Mean field reinforcement learning : an optimal control perspective

Athanasios Vasileiadis

UNIVERSITÉ CÔTE D'AZUR | ÉCOLE DOCTORALE SCIENCES FONDAMENTALES ET APPLIQUÉES

THÈSE DE DOCTORAT

APPRENTISSAGE PAR RENFORCEMENT À CHAMP MOYEN

Une perspective de contrôle optimal

Athanasios VASILEIADIS

Laboratoire de Mathématiques J.A. Dieudonné (LJAD)

Présentée en vue de l'obtention du grade de docteur en Mathématiques **d'Université Côte d'Azur**

Dirigée par : François DELARUE

Soutenue le : vendredi 23 février 2024

Devant le jury, composé de :
Noufel FRIKHA, Professeur, Univ. Panthéon Sorbonne
Huyên PHAM, Professeur, Univ. Paris Cité
Nicole BÄUERLE, Professeure, Karlsruhe Institute of Technology
Mireille BOSSY, Directrice de Recherche, INRIA Sophia-Antipolis
Zhenjie REN, Maître de Conférences, HDR, Univ. Paris-Dauphine
Patricia REYNAUD-BOURET, Directrice de Recherche, CNRS Univ. Côte d'Azur
Francisco J. SILVA ÁLVAREZ, Maître de Conférences, HDR, Univ. de Limoges

UNIVERSITÉ CÔTE D'AZUR | INSTITUT INTERDISCIPLINAIRE D'INTELLIGENCE ARTIFICIELLE 3IA CÔTE D'AZUR | FRANCE 2030

Apprentissage par renforcement à champ moyen

Une perspective de contrôle optimal

Mean Field Reinforcement Learning

An optimal control perspective

Athanasios Vasileiadis

Jury:

Rapporteurs

- **Noufel FRIKHA**, Professeur, Univ. Panthéon Sorbonne
- **Huyên PHAM**, Professeur, Univ. Paris Cité

Examinateurs

- **Nicole BÄUERLE**, Professeure, Karlsruhe Institute of Technology
- **Mireille BOSSY**, Directrice de Recherche, INRIA Sophia-Antipolis
- **Zhenjie REN**, Maître de Conférences, HDR, Univ. Paris-Dauphine
- **Patricia REYNAUD-BOURET**, Directrice de Recherche, CNRS Univ. Côte d'Azur
- **Francisco J. SILVA-ÁLVAREZ**, Maître de Conférences, HDR, Univ. de Limoges

Directeur de thèse

- **François DELARUE** Professeur, Univ. Côte d'Azur

Résumé

Dans cette thèse, nous étudions l'apprentissage par renforcement à champ moyen (MFRL) avec une perspective de contrôle optimal. Nous nous intéressons au cas où un grand nombre d'agents majoritairement symétriques interagissent via leurs états pour atteindre un objectif, en optimisant généralement un critère. Les agents peuvent être en concurrence ou en collaboration les uns avec les autres.

Dans le premier chapitre de la thèse, nous proposons une introduction générale à l'apprentissage par renforcement à champ moyen. Nous introduisons quelques idées fondamentales de l'apprentissage par renforcement, des jeux à champ moyen et du contrôle à champ moyen qui sont des prérequis pour la suite. Nous avons également besoin des notions d'équilibre de Nash, de fonction valeur et de limite champ moyen, que nous introduisons avec quelques techniques pour résoudre des modèles sans incertitude et sans apprentissage. Ensuite, nous motivons l'apprentissage multi-agent par renforcement dans la limite champ moyen, à la fois compétitif et collaboratif. Pour évaluer la portée de nos résultats, nous mettons en évidence les difficultés de cette voie de recherche. Enfin, une vue intuitive de nos principales contributions est donnée.

Le but du deuxième chapitre de la thèse est de démontrer que le bruit commun peut servir comme un bruit d'exploration pour apprendre la solution d'un jeu à champ moyen. Ce concept est illustré ici à travers un modèle linéaire quadratique, pour lequel une forme appropriée de bruit commun a déjà été prouvée d'être capable à restaurer l'existence et l'unicité. Nous allons encore plus loin et prouvons que la même forme du bruit commun peut forcer la convergence de l'algorithme d'apprentissage appelé «fictitious play», sans aucune structure potentielle ou monotone supplémentaire. Plusieurs exemples numériques sont fournis pour soutenir notre analyse théorique.

L'objet du troisième chapitre est de traiter une méthode Q-learning pour les processus de décision Markoviens, définis sur des espaces continus. L'algorithme s'appuie sur des idées de régression par noyau, qui ont été introduites à l'origine dans la littérature, mais nous améliorons l'analyse en rendant les arguments rigoureux et quantitatifs. Le cœur de notre travail est d'analyser la convergence de l'algorithme et, en particulier, de clarifier l'impact des propriétés de non-dégénérescence des transitions des processus de décision Markoviens sur le taux de convergence. À titre de résultat clé, nous parvenons à surmonter la malédiction de la dimension en supposant que la fonction valeur-action est suffisamment régulière (par rapport à la dimension effective). Comme application, nous illustrons comment cette approche s'applique à l'apprentissage par renforcement pour des problèmes de contrôle à champ moyen (ou processus de décision Markoviens à champ moyen) définis sur les espaces des états finis. Dans ce contexte, les transitions sont en effet non-dégénérées en présence d'une forme appropriée de bruit commun. En particulier, très similaire à l'analyse réalisée au chapitre précédent, le bruit commun peut servir comme un bruit d'exploration supplémentaire dans l'apprentissage à champ moyen pour des modèles sans bruit commun.

Nous concluons la thèse par deux problèmes ouverts à poursuivre en forme de pistes de travail, dans le même esprit que les résultats présentés jusqu'à ici.

Mots clés: jeux à champ moyen, contrôle à champ moyen, processus de décision Markoviens, équilibre de Nash, Q-learning, bruit d'exploration, bruit commun

Abstract

In this thesis we study Mean Field Reinforcement Learning via an optimal control perspective. We are interested in cases where a large number of mostly symmetrical agents interact through their states in order to achieve some objective, mostly optimizing some criterion. The agents can compete or collaborate with each other.

In the first chapter of the thesis, we offer a general introduction to mean field reinforcement learning. We introduce some basic ideas of Reinforcement Learning, Mean Field Games and Mean Field Control that are prerequisites for the rest. We also need the notions of Nash equilibrium, value function and mean field limit, that we introduce along with some techniques for solving models when there is no uncertainty and no learning. Next, we motivate Multi Agent Reinforcement Learning in the mean field limit both competitive and collaborative. To put our results in perspective we highlight the difficulties of this research path. Finally an intuitive overview of our main contributions is given.

The goal of this second chapter of the thesis, is to demonstrate that common noise may serve as an exploration noise for learning the solution of a mean field game. This concept is here exemplified through a toy linear-quadratic model, for which a suitable form of common noise has already been proven to restore existence and uniqueness. We here go one step further and prove that the same form of common noise may force the convergence of the learning algorithm called 'fictitious play', without any further potential or monotone structure. Several numerical examples are provided in order to support our theoretical analysis.

The purpose of the third chapter is to address a Q-learning method for Markov decision processes defined over continuous spaces. The algorithm relies upon ideas for kernel regression, that have originally been introduced in the literature but we refine the analysis making the arguments rigorous and quantitative. The very thrust of our work is to analyse the convergence of the algorithm and in particular to clarify the impact of the non-degeneracy properties of the transitions of the Markov decision processes onto the rate of convergence. As a key result, we succeed to overcome the curse of dimensionality by assuming that the action value function is regular enough (with respect to the effective dimension). As an application, we illustrate how this approach applies to reinforcement learning for mean field control problems (or mean field Markov decision processes) defined on finite state spaces. In this setting, the transitions are indeed non-degenerate in presence of a convenient form of common noise. In particular, very similar to the analysis achieved in the previous chapter, the common noise can serve as an additional exploration noise in mean field learning for models without common noise.

We conclude the thesis, with two open problems for future investigation in the same spirit as the results presented so far.

Key words: Mean Field Games, Mean Field Control, Markov Decision Process, Nash Equilibrium, Q-learning, Exploration Noise, Common Noise

In memory of Fani Aggelidou

Ιθάκη

Κ.Π ΚΑΒΑΦΗ

Σα βγεις στον πηγαιμό για την Ιθάκη,
να εύχεσαι να ΄ναι μακρύς ο δρόμος,
γεμάτος περιπέτειες, γεμάτος γνώσεις.
Τους Λαιστρυγόνας και τους Κύκλωπας,
τον θυμωμένο Ποσειδώνα μη φοβάσαι,
τέτοια στον δρόμο σου ποτέ σου δεν θα βρεις,
αν μέν΄ η σκέψις σου υψηλή, αν εκλεκτή
συγκίνησις το πνεύμα και το σώμα σου αγγίζει.
Τους Λαιστρυγόνας και τους Κύκλωπας,
τον άγριο Ποσειδώνα δεν θα συναντήσεις,
αν δεν τους κουβανείς μες στην ψυχή σου,
αν η ψυχή σου δεν τους στήνει εμπρός σου.

Να εύχεσαι να ΄ναι μακρύς ο δρόμος.
Πολλά τα καλοκαιρινά πρωιά να είναι
που με τι ευχαρίστηση, με τι χαρά
θα μπαίνεις σε λιμένας πρωτοϊδωμένους·
να σταματήσεις σ΄ εμπορεία Φοινικικά,
και τες καλές πραγμάτειες ν΄ αποκτήσεις,
σεντέφια και κοράλλια, κεχριμπάρια κι έβενους,
και ηδονικά μυρωδικά κάθε λογής,
όσο μπορείς πιο άφθονα ηδονικά μυρωδικά·
σε πόλεις Αιγυπτιακές πολλές να πας,
να μάθεις και να μάθεις απ΄ τους σπουδασμένους.

Πάντα στον νου σου να ΄χεις την Ιθάκη.
Το φθάσιμον εκεί είν΄ ο προορισμός σου.
Αλλά μη βιάζεις το ταξίδι διόλου.
Καλύτερα χρόνια πολλά να διαρκέσει·
και γέρος πια ν΄ αράξεις στο νησί,
πλούσιος με όσα κέρδισες στον δρόμο,
μη προσδοκώντας πλούτη να σε δώσει η Ιθάκη.

Η Ιθάκη σ΄ έδωσε τ΄ ωραίο ταξίδι.
Χωρίς αυτήν δεν θα ΄βγαινες στον δρόμο.
Άλλα δεν έχει να σε δώσει πια.
Κι αν πτωχική την βρεις, η Ιθάκη δεν σε γέλασε.
Έτσι σοφός που έγινες, με τόση πείρα,
ήδη θα το κατάλαβες οι Ιθάκες τι σημαίνουν.

Ithaque

C. P. CAVAFY
traduction de Marguerite Yourcenar

Quand tu partiras pour Ithaque,
souhaite que le chemin soit long,
riche en péripéties et en expériences.
Ne crains ni les Lestrygons, ni les Cyclopes,
ni la colère de Neptune.
Tu ne verras rien de pareil sur ta route si tes pensées restent hautes,
si ton corps et ton âme ne se laissent effleurer
que par des émotions sans bassesse.
Tu ne rencontreras ni les Lestrygons, ni les Cyclopes,
ni le farouche Neptune,
si tu ne les portes pas en toi-même,
si ton cœur ne les dresse pas devant toi.

Souhaite que le chemin soit long,
que nombreux soient les matins d'été,
où avec quelles délices, quelle joie, tu pénètreras
dans des ports vus pour la première fois.
Fais escale à des comptoirs phéniciens,
et acquiers de belles marchandises
nacre et corail, ambre et ébène,
et mille sortes d'entêtants parfums.
Acquiers le plus possible de ces entêtants parfums.
Visite de nombreuses cités égyptiennes,
et instruis-toi avidement auprès de leurs sages.

Garde sans cesse Ithaque présente à ton esprit.
Ton but final est d'y parvenir,
mais n'écourte pas ton voyage
mieux vaut qu'il dure de longues années,
et que tu abordes enfin dans ton île aux jours de ta vieillesse,
riche de tout ce que tu as gagné en chemin,
sans attendre qu'Ithaque t'enrichisse.

Ithaque t'a donné le beau voyage
sans elle, tu ne te serais pas mis en route.
Elle n'a plus rien d'autre à te donner.

Même si tu la trouves pauvre, Ithaque ne t'a pas trompé.
Sage comme tu l'es devenu à la suite de tant d'expériences,
tu as enfin compris ce que signifient les Ithaques.

Ithaka

C. P. CAVAFY
translated by Edmund Keeley

As you set out for Ithaka
hope your road is a long one,
full of adventure, full of discovery.
Laistrygonians, Cyclops,
angry Poseidon—don't be afraid of them
you'll never find things like that on your way
as long as you keep your thoughts raised high,
as long as a rare excitement
stirs your spirit and your body.
Laistrygonians, Cyclops,
wild Poseidon—you won't encounter them
unless you bring them along inside your soul,
unless your soul sets them up in front of you.

Hope your road is a long one.
May there be many summer mornings when,
with what pleasure, what joy,
you enter harbors you're seeing for the first time
may you stop at Phoenician trading stations
to buy fine things,
mother of pearl and coral, amber and ebony,
sensual perfume of every kind—
as many sensual perfumes as you can
and may you visit many Egyptian cities
to learn and go on learning from their scholars.

Keep Ithaka always in your mind.
Arriving there is what you're destined for.
But don't hurry the journey at all.
Better if it lasts for years,
so you're old by the time you reach the island,
wealthy with all you've gained on the way,
not expecting Ithaka to make you rich.

Ithaka gave you the marvelous journey.
Without her you wouldn't have set out.
She has nothing left to give you now.

And if you find her poor, Ithaka won't have fooled you.
Wise as you will have become, so full of experience,
you'll have understood by then what these Ithakas mean

Acknowledgements

This thesis would have never existed without my advisor François Delarue. This is hardly any news, but in our case this simple fact cannot be overstated! I am infinitely grateful for all the effort he has put throughout the years, the commitment and the dedication. He taught me maths but also he taught me ethos. His energy and love for "serious" math are inexhaustible and always a source of motivation for me to surpass my limits and become better. His habit of relentlessly questioning me has taught me scientific rigor and even thought in the beginning it was a bit scary finally seasoned me and developed my overall capacity in a far broader scope than merely mathematically or academically.

I would like to personally thank Noufel Frikha, and Huyên PHAM for taking the time to read my admittedly long thesis and writing the reports and Nicole Bäuerle, Mireille Bossy, Zhenjie Ren, Patricia Reynaud-Bouret and Francisco J. Silva for serving as members of my jury.

Next, within the community of MFGs I would like to thank my friends, Alekos Cecchin and Mathieu Laurière. Alekos has been a steady friend since day one of my journey in Nice. I always enjoy our inside jokes and appreciate the simplicity (without being simple) of discussing math with him. He has always been there for me and I look forward to our annual meetings. Mathieu is simply amazing, he is the embodiment of kindness, always curious about open problems and new advances, always on the move to explore new horizons. He is very inspiring and at the same time heartwarming. I will always keep the book he gifted me as a token of our friendship.

I would like to equally thank Gonçalo Dos Reis, for the last 6 months of collaboration. His relaxed style made me instantly feel like home in Edinburgh. I had an amazing time there and I look forward to visit again. It is always a pleasure to work with him.

My gratitude goes also to Cédric Bernardin and Patricia Reynaud-Bouret for encouraging me with pretty much everything from the courses I was teaching to speaking french and my cooking!

Special thanks to Sylvain Rubenthaler for the 3 years of teaching together Time Series in the masters, I appreciate the freedom and support he gave me to develop material for the students especially when I didn't believe I had something of worth to say.

I am indebted to the various students I had throughout the years, whose positive feedback helped me combat my imposter's syndrome and find self-worth and meaning into my daily job.

Last but not least, from the academic world, I would like to deeply thank Vassilis Papanicolaou, my lifelong mentor, advisor and I would say friend. Since the very first time we met, on my very first day of the first graduate course in Polythechnio I admire his character, his mathematical knowledge, his teaching style, his stories and his kindness. But most of all, I would like to thank him for his unconditional love and support that have really transformed my life. It was his trust, patience and

support that initiated me in the world of professional math and showed me the pleasure behind it.

Outside academia, I would like to wholeheartedly express my gratitude to my friend Apostolos Karagiannis for his patience and stoicism, listening to my daily complains or explanations of math, he had no idea. He has been an invaluable companion along my journey, pumping me up everyday with his faith and trust. He has been always the most positive part of my day.

I would like to thank my swim trainer Eleni Nikolakopoulou for teaching me perseverance, to never abandon the race and finally to not be afraid to make mistakes (in some cases to even intend), without her invaluable lessons in and out of the swimming pool this thesis would have never been finished.

Next, I would like to express my gratitude towards a very large group of friends whose contribution has been fundamental not only for supporting me throughout the thesis but mainly for teaching me how to live, without them I wouldn't have been here today and since the list is quite long, to attribute special thanks to each one, I would definitely have to write another thesis just for them (which I might do in another context), I hope they excuse me. I cherish every moment passed with Melina Kefalinou and Iakovos Kaparis, Charalampos Lampadaris, Maya Mina, Achileas Koufalexis, Ioanna and Kyriaki Stefa, Eirini Vandorou, Konstantina Diamanti, Faidra Drakonaki, Ioannis Oikonomakos, Melissa Sanabria Rossas, Guillaume Barnoin, Riccardo Di Dio, Clement and Jung, Boris Schminke and Sveta, Mehdi Zaïdi, Ryan Cotsakis, Dimitris Zekakos Xipolias, Dimitris Rountos, Thomas and Zeta Oikonomakou, Guilia Mezadri, Kleopatra Rizou, Sofia Athanasiou, Grigoris Liaskas, Panagiotis Kakalmanos and Nina Kazakou, Alejandro Orozco, Maria Protopsali, Pavlos and Evi Lampadari, Alexandros and Maria Petreli, Eleftheria Tsitsa and Charalampos Posonidis, Sotiris Mpourmpos, Giorgos Ritsos, Giorgos and Dimitris Asimakakis, Sensi Mattia and Alexandra Polymenopoulou.

Last but not least, my biggest, "bestest" ευχατιστώ goes to my family and Valia for the love and affection throughout my life. There are no words to describe my gratitude.

Preface

This thesis tries to target mathematical problems that are motivated by Reinforcement Learning but also on the opposite side to offer mathematical insights about solutions to Reinforcement Learning problems, especially in the Multi Agent case. Even though mean field models are all around us, people interact in a "mean field" way every day and examples can be found in other sciences rather easily, the mathematical treatment of the subject is quite advanced and technical and thus not accessible for the non-specialized readers. For this reason, while the contributions themselves might be quite technical, I provide a long, informative and non-technical introduction, focusing on the intuition behind the models and the "whys" rather than the "hows". My intension is to provide a useful document, reference for specialized and non-specialized readers.

In the introduction I try to give an overview of some main ideas, their connections as well as motivation for the research line presented in this thesis. To streamline the text, we start with problem definition, move on to description of the main tools available for solutions and conclude with the challenges and limitations that we face. Throughout this line, I look forward to highlight the answers that our contributions bring as well as the new questions they generate.

I hope, through reading this thesis the non-specialised reader would get an appreciation for the techniques employed and some perspective about potential applications and synergies of mathematics and artificial intelligence. For the experts we have reserved specific introductions in each chapter that makes them self-contained and as independent as possible.

Last, in the final chapter of the thesis I provide two research questions that I find interesting, along with references and some strategies. While the methods are not new for standard stochastic control problems and reinforcement learning, the adaptation to the space of probability measures is and thus opens new questions and requires new tools to be developed.

Contents

Chapter 1

Introduction

In this thesis we are going to discuss about mathematics related to Artificial Intelligence (AI) and more specifically to Mean Field Reinforcement Learning (MF-RL). We focus our attention in situations where there is a plethora of identical/ symmetrical agents that interact, either to compete in a game or collaborate to solve collectively a problem, we refer to the first case as a Mean Field Game and to the second one as a Mean Field Control.

In the next sections of this introductory chapter we will provide more details for the non specialist reader as a non technical introduction to the domain of our research. Our goal is to explain the great importance of MF-RL among AI paradigms especially for our modern world and elaborate on the main challenges that this line of research faces as we propose novel solutions and deepen our understanding of AI.

1.1 What are we talking about in this thesis?

1.1.1 Deus Ex Machina

In ancient greek drama (called *tragodia*), whenever there was an impasse in the plot it was a common playwriting technique to make a god appearing out of nowhere in order to give a solution to the problem. Nowadays, Artificial Intelligence seems to preform a similar task offering solutions to problems that previously were consider impossible. Consequently, there has been enormous interest for AI followed by an explosion of research articles, books, conferences. To satisfy interest from the general public, educational/review articles appear even on daily newspapers.

Here however, we do not attempt a general introduction to AI and refrain from entering into a discussion about defining what it is and what is not. The reader can think of all the different approaches and algorithms as generic that can be applied to specific problems and in general dealing with an ideal agent/robot that performs a task.

In our version of **Reinforcement Learning**, the aforementioned ideal agent/robot/computer tries to achieve a goal of either minimization of a cost or maximization of a reward while interacting with an unknown environment, key notions are the state of the agent and the action she chooses. Usually the interaction of the agent with the environment reveals her state (which is unknown and thus the greater difference with the other major paradigms of Machine Learning) and the

instantaneous reward or cost see Figure 1.1. One additional key notion is a policy that essentially is a function that tells the agent what to do on each given state.

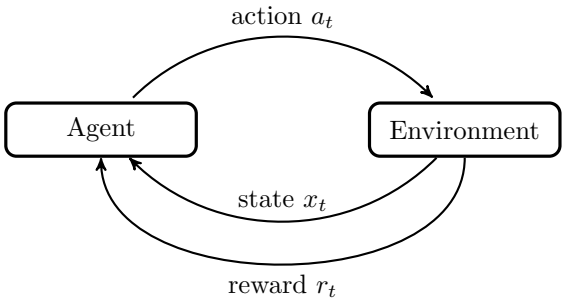


Figure 1.1: Standard Reinforcement Learning paradigm

We could roughly separate the development of RL in two 3 periods, even though the boundaries are not so distinct.

1. **Early Work (1950s-1960s)**: The foundations of reinforcement learning can be traced back to the work of Richard Bellman, who developed dynamic programming techniques in the 1950s. His work laid the mathematical foundations for solving sequential decision-making problems. However, the computational limitations of the time made it challenging to apply these methods to complex real-world problems.

2. **Temporal Difference Learning (1980s)**: In the 1980s, researchers such as Christopher Watkins, Andrew Barton, Richard Sutton, Dimitris Bertsekas and John Tsitsiklis made significant contributions to reinforcement learning leading to the introduction of the concept of temporal difference (TD) learning, which became a fundamental building block for RL algorithms.

3. **Deep Reinforcement Learning (2010s)**: The field of reinforcement learning gained significant attention and progress with the advent of deep learning techniques. Deep reinforcement learning (DRL) involves training neural networks as efficient functions approximations. This led to groundbreaking achievements, such as AlphaGo by DeepMind, which demonstrated superhuman performance in the game of Go.

Although impressive progress has been achieved in the last 20-30 years, still many challenges remain. We could articulate 3 major types:

1. **Curse of dimensionality**: as it was called by Bellman himself. When the state space and or the action space is large e.g. in chess the number of legal positions is 10^{43} and for the legal actions based on a position the average is $30 - 35$, it's clear that just keeping everything in memory and trying to test all possible paths for the evolution of the game is impossible and even if it was, it would have been highly inefficient.

2. **Sample efficiency**: RL algorithms often need a large number of interactions with the environment to learn effective policies, which can be impractical or costly in real-world scenarios.

When we use RL to solve games that can be simulated with minimum to no cost, sample efficiency is not really and issue but in robotic control like drones or autonomous vehicles samples can be scares or too costly to obtain.

3. **Generalization**: In this category we group various types of challenges. In the case of multi-agent systems we need to generalize the knowledge of the system and find policies that could be applied by the whole population. In the case of multi-environment systems we need to learn in one environment and generalize to others. Even in single environment and agent cases we might need to scale our solution or incorporate new elements for example in financial market models.

1.1.2 Reinforcement Learning as a Stochastic Control problem

During the early years of development RL was known mostly as stochastic optimal control and the term appeared later by Richard S. Sutton in the late 70s. In fact one of the most celebrated techniques that are being used to solve RL problems and inspired lot's of other methods, dynamic programming principle is due to R. Bellman from the 50s.

While stochastic control mainly aims to optimize a function given some underlying stochastic dynamics to achieve a specific target, when we control part of the drift and of volatility in the diffusion we need to take into account that our actions will not only affect how we steer the system but also the intensity of the future perturbations and this plays a role in the techniques employed to solve the problem. However, in the case of a reinforcement learning problem where we don't known the whole state space (or part of it) or the dynamics, or the cost/benefit function, the situation is quite different because we need to explore in order to get an optimal solution.

Our goal in this section is to provide an introduction to RL as a stochastic control problem and draw some interpretations that will be useful throughout the thesis. For more details on stochastic control we refer to the classical book of Jong and Zhou [119]. Dynamic Programming Principle is a fundamental technique we will use repetitively throughout the introduction and the thesis.

For a standard stochastic control problem we are given some dynamics for the state process X_t that is being controlled by a_t,

$$dX_t = b(X_t, a_t)dt + \sigma(X_t, a_t)dW_t, \quad X_0 = x,$$

where x is a random variable. Our target is to minimize the total discounted cost which we can represent in the form of a state value function over the set of all admissible controls[1],

$$V(t,x) = \inf_{a \in \mathcal{A}_{adm}} \mathbb{E}\Big[\int_t^T e^{-\gamma(s-t)} c(X_s, a_s)ds \mid X_t = x\Big]$$

with $\gamma \in [0,1]$ the discount rate, we suppose also that the terminal cost is zero.

[1] we call admissible controls all the controls that either respect some constrain or simply allow us to reach our target. Clearly the set can depend on the x

When the model is known, i.e. b, σ, c are known functions a well known approach to solve the problem is to use Bellman's Dynamic Programming Principle (DPP)

$$V(t,x) = \inf_{a \in \mathcal{A}_{adm}} \mathbb{E}\Big[\int_t^{t+h} e^{-\gamma s} c(X_s, a_s) ds + e^{\gamma h} V(t+h, X_{t+h}) \;\Big|\; X_t = x\Big] \quad \forall\, 0 \leqslant t \leqslant s \leqslant t+h \leqslant T,\ x \in \mathbb{R}$$
$$V(T,x) = 0 \quad \forall\, x \in \mathbb{R}$$

roughly DPP says that we can "construct" the value function by playing optimally step by step in a **backwards** manner. In each step we play optimally over $[t, t+h]$ and discount the value of $[t+h, T]$. Using Itô's formula on the dynamic programming equation we can derive a PDE that helps us "design" optimal controls, the famous Hamilton Jacobi Bellman Equation,

$$\gamma V(t,x) = \partial_t V(t,x) + \min_{a \in \mathcal{A}_{adm}} \Big\{ b(x,a)\partial_x V(t,x) + \frac{1}{2}\sigma^2(x,a)\partial^2_{xx} V(t,x) + c(x,a) \Big\},$$
$$V(T,x) = 0 \quad \forall\, x \in \mathbb{R}$$

which holds for all x, t, a in a backwards sense.

In this approach, the optimal control can be "synthesised" using the value function from the HJB equation as follows:

1. By solving HJB, we know for all pairs (x,t) the value of the control which minimizes the expression above. That is a deterministic function $a = \widetilde{a}(t,x)$ that we call a *feedback control* or *policy* as they are commonly called in RL.

2. Next, we solve the dynamics of X_t when she is controlled by $\widetilde{a}(t,x)$

$$dX_t^* = b\big(X_t^*, \widetilde{a}(t, X_t^*)\big)dt + \sigma\big(X_t^*, \widetilde{a}(t, X_t^*)\big)dW_t, \quad X_0^* = x.$$

   then $\widetilde{a}^*(t, X_t^*)$ is the optimal feedback control or optimal policy.

Unfortunately, when we don't know the model i.e. the functions b, σ, c, this procedure cannot work and we have to rely on trials to learn the values of the actions.

A central notion in this approach is the *exploratory policy*, which allows us to explore the states or the actions and thus gather samples for learning. In fact, at this point we are in a crossroads regarding the type of learning we want to achieve, we can either focus on learning the functions involved in the stochastic control problem and then use the standard approach described so far, which is proposed as a **model based** method where we form a model for the dynamics and for the instantaneous cost and update it accordingly. The first part of thesis is going to be in this direction and the introduction of the corresponding chapter will provide much more details.

One of the main critiques in the model based approach is **model bias** i.e. the assumption that the learned dynamics sufficiently and accurately approximates the real dynamics. An optimal control learned under dynamics b, σ and cost c is not guaranteed to be optimal under even mildly modified $b^\varepsilon, \sigma^\varepsilon, c^\varepsilon$. In simulated environments like games it's easy to overcome this restriction since there is stability in the simulator of the game e.g. the legal positions in chess are always the same.

However in real life situations when we want to use RL, the dynamics might change as structural changes happen in the environment and thus we need to incentivise the algorithm to continue to

adapt in the environment. This goes back to the **generalisation** issue described in the previous section. For RL to escape the "lab" and move towards real world situations there is still progress to be done in this direction.

Nevertheless, we have tools to partially answer this challenge in a satisfactory way for quite a vast range of examples, we can incorporate **model uncertainty**. There are several ways to do this and to lighten the exposition and keep as close as possible the reference to stochastic control formulation we choose to follow [112, 93]. Essentially, we consider uncertainty in the dynamics or the cost as a measure on the space of continuous functions. This is consistent with a weak formulation of the stochastic control problem. Learning is translated as changing the measure over the dynamics and the cost. This yields relaxed controls which are essentially distributions of actions [2].

Another way to interpret the use of relaxed controls for exploration is if we consider each action a_t^i as an trial for exploration that yields a state X_t^i, and in order to learn the evolution of the states or the cost we need to average a large number of trials. To explain further this idea consider that we are given a distribution U_t with a density $u_t(a)$ over the space of admissible controls $\mathcal{A}_{adm}$, then we sample $\{a_t^i\}_{i=1}^N \sim U_t$ with each control a^i being standard. Consequently we can define its controlled state X^i that should evolve according to an SDE similar to the first one for the time being $[0, T]$ and assuming that the Brownian Motions are independent of the state and action process for all i

$$dX_t^i = b(X_t^i, a_t^i)dt + \sigma(X_t^i, a_t^i)dW_t^i, \quad X_0^i = x, \quad \forall\, i = 1, ..., N,$$

and costs

$$\int_0^T e^{-\gamma t} c(X_t^i, a_t^i)dt \quad \forall\, i = 1, ..., N,$$

then to estimate the cost or the dynamics all we have to do is average the N samples

$$\frac{1}{N}\sum_{i=0}^{N} X_t^i = x + \frac{1}{N}\sum_{i=1}^{N}\int_0^t b(X_t^i, a_t^i)\, dt + \frac{1}{N}\sum_{i=1}^{N}\int_0^t \sigma(X_t^i, a_t^i)\, dW_t^i$$

and when $N \to \infty$

$$\mathbb{E}[X_t^u] = x + \mathbb{E}\Big[\int_0^t \int_{\mathcal{A}_{adm}} b(X_t^u, a)u_t(a)da\, dt\Big]$$

and similarly for the cost

$$\frac{1}{N}\sum_{i=1}^{N}\int_0^T e^{-\gamma t} c(X_t^i, a_t^i)dt \quad \underset{N\to\infty}{\to} \quad \mathbb{E}\Big[\int_0^T \int_{\mathcal{A}_{adm}} e^{-\gamma t} c(X_t^u, a)u_t(a)da\, dt\Big].$$

This prompts us to define the *exploratory formulation*

$$dX_t^u = \int_{\mathcal{A}_{adm}} b(X_t^u, a)u_t(a)da\, dt + \int_{\mathcal{A}_{adm}} \sigma(X_t^u, a)u_t(a)da\, dW_t, \quad X_0^u = x, \quad u \in \mathcal{P}(\mathcal{A}_{adm}),$$

[2] In game theory or RL they are very often called mixed strategies

and the corresponding value function

$$V(t,x) = \inf_{a \in \mathcal{A}_{adm}} \mathbb{E}\Big[\int_t^T \int_{\mathcal{A}_{adm}} e^{-\gamma(s-t)} c(X_s^u, a) u_t(a) da \, ds \mid X^u = x\Big].$$

Then we can follow with our strategy from classical control dealing with the HJB equation and defining the optimal controls. In [112] the authors solve the problem while using an entropy penalization to encourage exploration and in [93] they treat the unregularized problem.

