a continuous-time Markov Chain.

Since this structure corresponds to a pure-jump process on $\bar{M}_h^\kappa$, whose intensity is $1/\Delta t_h$, for $\Delta t_h = \mathcal{O}(h^2)$, we can use the same approximation methodology described above, up to some care in handing the boundary conditions of the scheme, to obtain convergence of order

$$\left| \Delta t_h \bar{\rho}_h^{\kappa,*} - \bar{\rho}^* \right| \leq C \left(h^\gamma + h^{-1} \left| \kappa h \right|^{1-p} \right),$$

in which $p \geq 3$ is the growth order of the Lyapunov function of the diffusion. We also study the construction of an approximately optimal control for (1.12) using $\bar{w}_h^\kappa$, in what seems to be a novel topic concerning the numerical resolution of control problems.

We then return to a bidding problem which proposes a variant of the one of Chapter 3. We use this numerical example to apply the numerical scheme of (1.13) and display the computational efficiency gains of the diffusive limit approximation.

1.2 Reinforcement Learning

Reinforcement Learning (RL) studies the problem of learning how to control a system such as (1.1) under uncertainty on its dynamics from the perspective of Machine Learning. There are two main ways to formalise this uncertainty, depending on whether one chooses to take the dynamics or the control as the primary object of study. These are known as the value-based and policy-search approaches respectively.

We focus on the former and adopt a model-based framework, which assumes that b in (1.1) is unknown but that one has access to a model class $\mathscr{F}_\Theta$ of possible dynamics, parametrised by $\theta \in \Theta \subset \mathbb{R}^{d_\Theta}$, $d_\Theta \in \mathbb{N}^*$. This paradigm assumes that all the uncertainty is captured, within $\mathscr{F}_\Theta$, by the lack of knowledge of a true parametrisation $\theta^* \in \Theta$. This leads us to extend the definition of $X^{t,x,\alpha}$ to incorporate a driving parameter $\theta \in \Theta$ by defining $X^{t,x,\alpha,\theta}$ as the solution to the Stochastic Differential Equation

$$X^{t,x,\alpha,\theta} = x + \int_t^{\cdot} \int_{\mathbb{R}^{d'}} b_\theta(X^{t,x,\alpha,\theta}_{s^-}, \alpha_s, e) N(\mathrm{d}e, \mathrm{d}s)\,. \tag{1.14}$$

We adopt the diffusive limit scaling of Section 1.1 for b_θ, and since the standard noise structure in Reinforcement Learning is an additive martingale, we will take $b_\theta(x, a, e) := \varepsilon\bar{\mu}_\theta(x, a) + \varepsilon^{1/2}\bar{\Sigma}e$, with regularity conditions on $(\bar{\mu}_\theta, \bar{\Sigma})$ from Section 1.1.2. This framework is meant to model deterministic systems under sub-Gaussian martingale perturbation, hence we will take ν to be a centred standard Gaussian measure on $\mathbb{R}^d$ for simplicity.

Regardless of the form of the control problem at hand, as a value-based method, we seek to explore the state-action space to learn b_θ and the reward function r in order to approximate the control problem successfully. This necessarily leads us to sample sub-optimal state-action pairs for the control problem, creating an inherent tension between exploration and control. Conversely, a greedy agent which focuses solely on exploiting the empirical best state at any given time will likely fail to learn. Indeed, it is liable to create a closed-loop sub-system in which it remains in one region of space forever due to unlucky observations outside of this region, unable to learn the rest of the system.

This tension is a consequence of learning from within a decision problem, as was observed by studying the bandit (stateless) case [82]. Its source is the nature of the feedback of the decisions: one only gets information about state-action pairs actually traversed.

Consequently, since ε and $\bar{\Sigma}$ are independent of actions taken, they are unaffected by this trade-off and can be estimated by simple statistical methods. To avoid unnecessary complexity, we simply take them as known.

Since X^{0,x,α,θ^*} of (1.14) is a stochastic process, there are two notions of samples one could consider: *episodic* in which we observe realisations of the process $X^{0,x,\alpha,\theta^*}(\omega)$ from different $\omega \in \Omega$, possibly up to a reset time $T \in \mathbb{R}_+$; or *online* in which we observe over time $t \in \mathbb{R}_+$ a trajectory $X^{0,x,\alpha,\theta^*}(\omega)$ for a single fixed ω, like in a time series.

Learning online, one cannot revisit past events by relying on the reset of the trajectory. Instead, the system has to be driven back into the relevant conditions. This characterises the learning difficulty in a way that is inherent to the dynamics of the process $X^{t,x,\alpha,\theta}$. At the same time, it excludes the study of nonstationary policies, including the finite horizon problem because beginning-of-the-world effects can't be reasonably learned along a single trajectory.

We choose to focus on the online problem, following the large literature of Online Reinforcement Learning Theory, and study the ergodic criterion

$$\rho^*_{\theta^*}(x) := \sup_{\alpha \in \mathcal{A}} \liminf_{T \to +\infty} \frac{1}{T} \mathbb{E}\left[\int_0^T r(X^{0,x,\alpha,\theta^*}_{s-}, \alpha_s) \mathrm{d}N_s \right],$$

in which $\mathcal{A}$ is defined as in Section 1.1 with respect to X^{0,x,α,θ^*}, because it captures the difficulty inherent in the process $X^{t,x,\alpha,\theta}$ in the control problem. As shown in Section 1.1.2, it also admits an optimal policy (see (1.10)) which is stationary, meaning it is possible to learn an optimal policy online.

The choice of an ergodic problem introduces the well-posedness issues of $\rho^*_{\theta^*}$ mentioned in Section 1.1.2, even in the discrete-state case see e.g. [103, § V.] and [12]. Until now, this ergodicity problem was a major hurdle in Reinforcement Learning with continuous states outside of the Linear Quadratic Regulator case [7, 81].

Indeed, the methods used to establish ergodicity in discrete systems, which are based on studying (discrete) transition matrices, see e.g. [102, § 8.], do not easily extend to continuous problems. In the general setting which was described in Section 1.1.2, we showed that the problem (1.9) was well-posed under moderate Lyapunov and coefficient conditions.

In Chapter 5, we specialise this analysis to the structure of b_θ of (1.14), which allows us to relax the Lyapunov assumptions of Chapter 4 to a single weaker condition from which the ergodicity and the stability of both the pure-jump and diffusive limit problems follow. This condition is a mixing (contraction) condition on the instantaneous dynamics, which takes the form of

$$\mathscr{V}(x + \varepsilon \bar{\mu}_\theta(x, a) - (x' + \varepsilon \bar{\mu}_\theta(x', a))) \leq (1 - \mathfrak{c}_{\mathscr{V}} \varepsilon) \mathscr{V}(x - x') \tag{1.15}$$

for any $(x, x', a, \theta) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{A} \times \Theta$, in which $\mathscr{V} \in \mathcal{C}^2(\mathbb{R}^d \setminus \{0\}, \mathbb{R}_+)$ is a suitable Lyapunov function and $\mathfrak{c}_{\mathscr{V}} > 0$ is a constant independent of ε. Using the additive sub-Gaussian noise we leverage this assumption beyond ergodicity

and show that $(X_s^{0,x,\alpha,\theta})_{s\in\mathbb{R}_+}$ behaves essentially like a sub-Gaussian process, including high-probability concentration. This shows that ρ_θ^* is well-posed for any $\theta \in \Theta$ by the arguments of Section 1.1.2.

Having shown that the problem is well-posed, we can seek to construct an agent $\alpha \in \mathcal{A}$ which maximises (1.9) absent knowledge of θ^*, i.e. an optimal learning algorithm. Unfortunately, this explicit maximisation problem has remained out of reach, even in simple (discrete-state or bandit) problems.

Instead, the Reinforcement Learning literature has focused on designing ad-hoc algorithms, and on understanding and comparing their performance using performance upper bounds. This methodology is complemented by problem hardness (minimax) lower bounds, which is a standard approach in Theoretical Computer Science.

The primary metric used in these bounds is the notion of the regret of an agent $\alpha \in \mathcal{A}$

$$\mathcal{R}_T(\alpha) := T\rho_{\theta^*}^* - \int_0^T r(X_{s-}^{0,x,\alpha,\theta^*}, \alpha_{s-})\mathrm{d}N_s\,, \tag{1.16}$$

defined for any $(x, T) \in \mathbb{R}^d \times \mathbb{R}_+$, which one bounds in the high-probability sense. This concept is naturally adapted to online decision-making problems, see e.g. [58, 109], as it neatly quantifies the cost of uncertainty about θ^* directly in terms of the reward function of the control task. As can be seen by dividing by T and letting $T \to +\infty$, the rate of growth of the regret quantifies the efficiency of α at learning to control and, in a minimax sense, the inherent difficulty of the problem.

In the absence of a formal optimum agent, many exploration agents have been proposed. These all add some exploratory behaviour to the greedy agent, which was discussed previously. One possibility is to force isotropic exploration, either by taking random actions some of the time (ϵ-greedy, and other forms of entropy-regularised policies), or by alternating phases of exploration and exploitation (Explore-Then-Commit, see e.g. [100]).

While these methods can be effective, calibrating the noise source generally requires prior knowledge about the system. In the absence of such information, their regrets are generally sub-optimal. To circumvent this issue requires structuring the exploratory behaviour to the problem, which is done by careful study and decomposition of the regret in order to identify and address key sub-tasks of the exploration-exploitation dilemma. Historic works have shown that it can be addressed via Thompson Sampling, see e.g. [6, 113], or by the use of the Optimism in the Face of Uncertainty paradigm, see e.g. [15, 16, 68].

This paradigm identifies four key sub-tasks of the problem: learning, i.e. the ability to estimate $\bar{\mu}_{\theta^*}$ and r and give δ-confidence sets for them, $\delta \in (0, 1)$; optimism, which is done by selecting a belief system $\tilde{\theta}$ which maximises ρ_θ^*

over θ in the confidence set; planning, which computes the optimal decision policy $\tilde{\pi}^*$ for $\tilde{\theta}$; and lazy-updates, which ensure the agent executes the policy $\tilde{\pi}^*$ for long enough to get useful feedback on $\tilde{\theta}$ by exploring the system.

The groundwork for planning was laid in Section 1.1.2, in which we showed how to use the Hamilton-Jacobi-Bellman equation (1.10) to solve the control problem ρ^*_θ, for any $\theta \in \Theta$. We also demonstrated an efficient approximate resolution method based on the diffusion limit and the numerical scheme of (1.13). This provides the tools for efficient planning in our near-diffusive setting.

Learning non-linear continuous dynamics has been addressed in the Reinforcement Learning literature through the establishment of the eluder dimension in the seminal articles [97, 107]. This framework however is only suited for bounded coefficients. In Chapter 5, we refine the original arguments to apply them to a stable process with unbounded drift (i.e. $\bar{\mu}_{\theta^*}$) on $\mathbb{R}^d$, which requires some finer care about measurability and covering arguments. As a result, the learning quantities become adaptive to the regions of space that are effectively visited by the trajectory $X^{0,x,\alpha,\theta^*}(\omega)$. We show that the resulting confidence sets are well calibrated and that this low fit error translates into a low prediction error using the eluder dimension. This allows us to use the point estimates and confidence sets of Non-Linear Least Squares regression to solve the learning problem for non-linear dynamics, under only a Lipschitz condition on $\bar{\mu}_\theta$ and the stability properties of X^{0,x,α,θ^*} derived from (1.15).

The need for lazy updates is an unintuitive consequence of the dynamical nature of (1.14). Since the policy $\tilde{\pi}^*$ optimises the ergodic criterion it may perform poorly in the short term until the system mixes. There is therefore a potential cost at each update of the policy and one way to control the resulting error is by only changing $\tilde{\theta}$ infrequently. The difficulty is in doing so without degrading the learning. We show, using refined functional inequalities for sequences, that the low ($\mathcal{O}(\log(T))$) prediction error can be used to derive an update criterion that triggers only $\mathcal{O}(\log(N_T))$ times by time T, at the cost of only a constant factor downgrade in the efficiency of the learning.

Having tackled the sub-tasks, we are able to give an algorithm that generates a control process $\varpi \in \mathcal{A}$ and to bound its regret. We show that

$$\mathbb{P}\left(\mathcal{R}_T(\varpi) \leq \mathcal{O}\left(\sqrt{T \mathrm{d}_{\mathrm{E},N_T} \log(\mathscr{N}^{\varepsilon}_{N_T}) \log(N_T) \log\left(\frac{1}{\delta}\right)}\right)\right) \geq 1-\delta\,, \qquad (1.17)$$

in which $\mathrm{d}_{\mathrm{E},N_T}$ and $\mathscr{N}^{\varepsilon}_{N_T}$ are the $2\sqrt{\varepsilon/N_T}$-eluder dimension (see [107, Def. 4.] and (5.56) in Section 5.5.2) and the $\varepsilon\|\bar{\Sigma}\|_{\mathrm{op}}/N_T$ covering number, respectively, of the class $\mathscr{F}_\Theta$ with the domain of its elements restricted to a ball of radius $\mathcal{O}(\sqrt{\log(T/\varepsilon)})$.

This matches (in T) the best known upper bounds for the regret in the discrete-state case [20, 68, 97], up to logarithmic factors. For comparison, known minimax lower bounds for the regret in the discrete-state case are of order[1] of $\Omega(\sqrt{T})$, e.g. [68], which is to say that there is no algorithm which can achieve regret growth slower than $\sqrt{T}$ on every instance.

Using the diffusion limit approximation of Chapter 4 we can also solve the planning task efficiently up to an additive term $\varepsilon^{\gamma/2}T$ which is linear in T. This shows that in very high-frequency systems, with $\varepsilon \ll 1$, the regret bound (1.17) is achievable in a meaningful way using this kind of approximation to save computational effort.

This work in the specific context of systems admitting a diffusive limit showcases how tools from Stochastic Control and continuous-time stochastic processes can be applied to continuous-state Reinforcement Learning problems. It provides a first step towards a general theory of Reinforcement Learning in continuous state-space based on these tools.

1.3 Online Auction Problems

High-frequency control systems of the type described in Sections 1.1 and 1.2 are common in real-world applications. They often model a system that monitors some outside event signals to which it reacts. A typical category of examples involves the purchase or sale of objects by algorithms via auctions, markets, or other mechanisms. At scale, these involve high-frequency events, each of which only contributes small changes to the relevant state process, which could be, for instance, price, inventory volume, or some more complex object such as an investment portfolio.

While financial markets are the most obvious example, they may not be well represented by arrivals from a Poisson process at all scales, for instance, because of self-excitation phenomena, see e.g. [52]. Instead, we focus throughout this thesis on the online display advertising auction application, which models the way that banner ads on webpages are sold during the loading time of the page. The overall goal of Chapter 6 is to demonstrate in an application how a control system of the type studied in Chapters 3 and 4 can arise as a natural response to the computational pressure of the high frequency of interactions.

The business model of most internet applications is based on advertising: either directly or indirectly via the trading of data. There are many different types of ads, each of which has its own trading mechanisms. We will focus on

[1] A function $T \in \mathbb{R}_+ \mapsto f(T) \in \mathbb{R}_+$ is of order $\Omega(\sqrt{T})$ if there are $(N, C) \in \mathbb{N} \times \mathbb{R}_+$ such that for any $T > N, f(T) > C\sqrt{T}$.

display ads, which are the ones that load in banners, typically at the top or on the sides of websites. The display ad slots on a webpage are sold in an online stream[2] by automated algorithms during the loading time of the page (about 50 ms) using auctions. This makes the auction times dependent only on the arrival of users, which should be reasonably modelled by a Poisson process. In terms of frequency, the order of magnitude is of 10^{10} to 10^{12} auctions per day, while individual costs and gains are very small due to the probability of a user clicking on an ad being small.

The key particularity of the display advertising market relative to other markets such, as financial ones, is concentration. Due to the computing infrastructure required by the scale of the automated auctions, the trading is undertaken by intermediaries on both sides, so that there are only about half a dozen participants in each auction. This has a large impact on the nature of the revenue maximisation problem for a seller.

In Auction Theory, the question of the optimal auction, given knowledge of the valuation distributions of bidders, was solved in the seminal 1981 article of Myerson [92]. Evidently, this assumption is unrealistic for display advertising. On the other hand, the high frequency of events means that lots of data are available to learn to maximise revenue from empirical observations. However, due to the complexity of the Myerson mechanism [92], it is impossible to learn it directly [90], and we will thus study the simpler second-price auction.

This format is both a reasonable approximation to the Myerson auction, see [104] and is the format which is historically used in display advertising[3]. In this format, the winning bid is simply the highest one, but the item is only sold if the winning bidder beats their *reserve price*, which is an individualised minimum sale price. If the reserve price is cleared, the winner pays the smallest bid which still wins them the item, that is: the maximum between their reserve price and the bids of the other bidders.

Because they are individualised, the problem of optimising revenue via the reserve prices is separable and boils down to finding the *monopoly price* r^* of each bidder. For a bidder with stationary bid distribution F, this is the maximiser of the monopoly revenue

$$\Psi^F : r \in \mathbb{R}_+ \mapsto \int_{\mathbb{R}_+} r 1_{r \geq B} F(\mathrm{d}B) \in \mathbb{R}_+ . \tag{1.18}$$

Several methods have been proposed to maximise (1.18), mostly based on bandits, e.g. [73], or on non-convex optimisation, e.g. [89]. Unfortunately, these methods all require storing many, if not all, past bids to update the price,

[2] That is, ads are sold sequentially and individually.

[3] Although it should be noted that this has been shifting in favour of the first-price auction and similar formats recently.

which will become computationally infeasible at the scales we are interested in.

The motivating question of Chapter 6 is whether it is possible to design an algorithm whose update is real-time (i.e. independent of the past number of bids), even at the cost of degrading the convergence speed, so that it can be applied online to the data stream.

The enticing candidate is a first-order method like Online Gradient Ascent (OGA). Unfortunately, because $(r, B) \in \mathbb{R}_+^2 \mapsto p(r, B) := r1_{r \geq B}$ is discontinuous (in r for any bid $B \in \mathbb{R}_+^*$), random gradients from it will not give an unbiased gradient estimate for Ψ^F. Consequently OGA need not converge to $r^* \in \operatorname{argmax} \Psi^F$.

The problem of biased gradients can be solved by smoothing p by convolution with a smooth kernel k and then applying OGA to the result. The resulting algorithm, Conv-OGA, is inevitably biased and we should like to reduce its bias over time by reducing smoothing. The main contribution of Chapter 6 is an algorithm, V-Conv-OGA, which does smoothing and OGA at the same time in order to trade off the biased gradients and the stability of OGA, thus retaining real-time updates.

We first give an analysis of the convexity[4] properties of Ψ^F under classical auction theory assumptions like monotone virtual valuation and increasing hazard rate. This allows us to control the bias and variance trade-off induced by the smoothing, which we use to show the almost sure convergence of iterates to r^* using the quasi-martingale method, see e.g. [32].

We strengthen this result by modifying arguments of [91] to obtain a convergence rate under a stronger assumption which eliminates arbitrarily small gradients. The full bounds exhibit the bias-variance trade-off, but upon balancing it they correspond to

$$\mathbb{E}\left[\|r_n - r^*\|^2\right] = \tilde{\mathcal{O}}\left(n^{-\frac{1}{2}}\right)$$

in which $(r_n)_{n \in \mathbb{N}^*}$ are the iterates of our algorithm and $\tilde{\mathcal{O}}$ hides polylogarithmic in n factors from the order notation. Since methods storing all bids, e.g. [89], can reach a rate of $\tilde{\mathcal{O}}(n^{-1})$, the corresponding gap can be viewed as the cost of using a real-time algorithm.

In practice, bidders in display ad-auctions are known to exhibit non-stationary or even strategic behaviour. By the nature of OGA, we can easily adapt Conv-OGA, with constant learning rate $\gamma_0 \in \mathbb{R}_+$, to deal with a non-stationary bidder. Given a reasonable number of bid distribution switches by time $N \in \mathbb{N}$, we show a dynamic regret relative to r^* of order $\mathcal{O}(\sqrt{N})$.

[4] Precisely log-concavity and pseudo-concavity.

Given a sequence of bids $(b_n)_{n\in\mathbb{N}^*} \subset \mathbb{R}_+$, the reserve prices $(r_n)_{n\in\mathbb{N}^*}$ generated by Conv-OGA are given by some initial $r_1 \in \mathbb{R}_+$ and the recursion

$$r_{n+1} = r_n + \gamma_0 f(r_n, b_n)\,, \tag{1.19}$$

in which $f(r, B) := \mathrm{D}[p(\cdot, B) \star k](r) \vee -r$, for a smoothing kernel k and $\gamma_0 > 0$.

In practice, numerous sources of noise affect the seller's reserve prices even with a deterministic update like (1.19), such as difficulty in attributing bids to the correct bidder. This calls for cautious updates and, thus, reserve price increments tend to be negligible, i.e. $\gamma_0 \ll 1$.

Meanwhile, from the point of view of the bidder, there are also other sources of noise, such as possible aggregation by the seller over unknown features. These multiple levels of noise explain why in practice noise dominates the dynamics of Conv-OGA, as in the diffusion limit regime of Section 1.1.

Since the auction arrival times are well modelled by a high-frequency Poisson process on $\mathbb{R}_+$, the bidder observes a reserve price process according to the controlled Stochastic Differential Equation

$$X^{0,x_0,\alpha} = x_0 + \int_0^{\cdot} \int_{\mathbb{R}^{d'}} \varepsilon b_1(X^{0,x_0,\alpha}_{s-}, \alpha_s, e) + \varepsilon^{\frac{1}{2}} b_2(X^{0,x_0,\alpha}_{s-}, \alpha_s, e) N(\mathrm{d}e, \mathrm{d}s), \tag{1.20}$$

in which $(\alpha_t)_{t\in\mathbb{R}_+}$ is the bid process. This matches Section 1.1. Thus, the bidder's objective of maximising its gains, which is known as a *bidding problem*, is an example of a stochastic control problem of the form considered in Section 1.1.

For the seller, $b_1(x, a, e) = f(x, a)$ represents an optimal choice of a stationary Markov pricing dynamic, in the sense outlined above. However, the definition of f via a convolution complicates the manipulation of the process when studying the bidding problem. This leads us to consider some simplifying heuristics in our applications.

Considering, for simplicity, that $b_2(x, a, e) = e$, the simplest heuristic is to take $b_1(x, a, e) := \beta(a - x)$, for $\beta \in \mathbb{R}_+$. In Section 4.5, we consider this dynamic with β random to model the uncertainty about the aggressivity of the seller's tracking of bids. In Section 3.4, we consider $b_1(x, a, e) := -x + qb + (1 - q)r_0$, $q \in (0, 1)$, which takes a convex combination of the last bid with a benchmark price $r_0 \in \mathbb{R}_+$.

While these heuristic models are not definitive models of the behaviour of sellers in display ad auctions, they are interesting because they exhibit an emergent (Markov) diffusive limit behaviour. This behaviour arose as a natural response to the problem of revenue maximisation in a highly concentrated auction format when faced with the heavy computational constraints of high-frequency events.

Chapter 6 thus motivates the examples adopted in Chapters 3 and 4 for numerical experiments, and grounds the overall diffusion limit method in the real world. Note that the monopoly price is not the only quantity of interest in online display ad-auctions and that many other state variables and control problems can be considered, see for instance [54] for other examples in inventory management and conversion maximisation.

Chapitre 2

Résumé de la thèse

Le développement de l'automatique dans le génie a placé les systèmes dynamiques au cœur de la plupart des technologies industrielles modernes. Les premiers succès se trouvèrent dans les chaînes de fabrication, le guidage des fusées ainsi que d'autres applications industrielles lourdes. Le développement des technologies de l'information et de l'internet a élargi la portée de cette tendance, les systèmes dynamiques automatisés devenant omniprésents dans la plupart, sinon tous, les aspects de notre vie quotidienne.

Dans nombreux de ces applications, les problèmes liés au bruit et à l'incertitude sont inévitables. Dans le contexte de l'automatisation des systèmes dynamiques, la prise en compte de dynamiques bruitées a conduit au développement de la théorie du contrôle stochastique, qui a fleuri et s'est ramifiée depuis la seconde moitié du vingtième siècle.

La branche du contrôle stochastique continu (en temps et en espace) s'est fleurie d'une théorie riche et fructueuse, basée sur les Équation aux Dérivées Partielles (EDP), à commencer par les travaux du groupe de Lev Pontriaguine dans les années 1960 ; voir par exemple [59] pour une perspective bibliographique. Cette approche basée sur les EDP fut considérablement enrichie dans les années 1980 par le développement de la théorie des solutions de viscosité [48, 49], permettant ainsi la résolution de nombreux problèmes complexes et irréguliers. Ces nouveaux outils ont conduit à des succès remarquables dans les domaines d'application de la théorie, notamment dans le domaine des mathématiques financières, voir par exemple [56].

Malgré ce succès dans l'incorporation de l'aléatoire, les systèmes de contrôle dans le monde réel doivent encore tenir compte du problème de l'incertitude dans les dynamiques du système. Dans le cadre du temps continu, ce problème a été abordé par les domaines du contrôle adaptatif et robuste, voir par exemple [30, 31]. Alors que le contrôle robuste cherche à être résistant à

l'incertitude, le contrôle adaptatif exploite les statistiques pour apprendre à contrôler le système. Les méthodes statistiques classiques utilisées dans ce domaine (par exemple, l'estimation du maximum de vraisemblance) et leur accent sur les résultats de convergence asymptotique contrastent avec les méthodes plus récentes de l'apprentissage automatique.

La branche du contrôle stochastique discret (en temps et en espace), voir par exemple [22, 102] a développé vis-à-vis du problème de l'incertitude dans les systèmes contrôlés la perspective de l'apprentissage automatique avec le paradigme de l'apprentissage par renforcement, voir par exemple [23, 112]. Cette perspective met l'accent sur la performance vis-à-vis du contrôle optimal, l'efficacité computationnelle des méthodes et les garanties de temps fini plutôt que sur la modélisation explicative.

Le domaine de l'apprentissage par renforcement à acquis une grande notoriété dans les années 2010, avec le développement des méthodes d'apprentissage par renforcement profond, voir par exemple [88]. Certaines applications pratiques ont été largement médiatisées, notamment AlphaGo [111] qui a atteint une performance supérieure à celle de l'humain au jeu de Go.

Ces méthodes d'apprentissage par renforcement profond exploitent intensivement les méthodes statistiques et nécessitent donc des efforts d'exploration intensifs pour générer des données. Cela implique l'utilisation de politiques de contrôle qui explorent de force le système. De telles politiques sont intrinsèquement en conflit avec l'objectif du problème de contrôle, ce qui conduit à un certain gâchis. L'étude de cette contrepartie entre exploration et exploitation est au cœur du domaine de la théorie de l'apprentissage par renforcement.

En utilisant la notion de regret pour quantifier ce compromis, la théorie de l'apprentissage par renforcement a développé une analyse fine des défis précis de l'exploration efficace des systèmes inconnus lorsque l'objectif final est le contrôle [16]. En nous permettant de fusionner les objectifs contradictoires du problème d'apprentissage statistique et du problème de contrôle, cette analyse a jeté les bases de la construction de méthodes d'exploration optimales pour l'apprentissage par renforcement [68].

Malheureusement, de par ses origines dans le contrôle discret, le domaine de la théorie de l'apprentissage par renforcement a eu du mal à étendre ses méthodes aux systèmes continus complexes, limitant ainsi son applicabilité dans le monde réel. Le principal obstacle à surmonter est l'analyse et la résolution numérique du problème de contrôle sur un espace d'états continu, tous deux étant des problèmes bien étudiés dans la théorie du contrôle stochastique continu.

Cette complémentarité observée entre les deux domaines, découlant de leur généalogie partagée suggère que le point de vue de l'apprentissage automatique en général, et de l'apprentissage par renforcement en particulier,

offre beaucoup de promesses pour le renouvellement du problème de l'apprentissage du contrôle des systèmes dynamiques. Inversement, les outils de la théorie du contrôle stochastique ont le potentiel de débloquer certains des défis auxquels la théorie de l'apprentissage par renforcement est confrontée dans le cadre continu.

L'objectif principal de cette thèse est de démontrer cette complémentarité en résolvant le problème de l'apprentissage par renforcement à états continus pour une catégorie de systèmes qui admettent une limite diffusive en tirant parti du contrôle stochastique continu, à la fois pour les outils analytiques et pour des approximations efficaces en termes de calculs. En raison de sa nature à l'interface de deux domaines d'étude, une partie de cette thèse est principalement axée sur le contrôle, que nous décrivons dans la Section 2.1, tandis qu'une autre partie se concentre plus résolument sur l'apprentissage par renforcement, que nous décrivons dans la Section 2.2, exploite ces résultats. Une troisième partie, qui est transversale aux deux autres, est décrite dans la Section 2.3 et se concentre sur un cas d'utilisation typique basé sur le problème des enchères séquentielles à haute fréquence dans la publicité, ancrant ainsi l'approche théorique dans un cadre pratique.

2.1 Approximations par la limite diffusive

La théorie de l'apprentissage est fondée sur l'observation d'événements singuliers (les échantillons), et est donc naturellement construite sur des processus en temps discret. D'autre part, la théorie du contrôle continu étudie par définition des processus définis sur $[0, T]$, pour $T > 0$, ou sur $\mathbb{R}_+$. Afin de concilier ces deux perspectives, nous choisissons de baser notre problème de contrôle sur un processus à sauts purs dont les événements arrivent selon un processus de Poisson.

Soit N une mesure aléatoire sur $\mathbb{R}_+ \times \mathbb{R}^{d'}$ admettant en guise de compensateur $\eta\nu(\mathrm{d}e)\mathrm{d}t$, pour une mesure de probabilité ν sur $\mathbb{R}^{d'}$, $d' \in \mathbb{N}^*$, $\eta > 0$. Étant donné un temps initial $t \in [0, T]$, une condition initiale $x \in \mathbb{R}^d$, $d \in \mathbb{N}^*$, et une application mesurable $(x, a, e) \in \mathbb{R}^d \times \mathbb{A} \times \mathbb{R}^{d'} \mapsto b(x, a, e) \in \mathbb{R}^d$, nous considérons la solution de l'Équation Différentielle Stochastique suivante

$$X^{t,x,\alpha} = x + \int_t^{\cdot} \int_{\mathbb{R}^{d'}} b(X_{s-}^{t,x,\alpha}, \alpha_s, e) N(\mathrm{d}e, \mathrm{d}s)\,, \tag{2.1}$$

dans laquelle le processus de contrôle α appartient à l'ensemble $\mathcal{A}$ des processus prévisibles à valeurs dans $\mathbb{A}$, pour $\mathbb{A} \subset \mathbb{R}^{d_{\mathbb{A}}}$ compact, $d_{\mathbb{A}} \in \mathbb{N}^*$.

Ce choix de dynamique d'état rend le mouvement du système évident : l'agent observe l'état actuel du système et choisit une action en fonction des

informations passées. Après avoir attendu un temps distribué exponentiellement, un événement se produit et l'état du système saute vers un point aléatoire dépendant de l'état actuel et de l'action choisie. Cette nature markovienne du système est cruciale.

Dans le même temps, ce type de temps d'attente aléatoire correspond naturellement à de nombreux systèmes numériques, dont les changements d'état sont pilotés par des événements exogènes. Des exemples typiques de ce type de système sont donnés par les problèmes de files d'attente et par les enchères séquentielles pour la publicité. De faibles valeurs de η font que le système se comporte davantage comme un système à temps discret (avec un nombre aléatoire d'événements N_T proche de ηT), tandis que $\eta \to +\infty$ correspond à un système entièrement continu en temps.

Le volume de certains systèmes digitaux basés sur l'internet a atteint des régimes pour lesquels η est suffisamment grand pour que le système puisse être raisonnablement approximé par un système diffusif en temps continu. C'est notamment le cas pour les enchères séquentielles pour la publicité, que nous utilisons comme exemple tout au long de la thèse et que nous présentons dans la Section 2.3.

2.1.1 Limites Diffusives en Contôle Stochastique

Considérons le problème de contrôle à horizon fini, défini pour $(t,x) \in [0,T] \times \mathbb{R}^d$ par

$$V_T(t,x) := \sup_{\alpha \in \mathcal{A}} \left[\int_t^T r(X_{s-}^{t,x,\alpha}, \alpha_s) \mathrm{d}N_s \right] \tag{2.2}$$

où $N_t := N(\mathbb{R}^{d'}, [0,t])$ et $(x,a) \in \mathbb{R}^d \times \mathbb{A} \mapsto r(x,a) \in \mathbb{R}$ est une fonction de récompense. Sous certaines conditions de régularité sur les coefficients b et r (par exemple, bornés et Lipschitziens en x uniformément en les autres arguments), nous pouvons montrer que V_T est la solution viscosité bornée unique de l'équation Hamilton-Jacobi-Bellman suivante

$$0 = \partial_t V_T + \sup_{a \in \mathbb{A}} \left\{ \mathcal{L}^a V_T + \eta r(\cdot, a) \right\} \text{ on } [0,T) \times \mathbb{R}^d, \tag{2.3}$$

avec la condition au bord $V_T(T,\cdot) = 0$ sur $\mathbb{R}^d$, où, étant donné une fonction $\phi \in \mathcal{C}_b^0([0,T) \times \mathbb{R}^d; \mathbb{R})$, l'opérateur

$$\mathcal{L}^a : \phi \mapsto \eta \int_{\mathbb{R}^{d'}} [\phi(\cdot, \cdot + b(\cdot, a, e)) - \phi] \nu(\mathrm{d}e) \in \mathcal{C}_b^0([0,T) \times \mathbb{R}^d; \mathbb{R}) \tag{2.4}$$

est le générateur infinitésimal (voir par exemple [72, § 9.7]) associé à (2.1) avec $\alpha \equiv a$.

Puisque (2.4), et donc (2.3), est non locale, l'analyse de (2.3) est difficile : tant en ce qui concerne les estimations de régularité que la résolution numérique. De plus, cet effet est amplifié lorsque η devient grand : l'erreur d'intégration numérique et $\|\partial_t V_T\|$ croissent avec η.

De nombreux systèmes admettent de larges valeurs de η, mais maintiennent malgré cela un certain degré de structure à l'échelle macroscopique, ceci puisque leurs incréments sont proportionnellement petits par rapport à η. L'exemple le plus connu de ce type de phénomène est celui des marchés financiers, voir par exemple [52, 67]. Dans le cas de (2.1), il existe plusieurs régimes de mise à l'échelle possibles, tels que le régime de limite fluide de [54], mais nous choisissons de nous concentrer sur la limite diffusive car elle préserve la stochasticité du processus lorsque $\eta \to +\infty$.

Dans ce régime, nous posons $\eta = \eta_\varepsilon := \varepsilon^{-1}$ pour $\varepsilon > 0$, et supposons que le coefficient est de la forme $b = \varepsilon b_1 + \varepsilon^{1/2} b_2$, pour des applications localement bornées b_1 et b_2 telles que $\int b_2(\cdot, e)\nu(\mathrm{d}e) = 0$ et $\inf \int b_2(\cdot, e)^2 \nu(\mathrm{d}e) > 0$. Lorsque $\varepsilon \downarrow 0$, le processus $X^{t,x,\alpha}$ de (2.1) converge vers un processus de diffusion contrôlé, voir [66]. Pour $(t, x) \in [0, T) \times \mathbb{R}^d$ et $\bar{\alpha} \in \bar{\mathcal{A}}$, avec $\bar{\mathcal{A}}$ l'ensemble des processus à valeurs dans $\mathbb{A}$ prévisibles par rapport à la filtration d'un processus de Wiener W à d dimensions, cette diffusion est donnée par la solution forte (unique) de l'Équation Différentielle Stochastique suivante

$$\bar{X}^{t,x,\alpha} = \int_t^\cdot \mu(\bar{X}_s^{t,x,\alpha}, \alpha_s)\mathrm{d}s + \int_t^\cdot \sigma(\bar{X}_s^{t,x,\alpha}, \alpha_s)\mathrm{d}W_s\,, \tag{2.5}$$

dont les coefficients sont donnés par

$$\mu := \int_{\mathbb{R}^{d'}} b_1(\cdot, e)\nu(\mathrm{d}e)\,,\ \sigma := \left(\int_{\mathbb{R}^{d'}} b_2(\cdot, e) b_2(\cdot, e)^\top \nu(\mathrm{d}e)\right)^{\frac{1}{2}},$$

qui sont des applications de $\mathbb{R}^d \times \mathbb{A}$ dans $\mathbb{R}^d$ et $\mathbb{R}^{d\times d}$, respectivement.

En termes du problème de contrôle, ceci se reflète dans l'Équation aux Dérivées Partielles (2.3) (en utilisant un développement de Taylor du second ordre en $\mathcal{L}^a$) par le fait que $\varepsilon V_T \to \bar{V}_T$ (voir par exemple [57]) où $\bar{V}_T$ est la solution de

$$0 = \partial_t \bar{V}_T + \sup_{a\in\mathbb{A}} \left\{\bar{\mathcal{L}}^a + r(\cdot, a)\right\} \text{ on } [0, T) \times \mathbb{R}^d, \tag{2.6}$$

avec la condition au bord $\bar{V}_T(T, \cdot) = 0$ sur $\mathbb{R}^d$, où, étant donné une fonction $\phi \in \mathcal{C}^{(1,2)}([0, T) \times \mathbb{R}^d; \mathbb{R})$, l'opérateur

$$\bar{\mathcal{L}}^a : \phi \mapsto \mu(\cdot, a)^\top \mathrm{D}_x\phi + \frac{1}{2}\,\mathrm{Tr}\left[\sigma(\cdot, a)\sigma(\cdot, a)^\top \mathrm{D}^2_{xx}\phi\right] \in \mathcal{C}^0([0, T) \times \mathbb{R}^d; \mathbb{R})$$

est le générateur infinitésimal de l'EDS (2.5) avec $\alpha \equiv a$. L'équation (2.6) est une équation de HJB diffusive standard dont l'analyse et la résolution numérique sont bien documentées, voir par exemple [78, 86] et les références qui y sont citées.

Cette méthode d'approximation via la limite diffusive a une histoire riche. Elle a été largement étudiée dans le contexte des réseaux de files d'attente, voir les livres [43, 77] ainsi que [62], mais a également été utilisée dans le contexte des sciences actuarielles [21, 45] et dans celui des mathématiques financières [95], entre autres. Néanmoins, il semble que, dans le contexte d'un problème général, le taux de convergence de cette approximation n'ait pas été étudié.

Le Chapitre 3 se concentre sur l'étude de ce taux de convergence vers la limite diffusive, avec $d = d' = 1$ pour simplifier. Nous considérons la famille renormalisée (indexée par $\varepsilon \in \mathbb{R}_+$) de problèmes prélimites

$$V_T^\varepsilon(t,x) := \sup_{\alpha \in \mathbb{A}} \mathbb{E}\left[\int_t^T \varepsilon r(X_{s-}^{t,x,\alpha}, \alpha_s) \mathrm{d}N_s^\varepsilon\right] \tag{2.7}$$

où N^ε est l'analogue de N quand $\eta = \eta_\varepsilon$. Afin de déterminer la convergence de V_T^ε vers $\bar{V}_T$, on utilise la méthodologie basée sur l'équation de HJB décrite ci-avant.

Nous débuterons par l'étude de la régularité de la solution $\bar{V}_T$ de (2.6) et montrons que $\bar{V}_T \in \mathcal{C}_b^{(1,2)}([0,T) \times \mathbb{R}; \mathbb{R}) \cap \mathcal{C}^0([0,T] \times \mathbb{R}; \mathbb{R})$ avec $\partial_{xx}\bar{V}_T$ étant γ-Hölder en espace pour un certain $\gamma \in (0,1)$. Nous montrons ensuite que l'ordre de convergence de V_T vers $\bar{V}_T$ est $\varepsilon^{\gamma/2}$ au sens où

$$\left|V_T^\varepsilon(t,x) - \bar{V}_T(t,x)\right| \leq C(T-t)\varepsilon^{\frac{\gamma}{2}}. \tag{2.8}$$

Nous montrerons ensuite comment utiliser (2.6) afin de construire un contrôle qui est $\varepsilon^{\gamma/2}$-optimal pour (2.7). Par la suite, nous montrerons comment construire des termes de correction d'erreur pour cette approximation, au premier ordre et à des ordres supérieurs, au prix d'une complexité accrue de l'approximation. Ceci est basé sur la construction d'une autre EDP, ou d'un système d'EDP, et sur l'utilisation extensive d'arguments de solutions de viscosité.

En utilisant un exemple inspiré par le problème des enchères séquentielles (voir aussi la Section 2.3), nous présenterons une démonstration de cette méthode d'approximation et montrons les gains computationnels qu'il est possible d'atteindre par rapport à la résolution directe de (2.3).

2.1.2 Contrôle Ergodique de la Limite Diffusive

Le Chapitre 3 a montré que l'approximation par limite diffusive peut conduire à des gains computationnels significatifs par rapport à la résolution directe de (2.3). Cependant, sur de nombreux problèmes concrets, l'horizon fini ne capture pas la nature réelle du problème de contrôle. Lorsque T est très grand, la question clé devient de savoir si le système entre dans un comportement stationnaire pour la plupart de la période étudiée, ou est dominé par un comportement transitoire.

Ces problèmes de contrôle à long terme, absents d'une forme naturelle d'escompte, sont mieux représentés par le problème de contrôle ergodique sur (2.1) défini par

$$\rho^*(x) := \sup_{\alpha \in \mathcal{A}} \liminf_{T \to +\infty} \frac{1}{T} \mathbb{E}\left[\int_0^T r(X_{s^-}^{0,x,\alpha}, \alpha_s) \mathrm{d}N_s \right] \tag{2.9}$$

pour tout $x \in \mathbb{R}^d$. Ce critère est aussi celui que l'on utilisera en apprentissage par renforcement, voir la Section 2.2. En raison de la limite inférieure, la faisabilité de (2.9) est intrinsèquement liée aux propriétés du processus $X^{0,x,\alpha}$ lui-même, et en particulier à la possibilité d'un comportement stationnaire (ergodicité).

Par exemple, si le système ne peut pas être stabilisé, c'est-à-dire s'il n'y a pas de contrôle qui puisse empêcher la masse du processus de s'échapper vers $+\infty$, alors ρ^* sera mal défini. Voir aussi [102, § 8.3.1] pour une discussion dans le cas du contrôle discret. Cette propriété illustre comment le problème ergodique capture certaines complexités fondamentales du problème de contrôle que d'autres cadres peuvent manquer, et souligne pourquoi il est particulièrement intéressant d'un point de vue théorique.

La résolution de la difficulté d'ergodicité de la limite inférieure, et donc la démonstration que (2.9) est bien posé, est réalisée grâce à des conditions de Lyapunov, voir par exemple [87]. Nous faisons des hypothèses de Lyapunov sur le processus dans lequel les fonctions de Lyapunov se comportent comme $\|\cdot\|^p$ pour un certain $p \geq 3$.

Au Chapitre 4, on montre d'abord l'ergodicité de (2.1) à partir de ses hypothèses de Lyapunov en utilisant des méthodes d'Équation Différentielle Ordinaire (EDO). Cette analyse montre des estimés uniformes de la Lipschitzité des fonctions de valeur du problème de contrôle escompté sur les processus à sauts purs (2.1), ce qui nous permet de montrer que le problème ergodique est bien posé via la méthode de la limite d'escompte nulle, voir par exemple [14, 25].

Cette méthode montre que le problème de contrôle est véritablement ergodique, en cela qu'il oublie la condition initiale : il existe une valeur $\rho^* \in \mathbb{R}$ telle que $\rho^*(x) = \rho^*$ pour tout $x \in \mathbb{R}^d$. De plus, la valeur ρ^* accompagnée

d'une fonction de décision auxiliaire $w : \mathbb{R}^d \to \mathbb{R}$, qui est Lipschitz, forment un couple de solution de viscosité à l'équation de HJB ergodique

$$0 = -\rho^* + \sup_{a \in \mathbb{A}} \left\{ \mathcal{L}^a w + \eta r(\cdot, a) \right\} \text{ on } \mathbb{R}^d. \tag{2.10}$$

qui est l'analogue ergodique de (2.3). Une application mesurable $\mathbb{R}^d \to \mathbb{A}$ maximisant ponctuellement le maximum dans (2.10) fournit une politique de décision stationnaire optimale et donc un contrôle de Markov optimal.

En utilisant une hypothèse de Lyapunov analogue et la même méthode d'EDO pour (2.5), on peut montrer le résultat analogue pour le problème limite diffusif. Notamment, il existe une solution Lipschitz $\bar{w} : \mathbb{R}^d \to \mathbb{R}$ à l'équation de HJB suivante

$$0 = -\bar{\rho}^* + \sup_{a \in \mathbb{A}} \left\{ \bar{\mathcal{L}}^a \bar{w} + r(\cdot, a) \right\} \text{ on } \mathbb{R}^d, \tag{2.11}$$

dans laquelle $\bar{\rho}^* \in \mathbb{R}$ est la valeur du problème de contrôle ergodique de (2.5), defini de manière analogue à (2.9).

Ayant montré que les deux problèmes sont bien posés et satisfont les équations de HJB requises, nous allons appliquer la méthodologie de la Section 2.1.1 à la résolution du problème de contrôle ergodique (2.9) et de l'équation de HJB (2.3). À cet effet, pour $\varepsilon \in \mathbb{R}_+$, nous définissons

$$\rho_\varepsilon^* := \sup_{\alpha \in \mathcal{A}} \liminf_{T \to +\infty} \frac{1}{T} \mathbb{E}\left[\int_0^T \varepsilon r(X_{s-}^{0,0,\alpha}, \alpha_s) \mathrm{d}N_s^\varepsilon \right] \tag{2.12}$$

où N^ε est définie comme dans Section 2.1.1.

On se concentre d'abord sur (2.11) et on s'appuie sur la Lipschitzité de $\bar{w}$ pour obtenir des estimée de régularité fortes de $\bar{w}$, pour $d \in \mathbb{N}$, en supposant que la volatilité σ est indépendante du contrôle $a \in \mathbb{A}$, bornée, et satisfait une condition d'ellipticité uniforme. Cette analyse montre que, pour μ Lipschitz (mais pas nécessairement bornée), $D^2\bar{w}$ est localement γ-Hölder continue pour tout $\gamma \in (0, 1)$, avec une constante croissant au plus linéairement en x.

Cette régularité estimée nous permet d'appliquer les arguments de la Section 2.1.1, et y ajoutant l'utilisation de conditions de moment (dérivées de conditions de Lyapunov) sur $X_t^{0,x,\alpha}$, pour tout $t \in \mathbb{R}_+$, afin de gérer la croissance linéaire de μ et la constante de Hölder de $\bar{w}$. Ceci nous donne la borne d'approximation suivante dans le cas ergodique avec dérive non bornée, ellipticité uniforme, et volatilité non contrôlée

$$|\rho_\varepsilon^* - \bar{\rho}^*| \leq C\varepsilon^{\frac{\gamma}{2}}.$$

On montre aussi comment construire un contrôle qui est $\varepsilon^{\gamma/2}$-optimal pour le problème de contrôle ergodique (2.12) en utilisant la politique de décision optimale obtenue à partir de (2.11).

Ensuite, on montre comment construire des termes de correction d'erreur pour cette approximation, mais en utilisant une méthode différente de celle du Chapitre 3. Dans ce cas, l'augmentation de la complexité vient de l'itération d'EDP supplémentaires plutôt que de la construction d'un système d'EDP de plus en plus grand. Cette méthode est plus simple que celle du Chapitre 3 et repose sur un argument de vérification.

Continuant vers l'objectif d'une résolution efficace de (2.9), nous nous plongeons dans la résolution numérique du problème de contrôle ergodique. Puisque les schémas numériques nécessitent généralement des constructions ad-hoc, nous prenons $d = 1$ une fois de plus, et renforçons les conditions de régularité et de vitesse de croissance dans nos hypothèses. Sous ces hypothèses, nous construisons un schéma numérique pour (2.11), dans lequel nous résolvons une équation de la forme

$$0 = -\bar{\rho}_h^{\kappa,*} + \sup_{a\in\mathbb{A}} \left\{ \bar{\mathcal{L}}_h^a \bar{w}_h^\kappa + r(\cdot, a) \right\} \tag{2.13}$$

sur une grille $\bar{M}_h^\kappa$ sur l'intervale $[-\kappa h, \kappa h]$ de finesse $h > 0$ contenant $N_\kappa := 2\kappa + 1$ points, $\kappa \in \mathbb{N}^*$, avec

$$\bar{\mathcal{L}}_h^a : v \in \mathbb{R}^{N_k} \mapsto P_h v \in \mathbb{R}^{N_k}$$

pour une matrice de transition $P_h \in \mathbb{R}^{N_\kappa \times N_\kappa}$ qui est obtenue via une approximation par différences finies de $\bar{\mathcal{L}}^a$. Le couple solution $(\bar{\rho}_h^{\kappa,*}, \bar{w}_h^\kappa)$ de (2.13) correspond à la valeur et à la fonction de décision, respectivement, d'un problème de contrôle ergodique sur une chaîne de Markov en temps continu.

Puisque cette structure correspond à un processus à sauts purs sur $\bar{M}_h^\kappa$, dont l'intensité est $1/\Delta t_h$, pour $\Delta t_h = \mathcal{O}(h^2)$, nous pouvons utiliser la même méthodologie d'approximation décrite ci-dessus, modulo une certaine prudence dans le traitement des conditions aux bords du schéma, pour obtenir une convergence d'ordre

$$\left| \Delta t_h \bar{\rho}_h^{\kappa,*} - \bar{\rho}^* \right| \leq C \left(h^\gamma + h^{-1} \left| \kappa h \right|^{1-p} \right),$$

où $p \geq 3$ est la puissance dans la vitesse de croissance de la fonction de Lyapunov. Nous étudions aussi la construction d'un contrôle approximativement optimal pour (2.12) en utilisant $\bar{w}_h^\kappa$, ce qui semble être un sujet nouveau concernant la résolution numérique des problèmes de contrôle.

Nous nous tournons ensuite de nouveau vers un problème d'enchères qui propose une variante de celui du Chapitre 3. Nous utilisons cet exemple numérique pour appliquer le schéma numérique de (2.13) et illustrer les gains d'efficacité computationnelle de l'approximation par limite diffusive.

2.2 Apprentissage par renforcement

L'apprentissage par renforcement étudie le problème du contrôle d'un système tel que (2.1) sous incertitude sur sa dynamique du point de vue de l'apprentissage automatique. Il y a deux façons principales de formaliser cette incertitude, selon que l'on choisit de prendre la dynamique ou le contrôle comme objet d'étude principal. Ces approches sont connues sous les noms d'approche basée sur la valeur et de recherche de politique, respectivement.

Nous nous concentrons sur l'approche basée sur la valeur et adoptons une perspective dite "model-based", c'est-à-dire basée sur une modélisation. Cette perspective considère que b dans (2.1) est inconnu mais que l'on a accès à une classe de modèles $\mathscr{F}_\Theta$ de dynamiques possibles, paramétrée par $\theta \in \Theta \subset \mathbb{R}^{d_\Theta}$, $d_\Theta \in \mathbb{N}^*$. Ce paradigme suppose que toute l'incertitude est capturée, dans $\mathscr{F}_\Theta$, par le manque de connaissance d'une vraie paramétrisation $\theta^* \in \Theta$. Cela nous amène à étendre la définition de $X^{t,x,\alpha}$ pour incorporer un paramètre de pilotage $\theta \in \Theta$ en définissant $X^{t,x,\alpha,\theta}$ comme la solution de l'Équation Différentielle Stochastique suivante

$$X^{t,x,\alpha,\theta} = x + \int_t^\cdot \int_{\mathbb{R}^{d'}} b_\theta(X_{s-}^{t,x,\alpha,\theta}, \alpha_s, e) N(\mathrm{d}e, \mathrm{d}s)\,. \tag{2.14}$$

Nous adoptons le régime de limite diffusive de la Section 2.1 pour b_θ, et puisque la structure de bruit standard en apprentissage par renforcement est une martingale additive, nous prendrons $b_\theta(x, a, e) := \varepsilon \bar{\mu}_\theta(x, a) + \varepsilon^{1/2} \bar{\Sigma} e$, avec les conditions de régularité sur $(\bar{\mu}_\theta, \bar{\Sigma})$ énoncées dans la Section 2.1.2. Ce cadre est destiné à modéliser des systèmes déterministes sous perturbation par une martingale sousgaussienne, nous prendrons donc ν comme une mesure gaussienne centrée standard sur $\mathbb{R}^d$ pour simplifier.

Quelle que soit la forme du problème de contrôle dont il est question, en tant qu'approche basée sur la valeur nous cherchons à explorer l'espace d'états-actions pour apprendre b_θ et la fonction de récompense r avec pour objectif de bien approximer le problème de contrôle. Cela nous conduit nécessairement à échantillonner des paires état-action sous-optimales pour le problème de contrôle, créant une tension inhérente entre exploration et contrôle. Inversement, un agent avide qui se concentre uniquement sur l'exploitation de l'état qui apparaît empiriquement le meilleur à un instant donné risque d'échouer à apprendre. En effet, il est susceptible de créer un sous-système en boucle fermée dans lequel il reste dans une région de l'espace pour toujours en raison d'observations malheureuses en dehors de cette région, incapable d'apprendre le reste du système.

Cette tension est une conséquence d'apprendre un système de décision de l'intérieur, tel que cela a été observé en étudiant les problèmes de bandits

manchots. Sa source est la nature de la rétroaction des décisions : on n'obtient des informations que sur les paires état-action effectivement traversées.

En conséquence, puisque ε et $\bar{\Sigma}$ sont indépendants des actions prises, ils ne sont pas affectés par ce compromis et peuvent être estimés par des méthodes statistiques simples. Pour éviter une complexité inutile, nous les prendrons simplement comme connus.

Puisque X^{0,x,α,θ^*} de l'équation (2.14) est un processus stochastique, il y a deux notions d'échantillons que l'on pourrait considérer : *épisodique* dans lequel on observe des réalisations du processus $X^{0,x,\alpha,\theta^*}(\omega)$ à partir de différents $\omega \in \Omega$, éventuellement jusqu'à un temps de réinitialisation $T \in \mathbb{R}_+$; ou bien *en ligne* dans lequel on observe au fil du temps $t \in \mathbb{R}_+$ une trajectoire $X^{0,x,\alpha,\theta^*}(\omega)$ pour un seul ω fixé, comme dans une série temporelle.

Dans l'apprentissage en ligne, on ne peut pas revisiter les événements passés en comptant sur la réinitialisation de la trajectoire. Au lieu de cela, le système doit être ramené dans les conditions pertinentes de l'intérieur. Cela caractérise la difficulté d'apprentissage d'une manière inhérente à la dynamique du processus $X^{t,x,\alpha,\theta}$. En même temps, cela exclut l'étude des politiques non stationnaires, y compris le problème à horizon fini car les effets de début du monde ne peuvent pas être raisonnablement appris le long d'une seule trajectoire.

Nous choisissons de nous concentrer sur le problème en ligne, suivant la vaste littérature de la théorie de l'apprentissage par renforcement en ligne, et étudierons le critère ergodique

$$\rho^*_{\theta^*}(x) := \sup_{\alpha\in\mathcal{A}} \liminf_{T\to+\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T r(X^{0,x,\alpha,\theta^*}_{s-}, \alpha_s)\mathrm{d}N_s\right],$$

dans lequel $\mathcal{A}$ est défini comme dans la Section 2.1 par rapport à X^{0,x,α,θ^*}, puisque celui-ci cerne la difficulté inhérente au processus $X^{t,x,\alpha,\theta}$ dans le problème de contrôle. Comme nous l'avons vu dans la Section 2.1.2, il admet également une politique optimale (voir (2.10)) qui est stationnaire, ce qui signifie qu'il est possible d'apprendre une politique optimale en ligne.

Choisir un problème ergodique introduit les problèmes de bien-poséité de $\rho^*_{\theta^*}$ mentionnés dans la Section 2.1.2, même dans le cas à états discrets, voir par exemple [103, § V.] et [12]. Jusqu'à présent, ce problème d'ergodicité était un obstacle majeur dans l'apprentissage par renforcement avec des états continus en dehors du cas de la commande linéaire quadratique [7, 81].

En effet les méthodes utilisées pour établir l'ergodicité dans les systèmes discrets, qui sont basées sur l'étude de leurs matrices de transition, voir par exemple [102, § 8.], ne s'étendent pas facilement aux problèmes continus. Dans le cadre général qui a été décrit dans la Section 2.1.2, nous avons mon-

tré que le problème (2.9) était bien posé sous des conditions modérées sur les coefficients d'une part et de type Lyapunov d'autre part.

Au Chapitre 5, nous spécialisons cette analyse à la structure que prends b_θ dans (2.14), ce qui nous permet de relâcher les hypothèses de Lyapunov du Chapitre 4 à une seule condition plus faible sous laquelle on a l'ergodicité et la stabilité du problème à sauts purs et celles du problème limite diffusif. Cette condition est une condition de mélange (contraction) sur la dynamique instantanée, qui prend la forme suivante

$$\mathscr{V}(x + \varepsilon\bar{\mu}_\theta(x,a) - (x' + \varepsilon\bar{\mu}_\theta(x',a))) \leq (1 - \mathfrak{c}_{\mathscr{V}}\varepsilon)\mathscr{V}(x - x') \tag{2.15}$$

pour tout $(x, x', a, \theta) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{A} \times \Theta$, dans laquelle $\mathscr{V} \in \mathcal{C}^2(\mathbb{R}^d \setminus \{0\}, \mathbb{R}_+)$ est une fonction de Lyapunov appropriée et $\mathfrak{c}_{\mathscr{V}} > 0$ est une constante indépendante de ε. En utilisant le bruit sous-gaussien additif, nous étendons cette hypothèse au-delà de l'ergodicité et montrons que $(X_s^{0,x,\alpha,\theta})_{s\in\mathbb{R}_+}$ se comporte essentiellement comme un processus sous-gaussien, y compris la concentration avec haute probabilité. Cela montre que ρ_θ^* est bien posé pour tout $\theta \in \Theta$ par les arguments de la Section 2.1.2.

Ayant montré que le problème est bien posé, nous pouvons chercher à construire un agent $\alpha \in \mathcal{A}$ qui maximise (2.9) en l'absence de connaissance de θ^*, c'est-à-dire un algorithme optimal d'apprentissage. Malheureusement, ce problème explicite de maximisation est resté hors de portée, même dans les problèmes simples (à états discrets ou de bandit).

En lieu et place, la littérature sur l'apprentissage par renforcement s'est concentrée sur la conception d'algorithmes ad hoc, et sur la compréhension et la comparaison de leurs performances en utilisant des bornes supérieures de performance. Cette méthodologie est complétée par des bornes inférieures de difficulté (minimax), ce qui est une méthodologie standard en informatique théorique.

La mesure principale utilisée dans ces bornes est la notion de regret d'un agent $\alpha \in \mathcal{A}$

$$\mathcal{R}_T(\alpha) := T\rho_{\theta^*}^* - \int_0^T r(X_{s-}^{0,x,\alpha,\theta^*}, \alpha_s)\mathrm{d}N_s\,, \tag{2.16}$$

définie pour tout $(x, T) \in \mathbb{R}^d \times \mathbb{R}_+$, que l'on borne dans le sens de la haute probabilité. Ce concept est naturellement adapté aux problèmes de décision en ligne, voir par exemple [58, 109], car elle quantifie de manière transparente le coût de l'incertitude sur θ^* directement en termes de la fonction de récompense du problème de contrôle. Comme on peut le voir en divisant par T et en laissant $T \to +\infty$, le taux de croissance du regret quantifie l'efficacité de α à apprendre à contrôler et, dans un sens minimax, la difficulté inhérente du problème.

En l'absence d'un agent optimal au sens formel, de nombreux agents d'exploration ont été proposés. Tous ajoutent un comportement exploratoire à l'agent avide, qui fut discuté précédemment. Une possibilité est de forcer une exploration isotrope, soit en prenant des actions aléatoires une partie du temps (agent ϵ-avide, et d'autres formes de politiques à entropie régularisée), soit en alternant des phases d'exploration et d'exploitation (Explore-Then-Commit, voir par exemple [100]).

Bien que ces méthodes puissent être efficaces, la calibration de leurs sources d'aléa requiert généralement une connaissance a priori du système. En l'absence de telles informations, leurs regrets sont généralement sous-optimaux. Pour éviter ce problème, il faut structurer le comportement exploratoire en fonction du problème, ce qui est fait par une étude et une décomposition minutieuses du regret afin d'identifier et de traiter les principales sous-tâches du dilemme exploration-exploitation. Des travaux ayant fait date ont montré que ce dernier pouvait être abordé par l'échantillonnage de Thompson, voir par exemple [6, 113], ou par l'utilisation du paradigme de l'optimisme devant l'incertain, voir par exemple [15, 16, 68].

Ce paradigme identifie quatre sous-tâches clés du problème : l'apprentissage, c'est-à-dire la capacité à estimer $\bar{\mu}_{\theta^*}$ et r et à donner des ensembles de confiance pour eux au niveau de confiance δ, $\delta \in (0,1)$; l'optimisme, qui est fait en sélectionnant un système de croyance $\tilde{\theta}$ qui maximise $\rho^*_{\tilde{\theta}}$ parmi les θ dans cet ensemble de confiance ; la planification, qui calcule la politique de décision optimale $\tilde{\pi}^*$ pour $\tilde{\theta}$; et les mises à jour parcimonieuses, qui garantissent que l'agent exécute la politique $\tilde{\pi}^*$ suffisamment longtemps pour obtenir des informations utiles sur $\tilde{\theta}$ en explorant le système.

Les fondations de la tâche de planification ont été posées à la Section 2.1.2, dans laquelle nous avons montré comment utiliser l'équation de Hamilton-Jacobi-Bellman (2.10) pour résoudre le problème de contrôle ρ^*_θ, pour tout $\theta \in \Theta$. Nous avons également montré une méthode de résolution approximative efficace basée sur la limite de diffusion et le schéma numérique de (2.13). Cela fournit les outils pour une planification efficace dans notre cadre proche de la limite diffusive.

Apprendre des dynamiques non-linéaires continues a été traité dans la littérature de l'apprentissage par renforcement via l'introduction de "l'eluder dimension", que l'on pourrait traduire par dimension d'évitement, dans les articles fondateurs [97, 107]. Ce cadre ne sied malheureusement qu'aux coefficients bornés. Au Chapitre 5, nous rafinerons les arguments originaux de ces travaux afin de les appliquer à un processus stable à la dérive (c.a.d. $\bar{\mu}_{\theta^*}$) non borné sur $\mathbb{R}^d$, ce qui requiert une attention plus fine à la mesurabilité et aux arguments de couverture. En conséquence, les quantités liées à la difficulté d'apprentissage deviennent adaptives aux régions de l'espace qui sont effectivement visitées par la trajectoire $X^{0,x,\alpha,\theta^*}(\omega)$.

Nous montrerons que les ensembles de confiance résultant de ces modifications sont bien calibrés, et que cela montre que l'erreur d'estimation est faible. Cela nous permet d'utiliser les estimations ponctuelles et les ensembles de confiance de la régression des moindres carrés non-linéaires pour résoudre le problème d'apprentissage pour les dynamiques non-linéaires, sous une condition de Lipschitz sur $\bar{\mu}_\theta$ et les propriétés de stabilité de X^{0,x,α,θ^*} dérivées de (2.15).

Le besoin de mises à jour parcimonieuses est une conséquence contre-intuitive de la nature dynamique de (2.14). Puisque la politique $\tilde{\pi}^*$ optimise le critère ergodique, elle peut générer peu de récompenses à court terme jusqu'à ce que le système se mélange. Il y a donc un coût potentiel à chaque mise à jour de la politique et une façon de contrôler l'erreur résultante est de ne changer $\tilde{\theta}$ qu'infréquemment. La difficulté est de le faire sans dégrader l'apprentissage. Nous montrerons, en utilisant des inégalités fonctionnelles sur mesure pour les suites, que l'erreur de prédiction faible ($\mathcal{O}(\log(T))$) peut être utilisée pour dériver un critère de mise à jour qui ne se déclenche que $\mathcal{O}(\log(N_T))$ fois jusqu'au temps T, au prix d'un facteur constant de dégradation de l'efficacité de l'apprentissage.

Ayant ainsi traité les sous-tâches, nous serons en mesure de donner un algorithme qui génère un processus de contrôle $\varpi \in \mathcal{A}$ et de borner son regret. Nous montrerons que

$$\mathbb{P}\left(\mathcal{R}_T(\varpi) \leq \mathcal{O}\left(\sqrt{T d_{E,N_T} \log(\mathcal{N}^{\varepsilon}_{N_T}) \log(N_T) \log\left(\frac{1}{\delta}\right)}\right)\right) \geq 1-\delta\,, \qquad (2.17)$$

où d_{E,N_T} et $\mathcal{N}^{\varepsilon}_{N_T}$ sont la $2\sqrt{\varepsilon/N_T}$-dimension d'évitement (voir [107, Def. 4.] et (5.56) à la Section 5.5.2) et le nombre de recouvrement, avec une finesse de $\varepsilon\|\bar{\Sigma}\|_{op}/N_T$, de la classe contenant les restrictions à une boule de rayon $\mathcal{O}(\sqrt{\log(T/\varepsilon)})$ des éléments de $\mathscr{F}_\Theta$.

Ceci concorde avec les meilleures bornes supérieures connues pour le regret dans le cas à états discrets [20, 68, 97], modulo des facteurs logarithmiques. En guise de comparaison, les bornes inférieures connues pour le regret dans le cas à états discrets sont de l'ordre de $\Omega(\sqrt{T})$, voir par exemple [68], ce qui signifie qu'il n'existe pas d'algorithme qui puisse atteindre une croissance du regret plus lente que $\sqrt{T}$ sur toutes les instances.

En utilisant l'approximation par la limite de diffusive du Chapitre 4, nous pouvons également résoudre le problème de planification efficacement au détriment d'un terme additif de l'ordre de $\varepsilon^{\gamma/2}T$, ce qui est linéaire en T, dans la borne de regret. Cela montre que dans les systèmes à très haute fréquence, avec $\varepsilon \ll 1$, la borne de regret (2.17) est atteignable de manière réaliste en utilisant ce type d'approximations pour économiser des efforts de calcul.

Ce travail, dans le contexte spécifique des systèmes admettant une limite diffusive, montre comment les outils du contrôle stochastique et des processus stochastiques en temps continu peuvent être appliqués aux problèmes d'apprentissage par renforcement à espace d'états et d'actions continu. Il fournit une première étape vers une théorie générale de l'apprentissage par renforcement en espace d'états et d'actions continu basée sur ces outils.

2.3 Problèmes d'enchères séquentielles

Les systèmes de contrôle à haute fréquence du type qui fut décrit aux Sections 2.1 et 2.2 sont courants dans les applications du monde réel. Ils modélisent souvent un système qui surveille certains signaux d'événements extérieurs auxquels il réagit. Une catégorie typique d'exemples implique l'achat ou la vente d'objets par des algorithmes via des enchères, des marchés ou d'autres mécanismes. À grande échelle, ceux-ci impliquent des événements à haute fréquence, dont chacun ne contribue que de petits changements au processus d'état pertinent, qui pourrait être, par exemple, le prix, le volume d'un inventaire de stocks ou un objet plus complexe tel qu'un portefeuille d'investissements.

Bien que les marchés financiers soient l'exemple le plus évident de systèmes à haute fréquence, ils ne sont pas bien représentés par des arrivées d'un processus de Poisson à toutes les échelles, notamment en raison de phénomènes d'auto-excitation, voir par exemple [52]. Au lieu de cela, nous nous concentrons tout au long de cette thèse sur l'application aux enchères séquentielles d'affichage publicitaire, qui modélise la façon dont les bannières publicitaires sur les pages Web sont vendues durant le temps de chargement de la page. L'objectif global du Chapitre 6 est d'illustrer par une application comment un système de contrôle du type étudié aux Chapitres 3 et 4 peut émerger comme une réponse naturelle à une pression de la difficulté de calcul issue de la haute fréquence des interactions.

Le modèle économique de la plupart des applications sur internet repose sur la publicité : soit directement, soit indirectement via la vente de données. Il existe de nombreux types de publicités, chacun ayant ses propres mécanismes de négociation. Nous nous concentrerons sur les publicités dites "*display*", qui sont celles qui se chargent dans des bannières, généralement situées en haut ou sur les côtés des sites Web. Les emplacements de bannières sur une page Web sont vendus séquentiellement et individuellement aux enchères par des algorithmes automatisés pendant le temps de chargement de la page (environ 50 ms). Cela rend les temps d'enchères dépendants uniquement de l'arrivée des utilisateurs, qui devrait être raisonnablement modélisée par un processus de Poisson. En termes de fréquence, l'ordre de grandeur est de 10^{10} à 10^{12} enchères par jour, tandis que les coûts et les gains indivi-

duels sont très faibles puisque la probabilité qu'un utilisateur clique sur une annonce est faible.

La particularité clé du marché de la publicité display en comparaison avec d'autres marchés, tels que les marchés financiers, est sa concentration. En raison de l'infrastructure informatique requise par l'échelle des enchères automatisées, la négociation est effectuée par des intermédiaires de part et d'autre, de sorte qu'il n'y a qu'une demi-douzaine de participants à chaque enchère. Cela a un impact important sur la nature du problème de maximisation des revenus pour un vendeur.

En théorie des enchères, la question de l'enchère optimale, étant donné la connaissance des distributions de valorisation des enchérisseurs, a été résolue dans l'article fondateur de Myerson [92]. Manifestement, cette hypothèse est irréaliste pour la publicité display. D'un autre côté, la haute fréquence des événements signifie que beaucoup de données sont disponibles pour apprendre à maximiser les revenus à partir d'observations empiriques. Cependant, en raison de la complexité du mécanisme de Myerson [92], il est impossible de l'apprendre directement [90], et nous étudierons donc les enchères au second prix.

Ce format est à la fois une approximation raisonnable de l'enchère de Myerson, voir [104], et le format qui a été historiquement utilisé dans la publicité display[1]. Dans ce format, la mise gagnante est simplement la plus élevée, mais l'objet n'est vendu que si l'enchérisseur gagnant dépasse son *prix de réserve*, qui est un prix de vente minimum individualisé. Si le prix de réserve est dépassé, le gagnant paie la plus petite mise qui lui aurait permis d'encore remporter l'objet, c'est-à-dire le maximum entre son prix de réserve et les mises des autres enchérisseurs.

Puisqu'ils sont individualisés, le problème de l'optimisation des prix de réserve pour maximiser le revenu est séparable, et se résume à trouver le *prix de monopole* r^* de chaque enchérisseur. Pour un enchérisseur dont les mises suivent une distribution stationnaire F, celui-ci est le maximiseur du revenu de monopole

$$\Psi^F : r \in \mathbb{R}_+ \mapsto \int_{\mathbb{R}_+} r 1_{r \geq B} F(\mathrm{d}B) \in \mathbb{R}_+ . \tag{2.18}$$

Plusieurs méthodes ont été proposées pour maximiser (2.18), principalement basées sur les bandits, par exemple [73], ou sur l'optimisation non convexe, par exemple [89]. Malheureusement, pour mettre à jour le prix ces méthodes nécessitent toutes de stocker de nombreuses enchères passées, voire leur totalité, ce qui deviendra informatiquement impossible aux échelles qui nous intéressent.

[1] Bien qu'il convienne de noter que le marché a évolué ces dernières années en faveur de l'enchère au premier prix et de formats similaires.

La question qui sous-tend le Chapitre 6 est de savoir s'il est possible de concevoir un algorithme dont la mise à jour est en temps réel (c'est-à-dire indépendante du nombre d'enchères passées), même au prix d'une dégradation de la vitesse de convergence, afin qu'il puisse être appliqué en ligne au flux de données.

Un candidat alléchant est une méthode du premier ordre comme la montée de gradients stochastique en ligne. Malheureusement, puisque $(r, B) \in \mathbb{R}_+^2 \mapsto p(r, B) := r1_{r \geq B}$ est discontinue (en r pour toute mise $B \in \mathbb{R}_+^*$), les gradients aléatoires issus de cette méthode ne donneront pas un estimateur non biaisé de gradient de Ψ^F. Par conséquent, une montée de gradients de ce type n'a aucune raison de converger vers $r^* \in \operatorname{argmax} \Psi^F$.

Le problème des gradients biaisés peut être résolu en lissant p par convolution avec un noyau lisse k, puis en appliquant la montée de gradients stochastique à ce résultat. L'algorithme résultant, Conv-OGA, est de ce fait biaisé et nous aimerions réduire son biais au fil du temps en réduisant le lissage. La contribution principale du Chapitre 6 est un algorithme, V-Conv-OGA, qui effectue le lissage et la montée de gradients stochastiques en même temps afin de faire un compromis entre le biais des gradients et la stabilité de la montée de gradients, le tout en conservant les mises à jour en temps réel.

Nous commencerons par une analyse des propriétés de convexité[2] de Ψ^F sous des hypothèses classiques de la théorie des enchères comme la monotonie de la valeur virtuelle ou la croissance du taux de défaillance. Cela nous permettra de contrôler le compromis biais-variance induit par le lissage, ce que nous utiliserons pour montrer la convergence presque sûre des itérés vers r^* en utilisant la méthode des quasi-martingales, voir par exemple [32].