The other option that we have instead of building a model, is to directly use observations of states, actions and rewards to learn the value function. This method is called **model free** and it is one of the most developed in the RL community so far. In this category falls the celebrated *Q-Learning* algorithm for which the second part of the thesis is devoted, where we present some new results that deepen our understanding for the properties of the convergence of the algorithm.

1.1.3 Mean Field Games

Mean Field Games started with the seminal works of Lasry Lions [82] and independently from Huang Cains Malhamé [72] in 2006 and ever since they have known tremendous success and popularity. Even though MFGs can describe a large variety of phenomena from physics to economics and biology, for the sake of clarity we will mostly focus on the game aspect, as it is the most relevant for what is going to follow.

In the first part of the thesis we will deal with the problem of learning equilibria of MFGs in the presence of uncertainty. We imagine a large number of symmetrical players playing a game, i.e. each agent optimizing a functional while competing with the rest of the population, this form of "competition" can be described as interactions either through states and/or actions. Usually we call such cases finite population games. In classical game theory we mostly study a game between 2 players (or a handful) that we try to gradually extend to several, a procedure known to be hard, especially in cases of games with continuous spaces. MFGs comes to tackle exactly that difficulty, instead of solving the game player by player, we look at the limit situation where the number of agents goes to infinity and a representative agent appears, thus resorting to aggregate quantities and population distribution instead of individually coupled optimization problems.

Since this part of the introduction is intended for the non-specialized we will keep the discussion mostly informal, focusing on the intuition and leave precise statements for the main body of the thesis, also in all stated problems the reader should always consider the most smooth case, assuming as smoothness and boundedness needed for the problems to be wellposed.

Individual agent's problem: fully general case

Suppose that we have N agents that play a dynamic continuous game i.e. a strategic game in continuous time with continuous state and action sets. Let us denote X_t^i the state of the individual agent and a_t^i her action, both taking values in compact subsets of the reals, and with the initial state of each agent x^i to be an $L^2(\mathbb{R})$ random variable and the evolution in time to be described by an SDE, where W_t^i is each players individual noise (a Brownian Motion) and W_t^0 is a systemic noise that affects all the players (again a Brownian Motion). Each agent has a running cost function f^i and a terminal one g^i both of them depend on the state and action of the player i but also on

the states and actions of the rest of the players $\{X_t^{-i}\}, \{a_t^{-i}\}$ (the notation here means all players except i). The goal of each player is to minimize the expected average cost over the time horizon $[0, T]$.

$$\inf_{a^i \in \mathcal{A}^i} J^i\big(a^i, \{a^{-i}\}\big) = \inf_{a^i \in \mathcal{A}^i} \mathbb{E}\Big[\int_0^T f^i\big(X_t^i, a_t^i, \{X_t^{-i}\}, \{a_t^{-i}\}\big)dt + g^i\big(X_T^i, \{X_T^{-i}\}\big)\Big]$$

subject to

$$dX_t^i = b^i\big(X_t^i, a_t^i, \{X_t^{-i}\}, \{a_t^{-i}\}\big)dt + \sigma^i\big(X_t^i, a_t^i, \{X_t^{-i}\}, \{a_t^{-i}\}\big)dW_t^i + \sigma_0\big(X_t^i, a_t^i, \{X_t^{-i}\}, \{a_t^{-i}\}\big)dW_t^0$$

$$X_0^i = x^i$$

for $i = 1, ..., N$

This problem is an extremely difficult one to solve and except for very specific cases there is no general solution. A classical technique to solve stochastic control problems like the aforementioned is to try to solve a PDE, the Hamilton - Jacobi - Bellman (HJB) equation through which we reconstruct optimal solutions. Unfortunately, this approach will yield a system of N fully coupled PDEs that there is no hope to solve.

Fortunately, not all hope is lost, within the previous class of problems we can identify a subclass of solvable ones that satisfy 2 hypotheses:

H1. Symmetry Each cost function $J^i(a^i, \{a^{-i}\})$ is a symmetric function of $\{a^{-i}\}$

H2. Weak interactions between players The influence of each player is diminishing as the $N \to \infty$

Whenever these two assumptions are satisfied we call the game an N-player symmetric game and for the rest of the thesis we refer to these type of games as N-player game, and abusing a bit the notation we could write a reformulation of the problem as

Individual agent's problem: symmetrical case

$$\inf_{a^i \in \mathcal{A}^i} \mathbb{E}\Big[\int_0^T f\big(X_t^i, a_t^i, \bar{\mu}_t^N\big)dt + g\big(X_T^i, \bar{\mu}_T^N\big)\Big]$$

subject to

$$dX_t^i = b\big(X_t^i, a_t^i, \bar{\mu}_t^N\big)dt + \sigma\big(X_t^i, a_t^i, \bar{\mu}_t^N\big)dW_t^i + \sigma_0\big(X_t^i, a_t^i, \bar{\mu}_t^N\big)dW_t^0$$

$$X_0^i = x^i$$

$$\bar{\mu}_t^N = \frac{1}{N}\sum_{i=0}^{N} \delta_{X_t^i}$$

We droped also the dependence of the empirical distribution on the actions for the sake of brevity in what follows but there is also a big body of research for MFGs of controls.

Remark 1.1.1. *Whenever we write $b(x, \mu)$ essentially me mean $b(x, \mu) = \int \widetilde{b}(x, dy)\mu(dy)$, moreover*

$$dX_t^i = b(X_t^i, \bar{\mu}_t^N)dt + \sigma(X_t^i, \bar{\mu}_t^N)dW_t^i = \frac{1}{N}\sum_{j=0}^{N} \widetilde{b}(X_t^i, X_t^j)dt + \frac{1}{N}\sum_{j=0}^{N} \widetilde{\sigma}(X_t^i, X_t^j)dW^i$$

and we carry over this notational convention to the rest of the functions.

Remark 1.1.2. *In principle we would like to exploit our assumptions to define limits when $N \to \infty$ for J^i and X^i since the problem at the limit is much easier to be solved, so the key question is:* ***how to pass to the limit?***

We will not attempt to answer the question rigorously since it is out of the scope of this introduction and we refer the interested reader to the original works of Snitzman [105] and McKean [88] or books about mean field games [26, 30, 31]. To provide a sketch of the basic idea behind what is call "propagation of chaos" we start with a standard uncontrolled diffusion of the form

$$\begin{aligned} dX_t^{i,N} &= \frac{1}{N}\sum_{j=0}^{N} \widetilde{b}(X_t^{i,N}, X_t^{j,N})dt + dW^i \\ X_0^{i,N} &= x_0 \end{aligned}$$

where we look at N particles and letting $N \to \infty$ each particle $X^{i,N}$ has a natural limit $\bar{X}_t^i$, it is an identical copy of a "nonlinear" process X_t, as it was called early in the literature, with

$$\begin{aligned} dX_t &= \Big\{ \int b(X_t, y)\mu_t(dt) \Big\} dt + dW_t \\ X_0 &= \mathrm{X}_0 \quad \textit{and } \mu_t \textit{ the law of } X_t \end{aligned}$$

Remark 1.1.3. *Whenever we have common noise in the dynamics as it was the case earlier, we need to modify the previous construction into what is called "conditional propagation of chaos" where μ_t is now the conditional law of the process given the common noise. The intuition remains the same about the distribution of particles just we add them in an ambient space that is subject it self to random perturbations, e.g. immersing the cloud of particles into a viscous fluid or a current inside a fluid, for more details see [31, Chapters 1, 2]*

Now to provide a limit for the cost functions on the N player game, we consider $J^{i,N}$ to be uniformly bounded and continuous and then by Ascoli-Arzela theorem we can find a convergent subsequence such that $J^{i_k} \to \bar{J}$ i.e.

$$\lim_{k\to\infty} \sup_{X \in S^N} \big| J^{i_k}(X) - \bar{J}(\bar{\mu}^X) \big| = 0$$

Thus, with another slight abuse of notation we can state the problem at the limit $N \to \infty$, as the problem of a representative agent that tries to minimize a cost that depends on the distribution of the other agents, given the dynamics that also depend on the distribution of the agents.

The representative's agent's problem

$$\begin{aligned} &\inf_{a\in\mathcal{A}} \bar{J}(a;\mu) = \inf_{a\in\mathcal{A}} \mathbb{E}\Big[\int_0^T f\big(\bar{X}, a_t, \mu_t\big)dt + g\big(\bar{X}_T, \mu_T\big) \Big] \\ &\textit{subject to} \\ &d\bar{X}_t = b\big(\bar{X}_t, a_t, \mu_t\big)dt + \sigma\big(\bar{X}_t, a_t, \mu_t\big)dW_t + \sigma_0\big(\bar{X}_t, a_t, \mu_t\big)dW_t^0 \\ &X_0 = x \\ &\mu_t = \mathcal{L}(X_t | \mathcal{F}^{W_t^0}) \quad \text{the conditional distribution of } X_t \text{ given } W_t^0 \end{aligned} \tag{1.1.1}$$

Since we have define our problem and before we discuss ways to solve the problem we should provide a notion of solution. Inspired by traditional game theory we call an action profile a^* a Nash equilibrium for the finite game if there is no other that yield a lower cost, i.e.

$$J^{i,N}(a^*) \leqslant J^{i,N}(a^i, a^{*,-i}) \quad \text{for all } a^i \in \mathcal{A}^i$$

And for the MFG we define an equilibrium as $\mu_t = \mathcal{L}(X_t^{a^*}|\mathcal{F}^{W_t^0})$ for all $t \in [0,T]$, we usually call this **consistency condition** because in the limit of infinitely many players *non of them can influence significantly the distribution* μ and thus consider it **fixed** when solving the minization problem.

Remark 1.1.4. *When there is no common noise $(\mu_t)_{0\leqslant t\leqslant T}$ is a deterministic flow of measures, while in the presence of common noise, the flow becomes stochastic and thus we need to take extra steps to ensure compatibility and for the corresponding σ-algebras, see [31, Chapter 1]*

To make our solution strategy more explicit we articulate the following steps:

1. Solve the optimal control problem $\min_a \bar{J}(a;\mu)$ subject to the dynamics of X_t when the environment μ_t is fixed.
2. Design the optimal control $\alpha^*(t,x)$ (possibly in feedback form depending on the formulation of the problem) and the optimal state process $X_t^{\alpha^*}$
3. Find a fixed point of the map $\Phi(\mu) = \big(\mathcal{L}(X_t^{\alpha^*})\big)_{0\leqslant t\leqslant T}$

Finally, we are ready to conclude our short description of the MFGs by reviewing some solution methods. They will provide a basis for the next chapter of the thesis. In a similar fashion to the stochastic control of the previous section, we can define the value function of our control problem $V(t,x)$ and with her a corresponding HJB equation when the flow $(\mu_t)_{0\leqslant t\leqslant T}$ is fixed. For the moment we will make a small detour to present an idea from deterministic optimal control that would serve as the basic intuitive guideline for the Stochastic HJB that we will need for our MFG solution.

1.1.3.1 Detour to Optimal Control and Potryangin's Maximum Principle

Let us say for a moment that we have a look at the following problem for a single agent/regulator under deterministic dynamics that evolve in $\mathbb{R}$

$$\begin{cases} \min\limits_{a\in\mathcal{A}_{adm}} J(a) = \displaystyle\int_0^T f^0(X_t, a_t)dt \\ \text{subject to} \\ \dot{X}_t = f(X_t, a_t), \quad X_0 = x \in \mathbb{R} \end{cases}$$

with $f, f^0 : \mathbb{R} \times \mathcal{A}_{adm} \to \mathbb{R}$ smooth nonlinear functions representing the dynamics and the running cost respectively. Consider also the dynamic cost

$$X_t^0 = \int_0^t f^0(X_s, a_s)ds.$$

Our goal is to solve the problem inspired by some geometrical intuition. First let's define the extended system

$$\dot{\boldsymbol{x}}_t = \begin{bmatrix} X_t^0 \\ X_t \end{bmatrix}, \quad \boldsymbol{f}(\boldsymbol{x}_t, a_t) = \begin{bmatrix} f^0(X_t^0, a_t) \\ f(X_t, a_t) \end{bmatrix},$$

the system now resides in $\mathbb{R}^2$ since we extended its dimension. If we assume that the control is constant u_t for all $t \in [0, T]$ then we can have a linearization of the extended system

$$\dot{b}_t = \partial_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}_t, \mathrm{a}_t), \quad b_0 = b, \quad \mathrm{a}_t \text{ constant} \tag{1.1.2}$$

that will describe the tangent vectors at $\boldsymbol{x}_t$ for all t. Next from this tangent vector we can study all the vectors p_t that have the property $\langle b_t, p_t \rangle = \textit{const.}$, in words their inner product is constant. It turns out that these vectors reside in a hyperplane and the vectors p_t are perpendicular to $\boldsymbol{x}_t$. The evolution of p_t can be described by

$$\begin{aligned} \dot{p}_t &= -\partial_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}_t, \mathrm{a}_\mathrm{t}) \\ p_T &= 0, \end{aligned} \tag{1.1.3}$$

we call vectors p_t *co-states* and the terminal condition is zero because there was no terminal cost in our example.

Remark 1.1.5. *When the control is not constant, system* (1.1.2) *is not anymore the linearization of the states but instead a first approximation to the evolution of the perturbation in the state.*

$$a_t^\varepsilon = \begin{cases} a_t^* & \textit{for } t \in [s-\varepsilon, s+\varepsilon] \\ b & \textit{otherwise} \end{cases}$$

$$\boldsymbol{x}_s^\varepsilon = \boldsymbol{x}_s^{a^*} + \varepsilon\big(\boldsymbol{f}(\boldsymbol{x}_s^{a^*}, b) - \boldsymbol{f}(\boldsymbol{x}_s^{a^*}, a_s^*)\big) + \mathcal{O}(\varepsilon),$$

the system will evolve from time s to T under (1.1.2) *and the costate system will continue to describe the evolution of the vectors in the "attached" hyperplane.*

Remark 1.1.6. *The equation of the co-states* (1.1.3) *should be understood in a backwards sense since we fix the terminal value that we need to obtain (the derivative of the terminal cost precisely) in ordinary differential equations time reversions are not a problem, we can always make the transformation $t \mapsto T - t$. The real difference will be apparent when we have a look at the stochastic version of the problem where we would need the solution to be adapted to the filtration generated by the noise.*

We made this short discussion and the remark so far, to motivate the definition of the Hamiltonian as

$$H(\boldsymbol{x}, a, p) = \langle p, \boldsymbol{f} \rangle = p_1 f(x, a) + p_2 f^0(x, a)$$

Remark 1.1.7. *In the usual form of the PMP the Hamiltonian appears with $p_2 = 1$ and this is because we can renormalize p_2 since it is a positive constant.*

Now we are ready to state Pontryangin's Maximum principle

Theorem 1.1.8. *If $(\boldsymbol{x}^*, a^*)$ is an optimal (extended) state-action pair, then there exits a continuous function p solving* (1.1.3) *with*

$$\begin{aligned}
\dot{\boldsymbol{x}}_t^* &= \nabla_p H(\boldsymbol{x}_t^*, a_t^*, p_t) \\
\dot{p}_t &= \nabla_{\boldsymbol{x}} H(\boldsymbol{x}_t^*, a_t^*, p_t) \\
H(\boldsymbol{x}_t^*, a_t^*, p_t) &= \max_{c \in \mathcal{A}_{adm}} H(\boldsymbol{x}_t^*, c, p_t) \equiv 0
\end{aligned}$$

and $p_t \neq 0$ for all $t \in [0, T]$ with $p_0 \leqslant 0$

The PMP helps identifying the optimal control within a class of optimal controls. The main difference with DPP from the previous section is that with PMP we assume existence of an optimal control while with DPP we prove existence in feedback form.

Remark 1.1.9. *In the present setting we could extend our model problem to include Browninan noise of constant intensity without hardly changing anything*

$$dx_t = f(x_t, a_t)dt + dW_t, \quad x_0 = x \in \mathbb{R}.$$

Obviously we lose the interpretation of the systems with the hyperplanes and the tangent vectors since there is no more smoothness because of the noise but nevertheless the rest that is based on a first order approximation works. The Hamiltonian is going to be of the same shape tho only major difference concerns the co-states that essentially encodes the shadow price of the states will be stochastic to reflect the stochastic nature of the forward dynamics. In this case it is no longer possible to reverse the time without destroying the adeptness of the solution to the filtration of the noise and thus we need to add an extra term that will account for that. Our problem remains the same we want p_T to be fixed at the terminal time, only that now p_t is a stochastic process that evolves according to the dynamics

$$\begin{aligned}
dp_t &= -\nabla_x H(x_t, p_t, a_t) + q_t dW_t \\
p_T &= 0,
\end{aligned} \tag{1.1.4}$$

to make up for this adaptiveness we add another process q_t that is part of the solution and is implicitly defined and we call (1.1.4) *a Backward Stochastic Differential Equation (BSDE). For more details see [119, Chapter 6]*

Remark 1.1.10. *The situation where we have control over the intensity of the Brownian motion (diffusion part) is substantially different both in terms of interpretation and in terms of applied techniques. To make things more concrete*

$$dx_t = f(x_t, a_t)dt + \sigma(x_t, a_t)dW_t, \quad x_0 = x \in \mathbb{R},$$

now our current decisions (control a_t) not only affect the direction we steer (drift) but also the intensity of the noise we are facing. So except for shadow price of the dynamics (steering) we need to account for the uncertainty (or risk) for the Brownian motion when we balance our decisions.

This makes the first order approximations insufficient, and we need to rely on second order ones. In this way, we add a second BSDEs in a similar fashion as in the previous remark.

$$dP_t = -\Big\{2\partial_x f(x_t,a_t)P_t + \partial_x\sigma^2(x_t,a_t)P_t + 2\partial_x\sigma^2(x_t,a_t)Q_t + \partial_{xx}H(x_t,a_t,p_t,q_t)\Big\}dt + Q_t dW_t.$$

For more details and a version of Stochastic PMP we refer to [119]

1.1.3.2 Solution strategies for MFGs with common noise

Coming to our MFG with common noise we would like to proceed as with classical dynamic programming. We need a HJB equation to design optimal feedback controls and an equation that describes the evolution of the conditional distribution (essentially a density). Given the dynamics of (1.1.1) we can write the stochastic Fokker-Planck equation using Itô's rule on a test function $\varphi(X_t,t)$ and taking the conditional expectation with respect to the initial condition and common noise W_t^0 this results in a Stochastic PDE since the dW_t^0 will survive the conditioning

$$\begin{aligned} d\mu_t &= \left\{-\partial_x\big(b(X_t,\mu_t,a_t)\mu_t\big) + \frac{1}{2}\partial_{xx}\Big(\big(\sigma^2(X_t,\mu_t,a_t)+\sigma_0^2(X_t,\mu_t,a_t)\big)\mu_t\Big)\right\}dt - \partial_x(\sigma_0(X_t,\mu_t,a_t)\mu_t)dW_t^0 \\ &\mu_0, \text{ given initial distribution on } \mathbb{R}, \end{aligned} \tag{1.1.5}$$

the way we should understand this equation is in the sense of distributions, for any test function $\varphi \in C_0^\infty(\mathbb{R}\times[0,T])$.

Next, we need to define a value function and a DPP, since as we explained earlier when the representative agent solves the optimization problem and the flow of measure is fixed, the value function depends only on t,x. But when we want to apply Itô's rule to expand the value function we notice that $u_t(X_t)_{0\leqslant t\leqslant T}$ is indeed a random field because of the random flow of measures μ_t[3] that enters in the coefficients of X_t and thus we need an adapted form of Itô's rule, it is known by various names Itô-Kunita formula or Itô-Wentzell and for precise statements we refer to [31, Section 1.4.2] [101] and [95]

$$u_t(x) = \operatorname*{essinf}_{a\in\mathcal{A}_{adm}} \mathbb{E}\Big[\int_t^T f\big(X_s,a_s,\mu_s\big)ds + g\big(X_T,\mu_T\big) \mid X_0 = x\Big],$$

with essential infimum to guarantee that the result remains a random variable, and for the DPP

$$u_t(x) = \operatorname*{essinf}_{a\in\mathcal{A}_{adm}} \mathbb{E}\Big[\int_t^{t+h} f\big(X_s,a_s,\mu_s\big)dt + u_{t+h}(X_{t+h}) \;\Big|\; X_t = x\Big],$$

we can derive the stochastic HJB in a heuristic way if we consider

$$u_t(x) \geqslant \mathbb{E}\Big[\int_t^{t+h} f\big(X_s,a_s,\mu_s\big)dt + u_{t+h}(X_{t+h}) \mid X_t = x\Big] \tag{1.1.6}$$

[3] that we still keep fixed as in the usual

and since $u_t(X_t)$ is indeed random and adapted to the filtration generated by the common noise W_t^0 for any fixed x, the best we can hope is for it to be a semimartingale

$$du_t(x) = \int_t^T \Phi_s(x)ds - \int_t^T \Psi_s(x)dW_t^0 \quad u_T(x) = g(x, \mu_T) \tag{1.1.7}$$

where in this BSDE the intuition behind term Ψ_t follows Remark 1.1.9 and it's part of the definition of the solution. We can now expand $u_t(X_t)$ by Itô-Wentzell's formula

$$\begin{aligned} du_t(X_t) =& \Big\{ \partial_x u_t(X_t) b(X_t, \mu_t, a_t) + \frac{1}{2}\partial_{xx} u_t(X_t)\Big(\sigma^2(X_t, \mu_t, a_t) + \sigma_0^2(X_t, \mu_t, a_t)\Big) \\ &+ \partial_x \Psi(X_t)\sigma_0(X_t, \mu_t, a_t) + \Phi_t(x)\Big\}dt + \partial_x u_t(X_t)\sigma(X_t, \mu_t, a_t)dW_t \\ &+ \partial_x u_t(X_t)\sigma_0(X_t, \mu_t, a_t)dW_t^0 + \Psi_t(x)dW_t^0, \end{aligned}$$

that means

$$\begin{aligned} u_{t+h}(X_{t+h}) - u_t(x) =& \int_t^{t+h} \Big\{ \partial_x u_t(X_t) b(X_t, \mu_t, a_t) + \frac{1}{2}\partial_{xx} u_t(X_t)\big(\sigma^2(X_t, \mu_t, a_t) \\ &+ \sigma_0^2(X_t, \mu_t, a_t)\big) + \partial_x \Psi(X_t)\sigma_0(X_t, \mu_t, a_t) + \Phi_t(x)\Big\}dt + \int_t^{h+h} \partial_x u_t(X_t)\sigma(X_t, \mu_t, a_t)dW_t \\ &+ \int_t^{h+h} \Big\{ \partial_x u_t(X_t)\sigma_0(X_t, \mu_t, a_t) + \Psi_t(x)\Big\}dW_t^0, \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}\Big[u_{t+h}(X_{t+h}) - u_t(x)\Big|\mathcal{F}_t^{W^0,x}\Big] \\ &= \mathbb{E}\Big[\int_t^{t+h} \Big\{ \partial_x u_t(X_t) b(X_t, \mu_t, a_t) + \frac{1}{2}\partial_{xx} u_t(X_t)\big(\sigma^2(X_t, \mu_t, a_t) + \sigma_0^2(X_t, \mu_t, a_t)\big) \\ &+ \partial_x \Psi(X_t)\sigma_0(X_t, \mu_t, a_t) + \Phi_t(x)\Big\}dt \;\Big|\; \mathcal{F}_t^{W^0,x}\Big], \end{aligned}$$

going back to (1.1.6)

$$\begin{aligned} 0 \leqslant \mathbb{E}\Big[\frac{1}{h}\int_t^{t+h} \Big\{ f\big(X_s, \mu_s, a_s\big) + \partial_t u_t(X_t) + \partial_x u_t(X_t) b(X_t, \mu_t, a_t) + \frac{1}{2}\partial_{xx} u_t(X_t)\big(\sigma^2(X_t, \mu_t, a_t) \\ + \sigma_0^2(X_t, \mu_t, a_t)\big) + \partial_x \Psi(X_t)\sigma_0(X_t, \mu_t, a_t) + \Phi_t(x)\Big\}dt \;\Big|\; \mathcal{F}_t^{W^0,x}\Big], \end{aligned}$$

letting $h \to 0$ and taking the inf we can identify process $\Phi_t(x)$ and get from (1.1.7) the form of SHJB equation

$$\begin{aligned} &du_t(x) = -H(X_t, a^*, \partial_x u_t(X_t), \partial_{xx} u_t(X_t), \partial_x \Psi_t(X_t))dt + \Psi_t(X_t)dW_t^0 \\ &u_T(x) = g(x, \mu_T) \end{aligned} \tag{1.1.8}$$

with the Hamlitonian H being

$$H(x, a^*, p, P, q) = \inf_{a \in \mathcal{A}_{adm}} \Big\{ f\big(x, \mu, a\big) + p \cdot b(x, \mu, a) + \frac{1}{2} P \cdot \big(\sigma^2(x, \mu, a) + \sigma_0^2(x, \mu, a)\big) + q \cdot \sigma_0(x, \mu, a) \Big\}$$

Nevertheless, this is not the only possible derivation for the stochastic HJB equation, in the next section we will see a HJB equation for a value function that depends on the measure argument, the so called master equation and we can derive the SHJB equation from the master equation following a strategy presented by P.L. Lions in his lectures at College de France [84] and appear in the notes [23, Section 7].

Remark 1.1.11. *Solving the system* (1.1.5)-(1.1.8) *apart from being notoriously difficult to be solved will give us an optimal control in* $a^* = \alpha(t, X_t^{a^*}, \mu_t, \partial_x u_t(X_t^{a^*})$ *that is going to be a* ***random field*** *and this is a very important remark for appreciation of the results that we want to obtain in this thesis. This will be more apparent in later parts of the introduction.*

Another approach to solve (1.1.1) would be using the Pontryagin's Maximum Principle but we refrain from providing all the details here since the first chapter of the thesis will follow this approach for a Linear Quadratic case and instead we refer to [31] for a general approach.

To conclude this Section we would like to recap our solution strategy. Finding Nash equilibrium for MFG is in a nutshell the same as finding a fixed point for the flow $(\mathcal{L}(X_t^{\alpha^*})_{0 \leqslant t \leqslant T}$ of conditional distributions from the optimal control problem and we derived a SPDE system to characterize these optimizers. For analysis of the equations themselves and their solvability we refer to classical references in the field [31, 26] in Section 1.3 we will highlight some challenges that mean field models face and especially models with common noise.

1.1.4 Mean Field Control

While MFGs mostly describe large competitive games we would like to use a similar framework to describe large collaborative games. We could imagine agents to collaborate in order to solve a common problem or perform a task such as drones delivering goods, putting in order warehouses or reaching specific areas. In domains like economics or finance, we could also imagine a central planner wanting to optimize some aspect of the economy, manipulating aggregate quantities, or a regulator for the financial markets.

There are two ways to interpret problems that involve the law of a process, either we can imagine some stochastic process that we control individually and its law appears in the cost functional that we want to minimize, that could be the easiest case. Or have a look at the mean field limit of a particle system, i.e. a MKV equation where we lose some of the nice properties enjoyed by standard SDEs with smooth coefficients. In either way the important remark here is that we are dealing with a "real" optimal control problem, set on the space of measures as we will demonstrate shortly while in contrast in MFGs we were looking mostly at a fix point problem for the measure flow $(\mu_t)_{0 \leqslant t \leqslant T}$.

For the sake of brevity and to minimize the technical burden we will start with a simplified version of an MFC problem to motivate our approach and then comment on the considerably more involved case of common noise. Let's have a look at the following control problem

$$\begin{aligned}&\inf_a\Big\{\mathbb{E}\int_0^T f(X_t,\mathcal{L}(X_t),a_t)dt\Big\}\\&\text{subject to}\\&dX_t=b(X_t,a_t)dt+\sigma(X_t)dW_t,\quad X_0=x,\end{aligned}\tag{1.1.9}$$

in equation (1.1.9) we notice that the law of X_t enters in the running cost and thus this is what really distinguish the MFC from a standard optimal control problem, adjustments of the drift affect the law of the process and thus the cost, so we need to take this into account when designing optimal controls.

In fact we can reformulate (1.1.9) to reflect more an optimal control of the law of the process. Assuming enough smoothness of $b(x,a),\sigma(x,a)$ so the law of X_t has a smooth density that satisfies the Fokker-Planck

$$\partial_t\mu_t(x)-\frac{1}{2}\partial^2_{xx}\big(\sigma(x)\mu_t(x)\big)+\partial_x\big(b(x,a)\mu_t(x)\big)=0,\quad \mu_0=\mathcal{L}(x)\tag{1.1.10}$$

we notice that the initial condition is being fixed as the law of the initial state (which is herself an $L_2(\mathbb{R})$ random variable). Now our problem reads as

$$\begin{aligned}&\inf_a\Big\{\int_0^T\int_{\mathbb{R}} f(x,\mu_t(x),\alpha(t,x))\mu_t(x)dxdt\Big\}\\&\text{subject to}\\&\partial_t\mu_t(x)-\frac{1}{2}\partial^2_{xx}\big(\sigma(x)\mu_t(x)\big)+\partial_x\big(b(x,a)\mu_t(x)\big)=0,\quad \mu_0=\mathcal{L}(x).\end{aligned}\tag{1.1.11}$$

It's a "standard" deterministic optimal control problem with a fixed initial condition just it is not classical because it is stated on the space of probability measures. We can define an value function for problem (1.1.11) staring at time t from μ[4]

$$u(t,\mu)=\inf_a\Big\{\int_t^T\int_{\mathbb{R}} f(x,\mu_s(x),\alpha(s,x))\mu_s(x)dx\,ds\;\Big|\;\mu_t=\mu\Big\}$$

next we can write a Dynamic programming principle for the value function

$$u(t,\mu)=\inf_a\Big\{\int_t^{t+h}\int_{\mathbb{R}} f(x,\mu_s(x),\alpha(s,x))\mu_s(x)dx\,ds+u(t+h,\mu_{t+h})\;\Big|\;\mu_t=\mu\Big\}.\tag{1.1.12}$$

Now using the classical strategy of deterministic optimal control we can derive a HJB equation by Taylor's expansion of $u(t+h,\mu_{t+h})-u(t,\mu)$ for which we need to define a notion of derivative on the space of probability measures.

[4] we remind the reader of our common notation abuse of measures and densities for this section

1.1.4.1 Derivatives of probability measures

Motivated by the DPP on the space of probability measures we would need to consider the introduction of derivatives for smooth functions of probability measures. Practically this is not an easy task as the space of probability measures, say for the sake of concreteness that we work on $\mathcal{P}_2(\mathbb{R})$ (i.e. measures with bounded second moments) is not a Hilbert space and thus we cannot simply apply generalized notions of derivatives from functional analysis that work on Hilbert spaces. We either need to work with the existing topological structure of the space, which is sufficiently nice because $(\mathcal{P}_2(\mathbb{R}), d_{W_1})$ is a compact topological space, with d_{W_1} the 1-Wasserstein distance. Or somehow "place" our smooth functions of measures into a suitable Hilbert space. The first strategy would lead to the notion of linear functional derivative, while the second to Lion's derivative.

Linear functional derivative

Definition 1.1.12. *We say that $u : \mathcal{P}_2 \to \mathbb{R}$ has a Linear Functional derivative, $\frac{\delta u}{\delta m} : \mathcal{P}_2(\mathbb{R}) \times \mathbb{R} \to \mathbb{R}$ that is a continuous and bounded differential operator with the property*

$$u(\mu) - u(\mu_0) = \int_0^1 \int_{\mathbb{R}} \frac{\delta u}{\delta m}\big(k\mu + (1-k)\mu_0\big)(x) d\big[\mu - \mu_0\big](x) dk$$

In plain words the definition says that the difference between $u(\mu)$ and $u(\mu_0)$ is the integral of the derivative evaluated in the segment between two measures μ, μ_0. For the definition to be better understood we could give a simple (trivial) example

Example 1.1.13. *Suppose $u(\mu) = \int_{\mathbb{R}} \varphi(x)\mu(dx)$ with $\varphi \in C^1(\mathbb{R})$ then*

$$\frac{\delta u}{\delta m}(\mu)(x) = \varphi(x) + constant$$

This example although its simplicity is educative because it illustrates that the notion of Linear functional derivative is not unique and is only determined up to a constant. This is something that the Lion's derivative takes into account.

First Order Lion's derivative

Following the ideas of P.L. Lions from his lectures in college de France we introduce the *lifting* of a function defined on the space of probability measures, $\mathcal{P}_2(\mathbb{R})$ to the space of random variables, so to $L_2(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R})$ over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, Ω being Polish and $\mathbb{P}$ atomless.

$$\widetilde{u}(X) = u(\mathcal{L}(X))$$

Consequently we can use Fréchet differentiation to define a notion of derivative (for a refresher of the notion see [121]). We adopt the definition from [23]

Definition 1.1.14. *We say that $u : \mathcal{P}_2 \to \mathbb{R}$ is differentiable at $\mu_0 \in \mathcal{P}_2$ if there exists $X_0 \in L_2(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathcal{L}(X_0) = \mu_0$ and $\widetilde{u}$ is Fréchet differentiable at X_0 with $D\widetilde{u}(X_0)$ its derivative and*

$$\widetilde{u}(X_0 + h) - \widetilde{u}(X_0) = D\widetilde{u}(X_0)(h) + \|h\|_{L_2}\mathcal{O}(h) \quad \text{for all } h \in L^2(\Omega, \mathcal{F}, \mathbb{P}).$$

Furthermore, u is C^1 in a neighbourhood of μ_0 if there exists $X_0 \in L_2(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathcal{L}(X_0) = \mu_0$ and $D\widetilde{u}(X_0) : L^2(\Omega, \mathcal{F}, \mathbb{P}) \to \big(L_2(\Omega, \mathcal{F}, \mathbb{P})\big)^* \equiv L_2(\Omega, \mathcal{F}, \mathbb{P})$ is continuous.

In fact, from Riesz representation theorem if $D\widetilde{u}(X_0)$ exists, we can identify it with an element of $L_2(\Omega, \mathcal{F}, \mathbb{P})$, that we should call $\mathcal{D}\widetilde{u}(X_0)$ with the property

$$D\widetilde{u}(X_0)(h) = \langle \mathcal{D}\widetilde{u}(X_0), h\rangle = \mathbb{E}[\mathcal{D}\widetilde{u}(X_0)h], \quad \text{for any } h \in L_2(\Omega, \mathcal{F}, \mathbb{P}).$$

Now Theorem 6.2 in [23] says that if $\widetilde{u}$ is differentiable at X_0 then it must be also at any $X \in L_2(\Omega, \mathcal{F}, \mathbb{P})$ in the neighbourhood of X_0 such that $\mathcal{L}(X) = \mu_0$, thus the law of $\mathcal{D}\widetilde{u}(X_0)$ is independent of X_0. This is what makes the definition of the derivative intrinsic (some people call this derivative the intrinsic one) . But we can say more about $\mathcal{D}\widetilde{u}(X_0)$, according to Theorem 6.5 [23]

$$\mathcal{D}\widetilde{u}(X_0) = \xi(X_0) \quad \text{for } \xi \in L_2(\Omega, \mathcal{F}, \mu_0), \quad \mu_0 - a.e.$$

we will call the representative of the equivalent class of ξ as ∂_μ and use the notation

$$u(\mu) - u(\mu_0) = \mathbb{E}[\partial_\mu u(\mathcal{L}(X_0))(X_0)(X - X_0)] + \|X - X_0\|_{L_2}\mathcal{O}(\|X - X_0\|_{L_2})$$

To illustrate the above construction we can revisit Example 1.1.13 to compute the Lion's derivative.

$$\psi(t) = \widetilde{u}(X + th) = \mathbb{E}[\varphi(X + th)], \quad \psi'(0) = D\widetilde{u}(X)(h) = \mathbb{E}[\varphi'(X)h].$$

Thus we can identify the $\partial_\mu u(\mu)$ with $\varphi'(x)$.

Remark 1.1.15. *Comparing the Linear functional derivative with Lion's derivative for Example 1.1.13 we can see a relationship of the form*

$$\partial_x \frac{\delta u}{\delta m}(\mu)(\cdot) = \partial_\mu u(\mu)(\cdot)$$

This relationship is more general than our simple example and is thoroughly analysed in [30, Chapter 1]. For a notion of gradient $\partial_x \partial_\mu u(\mu)$ ($\nabla_x \partial_\mu u(\mu)$ in higher dimensions) for see the following remarks.

Remark 1.1.16. *Very often we will use the notation $\partial_\mu u(\mu_t)$ omitting the second variable but the reader should always keep in mind that both measure derivatives are elements of the product space $\mathcal{P}_2 \times \mathbb{R}$ (or $\mathbb{R}^d$).*

Remark 1.1.17. *The Fréchet derivative $D\widetilde{u}(X_0)(h)$ should be interpreted as a generalized **directional derivative** in the direction of h. As it is a functional of with two arguments it is within our interest to define a form of gradient with respect to the directional variable h that we will refrain form using the common symbol ∇ and instead follow [30, 31] and denote it with $\partial_v \partial_\mu u(\mu)(v)$ to avoid confusion with state variables and its derivatives usually denoted by x.*

Second order Lion's derivative

Let us denote $D^2\widetilde{u}(X_0)$ the second order Fréchet derivative of $\widetilde{u}$ in a neighbourhood of X_0 then completely analogously to the first order case, we would need 2 directions, lets call them h_1, h_2 to

compute the derivative since $D^2\widetilde{u}(X_0)(h_1)(h_2) = D\big(D(\widetilde{u}(X_0)(h_1))\big)(h_2)$ with the operator D being symmetric. Then we need to repeat the steps we did before to properly define $\partial^2_{\mu\mu}u(\mu_t)(v)(v')$ but with some twist because of the directional sense that the derivative includes and the fact that we need to define μ_0 almost everywhere the first derivative in direction h_1 and then again to define $\partial^2_{\mu\mu}u(\mu_t)(h_1)$, μ_0-almost everywhere the second, for more details see [30, Section 5.6.2]. Finally, in a way that resembles a lot the chain rule we write

$$\frac{d}{dt}D\widetilde{u}(X_0 + th)\Big|_{t=0} = \partial_v\partial_\mu u\big(\mathcal{L}(X_0)\big)(X_0)h + \bar{\mathbb{E}}\Big[\partial^2_{\mu\mu}u\big(\mathcal{L}(X_0)\big)(X_0)(\bar{X}_0)\bar{h}\Big],$$

where the bar over expectation, X and h, denotes copies of the original ones in a copied space that was introduced to do deal with the identification second directional derivative as best explained in the aforementioned reference.

1.1.4.2 The master equation

Going back to our strategy of deriving a HJB equation of the value function in DPP (1.1.12), we can expand u using the first order derivative and assuming smoothness is both arguments

$$u(t + h, \mu_{t+h}) = u(t, \mu) + h\ \partial_t u(t, \mu) + h\ \int_{\mathbb{R}} \frac{\delta u}{\delta m}(t, \mu)(x)\partial_t\mu_t(x)dx$$

with the convention "$\partial_t\mu_t = \frac{d\mu_t}{dt}$" and using the Fokker-Planck equation

$$u(t + h, \mu_{t+h}) = u(t, \mu) + h\ \partial_t u(t, \mu) + h \int_{\mathbb{R}} \frac{\delta u}{\delta m}(t, \mu)(y)\Big(\frac{1}{2}\partial^2_{xx}\big(\sigma(x)\mu_t(x)\big) - \partial_x\big(b(x, a)\mu_t(x)\big)\Big)dx$$

then use integration by parts to pass the derivatives to $\frac{\delta u}{\delta m}(t, \mu)$ and Remark 1.1.15

$$u(t + h, \mu_{t+h}) = u(t, \mu) + h\ \partial_t u(t, \mu) + h \int_{\mathbb{R}} \left(\frac{1}{2}\sigma^2\partial_v\partial_\mu u(t, \mu)(x) + \partial_\mu u(t, \mu)(x)b(x, a)\right)\mu_t(x)dx$$

finally we can conclude using the common strategy that the HJB equations reads as

$$\begin{cases} \partial_t u(t, \mu) + \displaystyle\int_{\mathbb{R}} \left(H^*(t, x, \mu, \partial_\mu u(t, \mu), \alpha^*(t, x)) + \frac{1}{2}\sigma^2\partial_v\partial_\mu u(t, \mu)(x)\right)\mu(x)dx \\ u(T, \mu) = 0 \end{cases}$$

with the Hamiltonian

$$H(t, x, \mu, p, a) = f(x, \mu, a) + p \cdot b(x, a)$$

Solvability of this master equation on the space of probability measures is an important issue. This will be a major point of discussion later see Section 1.3

1.1.4.3 MFC with common noise

In order to discuss the MFC problem with common noise, we need to take into account the the stochastic nature of the Fokker-Planck equation and thus we need to apply a formal version of Itô's rule for smooth functions of measures. We will refrain from giving all the details or even the form of it and instead first write the value function in order to discuss a particularity that there exists because of the stochastic nature of the flow $(\mu_t)_{0\leqslant T}$ and then provide the form of the second order master equation. All the details can be found in [31, Section 3.3-4] and in [26, Chapter 4 and Appendix A].

We keep the our simplified version of dynamics and our smoothness assumptions

$$dX_t = b(X_t, a_t)dt + dW_t + dW_t^0, \quad X_0 = x, \tag{1.1.13}$$

the Stochastic Fokker- Planck is

$$\begin{aligned} d\mu_t^\alpha &= \Big\{-\partial_x\big(b(\cdot,\alpha(t,\cdot))\mu_t^\alpha\big) + \frac{1}{2}\Big(\big(\sigma^2+\sigma_0^2\big)\partial_{xx}\mu_t^\alpha\Big)\Big\}dt - \partial_x(\sigma_0\mu_t^\alpha)dW_t^0 \\ \mu_0^\alpha &= \mathcal{L}(x), \end{aligned} \tag{1.1.14}$$

and the value function for an initial condition of the state process ξ that is distributed according to μ

$$u(t,\mu) = \mathbb{E}[V(t,\xi,\mu)] = \inf_{\boldsymbol{a}} \mathbb{E}[\int_t^T f(X_t^a, \mathcal{L}(X_t^a \mid W_t), a_t)dt + g(X_t^a, \mathcal{L}(X_T^a \mid W_T))|X_t^a = \xi \sim \mu] =$$

with DPP for $V(t,\xi,\mu)$

$$V(t,\xi,\mu) = \inf_{\boldsymbol{a}}\Big\{\mathbb{E}\Big[\int_t^{t+h} f(X_s^a, \mathcal{L}(X_s^a \mid W_s), a_s)ds + u(t+h, \mathcal{L}(X_{t+h}|W_{t+h})) \;\Big|\; X_t^a = \xi \sim \mu\Big]\Big\},$$

Then supposing $V(t,\xi,\mu)$ is smooth enough in all 3 arguments we can use Itô's formula [31, Equation 4.37] to get a HJB equation on the space of probability measures. Neither the shape not the equation herself is in the scope of this thesis so do not even attempt to state her (since it would involve terms that need to be explained and make our previous analysis more refined). We are more interested in the philosophy behind the derivation and the principles to be followed.

1.2 Why Mean Field Reinforcement Learning is a great deal?

In this section we would like to talk about the motivation for this thesis. In recent years there has been big success for RL that attracted attention from a greater public outside academia. In 2006 AlphaGo developed by DeepMind was able to beat the world champion Go player, Lee Sedol, in a five-game match and later defeated the world's number one ranked player, Ke Jie. This of course fuelled the interest from the public as in can be seen from Figure 1.2, this was a starting point of

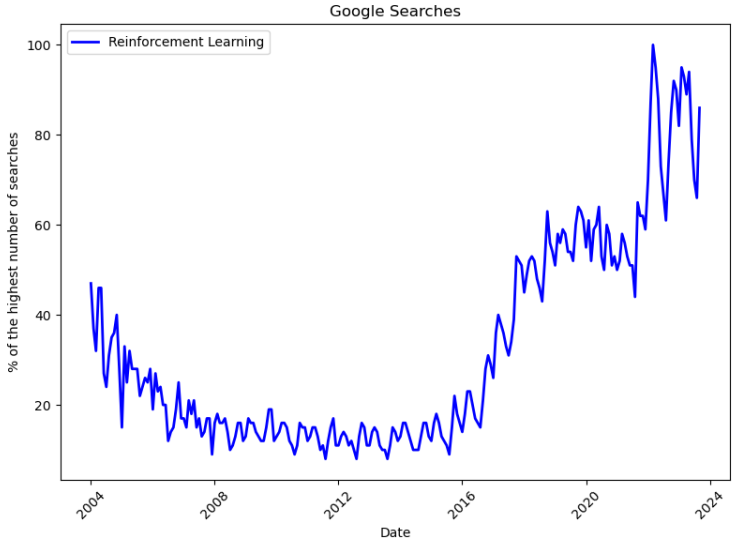


Figure 1.2: Google searches for the term reinforcement learning, as a percentage of the maximum number of clicks. With data from Google Trends, https://trends.google.com/trends/explore?hl=el

an exponential increase in the public's attention especially after the launch of ChatGPT in late 2022 and the general interest in artificial intelligence.

Naturally after 2016, the interests started to expand into various cases that could be solved by RL and extensions to multi agent system, see Figure 1.3. While MFGs have attracted a lot of attention and they have their own proper community there has been huge leverage from the side of RL to solve cases where under some of the assumptions of MFGs we saw in Section 1.1.3 considerable simplifications can be mades and thus solve the cases.

Modern technological advances call for coordination or competition between large numbers of identical agents, coordinating floats of transporting drones or drones performing auxiliary tasks in agriculture or robots putting in order warehouses, even agents agents avoiding congestion in communication networks or financial markets can be modelled as MARL.

On more reason that motivated us in this thesis was the opportunity to harness the power of approximation techniques from RL theory to investigate numerically issues arising directly from Mean Field theory, e.g. kernels approximation for solving the master equation in continuous space without space discretization. For more details in this direction see the next Section 1.4

Of course our main motivation for this thesis was to solve concrete problems that contribute in the general progress of AI, but before we dive into the challenges we aim at overcoming with our work we should give an informal but informative introduction into what we really mean by MARL. In order to avoid this introduction becoming a book herself we aim only at highlighting the main ideas in relation with the results presented later in the thesis, for a non technical introduction with applications we refer to [123], for a more technical review of the state of the art both in RL and and MARL we refer to [70] and [83]

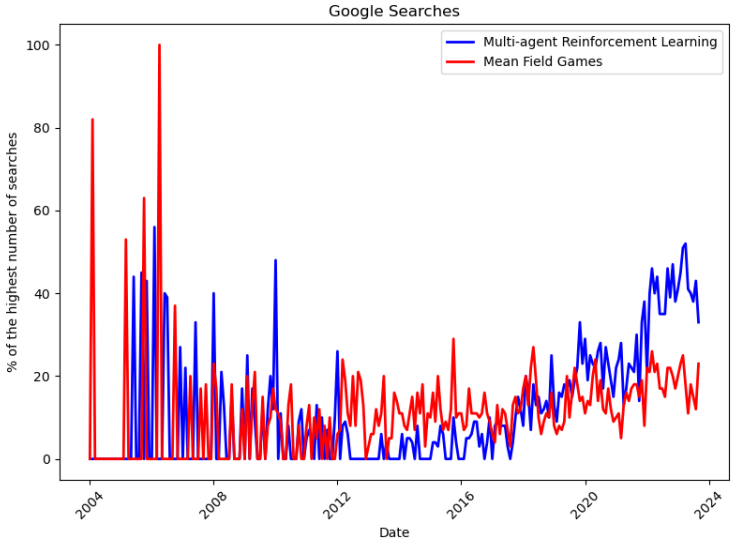


Figure 1.3: Comparison between Google searches for the terms: multi agent reinforcement learning and mean field games, as a percentage of the maximum number of clicks. With data from Google Trends, https://trends.google.com/trends/explore?hl=el

1.2.1 Learning in Games

Let's come back for a moment to a finite game, to clarify what we mean by learning before we advance into the mean field regime again. Suppose we are talking for a 2 player zero sum game $(R, C = -R)$, if the players are aware of the payoff functions, and the actions then all they have to do is to compute their min-max strategies and follow them forever to arrive in a Nash equilibrium if it exists. Now the question that concerns us is if we can arrive in a Nash equilibrium when the players interactions are distributed and thus take decisions when they don't know the payoff matrix of the game and can only estimate payoffs based on their interactions. To make things more concrete we will discuss a particular case of a learning algorithm, **fictitious play** that we will use also on the first part of the thesis. We borrow the following definition and example from Daskalakis [43]

Definition 1.2.1. *Fictitious play is the completely uncoupled player interaction in which in every round each player plays a best response to the opponent's historical (empirical) strategy. The following algorithm describes this procedure*

- ***Round 1***
  - *row player i_1, column player j_1, arbitrary*
  - *row player receives Re_{j_1}, column player receives $e_{i_1}^T C$ with e_{j_1}, e_{i_1} the coordinates on the payoff matrix*
- ***Round 2***

– *row player plays* $i_2 = \underset{i}{argmax}\{e_i^T R e_{j_1}\}$, *column player plays* $j_2 = \underset{j}{argmax}\{e_{i_1}^T C e_j\}$

– *row player receives* Re_{j_2}, *column player receives* $e_{i_2}^T C$

- ***Round n***

  – *row player plays* $i_n = \underset{i}{argmax}\{e_i^T R_{j_1} y_{n-1}\}$, $y_{n-1} = \frac{1}{n-1}\sum_{k\leqslant n} e_{j_k}$,

  column player plays $j_n = \underset{j}{argmax}\{x_{n-1}^T C e_j\}$, $x_{n-1} = \frac{1}{n-1}\sum_{k\leqslant n} e_{i_k}$,

  – *row player receives* Re_{j_n}, *column player receives* $e_{i_n}^T C$

Example Assume the payoff matrix

$$R = \begin{pmatrix} 2 & 1 & 0 \\ 2 & 0 & 3 \\ -1 & 3 & -3 \end{pmatrix}, \quad C = -R = \begin{pmatrix} -2 & -1 & 0 \\ -2 & 0 & -3 \\ 1 & -3 & 3 \end{pmatrix}$$

and at round 1 both row and column players choose arbitrarily $(i_1, j_1) = (1, 3)$ and both receive 0 as payoff. We summarise the first three rounds of the game in the following table

Round	i	j	$e_1^T R y_n$	$e_2^T R y_n$	$e_3^T R y_n$	$x_n^T C e_1$	$x_n^T C e_2$	$x_n^T C e_3$
1	1	3	0	3	-3	-2	-1	0
2	2	3	0	6	-6	-4	-1	-3
3	2	2	1	6	-3	-6	-1	-6

we would like to emphasize with this simple example that the players have to take decisions based on the estimates they have for the payoff of their actions and thus learning is translated into making more accurate predictions for the potential payoffs. An important variant of the deterministic fictitious play that we presented here is the *stochastic fictitious play* where the agent is allowed to randomize her action whenever she is indifferent between choices, an important consideration is the intensity of the noise in the system that have to be carefully chosen for more details see Fundenberg and Levine [58]. Reminiscent of that is Our version of fictitious play for MFGs in the first part is reminiscent of stochastic fictitious play, using a suitable randomization through the common noise see Section 1.4.3.

1.2.2 Introducing MARL

For this thesis, **Mean Field Reinforcement Learning refers to a (competitive or collaborative) game between a representative agent and an infinite population of agents as explained in Section 1.1.3.** Similar to the previous section we will separate the analysis into competitive and collaborative MARL.

We assume that we are already in the mean field regime as described in Section 1.1.3, consequently we have a representative agent and an infinite population of agents that either compete or collaborate. The representative agent is in a state $X \in \mathcal{X} \subseteq \mathbb{R}$ and the population at μ, then the agent picks actions a from an action set $\mathcal{A}$ and "nature" (or a black box as we will call it later)

returns the state of the agent X', the state of the population μ some aggregate cost c, thus we can have a sequence of transitions $(X_t, a_t, \mu_t) \to (X_{t+1}, a_{t+1}, \mu_{t+1}, c_{t+1})$ if the dynamics of X_t and the cost functional are known we can resort to techniques from Section 1.1.3 to solve the problem, otherwise as explained in Section 1.1.2 we have two ways to develop solutions, either **model based** where we construct and eventually learn a model for the dynamics and the cost or **model free** where we learn directly the optimal values from the sequence of the transitions.

1.2.2.1 Competitive MARL: the MFG with learning case

In this section, we consider the case of the competition between agents, and as in the MFG paradigm we need to find a fixed point for the flow of distributions of the states $(\mu_t)_{0\leqslant t\leqslant T}$ since each agent considers the μ_t fixed when solving the optimization problem. We opt for a model based approach since this is also the approach followed in the first part of the thesis, while in the next section we deploy a model free method to motivate the later parts of the thesis. Our model based approach is a natural extension of Section 1.1.2 and it was introduced by Guo et al [62].

To start let's consider the standard MFG problem without common noise.

$$\begin{aligned}
&\inf_{a\in\mathcal{A}} J(a;\mu) = \inf_{a\in\mathcal{A}} \mathbb{E}\Big[\int_0^T f\big(X_t, a_t, \mu_t\big)dt + g\big(X_T, \mu_T\big)\Big],\\
&\textit{subject to}\\
&dX_t = b\big(X_t, a_t, \mu_t\big)dt + \sigma\big(X_t, a_t, \mu_t\big)dW_t\\
&X_0 = x,\\
&\mu_t = \mathcal{L}(X_t) \quad \textit{the consistency condition},
\end{aligned} \tag{1.2.1}$$

obviously the case that we are concerned is when f, g, b, σ are unknown. One more point that merits discussion is the available information for the representative agent for designing (optimal) policies here again there are two possible ways, either the flow μ_t is being observed and the policy depends on it (**population dependent policy**) or **population agnostic** policies that does not depend on the mean field. The first one is more relevant into theoretical cases and offers better chances for generalization while the second one is more realistic since we cannot expect a single agent to be able to observe the whole population and is also the benefit of passing to the mean field regime in games.

Coming back to the question of how we can learn (1.2.1) we can sample actions from a distribution of actions as in Section 1.1.2 assume U_t a distribution of controls with N trials $\{a_t\}_{i=1}^N \sim U_t$ we call the density of U_t, $\pi_t : \mathbb{R} \to \mathcal{P}(\mathcal{A})$ a mixed strategy or a randomized policy (with some abuse of notation) for each i of the N trials we can describe the evolution of the dynamics as

$$dX_t^i = b(X_t^i, a_t^i, \mu_t)dt + \sigma(X_t^i, a_t^i, \mu_t)dW_t^i, \quad X_0^i = x \quad \text{for } i = 1, ..., N,$$

notice that the environment is the same for each sampled action and the Brownian Motions independent, the formulation is completely analogous for the cost. As in the Section 1.1.2 we are interested in the average over the N samples that permits us to write the exploratory formulation

$$dX_t^\pi = \int_{\mathcal{A}} b(X_t^\pi, a, \mu_t)\pi(a)da\ dt + \int_{\mathcal{A}} \sigma(X_t^\pi, a, \mu_t)\pi(a)da\ dW_t, \quad X_0^\pi = x$$

and the corresponding value function

$$V(t,x) = \inf_{\pi \in \mathcal{P}(\mathcal{A})} \mathbb{E}\Big[\int_t^T \int_{\mathcal{A}} f\big(X_s^\pi, a, \mu_s\big)\pi(a)da\, ds \;\Big|\; X_t^\pi = x\Big]$$
$$V(T,X) = g(X_T^\pi, \mu_T)$$

with the flow $(\mu_t)_{0\leqslant t\leqslant T}$ fixed. In [62] the value function is regularised by Shannon's and cross entropy on the randomized action to enhance exploration. Then to solve the MFG MARL we need to employ a similar strategy as in the MFG case from the previous Section

1. Solve the optimization problem using an appropriate method (DPP or SMP) under the fixed environment μ and initial state x to obtain the optimal randomized policy π^*
2. Implement π^* for all agents and update the state of population to $\mu' = \mathcal{L}(X^{\pi^*})$
3. Repeat the previous steps until the distance between μ' and μ becomes smaller than some threshold.

1.2.2.2 Collaborative MARL: the Mean Field Markov Decision Process derived from an MFC

In accordance with the point of view of this thesis, we will present the case where we are already in the mean field regime and we will neglect how we arrived there, however this point is thoroughly discussed in [91, 36, 61, 11] . As explained in Section 1.1.4 MFCs are the ideal framework to study collaborative games since the whole population have a common objective, are symmetrical and nobody can individually affect the mean of the population (weak interactions). If the players of the game are fully aware of the model data b, σ, f, g then they can resort to techniques presented in Section 1.1 in order to compute the optimal controls (even if for the moment we have said nothing about the solvability of the master equation). On the other hand when the model is not available and only observations of states, actions, costs and possibly of the mean field term (distribution) are given to the agents then they can either reconstruct a model (**model based**) as we explained so far or try directly to learn the optimal values for corresponding state-action-mean field pairs (**model free**).

In this subsection we will present the MFC with learning using a model free approach since this is our main motivation for the third part of the thesis. We will discuss cases both with common noise and without since we can use one united framework to describe both.

The MFC with learning

As usual we start with our state space $\mathcal{X} \subseteq \mathbb{R}$ being a compact subset of the reals, the same for the action space $\mathcal{A} \subseteq \mathbb{R}$ and $\mathcal{P}(\mathcal{X})$ the space of probability measures on $\mathcal{X}$. The space of probability measures for the actions require a separate, more delicate, treatment and cannot simply be $\mathcal{P}(\mathcal{A})$ as Remark 3.4 from [91] informs us. Instead we set

$$D(\mu) = \{\mathcal{L}(x,a) \in \mathcal{P}(\mathcal{X} \times \mathcal{A}) \;:\; \mathcal{L}(x) = \mu \in \mathcal{P}(\mathcal{X}) \text{ and } a \in \mathcal{A}\},$$

in words, the set of all probability measures on $\mathcal{X} \times \mathcal{A}$ that have first marginal μ then we can set

$$\Gamma = \{(\mu, \alpha) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{A}) : \mu \in \mathcal{P}(\mathcal{X}) \text{ and } \alpha \in D(\mu)\}. \tag{1.2.2}$$

To avoid confusion we will follow the nomenclature: (μ, α) *mean field state and action* that always respect (1.2.2), (x, a) representative's state and action or simply *state action*. We would like our **random** policies to be measurable functions from state distributions to mean field action distributions, $\pi_n : \mathcal{P}(\mathcal{X}) :\to \mathcal{P}\big(\mathcal{P}(\mathcal{A})\big)$ and in particular for a mean field state action $(\mu, \alpha) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{A})$ at we demand $\mathcal{L}(\alpha|\mu) = \pi(\mu)$ and $\pi(\mu)(D(\mu)) = 1$ i.e. the policy gives all its weight on μ. For the deterministic policies we have $\pi_n : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{A})$ with $\pi_n(\mu) \in D(\mu)$.

Based on our description of the mean field state - action pair we define the instantaneous cost function

$$f(\mu, \alpha) = \int_{\mathcal{X}\times\mathcal{A}} \widetilde{f}(x, a, \alpha)\alpha(dx, da)$$

where $\widetilde{f}$ is the cost for the representative agent, and f the aggregate cost for the system. It is obvious that f is symmetric for all agents. We will assume also that she is bounded and continuous.

Next, we deal with transitions for our representative agent and the mean field system. For the representative agent we have

$$x_{n+1} = F(x_n, a_n, \mathcal{L}(x_n|\mathcal{F}_n^{\varepsilon^0,\xi}), \varepsilon_n, \varepsilon_n^0), \quad x_0 = \xi \sim \bar{\mu} \in \mathcal{P}(\mathcal{X}), \tag{1.2.3}$$

where $(\varepsilon_i)_{0\leqslant i\leqslant n}$ is sequence of idiosyncratic noise and $(\varepsilon_i^0)_{0\leqslant i\leqslant n}$ is one of common noise. Finally $\mathcal{L}(x_n|\mathcal{F}_n^{\varepsilon^0,x_0})$ is the conditional law given the randomness of the initial condition and common noise. We should notice that (1.2.3) is a discrete time implicit analog of (1.1.13). For a more precise definition in terms of the technical details we refer to Section 2.2 of [36] . For the transitions of the mean field state we have

$$\mu_{n+1} = \bar{F}(\mu_n, \alpha_n, \varepsilon_n^0), \quad \mu_0 = \bar{\mu}, \tag{1.2.4}$$

with $\bar{F}$ being the pushforward of the distribution of common noise through the measurable function F. In fact we should interpret 1.2.4 as the analogue of (1.1.14). The interpretation is actually the same, the common noise in the dynamics makes the flow of conditional distributions random. Of course this creates implications in the definitions that again we refer to [36, 91]. In short, with absence of common noise we have a deterministic Mean Field MDP in the spirit of [61].

Example 1.2.2 (Two stage stochastic model)**.** *1. The population start from a initial distribution $\mu_0 \in \mathcal{P}(\mathcal{X})$*

2. The central planner chooses $\alpha_0 \in \mathcal{P}(\mathcal{A})$

3. The population transitions to $\mu_1 = \bar{F}(\mu_0, \alpha_0, \varepsilon_0^0)$

4. Having observed μ_1 the central planner chooses α_1

The problem is to find a feedback function $\pi(\mu_0, \mu_1) = \big(\pi_0(\mu_0), \pi_1(\mu_1)\big)$ such that given $\bar{F}(\mu_0, \alpha_0, \varepsilon_0^0)$ and the second stage cost function $f(\mu_1, \pi_1(\mu_1))$ the central planer minimizes the following quantity for every μ_0, α_0

$$V^{\pi}(\mu_0) = \int_{\mathcal{P}(\mathcal{X})} f(\mu_1, \pi_1(\mu_1)) P(\mu_0, \alpha_0, d\mu_1), \tag{1.2.5}$$

Equivalently with (1.2.5) we can break Dynamic programming into two steps, first we choose

$$V_1(\mu_1) = \inf_{\alpha_1} \{ f(\mu_1, \alpha_1) \}$$

and then

$$V_2(\mu_0) = \inf_{\alpha_0} \{ \int_{\mathcal{P}(\mathcal{X})} V_1(\mu_1) P(\mu_0, \alpha_0, d\mu_1) \},$$

then we would like

$$\begin{aligned} \inf_{\pi} \int_{\mathcal{P}(\mathcal{X})} V^{\pi}(\mu_0) &= \inf_{\pi_0} \inf_{\pi_1} \int_{\mathcal{P}(\mathcal{X})} f(\mu_1, \pi_1(\mu_1)) P(\mu_0, \alpha_0, d\mu_1) \\ &= \inf_{\pi_0} \int_{\mathcal{P}(\mathcal{X})} \inf_{\alpha_1} f(\mu_1, \alpha_1) P(\mu_0, \alpha_0, d\mu_1) \\ &= V_2(\mu_0). \end{aligned} \tag{1.2.6}$$

Example 1.2.2 with formulation (1.2.6) helps us articulate several issues that we need to address in our framework.

1. The transition kernel P should be understood as the push-forward of the distribution of the common noise though the measurable function $\bar{F}$, which itself asks for $\bar{F}$ to be measurable with respect to the filtration of the common noise.
2. $\bar{F}, f$ needs to be measurable with respect to the σ-algebra that we equip $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{A})$ for mean field states and actions respectively in order for the integral to be well defined.
3. The $\inf\{f(\mu_1, \alpha_1)\}$ is not necessarily Borel-measurable even if f is
4. In order to be able to exchange the integral and the inf in (1.2.6) we need to be able to find Borel-measurable policies such that

$$f(\mu_1, \pi_\delta(\mu_1)) \leqslant \inf_{\alpha_1} f(\mu_1, \alpha_1) + \delta$$

Putting together $(\Gamma, f, \bar{F}, \gamma)$ we can define a Mean Field Markov Decision Process for the Collaborative MARL in the spirit of classical Markov Decision Process for RL, [104, Chapter 3]

To conclude this section we give the state value function for the MFMDP and a DPP.

$$V^{\pi}(\bar{\mu}) = \mathop{\mathbb{E}}_{\alpha \sim \pi} [\sum_{n=0}^{N} \gamma^n f(\mu_n, \alpha_n) \mid \mu_0 = \bar{\mu}] \tag{1.2.7}$$

and the optimal state value function as

$$V^{*}(\bar{\mu}) = \inf_{\pi \in \Pi} \{ V^{\pi}(\bar{\mu}) \} \tag{1.2.8}$$

for the DPP we can introduce the Bellman optimality operator

$$\mathcal{T}V(\bar{\mu}) = \inf_{\alpha_0 \in D(\bar{\mu})} \left\{ f(\bar{\mu}, \alpha_0) + \gamma \mathbb{E}[V(\bar{F}\big(\bar{\mu}, \alpha_0, \varepsilon_0^0\big)) \mid \mu_0 = \bar{\mu}] \right\} \tag{1.2.9}$$

We can prove that $\mathcal{T}$ is a contraction on the space of lower semicontinuous value functions with Borel measurable policies see [36]

1.3 Main Challenges

In this section, we would like to zoom out and come back to the broader picture we have drawn so far, in order to articulate some major challenges that have been the main motivation for this thesis. In Section 1.1.1 we started with some general challenges that RL face and here we would like to specify a bit further in relation with the models presented so far.