Dans un second temps nous renforcerons ce résultat en modifiant les arguments de [91] pour obtenir un taux de convergence sous une hypothèse plus forte qui élimine les gradients arbitrairement petits. Les bornes complètes exhibent le compromis biais-variance, mais en optimisant le compromis elles correspondent à

$$\mathbb{E}\left[\|r_n - r^*\|^2\right] = \tilde{\mathcal{O}}\left(n^{-\frac{1}{2}}\right)$$

où $(r_n)_{n \in \mathbb{N}^*}$ sont les itérés de notre algorithme et $\tilde{\mathcal{O}}$ cache des facteurs polylogarithmiques en n dans la notation de l'ordre. Puisque les méthodes qui stockent toutes les mises, par exemple [89], peuvent atteindre un taux de $\tilde{\mathcal{O}}(n^{-1})$, l'écart correspondant peut être vu comme le coût d'utilisation d'un algorithme en temps réel.

En pratique, les enchérisseurs dans les enchères display sont connus pour présenter un comportement non stationnaire voire stratégique. Par la

[2] Plus précisément la log-concavité et la pseudo-concavité.

nature de la montée de gradients stochastique en ligne, nous pouvons facilement adapter Conv-OGA, avec un taux d'apprentissage constant $\gamma_0 \in \mathbb{R}_+$, pour gérer un enchérisseur non stationnaire. Étant donné un nombre raisonnable de changements de distribution de mise par le temps $N \in \mathbb{N}$, nous montrons un regret dynamique relatif à r^* d'ordre $\mathcal{O}(\sqrt{N})$.

Étant donné une séquence d'enchères $(b_n)_{n\in\mathbb{N}^*} \subset \mathbb{R}_+$, les prix de réserve $(r_n)_{n\in\mathbb{N}^*}$ générés par Conv-OGA sont donnés par un prix initial $r_1 \in \mathbb{R}_+$ et la récursion

$$r_{n+1} = r_n + \gamma_0 f(r_n, b_n)\,, \tag{2.19}$$

dans laquelle $f(r, B) := \mathrm{D}[p(\cdot, B) \star k](r) \vee -r$, pour un noyau de convolution k et $\gamma_0 > 0$.

Dans les faits, il existe de nombreuses sources de bruit qui affectent les prix de réserve du vendeur, même avec une mise à jour déterministe comme (2.19), telles que la difficulté d'attribuer les mises au bon enchérisseur. Cela appelle des mises à jour prudentes et, par conséquent, les incrémentations de prix de réserve ont tendance à être négligeables, c'est-à-dire $\gamma_0 \ll 1$.

En parallèle, du point de vue d'un enchérisseur, d'autres sources de bruit existent également, telles que l'agrégation possible par le vendeur sur des informations exogènes inconnues. Ces multiples niveaux de bruit expliquent pourquoi, en pratique, le bruit domine la dynamique de Conv-OGA, comme dans le régime de limite de diffusion de la Section 2.1.

Puisque les dates d'arrivées des enchères sont bien modélisées par un processus de Poisson, le prix de réserve est observé par l'enchérisseur selon l'Équation Différentielle Stochastique contrôlée

$$X^{0,x_0,\alpha} = x_0 + \int_0^{\cdot} \int_{\mathbb{R}^{d'}} \varepsilon b_1(X_{s-}^{0,x_0,\alpha}, \alpha_s, e) + \varepsilon^{\frac{1}{2}} b_2(X_{s-}^{0,x_0,\alpha}, \alpha_s, e) N(\mathrm{d}e, \mathrm{d}s), \tag{2.20}$$

dans laquelle $(\alpha_t)_{t\in\mathbb{R}_+}$ est le processus de mise. Cela correspond à ce qu'on a étudié à la Section 2.1. Ainsi, l'objectif de l'enchérisseur, qui est de miser de manière optimale afin de maximiser ses gains, est un exemple de problème de contrôle stochastique de la forme considérée dans la Section 2.1.

Pour le vendeur, $b_1(x, a, e) = f(x, a)$ représente un choix optimal d'une dynamique de prix de Markov stationnaire, au sens décrit ci-dessus. Néanmoins la définition de f via une convolution complique la manipulation du processus lors de l'étude du problème de l'enchérisseur. Cela nous amène à considérer des heuristiques simplificatrices dans nos applications.

Considérant, pour simplifier, que $b_2(x, a, e) = e$, l'heuristique la plus simple est de prendre $b_1(x, a, e) := \beta(a - x)$, pour $\beta \in \mathbb{R}_+$. Dans la Section 4.5,

nous considérons cette dynamique avec β aléatoire pour modéliser l'incertitude sur l'agressivité avec laquelle le vendeur pousse son prix à suivre les mises du vendeur. Dans la Section 3.4, nous considérons $b_1(x, a, e) := -x + qb + (1 - q)r_0$, $q \in (0, 1)$, qui prend une combinaison convexe de la dernière mise avec un prix de référence $r_0 \in \mathbb{R}_+$.

Bien que ces modèles heuristiques ne soient pas des modèles définitifs du comportement des vendeurs dans les enchères display, ils sont intéressants car ils présentent un exemple d'émergence d'un comportement (markovien) de limite diffusive. Ce comportement est apparu naturellement en réponse au problème de maximisation des revenus dans un format d'enchères hautement concentré, face aux contraintes de calcul lourdes des événements à haute fréquence.

Le Chapitre 6 motive ainsi les exemples adoptés dans les Chapitres 3 et 4 pour les expériences numériques, et resitue dans le monde réel la méthode de la limite diffusive. Notez que le prix de monopole n'est pas la seule quantité d'intérêt dans les enchères display, et que de nombreuses autres variables d'état et problèmes de contrôle peuvent être considérées, voir par exemple [54] pour d'autres exemples en gestion des stocks et maximisation de la conversion publicitaire.

Chapter 3

Diffusive Limit Approximation of Optimal Control Problems

We consider the diffusive limit of a typical pure-jump Markov control problem as the intensity of the driving Poisson process tends to infinity. We show that the convergence speed is provided by the Hölder exponent of the Hessian of the limit problem, and explain how correction terms can be constructed. This provides an alternative efficient method for the numerical approximation of the optimal control of a pure-jump problem in situations with very high jump intensity. We illustrate this approach in the context of a display advertising auction problem[a].

[a] This Chapter appeared as an article in the *Journal of Optimisation Theory and Applications*, see [4]

* * *

Contents

* * *

3.1 Introduction

Let N be a random point process with predictable compensator $\eta\nu(\mathrm{d}e)\mathrm{d}t$, for some probability measure ν on $\mathbb{R}$, $\eta > 0$, and let $X^{t,x,\alpha}$ be the solution of

$$X^{t,x,\alpha} = x + \int_t^{\cdot} \int b(X^{t,x,\alpha}_{s-}, \alpha_s, e) N(\mathrm{d}e, \mathrm{d}s)\,,$$

in which α belongs to the set $\mathcal{A}$ of predictable controls with values in some given set $\mathbb{A}$. Then, under mild assumptions, the value of the control problem

$$V_T(t,x) := \sup_{\alpha\in\mathcal{A}} \mathbb{E}\left[\int_t^T r(X^{t,x,\alpha}_{s-}, \alpha_s)\mathrm{d}N_s\right],$$

with $N_t := N(\mathbb{R}, [0,t])$, $t \geq 0$, solves the integro-differential equation

$$\partial_t V_T + \eta \sup_{a\in\mathbb{A}} \left(\int V_T(\cdot, \cdot + b(\cdot, a, e))\nu(\mathrm{d}e) - V_T + r(\cdot, a)\right) = 0 \tag{3.1}$$

on $[0,T)\times\mathbb{R}$, with boundary condition $V_T(T,\cdot) = 0$, possibly in the sense of viscosity solutions. From this characterization, standard numerical schemes follow that allow one to approximate both the value function V_T and the associated optimal control.

However, (3.1) is non-local and obtaining a precise approximation of the solution is highly time-consuming as soon as the intensity η of N is large. This is the case, for instance, for ad-auctions on the web, see e.g. [54] and Chapter 6, that are posted almost in continuous time, and on which one would typically like to apply reinforcement learning techniques based on the resolution of (3.1) for the current estimation of the parameters, leading to a possibly large number of resolutions for different sets of parameters. On the other hand, when η is very large, it is tempting to approximate the original jump diffusion control problem by its asymptotic as $\eta \to \infty$.

In this thesis, we consider the diffusive limit approximation. Namely, if one takes η of the form $\eta = 1/\varepsilon$, with ε small, and $b = \varepsilon b_1 + \sqrt{\varepsilon} b_2$ with $\int b_2(\cdot, e)\nu(\mathrm{d}e) = 0$, then a second order Taylor expansion on (3.1) implies that εV_T converges as $\varepsilon \to 0$ to the solution $\bar{V}_T$ of

$$0 = \partial_t \bar{V}_T + \sup_{\bar{a}\in\mathbb{A}} \left(\int b_1(\cdot, \bar{a}, e)\nu(\mathrm{d}e)\partial_x \bar{V}_T \right.$$
$$\left. + \frac{1}{2}\int |b_2|^2(\cdot, \bar{a}, e)\nu(\mathrm{d}e)\partial^2_{xx}\bar{V}_T + r(\cdot, \bar{a})\right), \tag{3.2}$$

$$0 = \bar{V}_T(T,\cdot). \tag{3.3}$$

The advantage of the above is that it is now a local equation that can be solved in a much more efficient way. Note that another possibility is to consider a first-order expansion as in [54], which corresponds to considering a fluid limit, but this is less precise.

For such a specification of the coefficients (η, b), the existence of a diffusive limit is expected, see e.g. [66] for general results on the convergence of stochastic processes. For control problems, the convergence of the value function can be proved by using the stability of viscosity solutions as in [57, §3], which considers the limit of discrete time zero-sum games, or by applying weak-convergence results. In particular, a large body of work on this subject exists within the insurance and queueing networks literatures, see e.g. [21, 43, 45]. However, it seems that there is no general result on the speed of convergence in the case of a (generic) optimal control problem as defined in Section 3.2 below.

In Section 3.3, we verify that the above intuition is correct. Unlike [57], we do not simply rely on the stability of viscosity solutions. Nor do we rely on the weak convergence of the underlying process. The reason is that weak convergence does not give access to the convergence speed in optimal control problems. Instead, we directly study the regularity of the solution to (3.2). Thanks to its vanishing terminal condition (otherwise it should be assumed smooth enough), we show that $\partial^2_{xx}\bar{V}_T$ is uniformly γ-Hölder in space, for some $\gamma \in (0, 1]$, whenever the coefficients of (3.2) are uniformly Lipschitz in space and under a uniform ellipticity condition. By a second order Taylor expansion, this allows us to pass from (3.2) to (3.1) up to an error term of order $\varepsilon^{\gamma/2}$, and therefore provides the required convergence rate. In general, this rate cannot be improved. As a by-product, we obtain an easy way to construct an $\varepsilon^{\gamma/2}$-optimal control for the original pure-jump control problem. We then study the limit $\varepsilon^{-\gamma/2}(V - \bar{V}_T)$ as $\varepsilon \to 0$. Under mild assumptions, we show that it solves a (possibly non-linear) PDE. This provides a first error correction term. To achieve higher orders of convergence, this approach can be generalised to a system of non-linear PDEs, upon its existence.

As an example of application, we consider in Section 3.4 a simplified repeated online auction bidding problem, where a buyer seeks to maximise its profit when facing both competition and a seller who adapts the price to incoming bids. Our numerical experiments show that our approximation permits a considerable gain in computation time.

For ease of exposition, we shall restrict to situations where the controlled process is of dimension one. This fact will be used explicitly only to derive our regularity results in Section 3.3.2. Similar results can be obtained in higher dimensions, by using standard regularity results for parabolic partial differential equations, see e.g. [79, 84].

Notations

Let us give some standard notations which will be used throughout. Given an open set $\mathcal{U} \subset \mathbb{R}^n$, $n \in \mathbb{N}$, we denote $\mathcal{C}^0(\mathcal{U})$, $\mathcal{C}^1(\mathcal{U})$, $\mathcal{C}^2(\mathcal{U})$ the spaces of real-valued functions which are respectively: continuous, once continuously differentiable, and twice continuously differentiable. Given $T > 0$, $\mathcal{C}^{(1,2)}([0,T)\times\mathcal{U})$ denotes the space of functions which are once continuously differentiable in time and twice in space. In addition, $\mathcal{C}_b^{(1,2)}([0,T)\times\mathcal{U})$ denotes the elements of $\mathcal{C}^{(1,2)}([0,T)\times\mathcal{U})$ with bounded derivatives of order up to one is time and up to two in space. When clear from context, we will omit $\mathcal{U}$ and $[0,T)$ for brevity.

3.2 The Pure-Jump Optimal Control Problem

In this section, we begin by defining our pure-jump control problem and state the well-known link with its associated Hamilton-Jacobi-Bellman equation. The properties stated below are elementary but will be useful for the derivation of our main approximation result of Section 3.3.

3.2.1 Definition

Let $\Omega = \mathbb{D}_T$ denote the space of one dimensional càdlàg[1] functions on $\mathbb{R}_+$ and $\mathcal{M}_+(\mathbb{R}\times\mathbb{R}_+)$ denote the collection of positive finite measures on $\mathbb{R}\times\mathbb{R}_+$. Consider a measure-valued map $N : \mathbb{D}_T \mapsto \mathcal{M}_+(\mathbb{R}\times\mathbb{R}_+)$ and a probability measure $\mathbb{P}$ on $\mathbb{D}_T$ such that N is a continuous real-valued $\mathbb{R}$-marked point process with compensator $\eta\nu(\mathrm{d}e)\mathrm{d}t$, in which $\eta > 0$ and ν is a probability measure on $\mathbb{R}$. See e.g. [37]. For ease of notations, we set $N_t := N(\mathbb{R},[0,t])$ for $t \geq 0$.

Let $\mathbb{F}^t = (\mathcal{F}_s^t)_{s\geq t}$ be the $\mathbb{P}$-augmentation of the raw filtration generated by the random measure N restricted to $[t,\infty)$, that is, for instance, by the process $\int_t^\cdot \int \exp(e)N(\mathrm{d}e,\mathrm{d}s)$. Given a compact subset $\mathbb{A}$ of $\mathbb{R}$, we let $\mathcal{A}^t$ be the collection of $\mathbb{F}^t$-predictable processes with values in $\mathbb{A}$. For ease of notation, we also define $\mathcal{A} := \cup_{t\geq 0}\mathcal{A}^t$. Throughout this chapter, unless otherwise stated we will work on the filtered probability space $(\Omega,\mathcal{F},\mathbb{F},\mathbb{P})$, where $\mathcal{F} = \mathcal{F}_T^0$ for $T > 0$ given and $\mathbb{F} = \mathbb{F}^0$.

We now consider a bounded measurable map $(x,a,e) \in \mathbb{R}\times\mathbb{A}\times\mathbb{R} \mapsto b(x,a,e)$. Given $(t,x) \in \mathbb{R}_+\times\mathbb{R}$ and $\alpha \in \mathcal{A}$, we define the càdlàg process

[1] Also known as right continuous with left limits.

$X^{t,x,\alpha}$ as the solution of

$$X^{t,x,\alpha} = x + \int_t^{\cdot} \int b(X_{s-}^{t,x,\alpha}, \alpha_s, e) N(\mathrm{d}e, \mathrm{d}s). \tag{3.4}$$

Given a bounded measurable map $(x, a) \in \mathbb{R} \times \mathbb{A} \mapsto r(x, a) \in \mathbb{R}$, we consider the expected gain function

$$(t, x, \alpha) \in [0, T] \times \mathbb{R} \times \mathcal{A} \mapsto J_T(t, x; \alpha) := \mathbb{E}\left[\int_t^T r(X_{s-}^{t,x,\alpha}, \alpha_s) \mathrm{d}N_s\right], \tag{3.5}$$

together with the value function

$$V_T(t, x) := \sup_{\alpha \in \mathcal{A}^t} J_T(t, x; \alpha), \ (t, x) \in [0, T] \times \mathbb{R}. \tag{3.6}$$

Throughout this chapter, we make the following standard assumption, which will in particular ensure that V_T is the unique (bounded) viscosity solution of the associated Hamilton-Jacobi-Bellman (HJB) equation, see Proposition 3.2.1 below.

Assumption 3.1.
For each $e \in \mathbb{R}$, $(x, a) \in \mathbb{R} \times \mathbb{A} \mapsto (b(x, a, e), r(x, a))$ is continuous. Furthermore, (b, r) is bounded.

***Remark* 3.2.1.** *Note that boundedness of the coefficients b and r is not essential in the following arguments. One could assume only linear growth in space, uniformly in the control. We make the above (strong) assumptions to avoid unnecessary complexities.*

3.2.2 Dynamic programming equation and optimal Markov control

Let us now state the well-known characterization of V_T in terms of the theory of viscosity solutions.

As usual, we say that a locally bounded lower-semicontinuous (resp. upper-semicontinuous) map $U : [0, T] \times \mathbb{R} \mapsto \mathbb{R}$ is a viscosity supersolution (resp. subsolution) of

$$\partial_t \varphi + \sup_{a \in \mathbb{A}} \left(\int \varphi(\cdot, \cdot + b(\cdot, a, e)) \nu(\mathrm{d}e) - \varphi + r(\cdot, a) \right) \eta = 0, \quad \text{on } [0, T) \times \mathbb{R}, \tag{3.7}$$

if for all $(t, x) \in [0, T) \times \mathbb{R}$ and all $\mathcal{C}^1$ functions $\varphi : [0, T] \times \mathbb{R} \mapsto \mathbb{R}$ such that (t, x) attains a minimum (resp. maximum) of $U - \varphi$ on $[0, T) \times \mathbb{R}$ we have

$$\kappa \left\{ \partial_t \varphi(t, x) + \eta \sup_{a \in \mathbb{A}} \left(\int U(t, x + b(x, a, e)) \nu(\mathrm{d}e) - U(t, x) + r(x, a) \right) \right\} \leq 0$$

with $\kappa = 1$ (resp. $\kappa = -1$).

Proposition 3.2.1.
V_T is a continuous and bounded viscosity solution of (3.7) such that

$$\lim_{t' \uparrow T, x' \to x} V_T(t', x') = 0, \ x \in \mathbb{R}. \tag{3.8}$$

Moreover, comparison holds for (3.7) in the class of bounded functions.

Proof. The argument being standard, we only sketch it.
First note that the continuity at T follows immediately from the fact that r is bounded, namely $|V_T(t,\cdot)| \leq \eta(T-t)\|r\|_\infty$ for $t \leq T$. Fix $h \in (0, T-t]$, $t \leq T$ and $x \in \mathbb{R}$. Let τ_1^t be the first jump of N after time t. Denote by V_{T*} and V_T^* the lower- and upper-semicontinuous envelopes of V_T, i.e.

$$V_{T*}(t', x') := \liminf_{(s,y)\to(t',x')} V_T(s,y)\,, \ V_T^*(t', x') := \limsup_{(s,y)\to(t',x')} V_T(s,y)\,.$$

It follows from the same arguments as in [35] that V_T satisfies the (weak) dynamic programming principle

$$\begin{aligned} &\sup_{\alpha \in \mathcal{A}^t} \mathbb{E}\left[V_{T*}(\tau_1^t \wedge h, X^{t,x,\alpha}_{\tau_1^t \wedge h}) + r(X^{t,x,\alpha}_{\tau_1^t -}, \alpha_{\tau_1^t}) 1_{\{\tau_1^t \leq h\}} \right] \\ &\leq V_T(t,x) \\ &\leq \sup_{\alpha \in \mathcal{A}^t} \mathbb{E}\left[V_T^*(\tau_1^t \wedge h, X^{t,x,\alpha}_{\tau_1^t \wedge h}) + r(X^{t,x,\alpha}_{\tau_1^t -}, \alpha_{\tau_1^t}) 1_{\{\tau_1^t \leq h\}} \right]. \end{aligned} \tag{3.9}$$

Following [35] again and using [33, Lemma 22], this implies that V_{T*} and V_T^* are, respectively, a super- and a subsolution in the viscosity sense of (3.7). Since b is bounded, the map $(t,x) \mapsto (1+x^2)e^{-Ct}$ is also a viscosity supersolution of the above with $r \equiv 0$, as soon as $C > 0$ is large enough. Standard arguments then imply that comparison holds for the above Hamilton-Jacobi-Bellman equation in the class of bounded functions (or even with linear growth), and therefore that $V_{T*} = V_T^*$, meaning that V_T is continuous. □

We next prove the existence of an optimal Markov control. In the following, we denote by $\mathscr{A}_T$ the collection of $\mathbb{A}$-valued Borel maps on $[0,T) \times \mathbb{R}$.

Proposition 3.2.2.
For all $(t,x) \in \mathbb{R}_+ \times \mathbb{R}$, there exists $\hat{\alpha}[t,x] \in \mathcal{A}^t$ such that $V_T(t,x) = J_T(t,x;\hat{\alpha}[t,x])$. It takes the form

$$\hat{\alpha}[t,x] = \sum_{i \geq 0} 1_{(\tau_i^t, \tau_{i+1}^t]} \hat{\mathrm{a}}(\cdot, X^{t,x,\hat{\alpha}[t,x]}_{\tau_i^t})$$

in which τ_i^t is the i-th jump of N after time t, for $i \geq 1$, with $\tau_0^t := t$, and $\hat{a} \in \mathscr{A}_T$ satisfies

$$\hat{a}(t', x') \in \underset{a \in \mathbb{A}}{\operatorname{argmax}} \left\{ \int V_T(t', x' + b(x', a, e)) \nu(\mathrm{d}e) + r(x', a) \right\},$$

for every $(t', x') \in [0, t) \times \mathbb{R}$.

Proof. Since V_T, b and r are continuous, by Proposition 3.2.1 and Assumption 3.1, and since $\mathbb{A}$ is compact, we can find a Borel measurable map $(t, x) \mapsto \hat{a}(t, x)$ such that $\hat{a}(t, x)$ belongs to $\operatorname{argmax}\{\int V_T(t, x + b(x, \cdot, e))\nu(de) + r(x, \cdot)\}$ for all $(t, x) \in [0, T) \times \mathbb{R}$, see e.g. [24, Proposition 7.33, p.153]. Let us fix $(t_0, x_0) \in [0, T] \times \mathbb{R}$. By the dynamic programming principle in (3.9), the continuity of V_T, and the definition of $\hat{a}$ above,

$$\begin{aligned}
V_T(t_0, x_0) &= \sup_{\alpha \in \mathcal{A}_0^t} \mathbb{E}\left[V_T\left(\tau_1^{t_0} \wedge T, X^{t_0, x_0, \alpha}_{\tau_1^{t_0} \wedge T}\right) + r\left(X^{t_0, x_0, \alpha}_{\tau_1^{t_0}-}, \alpha_{\tau_1^{t_0}}\right) 1_{\left\{\tau_1^{t_0} \leq T\right\}} \right] \\
&= \sup_{\alpha \in \mathcal{A}_0^t} \mathbb{E}\left[\left(\int V_T\left(\tau_1^{t_0} \wedge T, x_0 + b(x_0, \alpha_{\tau_1^{t_0} \wedge T}, e)\right) \nu(\mathrm{d}e) \right.\right. \\
&\qquad \left.\left. + r\left(x_0, \alpha_{\tau_1^{t_0}}\right)\right) 1_{\left\{\tau_1^{t_0} \leq T\right\}} \right] \\
&= \mathbb{E}\left[\left(\int V_T\left(\tau_1^{t_0} \wedge T, x_0 + b(x_0, \hat{\alpha}_{\tau_1^{t_0} \wedge T}, e)\right) \nu(\mathrm{d}e) \right.\right. \\
&\qquad \left.\left. + r\left(x_0, \hat{\alpha}_{\tau_1^{t_0}}\right)\right) 1_{\left\{\tau_1^{t_0} \leq T\right\}} \right] \\
&= \mathbb{E}\left[V_T(\tau_1^{t_0} \wedge T, X^{t_0, x_0, \hat{\alpha}}_{\tau_1^{t_0} \wedge T}) + r(X^{t_0, x_0, \hat{\alpha}}_{\tau_1^{t_0}-}, \hat{\alpha}_{\tau_1^{t_0}}) 1_{\{\tau_1^{t_0} \leq T\}} \right]
\end{aligned}$$

in which $\hat{\alpha} := \hat{a}(\cdot, x_0) 1_{(t_0, \tau_1^{t_0}]}$. For ease of notations, we now set $\vartheta_1 := \tau_1^{t_0} \wedge T$ and $X_1 := X^{t_0, x_0, \hat{\alpha}}_{\tau_1^{t_0} \wedge T}$. By the same reasoning as above, we have, for a fixed $\omega \in \Omega$,

$$\begin{aligned}
V(\vartheta_1(\omega), X_1(\omega)) = \mathbb{E}\left[V_T\left(\tau_1^{\vartheta_1(\omega)} \wedge T, X^{\vartheta_1(\omega), X_1(\omega), \hat{\alpha}(\omega)}_{\tau_1^{\vartheta_1(\omega)} \wedge T}\right) \right. \\
\left. + r\left(X^{\vartheta_1(\omega), X_1(\omega), \hat{\alpha}(\omega)}_{\tau_1^{\vartheta_1(\omega)}-}, \hat{\alpha}_{\tau_1^{\vartheta_1(\omega)}}(\omega)\right) 1_{\left\{\tau_1^{\vartheta_1(\omega)} \leq T\right\}} \right]
\end{aligned}$$

in which

$$\hat{\alpha}(\omega) := \hat{a}(\cdot, x_0) 1_{\left(t_0, \tau_1^{t_0}(\omega)\right]} + \hat{a}(\cdot, X_1(\omega)) 1_{\left(\tau_1^{t_0}(\omega), \tau_1^{\vartheta_1(\omega)}\right]}.$$

The right-hand side of the above coincides $\mathbb{P}$-a.e. with

$$\begin{aligned}&\mathbb{E}\left[V_T\left(\tau_1^{\vartheta_1}\wedge T, X^{\vartheta_1,X_1,\hat{\alpha}}_{\tau_1^{\vartheta_1}\wedge T}\right)+r\left(X^{\vartheta_1,X_1,\hat{\alpha}}_{\tau_1^{\vartheta_1}-},\hat{\alpha}_{\tau_1^{\vartheta_1}}\right)1_{\left\{\tau_1^{\vartheta_1}\le T\right\}}\middle|\mathcal{F}_{\vartheta_1}\right]\\&=\mathbb{E}\left[V_T\left(\tau_2^{t_0}\wedge T, X^{t_0,x_0,\hat{\alpha}}_{\tau_2^{t_0}\wedge T}\right)+r\left(X^{t_0,x_0,\hat{\alpha}}_{\tau_2^{t_0}-},\hat{\alpha}_{\tau_2^{t_0}}\right)1_{\left\{\tau_2^{t_0}\le T\right\}}\middle|\mathcal{F}_{\tau_1^{t_0}\wedge T}\right].\end{aligned}$$

Let us complete the definition of $\hat{\alpha}$ by now letting it be defined by

$$\hat{\alpha}=\sum_{i\ge 0}1_{\left(\tau_i^{t_0},\tau_{i+1}^{t_0}\right]}\hat{\mathrm{a}}\left(\cdot,X^{t_0,x_0,\hat{\alpha}}_{\tau_i^{t_0}}\right).$$

By iterating the above procedure, we have

$$V_T(t_0,x_0)=\mathbb{E}\left[V_T\left(\tau_n^{t_0}\wedge T, X^{t_0,x_0,\hat{\alpha}}_{\tau_n^{t_0}\wedge T}\right)+\int_{t_0}^{\tau_n^{t_0}\wedge T} r(X_{s-}^{t_0,x_0,\hat{\alpha}},\hat{\alpha}_s)\mathrm{d}N_s\right],\ n\ge 1.$$

Since $\tau_n^{t_0}\to\infty$ $\mathbb{P}$-a.s. as $n\to\infty$, it now follows from the dominated convergence theorem and (3.8) that

$$V_T(t_0,x_0)=\mathbb{E}\left[\int_{t_0}^{T} r(X_{s-}^{t_0,x_0,\hat{\alpha}},\hat{\alpha}_s)\mathrm{d}N_s\right].$$

□

3.3 Diffusive Approximation

As already mentioned, the characterization of Propositions 3.2.1 and 3.2.2 allows one to estimate numerically the value function and the associated optimal control. However, the integro-differential equation (3.7) is non-local and the computational cost of its numerical resolution increases as η grows. On the other hand, we can expect that our pure-jump problem admits a diffusive limit as $\eta\to\infty$ which is, by its local nature, much easier to solve numerically, and can serve as a good proxy of the original problem as soon as η is large enough.

In this section, we begin by defining the diffusion control problem that is the candidate for the diffusive limit of our pure-jump problem. We then study the regularity of the corresponding value function, from which we will be able to derive our main approximation result, see Theorem 3.3.1 below, and construct approximate optimal controls, see Proposition 3.3.2. Finally, we identify a first-order correction term in Section 3.3.5, which is extended to higher orders in Section 3.3.6.

3.3.1 The candidate diffusive limit

Given $\varepsilon \in (0,1)$, we now take as η the intensity

$$\eta_\varepsilon := \varepsilon^{-1}$$

so that it is large for $\varepsilon > 0$ small. To ensure the existence of a diffusive limit, we need to assume that the jump coefficient b introduced in Section 3.2 is of the form

$$b_\varepsilon = \varepsilon b_1 + \sqrt{\varepsilon} b_2$$

for two bounded measurable maps $b_1, b_2 : \mathbb{R} \times \mathbb{A} \times \mathbb{R} \mapsto \mathbb{R}$, each satisfying Assumption 3.1 (with b_i in place of b, $i = 1, 2$), and with b_2 satisfying the additional Assumption 3.2.

Assumption 3.2.
The function b_2 satisfies:

$$\int b_2(x,a,e)\nu(\mathrm{d}e) = 0 \quad \text{for all } (x,a) \in \mathbb{R} \times \mathbb{A}, \tag{3.10}$$

and there is $\varsigma > 0$ such that

$$\inf_{(x,a)\in\mathbb{R}\times\mathbb{A}} \int |b_2(x,a,e)|^2 \nu(\mathrm{d}e) \geq \varsigma > 0. \tag{3.11}$$

In the above, the coefficient b_1 should be interpreted as a drift term while b_2 is a volatility. The respective scaling in ε and $\sqrt{\varepsilon}$ together with Assumption 3.2 are required to ensure that our pure-jump problem actually admits a diffusive limit of the form (3.13) below. Problems in which this scaling of coefficient is appropriate, involve many jumps of small relative size, with a variance of the same order as their drift over time.

Likewise, we consider the value function

$$V_T^\varepsilon(t,x) := \sup_{\alpha\in\mathcal{A}^t} J_T^\varepsilon(t,x;\alpha) \text{ with } J_T^\varepsilon(t,x;\alpha) := \frac{1}{\eta_\varepsilon}\mathbb{E}\left[\int_t^T r(X_{s-}^{t,x,\alpha},\alpha_s)\mathrm{d}N_s\right] \tag{3.12}$$

for any $(t,x) \in [0,T] \times \mathbb{R}$. Note that the scaling by $1/\eta_\varepsilon$ means that (up to a constant factor $T-t$) we consider the gain by average unit of actions on the system. Indeed $\mathbb{E}[N_T - N_t] = \eta_\varepsilon(T-t)$ and the control applies only at jump times of N. Note that we omit the dependence of N on ε, for ease of notation.

We shall see that V_T^ε, together with the associated optimal policy, can be approximated by considering its diffusive limit as $\varepsilon \to 0$. The coefficients of the associated Brownian diffusion Stochastic Differential Equation (SDE) are given by:

$$\mu(x,a) := \int b_1(x,a,e)\nu(\mathrm{d}e),\ \sigma(x,a) := \left(\int |b_2(x,a,e)|^2\,\nu(\mathrm{d}e)\right)^{\frac{1}{2}},$$

for $(x,a) \in \mathbb{R} \times \mathbb{A}$.

From now on, we assume that they satisfy the following.

Assumption 3.3.
The maps $x \in \mathbb{R} \mapsto \mu(x,a)$, $x \in \mathbb{R} \mapsto \sigma(x,a)$ and $x \in \mathbb{R} \mapsto r(x,a)$ are Lipschitz, uniformly in $a \in \mathbb{A}$, with respective Lipschitz constants $\|\mu\|_{\mathrm{Lip}}$, $\|\sigma\|_{\mathrm{Lip}}$ and $\|r\|_{\mathrm{Lip}}$.

More precisely, let $\bar{\mathbb{P}}$ be a probability measure on $\mathbb{D}_T$ and let W be a stochastic process such that W is a $\bar{\mathbb{P}}$-Brownian motion, let $\bar{\mathbb{F}}^t = (\bar{\mathcal{F}}^t_s)_{s\geq 0}$ be the $\bar{\mathbb{P}}$-augmentation of the filtration generated by $(W_{\cdot\vee t} - W_t)$, and let $\bar{\mathcal{A}}^t$ be the collection of $\bar{\mathbb{F}}^t$-predictable processes. Given $\bar{\alpha} \in \bar{\mathcal{A}}^t$, we can then define $\bar{X}^{t,x,\bar{\alpha}}$ as the unique strong solution of

$$\bar{X}^{t,x,\bar{\alpha}} = x + \int_t^\cdot \mu(\bar{X}^{t,x,\bar{\alpha}}_s, \bar{\alpha}_s)\mathrm{d}s + \int_t^\cdot \sigma(\bar{X}^{t,x,\bar{\alpha}}_s, \bar{\alpha}_s)\mathrm{d}W_s. \tag{3.13}$$

The candidate diffusive limit problem is then defined as

$$\bar{V}_T(t,x) := \sup_{\bar{\alpha}\in\bar{\mathcal{A}}^t} \bar{\mathbb{E}}\left[\int_t^T r(\bar{X}^{t,x,\bar{\alpha}}_s, \bar{\alpha}_s)\mathrm{d}s\right],\ (t,x) \in [0,T]\times\mathbb{R}$$

where $\bar{\mathbb{E}}$ is the expectation operator under $\bar{\mathbb{P}}$.

3.3.2 Regularity properties

We first prove that $\bar{V}_T$ is a smooth solution of its associated Hamilton-Jacobi-Bellman equation. Most importantly, its second-order space derivative is γ-Hölder continuous, for some $\gamma \in (0,1]$. This will allow us, in Section 3.3.3 below, to prove that it actually coincides with the diffusive limit of V_T^ε as ε vanishes. The precise value of the Hölder exponent γ will be further discussed in Remark 3.3.1 below.

Proposition 3.3.1.

The function $\bar{V}_T$ belongs to $\mathcal{C}_b^{(1,2)}([0,T)\times\mathbb{R})\cap\mathcal{C}^0([0,T]\times\mathbb{R})$ and is the unique bounded solution of

$$\partial_t \bar{V}_T + \sup_{\bar{a}\in\mathbb{A}}\left(\mu(\cdot,\bar{a})\partial_x \bar{V}_T + \frac{1}{2}\sigma^2(\cdot,\bar{a})\partial^2_{xx}\bar{V}_T + r(\cdot,\bar{a})\right) = 0, \text{ on } [0,T)\times\mathbb{R}, \tag{3.14}$$

$$\bar{V}_T(T,\cdot) = 0, \text{ on } \mathbb{R}. \tag{3.15}$$

Moreover, there exists $\gamma\in(0,1]$, such that $\partial^2_{xx}\bar{V}_T$ is (uniformly) γ-Hölder continuous in space on $[0,T)\times\mathbb{R}$.

Proof.

1. We first show that $\bar{V}_T\in\mathcal{C}_b^{(1,2)}([0,T)\times\mathbb{R})\cap\mathcal{C}^0([0,T]\times\mathbb{R})$. Note that the continuity at T follows again from the fact that r is bounded: $|\bar{V}_T(t,\cdot)|\le(T-t)\|r\|_\infty$ for $t\le T$. Let us set

$$F(x,p,q) := \sup_{\bar{a}\in\mathbb{A}}\left(\mu(x,\bar{a})p + \frac{1}{2}\sigma^2(x,\bar{a})q + r(x,\bar{a})\right),\ (x,p,q)\in\mathbb{R}^3,$$

and observe that, by Assumptions 3.2 and 3.3,

$$\frac{1}{2}\varsigma\,|q-q'| \le |F(x,p,q)-F(x,p,q')| \le \frac{1}{2}\|\sigma\|^2_\infty\,|q-q'|, \tag{3.16}$$

$$vF(x,0,0) \le \|r\|_\infty\,(1+|v|^2), \tag{3.17}$$

and

$$\begin{aligned}|F(x,p,q)-F(x',p',q')| \le &(|p|\left\|\mu\right\|_{\text{Lip}} + |q|\,\|\sigma\|_\infty\,\|\sigma\|_{\text{Lip}} + \|r\|_{\text{Lip}})\,|x-x'|\\ &+ \left\|\mu\right\|_\infty|p-p'| + \frac{1}{2}\|\sigma\|^2_\infty\,|q-q'|\end{aligned} \tag{3.18}$$

for all $(x,x',p,p',q,q',v)\in\mathbb{R}^7$.

Let us assume for the moment that $q\mapsto F(x,p,q)$ is differentiable for all $(x,p)\in\mathbb{R}^2$. For $n\ge1$, existence of a $\mathcal{C}^{(1,2)}([0,T)\times\mathbb{R})$ solution $\bar{V}_{T,n}$ to (3.14) on $[0,T)\times(-n,n)$ with boundary condition $\bar{V}_{T,n}=0$ on $([0,T)\times\{-n,n\})\cup(\{T\}\times[-n,n])$ follows from [84, Thm. 14.24], (3.16), (3.17) and (3.18). It turns out that, using the notations of [84, Thm. 14.24], $\bar{V}_{T,n}$ is even in $H_{2+\theta_B}(B)$ for some $\theta_B\in(0,1)$, on each compact subset B of $[0,T)\times(-n,n)$. These $H_{2+\theta_B}$-norms depend only on the upper and

lower bounds on the derivative of $q \mapsto F(\cdot, q)$ and not on the fact that this map is differentiable. If it is not, one can thus first regularize F with respect to its last argument, by using a sequence of smooth kernels, and then pass to the limit. The corresponding sequence will be uniformly bounded in $H_{2+\theta_B}(B)$ on each compact subset B of $[0,T) \times (-n,n)$, so that the limit will keep these bounds. By stability, the limit solves the required equation with the appropriate boundary conditions. See also the discussion in the paragraph preceding [84, Theorem 14.24].

2. We now provide uniform estimates on the gradients. Note that, by the Feynman-Kac formula and a comparison argument,

$$\bar{V}_{T,n}(t,x) = \sup_{\bar{\alpha} \in \bar{\mathcal{A}}^t} \bar{\mathbb{E}}\left[\int_t^{T \wedge \tau_n^{t,x,\bar{\alpha}}} r(\bar{X}_s^{t,x,\bar{\alpha}}, \bar{\alpha}_s) \mathrm{d}s \right] \tag{3.19}$$

in which

$$\tau_n^{t,x,\bar{\alpha}} := \inf\{ s \geq t \,:\, \bar{X}_s^{t,x,\bar{\alpha}} \notin (-n,n) \}\,.$$

It follows that, for $h \in (0, T-t]$,

$$\bar{V}_{T,n}(t+h,x) = \sup_{\bar{\alpha} \in \bar{\mathcal{A}}^t} \bar{\mathbb{E}}\left[\int_t^{(T-h) \wedge \tau_n^{t,x,\bar{\alpha}}} r(\bar{X}_s^{t,x,\bar{\alpha}}, \bar{\alpha}_s) \mathrm{d}s \right]$$

which readily implies that

$$\left| \bar{V}_{T,n}(t+h,x) - \bar{V}_{T,n}(t,x) \right| \leq h \left\| r \right\|_\infty ,$$

and therefore

$$\frac{1}{T} \| \bar{V}_{T,n} \|_\infty \vee \| \partial_t \bar{V}_{T,n} \|_\infty \leq \| r \|_\infty \,. \tag{3.20}$$

Similarly, for $h \in (-1,1)$ such that $x + h \in [-n,n]$,

$$\begin{aligned} &\left| \bar{V}_{T,n}(t,x+h) - \bar{V}_{T,n}(t,x) \right| \\ &\quad \leq \sup_{\bar{\alpha} \in \bar{\mathcal{A}}^t} \bar{\mathbb{E}}\left[\| r \|_{\mathrm{Lip}} \int_t^T \left| \bar{X}_s^{t,x+h,\bar{\alpha}} - \bar{X}_s^{t,x,\bar{\alpha}} \right| \mathrm{d}s + \| r \|_\infty \left| \tau_n^{t,x+h,\bar{\alpha}} - \tau_n^{t,x,\bar{\alpha}} \right| \right]. \end{aligned}$$

The first term is handled by using the uniform Lipschitz continuity in space of (μ, σ):

$$\bar{\mathbb{E}}\left[\int_t^T \left| \bar{X}_s^{t,x+h,\bar{\alpha}} - \bar{X}_s^{t,x,\bar{\alpha}} \right| \mathrm{d}s \right] \leq C_1 \left| h \right| \tag{3.21}$$

in which $C_1 > 0$ does not depend on n. As for the second term, Assumption 3.3, (3.10) and our boundedness assumptions on (b_1, b_2), and

therefore on (μ, σ), allow us to apply [34, Thm. 2.3][2] with (in their notation) $\pi = 0$, $r = 1$ and for P of the form $\varphi(\bar{X}^{t,x+h,\bar{\alpha}})$ or $\varphi(\bar{X}^{t,x,\bar{\alpha}})$ for a smooth bounded function φ, with bounded first and second derivatives, such that $\varphi(y) = y + n$ for $y \in [-n, -n+1]$ and $\varphi(y) = n - y$ for $y \in [n-1, n]$. It implies that

$$\bar{\mathbb{E}}\left[\left|\tau_n^{t,x+h,\bar{\alpha}} - \tau_n^{t,x,\bar{\alpha}}\right|\right] \leq C_2 \bar{\mathbb{E}}\left[\left|\bar{X}^{t,x+h,\bar{\alpha}}_{\tau_n^{t,x+h,\bar{\alpha}} \wedge \tau_n^{t,x,\bar{\alpha}}} - \bar{X}^{t,x,\bar{\alpha}}_{\tau_n^{t,x+h,\bar{\alpha}} \wedge \tau_n^{t,x,\bar{\alpha}}}\right|\right] \leq C_2' |h|$$

for some positive constants C_2 and C_2' independent of n. Combined with (3.21), this leads to

$$\|\partial_x \bar{V}_{T,n}\|_\infty \leq \|r\|_{\mathrm{Lip}} C_1 + \|r\|_\infty C_2' . \tag{3.22}$$

The fact that $\bar{V}_{T,n}$ solves (3.14) combined with (3.16), (3.20), and (3.22) then proves that

$$\|\partial^2_{xx} \bar{V}_{T,n}\|_\infty \leq C_3 \tag{3.23}$$

for some $C_3 > 0$ that does not depend on n.

3. Let us now prove the uniform Hölder continuity of the gradients and second derivatives. As in 1. above, let us first assume that F is C^1. Given a neighbourhood $\mathcal{U} \subset [0,T] \times [-n, n]$ of a point (t, x), we derive as in [11, § 3.1] that there exists $C > 0$ and $\gamma \in (0, 1]$, depending only on the ellipticity constant ς and the Lipschitz constants of F with respect to its second and third arguments, such that

$$|\partial_t \bar{V}_{T,n}(t', x') - \partial_t \bar{V}_{T,n}(t, x)| \leq C\left(|t' - t|^{\frac{\gamma}{2}} + |x' - x|^\gamma\right) \sup_{\mathcal{U}} |\partial_t \bar{V}_{T,n}|,$$

for $(t', x') \in \mathcal{U}$. If F is not $\mathcal{C}^1$, one can first regularize it by using a sequence of kernels and then pass to the limit to obtain that the above still holds for the original F. In view of (3.20), this implies that

$$|\partial_t \bar{V}_{T,n}(t', x') - \partial_t \bar{V}_{T,n}(t, x)| \leq C\left(|t' - t|^{\frac{\gamma}{2}} + |x' - x|^\gamma\right) \|r\|_\infty , \tag{3.24}$$

for $(t', x') \in [0,T] \times \mathbb{R}$. Up to changing $\gamma \in (0, 1]$, one can prove similarly that

$$|\partial_x \bar{V}_{T,n}(t', x') - \partial_x \bar{V}_{T,n}(t, x)| \leq C\left(|t' - t|^{\frac{\gamma}{2}} + |x' - x|^\gamma\right), \tag{3.25}$$

for $(t', x') \in [0,T] \times \mathbb{R}$, for some $C > 0$ that does not depend on n. We now set $\Delta_h \bar{V}_{T,n} := h^{-\gamma}\left(\bar{V}_{T,n}(\cdot, \cdot + h) - \bar{V}_{T,n}\right)$, $h \in \mathbb{R}$. Again, up to

[2] Note that their Assumption (L) is not required since we are considering a finite time interval $[0, T]$, this can be easily seen from the proof of this theorem.

mollifying F with a smooth bounded kernel with derivatives bounded by 1, we can assume that F is $\mathcal{C}^1$. Then, for $t < T$ and $x \in (-n+h, n-h)$,

$$\begin{aligned}
&h^{-\gamma}\left\{F(x+h, \partial_x \bar{V}_{T,n}(t,x+h), \partial^2_{xx}\bar{V}_{T,n}(t,x+h))\right.\\
&\qquad \left. - F(x, \partial_x \bar{V}_{T,n}(t,x), \partial^2_{xx}\bar{V}_{T,n}(t,x))\right\}\\
&= h^{-\gamma}\Big\{\partial_x F(x^1_h, p^1_h, q^1_h)h + \partial_p F(x^2_h, p^2_h, q^2_h)[\partial_x \bar{V}_{T,n}(t,x+h) - \partial_x \bar{V}_{T,n}(t,x)]\\
&\quad + \partial_q F(x^3_h, p^3_h, q^3_h)[\partial^2_{xx}\bar{V}_{T,n}(t,x+h) - \partial^2_{xx}\bar{V}_{T,n}(t,x)]\Big\}
\end{aligned}$$

for some $x^i_h \in [x, x+h]$, $p^i_h \in [\partial_x \bar{V}_{T,n}(t,x+h) \wedge \partial_x \bar{V}_{T,n}(t,x), \partial_x \bar{V}_{T,n}(t,x+h) \vee \partial_x \bar{V}_{T,n}(t,x)]$ and $q^i_h \in [\partial^2_{xx}\bar{V}_{T,n}(t,x+h) \wedge \partial^2_{xx}\bar{V}_{T,n}(t,x), \partial^2_{xx}\bar{V}_{T,n}(t,x+h) \vee \partial^2_{xx}\bar{V}_{T,n}(t,x)]$, for $i = 1, 2, 3$. It follows that $\Delta_h \bar{V}_{T,n}$ satisfies a linearized equation of the form

$$0 = \partial_t \Delta_h \bar{V}_{T,n} + A_h \partial_x(\Delta_h \bar{V}_{T,n}) + B_h \partial^2_{xx}(\Delta_h \bar{V}_{T,n}) + C_h h^{1-\gamma}$$

at every point $(t,x) \in [0,T) \times \mathbb{R}$ such that $x + h \in (-n, n)$, in which, by Assumption 3.3, (3.10), and the estimates in 2/ above, $(A_h, C_h)_{h>0}$ is uniformly bounded and $\inf_{h>0} \inf_{[0,T]\times\mathbb{R}} B_h \geq \varsigma/2 > 0$. Hence,

$$\left|\partial^2_{xx}\Delta_h \bar{V}_{T,n}\right| \leq 2\varsigma^{-1}\left(\left|\partial_t \Delta_h \bar{V}_{T,n}\right| + |A_h|\left|\partial_x \Delta_h \bar{V}_{T,n}\right| + |C_h|\, h^{1-\gamma}\right)$$

We conclude from (3.24)-(3.25) that

$$\left|\partial^2_{xx}\bar{V}_{T,n}(t,x') - \partial^2_{xx}\bar{V}_{T,n}(t,x)\right| \leq C\,|x'-x|^{\gamma}\,,\ x, x' \in (-n,n),\ t < T, \tag{3.26}$$

for some $C > 0$ independent on n. If we now set $\Delta_h \bar{V}_{T,n} = h^{-\frac{\gamma}{2}}(\bar{V}_{T,n}(\cdot + h, \cdot) - \bar{V}_{T,n})$, then the same type of arguments leads to

$$\left|\partial^2_{xx}\bar{V}_{T,n}(t',x) - \partial^2_{xx}\bar{V}_{T,n}(t,x)\right| \leq C\,|t'-t|^{\frac{\gamma}{2}}\,,\ x \in (-n,n),\ t, t' < T, \tag{3.27}$$

for some $C > 0$ independent on n.

4. It follows from steps 2. and 3. that $(\bar{V}_{T,n})_{n\geq 1}$ is uniformly bounded in $H_{2+\gamma}([0,T)\times\mathbb{R})$, as defined in [84, § IV.1]. By the Arzelà-Ascoli theorem, it admits a subsequence that converges in $H_{2+\gamma}(B)$, for any compact set $B \subset [0,T)\times\mathbb{R}$, to a limit $\bar{V}_{T,\infty}$. This limit shares the same upper-bound in $H_{2+\gamma}([0,T)\times\mathbb{R})$ as $(\bar{V}_{T,n})_{n\geq 1}$. Since each $\bar{V}_{T,n}$ solves (3.14) on $[0,T)\times(-n,n)$ and satisfies the boundary condition (3.15) on $[-n,n]$, it follows that $\bar{V}_{T,\infty}$ solves (3.14) on $[0,T)\times\mathbb{R}$ and (3.15) on $\mathbb{R}$. As $\bar{V}_T$ is also a bounded solution of the same equation, comparison implies that $\bar{V}_{T,\infty} = \bar{V}_T$. □

***Remark* 3.3.1.**

a) *Let* $\bar{a} : [0,T) \times \mathbb{R} \mapsto \mathbb{A}$ *be a measurable map satisfying*

$$\bar{a} \in \operatorname*{argmax}_{a \in \mathbb{A}} \left(\mu(\cdot,a)\partial_x \bar{V}_T + \frac{1}{2}\sigma^2(\cdot,a)\partial^2_{xx}\bar{V}_T + r(\cdot,a) \right) \text{ on } [0,T) \times \mathbb{R},$$

see e.g. [24, Prop. 7.33, p.153]. Assume that there exists $\gamma_\circ \in (0,1)$ *such that* $(\mu,\sigma,r)(\cdot,\bar{a})$ *belongs to* $H_{\gamma_\circ}([0,T) \times \mathbb{R})$*, then we can take* $\gamma = \gamma_\circ$*. This follows from [79, §IV.14, p.390].*

b) *If* $(\mu(\cdot,\bar{a}), \sigma(\cdot,\bar{a}), r(\cdot,\bar{a}))$ *has more regularity, one can obviously obtain more regularity on* $\bar{V}_T$ *by, for instance, differentiating the associated partial differential equation.*

c) *In the case where* σ *does not depend on its a-argument, then one can appeal to [84, Thm. 12.16] to deduce that we can take* $\gamma = 1$*. This follows from the Lipschitz continuity of F.*

3.3.3 Convergence speed toward the diffusive limit

We now exploit the Hölder regularity stated above to prove that V^ε_T converges to $\bar{V}_T$ at a rate $\varepsilon^{\gamma/2}$ as ε vanishes. We shall see in Section 3.3.4 below that it provides an $\varepsilon^{\gamma/2}$-optimal control for the pure-jump problem. In general, this cannot be improved, see Example 3.3.1 in Section 3.3.5 below.

Theorem 3.3.1.
For all $(t,x) \in [0,T] \times \mathbb{R}$ and $\varepsilon > 0$,

$$\left| V^\varepsilon_T - \bar{V}_T \right|(t,x) \leq \sup_{\alpha \in \mathcal{A}^t} \mathbb{E}\left[\int_t^T \left| \delta r_\varepsilon(s, X^{t,x,\alpha}_s, \alpha_s) \right| \mathrm{d}s \right]$$

in which

$$\delta r_\varepsilon := \frac{1}{\varepsilon} \int \left(\bar{V}_T(\cdot, \cdot + b_\varepsilon) - \bar{V}_T \right) \nu(\mathrm{d}e) - \mu \partial_x \bar{V}_T - \frac{1}{2}\sigma^2 \partial^2_{xx} \bar{V}_T \tag{3.28}$$

satisfies

$$\|\delta r_\varepsilon\|_\infty \leq C^\varepsilon_K \varepsilon^{\frac{\gamma}{2}} \tag{3.29}$$

with

$$C^\varepsilon_K := \frac{1}{2}\|\partial^2_{xx}\bar{V}_T\|_\infty \left(\varepsilon^{1-\frac{\gamma}{2}} \|b_1\|^2_\infty + 2\varepsilon^{\frac{1-\gamma}{2}} \|b_1\|_\infty \|b_2\|_\infty \right)$$

$$+\frac{K}{2}\left(\varepsilon^{\frac{1}{2}}\|b_1\|_\infty+\|b_2\|_\infty\right)^{2+\gamma},$$

in which $K>0$ is the Hölder constant of $\partial^2_{xx}\bar{V}_T$ with respect to its space variable. In particular,

$$\limsup_{\varepsilon\downarrow 0}\varepsilon^{-\frac{\gamma}{2}}\|V^\varepsilon_T(t,\cdot)-\bar{V}_T(t,\cdot)\|_\infty\le\frac{1}{2}(T-t)K\|b_2\|_\infty^{2+\gamma},\ t\le T.$$

Proof. Since $\bar{V}_T\in\mathcal{C}_b^{(1,2)}([0,T)\times\mathbb{R})$, for any $(t,x,a,e)\in[0,T)\times\mathbb{R}\times\mathbb{A}\times\mathbb{R}$

$$\begin{aligned}\bar{V}_T(t,x+b_\varepsilon(x,a,e))-\bar{V}_T(t,x)=&\partial_x\bar{V}_T(t,x)b_\varepsilon(x,a,e)+\frac{1}{2}\partial^2_{xx}\bar{V}_T(t,x)|b_\varepsilon(x,a,e)|^2\\&+\frac{1}{2}(\partial^2_{xx}\bar{V}_T(t,x_\varepsilon)-\partial^2_{xx}\bar{V}_T(t,x))|b_\varepsilon(x,a,e)|^2\end{aligned}$$

for some x_ε that lies in the interval formed by x and $x+b_\varepsilon(x,a,e)$. By the left-hand side of (3.10), the definition of (μ,σ), and since $\partial^2_{xx}\bar{V}_T$ is γ-Hölder continuous in space with constant K,

$$\begin{aligned}&\left|\frac{1}{\varepsilon}\int\left(\bar{V}_T(t,x+b_\varepsilon(x,a,e))-\bar{V}_T(t,x)\right)\nu(\mathrm{d}e)\right.\\&\qquad\left.-\mu(x,a)\partial_x\bar{V}_T(t,x)-\frac{1}{2}\sigma^2(x,a)\partial^2_{xx}\bar{V}_T(t,x)\right|\\&\le\frac{1}{2}\|\partial^2_{xx}\bar{V}_T\|_\infty\left(\varepsilon\|b_1\|^2_\infty+2\varepsilon^{\frac{1}{2}}\|b_1\|_\infty\|b_2\|_\infty\right)+\varepsilon^{\frac{\gamma}{2}}\frac{K}{2}\left(\varepsilon^{\frac{1}{2}}\|b_1\|_\infty+\|b_2\|_\infty\right)^{2+\gamma}\end{aligned}$$

Hence,

$$\mu\partial_x\bar{V}_T+\frac{1}{2}\sigma^2\partial^2_{xx}\bar{V}_T+r=\frac{1}{\varepsilon}\int\left[\bar{V}_T(\cdot,\cdot+b_\varepsilon(\cdot,e))-\bar{V}_T+\varepsilon(r-\delta r_\varepsilon)\right]\nu(\mathrm{d}e)\tag{3.30}$$

in which δr_ε is the continuous function, defined in (3.28), and satisfies

$$\begin{aligned}\|\delta r_\varepsilon\|_\infty\le&\frac{1}{2}\|\partial^2_{xx}\bar{V}_T\|_\infty\left(\varepsilon\|b_1\|^2_\infty+2\varepsilon^{\frac{1}{2}}\|b_1\|_\infty\|b_2\|_\infty\right)\\&+\varepsilon^{\frac{\gamma}{2}}\frac{K}{2}\left(\varepsilon^{\frac{1}{2}}\|b_1\|_\infty+\|b_2\|_\infty\right)^{2+\gamma}.\end{aligned}$$

Combined with Proposition 3.3.1, this shows that $\bar{V}_T$ is a smooth solution of

$$0=\partial_t\bar{V}_T+\sup_{a\in\mathbb{A}}\frac{1}{\varepsilon}\int\left(\bar{V}_T(\cdot,\cdot+b_\varepsilon(\cdot,a,e))-\bar{V}_T+\varepsilon(r(\cdot,a)-\delta r_\varepsilon(\cdot,a))\right)\nu(\mathrm{d}e)\tag{3.31}$$

on $[0,T)\times\mathbb{R}$ with boundary condition $\bar{V}_T(T,\cdot)=0$ on $\mathbb{R}$. Applying Proposition 3.2.1 (with the appropriate coefficients), this implies that

$$\bar{V}_T(t,x)=\sup_{\alpha\in\mathcal{A}^t}\mathbb{E}\left[\int_t^T \varepsilon(r-\delta r_\varepsilon(s,\cdot))(X_{s-}^{t,x,\alpha},\alpha_s)\mathrm{d}N_s\right],$$

so that, by the definition of V_T^ε,

$$\begin{aligned}\left|V_T^\varepsilon-\bar{V}_T\right|(t,x)&\le\sup_{\alpha\in\mathcal{A}^t}\mathbb{E}\left[\int_t^T \varepsilon\left|\delta r_\varepsilon\right|(s,X_{s-}^{t,x,\alpha},\alpha_s)\mathrm{d}N_s\right]\\&=\sup_{\alpha\in\mathcal{A}^t}\mathbb{E}\left[\int_t^T \left|\delta r_\varepsilon\right|(s,X_{s}^{t,x,\alpha},\alpha_s)\mathrm{d}s\right].\end{aligned}$$

□

3.3.4 Constructing an $\varepsilon^{\gamma/2}$-optimal control for the pure-jump problem

We now show that an $\varepsilon^{\gamma/2}$-optimal control for (3.12) can be constructed by considering a measurable map $\bar{a}:[0,T)\times\mathbb{R}\mapsto\mathbb{A}$ satisfying

$$\bar{a}\in\operatorname*{argmax}_{\bar{a}\in\mathbb{A}}\left(\mu(\cdot,\bar{a})\partial_x\bar{V}_T+\frac{1}{2}\sigma^2(\cdot,\bar{a})\partial_{xx}^2\bar{V}_T+r(\cdot,\bar{a})\right)\ \text{ on }[0,T)\times\mathbb{R},\quad(3.32)$$

see e.g. [24, Prop. 7.33, p.153], and define $\bar{\alpha}^{t,x}\in\mathcal{A}^t$ by

$$\bar{\alpha}_s^{t,x}=\bar{a}(s,X_{s-}^{t,x,\bar{\alpha}^{t,x}}),\ s\in[t,T),$$

recall (3.4). As it is driven by a compound Poisson process, the couple of processes $(X^{t,x,\bar{\alpha}^{t,x}},\bar{\alpha}^{t,x})$ is well-defined.

Proposition 3.3.2.
For all $(t,x)\in[0,T)\times\mathbb{R}$ and $\varepsilon>0$, $\bar{\alpha}^{t,x}$ is $\varepsilon^{\gamma/2}$-optimal for V_T^ε. Namely,

$$\frac{1}{\eta_\varepsilon}\mathbb{E}\left[\int_t^T r(X_{s-}^{t,x,\bar{\alpha}^{t,x}},\bar{\alpha}_s^{t,x})\mathrm{d}N_s\right]\ge V_T^\varepsilon(t,x)-2(T-t)C_K^\varepsilon\varepsilon^{\frac{\gamma}{2}}.$$

Proof. It follows from Proposition 3.3.1, (3.32), and (3.30) that

$$\partial_t\bar{V}_T+\frac{1}{\varepsilon}\int\left(\bar{V}_T(\cdot,\cdot+b_\varepsilon(\cdot,\bar{a},e))-\bar{V}_T+\varepsilon r(\cdot,\bar{a})\right)\nu(\mathrm{d}e)\ge-\left\|\delta r_\varepsilon\right\|_\infty,$$

$$\partial_t \bar{V}_T + \sup_{a\in\mathbb{A}} \frac{1}{\varepsilon}\int \big(\bar{V}_T(\cdot,\cdot + b_\varepsilon(\cdot,a,e)) - \bar{V}_T + \varepsilon r(\cdot,a)\big)\nu(\mathrm{d}e) \le \|\delta r_\varepsilon\|_\infty ,$$

so that applying Itô's Lemma and using (3.15) leads to

$$\bar{V}_T(t,x) - (T-t)\,\|\delta r_\varepsilon\|_\infty \le \frac{1}{\eta_\varepsilon}\mathbb{E}\left[\int_t^T r(X_{s-}^{t,x,\bar{\alpha}^{t,x}}, \bar{\alpha}_s^{t,x})\mathrm{d}N_s\right]$$
$$\bar{V}_T(t,x) + (T-t)\,\|\delta r_\varepsilon\|_\infty \ge \sup_{\alpha\in\mathcal{A}^t}\frac{1}{\eta_\varepsilon}\mathbb{E}\left[\int_t^T r(X_{s-}^{t,x,\alpha}, \alpha_s)\mathrm{d}N_s\right] = V_T^\varepsilon(t,x).$$

We conclude by appealing to (3.29). □

3.3.5 First-order correction term

Under additional conditions, one can exhibit a first-order correction term to improve the convergence speed in Theorem 3.3.1 and Proposition 3.3.2. From now on, we assume the following.

Assumption 3.4.

(i) The map $(t,x,a) \in [0,T)\times\mathbb{R}\times\mathbb{A} \mapsto \varepsilon^{-\gamma/2}\delta r_\varepsilon(t,x,a)$ is continuous, uniformly in $\varepsilon \in (0,1)$.

(ii) The pointwise limit

$$r_1 := \lim_{\varepsilon\to 0} \varepsilon^{-\frac{\gamma}{2}}\delta r_\varepsilon, \tag{3.33}$$

is well-defined on $[0,T)\times\mathbb{R}\times\mathbb{A}$.

(iii) Given

$$\mathbb{A}_0 := \operatorname*{argmax}_{\bar{a}\in\mathbb{A}}\left(\mu(\cdot,\bar{a})\partial_x\bar{V}_T + \frac{1}{2}\sigma^2(\cdot,\bar{a})\partial_{xx}^2\bar{V}_T + r(\cdot,\bar{a})\right),$$

comparison holds in the sense of bounded discontinuous viscosity super- and subsolutions for

$$\begin{cases}\partial_t\varphi + \max_{\bar{a}\in\mathbb{A}_0}\left(\mu(\cdot,\bar{a})\partial_x\varphi + \frac{1}{2}\sigma(\cdot,\bar{a})^2\partial_{xx}^2\varphi + r_1(\cdot,\bar{a})\right) = 0, \text{ on } [0,T)\times\mathbb{R},\\ \varphi(T,\cdot) = 0 \text{ on } \mathbb{R}.\end{cases} \tag{3.34}$$

(iv) For all $(t_\circ, x_\circ) \in [0,T)\times\mathbb{R}$, $\bar{a}_\circ \in \mathbb{A}_0(t_\circ, x_\circ)$ and $(t_n, x_n)_{n\geq 1} \subset [0,T)\times\mathbb{R}$ such that $(t_n, x_n) \to (t_\circ, x_\circ)$ as $n \to \infty$, we can find $(\bar{a}_n)_{n\geq 1}$ such that $\bar{a}_n \in \mathbb{A}_0(t_n, x_n)$ for all $n \geq 1$ and $\bar{a}_n \to \bar{a}_\circ$ as $n \to \infty$.

***Remark* 3.3.2.** *Let us comment the above:*

a) *Note that r_1 is bounded, see* (3.29) *in Theorem 3.3.1. The right-hand side term in* (3.33) *therefore admits a limit superior and a limit inferior. The condition* (3.33) *implies that the limit is well-defined. This point will be further discussed in Remark 3.3.3 below.*

b) *If $\bar{V}_T$ admits a continuous bounded third-order space derivative $\partial^3_{xxx}\bar{V}_T$, then one easily checks that $\gamma = 1$ and*

$$r_1 = \frac{1}{2}\int \left[\frac{1}{3}b_2(\cdot,e)^3\partial^3_{xxx}\bar{V}_T + (b_1 b_2)(\cdot,e)\partial^2_{xx}\bar{V}_T\right]\nu(\mathrm{d}e),$$

by a simple Taylor expansion.

c) *Assume that one can find a continuous map $\bar{\mathrm{a}} : [0,T)\times\mathbb{R} \mapsto \mathbb{A}$ such that $\mathbb{A}_0(t,x) = \{\bar{\mathrm{a}}(t,x)\}$ for all $(t,x) \in [0,T)\times\mathbb{R}$, and $x \in \mathbb{R} \mapsto (\mu,\sigma)(x,\bar{\mathrm{a}}(t,x))$ is Lipschitz uniformly in $t \leq T$, then comparison holds, see e.g. [48, Section 8]. In general, this can be checked on a case-by-case basis.*

Under the above conditions, (3.34) admits a (unique) bounded viscosity solution, denoted by $\delta\bar{V}_T^{(1)}$, see below, and it is the first order term in the difference $V_T^\varepsilon - \bar{V}_T$, i.e. (3.36) below holds with

$$\bar{V}_T^{(1),\varepsilon} := \bar{V}_T + \varepsilon^{\frac{\gamma}{2}}\delta\bar{V}_T^{(1)}. \tag{3.35}$$

Theorem 3.3.2.

Let Assumption 3.4 hold. Then, (3.34) admits a (unique) bounded viscosity solution $\delta\bar{V}_T^{(1)}$ and, for all $(t,x) \in [0,T]\times\mathbb{R}$,

$$\lim_{\varepsilon\downarrow 0}\varepsilon^{-\frac{\gamma}{2}}(V_T^\varepsilon - \bar{V}_T)(t,x) = \delta\bar{V}_T^{(1)}(t,x)$$

and therefore

$$\limsup_{\varepsilon\downarrow 0}\varepsilon^{-\frac{\gamma}{2}}\left|V_T^\varepsilon(t,x) - \bar{V}_T^{(1),\varepsilon}(t,x)\right| = 0, \tag{3.36}$$

in which $\bar{V}_T^{(1),\varepsilon}$ is defined as in (3.35). If in addition $\delta\bar{V}_T^{(1)}$ is $\mathcal{C}^{(1,2)}([0,T)\times\mathbb{R})$ and $\partial^2_{xx}\delta\bar{V}_T^{(1)}$ is $\delta\gamma$-Hölder continuous in space, uniformly on $[0,T)\times\mathbb{R}$,

for some constant $\delta\gamma > 0$ such that

$$\limsup_{\varepsilon \downarrow 0} \varepsilon^{-\frac{\delta\gamma}{2}} \left\| \varepsilon^{-\frac{\gamma}{2}} \delta r_\varepsilon - r_1 \right\|_\infty < \infty, \tag{3.37}$$

then the control defined by

$$\check{\alpha}_s^{t,x} = \check{a}(s, X_{s-}^{t,x,\check{\alpha}^{t,x}}),\ s \in [t, T) \tag{3.38}$$

with

$$\check{a} \in \operatorname*{argmax}_{\bar{a} \in \mathbb{A}_0} \left\{ \mu(\cdot, \bar{a}) \partial_x \delta \bar{V}_T^{(1)} + \frac{1}{2} \sigma(\cdot, \bar{a})^2 \partial_{xx}^2 \delta \bar{V}_T^{(1)} + r_1(\cdot, \bar{a}) \right\},\ \text{on } [0, T) \times \mathbb{R}, \tag{3.39}$$

satisfies

$$\frac{1}{\eta_\varepsilon} \mathbb{E} \left[\int_t^T r(X_{s-}^{t,x,\check{\alpha}^{t,x}}, \check{\alpha}_s^{t,x}) \mathrm{d}N_s \right] \geq V_T^\varepsilon(t, x) - o\left(\varepsilon^{\frac{\gamma}{2}}\right),\ \text{ for all } \varepsilon > 0,$$

where $o : \mathbb{R}_+ \to \mathbb{R}$ is a continuous bounded function such that $o(y)/y \to 0$ as $y \downarrow 0$.

Proof. We split the proof into two steps.

1. Let us set $W_\varepsilon := \varepsilon^{-\gamma/2}(V_T^\varepsilon - \bar{V}_T)$ and consider its relaxed semi-limits

$$W^*(t, x) := \limsup_{\substack{(t', x') \to (t, x) \\ \varepsilon \downarrow 0}} W_\varepsilon(t', x'),\ W_*(t, x) := \liminf_{\substack{(t', x') \to (t, x) \\ \varepsilon \downarrow 0}} W_\varepsilon(t', x').$$

Note that Theorem 3.3.1 ensures that the above are well-defined and bounded. We claim that W^* and W_* are respectively bounded sub- and supersolutions of (3.34). For brevity, we will only include the details for the proof of the subsolution property, the supersolution property is proved similarly and we only mention how to adapt the arguments. Fix $\varphi \in \mathcal{C}_b^{(1,2)}$ and let $(t_\circ, x_\circ) \in [0, T) \times \mathbb{R}$ achieve a strict maximum of $W^* - \varphi$ on a ball $B_k := \{(t, x) \in [0, T) \times \mathbb{R} : |t_\circ - t'| \leq (T - t_\circ)/2,\ |x_\circ - x'| \leq k\} \subset [0, T) \times \mathbb{R}$, for some $k > 0$. Then, there exist a sequence $(t_{\varepsilon_n}, x_{\varepsilon_n})_{\varepsilon_n}$ such that $\varepsilon_n \to 0$, $W_{\varepsilon_n}(t_{\varepsilon_n}, x_{\varepsilon_n}) \to W^*(t_\circ, x_\circ)$, $(t_{\varepsilon_n}, x_{\varepsilon_n}) \to (t_\circ, x_\circ)$, and such that $(t_{\varepsilon_n}, x_{\varepsilon_n})$ is a maximum of $W_{\varepsilon_n} - \varphi$ in the interior of B_{2k}, see e.g. [17, Lemma 6.1]. For $k > \varepsilon_n^{1/2}(\|b_1\|_\infty + \|b_2\|_\infty)$, the viscosity subsolution property of $V_T^{\varepsilon_n}$, applying Proposition 3.2.1 to the test function $\bar{V}_T +$

$\varepsilon_n^{\gamma/2}\varphi$, implies that

$$0 \leq \partial_t\left(\bar{V}_T + \varepsilon_n^{\frac{\gamma}{2}}\varphi\right)(t_{\varepsilon_n}, x_{\varepsilon_n}) + \frac{1}{\varepsilon_n}\left\{\left(\bar{V}_T + \varepsilon_n^{\frac{\gamma}{2}}\varphi\right)(t_{\varepsilon_n}, x_{\varepsilon_n}) + \varepsilon_n r(x_{\varepsilon_n}, \bar{a}_n) - \int\left(\bar{V}_T + \varepsilon_n^{\frac{\gamma}{2}}\varphi\right)(t_{\varepsilon_n}, x_{\varepsilon_n} + b_{\varepsilon_n}(x_{\varepsilon_n}, \bar{a}_n, e))\nu(\mathrm{d}e)\right\}$$

for some $\bar{a}_n \in \mathbb{A}$. Since $\varphi \in \mathcal{C}_b^{(1,2)}$, a second order Taylor expansion combined with Assumption 3.2 implies that

$$\varepsilon_n^{\frac{\gamma}{2}}\left[\partial_t\varphi(t_{\varepsilon_n}, x_{\varepsilon_n}) + \frac{1}{\varepsilon_n}\left(\int \varphi(t_{\varepsilon_n}, x_{\varepsilon_n} + b_{\varepsilon_n}(x_{\varepsilon_n}, \bar{a}_n, e))\nu(\mathrm{d}e) - \varphi(t_{\varepsilon_n}, x_{\varepsilon_n})\right)\right]$$

goes to 0 as $n \to \infty$. Thus, if $\bar{a}$ is a limit point of $(\bar{a}_n)_{n\geq 1}$, we deduce from (3.28)-(3.29) and the above that

$$0 \leq \partial_t \bar{V}_T(t_\circ, x_\circ) + \mu(x_\circ, \bar{a})\partial_x \bar{V}_T(t_\circ, x_\circ) + \frac{1}{2}\sigma^2(x_\circ, \bar{a})\partial_{xx}^2 \bar{V}_T(t_\circ, x_\circ) + r(x_\circ, \bar{a}).$$

In view of Proposition 3.3.1, this shows that $\bar{a}_n$ converges to some element of $\bar{a} \in \mathbb{A}_0(t_\circ, x_\circ)$ as n goes to infinity, after possibly passing to a subsequence. By (3.31) and the above,

$$0 \leq \partial_t\varphi(t_{\varepsilon_n}, x_{\varepsilon_n}) + \frac{1}{\varepsilon_n}\int\Big(\varphi(t_{\varepsilon_n}, x_{\varepsilon_n} + b_{\varepsilon_n}(x_{\varepsilon_n}, \bar{a}_n, e)) - \varphi(t_{\varepsilon_n}, x_{\varepsilon_n}) + \varepsilon_n\varepsilon_n^{-\frac{\gamma}{2}}\delta r_{\varepsilon_n}(t_{\varepsilon_n}, x_{\varepsilon_n}, \bar{a}_n)\Big)\nu(\mathrm{d}e).$$

Sending $n \to \infty$ and using parts (i) and (ii) of Assumption 3.4 together with Assumption 3.2, this leads to

$$0 \leq \partial_t\varphi(t_\circ, x_\circ) + \mu(x_\circ, \bar{a})\partial_x\varphi(t_\circ, x_\circ) + \frac{1}{2}\sigma(x_\circ, \bar{a})\partial_{xx}^2\varphi(t_\circ, x_\circ) + r_1(t_\circ, x_\circ, \bar{a}),$$

so that the required subsolution property is proved on $[0, T) \times \mathbb{R}$. The fact that $W^*(T, \cdot) \leq 0$ follows from the last assertion of Theorem 3.3.1. To prove the supersolution property, it suffices to follow the same arguments but choose $\bar{a}_n \in \mathbb{A}_0(t_{\varepsilon_n}, x_{\varepsilon_n})$ that converges to some arbitrary $\bar{a}_\circ \in \mathbb{A}(t_\circ, x_\circ)$, see (iv) of Assumption 3.4. For a test function $\varphi \in \mathcal{C}_b^{(1,2)}$ for W_* at $(t_\circ, x_\circ) \in [0, T) \times \mathbb{R}$, keeping the same notations as above, this lead to

$$0 \geq \partial_t(\bar{V}_T + \varepsilon_n^{\frac{\gamma}{2}}\varphi)(t_{\varepsilon_n}, x_{\varepsilon_n}) + \frac{1}{\varepsilon_n}\left\{\varepsilon_n r(x_{\varepsilon_n}, \bar{a}_n) - \left(\bar{V}_T + \varepsilon_n^{\frac{\gamma}{2}}\varphi\right)(t_{\varepsilon_n}, x_{\varepsilon_n}) + \int\left(\bar{V}_T + \varepsilon_n^{\frac{\gamma}{2}}\varphi\right)(t_{\varepsilon_n}, x_{\varepsilon_n} + b_{\varepsilon_n}(x_{\varepsilon_n}, \bar{a}_n, e))\nu(\mathrm{d}e)\right\}$$

$$=\varepsilon_n^{\frac{\gamma}{2}}\left(\partial_t\varphi(t_{\varepsilon_n},x_{\varepsilon_n})+\frac{1}{\varepsilon_n}\int\Big[\varphi(t_{\varepsilon_n},x_{\varepsilon_n}+b_{\varepsilon_n}(x_{\varepsilon_n},\bar{a}_n,e))-\varphi(t_{\varepsilon_n},x_{\varepsilon_n})\right.$$
$$\left.+\varepsilon_n\varepsilon_n^{-\frac{\gamma}{2}}\delta r_{\varepsilon_n}(t_{\varepsilon_n},x_{\varepsilon_n},\bar{a}_n)\Big]\nu(\mathrm{d}e)\right)$$

by Proposition 3.3.1 and (3.28).

By comparison, $W := W^* = W_*$ is the unique bounded viscosity solution of (3.34) and is therefore equal to $\delta\bar{V}_T^{(1)}$.

2. We now assume that $\delta\bar{V}_T^{(1)}$ is $\mathcal{C}^{(1,2)}([0,T)\times\mathbb{R})$ and that $\partial^2_{xx}\delta\bar{V}_T^{(1)}$ is $\delta\gamma$-Hölder continuous in space, uniformly on $[0,T)\times\mathbb{R}$, for some $\delta\gamma>0$ such that (3.37) holds. Using (3.37) and the same arguments as in the proof of Theorem 3.3.1 lead to

$$\limsup_{\varepsilon\downarrow 0}\varepsilon^{-\frac{\delta\gamma}{2}}\left\|\delta r_\varepsilon^{(1)}\right\|_\infty<\infty, \tag{3.40}$$

in which

$$\delta r_\varepsilon^{(1)}:=\frac{1}{\varepsilon}\int\left(\delta\bar{V}_T^{(1)}(\cdot,\cdot+b_\varepsilon)-\delta\bar{V}_T^{(1)}\right)\nu(\mathrm{d}e)+\varepsilon^{-\frac{\gamma}{2}}\delta r_\varepsilon-r_1.$$
$$-\mu\partial_x\delta\bar{V}_T^{(1)}-\frac{1}{2}\sigma^2\partial^2_{xx}\delta\bar{V}_T^{(1)}$$

Moreover, direct computations using the above and (3.31) show that $\bar{V}_T^{(1),\varepsilon}$ defined in (3.35) solves

$$0=\partial_t\bar{V}_T^{(1),\varepsilon}+r(\cdot,\check{a})$$
$$+\frac{1}{\varepsilon}\int\left(\bar{V}_T^{(1),\varepsilon}(\cdot,\cdot+b_\varepsilon(\cdot,\check{a},e))-\bar{V}_T^{(1),\varepsilon}(t,x)-\varepsilon\varepsilon^{\frac{\gamma}{2}}\delta r_\varepsilon^{(1)}(\cdot,\check{a})\right)\nu(\mathrm{d}e)$$

on $[0,T)\times\mathbb{R}$, in which $\check{a}$ is defined as in (3.39). Together with (3.40), this implies that, for $\check{\alpha}^{t,x}$ defined as in (3.38), we have

$$\frac{1}{\eta_\varepsilon}\mathbb{E}\left[\int_t^T r(X_{s-}^{t,x,\check{\alpha}^{t,x}},\check{\alpha}_s^{t,x})\mathrm{d}N_s\right]\geq\bar{V}_T^{(1),\varepsilon}(t,x)-\varepsilon^{\frac{\gamma}{2}}O(\varepsilon),$$

in which $O:\mathbb{R}_+\to\mathbb{R}$ is a continuous function with $O(0)=0$. On the other hand, it follows from Part 1. that $|V_T^\varepsilon(t,x)-\bar{V}_T^{(1),\varepsilon}(t,x)|\leq o(\varepsilon^{\gamma/2})$. □

***Remark* 3.3.3.** *If the limit in* (3.33) *is not defined, one can still define its relaxed limit superior and limit inferior (recall that it is bounded). Let us denote them by r_1^* and r_{1*} respectively. Then, W^* defined in the above proof is simply a viscosity sub-solution of* (3.34) *with r_1^* in place of r_1. Similarly, W_* is a viscosity super-solution of the same equation but with r_{1*} in place of r_1. This still provides asymptotic upper and lower bounds for $\varepsilon^{-\gamma/2}(V_T^\varepsilon-\bar{V}_T)$.*

***Example* 3.3.1.** *To illustrate the above, we consider a toy model in which explicit solutions can be derived. Although it does not satisfy our general assumptions, e.g. of boundedness and Hölder regularity in space, we shall see that a similar approach can still be applied. We consider the dynamics*

$$X^{t,x,\alpha} = x + \int_t^{\cdot} X_{s-}^{t,x,\alpha} \int (\varepsilon b_1(\alpha_s, e) + \sqrt{\varepsilon} b_2(\alpha_s, e)) N(\mathrm{d}e, \mathrm{d}s),$$

in which b_1 and b_2 are bounded and continuous with respect to their first argument, uniformly in the second one. For $\beta \in (0, 1]$, the value function is defined as

$$V_T^{\varepsilon}(t,x) = \sup_{\alpha \in \mathcal{A}^t} \frac{1}{\eta_\varepsilon} \mathbb{E}\left[\int_t^T \int \left| X_{s-}^{t,x,\alpha} \right|^{\beta} r(\alpha_s) \mathrm{d}N_s \right],$$

for some continuous function r. Then, one easily checks that $\bar{V}_T(t,x) = \bar{f}(t)\,|x|^{\beta}$ in which $\bar{f}$ solves

$$\partial_t \bar{f} + \sup_{\bar{a} \in \mathbb{A}} \left(\bar{f}\{\beta\mu(\bar{a}) + \frac{1}{2}\beta(\beta-1)\sigma^2(\bar{a})\} + r(\bar{a}) \right) = 0, \text{ on } [0,T) \times \mathbb{R},$$

with $\bar{f}(T) = 0$. Because $|x|^{\beta}$ factorizes, the Hölder constant of $\partial_{xx}^2 \bar{V}_T$ can be considered around $x = 1$. Since the third-order space derivative of $\bar{V}_T$ is bounded in a neighbourhood of 1, Theorem 3.3.1 applies with $\beta = 1$. The convergence rate is therefore of order $\varepsilon^{1/2}$. Moreover, by direct computations, the first order correction term is of the form $\delta\bar{V}_T^{(1)}(t,x) = \delta\bar{f}(t)\,|x|^{\beta}$ where $\delta\bar{f} \not\equiv 0$ solves

$$\partial_t \delta\bar{f} + \sup_{\bar{a} \in \mathbb{A}_0} \left(\delta\bar{f} \left(\beta\mu(\bar{a}) + \frac{1}{2}\beta(\beta-1)\sigma^2(\bar{a}) \right) + r_1(\cdot, \bar{a}) \right) = 0 \text{ on } [0,T)$$

with $\delta\bar{f}(T) = 0$, in which

$$(t, \bar{a}) \in [0,T] \times \mathbb{A} \mapsto r_1(t, \bar{a}) := \beta(\beta-1)\ell \left(\int (b_1 b_2)(\bar{a}, e) \nu(\mathrm{d}e) \right) \bar{f}(t)$$

for some (explicit) continuous map ℓ with linear growth. In particular, this shows that the convergence rate in $\varepsilon^{1/2}$ proved in Theorem 3.3.1 is sharp.

3.3.6 Higher-order expansions

To conclude this section, note that higher order expansions can be obtained. As opposed to Section 3.3.5, we only provide here a verification argument, upon assuming the existence of an associated system of parabolic equations. Namely, let us assume the following.

Assumption 3.5.
There exists $(\delta\gamma_i)_{i=0,\dots,i_\circ} \subset (0,1]^{i_\circ+1}$ together with $\mathcal{C}^{(1,2)}([0,T)\times\mathbb{R}) \cap C^0([0,T]\times\mathbb{R})$ functions $(\delta\bar{V}_T^{(i)})_{i=0,\dots,i_\circ}$ such that, for $i = 0,\cdots,i_\circ$, $\partial^2_{xx}\delta\bar{V}_T^{(i)}$ is $\delta\gamma_i$-Hölder in space, uniformly on $[0,T)\times\mathbb{R}$, and $\delta\bar{V}_T^{(i)}$ solves

$$\partial_t\delta\bar{V}_T^{(i)} + \mu(\cdot,\check{a}_\varepsilon)\partial_x\delta\bar{V}_T^{(i)} + \frac{1}{2}\sigma(\cdot,\check{a}_\varepsilon)^2\partial^2_{xx}\delta\bar{V}_T^{(i)} + r_i(\cdot,\check{a}_\varepsilon) = 0$$

on $[0,T)\times\mathbb{R}$ with boundary condition $\delta\bar{V}_T^{(i)}(T,\cdot) = 0$ on $\mathbb{R}$, in which $\check{a}_\varepsilon$ is a Borel measurable map such that

$$\check{a}_\varepsilon \in \operatorname*{argmax}_{\bar{a}\in\mathbb{A}}\left(\mu(\cdot,\bar{a})\partial_x\bar{V}_T^{(i_\circ),\varepsilon} + \frac{1}{2}\sigma(\cdot,\bar{a})^2\partial^2_{xx}\bar{V}_T^{(i_\circ),\varepsilon} + r(\cdot,\bar{a})\right),$$

with

$$\bar{V}_T^{(i_\circ),\varepsilon} := \delta\bar{V}_T^{(0)} + \sum_{j=1}^{i_\circ}\varepsilon^{\frac{\gamma_{j-1}}{2}}\delta\bar{V}_T^{(j)}, \text{ in which } \gamma_i := \sum_{j=0}^{i}\delta\gamma_j \text{ for } i \le i_\circ,$$

and, using the conventions $\delta\gamma_{-1} := 0$ and $\delta r_\varepsilon^{(-1)} := r$, for $0 \le i \le i_\circ$,

$$\begin{aligned}\delta r_\varepsilon^{(i)} &:= \frac{1}{\varepsilon}\int\left(\delta\bar{V}_T^{(i)}(\cdot,\cdot+b_\varepsilon) - \delta\bar{V}_T^{(i)}\right)\nu(\mathrm{d}e) + \varepsilon^{-\frac{\delta\gamma_{i-1}}{2}}\delta r_\varepsilon^{(i-1)} \\ &\quad - \mu\partial_x\delta\bar{V}_T^{(i)} - \frac{1}{2}\sigma^2\partial^2_{xx}\delta\bar{V}_T^{(i)} - r_i \\ r_i &:= r1_{\{i=0\}} + 1_{\{i>0\}}\lim_{\varepsilon\to0}\varepsilon^{-\frac{\delta\gamma_{i-1}}{2}}\delta r_\varepsilon^{(i-1)} \text{ for } i \le i_\circ. \end{aligned} \tag{3.41}$$

The limits in (3.41) are well-defined on $[0,T)\times\mathbb{R}\times\mathbb{A}$, and

$$\limsup_{\varepsilon\downarrow0}\varepsilon^{-\frac{\delta\gamma_{i_\circ}}{2}}\|\varepsilon^{-\frac{\delta\gamma_{i_\circ-1}}{2}}\delta r_\varepsilon^{(i_\circ-1)} - r_{i_\circ}\|_\infty < \infty. \tag{3.42}$$

Proposition 3.3.3.
Let Assumption 3.5 hold. Then, for all $(t,x)\in[0,T]\times\mathbb{R}$,

$$\limsup_{\varepsilon\downarrow0}\varepsilon^{-\frac{\gamma_{i_\circ}}{2}}\left|V_T^\varepsilon - \bar{V}_T^{(i_\circ),\varepsilon}\right|(t,x) < \infty.$$

Moreover, the control defined by

$$\check{\alpha}^{t,x}_{\varepsilon,s} = \check{a}_\varepsilon(s, X^{t,x,\check{\alpha}^{t,x}_\varepsilon}_{s-}),\ s \in [t,T),$$

satisfies

$$\frac{1}{\eta_\varepsilon}\mathbb{E}\left[\int_t^T r(X^{t,x,\check{\alpha}^{t,x}_\varepsilon}_{s-}, \check{\alpha}^{t,x}_{\varepsilon,s})\mathrm{d}N_s\right] \geq V^\varepsilon_T(t,x) - C\varepsilon^{\frac{\gamma_{i_\circ}}{2}},\ \text{ for all } \varepsilon > 0,$$

for some constant $C > 0$.

Proof. With the above construction

$$\begin{aligned} 0 = \partial_t \bar{V}^{(i_\circ),\varepsilon}_T &+ \frac{1}{\varepsilon}\int \left(\bar{V}^{(i_\circ),\varepsilon}_T(\cdot, \cdot + b_\varepsilon(\cdot, \check{a}_\varepsilon, e)) - \bar{V}^{(i_\circ),\varepsilon}_T\right)\nu(\mathrm{d}e) \\ &- \varepsilon^{\frac{\gamma_{i_\circ}-1}{2}}\delta r^{(i_\circ)}_\varepsilon(\cdot, \check{a}_\varepsilon) + r(\cdot, \check{a}_\varepsilon) \end{aligned}$$

on $[0,T)\times\mathbb{R}$, while

$$\begin{aligned} 0 \geq \partial_t \bar{V}^{(i_\circ),\varepsilon}_T &+ \frac{1}{\varepsilon}\int \left(\bar{V}^{(i_\circ),\varepsilon}_T(\cdot, \cdot + b_\varepsilon(\cdot, a, e)) - \bar{V}^{(i_\circ),\varepsilon}_T\right)\nu(\mathrm{d}e) \\ &- \varepsilon^{\frac{\gamma_{i_\circ}-1}{2}}\delta r^{(i_\circ)}_\varepsilon(\cdot, a) + r(\cdot, a) \end{aligned}$$

on $[0,T)\times\mathbb{R}$ for all $a : [0,T)\times\mathbb{R} \to \mathbb{A}$. By (3.42) and the same arguments as in the proof of Theorem 3.3.1,

$$\limsup_{\varepsilon\downarrow 0} \varepsilon^{-\frac{\delta\gamma_{i_\circ}}{2}} \|\delta r^{(i_\circ)}_\varepsilon\|_\infty < \infty,$$

so that the required result follows by verification. □

3.4 Application to an Auction Problem

Repeated online auction bidding is a typical problem in which the real value of the parameters b, r and v are unknown, and on which reinforcement learning techniques are applied. The latter requires to estimate, very quickly, the optimal control for different sets of parameters. Being modelled as a discrete-time problem, with fixed auction times, or more realistically in the form of a pure-jump problem as in Section 3.2, see also [54], we face in any case the fact that auctions are issued almost continuously which corresponds to a very small time step in the discrete-time version or to a very large intensity in the pure-jump modelling. The numerical cost of a precise estimation of the optimal control is too important to combine it with a reinforcement learning approach.

3.4.1 Model and description of the optimal policy

We consider here a simple auction problem motivated by online advertising systems. A single ad campaign is provided several opportunities to buy ad space to display its ad over the course of the day. These ad spaces arrive at random, according to the point process N, since they are dependent on users from specific targeted audiences loading a website. In real-world display advertising, the kind encountered on the sides of web pages, these opportunities take the form of an auction between several bidders and an ad-exchange platform.

The format of the auction used is critical to the strategic behaviour of bidders and the revenue of the seller, see Chapter 6 for details. There is a large amount of literature in auction theory on the subject, see e.g. [92, 98, 99], and real-world auctions can take very complex formats. For simplicity, we consider an auctioneer who has implemented a lazy second-price auction [99, 115] with individualised reserve price. In this format, our bidding agent wins the ad slot if it submits a bid above its (henceforth the) reserve price and the competition, and if it wins it pays the maximum between the reserve price and the competition. For a given reserve price x, a bid $a \in (0, +\infty)$ and a random competition bid $B \geq 0$ following a smooth probability distribution F_B, the expected payoff $r(x, a)$ for an auction is thus expressible through a simple integration by parts as

$$r(x,a) = \mathbb{E}[(v - x \vee B)1_{a \geq x \vee B}] = 1_{a \geq x}\left((v-a)F_B(a) + \int_x^a F_B(b)\mathrm{d}b\right), \quad (3.43)$$

in which v is the value of the ad-slot for the bidder. Note that r is not continuous as it is assumed in the preceding sections. In practice, one can replace it with a smooth approximation. In the following, we shall construct a numerical scheme directly on r, without smoothing. It turns out that convergence still seems to be observed at the rate $\varepsilon^{1/2}$. Intuitively, this is due to the fact that the maximum values obtained in (3.7) and (3.14) are the same for r defined with $1_{a \geq x}$ and $1_{a > x}$ whenever $x < \sup \mathbb{A}$, which is true at each time with probability one for the controlled processes defined below.

As the right-hand side of (3.43) highlights, reserve prices are a mechanism put in place by sellers to compensate for lack of competition, which would drive down the price and their profits, see Chapter 6. It is well established that a reserve price is not as profitable as increasing the number of participants by one [41]. Consequently, when there are many bidders a control will have little effect on the system. To clearly demonstrate the use of controlling the reserve price, we study a strongly asymmetric setting, in which the agent has a value $v = 0.5$ much higher than the competition F_B, which we take uniform on $(0, 0.3)$. In this setting, it is directly competing against the seller for its extra value above the average competition. For the

purpose of this example, we do not want to go to the limit of this asymmetry, the posted-price auction in which there is no competition, as it could lead the control problem to degeneracy, such as negative prices and difficult boundary conditions.

There is a large literature on revenue maximisation algorithms in online auctions, or how to set the reserve price to maximise revenue, such as [27, 39, 42]. See also Sections 1.3 and 5.1. For the sake of simplicity, in this example, we will model the dynamics of the reserve price using a simple mean reverting process:

$$b_1(x, a, e) = qa + (1 - q)r_0 - x \text{ and } b_2(x, a, e) = e \text{ with } v \sim \texttt{Unif}(-0.1, 0.1),$$

with $q \in (0, 1)$ and $r_0 \in \mathbb{R}_+$. The reserve price process $X^{t,x,\alpha}$ is then defined from these coefficients as in (3.4), with $b := b_\varepsilon = \varepsilon b_1 + \sqrt{\varepsilon} b_2$ and $\eta := \eta_\varepsilon = \varepsilon^{-1}$. This corresponds to setting a minimum reserve price $(1-q)r_0$, and tracking the agent's bid with aggressiveness measured by q. Setting $r_0 = 0.15$ as the monopoly price of the competition guarantees the seller a better revenue against the competition, while qa allows him to pursue the agent's extra value. We set $q = 1/2$, for a balance between prudence and aggression.

The control problem consists in maximising the static auction revenue while considering the impact bids have on the system. In the static auction format, we can identify three domains the reserve price can be in: "non-competitive", "competitive", and "unprofitable". When the reserve price is below the competition's average[3] There is essentially no prejudice to the agent since the reserve price barely affects his profits. Therefore there is no need to compete with and control the reserve price. On the other hand, when the reserve price is in the range between 0.3 and $v = 0.5$, the reserve price becomes the dominant term in r and the agent has to compete with the seller over the value margin it has relative to other buyers. Finally, if the reserve price is above v, there is no possible profit so no reason to take part in the auction by bidding $a > 0$. For the same reason, we take $\mathbb{A} := [0, 0.5]$.

When the reserve price is dynamic, a good control seeks to maximise profit while pushing the reserve price to the non-competitive domain. One can see this in effect on Fig. 3.1. In the non-competitive regime (left), starting at a reserve price of 0.15, this policy recovers 85% of the best possible income of the static setting, where the reserve price is 0 for all t, and the average price is $0.5 - \mathbb{E}[B] = 0.35$. In the competitive regime (centre), the policy bids just above the reserve price to apply downward pressure until it reaches the non-competitive domain again. Finally, in the unprofitable regime (right), the agent boycotts the auction by bidding 0, bringing down the price. Notice how, when the agent stops boycotting, there is an inflection point in the downward trend of the price, schematically represented by the dotted line.

[3] Recall that the competition here models the distribution of the maximum bid of all other participants, so this average is the average of the maximum of other participants' bids.

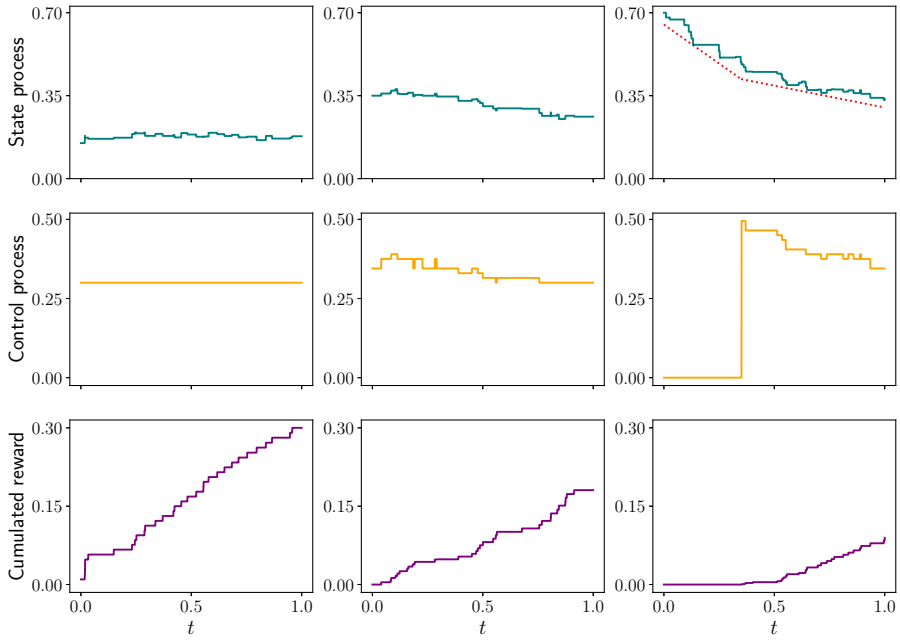

Figure 3.1: Selected sample realisations of the system for $\varepsilon = 10^{-1.5}$, starting from $x = 0.15$ (left), $x = 0.35$ (centre), and $x = 0.7$ (right).

3.4.2 Numerical implementation

Adapting (3.7) and (3.14), we normalise the horizon T to 1, and allow the reserve price to vary in $\mathbb{R}$. This allows us to easily set boundary conditions for the equation. When an auction happens with a price $x \leq 0$, the price is set by the competition, which will be a.s. positive. Thus as $x \to -\infty$, the reserve price becomes irrelevant and the value converges to the value of a single auction without reserve price. Conversely, for $\varepsilon < 1$, as $x \to +\infty$, the probability of $X_\cdot^{0,x,\alpha}$ descending below v by time T and generating any revenue decreases due to the noise. Hence, a Neumann boundary condition set to 0 is appropriate at $[0,1) \times \{-\infty, +\infty\}$. In numerical resolution, we will use Neumann boundary conditions equal to 0 on $[0,1) \times \{-1, 3\}$. Given this domain for the reserve price, we can set the controls on an even mesh in $\mathbb{A} = [0, 0.5]$, of fineness 0.01, denoted by $\mathbb{A}_n := \{10^{-2}k; k = 0, \dots, 50\}$.

We solve both problems numerically with an explicit finite difference solver, and for simplicity a Riemann sum using the same mesh for the numerical integration part of (3.7). This formulation is equivalent to a Markov Chain control problem, see e.g. [78]. Let $\mathrm{M}_t := \{k\Delta_t\,;\, k = 0, \dots, \lfloor 1/\Delta_t \rfloor\}$, $\mathrm{M}_x := \{-1 + k\Delta_x\,;\, k = 0, \dots, \lfloor 4/\Delta_x \rfloor\}$ be the time and space meshes, with

finenesses $\Delta_x = \varepsilon^{3/2}/2$, $\Delta_t = \Delta_x^{2/3}$. Denote $V_{T,n}^{\varepsilon,\Delta}(x_i)$ the output of the solver at time $t_n \in \mathrm{M}_t$ and position $x_i \in \mathrm{M}_x$. For the pure jump problem, we explicitly compute:

$$V_{T,n}^{\varepsilon,\Delta}(x_i) = V_{T,n+1}^{\varepsilon,\Delta}(x_i) + \frac{\Delta_t}{\varepsilon} \sup_{a \in \mathbb{A}_n} \left(\sum_{x_j \in \mathrm{M}_x} V_{T,n+1}^{\varepsilon,\Delta}(x_j) f_{x_i,a}^{\nu,\varepsilon}(x_j) \Delta_x - V_{T,n+1}^{\varepsilon,\Delta}(x_i) + r(x_i, a) \right)$$

in which $f_{x,a}^{\nu,\varepsilon}$ is the transition kernel induced by $b_1(x, a, \cdot)$, $b_2(x, a, \cdot)$, and ν. For the diffusion, we consider meshes $\bar{\mathrm{M}}_t = \{k d_t \,;\, k = 0, \dots, \lfloor 1/d_t \rfloor\}$, $\bar{\mathrm{M}}_x = \{-1 + k d_x \,;\, k = 0, \dots, \lfloor 4/d_x \rfloor\}$, with $d_x = 10^{-2}$, $d_t = d_x^2$ and solve recursively

$$\bar{V}_{T,n}^{\Delta}(x_i) = \bar{V}_{T,n}^{\Delta}(x_i) + d_t \sup_{a \in \mathbb{A}_n} \left\{ (qa + (1-q) r_0 - x_i) \delta_x^u \bar{V}_{T,n}^{\Delta}(x_i) + \frac{\sigma^2}{2} \delta_{xx} \bar{V}_{T,n}^{\Delta}(x_i) + r(x_i, a) \right\}$$

where δ_x^u and δ_{xx} are the uplift first order and centred second order finite differences on $\bar{\mathrm{M}}_x$ respectively.

To give some insight into the complexity trade-off, see that, when ε is large, there are relatively few jumps so the time iteration will not require many steps to get an accurate solution. This scaling is indicated by the Δ_t/ε term. At the same time, the jumps are large so even a coarse mesh in x will be sufficient for the numerical integration to approach the integral. Unfortunately as $\varepsilon \to 0$, one must refine both the time mesh, linearly with $1/\varepsilon$, and the integration mesh which is paid quadratically due to the non-local nature of the equation. In practice, this makes computations grow at a super-cubic rate with ε, which becomes prohibitively expensive quickly. In our example problem, the noise is supported on a bounded interval of size $\varepsilon^{1/2}$, and one thus saves some computation time. Figure 3.2 shows the computation cost (pictured with dots) still grows super-quadratically and overcomes the cost of our accurate diffusion mesh (solid horizontal line) even for large ε. Even though we computed the diffusive limit to very high precision, and with an explicit scheme, for ε of the order of 10^{-3} the CPU time spent on the resolution is already 6 times higher in the pure-jump problem. Note that, in the pure-jump case, if the control were to intervene in a non-linear way we might need to also refine the control mesh with ε, further increasing the computational burden.

Beyond gains in computation, Fig. 3.3 verifies that Theorem 3.3.1 holds with meaningful constants in finite time on this problem. Figure 3.3 shows that the error is very low even for large values of ε, and decreases at the correct rate of $\varepsilon^{1/2}$. Likewise, Fig. 3.4 shows the rate of Proposition 3.3.2 also holds even for large ε.

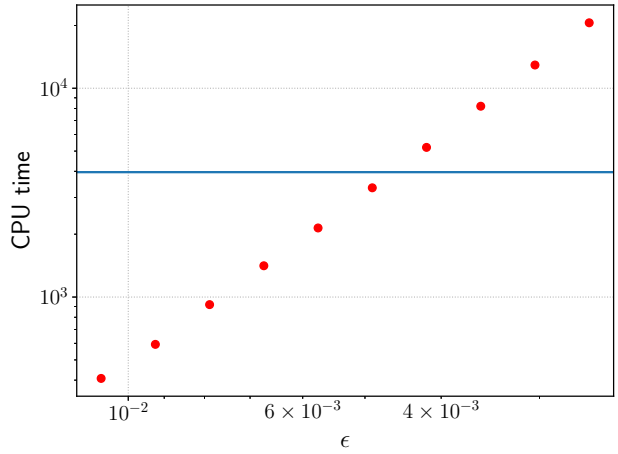

Figure 3.2: Numerical cost for V_T^ε (log scales).

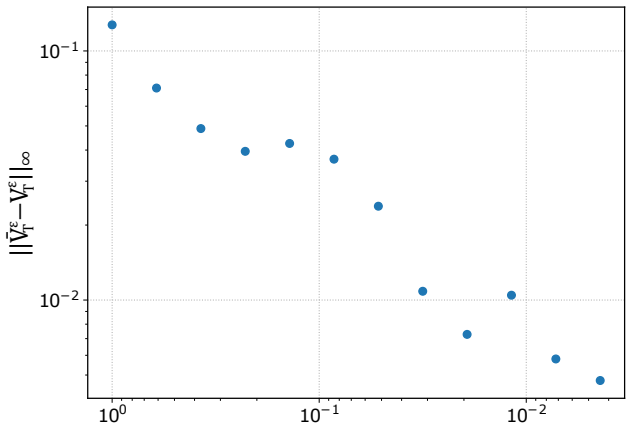

Figure 3.3: Limit value function error relative to V_T^ε, at $t = 0$ (log scales).

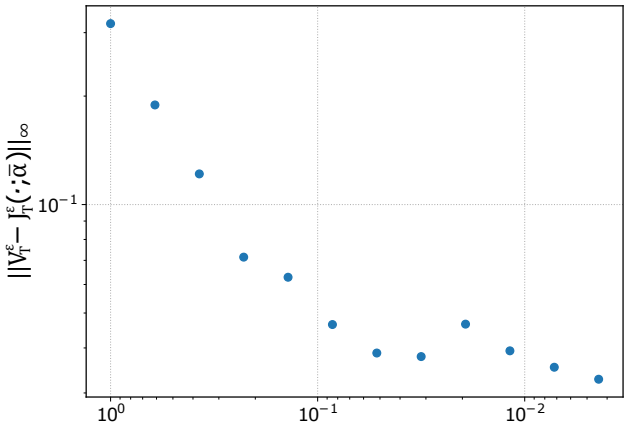

Figure 3.4: Limit policy error relative to V_T^ε, at $t = 0$ (log scales).

3.5 A Remark on the Diffusive Limit of Discrete-Time Problems

Instead of considering the diffusive limit of a continuous time pure-jump problem, one could similarly consider a sequence of pure discrete-time problems with actions at time $t_i^n := iT/n$, $i \leq n$:

$$V_T^n(t,x) := \sup_{\alpha \in \mathcal{A}} \frac{T}{n} \mathbb{E}\left[\sum_{i=1}^n 1_{\{t_i^n \geq t\}} r(\hat{X}_{t_i^n-}^{t,x,\alpha}, \alpha_{t_i^n})\right],$$

with $\hat{X}^{t,x,\alpha}$ defined by

$$\hat{X}^{t,x,\alpha} = x + \sum_{i=1}^n 1_{\{t_i^n \in (t,\cdot]\}} b(\hat{X}_{t_i^n-}^{t,x,\alpha}, \alpha_{t_i^n}, \xi_i^n)$$

and in which $(\xi_i^n)_{i\geq 1}$ is i.i.d. following the distribution ν and $\mathcal{A}$ is the collection of $\mathbb{A}$-valued predictable processes, with respect to the $\mathbb{P}$-augmented filtration generated by

$$\sum_{i=1}^n 1_{\{t_i^n \in [0,\cdot]\}} \exp(\xi_i^n).$$

Upon taking b of the form

$$b_n = \frac{T}{n} b_1 + \sqrt{\frac{T}{n}} b_2, \quad \text{with } \mathbb{E}[b_2(\cdot, \xi_1^n)] = 0,$$

one would obtain the same diffusive limit as in Section 3.3.3 when letting $n \to \infty$. Namely, the same arguments as in [57, § 3] combined with Proposition 3.3.1 and the fact that comparison holds for (3.14) imply that $\lim_{n\to\infty} V_T^n$ is well-defined and is equal to $\bar{V}_T$.

One can also check that the convergence holds at a speed $n^{-\gamma/2}$. Let us sketch the proof. First, the same arguments as in the proof of Theorem 3.3.1 imply that

$$\delta r_n := \frac{n}{T}\mathbb{E}\left[\bar{V}_T(\cdot,\cdot + b_n(\cdot,\xi_i^n)) - \bar{V}_T\right] - \mu\partial_x\bar{V}_T - \frac{1}{2}\sigma^2\partial_{xx}\bar{V}_T$$

satisfies

$$\|\delta r_n\|_\infty \le Cn^{-\frac{\gamma}{2}} \tag{3.44}$$

for some $C > 0$ independent on n. Thus, by Proposition 3.3.1

$$\begin{aligned} 0 = {} & \partial_t\bar{V}_T(t,x)\frac{T}{n} \\ & + \sup_{a\in\mathbb{A}}\mathbb{E}\left[\bar{V}_T(t,x + b_n(x,a,\xi_1^n)) - \bar{V}_T(t,x) + \frac{T}{n}(r(x,a) - \delta r_n(t,x,a))\right] \end{aligned}$$

so that

$$\begin{aligned} & \bar{V}_T(t_i^n,x) \\ & = \sup_{a\in\mathbb{A}}\mathbb{E}\left[\int_{t_i^n}^{t_{i+1}^n}\partial_t\bar{V}_T(t_i^n,x)\mathrm{d}s + \bar{V}_T(t_i^n,x + b_n(x,a,\xi_{i+1}^n))\right. \\ & \qquad\qquad \left. + \frac{T}{n}(r(x,a) - \delta r_n(t_i^n,x,a))\right] \\ & = \sup_{a\in\mathbb{A}}\left(\mathbb{E}\left[\bar{V}_T(t_{i+1}^n,x + b_n(x,a,\xi_{i+1}^n)) + \frac{T}{n}r(x,a)\right]\right. \\ & \qquad \left. + \mathbb{E}\left[\int_{t_i^n}^{t_{i+1}^n}[\partial_t\bar{V}_T(t_i^n,x) - \partial_t\bar{V}_T(s,x + b_n(x,a,\xi_{i+1}^n)) - \delta r_n(t_i^n,x,a)]\mathrm{d}s\right]\right). \end{aligned}$$

We then use (3.24) and (3.44) to obtain that

$$\begin{aligned} & \bar{V}_T(t_i^n,x) \\ & \quad = \sup_{a\in\mathbb{A}}\mathbb{E}\left[\bar{V}_T(t_{i+1}^n,x + b_n(x,a,\xi_{i+1}^n)) + \frac{T}{n}r(x,a) + \int_{t_i^n}^{t_{i+1}^n}\varpi_n(s,x,a)\mathrm{d}s\right] \end{aligned}$$

in which $\|\varpi_n\|_\infty \le Cn^{-\frac{\gamma}{2}}$, for some $C > 0$ independent of n. It follows that

$$\bar{V}_T(t_i^n,x) = \sup_{\alpha\in\mathcal{A}}\mathbb{E}\left[\frac{T}{n}\sum_{j=i}^{n} r(\hat{X}_{t_j^n-}^{t,x,\alpha},\alpha_{t_i^n}) + \int_{t_i^n}^{T}\varpi_n(s,\hat{X}_s^{t,x,\alpha},\alpha_s)\mathrm{d}s\right],$$

which provides the expected result since $\partial_t \bar{V}_T$ is bounded.

Likewise, the Markov control defined through (3.32) can be shown to be $n^{-\gamma/2}$-optimal for V_T^n, see the proof of Proposition 3.3.2.

Conclusion

We studied the diffusion limit of a pure-jump control problem as the jump intensity goes to infinity, upon assuming a correct scaling of the coefficients. Under appropriate conditions, we showed that the second-order derivative of the value function associated with the limiting diffusing problem is Hölder continuous and that its Hölder exponent drives the convergence rate. Convergence can even be improved by using a first (or even higher) order correction scheme. This approach is particularly efficient for the numerical approximation of the optimal control associated with a pure jump process with large intensity, as is the case in auctions associated with online advertising systems.

Chapter 4

Diffusive Limit Approximation of Optimal Ergodic Control Problems

Motivated by the design of fast reinforcement learning algorithms, we study the diffusive limit of a class of pure jump ergodic stochastic control problems. We show that, whenever the intensity of jumps is large enough, the approximation error is governed by the Hölder continuity of the Hessian matrix of the solution to the limit ergodic partial differential equation. This extends to this context the results of Chapter 3. We also explain how to construct a first-order error correction term under appropriate smoothness assumptions. Finally, we quantify the error induced by the use of the Markov control policy constructed from the numerical finite difference scheme associated with the limit diffusive problem, which seems to be new in the literature and of its own interest. This approach permits a very significant reduction in numerical resolution cost[a].

[a] This Chapter is under review as an article in *Stochastic Processes and Applications*, see [3]

* * *

Contents

* * *

4.1 Introduction

Let N be a random point process $\eta\nu(\mathrm{d}e)\mathrm{d}t$, for some finite probability measure ν on $\mathbb{R}^{d'}$, $d' \in \mathbb{N}$, $\eta > 0$, and let $X^{x,\alpha}$ be the solution of

$$X^{x,\alpha} = x + \int_0^{\cdot} \int_{\mathbb{R}^{d'}} b(X^{x,\alpha}_{s-}, \alpha_s, e) N(\mathrm{d}e, \mathrm{d}s)\,,$$

in which α belongs to the set $\mathcal{A}$ of predictable controls with values in some given compact set $\mathbb{A} \subset \mathbb{R}^m$ and the initial data $x \in \mathbb{R}^d$, $m \in \mathbb{N}$. Under some standard stability assumptions, the value of the ergodic optimal control problem

$$\rho^* := \sup_{\alpha \in \mathcal{A}} \liminf_{T \to +\infty} \frac{1}{\eta T} \mathbb{E}\left[\int_0^T r(X^{0,\alpha}_{s-}, \alpha_s) \mathrm{d}N_s \right]$$

with $N_t := N(\mathbb{R}^{d'}, [0,t])$, $t \geq 0$, along with some continuous function w, solves the integro-differential equation

$$\rho^* + \sup_{a \in \mathbb{A}} \left\{ \eta \int_{\mathbb{R}^{d'}} [\mathrm{w}(\cdot + b(\cdot, a, e)) - \mathrm{w}] \nu(\mathrm{d}e) + r(\cdot, a) \right\} = 0 \text{ on } \mathbb{R}^d \tag{4.1}$$

possibly in the viscosity solution sense. This characterisation leads to numerical schemes for approximating the value of the problem and the Markov optimal control.

However, (4.1) is non-local in nature which means that, unless ν is concentrated on a small number of points, the cost of numerical approximation is large, in particular when the intensity η is. This is a problem, e.g., for bidding problems (see Section 1.3) in online display-ad auctions, where the system moves near-continuously in time, meaning that η is very large, and where unknown system parameters motivate the use of reinforcement learning to solve the control problem (see Section 1.2). Reinforcement learning compounds the cost by requiring the computation of ρ^* for many plausible values of the parameters.

On the other hand, when η is very large, asymptotic regimes exist which offer an alternative approximation path, notably the diffusive limit on which this thesis focuses. Indeed, taking $\eta = \varepsilon^{-1}$ and $b(x,a,e) = \varepsilon b_1(x,a,e) + \varepsilon^{1/2} b_2(x,e)$, with $\int_{\mathbb{R}^{d'}} b_2(\cdot, e)\nu(\mathrm{d}e) = 0$, an immediate second order expansion shows that (ρ^*, w) converges as $\varepsilon \to 0$ to the solution $(\bar{\rho}^*, \bar{\mathrm{w}})$ of

$$\bar{\rho}^* + \sup_{\bar{a} \in \mathbb{A}} \left\{ \mathrm{D}\bar{\mathrm{w}} \int_{\mathbb{R}^{d'}} b_1(\cdot, \bar{a}, e)\nu(\mathrm{d}e) + \mathrm{Tr}\left[\int_{\mathbb{R}^{d'}} b_2 b_2^{\mathsf{T}}(\cdot, e)\nu(\mathrm{d}e) \mathrm{D}^2\bar{\mathrm{w}} \right] + r(\cdot, a) \right\} = 0 \tag{4.2}$$

on $\mathbb{R}^d$.

Unlike (4.1), (4.2) is a local equation and much more easily solved numerically. Note that another possible limit regime, albeit less precise, is obtained via a first-order expansion as in [54], which corresponds to considering a fluid limit.

For such a specification of the coefficients (η, b), the existence of a diffusive limit is expected, see e.g. [66] for general results on the convergence of stochastic processes. Stability of viscosity solutions, see e.g. [57, § 3], can also be used to prove the convergence of the value function of stochastic control problems. This has been a subject of particular interest in the insurance and queueing network literatures, see e.g. [21, 43, 45]. Nonetheless, these approaches do not permit to characterize the speed of convergence in the case of a (generic) ergodic optimal control problem as defined in Section 4.2 below, which is essential for studying general reinforcement learning problems, see Chapter 5.

The aim of this chapter is to characterize this convergence speed and explain how to numerically construct, in an efficient way, an approximation of the optimal control. A first step in this direction was made in Chapter 3 by considering finite time horizon problems. Such problems are easier to handle from a mathematical point of view but are unfortunately less adapted to reinforcement learning algorithms, see Section 1.2.

Still, a similar approach can be used, up to additional technicalities. As in Chapter 3, we study the regularity of $\bar{w}$ in the solution couple to (4.2). We show that its second order derivative is (locally) γ-Hölder with a constant of at most linear growth in x, for some $\gamma \in (0, 1]$, whenever the coefficients of (4.2) are uniformly Lipschitz in space, $\int b_1(\cdot, e)\nu(\mathrm{d}e)$ has linear growth, b_2 and r are continuous and bounded and under a uniform ellipticity condition. By a second-order Taylor expansion, this allows us to pass (rigorously) from (4.2) to (4.1) up to an error term of order $\varepsilon^{\gamma/2}$ (locally), and therefore provides the required convergence rate by verification. In general, this rate cannot be improved. As a by-product, the Markov control taken from the Hamilton-Jacobi-Bellman equation of the diffusive limit problem provides an $\varepsilon^{\gamma/2}$-optimal control for the original pure-jump control problem. Under additional regularity assumptions, it can even be improved by constructing a first-order correction term.

In principle, this provides an efficient way of constructing an almost optimal Markov control . However, it still remains to build up a pure numerical scheme. To complete the picture we therefore derive a convergence rate for a finite difference method for the numerical estimation of $\bar{\rho}^*$, depending again on γ. More importantly, we explain how to numerically construct an almost optimal Markov control process based on a smoothed version of the numerical approximation of $\bar{w}$ and we obtain a convergence rate towards $\bar{\rho}^*$, and therefore ρ^*, of the expected average gain associated to such a control. The latter seems to be (surprisingly) completely new and of own interest in the

optimal control literature.

As an example of application, we consider in Section 4.5 a simplified repeated online auction bidding problem, where a buyer seeks to maximise its profit when facing both competition and a seller who adapts its price to incoming bids. This example extends the one of Section 3.4. Our numerical experiments show that our approximation permits a considerable gain in computation time (as expected).

Note that we restrict here to the case where b_2 does not depend on the value of the control, meaning that $\bar{w}$ solves a semi-linear equation. In principle, the fully non-linear case could be studied along the same lines of arguments but the required regularity of the corresponding function $\bar{w}$ would be much more complex to derive. We avoid considering this more general case for the sake of simplicity (note that standard reinforcement learning problems use simple additive noises, see Chapter 5).

Notations

We collect here some standard notations that will be used throughout this chapter. Any element x of $\mathbb{R}^d$ is viewed as a column vector. $\mathbb{M}^d$ (resp. $\mathbb{S}^d$) denotes the collection of (resp. symmetric) d-dimensional matrices. On $\mathbb{R}^d$ or $\mathbb{M}^d$, the superscript $^\mathsf{T}$ denotes transposition, we set $xy := x^\mathsf{T} y$ and $\|x\| := \sqrt{xx}$ for $x, y \in \mathbb{R}^d$. We let $\mathrm{Tr}[M]$ denote the trace of $M \in \mathbb{M}^d$ and $\|M\|$ be the Euclidean norm of M viewed as a vector of $\mathbb{R}^{d\times d}$. We denote by $\mathcal{B}_\ell(x)$ the open ball centred at $x \in \mathbb{R}^d$ of radius $\ell > 0$. Given an open set $\mathcal{U} \subset \mathbb{R}^n$, $n \geq 1$, $p \in \{0, 1, 2\}$, we use the standard notation $\mathcal{C}^p(\mathcal{U})$ to denote the space of p-times continuously differentiable real-valued maps u on $\mathcal{U}$, and $\mathcal{C}^p_b(\mathcal{U})$ to denote the subspace of functions $u \in \mathcal{C}^p(\mathcal{U})$ such that

$$\|u\|_{\mathcal{C}^p_b(\mathcal{U})} := \sum_{j=0}^{p} \sup_{x\in\mathcal{U}} |\mathrm{D}^j u(x)| < \infty$$

in which $\mathrm{D}^0 u := u$, $\mathrm{D}^1 u$ is the gradient of u, as a line vector, $\mathrm{D}^2 u$ is the Hessian matrix of u. When $p = 0$, and even if u is not continuous, we simply write $\|u\|_{\infty,\mathcal{U}}$ to denote the sup of $|u|$ on $\mathcal{U}$. Given $\gamma \in [0, 1]$, we denote the γ-Hölder modulus of $u \in \mathcal{C}^0(\mathcal{U})$ on $\mathcal{U}$ as

$$[u]^\gamma_{\mathcal{C}^0(\mathcal{U})} := \sup_{x,x'\in\mathcal{U}} \frac{|u(x') - u(x)|}{|x' - x|^\gamma},$$

where we use the convention $0/0 = 0$. If $u = (u^1, \cdots, u^d)$ takes values in $\mathbb{R}^d$, $d \geq 1$, we use the same notation to denote the sum of the elements $\{[u^i]^\gamma_{\mathcal{C}^0(\mathcal{U})}, i \leq d\}$. We write $u \in \mathcal{C}^{p,\gamma}(\mathcal{U})$ if $\mathrm{D}^p u$ is γ-Hölder on each compact

subset of $\mathcal{U}$, and $u \in \mathcal{C}_b^{p,\gamma}(\mathcal{U})$ if

$$\|u\|_{\mathcal{C}_b^{p,\gamma}(\mathcal{U})} := \|u\|_{\mathcal{C}_b^p(\mathcal{U})} + [\mathrm{D}^p u]^{\gamma}_{\mathcal{C}^0(\mathcal{U})} < \infty.$$

In particular $\mathcal{C}^{0,1}(\mathcal{U})$ denotes the set of Lipschitz functions on $\mathcal{U}$.

If u is restricted to take values in a subset $\mathcal{U}'$ of $\mathbb{R}$, we write $\mathcal{C}^p(\mathcal{U};\mathcal{U}')$, $\mathcal{C}_b^p(\mathcal{U};\mathcal{U}')$, $\mathcal{C}^{p,\gamma}(\mathcal{U};\mathcal{U}')$ or $\mathcal{C}_b^{p,\gamma}(\mathcal{U};\mathcal{U}')$ for the corresponding sets. We also use the notation $\mathcal{C}^0_{\mathrm{lin}}(\mathcal{U})$ to denote the collection of continuous real-valued function u such that

$$[u]_{\mathcal{C}^0_{\mathrm{lin}}(\mathcal{U})} := \sup_{x\in\mathcal{U}} \frac{|u(x)|}{1+\|x\|} < \infty.$$

In all the above notations, we omit $\mathcal{U}$ if it is equal to $\mathbb{R}^d$ and $\mathcal{U}'$ if it is clear.

4.2 Pure-Jump Ergodic Optimal Control

In order to alleviate notations, we first consider the case where the intensity of the jump process is given, and recall rather standard results from the ergodic control literature.

Let $\Omega = \mathbb{D}$ denote the space of d-dimensional càdlàg functions on $\mathbb{R}_+$ and $\mathcal{M}(\mathbb{R}^{d'}\times\mathbb{R}_+)$ denote the collection of positive finite measures on $\mathbb{R}^{d'}\times\mathbb{R}_+$, for some $d, d' \in \mathbb{N}^*$. Consider a measure-valued map $N : \mathbb{D} \mapsto \mathcal{M}(\mathbb{R}^{d'}\times\mathbb{R}_+)$ and a probability measure $\mathbb{P}$ on $\mathbb{D}$ such that N is a right-continuous real-valued $\mathbb{R}^{d'}$-marked point process with compensator $\eta\nu(\mathrm{d}e)\mathrm{d}t$, in which $\eta > 0$ and ν is a probability measure on $\mathbb{R}^{d'}$. See Chapter 3 or, e.g. , [37]. For ease of notations, we set $N_t := N(\mathbb{R}^{d'}, [0,t])$ for $t \geq 0$.

Let $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ denote the $\mathbb{P}$-augmentation of the filtration generated by $(\int_0^t \int_{\mathbb{R}^{d'}} \exp(e) N(\mathrm{d}e, \mathrm{d}r))_{t\geq 0}$. Given a compact set $\mathbb{A} \subset \mathbb{R}^m$, $m \in \mathbb{N}$, let $\mathcal{A}$ be the collection of $\mathbb{F}$-predictable processes with values in $\mathbb{A}$. Throughout this chapter, unless otherwise stated, we will work on the filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, where $\mathcal{F} = \mathcal{F}_\infty$.

Given $(t,x) \in \mathbb{R}_+ \times \mathbb{R}^d$, $\alpha \in \mathcal{A}$, and a measurable map $(x,a,e) \in \mathbb{R}^d \times \mathbb{A} \times \mathbb{R}^{d'} \mapsto b(x,a,e) \in \mathbb{R}^d$, we define the càdlàg process $X^{x,\alpha}$ as the solution of

$$X^{x,\alpha} = x + \int_0^{\cdot} \int_{\mathbb{R}^{d'}} b(X^{x,\alpha}_{s-}, \alpha_s, e) N(\mathrm{d}e, \mathrm{d}s)\,. \tag{4.3}$$

We then consider the ergodic gain functional

$$\rho(x,\alpha) := \liminf_{T\to\infty} \frac{1}{\eta T}\mathbb{E}\left[\int_0^T r(X_{s-}^{x,\alpha},\alpha_s)\mathrm{d}N_s\right],\ (x,\alpha)\in\mathbb{R}^d\times\mathcal{A}, \tag{4.4}$$

for some bounded measurable map $(x,a)\in\mathbb{R}^d\times\mathbb{A}\mapsto r(x,a)\in\mathbb{R}$. Note that this actually also pertains to the case where the reward function r depends on the mark e, by arguing as in Remark 4.2.1 below. By the same remark, the cost could have an extra component given in terms of the Lebesgue measure.

In the above the scaling by $1/(\eta T)$ means that we consider the gain by average unit of time the controller acts on the system. Indeed, $\mathbb{E}[N_T]=\eta T$ and the control applies only at jump times of N.

This functional induces an infinite horizon control problem corresponding to finding the value function

$$\rho^* := \sup_{\alpha\in\mathcal{A}}\rho(\cdot,\alpha). \tag{4.5}$$

This problem is meaningfully ergodic when ρ^* is constant over $\mathbb{R}^d$, i.e. the initial condition does not play any role.

Throughout this chapter, we make the following assumptions. First, we impose some control on the coefficients (b,r).

Assumption 4.1.
The map (b,r) is continuous. Moreover, there exists $L_{b,r}>0$ such that

$$[b(\cdot,a,e)]_{\mathcal{C}^0_{\mathrm{lin}}} + \|r(\cdot,a)\|_{\mathcal{C}_b^{0,1}} \le L_{b,r},\ \text{ for all } (a,e)\in\mathbb{A}\times\mathbb{R}^{d'}.$$

The next assumption, known as asymptotic flatness, guarantees that each control process contracts all possible paths of (4.3) exponentially fast to a single trajectory. This is a sufficient condition to ensure that ρ^* does not depend on the initial condition, refer to the proof of Lemma 4.A.1 in Section 4.A. It can be compared to standard assumptions used in the Brownian diffusion case as in e.g. [13, Pf. of Lemma 7.3.4], up to a more abstract statement.

Assumption 4.2.
There is $\zeta\in\mathcal{C}^0(\mathbb{R}^d\times\mathbb{R}^d;\mathbb{R}_+)$ such that

(i) There exists $(\ell_\zeta, L_\zeta)\in(\mathbb{R}_+^*)^2$ and $p_\zeta\ge 1$ for which

$$\ell_\zeta\|x-x'\|^{p_\zeta} \le \zeta(x,x') \le L_\zeta\|x-x'\|^{p_\zeta},\ \text{ for all } x,x'\in\mathbb{R}^d.$$

(ii) There exists $C_\zeta > 0$ such that for all $x, x' \in \mathbb{R}^d$, $a \in \mathbb{A}$ and $\iota > 0$

$$\eta \int_{\mathbb{R}^{d'}} \{\zeta(x + b(x,a,e), x' + b(x',a,e)) - \zeta(x,x')\} \nu(\mathrm{d}e) \leq -C_\zeta \zeta(x,x') \,. \tag{4.6}$$

Our last assumption is typically required to control the long-time behaviour of solutions of (4.3), see Lemma 4.A.2 in the Appendix. It is a form of Lyapunov stability assumption, see e.g. [29, 63] for comparison.

Assumption 4.3.
There is $\xi \in \mathcal{C}^0(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}_+)$ such that

(i) There exists $(\ell_\xi, L_\xi) \in (\mathbb{R}_+^*)^2$ and $p_\xi \geq 1$ for which

$$\ell_\xi |x|^{p_\xi} \leq \xi(x) \leq L_\xi \|x\|^{p_\xi}, \quad \text{for all } x \in \mathbb{R}^d.$$

(ii) There exists $C_\xi^1 > 0$ and $C_\xi^2 \in \mathbb{R}$ such that for all $x \in \mathbb{R}^d$, $a \in \mathbb{A}$

$$\eta \int_{\mathbb{R}^{d'}} \{\xi(x + b(x,a,e)) - \xi(x)\} \nu(\mathrm{d}e) \leq -C_\xi^1 \xi(x) + C_\xi^2. \tag{4.7}$$

***Example* 4.2.1.** *Consider a bidding problem in a repeated auction with reserve (see Chapter 6 or [75] for an introduction), in which X stands for the current reserve price and α is the bid. We set $e = (e_1, e_2, e_3, e_4) \in \mathbb{R}^4$ and consider the dynamic induced by $b(x,a,e) := e_1(ae_2 + e_3 - x)$ for $\mathbb{A} := [\underline{a}, \bar{a}] \subset \mathbb{R}_+$. This means that the dynamic is mean-reverting around the level $ae_2 + e_3$. In this formula, e_2 correspond to the retail value (the price at which the bidder will sell to the final client the product he bought) so that the value a of the control is the so-called shading factor. Then, $e_1 \geq 0$ is the realization of a random mean-reversion speed and e_3 is an exogenous noise. If the reserve price value x is smaller than the bid price ae_2 (up to the additional noise e_3) then it moves up for the next auction, and the other way around if it is bigger. In a second-price auction, with e_4 as the value of the competition bid, the natural reward function is*

$$r(x,a) = \int_{\mathbb{R}^4} (e_2 - x \vee e_4) 1_{\{ae_2 \geq x \vee e_4\}} \nu(\mathrm{d}e) \,.$$

We assume that $\nu([0,1] \times \mathbb{R}_+ \times \mathbb{R}^2) = 1$, $1 - \int_{\mathbb{R}^4} (1 - e_1)^{2p} \nu(de) =: m_1 \in (0,1]$ and that $\int_{\mathbb{R}^4} \sup_{a \in \mathbb{A}} |ae_1 e_2 + e_1 e_3|^{2p} \nu(de) < \infty$, for some integer $p \geq 1$. Then, Assumption 4.2 holds with $\zeta(x,x') := |x - x'|^{2p}$ and $C_\zeta = \eta m_1$, while Assumption 4.3 holds with $\xi(x) = |x|^{2p}$, $C_\xi^1 = \frac{1}{2}\eta m_1$ and $C_\xi^2 = \eta C_e$ for some $C_e > 0$ that does not depend on η.

Under a standard log-normal model for valuations (see e.g. [98]), and a uniform competition on $[0, \bar{c}]$ for some $\bar{c} > 0$, it is easily verified that Assumption 4.1 holds. This example is developed further in Section 4.5.

Under the above assumptions, we obtain the following classical result, Theorem 4.2.1 below whose proof is rather standard, but produced in Section 4.A by lack of an appropriate reference. To state it, we first need to introduce the following auxiliary optimal control problems, defined for all $x \in \mathbb{R}^d$, $\lambda, T > 0$, and $t \leq T$:

$$V_\lambda(x) := \sup_{\alpha \in \mathcal{A}} J_\lambda(x, \alpha) \text{ with } J_\lambda(x, \alpha) := \frac{1}{\eta} \mathbb{E}\left[\int_0^\infty e^{-\lambda s} r(X_{s-}^{x,\alpha}, \alpha_s) \mathrm{d}N_s \right] \quad (4.8)$$

and

$$V_T(t, x) := \sup_{\alpha \in \mathcal{A}} J_T(t, x, \alpha) \text{ with } J_T(t, x, \alpha) := \frac{1}{\eta} \mathbb{E}\left[\int_t^T r(X_{s-}^{t,x,\alpha}, \alpha_s) \mathrm{d}N_s \right] \quad (4.9)$$

in which $X^{t,x,\alpha}$ is defined as in (4.3) starting from t, see also (3.4)

***Remark* 4.2.1.** *Note that Assumption 4.1 implies that $\sup_{[0,t]} |X^{x,\alpha}|$ has moments of any order, for all $t \geq 0$, $(x, \alpha) \in \mathbb{R}^d \times \mathcal{A}$. Also, it follows from Assumption 4.1 again and the fact that ν is a probability measure that*

$$\rho(x, \alpha) = \liminf_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\int_0^T r(X_s^{x,\alpha}, \alpha_s) ds \right],$$

$$J_\lambda(x, \alpha) = \mathbb{E}\left[\int_0^\infty e^{-\lambda s} r(X_s^{x,\alpha}, \alpha_s) ds \right], \text{ and}$$

$$J_T(t, x, \alpha) = \mathbb{E}\left[\int_t^T r(X_s^{t,x,\alpha}, \alpha_s) ds \right].$$

For the same reason, we could consider expected gains of the more general form

$$\frac{1}{\eta T} \mathbb{E}\left[\int_0^T \int_{\mathbb{R}^{d'}} \tilde{r}(X_{s-}^{x,\alpha}, \alpha_s, e) N(\mathrm{d}e, \mathrm{d}s) \right] = \frac{1}{T} \mathbb{E}\left[\int_0^T \int_{\mathbb{R}^{d'}} \tilde{r}(X_s^{x,\alpha}, \alpha_s, e) \nu(\mathrm{d}e) \mathrm{d}s \right]$$

upon replacing r by $(x, a) \in \mathbb{R}^d \times \mathbb{A} \mapsto \int_{\mathbb{R}^{d'}} \tilde{r}(x, a, e) \nu(\mathrm{d}e)$.

Theorem 4.2.1.
Let Assumptions 4.1 to 4.3 hold. Then, there exists sequences $(\lambda_n)_{n \geq 1}$ going to 0 and $(T_n)_{n \geq 1}$ going to $+\infty$ such that the sequences $(\lambda_n V_{\lambda_n})_{n \geq 1}$ and $(T_n^{-1} V_{T_n}(0, \cdot))_{n \geq 1}$ converge uniformly on compact sets to $\rho^*(0)$, and such that $(V_{\lambda_n} - V_{\lambda_n}(0))_{n \geq 1}$ converges uniformly on compact sets to a function

$w \in \mathcal{C}^{0,1}$ which solves

$$\rho^* = \sup_{a\in\mathbb{A}}\left\{\eta\int_{\mathbb{R}^{d'}}[w(\cdot + b(\cdot,a,e)) - w]\,\nu(de) + r(\cdot,a)\right\},\ \text{ on } \mathbb{R}^d. \quad (4.10)$$

Moreover, ρ^* is constant over $\mathbb{R}^d$, and, if $(\tilde{w},\tilde{\rho}) \in \mathcal{C}^0_{\text{lin}} \times \mathbb{R}$ solves the ergodic Hamilton-Jacobi-Bellman equation

$$\tilde{\rho} = \sup_{a\in\mathbb{A}}\left\{\eta\int_{\mathbb{R}^{d'}}[\tilde{w}(\cdot + b(\cdot,a,e)) - \tilde{w}]\nu(\mathrm{d}e) + r(\cdot,a)\right\},\ \text{ on } \mathbb{R}^d, \quad (4.11)$$

then $\tilde{\rho} = \rho^*$.

***Remark* 4.2.2.** *As a by-product of Theorem 4.2.1 and the first part of the proof of Lemma 4.A.4, for all $x \in \mathbb{R}^d$, there exists an optimal Markov control defined by $\hat{\alpha} := \hat{a}(X^{x,\hat{\alpha}}_{\cdot-})$ in which $\hat{a}$ is a measurable map satisfying*

$$\eta\int_{\mathbb{R}^{d'}} w(\cdot + b(\cdot,\hat{a}(\cdot),e))\nu(\mathrm{d}e) + r(\cdot,\hat{a}(\cdot)) = \max_{a\in\mathbb{A}}\left\{\eta\int_{\mathbb{R}^{d'}} w(\cdot + b(\cdot,a,e))\nu(\mathrm{d}e) + r(\cdot,a)\right\}$$

on $\mathbb{R}^d$. Moreover,

$$\rho^* = \lim_{T\to\infty}\frac{1}{\eta T}\mathbb{E}\left[\int_0^T r(X^{x,\hat{\alpha}}_{s-},\hat{\alpha}_s)\mathrm{d}N_s\right].$$

4.3 Approximation for Models with Large Activity

Given an $\varepsilon \in (0,1)$, we now replace η by

$$\eta_\varepsilon := \varepsilon^{-1}.$$

In the following, we omit the dependence of N and $X^{x,\alpha}$ on ε for ease of notations and set

$$\rho^*_\varepsilon := \sup_{\alpha\in\mathcal{A}}\liminf_{T\to\infty}\frac{1}{\eta_\varepsilon T}\mathbb{E}\left[\int_0^T r(X^{0,\alpha}_{t-},\alpha_t)\mathrm{d}N_t\right].$$

We shall see that ρ^*_ε, together with the associated optimal policy, can be approximated by considering its diffusive limit as $\varepsilon \to 0$, upon assuming that the jump coefficient $b := b_\varepsilon$ introduced in Section 4.2 is of the form

$$b_\varepsilon = \varepsilon b_1 + \sqrt{\varepsilon}b_2,$$

and making the following assumption.

Assumption 4.4.
We have $b = \varepsilon b_1 + \sqrt{\varepsilon} b_2$ for some continuous functions $b_1 : \mathbb{R}^d \times \mathbb{A} \times \mathbb{R}^{d'} \mapsto \mathbb{R}^d$ and $b_2 : \mathbb{R}^d \times \mathbb{R}^{d'} \mapsto \mathbb{R}^d$ such that:

(i) There exists $L_{b_1,b_2} > 0$ such that

$$[b_1(\cdot, a, e)]_{\mathcal{C}^0_{\text{lin}}} + \|b_2(\cdot, e)\|_{\mathcal{C}^0_b} \leq L_{b_1,b_2} \text{ for all } (a, e) \in \mathbb{A} \times \mathbb{R}^{d'}.$$

(ii) There exists $\varsigma > 0$ such that

$$\int_{\mathbb{R}^{d'}} b_2(\cdot, e)\nu(\mathrm{d}e) = 0 \text{ and } \int_{\mathbb{R}^{d'}} b_2(\cdot, e) b_2(\cdot, e)^\top \nu(\mathrm{d}e) \geq \varsigma \, \mathrm{I}_d \text{ on } \mathbb{R}^d,$$

in which I_d is the identity matrix.

(iii) The map

$$(x, a) \in \mathbb{R}^d \times \mathbb{A} \mapsto \mu(x, a) := \int_{\mathbb{R}^{d'}} b_1(x, a, e)\nu(\mathrm{d}e)$$

is Lipschitz in x uniformly in a, and there exists a Lipschitz $\mathbb{R}^{d\times d}$-valued function σ defined on $\mathbb{R}^d$ such that

$$\sigma\sigma^\top = \int_{\mathbb{R}^{d'}} b_2(\cdot, e) b_2(\cdot, e)^\top \nu(\mathrm{d}e).$$

(iv) The estimates of Assumptions 4.1 to 4.3 hold for each $(\eta_\varepsilon, b_\varepsilon, r)$ in place of (η, b, r), uniformly in $\varepsilon > 0$.

***Example* 4.3.1.** *Consider the context of Example 4.2.1 in which $\eta = \varepsilon^{-1}$ and*

$$b_\varepsilon(x, a, e) = e_1 \left(\varepsilon(e_2 a - x) + \varepsilon^{\frac{1}{2}} e_3 \right), \ (x, a, e) \in \mathbb{R}^d \times \mathbb{A} \times \mathbb{R}^4$$

with ν as in Example 4.2.1 such that in addition $\int_{\mathbb{R}^4} e_1 e_3 \nu(\mathrm{d}e) = 0$. In this context, we obtain $\mu(x, a) = n_2 a - n_1 x$, with $n_1 := \int_{\mathbb{R}^4} e_1 \nu(de)$ and $n_2 := \int_{\mathbb{R}^4} e_1 e_2 \nu(\mathrm{d}e)$, and $\sigma(x)^2 = \int_{\mathbb{R}^4} |e_1 e_3|^2 \nu(\mathrm{d}e)$.
Assume that $n_1 > 0$. Using a second order Taylor expansion around $\varepsilon = 0$, one easily checks that Assumption 4.3 holds with $\xi(x) = \|x\|^{2p}$, $p \geq 1$, for some C^1_ξ and C^2_ξ that do not depend on $\varepsilon > 0$. Similarly, Assumption 4.2 holds with $\zeta(x, x') = \|x - x'\|^{2p}$, $p \geq 1$, for some $C_\zeta > 0$, uniformly in $\varepsilon \in (0, \varepsilon_\circ)$, for some $\varepsilon_\circ > 0$ small enough.

4.3.1 Candidate diffusion limit

Let $\bar{\mathbb{P}}$ be a probability measure on $\mathbb{D}$ and let W be a stochastic process such that W is a $\bar{\mathbb{P}}$-Brownian motion, let $\bar{\mathbb{F}} = (\bar{\mathcal{F}}_s)_{s\geq 0}$ be the $\bar{\mathbb{P}}$-augmentation of the filtration generated by W, and let $\bar{\mathcal{A}}$ be the collection of $\bar{\mathbb{F}}$-predictable processes. Given $\bar{\alpha} \in \bar{\mathcal{A}}$, we can then define $\bar{X}^{x,\bar{\alpha}}$ as the unique strong solution (see [114, Thm. 1]) of

$$\bar{X}^{x,\bar{\alpha}} = x + \int_0^\cdot \mu(\bar{X}_s^{x,\bar{\alpha}}, \bar{\alpha}_s)\mathrm{d}s + \int_0^\cdot \sigma(\bar{X}_s^{x,\bar{\alpha}})\mathrm{d}W_s\,. \tag{4.12}$$

The corresponding ergodic control problem is defined by

$$\bar{\rho}^*(x) := \sup_{\bar{\alpha}\in\bar{\mathcal{A}}} \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T r(\bar{X}_s^{x,\bar{\alpha}}, \bar{\alpha}_s)\mathrm{d}s\right],\ x \in \mathbb{R}^d.$$

As in Section 4.2, we define for $\lambda > 0$ and $x \in \mathbb{R}^d$

$$\bar{V}_\lambda(x) := \sup_{\bar{\alpha}\in\bar{\mathcal{A}}} \bar{J}_\lambda(x,\bar{\alpha}) \text{ with } \bar{J}_\lambda(x,\bar{\alpha}) := \mathbb{E}\left[\int_0^\infty e^{-\lambda s} r(\bar{X}_s^{x,\bar{\alpha}}, \bar{\alpha}_s)\mathrm{d}s\right],$$

and we impose conditions corresponding to the estimates obtained in Lemmas 4.A.1 and 4.A.2.

Assumption 4.5.
There exists $L_{\bar{V}}, C_{\bar{X}} > 0$ and $p_{\bar{X}} \geq 1$ such that:

(i) For all $x, x' \in \mathbb{R}^d$ and $\lambda \in (0,1)$,

$$|\bar{V}_\lambda(x) - \bar{V}_\lambda(x')| \leq L_{\bar{V}} \|x - x'\|$$

(ii) For all $x \in \mathbb{R}^d$ and $\bar{\alpha} \in \bar{\mathcal{A}}$,

$$\mathbb{E}\left[\|\bar{X}_t^{x,\bar{\alpha}}\|^{p_{\bar{X}}}\right] \leq C_{\bar{X}}\left\{e^{-\frac{t}{C_{\bar{X}}}}\|x\|^{p_{\bar{X}}} + 1\right\},\ t \geq 0.$$

***Remark* 4.3.1.**

a) *The condition* (i) *of Assumption 4.5 holds for instance under [13, Assumption 7.3.1]. Indeed, the latter implies a similar bound as* (4.50)*, see [13, Lemma 7.3.4], and the estimate of* (i) *then follows from the same arguments as in the proof of Lemma 4.A.1. More generally, it suffices to find a family of* $\mathcal{C}^2(\mathbb{R}^d \times$

$\mathbb{R}^d;\mathbb{R})$*-functions* $(\bar{\zeta}_\iota)_{\iota>0}$ *that is locally bounded, satisfies*

$$\mathrm{D}\bar{\zeta}_\iota(x,x')\begin{pmatrix}\mu(x,a)\\ \mu(x',a)\end{pmatrix}+\frac{1}{2}\operatorname{Tr}\left[\Sigma(x,x')\mathrm{D}^2\bar{\zeta}_\iota(x,x')\right]\le -C_{\bar{\zeta}}\bar{\zeta}_\iota(x,x')+\varrho_\iota\,, \tag{4.13}$$

for any $x,x'\in\mathbb{R}^d$, $a\in\mathbb{A}$, *and* $\iota>0$, *and in which* $C_{\bar{\zeta}}>0$, $\lim_{\iota\to 0}\varrho_\iota=0$ *and*

$$\Sigma(x,x'):=\begin{pmatrix}\sigma(x)\\ \sigma(x')\end{pmatrix}\begin{pmatrix}\sigma(x)\\ \sigma(x')\end{pmatrix}^{\mathsf{T}},$$

and such that $(\bar{\zeta}_\iota)_{\iota>0}$ *converges pointwise as* $\iota\to 0$ *to a map* $\bar{\zeta}:\mathbb{R}^d\times\mathbb{R}^d\mapsto\mathbb{R}$ *satisfying*

$$\frac{1}{C_{\bar{\zeta}}}\|x-x'\|^{p_{\bar{\zeta}}}\le\bar{\zeta}(x,x')\le C_{\bar{\zeta}}\|x-x'\|^{p_{\bar{\zeta}}}\,,\ \textit{for all } x,x'\in\mathbb{R}^d,$$

for some $p_{\bar{\zeta}}\ge 1$. *This follows from the arguments that are used in the proof of Lemma 4.A.1 upon first applying Itô's lemma to* $\bar{\zeta}_\iota$ *and then sending* $\iota\to 0$ *to deduce the counterpart of* (4.49) *before using the inequalities just above.*

b) *The condition* (ii) *of Assumption 4.5 holds for instance if we can find a smooth function* $\bar{\xi}$ *and constants* $C^1_{\bar{\xi}}>0$ *and* $C^2_{\bar{\xi}}$ *such that*

$$\mathrm{D}\bar{\xi}(x)\mu(x,a)+\frac{1}{2}\operatorname{Tr}\left[\sigma\sigma^{\mathsf{T}}(x)\mathrm{D}^2\bar{\xi}(x)\right]\le -C^1_{\bar{\xi}}\bar{\xi}(x)+C^2_{\bar{\xi}}, \tag{4.14}$$

and

$$\frac{1}{C^2_{\bar{\xi}}}\|x\|^{p_{\bar{\xi}}}\le\bar{\xi}(x)\le C^2_{\bar{\xi}}|x|^{p_{\bar{\xi}}}, \tag{4.15}$$

for all $x\in\mathbb{R}^d$, *for some* $p_{\bar{\xi}}\ge 1$. *This follows from the same arguments as in the proof of Lemma 4.A.2, taking* $p_{\bar{\xi}}=p_{\bar{X}}$. *As in* a) *above, it suffices that* (4.14) *holds for a sequence of approximating smooth functions. In particular, condition* (ii) *of Assumption 4.5 holds under [13, Assumption 7.3.1], see [13, Lemma 7.6.3].*

***Example* 4.3.2.** *Consider the context of Example 4.3.1 with* σ *constant, then it satisfies [13, Assumption 7.3.1], and therefore Assumption 4.5, by [13, Example 7.3.3].*

In order to state the counterpart of Theorem 4.2.1 for the diffusive limit ergodic control problem, we also define, for $T>0$, $t\le T$ and $x\in\mathbb{R}^d$,

$$\bar{V}_T(t,x):=\sup_{\bar{\alpha}\in\bar{\mathcal{A}}}\bar{J}_T(t,x,\bar{\alpha})\ \text{ with }\ \bar{J}_T(t,x,\bar{\alpha}):=\mathbb{E}\left[\int_t^T r(\bar{X}^{t,x,\bar{\alpha}}_s,\bar{\alpha}_s)\mathrm{d}s\right],$$

and set

$$\bar{\mathcal{L}}^{\bar{a}}\varphi=\mathrm{D}\varphi^{\mathsf{T}}\mu(\cdot,\bar{a})+\frac{1}{2}\operatorname{Tr}[\sigma\sigma^{\mathsf{T}}\mathrm{D}^2\varphi],\ \bar{a}\in\mathbb{A},$$

for a smooth function $\varphi:\mathbb{R}^d\to\mathbb{R}$.