1.3.1 Curse of Dimensionality

It has been evident that the general case of N player continuous games is almost impossible to solve except in very specific cases, since we are dealing with a system of N coupled HJB equations. By solving usually in this thesis we mean **computing Nash equilibria**. Nevertheless we can still rely on the theory of MFGs - MFCs for the case of symmetrical weakly interacting agents. The major issue we face is solving numerically the master equation. This is challenging for two main reasons:

1. The equation is a PDE on the state of probability measures, involving derivatives of measures. The space itself might have nice properties, such that a form of differential calculus can be introduced but it lacks a Lebesgue measure that could make easier a standard theory of integration.

2. The space of probability measures is infinite dimensional and necessarily we need some finite dimensional approximation for probability measures since there is not computer that can run algorithms involving infinite dimensional quantities. This finite dimensional approximation can yield a problem that can be quite high dimensional, introducing a **discretisation cost** see Chapter 4

The same interpretation holds for MARL whether we take a model based approach where we write a master equation involving our approximated problem data b, σ, f or for a model free for iterations of the Bellman equation.

In a nutshell, the success of the mean field approach is in reducing the mathematical complexity of problems using the distribution of the agents but one way or another we have to make some choices when we design models to take into account the curse of dimensionality.

1.3.2 Efficient exploration

In order to learn the value function in any form of RL problem model based or model free we need to explore the state and action space, and we need to do it in an optimal way since usually

we cannot wait until we have fully explore to start exploiting our knowledge. Also in most real life cases exploration can have a "physical" cost that we don't necessarily take into account in simulated environments or we want to mix learning in simulated environments but incorporating online learning from real data as is most often seen in finance.

So far, in Sections 1.1.2, 1.2.2.1 we presented an exploratory stochastic control framework for RL problems that in literature so far appears to relay on entropy to randomize the choice of actions. In the case of Linear Quadratic models that we will mostly discuss in this thesis the addition of entropy destroys the convex structure of the cost and thus the uniqueness of the solution.

In contrast, we know from training of neural networks that artificial noise can be beneficial for training and robustness of the network see for example [122]. Similar ideas have already been extended to (non-mean field) deep reinforcement learning in [102]. Drawing inspiration from these works we propose our scheme of exploration by noise for the MFGs using common noise as an exploration noise. The main challenge in this approach is to find the right type of noise to add to the system. When we are dealing with Linear Quadratic models the choice of a finite dimensional additive noise is almost inevitable , however in more complicated models the choice of the appropriate noise is a serious challenge.

Another important issue that is connected to the efficient exploration is **uniqueness** of solution or Nash equilibria for competitive games and social (Pareto) optima for collaborative games. Especially in the case of MARL that this thesis is concerned we know that the rule is the existence of multiple equilibria that can potentially create instability of approximation methods. It is also known that uniqueness when relaxed controls (random policies) are used holds only under stringent convexity assumptions [78].

Last but not least, we could comment that various forms of regularization have been proposed either by entropy (to smoothen the max or min of actions) or to penalise the growth of value functions or convexify the cost (or reward) function.

1.3.3 Observability - population dependent / population agnostic policies

In classical control, the observability problem pertains to whether or not the states of a system can be fully determined or observed based on the information available from the system's output or measurements. In MARL there is one major question to be answered *Which information is available to agent to take decisions?*

The major benefit of MFGs and MFCs presented in Section 1.1 is that they simplify this information structure since the representative agent has only need to observe the mean field state of the population to design optimal actions on a given state. The optimal feedback functions of the mean field regime form an approximate ε-Nash equilibrium for the symmetric N-player game.

Formally, if we assume that each agent has *global* information we define

$$\bar{\alpha}_t^j = \frac{1}{N}\sum_{k\neq j}^{N}\delta_{\alpha_t^k}, \quad \alpha_t^k \sim \pi^k(\cdot|s_t, \bar{\mu}_t^k, \bar{\alpha}_{t-1}^k)$$

where $\bar{\mu}$ is the empirical mean state of the population, and $\bar{\alpha}$ of actions. However, this benefit is no longer granted if we assume that each agent has only *local* information and we come to the question raised in Section 1.2.2.1 of population dependent versus population agnostic policies.

1.3.4 Model based vs Model free Reinforcement Learning

Throughout this introduction, we exemplified both model based and model free RL methods to solve problems, here we would like to compare the two and articulate some possible shortcomings of each one that might make it more adapted to specific problems than other.

Model based

1. **Efficient Exploration**: Once the model is learned, the agent can use it for planning. It can simulate possible trajectories in the model and evaluate different actions and policies without interacting with the real environment. This makes the method more sample efficient reducing the need for a large number of real-world interactions. The learned model can be used to plan exploratory actions more efficiently.
2. **Stability**: Model-based methods can be more stable and require less fine-tuning, as they rely on known dynamics, which can lead to more predictable learning.
3. **Major Challenge: Model bias,** As explained in Section 1.1.2 building an accurate model can be challenging, especially in complex, high-dimensional, or unknown environments. Errors in the model can lead to suboptimal policies.

Model free

1. **Robustness to Model Errors**: Model-free methods are often more robust to errors in the model because they do not rely on a perfect representation of the environment dynamics.
2. **Simplicity and Applicability**: Model-free methods are relatively simple and can be applied to a wide range of RL problems without needing a model of the environment.
3. **Major Challenge: Exploration Efficiency,** Model-free approaches can require a large number of real-world interactions to learn a good policy, making them less sample-efficient compared to model-based methods.

1.4 The contributions of this thesis

At this point, after we explained what are the objects, the types of problems we aim at dealing with in this thesis and why we think they are important, it's high time to elaborate on our point of view for the way we will address the problems.

On the one hand we explained intuitively a mathematical approach to games with many symmetrical and weakly interacting players, collaborative or competitive. these represent obviously cases where the "model" is known and we can do computations based on it. However, the (theoretical) analysis that needs to be done and all the notions introduced so far e.g. HJB equation on the space of probability measures or infinite dimensional FBSDEs, appear naturally and correspond to "physical" objects that encode and decode hypotheses for problems coming from "real life".

On the other hand, starting from an algorithmic point of view, we aim to solve concrete problems, develop successful algorithms and then follow with a theory that generalizes or sets the foundations for a more systematic approach of a broader class of problems.

In some sense, we try to marry these two extremes and advocate for a systematic and rigorous approach to Mean Field Reinforcement Learning. This thesis reflects the effort to translate the intuition that we gain studying the theoretical aspects of the problem into actionable plans for algorithms or in general solution methods.

1.4.1 Towards explainable and trustworthy AI

The first contribution of this thesis is a conceptual one since using mathematics we bring transparency to optimal decision making in MARL.

Rigorous mathematical analysis provides foundations for **safe and explainable algorithms**. When we discuss Nash equilibria or social optima in MARL essentially we study where the system will converge and under which assumptions that rule out dangerous situations for humans, e.g. for floats of delivery drones to collide with humans or collide with themselves. Furthermore, by studying the error bounds of the algorithms we can define the precision we want to obtain when executing tasks by artificial agents.

The study of common noise, as part of the model or as artificial noise for exploration can enhance model's **robustness** in unexpected or unusual situations.

1.4.2 AI inspiring new solutions to mathematical problems

One key motivation for using common noise in our competitive MARL comes from inspection of a simple control problem with noise. The most straightforward thing we could do, when we are given a control problem would be to approximate the feedback function of the inputs. This approximation can be done by a neural network as it was first done by the seminal work of Han and E [67].

To make things more concrete we focus on a linear quadratic example, adopting the same notation as in Section 1.1.2

$$\begin{cases} \min\limits_{a\in\mathcal{A}_{adm}} \mathbb{E}\Big[\int_0^T \frac{1}{2}(X_t^2 + a_t^2)dt + \frac{1}{2}X_T^2\Big] \\ \text{subject to} \\ dX_t = a_t dt + \sigma dW_t, \quad X_0 = x \end{cases} \tag{1.4.1}$$

We know already the shape of the Value function and of the optimal feedback, that is going to be affine. In order to profit from the approximative power of Artificial Neural Networks (ANN) there are two possible strategies, either we approximate directly the controls in what is usually referred to as **Policy Iteration Methods** or approximate the value function that usually is called **Value Iteration Methods**. For the first part of the thesis we choose a Policy iteration method while later we switch for Value Iteration. The reasons for this choice will be clear soon.

In the case of (1.4.1) we can follow 2 paths to mimic the strategy of [67]

1. Suppose that we solve the problem in a class of affine functions and then try to learn the coefficients, η_t, h_t
$$a_t^\star = -\eta_t X_t^\star - h_t$$
2. Approximate the whole feedback function as an nonlinear function of the input.
$$a_t = \alpha(t, X_t)$$

For both strategies we discretise in time problem (1.4.1) and then use sequentially in time the states to pass them through the neural network (ANN) to synthesise the control and then calculate the cost as shown in Figure 1.4

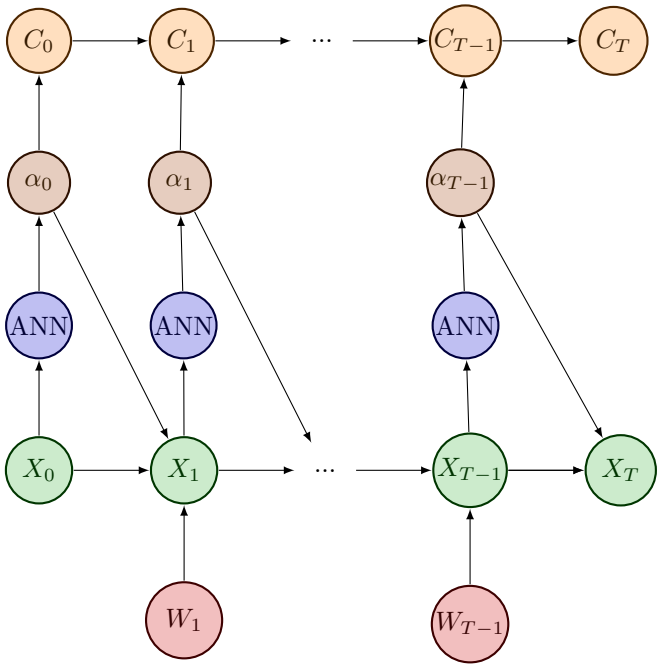


Figure 1.4: Sequential structure of Policy Iteration for Stochastic Control

The reader should note that by ANN we can use either strategy, postulating the affine form or approximating the whole feedback function. Now to train ANN[5] we need Monte Carlo samples that we denote by $\{X^i\}_{i=0}^M$ and in vector form with some abuse of notation the training should happen as in the Figure 1.5

The effect of the noise is very important for the method and it is a crucial observation for solving MFGs by Policy Iteration methods. As it will be more clear in the Introduction of Chapter 2 we need a noise that will randomize sufficiently the flow of (conditional) probability measures to be able to catch the fixed point.

[5]at this stage we don't assume any specific architecture and for illustration purposes we use fully connected feed forward NN. Furthermore, the network can be the same at each time step or different

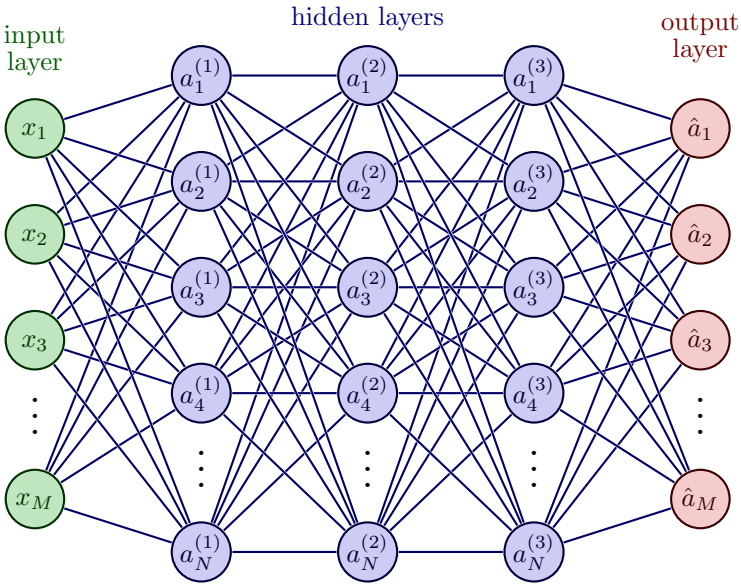


Figure 1.5: Feedforward NN for feedback controls

1.4.3 A novel Fictitious Play-type scheme for a model with common noise

Going back to our initial intention to tackle first competitive MARL we propose a novel Fictitious Play scheme for a Linear Quadratic variant of (1.2.1) with the law of the process replaced by a conditional mean given the common noise. We switch notation to agree more with Chapter 2 and let B_t be the idiosyncratic noise and W_t the common noise

$$\begin{cases} \min\limits_{a\in\mathcal{A}_{adm}} \mathbb{E}\Big[\int_0^T \frac{1}{2}\Big(\big(X_t + f(m_t)\big)^2 + a_t^2\Big)dt + \frac{1}{2}\big(X_T + g(m_T)\big)^2\Big] \\ \text{subject to} \\ dX_t = a_t dt + \sigma dB_t + \varepsilon dW_t, \quad X_0 = x \\ \text{with consistency condition} \\ m_t = \mathbb{E}[X_t|\mathcal{F}_t^W]. \end{cases} \tag{1.4.2}$$

Now again as in Sections 1.1.3, 1.2.2.1 our goal is to solve the optimal control when then m_t is fixed and then find a fixed point for the flow of conditional expectations. To this end we can transform (1.4.2) to an equivalent FBSDE system using Pontryagin's Maximum Principle and taking conditional expectation. Using the affine shape of the optimal control $a_t^\star = -\eta_t X_t + h_t$ we get

$$\begin{aligned} dm_t &= \big\{-\eta_t m_t - h_t\big\}dt + \varepsilon dW_t, \quad m_0 = \mathbb{E}[X_0] \\ dh_t &= \big\{-f(m_t) + \eta_t h_t\big\}dt + \varepsilon k_t dW_t, \quad h_T = g(m_T), \end{aligned}$$

where k_t is the process that makes h_t adapted to the filtration of Brownian Motion as briefly explained in Section 1.1.3.1 and Remark 1.1.9

Next we can introduce a shifted Brownian Motion that will help us define the our Fictitious Play

$$dW_t^{\boldsymbol{h}/\varepsilon} = \frac{1}{\varepsilon} h_t dt + dW_t,$$

and the probability measure $\mathbb{P}^h$ whose density is

$$\mathcal{E}(\frac{1}{\varepsilon}h) = \exp\Big(-\frac{1}{\varepsilon}\int_0^T h_t dW_t - \frac{1}{2\varepsilon^2}\int_0^T h_t^2 dt\Big).$$

Now, when m_t and h_t are two progressively-measurable processes with respect to the filtration generated by εW_t, we have a look at the new cost functional:

$$\begin{aligned} &\mathbb{E}^{\boldsymbol{h}/\varepsilon}\Big[\int_0^T \frac{1}{2}\Big(\big(X_t + f(m_t)\big)^2 + a_t^2\Big)dt + \frac{1}{2}\big(X_T + g(m_T)\big)^2\Big] \\ &\text{subject to} \\ &dX_t = a_t dt + \sigma dB_t + \varepsilon dW_t^{\boldsymbol{h}/\varepsilon}, \end{aligned} \tag{1.4.3}$$

where the common noise, $W_t^{\boldsymbol{h}/\varepsilon}$ is the shifted one. For reasons that will become clear when we want to compare the costs for the optimal control we also introduce the notation

$$J^\varepsilon(\boldsymbol{\alpha}; \boldsymbol{m}; \boldsymbol{h}) := \mathbb{E}^{\boldsymbol{h}/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}; \boldsymbol{m}; \varepsilon \boldsymbol{W}^{\boldsymbol{h}/\varepsilon}\big)\Big], \tag{1.4.4}$$

with

$$\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}; \boldsymbol{m}; \varepsilon \boldsymbol{W}^{\boldsymbol{h}/\varepsilon}\big) = \int_0^T \frac{1}{2}\Big(\big(X_t + f(m_t)\big)^2 + a_t^2\Big)dt + \frac{1}{2}\big(X_T + g(m_T)\big)^2 \tag{1.4.5}$$

We call (1.4.3) **original MFG problem characterized by the FBSDE system**

$$\begin{aligned} dm_t &= \big\{-\eta_t m_t - h_t\big\}dt + \varepsilon dW_t^{\boldsymbol{h}/\varepsilon}, \quad m_0 = \mathbb{E}[X_0] \\ dh_t &= \big\{-f(m_t) + \eta_t h_t\big\}dt + \varepsilon k_t dW_t^{\boldsymbol{h}/\varepsilon}, \quad h_T = g(m_T), \end{aligned} \tag{1.4.6}$$

to distinguish it from the one coming from our learning method.

We are now ready to introduce our first version of fictitious play as an adaptation of the one presented in Section 1.2.1 due to [24]. Consider the two step iterative learning procedure, whose description at rank n goes as follows:

Harmonic best action For a proxy $\overline{\boldsymbol{m}}^n := (\overline{m}_t^n)_{0\leqslant t\leqslant T}$ of the conditional mean $(m_t)_{0\leqslant t\leqslant T}$ of the in-equilibrium population and a proxy $\boldsymbol{h}^n := (h_t^n)_{0\leqslant t\leqslant T}$ of the opposite[6] of the ${\mathcal{F}_t}^W$-adapted intercept of the equilibrium feedback in $a_t^\star = -\eta_t X_t - h_t$, solve the stochastic control problem in the fixed environment $(\overline{m}_t^n, h_t^n)$ to obtain the best response $\boldsymbol{\alpha}^{n+1} = \alpha^*(t, x)$

[6]The opposite comes from the sign $-$ in the formula of the optimal feedback.

$$\begin{aligned}&\min_{a\in\mathcal{A}} \mathbb{E}^{\boldsymbol{h}^n/\varepsilon}\Big[\int_0^T \frac{1}{2}\Big(\big(x_t + f(\overline{m}_t^n)\big)^2 + a_t^2\Big)dt + \frac{1}{2}\big(x_T + g(\overline{m}_T^n)\big)^2\Big]\\ &\text{subject to}\\ &dx_t = a_t dt + \sigma dB_t + \varepsilon dw_t^{\boldsymbol{h}^n/\varepsilon}, \quad x_0 = x\end{aligned} \tag{1.4.7}$$

Harmonic update Given $\boldsymbol{h}^{n+1}$, the optimal trajectory of the above minimization problem is

$$X_t^{n+1} = X_0 - \int_0^t \big(\eta_s X_s^{n+1} + h_s^{n+1}\big)ds + \sigma B_t + \varepsilon W_t^{\boldsymbol{h}^n/\varepsilon}. \tag{1.4.8}$$

We then let

$$m_t^{n+1} = \mathbb{E}\big[X_t^{n+1} \,|\, \mathcal{F}_t^W\big], \tag{1.4.9}$$

together with

$$\overline{m}_t^{n+1} = \frac{1}{n+1}\sum_{k=1}^{n+1} m_t^k = \frac{1}{n+1} m_t^{n+1} + \frac{n}{n+1}\overline{m}_t^n. \tag{1.4.10}$$

The reader can identify the form of (2.1.22) as the standard fictitious play update rule introduced at Section 1.2.1 for the n^{th} round.

The rationale behind this strategy relies on Girsanov's transformation, to change dynamically the form of the common noise (with respect to the rank of the iteration of fictitious play). Precisely, this permits to decouple the two forward and backward equations, as clearly shown if we write the equation for $\overline{\boldsymbol{m}}^{n+1}$ as an equation with respect to the historical common noise. To see this we start from (2.1.20) changing back to the original common noise W_t and taking conditional expectation :

$$dm_t^{n+1} = -\big(\eta_t m_t^{n+1} + h_t^{n+1} - h_t^n\big)dt + \varepsilon dW_t, \quad m_0 = \mathbb{E}[X_0],$$

then we sum and divide by $n+1$ to arrive in

$$d\overline{m}_t^{n+1} = -\Big(\eta_t \overline{m}_t^{n+1} + \frac{1}{n+1}\big[h_t^{n+1} - h_t^0\big]\Big)dt + \varepsilon dW_t, \quad m_0 = \mathbb{E}[X_0]. \tag{1.4.11}$$

The forward equation then becomes asymptotically autonomous provided that $\boldsymbol{h}^{n+1}$ can be bounded independently of n, which we succeed to prove in Section 2. Following our standard strategy so far we can use Pontryagin's Maximum Principle for problem (2.1.19) to write a backward equation for $\boldsymbol{h}^n$

$$dh_t^{n+1} = \big\{-f(\overline{m}_t^n) + \eta_t h_t^{n+1}\big\}dt + \varepsilon k_t dW_t^{\boldsymbol{h}^n/\varepsilon}, \quad h_T^{n+1} = g(\overline{m}_T^n), \tag{1.4.12}$$

Then our first major results reads as follows

Theorem 1.4.1. *The **learning FBSDE*** (2.1.23)-(1.4.12) *converges to the decoupled **original FBSDE system** under the historical probability measure* $\mathbb{P}$*,*

$$\begin{aligned} dm_t &= \big\{-\eta_t m_t\big\}dt + \varepsilon dW_t, \quad m_0 = \mathbb{E}[X_0] \\ dh_t &= \big\{-f(m_t) + \eta_t h_t + k_t h_t\big\}dt + \varepsilon k_t dW_t, \quad h_T = g(m_T), \end{aligned} \tag{1.4.13}$$

with an explicit bound on rate of convergence

$$\mathbb{E}^{\boldsymbol{h}/\varepsilon}\Big[\sup_{0\leqslant t\leqslant T}\Big(\big|m_t - \overline{m}_t^n\big|^2 + \big|h_t - h_t^n\big|^2\Big)\Big] \leqslant C\frac{1}{n^2}\exp(\frac{C}{\varepsilon^2})$$

for a constant C that depends on the dimension, terminal time T, and the regularity of f and g.

In plain words, Theorem 1.4.1 addresses the *strong error* of the scheme under the measure $\mathbb{P}^{\boldsymbol{h}}$, we can also addresses the *weak error* of the scheme by comparing the law of the learning scheme with the law of the original MFG equilibrium

Theorem 1.4.2. *The weak error of the scheme for the Fortet-Mourier distance satisfies:*

$$\sup_F\Big|\mathbb{E}^{\boldsymbol{h}^n/\varepsilon}\Big[F\big(\overline{\boldsymbol{m}}^n, \boldsymbol{h}^n\big)\Big] - \mathbb{E}^{\boldsymbol{h}/\varepsilon}\Big[F\big(\boldsymbol{m}, \boldsymbol{h}\big)\Big]\Big| \leqslant \frac{1}{n}\exp\big(\tfrac{C}{\varepsilon^2}\big), \tag{1.4.14}$$

the supremum in the left-hand side being taken over all the functions F on $\mathcal{C}([0,T];\mathbb{R}\times\mathbb{R})$ that are 1-Lipschitz continuous and bounded by 1. The constant C depends on the same parameters as the one in Theorem 1.4.1

Observing closely both bounds of Theorems 1.4.1 and 1.4.2 we see that they depend dramatically on ε and they completely deteriorate when $\varepsilon \to 0$ which interests us in practice since we want our scheme to provide a selection of equilibria. Also in practice, for what is going to come next we want to be able to lower the intensity of our exploration noise once we have learned the system without destroying the structure of our method.

For these reasons we need to compensate in our scheme with a parameter that we can tune according to the intensity of noise meaning that away from zero it wouldn't cause problems and close to zero it would compensate for the blow-up of the bounds. It turns out that the most effective way to achieve this goal is to modify the learning rate of the scheme where we can pass from a **Harmonic** version in (2.1.22) to a **Geometric** one as follows

$$\overline{m}_t^{n+1} = \frac{\varpi(1-\varpi^{-1})}{1-\varpi^{-(n+1)}}\sum_{k=1}^{n+1}\varpi^{-k}m_t^k = \frac{\varpi^{-n}(1-\varpi^{-1})}{1-\varpi^{-(n+1)}}m_t^{n+1} + \Big(1 - \frac{\varpi^{-n}(1-\varpi^{-1})}{1-\varpi^{-(n+1)}}\Big)\overline{m}_t^n, \tag{1.4.15}$$

with the obvious convention

$$\frac{1}{n+1} = \frac{\varpi^{-n}(1-\varpi^{-1})}{1-\varpi^{-(n+1)}}.$$

For our compensation to be effective we need to scale with ϖ also the initial condition ϖX_0 and the intercept ϖh^n. The new tilted harmonic fictitious play, along with the rest of the details and the new bounds for the main results are presented in 2.

Now we are ready to introduce one of the major contributions of the thesis, restoration of a common noise for an MFG without common noise as action randomization that can help explore possible solutions of the game.

1.4.4 Exploration noise

Our set-up is going to be similar to the previous Section with the difference that our **original MFG** is the one **without common noise**, i.e.

$$\begin{cases} \min\limits_{a\in\mathcal{A}_{adm}} \mathbb{E}\Big[\int_0^T \frac{1}{2}\Big(\big(X_t+f(m_t)\big)^2+a_t^2\Big)dt+\frac{1}{2}\big(X_T+g(m_T)\big)^2\Big] \\ \text{subject to} \\ dX_t=\beta_t dt+\sigma dB_t, \quad X_0=x \\ \beta_t=a_t+\varepsilon\dot{W}_t \\ \text{with consistency condition} \\ m_t=\mathbb{E}[X_t|\mathcal{F}_t^W]. \end{cases} \tag{1.4.16}$$

To elaborate on our model we need to compare (1.4.2) with (1.4.16), in both cases the individual (representative) player choses her actions a_t and in case of (1.4.2) the resulting state dynamics are subjected to common noise while in (1.4.16) the action themselves are subject to noise. This major difference yields two major revisions that have to be done in our fictitious play scheme based on action randomization.

First, $\dot{W}_t$ is not differentiable and we need to consider a mollification of the noise. We consider a piecewise affine interpolation with time discretisation parameter p, for reasons that are explained in 2 so we end up with a time discretised p-MFG.

Second, $(a_t)_{0\leqslant t\leqslant T}\to(a_t+\varepsilon\dot{W}_t)_{0\leqslant t\leqslant T}$ turns the control into an unbounded process and even if we are able to solve any anticipativity and measurability issues by our piecewise interpolation we still need to modify the cost to keep everything in order.

One last comment about the comparison of (1.4.2) with (1.4.16) is the fact that we need to keep the same consistency condition for our MFG solutions otherwise the system will not feel the common noise, we have various comments along 2 for the cases when the common noise is zero in the **original MFG**. Figure 1.6 illustrates the relationship between **original MFG (without common noise)**, **learning MFGs (with common noise)** and the **approximate solution** that comes as output of the fictitious play algorithm

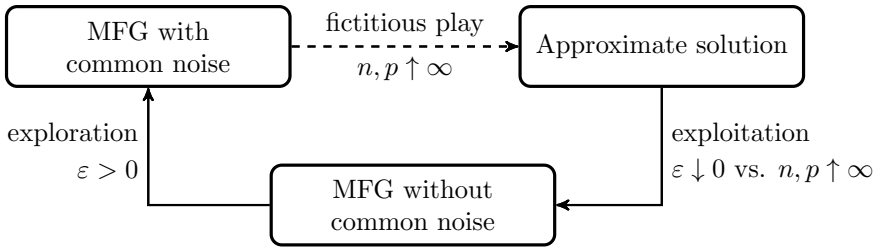


Figure 1.6: Exploration vs. exploitation.

To make it clear, we introduce a family of regular processes $((W_t^p)_{0\leqslant t\leqslant T})_{p\geqslant 1}$ such that, almost

surely (under $\mathbb{P}$),

$$\lim_{p\to\infty}\sup_{0\leqslant t\leqslant T}|W_t - W_t^p| = 0,$$

and, for any $p \geqslant 1$, the paths of $(W_t^p)_{0\leqslant t\leqslant T}$ are continuous and piecewise continuously differentiable and assume the following piecewise affine interpolation:

$$W_t^p := W_{\tau_p(t)} + \frac{p(t-\tau_p(t))}{T}\big(W_{\tau_p(t)+T/p} - W_{\tau_p(t)}\big), \tag{1.4.17}$$

where, by definition, $\tau_p(t) := \lfloor pt/T\rfloor(T/p)$. In words $\tau_p(t)$ is the unique element of $(T/p)\cdot\mathbb{N}$ such that $\tau_p(t) \leqslant t < \tau_p(t) + T/p = \tau_p(t + T/p)$, namely $\tau_p(t) = \ell T/p$, for $t \in [\ell T/p, (\ell+1)T/p)$ and $\ell \in \{0, \cdots, p-1\}$. The initialization is similar to the previous section $(m_t^0, h_t^0 = (\mathbb{E}[X_0], 0)_{-\leqslant t\leqslant T}$

For an $\mathcal{F}_t^W$-adapted and continuous environment $\boldsymbol{m} = (m_t)_{0\leqslant t\leqslant T}$, the cost functional is turned into the following discrete time version

$$\widetilde{J}^p(\boldsymbol{\alpha};\boldsymbol{m}) = \frac{1}{2}\mathbb{E}\bigg[\big|X_T + g(m_T)\big|^2 + \int_0^T\Big\{\big|X_{\tau_p(t)} + f(m_{\tau_p(t)})\big|^2 + |\alpha_t + \varepsilon\dot{W}_t^p|^2\Big\}dt\bigg], \tag{1.4.18}$$

where the expectation is taken over both the idiosyncratic and exploration (common) noises.

In the presence of the randomization, the cost functional might become very large as p tends to ∞, even for a control $\boldsymbol{\alpha}$ of finite energy. This prompts us to renormalise $\widetilde{J}^p$. As a result of the adaptability constraint, we indeed have

$$\mathbb{E}\int_0^T \alpha_t\cdot\dot{W}_t^p dt = 0, \quad \text{and} \quad \mathbb{E}\int_0^T |\dot{W}_t^p|^2 dt = d\,p,$$

So, we must subtract to the cost a diverging term to recover the original cost functional. The effective cost must be equation (1.4.18) $-\frac{1}{2}\varepsilon^2 d\,p$, so we recover the right cost

$$\begin{aligned} J^p(\boldsymbol{\alpha};\boldsymbol{m}) :=& \widetilde{J}^p(\boldsymbol{\alpha};\boldsymbol{m}) - \tfrac{1}{2}\varepsilon^2 d\,p \\ =& \frac{1}{2}\mathbb{E}\bigg[\big|X_T + g(m_T)\big|^2 + \int_0^T\Big\{\big|X_{\tau_p(t)} + f(m_{\tau_p(t)})\big|^2 + |\alpha_{\tau_p(t)}|^2\Big\}dt\bigg]. \end{aligned} \tag{1.4.19}$$

We can define again the shifted Brownian Motion as in the previous subsection, let $\boldsymbol{W}^{\boldsymbol{h}/\varepsilon}$:

$$W_t^{\boldsymbol{h}/\varepsilon} = W_t + \frac{1}{\varepsilon}\int_0^t h_s ds.$$

We compute the p-piecewise linear interpolation $\boldsymbol{W}^{p,\boldsymbol{h}/\varepsilon}$ of $\boldsymbol{W}^{\boldsymbol{h}/\varepsilon}$:

$$W_t^{p,\boldsymbol{h}/\varepsilon} = W_{\tau_p(t)}^{p,\boldsymbol{h}/\varepsilon} + \int_{\tau_p(t)}^t dW_s^p + \frac{1}{\varepsilon}\int_{\tau_p(t)}^t h_s ds,$$

from which we deduce that

$$W_t^{p,\boldsymbol{h}/\varepsilon} = W_t^p + \frac{1}{\varepsilon}\int_0^t h_s ds, \quad t\in[0,T]. \tag{1.4.20}$$

For the sake of brevity we will omit the rest of the details to define properly our Fictitious Play, it can be found in the corresponding Section 2.2.2. Instead we give the main theorem for the convergence of the learning scheme to the original (decoupled) MFG system and the exploitability of the output of the algorithm for defining the approximate solution.

Theorem 1.4.3. *There exists a threshold $c > 0$, depending on d, T, the norms $\|f\|_{1,\infty}$ and $\|g\|_{1,\infty}$ such that, for $p\varepsilon^2 \geqslant c$, the scheme*

$$\begin{cases} h^{n+1}_{\ell T/p} = \widetilde{h}^{n+1}_{(\ell+1)T/p} + \frac{T}{\varpi p} f(\overline{m}^n_{(\ell+1)T/p}) \mathbf{1}_{\{\ell \leqslant p-2\}} \\ \qquad - \frac{T}{p}\Big(\eta^{(p)}_{(\ell+1)T/p} + \frac{T}{p}\mathbf{1}_{\{\ell\leqslant p-2\}}\Big)\widetilde{h}^{n+1}_{\ell T/p} - \varepsilon \int_{\ell T/p}^{(\ell+1)T/p} k^{n+1}_s dW^{\varpi \boldsymbol{h}^n/\varepsilon}_s, \\ m^{n+1}_{(\ell+1)T/p} = \mathbb{E}\big[X^{(p),n+1,\varpi}_{\ell T/p} \,|\, \sigma(\boldsymbol{W})\big] = \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\big[X^{(p),n+1,\varpi}_{\ell T/p} \,|\, \sigma(\boldsymbol{W})\big], \\ \text{together with} \\ \overline{m}^{n+1}_t = \frac{\varpi(1-\varpi^{-1})}{1-\varpi^{-(n+1)}} \sum_{k=1}^{n+1} \varpi^{-k} m^k_t = \frac{\varpi^{-n}(1-\varpi^{-1})}{1-\varpi^{-(n+1)}} m^{n+1}_t + \Big(1 - \frac{\varpi^{-n}(1-\varpi^{-1})}{1-\varpi^{-(n+1)}}\Big)\overline{m}^n_t. \end{cases} \tag{1.4.21}$$

converges to $(\boldsymbol{m}^{(p)}, \widetilde{\boldsymbol{h}}^{(p)}/\varpi, \boldsymbol{k}^{(p)}/\varpi)$, where $(\boldsymbol{m}^{(p)}, \widetilde{\boldsymbol{h}}^{(p)}, \boldsymbol{k}^{(p)})$ is the unique solution of the decoupled discrete-time FBSDE system:

$$\begin{aligned} m^{(p)}_{(\ell+1)T/p} &= m^{(p)}_{\ell T/p} - \tfrac{T}{p}\eta^{(p)}_{\ell T/p} m^{(p)}_{\ell T/p} + \varepsilon\big(W_{(\ell+1)T/p} - W_{\ell T/p}\big), \\ \widetilde{h}^{(p)}_{\ell T/p} &= \widetilde{h}^{(p)}_{(\ell+1)T/p} + \tfrac{T}{p} f\big(m^{(p)}_{(\ell+1)T/p}\big)\mathbf{1}_{\{\ell\leqslant p-2\}} \\ &\quad - \tfrac{T}{p}\Big(\eta^{(p)}_{(\ell+1)T/p} + \tfrac{T}{p}\mathbf{1}_{\{\ell\leqslant p-2\}}\Big)\widetilde{h}^{(p)}_{\ell T/p} - \Big(\int_{\ell T/p}^{(\ell+1)T/p} k^{(p)}_s ds\Big)\widetilde{h}^{(p)}_{\ell T/p} \\ &\quad - \varepsilon \int_{\ell T/p}^{(\ell+1)T/p} k^{(p)}_s dW_s, \quad \ell \in \{0, \cdots, p-1\}, \\ m^{(p)}_0 &= \mathbb{E}(X_0), \quad \widetilde{h}^{(p)}_T = g\big(m^{(p)}_T\big), \end{aligned} \tag{1.4.22}$$

with an explicit bound on the rate of convergence, namely

$$\mathrm{essup}_{\omega\in\Omega}\Big[\sup_{\ell=0,\cdots,p}\Big(|m^{(p)}_{\ell T/p} - \overline{m}^n_{\ell T/p}|^2 + |\varpi^{-1}\widetilde{h}^{(p)}_{\ell T/p} - h^n_{\ell T/p}|^2\Big)\Big] \leqslant \varpi^{-2n}\exp\big(C\varepsilon^{-2}\big), \tag{1.4.23}$$

for a constant C that also depends on d, T, $\|f\|_{1,\infty}$, $\|g\|_{1,\infty}$.

Moreover, if we extend $\boldsymbol{m}^{(p)}$ by continuous interpolation to the entire $[0,T]$ and if we call $\boldsymbol{h}^{(p)}$ the piecewise constant extension of $\widetilde{\boldsymbol{h}}^{(p)}$ to the entire $[0,T]$, then, up to a modification of the constant C, the weak error of the scheme for the Fortet-Mourier distance satisfies

$$\sup_F \Big|\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[F\big(\overline{\boldsymbol{m}}^n, \boldsymbol{h}^n\big)\Big] - \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[F\big(\boldsymbol{m}^{(p)}, \varpi^{-1}\boldsymbol{h}^{(p)}\big)\Big]\Big| \leqslant \varpi^{-n}\exp\big(C\varepsilon^{-2}\big), \tag{1.4.24}$$

the supremum being taken over all the functions F on $\mathcal{C}([0,T];\mathbb{R}^d\times\mathbb{R}^d)$ that are bounded by 1 and 1-Lipschitz continuous.

We come now to the discussion on the *exploitability* of the output of the learning scheme. It refers to the extent to which an agent can take advantage of the behaviors of other agents in the environment to achieve its own objectives. Exploitability becomes relevant when considering how well an agent can adapt and respond to the strategies employed by its peers. An agent with low exploitability is less susceptible to being taken advantage of by others, as it can effectively anticipate and counter their actions. Reducing exploitability is a common goal in the design of MARL algorithms, as it contributes to the stability and effectiveness of multi-agent systems.

To make things precise we consider the strategy $\boldsymbol{\alpha}^{(p),n,\diamond}$ defined by

$$\alpha_t^{(p),n,\diamond} := -\eta^{(p)}_{\tau_p(t)} X^{(p),\star}_{\tau_p(t)} - \varpi h^n_t, \tag{1.4.25}$$

where we recall

$$\begin{aligned} dX_t^{(p),\star} &= -\eta^{(p)}_{\tau_p(t)} X^{(p),\star}_{\tau_p(t)} dt + \sigma dB_t + \varepsilon dW^p_t \\ &= \alpha_t^{(p),n,\diamond} dt + \sigma dB_t + \varepsilon dW^{p,\varpi \boldsymbol{h}^{n/\varepsilon}}_t, \quad X_0^{(p),n,\diamond} = X_0. \end{aligned} \tag{1.4.26}$$

we have also

$$\mathbb{E}^{\varpi \boldsymbol{h}^{n/\varepsilon}}\Big[X_t^{(p),\star} \,|\, \sigma(\boldsymbol{W}) \Big] = \mathbb{E}\Big[X_t^{(p),\star} \,|\, \sigma(\boldsymbol{W}) \Big] = m_t^{(p)}. \tag{1.4.27}$$

Theorem 1.4.4. *Assume that the law of the initial condition has sub-Gaussian tails, i.e.* $\mathbb{P}(\{|X_0| \geqslant r\}) \leqslant a^{-1}\exp(-ar^2)$, *for some* $a > 0$ *and for any* $r > 0$. *Then, there exist two positive constants* c *and* C, *only depending on the parameters* a, d, T, $\|f\|_{1,\infty}$, $\|g\|_{1,\infty}$, *such that, for any* $\varepsilon \in (0,1]$, *any integer* $p \geqslant 1$ *satisfying* $p\varepsilon^2 \geqslant c$ *and any integer* $n \geqslant 1$,

$$\begin{aligned} \inf_{\boldsymbol{\alpha}} \mathbb{E}^{\varpi \boldsymbol{h}^{n/\varepsilon}}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}; \boldsymbol{m}^{(p)}; 0\big) \Big] &\geqslant \mathbb{E}^{\varpi \boldsymbol{h}^{n/\varepsilon}}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}^{(p),n,\diamond}; \boldsymbol{m}^{(p)}; 0\big) \Big] \\ &\quad - C\varepsilon - Cp^{-1} - \exp(C\varepsilon^{-2})\varpi^{-n}, \end{aligned} \tag{1.4.28}$$

the infimum in the left-hand side being taken over $\mathcal{F}_t^{X_0,\boldsymbol{B},\boldsymbol{W}}$*-progressively measurable and square-integrable processes* $(\boldsymbol{\alpha}_t)_{0\leqslant t\leqslant T}$.

The difference between the term in the left-hand side and the first term in the right-hand side of (2.2.81) *is called the* $\mathbb{P}^{\varpi \boldsymbol{h}^{n/\varepsilon}}$*-mean exploitability of the policy* $\boldsymbol{\alpha}^{(p),n,\diamond}$. *(Obviously, this difference is non-positive.)*

We would like to make the following comments to clarify Theorem 2.2.27

Remark 1.4.5. *Both the costs of* $(\boldsymbol{\alpha}, \boldsymbol{m}^{(p)})$ *and* $(\boldsymbol{\alpha}^{(p),n,\diamond}, \boldsymbol{m}^{(p)})$ *are computed according to the time-continuous original model without common noise (even though the control* $\boldsymbol{\alpha}^{(p),n,\diamond}$ *is piecewise constant and random as an output of the scheme).*

Remark 1.4.6. *The bound that is given for the mean exploitability depends on the three parameters* ε, n *and* p. *Typically, we want to choose* ε *small and* p *large, which is well understood: the equilibrium* $(\boldsymbol{m}^{(p)}, \boldsymbol{\alpha}^{(p),n,\diamond})$ *is learnt for the* p*-discrete MFG with an* ε*-common noise; if* ε *is large or* p *is small, the equilibrium that is learnt cannot be a 'good' approximate equilibrium of the original MFG. This is exactly what the terms* $-C\varepsilon$ *and* $-C/p$ *say in* (2.2.81). *As for the last term in* (2.2.81), *it becomes smaller and smaller as* n *increases.*

1.4.5 Quantitative Bounds for Kernel Based Q-Learning

Now we shift our attention to a model-free approach, namely Q-learning. Our ultimate goal being to address learning a Mean Field MDP in the spirit of 1.2.2.2 we will develop first some bounds that help us to analyse the exploration-exploitation tradeoff.

We start with our state and actions spaces $\overline{S}$ and $\overline{A}$ respectively, being compact subsets of $\mathbb{R}^d$ with dimensions $d_{\overline{S}}$ and $d_{\overline{A}}$, $D = d_{\overline{S}} + d_{\overline{A}}$. Switching to a reward instead of a cost, we assume a reward function $R : S \times A \to \mathbb{R}$ that is bounded and has bounded derivatives of order D. Assume also a transition kernel $\overline{\mathsf{P}} : \overline{S} \times \overline{A} \to \mathcal{P}\big(\overline{S}\big)$, where we require the mapping $\overline{\mathsf{P}}$ to be measurable (which means here that $(s,a) \mapsto \overline{\mathsf{P}}((s,a),E)$ is measurable for the standard Borel σ-fields and for any $E \in \mathcal{B}(S)$).

Assumptions (Cost and Transition Kernel).

1. *The function R is bounded and has bounded derivatives of any order up to $5(\lfloor d_S/2 \rfloor + 1)$.*
2. *For any given $h > 0$, there exists $\eta, \eta' > 0$ such that, for any balls B_S of $\mathbb{R}^{d_S}$ and B_A of $\mathbb{R}^{d_A}$, of radius greater than h each and respectively included in $\overline{S}$ and $\overline{A}$,*
$$\eta h^D \leqslant \mathbb{P}\Big(\big\{\big(\mathrm{s}_{n+1}, \mathrm{a}_{n+1}\big) \in B_S \times B_A\big\} \,|\, \mathcal{F}_n\Big) \leqslant \eta' h^D. \tag{1.4.29}$$
3. *For any function $\varphi : \overline{S} \to \mathbb{R}$, whose derivatives up to a certain order $k \in \{1, \cdots, 5(\lfloor d_S/2 \rfloor + 1)\}$ are bounded by a certain constant C, the function*
$$(s,a) \in \overline{S} \times \overline{A} \mapsto \int_{\overline{S}} \varphi(s') \overline{\mathsf{P}}\big((s,a), ds'\big)$$
also has bounded derivatives up to k, with bounds only depending on C.

Remark 1.4.7. *The assumptions are to guarantee the contraction for the Bellman operator that will be defined later.*

Remark 1.4.8. *The order $5(\lfloor d_S/2 \rfloor + 1)$ is here found from Sobolev embedding theorems, which play a great role in the proof of our final result especially when we compare the distance between MDPs and which draw a clear connection between the dimension of the state and action spaces and the required regularity on the data.*

A policy π is a measurable function $\pi : \overline{S} \to \overline{A}$. Under a policy π, the value of a state s is

$$V^\pi(s) := \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k R\big(\mathrm{s}_k, \pi(\mathrm{s}_k)\big) \;\big|\; \mathrm{s}_0 = s\Big],$$

where $(\mathrm{s}_n)_{n \geqslant 0}$ is the Markov chain associated with the transition kernel $s \in \overline{S} \mapsto \overline{\mathsf{P}}((s, \pi(s)), \cdot) \in \mathcal{P}(\overline{S})$, and the action value function of a pair $(s,a) \in \overline{S} \times \overline{A}$ is

$$Q^\pi(s,a) := R(s,a) + \mathbb{E}\Big[\sum_{k=1}^{\infty} \gamma^k R\big(\mathrm{s}_k, \pi(\mathrm{s}_k)\big) \;\big|\; \mathrm{s}_0 = s\Big],$$

with the optimal ones being respectively

$$V^\star(s) := \sup_\pi V^\pi(s), \quad \text{and} \quad V^\star(s) = \sup_{a\in\overline{A}} Q^\star(s,a), \quad \forall s \in \overline{S},$$

with

$$Q^\star(s,a) = R(s,a) + \gamma \int_{\overline{S}} V^\star(s')\overline{\mathsf{P}}\big((s,a), ds'\big), \quad \forall (s,a) \in \overline{S}\times\overline{A}.$$

In order to learn the optimal value we can repeat the usual strategy we have mentioned throughout this introduction, Dynamic programming. In RL this is usually known as value iteration and for discrete time, space problem over a finite horizon it can be solved explicitly by iterations of the form

$$V_n^\star(s) = \sup_{a\in\overline{A}}\Big\{ R(s,a) + \gamma\mathbb{E}[V_{n+1}^\star(\mathrm{s}_{n+1}) \mid (\mathrm{s}_n, \mathrm{a}_n) = (s,a)]\Big\}, \quad \text{for } \ n = N-1, ..., 1, 0,$$

starting from a terminal value V_N. When the time horizon is infinite as in our case we let $N \to \infty$ and get $V^\star = T^\star V^\star$ for T being the Bellman operator

$$\begin{aligned} \big(\mathcal{T}^\pi U\big)(s) &= R\big(s,\pi(s)\big) + \gamma \int_{\overline{S}} U(s')\overline{\mathsf{P}}\big((s,\pi(s), ds'\big), \\ \big(\mathcal{T}^\star U\big)(s) &= \sup_\pi\big\{\mathcal{T}^\pi U(s)\big\}, \end{aligned} \tag{1.4.30}$$

where U is a test function from $\overline{S}$ to $\mathbb{R}$.

Assumptions (Cost and Transition Kernel). *We assume that the value function has bounded derivatives of order up to* $5(\lfloor d_S/2\rfloor + 1)$.

Except for discrete time-space this strategy is also very much dependent on the dimension, since as the spaces grow larger so does the computational requirements, in a phenomenon called "curse of dimensionality" that we have mentioned several times already. For the method to become feasible we need

1. Temporal Difference (TD) learning
2. Function approximation to deal with the continuous nature of the underlying spaces.

We refrain from giving all the details here and instead refer to the classical textbook [104] and the Introduction of Chapter 3

For the TD part we make the choice for Q-Learning algorithm [115, 104], central to this method is the TD-update or *target* that we use to update the value function. Our target is $R(s,a) + \gamma \sup_{a'\in\overline{A}} Q(\mathrm{s}'[s], a')$, where $\mathrm{s}'[s]$ is the random state that is reached from s and is being sampled from the distribution $\overline{\mathsf{P}}((s,a),\cdot)$. Then, the update rule takes the form

$$Q^{n+1}(s,a) = Q^n(s,a) + \alpha\bigg(R(s,a) + \gamma \sup_{a'\in\overline{A}} Q^n(\mathrm{s}'[s], a') - Q^n(s,a)\bigg), \tag{1.4.31}$$

where α is a learning rate in $(0,1)$.

For the function approximation we choose kernel regression [69, Chapter 6] . The main idea is that instead of discretising the spaces over some grid, we can use kernel to average the information of the nearby points. In fact kernel regression is a smooth variant of k-nearest neighbours algorithm where we define a kernel that provides the smooth averaging of our data points. Central to this method is the choice of kernel, since it dictates the shape based on which we select the points that influence our estimator and its size which we denote as h, called bandwidth.

To explain the idea a bit further, say that we interact with the (unknown) environment and obtain a sample $(\underline{\mathrm{s}},\underline{\mathrm{a}})_n = \{\mathrm{s}_0,\mathrm{a}_0,\mathrm{s}_1,...,\mathrm{s}_n,\mathrm{a}_n\}$ of size n, we want to define $\mathscr{K} : \mathbb{R}^{d_S}\times\mathbb{R}^{d_A}\to\mathbb{R}$ a smooth non-negative compactly supported function such that based on our sample $(\underline{\mathrm{s}},\underline{\mathrm{a}})_n$ and for any function $f : \overline{S}\times\overline{A}\to\mathbb{R}$ at a point (s,a) at depth n we can define the operation

$$\mathcal{A}_{h,n}f(s,a) := \sum_{k=0}^{n}\frac{\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)f(\mathrm{s}_k,\mathrm{a}_k)}{\sum_{k=0}^{n}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)}, \tag{1.4.32}$$

provided the denominator is not zero and if the denominator is zero, we return 0 for $\mathcal{A}_{h,n}f(s,a)$, more details an motivation about the choice and the use of $\mathcal{A}_{h,n}$ see Introduction of Chapter 3, for examples of regression kernels see [69, Chapter 6]. For the moment what is important for us to motivate our choice is that $\mathcal{A}_{h,n}$ preserves the smoothness of the functions it is applied to f and the contraction property of (1.4.30) when used on top of $\mathcal{T}$.

To apply the kernel approximation on the action value function we notice the recursion

$$\mathcal{A}_{h,n}f(s,a) = \alpha_n(s,a)f(\mathrm{s}_n,\mathrm{a}_n) + \big(1-\alpha_n(s,a)\big)\mathcal{A}_{h,n-1}f(s,a), \tag{1.4.33}$$

with

$$\alpha_n(s,a) = \begin{cases} \dfrac{\mathscr{K}_h\big(s-\mathrm{s}_n,a-\mathrm{a}_n\big)}{\sum_{j=0}^{n}\mathscr{K}_h\big(s-\mathrm{s}_j,a-\mathrm{a}_j\big)} & \text{if} \quad \displaystyle\sum_{j=0}^{n}\mathscr{K}_h\big(s-\mathrm{s}_j,a-\mathrm{a}_j\big) > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{1.4.34}$$

then we adapt the update rule (1.4.31) and let (for $n\geqslant 1$)

$$\widehat{Q}_n(s,a) = \widehat{Q}_{n-1}(s,a) + \alpha_{n-1}(s,a)\Big(R(\mathrm{s}_{n-1},\mathrm{a}_{n-1}) + \gamma\mathrm{sup}_{a'\in A}\widehat{Q}_{n-1}(\mathrm{s}_n,a') - \widehat{Q}_{n-1}(s,a)\Big). \tag{1.4.35}$$

Thus, we arrive in

$$\widehat{Q}_n(s,a) = \begin{cases} \dfrac{\sum_{j=0}^{n-1}\mathscr{K}_h\big(s-\mathrm{s}_j,a-\mathrm{a}_j\big)y_j}{\sum_{j=1}^{n-1}\mathscr{K}_h\big(s-\mathrm{s}_j,a-\mathrm{a}_j\big)} & \text{if} \quad \sum_{j=1}^{n-1}\mathscr{K}_h\big(s-\mathrm{s}_j,a-\mathrm{a}_j\big) > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{1.4.36}$$

where

$$y_j = R(\mathrm{s}_j,\mathrm{a}_j) + \gamma\sup_{a\in\overline{A}}\widehat{Q}_j(\mathrm{s}_{j+1},a),$$

is the TD-target.

Remark 1.4.9. *We can refine the upper boundedness assumption of the transition kernel. We let $\rho > 1$ be the smallest real such that the support of $\mathscr{K}$ is in $B(o, \rho) = \rho B(0, 1)$, we assume that there exists another constant $\eta' > 1$ such that, for any D-dimensional ball of radius $3\rho h$,*

$$\mathbb{P}\Big(\big\{\big(\mathrm{s}_{n+1}, \mathrm{a}_{n+1}\big) \in B_{2\rho h}\big\} \,|\, \mathcal{F}_n\Big) \leqslant \eta' h^D. \tag{1.4.37}$$

To wrap-up everything, we present a pseudo code of the algorithm

Algorithm 1: Kernel Based Q-Learning

input : Type of kernel $\mathscr{K}$, bandwidth h, discount γ, number of iteration n
output: Approximate Action Value Function $\widehat{Q}_n$

initialization:;
$\widehat{Q}_0(s, a) = 0 \ \forall (s, a)$;
set inital state s_0;
for k *in* n **do**
  Choose action a_k;
  Get: reward r_k, next state s';
  Compute $y_k = r_k + \gamma \max_{a \in \overline{A}} \widehat{Q}_k(s', a)$ using (1.4.35) ;
  Store in memory (s_k, a_k, y_k);
  $k = k + 1$;
  $s_k = s'$;

Given the nature of memory based approximations, essentially all we need is just successive iterations $\mathrm{s}_k, \mathrm{a}_k, \mathrm{y}_k, \mathrm{s}_{k+1}, \cdots$

Our main result is the rate of convergence, $\widehat{Q}_{h,n}(s, a)$ to $Q^\star(s, a)$ in probability when n becomes large and h small, for all s, a in $\overline{S} \times \overline{A}$

Theorem 1.4.10. *There exists a constant C, depending on the various parameters underpinning the aforementioned assumptions, such that, for an error threshold $\varepsilon > 0$, we can find h_ε and n_ε such that the* sup *distance between $\widehat{Q}_{h_\varepsilon, n_\varepsilon}$ and $Q^\star$ is less than $C(1 + |\ln(\varepsilon)|)\varepsilon$ on an event of probability greater than $1 - C\varepsilon^D$. The number n_ε that is necessary to do so is less than $\exp(C|\ln(\varepsilon)|^3)$.*

To highlight our contribution and explain our result we provide a short description of the proof strategy. The result is based on a proof device, the Action Replay Process, first conceived by Watkins [115], which is an artificial MDP constructed based on the original one in such a way that for n big enough the transitions and expected rewards of the two MDPs are close. For the sake of brevity we will omit all technical details and focus on the intuition, in Section 3.2.2 we provide all the necessary details.

To introduce ARP, assume we have a realization of the original MDP, $(\underline{\mathrm{s}}, \underline{\mathrm{a}})_{n \in \mathbb{N}}$[7], then we construct a homogeneous Markov process $(\Lambda_k, \Sigma_k, B_k)_{k \in \mathbb{N}}$ on an auxiliary probability space $(\Xi, \mathcal{G})$ with values in $\mathbb{N} \times \overline{S} \times \overline{A}$ such that the first component denotes the level at which the process starts and the transitions only allow the time component to decrease. For the transitions assume that on a given level $(n, s, a) \in \mathbb{N} \times \overline{S} \times \overline{A}$ we flip a "bias" coin and with probability $\alpha_n(s, a)$ as in (1.4.34),

[7] in fact, the realization should include rewards, but for simplicity we omit them

we accept the level and record it, i.e. we get reward r_n while with probability $1-\alpha_n(s,a)$ we discard the level, decrease the time component by one and flip the bias coin for the next level $(n-1,s',a')$. This process is repeated until the time component reaches zero where the ARP terminates at an arbitrary state that might differ at each realization.

Formally we call $\mathbf{P}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}$ the probability measure under which, the process $(\Lambda_k,\Sigma_k,B_k)_{k\in\mathbb{N}}$ is a homogeneous Markov chain with transition probabilities $\big(\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\big((n,s,a),\cdot\big)_{n\in\mathbb{N},s\in\overline{S},a\in\overline{A}}$, namely

$$\mathbf{P}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\Big(\Big\{(\Lambda_{k+1},\Sigma_{k+1},B_{k+1})\in E\Big\}\,|\,\mathcal{G}_k^{(\Lambda,\Sigma,B)}\Big)=\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\big((\Lambda_k,\Sigma_k,B_k),E\big),\tag{1.4.38}$$

for E a Borel subset of $\mathbb{N}\times\overline{S}\times\overline{A}$, where $\mathbb{G}^{(\Lambda,\Sigma,B)}=(\mathcal{G}_k^{(\Lambda,\Sigma,B)})_{k\in\mathbb{N}}$ is the filtration generated by $(\Lambda_k,\Sigma_k,B_k)_{k\in\mathbb{N}}$. The reader can find the precise description of $\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}$ describing the " biased coin flips" in Section 3.2.2 and also in [114].

To conclude the definition of ARP as an MDP we give its state and action value functions

$$V^{\mathrm{ARP},\star}(n,s)=\sup_{\pi}\mathbf{E}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\Big[\mathbf{1}_{\{\tau\geqslant 1\}}\sum_{k=0}^{\infty}\gamma^k R\big(\Sigma_k,\pi(\Lambda_k,\Sigma_k)\big)\,|\,(\Lambda_0,\Sigma_0)=(n,s)\Big],$$

with the supremum being taken over time dependent strategies π from $\mathbb{N}\times\overline{S}$ into $\overline{A}$. The variable τ should be implicitly understood as the ARP having at least one transition before termination. Finally, following Lemma (3.2.2) of Section 3.2.2 the expectation $\mathbf{E}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}$ should be understood as

$$\mathbf{E}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\Big[R\big(\Lambda_1,\Sigma_1,B_1\big)\,|\,(\Lambda_0,\Sigma_0,B_0)=(n,s,a)\Big]=\sum_{k=0}^{n-1}\frac{\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)R\big(k,\mathrm{s}_{k+1},\mathrm{a}_{k+1}\big)}{\sum_{l=1}^{n}\mathscr{K}_h\big(s-\mathrm{s}_l,a-\mathrm{a}_l\big)}.$$

The optimal action-value function $Q^{\mathrm{ARP},\star}$ is the solution of

$$\begin{aligned}Q^{\mathrm{ARP},\star}\big(n,s,a\big)&=R(s,a)\mathbf{P}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\big(\{\tau\geqslant 1\}\,|\,(\Lambda_0,\Sigma_0,B_0)=(n,s,a)\big)\\&\quad+\gamma\,\mathbf{E}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\Big[\sup_{a'\in\bar{A}}Q^{\mathrm{ARP},\star}\big(\Lambda_1,\Sigma_1,a'\big)\,|\,(\Lambda_0,\Sigma_0,B_0)=(n,s,a)\Big],\end{aligned}\tag{1.4.39}$$

Since our description of the ARP is complete we can articulate the steps we take to prove our main theorem and highlight our contributions

1. **ARP and kernel based action value function have the same optimal values**.

2. **Maximal inequalities for the covering times of ARP**. We cover $\overline{S}\times\overline{A}$ by J subsets $B_1,...,B_J$ called cells, the covering time is the time the MDP need to visit all cells. Let T_k be the first time after ℓ when each set has been visited at least k times by the process $(\mathrm{s}_n,\mathrm{a}_n)_{n\geqslant 0}$:

$$T_k=\min\left\{n\geqslant\ell:\min_{1\leqslant j\leqslant J}\sum_{i=\ell}^{n}\mathbf{1}_{B_j}\big(\mathrm{s}_i,\mathrm{a}_i\big)\geqslant k\right\},$$

with the convention that $T_0=\ell-1$. We use the coupon collector to obtain upper and lower bounds for T_1. *This is a novel interpretation used in our proof.*

3. **If ARP starts at a level ℓ_n the probability of going below ℓ_0 is small**. Identifying the event where that happens is crucial for the final convergence in probability because intuitively when n increases the ARP has sufficiently many levels to learn the transitions of the original MDP. To see this consider the ARP making the transition from level n at state s to E a Borel subset of $\overline{S}$ using a, at a level m that is

$$\overline{\Pi}_{(\underline{s},\underline{a})}\Big((n,s,a),E\Big) = \sum_{m=0}^{n-1} \overline{\Pi}_{(\underline{s},\underline{a})}\Big((n,s,a),(m,E)\Big)$$

so at the aforementioned event we can concentrate our attention on $\overline{\Pi}_{(\underline{s},\underline{a})}\Big((n,s,a),ds'\Big)$ instead of looking at every transition of the ARP.

4. **Control the distance between the transition kernels of ARP and MDP on a Sobolev space of smooth test functions**. In fact, we don't need just the supremum over all test functions, but also the supremum over all initial point s,a, which makes the dimension of the space coming into play. To preserve the rate of convergence in higher dimensions we demand extra smoothness from our test functions. *The novelty of our contribution here lies in the rigorous treatment of the distance between MDPs and in tracking the influence of the choice of tests functions to the final rate of convergence* .

1.4.6 Application to finite state MFMDP

We come now to a mean field toy model with finite states to apply the result of the previous section.

Suppose that $\overline{S}$ and $\overline{A}$ are finite sets with 3 nodes each. Then we identify the space of probability measures $\mathcal{P}(\overline{S})$ with the 3-dim probability simplex $\mathbb{S}^3$. In the mean field limit, as explained already in Section 1.2.2.2 we are dealing with a "lifted" MDP, set on the space probability measures $\mathcal{P}(\overline{S}) \equiv \mathbb{S}^3$, i.e. a continuous space that fits the description of the previous section. If the transitions of the representative agent are given by

$$\begin{aligned} &X_{n+1} = F(X_n, \mu_n, \alpha_n, U^1_{n+1}, U^0_{n+1}) \\ &\text{with} \\ &\mu_n = \mathcal{L}(X_n \,\big|\, U^0_n) \end{aligned} \tag{1.4.40}$$

where we denote $(U^1_k)_{k\geqslant 0}$ and $(U^0_k)_{k\geqslant 0}$ two independent sequences of iid random variables from a Uniform over $[0,1]$, with U^0 being a common noise.

Now let us describe some possible models for transition from (μ, a) to $E \subset \mathbb{S}^3$

1. Suppose that X_{n+1} is distributed conditional on the common noise, according to a random measure $\mu \in \mathbb{S}^3$ that is sampled uniformly.

2. Suppose a sequence $(\varepsilon^0_n)_{n\geqslant 1}$, independent of the two sequences $(U^0_n)_{n\geqslant 1}$ and $(U^1_n)_{n\geqslant 1}$, such that each ε^0_n is Bernoulli(p) distributed for $p \in (0,1)$. Then, we may consider the case where, with probability $1-\varepsilon_n$, the agent follows the 'original' dynamics (1.4.40) but without common noise, i.e. $X_{n+1} = F(X_n, \mu_n, \alpha_n, U_{n+1})$ and with probability ε_0, it is resampled uniformly on the whole space.

3. Suppose $\mu \in \mathbb{S}^3$ and Z^0 random variable with a smooth density φ such that $\mu + Z^0$ to belong in the prob simplex, then X_{n+1} is resampled according to $\mu + Z^0$.

Now, for a measurable kernel $\overline{\mathsf{P}}^0 : \mathbb{S}^3 \times \mathbb{S}^3 \to \mathcal{P}(\mathbb{S}^3)$ that would correspond to the dynamics of a mean field (1.4.40), we can construct the following (new) kernel $\overline{\mathsf{P}}$ (by means of the same principles as those underpinning the aforementioned examples 1 and 3):

$$\overline{\mathsf{P}}\big((\mu,\alpha),E\big) = p\frac{\lambda(E)}{\lambda(\overline{S})} + (1-p)\int_{\mathbb{R}^{d_S}}\left(\int_{\mathbb{S}^3} 1_E\big(q+(1-q)z+y\big)\overline{\mathsf{P}}^0\big((\mu,\alpha),dz\big)\right)\varphi(y)dy, \quad (1.4.41)$$

where λ the Lebesgue measure and $p \in (0,1)$. One may start from $\overline{\mathsf{P}}^0$ and then consider $\overline{\mathsf{P}}$ as a randomized version of it, just used for the purpose of learning.