Theorem 4.3.1.
Let Assumptions 4.4 and 4.5 hold. Then, there exists sequences $(\lambda_n)_{n\geq 1}$ going to 0 and $(T_n)_{n\geq 1}$ going to $+\infty$ such that the sequences $(\lambda_n \bar{V}_{\lambda_n})_{n\geq 1}$ and $(T_n^{-1}\bar{V}_{T_n}(0,\cdot))_{n\geq 1}$ converge uniformly on compact sets to $\bar{\rho}^*(0)$, and such that $(\bar{V}_{\lambda_n} - \bar{V}_{\lambda_n}(0))_{n\geq 1}$ converges uniformly on compact sets to a function $\bar{w} \in \mathcal{C}^2 \cap \mathcal{C}^0_{\text{lin}}$ that satisfies

$$\bar{\rho}^* = \sup_{\bar{a}\in\mathbb{A}} \left\{ \bar{\mathcal{L}}^{\bar{a}} \bar{w} + r(\cdot, \bar{a}) \right\}, \text{ on } \mathbb{R}^d, \tag{4.16}$$

and

$$[\bar{w}]^1_{\mathcal{C}^0} \leq L^\gamma_{\bar{w}} \text{ and } \|\bar{w}\|_{\mathcal{C}^{2,\gamma}_b(B_1(x))} \leq L^\gamma_{\bar{w}}(1 + \|x\|), \quad \text{for all } x \in \mathbb{R}^d, \tag{4.17}$$

for some $L^\gamma_{\bar{w}} > 0$, for all $\gamma \in (0,1)$.

Moreover, $\bar{\rho}^*$ is constant over $\mathbb{R}^d$, and, if $(\tilde{w}, \tilde{\rho}) \in (\mathcal{C}^2 \cap \mathcal{C}^0_{\text{lin}}) \times \mathbb{R}$ solves the ergodic equation

$$\tilde{\rho} = \sup_{\bar{a}\in\mathbb{A}} \left\{ \bar{\mathcal{L}}^{\bar{a}} \tilde{w} + r(\cdot, \bar{a}) \right\}, \quad \text{on } \mathbb{R}^d, \tag{4.18}$$

then $\tilde{\rho} = \bar{\rho}^*$.

Proof. The proof is exactly the same as the one of Theorem 4.2.1 upon replacing the estimates of Lemmas 4.A.1 and 4.A.2 by the ones of Assumption 4.5. See Section 4.A. The only significant difference is that we have to show the estimate (4.17).

1. The fact that, for an appropriate sequence $(\lambda_n)_{n\geq 0}$ that converges to 0, $\lambda_n \bar{V}_{\lambda_n}(0) \to c \in \mathbb{R}$ and $\bar{V}_{\lambda_n} - \bar{V}_{\lambda_n}(0) \to \bar{w}$ uniformly on compact sets for some $\bar{w} \in \mathcal{C}^{0,1}$ follows from Assumption 4.5 and the same arguments as in the first part of the proof of Lemma 4.A.3 below.

2. We now argue as in the proof of [13, Thm. 3.5.6]. Fix $n \geq 1$, let $\bar{\tau}_n^{x,\bar{\alpha}}$ be the first exit time of $\bar{X}^{x,\bar{\alpha}}$ from $B_n(0)$, for $(x, \bar{\alpha}) \in \mathbb{R}^d \times \bar{\mathcal{A}}$, and set

$$\bar{V}^n_\lambda(x) := \sup_{\bar{\alpha}\in\bar{\mathcal{A}}} \mathbb{E}\left[\int_0^{\bar{\tau}_n^{x,\bar{\alpha}}} e^{-\lambda s} r(\bar{X}^{x,\bar{\alpha}}_s, \bar{\alpha}_s) \mathrm{d}s \right].$$

Then, $\bar{V}^n_\lambda \in \mathcal{C}^2(B_n(0))$ by the arguments in the proof of [13, Thm. 3.5.6]. Moreover, Assumption 4.5 and the linear growth of r (recall that it is

assumed Lipschitz) imply that

$$\sup_{n\geq 1} [\bar{V}_\lambda^n]_{\mathcal{C}^0_{\text{lin}}} \leq C_\lambda$$

for some $C_\lambda > 0$. Then, arguing as in the proof of [13, Thm. 3.5.6], we obtain that, for all $\lambda > 0$, $(\bar{V}_\lambda^n)_{n\geq 1}$ converges as $n \to \infty$ to a map $\psi_\lambda \in \mathcal{C}^2$ that solves

$$\lambda\psi_\lambda = \sup_{\bar{a}\in\mathbb{A}} \left\{ \bar{\mathcal{L}}^{\bar{a}}\psi_\lambda + r(\cdot,\bar{a}) \right\}, \quad \text{on } \mathbb{R}^d,$$

and has at most linear growth. Using this linear growth property, Assumption 4.5 and a verification argument, we deduce that $\psi_\lambda = \bar{V}_\lambda$.

Since $\bar{V}_\lambda \in \mathcal{C}^{0,1}$ by Assumption 4.5, it follows from Assumption 4.4 and Lemma 4.B.2 that, given $\gamma \in (0,1)$, $\bar{V}_\lambda \in \mathcal{C}^{2,\gamma}$ and that there is $K > 0$ (depending on γ but not on $\lambda \in (0,1)$) such that

$$\|\Delta\bar{V}_\lambda\|_{\mathcal{C}_b^{2,\gamma}(B_1(x))} \leq K(1+|x|), \quad \text{for all } (x,\lambda) \in \mathbb{R}^d \times (0,1), \tag{4.19}$$

where $\Delta\bar{V}_\lambda := \bar{V}_\lambda - \bar{V}_\lambda(0)$ solves

$$\lambda\bar{V}_\lambda(0) + \lambda\Delta\bar{V}_\lambda = \sup_{\bar{a}\in\mathbb{A}} \left\{ \bar{\mathcal{L}}^{\bar{a}}\Delta\bar{V}_\lambda + r(\cdot,\bar{a}) \right\}, \quad \text{on } \mathbb{R}^d.$$

Let $(\lambda_n)_{n\geq 0}$ be as in step 1. Passing to the limit in the above leads to (4.16), with c defined in step 1. in place of $\bar{\rho}^*$, and to (4.17).

3. By the same arguments as in Lemma 4.A.4, if $(\tilde{w},\tilde{\rho}) \in (\mathcal{C}^2 \cap \mathcal{C}^0_{\text{lin}}) \times \mathbb{R}$ solves (4.11) then $\tilde{\rho} = \bar{\rho}^*$. In particular, $\bar{\rho}^*$ is constant and $c = \bar{\rho}^*$ by step 2.

4. The existence of $(T_n)_{n\geq 1}$ going to $+\infty$ such that $(T_n^{-1}\bar{V}_{T_n}(0,\cdot))_{n\geq 1}$ converge uniformly on compact sets to $\bar{\rho}^*(0)$ then follows from the same arguments as in Lemma 4.A.5. □

4.3.2 First-order approximation guarantees

We can now turn to the main part of this chapter and quantify the approximation error due to passing to the diffusive limit in the original pure jump problem. We will show below that it is controlled by the Hölder regularity of $D^2\bar{w}$, namely that the approximation error is of the order of $\varepsilon^{\gamma/2}$ for all $\gamma \in (0,1)$. In Section 4.3.3, we will see that it can be improved by considering appropriate correction terms.

The cornerstone of the analysis is the residual term of a second-order Taylor expansion of $\bar{w}$ performed on the Dynkin operator of the pure-jump process (4.3), namely:

$$\delta r_\varepsilon(x,a) := \frac{1}{\varepsilon}\int_{\mathbb{R}^{d'}} [\bar{w}(x + b_\varepsilon(x,a,e)) - \bar{w}(x)]\,\nu(de) \tag{4.20}$$

$$-D\bar{w}(x)^\top \mu(x,a) - \frac{1}{2}\operatorname{Tr}\left[\sigma\sigma^\top(x) D^2\bar{w}(x)\right], \tag{4.21}$$

defined for $(x,a) \in \mathbb{R}^d \times \mathbb{A}$. The function δr_ε measures the error of the diffusion approximation explicitly in terms of the control problem, and thus will be shown to effectively control the error in all quantities of interest. Leveraging the regularity results in (4.17), the Hölder regularity of $D^2\bar{w}$ yields Proposition 4.3.1, which in turn yields Theorem 4.3.2.

Proposition 4.3.1.
Let Assumptions 4.4 and 4.5 hold with $p_\xi \geq 3$. Fix $\gamma \in (0,1)$. Then, there exists $L^{\gamma,1}_{\delta r}, L^{\gamma,2}_{\delta r} > 0$ such that, for each $0 < \varepsilon \leq \varepsilon_\circ := (L_{b_1,b_2})^{-2}$ and $(x,a) \in \mathbb{R}^d \times \mathbb{A}$,

$$\|\delta r_\varepsilon(x,a)\| \leq \varepsilon^{\frac{\gamma}{2}} L^{\gamma,1}_{\delta r}(1 + \|x\|^3), \tag{4.22}$$

and

$$\sup_{t\geq 0}\sup_{\alpha\in\mathcal{A}} \mathbb{E}\left[\|\delta r_\varepsilon(X^{x,\alpha}_t, \alpha_t)\|\right] \leq \varepsilon^{\frac{\gamma}{2}} L^{\gamma,2}_{\delta r}(1 + \|x\|^3)\,. \tag{4.23}$$

Proof.

1. We first prove the estimate (4.22) using (4.17). Namely,

$$\begin{aligned}\bar{w}(x + b_\varepsilon(x,a,e)) - \bar{w}(x) = \bar{w}(x + \varepsilon b_1(x,a,e) + \varepsilon^{\frac{1}{2}} b_2(x,e)) \\ - \bar{w}(x + \varepsilon^{\frac{1}{2}} b_2(x,e)) \\ + \bar{w}(x + \varepsilon^{\frac{1}{2}} b_2(x,e)) - \bar{w}(x)\end{aligned}$$

where

$$\begin{aligned}\bar{w}(x + \varepsilon b_1(x,a,e) + \varepsilon^{\frac{1}{2}} b_2(x,e)) - \bar{w}(x + \varepsilon^{\frac{1}{2}} b_2(x,e)) \\ = \varepsilon D\bar{w}(x + \varepsilon^{\frac{1}{2}} b_2(x,e)) b_1(x,a,e)\end{aligned}$$

$$+\int_0^1 \frac{\varepsilon^2}{2} b_1(x,a,e)^\top \mathrm{D}^2\bar{\mathrm{w}}(\hat{x}_1^\varepsilon(u))b_1(x,a,e)\mathrm{d}u$$

in which

$$\hat{x}_1^\varepsilon(u) := x + \varepsilon^{\frac{1}{2}} b_2(x,e) + u\varepsilon b_1(x,a,e)$$

is such that

$$\sup_{u\in[0,1]} |\hat{x}_1^\varepsilon(u)| \le |x| + \varepsilon^{\frac{1}{2}} L_{b_1,b_2} + \varepsilon L_{b_1,b_2}(1+|x|),$$

by definition of L_{b_1,b_2} in Assumption 4.4. By (4.17) and Assumption 4.4, this implies that

$$\begin{aligned}
&\frac{\varepsilon^2}{2}\left|b_1(x,a,e)^\top \mathrm{D}^2\bar{\mathrm{w}}(\hat{x}_1^\varepsilon(u))b_1(x,a,e)\right| \\
&\quad\le \frac{\varepsilon^2}{2}\big(L_{b_1,b_2}\big)^2(1+\|x\|)^2 L_{\bar{\mathrm{w}}}^\gamma\left(1+\|x\|+\varepsilon^{\frac{1}{2}}L_{b_1,b_2}+\varepsilon L_{b_1,b_2}(1+\|x\|)\right).
\end{aligned}$$

Moreover, since $\varepsilon^{1/2}L_{b_1,b_2}\le 1$, we have

$$\left\|\mathrm{D}\bar{\mathrm{w}}(x+\varepsilon^{\frac{1}{2}}b_2(x,e)) - \mathrm{D}\bar{\mathrm{w}}(x)\right\| \le L_{\bar{\mathrm{w}}}^\gamma(1+\|x\|)\varepsilon^{\frac{1}{2}}L_{b_1,b_2}$$

by (i) of Assumption 4.4 and (4.17). Using (ii) of Assumption 4.4, we next obtain that

$$\begin{aligned}
&\int_{\mathbb{R}^{d'}}\left[\bar{\mathrm{w}}(x+\varepsilon^{\frac{1}{2}}b_2(x,e)) - \bar{\mathrm{w}}(x)\right]\nu(\mathrm{d}e) \\
&\qquad = \int_{\mathbb{R}^{d'}}\int_0^1 \frac{\varepsilon}{2} b_2(x,e)^\top \mathrm{D}^2\bar{\mathrm{w}}(\hat{x}_2^\varepsilon(u,e))b_2(x,e)\mathrm{d}u\,\nu(\mathrm{d}e)
\end{aligned}$$

in which

$$\hat{x}_2^\varepsilon(u,e) := x + u\varepsilon^{\frac{1}{2}}b_2(x,e) \in B_1(x)$$

since $\varepsilon^{1/2}L_{b_1,b_2}\le 1$ by assumption and (i) of Assumption 4.4. Then, by (4.17) again and (iii) of Assumption 4.4

$$\begin{aligned}
&\left|\int_{\mathbb{R}^{d'}}\left[\bar{\mathrm{w}}\left(x+\varepsilon^{\frac{1}{2}}b_2(x,e)\right) - \bar{\mathrm{w}}(x)\right]\nu(\mathrm{d}e) - \frac{\varepsilon}{2}\mathrm{Tr}[\sigma\sigma^\top(x)\mathrm{D}^2\bar{\mathrm{w}}(x)]\right| \\
&= \left|\int_{\mathbb{R}^{d'}}\left[\bar{\mathrm{w}}(x+\varepsilon^{\frac{1}{2}}b_2(x,e)) - \bar{\mathrm{w}}(x)\right]\nu(\mathrm{d}e)\right. \\
&\qquad\qquad \left. -\frac{\varepsilon}{2}\int_{\mathbb{R}^{d'}} b_2(x,e)^\top \mathrm{D}^2\bar{\mathrm{w}}(x)b_2(x,e)\nu(\mathrm{d}e)\right| \\
&\le \frac{\varepsilon}{2}\big(L_{b_1,b_2}\big)^2 L_{\bar{\mathrm{w}}}^\gamma(1+\|x\|)\left(\varepsilon^{\frac{1}{2}}L_{b_1,b_2}\right)^\gamma.
\end{aligned}$$

The estimate (4.22) is obtained by combining the above.

2. The estimate (4.23) follows from (4.22), Lemma 4.A.2, and the fact that $p_\xi \geq 3$. □

We are now in a position to state the main result of this section.

Theorem 4.3.2.
Let Assumptions 4.4 and 4.5 hold with $p_\xi \geq 3$. Then, for all $\gamma \in (0,1)$, there exists $L^\gamma_{\delta\rho} > 0$ such that

$$\left|\bar\rho^* - \rho^*_\varepsilon\right| \leq \varepsilon^{\frac{\gamma}{2}} L^\gamma_{\delta\rho} \text{ for all } \varepsilon \in (0,1).$$

Moreover, there exists a measurable map $\hat a : \mathbb{R}^d \mapsto \mathbb{A}$ such that

$$\bar{\mathcal{L}}^{\hat a}\bar w + r(\cdot,\hat a) = \sup_{\bar a \in \mathbb{A}} \left\{\bar{\mathcal{L}}^{\bar a}\bar w + r(\cdot,\bar a)\right\}, \text{ on } \mathbb{R}^d$$

and

$$\rho^*_\varepsilon - \varepsilon^{\frac{\gamma}{2}} L^\gamma_{\delta\rho} \leq \liminf_{T\to\infty} \frac{1}{\eta_\varepsilon T}\mathbb{E}\left[\int_0^T r(X^{\hat a}_{s-}, \hat a(X^{\hat a}_{s-}))\mathrm{d}N_s\right], \text{ for all } \varepsilon \in (0,1),$$

in which $X^{\hat a}$ solves

$$X^{\hat a}_\cdot = \int_0^\cdot \int_{\mathbb{R}^{d'}} b_\varepsilon(X^{\hat a}_{s-}, \hat a(X^{\hat a}_{s-}), e) N(\mathrm{d}e, \mathrm{d}s)\,.$$

Proof. Fix $\gamma \in (0,1)$. Hereafter, we denote by w_ε the function w introduced in Theorem 4.2.1 for $\eta = \eta_\varepsilon = \varepsilon^{-1}$. By Theorems 4.2.1 and 4.3.1, $\Delta^\varepsilon := \bar w - w_\varepsilon$ solves

$$\bar\rho^* - \rho^*_\varepsilon \leq \sup_{a\in\mathbb{A}}\left\{\frac{1}{\varepsilon}\int_{\mathbb{R}^{d'}} [\Delta^\varepsilon(\cdot + b_\varepsilon(\cdot,a,e)) - \Delta^\varepsilon]\,\nu(\mathrm{d}e) - \delta r_\varepsilon(\cdot,a)\right\}, \text{ on } \mathbb{R}^d.$$

By the same arguments as in the proof of Lemma 4.A.4, (4.23) applied with $x = 0$, (4.17), (4.51), and Lemma 4.A.2, we deduce that

$$\bar\rho^* - \rho^*_\varepsilon \leq L^1_{\delta\rho}\varepsilon^{\frac{\gamma}{2}}$$

for some $L^1_{\delta\rho} > 0$ that does not depend on $\varepsilon \in (0,1)$. Replacing Δ^ε by $-\Delta^\varepsilon$ in this argument implies that

$$\rho^*_\varepsilon - \bar\rho \leq L^2_{\delta\rho}\varepsilon^{\frac{\gamma}{2}}$$

for some $L^2_{\delta\rho} > 0$ that does not depend on $\varepsilon \in (0,1)$.

The second assertion of the theorem is then proved by following the arguments in the first part of the proof of Lemma 4.A.4 and using the above. □

4.3.3 Higher-order expansions

Under additional conditions, one can exhibit a first-order correction term to improve the convergence speed in Theorem 4.3.2. It is in the spirit of the correction term introduced in Section 3.3.5 but is formulated differently. In particular, the function $\delta\bar{w}_\varepsilon$ introduced below depends on ε and the optimization in (4.24) is performed over the whole set $\mathbb{A}$. This approach can be iterated to higher order correction terms in an obvious manner, upon additional regularity conditions, without considering a coupled system of PDEs as in Section 3.3.6.

From now on, we assume the following.

Assumption 4.6.
There exists $\gamma_\circ \in (0,\gamma]$ and $(\delta\gamma, \delta C) \in (0,1) \times \mathbb{R}$ such that, for each $\varepsilon \in (0,1)$, we can find $\delta\bar{\rho}^*_\varepsilon \in \mathbb{R}$ and $\delta\bar{w}_\varepsilon \in \mathcal{C}^0_{\mathrm{lin}}$ satisfying $\|\delta\bar{w}_\varepsilon\|_{\mathcal{C}_b^{2,\delta\gamma}(B_1(x))} \leq \delta C(1+\|x\|)$ for all $x \in \mathbb{R}^d$ and

$$\delta\bar{\rho}^*_\varepsilon = \sup_{\bar{a}\in\mathbb{A}} \left\{ \bar{\mathcal{L}}^{\bar{a}} \delta\bar{w}_\varepsilon + \varepsilon^{-\frac{\gamma_\circ}{2}} [\delta r_\varepsilon + r](\cdot,\bar{a}) \right\} \text{ on } \mathbb{R}^d, \tag{4.24}$$

in which

$$f(\cdot,\bar{a}) := \bar{\mathcal{L}}^{\bar{a}} \bar{w} + r(\cdot,\bar{a}) - \bar{\rho}^*.$$

Theorem 4.3.3.
Let the conditions of Theorem 4.3.2 and Assumption 4.6 hold. Assume further that $p_{\bar{X}} \geq 3$. Then,

$$\limsup_{\varepsilon\downarrow 0} \|\delta\bar{\rho}^*_\varepsilon\| < \infty$$

and

$$\bar{\rho}^{*(1)}_\varepsilon := \bar{\rho}^* + \varepsilon^{\frac{\gamma_\circ}{2}} \delta\bar{\rho}^*_\varepsilon, \ \varepsilon \in (0,1),$$

satisfies

$$\limsup_{\varepsilon\downarrow 0} \varepsilon^{-\frac{\gamma_\circ+\delta\gamma}{2}} |\rho^*_\varepsilon - \bar{\rho}^{*(1)}_\varepsilon| < \infty.$$

Moreover, for each $\varepsilon \in (0,1)$, there exists a measurable map $\hat{a}_\varepsilon : \mathbb{R}^d \mapsto \mathbb{A}$ such that

$$\bar{\mathcal{L}}^{\hat{a}_\varepsilon}\delta\bar{w}_\varepsilon + \varepsilon^{-\frac{\gamma_\circ}{2}}[\delta r_\varepsilon + f](\cdot,\hat{a}_\varepsilon) = \sup_{\bar{a}\in\mathbb{A}}\left\{\bar{\mathcal{L}}^{\bar{a}}\delta\bar{w}_\varepsilon + \varepsilon^{-\frac{\gamma_\circ}{2}}[\delta r_\varepsilon + f](\cdot,\bar{a})\right\} \text{ on } \mathbb{R}^d$$

and

$$\limsup_{\varepsilon\downarrow 0}\varepsilon^{-\frac{\gamma_\circ+\delta\gamma}{2}}|\rho^*_\varepsilon - \rho_\varepsilon(0,\hat{a}_\varepsilon(X^{\hat{a}_\varepsilon}_{\cdot-}))| < \infty,$$

in which $X^{\hat{a}_\varepsilon}$ solves

$$X^{\hat{a}_\varepsilon} = \int_0^\cdot\int_{\mathbb{R}^{d'}} b_\varepsilon(X^{\hat{a}_\varepsilon}_{s-},\hat{a}_\varepsilon(X^{\hat{a}_\varepsilon}_{s-}),e)N(\mathrm{d}e,\mathrm{d}s)\,.$$

Proof. It follows from the same arguments as in Lemma 4.A.4 and the fact that $f \le 0$ (by (4.16)) that

$$\begin{aligned}\delta\bar{\rho}^*_\varepsilon &= \sup_{\bar{\alpha}\in\bar{\mathcal{A}}}\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\int_0^T \varepsilon^{-\frac{\gamma_\circ}{2}}[\delta r_\varepsilon + f](\bar{X}^{0,\bar{\alpha}}_s,\bar{\alpha}_s)\mathrm{d}s\right]\\ &\le \sup_{\bar{\alpha}\in\bar{\mathcal{A}}}\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\int_0^T \varepsilon^{-\frac{\gamma_\circ}{2}}\delta r_\varepsilon(\bar{X}^{0,\bar{\alpha}}_s,\bar{\alpha}_s)\mathrm{d}s\right].\end{aligned}$$

Let $\hat{a}$ be as in Theorem 4.3.2. Then, $f(\cdot,\hat{a}) = 0$ by (4.16). Hence,

$$\delta\bar{\rho}^*_\varepsilon \ge \liminf_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\int_0^T \varepsilon^{-\frac{\gamma_\circ}{2}}\delta r_\varepsilon(\bar{X}^{\hat{a}}_s,\hat{a}(\bar{X}^{\hat{a}}_s))\mathrm{d}s\right]$$

in which $\bar{X}^{\hat{a}}$ solves

$$\bar{X}^{\hat{a}} = \int_0^\cdot \mu(\bar{X}^{\hat{a}}_s,\hat{a}(\bar{X}^{\hat{a}}_s))\mathrm{d}s + \int_0^\cdot \sigma(\bar{X}^{\hat{a}}_s)\mathrm{d}W_s.$$

Note that the existence of a solution of the above is guaranteed, upon considering another probability space and Brownian motion. Combining the above inequalities with (4.22), the fact that $\gamma_\circ \le \gamma$, and the second assertion of Assumption 4.5 with $p_{\bar{X}} \ge 3$ shows that $\|\delta\bar{\rho}^*_\varepsilon\| \le C'$ for some $C' > 0$ that does not depend on $\varepsilon \in (0,\varepsilon_\circ]$.

Moreover, by Assumption 4.6 and the same arguments as in the proof of Proposition 4.3.1,

$$\delta r'_\varepsilon(x,a) := \frac{1}{\varepsilon}\int_{\mathbb{R}^{d'}}[\delta\bar{w}_\varepsilon(x + b_\varepsilon(x,a,e)) - \delta\bar{w}_\varepsilon(x)]\,\nu(\mathrm{d}e) - \bar{\mathcal{L}}^a\delta\bar{w}_\varepsilon$$

satisfies

$$\|\delta r'_\varepsilon(x,\cdot)\|_\infty \le \varepsilon^{\frac{\delta\gamma}{2}} C''(1+\|x\|^3),\ x\in\mathbb{R}^d,$$

for some $C''>0$ that does not depend on $\varepsilon\in(0,\varepsilon_\circ]$. Since, by construction, $\bar{w}^{(1)}_\varepsilon := \bar{w}+\varepsilon^{\frac{\gamma_\circ}{2}}\delta\bar{w}_\varepsilon$ solves

$$\bar{\rho}^{*(1)}_\varepsilon = \sup_{a\in\mathbb{A}}\left[\frac{1}{\varepsilon}\int_{\mathbb{R}^{d'}}\left[\bar{w}^{(1)}_\varepsilon(\cdot+b_\varepsilon(\cdot,a))-\bar{w}^{(1)}_\varepsilon\right]\nu(de)-\varepsilon^{\frac{\gamma_\circ}{2}}\delta r'_\varepsilon(\cdot,a)+r(\cdot,a)\right]$$

on $\mathbb{R}^d$, the same arguments as in the proof of Theorem 4.3.2 then imply that $\|\bar{\rho}^{*(1)}_\varepsilon-\rho^*_\varepsilon\|\le L\varepsilon^{(\gamma_\circ+\delta\gamma)/2}$, for some $L>0$ that does not depend on $\varepsilon\in(0,\varepsilon_\circ]$, and also lead to the last assertion of the theorem. □

4.4 Numerical Resolution of the Ergodic Diffusive Problem

4.4.1 Numerical resolution of (4.16)

The numerical resolution of (4.16) can be done by using standard finite difference schemes as explained in [78, Ch. 7]. We focus on the one-dimensional case $d=1$ for simplicity, and also because similar schemes in higher dimensions often have to be constructed on a case-by-case basis, see e.g. [78, Ch. 5].

Given $\kappa\in\mathbb{N}$, $\kappa\ge 3$, and $h>0$, we consider the space grid $\bar{M}^\kappa_h := \{z_i := -\kappa h+(i-1)h,\ 1\le i\le 2\kappa+1\}$. We use the notation $\breve{M}^\kappa_h := \bar{M}^\kappa_h\setminus\{z_1,z_{2\kappa+1}\}$ and denote by L^κ_h the collection of real-valued maps φ defined on $\bar{M}^\kappa_h$. For $\varphi\in L^\kappa_h$, we define the usual finite (central) differences operators:

$$\Delta_h\varphi(x) := \frac{\varphi(x+h)-\varphi(x-h)}{2h},\ \Delta^2_h\varphi(x) := \frac{\varphi(x+h)+\varphi(x-h)-2\varphi(x)}{h^2},$$

for $x\in\breve{M}^\kappa_h$, and set

$$\bar{\mathcal{L}}^{\bar{a}}_h\varphi := \Delta_h\varphi\mu(\cdot,\bar{a})+\frac{1}{2}\sigma^2\Delta^2_h\varphi,\ \bar{a}\in\mathbb{A}. \tag{4.25}$$

Then, we approximate the solution $(\bar{\rho}^*,\bar{w})$ of (4.16) by a solution $(\bar{\rho}^{\kappa,*}_h,\bar{w}^\kappa_h)\in\mathbb{R}\times L^\kappa_h$ of

$$\bar{\rho}^{\kappa,*}_h = \sup_{\bar{a}\in\mathbb{A}}\left\{\bar{\mathcal{L}}^{\bar{a}}_h\bar{w}^\kappa_h+r(\cdot,\bar{a})\right\},\ \text{on } \breve{M}^\kappa_h, \tag{4.26}$$

with a suitable reflecting boundary at z_1 and $z_{2\kappa+1}$, see below. Note that $\bar{w}^\kappa_h$ is defined only up to a constant and that we can, and will, set $\bar{w}^\kappa_h(0)=\bar{\rho}^{\kappa,*}_h$

in the following. Let us now denote by $\mathscr{A}$ the collection of measurable maps from $\mathbb{R}$ to $\mathbb{A}$ and identify, given $\bar{a} \in \mathscr{A}$, $\bar{w}_h^\kappa$ and $r(\cdot, \bar{a}(\cdot))$ on $\bar{M}_h^\kappa$ to column vectors $\bar{\mathcal{W}}_h^\kappa := (\bar{w}_h^\kappa(z_i))_{1\le i\le 2\kappa+1}$ and $R_h^\kappa(\bar{a}) := (r(z_i, \bar{a}(z_i)))_{1\le i\le 2\kappa+1}$ of $\mathbb{R}^{2\kappa+1}$. Let us fix

$$\Delta t_h := \frac{h^2}{(L_{b_1,b_2})^2}.$$

Then, to solve (4.26) on $\bar{M}_h^\kappa$ with $\bar{w}_h^\kappa(0) = \bar{\rho}_h^{\kappa,*}\Delta t_h$, for including a suitable reflection term on the boundary $\{z_1, z_{2\kappa+1}\}$, we search for $(\bar{\rho}_h^{\kappa,*}, \bar{\mathcal{W}}_h^\kappa) \in \mathbb{R} \times \mathbb{R}^{2\kappa+1}$ that satisfies

$$\bar{\mathcal{W}}_h^\kappa = \sup_{\bar{a}\in\mathscr{A}} \bar{Q}_h^{\bar{a}} \left\{ \bar{\mathcal{W}}_h^\kappa - e\bar{\rho}_h^{\kappa,*}\Delta t_h + R_h^\kappa(\bar{a})\Delta t_h \right\}, \text{ on } \bar{M}_h^\kappa \tag{4.27}$$

$$\bar{w}_h^\kappa(0) = \bar{\rho}_h^{\kappa,*}\Delta t_h \tag{4.28}$$

in which e is the column vector of $\mathbb{R}^{2\kappa+1}$ with all entries equal to 1, and $\bar{Q}_h^{\bar{a}} = ((\bar{Q}_h^{\bar{a}})^{i,j})_{1\le i,j\le 2\kappa+1}$ is the matrix with all entries null except for

$$(\bar{Q}_h^{\bar{a}})^{i,i-1} := q_h^-(z_i, \bar{a}(z_i)), \ (\bar{Q}_h^{\bar{a}})^{i,i} := q_h(z_i, \bar{a}(z_i)), \ (\bar{Q}_h^{\bar{a}})^{i,i+1} := q_h^+(z_i, \bar{a}(z_i)),$$

for $1 < i < 2\kappa + 1$, with

$$q_h := 1 - \frac{\sigma^2}{(L_{b_1,b_2})^2}, \ q_h^+ := \frac{\mu h + \sigma^2}{2(L_{b_1,b_2})^2}, \text{ and } q_h^- := \frac{-\mu h + \sigma^2}{2(L_{b_1,b_2})^2},$$

and except for

$$(\bar{Q}_h^{\bar{a}})^{1,j} := (\bar{Q}_h^{\bar{a}})^{3,j} \text{ for } j = 2, 3, 4$$
$$(\bar{Q}_h^{\bar{a}})^{2\kappa+1,j} := (\bar{Q}_h^{\bar{a}})^{2\kappa-1,j} \text{ for } j = 2\kappa-2, 2\kappa-1, 2\kappa.$$

The above scheme is of the form of [78, Ch. 7, (2.3)].

Without loss of generality, one can assume from now on that

$$L_{b_1,b_2} > \|\sigma\|_\infty.$$

Then, recalling (i)–(ii) of Assumption 4.4, $\bar{Q}_h^{\bar{a}}$ defines a Transition Probability Matrix satisfying

$$\min_{1\le i\le 2\kappa+1} \min_{1\vee(i-1)\le j\ne i\le(2\kappa+1)\wedge(i+1)} (\bar{Q}_h^{\bar{a}})^{i,j} =: \underline{p}_h > 0$$

whenever

$$L_{b_1,b_2}(1 + \kappa h)h < \varsigma. \tag{4.29}$$

Given $\bar{a} \in \mathscr{A}$, let $(Z_t^{x,\bar{a}})_{t\in\mathbb{N}}$ be the Markov chain starting from $x \in \bar{\mathrm{M}}_h^\kappa$ and such that

$$\mathbb{P}\left[Z_{t+1}^{x,\bar{a}} = z_j | Z_t^{x,\bar{a}} = z_i\right] = (\bar{Q}_h^{\bar{a}})^{i,j},\ 1 \le i,j \le 2\kappa+1,\ t \in \mathbb{N},$$

then

$$\mathbb{P}[Z_\kappa^{x,\bar{a}} = 0] \ge (\underline{p}_h)^\kappa > 0, \tag{4.30}$$

under (4.29). Then, assuming further that

$$(b,r)(x,\cdot) : \mathbb{A} \mapsto \mathbb{R}^2 \text{ is continuous for all } x \in \mathbb{R}, \tag{4.31}$$

it follows that the conditions of [78, Ch. 7, Thm. 2.1] hold so that $(\bar{\rho}_h^{\kappa,*}, \bar{\mathcal{W}}_h^\kappa)$ is well-defined and can be computed by using the iterative scheme of [78, Ch. 7, (2.3)].

Under the following conditions, one can exhibit an upper bound on the convergence rate of the above numerical scheme.

Assumption 4.7.
There exists a function $\bar{\xi} \in \mathcal{C}^3(\mathbb{R})$, $p_{\bar{\xi}} \ge 2$, and constants $C_{\bar{\xi}}^1 > 0$ and $C_{\bar{\xi}}^2 \in \mathbb{R}$ such that (4.14) and (4.15) hold for all $x \in \mathbb{R}^d$. Moreover, there are constants $L > 0$, $\Upsilon > 0$, and $C_\Upsilon > 0$, such that $\|\mathrm{D}^2\bar{\xi}(x)\| + \|\mathrm{D}^3\bar{\xi}(x)\| \le L(1+\|x\|^{p_{\bar{\xi}}-1})$ for all $x \in \mathbb{R}$, and $\mathrm{sgn}(x)\mathrm{D}\bar{\xi}(x) \ge C_\Upsilon \|x\|^{p_{\bar{\xi}}-1}$ for all $\|x\| \ge \Upsilon$, where $\mathrm{sgn}(\cdot)$ is the sign function.

Proposition 4.4.1.
Let Assumptions 4.4, 4.5 and 4.7 hold with $p_{\bar{X}} = p_{\bar{\xi}} \ge 3$. Assume further that (4.31) is satisfied. Then, there exists $L_{\mathrm{num}} > 0$ and $h_{\mathrm{num}} > 0$ such that, for all $(h,\kappa) \in (0, h_{\mathrm{num}}) \times \mathbb{N}$, satisfying (4.29), $\kappa h^2 \le 1$ and $(\kappa-3)h \ge \Upsilon$, we have

$$|\bar{\rho}_h^{\kappa,*} - \bar{\rho}^*| \le L_{\mathrm{num}}\left(h^\gamma + h^{-1}\,|\kappa h|^{-|p_{\bar{\xi}}-1|}\right).$$

In particular,

$$|\bar{\rho}_h^{\kappa,*} - \rho_\varepsilon^*| \le L_{\mathrm{num}}\left(h^\gamma + h^{-1}\,|\kappa h|^{-|p_{\bar{\xi}}-1|}\right) + \varepsilon^{\frac{\gamma}{2}} L_{\delta\rho}^\gamma \text{ for all } \varepsilon \in (0,1).$$

Proof. Given $\bar{a} \in \mathscr{A}$ and $x \in \bar{\mathrm{M}}_h^\kappa$, let $\tilde{X}^{x,\bar{a}^{x,\bar{a}}}$ be the pure jump continuous-time Markov chain defined by a sequence of jump times $(\tau_n)_{n\ge1}$ such that the increments $(\tau_{n+1} - \tau_n)_{n\ge0}$ (with the convention $\tau_0 = 0$) are independent

and identically distributed according to the exponential law of mean Δt_h and such that, for $n \geq 1$,

$$\mathbb{P}\left[\tilde{X}^{x,\bar{a}}_{\tau_n} = z_i | (\tilde{X}^{x,\bar{a}}_0, \tau_0), \dots, (\tilde{X}^{x,\bar{a}}_{\tau_{n-1}}, \tau_{n-1}), \tau_n\right] = (\bar{Q}^{\bar{a}}_h)^{i,j(\tilde{X}^{x,\bar{a}}_{\tau_{n-1}})},$$

with

$$j(\tilde{X}^{x,\bar{a}}_{\tau_{n-1}}) \in \mathbb{N} \text{ s.t. } z_{j(\tilde{X}^{x,\bar{a}}_{\tau_{n-1}})} = \tilde{X}^{x,\bar{a}}_{\tau_{n-1}},$$

and $\tilde{X}^{x,\bar{a}} = \tilde{X}^{x,\bar{a}}_{\tau_{n-1}}$ on $[\tau_{n-1}, \tau_n)$.

1. First note that, by construction, $\bar{w}^\kappa_h$ is bounded on the finite set $\bar{M}^\kappa_h$. Then, by the arguments in the proof of Lemma 4.A.4 and (4.27), we have

$$\bar{\rho}^{\kappa,*}_h = \sup_{\bar{a} \in \mathscr{A}} \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\int_0^T r(\tilde{X}^{x,\bar{a}}_s, \bar{a}(\tilde{X}^{x,\bar{a}}_s)) ds\right]. \tag{4.32}$$

2. We now prove that there exists $C^{1'}_{\bar{\xi}}, C^{2'}_{\bar{\xi}}, h_{\text{num}} > 0$ such that, for all $x \in \mathbb{R}$, $\bar{a} \in \mathscr{A}$, $0 < h \leq h_{\text{num}}$ and κ such that (4.29) holds, $\kappa h^2 \leq 1$ and $(\kappa - 3)h \geq \Upsilon$, we have

$$\mathbb{E}\left[\|\tilde{X}^{x,\bar{a}}_t\|^{p_{\bar{\xi}}}\right] \leq C^2_{\bar{\xi}} \left\{ e^{-C^{1'}_{\bar{\xi}} t} C^2_{\bar{\xi}} \|x\|^{p_{\bar{\xi}}} + \frac{C^{2'}_{\bar{\xi}}}{C^{1'}_{\bar{\xi}}} \left(1 - e^{-C^{1'}_{\bar{\xi}} t}\right) \right\}, \ t \geq 0. \tag{4.33}$$

Using Assumptions 4.4 and 4.7, and Taylor expansions of the first and second orders, we first deduce that, for $x \in \breve{M}^\kappa_h$,

$$D\bar{\xi}(x)\mu(x, \bar{a}(x)) + \frac{1}{2}\sigma^2(x) D^2 \bar{\xi}(x) = \frac{1}{\Delta t_h} \mathbb{E}\left[\bar{\xi}(\tilde{X}^{x,\bar{a}}_{\tau_1}) - \bar{\xi}(x)\right] - c(x)h,$$

in which $\|c(x)\| \leq C(1 + \|x\|^{p_{\bar{\xi}}}) \leq C(1 + C^2_{\bar{\xi}} \bar{\xi}(x))$ for some $C > 0$ independent on $x \in \mathbb{R}$, $\bar{a} \in \mathscr{A}$, h and κ. Using (4.14), this implies that, for $x \in \breve{M}^\kappa_h$,

$$\frac{1}{\Delta t_h} \mathbb{E}\left[\bar{\xi}(\tilde{X}^{x,\bar{a}}_{\tau_1}) - \bar{\xi}(x)\right] \leq -\left(C^1_{\bar{\xi}} - hCC^2_{\bar{\xi}}\right) \bar{\xi}(x) + C^2_{\bar{\xi}} + Ch. \tag{4.34}$$

Consider now the case $x = z_1$, the other boundary being symmetric. Let Ξ be a discrete random variable taking value $k \in \{1, 2, 3\}$ with probability $(\bar{Q}^{\bar{a}}_h)^{1,k}$. Using Assumption 4.7 and (4.15), we obtain that, for some random variable $\hat{z}_\Xi$ such that $\hat{z}_\Xi \in [z_1, z_1 + \Xi h]$ a.s.,

$$\frac{1}{\Delta t_h} \mathbb{E}\left[\xi(z_1 + \Xi h) - \xi(z_1)\right] = \frac{1}{\Delta t_h} \mathbb{E}[\Xi h D\xi(\hat{z}_\Xi)]$$

$$\begin{aligned}
&\leq -\frac{(L_{b_1,b_2})^2}{h} C_\Upsilon \mathbb{E}\left[\Xi |\hat{z}_\Xi|^{p_{\bar{\xi}}-1}\right] \\
&\leq -(L_{b_1,b_2})^2 C_\Upsilon \mathbb{E}\left[\kappa h |\hat{z}_\Xi|^{p_{\bar{\xi}}-1}\right] \\
&\leq -C' \bar{\xi}(z_1) \qquad (4.35)
\end{aligned}$$

when $|(\kappa - 3)h| \geq \Upsilon$ and $\kappa h^2 \leq 1$, in which $C' > 0$ does not depend on κ nor h. The above also holds with $z_{2\kappa+1}$ in place of z_1. Combining (4.34)–(4.35), we obtain

$$\begin{aligned}
\frac{1}{\Delta t_h} \mathbb{E}\left[\bar{\xi}(\tilde{X}^{x,\bar{a}}_{\tau_1}) - \bar{\xi}(x)\right] &\leq -\left((C^1_{\bar{\xi}} - hCC^2_{\bar{\xi}}) \wedge C'\right) \bar{\xi}(x) + C^2_{\bar{\xi}} + Ch \\
&\leq -C^{1'}_{\bar{\xi}} \bar{\xi}(x) + C^{2'}_{\bar{\xi}},
\end{aligned}$$

for all $x \in \bar{M}^\kappa_h$, whenever $h \leq h_{\text{num}}$, in which $C^{1'}_{\bar{\xi}}, C^{2'}_{\bar{\xi}}, h_{\text{num}} > 0$ do not depend on κ nor h. One can then argue as in the proof of Lemma 4.A.2 to obtain (4.33).

3. From now on, we denote by $C > 0$ a generic constant, which may change from line to line, but does not depend on κ or h. We now appeal to (4.17) and the Lipschitz continuity of (μ, σ), and use the fact that $h \leq 1$ to deduce by consistency arguments that, for $x \in \bar{M}^\kappa_h$,

$$\bar{\mathcal{L}}^{\bar{a}(x)} \bar{w}(x) = \frac{1}{\Delta t_h} \mathbb{E}\left[\bar{w}(\tilde{X}^{x,\bar{a}}_{\tau_1}) - \bar{w}(x)\right] + \delta r_h(x, a(x))$$

in which

$$|\delta r_h(x, \bar{a}(x))| \leq C((1 + |x|)h + h^\gamma)(1 + |x|) + Ch^{-1}(1 + |x|)1_{\{|x| = \kappa h\}}.$$

The above combined with (4.16) implies that

$$\bar{\rho}^* = \frac{1}{\Delta t_h} \sup_{\bar{a} \in \mathscr{A}} \mathbb{E}\left[\bar{w}(\tilde{X}^{x,\bar{a}}_{\tau_1}) - \bar{w}(x) + (r(x, \bar{a}) + \delta r_h(x, \bar{a}))\Delta t_h\right] \quad \text{for } x \in \bar{M}^\kappa_h.$$

Arguing again as in the proof of Lemma 4.A.4, recalling (4.32), and combining (4.33) with the inequalities of Hölder and Markov, we deduce that we can find $C, C', C'' > 0$ independent of h such that

$$\begin{aligned}
&\left|\bar{\rho}^{\kappa,*}_h - \bar{\rho}^*\right| \\
&\leq \sup_{\bar{a} \in \mathscr{A}} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\int_0^T |\delta r_h|\left(\tilde{X}^{0,\bar{a}}_s, \bar{a}(\tilde{X}^{0,\bar{a}}_s)\right)\left(1_{\{|\tilde{X}^{0,\bar{a}}_s| < \kappa h\}} + 1_{\{|\tilde{X}^{0,\bar{a}}_s| = \kappa h\}}\right) ds\right] \\
&\leq \sup_{\bar{a} \in \mathscr{A}} \limsup_{T \to \infty} \frac{C}{T} \int_0^T \mathbb{E}\Big[(h + h|\tilde{X}^{0,\bar{a}}_s|^2 + h^\gamma |\tilde{X}^{0,\bar{a}}_s|) \\
&\qquad\qquad + h^{-1}(1 + |\tilde{X}^{0,\bar{a}}_s|)1_{\{|\tilde{X}^{0,\bar{a}}_s| = \kappa h\}}\Big] ds
\end{aligned}$$

$$\leq \sup_{\bar{a}\in\mathscr{A}} \limsup_{T\to\infty} h^{-1}\frac{C}{T}\int_0^T \mathbb{E}\Big[(1+|\tilde{X}_s^{0,\bar{a}}|)^{p_{\bar{\xi}}}\Big]^{\frac{1}{p_{\bar{\xi}}}}\left(\frac{\mathbb{E}[|\tilde{X}_s^{0,\bar{a}}|^{p_{\bar{\xi}}}]}{(\kappa h)^{p_{\bar{\xi}}}}\right)^{\frac{p_{\bar{\xi}}-1}{p_{\bar{\xi}}}} \mathrm{d}s$$
$$+ C'(h+h^\gamma)$$
$$\leq C''\left(h^\gamma + h^{-1}|\kappa h|^{-|p_{\bar{\xi}}-1|}\right).$$

It remains to appeal to Theorem 4.3.2 to complete the proof. □

4.4.2 Construction of a near-optimal control

One can also construct from the above scheme an almost optimal control for the original pure jump problem. For this purpose, let ϕ be a smooth density function with support $(-1,1)$ such that $\|\phi\|_{\mathcal{C}_b^2} \leq 1$. Given $n \geq 1$, let

$$\bar{w}_h^{\kappa,n}(x) := \int \big(\bar{w}_h^\kappa(y) - \bar{\rho}_h^{\kappa,*}\Delta t_h\big)\phi(n(y-x))\mathrm{d}y,\ x\in\mathbb{R},$$

with the convention that $\bar{w}_h^\kappa = \bar{w}_h^\kappa(z_1) - \bar{\rho}_h^{\kappa,*}\Delta t_h$ on $(-\infty, z_1)$ and $\bar{w}_h^\kappa = \bar{w}_h^\kappa(z_{2\kappa+1}) - \bar{\rho}_h^{\kappa,*}\Delta t_h$ on $(z_{2\kappa+1},\infty)$.

Let $\bar{a}_h^{\kappa,n} \in \mathscr{A}$ be such that

$$\bar{a}_h^{\kappa,n} \in \operatorname*{argmax}_{a\in\mathbb{A}}[\bar{\mathcal{L}}^a \bar{w}_h^{\kappa,n} + r(\cdot,a)],\quad \text{on } \mathbb{R}, \tag{4.36}$$

and set $\hat{\alpha}_h^{\kappa,n} = \bar{a}_h^{\kappa,n}(\hat{X}^{\kappa,n,h})$ with

$$\hat{X}_\cdot^{\kappa,n,h} = \int_0^\cdot \int_{\mathbb{R}^{d'}} b_\varepsilon(\hat{X}_{s-}^{\kappa,n,h}, \bar{a}_h^{\kappa,n}(\hat{X}_{s-}^{\kappa,n,h}), e)N(\mathrm{d}e,\mathrm{d}s).$$

The control $\bar{a}_h^{\kappa,n}$ can be computed numerically at low cost, e.g. via first-order conditions; Proposition 4.4.2 gives the associated error bounds. It appears that this approach is novel in the literature and it seems of independent methodological interest.

Proposition 4.4.2.
Let the conditions of Proposition 4.4.1 hold. Then, there exists $C>0$ such that, for all $K>0$, $n\geq 1$ and $\varepsilon\in(0,1)$,

$$|\rho_\varepsilon(0,\hat{\alpha}_h^{\kappa,n}) - \rho_\varepsilon^*|$$
$$\leq C\left(n^{-\gamma} + \varepsilon^{\frac{\gamma}{2}} + n\sup_{x\in B_K(0)}\left|\bar{w}_h^\kappa - \bar{\rho}_h^{\kappa,*}\Delta t_h - \bar{w}\right|(x) + nK^{-1}\right). \tag{4.37}$$

If, moreover,

(i) σ is constant,

(ii) there exists $c_\mu > 0$ such that $\mu(x) - \mu(x') \le -c_\mu(x - x')$ if $x \ge x' \in \mathbb{R}$,

(iii) there exists $R > 0$ such that

$$\sup_{|x|>R} \sup_{\bar{a}\in\mathbb{A}} \mu(x, \bar{a})x < -\frac{1}{2}\sigma^2, \tag{4.38}$$

then

$$\limsup_{h\to 0} \sup_{x\in B_K(0)} \left|\bar{w}_h^{\kappa_h} - \bar{\rho}_h^{\kappa_h,*}\Delta t_h - \bar{w}\right|(x) = 0$$

for any family $(\kappa_h)_{h>0} \subset (2\mathbb{N}+1)$ satisfying

$$\lim_{h\downarrow 0} \kappa_h h^2 = 0 \text{ and } \lim_{h\downarrow 0} \kappa_h h^{\frac{p_\xi}{p_\xi - 1}} = \infty.$$

Proof.

1. We first note that

$$\begin{aligned} \mathrm{D}\bar{w}_h^{\kappa,n}(x) = &\int \mathrm{D}\bar{w}(y)\phi(n(y-x))dy \\ &- \int (\bar{w}_h^\kappa - \bar{\rho}_h^{\kappa,*}\Delta t_h - \bar{w})(y) n\phi'(n(y-x))dy \end{aligned}$$

and

$$\begin{aligned} \mathrm{D}^2\bar{w}_h^{\kappa,n}(x) = &\int \mathrm{D}^2\bar{w}(y)\phi(n(y-x))dy \\ &+ \int (\bar{w}_h^\kappa - \bar{\rho}_h^{\kappa,*}\Delta t_h - \bar{w})(y) n^2\phi''(n(y-x))dy \end{aligned}$$

for $x \in \mathbb{R}$, in which ϕ' and ϕ'' stand for the first and second order derivatives of ϕ. Hence, it follows from (4.17), (i) of Assumption 4.4, and (4.16) that

$$\begin{aligned} \bar{\mathcal{L}}^{\bar{a}_h^{\kappa,n}(\cdot)}\bar{w}_h^{\kappa,n} + r(\cdot, \bar{a}_h^{\kappa,n}(\cdot)) &= \max_{a\in\mathbb{A}}[\bar{\mathcal{L}}^a\bar{w}_h^{\kappa,n} + r(\cdot, a)] \\ &\ge \max_{a\in\mathbb{A}}[\bar{\mathcal{L}}^a\bar{w} + r(\cdot, a)] - \frac{1}{2}\delta r_h^{\kappa,n} \\ &= \bar{\rho}^* - \frac{1}{2}\delta r_h^{\kappa,n} \end{aligned}$$

in which $\delta r_h^{\kappa,n}$ satisfies, for some $C > 0$ independent on n, κ and h,

$$0 \le \delta r_h^{\kappa,n}(x) \le C(1+|x|)\left[n^{-\gamma} + 2n^2 \int_{\mathcal{B}_{\frac{1}{n}}(x)} \left| \bar{w}_h^\kappa - \bar{\rho}_h^{\kappa,*} \Delta t_h - \bar{w} \right|(y) dy \right].$$

Similarly,

$$\begin{aligned} \bar{\rho}^* - \frac{1}{2}\delta r_h^{\kappa,n} &\le \bar{\mathcal{L}}^{\bar{a}_h^{\kappa,n}(\cdot)} \bar{w}_h^{\kappa,n} + r(\cdot, \bar{a}_h^{\kappa,n}(\cdot)) \\ &\le \bar{\mathcal{L}}^{\bar{a}_h^{\kappa,n}(\cdot)} \bar{w} + r(\cdot, \bar{a}_h^{\kappa,n}(\cdot)) + \frac{1}{2}\delta r_h^{\kappa,n}. \end{aligned}$$

Recalling (4.21) and Theorem 4.3.2, we deduce that

$$\begin{aligned} \rho_\varepsilon^* - \varepsilon^{\frac{\gamma}{2}} L_{\delta\rho}^\gamma \le \frac{1}{\varepsilon} \int \big[\bar{w}(x + b_\varepsilon(x, \bar{a}_h^{\kappa,n}(x), e)) - \bar{w}(x) \big] \nu(de) \\ + r(x, \bar{a}_h^{\kappa,n}(x)) + \delta r_h^{\kappa,n}(x) - \delta r_\varepsilon(x, \bar{a}_h^{\kappa,n}(x)) \end{aligned}$$

for all $x \in \mathbb{R}$. We then deduce (4.37) by the same arguments as in the proof of Theorem 4.3.2.

2. It remains to prove the second assertion of the proposition. For ease of notations, we do not write the dependence of κ with respect to h, but we keep in mind that we can consider h small and that κ can be adjusted as soon as the following results can apply to sequences such that $\lim_{h\downarrow 0} \kappa_h h^2 = 0$ and $\lim_{h\downarrow 0} \kappa_h h^{p_{\bar{\xi}}/(p_{\bar{\xi}}-1)} = \infty$.

2.a. We first prove that $[\bar{w}_h^\kappa]_{\mathcal{C}^0_{\rm lin}(\bar{\mathbb{M}}_h^\kappa)}$ does not depend on κ nor h. To this end, we adapt the arguments of Lemma 4.A.1 and Theorem 4.2.1, and actually prove that it is Lipschitz, uniformly in κ and h.

Let $(\xi_j)_{j\ge 1}$ be a sequence of i.i.d. random variables following the uniform distribution on $[0,1]$ and let $(\tau_n)_{n\ge 1}$ be a random sequence, independent of $(\xi_j)_{j\ge 1}$, such that the increments $(\tau_{n+1} - \tau_n)_{n\ge 0}$ (with the convention $\tau_0 = 0$) are independent and identically distributed according to the exponential law of mean Δt_h. Given $(x, \bar{a}, y) \in \mathbb{R} \times \mathbb{A} \times \mathbb{R}$, set

$$\Delta x(x, \bar{a}, y) := h 1_{\{y \le q_h^+(x,\bar{a})\}} - h 1_{\{q_h^+(x,a) < y \le (q_h^+ + q_h^-)(x,\bar{a})\}}, \text{ if } x \in \breve{\mathbb{M}}_h^\kappa,$$

and

$$\Delta x(z_1, \bar{a}, y) := 2h + \Delta x(z_3, \bar{a}, y), \ \Delta x(z_{2\kappa+1}, \bar{a}, y) := -2h + \Delta x(z_{2\kappa-1}, \bar{a}, y).$$

Let $\breve{\mathcal{A}}$ denote the collection of $\mathbb{A}$-valued processes that are predictable with respect to the filtration generated by $t \mapsto \sum_{i\ge 1} \exp(\xi_i) 1_{\{\tau_i \le t\}}$ Given

$\check{\alpha} \in \check{\mathcal{A}}$ and $x \in \bar{\mathrm{M}}_h^\kappa$, let $\check{X}^{x,\check{\alpha}}$ be the pure jump continuous time Markov chain defined by

$$\check{X}^{x,\check{\alpha}}_{\tau_{i+1}} = \check{X}^{x,\check{\alpha}}_{\tau_i} + \Delta x(\check{X}^{x,\bar{a}}_{\tau_i}, \check{\alpha}_{\tau_i}, \xi_{i+1})$$

and $\check{X}^{x,\check{\alpha}} = \check{X}^{x,\check{\alpha}}_{\tau_i}$ on $[\tau_i, \tau_{i+1})$, $i \geq 0$. It has the same law as the process $\tilde{X}^{x,\check{\alpha}}$ introduced at the beginning of the proof of Proposition 4.4.1, and in particular

$$\bar{\rho}_h^{\kappa,*} = \sup_{\check{\alpha}\in\check{\mathcal{A}}} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T r(\check{X}^{x,\bar{a}}_s, \check{\alpha}_s)ds\right].$$

We set

$$\check{V}_\lambda(x) := \sup_{\check{\alpha}\in\check{\mathcal{A}}} \mathbb{E}\left[\int_0^\infty e^{-\lambda s} r(\check{X}^{x,\check{\alpha}}_s, \check{\alpha}_s)ds\right].$$

2.a.i. We first need to obtain contraction estimates similar to the ones obtained in the proof of Lemma 4.A.1. We restrict for the moment to the case where the distance between the initial data is in $2h\mathbb{Z}$.

Let us first observe that, for h small enough for condition (4.29) to hold, we have $q_h^+(x) < (q_h^+ + q_h^-)(x') = \sigma^2(L_{b_1,b_2})^{-2} =: m$ and conversely. Although recall that, by assumption, $\mu(x) - \mu(x') \leq -c_\mu(x - x') \leq 0$ and therefore $q_h^+(x) \leq q_h^+(x')$ if $x \geq x' \in \mathbb{R}$. Keeping this in mind, direct computations show that, if $x - x' \in 2h\mathbb{Z}$ and $x, x' \in \check{\mathrm{M}}_h^\kappa$, and $\bar{a} \in \mathbb{A}$, then

$$\begin{aligned}
&\frac{1}{\Delta t_h}\mathbb{E}\left[|x + \Delta x(x,\bar{a},\xi_1) - x' - \Delta x(x',\bar{a},\xi_1)| - |x - x'|\right] \\
&\qquad = (|x - x' + 2h| - |x - x'|)\,\frac{q_h^+(x) \wedge m - q_h^+(x) \wedge q_h^+(x')}{\Delta t_h} \\
&\qquad\quad + (|x - x' - 2h| - |x - x'|)\,\frac{q_h^+(x') \wedge m - q_h^+(x') \wedge q_h^+(x)}{\Delta t_h} \\
&\qquad = (|x - x' + 2h| - |x - x'|)\,\frac{\mu(x) - \mu(x) \wedge \mu(x')}{2h} \\
&\qquad\quad + (|x - x' - 2h| - |x - x'|)\,\frac{\mu(x') - \mu(x) \wedge \mu(x')}{2h} \\
&\qquad = 1_{\{x\geq x'\}}(\mu(x) - \mu(x')) + 1_{\{x'>x\}}(\mu(x') - \mu(x)) \\
&\qquad \leq -c_\mu\,|x - x'|\,.
\end{aligned}$$

On the other hand, if $x = z_1$, $x' \in \check{\mathrm{M}}_h^\kappa$ and $z_1 - x' \in 2h\mathbb{Z}$, then

$$\frac{1}{\Delta t_h}\mathbb{E}\left[|x + \Delta x(x,\bar{a},\xi_1) - x' - \Delta x(x',\bar{a},\xi_1)| - |x - x'|\right]$$

$$
\begin{aligned}
&= \frac{1}{\Delta t_h}\Big(- 2hq_h^+(x',\bar a) - 4h(q_h^+(z_3,\bar a)) - q_h^+(x',\bar a) \\
&\quad - 2h(1-q_h^+(z_3,\bar a))\Big)1_{\{x'\geq z_1+4h\}} - \frac{1}{\Delta t_h}\,|x-x'|\,1_{\{x'<z_1+4h\}} \\
&\leq -c_\mu\,|z_3-x'| \\
&\leq -\frac{1}{2}c_\mu\,|x-x'|\,.
\end{aligned}
$$

In the case, $x' = z_{2\kappa+1}$ (with $\kappa \geq 4$ which we can assume here w.l.o.g.), then

$$
\begin{aligned}
&\frac{1}{\Delta t_h}\mathbb{E}\left[|x+\Delta \mathrm{x}(x,\bar a,\xi_1) - x' - \Delta \mathrm{x}(x',\bar a,\xi_1)| - |x-x'|\right] \\
&= \frac{1}{\Delta t_h}\left[-4hq_h^+(x',\bar a) - 6h(q_h^+(z_3,\bar a)) - q_h^+(z_{2\kappa-1},\bar a) - 4h(1-q_h^+(z_3,\bar a))\right] \\
&\leq -c_\mu\,|z_3 - z_{2\kappa-1}| \\
&\leq -\frac{1}{2}c_\mu\,|x-x'|\,,
\end{aligned}
$$

in which the last inequalities follow from the fact that $\kappa \geq 4$. A similar analysis can be done when $x' = z_{2\kappa+1}$ and $x \in \bar{\mathrm{M}}_h^\kappa$. The above implies that, for h small enough,

$$
\frac{1}{\Delta t_h}\mathbb{E}\left[|x+\Delta \mathrm{x}(x,\bar a,\xi_1) - x' - \Delta \mathrm{x}(x',\bar a,\xi_1)| - |x-x'|\right] \leq -\frac{1}{2}c_\mu\,|x-x'|
$$

for any $(x,x') \in \bar{\mathrm{M}}_h^\kappa \times \bar{\mathrm{M}}_h^\kappa$ such that $x - x' \in 2h\mathbb{Z}$, which is the required contraction property, whenever $x - x' \in 2h\mathbb{Z}$. The key property is that $\check X^{x,\check\alpha} - \check X^{x',\check\alpha}$ remains in $2h\mathbb{Z}$ whenever $x - x' \in 2h\mathbb{Z}$ (by the above calculations jumps of $\check X^{x,\check\alpha} - \check X^{x',\check\alpha}$ lie in $\{-6h,-4h,-2h,0,2h,4h,6h\}$). Then, the same arguments as in the proof of Lemma 4.A.1 imply that one can find $\check L > 0$, that only depends on c_μ, such that

$$
\left|\check V_\lambda(x) - \check V_\lambda(x')\right| \leq \check L\,|x-x'| \tag{4.39}
$$

for any $(x,x') \in \bar{\mathrm{M}}_h^\kappa \times \bar{\mathrm{M}}_h^\kappa$ such that $x - x' \in 2h\mathbb{Z}$. In particular,

$$
\left|\check V_\lambda(x) - \check V_\lambda(0)\right| \leq \check L\,|x|\,, \quad \text{for } x \in \bar{\mathrm{M}}_h^\kappa \cap (2h\mathbb{Z}). \tag{4.40}
$$

2.a.ii. We now turn to the general case in which the distance between the initial data does not belong to $2h\mathbb{Z}$. Take $x \in \{x_\circ - h, x_\circ + h\} \cap \breve{\mathrm{M}}_h^\kappa$, for some $x_\circ \in \bar{\mathrm{M}}_h^\kappa \cap (2h\mathbb{Z})$. Let θ_1 be the first time at which $|\check X_{\theta_1}^{x,\check\alpha} - x| = h$. By the Dynamic Programming Principle,

$$
\left|\check V_\lambda(x) - \check V_\lambda(x_\circ)\right| \leq \sup_{\check\alpha\in\check{\mathcal{A}}}\mathbb{E}\left[\frac{1-e^{-\lambda\theta_1}}{\lambda}\,\|r\|_\infty + e^{-\lambda\theta_1}\left|\check V_\lambda(\check X_{\theta_1}^{x,\check\alpha}) - \check V_\lambda(x_\circ)\right|\right]
$$

$$+\mathbb{E}\left[(1-e^{-\lambda\theta_1})\left|\check{V}_\lambda(x_\circ)\right|\right]$$

in which $\check{X}^{x,\check{\alpha}}_{\theta_1} - x_\circ \in \{-2h, 0, 2h\}$ and therefore $\left|\check{V}_\lambda(\check{X}^{x,\check{\alpha}}_{\theta_1}) - \check{V}_\lambda(x_\circ)\right| \leq 2\check{L}\,|h|$ by (4.39). By exhaustive enumeration, one can compute

$$\begin{aligned}
\mathbb{E}[e^{-\lambda\theta_1}] &= \sum_{k\geq 1} q_h(x)^{k-1}(1-q_h(x))\left(\int_0^\infty e^{-\lambda y}\frac{1}{\Delta t_h}e^{-\Delta t_h^{-1}y}dy\right)^k \\
&= \sum_{k\geq 1} q_h(x)^{k-1}(1-q_h(x))(\lambda\Delta t_h+1)^{-k} \\
&= (\lambda\Delta t_h+1)^{-1}(1-q_h(x))\frac{\lambda\Delta t_h+1}{\lambda\Delta t_h+1-q_h(x_\circ)} \\
&= \frac{1-q_h(x)}{\lambda\Delta t_h+1-q_h(x)} \leq 1.
\end{aligned}$$

Since $1-q_h(x) \geq (\sigma/L_{b_1,b_2})^2 \geq (\varsigma/L_{b_1,b_2})^2 > 0$ for all h, by Assumption 4.4, the above implies that, for some $C > 0$, independent on λ, κ and h,

$$\begin{aligned}
\left|\check{V}_\lambda(x) - \check{V}_\lambda(x_\circ)\right| \leq & \sup_{a\in\mathbb{A}} \mathbb{E}\left[\frac{\Delta t_h \left\|r\right\|_\infty}{\lambda\Delta t_h+1-q_h(x)} + \frac{2\check{L}\,|h|\,(1-q_h(x))}{\lambda\Delta t_h+1-q_h(x)}\right] \\
&+ \mathbb{E}\left[\frac{\lambda\Delta t_h}{\lambda\Delta t_h+1-q_h(x)}\left|\check{V}_\lambda(x_\circ)\right|\right] \\
=& C\left(\Delta t_h + h + \lambda\Delta t_h\left|\check{V}_\lambda(x_\circ)\right|\right).
\end{aligned}$$

Note that $\lambda\check{V}_\lambda$ is bounded by $\left\|r\right\|_\infty < \infty$, while $\Delta t_h \leq h \leq |x|$, for $x \neq 0$ and h small enough. Since $x_\circ \in \bar{\mathrm{M}}^\kappa_h \cap (2h\mathbb{Z})$, the above, combined with (4.40), thus shows that

$$\left|\check{V}_\lambda(x) - \check{V}_\lambda(0)\right| \leq \check{L}'\,|x|\,,\ \forall\, x \in \check{\mathrm{M}}^\kappa_h, \tag{4.41}$$

for some $\check{L}' > 0$ that does not depend on λ, h nor κ. In the case where $x \in \{z_1, z_{2\kappa+1}\}$, we can conduct a similar analysis by considering the first time θ_1 at which $\check{X}^{x,\check{\alpha}}$ jumps. In this case, $\check{X}^{x,\check{\alpha}}_{\theta_1} \in \check{\mathrm{M}}^\kappa_h$ by construction and $|\check{X}^{x,\check{\alpha}}_{\theta_1} - x_\circ| \leq 2h$. Given (4.39), we retrieve a similar estimate as (4.41). Hence,

$$\left|\check{V}_\lambda(x) - \check{V}_\lambda(0)\right| \leq \check{L}'\,|x|\,,\ \forall\, x \in \bar{\mathrm{M}}^\kappa_h, \tag{4.42}$$

for some $\check{L}' > 0$ that does not depend on λ, h nor κ.

2.a.iii. We are now in position to show that $[\bar{\mathrm{w}}^\kappa_h]_{\mathcal{C}^0_{\mathrm{lin}}(\bar{\mathrm{M}}^\kappa_h)}$ does not depend on κ nor h. Using (4.42) and the arguments of Lemma 4.A.3, we obtain that, after possibly passing to a subsequence, $(\check{V}_\lambda - \check{V}_\lambda(0))_{\lambda>0}$ converges pointwise, as $\lambda \to 0$, to $\bar{\mathrm{w}}^\kappa_h - \bar{\rho}^{\kappa,*}_h \Delta t_h$ and that the latter satisfies

$$\left|\bar{\mathrm{w}}^\kappa_h(x) - \bar{\rho}^{\kappa,*}_h \Delta t_h\right| \leq \check{L}'\,|x|\,,\ x \in \bar{\mathrm{M}}^\kappa_h. \tag{4.43}$$

2.b. To complete the proof, it remains to appeal to the stability of viscosity solutions, and use comparison results in the class of semi-continuous super/sub-solutions with linear growth. Let $(\kappa_h)_{h>0}$ be as in the statement of the Proposition. By (4.43), $(\bar{w}_h^{\kappa_h} - \bar{\rho}_h^{\kappa_h,*}\Delta t_h)_{h>0}$ admits locally bounded relaxed semi-limits

$$\bar{w}_0^{\infty *}(x) := \limsup_{x' \to x,\, h \downarrow 0} \bar{w}_h^{\kappa_h}(x') - \bar{\rho}_h^{\kappa_h,*}\Delta t_h$$

and

$$\bar{w}_{0*}^{\infty}(x) := \liminf_{x' \to x,\, h \downarrow 0} \bar{w}_h^{\kappa_h}(x') - \bar{\rho}_h^{\kappa_h,*}\Delta t_h.$$

which take the value 0 at 0, recall (4.28), and have linear growth. We can then use (4.27), Proposition 4.4.1 and standard stability arguments for viscosity solutions, see e.g. [57, § 3], to deduce that $\bar{w}_{0*}^{\infty}$ and $\bar{w}_0^{\infty *}$ are respectively viscosity super- and subsolutions of (4.11). We claim that $\bar{w}_0^{\infty *} = \bar{w} + g$ for some $g \in \mathbb{R}$. Then, we will deduce that $g = \bar{w}_0^{\infty *}(0) - \bar{w}(0) = 0$ by construction. The same argument can be used to prove that $\bar{w}_{0*}^{\infty} = \bar{w}$. To prove the above, we follow the arguments of [18, Pf. of Thm. 3.1]. We first fix $R > 0$ and let $B_R := B_R(0)$ be the open ball of radius R centred at 0. Set $g := \max_{\partial B_R}(\bar{w}_0^{\infty *} - \bar{w})$. Since $\Phi := \bar{w}_0^{\infty *} - \bar{w} - g$ has linear growth, see (4.17) and above, we can fix $\iota > 0$, independently of R, such that $x \mapsto \Phi(x) - \iota|x|^2$ has a maximum point $\hat{x}_R$ on $(B_R)^c$. If $\sup_{(B_R)^c} \Phi > 0$, then, for $\iota > 0$ small enough, we have $\Phi(\hat{x}_R) - \iota|\hat{x}_R|^2 > 0$ and therefore $\hat{x}_R$ lies in the interior of $(B_R)^c$. We now use the subsolution property of $\bar{w}_0^{\infty *}$ and the fact that $\bar{w}$ is a smooth solution of (4.11) to obtain

$$\begin{aligned} 0 \leq & \sup_{\bar{a} \in \mathbb{A}} \left\{ \bar{\mathcal{L}}^{\bar{a}} \bar{w}(\hat{x}_R) + r(\hat{x}_R, \bar{a}) - \bar{\rho}^* + \iota \left(2\mu(\hat{x}_R, \bar{a})\hat{x}_R + \sigma^2 \right) \right\} \\ \leq & \iota \sup_{\bar{a} \in \mathbb{A}} \left\{ 2\mu(\hat{x}_R, \bar{a})\hat{x}_R + \sigma^2 \right\}. \end{aligned}$$

Using (4.38), we get a contradiction for R large enough. This shows that $\sup_{(B_R)^c} \Phi \leq 0$. Now the fact that $\max_{B_R \cup \partial B_R} \Phi = 0$ follows by the maximum principle applied to (4.11) on B_R with Dirichlet boundary conditions on ∂B_R. Moreover, Φ is a viscosity subsolution of

$$0 \leq \sup_{\bar{a} \in \mathbb{A}} \bar{\mathcal{L}}^{\bar{a}} \Phi.$$

We can thus now appeal to the strong maximum principle, see e.g. [70, Thm. 1], to deduce that $\bar{w}_0^{\infty} - \bar{w} - g = \Phi \equiv 0$, which concludes the proof. □

4.5 Application to High-Frequency Auctions

4.5.1 Motivation and setting

Web display advertising is a typical example of real-world high-frequency pure jump control problems [54]. The ad-spaces are sold by algorithmic platforms in automated auctions which occur at the dozen microsecond scale [98]. The frequency imposes computational issues on optimisation problems in this industry, while at the same time, the volume creates a significant monetary incentive for all parties to engage in revenue maximisation.