For this example, the Bellman equation reads as

$$\begin{aligned} V(\mu) &= \sup_{\alpha\in\mathbb{S}^3}\Big[R(\mu,\alpha) + \gamma\int_{\mathbb{S}^3} V(\mu')\overline{\mathsf{P}}\big((\mu,\alpha),d\mu'\big)\Big] \\ &= \sup_{\alpha\in\mathbb{S}^3}\Big[R(\mu,\alpha) + \gamma(1-p)\int_{\mathbb{R}^{|\mathscr{S}|}}\left(\int_{\mathbb{S}^3} V\big(q+(1-q)z+y\big)\overline{\mathsf{P}}^0\big((\mu,\alpha),dz\big)\right)\varphi(y)dy\Big] \\ &\quad + \frac{\gamma p}{\lambda(\mathbb{S}^3)}\int_{\mathbb{S}}^3 V(\mu')d\mu'. \end{aligned}$$

There may be several types of conditions under which the value function V is regular in μ (which is a prerequisite in our main theorem). In particular we show in the end of Chapter 3 how we can apply our result to learn the value function associated with a transition kernel $\overline{\mathsf{P}}^0$.

Chapter 2

Exploration noise for learning linear quadratic Mean Field Games

2.1 Introduction

2.1.1 General context.

Since their inception fifteen years ago in the seminal works of Lasry and Lions [80, 81, 82], Lions [85] and Huang et al. [72, 74, 73], mean field games (MFGs) have met a tremendous success, inspiring mathematical works from different areas like partial differential equations (PDEs), control theory, stochastic analysis, calculus of variation, toward a consistent and expanded theory for games with many weakly interacting rational agents. Meanwhile, the increasing number of applications has stimulated a long series of works on discretization and numerical methods for approximating the underlying equilibria (which we also call solutions); see for instance Achdou et al. [2, 1], Achdou and Capuzzo-Dolcetta [3] and Achdou and Laurière [5] for discretization with finite differences; Carlini and Silva [28, 29] for semi-Lagrangian schemes; Benamou and Carlier [17] and Bricenõ Arias et al. [9, 8] for variational discretization; and Achdou and Laurière [4] for an overview. These contributions often include numerical methods for solving the discrete schemes such as the Picard method, Newton method, fictitious play. We review the latter one in detail in the sequel. Recently, other works have also demonstrated the possible efficiency of tools from machine learning within this complex framework: standard equations for characterizing the equilibria may be approximately solved by means of a neural network; see for instance Carmona and Laurière [33, 34]. Last (but not least), motivated by the desire to develop methods for models with partially unknown data, several authors have integrated important concepts from reinforcement learning in their studies; we refer to Carmona et al. [35, 36], Elie et al. [54] and Guo et al. [63].

The aim of our work is to make one new step forward in the latter direction with a study at the frontier between the theoretical analysis of MFGs and the development of appropriate forms of learning. In particular, our main objective is to provide a proof of concept for the notion of exploration noise, which is certainly a key ingredient in machine learning. For the reader who is aware of the MFG theory, our basic idea is to prove that common noise may indeed serve for the exploration of the space of solutions and, in the end, for the improvement of the existing

learning algorithms. For sure, this looks a very ambitious program since there has not been, so far, any complete understanding of the impact of the common noise onto the shape of the solutions. However, several recent works clearly indicate that noise might be helpful for numerical purposes. Indeed, in a series of works, Bayraktar et al. [15], Delarue [47], and Foguen Tchuendom [56], it was shown that common noise could help to force uniqueness of equilibria in a certain number of MFGs. This is a striking fact because nonuniqueness is the typical rule for MFGs, except in some particular classes with some specific structure; see for instance the famous uniqueness result of Lasry and Lions [80] for games with monotonous coefficients and Carmona and Delarue [30, chapter 3]. Conceptually, the key condition for forcing uniqueness is that, under the action of the (hence common) noise, the equilibria can visit sufficiently well the state space in which they live. This makes the whole rather subtle because, in full theory, the state space is the space of probability measures, which is typically infinite-dimensional. In this respect, the main examples for which such a smoothing effect has been rigorously established so far are (*i*) a linear quadratic model with possibly nonlinear mean field interaction terms in the cost functional (Foguen Tchuendom [56]), (*ii*) a general model with an infinite dimensional noise combined with a suitable form of local interaction in the dynamics (Delarue [47]); and (*iii*) models defined on finite state spaces and forced by a variant of the Wright-Fisher noise used in population genetics literature (Bayraktar et al. [15]).

In order to prove the possible efficiency of our approach, we here restrict the whole discussion to the first of the three aforementioned instances. We provide below a long informative introduction in which we present the model in detail together with the related literature and our own results. The guideline of this introduction is the following. We specify the model in Subsection 2.1.2, both without and with common noise. In particular, we recall therein the various known characterizations of the equilibria in both cases. In Subsection 2.1.3, we provide a brief review of the existing learning procedures and we exemplify the form of the so-called fictitious play for the linear-quadratic model studied in the paper. The thrust of our paper is to introduce a variant of the fictitious play, which we refer to as the *tilted fictitious play* and whose convergence can be proven in the presence of common noise under more general conditions than those of the standard fictitious play (within the class of linear-quadratic MFGs introduced in Subsection 2.1.2). This tilted fictitious play is defined in Subsection 2.1.4. Although the tilted fictitious play can be regarded as a theoretical learning scheme, we explain in Subsection 2.1.5 how it can be adapted to statistical learning. Notably, in this adaptation, the common noise serves as an exploration noise for learning the equilibria of the underlying MFG. This interpretation of the common noise in terms of exploration is in fact one key point of the paper. Exploitation is then briefly addressed in Subsection 2.1.6. Therein, we present the main bounds that we are able to prove for the mean exploitability of the policy that is returned by the tilted fictitious play. Finally, we provide an overview of our numerical results in Subsection 2.1.7, and we conclude the introduction by providing a comparison with the existing literature in Subsection 2.1.8. The main assumption and notation are given in Subsection 2.1.9. The organization of the paper is also presented in Subsection 2.1.9.

2.1.2 Our model.

In this subsection, we expose the linear-quadratic class of MFGs that is addressed in the paper, without and with common noise. We insist in both cases on the form of the equilibrium feedbacks.

As made clear in (2.1.5), those feedbacks are necessarily affine. Moreover, for any of them, the intercept in the affine structure is characterized by a backward stochastic differential equation (BSDE), which is recalled in (2.1.6) and which plays a key role in the subsequent analysis.

To make it clear, the MFG that we consider below comprises the state equation

$$\mathrm{d}X_t = \alpha_t \mathrm{d}t + \sigma \mathrm{d}B_t + \varepsilon \mathrm{d}W_t, \quad t \in [0,T], \tag{2.1.1}$$

together with the cost functional

$$J(\boldsymbol{\alpha};\boldsymbol{m}) = \frac{1}{2}\mathbb{E}\Big[\big|RX_T + g(m_T)\big|^2 + \int_0^T \Big\{\big|QX_t + f(m_t)\big|^2 + |\alpha_t|^2\Big\}\mathrm{d}t\Big]. \tag{2.1.2}$$

Above, X_t is the state at time t of a representative agent evolving within a continuum of other agents. It takes values in $\mathbb{R}^d$, for some integer $d \geqslant 1$, and evolves from the initial time 0 to the terminal time T according to the equation (2.1.1). Therein, $\boldsymbol{B} = (B_t)_{0\leqslant t\leqslant T}$ and $\boldsymbol{W} = (W_t)_{0\leqslant t\leqslant T}$ are two independent d-dimensional Brownian motions on a complete probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and $\sigma \geqslant 0$ and $\varepsilon \geqslant 0$ account for their respective intensity in the dynamics. Whereas $\boldsymbol{B}$ is thought as a private (or idiosyncratic) noise felt by the representative player only (and not by the others), the process $\boldsymbol{W}$ is said to be a *common* or *systemic* noise as all the others in the continuum also feel it. The initial condition (also called initial private state) X_0 may be random and, in any case, X_0 is independent of the pair $(\boldsymbol{B},\boldsymbol{W})$. The process $\boldsymbol{\alpha} = (\alpha_t)_{0\leqslant t\leqslant T}$ is a so-called control process, which is usually assumed to be progressively-measurable with respect to the augmented filtration $\mathbb{F}$ generated by $(X_0, \sigma\boldsymbol{B}, \varepsilon\boldsymbol{W})$ (when σ or ε are zero, the corresponding process is no longer used to generate the filtration). The key fact in MFG theory is that the representative player aims at choosing the best possible $\boldsymbol{\alpha}$ in order to minimize the cost functional J in (2.1.2). The leading symbol $\mathbb{E}$ in the definition of J is understood as an expectation with respect to all the inputs $(X_0, \boldsymbol{B}, \boldsymbol{W})$. If there were no dependence on the extra term $\boldsymbol{m} = (m_t)_{0\leqslant t\leqslant T}$ in f and g, the minimization of J would reduce to a mere linear-quadratic stochastic control problem driven by Q and R, with the latter being two d-square matrices[1]. The essence of MFGs is that $(m_t)_{0\leqslant t\leqslant T}$ accounts for the flow of marginal statistical states of the continuum of players surrounding the representative agent. In full generality, each m_t should be regarded as a probability measure hence describing the statistical distribution of the other agents at time t. For simplicity, we here just assume m_t to be the d-dimensional mean of the other agents at time t; consistently, $f = (f^1, \cdots, f^d)$ and $g = (g^1, \cdots, g^d)$ are $\mathbb{R}^d$-valued functions defined on $\mathbb{R}^d$. However, the notion of *mean* should be clarified, because of the distinction between the two *private* and *common* noises. From a modelling point of view, this *mean* should result from a law of large numbers taken over players that would be subjected to independent and identically distributed (i.i.d.) initial and private noises (consistently with our former description of $\boldsymbol{B}$) but to the same common noise $\boldsymbol{W}$. Because of this, $(m_t)_{0\leqslant t\leqslant T}$ is itself required to be a stochastic process, progressively-measurable with respect to the filtration generated by $\varepsilon\boldsymbol{W}$. When $\varepsilon = 0$, the filtration becomes trivial and, accordingly, $(m_t)_{0\leqslant t\leqslant T}$ is assumed to be deterministic. The notion of Nash equilibrium or MFG solution then comes through a fixed point argument. In short, $(m_t)_{0\leqslant t\leqslant T}$ is said to be an equilibrium if the minimizer $(X_t^\star)_{0\leqslant t\leqslant T}$

[1] We could take Q and R as $e \times d$ matrices, for a general $e \geqslant 1$ and then f and g as being e-dimensional. For simplicity, we feel easier to work with $e = d$.

of (2.1.1)–(2.1.2) satisfies:

$$m_t = \mathbb{E}\big[X_t^\star \,|\, \varepsilon \boldsymbol{W}\big], \quad t \in [0,T], \tag{2.1.3}$$

with the conditional expectation becoming an expectation if $\varepsilon = 0$ (whence our choice to use $\varepsilon \boldsymbol{W}$ as random variable in the conditioning, as this notation allows us to distinguish easily between the two cases without and with common noise). Throughout, f and g are typically assumed to be bounded and Lipschitz continuous functions. When $\varepsilon = 0$, this is enough for ensuring the existence of solutions to the fixed point (2.1.3), but uniqueness is known to fail in general, except under some additional conditions. For instance, the Lasry-Lions monotonicity condition, when adapted to this setting, says that uniqueness indeed holds true if $Q^\dagger f$ and $R^\dagger g$ (with $\dagger$ denoting the transpose) are non-decreasing in the sense that (see [45])

$$\forall x, x' \in \mathbb{R}^d, \quad (x - x') \cdot \big(Q^\dagger f(x) - Q^\dagger f(x')\big) \geqslant 0, \quad (x - x') \cdot \big(R^\dagger g(x) - R^\dagger g(x')\big) \geqslant 0, \tag{2.1.4}$$

with $\cdot$ denoting the standard inner product in $\mathbb{R}^d$. When $\varepsilon > 0$, existence and uniqueness hold true, even though the coefficients are not monotone (see [56]). This is a very clear instance of the effective impact of the noise onto the search of equilibria.

The reason why the MFG (2.1.1)–(2.1.2)–(2.1.3) becomes uniquely solvable under the action of the common noise may be explained as follows. In short (and this is the rationale for working in this set-up), the linear-quadratic structure of (2.1.1) forces uniqueness of the minimizer to (2.1.2) when $\boldsymbol{m}$ is fixed. Even more the optimal control is given in the Markovian form

$$\alpha_t^\star = -\eta_t X_t^\star - h_t, \quad t \in [0,T], \tag{2.1.5}$$

where $\boldsymbol{\eta} = (\eta_t)_{0 \leqslant t \leqslant T}$ is the $d \times d$-matrix valued solution of an *autonomous* Riccati equation that only depends on Q and R and that is in particular independent of the input $\boldsymbol{m}$. In other words, only the intercept $(\boldsymbol{h}_t)_{0 \leqslant t \leqslant T}$ in the above formula depends on the inputs f, g, R, Q and $\boldsymbol{m}$. The characterization of $\boldsymbol{h}$ is usually obtained by the (stochastic if $\varepsilon > 0$) Pontryagin principle, namely $\boldsymbol{h}$ solves the Backward Stochastic Differential Equation (BSDE)

$$h_t = R^\dagger g(m_T) + \int_t^T \big\{Q^\dagger f(m_s) - \eta_s h_s\big\} \mathrm{d}s - \varepsilon \int_t^T k_s \mathrm{d}W_s, \quad t \in [0,T]. \tag{2.1.6}$$

When $\varepsilon = 0$, the above stochastic integral disappears and the equation (2.1.6) becomes a mere Ordinary Differential Equation (ODE) but set backwards in time. When $\varepsilon > 0$, the solution is the pair $(\boldsymbol{h}, \boldsymbol{k})$, which is required to be progressively measurable with respect to the filtration generated by $\boldsymbol{W}$. Existence and uniqueness are well-known facts in BSDE theory. By taking the conditional mean given $\boldsymbol{W}$ in (2.1.1), we deduce that solving the MFG problem thus amounts to find a pair $(\boldsymbol{m}, \boldsymbol{h})$ satisfying the forward-backward stochastic differential equation (FBSDE):

$$\begin{aligned} \mathrm{d}m_t &= -\big(\eta_t m_t + h_t\big)\mathrm{d}t + \varepsilon \mathrm{d}W_t, \quad m_0 = \mathbb{E}(X_0), \\ \mathrm{d}h_t &= -\big(Q^\dagger f(m_t) - \eta_t h_t\big)\mathrm{d}t + \varepsilon k_t \mathrm{d}W_t, \quad h_T = R^\dagger g(m_T). \end{aligned} \tag{2.1.7}$$

Unique solvability was proven in [44, 86]. The smoothing effect of the noise manifests at the level of the related system of Partial Differential Equations (PDEs), which is sometimes called the *master equation* of the game:

$$\partial_t \theta_\varepsilon(t,x) + \tfrac{\varepsilon^2}{2} \Delta_{xx}^2 \theta_\varepsilon(t,x) - \big(\eta_t x + \theta_\varepsilon(t,x)\big) \cdot \nabla_x \theta_\varepsilon(t,x) + Q^\dagger f(x) - \eta_t \theta_\varepsilon(t,x) = 0, \tag{2.1.8}$$

with $\theta_\varepsilon(T,\cdot) = R^\dagger g(\cdot)$ as a boundary condition at the terminal time, θ_ε being a function from $[0,T]\times\mathbb{R}^d \to \mathbb{R}^d$. When $\varepsilon > 0$, the PDE is uniformly parabolic and hence has a unique classical solution (with bounded derivatives). When $\varepsilon = 0$, it becomes an hyperbolic equation and singularities may emerge, precisely when the solutions to (2.1.7) (which are nothing but the characteristics of (2.1.8)) cease to be unique.

2.1.3 Learning procedures

We now provide an overview of the existing methods that can be used for solving the standing MFG numerically, or at least for decoupling the two forward and backward equations in (2.1.7). In particular, we spend some time here recalling the definition of the so-called *fictitious play*, see (2.1.9) and (2.1.10).

In our setting, numerical solutions to the MFG may be found by solving either the FBSDE (2.1.7) or the nonlinear equation (2.1.8). For sure, independently of any applications to MFGs, there have been well-known numerical methods for the two objects, see for instance (and for a tiny example) Beck et al. [16], Bender and Zhang [18], Cvitanić and Zhang [42], Delarue and Menozzi [46], E et al. [53], Huijskens et al. [75], Ma et al. [87], and Milstein and Tretyakov [90]. In Delarue and Menozzi [46], Huijskens et al. [75], Ma et al. [87], and Milstein and Tretyakov [90], the problem is solved by constructing an approximation of θ_ε by means of a backward induction. This is a typical strategy in the field, which is fully consistent with the dynamic programming principle that holds true for uniquely solvable mean field games (see [31, chapter 4]). Although formulated differently, Bender and Zhang [18] also relies on a backward induction. In comparison with Bender and Zhang [18], Delarue and Menozzi [46], Huijskens et al. [75], Ma et al. [87], and Milstein and Tretyakov [90], the works Beck et al. [16], Cvitanić and Zhang [42], and E et al. [53] proceed in a completely different manner because the backward equation therein is reformulated into a forward equation with an unknown initial condition; the goal is then to tune both the initial condition and the martingale representation term of the backward equation in order to minimize, at terminal time, the distance to the prescribed boundary condition. Those methods have the following main limitation within a learning prospect for MFGs: they require the coefficients f, g, Q and R entering the model to be explicitly known. Even more, they make no real use of the control structure (2.1.2) that underpins the game.

In fact, Delarue and Menozzi [46], Huijskens et al. [75], Ma et al. [87], and Milstein and Tretyakov [90] suffer from another drawback because all these works involve a space discretization that consists in approximating the function θ_ε at the nodes of a spatial grid. Accordingly, the complexity increases with the physical dimension d of the state variable. In particular, similar strategies would become even more costly for a more general mean field dependence than the one addressed in (2.1.2). Indeed, in the general case, the spatial variable is no longer the mean but the whole statistical distribution (which is an infinite dimensional object). For sure, the latter raises challenging questions that go beyond the scope of this paper because we cannot guess of a numerical method that would directly allow to handle the infinite dimensional statistical distribution of the solution. However, this is an objective that should be kept in mind. Say for instance that particle or quantization methods would be natural candidates to overcome such an issue, see for instance Chaudru de Raynal and Garcia Trillos [100] and Crisan and McMurray [41].

Whatever the method, the true difficulty is to decouple (2.1.7), because the two forward and

backward time directions are conflicting. One of our basic concern here is thus to take benefit of the noise in order to define an iterative scheme that decouples (2.1.7) efficiently. At the same time, because the very structure of a MFG corresponds very naturally to the precepts of reinforcement learning, we want to have a method that may work with unknown coefficients f, g, Q and R, and that may benefit from the observation of the cost if it is available. In this regard, we ask our scheme to have a learning structure that should manifest in a sequence of steps of the form

1. computation of a best action,

2. update of the state variable,

and hence that would be adapted to real data. Surprisingly, this is not an easy question, even though the equation (2.1.7) is well-posed. As demonstrated in the recent work [38], naive Picard iterations may indeed fail. They would consist in solving inductively the backward equation

$$\mathrm{d}h_t^{n+1} = -\big(Q^\dagger f(\overline{m}_t^n) - \eta_t h_t^{n+1}\big)\mathrm{d}t + \varepsilon k_t^{n+1}\mathrm{d}W_t, \quad h_T^{n+1} = R^\dagger g(\overline{m}_T^n). \tag{2.1.9}$$

for a given proxy $\overline{\boldsymbol{m}}^n := (\overline{m}_t^n)_{0\leqslant t\leqslant T}$ of the forward equation and then in plugging the solution $\boldsymbol{h}^{n+1} := (h_t^{n+1})_{0\leqslant t\leqslant T}$ into the forward equation

$$\mathrm{d}\overline{m}_t^{n+1} = -\big(\eta_t \overline{m}_t^{n+1} + h_t^{n+1}\big)\mathrm{d}t + \varepsilon \mathrm{d}W_t, \quad m_0^{n+1} = \mathbb{E}(X_0). \tag{2.1.10}$$

Intuitively, the reason why it may fail is that the updating rule $\overline{\boldsymbol{m}}^n \mapsto \overline{\boldsymbol{m}}^{n+1}$ is too ambitious. In other words, the increment may be too high and smaller steps are needed to guarantee the convergence of the procedure.

A more successful strategy is known in MFG theory (and more generally in game theory) under the name of *fictitious play*, see [24, 54, 59, 65, 66, 97, 96]. It is a learning procedure with a decreasing harmonic step of size $1/n$ at rank n of the iteration. Having as before a proxy $\overline{\boldsymbol{m}}^n$ for the state of the population at rank n of the learning procedure, $\boldsymbol{h}^{n+1}$ is computed as above. Next, the same forward equation as before is also solved, but the solution is denoted by $\boldsymbol{m}^{n+1}$, namely

$$\mathrm{d}m_t^{n+1} = -\big(\eta_t m_t^{n+1} + h_t^{n}\big)\mathrm{d}t + \varepsilon \mathrm{d}W_t, \quad m_0^{n+1} = \mathbb{E}(X_0), \tag{2.1.11}$$

and then the updating rule is given by

$$\overline{m}_t^{n+1} = \frac{1}{n+1} m_t^{n+1} + \frac{n}{n+1}\overline{m}_t^n, \quad t \in [0,T]. \tag{2.1.12}$$

Very importantly, both the Picard iteration and the fictitious play have a learning interpretation. In both cases, the backward equation (2.1.9) provides the best response $\boldsymbol{\alpha}^{n+1,\star}$ to the minimization of the functional $\boldsymbol{\alpha} \mapsto J(\boldsymbol{\alpha}; \overline{\boldsymbol{m}}^n)$, which has indeed the same form as in (2.1.5):

$$\alpha_t^{n+1} = -\eta_t X_t^{n+1} - h_t^{n+1}, \quad t \in [0,T],$$

where $(X_t^{n+1,\star})_{0\leqslant t\leqslant T}$ is obtained implicitly by solving the corresponding state equation (2.1.1). Equivalently, the above formula gives the form of the optimal feedback function to the minimization of the functional $\boldsymbol{\alpha} \mapsto J(\boldsymbol{\alpha}; \overline{\boldsymbol{m}}^n)$ (bearing in mind that the feedback function may be here random because of the common noise).

In fact, the fictitious play has been addressed so far in the following two main cases: potential MFGs ([24]) and MFGs with monotone coefficients (here, $Q^\dagger f$ and $R^\dagger g$ are non-decreasing; see [54, 59, 65, 66, 96] within a general setting). Also, except in the recent work [96] (whose framework is a bit different because the fictitious play is in continuous time), the analysis has just been carried out in the case $\varepsilon = 0$, i.e., when there is no common noise. Here, it is certainly useful to recall that potential MFGs are a class of MFGs for which there exists a potential, namely a functional $\mathcal{J}(\boldsymbol{\alpha})$ associated with the same state dynamics as in (2.1.1)–(2.1.2), such that any minimizer of $\mathcal{J}$ is a solution of the MFG. In our setting, the shape of $\mathcal{J}(\boldsymbol{\alpha})$ is

$$\mathcal{J}\big(\boldsymbol{\alpha}\big) = \mathbb{E}\bigg[\mathcal{G}\big(\mathcal{L}(X_T|\varepsilon\boldsymbol{W})\big) + \int_0^T \frac{1}{2}\Big[\mathcal{F}\big(\mathcal{L}(X_t|\varepsilon\boldsymbol{W})\big) + |\alpha_t|^2\Big]\mathrm{d}t\bigg], \tag{2.1.13}$$

where $\mathcal{L}(X_t|\varepsilon\boldsymbol{W})$ denotes the conditional law of X_t given the common noise and

$$\mathcal{G}(\mu) = \frac{1}{2}\int_{\mathbb{R}^d} |Rx|^2 \mathrm{d}\mu(x) + G\big(\bar{\mu}\big), \quad \mathcal{F}(\mu) = \frac{1}{2}\int_{\mathbb{R}^d} |Qx|^2 \mathrm{d}\mu(x) + F\big(\bar{\mu}\big), \tag{2.1.14}$$

with G and F denoting primitives of $R^\dagger g$ and $Q^\dagger f$ (if any). In particular, the model is always potential when $d = 1$, but there is no potential structure when $d \geqslant 2$, unless $Q^\dagger f$ and $R^\dagger g$ both derive from (Euclidean) potentials, meaning that $(Q^\dagger f)^i = \partial F/\partial x_i$ and $(R^\dagger g)^i = \partial G/\partial x_i$.

2.1.4 Tilted harmonic and geometric fictitious plays

This subsection contains one of the main novelties of the paper. For reasons that are explained below, we introduce two variants of the fictitious play, which we call 'tilted harmonic' and 'tilted geometric'. The definition of the geometric version (which is the version that is effectively addressed in the paper) consists of the four equations (2.1.24)–(2.1.20)–(2.1.21)–(2.1.25) below and differs significantly from (2.1.11)–(2.1.12).

Consistently with our agenda, our goal is indeed to prove that the fictitious play may converge thanks to the presence of the common noise (i.e., $\varepsilon > 0$). Seemingly, the above discussion about the potential structure of our model in dimension $d = 1$ demonstrates that this question becomes especially relevant in dimension greater than or equal to 2 (the results from [24] could be easily adapted to this setting, even in the presence of the common noise). In fact, as shown in Subsection 2.3.5, the question is also interesting in dimension $d = 1$ in cases when equilibria are not unique.

However, this program is more challenging than it seems because we are not able, even in the presence of the common noise, to prove the convergence of the fictitious play stated in (2.1.9) and (2.1.11). Instead, we take benefit of the noise in order to reformulate the two equations (2.1.9) and (2.1.11) into a new system obtained by a mere shift of the common noise. By Girsanov theorem, the new shifted common noise has the same law as the original one but under a tilted probability measure. Using the harmonic updating rule (2.1.12), this so-called *tilted harmonic* scheme is then shown to converge (in the case $\varepsilon > 0$).

In order to state the new fictitious play properly, we thus need to allow for another form of common noise in (2.1.1). For a process $\boldsymbol{h} = (h_t)_{0\leqslant t\leqslant T}$, progressively-measurable with respect to the filtration generated by $\boldsymbol{W}$, we thus introduce the *shifted* Brownian motion

$$\boldsymbol{W}^{\boldsymbol{h}/\varepsilon} = \bigg(W_t^{\boldsymbol{h}/\varepsilon} := W_t + \frac{1}{\varepsilon}\int_0^t h_s \mathrm{d}s\bigg)_{0\leqslant t\leqslant T},$$

together with the tilted probability measure $\mathbb{P}^{\boldsymbol{h}}$ whose density with respect to $\mathbb{P}$ is

$$\mathcal{E}\big(\tfrac{1}{\varepsilon}\boldsymbol{h}\big) := \exp\bigg(-\frac{1}{\varepsilon}\int_0^T h_t \cdot \mathrm{d}W_t - \frac{1}{2\varepsilon^2}\int_0^T |h_t|^2 \mathrm{d}t\bigg). \tag{2.1.15}$$

Accordingly, for any two frozen continuous paths $\boldsymbol{n} := (n_t)_{0\leqslant t\leqslant T}$ and $\boldsymbol{w} := (w_t)_{0\leqslant t\leqslant T}$, we define the *non-averaged* cost functional

$$\mathcal{R}^{x_0}\big(\boldsymbol{\alpha}; \boldsymbol{n}; \varepsilon\boldsymbol{w}\big) := \frac{1}{2}\bigg[\big|Rx_T + g(n_T)\big|^2 + \int_0^T \Big\{\big|Qx_t + f(n_t)\big|^2 + |\alpha_t|^2\Big\}\mathrm{d}t\bigg], \tag{2.1.16}$$

where

$$x_t = x_0 + \int_0^t \alpha_s \mathrm{d}s + \sigma B_t + \varepsilon w_t, \quad t \in [0,T], \tag{2.1.17}$$

the latter being nothing but the integral version of (2.1.1), when $\boldsymbol{W}$ is replaced by the frozen trajectory $\boldsymbol{w}$. Now, when $\boldsymbol{m}$ and $\boldsymbol{h}$ are two progressively-measurable processes with respect to the filtration generated by $\varepsilon\boldsymbol{W}$, we have a look at the new cost functional[2]:

$$J^{\varepsilon}(\boldsymbol{\alpha}; \boldsymbol{m}; \boldsymbol{h}) := \mathbb{E}^{\boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}; \boldsymbol{m}; \varepsilon\boldsymbol{W}^{\boldsymbol{h}/\varepsilon}\big)\Big]. \tag{2.1.18}$$

We now define the first version of our titled fictitious play according to a two step iterative learning procedure, whose description at rank n goes as follows:

Harmonic best action For a proxy $\overline{\boldsymbol{m}}^n := (\overline{m}_t^n)_{0\leqslant t\leqslant T}$ of the conditional mean $\boldsymbol{m} = (m_t)_{0\leqslant t\leqslant T}$ of the in-equilibrium population (as given by the forward component of (2.1.7)) and a proxy $\boldsymbol{h}^n = (h_t^n)_{0\leqslant t\leqslant T}$ of the opposite[3] of the $\mathbb{F}^{\boldsymbol{W}}$-adapted intercept of the equilibrium feedback in (2.1.5) (as given by the backward component of (2.1.7)), solve

$$\boldsymbol{\alpha}^{n+1} = \mathrm{argmin}_{\boldsymbol{\alpha}}\mathbb{E}^{\boldsymbol{h}^n/\varepsilon}\big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}; \overline{\boldsymbol{m}}^n; \varepsilon\boldsymbol{W}^{\boldsymbol{h}^n/\varepsilon}\big)\big], \tag{2.1.19}$$

the infimum being taken over all $\mathbb{F}$-progressively measurable ($\mathbb{R}^d$-valued) controls $\boldsymbol{\alpha}$.

The optimal feedback being of the same linear form as in (2.1.5) (the proof is given below), we may call $\boldsymbol{h}^{n+1} = (h_t^{n+1})_{0\leqslant t\leqslant T}$ the opposite of the resulting intercept.

Harmonic update Given $\boldsymbol{h}^{n+1}$, the optimal trajectory of the above minimization problem is

$$X_t^{n+1} = X_0 - \int_0^t \big(\eta_s X_s^{n+1} + h_s^{n+1}\big)\mathrm{d}s + \sigma B_t + \varepsilon W_t^{\boldsymbol{h}^n/\varepsilon}, \quad t \in [0,T]. \tag{2.1.20}$$

We then let[4]

$$m_t^{n+1} = \mathbb{E}\big[X_t^{n+1} \,|\, \boldsymbol{W}\big], \quad t \in [0,T], \tag{2.1.21}$$

[2]The reader should observe that, when there is no common noise and when $\boldsymbol{\alpha}$ is progressively-measurable with respect to $\mathbb{F}^{X_0,\boldsymbol{B}}$ and $\boldsymbol{m}$ is deterministic, $\mathbb{E}[\mathcal{R}(\boldsymbol{\alpha}; \boldsymbol{m}; 0)]$ coincides with the original cost $J(\boldsymbol{\alpha}; \boldsymbol{m})$ in (2.1.2).

[3]The opposite comes from the sign $-$ in the formula (2.1.5) of the optimal feedback.

[4]The reader will find in Remark 2.2.2 a useful comment about the definition of $(m_t^{n+1})_{0\leqslant\leqslant T}$: the conditional expectation can be equivalently taken under $\mathbb{P}^{\boldsymbol{h}^n}$. This comes from the fact the Girsanov density is measurable with respect to the σ-field $\sigma(\boldsymbol{W})$ generated by $\boldsymbol{W}$.

together with

$$\overline{m}_t^{n+1} = \frac{1}{n+1}\sum_{k=1}^{n+1} m_t^k = \tfrac{1}{n+1} m_t^{n+1} + \tfrac{n}{n+1}\overline{m}_t^n, \quad t \in [0,T]. \tag{2.1.22}$$

For sure the rationale behind this strategy relies on Girsanov's transformation. Under the tilted probability measure $\mathbb{P}^{\boldsymbol{h}/\varepsilon}$, the process $\boldsymbol{W}^{\boldsymbol{h}/\varepsilon}$ is a new Brownian motion, with the same law as the original (or *historical*) common noise under $\mathbb{P}$. However, the main trick here is to dynamically change the form of the common noise (dynamically with respect to the rank of the iteration in the fictitious play). Precisely, this permits to decouple the two forward and backward equations, as clearly shown if we write the equation for $\overline{\boldsymbol{m}}^{n+1}$ as an equation with respect to the historical common noise (the proof is given in (2.2.6) below):

$$\overline{m}_t^{n+1} = m_0 - \int_0^t \Big(\eta_s m_s^{n+1} + \tfrac{1}{n+1}\big[h_s^{n+1} - h_s^0\big]\Big)\mathrm{d}s + \varepsilon W_t, \quad t \in [0,T]. \tag{2.1.23}$$

The forward equation then becomes asymptotically autonomous provided that $\boldsymbol{h}^{n+1}$ can be bounded independently of n, which we succeed to prove in Section 2.2. In turn, the scheme can be easily shown to converge (notice however that we are not able to prove the convergence of the standard non-tilted fictitious play outside any further potential or monotonicity assumption, even in the presence of the common noise). Moreover, the rate can be proved to decay (at least) like $\mathcal{O}(1/n)$, with $\mathcal{O}(\cdot)$ standing for the 'big $\mathcal{O}$ Landau notation'. Unfortunately, the best estimate we have for the constant driving the term $\mathcal{O}(1/n)$ blows up exponentially fast with $1/\varepsilon^2$. Although the numerical experiments that are reported below indicate that the rate may decrease faster than $1/n$ and grow slower than $\exp(\mathcal{O}(1/\varepsilon^2))$, the theoretical guarantee that is hence available for the version of the fictitious play comprising the four equations (2.1.19)–(2.1.20)–(2.1.21)–(2.1.22) is thus rather poor when the viscosity parameter ε tends to 0. This prompts us to provide a variant of the scheme, with an updating step (in equation (2.1.21)) that is different from $1/(n+1)$ and that yields a more favourable trade-off between n and ε^{-2} (or equivalently that allows for a cheaper choice of n when ε is small). In order to clarify things, we refer below to the scheme (2.1.19)–(2.1.20)–(2.1.21)–(2.1.22) as the 'tilted harmonic fictitious play'.