Consequently, the question of the strategic behaviour of bidders in repeated auctions in the face of learning sellers has been a popular topic in contemporary auction theory, see e.g. [93, § 4] for a survey. A rich line of work has focused on asymmetric problems where one player is significantly more patient than the other [9, 94]. This asymmetry reduces game theoretic considerations in the analysis to optimisation or control problems. In this example, we take interest in the case where the buyer is infinitely patient (it optimises an ergodic objective), while the seller's algorithm has an effectively finite memory of bidder behaviour.

Given these horizons, the format of the auction will strongly influence the behaviour of bidders and sellers when they seek to maximise their profit, see e.g. [75] for some generic examples. While it is a sub-optimal auction format for the seller [92], we choose to focus on the second-price auction format here. Indeed, there are unsurmountable difficulties in learning the optimal auction format [90], and second-price is in practice a common compromise between tractability and optimality [105].

Recalling the notations introduced in Example 4.2.1, in a second-price auction (with reserve) the bidder wins if it outbids the competition e_4 and the reserve price x, and pays the smallest bid which still wins the auction, i.e. $x \vee e_4$. As a result of the time scale there is little time in practice to perform computations to determine the bid, and one typically relies on using a precomputed function of the value to bid when an auction arrives and e_3 is revealed. More formally, the bid should be predictable. For simplicity, in this example, we consider a linear shading of the value: ae_2, where the control input value a is the shading factor. Consequently, we have the (expected) reward function

$$r(x,a) := \int (e_2 - x \vee e_4) 1_{\{ae_2 \geq x \vee e_4\}} \nu(\mathrm{d}e)\,. \tag{4.44}$$

Such auctions are well-defined only for positive bids. Thus, we impose $a \in \mathbb{R}_+$.

Within the constraints of a second-price auction, maximisation of profits

corresponds to tuning the reserve price x. Dynamically optimising the reserve price is a difficult problem even for a stationary bidder, see Chapter 6. To simplify, we consider the mean-reverting dynamic which we introduced in Example 4.2.1. For some $\eta = \varepsilon^{-1}$ fixed, this dynamic is given by (4.3) with $b := b_\varepsilon = \varepsilon b_1 + \sqrt{\varepsilon} b_2$, in which

$$b_1(x,a,e) := e_1(ae_2 - x) \text{ and } b_2(x,a,e) := e_1 e_3. \tag{4.45}$$

In the above framework, the noise e_1 models seller aggressivity as exogenous randomness, while the noise e_3 models the seller's internal randomisation whose aim is to increase robustness to strategic play. Under the conditions outlined in Example 4.2.1, we can choose for simplicity

$$\nu(\mathrm{d}e) = \prod_{i=1}^{4} f_i(e_i)\mathrm{d}e_i$$

in which

$$f_1 \sim \texttt{Unif}(0,1) \quad \& \quad f_3 \sim \mathcal{N}(0, \sigma_0^2)$$

with $\sigma_0 = \frac{1}{2}$.

Second-price auctions without reserve leave the most revenue on the table when the buyers are highly asymmetrical, we therefore study

$$f_2 \sim \texttt{LogNorm}(\mu_1, \sigma_1) \quad \& \quad f_4 \sim \texttt{Unif}(0,1)$$

with $\mu_1 = 0$ and $\sigma_1 = \frac{1}{2}$. Note that empirical observations [98] suggest log-normal is a realistic statistical model for values.

Assumption 4.1, and the remaining conditions in Example 4.2.1 for Assumptions 4.2 and 4.3 are easily seen to hold under the above choices. Therefore, this pure jump process admits, and converges to, a diffusion limit by Theorem 4.3.2, in particular it is easily checked that the coefficients of the limit diffusion are given by

$$\mu(x,a) := \frac{1}{2}(aC - x) \text{ and } \sigma(x) := \frac{\sigma_0}{\sqrt{3}}, \tag{4.46}$$

where $C := \exp\left(\mu_1 + \frac{\sigma_1^2}{2}\right)$. It is clear from (4.44) and (4.46) that values of a larger than 1 cannot be optimal, therefore we fix $\mathbb{A} = [0,1]$.

4.5.2 Numerical resolution of the HJB equations

Using this example motivated by high-frequency auctions we illustrate in this section the benefits of the diffusion limit problem in regard to numerical

computation. We use the method detailed in Section 4.4 to solve numerically (4.16), with coefficients μ and σ given by (4.46). Throughout, we will take $\kappa_h := h^{-1/4}$, for which $h \leq (\sigma/2)^{8/3}$ suffices to uphold condition (4.29) since we have $[\mu]_{\mathcal{C}^0_{\text{lin}}} \leq (1 + e^{1/8})/2$. Note that, with f_1, f_2, f_3 as above, p in Example 4.2.1 and Example 4.3.2 can be taken to be any positive real number.

In comparison to (4.16), solving (4.10) with coefficients given by (4.45) is complicated by the computation of the integral term. In many situations, when ν is a non-atomic measure with a known closed form, quadrature would be the preferred method for resolution, see e.g. [47]. In this example, this quadrature would be 4-dimensional, which is somewhat expensive.

In contrast, the relatively simple form of the combination of independent noise sources makes Monte Carlo simulation competitive in this specific example. Fixing a grid M_ε analogous to the one in Section 4.4, we compute the empirical transition distribution $p^{\mathrm{a}}_\varepsilon : (x, a) \in \mathrm{M}_\varepsilon \times \mathbb{A} \to p^{\mathrm{a}}_\varepsilon(\cdot; x, a) \in \Delta_{2\kappa_\varepsilon+1}$, where $\Delta_{2\kappa_\varepsilon+1}$ is the $2\kappa_\varepsilon + 1$-dimensional probability simplex, based on N_ε independent samples from each law, by projecting sample transitions onto M_ε. We then approximate for (4.10) by solving the analogue of (4.27), i.e. finding $(\rho^{\kappa_\varepsilon,*}_{\varepsilon,h_\varepsilon}, \mathcal{W}^{\kappa_\varepsilon}_{h_\varepsilon})$, $\mathcal{W}^{\kappa_\varepsilon}_{h_\varepsilon} := (\bar{\mathrm{w}}^{\kappa_\varepsilon}_{h_\varepsilon}(z_i))_{1\leq i\leq 2\kappa_\varepsilon+1}$, solving

$$0 = \max_{\mathrm{a}\in\mathscr{A}} \left\{ \frac{1}{\varepsilon}(P^{\mathrm{a}}_\varepsilon - \mathrm{I}_{2\kappa_\varepsilon+1})\mathcal{W}^{\kappa_\varepsilon}_{h_\varepsilon} - e\mathcal{W}^{\kappa_\varepsilon}_{h_\varepsilon}(0) + \mathrm{R}^{\kappa_\varepsilon}_{h_\varepsilon}(\mathrm{a}) \right\} \tag{4.47}$$

$$\rho^{\kappa_\varepsilon,*}_{\varepsilon,h_\varepsilon} = \mathcal{W}^{\kappa_\varepsilon}_{h_\varepsilon}(0) \tag{4.48}$$

by policy iteration, where $P^{\mathrm{a}}_\varepsilon = (p^{\mathrm{a}}_\varepsilon(z_j; z_i, \mathrm{a}(z_i)))_{1\leq i,j\leq 2\kappa_\varepsilon+1}$, $\mathrm{I}_{2\kappa_\varepsilon+1}$ is the $2\kappa_\varepsilon+1$-dimensional identity matrix, and $\mathrm{R}^{\kappa_\varepsilon}_{h_\varepsilon}(\mathrm{a})$ is as in Section 4.4.

As $\varepsilon \to 0$, all the transitions concentrate into a ball of size $\varepsilon^{\frac{1}{2}}$ with a drift of size ε, meaning that the mesh must refine faster than ε, in order to avoid degeneracy. Therefore, we consider the sequence of grids $(\mathrm{M}_\varepsilon)_{\varepsilon\geq 0}$, with $\mathrm{M}_\varepsilon := \{y_i = -10 + (i-1)h_\varepsilon : 1 \leq i \leq 2\kappa_\varepsilon + 1\}$ with $h_\varepsilon = \varepsilon^{3/2}$ and $\kappa_\varepsilon = \mathrm{N}_\varepsilon = 20\varepsilon^{3/2}$. Note that the refinement of the grid M_ε as $\varepsilon \to 0$ does not imply that the accuracy of the scheme increases as $\varepsilon \to 0$, the increasingly fine resolution is a *cost* incurred due to η_ε. The increase in this cost becomes impossible to maintain as ε becomes small, this is illustrated by Fig. 4.1: it rises at a rate $\varepsilon^{-3/2}$.

In contrast, using the diffusion limit by combining Sections 4.3 and 4.4, allows us to solve the problem to a high precision for relatively cheap. Figure 4.2 demonstrates the convergence in value of Theorem 4.3.2, with a rate of $\varepsilon^{1/2}$.

Explicit computation for an approximately optimal control using (4.36) is impractical for the r given in (4.44), due to its lack of a closed form derivative to apply first order conditions. Nevertheless, in order to illustrate the bounds

in Proposition 4.4.2, we resort to numerical approximation. We fix a grid $\mathbb{A}_\Gamma := \{i\Gamma^{-1}, 0 \le i \le \Gamma\}$ on $\mathbb{A} = [0,1]$, fixing $\Gamma = 100$, and then solve the maximum in (4.36) on $\mathbb{A}_\Gamma$ instead of $\mathbb{A}$. Contrary to Section 4.4.2, we only compute it on $\bar{\mathrm{M}}_h^{\kappa_h}$. This yields a map $\mathrm{a}_h^\Gamma : \bar{\mathrm{M}}_h^{\kappa_h} \to \mathbb{A}_\Gamma$, which can be viewed as a vector of controls associated with $\bar{\mathrm{M}}_h^{\kappa_h}$.

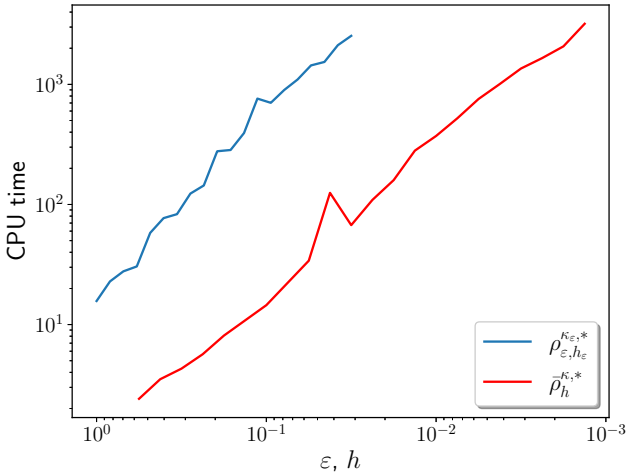

Figure 4.1: Comparison of computation costs for (4.47) (ρ^*_ε) and (4.27) ($\bar{\rho}^{\kappa,*}_h$).

From here, we define $\check{\mathrm{a}}_h^\Gamma \in \mathscr{A}$ by $\check{\mathrm{a}}_h^\Gamma := \mathrm{a}_h^\Gamma(\Pi_{\bar{\mathrm{M}}_h^{\kappa_h}}(\cdot))$ where $\Pi_{\bar{\mathrm{M}}_h^{\kappa_h}} : \mathbb{R} \to \bar{\mathrm{M}}_h^{\kappa_h}$ is the projector onto the grid $\bar{\mathrm{M}}_h^{\kappa_h}$; consider the solution $\check{X}^{\kappa_h,h,\Gamma}$ of

$$\check{X}^{\kappa_h,h,\Gamma} = \int_0^\cdot \int_{\mathbb{R}^{d'}} b_\varepsilon(\check{X}^{\kappa_h,h,\Gamma}_{s-}, \check{\mathrm{a}}_h^\Gamma(\check{X}^{\kappa_h,h,\Gamma}_{s-}), e) N(\mathrm{d}e, \mathrm{d}s)\,;$$

and evaluate $\rho_\varepsilon(0, \check{\alpha}_h^\Gamma)$ for each ε, with $\check{\alpha}_h^\Gamma := \check{\mathrm{a}}_h^\Gamma(\check{X}^{\kappa_h,h,\Gamma}_{\cdot-})$. In practice, we fix $T = 1000$ and compute

$$\frac{\varepsilon}{T}\mathbb{E}\left[\int_0^T r(\check{X}^{\kappa_h,h,\Gamma}_{t-}, \check{\mathrm{a}}_h^\Gamma(\check{X}^{\kappa_h,h,\Gamma}_{t-}))\mathrm{d}N_t\right]$$

by Monte Carlo with 1000 trajectories[1] of $\check{X}^{\kappa_h,h,\Gamma}$.ß

Despite the noise and the simple approximate control scheme, we recover the bounds of Proposition 4.4.2, in terms of ε in Fig. 4.3, with $h = 0.002667$ being the smallest h on Fig. 4.1. Note that this convergence rate matches the one of Fig. 4.2.

[1]Computing an ergodic average over each trajectory is very numerically expensive for small values of ε, reducing the feasible amount of samples. In spite of the noise, the slope $1/2$ is still visible in Fig. 4.3.

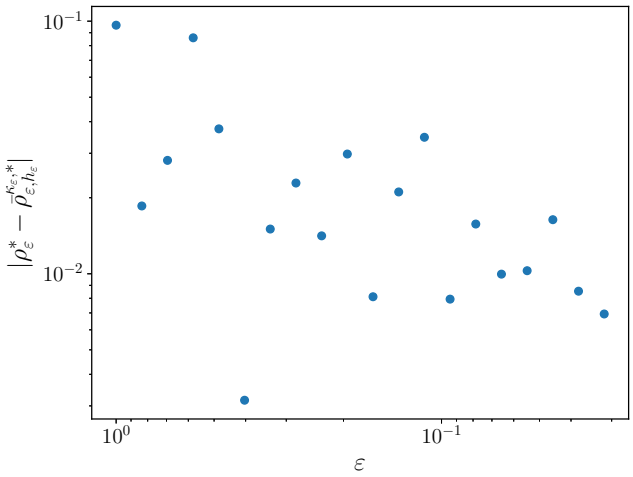

Figure 4.2: Approximation error of ρ^*_ε by $\bar{\rho}^{\kappa_h,*}_h$.

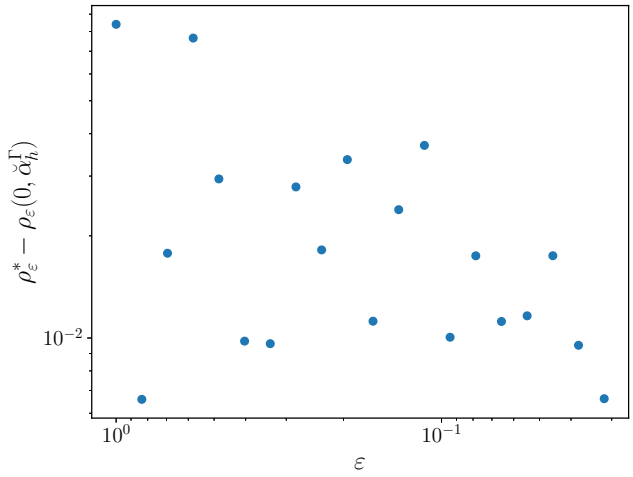

Figure 4.3: Suboptimality of the diffusive control relative to ρ^*_ε.

Conclusion

In this chapter, we studied the diffusion limit of a pure-jump *ergodic* control problem as the jump intensity goes to infinity, upon assuming a correct scaling of the coefficients. Unlike in Chapter 3, ergodicity requires us to study Lyapunov properties of the pre-limit and limit processes to establish the well-posedness of the ergodic control problems.

With this done, we applied a similar methodology to Chapter 3: we first studied the regularity of the solution to the diffusive Hamilton-Jacobi-Bellman equation, which we showed is locally γ-Hölder continuous for any $\gamma \in (0, 1)$ with a Hölder constant which grows linearly in the state x. This growth is inherited from the linear growth of the drift b_1. This Hölder exponent drives the convergence rate, which can be improved by error correction schemes.

We then studied numerical schemes of the diffusive HJB equation, showing a convergence rate via a similar methodology as the diffusive limit. We also studied the question of obtaining an approximately optimal control from the numerical solution of the diffusive HJB equation, which appears novel.

This overall methodology is effective at obtaining numerical approximations for high-frequency pure-jump ergodic control problems, as exemplified by those seen in online advertising auctions.

4.A Proof of Theorem 4.2.1

In this appendix, we first provide the proof of Theorem 4.2.1. It follows a standard route. We adapt arguments of [14] and [13] to our context.

We first show that $(V_\lambda)_{\lambda\in(0,1)}$ is equi-Lipschitz continuous, under the contraction condition of Assumption 4.2.

Lemma 4.A.1.
Let Assumptions 4.1 and 4.2 hold, then

$$|V_\lambda(x) - V_\lambda(x')| \le L_V|x - x'|, \text{ for } x, x' \in \mathbb{R}^d, \lambda \in (0,1),$$

in which

$$L_V := \frac{L_{b,r}p_\zeta}{C_\zeta}\left(\frac{L_\zeta}{\ell_\zeta}\right)^{\frac{1}{p_\zeta}}.$$

Proof. Fix $x, x' \in \mathbb{R}^d$, together with $\alpha \in \mathcal{A}$. By Assumption 4.2,

$$\begin{aligned}\eta \int_{\mathbb{R}^{d'}} &[\zeta(X_{\cdot-} + b(X_{\cdot-}, \alpha, e), X'_{-} + b(X'_{\cdot-}, \alpha, e)) - \zeta(X_{\cdot-}, X'_{\cdot-})]\, \nu(\mathrm{d}e) \\ &\le -C_\zeta \zeta(X_{\cdot-}, X'_{\cdot-})\end{aligned}$$

in which $(X, X') := (X^{x,\alpha}, X^{x',\alpha})$ and $(X_{\cdot-}, X'_{\cdot-})$ is its left-limit. Applying Itô's Lemma then implies that

$$\zeta(X_t, X'_t) \le \zeta(x, x') - C_\zeta \int_0^t \zeta(X_s, X'_s)\mathrm{d}s + M_t,\ t \ge 0,$$

where M is a local martingale. Upon using a localisation argument, recall (i) of Assumption 4.2, taking the expectation and using an immediate comparison result for ODEs leads to

$$\mathbb{E}[\zeta(X_t, X'_t)] \le \zeta(x, x')e^{-C_\zeta t},\ t \ge 0. \tag{4.49}$$

It remains to use (i) of Assumption 4.2 to deduce that

$$\mathbb{E}\left[\|X_t - X'_t\|^{p_\zeta}\right] \le \frac{L_\zeta}{\ell_\zeta}\|x - x'\|^{p_\zeta}e^{-C_\zeta t}, t \ge 0. \tag{4.50}$$

Combining the above with Remark 4.2.1, the Lipschitz continuity assumption on r, Assumption 4.1, and using Jensen's inequality then leads to

$$|J_\lambda(x, \alpha) - J_\lambda(x', \alpha)| \le L_{b,r}\int_0^\infty e^{-\lambda t}\mathbb{E}\left[\|X_t - X'_t\|\right]\mathrm{d}t$$

$$
\begin{aligned}
&\leq L_{b,r} \int_0^\infty e^{-\lambda t} \mathbb{E}\big[\|X_t - X_t'\|^{p_\zeta}\big]^{\frac{1}{p_\zeta}} \mathrm{d}t \\
&\leq L_{b,r} \left(\frac{L_\zeta}{\ell_\zeta}\right)^{\frac{1}{p_\zeta}} \int_0^\infty \|x - x'\| \, e^{-\lambda t - \frac{C_\zeta}{p_\zeta} t} \, \mathrm{d}t \\
&\leq \frac{L_{b,r} p_\zeta}{C_\zeta + \lambda p_\zeta} \left(\frac{L_\zeta}{\ell_\zeta}\right)^{\frac{1}{p_\zeta}} \|x - x'\| \, .
\end{aligned}
$$

Since $|V_\lambda(x) - V_\lambda(x')| \leq \sup_{\alpha \in \mathcal{A}} |J_\lambda(x,\alpha) - J_\lambda(x',\alpha)|$ and $\lambda p_\zeta \geq 0$, this completes the proof. □

We now use Assumption 4.3 to provide a uniform (in time and the control) estimate on the diffusion (4.3).

Lemma 4.A.2.
Let Assumptions 4.1 and 4.3 hold. Then,

$$
\mathbb{E}\left[\|X_t^{x,\alpha}\|^{p_\xi}\right] \leq \frac{1}{\ell_\xi} \left\{ e^{-C_\xi^1 t} L_\xi \|x\|^{p_\xi} + \frac{C_\xi^2}{C_\xi^1} \left(1 - e^{-C_\xi^1 t}\right) \right\}, \; t \geq 0,
$$

for any $(x, \alpha) \in \mathbb{R}^d \times \mathcal{A}$.

Proof. Fix $(x, \alpha) \in \mathbb{R}^d \times \mathcal{A}$ and let us write X for $X^{x,\alpha}$. By (4.7) and the same arguments as in the proof of 4.A.1,

$$
\mathbb{E}[\xi(X_t)] \leq \xi(x) + \int_0^t \mathbb{E}[-C_\xi^1 \xi(X_s) + C_\xi^2] ds, \; t \geq 0,
$$

which implies that

$$
\mathbb{E}[\xi(X_t)] \leq e^{-C_\xi^1 t} \xi(x) + \frac{C_\xi^2}{C_\xi^1} \left(1 - e^{-C_\xi^1 t}\right), \; t \geq 0.
$$

We conclude with (i) of Assumption 4.3. □

We can now prove a first convergence result.

Lemma 4.A.3.
Let Assumptions 4.1 and 4.2 hold. Then there is $c \in \mathbb{R}$ *and a sequence* $(\lambda_n)_{n \geq 1}$ *going to* 0 *such that* $(\lambda_n V_{\lambda_n})_{n \geq 1}$ *converges uniformly on compact sets to* c*, and such that* $(V_{\lambda_n} - V_{\lambda_n}(0))_{n \geq 1}$ *converges uniformly on compact sets to a function*

$\mathrm{w} \in \mathcal{C}^{0,1}$ *that solves*

$$c = \sup_{a\in\mathbb{A}}\left\{\eta \int_{\mathbb{R}^{d'}} [\mathrm{w}(\cdot + b(\cdot, a, e)) - \mathrm{w}]\, \nu(de) + r(\cdot, a)\right\}, \quad \textit{on } \mathbb{R}^d,$$

and satisfies

$$|\mathrm{w}(x)| \le L_V\, |x|\,, \; x \in \mathbb{R}^d. \tag{4.51}$$

Proof. The proof applies classical arguments from [14] to the pure jump setting. By Lemma 4.A.1, $(V_\lambda - V_\lambda(0))_{\lambda>0}$ is equicontinuous in the Lipschitz sense and, in particular, $|V_\lambda(x) - V_\lambda(0)| \le L_V\,|x|$ for all $x \in \mathbb{R}^d$ and $\lambda > 0$. Hence, $(\lambda(V_\lambda - V_\lambda(0)))_{\lambda\ge 0}$ converges uniformly on compact sets to 0 as $\lambda \to 0$. Since $(\lambda V_\lambda(0))_{\lambda\ge 0}$ is bounded (recall Lemma 4.A.2 and Assumption 4.1) there is a sequence $(\lambda_n)_{n\ge 1}$ converging to 0 such that $\lambda_n V_{\lambda_n}(0) \to c \in \mathbb{R}$ as $n \to \infty$. Thus, $\lambda_n V_{\lambda_n} \to c$ uniformly on compact sets.

By Lemma 4.A.1, $(V_\lambda - V_\lambda(0))_{\lambda>0}$ is locally bounded. Then, a diagonalisation argument allows one to extract a further subsequence (also denoted $(\lambda_n)_{n\ge 0}$) such that $V_{\lambda_n} - V_{\lambda_n}(0) \to \mathrm{w}$ on $\mathbb{Q}^d$ for some $\mathrm{w} : \mathbb{Q}^d \to \mathbb{R}$. By the uniform equicontinuity of $(V_\lambda)_{\lambda\in(0,1)}$, w can be extended to $\mathbb{R}^d$ and $V_{\lambda_n} - V_{\lambda_n}(0) \to \mathrm{w}$ uniformly on compact sets. Moreover, w is L_V-Lipschitz and $\mathrm{w}(0) = 0$, which implies (4.51).

Next, it follows from standard arguments, see e.g. [35], that V_{λ_n} solves for each $n \ge 1$

$$0 = \sup_{a\in\mathbb{A}}\left\{\eta \int_{\mathbb{R}^{d'}} [V_{\lambda_n}(\cdot + b(\cdot, a, e)) - V_{\lambda_n}]\nu(\mathrm{d}e) + r(\cdot, a)\right\} - \lambda_n V_{\lambda_n}, \quad \text{on } \mathbb{R}^d. \tag{4.52}$$

Hence,

$$\begin{aligned}
&\lambda_n V_{\lambda_n}(0)\\
&= -\lambda_n(V_{\lambda_n} - V_{\lambda_n}(0))\\
&\quad + \sup_{a\in\mathbb{A}}\left\{\eta \int_{\mathbb{R}^{d'}} [V_{\lambda_n}(\cdot + b(\cdot, a, e)) - V_{\lambda_n}(0) - (V_{\lambda_n} - V_{\lambda_n}(0))]\nu(\mathrm{d}e) + r(\cdot, a)\right\}
\end{aligned}$$

on $\mathbb{R}^d$, and passing to the limit (recall Assumption 4.1 and that ν is a probability measure) implies that

$$c = \sup_{a\in\mathbb{A}}\left\{\eta \int_{\mathbb{R}^{d'}} [\mathrm{w}(\cdot + b(\cdot, a, e)) - \mathrm{w}]\nu(\mathrm{d}e) + r(\cdot, a)\right\}, \quad \text{on } \mathbb{R}^d.$$

□

We now have to prove that the constant c defined above equals $\rho^*(0)$ and that only $(w, \rho^*(0))$ solves (4.10), up to restricting to functions with linear growth taking the value 0 at 0.

Lemma 4.A.4.
Let Assumptions 4.1 to 4.3 hold. Let $(\tilde{w}, \tilde{\rho}) \in \mathcal{C}^0_{\text{lin}} \times \mathbb{R}$ be a solution of the ergodic equation

$$\tilde{\rho} = \sup_{a \in \mathbb{A}} \left\{ \eta \int_{\mathbb{R}^{d'}} [\tilde{w}(\cdot + b(\cdot, a, e)) - \tilde{w}] \nu(de) + r(\cdot, a) \right\}, \quad \text{on } \mathbb{R}^d.$$

Then, ρ^* is constant and equal to $\tilde{\rho}$. In particular, the constant c of Lemma 4.A.3 is equal to ρ^*.

Proof. Let us fix $x \in \mathbb{R}^d$.

1. By Lemma 4.A.3 and [24, Prop. 7.33, p.153], we can find a measurable map $x' \in \mathbb{R}^d \mapsto \hat{a}(x') \in \mathbb{A}$ such that

$$\tilde{\rho} = \eta \int_{\mathbb{R}^{d'}} [\tilde{w}(\cdot + b(\cdot, \hat{a}(\cdot), e)) - \tilde{w}] \nu(de) + r(\cdot, \hat{a}(\cdot)), \quad \text{on } \mathbb{R}^d.$$

   Let $\hat{X}$ denote the solution of (4.3) associated to $\hat{\alpha} := \hat{a}(\hat{X}_{\_})$ and the initial condition x. Then, Itô's Lemma implies that

$$\mathbb{E}\left[\tilde{w}(\hat{X}_t) - \tilde{w}(x) + \frac{1}{\eta} \int_0^t r(\hat{X}_{s-}, \hat{\alpha}_s) dN_s \right] = \tilde{\rho} t, \ t \geq 0.$$

   Moreover, since $\tilde{w}$ has linear growth, there exists $C > 0$ such that

$$\mathbb{E}\left[|\tilde{w}(\hat{X}_t) - \tilde{w}(x)| \right] \leq C \mathbb{E}\left[\|\hat{X}_t\| + \|x\| \right].$$

   By Lemma 4.A.2, $\mathbb{E}[\|\hat{X}_t\|]/t \to 0$ as $t \to \infty$ since $p_\xi \geq 1$. Then, the above implies that

$$\lim_{t \to \infty} \frac{1}{\eta t} \mathbb{E}\left[\int_0^t r(\hat{X}_{s-}, \hat{\alpha}_s) dN_s \right] = \tilde{\rho}.$$

2. Conversely, for any $\alpha \in \mathcal{A}$,

$$\mathbb{E}\left[\tilde{w}(X_t^{x,\alpha}) - \tilde{w}(x) + \frac{1}{\eta} \int_0^t r(X_{s-}^{x,\alpha}, \alpha_s) dN_s \right] \leq \tilde{\rho} t, \ t \geq 0.$$

By Lemma 4.A.2 and the linear growth of $\tilde{w}$ again, we deduce that

$$\limsup_{t\to\infty} \frac{1}{\eta t}\mathbb{E}\left[\int_0^t r(X_{s-}^{x,\alpha}, \alpha_s)\mathrm{d}N_s\right] \leq \tilde{\rho}.$$

3. Combining 1. and 2. implies that $\tilde{\rho} = \rho^*(x)$. By arbitrariness of $x \in \mathbb{R}^d$, ρ^* is constant. □

We are now in a position to prove our second convergence result, and therefore to complete the proof of Theorem 4.2.1.

Lemma 4.A.5.
Let Assumptions 4.1 to 4.3 hold. Then, there exists a sequence $(T_n)_{n\geq 1}$ going to $+\infty$ such that $(T_n^{-1}V_{T_n}(0,\cdot))_{n\geq 1}$ converges uniformly on compact sets to $\rho^*(0)$.

Proof. The proof follows from the same arguments as in [14, Prop. VI.1] except that in their case the convergence holds uniformly on $\mathbb{R}^d$. Let $(\lambda_n)_{n\geq 1}$ be as in Lemma 4.A.3 and set $T_n := \delta/\lambda_n$ for some $\delta \in (0,1)$, so that $\lambda_n \to 0$ and $T_n \to \infty$ as $n \to \infty$. Fix $x \in \mathbb{R}^d$. By Lemmas 4.A.1 and 4.A.2, we can find $C > 0$ such that $\mathbb{E}[|V_{\lambda_n}(X_t^{x,\alpha}) - V_{\lambda_n}(x)|] \leq C(1+|x|)$ uniformly in $\alpha \in \mathcal{A}$ and for all $x \in \mathbb{R}^d$, and $t \geq 0$. Arguing as in the proof of [14, Prop. VI.1], we then deduce from the Dynamic Programming Principle applied to V_{λ_n}, see e.g. [35], Lemmas 4.A.1 and 4.A.2 and Assumption 4.1 that, for some $C' > 0$ that does not depend on n,

$$\left|\rho^*(1-e^{-\delta}) - \frac{\delta}{T_n}V_{T_n}(0,x)\right| \leq 2\left|\lambda_n V_{\lambda_n}(x) - \rho_\varepsilon^*\right| + \lambda_n C'(1+|x|).$$

It remains to divide the above by δ, send $n \to \infty$ and use Lemmas 4.A.3 and 4.A.4 to obtain that

$$\rho^*\frac{(1-e^{-\delta})}{\delta} \leq \liminf_{n\to\infty}\frac{1}{T_n}V_{T_n}(0,x) \leq \limsup_{n\to\infty}\frac{1}{T_n}V_{T_n}(0,x) \leq \rho^*\frac{(1-e^{-\delta})}{\delta},$$

and we conclude by arbitrariness of $\delta \in (0,1)$. The fact that the convergence is uniform on compact sets follows from the above and Lemma 4.A.3. □

4.B Estimates for Elliptic HJB Equations Without Control of the Volatility Part

In this section, we collect standard estimates on elliptic HJB equations associated with infinite horizon optimal control problems of a diffusion, in which

there is no control on the volatility part. This is a specific class of quasi-linear equations whose analysis is standard. Our focus here is on the growth rate of local $\mathcal{C}_b^{2,\gamma}$-estimates in the case where the solution is already known to be Lipschitz. We follow closely the arguments of [61] that consider compact domains and insist only on the points where the Lipschitz continuity property is used.

As usual, we first consider linear equations of the form

$$0 = \langle \mathbf{b}, \mathrm{D}u^{\mathsf{T}} \rangle + \frac{1}{2} \mathrm{Tr}\left[\mathfrak{a}\mathrm{D}^2 u\right] - \lambda u - \mathfrak{f} \text{ on } \mathbb{R}^d. \tag{4.53}$$

We fix $M > 0$ and a modulus of continuity ϱ (i.e. a real valued map on $\mathbb{R}^d$ that is continuous at 0 and such that $\varrho(0) = 0$). We let $\mathfrak{S}(M, \varrho)$ denote the collections of real-valued maps $u \in \mathcal{C}^2$ such that $u(0) = 0$, $\|\mathrm{D}u\|_\infty \le M$ and that are strong solutions of (4.53) with coefficients satisfying:

(a) $\lambda \in [0, 1]$,

(b) $(\mathbf{b}, \mathfrak{f}) : \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}$ is measurable and $[\mathbf{b}]_{\mathcal{C}^0_{\mathrm{lin}}} + \|\mathfrak{f}\|_\infty \le M$,

(c) $\mathfrak{a} : \mathbb{R}^d \to \mathbb{S}^d$ is bounded by M and admits ϱ as a modulus of continuity,

(d) $\inf\{\xi^{\mathsf{T}} \mathfrak{a} \xi : \xi \in \mathbb{R}^d, \|\xi\| = 1\} \ge 1/M$.

Hereafter, we use the convention $0/0 = 0$.

Lemma 4.B.1.
For each $\gamma \in (0, 1)$, there exists $K^\gamma_{M,\varrho} > 0$ such that any $u \in \mathfrak{S}(M, \varrho)$ satisfies

$$\|u\|_{\mathcal{C}_b^{1,\gamma}(B_2(x))} \le K^\gamma_{M,\varrho}(1 + \|x\|), \quad \textit{for all } x \in \mathbb{R}^d.$$

Proof.

1. Given $p > 1$, we first estimate $\|u\|_{W^{2,p}(B_2(x))}$ in which $\|\cdot\|_{W^{2,p}(B_2(x))}$ denotes the norm associated to the Sobolev space $W^{2,p}(B_2(x))$. We follow the proof of [61, Thm. 9.11]. Fix $x_0 \in B_2(x)$. By [61, (9.37)], for any $v \in W^{2,p}(B_3(x_0))$ supported in some $B_R(x_0) \subset B_3(x)$, $R > 0$, there is $C_1 > 0$, that depends only p, such that

$$\begin{aligned} &\|\mathrm{D}^2 v\|_{L^p(B_R(x_0))} \\ &\le C_1 M \left(\sup_{B_R(x_0)} \|\mathfrak{a} - \mathfrak{a}(x_0)\| \, \|\mathrm{D}^2 v\|_{L^p(B_R(x_0))} + \|\mathrm{Tr}[\mathfrak{a}\mathrm{D}^2 v]\|_{L^p(B_R(x_0))} \right), \end{aligned}$$

in which $\|\cdot\|_{L^p(B_R(x_0))}$ denotes the usual norm of the L^p-space associated to the Lebesgue measure on $B_R(x_0)$.

The uniform continuity of $\mathfrak{a}$ implies that there exists an $R > 0$ small enough, which only depends on p, M and ϱ, such that $|\mathfrak{a} - \mathfrak{a}(x_0)| \leq 1/2MC_1$ on $B_R(x_0)$, so that the above implies that

$$\|\mathrm{D}^2 v\|_{L^p(B_R(x_0))} \leq 2C_1 M \|\mathrm{Tr}[\mathfrak{a}\mathrm{D}^2 v]\|_{L^p(B_R(x_0))}. \tag{4.54}$$

Take $u \in \mathfrak{S}(M, \varrho)$ a solution to (4.53) in $B_3(x)$, applying (4.54) yields

$$\begin{aligned}&\|\mathrm{D}^2 u\|_{L^p(B_R(x_0))}\\ &\leq C_2 \left(\|\mathfrak{f}\|_{\mathcal{C}_b^0(B_3(x))} + \lambda \|u\|_{\mathcal{C}_b^0(B_3(x))} + \|\mathbf{b}\|_{\mathcal{C}_b^0(B_3(x))} \|\mathrm{D}u^{\mathsf{T}}\|_{\mathcal{C}_b^0(B_3(x))} \right)\end{aligned}$$

for some $C_2 > 0$ that only depends on M, p and ϱ. From the definition of $\mathfrak{S}(M, \varrho)$, it follows that there is $C_3 > 0$, independent of x_0, such that

$$\|u\|_{W^{2,p}(B_R(x_0))} \leq C_3(1 + \|x\|),$$

and, by covering $B_2(x)$ with finitely many balls of radius less that R, one obtains

$$\|u\|_{W^{2,p}(B_2(x))} \leq C_4(1 + \|x\|)$$

for some C_4 that depends only on p, M and ϱ.

2. Using an embedding theorem, see e.g. [61, Thm. 7.26], we can find $\bar{K}^{\gamma,p} > 0$ such that

$$\|u\|_{\mathcal{C}_b^{1,\gamma}(B_2(x))} \leq \bar{K}^{\gamma,p} \|u\|_{W^{2,p}(B_2(x))}, \ \forall\, u \in \mathfrak{S}(M, \varrho),\ x \in \mathbb{R}^d,$$

for all $p \in \mathbb{N}$ such that $0 < d/p < 1$ and $\gamma \in (0, 1 - d/p)$. Given $\gamma \in (0, 1)$, the required result follows by combining the above for some p large enough. □

Let us now turn to the quasilinear case

$$0 = \hat{\mathbf{b}}(\cdot, \mathrm{D}u^{\mathsf{T}}) + \frac{1}{2}\mathrm{Tr}[\mathfrak{a}\mathrm{D}^2 u] - \lambda u \text{ on } \mathbb{R}^d, \tag{4.55}$$

in which

$$\hat{\mathbf{b}}(x, y) := \langle \mathbf{b}(x, y), y \rangle - \mathfrak{f}(x, y),\ (x, y) \in \mathbb{R}^d \times \mathbb{R}^d.$$

We again fix $M > 0$, and take $\rho = (\rho_1, \rho_2) \in (0, 1]^2$, and let $\tilde{\mathfrak{S}}(M, \rho)$ denote the collection of real-valued maps $u \in \mathcal{C}^2$ such that $u(0) = 0$, $\|\mathrm{D}u\| \leq M$, and that are solutions of (4.55) for some coefficients satisfying:

(a) $\lambda \in [0, 1]$,

(b) $(\mathbf{b}, \mathfrak{f}) : \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}$ is measurable and $[\mathbf{b}]_{\mathcal{C}^0_{\rm lin}(\mathbb{R}^{2d})} + \|\mathfrak{f}\|_{\mathcal{C}^0_b(\mathbb{R}^{2d})} \le M$,

(c) $\mathfrak{a} : \mathbb{R}^d \to \mathbb{S}^d$ is measurable and bounded by M.

(d) $\inf\{\xi^{\mathsf{T}} \mathfrak{a}\, \xi : \xi \in \mathbb{R}^d, \|\xi\| = 1\} \ge 1/M$,

(e) for all $x, x' \in \mathbb{R}^d$ such that $\|x - x'\| \le 1$ and all $y, y' \in \mathbb{R}^d$:

$$\|\mathfrak{a}(x) - \mathfrak{a}(x')\| + \left|\hat{\mathbf{b}}(x, y) - \hat{\mathbf{b}}(x', y')\right| \le M\left(\|x - x'\|^{\rho_1} + \|y - y'\|^{\rho_2}\right).$$

Lemma 4.B.2.
Fix $\gamma \in (0, \rho_1 \wedge \rho_2)$. Then, there exists $\tilde{K}^\gamma_{M,\rho} > 0$ such that any $u \in \tilde{\mathfrak{S}}(M, \rho)$ satisfies

$$\|u\|_{\mathcal{C}^{2,\gamma}_b(B_1(x))} \le \tilde{K}^\gamma_{M,\rho}(1 + \|x\|), \quad \textit{for all } x \in \mathbb{R}^d.$$

Proof. Fix $x \in \mathbb{R}^d$. Since $\|\mathrm{D}u\| \le M$, by Lemma 4.B.1 applied to the coefficient $x' \in \mathbb{R}^d \mapsto (\mathbf{b}(x', \mathrm{D}u^{\mathsf{T}}(x')), \mathfrak{a}(x'), \mathfrak{f}(x', \mathrm{D}u^{\mathsf{T}}(x')))$ in place of $(\mathbf{b}, \mathfrak{a}, \mathfrak{f})$, for each $\gamma \in (0, 1)$, we can find $C_\gamma > 0$ such that

$$\|u\|_{\mathcal{C}^{1,\gamma}_b(B_2(x))} \le C_\gamma(1 + \|x\|) \text{ for all } x \in \mathbb{R}^d. \tag{4.56}$$

It then follows from [61, Thm. 9.19] that $u \in \mathcal{C}^{2,\gamma}_b(B_2(x))$ for any $\gamma \in (0, \rho_1 \wedge \rho_2)$.

To obtain an associated estimate, we turn to the proof of [61, Thm. 6.2] which we apply to the solution $w = u$ of the linear equation

$$Lw := \frac{1}{2}\mathrm{Tr}\left[\mathfrak{a}\mathrm{D}^2 w\right] = -\hat{\mathbf{b}}\left(\cdot, \mathrm{D}w^{\mathsf{T}}\right) + \lambda w,$$

in our particular setting. Fix $x_0 \in B_2(x)$, and consider the constant coefficient equation $L_0 w := (1/2) \times \mathrm{Tr}[\mathfrak{a}(x_0)\mathrm{D}^2 w] = F$ with

$$F(z) := \frac{1}{2}\mathrm{Tr}\left[(\mathfrak{a}(x_0) - \mathfrak{a}(z))\mathrm{D}^2 u(z)\right] - \hat{\mathbf{b}}\left(z, \mathrm{D}u^{\mathsf{T}}(z)\right) + \lambda u(z),\ z \in \mathbb{R}^d.$$

We first introduce some notations. For $\mathfrak{O} \subset \mathbb{R}^d$, $\gamma \in (0, 1)$, and $f \in \mathcal{C}^{2,\gamma}(\mathfrak{O})$ define the following norm and Schauder semi-norm respectively as follows:

$$\begin{aligned}
\|f\|^{(2)}_{0,\gamma,\mathfrak{O}} &:= \sup_{z \in \mathfrak{O}} d_z^2 \|f(z)\| + \sup_{(z,z') \in \mathfrak{O}^2} d_{z,z'}^{2+\gamma} \frac{\|f(z) - f(z')\|}{\|z - z'\|^\gamma} \\
[f]^*_{2,\gamma,\mathfrak{O}} &:= \sup_{(z,z') \in \mathfrak{O}^2} d_{z,z'}^{2+\gamma} \frac{\|\mathrm{D}^2 f(z) - \mathrm{D}^2 f(z')\|}{\|z - z'\|^\gamma}
\end{aligned} \tag{4.57}$$

$$[f]^*_{2,\mathfrak{O}} := \sup_{z\in\mathfrak{O}} d_z^2 \|\mathrm{D}^2 f(z)\|\,, \tag{4.58}$$

where d_z is the distance of z to the boundary of $\mathfrak{O}$ and $d_{z,z'} := d_z \wedge d_{z'}$ for any $(z,z') \in \mathfrak{O}^2$.

We now fix $\gamma \in (0, \rho_1 \wedge \rho_2)$. Let $\mu \in (0, 1/2]$ and set $\mathfrak{O} := B_2(x)$. Fix $y_0 \in B_2(x)$ such that $d_{x_0} \le d_{y_0}$ (without loss of generality) and set $B := B_{\mu d_{x_0}}(x_0)$. Then, [61, Lemma 6.1 (a.)] (see [61, (6.16)] for details) applied to $L_0 w = F$ implies that

$$\begin{aligned} d_{x_0,y_0}^{2+\gamma} \frac{\|\mathrm{D}^2u(x_0) - \mathrm{D}^2u(y_0)\|}{\|x_0 - y_0\|^\gamma} &= d_{x_0}^{2+\gamma} \frac{\|\mathrm{D}^2u(x_0) - \mathrm{D}^2u(y_0)\|}{\|x_0 - y_0\|^\gamma} \\ &\le \frac{C_1^\gamma}{\mu^{2+\gamma}} \left(\|u\|_{\mathcal{C}_b^0(B_2(x))} + \|F\|^{(2)}_{0,\gamma,B} \right) + \frac{4}{\mu^\gamma} [u]^*_{2,B_2(x)} \end{aligned}$$

for some $C_1^\gamma > 0$, which only depends on $\gamma \in (0, \rho_1 \wedge \rho_2)$. Then, using [61, (6.8)] yields

$$\begin{aligned} d_{x_0,y_0}^{2+\gamma} \frac{\|\mathrm{D}^2u(x_0) - \mathrm{D}^2u(y_0)\|}{\|x_0 - y_0\|^\gamma} \le& \frac{C_1^\gamma}{\mu^{2+\gamma}} \left(\|u\|_{\mathcal{C}_b^0(B_2(x))} + \|F\|^{(2)}_{0,\gamma,B} \right) \\ &+ 4 \left(C_1(\mu) \|u\|_{\mathcal{C}_b^0(B_2(x))} + \mu^\gamma [u]^*_{2,\gamma,B_2(x)} \right) \end{aligned}$$

for some $C_1(\mu) > 0$ that only depends on μ. The Schauder estimate then comes from bounding term by term $\|F\|^{(2)}_{0,\gamma,B}$. First, we argue as for [61, (6.19)], using (c) and (e) in the definition of $\tilde{\mathfrak{S}}(M,\rho)$, to obtain

$$\|\mathrm{Tr}[(\mathfrak{a}(x_0) - \mathfrak{a})\mathrm{D}^2u]\|^{(2)}_{0,\gamma,B} \le C_2^\gamma \mu^{2+\gamma} \left[C_2(\mu) \|u\|_{\mathcal{C}_b^0(B_2(x))} + \mu^\gamma [u]^*_{2,\gamma,B_2(x)} \right]$$

for some $C_2^\gamma, C_2(\mu) > 0$ which only depend on γ and μ. Second, we combine (4.56) with items (a) and (e) in the definition of $\tilde{\mathfrak{S}}(M,\rho)$ to obtain that

$$\|\hat{\mathbf{b}}(\cdot, \mathrm{D}u^\mathsf{T}) - \lambda u\|^{(2)}_{0,\gamma,B_2(x)} \le C_3^\gamma (1 + \|x\|)$$

for some $C_3^\gamma > 0$, that only depends on γ.

Combining the above with (4.56) and using the arbitrariness of $x_0, y_0 \in B_2(x)$ leads to

$$\begin{aligned} [u]^*_{2,\gamma,B_2(x)} \le& \frac{C_1^\gamma}{\mu^{2+\gamma}} \left(\|u\|_{\mathcal{C}_b^0(B_2(x))} + C_3^\gamma (1 + \|x\|) \right) \\ &+ \frac{C_1^\gamma C_2^\gamma}{2} \left[C_2(\mu) \|u\|_{\mathcal{C}_b^0(B_2(x))} + \mu^\gamma [u]^*_{2,\gamma,B_2(x)} \right] \end{aligned}$$

$$+ 4\left(C_1(\mu)\,\|u\|_{\mathcal{C}_b^0(B_2(x))} + \mu^\gamma [u]^*_{2,\gamma,B_2(x)}\right)$$

We now take $\mu > 0$ small enough and recall (4.56) to obtain, for each $0 < \gamma < \rho_1 \wedge \rho_2$, a constant $C_4^\gamma > 0$, independent on x, such that

$$[u]^*_{2,\gamma,B_2(x)} \leq C_4^\gamma\,(1 + \|x\|)$$

and we conclude by using [61, (6.9)] and the fact that the distance between a point of $B_1(x)$ and the boundary of $B_2(x)$ is a least 1. □

Chapter 5

Near-Continuous Time Reinforcement Learning with Continuous States

To design Reinforcement Learning algorithms for high-frequency systems with continuous state-action spaces, we consider interactions with a pure-jump process of arbitrary transition kernel driven by a Poisson clock of frequency ε^{-1}. This model captures arbitrary time scales, from discrete ($\varepsilon = 1$) to continuous time ($\varepsilon \downarrow 0$), and richer systems than the rigid discrete and Linear-Quadratic frameworks. We show that the celebrated optimism protocol applies when the sub-tasks (learning and planning) can be performed effectively. Learning is tackled within the eluder dimension framework and we propose an approximate planning method based on a diffusive limit approximation of the jump process. Overall, our algorithm enjoys a regret of order $\tilde{\mathcal{O}}\big(\varepsilon^{1/2}T + \sqrt{T}\big)$. As the frequency of interactions blows up, the approximation error $\varepsilon^{1/2}T$ vanishes, showing that $\tilde{\mathcal{O}}\big(\sqrt{T}\big)$ is attainable in near-continuous time[a].

[a] This Chapter is under review as an article at the $3R^{\text{th}}$ *International conference on Algorithmic Learning Theory* (ALT), and appeared at the 16^{th} European Workshop on Reinforcement Learning (EWRL), see [5]

* * *

Contents

* * *

5.1 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. [82]) 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. [112, § 4.2]).

Unfortunately, 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. See also Sections 1.1 and 1.3.

A natural approach to planning in such systems is to directly model the problem in continuous time. This is the common approach in finance, see for instance [43, 47, 96]. 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 parametrised 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 exploration-exploitation trade-off.

For discrete-time dynamics on finite state-action spaces, the planning problem falls under the umbrella of Markov Decision Processes (MDPs), an extensive review of which is available in [102]. 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. [22, § 4.6] for a review of the challenges. Perhaps the only exception that retains those nice theoretical and computational properties is the celeb-

rated Linear Quadratic (LQ) framework [69]. 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 [13, 14, 85], both on the theoretical aspects as well as numerical solvers based on Partial Differential Equations (PDEs), another storied field [19, 28, 78]. The near-continuous time framework lies between the two theories, and the results of Chapter 4 show how to navigate between them and approximately solve the planning problem in the high-frequency interactions regime by solving its diffusive counterpart.

Similar to the planning problem, the natural way to push learning 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. [1]. Subsequent work, notably [97, 107], 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 [107], 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.

Optimism in the Face of Uncertainty (OFU) has proved highly successful in sequential decision-making from bandits to RL. The works of [16, 20, 68] showed how to extend the celebrated Upper Confidence Bound (UCB) [15] algorithm from bandits to finite MDPs; later, extensions were made to continuous state in the LQ setting, see e.g. [2, 7, 44] and references therein. Extension from bandit to MDPs 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 Chapter 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 [107] 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 sta-

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

bility arguments. Overall, our algorithm enjoys near-optimal performance as its regret scales[2] 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.

5.2 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 $a \in \mathbb{A} \subset \mathbb{R}^{d_{\mathbb{A}}}$, $d_{\mathbb{A}} \in \mathbb{N}^*$, 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. 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 parametrises the mean wait time between events and gives us direct control of the frequency of interactions.

***Remark* 5.2.1.** *While the additive noise structure is a design choice that simplifies the analysis, the choice of parametrising 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.*

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 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 $X^{\alpha,\theta^*} := X^{x_0,\alpha,\theta^*}$ denote the pathwise-unique solution of

$$\begin{cases} X^{\alpha,\theta^*}_{\tau_n} = X^{\alpha,\theta^*}_{\tau_{n-1}} + \mu_{\theta^*}(X^{\alpha,\theta^*}_{\tau_{n-1}}, \alpha_{\tau_{n-1}}) + \Sigma\xi_n \\ X^{\alpha,\theta^*}_{\tau_0} = x_0 \end{cases} . \tag{5.1}$$

[2] Given maps $g : \mathbb{R}_+ \to \mathbb{R}_+^*$ and $f : \mathbb{R}_+ \to \mathbb{R}_+$, $f = \tilde{\mathcal{O}}(g)$ if there is a finite $n \in \mathbb{N}^*$ such that $f/\log(\cdot)^n = \mathcal{O}(g)$. In our case, these logarithmic factors in T appear only in the $\tilde{\mathcal{O}}(\sqrt{T})$ term.

See (4.3) for an alternative definition[3].

In (5.1), 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.

In our model-based paradigm, ignorance about the system is condensed to a single parameter set $\Theta \subset \mathbb{R}^{d_\Theta}$, $d_\Theta \in \mathbb{N}^*$ containing the unknown nominal parameter θ^*. To single out the RL challenges, we further assume that θ^* only affects the drift assuming other quantities (Σ, ε, 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

$$\rho^{\alpha}_{\theta^*}(x_0) := \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{n=1}^{N_T} r(X^{\alpha,\theta^*}_{\tau_n}, \alpha_{\tau_n})\right]. \tag{5.2}$$

The goal of the agent is to accumulate as much reward as possible, i.e. to try 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 an online learning algorithm generating $\alpha \in \mathcal{A}$ according to its regret.

Definition 5.1.
For any $T \in \mathbb{R}_+^*$, $x_0 \in \mathbb{R}^d$, and $\alpha \in \mathcal{A}$, the regret of α is

$$\mathcal{R}_T(\alpha) := T\rho^*_{\theta^*}(x_0) - \sum_{n=1}^{N_T} r(X^{\alpha,\theta^*}_{\tau_n}, \alpha_{\tau_n}). \tag{5.3}$$

Noticing that N_T is the number of events up to time T, the definitions of the optimal performance (5.2) and the regret (5.3) 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.

5.2.1 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 dynamics and reward. This

[3]Note that relative to Chapters 3 and 4 we choose to work here with adapted controls instead of predictable ones, up to a modification in the reward function

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

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

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 $\Sigma\Sigma^\intercal$ to be linear in ε. Known as the diffusive regime, this preserves stochasticity[4] 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 Euclidean norm on $\mathbb{R}^d$ and $\|\cdot\|_{\mathrm{op}}$ for the operator norm on $\mathbb{R}^{d\times d}$ associated to $\|\cdot\|$.

Assumption 5.1.
The map $(\bar{\mu}, \bar{r})$ is continuous, and there is $L_0 > 0$ such that for all $(\theta, a) \in \Theta \times \mathbb{A}$

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

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

Assumption 5.1 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 5.2). Finally, we assume non-degeneracy of the noise by requiring $\bar{\Sigma}$ to be full rank.

We conclude with Assumption 5.2 to ensure the stability of the state process. Let $\mathbb{R}^d_* := \mathbb{R}^d \backslash \{0\}$ 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.

[4] 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 [54].

Assumption 5.2.

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 $\varepsilon \in (0, 1)$:

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

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

$$\text{(iii)} \quad \mathscr{V}(\psi^\varepsilon_\theta(x, a) - \psi^\varepsilon_\theta(x', a)) \leq (1 - \varepsilon \mathfrak{c}_{\mathscr{V}}) \mathscr{V}(x - x') \,. \tag{5.4}$$

in which $\psi^\varepsilon_\theta(x, a) := x + \varepsilon \bar{\mu}(x, a)$.

Assumption 5.2 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 jumps.

Stability theory has been extensively studied in the special case of linear dynamics. In this case, we recover Assumption 5.2 from the Continuous Algebraic Riccati Equation (CARE), see e.g. [81, § 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 5.2. Indeed, conditions (i) and (ii) follow as $\mathscr{V}$ is a norm and, for $\varepsilon \leq 1/2\lambda_{\max}(P)$, we have

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

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

5.3 Contributions

5.3.1 Algorithm and guarantees

Our main contribution is a demonstration of the Optimism in the Face of Uncertainty protocol in the near-continuous time continuous state-action RL

problem. The ingredients of OFU are: learning from accumulated data to design confidence sets; using lazy updates to trade-off policy revision and learning guarantees; and planning amongst plausible parametrisations. We summarise this protocol in Algorithm 1.

Algorithm 1 `OFU-Diffusion`

Input: confidence level δ, initial state x_0, initial control ϖ_0
for $n \in \mathbb{N}^*$ **do**
  At time τ_n, receive $r(X^{\varpi,\theta^*}_{\tau_{n-1}}, \varpi_{\tau_{n-1}})$ and $X^{\varpi,\theta^*}_{\tau_n}$.
  if n satisfies (5.7) **then**
    $n_k \leftarrow n$, $k \leftarrow k+1$,
    Compute $\hat{\theta}_{n_k}$ using (5.5) and $\mathcal{C}_{n_k}(\delta/3)$ with (5.6).
    $\tilde{\theta}_k \leftarrow \mathrm{argmax}_{\theta \in \mathcal{C}_{n_k}(\delta/3)} \bar{\rho}^*_\theta$
    $\pi_k \leftarrow \bar{\pi}^*_{\tilde{\theta}_k}$ using (5.10)
  end if
  Play $\varpi_{\tau_n} := \pi_k(X^{\varpi,\theta^*}_{\tau_n})$.
end for

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 Squares fit and the associated confidence set $\mathcal{C}_{n_k}(\delta)$, defined (for $\beta_n(\delta)$ given in (5.14) for all $n \in \mathbb{N}$) by

$$\hat{\theta}_{n_k} \in \operatorname*{argmin}_{\theta \in \Theta} \sum_{n=0}^{n_k-1} \left\| X^{\varpi,\theta^*}_{\tau_{n+1}} - X^{\varpi,\theta^*}_{\tau_n} - \mu_\theta(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) \right\|^2, \tag{5.5}$$

$$\mathcal{C}_{n_k}(\delta) := \left\{ \theta \in \Theta : \sqrt{\sum_{n=0}^{n_k-1} \left\| \mu_\theta(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) - \mu_{\hat{\theta}_{n_k}}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) \right\|^2} \le \beta_{n_k}(\delta) \right\}. \tag{5.6}$$

Our episodic scheme follows the same rationale as in [2, 68], and triggers updates as soon as enough information is collected. Formally, it constructs a sequence of episodes whose starting times are defined by $n_0 := 0$ and, for any $k \in \mathbb{N}$, n_{k+1} is the first time n satisfying (5.7)

$$\sqrt{\sup_{\theta \in \mathcal{C}_{n_k}(\delta)} \sum_{i=0}^{n} \left\| \mu_\theta(X^{\varpi,\theta^*}_{\tau_i}, \varpi_{\tau_i}) - \mu_{\hat{\theta}_{n_k}}(X^{\varpi,\theta^*}_{\tau_i}, \varpi_{\tau_i}) \right\|^2} > 2\beta_n(\delta). \tag{5.7}$$

At the heart of our proposal is the way in which we address the optimistic planning, detailed in Section 5.3.4. 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

$$\bar{\rho}_\theta^{\bar{\alpha}}(x_0) := \liminf_{T\to+\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T \bar{r}(\bar{X}_s^{\bar{\alpha},\theta}, \bar{\alpha}_t)\mathrm{d}s\right] \tag{5.8}$$

in which

$$\begin{cases} \mathrm{d}\bar{X}_t^{\bar{\alpha},\theta} = \bar{\mu}_\theta(\bar{X}_t^{\bar{\alpha},\theta}, \bar{\alpha}_t)\mathrm{d}t + \bar{\Sigma}\mathrm{d}W_t \\ \bar{X}_0^{\bar{\alpha},\theta} = x_0 \end{cases} \tag{5.9}$$

in which W denotes a $\mathbb{P}$-Brownian motion, $\bar{\mathbb{F}}$ its filtration and $\bar{\mathcal{A}}$ the set of $\mathbb{A}$-valued $\bar{\mathbb{F}}$-predictable processes. This diffusive problem gives us an optimality criterion and associated optimal control[5]:

$$\bar{\rho}_\theta^*(x_0) := \sup_{\alpha\in\mathcal{A}} \bar{\rho}_\theta^{\bar{\alpha}}(x_0) \text{ and } \bar{\pi}_\theta^* \circ \bar{X}^{\bar{\pi}_\theta^*,\theta} \in \operatorname*{argmax}_{\bar{\alpha}\in\bar{\mathcal{A}}} \bar{\rho}_\theta^{\bar{\alpha}}(x_0) \tag{5.10}$$

which approximates the original jump-process problem $\rho_\theta^*(x_0)$. This problem admits a Hamilton-Jacobi-Bellman equation (given in (5.19) below) characterising an optimal policy $\bar{\pi}_\theta^* : \mathbb{R}^d \to \mathbb{A}$ which yields a computable optimal Markov control for (5.10).

Theorem 5.3.1.
Under Assumptions 5.1 and 5.2, 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

$$\mathcal{R}_T(\varpi) \leq 2C_\gamma \varepsilon^{\frac{\gamma}{2}} T + C\sqrt{\mathrm{d}_{\mathrm{E},T\varepsilon^{-1}} \log(\mathscr{N}^{\varepsilon}_{T\varepsilon^{-1}}) T \log(T\delta^{-1})} \tag{5.11}$$

with probability at least $1-\delta$, in which $\mathrm{d}_{\mathrm{E},T\varepsilon^{-1}}$ is the $2\varepsilon/\sqrt{T}$-eluder dimension (see [107, Def. 4.] and (5.56) in Section 5.5.2) 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 $\varepsilon^2\|\bar{\Sigma}\|_{\mathrm{op}}^2/T$-log-covering number of this same restricted class.

Theorem 5.3.1 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 Chapter 4 and $C_\gamma < +\infty$ holds for $\gamma < 1$ (see Lemma 4.B.1). Quantifying the behaviour of C_γ as $\gamma \uparrow 1$ for arbitrary dynamics is technically intricate. Nevertheless, our bound indicates that the long run approximation error vanishes

[5]Henceforth, we will use the obvious notational confusion between the policy $\bar{\pi}_\theta^*$ and the control process $\bar{\pi}_\theta^* \circ \bar{X}^{\bar{\pi}_\theta^*,\theta}$ it generates.

as $\varepsilon \downarrow 0$ almost as fast as $\varepsilon^{1/2}$. The second term quantifies all other sources of error and exhibits the expected scaling in the complexity measures of [107], in terms of both eluder dimension and log-covering numbers, as well as the $\sqrt{T}$ horizon dependency.

5.3.2 Stability results

Working with unbounded processes and generic drifts requires us to prevent state blow-up, which could degrade regret regardless of learning. In Proposition 5.3.1 we combine the Lyapunov stability of (5.4) with concentration arguments to show that unstable trajectories can only happen with low probability. A detailed proof is given in Section 5.4.

Proposition 5.3.1.
Let Assumptions 5.1 and 5.2 hold. Then, 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 $\theta \in \Theta$

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}_+} \frac{\|X_t^{\alpha,\theta}\|}{H_\delta(N_t)} \geq 1\right) \leq \delta. \tag{5.12}$$

Working on the high-probability event of Proposition 5.3.1 allows us to handle the unbounded state in the learning, planning, and optimism.

5.3.3 Learning results

The crux of our analysis is incorporating Proposition 5.3.1 into the NLLS method of [107] by refining it to be adaptive to the norm of the state process. For $R > 0$, let $\mathcal{B}(R) \subset \mathbb{R}^d$ denotes the Euclidean ball of radius R at 0. To adapt the log-covering number, we can work with H_δ by formally defining $\mathscr{N}_n^\varepsilon$ as the size of the smallest cover $\mathscr{C}_n^\varepsilon$ of $\mathscr{F}_\Theta := (\mu_\theta)_{\theta\in\Theta}$ such that

$$\sup_{\mu_1\in\mathscr{F}_\Theta} \min_{\mu_2\in\mathscr{C}_n^\varepsilon} \sup_{x\in\mathcal{B}(H_\delta(n))} \|\mu_1(x) - \mu_2(x)\| \leq \frac{\varepsilon\|\bar{\Sigma}\|_{\mathrm{op}}^2}{n}. \tag{5.13}$$

Restricting the domain of $\mathscr{F}_\Theta$ allows us to handle the richness of unbounded models and states while following [107] to define confidence sets. Let $\delta \in (0,1)$, set $\beta_0 := \varepsilon^{1/2}$, and define $\beta_n(\delta)$ as

$$\beta_0 \vee 2\varepsilon^{\frac{1}{2}}\|\bar{\Sigma}\|_{\mathrm{op}}\left(\sqrt{1 + 2\left(\sqrt{2\log\left(\frac{4\pi^2 n^3}{3\delta}\right)} + \sqrt{\kappa_n(\delta)}\right)} + \sqrt{\kappa_n(\delta)}\right) \tag{5.14}$$

for any $n \in \mathbb{N}^*$, in which

$$\kappa_n(\delta) := \log\left(\frac{2\pi^2 n^2 \varepsilon \mathscr{N}_n^\varepsilon}{3\delta}\left(\|\bar{\Sigma}\|_{\mathrm{op}}^2 + 8nL_0^2(1+H_\delta(n))\right)\right).$$

Using this choice $(\beta_n)_{n\in\mathbb{N}}$ and replacing n_k by n in (5.6) 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 5.3.2.

Proposition 5.3.2. **(Adapted from [97, Prop. 5])**
Under Assumptions 5.1 and 5.2, for any $x_0 \in \mathbb{R}^d$, and $\delta > 0$,

$$\mathbb{P}\left(\left\{\theta^* \in \bigcap_{n=1}^{\infty} \mathcal{C}_n(\delta)\right\} \cap \left\{\sup_{n\in\mathbb{N}^*} \frac{\left\|X_{\tau_n}^{\varpi,\theta^*}\right\|}{H_\delta(n)} \le 1\right\}\right) \ge 1-\delta, \tag{5.15}$$

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.

To adapt the ϵ-eluder dimension (defined for $\epsilon > 0$ in [97, Def. 3.]), which we denote $\dim_{\mathrm{E}}(\cdot,\epsilon)$, to our unbounded state we proceed along 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 $B_n := \mathcal{B}(\sup_{t\le\tau_n}\|X_t^{\varpi,\theta^*}\|)$, denoted by $\mathrm{d}_{\mathrm{E},n} := \dim_{\mathrm{E}}((f|_{B_n})_{f\in\mathscr{F}_\Theta}, 2\sqrt{\varepsilon/n})$. In Proposition 5.3.3 we obtain first and second-order prediction error bounds from this eluder dimension. In Proposition 5.3.3 the order notation $\tilde{\mathcal{O}}$ hides terms that are poly-logarithmic in N_t and $\mathrm{d}_{\mathrm{E},N_T}$ whose full details are given in Section 5.5.2.

Proposition 5.3.3.
Under Assumptions 5.1 and 5.2, 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 $1-\delta$

$$\sum_{n=1}^{N_t} \|\mu_{\hat{\theta}_n} - \mu_{\theta^*}\|(X_{\tau_n}^{\alpha,\theta^*}, \alpha_{\tau_n}) \le \tilde{\mathcal{O}}\left(\sqrt{\varepsilon \mathrm{d}_{\mathrm{E},N_T} \log(\mathscr{N}_{N_T}^\varepsilon) N_t} + \varepsilon \mathrm{d}_{\mathrm{E},N_T}\right), \tag{5.16}$$

and

$$\sum_{n=1}^{N_t} \|\mu_{\hat{\theta}_n} - \mu_{\theta^*}\|^2(X_{\tau_n}^{\alpha,\theta^*}, \alpha_{\tau_n}) \leq \tilde{\mathcal{O}}\left(d_{E,N_T} \log(\mathscr{N}_{N_T}^{\varepsilon})\right). \tag{5.17}$$

We leverage the second order bound (5.17) of Proposition 5.3.3 to define our lazy-update scheme (5.7). We show in Section 5.7 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.

5.3.4 Planning results

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 $(x_0, \alpha, \theta) \in \mathbb{R}^d \times \mathcal{A} \times \Theta$ by replacing θ^* by θ in (5.1) and (5.2). 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 5.3.4. (Adapted from Theorem 4.2.1 and Remark 4.2.2) Under Assumptions 5.1 and 5.2, 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

$$\varepsilon\rho_\theta^* = \max_{a \in \mathbb{A}} \left\{\mathbb{E}[W_\theta^*(x + \mu_\theta(x,a) + \Sigma\xi)] - W_\theta^*(x) + r(x,a)\right\}, \tag{5.18}$$

for any $x \in \mathbb{R}^d$;

(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 of (5.18), and $\pi_\theta^* \circ X^{\pi_\theta^*,\theta}$ is an optimal Markov control, i.e. $\rho_\theta^{\pi_\theta^*}(\cdot) \equiv \rho_\theta^*$.

Proposition 5.3.4.(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 (5.4). Points (ii) and (iii) show that there is an optimal policy, which can be computed by solving the Hamilton-Jacobi-Bellman equation (5.18). 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, (5.18) 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.

In the limit regime of interest, i.e. as $\varepsilon \downarrow 0$, the non-local behaviour of (5.18) 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. [13, 14].

Proposition 5.3.5. **(Adapted from Theorem 4.3.1)**
Under Assumptions 5.1 and 5.2, 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

$$\bar{\rho}_\theta^* = \max_{a \in \mathbb{A}} \left\{ \bar{\mu}_\theta(x,a)^\intercal \nabla \bar{W}_\theta^*(x) + \bar{r}(x,a) \right\} + \frac{1}{2} \operatorname{Tr}\left[\bar{\Sigma}\bar{\Sigma}^\intercal \nabla^2 \bar{W}_\theta^*(x) \right] \tag{5.19}$$

for any $x \in \mathbb{R}^d$.

(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 (5.19), and $\bar{\pi}_\theta^* \circ \bar{X}^{\bar{\pi}_\theta^*,\theta}$ is an optimal Markov control, i.e. $\bar{\rho}_\theta^{\bar{\pi}_\theta^*}(\cdot) \equiv \bar{\rho}_\theta^*$.

Proposition 5.3.5 ensures that the diffusive problem satisfies all the properties of Proposition 5.3.4 (ergodicity, optimal policy, and HJB equation). However, the HJB (5.19) is now a second-order (local) Partial Differential Equation 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 (5.18). Moreover, this diffusivePDE belongs to a well-studied family, both from the points of view of theory [61, 80] and of numerics [74, 76]. These facts motivate the use of these tools to construct approximate planning methods for (5.18) in the near-continuous time regime as $\varepsilon \downarrow 0$.

Proposition 5.3.6. **(Adapted from Theorem 4.3.2)**
Under Assumptions 5.1 and 5.2, for any $\gamma \in (0,1)$, there is a constant $C_\gamma > 0$, independent of ε, such that, for any $\theta \in \Theta$,

$$\left|\bar{\rho}_\theta^* - \rho_\theta^*\right| \leq C_\gamma \varepsilon^{\frac{\gamma}{2}} \text{ and } \rho_\theta^* - \rho_\theta^{\bar{\pi}_\theta^*}(0) \leq C_\gamma \varepsilon^{\frac{\gamma}{2}}. \tag{5.20}$$

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

$$\varepsilon \rho_\theta^{\bar{\pi}^*_\theta}(0) = \mathbb{E}[\bar{W}^*_\theta(x + \mu_\theta(x,a) + \Sigma\xi)] - \bar{W}^*_\theta(x) + r(x, \bar{\pi}^*_\theta(x)) + e_\theta(x) \tag{5.21}$$

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

Proposition 5.3.6, combined with (5.19) provides a certifiable approximation for solving the control problem (5.2) with off-the-shelf diffusive HJB solvers, at a cost independent of ε. An example of this methodology is seen in Section 4.4, in which Fig. 4.1 shows the reduction in computational effort. Proposition 5.3.6 also provides in (5.21) an HJB-like representation of the approximation, which provides a key with which to analyse the regret incurred when using this approximation.

5.3.5 Regret decomposition

To sketch the proof of Theorem 5.3.1, we will work on the high-probability event of Proposition 5.3.2, and omit martingale measurability issues which 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. Section 5.7 is dedicated to a complete proof.

Sketch of the proof of Theorem 5.3.1. 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 (5.21) of Proposition 5.3.6.(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

$$\begin{aligned}
\mathcal{R}_T(\varpi) = T\rho^*_{\theta^*} - \varepsilon \sum_{n=1}^{N_T} \rho_{\theta_n}^{\bar{\pi}^*_{\theta_n}}(0) && (R_1)\\
+ \sum_{n=1}^{N_T} \mathbb{E}[\bar{W}^*_{\theta_n}(\tilde{X}^{\varpi,\theta_n}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] - \bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}) && (R_2)\\
+ \sum_{n=1}^{N_T} e_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}) && (R_3)
\end{aligned}$$

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 5.3.2, 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 5.3.6, show that (R_1) decomposes into

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

Also by Proposition 5.3.6, $R_3 \leq \varepsilon^{1+\gamma/2} N_T(1 + H_\delta(N_T)^3)$. Thus $R_1 + R_3 \lesssim C_\gamma \varepsilon^{\gamma/2} T$.

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 5.3.5, yields

$$R_2 \leq L_{\bar{W}} \sum_{n=1}^{N_T} \left\|\mu_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) - \mu_{\theta^*}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n})\right\| \tag{R_4}$$

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

$$+ \sum_{n=1}^{N_T} \mathbb{E}[\bar{W}^*_{\theta_{n+1}}(X^{\varpi,\theta^*}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] - \bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}), \tag{R_6}$$

by adding and subtracting $\mathbb{E}[\bar{W}^*_{\theta_{n+1}}(\tilde{X}^{\varpi,\theta^*}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] = \mathbb{E}[\bar{W}^*_{\theta_{n+1}}(X^{\varpi,\theta^*}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}]$. The term (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 5.3.3 in Section 5.5 which yields

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

□

5.4 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 5.2, 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 5.3.1. Section 5.4.1 is dedicated to its proof, and it will be used in the proofs of learning results and high-probability regret bounds (Sections 5.5 and 5.7).

Proposition 5.3.1.
Let Assumptions 5.1 and 5.2 hold. Then, 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 $\theta \in \Theta$

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}_+} \frac{\|X_t^{\alpha,\theta}\|}{H_\delta(N_t)} \geq 1\right) \leq \delta. \tag{5.12}$$

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 5.3.1 in this case is a moment result, given by Lemma 5.4.4, which is proved in Section 5.4.2 and will then be used in Section 5.6.

Lemma 5.4.4.
Under Assumptions 5.1 and 5.2, for any $p \geq 2$, there is a constant $\mathfrak{c}'_p > 0$ independent of ε such that

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

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

5.4.1 Proof of Proposition 5.3.1

This section is dedicated to the proof of Proposition 5.3.1 which is a high-probability bound on the state process. This proof follows the classical path of the Chernoff method. We will derive an exponential moment bound for the state process in Lemma 5.4.2. First, we must obtain a stochastic stability condition in expectation in Lemma 5.4.1. In what follows, let $R_\varepsilon := \sqrt{8d\log(1/\varepsilon)}$ and $\xi \sim \nu$.

Lemma 5.4.1.
Let Assumptions 5.1 and 5.2 hold. Then,

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

$$\mathscr{V}(\psi_\theta^\varepsilon(x,a) - \sqrt{\varepsilon}\eta) \le (1 - \varepsilon \mathfrak{c}_{\mathscr{V}})\mathscr{V}(x - \sqrt{\varepsilon}\eta) + \varepsilon M_{\mathscr{V}} L_0(1 + \|\eta\|)\,; \tag{5.22}$$

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

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

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

Proof.

(i) By Lipschitzness of $\mathscr{V}$ and (5.4), for any $(\eta, x, a, \theta) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{A} \times \Theta$, we have

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

from which (5.22) follows by using Assumption 5.1 which itself implies $\|\bar{\mu}_\theta(\sqrt{\varepsilon}\eta, a)\| \le L_0(1 + \sqrt{\varepsilon}\|\eta\|) \le L_0(1 + \|\eta\|)$ since $\varepsilon \in (0, 1)$.

(ii) By the symmetry of the law of $\bar{\Sigma}\xi$, by (5.22) applied for $\eta = \bar{\Sigma}\xi$, and by taking expectation, we have that for any $x \in \mathbb{R}^d$

$$\begin{aligned}
\mathbb{E}[\mathscr{V}(\psi_\theta^\varepsilon(x,a) + \Sigma\xi)] &= \mathbb{E}[\mathscr{V}(\psi_\theta^\varepsilon(x,a) - \sqrt{\varepsilon}\bar{\Sigma}\xi)] \\
&\le (1 - \varepsilon \mathfrak{c}_{\mathscr{V}})\mathbb{E}[\mathscr{V}(x - \sqrt{\varepsilon}\bar{\Sigma}\xi)] \\
&\quad + \varepsilon M_{\mathscr{V}} L_0(1 + \|\bar{\Sigma}\|_{\mathrm{op}}\mathbb{E}[\|\xi\|])\,. \qquad (5.23)
\end{aligned}$$

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 $\mathbb{E}[\|\xi\|] \le \sqrt{d}$. Thus, the second term in (5.23) 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 $\|\xi\| \ge R_\varepsilon$, corresponding to the tails.

(ii).1. For the first event we consider (on which $\|\xi\| < R_\varepsilon$), we must have $0 \notin \mathcal{B}(x, \|\Sigma\xi\|)$ for any $x \notin \mathcal{B}(\|\Sigma\|_{\mathrm{op}} R_\varepsilon)$, 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

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

for some $\hat{x} \in \{x + \Delta\Sigma\xi\}_{\Delta\in[0,1]}$. By the Cauchy-Schwartz inequality and the derivative bounds of Assumption 5.2, we obtain

$$\begin{aligned}\mathbb{E}[\mathscr{V}(x+\Sigma\xi)1_{\{\|\xi\|<R_\varepsilon\}}] &\le \mathscr{V}(x)+\mathbb{E}[\xi^\mathsf{T}1_{\{\|\xi\|<R_\varepsilon\}}]\Sigma^\mathsf{T}\nabla\mathscr{V}(x)+\frac{\varepsilon}{2}M'_{\mathscr{V}}\|\bar{\Sigma}\|_{\mathrm{op}}^2\\&\le \mathscr{V}(x)+\frac{\varepsilon}{2}M'_{\mathscr{V}}\|\bar{\Sigma}\|_{\mathrm{op}}^2,\end{aligned}$$

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

(ii).2. On the second event (on which $\|\xi\| \ge R_\varepsilon$), 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. [83, (3.5)]):

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

(ii).3. To complete the proof, we combine both cases in (5.23), and let

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

□

Lemma 5.4.2.

Let Assumptions 5.1 and 5.2 hold. Then, for any $\lambda \in \mathbb{R}_+$, we have

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

for all $(x_0, \alpha, \theta) \in \mathbb{R}^d \times \mathcal{A} \times \Theta$ *and* $n \in \mathbb{N}$.

Proof. For $n \in \mathbb{N}^*$, let us define the following events for any $0 \leq i < n$:

$$E_{i,n-1} := \{i = \sup\{j \in \{0, \dots, n-1\} : \|X^{\alpha,\theta}_{\tau_j}\| \leq \|\Sigma\|_{\mathrm{op}} R_\varepsilon\}\}$$

and

$$\bar{E}_{n-1} := \{\min_{j \in \{0,\dots,n-1\}} \|X^{\alpha,\theta}_{\tau_j}\| > \|\Sigma\|_{\mathrm{op}} R_\varepsilon\}.$$

Note that both these events are $\mathcal{F}_{\tau_{n-1}}$-measurable and $\bar{E}^c_{n-1} = \cup_{i \leq n-1} E_{i,n-1}$, so that $\{\bar{E}_{n-1}, E_{0,n-1}, \dots, E_{n-1,n-1}\}$ induces a partition of Ω. We begin by working conditionally on each of these events, and in a second part we will collect them to bound $\mathbb{E}[\exp(\lambda \mathscr{V}(X^{\alpha,\theta}_{\tau_n}))]$.

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

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

Using a result for Lipschitz functions of Gaussian random variables (see e.g. [36, Thm 5.5]) applied to $\mathscr{V}$ and ξ, we obtain

$$\begin{aligned}
&\mathbb{E}[e^{\lambda \mathscr{V}(X^{\alpha,\theta}_{\tau_n})} 1_{E_{i,n-1}}] \\
&\quad \leq e^{\frac{\lambda^2}{2} M^2_{\mathscr{V}} \|\Sigma\|^2_{\mathrm{op}}} \mathbb{E}\Big[\exp\left(\mathbb{E}[\lambda \mathscr{V}(X^{\alpha,\theta}_{\tau_n})|\mathcal{F}_{\tau_{n-1}}]\right) 1_{E_{i,n-1}} \Big] \\
&\quad = e^{\frac{\lambda^2}{2} M^2_{\mathscr{V}} \|\Sigma\|^2_{\mathrm{op}}} \mathbb{E}\Big[\exp\left(\mathbb{E}[\lambda \mathscr{V}(\psi^\varepsilon_\theta(X^{\alpha,\theta}_{\tau_{n-1}}, \alpha_{\tau_{n-1}}) + \Sigma \xi_n)|\mathcal{F}_{\tau_{n-1}}]\right) 1_{E_{i,n-1}} \Big].
\end{aligned} \tag{5.24}$$

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

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

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

$$C_H := L_{\mathscr{V}} \left((1 + L_0)\|\bar{\Sigma}\|_{\mathrm{op}} \sqrt{8de^{-1}} + 1 + \|\bar{\Sigma}\|_{\mathrm{op}} \sqrt{d} \right). \tag{5.25}$$

Combining this with (5.24) yields

$$\mathbb{E}\left[e^{\lambda \mathscr{V}(X^{\alpha,\theta}_{\tau_n})} 1_{E_{i,n-1}} \right] \leq \exp\left(\frac{\lambda^2}{2} M^2_{\mathscr{V}} \|\Sigma\|^2_{\mathrm{op}} + \lambda C_H \right), \tag{5.26}$$

in the case $i = n - 1$.

If $i < n - 1$, we can apply the same methodology, and continuing from (5.24) apply Lemma 5.4.1 to obtain

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda \mathscr{V}(X^{\alpha,\theta}_{\tau_n})} 1_{E_{i,n-1}} \right] \\
&\quad \leq e^{\frac{\lambda^2}{2} M^2_{\mathscr{V}} \|\Sigma\|^2_{\mathrm{op}}} \mathbb{E}\Big[\exp\big(\mathbb{E}\big[\lambda \mathscr{V}(\psi^{\varepsilon}_{\theta}(X^{\alpha,\theta}_{\tau_{n-1}}, \alpha_{\tau_{n-1}}) + \Sigma \xi_n) | \mathcal{F}_{\tau_{n-1}} \big] \big) \\
&\qquad\qquad \times 1_{\{X^{\alpha,\theta}_{\tau_{n-1}} > \|\Sigma\|_{\mathrm{op}} R_\varepsilon\}} 1_{E_{i,n-2}} \Big], \qquad (5.27) \\
&\quad \leq e^{\frac{\lambda^2}{2} M^2_{\mathscr{V}} \|\Sigma\|^2_{\mathrm{op}} + \lambda \varepsilon \mathfrak{c}'_{\mathscr{V}}} \mathbb{E}[\exp((1 - \varepsilon \mathfrak{c}_{\mathscr{V}}) \lambda \mathscr{V}(X^{\alpha,\theta}_{\tau_{n-1}})) 1_{E_{i,n-2}}].
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda \mathscr{V}(X^{\alpha,\theta}_{\tau_n})} 1_{E_{i,n-1}} \right] \\
&\quad \leq \exp\left(\lambda C_H + \lambda \varepsilon \mathfrak{c}'_{\mathscr{V}} \sum_{k=0}^{n-1-i} (1 - \varepsilon \mathfrak{c}_{\mathscr{V}})^k + \frac{\lambda^2 M^2_{\mathscr{V}} \|\Sigma\|^2_{\mathrm{op}}}{2} \sum_{k=0}^{n-1-i} (1 - \varepsilon \mathfrak{c}_{\mathscr{V}})^{2k} \right) \\
&\quad \leq \exp\left(\lambda C_H + \lambda \frac{\mathfrak{c}'_{\mathscr{V}}}{\mathfrak{c}_{\mathscr{V}}} + \frac{\lambda^2 M^2_{\mathscr{V}} \|\bar{\Sigma}\|^2_{\mathrm{op}}}{2\mathfrak{c}_{\mathscr{V}}} \right). \qquad (5.28)
\end{aligned}$$

On the event $\bar{E}_{n-1}$, that is if the process is never in the ball $\mathcal{B}(\|\Sigma\|_{\mathrm{op}} R_\varepsilon)$ before time τ_n, we use the fact that (5.27) 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

$$\mathbb{E}\left[e^{\lambda \mathscr{V}(X^{\alpha,\theta}_{\tau_n})} 1_{\bar{E}_{n-1}} \right] \leq \exp\left(\lambda L_{\mathscr{V}} \|x_0\| + \lambda \frac{\mathfrak{c}'_{\mathscr{V}}}{\mathfrak{c}_{\mathscr{V}}} + \frac{\lambda^2 M^2_{\mathscr{V}} \|\bar{\Sigma}\|^2_{\mathrm{op}}}{2\mathfrak{c}_{\mathscr{V}}} \right). \tag{5.29}$$

Using our partition and combining (5.26), (5.28), and (5.29) we can thus write, for any $n \in \mathbb{N}$

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

which concludes the proof. □

With these two lemmas, we can now prove Proposition 5.3.1, the main result of this section. First, let us give the exact definition of $H_\delta(n)$:

$$H_\delta(n) := \frac{1}{\ell_{\mathscr{V}}}\left(C_H + L_{\mathscr{V}}\left\|x_0\right\|\right) + \frac{\mathfrak{c}'_{\mathscr{V}}}{\ell_{\mathscr{V}}\mathfrak{c}_{\mathscr{V}}} + \frac{M_{\mathscr{V}}}{\ell_{\mathscr{V}}}\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{\frac{2}{\mathfrak{c}_{\mathscr{V}}}\log\left(\frac{\pi^2(n+1)^3}{6\delta}\right)} \tag{5.30}$$

in which C_H is defined in (5.25), so that $H_\delta(n) = \mathcal{O}\big(\sqrt{\log(n\delta^{-1})}\big)$.

Proposition 5.3.1.
Let Assumptions 5.1 and 5.2 hold. Then, 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 $\theta \in \Theta$

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}_+} \frac{\|X_t^{\alpha,\theta}\|}{H_\delta(N_t)} \geq 1\right) \leq \delta. \tag{5.12}$$

Proof. Fix $n \in \mathbb{N}$, by Markov's inequality and Assumption 5.2, for any $u > 0$, we have

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

which implies that

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

Applying Lemma 5.4.2, and taking $\lambda = \mathfrak{c}_{\mathscr{V}}\ell_{\mathscr{V}} u/(M_{\mathscr{V}}^2\|\bar{\Sigma}\|_{op}^2)$, we obtain

$$\mathbb{P}\left(\|X_{\tau_n}^{\alpha,\theta}\| > u + \frac{\mathfrak{c}'_{\mathscr{V}}}{\ell_{\mathscr{V}}\mathfrak{c}_{\mathscr{V}}} + \varepsilon^{\frac{1}{2}}\frac{L_{\mathscr{V}}}{\ell_{\mathscr{V}}}\|\bar{\Sigma}\|_{\mathrm{op}} R_\varepsilon + \frac{L_{\mathscr{V}}}{\ell_{\mathscr{V}}}\left\|x_0\right\|\right)$$

$$\leq (n+1)\exp\left(-\lambda\ell_{\mathscr{Y}}u + \lambda^2\frac{M_{\mathscr{Y}}^2\|\bar{\Sigma}\|_{\mathrm{op}}^2}{2\mathfrak{c}_{\mathscr{Y}}}\right)$$
$$= (n+1)\exp\left(-\frac{\mathfrak{c}_{\mathscr{Y}}\ell_{\mathscr{Y}}^2}{2M_{\mathscr{Y}}^2\|\bar{\Sigma}\|_{\mathrm{op}}^2}u^2\right).$$

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

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

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

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

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

5.4.2 Expectation bounds of higher orders

In this section, we will focus on higher moment conditions of the state process, which will be used in the control results of Section 5.6. In Lemma 5.4.3 and Corollary 5.4.1 we work to raise the stochastic stability condition from Lemma 5.4.1 to a power $p \geq 2$. Lemma 5.4.4, the main result of this section, will follow from this by arguments of Chapter 4.

Lemma 5.4.3.
Let Assumptions 5.1 and 5.2 hold. Then, 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

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

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

$$\mathscr{V}\left(\psi_\theta^\varepsilon(x,a) - \sqrt{\varepsilon}\eta\right)^p \leq (1 - \varepsilon\mathfrak{c}_{\mathscr{Y}})\mathscr{V}\left(x - \sqrt{\varepsilon}\eta\right)^p + g(x,\eta). \tag{5.31}$$

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 (5.22) to the power p

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

We will now expand the right-hand side. Let $a = (1 - \varepsilon \mathfrak{c}_{\mathscr{V}})\mathscr{V}\left(x - \sqrt{\varepsilon}\eta\right)$ 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 $(1 - \varepsilon \mathfrak{c}_{\mathscr{V}}) \in (0,1)$, $\varepsilon \leq 1$, $b \leq 1 + b$, $\sum_{k=0}^{p-1} \binom{p}{k} \leq 2^p$, and 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

$$\begin{aligned}\mathscr{V}\left(\psi_\theta^\varepsilon(x,a) - \sqrt{\varepsilon}\eta\right)^p \leq (1 - \varepsilon \mathfrak{c}_{\mathscr{V}})\mathscr{V}\left(x - \sqrt{\varepsilon}\eta\right)^p \\ + \varepsilon(1 + M_{\mathscr{V}} L_0 (1 + \|\eta\|))^p \left(1 + \mathscr{V}\left(x - \sqrt{\varepsilon}\eta\right)^{p-1}\right) 2^{p-2+p}.\end{aligned} \tag{5.32}$$

Finally, we have

$$\begin{aligned}(1 + M_{\mathscr{V}} L_0 (1 + \|\eta\|))^p &= (1 + M_{\mathscr{V}} L_0 + M_{\mathscr{V}} L_0 \|\eta\|)^p \\ &\leq \left(1 + M_{\mathscr{V}} L_0 + (1 + M_{\mathscr{V}} L_0)\|\eta\|\right)^p \\ &= (1 + M_{\mathscr{V}} L_0)^p (1 + \|\eta\|)^p \\ &\leq (1 + M_{\mathscr{V}} L_0)^p (1 + \|\eta\|^p) 2^{p-1}. \end{aligned} \tag{5.33}$$

Combining (5.32) and (5.33), leads to the required result. □

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

***Corollary* 5.4.1.** *Under Assumptions 5.1 and 5.2, for any $p \geq 2$, there is a constant $\mathfrak{c}_p > 0$ independent of ε such that*

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

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

Proof.

1. Taking the expectation of the bound on g from Lemma 5.4.3 and applying Hölder's inequality yields

$$\mathbb{E}[g(x,\xi)] \leq \varepsilon C_p \mathbb{E}\left[\left(1 + \mathscr{V}\left(x - \sqrt{\varepsilon}\xi\right)^{p-1}\right)\left(1 + \|\xi\|^p\right)\right]$$

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

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

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

is a finite constant and we have

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

2. Recalling Lemma 5.4.3, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathscr{V}\left(\psi_\theta^\varepsilon(x,a)+\Sigma\xi\right)^p\right] \\
&\leq (1-\varepsilon \mathfrak{c}_{\mathscr{V}})\,\mathbb{E}\left[\mathscr{V}\left(x-\sqrt{\varepsilon}\xi\right)^p\right] + \mathbb{E}[g(x,\xi)] \\
&\leq \left(1-\varepsilon\frac{\mathfrak{c}_{\mathscr{V}}}{2}\right)\mathbb{E}\left[\mathscr{V}\left(x-\sqrt{\varepsilon}\xi\right)^p\right] \\
&\quad + \varepsilon\mathbb{E}\left[C_p'\left(1+\mathscr{V}\left(x-\sqrt{\varepsilon}\xi\right)^{p-\frac{1}{p}}\right) - \frac{\mathfrak{c}_{\mathscr{V}}}{2}\mathscr{V}\left(x-\sqrt{\varepsilon}\xi\right)^p\right]. \qquad (5.34)
\end{aligned}$$

3. For any $p \geq 1$, the function

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

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

$$C_p' \mathscr{V}(z)^{p-\frac{1}{p}} - \frac{\mathfrak{c}_{\mathscr{V}}}{2}\mathscr{V}(z)^p \leq C_p''.$$

Applying this to the expectation in (5.34), we have

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

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

Lemma 5.4.4.
Under Assumptions 5.1 and 5.2, for any $p \geq 2$, there is a constant $\mathfrak{c}'_p > 0$ independent of ε such that

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

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

Proof. Recall from Corollary 5.4.1 that we have

$$\mathbb{E}\left[\mathscr{V}\left(\psi_\theta^\varepsilon(x,a) + \Sigma\xi\right)^p\right] \leq \left(1 - \varepsilon\frac{\mathfrak{c}_{\mathscr{V}}}{2}\right)\mathbb{E}\left[\mathscr{V}(x + \Sigma\xi)^p\right] + \varepsilon\mathfrak{c}_p \tag{5.35}$$

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 $\mathscr{V}(\cdot)^p \in \mathcal{C}^2(\mathbb{R}^d; [0, +\infty))$ and use the fact that $\mathbb{E}[\xi] = 0$ to obtain

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

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

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

By the binomial theorem as in the proof of Lemma 5.4.3, and as $|\Delta| \leq 1$, we have

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

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 (5.35), we thus have

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

As in part 3. of the proof of Corollary 5.4.1, 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

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

for all $x \in \mathbb{R}^d$. Therefore, we have the desired Lyapunov generator condition

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

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

$$\frac{1}{\varepsilon}\int\left(\mathscr{V}\left(\psi_\theta^\varepsilon(x,a)+\Sigma e\right)^p-\mathscr{V}(x)^p\right)\nu(\mathrm{d}e)\leq-\frac{\mathfrak{c}_{\mathscr{V}}}{4}\mathscr{V}(x)^p+\mathfrak{c}_p'. \tag{5.36}$$

By Itô's Lemma, (5.36), and a localisation argument, we have, for any $t \geq t_0 \geq 0$, that

$$\begin{aligned}\mathbb{E}[\mathscr{V}(X_t^{x_0,\alpha,\theta})^p] &= \mathbb{E}\left[\mathscr{V}\left(X_{t_0}^{x_0,\alpha,\theta}\right)^p\right] \\ &\quad+\mathbb{E}\left[\int_{t_0}^t\frac{1}{\varepsilon}\int\left[\mathscr{V}\left(\psi_\theta^\varepsilon\left(X_s^{x_0,\alpha,\theta},\alpha_s\right)+\Sigma e\right)^p-\mathscr{V}(X_s^{x_0,\alpha,\theta})^p\right]\nu(\mathrm{d}e)\mathrm{d}s\right] \\ &\leq \mathbb{E}\left[\mathscr{V}\left(X_{t_0}^{x_0,\alpha,\theta}\right)^p\right]-\frac{\mathfrak{c}_{\mathscr{V}}}{4}\int_{t_0}^t\mathbb{E}\left[\mathscr{V}\left(X_s^{x_0,\alpha,\theta}\right)^p\right]\mathrm{d}s+(t-t_0)\mathfrak{c}_p'.\end{aligned}$$

By a simple comparison argument for ODEs, we then obtain

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

Finally, using Assumption 5.2, we obtain

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

□

5.5 Learning: Concentration & Online Prediction Error

The key result of this section, Proposition 5.3.2, extends and builds heavily on [107, Prop. 5]. Proposition 5.3.2 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 [107] and to incorporate some self-normalised inequality arguments as well as high-probability bounds on the state from Section 5.4. While many of the original ideas are still used, the way they link together has changed and thus we will include, in Section 5.5.1, a complete derivation for the sake of clarity. In this spirit, we will prove a generic result (Theorem 5.5.1), which itself implies Proposition 5.3.2.

Proposition 5.3.2. **(Adapted from [97, Prop. 5])**
Under Assumptions 5.1 and 5.2, for any $x_0 \in \mathbb{R}^d$, and $\delta > 0$,

$$\mathbb{P}\left(\left\{\theta^* \in \bigcap_{n=1}^{\infty} \mathcal{C}_n(\delta)\right\} \cap \left\{\sup_{n\in\mathbb{N}^*} \frac{\left\|X^{\varpi,\theta^*}_{\tau_n}\right\|}{H_\delta(n)} \leq 1\right\}\right) \geq 1-\delta\,, \tag{5.15}$$

Proposition 5.3.2 ensures that the sets $(\mathcal{C}_n(\delta))_{n\in\mathbb{N}}$ defined in (5.6) are valid confidence sets. In order to bound the regret, we need to go further and bound the online prediction error of functions within these confidence sets along the trajectory (see. (5.57)).

For any $n \in \mathbb{R}$, let $\mathrm{d}_{\mathrm{E},n}$ denote the $2\sqrt{\varepsilon/n}$-eluder dimension of the model class $\mathscr{F}_\Theta$, elements of which have been restricted to $B_n := \mathcal{B}(\sup_{s\leq\tau_n}\|X^{\varpi,\theta^*}_s\|)$, i.e. $\mathrm{d}_{\mathrm{E},n} := \dim_{\mathrm{E}}((f|_{B_n})_{f\in\mathscr{F}_\Theta}, 2\sqrt{\varepsilon/n})$. In Section 5.5.2, we derive a general result (Proposition 5.5.6) from which Proposition 5.3.3 follows.

Proposition 5.3.3.
Under Assumptions 5.1 and 5.2, 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 $1-\delta$

$$\sum_{n=1}^{N_t} \|\mu_{\hat{\theta}_n} - \mu_{\theta^*}\|(X_{\tau_n}^{\alpha,\theta^*}, \alpha_{\tau_n}) \leq \tilde{\mathcal{O}}\left(\sqrt{\varepsilon \mathrm{d}_{\mathrm{E},N_T} \log(\mathscr{N}_{N_T}^{\varepsilon}) N_t} + \varepsilon \mathrm{d}_{\mathrm{E},N_T}\right), \tag{5.16}$$

and

$$\sum_{n=1}^{N_t} \|\mu_{\hat{\theta}_n} - \mu_{\theta^*}\|^2(X_{\tau_n}^{\alpha,\theta^*}, \alpha_{\tau_n}) \leq \tilde{\mathcal{O}}\left(\mathrm{d}_{\mathrm{E},N_T} \log(\mathscr{N}_{N_T}^{\varepsilon})\right). \tag{5.17}$$

5.5.1 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 [97, 107] 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 a 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 $\mathbb{H} = (\mathcal{H}_i)_{i\in\mathbb{N}}$.

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

$$\langle Z|Z'\rangle_n := \sum_{i=0}^{n-1} \langle Z_i|Z'_i\rangle \text{ and } \|Z\|_n := \sqrt{\langle Z|Z\rangle_n}. \tag{5.37}$$

While $\|\cdot\|_n$ is not a norm, it plays this role and we follow here the notational convention of [107]. 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 5.3, which in our end-goal application is subsumed by Assumptions 5.1 and 5.2 and Proposition 5.3.1.

Assumption 5.3.
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 $n \in \mathbb{N}^*$

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

The definition of this cover corresponds to the one used in [107] 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

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

and the event

$$\mathcal{E}_n^0(\delta) :=$$
$$\left\{ \|\hat{f}_n - f^*\|_n \le 2\Gamma \sqrt{L_n^1\left(\frac{3\delta}{\pi^2 n^2}\right)} \right.$$
$$\left. + 2\sqrt{\Gamma^2 + 2\Gamma\left(n \sup_{g \in \mathscr{C}_n^\Gamma} C_n^2(g) \sqrt{2\log\left(\frac{4\pi^2 n^3}{3\delta}\right)} + \sqrt{2n \sup_{g \in \mathscr{C}_n^\Gamma} C_n^2(g) L_n^1\left(\frac{3\delta}{\pi^2 n^2}\right)}\right)} \right\}. \tag{5.38}$$

Building upon the proof method of [107], the cornerstone of this section is Lemma 5.5.2, 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 5.5.2.
Under Assumption 5.3, for $n \in \mathbb{N}^*$ and $\delta \in (0,1)$, we have

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

We begin the proof of Lemma 5.5.2 by giving the concentration inequality of Lemma 5.5.1.

Lemma 5.5.1.
Let Assumption 5.3 hold. Then, for all $n \in \mathbb{N}^*$, $\delta \in (0,1)$, and $f \in \mathscr{F}$

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

Proof. This proof relies on extensively studied arguments for self-normalised inequalities, but we include it for completeness because it uses nonstandard 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

$$\begin{aligned}&\mathbb{E}\left[M_{n+1}^\lambda(f)|\mathcal{H}_n\right]\\&\quad= M_n^\lambda(f)\mathbb{E}\left[\exp\left(\lambda\langle\xi_{n+1}|f(X_n)-f^*(X_n)\rangle_n\right)\middle|\mathcal{H}_n\right]e^{-\frac{\lambda^2\Gamma^2}{2}\|f(X_n)-f^*(X_n)\|_n^2}.\end{aligned} \tag{5.39}$$

By the Cauchy-Schwartz inequality

$$\left|\langle\xi_n|f(X_n)-f^*(X_n)\rangle_n\right| \le \|\xi_n\|_n \, \|f(X_n)-f^*(X_n)\|_n$$

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

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

and thus, by (5.39), $M_n^\lambda(f)$ is a supermartingale. By definition of $\langle\cdot|\cdot\rangle_0$ and $\|\cdot\|_0$, $M_0^\lambda(f) = 1$, so that $\mathbb{E}[M_n^\lambda(f)] \le 1$ 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, for every $f \in \mathscr{F}$, the process $(M_n(f))_{n\in\mathbb{N}}$ by

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

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

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

□

We will turn to the proof of Lemma 5.5.2. Recall (5.38), which defined for $\delta \in (0,1)$ and $n \in \mathbb{N}^*$, the event

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

Lemma 5.5.2.
Under Assumption 5.3, 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 [107]. We begin by giving two small auxiliary results which we will use.

1. 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
$$\mathbb{P}\left(\sup_{i \leq n} \|\xi_i\| \leq \Gamma \sqrt{2\log\left(\frac{2n}{\delta}\right)}\right) \geq 1 - \delta \tag{5.40}$$

2. For any $f \in \mathscr{F}$, and $n \in \mathbb{N}^*$ we have
$$\begin{aligned}\|f^* - Y\|_n^2 - \|f - Y\|_n^2 &= \langle f^* - Y | f^* - Y\rangle_n \\ &\quad - \langle f - f^* + f^* - Y | f - f^* + f^* - Y\rangle_n \\ &= \langle f^* - Y | f^* - Y\rangle_n - \langle f - f^* | f - f^*\rangle_n \\ &\quad + 2\langle Y - f^* | f - f^*\rangle_n - \langle Y - f^* | Y - f^*\rangle_n \\ &= -\|f - f^*\|_n^2 + 2\langle \xi | f - f^*\rangle_n\,. \end{aligned} \tag{5.41}$$
Applying (5.41) with $f := \hat{f}_n$, the n-point non-linear least-square fit, leads to a non-negative left-hand side and thus
$$\|\hat{f}_n - f^*\|_n^2 \leq 2\left|\langle \xi | f - f^*\rangle_n\right|\,.$$
At the same time, for all $n \in \mathbb{N}^*$, by definition of $\mathscr{C}_n^\Gamma$, it holds that for all $g \in \mathscr{C}_n^\Gamma$
$$\begin{aligned}\|\hat{f}_n - f^*\|_n^2 &\leq 2\left|\langle \xi | g - f^*\rangle_n\right| + 2\left|\langle \xi | \hat{f}_n - g\rangle_n\right| \\ &\leq 2\left|\langle \xi | g - f^*\rangle_n\right| + 2n \sup_{i \leq n} \|\xi_i\|_2\, C_n^2(g)\,. \end{aligned} \tag{5.42}$$

Combining (5.40) and (5.42), we obtain, for all $\delta \in (0,1)$, $n \in \mathbb{N}^*$, and $g \in \mathscr{C}_n^\Gamma$, that

$$\mathbb{P}\left(\|\hat{f}_n - f^*\|_n^2 \geq 2\left|\langle \xi | g - f^* \rangle_n\right| + 2nC_n^2(g)\Gamma\sqrt{2\log\left(\frac{2n}{\delta}\right)}\right) \leq \delta \tag{5.43}$$

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

$$C_n^1(g) \leq \Gamma^2 + 8nL^2(1 + \sup_{i \leq n} \|X_i\|^2)\,. \tag{5.44}$$

$$C_n^1(g) \leq \Gamma^2 + \|\hat{f}_n - f^*\|_n^2 + \|g - \hat{f}_n\|_n^2 \leq C_n^1(\hat{f}_n) + nC_n^2(g)\,, \tag{5.45}$$

Applying Lemma 5.5.1 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 (5.44) and (5.45) this becomes

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

in which $C_{\Gamma,L,X,n} := (Gamma^2 + 8L^2(1 + \sup_{i \leq n} \|X_i\|^2))\mathscr{N}_n^\Gamma$ and thus

$$\delta_0(n) \geq \mathbb{P}\left(\sup_{g \in \mathscr{C}_n^\Gamma} \left|\langle \xi | g - f^* \rangle_n\right| \geq \Gamma\sqrt{2L_n^1(\delta_0(n))}\left(\sqrt{C_n^1(\hat{f}_n)} + \sqrt{n \sup_{g \in \mathscr{C}_n^\Gamma} C_n^2(g)}\right)\right). \tag{5.46}$$

Combining (5.43) and (5.46) by a union bound gives us

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

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

$$C_n^1(\hat{f}_n) \leq \Gamma^2 + \Gamma\sqrt{2C_n^1(\hat{f}_n)L_n^1(\delta_0(n)/2)} + h_n^\Gamma\,, \tag{5.47}$$

in which

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

Viewing (5.47) as a second order polynomial in $\sqrt{C_n^1(\hat{f}_n)}$, we obtain via its roots that

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

Since $\|\hat{f}_n - f^*\|_n \le \sqrt{C_n^1(\hat{f}_n)}$ by definition of $C_n^1(\hat{f}_n)$, $\|\hat{f}_n - f^*\|_n$ is upper bounded by

$$\begin{aligned}&2\sqrt{\Gamma^2 + 2\Gamma\left(n \sup_{g\in\mathscr{G}_n^\Gamma} C_n^2(g)\sqrt{2\log\left(\frac{4n}{\delta_0(n)}\right)} + \sqrt{2n \sup_{g\in\mathscr{G}_n^\Gamma} C_n^2(g) L_n^1\left(\frac{\delta_0(n)}{2}\right)}\right)} \\ &\quad + 2\Gamma\sqrt{L_n^1(\delta_0(n)/2)}\,.\end{aligned}$$

Therefore, letting

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

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

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

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 5.5.2, we used self-normalised inequalities to generalise the results of [107] to unbounded states. We now incorporate the high probability bound of Assumption 5.3 and formalise confidence sets, which

will prove Theorem 5.5.1. Theorem 5.5.1 can then be specified for our setting by merging it with the results of Section 5.4 in Proposition 5.3.2.

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

$$\mathcal{C}_n(\delta) := \left\{f \in \mathscr{F} \,:\, \|f - \hat{f}_n\|_n \leq \beta_n\right\} \tag{5.48}$$

with

$$\beta_n(\delta) := \beta_0 \vee 2\Gamma\left(\sqrt{1 + 2\left(\sqrt{2\Gamma \log\left(\frac{8n}{\delta}\right)} + \sqrt{2L_n^0\left(\frac{\delta}{4}\right)}\right)} + \sqrt{L_n^0\left(\frac{\delta}{4}\right)}\right). \tag{5.49}$$

Theorem 5.5.1.
Let Assumption 5.3 hold. We then have

$$\mathbb{P}\left(\left\{\bigcap_{n\in\mathbb{N}^*}\{f^* \in \mathcal{C}_n(\delta)\}\right\}\bigcap\left\{\sup_{n\in\mathbb{N}^*}\frac{\|X_n\|}{H_\delta(n)} \leq 1\right\}\right) \leq \delta \text{ for all } \delta \in (0,1).$$

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 $\mathcal{B}(H_\delta(n))$. 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

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

By Lemma 5.5.2, Assumption 5.3, and a union bound, $\mathbb{P}\left(\mathcal{E}^0(\delta)\right) \geq 1-2\delta$, and we obtain the result by (5.48) and (5.49), i.e. by definition of $\mathcal{C}_n(\delta)$. □

Proposition 5.3.2. **(Adapted from [97, Prop. 5])**
Under Assumptions 5.1 and 5.2, for any $x_0 \in \mathbb{R}^d$, and $\delta > 0$,

$$\mathbb{P}\left(\left\{\theta^* \in \bigcap_{n=1}^{\infty} \mathcal{C}_n(\delta)\right\} \cap \left\{\sup_{n\in\mathbb{N}^*} \frac{\left\|X_{\tau_n}^{\varpi,\theta^*}\right\|}{H_\delta(n)} \leq 1\right\}\right) \geq 1-\delta\,, \tag{5.15}$$

Proof. The proof follows by applying Theorem 5.5.1 to this setting. That is, taking $(X_i)_{i\in\mathbb{N}} := ((X_{\tau_i}^{\varpi,\theta^*}, \varpi_{\tau_i}))_{i\in\mathbb{N}}$, $(Y_i)_{i\in\mathbb{N}} := (X_{\tau_{i+1}}^{\varpi,\theta^*} - X_{\tau_i}^{\varpi,\theta^*})_{i\in\mathbb{N}}$, $\mathscr{F} := \mathscr{F}_\Theta$ and with $(\xi_{n+1})_{n\in\mathbb{N}}$ and $(\beta_n(\delta))_{n\in\mathbb{N}^*}$ as defined in Section 5.2 and (5.14) respectively. This sets $\Gamma = \|\Sigma\|_{\mathrm{op}} = \varepsilon^{1/2}\|\bar{\Sigma}\|_{\mathrm{op}}$. 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. □

5.5.2 Widths of confidence sets

In Section 5.5.1, 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 $(\mu_{\theta_n})_{n\in\mathbb{N}}$.

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 $k(n) \leq n$, for all $n \in \mathbb{N}$. Therefore, as in Section 5.5.1, we begin by showing a general result (Proposition 5.5.4) in the learning framework of [107], then we apply it to our setting to prove Proposition 5.3.3. Using the notation of Section 5.5.1, 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 [107] is a notion of dimension which is perfectly tailored to converting fit errors into prediction errors. We defer to [107] for its technical definition. Unlike [107], 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 some set $S \subset \mathcal{X}$ of elements of $\mathscr{F}$, which we denote by $\dim_{\mathrm{E}}^{S}(\mathscr{F}, \epsilon)$. In this way, the eluder dimension of [107] is $\dim_{\mathrm{E}}^{\mathcal{X}}(\mathscr{F}, \epsilon)$. For $n \in \mathbb{N}^*$, let $B_n := \mathcal{B}(\sup_{i\in[n]} \|X_i\|)$ 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 $u \in \mathbb{R}_+$.

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 $n \in \mathbb{N}$. The role of the eluder dimension is evident in the key result of [107] which we reproduce in Lemma 5.5.3 (up to incorporating our extended definition of eluder dimension).

To prove Proposition 5.5.4, the result of [107] we leverage is Lemma 5.5.3 which we combine with two functional inequalities given in Lemma 5.5.5.

Lemma 5.5.3. ***([107, 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

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

Proof. Following the proof of [107, Prop. 3], the only modification involves the bound $\|\bar{f} - \underline{f}\|_n \le \tilde{\beta}_n$, for any $(\bar{f}, \underline{f}) \in \mathscr{F}_n^2$, which holds by assumption. □

This result combined with some functional inequalities for sequences (see Lemma 5.5.5 below) transforms a bound on fit error into a bound on prediction error in Proposition 5.5.4.

Proposition 5.5.4.
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 $\mathscr{F}_i := \{f \in \mathscr{F} : \|f - \tilde{f}_i\|_i \leq \tilde{\beta}_i\}$. Under Assumption 5.3, for any $n \in \mathbb{N}^*$, we have

$$\sum_{i=1}^{n} \sup_{(f,f')\in\mathscr{F}_n^2} \|f(X_i) - f'(X_i)\| \leq 2\tilde{\beta}_n \sqrt{\mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde{\beta}_0) n} + \mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde{\beta}_0) 2L(1 + \sup_{i\in[n]} \|X_i\|), \tag{5.50}$$

and

$$\sum_{i=1}^{n} \sup_{(f,f')\in\mathscr{F}_n^2} \|f(X_i) - f'(X_i)\|^2 \leq 4\tilde{\beta}_n^2 \mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde{\beta}_0) \left(3 + \log\left(\frac{8nL^2(1 + \sup_{i\in[n]} \|X_i\|)}{16\tilde{\beta}_n^4 (\mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde{\beta}_0))^2}\right)\right) + 2\mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(1 + 2\tilde{\beta}_n^2 \mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}(\tilde{\beta}_0)) \left(1 + 8L^2 \left(1 + \sup_{i\in[n]} \|X_i\|^2\right)\right). \tag{5.51}$$

Proof. The proof consists in applying Lemma 5.5.5 to Lemma 5.5.3, by letting $x_i = \Lambda(\mathscr{F}_i, X_i)$, $\zeta_n^\epsilon = 4\tilde{\beta}^2 \dim_{\mathrm{E}}^{B_n}(\mathscr{F}, \epsilon)$, recall $B_n := \mathcal{B}(\sup_{i\in[n]} \|X_i\|)$, and $\chi^\epsilon = \dim_{\mathrm{E}}^{B_n}(\mathscr{F}, \epsilon)$. When we set the value of ϵ in the proof of Lemma 5.5.5, χ^ϵ becomes

$$\dim_{\mathrm{E}}^{B_n}\left(\mathscr{F}, \sqrt{\frac{4(\tilde{\beta}_n)^2}{n}}\right) \leq \dim_{\mathrm{E}}^{B_n}\left(\mathscr{F}, \sqrt{\frac{4(\tilde{\beta}_0)^2}{n}}\right)$$

as $(\tilde{\beta}_n)_{n\in\mathbb{N}}$ is non-decreasing and the eluder dimension is decreasing in its third argument. An analog 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 (5.53) and (5.54), which gives the result. □

Before returning to the consequences of Proposition 5.5.4, let us now prove Lemma 5.5.5.

Lemma 5.5.5.
Let $(x_i)_{i\in\mathbb{N}^*} \subset \mathbb{R}_+$ and assume the existence of 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

$$\sum_{i=1}^{n} 1_{\{x_i>\epsilon\}} \leq \frac{\zeta_n^\epsilon}{\epsilon^2} + \chi^\epsilon \text{ for all } (n,\epsilon) \in \mathbb{N}^* \times \mathbb{R}_+^*. \tag{5.52}$$

Then, the following two inequalities hold

$$\sum_{i=1}^{n} x_i \leq 2\sqrt{n\zeta_n^\epsilon} + \chi^\epsilon \sup_{i\in[n]} x_i \tag{5.53}$$

$$\sum_{i=1}^{n} x_i^2 \leq \zeta_n^\epsilon \left(3 + \log\left(\frac{n \sup_{i\in[n]} x_i^2}{(\zeta_n^\epsilon)^2}\right)\right) + \chi^\epsilon(2+\zeta_n^\epsilon)\left(1 + \sup_{i\in[n]} x_i^2\right). \tag{5.54}$$

Proof.

1. For $\epsilon > 0$, we have by (5.52)

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

   and thus

$$\sum_{i=1}^{n}(x_i-\epsilon)1_{\{x_i>\epsilon\}} \leq \frac{\zeta_n^\epsilon}{\epsilon} + \chi^\epsilon \sup_{i\in[n]} x_i\,. \tag{5.55}$$

   Combining (5.55) with

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

   yields

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

Setting $\epsilon = \sqrt{\zeta_n^\epsilon / n}$ yields (5.53).

2. To prove (5.54), we iterate the bound (5.55)

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

Now, by some algebraic manipulations of $\sum_{i=1}^n x_i^2$, then completing the square, discarding negative terms, and using (5.55) in the third step, we get

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

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

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

□

To complete this section, we apply Proposition 5.5.4 to our setting. For $n \in \mathbb{N}^*$, let us recall the shorthand notation

$$\mathrm{d}_{\mathrm{E},n} := \dim_{\mathrm{E}}^{B_n} \left(\mathscr{F}_\Theta, 2\sqrt{\frac{\varepsilon}{n}} \right) \tag{5.56}$$

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

Proposition 5.5.6.
Under Assumptions 5.1 and 5.2, 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 $(\mathscr{F}_{\Theta,n})_{n\in\mathbb{N}}$ of subsets of $\mathscr{F}_\Theta$ of the form

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

we have

$$\sum_{n=1}^{N_t} \sup_{\substack{\mu_1\in\mathscr{F}_{\Theta,n}\\ \mu_2\in\mathscr{F}_{\Theta,n}}} \|\mu_1 - \mu_2\|(X_{\tau_n}^{\alpha,\theta}, \alpha_{\tau_n}) \le 2\beta_{N_t}\sqrt{\mathrm{d}_{\mathrm{E},N_T}} + 2L\mathrm{d}_{\mathrm{E},N_T}(1 + \sup_{s\le t}\|X_s^{\alpha,\theta}\|), \tag{5.57}$$

and

$$\begin{aligned}\sum_{n=1}^{N_t} \sup_{\substack{\mu_1\in\mathscr{F}_{\Theta,n}\\ \mu_2\in\mathscr{F}_{\Theta,n}}} &\|\mu_1 - \mu_2\|^2(X_{\tau_n}^{\alpha,\theta}, \alpha_{\tau_n}) \\ &\le 4\beta_{N_T}^2\mathrm{d}_{\mathrm{E},N_T}\left(3+\log\left(\frac{8\varepsilon^2 N_t L_0^2\left(1+\sup_{s\le t}\|X_s^{\alpha,\theta}\|\right)}{16\beta_{N_t}^4\mathrm{d}_{\mathrm{E},N_T}^2}\right)\right) \\ &\quad + 2\mathrm{d}_{\mathrm{E},N_t}(1+2\beta_{N_t}^2\mathrm{d}_{\mathrm{E},N_t})\left(1+8\varepsilon^2L_0^2\left(1+\sup_{s\le t}\|X_s^{\alpha,\theta}\|^2\right)\right).\end{aligned} \tag{5.58}$$

Proof. Immediate by applying Proposition 5.5.4 to our setting, as we did in the proof of Proposition 5.3.2. □

Under the event of Proposition 5.3.2, which ensures that $\theta^* \in \cap_{n\in\mathbb{N}}\mathcal{C}_n(\delta)$, we can derive from Proposition 5.5.6 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 5.3.3.
Under Assumptions 5.1 and 5.2, 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 $1-\delta$

$$\sum_{n=1}^{N_t} \|\mu_{\hat{\theta}_n} - \mu_{\theta^*}\|(X_{\tau_n}^{\alpha,\theta^*}, \alpha_{\tau_n}) \leq \tilde{\mathcal{O}}\left(\sqrt{\varepsilon d_{E,N_T} \log(\mathscr{N}_{N_T}^{\varepsilon}) N_t} + \varepsilon d_{E,N_T}\right), \tag{5.16}$$

and

$$\sum_{n=1}^{N_t} \|\mu_{\hat{\theta}_n} - \mu_{\theta^*}\|^2(X_{\tau_n}^{\alpha,\theta^*}, \alpha_{\tau_n}) \leq \tilde{\mathcal{O}}\left(d_{E,N_T} \log(\mathscr{N}_{N_T}^{\varepsilon})\right). \tag{5.17}$$

Proof. This follows from Proposition 5.5.6 by choosing $(\tilde{\beta}_n)_{n\in\mathbb{N}} := (\beta_n(\delta))_{n\in\mathbb{N}}$ and $(\mathscr{F}_{\Theta,n})_{n\in\mathbb{N}} := (\mathcal{C}_n(\delta))_{n\in\mathbb{N}}$, 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 $(\tilde{\beta}_n)_{n\in\mathbb{N}}$, $(\mathscr{F}_{\Theta,n})_{n\in\mathbb{N}}$, and $(\tilde{\mu}_n)_{n\in\mathbb{N}}$ are adapted to $\mathbb{F}$, and therefore we can apply Proposition 5.5.6 on the event of Proposition 5.3.2 without issues. This yields

$$\sum_{n=1}^{N_t} \|\mu_{\hat{\theta}_n} - \mu_{\theta^*}\|(X_{\tau_n}^{\alpha,\theta^*}, \alpha_{\tau_n}) \leq 2\beta_{N_t}(\delta)\sqrt{d_{E,N_T}} + 2\varepsilon L_0 d_{E,N_T}(1 + H_\delta(N_T)),$$

and

$$\begin{aligned}\sum_{n=1}^{N_t} &\|\mu_{\hat{\theta}_n} - \mu_{\theta^*}\|^2(X_{\tau_n}^{\alpha,\theta^*}, \alpha_{\tau_n}) \\ &\leq 4\beta_{N_T}(\delta)^2 d_{E,N_T}\left(3 + \log\left(\frac{8\varepsilon^2 N_t L_0^2(1 + H_\delta(N_t))}{16\beta_{N_t}(\delta)^4 d_{E,N_T}^2}\right)\right) \\ &\quad + 2d_{E,N_t}(1 + 2\beta_{N_t}(\delta)^2 d_{E,N_t})(1 + 8\varepsilon^2 L_0^2(1 + H_\delta(N_t)^2)).\end{aligned}$$

To obtain the bounds of (5.16)–(5.17), it suffices to recall the definitions of $\beta_n(\delta)$ (i.e. (5.14)) and $H_\delta(n)$ (i.e. (5.30)). □

5.6 Planning and Diffusive Limit Approximation

The results of this section build upon those of Chapter 4 but with specialised results for our setting. In Chapter 4 the key results of this section (Propositions 5.3.4 to 5.3.6) are shown under a stronger and more abstract set of assumptions (Assumptions 4.2, 4.3 and 4.5). For the comfort of the reader, we thus present only the necessary steps to extend its results to our assumptions. Since our assumptions do not directly subsume theirs, we exhibit in

each case from Assumptions 5.1 and 5.2 how to recover the keystone results that underpin the technical arguments of Chapter 4.

We begin with the well-posedness results for the pure jump case (Proposition 5.3.4) and the diffusive limit case (Proposition 5.3.5), and then focus on the approximation result linking the two regimes (Proposition 5.3.6). In Chapter 4, Proposition 5.3.4 corresponds to Theorem 4.2.1 and Remark 4.2.2. In Section 5.6.1, we show how it follows from Assumptions 5.1 and 5.2 by proving the two intermediary results used in Chapter 4 to prove the result.

5.6.1 Proof of Proposition 5.3.4

Proposition 5.3.4. (Adapted from Theorem 4.2.1 and Remark 4.2.2) Under Assumptions 5.1 and 5.2, 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

$$\varepsilon\rho_\theta^* = \max_{a\in\mathbb{A}} \left\{ \mathbb{E}[W_\theta^*(x + \mu_\theta(x,a) + \Sigma\xi)] - W_\theta^*(x) + r(x,a) \right\}, \quad (5.18)$$

for any $x \in \mathbb{R}^d$;

(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 of (5.18), and $\pi_\theta^* \circ X^{\pi_\theta^*,\theta}$ is an optimal Markov control, i.e. $\rho_\theta^{\pi_\theta^*}(\cdot) \equiv \rho_\theta^*$.

In Chapter 4, Theorem 4.2.1 and Remark 4.2.2 follow from Lemmas 4.A.1 and 4.A.2, which respectively give a mixing condition and a moment bound for $X^{\alpha,\theta}$. We already proved Lemma 4.A.2 with Lemma 5.4.4 (reproduced below for reference). Moreover, Lemma 5.6.1 which reproduces Lemma 4.A.1 holds with only minor modifications of the proof from Chapter 4.

Lemma 5.4.4.
Under Assumptions 5.1 and 5.2, for any $p \geq 2$, there is a constant $\mathfrak{c}'_p > 0$ independent of ε such that

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

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

Lemma 5.6.1.
For any $(x, x') \in \mathbb{R}^d \times \mathbb{R}^d$, $\theta \in \Theta$, *and* $\alpha \in \mathcal{A}$,

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

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

Proof. We can follow the proof of Lemma 4.A.1 using Assumption 5.2 directly without resorting to the higher order Lyapunov function ζ which is used therein. □

5.6.2 Proof of Proposition 5.3.5

In Chapter 4, Proposition 5.3.5 corresponds to Theorem 4.3.1. In Section 5.6.2, we show that it also follows from Assumptions 5.1 and 5.2 by proving that Assumption 4.5 holds under Assumptions 5.1 and 5.2.

Proposition 5.3.5. (Adapted from Theorem 4.3.1)
Under Assumptions 5.1 and 5.2, 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

$$\bar{\rho}_\theta^* = \max_{a \in \mathbb{A}} \left\{ \bar{\mu}_\theta(x, a)^\intercal \nabla \bar{W}_\theta^*(x) + \bar{r}(x, a) \right\} + \frac{1}{2} \operatorname{Tr}\left[\bar{\Sigma} \bar{\Sigma}^\intercal \nabla^2 \bar{W}_\theta^*(x) \right] \tag{5.19}$$

for any $x \in \mathbb{R}^d$.

(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 (5.19), and $\bar{\pi}_\theta^* \circ \bar{X}^{\bar{\pi}_\theta^*, \theta}$ is an optimal Markov control, i.e. $\bar{\rho}_\theta^{\bar{\pi}_\theta^*}(\cdot) \equiv \bar{\rho}_\theta^*$.

***Remark* 5.6.1.** *Proposition 5.3.5.*(iii) *is not stated as is in Theorem 4.3.1, but it follows from it by the same arguments as Remark 4.2.2.*

Proposition 5.3.5, such as it is stated in Theorem 4.3.1 relies on Assumption 4.5. This assumption contains two conditions, which we will show respectively in Lemmas 5.6.2 and 5.6.3.

As detailed in Remark 4.3.1.a), the first condition can be shown by proving an analog of Lemma 4.A.1 for the diffusive limit process (5.8). In terms of arguments of the proof, this analog requires only a change in the infinitesimal generator used in Itô's Lemma[6]. In the proof of Lemma 5.6.2, we, therefore, show how to adapt Lemma 4.A.1 to the generator of the diffusion under Assumptions 5.1 and 5.2.

In the proof of Lemma 4.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 5.6.2. Then one takes the vanishing discount limit in the Hamilton-Jacobi-Bellman equation using the theory of viscosity solutions to complete the proof.

Lemma 5.6.2.
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 (5.4), we obtain as $\varepsilon \to 0$

$$(\bar{\mu}_\theta(x_1, a) - \bar{\mu}_\theta(x_2, a))^\intercal \nabla \mathscr{V}(x_1 - x_2) \leq -\mathfrak{c}_{\mathscr{V}} \mathscr{V}(x_1 - x_2). \tag{5.59}$$

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 $\mathscr{V}_\iota := \mathscr{V}\,\mathrm{erf}(\iota\mathscr{V})$ 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 (5.59) everywhere.

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

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

By Assumption 5.2, this implies that

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

[6] For a general overview of this sort of stability results and of Stochastic Lyapunov conditions in the diffusive case, see e.g. [71, § 5.7].

$$\leq -\mathfrak{c}_{\mathscr{V}} \mathscr{V}_\iota(z)\,. \tag{5.60}$$

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 (5.60), this yields, for $t \geq t_0 \geq 0$,

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

We then conclude by the same ODE comparison argument as in the proof of Lemma 5.4.4 and then pass to the limit as $\iota \to 0$ to obtain the claimed result using Assumption 5.2.(i). □

While Lemma 5.6.2 showed that Assumption 4.5.(i) is implied by Assumptions 5.1 and 5.2. It remains now to verify Assumption 4.5.(ii). Note that by Remark 4.3.1.b), an equation of the form of (4.14) is sufficient to do so. Lemma 5.6.3 gives exactly this result with (5.61), by noting that (4.15) holds by Assumption 5.2.

Lemma 5.6.3.
Let Assumptions 5.1 and 5.2 hold. Then, for any $p \geq 2$, there are $(\bar{\mathfrak{c}}_p, \bar{\mathfrak{c}}_p') \in \mathbb{R}_+^2$ such that

$$\bar{\mu}_\theta(x,a)^\intercal \nabla \mathscr{V}(x)^p + \mathrm{Tr}\left[\bar{\Sigma}\bar{\Sigma}^\intercal \nabla^2 \mathscr{V}(x)^p\right] \leq -\bar{\mathfrak{c}}_p \mathscr{V}(x)^p + \bar{\mathfrak{c}}_p' \tag{5.61}$$

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 (5.4), which gives

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

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

$$\begin{aligned}&(\bar{\mu}_\theta(x,a) - \bar{\mu}_\theta(x',a))^\mathsf{T} \nabla \mathscr{V}(x - x') \\ &\leq -\mathfrak{c}_{\mathscr{V}} \mathscr{V}(x - x') - \frac{\varepsilon}{2} (\bar{\mu}_\theta(x,a) - \bar{\mu}_\theta(x',a))^\mathsf{T} \nabla^2 \mathscr{V}(\hat{x}) (\bar{\mu}_\theta(x,a) - \bar{\mu}_\theta(x',a)) \,.\end{aligned}$$

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

$$\begin{aligned}&(\bar{\mu}_\theta(x,a) - \bar{\mu}_\theta(x',a))^\mathsf{T} \nabla \mathscr{V}(x - x') + \frac{1}{2} \operatorname{Tr}\left[\bar{\Sigma}\bar{\Sigma}^\mathsf{T} \nabla^2 \mathscr{V}(x - x')\right] \\ &\quad \leq -\mathfrak{c}_{\mathscr{V}} \mathscr{V}(x - x') + \frac{d}{2} \|\bar{\Sigma}\|_{\mathrm{op}}^2 M'_{\mathscr{V}} \,.\end{aligned}$$

Taking $x' = 0$ implies that

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

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 $\mathscr{V} \in \mathcal{C}^2(\mathbb{R}^d \setminus \{0\}; \mathbb{R}_+)$ and vanishes at 0 (see Assumption 5.1), $\mathscr{V}(\cdot)^p$ 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

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

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

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

for all $x \in \mathbb{R}^d$. □

5.6.3 Proof of Proposition 5.3.6

Propositions 5.3.4 and 5.3.5 together ensure that both the prelimit and limit regimes are well posed, while Proposition 5.3.6 gives the rate of convergence

of the control problems along this limit. This result is essentially contained in the proof of Theorem 4.3.2, but since its statement is different, we include a proof for completeness in Section 5.6.3.

Proposition 5.3.6. (Adapted from Theorem 4.3.2)
Under Assumptions 5.1 and 5.2, for any $\gamma \in (0,1)$, there is a constant $C_\gamma > 0$, independent of ε, such that, for any $\theta \in \Theta$,

$$|\bar{\rho}^*_\theta - \rho^*_\theta| \leq C_\gamma \varepsilon^{\frac{\gamma}{2}} \text{ and } \rho^*_\theta - \rho^{\bar{\pi}^*_\theta}_\theta(0) \leq C_\gamma \varepsilon^{\frac{\gamma}{2}}. \tag{5.20}$$

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

$$\varepsilon \rho^{\bar{\pi}^*_\theta}_\theta(0) = \mathbb{E}[\bar{W}^*_\theta(x + \mu_\theta(x,a) + \Sigma\xi)] - \bar{W}^*_\theta(x) + r(x, \bar{\pi}^*_\theta(x)) + e_\theta(x) \tag{5.21}$$

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

Proposition 5.3.6 can be proven by modification of the proof of Theorem 4.3.2 to which it corresponds. Below, we produce a self-contained proof in order to clarify how (5.21) is derived from the proof.

Proof. The first part of Proposition 5.3.6, that is (5.20), corresponds to Theorem 4.3.2, which we previously showed holds in our setting by verifying its assumptions. We now prove the second claim. Let

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

for all $(x,a) \in \mathbb{R}^d \times \mathbb{A}$. From (5.19), and Proposition 5.3.5.(iii) we have

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

which implies

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

for every $x \in \mathbb{R}^d$. Note that $|\delta r_\varepsilon^\theta(x, \bar{\pi}_\theta^*(x))| \leq \sup_{a\in\mathbb{A}}|\delta r_\varepsilon^\theta(x,a)|$, which is bounded by $c_\gamma \varepsilon^{\gamma/2}(1 + \|x\|^3)$ for some constant $c_\gamma > 0$, by Theorem 4.3.2. An application of (5.20) yields

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

and, at the same time, $\bar{\rho}_\theta^* - \rho_\theta^{\bar{\pi}_\theta^*}(0) \geq \bar{\rho}_\theta^* - \rho_\theta^* \geq -C_\gamma \varepsilon^{\gamma/2}$. Therefore, there is a function $e_\theta : \mathbb{R}^d \to \mathbb{R}$ such that (5.21) holds, which also satisfies

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

□

5.7 Regret Analysis

In this section, we complete the analysis of the regret of Algorithm 1 and prove Theorem 5.3.1. First, we will give the regret decomposition, and then in the later subsections, we will bound terms one by one calling upon the results of the previous appendices.

Theorem 5.3.1.
Under Assumptions 5.1 and 5.2, 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

$$\mathcal{R}_T(\varpi) \leq 2C_\gamma \varepsilon^{\frac{\gamma}{2}} T + C\sqrt{\mathrm{d}_{\mathrm{E},T\varepsilon^{-1}} \log(\mathscr{N}^{\varepsilon}_{T\varepsilon^{-1}}) T \log(T\delta^{-1})} \tag{5.11}$$

with probability at least $1-\delta$, in which $\mathrm{d}_{\mathrm{E},T\varepsilon^{-1}}$ is the $2\varepsilon/\sqrt{T}$-eluder dimension (see [107, Def. 4.] and (5.56) in Section 5.5.2) 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 $\varepsilon^2\|\bar{\Sigma}\|_{\mathrm{op}}^2/T$-log-covering number of this same restricted class.

5.7.1 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 5.3.5, 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, $\varpi_{\tau_n} = \bar{\pi}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n})$, so that

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

At the heart of the regret decomposition is the use of the HJB-type equation (5.21) applied for each n at the point $X^{\varpi,\theta^*}_{\tau_n}$. For clarity, let us introduce for all $n \in \mathbb{N}$ the random variable $\tilde{X}^{\varpi,\theta_n}_{\tau_{n+1}}$ equal in distribution, conditionally on $\mathcal{F}_{\tau_n}$, to the random variable $\psi^{\varepsilon}_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) + \Sigma\xi_{n+1}$. With this notation (5.21) becomes

$$\begin{aligned}&\varepsilon\rho^{\bar{\pi}^*_{\theta_n}}_{\theta_n}(0)\\ &= \mathbb{E}[\bar{W}^*_{\theta_n}(\tilde{X}^{\varpi,\theta_n}_{\tau_{n+1}})|\mathcal{F}_{\tau_n}] - \bar{W}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}) + r(X^{\varpi,\theta^*}_{\tau_n}, \bar{\pi}^*_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n})) + e_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n}).\end{aligned} \tag{5.62}$$

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 (5.62) yields

$$\begin{aligned}\mathcal{R}_T(\varpi) &= T\rho^*_{\theta^*} - \sum_{n=1}^{N_T} \varepsilon\rho^{\bar{\pi}^*_{\theta_n}}_{\theta_n}(0) + \sum_{n=1}^{N_T} e_{\theta_n}(X^{\varpi,\theta^*}_{\tau_n})\\ &\quad + \sum_{n=1}^{N_T} \mathbb{E}\left[\bar{W}^*_{\theta_n}\left(\tilde{X}^{\varpi,\theta_n}_{\tau_{n+1}}\right)|\mathcal{F}_{\tau_n}\right] - \bar{W}^*_{\theta_n}\left(X^{\varpi,\theta^*}_{\tau_n}\right).\\ &= (T - \varepsilon N_T)\rho^*_{\theta^*} \qquad (R_1)\\ &\quad + \varepsilon\sum_{n=1}^{N_T}\left(\rho^*_{\theta^*} - \rho^{\bar{\pi}^*_{\theta_n}}_{\theta_n}(0)\right) + \sum_{n=1}^{N_T} e_{\theta_n}\left(X^{\varpi,\theta^*}_{\tau_n}\right) \qquad (R_2)\\ &\quad + \sum_{n=1}^{N_T} \mathbb{E}\left[\bar{W}^*_{\theta_n}\left(\tilde{X}^{\varpi,\theta_n}_{\tau_{n+1}}\right)|\mathcal{F}_{\tau_n}\right] - \bar{W}^*_{\theta_n}\left(X^{\varpi,\theta^*}_{\tau_n}\right). \qquad (5.63)\end{aligned}$$

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, (5.63), 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 $(\tilde{X}^{\varpi,\theta^*}_{\tau_{n+1}})_{n\in\mathbb{N}}$ while the random terms use the real transitions $(X^{\varpi,\theta^*}_{\tau_{n+1}})_{n\in\mathbb{N}}$.

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

$$\tilde{X}^{\varpi,\theta}_{\tau_{n+1}} \overset{\mathrm{d}}{=} X^{\varpi,\theta^*}_{\tau_{n+1}} - \mu_{\theta^*}(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}) + \mu_\theta(X^{\varpi,\theta^*}_{\tau_n}, \varpi_{\tau_n}), \tag{5.64}$$

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:

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

Using (5.64), and the uniform L_W-Lipschitzness of $(\bar{W}^*_{\theta_n})_{n\in\mathbb{N}}$, we get for each $n \in \mathbb{N}$

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

and thus the regret term (5.63) is bounded by

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

in which

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

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

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

At the end of this decomposition, we have constructed a true martingale in (R_5), which we bound in Section 5.7.6. The first term (R_3) accumulates

the fit error described in Proposition 5.3.3, up to the lazy updates, which we study in Section 5.7.4. The term (R_4) is bounded by the number of effective updates of θ_n (namely, $\sum_{n=1}^{N_T} 1_{\{\theta_{n+1}\neq\theta_n\}}$) in Section 5.7.5. Finally, the bounds on (R_1) and (R_2) are given in Sections 5.7.2 and 5.7.3 respectively.

To combine the high-probability events used to bound (R_1) and (R_5), with the event of Proposition 5.3.2 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.

5.7.2 The Poisson clock variation term (R_1)

We bound (R_1) using Lemma 5.7.1 which is a standard sub-exponential concentration result, see e.g. [40, Lemma 4.1]. It implies

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

Lemma 5.7.1.
For any $T \in \mathbb{R}_+^*$ and $\delta \in (0,1)$,

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

Proof. Let $v := \varepsilon^{-1}T$. For any $\lambda \in [-1,1]$, $\mathbb{E}[e^{\lambda(N_T - v)}] = \exp(v(e^\lambda - 1 - \lambda)) \leq e^{\lambda^2 v}$. Therefore, N_T is $(\sqrt{2}v, 1)$-sub-exponential (see [40] once more) and therefore,

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

which implies

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

□

5.7.3 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_n} of Proposition 5.3.6, which yields

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

We decompose the remaining term of (R_2) into

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

by applying Proposition 5.3.6 to all but the second pair of terms.

On the event of Proposition 5.3.2, 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

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

5.7.4 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 5.5.6. Nevertheless, the lazy update scheme is designed not to degrade the overall learning performance by more than a constant factor. Leveraging (5.7), we get

$$\sum_{i=1}^{n-1} \left\|\mu_{\theta_n}(X_{\tau_i}^{\varpi,\theta^*}, \varpi_{\tau_i}) - \mu_{\theta^*}(X_{\tau_i}^{\varpi,\theta^*}, \varpi_{\tau_i})\right\| \leq \begin{cases} 2\beta_n(\delta/3) \text{ if } n < n_k \\ \beta_n(\delta/3) \text{ if } n = n_k \end{cases}. \quad (5.65)$$

As a result, μ_{θ_n} is chosen within an inflated version of $\mathcal{C}_n(\delta/3)$, defined as in (5.6) 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 5.3.3, by applying Proposition 5.5.6 to the inflated confidence sets, up to the constant factor 2 in the bounds. Therefore, on the event of Proposition 5.3.2, we have

$$R_3 = L_W \sum_{n=1}^{N_T} \|\mu_{\theta_n} - \mu_{\theta^*}\|(X_{\tau_n}^{\varpi,\theta^*}, \varpi_{\tau_n})$$

$$\leq L_W \sum_{n=1}^{N_T} \|\mu_{\theta_n} - \mu_{\hat{\theta}_{n_k}}\|(X_{\tau_n}^{\varpi,\theta^*}, \varpi_{\tau_n}) + L_W \sum_{n=1}^{N_T} \|\mu_{\hat{\theta}_{n_k}} - \mu_{\theta^*}\|(X_{\tau_n}^{\varpi,\theta^*}, \varpi_{\tau_n})$$
$$\leq 3L_W \left(2\beta_{N_T}(\delta/3)\sqrt{\mathrm{d}_{\mathrm{E},N_t}} + \mathrm{d}_{\mathrm{E},N_t} H_{\delta/3}(N_T)\right).$$

5.7.5 The lazy-update term (R_4)

We observe that (R_4) is bounded by

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

Thus, bounding the number of updates with Lemma 5.7.2 bounds (R_4).

Lemma 5.7.2.
Let Assumptions 5.1 and 5.2 hold. Then, Algorithm 1 generates episodes which satisfy for all $T \in \mathbb{R}_+$ and $\delta \in (0,1)$

$$\sum_{n=1}^{N_T} !1_{\{\theta_n \neq \theta_{n+1}\}} \leq 4\beta_{N_T}\left(\frac{\delta}{3}\right)^2 \mathrm{d}_{\mathrm{E},N_T}\left(3 + \log\left(\frac{8\varepsilon^2 N_t L_0^2(1+\sup_{s\leq t}\|X_s^{\varpi,\theta^*}\|)}{16\beta_{N_t}\left(\frac{\delta}{3}\right)^4 \mathrm{d}_{\mathrm{E},N_T}^2}\right)\right)$$
$$+ 2\mathrm{d}_{\mathrm{E},N_t}\left(1 + 2\beta_{N_t}(\delta/3)^2\mathrm{d}_{\mathrm{E},N_t}\right)\left(1 + 8\varepsilon^2 L_0^2\left(1 + \sup_{s\leq t}\|X_s^{\varpi,\theta^*}\|^2\right)\right).$$

Proof. Consider $k \in \mathbb{N}^*$, by (5.7), each time we trigger an update we have

$$2\beta_{n_k}(\delta/3)^2 < \sup_{\mu_\theta \in \mathcal{C}_{n_{k-1}}(\delta)} \|\mu_\theta - \mu_{\hat{\theta}_{n_{k-1}}}\|_{n_k}^2$$
$$\leq \sup_{\mu_\theta \in \mathcal{C}_{n_{k-1}}(\delta)} \|\mu_\theta - \mu_{\hat{\theta}_{n_{k-1}}}\|_{n_{k-1}}^2$$

$$+ \sup_{\mu_\theta \in \mathcal{C}_{n_{k-1}}(\delta)} \sum_{n=n_{k-1}+1}^{n_k} \left\| \mu_\theta(X^{\varpi,\theta}_{\tau_n}, \varpi_{\tau_n}) - \mu_{\hat{\theta}_{n_{k-1}}}(X^{\varpi,\theta}_{\tau_n}, \varpi_{\tau_n}) \right\|^2$$

$$\leq \beta_{n_k}(\delta/3)^2 + \sum_{n=n_{k-1}+1}^{n_k} \Lambda(\mathcal{C}_{n_{k-1}}(\delta/3); X^{\varpi,\theta}_{\tau_n}, \varpi_{\tau_n})^2 .$$

Since the sequence $(\beta_n(\delta/3))_{n\in\mathbb{N}}$ is non-decreasing, by summing over all episodes we have that

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

for all $T \in \mathbb{R}_+$, 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 5.5.6, i.e. (5.58) now yields the desired result as $\beta_0(\delta/3)^2 = \varepsilon$. □

5.7.6 The martingale term (R_5)

For $n \in \mathbb{N}^*$, 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

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

On the one hand, Z_n is a $L_W\|\Sigma\|_{\text{op}}$-Lipschitz function of ξ_n, which is Gaussian and of mean 0. Therefore, by [36, Thm 5.5], Z_n is $L_W \|\Sigma\|_{\text{op}}$-sub-Gaussian and

$$\mathbb{P}\left(\sum_{n=1}^{N_T} Z_n > L_W \|\bar{\Sigma}\|_{\text{op}} \sqrt{2\varepsilon N_T \log\left(\frac{1}{\delta}\right)} \right) \leq \delta . \tag{5.66}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\bar{W}^*_{\theta_{N_T+1}}\left(X^{\varpi,\theta^*}_{\tau_{N_T+1}}\right) | \mathcal{F}_{\tau_{N_T}} \right] \\
&\quad \leq L_W \left(1 + \mathbb{E}\left[\left\| X^{\varpi,\theta^*}_{\tau_{N_T+1}} \right\| | \mathcal{F}_{\tau_{N_T}} \right]\right) \\
&\quad \leq L_W \left(1 + \varepsilon L_0 + (1+\varepsilon L_0) \left\| X^{\varpi,\theta^*}_{\tau_{N_T}} \right\| + \varepsilon^{\frac{1}{2}} \|\bar{\Sigma}\|_{\text{op}} \mathbb{E}\left[\|\xi_{N_T+1}\| | \mathcal{F}_{\tau_{N_T}} \right]\right) \\
&\quad \leq L_W (1+\varepsilon L_0) \left(1 + \sup_{s\leq T} \left\| X^{\varpi,\theta^*}_s \right\|_2 + \varepsilon^{\frac{1}{2}} \|\bar{\Sigma}\|_{\text{op}} \sqrt{d} L_W \right).
\end{aligned} \tag{5.67}$$

Combining (5.66) and (5.67) yields

$$R_5 \leq 2L_W(1+\varepsilon L_0)\left(1+\sup_{s\leq T}\left\|X_s^{\varpi,\theta^*}\right\|+\varepsilon^{\frac{1}{2}}\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{d}L_W\right) + L_W\left\|\bar{\Sigma}\right\|_{\mathrm{op}}\sqrt{2\varepsilon N_T\log\left(\frac{3}{\delta}\right)} \tag{5.68}$$

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

5.7.7 Collecting the bounds

We conclude the proof of Theorem 5.3.1 by collecting all the terms from Sections 5.7.2–5.7.6 and simplifying them. By a union bound over the events listed in steps Sections 5.7.2, 5.7.4 and 5.7.6, with probability at least $1-\delta$

$$\begin{aligned}
&\mathcal{R}_T(\varpi)\\
&\leq 2L_0\left(\sqrt{\varepsilon T\log\left(\frac{6}{\delta}\right)}\vee 2\varepsilon\log\left(\frac{6}{\delta}\right)\right)\\
&+4N_TC_\gamma\varepsilon^{1+\frac{\gamma}{2}}+2C'_\gamma N_T\varepsilon^{1+\frac{\gamma}{2}}(1+H^3_{\delta/3}(N_T))\\
&+6L_W\beta_{N_T}\left(\frac{\delta}{3}\right)\sqrt{\mathrm{d}_{\mathrm{E},N_T}}+2\varepsilon L_0L_W\mathrm{d}_{\mathrm{E},N_T}(1+H_{\delta/3}(N_T))\\
&+2L_W(1+\varepsilon L_0)\left(1+H_{\delta/3}(N_T)+3d\varepsilon^{\frac{1}{2}}\|\bar{\Sigma}\|_{\mathrm{op}}\right)\\
&\quad\times\left(4\beta_{N_T}\left(\frac{\delta}{3}\right)^2\mathrm{d}_{\mathrm{E},N_T}\times\left(3+\log\left(\frac{8\varepsilon^2N_tL_0^2(1+H_{\delta/3}(N_T))}{16\beta_{N_T}(\delta/3)^4\mathrm{d}^2_{\mathrm{E},N_T}}\right)\right)\right.\\
&\qquad\left.+2\mathrm{d}_{\mathrm{E},N_T}\left(1+2\beta_{N_t}\left(\frac{\delta}{3}\right)^2\mathrm{d}_{\mathrm{E},N_T}\right)\left(1+8\varepsilon^2L_0^2\left(1+H_{\delta/3}(N_T)^2\right)\right)\right)\\
&+L_W\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{2\varepsilon N_T\log\left(\frac{3}{\delta}\right)}+2L_W(1+\varepsilon L_0)\left(1+H_{\delta/3}(N_T)+\varepsilon^{\frac{1}{2}}\|\bar{\Sigma}\|_{\mathrm{op}}\sqrt{d}L_W\right),
\end{aligned}$$

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

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

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

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

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

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 that 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.