In a nutshell, the key point is to perform the following modifications in each of the two steps of the fictitious play, with the new version being called 'geometric' for reasons that become obvious in the next few lines:

Geometric best action With the same notation as before, use ϖX_0 as initial private state, $\varpi\boldsymbol{\alpha}$ as control, and $\mathbb{P}^{\varpi\boldsymbol{h}^n/\varepsilon}$ and $\boldsymbol{W}^{\varpi\boldsymbol{h}^n/\varepsilon}$ as tilted probability measure and tilted noise, where the rate ϖ is a fixed real that is typically chosen in the interval $(1,\sqrt{2}]$. In words, replace (2.1.19) by

$$\boldsymbol{\alpha}^{n+1} = \mathrm{argmin}_{\boldsymbol{\alpha}}\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\big[\mathcal{R}^{\varpi X_0}\big(\varpi\boldsymbol{\alpha}; \overline{\boldsymbol{m}}^n; \varpi\varepsilon\boldsymbol{W}^{\varpi\boldsymbol{h}^n/\varepsilon}\big)\big]. \tag{2.1.24}$$

Geometric update Accordingly, in (2.1.20), replace $\varepsilon W_t^{\boldsymbol{h}^n/\varepsilon}$ by $\varepsilon W_t^{\varpi\boldsymbol{h}^n/\varepsilon}$, and next use the updating formula:

$$\overline{m}_t^{n+1} = \frac{\varpi(1-\varpi^{-1})}{1-\varpi^{-(n+1)}}\sum_{k=1}^{n+1}\varpi^{-k} m_t^k = \tfrac{\varpi^{-n}(1-\varpi^{-1})}{1-\varpi^{-(n+1)}} m_t^{n+1} + \Big(1 - \tfrac{\varpi^{-n}(1-\varpi^{-1})}{1-\varpi^{-(n+1)}}\Big)\overline{m}_t^n, \tag{2.1.25}$$

for $t \in [0,T]$.

Here is the main idea. In comparison with (2.1.20), the optimal trajectory of (2.1.24) is

$$X_t^{n+1} = X_0 - \int_0^t \big(\eta_s X_s^{n+1} + h_s^{n+1}\big)\mathrm{d}s + \tfrac{1}{\varpi}\sigma B_t + \varepsilon W_t^{\varpi \boldsymbol{h}^n/\varepsilon}, \quad t \in [0,T]. \tag{2.1.26}$$

The above differs from (2.1.20) because the intensity of the idiosyncratic noise is σ/ϖ. Fortunately, this term disappears when taking conditional expectations given the common noise, as done in the formula (2.1.21). (In fact, one can also recover σ as intensity by considering $\varpi \boldsymbol{X}^{n+1}$ as optimal trajectory, but with ϖX_0 as initial condition.) Moreover, it must be stressed that the common noise right above is $\varepsilon \boldsymbol{W}^{\varpi \boldsymbol{h}^n/\varepsilon}$, whereas it is $\varepsilon \boldsymbol{W}^{\boldsymbol{h}^n/\varepsilon}$ in (2.1.20). This says that, under the historical probability measure $\mathbb{P}$, (2.1.26) rewrites

$$X_t^{n+1} = X_0 - \int_0^t \big(\eta_s X_s^{n+1} + h_s^{n+1} - \varpi h_s^n\big)\mathrm{d}s + \tfrac{1}{\varpi}\sigma B_t + \varepsilon W_t, \quad t \in [0,T]. \tag{2.1.27}$$

The presence of the factor ϖ in the last term of the drift makes it possible to use the geometric updating formula (2.1.25). In particular, this is our result to show (see §2.2.1.2 below) that, under the initialization $\boldsymbol{h}^0 = 0$, (2.1.23) becomes

$$\overline{m}_t^{n+1} = m_0 - \int_0^t \Big(\eta_s m_s^{n+1} + \tfrac{(1-\varpi)\varpi^{-n}}{1-\varpi^{-(n+1)}} h_s^{n+1}\Big)\mathrm{d}s + \varepsilon W_t, \quad t \in [0,T]. \tag{2.1.28}$$

The reader may easily compare with (2.1.23). Whereas the difference (with a standard decoupled Ornstein-Uhlenbeck process) decreases like $\mathcal{O}(1/n)$ in (2.1.23), it decreases like $\mathcal{O}(\varpi^{-n})$ in (2.1.28). In the end, this gives a geometric rate of convergence (in the parameter n). Although this does not change the presence of a leading constant of size $\exp(\mathcal{O}(1/\varepsilon^2))$, this makes it possible, in order to reach a given theoretical guarantee, to choose a value of n (much) lower than in the tilted harmonic fictitious play. Our main statement in this regard is Theorem 2.2.4. Also, it is worth adding that, under the obvious convention that

$$\frac{(1-\varpi^{-1})\varpi^{-n}}{1-\varpi^{-(n+1)}} = \frac{1}{n+1} \tag{2.1.29}$$

when $\varpi = 1$, the updating rule (2.1.22) coincides with (2.1.25). For this reason, we sometimes speak about the harmonic scheme (2.1.19)–(2.1.20)–(2.1.21)–(2.1.22) as a particular case of the geometric scheme when $\varpi = 1$. (Notice however that, although it could be easily adapted to the case $\varpi = 1$, see Remark 2.2.6, the statement of Theorem 2.2.4 does not formally apply to $\varpi = 1$ and even if it did, the statement would be in fact trivial.) Last but not least, we stress that, in the definition of the geometric best action, the tilted noise is biased, with $\varpi - 1$ as bias. This is an important feature of the scheme and this is the reason why we feel important to distinguish the best action (2.1.24) from the best action (2.1.19) and to call the former 'geometric' and the second 'harmonic'.

For sure, the reader may want to reformulate the tilted fictitious play for a more general MFG, with a structure that would no longer be linear-quadratic. Whereas the theory of MFG with an additive finite-dimensional common noise (such as $\boldsymbol{W} = (W_t)_{0 \leqslant t \leqslant T}$) is by now well-established (see

for instance the book Carmona and Delarue [31]), the real interest for addressing the same tilted form of the fictitious play but in a wider setting is however not clear to us. Indeed, aforementioned known theoretical results on the smoothing effect of the common noise (see Delarue [47]) require in general an infinite dimensional noise of a much more complicated structure than the additive finite-dimensional noise $(W_t)_{0\leqslant t\leqslant T}$ used in (2.1.28). For this reason, we have decided to restrict the whole exposition to the linear-quadratic setting, even though (2.1.24), (2.1.20), (2.1.21) and (2.1.25) could be recast, for the same additive finite-dimensional noise, within a larger framework. Outside the class of linear-quadratic games, the main changes are the following ones. First, the optimal feedback in (2.1.20) is no longer affine. Second, the mean field fixed point can no longer be formulated in terms of the sole conditional expectation, as it is done in (2.1.20), but involves the full statistical law of X_t^{n+1} given $\boldsymbol{W}$. Third (and subsequently), the rule (2.1.25) must be reformulated in terms of the full statistical distributions and not only in terms of their means. We leave the details to the reader. Needless to say, obtaining a relevant extension of the tilted fictitious play for general MFGs remains a very interesting but highly challenging objective.

2.1.5 Exploration

We now explain how the common noise in the tilted fictitious play can be regarded as an exploration noise for the original MFG without common noise.

For sure, changing the common noise as we have done in the cost functional (2.1.24) (see also (2.1.19)) raises indeed many practical questions[5]. In order to understand this properly, we should follow the presentation given in Carmona et al. [35, 36] and think of $\mathcal{R}$ as being a black-box representing a decentralized unit. For instance, the box may regulate the consumption/production/storage of energy of a single individual connected to a smart grid, see for instance Alasseur [7] and the references therein; the box may also be an autonomous car moving in a flock of vehicles, see for instance Huang et al. [71].

The black-box operation is described in Figure 2.1. In this picture, the decentralized black-box receives four inputs: (i) the control, as tuned by the individual operating the black-box; (ii) the initial private state X_0; (iii) the two idiosyncratic and independent noises; (iv) the state of the population. In this representation, the only input that can be tuned by the individual operating the black-box is the control itself.

However, this picture makes sense only if the original model itself is subjected to a common noise. In the absence of common noise, we thus need to restore a form of common noise in order to conciliate our tilted fictitious play with the above picture. This comes through the notion of exploration. Below, we thus regard the common noise as a way to explore the space of possible solutions. This amounts to say that the original mean field game is no longer the mean field game with common noise, but the mean field game without common noise, i.e. $\varepsilon = 0$. Consistently, the individual operating the black-box can restore the presence of a common noise by modifying her/his control accordingly. Formally, any control $\boldsymbol{\alpha}$, as chosen above by a tagged individual, is

[5]The reader may find it reminiscent of the weak formulation of MFGs introduced in Carmona and Lacker [32], but this is substantially different because the Girsanov transformation is here applied to the common noise (and not to the idiosyncratic one).

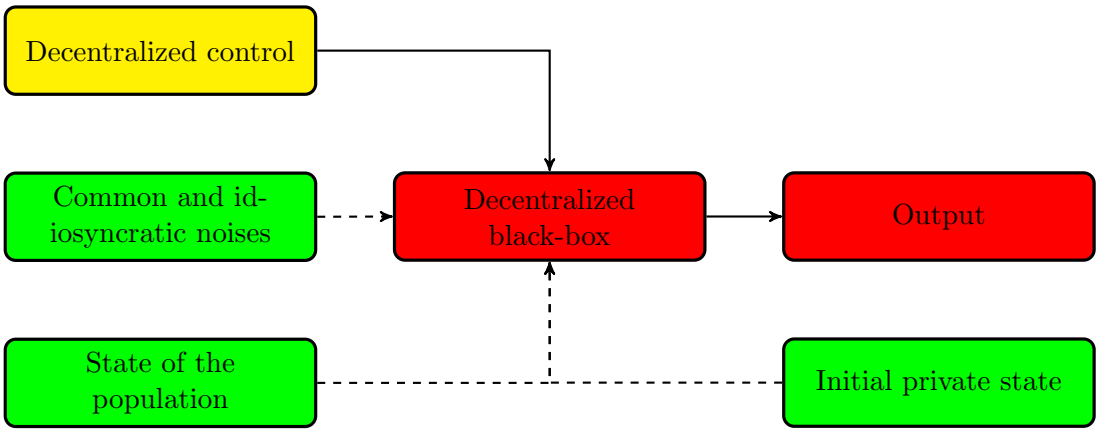


Figure 2.1: Black-box operation for an MFG with a common noise. In the four input arrows, only the plain line can be tuned. Given the input state of the population, the private initial state X_0 and the realizations of the two noises, the decentralized black-box returns an output, in the form of a cost depending on the input state and on the realizations of the noises.

then subjected to an additional randomization of the form

$$\boldsymbol{\alpha} = (\alpha_t)_{0\leqslant t\leqslant T} \mapsto \left(\alpha_t + \varepsilon \dot{W}^{\boldsymbol{h}}_t\right)_{0\leqslant t\leqslant T},$$

where $\boldsymbol{h}$ is given as an information on the whole state of the population, in addition to $\boldsymbol{m}$. Figure 2.1 becomes the new Figure 2.2 below. In this new figure, the decentralized black-box is exactly the same as the decentralized black-box appearing in Figure 2.1 in the absence of the common noise therein. In clear, the black-box takes as inputs the following three features: a time-dependent flow of actions impacting the dynamics of a single individual, the initial private state X_0 and the time-dependent flow of probability measures characterizing the state of the environment. As for the output, the black-box returns the realization $\mathcal{R}^{X_0}$ of the cost to a single individual, for the given action, the given realization of the initial condition and the given environment (see (2.1.16)). In particular, it is worth mentioning that the common randomization does not impact the operation performed by the black-box, which is an important fact from the practical point of view. What changes in the presence of the common randomization is the form of the input that is inserted in the black-box: the input at time t is 'corrupted' by $\varepsilon \dot{W}^{\boldsymbol{h}}_t$. In particular, the reader should agree that Figure 2.2 is consistent with the definitions (2.1.19) and (2.1.24) of the best actions in our two tilted fictitious plays, up to a slight but subtle difference between (2.1.19) and (2.1.24): in order to compute (2.1.24), the initial private state X_0 in the black-box must be free (as made clear in the caption), in the sense that it can be changed for the purpose of the experiment. This is an important feature because the non-averaged cost functional $\mathcal{R}$ in the geometric fictitious play is initialized from ϖX_0 (and not X_0), see (2.1.24). We feel that this additional assumption on the model is affordable in practice.

Now, if we had to think of a (possibly infinite) cloud of players, the actions of all of them would be corrupted by the same realization of the noise. Assume indeed that, given the same two proxies $\overline{\boldsymbol{m}}^n$ and $\boldsymbol{h}^n$ as in (2.1.19) and (2.1.24), the players choose some common feedback function (which

is exactly what happens for an MFG equilibrium). The resulting action of each of them is then randomized by the same realization of the common noise and, subsequently, the black-box returns the (non-averaged) cost to each player. If the players are driven by independent copies of $(B_t)_{0\leqslant t\leqslant T}$ (which fits the fact that $(B_t)_{0\leqslant t\leqslant T}$ is an idiosyncratic noise) and by independent copies of ϖX_0 (which fits the fact that ϖX_0 is the new initial private condition in the ϖ-geometric scheme), then the empirical mean cost to all the players should be regarded as the conditional expectation of the cost $\mathcal{R}^{\varpi X_0}$ given $\boldsymbol{W}$. Assuming that the common noise is observable (which makes perfect sense if it is sampled by some experimenter), we can compute $\mathcal{E}(\varpi \boldsymbol{h}^n/\varepsilon)$ and then multiply it with the conditional expectation of the cost. Sampling the common noise as many times as desired, we get an empirical approximation of the mean cost under $\mathbb{P}^{\varpi \boldsymbol{h}^n/\varepsilon}$ in (2.1.24). Although this picture may look rather naive, it is in fact the basis of a numerical method that is detailed next, see Figure 2.3 for a primer.

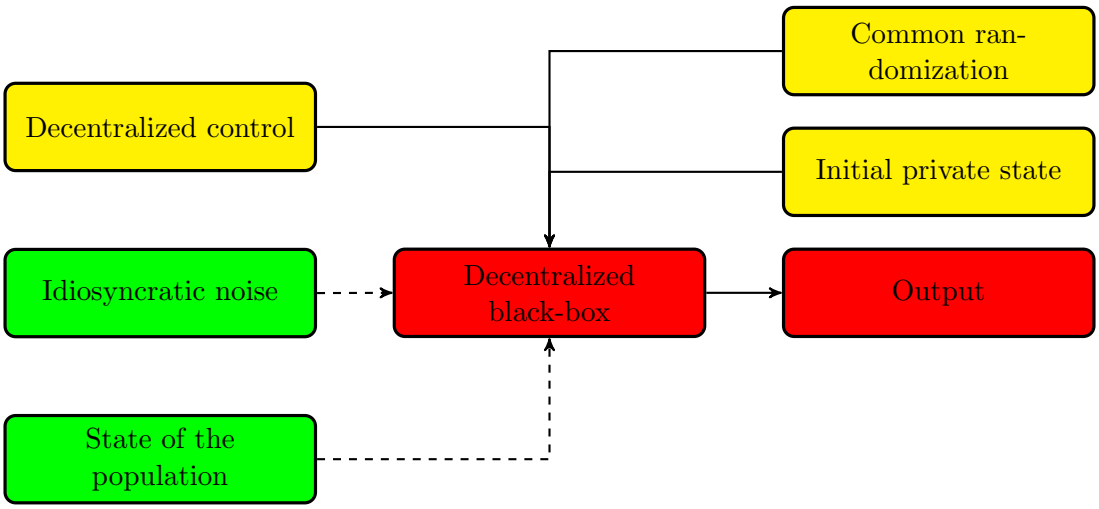


Figure 2.2: Black-box operation for an MFG without a common noise, but subjected to a common randomization of the control. In the five input arrows, only the plain lines (corresponding to yellow boxes) can be tuned. Given the input state of the population and the realizations of the two noises, the decentralized black-box returns an output, in the form of a (random) cost depending on the input state and on the realizations of the noises.

This concept faces however obvious mathematical difficulties, because the new control after randomization is no longer of finite energy. We solve this issue by replacing the time derivative of $\boldsymbol{W}$ by finite differences or, equivalently, by replacing $\boldsymbol{W} = (W_t)_{0\leqslant t\leqslant T}$ by its piecewise linear interpolation along a mesh of $[0,T]$, say of uniform step T/p, for a given integer $p \geqslant 1$. We denote this interpolation by $\boldsymbol{W}^p = (W_t^p)_{0\leqslant t\leqslant T}$. Accordingly, the return of the black-box should be *renormalized*, letting

$$\mathcal{R}^{p,x_0}(\boldsymbol{\alpha};\boldsymbol{m};\varepsilon\boldsymbol{w}) := \mathcal{R}^{x_0}(\boldsymbol{\alpha}+\varepsilon\dot{\boldsymbol{w}};\boldsymbol{m};0) - \tfrac{1}{2}\varepsilon^2 p, \tag{2.1.30}$$

for a piecewise-affine path $\boldsymbol{w}$, affine on each $[\ell T/p,(\ell+1)T/p)$ for $\ell \in \{0,\cdots,p-1\}$. Our second main statement, see Theorem 2.2.17, is to prove that the geometric fictitious play that is hence

obtained by replacing $\boldsymbol{W}$ by $\boldsymbol{W}^p$ and $\mathcal{R}^{x_0}(\boldsymbol{\alpha}; \overline{\boldsymbol{m}}^n; \varepsilon \boldsymbol{W}^{\boldsymbol{h}^n})$ by $\mathcal{R}^{p,x_0}(\boldsymbol{\alpha}; \overline{\boldsymbol{m}}^n; \varepsilon \boldsymbol{W}^{p,\boldsymbol{h}^n})$, with

$$W_t^{p,\boldsymbol{h}^n} = W_t + \frac{1}{\varepsilon} \int_0^t h_s^n \mathrm{d}s, \quad t \in [0,T], \tag{2.1.31}$$

converges, when the number n of iterations tends to ∞, to the solution of the discrete-time version, with T/p as step size, of the mean field game with $\varepsilon \boldsymbol{W}^p$ as common noise. Implicitly, this requires to force $\boldsymbol{\alpha}$ and $\boldsymbol{h}^n$ to be constant on any subdivision of the mesh, but we feel more appropriate not to detail all the ingredients here. We refer the reader to Subsection 2.2.2 for a complete description. Importantly, the state dynamics over which the return $\mathcal{R}^p$ is computed write

$$\begin{aligned} X_t &= \varpi X_0 + \int_0^t \varpi \alpha_s \mathrm{d}s + \sigma B_t + \varepsilon W_t^{p,\varpi \boldsymbol{h}^n/\varepsilon} \\ &= \varpi X_0 + \int_0^t \Big(\varpi \alpha_s + \varepsilon \dot{W}_s^{p,\varpi \boldsymbol{h}^n/\varepsilon} \Big) \mathrm{d}s + \sigma B_t, \quad t \in [0,T]. \end{aligned} \tag{2.1.32}$$

In this approach, the control after randomization is thus given by $\varpi \boldsymbol{\alpha} + \varepsilon \dot{\boldsymbol{W}}^{p,\varpi \boldsymbol{h}^n/\varepsilon}$, which clarifies the meaning of the common randomization in Figure 2.2. In the end, this fits well the concept of exploration, as stated by Sutton and Barto [104, chapter 1, p. 1]: '*The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them*'. Noticeably, our form of randomization (2.1.32) can be restricted to controls in semi-feedback form, meaning that the instantaneous control in (2.1.20) can be chosen as the image of the current state of $\boldsymbol{X}^{n+1}$ by a time-space random function adapted to the filtration generated by $\boldsymbol{W}$. In particular, thinking of the mean field game as a game with infinitely many agents, all of them are then understood to play (at each iteration of the fictitious play) the same random semi-feedback function. This makes a subtle difference with standard exploration methods for multi-agent reinforcement learning in which the random control of each agent carries its own noise. Last but not least, it must be noticed that the algorithm runs from the sole observations of the returns of the black-box, and in particular without any further detailed of the cost coefficients f, g, Q and R, provided we use a reinforcement learning method to solve the black-box.

Importantly, it must be emphasized again that in the two Figures 2.1 and 2.2 the decentralized black-box does not return the expected cost, but only the realization of the cost for the given realizations of the noises. At each step of the fictitious play, the optimization of the expected cost is formally performed over all the possible trajectories of the independent and common noises, bearing in mind that the control is adapted to both noises and that the state of the population is adapted to the common noise. This is stylized in the form of Figure 2.3. Therein, this is our choice to represent the possible trajectories of the two noises in the form of an infinite sequence of realizations from an i.i.d. sample of the idiosyncratic and common noises (and similarly for the initial condition), this formal representation being very convenient for introducing next the numerical implementation. In clear, assuming that we have two independent families $(\boldsymbol{B}^i = (B_t^i)_{0 \leqslant t \leqslant T})_{i \geqslant 1}$ and $(\boldsymbol{W}^j = (W_t^j)_{0 \leqslant t \leqslant T})_{j \geqslant 1}$ of (d-dimensional) independent Brownian motions and, independently, a family $(X_0^i)_{i \geqslant 1}$ of i.i.d. initial conditions, we are given at rank n of the iterative process represented in Figure 2.3 two collections of proxies $(\overline{\boldsymbol{m}}^{n,j})_{j \geqslant 1}$ and $(\boldsymbol{h}^{n,j})_{j \geqslant 1}$ of i.i.d. continuous paths with values in $\mathbb{R}^d$, with each $\overline{\boldsymbol{m}}^{n,j}$ and $\boldsymbol{h}^{n,j}$ being assumed to be adapted with respect to $\boldsymbol{W}^j$. We then consider a collection of i.i.d. $\mathbb{R}^d$-valued control processes $\boldsymbol{\alpha}^{i,j} = (\alpha_t^{i,j})_{0 \leqslant t \leqslant T}$, with each $\boldsymbol{\alpha}^{i,j}$

being assumed to be adapted with respect to $(X_0^i, \boldsymbol{B}^i, \boldsymbol{W}^j)$. Each $\boldsymbol{\alpha}^{i,j}$ and each $\boldsymbol{h}^{n,j}$ are assumed to be constant on each $[\ell T/p, (\ell+1)T/p)$ for $\ell \in \{0, \cdots, p-1\}$, where p stands as before for the discretization parameter. For a given outcome ω in the probability space carrying all the processes, we then choose as inputs of the black-box (corresponding to $\mathcal{R}^{p,\varpi X_0}$ in (2.1.30)) the noise $\boldsymbol{B}^i(\omega)$, the environment $\overline{\boldsymbol{m}}^{n,j}(\omega)$ and the control $\varpi\boldsymbol{\alpha}^{i,j}(\omega)$ perturbed by the additional randomization $\varepsilon\dot{\boldsymbol{W}}^{p,\varpi\boldsymbol{h}^{n,j}/\varepsilon,j}(\omega)$. All these inputs are represented by the green boxes in Figure 2.3. As made clear by the red boxes on Figure 2.3, the black-box returns a cost that depends on i, j (in Figure 2.3, we use the notation $\sharp i$ and $\sharp j$ to refer to the various inputs and outputs associated with $\boldsymbol{B}^i(\omega)$ and $\boldsymbol{W}^j(\omega)$). Multiplying by $\mathcal{E}(\varpi\boldsymbol{h}^{n,j}/\varepsilon)(\omega)$ and averaging over i, j, we hence get the averaged cost that appears in the right-hand side of (2.1.24). This makes it possible to optimize with respect to the control $\boldsymbol{\alpha}$ (or equivalently with respect to $(\boldsymbol{\alpha}^{i,j})_{i,j\geqslant 1}$) and thus to compute the optimal control $\boldsymbol{\alpha}^{n+1}$ that appears in the left-hand side of (2.1.24) (up to the time-discretization procedure that we feel better not to address at this early stage of the paper): this is the yellow box in Figure 2.3. Once the optimal control has been computed, we may follow the arrows connecting the three blue boxes in Figure 2.3: for each pair (i, j), we can compute the realization $\boldsymbol{X}^{n+1,i,j}(\omega)$ of the optimal state. Averaging with respect to i, we get $\boldsymbol{m}^{n+1,j}(\omega)$ and then, following the updating rule (2.1.25), we can update the state of the environment: the new state is $\overline{\boldsymbol{m}}^{n+1,j}(\omega)$. The new value of $\boldsymbol{h}^{n+1,j}(\omega)$ is directly taken from the affine form of $\boldsymbol{\alpha}^{n+1,j}(\omega)$, see (2.1.5).

2.1.6 Exploitation

We now summarize the outline of the exploitation analysis that is achieved in the paper. In particular, we present the main bound that we can prove next for the so-called exploitability of the (geometric) tilted fictitious play (the meaning of which is explained below).

In reinforcement learning, this is indeed a common practice to distinguish exploration from exploitation. Whereas exploration is intended as a way to visit the space of actions, exploitation is related to the error that is achieved by the learning method. When the learning addresses a stochastic control problem, the analysis goes through the loss, which is the (absolute value of the) difference between the best possible cost and the cost to the strategy returned by the algorithm. Because we are dealing with a game, we use here the concept of approximated Nash equilibrium to define the exploitability. In brief, the point is to prove that the output of the algorithm is a ϱ-Nash equilibrium, for $\varrho > 0$ and to define the exploitability as the infimum of those ϱ. In our case, the situation is a bit more subtle because the learning procedure returns a random approximated equilibrium; ideally, we should associate with it a random exploitability. However, the analysis would be too difficult. Instead, we follow (2.1.24) and average the cost with respect to the common noise. We then define the notion of approximated equilibrium with respect to this averaged cost. The resulting exploitability can then be decomposed as the sum of two terms:

- The 'error' resulting from the approximation, learnt by our fictitious play, of the discrete-time mean field game with step size T/p and with common noise of intensity ε. The implementation of the algorithm involves n iterations and a piecewise linear approximation of the Brownian motion $\boldsymbol{W}$ with T/p steps. For fixed[6] $\varepsilon \in (0, 1]$ and $p \in \mathbb{N}$, this error tends to 0 as n tends to

[6]For convenience, we assume ε to be less than 1 in the analysis. Obviously, the results would remain true for $\varepsilon > 1$, but the various constants in our analysis could then depend on any *a priori* upper bound on ε.

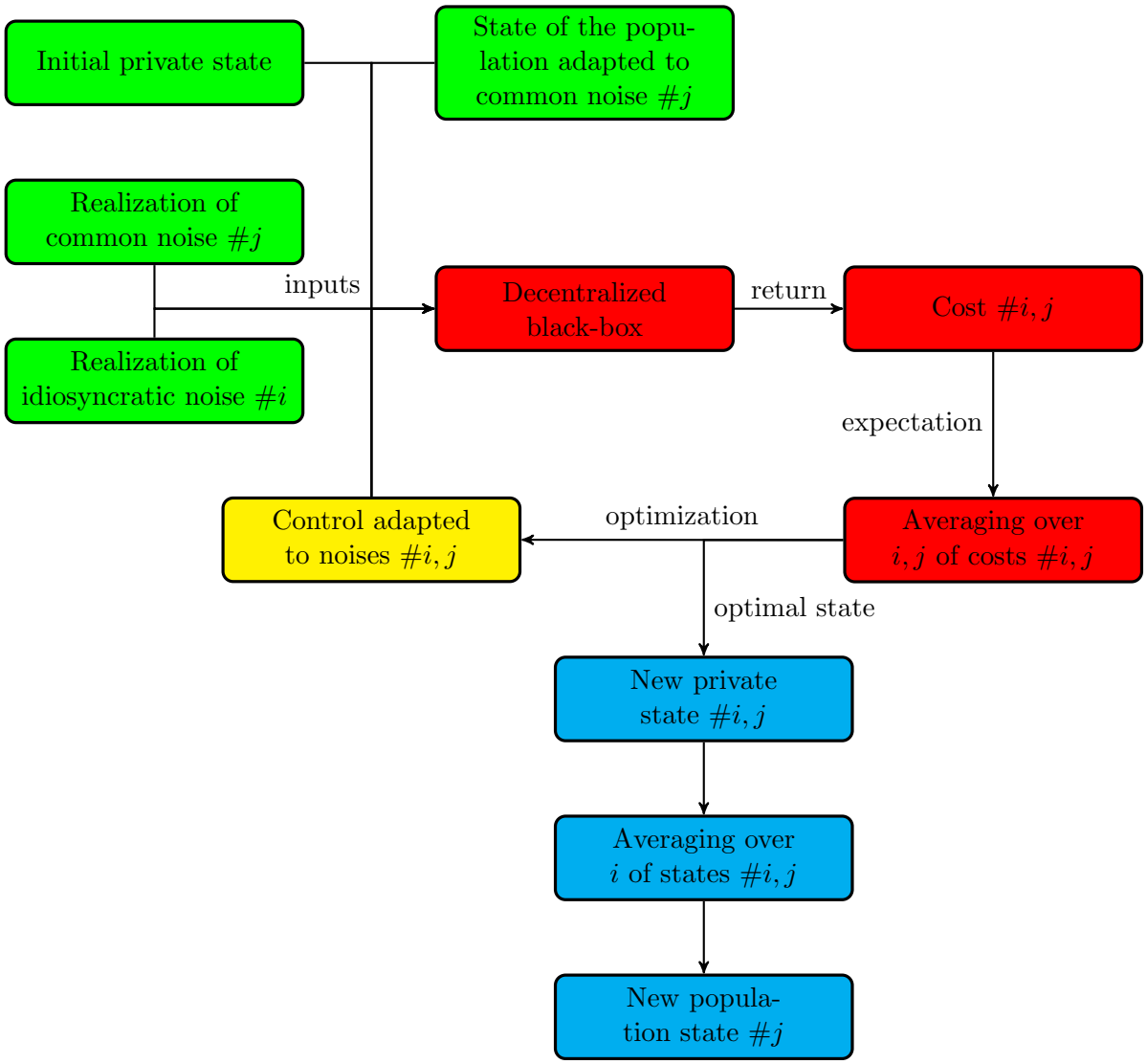


Figure 2.3: Black-box (in red) inserted at the core of a learning step. Expectations are formally written as means over a sequence of realizations from an i.i.d. sample of initial conditions and of idiosyncratic and common noises. For any respective realizations $\#i$ ($\#i$ reading 'number i') and $\#j$ ($\#j$ reading 'number j') of the idiosyncratic and common noises plugged into the black-box, we compute the corresponding state of the population (that depends on the realization $\#j$ of the common noise) and we choose the control (that is adapted to the realizations $\#i, j$ of the two noises and of the initial condition). The inputs are thus in green, except the control which is subjected to optimization (hence in yellow, as a mixture of green and red). The return (in red) of the black-box depends on the realizations $\#i, j$. The expectation is formally obtained by averaging with respect to $\#i, j$. Once the optimizer has been found, we compute (in blue) the optimal states, depending on the realizations $\#i, j$. The output state of the population after one learning step is obtained by averaging over i.

∞, with an explicit rate $\varrho_{\varepsilon,p}(n)$.

- The 'error' resulting from the common noise and discrete-time approximation of the original time-continuous mean field game without common noise. Intuitively, the (unique) solution of the time-discrete mean field game with common noise produces a random approximated Nash equilibrium of the original mean field game whose accuracy gets finer and finer as p increases and ε decreases.