The optimistic step of Algorithm 1 chooses $\tilde{\theta}_n$ in an inefficient manner. Like in the `UCRL2` algorithm [68], optimistic exploration can be performed at the same time as planning by solving an expanded Hamilton-Jacobi-Bellman equation, that is (5.19) in which the maximum would now be 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 [68].

The with which we quantify learning progress in order to design the lazy update scheme (see (5.7)) remains fundamentally discrete. Through simpler heuristics, it might be possible to obtain computationally cheaper lazy update schemes. 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.

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 (5.19) which are built on a case-by-case basis for $d > 1$, see [78].

Chapter 6

Real-Time Optimisation for Online Learning in Auctions

In display advertising, a small group of sellers and bidders face each other in up to 10^{12} auctions a day. In this context, revenue maximisation via monopoly price learning is a high-value problem for sellers. By nature, these auctions are online and produce a very high-frequency stream of data. This results in a computational strain that requires algorithms to be real-time. Unfortunately, existing methods inherited from the batch setting suffer $\mathcal{O}\left(\sqrt{n}\right)$ time/memory complexity at each update, prohibiting their use. In this chapter[a], we provide the first algorithm for online learning of monopoly prices in online auctions whose update is constant in time and memory.

[a] This Chapter appeared as an article in the proceedings of the 37th *International Conference on Machine Learning* (ICML), see [50].

* * *

Contents

* * *

6.1 Introduction

Over the last two decades, online display advertising has become a key monetisation stream for many businesses. The market for the trading of these ads is controlled by a very small number of large intermediaries (less than ten) who buy and sell at auction, which means that a seller-buyer pair might trade together in 10^{10} to 10^{12} auctions per day. Repeated auctions on this scale raise the stakes of revenue maximisation, while making computational efficiency a key consideration. In his 1981 seminal work [92] on revenue maximisation, Myerson described *the* revenue-maximising auction when the bid distributions of buyers are known. In the context of online display ads these distributions are private, but the large volume of data collected by sellers on buyers paves the way to learning revenue maximising auctions.

The learning problem associated with the Myerson auction has infinite pseudo-dimension [90], making it impossible to learn [101]. Second-price auctions with personalised reserve prices (i.e. different for each bidder) stand as the commonly accepted compromise between optimality and tractability. They provide a 2-approximation to the revenue of the Myerson auction while securing finite pseudo-dimension [105].

Second-price auctions with personalised reserves can be either *eager* or *lazy*. In the eager format, the item goes to the highest bidder *amongst* those who cleared their reserve prices and goes unsold if none of them did. In the lazy format, the item goes to the highest bidder *if* he cleared his reserve price and goes unsold if he did not. While an optimised *eager* version leads to better revenue than an optimised *lazy* version, solving the eager auction's associated Empirical Risk Minimisation (ERM) problem is NP-hard [99], and even APX-hard [105]. In contrast, solving the ERM problem for the *lazy* version can be done in polynomial time [105]: it amounts to computing a bidder-specific quantity called the *monopoly price*. Not only is the monopoly price the optimal reserve in the lazy second-price auction, but it is also a provably good reserve in the eager one [105], and the optimal reserve in posted-price [99]. This makes learning monopoly prices for revenue maximisation an important and popular research direction.

Finding the monopoly price in a repeated second-price auction is a natural sequential decision making problem based on the incoming bids. All three aforementioned settings relating to the monopoly price have been studied: posted-price [10, 27], eager [42, 73, 105], and lazy which we study [26, 27, 38, 89, 106, 110]. Each setting also corresponds to a different observability structure. The offline problems are well understood, but no online method offers the $\mathcal{O}(1)$ efficiency crucial for real-world settings. We focus, therefore, on the key problem of learning monopoly prices, online and efficiently, in stationary and nonstationary cases.

We propose a real-time first-order algorithm that makes online learning of monopoly prices computationally feasible when interacting with stationary and nonstationary buyers. After detailing the setting in Section 6.1.1, we detail the setting and problem we consider. We review, in Section 6.2, the existing approaches and stress the challenges of the problem including overcoming computational complexities. Our approach, based on convolution and the $\mathcal{O}(1)$ Online Gradient Ascent algorithm, is described in Section 6.3. We study performance for stationary bidders in Section 6.4 with $\mathcal{O}(n^{-1/2})$ convergence rate to the monopoly price, and for nonstationary bidders in Section 6.5 with $\mathcal{O}(\sqrt{N})$ dynamic regret.

6.1.1 Setting

A key property of a personalised reserve price in a lazy second-price auction is that it can be optimised separately for each bidder [99]. For a bidder whose bids are sampled independently and identically from a distribution with Cumulative Distribution Function (CDF) F, the optimal reserve price is the monopoly price r^*, i.e. the maximiser of the monopoly revenue defined as

$$r \in \mathbb{R}_+ \mapsto \Psi^F(r) := r(1 - F(r)) \in \mathbb{R}_+ . \tag{6.1}$$

Thus, without loss of generality, we study each bidder separately in the following repeated game: the seller sets a reserve price r and simultaneously the buyer submits a bid $b \in [0, \bar{b}]$ drawn from his private distribution F, whose Probability Density Function (PDF) is denoted by f. The seller then observes b which determines the instantaneous revenue

$$(r, b) \in \mathbb{R}_+^2 \mapsto p(r, b) := r1_{r \leq b} \in \mathbb{R}_+ \tag{6.2}$$

which satisfies $\mathbb{E}_F[p(r, b)] = \Psi^F(r)$ for all $r \in \mathbb{R}_+$. In this work, we consider two settings, depending on whether the bid distribution is stationary or not.

Stationarity here means F is fixed for the whole game. We thus have a stream of i.i.d. bids from F, where the seller aims to maximise her long-term revenue Ψ^F. Or, equivalently, tries to construct a sequence of reserve prices $(r_n)_{n \in \mathbb{N}^*}$ adapted to the filtration $\mathbb{F} := (\mathcal{F}_n)_{n \in \mathbb{N}}$ in which $\mathcal{F}_n = \sigma((b_i)_{i=1}^{n-1})$ such that $\Psi^F(r_n) \to \Psi^F(r^*)$ as fast as possible.

In real-world applications, bid distributions may change over time based on the current context. For example, near Christmas, the overall value of advertising might go up since customers spend more readily, and thus bids might increase. The bidder could also refactor his bidding policy for reasons entirely independent of the seller. We relax the stationarity assumption by allowing bids to be drawn according to a sequence of distributions $(F_n)_{n \in \mathbb{N}^*}$

that varies over time. As a result, the monopoly prices $(r_n^*)_{n\in\mathbb{N}^*}$ and optimal monopoly revenues $(\Psi^{F_n})_{n\in\mathbb{N}}$ fluctuate and convergence is no longer defined. Instead, we evaluate the performance of an adaptive sequence of reserve prices $\mathbf{r} := (\mathbf{r}_n)_{n\in\mathbb{N}} \subset \mathbb{R}_+$ by its expected dynamic regret

$$\mathcal{R}_N(\mathbf{r}) = \mathbb{E}\left[\sum_{n=1}^{N} \Psi^{F_n}(r_n^*) - \Psi^{F_n}(\mathbf{r}_n)\right], \tag{6.3}$$

and our objective is to track the monopoly price as fast as possible to minimise dynamic regret.

6.2 Related Work and Challenges

6.2.1 Related work

Lazy second-price auctions have been studied both in batch [89, 99, 106, 110] and online [26, 27, 38] settings. All existing approaches aim to optimise, at least up to a precision of $1/\sqrt{n}$, the Empirical Risk Minimisation objective

$$r \in \mathbb{R}_+ \mapsto \Psi^{\hat{F}_n}(r) := r(1 - \hat{F}_n(r)) = \frac{1}{n}\sum_{i=1}^{n} r1_{r\leq b_i}. \tag{6.4}$$

However, regardless of how well-behaved Ψ^F is, $\Psi^{\hat{F}_n}$ is very poorly behaved for optimisation: it is non-smooth, non-quasi-concave, discontinuous, and is increasing almost everywhere (see Fig. 6.1, center, dashed). Thus, direct optimisation with first order methods is not applicable. Some attempts have been made in the batch setting to optimise surrogate objectives but ended up with an irreducible bias [106] or with hyper-parameters whose tuning is as hard as the initial problem [110]. The classical approach relies on sorting the bids $(b_i)_{i=1}^n$ to be able to enumerate $\Psi^{\hat{F}_n}$ linearly over $(b_i)_{i=1}^n$, see e.g. [89, 99]. A popular improvement in terms of complexity, especially used in online approaches [26, 27] consists in applying the same principle on a regular grid of resolution $1/\sqrt{n}$, which in the end provides an update with complexity $\mathcal{O}(\sqrt{n})$ and a memory requirement of $\mathcal{O}(\sqrt{n})$.

This idea of discretising the bid space was widely adopted in partially observable settings, for instance online eager and posted-price auctions, as it reduces these problems to multi-armed bandits with their well-studied algorithms [42, 73, 104] at the price of still suffering the same update and memory complexities of $\mathcal{O}(\sqrt{n})$.

Numerous approaches with adversarial bandits also followed this discretisation approach to adapt the `Exp3` or `Exp4` algorithms to all of these settings, see, amongst others, [38, 46, 73]. Furthermore, bandit algorithms also

allow for the handling of nonstationary bidders. However, the work of [10] stresses that bidders cannot behave in an arbitrary way, as they optimise their own objective that is not incompatible with the seller's[1]. Hence, the non-stationarity mostly comes from the item's value changing over time. This suggest adapting a regular stochastic algorithms (ERM, the Upper Confidence Bound algorithm, …), e.g. using sliding windows [60, 82].

Non-smooth or non-differentiable objectives such as $\Psi^{\hat{F}_n}$, for any $n \in \mathbb{N}^*$ have been studied in both stochastic and zeroth-order optimisation. In both, convolution smoothing has been employed to circumvent these problems. In [51] stochastic gradient with decreasing convolution smoothing is studied for the convex case. Unfortunately, very few distributions yield a concave Ψ^F. In zeroth-order optimisation, the only feedback received for an input is the value of the objective at that input. In this setting, Flaxman, Kalai, and Mac Mahan [55] perturb their inputs to estimate a convolved gradient. In contrast, we obtain a closed form and do not need to perturb inputs.

6.2.2 Challenges

The directing challenge of our line of work is to devise an online learning algorithm for monopoly prices with minimal cost, to handle very large real-world data streams. With 10^{10} daily interactions in one seller-bidder pair, it is acceptable to forfeit some convergence speed in exchange for the feasibility of the algorithm. It is not possible to accept update complexity or memory requirement scaling with n. Our objective is thus to find a method which: converges to r^* in the stationary setting or has a low regret $\mathcal{R}_N$ in the non-stationary one; has $\mathcal{O}(1)$ memory footprint; and computes the next reserve price r_{n+1} with $\mathcal{O}(1)$ computations.

Unfortunately, none of the previously proposed methods fit these requirements. On one hand, all methods based on solving ERM by sorting [42, 104] need to keep all past bids in memory ($\mathcal{O}(n)$ dependence) and their update steps require at best $\mathcal{O}\left(\sqrt{n}\right)$ computations. On the other hand, methods such as `Exp3` or `Exp4`, see e.g. [38, 46], which are adversarial, are designed for finite action spaces and thus need to discretise $[0, \bar{b}]$ into $\sqrt{n}$ intervals to keep their regret guarantees which also leads to a complexity of $\mathcal{O}\left(\sqrt{n}\right)$ from sampling over the discretisation to compute r_{n+1}.

First-order methods such as Online Gradient Ascent are standard tools in online learning and enjoy $\mathcal{O}(1)$ update and memory. This makes them great candidates for our problem. OGA requires three ingredients to converge: an objective whose gradients always point towards the optimum[2], and

[1] An auction is not a zero-sum game: if the item goes unsold, neither player receives payoff.

[2] A condition known as pseudo-concavity or variational coherency, see also Section 6.A.3.

a gradient estimator which has bounded variance and is unbiased. Unfortunately, discontinuity of p makes $\mathrm{D}p$ a biased estimator of $\mathrm{D}\Psi^F$. A natural approach is to construct a surrogate for p which has unbiased gradients and preserves the other two conditions.

Optimising a surrogate objective inherently creates a bias, which has to be reduced over time. To do so without breaking the convergence of OGA, we must conduct a careful finite-time analysis of the algorithm, which is an analytical challenge. We must re-analyse classical results (e.g. [91]) for varying objectives: the challenge is to design a bias reduction procedure, and then integrate it into these proofs to show we preserve consistency.

Resolving the above challenges is sufficient to achieve efficient convergence in the stationary setting. However, it is not sufficient in order to track nonstationary bid distributions. Taking a constant surrogate and learning rate, it is possible to adapt the stationary solution to the nonstationary case and keep its computational efficiency. The challenge is to devise this adaptation and then to derive (sub-linear) regret for it.

We propose a method based on convolutional smoothing to design surrogates in pseudo-concave problems with biased gradients. We use it to create a first-order real-time optimisation algorithm which reduces the surrogate's bias during optimisation. We prove convergence and give rates in the stationary setting and dynamic regret bounds for tracking.

We first translate standard auction theory assumptions (e.g. increasing virtual value) into properties of generalised concavity of the monopoly revenue, see Proposition 6.3.1. Next, we introduce our smoothing method and show, in Proposition 6.3.2, that it preserves the properties of Proposition 6.3.1 while offering arbitrary smoothness, which fixes the biased gradient problem. Finally, we provide controls on the bias and variance of the gradient estimates of our surrogate in terms of the chosen kernel in Proposition 6.3.3. This makes our surrogate compatible with Online Gradient Ascent.

We construct an algorithm (V-Conv-OGA) that performs gradient ascent while simultaneously decreasing the strength of the smoothing over time, reducing the bias to zero. As a result our algorithm almost surely converges to the monopoly price (Theorem 6.4.1) while enjoying computational efficiency. Further, under a minimum curvature assumption, we provide the rate of convergence and optimal tuning parameters (Theorem 6.4.2 and Corollary 6.4.1). At the cost of a slight degradation in convergence speed (from $\mathcal{O}(n^{-1})$ to $\mathcal{O}(n^{-1/2})$), our algorithm has update and memory complexity of $\mathcal{O}(1)$ which is vital for real-world applications. These Results are summarised in Table 6.1.

Contrary to the stationary setting, when tracking a nonstationary bid distribution we do not decrease the strength of the smoothing over time. When the bias created is smaller than the level of noise, our algorithm can still

achieve sub-linear dynamic regret when tracking changing bid distributions. For reasonably varying distribution, we show a regret bound of $\mathcal{O}(\sqrt{N})$ after N steps, see Theorem 6.5.1 and Corollary 6.5.1.

	Update	Memory	Convergence
ERM	$\mathcal{O}(n)$	$\mathcal{O}(n)$	$\mathcal{O}(n^{-1})$
Discretised ERM	$\mathcal{O}(n^{1/2})$	$\mathcal{O}(n^{1/2})$	$\mathcal{O}(n^{-1})$
V-Conv-OGA	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\mathcal{O}(n^{-1/2})$

Table 6.1: Comparison of our method (V-Conv-OGA) against solving ERM at each step and ERM discretised on a grid of fineness $n^{-1/2}$, in terms of complexity and the convergence (i.e. $|\Psi^F(r_n) - \Psi^F(r^*)|$).

6.3 Smooth Surrogate for First-Order Methods

Our objective, to reiterate, is to design an online optimisation procedure to learn or track the optimal reserve price whose updates require $\mathcal{O}(1)$ computational and memory cost. To this end, we focus on first-order methods and consider classical Online Gradient Ascent. Unfortunately, the specific problem of learning a monopoly price does not provide a way to compute unbiased gradient estimates for Ψ^F from bid samples. We therefore want to design a surrogate that makes p sufficiently smooth so that differentiation and integration (expectation) commute. This is a well-known property of convolutional smoothing, suggesting its use. In addition, we must ensure our surrogate preserves the optimisation properties that Ψ^F already has. These must thus be studied first, before smoothing to obtain a surrogate.

6.3.1 Properties of the monopoly revenue

The standard assumptions of auction theory are made to guarantee that the monopoly price exists and, thus, that the optimisation problem is well-posed. These assumptions generally take the form of assuming quasi-concavity of the monopoly revenue. We refine this characterisation by translating the assumptions we make on F into specific concavity properties of the monopoly revenue in Proposition 6.3.1. For ease of exposition, we will make a smoothness assumption of F in Assumption 6.1.

Assumption 6.1.
$F \in \mathcal{C}^2\big([0,\bar{b}];[0,1]\big)$ and f is supported on $(0,\bar{b})$.

The concavity properties of Ψ^F critically depend on the regularity of its hazard rate function, which we will denote by

$$h_F : b \in [0,\bar{b}) \mapsto \frac{f(b)}{1-F(b)} \in \mathbb{R},$$

notably via the virtual value function ψ_F defined by

$$\psi_F : b \in \mathbb{R}_+ \mapsto b - \frac{1}{h_F(b)} \in \mathbb{R},$$

with the convention that $1/h_F(b) = 0$ for $b \geq \bar{b}$.

Assumption 6.2.
F is strictly regular, i.e. the virtual value ψ_F is strictly increasing on $\mathbb{R}_+$.

Assumption 6.2 is a classical assumption in auction theory, see [75] for a review, and implies a pseudo-concave[3] revenue as will be shown by Proposition 6.3.1. This assumption is satisfied by common distributions, exhaustively listed in [53] as well as by real-world data, see e.g. [98]. On the other hand, Assumption 6.3 strengthens Assumption 6.2 by requiring a minimum curvature around the maximum.

Assumption 6.3.
The hazard rate h_F of F is strongly increasing on $[0,\bar{b})$, i.e. there is $\mu_F > 0$ such that

$$h_F(b') - h_F(b) \geq \mu_F(b'-b) \text{ for any } (b,b') \in [0,\bar{b})^2 \text{ such that } b \leq b'.$$

Proposition 6.3.1.
Let Assumption 6.1 hold, then $\Psi^F \in \mathcal{C}^2\left([0,\bar{b}];\mathbb{R}_+\right)$, $\Psi^F > 0$ on $(0,\bar{b})$ and

(i) if, additionally, Assumption 6.2 holds, then Ψ^F is strictly pseudo-concave;

[3]The definition of this, and other, notions of convex analysis are recalled in Section 6.A.3 for the convenience of the reader.

(ii) if, additionally, Assumption 6.3 holds, then Ψ^F is μ_F-strongly log-concave for some $\mu_F \in \mathbb{R}_+$.

Proof.

1. Under Assumption 6.2, $\psi_F(r)$ is strictly increasing. Moreover, for all $r \in [0, \bar{b}]$,

$$\mathrm{D}\Psi^F(r) = 1 - F(r) - rf(r) = -\psi_F(r)f(r). \tag{6.5}$$

In view of Definition 6.1, we must first show that $\mathrm{D}\Psi^F(r_1)(r_1 - r_2) \geq 0$ implies $\Psi^F(r_1) \geq \Psi^F(r_2)$ for any $(r_1, r_2) \in [0, \bar{b}]^2$. Without loss of generality, let us assume $r_1 \leq r_2$.

Since ψ_F is strictly increasing, $\psi_F(r_1) \leq \psi_F(r_2)$ and as a result $\mathrm{D}\Psi^F(r_1) \leq 0$ if and only if $\psi_F(r_1) \geq 0$, which implies that $\psi_F(r) \geq 0$ if and only if $\mathrm{D}\Psi^F(r) \leq 0$ for all $r \in [r_1, r_2]$. Therefore,

$$0 \geq \int_{r_1}^{r_2} \mathrm{D}\Psi^F(r)\mathrm{d}r = \Psi^F(r_2) - \Psi^F(r_1).$$

The case $\mathrm{D}\Psi^F(r_1) \geq 0$ is treated similarly.

Now, given Definition 6.2, we must show that Ψ^F has exactly one critical point. This follows immediately from (6.5), $f > 0$ on $(0, \bar{b})$ and the fact that ψ_F is strictly increasing, meaning it can only cross 0 once.

2. Under Assumption 6.3, the hazard rate h_F satisfies $h_F(r_2) - h_F(r_1) \geq \mu_F(r_2 - r_1)$, for any $(r_1, r_2) \in [0, \bar{b}]^2$ such that $r_1 \leq r_2$. To show that $\log \Psi^F(r) = \log(r) + \log(1 - F(r))$ is μ_F-strongly concave, it suffices to show that $\log(1 - F(r))$ is μ_F-strongly concave since $\log(r)$ is concave.

Since $F \in \mathcal{C}^1([0, \bar{b}]; [0, 1])$, we can use a characterisation of the strong concavity of $G \log(1 - F(\cdot))$ based on its derivative. Indeed, G is μ_F-strongly concave if and only if

$$(\mathrm{D}G(r_2) - \mathrm{D}G(r_1))(r_2 - r_1) \leq -\mu_F |r_2 - r_1|^2, \text{ for all } (r_1, r_2) \in (0, \bar{b})^2.$$

Consider $(r_1, r_2) \in (0, \bar{b})^2$, and, without loss of generality, let $r_1 \leq r_2$, we have

$$\begin{aligned} \mathrm{D}G(r_2) - \mathrm{D}G(r_1) &= \frac{-f(r_2)}{1 - F(r_2)} - \frac{-f(r_1)}{1 - F(r_1)} = h_F(r_1) - h_F(r_2) \\ &\leq -\mu_F(r_2 - r_1). \end{aligned}$$

Thus G and hence ψ_F are μ_F-strongly concave on $(0, \bar{b})$. □

6.3.2 A method based on smoothing

Proposition 6.3.1 ensures that the first condition for the convergence of Online Gradient Ascent is met under standard assumptions, i.e. Assumption 6.2 or Assumption 6.3. The main difficulty standing in the way of using OGA for revenue optimisation lies in the undesirable shape of the instantaneous revenue p. Indeed, p is non-smooth (discontinuous even) and cannot be used to construct an unbiased estimate of $\mathrm{D}\Psi^F$, which is necessary for first-order methods.

Mohri and Medina [89] suggest replacing $p(\cdot, b)$ by a continuous upper bound. This surrogate can be used for OGA, but it has potentially large areas of zero-gradient, which means it does not learn from all samples. We give a general surrogate construction (based on convolutional smoothing) which can approximate the original monopoly revenue to arbitrary accuracy while preserving the concavity properties of Ψ^F and offering the desired level of smoothness, and which exhibits no areas of zero-gradient.

Formally, given a kernel k (considered a metaparameter), we use convolution smoothing to create surrogates for p and Ψ^F

$$p_k(\cdot, b) = p(\cdot, b) \star k \text{ and } \Psi^F_k = \Psi^F \star k, \text{ respectively.} \tag{6.6}$$

This smoothing guarantees that $\mathrm{D}p_k(\cdot, b)$ is an unbiased estimate of $\mathrm{D}\Psi^F_k$. On Fig. 6.1, we illustrate the effect of this smoothing on p, $\Psi^{\hat{F}_n}$, and Ψ^F.

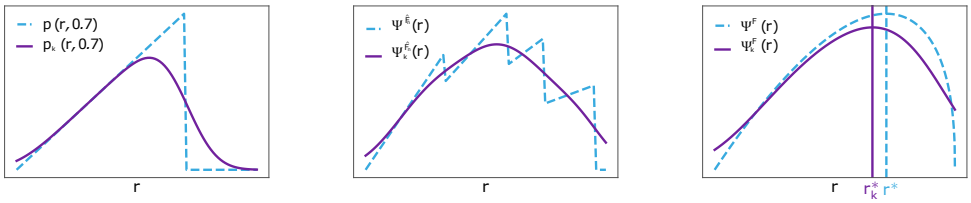

Figure 6.1: The effect of smoothing the monopoly revenue of a bidder with F a Kumaraswamy[4] (1,0.4) distribution with a Gaussian kernel. Left: smoothing of $p(\cdot, b)$ for $b = 0.7$. Center: smoothing of the empirical revenue $\Psi^{\hat{F}_n}$ for some randomly drawn values of b_n. Right: smoothing of the monopoly revenue Ψ^F. Note the differences in the optima (i.e. the surrogate bias).

Let $\mathcal{L}^1(\mathbb{R}; \mathbb{R}_+)$ denote the Lebesgue space of absolutely integrable positive functions on $\mathbb{R}$, $\|\cdot\|_1$ its standard norm, and let $\mathcal{L}^1_1(\mathbb{R}; \mathbb{R}_+)$ denote the set of $\mathcal{L}^1(\mathbb{R}; \mathbb{R}_+)$ functions k such that $\int_{\mathbb{R}} k(x)\mathrm{d}x = 1$. We introduce the set

[4]This distribution satisfies our concavity assumptions and can display highly eccentric behaviour for easy visualisation of the impact of the surrogate. It is similar to a Beta distribution.

$\mathcal{K}$ of admissible kernels which contains all strictly log-concave elements of $\mathcal{C}^1(\mathbb{R};\mathbb{R}_+) \cap \mathcal{L}_1^1(\mathbb{R};\mathbb{R}_+)$. This set contains a large family of kernels, including standard smoothing ones such as Gaussian kernels and mollifiers. Proposition 6.3.2 shows that convolution with elements of $\mathcal{K}$ preserves pseudo- and log-concavity.

Proposition 6.3.2.
Let Assumption 6.1 hold and let $k \in \mathcal{K}$, then:

(i) Ψ_k^F and p_k are in $\mathcal{C}^1(\mathbb{R};\mathbb{R}_+)$,

(ii) $\Psi_k^F = \int p_k(\cdot, b)\mathrm{d}F(b)$ and $\mathrm{D}\Psi_k^F = \int \mathrm{D}p_k(\cdot, b)\mathrm{d}F(b)$,

(iii) if, additionally, Assumption 6.2 holds, Ψ_k^F is strictly pseudo-concave on $(0, \bar{b})$,

(iv) if, additionally, Assumption 6.3 holds, Ψ_k^F is strictly log-concave on $\mathbb{R}$.

Proof. The proof is a straightforward application of the properties of the convolution, the Fubini-Tonelli theorem and of the stability of concavity under convolution detailed in Section 6.A.3. Precisely, we have

(i) Since $k \in \mathcal{C}^1(\mathbb{R};\mathbb{R}_+) \cap \mathcal{L}^1(\mathbb{R};\mathbb{R}_+)$, and since Ψ^F and p are in $\mathcal{L}^1(\mathbb{R};\mathbb{R}_+)$, Ψ_k^F and p_k are in $\mathcal{C}^1(\mathbb{R};\mathbb{R}_+) \cap \mathcal{L}^1(\mathbb{R};\mathbb{R}_+)$.

(ii) Since p, Ψ^F, and k are positive, so are Ψ_k^F and p_k. By the Fubini-Tonelli theorem $\Psi_k^F = \int p_k(\cdot, b)\mathrm{d}F(b)$ and $\mathrm{D}\Psi_k^F = \int \mathrm{D}p_k(\cdot, b)\mathrm{d}F(b)$.

(iii) Under Assumption 6.2, by Proposition 6.3.1, the monopoly revenue Ψ^F is strictly pseudo-concave. Since $\Psi^F(0) = \Psi^F(\bar{b}) = 0$ and $k \in \mathcal{K}$ is strictly log-concave, we can apply Lemma 6.A.2 to guarantee the strict pseudo-concavity of Ψ_k^F.

(iv) Under Assumption 6.3, Ψ^F is strictly log-concave. Since $k \in \mathcal{K}$ is strictly log-concave, convolution with it preserves strict log-concavity (see Theorem 6.A.2), Ψ_k^F is strictly log-concave. □

Proposition 6.3.2 guarantees that the surrogate satisfies the unbiased gradient and pseudo-concavity conditions OGA. Applying OGA to the surrogate Ψ_k^F gives Algorithm 2, in which $\Pi_{\mathcal{I}}$ denote the orthogonal projection onto the interval $\mathcal{I}$. Notice that the usual convolution properties imply that $\mathrm{D}p_k(\cdot, b) = p(\cdot, b) \star \mathrm{D}k$, which is a simple (generally closed-form) computation. For any $k \in \mathcal{K}$, let us denote the variance of the surrogate gradient by

$$V_k := \sup_{r \geq 0} \int |\mathrm{D}p_k(\cdot, b)|^2 \mathrm{d}F(b)$$

By controlling V_k in terms of $\|k\|_\infty$, Proposition 6.3.3 shows that the bounded variance condition of OGA also holds.

Since OGA's three conditions are satisfied, we can guarantee convergence to the maximum r_k^* of Ψ_k^F (see e.g. [32]). This defines Conv-OGA, see Algorithm 2. However, in general, r_k^* is not the monopoly price r^* and the surrogate is biased. For any $k \in \mathcal{K}$, this bias is denoted by

$$B_k := \left|\Psi^F(r^*) - \Psi^F(r_k^*)\right|,$$

on which Proposition 6.3.3 gives a control in terms of the L^1 distance between the Cumulative Distribution Function K of k and the CDF $1_{\mathbb{R}+}$ of the Dirac mass at 0, which we denote δ_0, which is the only kernel to guarantee $r_{\delta_0}^* = r^*$.

For any interval $\mathcal{I} \subset [0, \bar{b}]$, let $\Pi_{\mathcal{I}}$ denote the orthogonal projection onto $\mathcal{I}$.

Algorithm 2 Conv-OGA

Input: r_0, $(\gamma_n)_{n\in\mathbb{N}}$, $k \in \mathcal{K}$, $\mathcal{I} \subset [0, \bar{b}]$.
for $n \in \mathbb{N}$ **do**
  Observe b_n,
  $r_n \leftarrow \Pi_{\mathcal{I}}\big(r_{n-1} + \gamma_t \mathrm{D}p_k(r_{n-1}, b_n)\big)$.
end for

Proposition 6.3.3.
Let Assumption 6.1 hold and let $k \in \mathcal{K}$. Then,

(i) $B_k \leq 2\|\mathrm{D}\Psi^F\|_\infty \|K - 1_{\mathbb{R}+}\|_1$,

(ii) $V_k \leq 1 + \bar{b}\left(1 + \|\mathrm{D}\Psi^F\|_\infty\right)\|k\|_\infty$.

Proof.

(i) The bound on B_k relies on Lemma 6.3.4, which guarantees that for all $r \geq 0$,

$$\left|\Psi^F(r) - \Psi_k^F(r)\right| \leq \|\mathrm{D}\Psi^F\|_\infty \int_{\mathbb{R}} |r|\, k(r)\mathrm{d}r.$$

Decomposing B_k as

$$B_k \leq \Psi^F(r^*) - \Psi_k^F(r^*) + \Psi_k^F(r^*) - \Psi_k^F(r_k^*) + \Psi_k^F(r_k^*) - \Psi^F(r_k^*)$$

$$\leq \Psi^F(r^*) - \Psi_k^F(r^*) + \Psi_k^F(r_k^*) - \Psi^F(r_k^*)$$

and applying (6.9) from Lemma 6.3.4 (see below) twice proves the desired result.

(ii) For all $(r, b) \in \mathbb{R}_+ \times [0, \bar{b}]$, using properties of the convolution and integrating by parts yields

$$\begin{aligned}\mathrm{D}p_k(r,b) &= \int_{\mathbb{R}} p(s,b)\mathrm{D}k(r-s)\mathrm{d}s = \int_0^b s\mathrm{D}k(r-s)\mathrm{d}s \\ &= -\left[sk(r-s)\right]_0^b + \int_0^b k(r-s)\mathrm{d}s = \int_0^b k(r-s)\mathrm{d}s - bk(r-b).\end{aligned}$$

Since $k > 0$, and $\|k\|_1 = 1$ it is clear that $1 \geq \mathrm{D}p_k(r,b) \geq -bk(r-b)$, and thus

$$\left|\mathrm{D}p_k(r,b)\right|^2 \leq \max\left(1, b^2k(r-b)^2\right) \leq 1 + b^2k(r-b)^2.$$

Integrating with respect to F, one obtains:

$$\begin{aligned}\int_{\mathbb{R}} \left|\mathrm{D}p_k(r,b)\right|^2 \mathrm{d}F(b) &\leq 1 + \int_0^{\bar{b}} b^2k(r-b)^2 f(b)\mathrm{d}b \\ &\leq 1 + \|k\|_\infty \int_0^{\bar{b}} b^2 f(b)k(r-b)\mathrm{d}b. \qquad (6.7)\end{aligned}$$

By Assumption 6.1 $bf(b) = 1 - F(b) - \mathrm{D}\Psi^F(b)$, for all $b \in (0, \bar{b})$, which implies

$$bf(b) \leq 1 + \|\mathrm{D}\Psi^F\|_\infty < \infty. \qquad (6.8)$$

Combining (6.7) and (6.8) yields

$$\begin{aligned}\int_{\mathbb{R}} \left|\mathrm{D}p_k(r,b)\right|^2 \mathrm{d}F(b) &\leq 1 + \left(1 + \|\mathrm{D}\Psi^F\|_\infty\right)\|k\|_\infty \int_0^{\bar{b}} bk(r-b)\mathrm{d}b \\ &\leq 1 + \bar{b}\left(1 + \|\mathrm{D}\Psi^F\|_\infty\right)\|k\|_\infty .\end{aligned}$$

□

Lemma 6.3.4.
Let Assumption 6.1 hold and $k \in \mathcal{K}$, then

$$|\Psi^F(r) - \Psi_k^F(r)| \leq \|\mathrm{D}\Psi^F\|_\infty \|K - 1_{\mathbb{R}^+}\|_1 \qquad (6.9)$$

$$\left|\mathrm{D}\Psi^F(r) - \mathrm{D}\Psi_k^F(r)\right| \leq \|\mathrm{D}^2\Psi^F\|_\infty \|K - 1_{\mathbb{R}^+}\|_1 , \qquad (6.10)$$

for any $r \in [0, \bar{b}]$. *Moreover,*

$$\|K - 1_{\mathbb{R}^+}\|_1 \leq \int_{\mathbb{R}} |r|\, k(r)\mathrm{d}r \tag{6.11}$$

Proof. First, note that $\mathrm{D}\Psi^F$ is bounded on $[0, \bar{b}]$, since it is continuous on this closed interval. Thus, integrating by parts leads to

$$\begin{aligned} \left|\Psi^F(r) - \Psi^F_k(r)\right| &= \left|[\Psi^F \star (\delta_0 - k)](r)\right| \\ &\leq \left|\Big[\Psi^F(t)[1_{\mathbb{R}^+} - K](r-t)\Big]_{-\infty}^{+\infty} - [\mathrm{D}\Psi^F \star (1_{\mathbb{R}^+} - K)](r)\right| \end{aligned} \tag{6.12}$$

The first term equals 0 since $K(s) \to 1$ as $s \to +\infty$. Thus, by an application of Young's convolution inequality, we obtain

$$\left|\Psi^F(r) - \Psi^F_k(r)\right| \leq \|\mathrm{D}\Psi^F\|_\infty \|K - 1_{\mathbb{R}^+}\|_1$$

This proves (6.9), and by the same argument we obtain (6.10). Finally, integrating by parts we obtain

$$\begin{aligned} \|K - 1_{\mathbb{R}^+}\|_1 &= \int_{-\infty}^0 K(r)\mathrm{d}r + \int_0^\infty [1 - K(r)]\mathrm{d}r \\ &= \Big[rK(r)\Big]_{-\infty}^0 - \int_{-\infty}^0 rk(r)\mathrm{d}r + \Big[r(1-K(r))\Big]_0^\infty + \int_0^\infty rk(r)\mathrm{d}r \\ &= \int_{\mathbb{R}} |r|\, k(r)\mathrm{d}r. \end{aligned}$$

which completes the proof with (6.11). □

***Remark* 6.3.1.** *If one chooses a family of kernels, these bounds can be expressed in terms of its parameters. For instance, when* k *is zero-mean Gaussian with variance* σ^2*, one easily recovers:*

$$\|K - 1_{\mathbb{R}^+}\|_1 = \sigma\sqrt{2/\pi} \quad \|k\|_\infty = \left(\sqrt{2\pi}\sigma\right)^{-1}. \tag{6.13}$$

Conv-OGA converges only to r_k^*. To remedy this, we would like to decrease B_k over time by letting[5] $k \to \delta_0$. However, since $\|k\|_\infty \to +\infty$ as $k \to \delta_0$, we will have to tread carefully in our analysis which occupies the next section.

[5] In the L^1 sense ($\|K - 1_{\mathbb{R}_+}\| \to 0$), which in view of Remark 6.3.1 can also be done parametrically. For instance in the Gaussian case of Remark 6.3.1 by sending $\sigma \to 0$.

6.4 Convergence with a Stationary Bidder

To decrease the bias B_k over time, we introducing a decaying kernel sequence $(k_n)_{n\in\mathbb{N}}$ into Conv-OGA, giving V-Conv-OGA (Algorithm 3). This section will demonstrate its consistency and convergence by controlling the trade-off between bias B_k and variance V_k, as B_k is reduced to zero over time. This trade-off decomposes the total error as:

$$\Psi^F(r^*) - \Psi^F(r_n) = \underbrace{\Psi^F(r^*) - \Psi^F(r_k^*)}_{\text{(surrogate bias)}} + \underbrace{\Psi^F(r_k^*) - \Psi^F(r_n)}_{\text{(estimation)}}.$$

This decomposition highlights that the kernel should converge to δ_0 *quickly enough* to cancel the bias B_k, yet *slowly enough* to control V_k and preserve the convergence speed of Online Gradient Ascent.

Algorithm 3 V-Conv-OGA

Input: r_0, $(\gamma_n)_{n\in\mathbb{N}}$, $(k_n)_{n\in\mathbb{N}} \subset \mathcal{K}$, $\mathcal{I} \subset [0, \bar{b}]$.
for $n \in \mathbb{N}$ **do**
  Observe b_n,
  $r_n \leftarrow \Pi_{\mathcal{I}}\big(r_{n-1} + \gamma_n \mathrm{D}p_{k_n}(r_{n-1}, b_n)\big)$.
end for

6.4.1 General convergence result

Theorem 6.4.1 provides sufficient conditions on the schedules of $(k_n)_{n\in\mathbb{N}^*}$ and $(\gamma_n)_{n\in\mathbb{N}^*}$ that guarantees V-Conv-OGA converges almost surely to r^*. It is derived by adapting stochastic optimisation results, see e.g. [32], to the changing objective $\Psi^F_{k_n}$.

Theorem 6.4.1.
Let Assumptions 6.1 and 6.2 hold and $(k_n)_{n\in\mathbb{N}} \subset \mathcal{K}$. Then, by running V-Conv-OGA with $\mathcal{I} = [0, \bar{b}]$, we have

$$r_n \overset{a.s.}{\to} r^* \text{ as } n \to \infty,$$

for any family $(\gamma_n)_{n\in\mathbb{N}}$ such that $\sum_{n=1}^{+\infty} \gamma_n = +\infty$, $\sum_{n=1}^{+\infty} \gamma_n \|K_n - 1_{\mathbb{R}^+}\|_1 < +\infty$, and $\sum_{n=1}^{+\infty} \gamma_n^2 \|k_n\|_\infty < +\infty$.

Proof. The proof inherits a lot from classical methods, see e.g. [32]. The main difference lies in the type of concavity required. The proof in [32] is derived

for variationally coherent functions, i.e. functions $(\Psi^F_{k_n})_{n\in\mathbb{N}}$ such that

$$\sup_{|r-r^*_{k_n}|>\epsilon} \left(r - r^*_{k_n}\right) \mathrm{D}\Psi^F_{k_n}(r) < 0, \text{ for any } r \in \mathbb{R},\ n \in \mathbb{N}, \text{ and } \epsilon > 0.$$

However, this assumption is clearly violated here since $\mathrm{D}\Psi^F_{k_n}(r) \to 0$ as $r \to \infty$. Nevertheless, since $\Psi^F_{k_n}$ is strictly pseudo-concave and strictly positive, one can obtain a similar result.

Following [32], we introduce the Lyapunov process $v_n = \|r_n - r^*\|^2$. Using the fact that the projection operator $\Pi_{\mathcal{I}}$ is a contraction, one obtains

$$\begin{aligned} v_{n+1} &= \left(\Pi_{\mathcal{I}}\left(r_n + \gamma_n \mathrm{D}p_{k_n}(r_n, b_n)\right) - r_*\right)^2 \\ &\leq \left(r_n + \gamma_n \mathrm{D}p_{k_n}(r_n, b_n) - r^*\right)^2 \\ &\leq v_n + 2\gamma_n (r_n - r^*)\mathrm{D}p_{k_n}(r_n, b_n) + \gamma_n^2 \left(\mathrm{D}p_{k_n}(r_n, b_n)\right)^2. \end{aligned}$$

Hence, v_n satisfies the recursion:

$$v_{n+1} - v_n \leq 2\gamma_n (r_n - r^*)\mathrm{D}p_{k_n}(r_n, b_n) + \gamma_n^2 \left(\mathrm{D}p_{k_n}(r_n, b_n)\right)^2$$

Taking the conditional expectation with respect to $\mathcal{F}_n$ yields

$$\begin{aligned} &\mathbb{E}\left[v_{n+1} - v_n | \mathcal{F}_n\right] \\ &\quad \leq 2\gamma_n (r_n - r^*)\mathbb{E}\left[\mathrm{D}p_{k_n}(r_n, b_n)|\mathcal{F}_n\right] + \gamma_n^2 \mathbb{E}\left[\left(\mathrm{D}p_{k_n}(r_n, b_n)\right)^2 | \mathcal{F}_n\right] \\ &\quad \leq 2\gamma_n (r_n - r^*)\mathrm{D}\Psi^F_{k_n}(r_n) + \gamma_n^2 \mathbb{E}\left[\left(\mathrm{D}p_{k_n}(r_n, b_n)\right)^2 | \mathcal{F}_n\right] \end{aligned} \tag{6.14}$$

as Proposition 6.3.2 implies that $\mathbb{E}[\mathrm{D}p_{k_n}(r_n, b_n)|\mathcal{F}_n] = \mathrm{D}\Psi^F_{k_n}(r_n)$.

We decompose the first term to isolate the gradient bias, obtaining

$$\begin{aligned} \mathbb{E}\left[v_{n+1} - v_n | \mathcal{F}_n\right] \leq\ & 2\gamma_n (r_n - r^*)\mathrm{D}\Psi^F(r_n) && (U_1) \\ &+ 2\gamma_n (r_n - r^*)(\mathrm{D}\Psi^F_{k_n}(r_n) - \mathrm{D}\Psi^F(r_n)) && (U_2) \\ &+ \gamma_n^2 \mathbb{E}\left[\left(\mathrm{D}p_{k_n}(r_n, b_n)\right)^2 | \mathcal{F}_n\right]. && (U_3) \end{aligned}$$

The first term, that is (U_1), is negative by the pseudo-concavity of Ψ^F, see Proposition 6.3.1. The second term, i.e. (U_2) is the gradient bias term, and it is bounded by $2\gamma_n \|\mathrm{D}^2\Psi^F\|_\infty \|K - 1_{\mathbb{R}^+}\|_1$ by (6.10). Finally, (U_3) is bounded by $\gamma_n^2 \left(1 + \bar{b}(1 + \|\mathrm{D}\Psi^F\|_\infty)\|k\|_\infty\right)$ by Proposition 6.3.3.

Using the same quasi-martingale argument as in [32], we have that (U_1)–(U_3), together with $\sum_{n=1}^\infty \gamma_n \|K_n - 1_{\mathbb{R}^+}\|_1 < \infty$ and $\sum_{n=1}^\infty \gamma_n^2 \|k_n\|_\infty < \infty$, imply that

$(v_n)_{n\in\mathbb{N}}$ converges, as $n \to \infty$, almost surely to some finite limit h_∞. Moreover, we have that

$$\sum_{n=1}^{\infty} \mathbb{E}\left[v_{n+1} - v_n | \mathcal{F}_n\right] < \infty.$$

Thus, using (6.14), we have that

$$0 \leq \sum_{n=1}^{\infty} \gamma_n (r^* - r_n) \mathrm{D}\Psi^F(r_n) < \infty. \tag{6.15}$$

It remains to prove that $h_\infty = 0$. Suppose for a contradiction that $h_\infty > 0$. In this case, for $\epsilon > 0$ there is a time n_ϵ such that for any $n \geq n_\epsilon$, $(r^* - r_n)\mathrm{D}\Psi^F(r_n) \geq \epsilon > 0$. This implies that

$$\sum_{n=1}^{\infty} \gamma_n (r^* - r_n) \mathrm{D}\Psi^F(r_n) = +\infty,$$

which is in contradiction with (6.15). The same argument applies to the case $h_\infty < 0$, and thus we have that $v_n \to 0$ almost surely as $n \to +\infty$ and $\Psi^F(r_n) \to 0$ almost surely as $n \to \infty$. □

If a constant kernel sequence were to be used in V-Conv-OGA, we would recover the usual stochastic approximation conditions on the step size γ_n, namely that $\sum_{n=1}^{\infty} \gamma_n = +\infty$ and $\sum_{n=1}^{\infty} \gamma_n^2 < +\infty$. This suggests setting $\gamma_n \propto 1/n$. For such a choice of step-size, Theorem 6.4.1 asserts convergence if $\sum_{n=1}^{\infty} \gamma_n \|K_n - 1_{\mathbb{R}^+}\|_1 < +\infty$, which is guaranteed by $k_n \to \delta_0$. This means $\|k_n\|_\infty \to \infty$ as $n \to \infty$, but $\sum_{n=1}^{\infty} \gamma_n^2 \|k_n\|_\infty < +\infty$ tells us explicitly how slow our decay must be in terms of the family of kernels. For example in the case of a Gaussian kernel, (6.13) implies that a suitable choice of kernel variance is $\sigma_n \propto n^{-\alpha}$ for $\alpha \in (0, 1)$.

6.4.2 Finite-time convergence rates

While Theorem 6.4.1 provides sufficient conditions on the kernel sequence $(k_n)_{n\in\mathbb{N}}$ for V-Conv-OGA to be consistent, it does not characterise the rate of the convergence, and thus cannot be leveraged to optimise the step size γ_n and the decay rate of the kernel.

To obtain finite time guarantees on the rate of convergence, we must impose stronger conditions on the monopoly revenue Ψ^F. Recall that under Assumption 6.2, Ψ^F is strictly pseudo-concave. It is well known that such functions can have large areas of arbitrarily small gradients. Since these

can make first-order methods arbitrarily slow, no meaningful rate can be obtained for them. Strengthening the assumption to Assumption 6.3, i.e. excluding vanishing gradients by ensuring Ψ^F is μ_h-strongly log-concave (see Proposition 6.3.1), will give a rate in Theorem 6.4.2 under the further technical assumption Assumption 6.4.

Assumption 6.4.
The seller is given a compact subset $\mathcal{I} \subseteq [0, \bar{b}]$ and a constant $\underline{\Psi}^F > 0$ such that $r^* \in \mathcal{I}$ and for all $r \in \mathcal{I}$, $\Psi^F(r) \geq \underline{\Psi}^F$.

Assumption 6.4 ensures that the seller can lower bound revenue on a compact subset of $[0, \bar{b}]$. It should be understood as prior knowledge of the seller based on the format of the auction and the type of item sold. The interval $\mathcal{I}$ exists for any $\underline{\Psi}^F < \Psi^F(r^*)$, so this hypothesis is not restrictive relative to Assumption 6.3.

Theorem 6.4.2.
Let Assumptions 6.1, 6.3 and 6.4 hold. Let $(k_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{K}$ be such that $\|K_n - 1_{\mathbb{R}^+}\|_1 \leq v_1 n^{-\alpha_1}$ and $\|k_n\|_\infty \leq v_\infty n^{\alpha_\infty}$ with α, α_1, α_∞ satisfying the condition of Theorem 6.4.1, i.e. $\alpha \leq 1$, $\alpha + \alpha_1 > 1$, and $2\alpha - \alpha_\infty > 1$.

Running V-Conv-OGA on $\mathcal{I}$ with $\gamma_n := v_\gamma n^{-\alpha}$ for $n \in \mathbb{N}^*$, in which $v_\gamma \leq 1/\Gamma$ for $\Gamma := 2\underline{\Psi}^F \mu_h$, yields

$$\mathbb{E}\left[\|r_n - r^*\|^2\right] \leq \left(\bar{b}^2 + 2Cv_\gamma v_1 \varphi_{\Gamma v_\gamma - \alpha_1}(n) + 2C_\infty v_\gamma^2 v_\infty \varphi_{\Gamma v_\gamma + \alpha_\infty - 1}(n)\right) n^{-\Gamma v_\gamma}$$

for all $n \geq 2$, when $\alpha = 1$. If $\alpha \in (0, 1)$, on the other hand, it yields

$$\begin{aligned}
&\mathbb{E}\left[\|r_n - r^*\|^2\right] \\
&\quad \leq \left(\bar{b}^2 + C_1 v_\gamma v_1 \varphi_{1-\alpha-\alpha_1}(n) + C_\infty v_\gamma^2 v_\infty \varphi_{1+\alpha_\infty - 2\alpha}(n)\right) \exp\left(-\frac{2v_\gamma}{\Gamma} n^{1-\alpha}\right) \\
&\qquad + C_1 \frac{2v_1}{\Gamma} n^{-\alpha_1} + C_\infty \frac{2v_\gamma v_\infty}{\Gamma} n^{\alpha_\infty - \alpha}
\end{aligned}$$

for all $n \geq 2$.

The function φ and the constants C_1, C_∞ are given by

$$\varphi_\beta(n) := \log(n) 1_{\{\beta = 0\}} + \frac{n^\beta - 1}{\beta} 1_{\{\beta \neq 0\}},$$

$$C_1 := 2\bar{b}\|\mathrm{D}^2\Psi^F\|_\infty, \text{ and } C_\infty := 1 + \bar{b}\left(1 + \|\mathrm{D}\Psi^F\|_\infty\right),$$

for any $(\beta, n) \in \mathbb{R}_+ \times \mathbb{N}^*$.

Proof. The proof builds on [91, Thm. 2]. There are three main differences: first, we do not require the local function p_{k_n} to be concave; nor do we rely on the strong concavity of $\Psi^F_{k_n}$ but instead on its strong log-concavity and its lower bound $\underline{\Psi}^F$; finally, our objective function varies over time because of the sequence of convolution kernels $(k_n)_{t\geq 1}$.

Assumption 6.3 together with the lower bounded revenue $\Psi^F_{k_n}$ leads to a form of local strong concavity of Ψ^F. From Proposition 6.3.1, the strongly increasing hazard rate ensures that Ψ^F is strongly log-concave with parameter μ_h. Further, Ψ^F admits a unique maximum r^* such that $r^* \in \mathcal{I}$ by assumption. As a result, for any $r \in \mathcal{I}$,

$$\begin{aligned}-\mu_h \|r^* - r\|^2 &\geq (r - r^*)\left(\mathrm{D}\left[\log \Psi^F\right](r) - \mathrm{D}\left[\log \Psi^F\right](r^*)\right)\\ &= (r - r^*)\left(\frac{\mathrm{D}\Psi^F(r)}{\Psi^F(r)} - \frac{\mathrm{D}\Psi^F(r^*)}{\Psi^F(r^*)}\right).\end{aligned}$$

By the first-order optimality condition of Ψ^F, this implies that

$$(r - r^*)\mathrm{D}\Psi^F(r) \leq -\Psi^F(r)\mu_h \|r^* - r\|^2 \leq -\mu_h \underline{\Psi}^F \|r^* - r\|^2.$$

Let us introduce $\tilde{\mu}_h := \Gamma/2 = \mu_h \underline{\Psi}^F$ the quantity which plays the role of the strong-concavity parameter in [91, Thm. 2].

As in the proof of Theorem 6.4.1, we introduce the Lyapunov process $v_n := \|r_n - r^*\|_2^2$ and its expectation $\bar{v}_n := \mathbb{E}[v_n]$. From (U_1)–(U_3), we have

$$\begin{aligned}\mathbb{E}[v_{n+1} - v_n|\mathcal{F}_n] \leq 2\gamma_n(r_n - r^*)\mathrm{D}\Psi^F(r_n) + 2\gamma_n\bar{b}\|\mathrm{D}^2\Psi^F\|_\infty \|K_n - 1_{\mathbb{R}^+}\|_1\\ + \gamma_n^2\left(1 + \bar{b}\left(1 + \|\mathrm{D}\Psi^F\|_\infty\right)\|k\|_\infty\right).\end{aligned}$$

Combining the local strong-concavity of Ψ^F and the kernel conditions[6] of the statement ($\|K_n - 1_{\mathbb{R}^+}\|_1 \leq \nu_1 n^{-\alpha_1}$, $\|k_n\|_\infty \leq \nu_\infty n^{\alpha_\infty}$), yields

$$\mathbb{E}[v_{n+1} - v_n|\mathcal{F}_n] \leq 2\tilde{\mu}_h\gamma_n v_n + C_1\gamma_n\nu_1 n^{-\alpha_1} + C_\infty\gamma_n^2\nu_\infty n^{\alpha_\infty},$$

in which $C_1 := 2\bar{b}\|\mathrm{D}^2\Psi^F\|_\infty$ and $C_\infty := 1 + \bar{b}\left(1 + \|\mathrm{D}\Psi^F\|_\infty\right)$.Taking the expectation leads to

$$\bar{v}_{n+1} \leq (1 - 2\tilde{\mu}_h\gamma_n)\bar{v}_n + C_1\gamma_n\nu_1 n^{-\alpha_1} + C_\infty\gamma_n^2\nu_\infty n^{\alpha_\infty}. \tag{6.16}$$

In line with [91], we split the proof depending whether $\alpha = 1$ or $\alpha \in (0, 1)$.

[6]Without loss of generality, we assume that $\nu_\infty \geq 1$.

1. Let us consider the case in which $\alpha = 1$. Using the identity $1 - x \leq \exp(-x)$, for all $x \in \mathbb{R}$, and applying the recursion n times in (6.16), we have

$$\begin{aligned}
\bar{v}_n &\leq \bar{v}_1 \exp\left(-2\tilde{\mu}_h \sum_{i=1}^{n-1} \gamma_i\right) + C_1 \nu_1 \sum_{i=1}^{n-1} \gamma_i i^{-\alpha_1} \exp\left(-2\tilde{\mu}_h \sum_{j=i+1}^{n-1} \gamma_j\right) \\
&\quad + C_\infty \nu_\infty \sum_{i=1}^{n-1} \gamma_i^2 i^{\alpha_\infty} \exp\left(-2\tilde{\mu}_h \sum_{j=i+1}^{n-1} \gamma_j\right) \\
&\leq \left(C_1 \nu_\gamma \nu_1 \sum_{i=1}^{n-1} i^{-\alpha-\alpha_1} + C_\infty \nu_\gamma^2 \nu_\infty \sum_{i=1}^{n-1} i^{-2\alpha+\alpha_0}\right) \exp\left(-2\tilde{\mu}_h \nu_\gamma \sum_{j=i+1}^{n-1} j^{-1}\right) \\
&\quad + \bar{v}_1 \exp\left(-2\tilde{\mu}_h \nu_\gamma \sum_{i=1}^{n-1} i^{-1}\right).
\end{aligned}$$

Since,

$$\begin{aligned}
\sum_{i=1}^{n-1} i^{-1} &\geq \log(n) \\
\sum_{j=\ell+1}^{n-1} j^{-1} &\geq \log(n/\ell + 1),
\end{aligned}$$

for all $n \geq 2$ and $1 \leq \ell \leq n-1$, we obtain that (assuming $\tilde{\mu}_h \nu_\gamma \leq 1/2$):

$$\begin{aligned}
\bar{v}_n &\leq \bar{v}_1 n^{-2\tilde{\mu}_h \nu_\gamma} + C_1 \nu_\gamma \nu_1 n^{-2\tilde{\mu}_h \nu_\gamma} \sum_{i=1}^{n-1} i^{-1-\alpha_1} (i+1)^{2\tilde{\mu}_h \nu_\gamma} \\
&\quad + C_\infty \nu_\gamma^2 \nu_\infty n^{-2\tilde{\mu}_h \nu_\gamma} \sum_{i=1}^{n-1} i^{-2+\alpha_\infty} (i+1)^{2\tilde{\mu}_h \nu_\gamma} \\
&\leq \bar{v}_1 n^{-2\tilde{\mu}_h \nu_\gamma} + 2C_1 \nu_\gamma \nu_1 n^{-2\tilde{\mu}_h \nu_\gamma} \sum_{i=1}^{n-1} i^{2\tilde{\mu}_h \nu_\gamma - 1 - \alpha_1} \\
&\quad + 2C_\infty \nu_\gamma^2 \nu_\infty n^{-2\tilde{\mu}_h \nu_\gamma} \sum_{i=1}^{n-1} i^{-2+\alpha_\infty + 2\tilde{\mu}_h \nu_\gamma} \\
&\leq \left(\bar{b}^2 + 2C_1 \nu_\gamma \nu_1 \varphi_{2\tilde{\mu}_h \nu_\gamma - \alpha_1}(n) + 2C_\infty \nu_\gamma^2 \nu_\infty \varphi_{2\tilde{\mu}_h \nu_\gamma + \alpha_\infty - 1}\right) n^{-2\tilde{\mu}_h \nu_\gamma}.
\end{aligned}$$

2. Consider now the case in which $\alpha \in (0,1)$. By applying the recursion of (6.16) n times, we have

$$\bar{v}_n \leq \bar{v}_1 \prod_{i=1}^{n-1}(1-2\tilde{\mu}_h\gamma_i) \tag{A_n^1}$$

$$+ C_1 \sum_{i=1}^{n-1} \gamma_i v_1 n^{-\alpha_1} \prod_{j=i+1}^{n-1}(1-2\tilde{\mu}_h\gamma_j) \tag{A_n^2}$$

$$+ C_\infty \sum_{i=1}^{n-1} \gamma_i^2 v_\infty n^{\alpha_\infty} \prod_{j=i+1}^{n-1}(1-2\tilde{\mu}_h\gamma_j) \tag{A_n^3}$$

The derivation slightly differs from the case $\alpha = 1$. Following [91], one has:

$$A_n^1 \leq \bar{v}_1 \exp\left(-2\tilde{\mu}_h \sum_{i=1}^{n-1}\gamma_i\right)$$

$$A_n^2 \leq \frac{C_1 v_1 \lfloor n/2 \rfloor^{\alpha_1}}{2\tilde{\mu}_h} + C_1 v_1 \exp\left(-2\tilde{\mu}_h \sum_{j=\lfloor n/2 \rfloor}^{n-1}\gamma\right)\sum_{i=1}^{n-1}\gamma_i n^{-\alpha_1}$$

$$A_n^3 \leq C_\infty v_\infty \frac{\gamma_{\lfloor n/2 \rfloor}\lfloor n/2 \rfloor^{\alpha_\infty}}{2\tilde{\mu}_h} + C_\infty v_\infty \exp\left(-2\tilde{\mu}_h \sum_{j=\lfloor t/2 \rfloor}^{n-1}\gamma_j\right)\sum_{i=1}^{n-1}\gamma_i^2 n^{\alpha_\infty}.$$

Using the expression of γ_n, and the fact that $\alpha_1 < 1$, together with the identities $\varphi_{1-\alpha}(t) - \varphi_{1-\alpha}(t/2) \geq t^{1-\alpha}/2$ and $\varphi_{1-\alpha}(t) \geq t^{1-\alpha}/2$ for $t \geq 1$, yields

$$\begin{aligned} A_n^1 &\leq \bar{v}_1 \exp\left(-2\tilde{\mu}_h v_\gamma \varphi_{1-\alpha}(n)\right)\bar{v}_1 \leq \exp\left(-\tilde{\mu}_h v_\gamma n^{1-\alpha}\right), \\ A_n^2 &\leq \frac{C_1 v_1}{\tilde{\mu}_h} n^{-\alpha_1} + C_1 v_\gamma v_1 \exp\left(-\tilde{\mu}_h v_\gamma n^{1-\alpha}\right)\sum_{i=1}^{n-1} i^{-\alpha-\alpha_1} \\ &\leq \frac{C_1 v_1}{\tilde{\mu}_h} n^{-\alpha_1} + C_1 v_\gamma v_1 \exp\left(-\tilde{\mu}_h v_\gamma n^{1-\alpha}\right)\varphi_{1-\alpha-\alpha_1}(n), \end{aligned}$$

and

$$\begin{aligned} A_n^3 &\leq C_\infty \frac{v_\gamma v_\infty}{\tilde{\mu}_h} n^{\alpha_\infty-\alpha} + C_\infty v_\gamma^2 v_\infty \exp\left(-\tilde{\mu}_h v_\gamma n^{1-\alpha}\right)\sum_{i=1}^{n-1} i^{\alpha_\infty-2\alpha} \\ &\leq C_\infty \frac{v_\gamma v_\infty}{\tilde{\mu}_h} n^{\alpha_\infty-\alpha} + C_\infty v_\gamma^2 v_\infty \exp\left(-\tilde{\mu}_h v_\gamma n^{1-\alpha}\right)\varphi_{1+\alpha_\infty-2\alpha}(n). \end{aligned}$$

Putting everything together, the final bound we obtain is

$$\bar{v}_n \leq \left(\bar{b}^2 + C_1 v_\gamma v_1 \varphi_{1-\alpha-\alpha_1}(n) + C_\infty v_\gamma^2 v_\infty \varphi_{1+\alpha_\infty-2\alpha}(n)\right) \exp\left(-\tilde{\mu}_h v_\gamma n^{1-\alpha}\right) + \frac{C_1 v_1}{\tilde{\mu}_h} n^{-\alpha_1} + C_\infty \frac{v_\gamma v_\infty}{\tilde{\mu}_h} n^{\alpha_\infty - \alpha}.$$

□

Theorem 6.4.2 shows the existence of two distinct regimes which are shared by either choice of α: a transient regime (whose rate is $n^{-\Gamma v_\gamma}$ if $\alpha = 1$ and $\exp(-\mu_h \underline{\Psi}^F v_\gamma n^{1-\alpha})$ if $\alpha \in (0,1)$), and a stationary regime ($n^{-\alpha_1} + n^{\alpha_\infty - \alpha}$ and $n^{-\alpha_1} + n^{\alpha_\infty - 1}$, respectively).

On Fig. 6.2, the transient phase is visible up to 2×10^3 steps. Since the rate of the transient regime depends only on $v_\gamma = \gamma_0$, $\underline{\Psi}^F$ known from Assumption 6.4, and μ_h known from Assumption 6.3, we can set v_γ to make the *stationary* regime the driver of the rate.

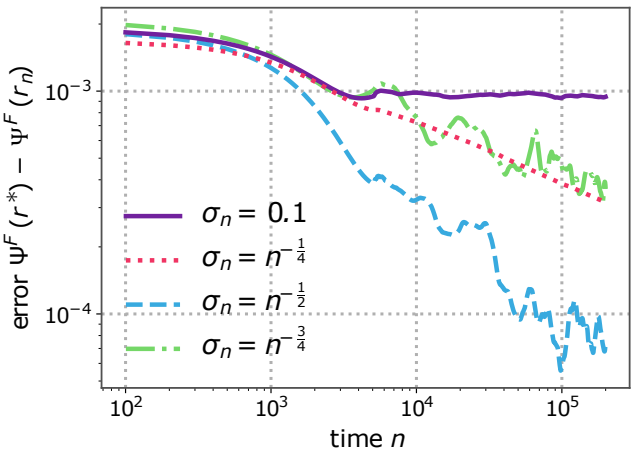

Figure 6.2: Expected error of V-Conv-OGA for different schedules of $(\sigma_n)_{n\in\mathbb{N}}$ on i.i.d samples from a Kumaraswamy (1, 0.4) distribution (log-log scale). Bottom: representative reserve price trajectories.

To optimise the stationary regime we face a bias-variance trade-off. Like Theorem 6.4.1, Theorem 6.4.2 requires that $k \to \delta_0$ (via the condition $\alpha_1 > 0$) while imposing a bound on the growth speed of the variance bound V_k (via $\alpha_\infty < \alpha$). Unlike Theorem 6.4.1, however, the detailed rates of Theorem 6.4.2 can be used to determine optimal parameters for the trade-off, taking into account the antagonistic effects of α_1 and α_∞.

From Theorem 6.4.2, we recover that the optimal learning rate is $\gamma_n \propto 1/n$. To tune the kernels it is sensible to fix a parametric family and tune its

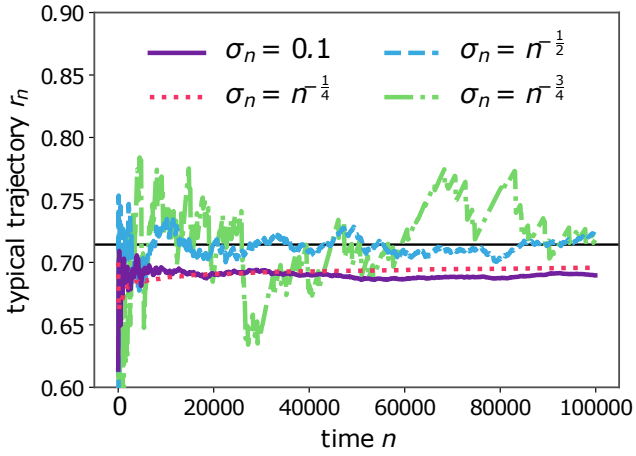

Figure 6.3: Representative sample trajectories of V-Conv-OGA for different schedules of $(\sigma_n)_{n\in\mathbb{N}}$ ($\gamma_n := n^{-1}$ for $n \in \mathbb{N}^*$) on i.i.d. samples from a Kumaraswamy (1, 0.4) distribution.

parameters. For zero-mean Gaussian kernels (recall Remark 6.3.1), we have Corollary 6.4.1.

***Corollary* 6.4.1.** *If we fix $\gamma_n \propto 1/n$, and let $(k_n)_{n\in\mathbb{N}}$ be Gaussian $(0, 1/n)$ kernels in V-Conv-OGA, Theorem 6.4.2 becomes*

$$\mathbb{E}\left[\|r_n - r^*\|^2\right] = \tilde{\mathcal{O}}\left(n^{-1/2}\right).$$

This rate is optimal up to logarithmic factors.

Figure 6.3 demonstrates the optimality in the bias-variance trade-off of this choice: the $\sigma_n := n^{-1/2}$ (blue) curve is the optimal rate on the top pane, and attains the rate of Corollary 6.4.1. The bottom pane illustrates the bias-variance trade-off at hand in Theorem 6.4.2. If the kernel decays slower than $n^{-1/2}$ (red), the learning rate shrinks much faster and convergence is very slow but very smooth. If σ_n decreases too fast (green) the variance becomes overwhelming and noise swallows the performance.

Our novel analysis of V-Conv-OGA showed its almost sure convergence under Assumption 6.2, and that with the curvature condition of Assumption 6.3 and the technical Assumption 6.4 we could fully characterise its convergence rates. We could thus derive optimal learning rates and place conditions on optimal kernel decay rates. We made the optimal decay rate explicit for Gaussian kernels. This concludes the discussion of V-Conv-OGA in the stationary bid distribution problem, and we now move to the nonstationary setting.

6.5 Tracking a Nonstationary Bidder

In practical applications of online auctions, such as display advertising, bidders might change their bid distribution over time. These changes often result from non-stationarity in the private information of bidders. It is therefore beneficial to be able to effectively adapt one's reserve price over time to track changing bid distributions $(F_n)_{n\in\mathbb{N}}$. We use the dynamic regret $\mathcal{R}_N$ to measure the quality of an algorithm's tracking.

The difficulty in the nonstationary setting is to trade-off adaptability (how fast a change in the bid distribution is detected) *vs*. accuracy (proximity to the monopoly price between switches). Convergent algorithms like ERM or V-Conv-OGA will have high accuracy in the first phase, but then suffer as they try to adapt to changes later on when their learning rate is very small. Windowed methods are more adaptable but still carry with them a lag, directly dependent on their window size. First-order methods like Conv-OGA (with constant learning rate γ_0) are much more adaptable, but their convergence rate ($\mathcal{O}(n^{-1/2})$) hurts their accuracy. Nevertheless, we show that Conv-OGA is effective, with $\mathcal{O}(\sqrt{N})$ regret after N steps.

The dynamic regret $\mathcal{R}_N$ cannot be meaningfully controlled for arbitrary sequences $(F_n)_{n\in\mathbb{N}}$. As such it is customary to assume Assumption 6.5 that $(F_n)_{n\in\mathbb{N}}$ contains at most $\kappa_N - 1$ switches up to a horizon $N \in \mathbb{N}^*$, see e.g. [60, 82]. This corresponds to approximating a slowly changing sequence of distributions F_n by a piece-wise constant sequence.

Assumption 6.5.
Given some horizon N, there exists $\kappa_N \leq N$ such that $\sum_{t=1}^{N-1} 1_{\{F_n \neq F_{t+1}\}} \leq \kappa_N - 1$.

Under Assumption 6.5, the game (up to time N) decomposes into κ_N phases. The first step towards controlling the regret is to bound the tracking performance in each phase. We do this in Theorem 6.5.1, which shows an incompressible asymptotic error (the bias of our surrogate plus the variance) and a transient regime with exponential decay.

Theorem 6.5.1.
Let Assumptions 6.1, 6.3 and 6.4 and $k \in \mathcal{K}$. Then, by running Conv-OGA on $\mathcal{I}$ with a constant step-size $\gamma_0 > 0$, we have

$$\mathbb{E}\left[\|r_n - r_k^*\|^2\right] \leq \left(\bar{b}^2 + C(\gamma_0, k)(n-1)\right)e^{-\frac{\mu_h \Psi^F \gamma_0}{2} n} + \frac{2C(\gamma_0, k)}{\mu_h \Psi^F}$$

for in $n \in \mathbb{N}^*$, in which $C(\gamma_0, k) = \mathcal{O}\left(\gamma_0 \|K - 1_{\mathbb{R}^+}\|_1 + \gamma_0^2 \|k\|_\infty\right)$.

Proof. Similarly to the one of Theorem 6.4.2, this proof builds on [91]. Since Ψ^F is μ_h-strongly log-concave, one has for all $r \in \mathcal{I}$,

$$(r_n - r^*)\mathrm{D}\Psi^F(r) \leq -\Psi^F(r)\mu_h \|r - r^*\|^2 \leq -\tilde{\mu}_h \|r - r^*\|^2 \tag{6.17}$$

in which $\tilde{\mu}_h := \mu_h \Psi^F$. As a result, although we do not assume the function to be strongly concave, it still enjoys a similar property near r^* on the bounded subset $\mathcal{I}$. Let $v_n = \|r_n - r^*\|^2$ be the same Lyapunov process as in the proof of Theorem 6.4.1. Since the projection operator over $\mathcal{I}$ is $1-$Lipschitz and from (U_1)—(U_3), one has

$$\begin{aligned}\mathbb{E}\left[v_{n+1} - v_n | \mathcal{F}_n\right] \leq{}& 2\gamma_0 (r_n - r^*)\mathrm{D}\Psi^F(r_n) + 2\gamma_0 \bar{b} \|\mathrm{D}^2\Psi^F\|_\infty \|K - 1_{\mathbb{R}^+}\|_1 \\ &+ \gamma_n^2\left(1 + \bar{b}\left(1 + \|\mathrm{D}\Psi^F\|_\infty\right)\|k\|_\infty\right).\end{aligned} \tag{6.18}$$

Letting

$$C(\gamma_0, k) := 2\gamma_0 \bar{b} \|\mathrm{D}^2\Psi^F\|_\infty \|K - 1_{\mathbb{R}^+}\|_1 + \gamma_0^2\left(1 + \bar{b}\left(1 + \|\mathrm{D}\Psi^F\|_\infty\right)\|k\|_\infty\right)$$

and $\bar{v}_n := \mathbb{E}[v_n]$, and taking the expectation in (6.18), yields

$$\bar{v}_{n+1} \leq (1 - 2\gamma_n \tilde{\mu}_h)\bar{v}_n + C(\gamma_0, k). \tag{6.19}$$

Further, (6.19) is exactly the same as [91, (25)] with different definitions for the constants, and the rest of the proof follows. As a result, we have

$$\bar{v}_n \leq \left(\bar{v}_0 + C(\gamma_0, k)(n-1)\right)\exp\left(-\frac{\tilde{\mu}_h \gamma_0}{2} n\right) + \frac{2C(\gamma_0, k)}{\tilde{\mu}_h},$$

for any $n \in \mathbb{N}^*$. □

Theorem 6.5.1 shows that, immediately after a bid distribution switch, there will be a transient regime of order $n\exp(-\mu_h \Psi^F \gamma_0 n/2)$, but afterward r_n will oscillate in a band of size $2C(\gamma_0, k)/\mu_h \Psi^F$ around r_k^*. We can then use Theorem 6.5.1 to derive a sub-linear regret bound given N, κ_N in Corollary 6.5.1.

***Corollary* 6.5.1.** *Let Assumptions 6.1 and 6.3 to 6.5 hold and $k \in \mathcal{K}$. Then, there are positive functions $\Xi(k, \gamma_0)$ and $\Lambda(k, \gamma_0)$, independent of N and κ_N, such that* Conv-OGA *has a nonstationary regret of*

$$\mathcal{R}_N \leq \Xi(k, \gamma_0) N + \Lambda(k, \gamma_0) \kappa_N,$$

for all $N \geq 0$. Further, if the horizon N is known in advance, running Conv-OGA *with $\gamma_0 = N^{-1/2}$ and k a kernel satisfying $\|K - 1_{\mathbb{R}^+}\|_1 \leq N^{-1/2}$ and $\|k\|_\infty \leq N^{1/2}$, then $\mathcal{R}_N = \mathcal{O}\left(\sqrt{N}\right)$.*

Proof. Denoting $(s_i)_{i=1}^{\kappa_N}$ and $(t_i)_{i=1}^{\kappa_N}$ respectively the start and the end of the intervals in $\mathbb{N}^*$ on which the distribution is constant, we have

$$\begin{aligned}
\mathcal{R}_N &= \mathbb{E}\left[\sum_{n=1}^{N} \Psi^{F_n}(r_n^*) - \Psi^{F_n}(r_n)\right] \\
&\leq \mathbb{E}\left[\sum_{n=1}^{N} \frac{\|\mathrm{D}^2\Psi^{F_n}\|_\infty}{2} \|r_n^* - r_n\|_2^2\right] \\
&\leq \sum_{i=1}^{\kappa_N} \frac{\|\mathrm{D}^2\Psi^{F_{s_i}}\|_\infty}{2} \sum_{n=s_i}^{t_i} \mathbb{E}\left[\|r_{s_i}^* - r_n\|_2^2\right] \\
&\leq \sum_{i=1}^{\kappa_N} \frac{\|\mathrm{D}^2\Psi^{F_{s_i}}\|_\infty}{2} \sum_{n=1}^{t_i - s_i + 1} \mathbb{E}\left[\|r_{s_i}^* - r_{n+s_i-1}\|_2^2\right]
\end{aligned}$$

Let us define $\tilde{\Gamma} := \mu_h \underline{\Psi}^F \gamma_0 / 2$, $M_\Psi := \max_{i \in [\kappa_N]} \|\mathrm{D}^2\Psi^{F_{s_i}}\|_\infty / 2$, and

$$\bar{C}(\gamma_0, k) := 4 M_\Psi \|K - 1_{\mathbb{R}^+}\|_1 + \gamma_0^2 \left(1 + \bar{b}\,(1 + 2M_\Psi)\,\|k\|_\infty\right).$$

Applying Theorem 6.5.1 on $\mathbb{E}\left[\|r_{s_i}^* - r_{n+s_i-1}\|_2^2\right]$, we obtain

$$\begin{aligned}
\mathcal{R}_N &\leq \sum_{i=1}^{\kappa_N} \frac{\|\mathrm{D}^2\Psi^{F_{s_i}}\|_\infty}{2} \sum_{n=1}^{t_i - s_i + 1} \left(\left(\bar{b}^2 + \bar{C}(\gamma_0, k)(n-1)\right) e^{-\tilde{\Gamma} n} + \frac{2\bar{C}(\gamma_0, k)}{\mu_h \underline{\Psi}^F}\right) \\
&\leq M_\Psi \left(\frac{2\bar{C}(\gamma_0, k)}{\mu_h \underline{\Psi}^F} N + \sum_{i=1}^{\kappa_N} \sum_{n=1}^{t_i - s_i + 1} \bar{b}^2 e^{-\tilde{\Gamma} n} + \bar{C}(\gamma_0, k)(n-1) e^{-\tilde{\Gamma} n}\right) \\
&\leq M_\Psi \left(\frac{2\bar{C}(\gamma_0, k)}{\mu_h \underline{\Psi}^F} N + \sum_{i=1}^{\kappa_N} \sum_{n=0}^{t_i - s_i} \bar{b}^2 e^{-\tilde{\Gamma}(n+1)} + \bar{C}(\gamma_0, k) n e^{-\tilde{\Gamma}(n+1)}\right) \\
&\leq M_\Psi \left(\frac{2\bar{C}(\gamma_0, k)}{\mu_h \underline{\Psi}^F} N + \sum_{i=1}^{\kappa_N} \left(\frac{\bar{b}^2 e^{-\tilde{\Gamma}}}{1 - e^{-\tilde{\Gamma}}} + \sum_{n=0}^{t_i - s_i} \bar{C}(\gamma_0, k) n e^{-\tilde{\Gamma}(n+1)}\right)\right) \\
&\leq M_\Psi \left(\frac{2\bar{C}(\gamma_0, k)}{\mu_h \underline{\Psi}^F} N + \left(\frac{\bar{b}^2 \kappa_N}{e^{\tilde{\Gamma}} - 1} +\right.\right.
\end{aligned}$$

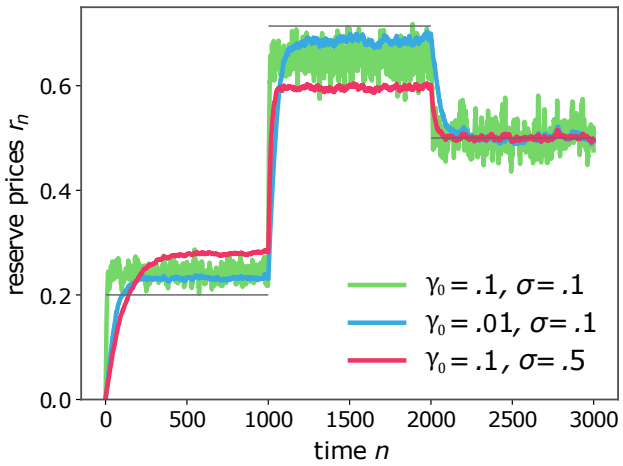

Figure 6.4: Tracking by Conv-OGA of three Kumaraswamy distributions (with parameters $(1,4)$, $(1,0.4)$, and $(1,1)$ resp.) with different Gaussian kernels and learning rates.

$$\bar{C}(\gamma_0,k)e^{-2\tilde{\Gamma}}\sum_{i=1}^{\kappa_N}\frac{1-e^{-\tilde{\Gamma}(t_i-s_i)}\left(1+(t_i-s_i)\left(1-e^{-\tilde{\Gamma}}\right)\right)}{\left(1-e^{-\tilde{\Gamma}}\right)^2}\Bigg)\Bigg)$$
$$\leq M_\Psi\left(\frac{2\bar{C}(\gamma_0,k)}{\mu_h\underline{\Psi}^F}N+\left(\frac{\bar{b}^2}{e^{\tilde{\Gamma}}-1}+\frac{\bar{C}(\gamma_0,k)}{\left(e^{\tilde{\Gamma}}-1\right)^2}\right)\kappa_N\right)$$

Getting $\mathcal{R}_N = \mathcal{O}\left(\sqrt{N}\right)$ when N is known in advance just amounts to plugging $\gamma_0 = N^{-1/2}$, $\|K - 1_{\mathbb{R}^+}\|_1 \propto N^{-1/2}$ and $\|k\|_\infty \propto \sqrt{N}$ in the last equation. □

Figure 6.4 illustrates the behaviour of Conv-OGA in a nonstationary environment. In agreement with Theorem 6.5.1 and Corollary 6.5.1, γ_0 controls the length of the transient regime due to the $\exp(-\mu_h\underline{\Psi}^F\gamma_0 n/2)$ term. Increasing γ_0 shortens it but increases the width of the band of the asymptotic regime as $C(\gamma_0,k)$ increases with γ_0 (blue *vs.* green curves). For a fixed γ_0, the stationary regime in terms of k exhibits a bias-variance trade-off: $\|K - 1_{\mathbb{R}^+}\|_1$ corresponds to the bias and $\|k\|_\infty$ to the variance (see Proposition 6.3.3). In the case of a Gaussian kernel, increasing the kernel variance σ^2 reduces the trajectory variance but increases bias (green *vs.* red curves).

Conv-OGA using a constant learning rate is an effective and efficient (i.e. real-time) algorithm for tracking monopoly prices of nonstationary bidders. It incurs $\mathcal{O}(\sqrt{N})$ regret given the horizon and κ_N, by tuning γ_0 and k, while maintaining the computational efficiency of online methods.

Conclusion

In this chapter we introduced V-Conv-OGA, the first real-time ($\mathcal{O}(1)$ update-time and memory) method for monopoly price learning. We first gave some theoretical results bridging auction theory and optimisation. Next, we proceeded to show how to fix the biased gradient problem with smooth surrogates, giving Conv-OGA. Next, we let the smoothing decrease over time in V-Conv-OGA, for whom we showed convergence of $\mathcal{O}(n^{-1/2})$. Finally, we adapted Conv-OGA to perform tracking of nonstationary bid distributions and obtained $\mathcal{O}(\sqrt{N})$ dynamic regret.

In the context of high-frequency auctions, computational efficiency precedes numerical precision, thus we traded $\mathcal{O}(n^{1/2})$ complexity and $\mathcal{O}(n^{-1})$ speed for $\mathcal{O}(1)$ complexity and $\mathcal{O}(n^{-1/2})$ speed. Whether or not it is possible to reach the optimal rate with a real-time algorithm remains an open question. We conjecture this to be impossible in general, but we know it is possible in some instances. If F is a symmetric distribution, then Conv-OGA with a constant *symmetric* kernel has no bias and $\mathcal{O}(n^{-1})$ convergence. Adapting the chosen kernel to some a priori knowledge about F is a possible direction to match the optimal offline rate.

The second question concerns the extension to partially observable settings, such as online eager auctions, when the seller does not observe bids under the reserve. Obviously, extensions using a reduction to multi-armed bandits (UCB, `Exp3`, `Exp4`, etc.) via a discretisation of the bid space cannot be real-time: the discretisation creates a need for $\mathcal{O}(\sqrt{n})$ in memory and the same for the update. Yet, it is possible to obtain a strait-forward extension of V-Conv-OGA in this setting, by plugging it into an Explore-Then-Commit (ETC) algorithm [100]: V-Conv-OGA learns an estimate of the monopoly price during the exploration period, which is then used during the exploitation period. As for other algorithms, by using a doubling trick to handle an unknown horizon, ETC+V-Conv-OGA exhibits a sub-linear regret. Unfortunately, like in the lazy auction setting, the regret is not optimal and the question of whether a real-time algorithm can match this optimal regret is still open.

The question of partial observability also applies to nonstationary bidders. In this case, extending Conv-OGA with ETC is no longer straightforward, as the switching times are unknown. Thus, it is not obvious when to re-trigger an exploration phase of ETC to adapt to the change of the bidder's distribution. A potential way to tackle this problem could be to use randomised resets for the algorithm [8] or change-point detection algorithms to trigger exploration [64].

6.A Pseudo- and Log-Concavity

We recall in this appendix some results of convex analysis on log-concave and pseudo-concave functions. Notations are independent from the rest of this chapter. Let $\mathcal{I} \subset \mathbb{R}$, be an interval and $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$ be the extended real number line.

6.A.1 Pseudo-concavity

Definition 6.1.
A function $f : \mathcal{I} \to \mathbb{R}, f \in \mathcal{C}^1(\mathcal{I})$, is pseudo-concave on $\mathcal{I}$ if

$$\mathrm{D}f(x)(x-y) \geq 0 \Rightarrow f(x) \geq f(y) \text{ for any } (x,y) \in \mathcal{I}^2.$$

Definition 6.2.
A function $f : \mathcal{I} \to \mathbb{R}, f \in \mathcal{C}^1(\mathcal{I})$, is strictly pseudo-concave on $\mathcal{I}$ if it is pseudo-concave and has at most one critical point.

6.A.2 Log-concavity

Definition 6.3.
A function $f : \mathcal{I} \to \bar{\mathbb{R}}$ is log-concave on $\mathcal{I}$ if for any $\alpha \in [0,1]$,

$$f(\alpha x + (1-\alpha)y) \geq f(x)^{\alpha} f(y)^{1-\alpha} \tag{6.20}$$

for all $(x,y) \in \mathcal{I}^2$. Note that, if f is a map from $\mathcal{I}$ to $\mathbb{R}_*^+$, this is equivalent to $f = e^{-\varphi}$ for some function φ which is convex on $\mathcal{I}$.

A log-concave function is strictly log-concave if (6.20) is strict for any $(x,y) \in \mathcal{I}^2$ such that $x \neq y$. If $f : \mathcal{I} \to \mathbb{R}_*^+$, this is equivalent to $f = e^{-\varphi}$ for some function φ which is strictly convex on $\mathcal{I}$.

Definition 6.4.
A function $f : \mathcal{I} \to \bar{\mathbb{R}}$ is μ-strongly log-concave on $\mathcal{I}$, for $\mu > 0$, if $x \mapsto f(x)e^{-\mu x^2}$ is log-concave.
Note that if $f : \mathcal{I} \to \mathbb{R}_*^+$ this is equivalent to saying $f = e^{-\varphi}$ for some function φ which is μ-strongly convex on $\mathcal{I}$.

We also recall a useful technical result for any log-concave function f, which is a straightforward consequence of the concavity characterization of $\log(f)$.

Proposition 6.A.1.
Let $f : \mathbb{R} \to \mathbb{R}_+$ be a strictly log-concave function. Then

$$\frac{f(v+\delta)}{f(v)} > \frac{f(u+\delta)}{f(u)}$$

for all $\delta > 0$, for any $u > v > 0$.

Proof of Proposition 6.A.1. The proof is a straightforward application of properties of strictly concave functions applied to $\log(f)$. Let $F(x,y) = (\log f(x) - \log f(y))/(x-y)$, then F is strictly decreasing in x for every fixed y (and vice-versa). Thus,

$$F(v+\delta, v) > F(v+\delta, u) > F(u+\delta, u)$$

which implies

$$\log f(v+\delta) - \log f(v) > \log f(u+\delta) - \log f(u)$$
$$\frac{f(v+\delta)}{f(v)} > \frac{f(u+\delta)}{f(u)}.$$

□

6.A.3 Stability under convolution

Theorem 6.A.1. **([65])**
Let $f \in \mathcal{C}^1(\mathcal{I}; \mathbb{R}_+) \cap \mathcal{L}^1(\mathcal{I}; \mathbb{R}_+)$ be pseudo-concave on $\mathcal{I}$ and $g \in \mathcal{L}^1(\mathbb{R}; \mathbb{R}_+)$ be log-concave. Then, $f \star g$ is pseudo-concave on $\mathbb{R}$.

We extend this theorem to strict pseudo-concavity.

Lemma 6.A.2. ***(Extension of [65])***
Let $f \in \mathcal{C}^1(\mathcal{I}; \mathbb{R}_+) \cap \mathcal{L}^1(\mathcal{I}; Rb_+)$ be strictly pseudo-concave on $\mathcal{I} := [x_1, x_2]$. Assume $\lim_{x \to x_1+} f(x) = \lim_{x \to x_2-} f(x) = 0$ and $g \in \mathcal{L}^1(\mathbb{R}; \mathbb{R}_+)$ is strictly log-concave. Then, $f \star g$ is strictly pseudo-concave on $\mathbb{R}$.

Proof. For simplicity, let us consider f to be extended by 0 on $\mathbb{R} \setminus \mathcal{I}$. The proof is conducted in two steps. First, we show $f \star g$ admits a maximum on

the interior of its domain (which is a critical point) and we denote it by x^*. Second, we show that $f \star g$ is strictly increasing on $(-\infty, x^*)$ and strictly decreasing on $(x^*, +\infty)$ which immediately proves the strict pseudo-concavity (including unicity of x^*).

1. Since $f \in \mathcal{C}^1(\mathcal{I}; \mathbb{R}_+) \cap \mathcal{L}^1(\mathcal{I}; Rb_+)$ and $g \in \mathcal{L}^1(\mathbb{R}; Rb_+)$, the convolution $f \star g$ is well defined and in $\mathcal{C}^1(\mathbb{R}; \mathbb{R}_+) \cap \mathcal{L}^1(\mathbb{R}; \mathbb{R}_+)$. As a result, $[f \star g](x) \to 0$ as $|x| \to \infty$ and Rolle's theorem guarantees that there exists at least one point $x^* \in \mathbb{R}$ such that $[f \star g](x^*) \geq [f \star g](x)$ for all $x \in \mathbb{R}$. Furthermore, Theorem 6.A.1 ensures that $f \star g$ is pseudo-concave, hence $\mathrm{D}[f \star g](x^*) = 0$.

2. Using the differentiation property of the convolution, one has that for all $x \in \mathbb{R}$,

$$\mathrm{D}[f \star g](x) = \int_{\mathbb{R}} f(t)\mathrm{D}g(x-t)\mathrm{d}t = \int_{x_1}^{x_2} \mathrm{D}f(t)g(x-t)\mathrm{d}t, \tag{6.21}$$

as $\lim_{x \to x_1} f(x) = \lim_{x \to x_2} = 0$. Let x^* be a critical point of $f \star g$, applying (6.21) at $x = x^*$ leads to

$$0 = \int_{x_1}^{x_2} \mathrm{D}f(t)g(x^*-t)\mathrm{d}t. \tag{6.22}$$

Moreover, let y^* be the unique critical point of f, which is in (x_1, x_2) by Rolle's theorem. Splitting the integral in (6.21) at y^* yields

$$\mathrm{D}[f \star g](x) = \int_{x_1}^{y^*} \mathrm{D}f(t)g(x-t)\mathrm{d}t + \int_{y^*}^{x_2} \mathrm{D}f(t)g(x-t)\mathrm{d}t. \tag{6.23}$$

The crux of the proof consists in proving that $\mathrm{D}[f \star g](x^* + \delta) > 0$ for all $\delta > 0$ and $\mathrm{D}[f \star g](x^* + \delta) < 0$ for all $\delta < 0$. Since the derivation is similar in both cases, we only display here the case $\delta > 0$. From (6.23), we have:

$$\begin{aligned}
&\mathrm{D}[f \star g](x^* + \delta) \\
&= \int_{x_1}^{y^*} \mathrm{D}f(t)g(x^* + \delta - t)\mathrm{d}t + \int_{y^*}^{x_2} \mathrm{D}f(t)g(x^* + \delta - t)\mathrm{d}t \\
&= \int_{x_1}^{y^*} \mathrm{D}f(t)g(x^* - t)\frac{g(x^* + \delta - t)}{g(x^* - t)}\mathrm{d}t + \int_{y^*}^{x_2} \mathrm{D}f(t)g(x^* - t)\frac{g(x^* + \delta - t)}{g(x^* - t)}\mathrm{d}t.
\end{aligned}$$

We now provide upper and lower bounds for $g(x^* + \delta - t)/g(x^* - t)$, respectively on $[x_1, y^*]$ and $[y^*, x_2]$. Let

$$t^* = \underset{t \in [x_1, y^*]}{\operatorname{argmax}} \frac{g(x^* + \delta - t)}{g(x^* - t)}$$

which is well-defined under our assumptions. Then,

$$\frac{g(x^* + \delta - t)}{g(x^* - t)} \leq \frac{g(x^* + \delta - t^*)}{g(x^* - t^*)} \text{ for } t \in [x_1, y^*].$$

Moreover, applying Proposition 6.A.1, we have

$$\frac{g(x^* + \delta - t)}{g(x^* - t)} > \frac{g(x^* + \delta - t^*)}{g(x^* - t^*)},$$

for (Lebesgue) almost every $t \in [y^*, x_2)$. Since f is strictly pseudo-concave, $\mathrm{D}f(t) > 0$ on $[x_1, y^*)$ and $\mathrm{D}f(t) < 0$ on $(y^*, x_2]$, which implies that

$$\begin{aligned}
\mathrm{D}[f \star g](x^* + \delta) &= \int_{x_1}^{y^*} \mathrm{D}f(t) g(x^* - t) \frac{g(x^* + \delta - t)}{g(x^* - t)} \mathrm{d}t \\
&\quad + \int_{y^*}^{x_2} \mathrm{D}f(t) g(x^* - t) \frac{g(x^* + \delta - t)}{g(x^* - t)} \mathrm{d}t \\
&< \int_{x_1}^{y^*} \mathrm{D}f(t) g(x^* - t) \frac{g(x^* + \delta - t^*)}{g(x^* - t^*)} \mathrm{d}t \\
&\quad + \int_{y^*}^{x_2} \mathrm{D}f(t) g(x^* - t) \frac{g(x^* + \delta - t^*)}{g(x^* - t^*)} \mathrm{d}t \\
&< \frac{g(x^* + \delta - t^*)}{g(x^* - t^*)} \mathrm{D}[f \star g](x^*) = 0,
\end{aligned}$$

by (6.22), which proves the desired result. □

Similar stability properties under convolution are asserted for strictly and strongly log-concave functions. The first result, Proposition 6.A.3, is standard and can be derived from the Prépoka-Leindler inequality, while Theorem 6.A.2 is retrieved from [108].

Proposition 6.A.3.
Let $f : \mathcal{I} \subset \mathbb{R} \to \mathbb{R}^+$ and $g : \mathbb{R} \to \mathbb{R}^+$ be log-concave. Then, $f \star g$ is log-concave.

Theorem 6.A.2. **([108, Thm. 6.6])**
Let $f : \mathcal{I} \subset \mathbb{R} \to \mathbb{R}^+$ and $g : \mathbb{R} \to \mathbb{R}^+$ be $\mu \in \mathbb{R}_+$ and $\mu' \in \mathbb{R}_+$ strongly log-concave, respectively. Then, $f \star g$ is $\mu\mu'/\sqrt{\mu^2 + \mu'^2}$ strongly log-concave. Furthermore, the convolution of two strictly log-concave functions is strictly log-concave.

Index

Bold page numbers are used to indicate the main reference for an entry.

Symbols

$\langle \cdot | \cdot \rangle_n$ The inner product defined in (5.37). 149
$\| \cdot \|_n$ The inner product defined in (5.37). 149

$\mathbb{A}_0$ For any $(t,x) \in [0,T) \times \mathbb{R}$ the set of maximisers of the maximum in the diffusive HJB equation,see Assumption 3.4. 57
$\bar{\mathcal{A}}^t$ The set of admissible controls starting from time $t \in \mathbb{R}_+$, i.e. all $\bar{\mathbb{F}}^t$-predictable processes taking values in $\mathbb{A}$. 49
$\bar{\mathcal{A}}$ The set of admissible controls, i.e. all $\bar{\mathbb{F}}$-predictable processes taking values in $\mathbb{A}$. 84, 130
$\mathscr{A}_T$ The set of Borel-measurable maps from $[0,T) \times \mathbb{R}^d$ to $\mathbb{A}$. 45
$\mathscr{A}$ The set of Borel-measurable maps from $\mathbb{R}^d$ to $\mathbb{A}$. 94, 133
$\bar{a}$ A pointwise maximiser in the HJB equation. 54
$\check{a}_\varepsilon$ The maximiser map in Assumption 3.5 used in higher order error correction. 63
$\check{a}$ The first-order corrected decision rule, defined in (3.39). 59
$\hat{a}$ A pointwise maximiser in the HJB equation (3.7), see Proposition 3.2.2. 45
$\mathbb{A}$ The set of admissible actions, a compact subset of $\mathbb{R}^{d_{\mathbb{A}}}$. 43, 78, 125
$\mathcal{A}$ The set of admissible controls, i.e all $\mathbb{F}$-predictable processes taking values in $\mathbb{A}$. 43, 78, 125
$\bar{\alpha}^{t,x}$ The control in $\mathcal{A}^t$ defined by using the decision rule $\bar{a}$, see (3.32). 56
$\bar{\alpha}$ An arbitrary control in $\bar{\mathcal{A}}$. 84
$\check{\alpha}^{t,x}$ The first-order corrected control, defined in (3.38). 59
$\hat{\alpha}$ A piecewise-constant optimal control in the pure-jump problem, see Proposition 3.2.2. 45
$\mathcal{A}^t$ The set of admissible controls starting from time $t \in \mathbb{R}_+$, i.e all $\mathbb{F}^t$-predictable processes taking values in $\mathbb{A}$. 43

B_n The random set $\mathcal{B}(\sup_{t \le \tau_n} \|X_t^{\varpi,\theta^*}\|)$. 132

b_1 The drift-like component of the pre-limit pure jump process. 48, 83

b_2 The volatility-like component of the pre-limit pure jump process. 48, 83

$\bar{b}$ The maximum of the support of F in Chapter 6. 184

b The jump function of the pure-jump system. 43

b_ε The jump function of the pre-limit problem for $\varepsilon > 0$. 48

β_{n_k} The radius of the lazily updated confidence set $\mathcal{C}_{n_k}$. 129

β_n The radius of the theoretical confidence set $\mathcal{C}_n$. 131

B_k The bias of the smoothed monopoly revenue Ψ_k^F. 193

$C_{\bar{X}}$ The constant in the moment condition on $\bar{X}^{x,\bar{\alpha}}$, see Assumption 4.5. 84

$\mathcal{C}_{n_k}$ The confidence set of Algorithm 1, in episode k. 129

$\mathscr{C}_n^\varepsilon$ The $\varepsilon\|\bar{\Sigma}\|_{\mathrm{op}}/n$ cover of $\mathscr{F}_\Theta$, restricted to the ball $\mathcal{B}(H_\delta(n))$. 131

$\mathcal{C}_n$ The confidence set considered in Section 5.5.1. 156

$\mathscr{C}_n^\Gamma$ The cover of $\mathscr{F}$ considered in Section 5.5.1. 150

C_Υ The coercivity constant of Assumption 4.7. 95

C_ξ^1 The contraction coefficient under ξ of the stochastic generator in Assumption 4.3. 80

C_ξ^2 The additive constant in condition (4.7) of Assumption 4.3. 80

$C_{\bar{\xi}}^1$ The contraction constant of (4.14). 85, 95

$C_{\bar{\xi}}^2$ The additive constant of (4.14) and growth rate constant of (4.15). 85, 95

C_ζ The contraction coefficient under ζ of the stochastic generator in Assumption 4.2. 80

$\mathcal{C}^0$ The space of continuous functions. 43

$\mathcal{C}^{0,1}$ The space of Lipschitz functions. 78

$\mathcal{C}^1$ The space of functions which are once continuously differentiable. 43

$\mathcal{C}^{(1,2)}$ The set of functions which are once continuously differentiable in time and twice in space, with bounded derivatives up to these orders. 43

$\mathcal{C}_b^{(1,2)}$ The space of functions which are once continuously differentiable in time and twice in space, with bounded derivatives up to these orders. 43

$\mathcal{C}^2$ The space of functions which are twice continuously differentiable. 43

$\bar{\mathfrak{c}}_p$ The contraction coefficient of the diffusive system in (5.61). 168

$\mathfrak{c}_p$ The additive constant in Corollary 5.4.1. 145

$\mathfrak{c}'_{\mathscr{V}}$ The additive constant in the stochastic Lyapunov condition of Lemma 5.4.1.(ii). 138

$\mathfrak{c}_{\mathscr{V}}$ The contraction coefficient of the pure-jump system in Assumption 5.2. 128

d_Θ The dimension of the parameter space Θ. 126

$d_\mathbb{A}$ The dimension of the action space $\mathbb{A}$. 125

$\mathbb{D}_T$ The Skorohod space of càdlàg functions defined on $[0, T]$ with values in $\mathbb{R}^d$. 43

$\mathbb{D}$ The Skorohod space of càdlàg functions defined on $[0, +\infty)$ with values in $\mathbb{R}^d$. 78, 84, 125

$\mathrm{d}_{\mathrm{E},N_T}$ The N_T-point eluder dimension of $\mathscr{F}_\Theta$, restricted to the ball $\mathcal{B}(H_\delta(N_T))$. 132

$\mathrm{d}_{\mathrm{E},n}$ The n-point eluder dimension of $\mathscr{F}_\Theta$, restricted to the ball $\mathcal{B}(H_\delta(n))$. 132, 161

$\mathrm{d}_{\mathrm{E},T\varepsilon^{-1}}$ The $2\varepsilon/\sqrt{T}$-eluder dimension of $\mathscr{F}_\Theta$, restricted to the ball $\mathcal{B}(H_\delta(T/\varepsilon))$. 130, 170

$\delta\bar{V}_T^{(1)}$ The unique bounded viscosity solution of (3.34). 58

$\delta\bar{V}_T^{(i)}$ The higher order error correction functions, see Assumption 3.5. 63

$\delta\gamma_i$ The Hölder regularity exponent of $\partial_{xx}\delta\bar{V}_T^{(i)}$, if it exists. 63

$\delta\gamma$ The Hölder regularity exponent of $\partial_{xx}\delta\bar{V}_T^{(1)}$, if it exists. 58

δ The confidence level parameter of Algorithm 1. 129

$\delta r_\varepsilon^{(1)}$ The first order controlled process approximation correction function, see 2. of the proof of Theorem 3.3.2. 61

$\delta r_\varepsilon^\theta$ The controlled process approximation error for the pre-limit problem associated to $\theta \in \Theta$, for $\varepsilon > 0$. 169

δr_ε The controlled process approximation error for the pre-limit problem for $\varepsilon > 0$. 54, 88

δr_n The controlled process approximation error for the pre-limit problem in the discrete-time case, see Section 3.5. 71

$\mathrm{d}_{\mathrm{E},n}^{\mathscr{F}}$ The $2\cdot/\sqrt{n}$-eluder dimension of $(f|_{B_n})_{f\in\mathscr{F}}$ considered in Section 5.5.1. 158

e_θ The error function in the approximation of Proposition 5.3.6 for the ergodic control problem with parameter θ. 135, 169

$\ell_{\mathscr{V}}$ The lower growth rate of $\mathscr{V}$ in Assumption 5.2. 128

η_ε The intensity of the pre-limit problem $\eta_\varepsilon := \varepsilon^{-1}$. 48, 82

η The intensity of the Poisson random measure N. 43, 78

$\mathbb{F}^t$ The $\mathbb{P}$-augmentation of the raw filtration generated by the restriction of N to $[t, +\infty)$. 43

$\mathbb{F}$ The $\mathbb{P}$-augmentation of the raw filtration generated by N from 0 onwards in Chapters 3 to 5; The filtration generated by the bid process in Chapter 6, see $\mathcal{F}_n$. 43, 78, 125, 184

$\mathcal{F}_t$ equal to $\mathcal{F}_s^0$; The sigma-algebra generated by the past of N (considered on $[0,+\infty)\times\mathbb{R}^d$) at time t. 78, 125

$\mathcal{F}_s^t$ The sigma-algebra generated by the past of the restriction of N to $[t,+\infty)\times\mathbb{R}^d$ at time $s \geq t \geq 0$. 43

$\mathcal{F}$ The σ-algebra of the considered filtered probability space, either $\mathcal{F}_T^0$ in Chapter 3 or $\mathcal{F}_\infty$ in Chapters 4 and 5. 43, 78

F The bid distribution considered in Chapter 6. 184

$\bar{\mathbb{F}}^t$ The $\bar{\mathbb{P}}$-augmentation of the raw filtration generated by the restriction of W to $[t,+\infty)$. 49

$\bar{\mathbb{F}}$ The $\bar{\mathbb{P}}$-augmentation of the raw filtration generated by W. 84, 130

$\bar{\mathcal{F}}_s^t$ The σ-algebra generated by the past of the restriction of W to $[t,+\infty)$ at time $s \geq t \geq 0$. 49

$\bar{\mathcal{F}}_s$ Equals $\bar{\mathcal{F}}_s^0$; The σ-algebra generated by the past of W at time $s \geq 0$. 84

$\mathscr{F}$ The function class considered in Section 5.5.1. 149

f^* The function to be estimated in Section 5.5.1. 149

$f_{x,a}^{\nu,\varepsilon}$ The transition kernel induced on the pure-jump problem by the numerical scheme of Section 4.4 68

$\mathcal{F}_n$ The σ-algebra engendered by the bid process up to time $n-1 \in \mathbb{N}^*$ in Chapter 6. 184

$\mathscr{F}_\Theta$ The function class of drifts of the RL problem. 131

γ_0 The constant learning rate of Conv-OGA used in the nonstationary setting. 206

γ The Hölder regularity exponent of the second derivative of the value function. 50, 86, 130, 170

$\mathbb{H}$ Equals $(\mathcal{H}_i)_{i\geq 0}$; The filtration generated by $(\xi)_{i\in\mathbb{N}}$ considered in Section 5.5. 149

$\mathcal{H}_i$ The sigma-algebra generated by the past of the process $(\xi)_{i\in\mathbb{N}}$ at time $i \in \mathbb{N}$. 149

H_δ The high probability band in which the stable process lives 142

h The mesh size of $\bar{\mathrm{M}}_h^\kappa$ in Chapter 4. 93

h_{num} The maximum value of h for which the approximation result of Proposition 4.4.1 holds. 95

$\mathcal{I}$ The given interval containing the monopoly price. 193

J_T^ε The gain functional of the pre-limit finite-horizon control problem for ε^{-1}. 48

J_T The gain functional of the pure-jump finite-horizon control problem. 44, 81

J_λ The gain functional of the pure-jump ergodic control problem with discount factor λ. 81

$\bar{J}_T$ The gain functional of the finite-horizon control problem (3.5). 85

$\bar{J}_\lambda$ The gain functional of the diffusive ergodic control problem with discount factor λ. 84

$\mathcal{K}$ The set of admissible smoothing kernels, i.e. log-concave elements of $\mathcal{C}^1(\mathbb{R};\mathbb{R}_+) \cap \mathcal{L}^1_1(\mathbb{R};\mathbb{R}_+)$. 192

k In Chapter 6, a convolution kernel in $\mathcal{K}$. 191

κ The number of points in $\bar{\mathrm{M}}_h^\kappa \cap [z_1, 0)$, i.e. $|\bar{\mathrm{M}}_h^\kappa| = 2\kappa + 1$. 93

κ_N The number of switches in the nonstationary setting, see Assumption 6.5. 205

$L_{\mathscr{V}}$ The upper growth rate of $\mathscr{V}$ in Assumption 5.2. 128

L_W The Lipschitz constant of the family of functions $(W^*_\theta)_{\theta\in\Theta}$. 133, 164

$L_{\bar{V}}$ The Lipschitz constant of $\bar{V}_\lambda$. 84

L_{b_1,b_2} The coefficient growth constant of b_1 and b_2, see Assumption 4.4. 83

$L^\gamma_{\bar{\mathrm{w}}}$ The regularity constant of $\bar{\mathrm{w}}$ in Theorem 4.3.1. 86

L_0 The coefficient regularity constant defined of Assumption 5.1. 127

$\bar{\mathcal{L}}^{\bar{a}}_h$ The generator of the controlled process induced by centred finite differences with fineness h, see (4.25). 93

$\bar{\mathcal{L}}$ The generator of the diffusion process in (4.12). 85

$L_{b,r}$ The regularity constant of the coefficients, see Assumption 4.1. 79

L_{num} The constant in the approximation result for the numerical scheme of Section 4.4, see Proposition 4.4.1. 95

L_V The Lipschitz constant of V_λ, see Lemma 4.A.1. 111

L_ξ The upper growth rate of ξ in Assumption 4.3. 80

ℓ_ξ The lower growth rate of ξ in Assumption 4.3. 80

L_ζ The upper growth rate of ζ in Assumption 4.2. 79

ℓ_ζ The lower growth rate of ζ in Assumption 4.2. 79

$\mathcal{M}_+$ The restriction of $\mathcal{M}$ to positive measures. 43

$\mathcal{M}$ The space of finite measures on the measure space given by the argument space and its Borel σ-algebra. 78, 223

$\bar{\mathrm{M}}_t$ The mesh of the time interval $[0, T]$ for. 68

$\bar{\mathrm{M}}_x$ The mesh on the space domain $[-1, 3]$ for. 68

L^κ_h The set of maps from $\bar{\mathrm{M}}^\kappa_h$ to $\mathbb{R}$. 93

M_t The mesh of the time interval $[0, T]$ for (3.7). 67

M_ε The mesh for the numerical scheme of the pure-jump problem in Section 4.5 107

M_x The mesh on the space domain $[-1, 3]$ for (3.7). 67

$\bar{\mathrm{M}}^\kappa_h$ The mesh for the numerical scheme of Section 4.4. 93

$\check{\mathrm{M}}_h^\kappa$ The mesh for the numerical scheme $\bar{\mathrm{M}}_h^\kappa$ without its boundary points. 93

$M_{\mathscr{V}}$ The Lipschitz regularity upper bound on $\mathscr{V}$ in Assumption 5.2. 128

$M'_{\mathscr{V}}$ The hessian regularity upper bound on $\mathscr{V}$ in Assumption 5.2. 128

$\bar{\mu}_\theta$ The drift given the parametrisation $\theta \in \Theta$. 127

μ The drift of the diffusion limit SDE (3.13). 49, 83

μ_F The strong log-concavity constant of Ψ^F. 189

N An $\mathbb{R}^d$-marked Poisson process on $\mathbb{R}_+$ ($d = 1$ in Chapter 3, $d \in \mathbb{N}$ otherwise) with compensator $\eta\nu(\mathrm{d}e)\mathrm{d}t$. 43, 78, 125

$\mathscr{N}_n^\varepsilon$ The $\varepsilon\|\bar{\Sigma}\|_{\mathrm{op}}/n$ covering number of $\mathscr{F}_\Theta$, restricted to the ball $\mathcal{B}(H_\delta(n))$. 131

$\mathscr{N}_{T\varepsilon^{-1}}^\varepsilon$ The $\varepsilon^2\|\bar{\Sigma}\|_{\mathrm{op}}^2/T$ covering number of $\mathscr{F}_\Theta$, restricted to the ball $\mathcal{B}(H_\delta(T/\varepsilon))$. 130, 170

$\mathscr{N}_n^\Gamma$ The Γ^2/n-covering number of $\mathscr{F}$ considered in Section 5.5.1. 150

ν A measure on $\mathbb{R}^d$ such that $\nu(\mathrm{Id}) = 0$ and $\nu(\mathrm{Id}(\cdot)^2) < +\infty$. 43, 78, 125

$\mathbb{P}$ A probability measures on $\mathbb{D}$ which renders N a $\mathbb{P}$-Poisson random measure. 43, 78, 125

$\bar{\mathbb{E}}$ The expectation under $\bar{\mathbb{P}}$. 49

$\bar{\mathbb{P}}$ A probability measure on $\mathbb{D}$ which renders W a $\bar{\mathbb{P}}$-Wiener Process. 49, 84

$p(r, b)$ The instant revenue of a single-buyer second-price auction with reserve price r when the bid has value b Chapter 6. 184

$p_{\bar{X}}$ The exponent in the moment condition on $\bar{X}^{x,\bar{\alpha}}$, see Assumption 4.5. 84

Ψ_k^F Equals $\Psi^F \star k$; the monopoly revenue smoothed by $k \in \mathcal{K}$. 191

p_ξ The growth order of ξ in Assumption 4.3. 80

p_ζ The growth order of ζ in Assumption 4.2. 79

$\bar{\pi}_\theta^*$ An optimal Markov policy of the diffusive control problem (5.10) parametrised by θ. 130

π_θ^* An optimal Markov policy of the control problem (5.2), according to the parametrisation θ. 133, 164

π_k The policy of Algorithm 1, in episode k. 129

p_k Equals $p \star k$. 191

ψ_θ^ε The dynamic of the system $\psi_\theta^\varepsilon(x, a) := x + \bar{\mu}_\theta(x, a)$. 128

$\Psi^{\hat{F}_n}$ The Monopoly revenue of the empirical distribution $\hat{F}_n$ given n samples of F. 185

Ψ^F The Monopoly revenue of a distribution F on $\mathbb{R}_+$. 184

ψ_F The virtual value function of the distribution F. 189

$p_{\bar{\xi}}$ The growth order of $\bar{\xi}$ in Assumption 4.7. 85, 95

$\bar{Q}_h^{\bar{a}}$ The transition matrix of the numerical scheme of Section 4.4. 94

q_h Equals $1-(\sigma/L_{b_1,b_2})^2$; The diagonal entries of $\bar{Q}_h^{\bar{a}}$ (except at the boundary). 94

q_h^- Equals $(-\mu h+\sigma^2)/(2L_{b_1,b_2}^2)$; The subdiagonal entries of $\bar{Q}_h^{\bar{a}}$ (except at the boundary). 94

q_h^+ Equals $(\mu h+\sigma^2)/(2L_{b_1,b_2}^2)-\Delta t_h\lambda$; The supradiagonal entries of $\bar{Q}_h^{\bar{a}}$ (except at the boundary). 94

$R_{h_\varepsilon}^{\kappa_\varepsilon}$ The reward approximated on the mesh M_ε. 107

R_h^κ The reward approximated on the mesh $\bar{M}_h^\kappa$. 94

R_ε equal to $\sqrt{8d\log 1/\varepsilon}$. 137

$\mathcal{R}_T(\alpha)$ The regret of the control α at time T. 126

$\bar{r}$ The reward function of the diffusion limit control problem. 127

r_1 Equals $\lim_{\varepsilon\to 0}\varepsilon^{-\frac{\gamma}{2}}\delta r_\varepsilon$; see Assumption 3.4. 57

r The reward function of the control problems. 44

r^* The monopoly price of F. 184

r_k^* The monopoly price of Ψ_k^F in the nonstationary setting of Section 6.5. 193

r_n^* The monopoly price of F_n in the nonstationary setting. 185

$\bar{\rho}_\theta^{\bar{\alpha}}$ The ergodic gain functional, evaluated at $\bar{\alpha}$ of the RL problem according to the parametrisation θ. 130

$\bar{\rho}_h^{\kappa,*}$ The value of the Diffusion ergodic control problem, approximated on the mesh $\bar{M}_h^\kappa$. 93

$\bar{\rho}_\theta^*$ The value of the diffusive ergodic control problem parametrised by θ. 134, 165

$\bar{\rho}^*$ The value of the pure-jump ergodic control problem. 84

$\bar{\rho}_\theta^*$ The value of the ergodic control problem in the diffusive limit corresponding to the parametrisation θ. 130

$\delta\bar{\rho}_\varepsilon^*$ The correction terms for the ergodic control problem see (4.24). 91

$\rho_{\theta*}^\alpha$ The ergodic gain functional, evaluated at α of the RL problem according to the true parametrisation θ^*. 126

ρ_θ^* The value of the pure-jump ergodic control problem parametrised by θ 133, 164

$\rho_{\theta*}^*$ The value of the ergodic control problem in the system corresponding to the true parametrisation θ^*. 126

ρ_ε^* The value of the pre-limit pure-jump ergodic control problem for $\varepsilon>0$. 82

ρ^* The value of the pure-jump ergodic control problem. 79, 82

ρ The gain functional of the pure-jump ergodic control problem. 79

$\mathbb{S}^d$	The set of symmetric matrices of size $d \times d$. 77
Σ	The covariance of the additive martingale noise in Chapter 5. 125
$\bar{\Sigma}$	The volatility of the SDE (5.9). 127
σ	The volatility of the diffusion limit SDE (3.13). 49, 83
Δt_h	Equals $(h/L_{b_1,b_2})^2$; the intensity of the Poisson process of the Markov chain scheme of Section 4.4. 94
τ_n	The n-th arrival time of the process N, unless otherwise specified. 125
$\hat{\theta}_{n_k}$	The NLLS estimate of the parameter θ^* with n_k points. 129
Θ	The parameter space indexing $\mathscr{F}_\Theta$. 126
θ_n	The chosen parametrisation of the system on $[\tau_n, \tau_{n+1}]$. 135, 170
θ^*	The nominal (true) parametrisation of the system. 126
Υ	The radius of the ball in the coercivity condition of Assumption 4.7. 95
$V^{\varepsilon,\Delta}_{T,n}$	The numerical solution to (3.7). 68
V^ε_T	The value function of the pre-limit finite-horizon control problem for ε^{-1}. 48
V_T	The value function of the pure-jump finite-horizon control problem. 44, 81
V^n_T	The discrete time control problem with n jumps in $[0, T]$. 70
V_λ	The value function of the pure-jump ergodic control problem with discount factor λ. 81
$\bar{V}^{(1),\varepsilon}_T$	The first order corrected approximation, i.e. $\bar{V}^{(1),\varepsilon}_T := \bar{V}_T + \varepsilon^{\gamma/2}\delta\bar{V}^{(1)}_T$. 58
$\bar{V}^\Delta_{T,n}$	The numerical solution to. 68
$\bar{V}_T$	The value function of the Finite horizon diffusion limit control problem. 49, 85
$\bar{V}_\lambda$	The value function of the diffusive discounted control problem with discount factor λ. 84
$\mathscr{V}$	The Lyapunov function of Assumption 5.2. 128
v	The value for the auctioned good in the auction example of Section 3.4 65
ς	The ellipticity constant of the diffusion limit, see Eq. (3.10). 48, 83, 127
$\varepsilon_\circ$	The constant defined as $(L_{b_1,b_2})^{-2}$ in Proposition 4.3.1. 88
ε	The approximation parameter of the diffusion limit; the mean inter-arrival time of N^ε_s. 48, 82, 125
V_k	The variance of the gradient of the smoothed revenue Ψ^F_k. 193

$\mathcal{W}^{\kappa_\varepsilon}_{h_\varepsilon}$ The solution of the diffusive PDE (4.16) $\bar{w}^{\kappa}_{h}$ approximated on the mesh M_ε. 107

$\bar{\mathcal{W}}^{\kappa}_{h}$ The solution of the diffusive PDE (4.16) $\bar{w}^{\kappa}_{h}$ approximated on the mesh $\bar{M}^{\kappa}_{h}$. 94

W_ε Defined as $\varepsilon^{-\gamma/2}(V^\varepsilon_T - \bar{V}_T)$. 59

W A Wiener process on $\mathbb{R}$ in Chapter 3 $\mathbb{R}^d$ in Chapter 4. 49, 84, 130

$\bar{w}^{\kappa,n}_{h}$ The smoothed version of $\bar{w}^{\kappa}_{h}$ used in Section 4.4.2. 98

$\bar{w}^{\kappa}_{h}$ The decision part of the solution of the HJB equation for the Diffusion ergodic control problem, approximated on the mesh $\bar{M}^{\kappa}_{h}$. 93

$\bar{w}$ The decision part of the solution of the HJB equation for the Diffusion ergodic control problem. 86

$\bar{W}^*_\theta$ The decision function for the diffusive ergodic control problem, with parameter θ. 134, 165

w The decision part of the solution of the HJB equation for the pure-jump ergodic control problem. 75, 82

W^*_θ The decision function for the ergodic control problem, with parameter θ. 133, 164

$\hat{X}$ The discrete time controlled process with n jumps in $[0, T]$. 70

$X^{x,\alpha}$ The unique solution to the SDE (4.3)with initial condition $x \in \mathbb{R}^d$ at time 0 and control $\alpha \in \mathcal{A}$. 78

$X^{\alpha,\theta*}$ The true state process of the system in Chapter 5 for a control $\alpha \in \mathcal{A}$. 125

$X^{t,x,\alpha}$ The unique solution to the SDE (3.4) with initial condition $x \in \mathbb{R}^d$ at time $t \in \mathbb{R}_+$ and control $\alpha \in \mathcal{A}$. 44

$\bar{X}^{x,\bar{\alpha}}$ The unique solution to the SDE (4.12) with initial condition $x \in \mathbb{R}^d$ at time 0 and control $\bar{\alpha} \in \bar{\mathcal{A}}$. 84

$\bar{X}^{t,x,\bar{\alpha}}$ The unique solution to the SDE (4.12) with initial condition $x \in \mathbb{R}^d$ at time $t \in \mathbb{R}_+$ and control $\bar{\alpha} \in \bar{\mathcal{A}}$. 49

$(X_i)_{i\in\mathbb{N}}$ The feature process considered in Section 5.5.1. 149

$\tilde{X}^{\varpi,\theta}_{\tau_{n+1}}$ The counterfactual (in the system θ) of the state at time τ_{n+1}, given $X^{\varpi,\theta^*}_{\tau_n}$, according to the control ϖ. 135

x_0 The initial condition of the state process in Chapter 5. 125

$\tilde{X}^{x,\bar{a}}$ The controlled Markov chain corresponding to the numerical scheme of Section 4.4. 95

$\bar{\xi}$ The Lyapunov-like function of Assumption 4.7. 85, 95

$(\xi_i)_{i\in\mathbb{N}}$ The noise process considered in Section 5.5.1, iid centred Gaussian random variables on $\mathbb{R}^d$. 149

ξ A standard centred Gaussian random variable on $\mathbb{R}^d$. 125

ξ The Lyapunov-like function of Assumption 4.3. 80

$(Y_i)_{i\in\mathbb{N}}$	The regresand process considered in Section 5.5.1. 149
ζ	The Lyapunov-like function of Assumption 4.2. 79

Abbreviations

CARE	Continuous Algebraic Riccati Equation 128
CDF	Cumulative Distribution Function 184, 193
DPP	Dynamic Programming Principle 115
DRL	Deep Reinforcement Learning 2
EDO	Équation Différentielle Ordinaire 25, 26
EDP	Équation aux Dérivées Partielles 19, 23, 24, 27
EDS	Équation Différentielle Stochastique 21, 23, 24, 28, 36
ERM	Empirical Risk Minimisation 183, 185, 186, 188, 205
ETC	Explore-Then-Commit 209
HJB	Hamilton-Jacobi-Bellman 4, 5, 7, 12, 22, 24, 26, 31, 44, 49, 110, 115, 130, 133–135, 166, 171, 178, 219
LQ	Linear Quadratic 10, 124
MDPs	Markov Decision Processes 123, 124
ML	Machine Learning 1, 2, 9
NLLS	Non-Linear Least Squares 12, 124, 129, 131, 132, 157, 163, 178
ODE	Ordinary Differential Equation 6, 7, 111
OFU	Optimism in the Face of Uncertainty v, 124, 128, 129, 132, 248
OGA	Online Gradient Ascent 15, 186–188, 191–193, 196
PDE	Partial Differential Equation 1, 5, 134, 179
PDF	Probability Density Function 184
PSD	Positive Semi-Definite 128
RL	Reinforcement Learning 2, 6, 9–13, 121, 123, 126, 128, 178
SDE	Stochastic Differential Equation 3, 4, 9, 16, 49
TPM	Transition Probability Matrix 94
UCB	Upper Confidence Bound 124, 186, 209

Bibliography

[1] 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.
124

[2] 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.
124, 129

[3] Marc Abeille, Bruno Bouchard, and Lorenzo Croissant.
Diffusive limit approximation of pure-jump optimal ergodic control problems, September 2022.
arXiv:2209.15284 [math].
73

[4] Marc Abeille, Bruno Bouchard, and Lorenzo Croissant.
Diffusive limit approximation of pure-jump optimal stochastic control problems.
Journal of Optimization Theory and Applications, 196:147–176, 2023.
39

[5] Marc Abeille, Bruno Bouchard, and Lorenzo Croissant.
Reinforcement Learning in near-continuous time with continuous states, 2023.
Under review at the *35th International Conference on Algorithmic Learning Theory*.
121

[6] Marc Abeille and Alessandro Lazaric.

Linear Thompson sampling revisited.
In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 176–184. PMLR, 20–22 Apr 2017.
11, 31

[7] 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.
10, 29, 124

[8] Robin Allesiardo, Raphaël Féraud, and Odalric-Ambrym Maillard.
The non-stationary stochastic multi-armed bandit problem.
International Journal of Data Science and Analytics, 3(4):267–283, Jun 2017.
209

[9] Kareem Amin, Afshin Rostamizadeh, and Umar Syed.
Learning prices for repeated auctions with strategic buyers.
In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
105

[10] Kareem Amin, Afshin Rostamizadeh, and Umar Syed.
Repeated contextual auctions with strategic buyers.
In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.
183, 186

[11] Ben Andrews.
Fully nonlinear parabolic equations in two space variables.
arXiv preprint math/0402235, 2004.
52

[12] Ari Arapostathis, Vivek S. Borkar, Emmanuel Fernández-Gaucherand, Mrinal K. Ghosh, and Steven I. Marcus.
Discrete-time controlled Markov processes with average cost criterion: A survey.
SIAM Journal on Control and Optimization, 31(2):282–344, March 1993.
Society for Industrial and Applied Mathematics.
10, 29

[13] 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.
79, 84, 85, 86, 87, 111, 124, 134

[14] Mariko Arisawa and Pierre-Louis Lions.
On ergodic stochastic control.
Communications in partial differential equations, 23(11-12):2187–2217, 1998.
7, 25, 111, 113, 115, 124, 134

[15] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer.
Finite-time analysis of the multiarmed bandit problem.
Machine learning, 47:235–256, 2002.
11, 31, 124

[16] 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.
2, 11, 20, 31, 124

[17] Guy Barles.
An introduction to the theory of viscosity solutions for first-order Hamilton–Jacobi equations and applications.
In *Hamilton–Jacobi equations: approximations, numerical analysis and applications*, pages 49–109. Springer, 2013.
59

[18] Guy Barles and João Meireles.
On unbounded solutions of ergodic problems in $\mathbb{R}^m$ for viscous Hamilton-Jacobi equations.
Communications in Partial Differential Equations, 41(12):1985–2003, December 2016.
104

[19] Guy Barles and Panagiotis E. Souganidis.
Convergence of approximation schemes for fully nonlinear second order equations.
Asymptotic analysis, 4(3):271–283, 1991.
124

[20] 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, pages 35–42, Arlington, VA, 2009. AUAI Press.
13, 32, 124

[21] N. Bäuerle.
Approximation of optimal reinsurance and dividend payout policies.
Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics, 14(1):99–113, 2004.
5, 24, 42, 76

[22] Dimitri P. Bertsekas.
Dynamic Programming and Optimal Control, volume II.
Athena Scientific, Belmont, MA, 3rd edition, 2011.
2, 20, 123

[23] Dimitri P. Bertsekas.
Reinforcement Learning and Optimal Control.
Athena Scientific, Belmont, Massachusetts, 2nd printing (includes editorial revisions) edition, 2019.
2, 20

[24] Dimitri P. Bertsekas and Steven E. Shreve.
Stochastic Optimal Control. The Discrete-Time Case.
Academic Press, New York, 1978.
46, 54, 56, 114

[25] David Blackwell.
Discrete Dynamic Programming.
The Annals of Mathematical Statistics, 33(2):719–726, 1962.
Publisher: Institute of Mathematical Statistics.
7, 25

[26] Avrim Blum and Jason D. Hartline.
Near-optimal online auctions.
In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '05, pages 1156–1163, USA, 2005. Society for Industrial and Applied Mathematics.
183, 185

[27] Avrim Blum, Vijay Kumar, Atri Rudra, and Felix Wu.
Online Learning in online auctions.
Theoretical Computer Science, 324(2-3):137–146, 2004.
66, 183, 185

[28] 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.
Society for Industrial and Applied Mathematics.
124

[29] Vivek S. Borkar, Shuhang Chen, Adithya Devraj, Ioannis Kontoyiannis, and Sean P. Meyn.
The ODE method for asymptotic statistics in Stochastic Approximation and Reinforcement Learning.
arXiv:2110.14427 [cs, math, stat], December 2021.
arXiv: 2110.14427.
80

[30] Vivek S. Borkar and Mrinal K. Ghosh.
Ergodic control of multidimensional diffusions, II: Adaptive control.
Applied Mathematics and Optimization, 21(1):191–220, January 1990.
1, 19

[31] Vivek S. Borkar and Pravin P. Varaiya.
Adaptive control of Markov chains, I: Finite parameter set.
In *IEEE Transactions on Automatic Control*, volume 24, pages 953–957, December 1979.
IEEE Transactions on Automatic Control.
1, 19

[32] Léon Bottou.
Online Learning and Stochastic Approximations.
On-line learning in neural networks, 17(9):142, 1998.
15, 35, 193, 196, 197

[33] Bruno Bouchard.
Stochastic targets with mixed diffusion processes and viscosity solutions.
Stochastic processes and their applications, 101(2):273–302, 2002.
45

[34] Bruno Bouchard, Stefan Geiss, and Emmanuel Gobet.
First time to exit of a continuous itô process: General moment estimates and L_1-convergence rate for discrete time approximations.
Bernoulli, 23(3):1631–1662, 2017.
52

[35] Bruno Bouchard and Nizar Touzi.
Weak Dynamic Programming Principle for viscosity solutions.
SIAM Journal on Control and Optimization, 49(3):948–962, 2011.
45, 113, 115

[36] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart.

Concentration inequalities: A nonasymptotic theory of independence.
Oxford University Press, Oxford, 1st edition, 2013.
140, 176

[37] Pierre Brémaud.
Point processes and queues: Martingale dynamics, volume 66.
Springer, 1981.
43, 78

[38] Sebastien Bubeck, Nikhil R. Devanur, Zhiyi Huang, and Rad Niazadeh.
Online auctions and multi-scale Online Learning.
In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 497–514. ACM, 2017.
183, 185, 186

[39] Sebastien Bubeck, Nikhil. R. Devanur, Zhiyi Huang, and Rad Niazadeh.
Multi-scale Online Learning: Theory and applications to online auctions and pricing.
The Journal of Machine Learning Research, 20(1):2248–2284, 2019.
66

[40] Valerii Vladimirovich Buldygin and Yuriy Vasyliovych Kozachenko.
Metric characterization of random variables and random processes, volume 188 of *Translations of mathematical monographs*.
American Mathematical Society, Providence, RI, 2000.
173

[41] Jeremy Bulow and Paul Klemperer.
Auctions versus negotiations.
American Economic Review, 86(1):180–94, 1996.
65

[42] Nicolò. Cesa-Bianchi, Claudio Gentile, and Yishay Mansour.
Regret minimization for reserve prices in second-price auctions.
IEEE Transactions on Information Theory, 61(1):549–564, 2014.
66, 183, 185, 186

[43] 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.
5, 24, 42, 76, 123

[44] 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.
124

[45] Asaf Cohen and Virginia R. Young.
Rate of convergence of the probability of ruin in the Cramér–Lundberg model to its diffusion approximation.
Insurance: Mathematics and Economics, 93:333–340, 2020.
5, 24, 42, 76

[46] Maxime C. Cohen, Ilan Lobel, and Renato Paes Leme.
Feature-based dynamic pricing.
In *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, page 817, New York, NY, USA, 2016. Association for Computing Machinery.
185, 186

[47] Rama Cont and Peter Tankov.
Financial modelling with jump processes.
Chapman & Hall/CRC financial mathematics series. Chapman & Hall/CRC, Boca Raton, FL, 2004.
107, 123

[48] Michael G. Crandall, Hitoshi Ishii, and Pierre-Louis Lions.
User's guide to viscosity solutions of second order partial differential equations.
Bull. Amer. Math. Soc. (N.S.), 27(1):1–67, 1992.
1, 19, 58

[49] Michael G. Crandall and Pierre-Louis Lions.
Viscosity solutions of Hamilton-Jacobi equations.
Transactions of the American Mathematical Society, 277(1):1–42, 1983.
1, 19

[50] Lorenzo Croissant, Marc Abeille, and Clément. Calauzènes.
Real-time optimisation for Online Learning in auctions.
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 2217–2226. PMLR, 13–18 Jul 2020.
181

[51] John C. Duchi, Peter L. Bartlett, and Martin J. Wainwright.
Randomized smoothing for stochastic optimization.
SIAM Journal on Optimization, 22(2):674–701, 2012.
186

[52] Omar El Euch, Masaaki Fukasawa, and Mathieu Rosenbaum.

The microstructural foundations of leverage effect and rough volatility.
Finance and Stochastics, 22(2):241–280, April 2018.
4, 13, 23, 33

[53] Christian Ewerhart.
Regular type distributions in Mechanism Design and ρ-concavity.
Economic Theory, 53(3):591–603, 2013.
189

[54] Joaquin Fernandez-Tapia, Olivier Guéant, and Jean-Michel Lasry.
Optimal Real-Time Bidding Strategies.
Applied Mathematics Research eXpress, September 2016.
4, 17, 23, 37, 41, 42, 64, 76, 105, 127

[55] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan.
Online Convex Optimization in the bandit setting: Gradient descent without a gradient.
In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, pages 385–394, USA, 2005. Society for Industrial and Applied Mathematics.
186

[56] Wendell H. Fleming and H. Mete Soner.
Controlled Markov processes and viscosity solutions.
Number 25 in Applications of mathematics. Springer, New York, 2nd ed edition, 2006.
1, 19

[57] Wendell. H. Fleming and Panagiotis E. Souganidis.
On the existence of value functions of two-player, zero-sum stochastic differential games.
Indiana University Mathematics Journal, 38(2):293–314, 1989.
5, 23, 42, 71, 76, 104

[58] Dean P. Foster and Rakesh Vohra.
Regret in the on-line decision problem.
Games and Economic Behavior, 29(1):7–35, October 1999.
11, 30

[59] A. T. Fuller.
Bibliography of Pontryagin's maximum principle.
Journal of Electronics and Control, 15(5):513–517, November 1963.
1, 19

[60] Aurélien Garivier and Eric Moulines.
On Upper-Confidence Bound Policies for Switching Bandit Problems.

In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 174–188, Berlin, Heidelberg, 2011. Springer.
186, 205

[61] David Gilbarg and Neil S. Trudinger.
Elliptic partial differential equations of second order.
Classics in mathematics. Springer, Berlin ; New York, 2nd ed., rev. 3rd printing edition, 2001.
116, 117, 118, 119, 120, 134

[62] Peter W. Glynn.
Chapter 4: Diffusion approximations.
In *Handbooks in Operations Research and Management Science*, volume 2, pages 145–198. Elsevier, 1990.
5, 24

[63] Sigurdur Hafstein.
Lyapunov functions for linear stochastic differential equations: BMI formulation of the conditions:.
In *Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics*, pages 147–155, Prague, Czech Republic, 2019. SCITEPRESS - Science and Technology Publications.
80

[64] Cédric Hartland, Sylvain Gelly, Nicolas Baskiotis, Olivier Teytaud, and Michèle Sebag.
Multi-armed Bandit, Dynamic Environments and Meta-Bandits.
working paper or preprint, November 2006.
209

[65] Il'dar Abdullovich Ibragimov.
On the composition of unimodal distributions.
Theory of Probability & Its Applications, 1(2):255–260, 1956.
211

[66] Jean Jacod and Albert Nikolayevich Shiryaev.
Limit theorems for stochastic processes, volume 288.
Springer Science & Business Media, 2013.
4, 23, 42, 76

[67] Thibault Jaisson and Mathieu Rosenbaum.
Limit theorems for nearly unstable Hawkes processes.
The Annals of Applied Probability, 25(2):600–631, April 2015.
Publisher: Institute of Mathematical Statistics.
4, 23

[68] Thomas Jaksch, Ronald Ortner, and Peter Auer.
Near-optimal regret bounds for Reinforcement Learning.
Journal of Machine Learning Research, 11(51):1563–1600, 2010.
2, 11, 13, 20, 31, 32, 124, 129, 178

[69] 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.
124

[70] Bernd Kawohl and Nikolai Kutev.
Strong maximum principle for semicontinuous viscosity solutions of nonlinear partial differential equations.
Archiv der Mathematik, 70:470–478, 1998.
104

[71] Rafail Khasminskii.
Stochastic Stability of Differential Equations, volume 66 of *Stochastic Modelling and Applied Probability*.
Springer, Berlin, Heidelberg, 2012.
166

[72] Fima C. Klebaner.
Introduction to Stochastic Calculus with applications.
ICP, Imperial College Press, London, 3. ed edition, 2012.
4, 22

[73] Robert Kleinberg and Tom Leighton.
The value of knowing a demand curve: Bounds on regret for online posted-price auctions.
In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE, 2003.
14, 34, 183, 185

[74] 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.
134

[75] Vijay Krishna.
Auction Theory.
Academic press, 2009.
80, 105, 189

[76] Harold J. Kushner.

Probability Methods for Approximations in Stochastic Control and for Elliptic Equations, volume 129 of *Mathematics in Science and Engineering*.
Elsevier, 1977.
134

[77] Harold J. Kushner.
Heavy Traffic Analysis of Controlled Queueing and Communication Networks, volume 47 of *Stochastic Modelling and Applied Probability*.
Springer New York, New York, NY, 2001.
5, 24

[78] 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.
5, 24, 67, 93, 94, 95, 124, 179

[79] Olga Aleksandrovna Ladyzhenskaya, Vsevolod Alexeyevich. Solonnikov, and Nina Nikolaevna Ural'tseva.
Linear and quasi-linear equations of parabolic type, volume 23.
American Mathematical Soc., 1988.
42, 54

[80] Olga Aleksandrovna Ladyzhenskaya and Nina Nikolaevna Ural'tseva.
Linear and quasilinear elliptic equations, volume 46 of *Mathematics in Science and Engineering*.
Elsevier, 1968.
134

[81] Peter Lancaster and Leiba Rodman.
Algebraic Riccati equations.
Clarendon press, 1995.
10, 29, 128

[82] Tor Lattimore and Csaba Szepesvári.
Bandit algorithms.
Cambridge University Press, 2020.
9, 123, 186, 205

[83] Michel Ledoux and Michel Talagrand.
Probability in Banach spaces, volume 23 of *Classics in Mathematics*.
Springer, Berlin, Heidelberg, 1991.
139

[84] Gary M. Lieberman.
Second order parabolic differential equations.
World Scientific, Singapore, River Edge (N.J.), 1996.

Reprint : 1998.
42, 50, 51, 53, 54

[85] 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.
124

[86] Pierre-Louis Lions.
Some recent results in the optimal control of diffusion processes.
In Kiyosi Itô, editor, *North-Holland Mathematical Library*, volume 32 of *Stochastic Analysis*, pages 333–366. Elsevier, January 1984.
5, 24

[87] Sean P. Meyn and Richard L. Tweedie.
Markov Chains and Stochastic Stability.
Cambridge University Press, Cambridge, 2nd ed edition, 2009.
OCLC: 667096030.
6, 25

[88] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al.
Human-level control through deep reinforcement learning.
nature, 518(7540):529–533, 2015.
2, 20

[89] Mehryar Mohri and Andrés Muñoz Medina.
Learning algorithms for second-price auctions with reserve.
The Journal of Machine Learning Research, 17(1):2632–2656, 2016.
14, 15, 34, 35, 183, 185, 191

[90] Jamie H. Morgenstern and Tim Roughgarden.
On the pseudo-dimension of nearly optimal auctions.
In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.
14, 34, 105, 183

[91] Eric Moulines and Francis Bach.
Non-asymptotic analysis of stochastic approximation algorithms for Machine Learning.
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.
15, 35, 187, 200, 202, 206

[92] Roger B. Myerson.
Optimal auction design.
Mathematics of operations research, 6(1):58–73, 1981.
14, 34, 65, 105, 183

[93] Thomas Nedelec, Clément Calauzènes, Noureddine El Karoui, and Vianney Perchet.
Learning in Repeated Auctions.
Foundations and Trends® in Machine Learning, 15(3):176–334, February 2022.
Now Publishers, Inc.
105

[94] Thomas Nedelec, Noureddine El Karoui, and Vianney Perchet.
Learning to bid in revenue-maximizing auctions.
In *Proceedings of the 36th International Conference on Machine Learning*, pages 4781–4789. PMLR, May 2019.
105

[95] Daniel B. Nelson.
ARCH models as diffusion approximations.
Journal of Econometrics, 45(1):7–38, July 1990.
5, 24

[96] Anna A. Obizhaeva and Jiang Wang.
Optimal trading strategy and supply/demand dynamics.
Journal of Financial Markets, 16(1):1–32, 2013.
123

[97] 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.
12, 13, 31, 32, 124, 132, 148, 149, 157

[98] Michael Ostrovsky and Michael Schwarz.
Reserve prices in internet advertising auctions: A field experiment.
In *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, pages 59–60, New York, NY, USA, 2011. Association for Computing Machinery.
65, 81, 105, 106, 189

[99] Renato Paes Leme, Martin Pál, and Sergei Vassilvitskii.
A field guide to personalized reserve prices.
In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 1093–1102, Republic and Canton of Geneva, CHE,

2016. International World Wide Web Conferences Steering Committee.
65, 183, 184, 185

[100] Vianney Perchet and Philippe Rigollet.
The multi-armed bandit problem with covariates.
The Annals of Statistics, 41(2):693–721, 04 2013.
11, 31, 209

[101] David Pollard.
Convergence of Stochastic Processes.
Springer Series in Statistics. Springer, New York, NY, 1984.
183

[102] Martin L. Puterman.
Markov Decision Processes: Discrete stochastic Dynamic Programming.
Wiley series in probability and statistics. Wiley-Interscience, Hoboken, NJ, 2005.
2, 6, 10, 20, 25, 29, 123

[103] Sheldon Ross.
Introduction to Stochastic Dynamic Programming.
Academic Press, New York, London, 1983.
10, 29

[104] Tim Roughgarden and Joshua R. Wang.
Minimizing Regret with Multiple Reserves.
In *Proceedings of the 2016 ACM Conference on Economics and Computation - EC '16*, volume 9, pages 601–616, 2016.
14, 34, 185, 186

[105] Tim Roughgarden and Joshua R Wang.
Minimizing regret with multiple reserves.
ACM Transactions on Economics and Computation (TEAC), 7(3):1–18, 2019.
105, 183

[106] Maja R. Rudolph, Joseph G. Ellis, and David M. Blei.
Objective variables for probabilistic revenue maximization in second-price auctions with reserve.
In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 1113–1122, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.
183, 185

[107] Daniel Russo and Benjamin Van Roy.
Eluder dimension and the sample complexity of optimistic exploration.

In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 2256–2264. Curran Associates, Inc., 2013.
12, 31, 32, 124, 130, 131, 148, 149, 150, 151, 153, 155, 157, 158, 170

[108] Adrien Saumard and Jon A. Wellner.
Log-concavity and strong log-concavity: A review.
Statistics Surveys, 8:45–114, 2014.
213, 214

[109] Leonard J. Savage.
The theory of statistical decision.
Journal of the American Statistical Association, 46(253):55–67, March 1951.
11, 30

[110] Weiran Shen, Sebastien Lahaie, and Renato Paes Leme.
Learning to clear the market.
In *International Conference on Machine Learning*, pages 5710–5718, 2019.
183, 185

[111] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al.
Mastering the game of go with deep neural networks and tree search.
nature, 529(7587):484–489, 2016.
2, 20

[112] Csaba Szepesvári.
Algorithms for Reinforcement Learning.
Number 4 in Synthesis lectures on artificial intelligence and machine learning. Morgan & Claypool Publishers, 2010.
2, 20, 123

[113] William R. Thompson.
On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.
Biometrika, 25(3/4):285–294, 1933.
11, 31

[114] Alexander Yurievich Veretennikov.
On strong solutions and explicit formulas for solutions of stochastic integral equations.
Mathematics of the USSR-Sbornik, 39(3):387–403, April 1981.
84

[115] W. Vickrey.

Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.
65

RÉSUMÉ

On considère la limite diffusive d'un problème de contrôle Markovien à sauts purs quelconque lorsque l'intensité de son processus de Poisson tend vers l'infini. On quantifie la vitesse de convergence en fonction de l'exposant de Hölder de la Hessienne du problème limite. On montre ensuite comment construire des termes de correction pour cette approximation, selon deux méthodologies différentes. Notre analyse couvre le problème à horizon fini, escompté et ergodique. Dans le cas ergodique, on quantifie l'erreur induite par l'utilisation de la politique de contrôle Markovienne construite à partir du schéma numérique de différences finies associé au problème diffusif limite. Cette approche permet une réduction très significative du coût de résolution numérique des problèmes de contrôle à sauts purs lorsque l'intensité des sauts est grande.
On s'attache ensuite au problème de l'incertitude dans les systèmes de contrôle, et on étend notre étude au contexte de l'apprentissage par renforcement en ligne. Dans le paradigme de l'optimisme devant l'incertain, on exploite le carcan de la dimension d'eluder pour gérer l'apprentissage et la limite diffusive pour résoudre approximativement le sous-problème de planification. Notre algorithme étend la théorie existante des problèmes discrets aux problèmes avec états et actions continus. L'utilisation d'outils issus de la théorie des processus stochastiques à temps continu nous permet également d'étudier une classe de coefficients plus générique que les travaux précédents.
Notre étude des systèmes à limite diffusive est motivée et illustrée par le problème d'enchérir dans une enchère séquentielle à haute fréquence contre un vendeur qui maximise son revenu sous contrainte d'utiliser une règle de mise à jour en temps réel.

MOTS CLÉS

Limite diffusive, contrôle stochastique, apprentissage par renforcement, enchères séquentielles, équations de Hamilton-Jacobi-Bellman, optimisme devant l'incertain.

ABSTRACT

We consider the diffusive limit of a generic pure-jump Markov control problem as the intensity of the driving Poisson process tends to infinity. We quantify the convergence speed in terms of the Hölder exponent of the Hessian of the limit problem. We then explain how correction terms can be constructed for this limit approximation, according to two different methodologies. Our analysis covers the finite-horizon, discounted, and ergodic problems. In the ergodic case, we quantify the error induced by the use of the Markov control policy constructed from the numerical finite difference scheme associated with the limit diffusive problem. This approach permits a very significant reduction in numerical resolution cost for pure-jump control problems when the intensity of jumps is large.
Considering the problem of uncertainty in control systems, we study these high-frequency pure-jump problems in the context of online Reinforcement Learning. Using the Optimism in the Face of Uncertainty paradigm, we leverage the eluder dimension framework for learning and lazy updates, as well as the diffusive limit for approximate resolution of the planning sub-problem. This extends existing theory from discrete processes to continuous states and actions. The use of tools for continuous-time stochastic processes also permits us to study a more generic class of coefficients than previous work.
Our study of diffusion limit systems is motivated and illustrated by the bidding problem in a high-frequency online auction against a seller who maximises its revenue under the constraints of using a real-time update rule.

KEYWORDS

Diffusive limit, Stochastic Control, Reinforcement Learning, online auctions, Hamilton-Jacobi-Bellman equations, Optimism in the Face of Uncertainty.