In fact, this principle can be put in the form of a more general stability result that permits to evaluate in the end the trade-off between exploration and exploitation. When X_0 has sub-Gaussian tails, the exploitability is bounded by

$$O\Big(\exp(C\varepsilon^{-2})\varpi^{-n} + \varepsilon + \tfrac{1}{p}\Big), \tag{2.1.33}$$

see Theorem 2.2.27. For fixed values of n and p (the latter two parametrizing the complexity of the algorithm and the required memory[7]), we are hence able to tune the intensity of the noise.

To our mind, this result demonstrates the interest of our concept, even though it says nothing about the equilibria that are hence selected in this way when ε tends to 0. In fact, the latter is a difficult question, which has been addressed for instance in [45] (for the same model as in (2.1.1)–(2.1.2), but with $d = 1$ and for some specific choices of f and g) and [37] (for finite state potential games); generally speaking, this problem raises many theoretical questions that are out of the scope of this paper and we just address it here through numerical examples (see the subsection below). The diagram (2.4) right below hence summarizes the balance between exploitation and exploration in our case. In words, the geometric tilted fictitious play allows us to learn a solution of the discrete-time MFG with step size T/p and with common noise (or, equivalently, *exploration* noise) of intensity $\varepsilon > 0$ by choosing n large enough. The equilibrium that is hence learnt is an approximate equilibrium of the original MFG (in continuous time and without common noise). For a small intensity ε, the exploitability is small if n and p are large enough, hence the trade-off between ε and (n, p).

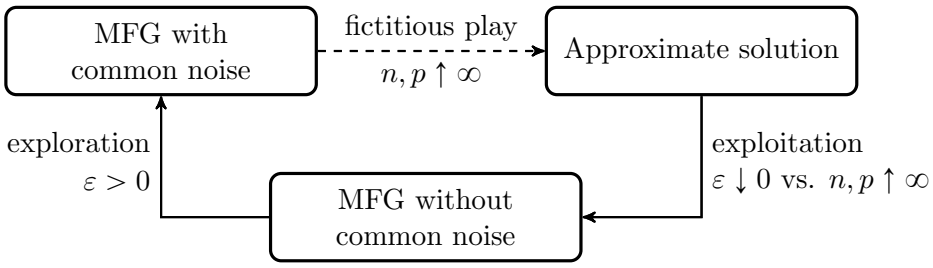


Figure 2.4: Exploration vs. exploitation.

For sure, our analysis of exploitation is carried out in the ideal case when the underlying expectations in (2.1.24) are understood in the theoretical sense and when the optimal control

[7]We feel better not to give any order of complexity and memory in terms of n and p. This would have no sense because the integration and optimization steps are not discretized here.

problem at each step can be computed perfectly. We do not address here the approximation of those theoretical expectations by empirical means nor the numerical approximation of the optimizers.

2.1.7 Numerical examples

We complete the paper with some numerical examples that demonstrate the relevance of our concept. The numerical implementation requires additional ingredients that are explained in detail in Section 2.3. Obviously, the main difficulty is the encoding of the decentralized black-box, as represented in Figures 2.2 and 2.3. As clearly suggested by the latter figure, expectations are then approximated by averaging the costs over realizations from a finite i.i.d. sample of idiosyncratic and common noises. In this regard, the principle highlighted by Figure 2.3 is the same, but the optimization step needs to be clarified. Although we do not provide any further theoretical justification of the accuracy of the numerical optimization that is hence performed, we feel useful to stress that controls are chosen in a semi-feedback form, namely of the same linear form as in (2.1.5). Numerically, the coefficient η_t is hence parameterized in the form of a coefficient that is only allowed to depend on time; and the intercept h_t is sought as a function of the current mean state of the population. This latter function is parameterized in the form of a finite expansion along an Hermite polynomial basis, our choice for Hermite polynomials being dictated by the Gaussian nature of the trajectories in (2.1.28), or in the form of a neural network. In our numerical experiments, both the linear coefficient and the coefficients in the regression of the intercept along the given class of functions are found by ADAM optimization method. The results exposed in Section 2.3 demonstrate the following features:

1. For a given value of the intensity of the common noise (say $\varepsilon = 1$), we run examples in dimension $d = 2$. Both the (tilted) harmonic and geometric fictitious play converge well, without any significant differences between the two of them on the examples under study. Solutions are compared to numerical solutions of (2.1.7) found by a BSDE solver that uses explicitly the shape of the coefficients f, g, Q and R and that does not use the observations of the cost.

2. We also run examples in higher dimension, namely $d = 12$ and $d = 20$. Obviously, this raises challenging questions in terms of complexity, in particular for regressing $\boldsymbol{h}^{n+1}$ in (2.1.20) by means of a suitable basis. Although we show numerically that neural networks may behave well in these examples, the main difficulty in the higher dimensional setting comes from the various Monte-Carlo estimations on which Figure 2.3 implicitly relies. Indeed, the variance of the cost in (2.1.24) may increase fast with the dimension. This phenomenon has an impact on the behavior of the algorithm, which may fail to converge as clearly illustrated in some of the examples below. Next, we address one simple reduction variance method, which returns much better results in dimension $d = 12$ and $d = 20$ and which demonstrates that the algorithm may remain relevant in this more challenging framework at the price of some marginal modifications. In light of these results, we believe that it would be highly valuable to provide a more exhaustive analysis of the possible strategies for reducing the variance underpinning the various Monte-Carlo computations. We hope to address this point in future contributions.

3. The standard fictitious play, with idiosyncratic noise but without common noise, may fail to converge, meaning that, not only there may not be any known mathematical guarantee supporting the convergence, but even more, for one of the examples we treat below in dimension $d = 2$, the cost returned by the algorithm has an oscillatory behavior that is not observed in presence of the common noise. Even though we do not have a mathematical explanation for these observations, this is a crucial point of the paper as it demonstrates numerically that the common noise helps the algorithm to converge. From a conceptual point of view, this is a key observation. Of course, it would be very much desirable to have a table of comparison, with a mathematical description of the behavior of the algorithm without and with common noise and for various types coefficients. This looks however out of reach of the existing literature. Still, it is worth observing that, in dimension 1, the usual algorithm is known to converge because the model is potential, see [24], and that, in this setting, we have not observed any numerical oscillatory behavior similar to the two dimensional one (for the same types of coefficients). We elaborate on this point in §2.3.3.1. We also feel useful to recall from the previous Subsection 2.1.6 that, from a theoretical point of view, we are able to provide explicit bounds for the rate of convergence, which is another substantial contribution of our work. Indeed, as explained in the forthcoming §2.2.1.3, we know a few results only in which the rate of convergence of the fictitious play without common noise is addressed explicitly.

4. In order to study the behavior of our fictitious play when the intensity ε of the common noise becomes small, we focus on a one-dimensional MFG that has multiple equilibria when there is no common noise. This case is highly challenging. Not only theoretical bounds like (2.1.33) are especially bad when ε is small, but also additional numerical issues arise in the small noise regime. In particular, the variance of the various estimators suffer from the same drawback as in the high-dimensional setting and may be very large. However, we here show that, numerically, the geometric tilted fictitious play run with a high rate ϖ and a small intensity ε is able to select quite quickly the same equilibrium as the one predicted in [45, 37]. In contrast, the standard fictitious play (without common noise) also converges but may not select the right equilibrium (for the same choice of parameters). Also, it is worth observing that, in this experiment, the geometric variant of the titled fictitious play performs better than the harmonic one, which is consistent with our theoretical analysis. By the way, in the first arXiv version [48] of this work (in which we just studied the harmonic variant), we complemented the numerical analysis with a preferential sampling method in order to reduce the underlying variance and hence obtain a better accuracy in the selection of an equilibrium. This would be an interesting question to address the possible interest of such a preferential sampling method when combined with the geometric variant of the algorithm. We leave this for future works.

In the end, the numerical experiments carried out here confirm the relevance of the tilted scheme. However, it is fair to say that it is more subtle to demonstrate the superiority of the geometric variant over the harmonic variant, even if the experiment (4) reported above clearly points in that direction. In our opinion, the geometric variant has the great merit of offering theoretical convergence guarantees that are affordable numerically. However, the experiments described in Section 2.3 show that the harmonic variant is also numerically relevant. Numerical results could

probably be optimized by combining the two approaches, so as to benefit from the advantages of both.

It should be stressed that our numerical experiments are run under `Tensorflow`, using a pre-implemented version of ADAM optimization method in order to compute an approximation of the best response in (2.1.24). Accordingly, the optimization algorithm itself relies explicitly on the linear-quadratic structure of the mean field game through the internal automatic differentiation procedure (used for computing gradients in descents). In this sense, our numerical experiments use in fact more than the sole observations of the costs. Anyhow, this does not change the conclusion: descent methods, based on accurate approximations of the gradients, do benefit from the presence of the common noise. For instance, the construction of accurate approximations of the gradient is addressed in Carmona and Laurière [35, 36] (within a slightly different setting), in which a model-free reinforcement learning method is fully implemented[8].

2.1.8 Comparison with existing works

Exploration and exploitation are important concepts in reinforcement learning and related optimal control.

In comparison with the discrete-time literature, there have been less papers on the analysis of exploration/exploitation in the time continuous setting. In both Murray and Paladino [93] and Wang et al. [111], the randomization of the actions goes through a formulation of the corresponding control problem in terms of relaxed controls. In Murray and Paladino [93], the authors address questions that are seemingly different from ours, as the objective is to allow for a model with some uncertainty on the state dynamics. Accordingly, the cost functional is averaged out with respect to some prior probability measure on the vector field driving the dynamics. Under suitable assumptions on this prior probability measure, a dynamic programming principle and a then a Hamilton-Jacobi-Bellman are derived. In fact, the paper Wang et al. [111], which addresses stochastic optimal controls, is closer to the spirit of our work. Therein, relaxed controls are combined with an additional entropic regularization that forces exploration. In case when the control problem has a linear-quadratic structure, quite similar to the one we use here (except that there is no mean field interaction), the entropic regularization is shown to work as a Gaussian exploration. Although our choice for working with a Gaussian exploration looks consistent with the result of Wang et al. [111], there remain however some conceptual differences between the two approaches: In the theory of relaxed controls, the drifts in the dynamics are averaged out with respect to the distribution of the controls; In our paper, the dynamics are directly subjected to the randomized action. In this respect, our work is closer to the earlier contribution of Doya [52].

Recently, the approach initiated in Wang et al. [111] has been extended to mean field games. In Guo et al. [118] and Firoozi and Jaimungal [55], the authors study the impact of an entropic regularization onto the shape of the equilibria. In both papers, the models under study are linear-quadratic and subjected to a sole idiosyncratic noise (i.e., there is no common noise). However, they differ on the following important point: in Firoozi and Jaimungal [55], the intensity of the idiosyncratic noise is constant, whilst it depends on the standard deviation of the control in Guo et

[8]The reader may find in Munos [92] a nice explanation about the distinction between model-free and model-based reinforcement learning. In any case, our algorithm is not model-based: It would be model-based if we tried to learn first f, g, Q or R. We refer to §2.3.3.3 for a discussion about the possible numerical interest to learn Q and R first.

al. [118]. In this sense, Firoozi and Jaimungal [55] is closer to the set-up that we investigate here. Accordingly, the presence of the entropic regularization leads to different consequences: In Guo et al. [118], the effective intensity of the idiosyncratic noise grows up under the action of the entropy and this is shown to help numerically in some learning method (of a quite different spirit than ours). In Firoozi and Jaimungal [55], the entropy plays no role on the structure of the equilibria, which demonstrates, if needed, that our approach here is substantially different.

Within the mean field framework, there have been several recent contributions on reinforcement learning for models featuring a common noise. In Carmona et al. [35], the authors investigate the convergence of a policy gradient method for a discrete time linear quadratic mean field control problem (and not an MFG) with a common noise. In comparison with (2.1.2), the cost functional itself is quadratic with respect to the mean field interaction. The linear quadratic structure then allows us to simplify the search for the optimal feedbacks, in the form of two linear functions, one linear function of the mean state of the population and one linear function of the deviation to the mean state. Accordingly, the problem is rewritten in terms of two separate (finite-dimensional) linear-quadratic control problems, one with each of the two linear factors. Convergence of the descent for finding the optimizers is studied for a model free method using a black-box simulating the evolution of the population. This black-box is comparable to ours. Importantly, non-degeneracy of the very first inputs of the common noise is used in the convergence analysis. In another work (Carmona et al. [36]), the same three authors have developed a model free Q-learning method for a mean field control problem in discrete time and finite space. The model may feature a common noise, but the latter has no explicit impact onto the convergence analysis carried out in the paper. Last, in Elie et al. [54] and Perrin et al. [96], the authors deal with discrete and continuous time learning for MFGs using fictitious play and introducing a form of common noise. Their analysis is supported by various applications and numerical examples (including a discussion on the tools from deep learning to compute the best responses at any step of the fictitious play). The analysis also relies on the notion of exploitability. As the common noise therein is not used for exploratory reasons, the coefficients are assumed to satisfy the monotonicity Lasry-Lions condition in order to guarantee the convergence of the fictitious play.

After the publication of the first arXiv version [48] of our paper, several related works by other authors were released. For instance, the authors of Muller et al. [**mueller**] address Policy Space Response Oracles (PSRO) within the mean field framework in order to approximate Nash equilibria but also relaxed equilibria that are said to be correlated. In particular, the authors explain how to implement PSRO via regret minimization in order approximate correlated equilibria. In Hu and Laurière [**hu:hal-03656245**], the authors provide a very nice survey of recent developments in machine learning for stochastic control and games. Last, it is also fair to quote Hambly et al. [**hambly**], in which the authors address the global convergence of the natural policy gradient method to the Nash equilibrium in a general N-player linear-quadratic game. Noticeably, the proof of convergence therein requires a certain amount of noise.

2.1.9 Main assumption, useful notation and organization.

2.1.9.1 Assumption

Throughout the analysis, σ in (2.1.1) is a fixed non-negative real. The intensity ε of the common noise is taken in $[0,1]$. Most of the time, it is implicitly required to be strictly positive, but we sometimes refer to the case $\varepsilon = 0$ in order to compare with the situation without common noise. As we are mainly interested with the case when ε is small, we assume ε to be in $[0,1]$ in any case. As for the coefficients f and g, they are assumed to be bounded and Lipschitz continuous. We write

$$\|f\|_{1,\infty} = \sup_{x\in\mathbb{R}^d} |f(x)| + \sup_{x,x'\in\mathbb{R}^d : x\neq x'} \frac{|f(x)-f(x')|}{|x-x'|} < \infty,$$

and similarly for g.

2.1.9.2 Notation

The notation I_d stands for the d-dimensional identity matrix. For a random variable Z with values in a Polish space $\mathcal{S}$, we denote by $\sigma(Z)$ the σ-field generated by Z. For a process $\boldsymbol{Z} = (Z_t)_{0\leqslant t\leqslant T}$ with values in a Polish space $\mathcal{S}$, we denote by $\mathbb{F}^{\boldsymbol{Z}}$ the augmented filtration generated by $\boldsymbol{Z}$. In particular, the notation $\mathbb{F}^{\boldsymbol{W}}$ is frequently used to denote the augmented filtration generated by the common noise when $\varepsilon \in (0,1]$. We recall that $\mathbb{F} = \mathbb{F}^{(X_0,\sigma\boldsymbol{B},\varepsilon\boldsymbol{W})}$.

Also, for a filtration $\mathbb{G}$, we write $\mathbb{S}^d(\mathbb{G})$ for the space of continuous and $\mathbb{G}$-adapted processes $\boldsymbol{Z} = (Z_t)_{0\leqslant t\leqslant T}$ with values in $\mathbb{R}^d$ that satisfy

$$\mathbb{E}\Big[\sup_{0\leqslant t\leqslant T} |Z_t|^2\Big] < \infty.$$

2.1.9.3 Organization of the paper

The mathematical analysis is carried out in Section 2.2. Subsection 2.2.1 addresses the error associated with the scheme (2.1.24)–(2.1.20)–(2.1.21)–(2.1.25) without any time discretization, see Theorem 2.2.4. Similar results, but with the additional time discretization that makes the exploration possible, are established in Subsection 2.2.2, see in particular Theorem 2.2.17. In Subsection 2.2.3, we make the connection with the original mean field game without common noise. In particular, we prove that our learning procedure permits to construct approximate equilibria to the original problem. The bound (2.1.33) for the exploitability is established in Theorem 2.2.27. Following our agenda, we provide the results of some numerical experiments in Section 2.3. The method is tested on some benchmark examples that are presented in Subsection 2.3.1. The implemented version of the algorithm is explained in Subsection 2.3.2. The results, for a fixed intensity of the common noise, are exposed in Subsection 2.3.3. In the final Subsection 2.3.5, we provide an example that illustrates the behavior of the algorithm for a decreasing intensity of the common noise.

2.2 Theoretical results

The theoretical results are presented in three main steps. The general philosophy, as exposed in Figure 2.1, is addressed in Subsection 2.2.1. The analysis of the algorithm under a time-discrete randomization of the actions is addressed in Subsection 2.2.2. Lastly, the dilemma between exploration and exploitation is investigated in Subsection 2.2.3.

2.2.1 Tilted fictitious play with common noise

Throughout the subsection, the intensity $\varepsilon \in (0,1]$ of the common noise and the learning parameter ϖ are fixed. We take ϖ in $(1,\sqrt{2}]$ (we refer to Remark 2.2.11 for the need to have ϖ close to 1).

2.2.1.1 Construction of the learning sequence

We here formalize the scheme introduced in (2.1.20)–(2.1.21)–(2.1.24)–(2.1.25). We hence construct a sequence of proxies $(\boldsymbol{m}^n)_{n\geqslant 0}$ and $(\boldsymbol{h}^n)_{n\geqslant 0}$. The two initial processes $\boldsymbol{m}^0$ and $\boldsymbol{h}^0$ are two $\mathbb{F}^{\boldsymbol{W}}$-adapted continuous processes with values in $\mathbb{R}^d$. Typically, we choose $\boldsymbol{m}^0 = (m_t^0 = \mathbb{E}(X_0))_{0\leqslant t\leqslant T}$ and $\boldsymbol{h}^0 = (h_t^0 = 0)_{0\leqslant t\leqslant T}$. Assuming that, at some rank n, we have already defined $(\boldsymbol{m}^1,\cdots,\boldsymbol{m}^n)$ and $(\boldsymbol{h}^1,\cdots,\boldsymbol{h}^n)$, each in $\mathbb{S}^d(\mathbb{F}^{\boldsymbol{W}})$, we call

$$\boldsymbol{\alpha}^{n+1,\varpi} := \mathrm{argmin}_{\boldsymbol{\alpha}} \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{\varpi X_0}\Big(\varpi\boldsymbol{\alpha}; \overline{\boldsymbol{m}}^n; \varpi\varepsilon \boldsymbol{W}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big)\Big], \tag{2.2.1}$$

the infimum being taken over controlled processes $\boldsymbol{\alpha}$ that are $\mathbb{F}$-progressively and that satisfy $\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\int_0^T |\alpha_t^{n+1,\varpi}|^2 \mathrm{d}t < \infty$. Above, the process $\overline{\boldsymbol{m}}^n = (\overline{m}_t^n)_{0\leqslant t\leqslant T}$ is defined in terms of $(\boldsymbol{m}^0,\cdots,\boldsymbol{m}^n)$ through the formulas $\overline{\boldsymbol{m}}^0 := \boldsymbol{m}^0$ (if $n=0$) and

$$\overline{m}_t^n := \frac{\varpi(1-\varpi^{-1})}{1-\varpi^{-n}} \sum_{k=1}^n \varpi^{-k} m_t^k, \quad t\in[0,T], \tag{2.2.2}$$

if $n \geqslant 1$, with the latter being consistent with the geometric updating rule (2.1.25). The following lemma, the proof of which is deferred to Subsection 2.2.1.5, explains how the next proxy $\boldsymbol{m}^{n+1}$ can be computed through the best response of the control problem (2.2.1):

Lemma 2.2.1. *Under the above assumptions, the process $\boldsymbol{\alpha}^{n+1,\varpi}$ writes*

$$\alpha_t^{n+1,\varpi} = -\Big(\eta_t X_t^{n+1,\varpi} + h_t^{n+1}\Big), \quad t\in[0,T],$$

where

(i) $\boldsymbol{\eta} = (\eta_t)_{0\leqslant t\leqslant T}$ solves the Riccati equation:

$$\dot{\eta}_t - \eta_t^2 + Q^\dagger Q = 0, \quad t\in[0,T]\,; \quad \eta_T = R^\dagger R; \tag{2.2.3}$$

(ii) $\boldsymbol{h}^{n+1} = (h_t^{n+1})_{0\leqslant t\leqslant T} \in \mathbb{S}^d(\mathbb{F}^{\boldsymbol{W}})$ solves the backward SDE:

$$\begin{aligned} &\mathrm{d}h_t^{n+1} = \big(-\tfrac{1}{\varpi}Q^\dagger f(\overline{m}_t^n) + \eta_t h_t^{n+1}\big)\mathrm{d}t + \varepsilon k_t^{n+1}\mathrm{d}W_t^{\varpi \boldsymbol{h}^n/\varepsilon}, \quad t\in[0,T], \\ &h_T^{n+1} = \tfrac{1}{\varpi}R^\dagger g\big(\overline{m}_T^n\big); \end{aligned} \tag{2.2.4}$$

(iii) $\boldsymbol{X}^{n+1,\varpi} = (X_t^{n+1,\varpi})_{0\leqslant t\leqslant T}$ solves the forward SDE:

$$\mathrm{d}X_t^{n+1,\varpi} = -\Big(\eta_t X_t^{n+1,\varpi} + h_t^{n+1}\Big)\mathrm{d}t + \tfrac{1}{\varpi}\sigma \mathrm{d}B_t + \varepsilon \mathrm{d}W_t^{\varpi \boldsymbol{h}^n/\varepsilon}, \quad t\in[0,T]; \quad X_0^{n+1,\varpi} = X_0.$$

The statement makes it possible to let

$$m_t^{n+1} := \mathbb{E}\big[X_t^{n+1,\varpi} \,|\, \sigma(\boldsymbol{W})\big], \quad t \in [0,T]; \quad m_0^{n+1} = \mathbb{E}(X_0), \tag{2.2.5}$$

and then, consistently with (2.2.2),

$$\overline{m}_t^{n+1} := \tfrac{\varpi(1-\varpi^{-1})}{1-\varpi^{-(n+1)}} \sum_{k=1}^{n+1} \varpi^{-k} m_t^k = \tfrac{\varpi^{-n}(1-\varpi^{-1})}{1-\varpi^{-(n+1)}} m_t^{n+1} + \Big(1 - \tfrac{\varpi^{-n}(1-\varpi^{-1})}{1-\varpi^{-(n+1)}}\Big)\overline{m}_t^n. \tag{2.2.6}$$

Remark 2.2.2. *Notice that because the density $\mathcal{E}(\varpi \boldsymbol{h}^n/\varepsilon)$ is $\sigma(\boldsymbol{W})$-measurable, we also have*

$$m_t^{n+1} = \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\big[X_t^{n+1,\varpi} \,|\, \sigma(\boldsymbol{W})\big], \quad t \in [0,T],$$

*which is a direct consequence of Bayes' rule, see [***JacodProtter***].*

Remark 2.2.3. *Although each $\overline{\boldsymbol{m}}^n$ depends in an obvious manner on ϖ, this is our choice not to add ϖ as a label in the notation. The reason is that the limit is independent of ϖ, as clarified in §2.2.1.2 below. Similarly, we do not put ϖ in the notation $\boldsymbol{h}^n$: as shown in Theorem 2.2.4, the limit is independent of ϖ up to a rescaling by ϖ. This is in contrast with the quantities $\boldsymbol{X}^{n,\varpi}$ and $\boldsymbol{\alpha}^{n,\varpi}$: the limits depend on ϖ in a non-trivial manner, which we also make clear in the next paragraph.*

On another matter, it must be noticed that Lemma 2.2.1 remains valid when $\varpi = 1$. As explained in Introduction, see (2.1.29)*, the updating rate in* (2.2.6) *must then be understood as $1/(n+1)$ and, accordingly, the formula* (2.2.2) *coincides with the harmonic updating rule* (2.1.22)*. This remark is important in order to compare the two harmonic and geometric schemes.*

2.2.1.2 Main statement

As noticed in the introduction, it is especially convenient to reformulate the dynamics of $\boldsymbol{X}^{n+1,\varpi}$ under the historical probability. Quite clearly, we can write:

$$\mathrm{d}X_t^{n+1,\varpi} = -\Big(\eta_t X_t^{n+1,\varpi} + h_t^{n+1} - \varpi h_t^n\Big)\mathrm{d}t + \tfrac{1}{\varpi}\sigma \mathrm{d}B_t + \varepsilon \mathrm{d}W_t, \quad t \in [0,T]; \quad X_0^{n+1,\varpi} = X_0, \tag{2.2.7}$$

and then

$$\mathrm{d}m_t^{n+1} = -\Big(\eta_t m_t^{n+1} + h_t^{n+1} - \varpi h_t^n\Big)\mathrm{d}t + \varepsilon \mathrm{d}W_t, \quad t \in [0,T]; \quad m_0^{n+1} = \mathbb{E}(X_0). \tag{2.2.8}$$

By dividing (2.2.8) by ϖ^{n+1}, then summing over the index n and eventually multiplying by $\varpi(1-\varpi^{-1})/(1-\varpi^{-(n+1)})$, we get:

$$\mathrm{d}\overline{m}_t^{n+1} = -\Big(\eta_t \overline{m}_t^{n+1} + \tfrac{(1-\varpi^{-1})\varpi^{-n}}{1-\varpi^{-(n+1)}} h_t^{n+1}\Big)\mathrm{d}t + \varepsilon \mathrm{d}W_t, \quad t \in [0,T]. \tag{2.2.9}$$

By coupling with the backward equation in the statement of Lemma 2.2.1 (when reformulated under the historical probability), we obtain the forward-backward system:

$$\begin{aligned}
&\mathrm{d}\overline{m}_t^{n+1} = -\Big(\eta_t \overline{m}_t^{n+1} + \tfrac{(1-\varpi^{-1})\varpi^{-n}}{1-\varpi^{-(n+1)}} h_t^{n+1}\Big)\mathrm{d}t + \varepsilon \mathrm{d}W_t, \\
&\mathrm{d}h_t^{n+1} = \Big(-\tfrac{1}{\varpi} Q^\dagger f\big(\overline{m}_t^n\big) + \eta_t h_t^{n+1}\Big)\mathrm{d}t + \varpi k_t^{n+1} h_t^n \mathrm{d}t + \varepsilon k_t^{n+1} \mathrm{d}W_t, \quad t \in [0,T], \\
&h_T^{n+1} = \tfrac{1}{\varpi} R^\dagger g\big(\overline{m}_T^n\big).
\end{aligned} \tag{2.2.10}$$

Our main result in this regard is the following statement:

Theorem 2.2.4. *The scheme* (2.2.10) *converges to the decoupled version of the FBSDE system:*

$$\begin{aligned}
&\mathrm{d}m_t = -\eta_t m_t \mathrm{d}t + \varepsilon \mathrm{d}W_t, \\
&\mathrm{d}h_t^{\varpi} = \Big(-\tfrac{1}{\varpi} Q^\dagger f(m_t) + \eta_t h_t^{\varpi} + \varpi k_t^{\varpi} h_t^{\varpi}\Big)\mathrm{d}t + \varepsilon k_t^{\varpi} \mathrm{d}W_t, \quad t \in [0,T], \\
&m_0 = \mathbb{E}(X_0), \quad h_T = \tfrac{1}{\varpi} R^\dagger g(m_T),
\end{aligned} \tag{2.2.11}$$

with an explicit bound on the rate of convergence, namely

$$\mathrm{essup}_{\omega\in\Omega}\Big[\sup_{0\leqslant t\leqslant T}\Big(|m_t - \overline{m}_t^n|^2 + |h_t^{\varpi} - h_t^n|^2\Big)\Big] \leqslant \varpi^{-2n}\exp\big(\tfrac{C}{\varepsilon^2}\big), \tag{2.2.12}$$

for a constant C that depends on d, T and the norms $\|Q^\dagger f\|_{1,\infty}$ and $\|R^\dagger g\|_{1,\infty}$.

Moreover, up to a modification of the constant C, the weak error of the scheme for the Fortet-Mourier distance satisfies:

$$\sup_F \Big|\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[F\big(\overline{\boldsymbol{m}}^n, \boldsymbol{h}^n\big)\Big] - \mathbb{E}^{\varpi \boldsymbol{h}^{\varpi}/\varepsilon}\Big[F\big(\boldsymbol{m}, \boldsymbol{h}\big)\Big]\Big| \leqslant \varpi^{-n}\exp\big(\tfrac{C}{\varepsilon^2}\big), \tag{2.2.13}$$

the supremum in the left-hand side being taken over all the functions F on $\mathcal{C}([0,T];\mathbb{R}^d\times\mathbb{R}^d)$ that are bounded by 1 *and* 1*-Lipschitz continuous.*

Importantly (and this is the interest of the result), the solution $(m_t, \varpi h_t^{\varpi})_{0\leqslant t\leqslant T}$ of the system (2.2.11) should be regarded as a solution of the (original) MFG with common noise (2.1.1)–(2.1.2)–(2.1.3), whenever the latter is formulated in the weak form. Indeed, for $\boldsymbol{m} = (m_t)_{0\leqslant t\leqslant T}$ and $\boldsymbol{h} := \varpi \boldsymbol{h}^{\varpi} = (\varpi h_t^{\varpi})_{0\leqslant t\leqslant T}$ as in (2.2.11), the optimal path $(X_t)_{0\leqslant t\leqslant T}$ associated with the minimization problem

$$\inf_{\boldsymbol{\alpha}} \mathbb{E}^{\boldsymbol{h}/\varepsilon}\Big[\mathcal{R}^{X_0}\Big(\boldsymbol{\alpha}; \boldsymbol{m}; \varepsilon \boldsymbol{W}^{\boldsymbol{h}/\varepsilon}\Big)\Big] \tag{2.2.14}$$

has exactly $\boldsymbol{m} = (m_t)_{0\leqslant t\leqslant T}$ as conditional expectation given the common noise, which follows from an obvious adaptation of Lemma 2.2.1. Namely,

$$m_t = \mathbb{E}\big[X_t \,|\, \sigma(\boldsymbol{W})\big], \quad t \in [0,T],$$

where $(X_t)_{0\leqslant t\leqslant T}$ is the optimal trajectory to the latter cost functional. As before, this can be rewritten as

$$m_t = \mathbb{E}^{\boldsymbol{h}/\varepsilon}\big[X_t \,|\, \sigma(\boldsymbol{W})\big], \quad t \in [0,T].$$

Indeed, under $\mathbb{P}^{\boldsymbol{h}/\varepsilon}$, the process $(\boldsymbol{m}, \boldsymbol{h})$ satisfies the following forward-backward system, which characterizes the conditional expectation of the optimal trajectory to (2.2.14):

$$\begin{aligned}
&\mathrm{d}m_t = -\eta_t m_t \mathrm{d}t - h_t \mathrm{d}t + \varepsilon \mathrm{d}W_t^{\boldsymbol{h}/\varepsilon}, \\
&\mathrm{d}h_t = \big(-Q^\dagger f(m_t) + \eta_t h_t\big)\mathrm{d}t + \varepsilon k_t \mathrm{d}W_t^{\boldsymbol{h}/\varepsilon}, \quad t \in [0,T], \\
&m_0 = \mathbb{E}(X_0), \quad h_T = R^\dagger g(m_T).
\end{aligned} \tag{2.2.15}$$

To derive the above system, it suffices to write it first under $\mathbb{P}$:

$$\begin{aligned} &\mathrm{d}m_t = -\eta_t m_t \mathrm{d}t + \varepsilon \mathrm{d}W_t, \\ &\mathrm{d}h_t = \big(-Q^\dagger f(m_t) + \eta_t h_t + k_t h_t\big)\mathrm{d}t + \varepsilon k_t \mathrm{d}W_t, \quad t \in [0,T], \\ &m_0 = \mathbb{E}(X_0), \quad h_T = R^\dagger g(m_T). \end{aligned} \tag{2.2.16}$$

By the changes of variable (with the first one being already defined in the paragraph preceding (2.2.14))

$$h_t^\varpi = \tfrac{1}{\varpi} h_t, \quad k_t^\varpi := \tfrac{1}{\varpi} k_t, \quad t \in [0,T], \tag{2.2.17}$$

we indeed recover (2.2.11). As we claimed, this identifies the pair $(\boldsymbol{m}, \boldsymbol{h} = \varpi \boldsymbol{h}^\varpi)$ as an equilibrium of the original MFG (2.1.1)–(2.1.2)–(2.1.3).

Noticeably, the two systems (2.2.15) and (2.2.16) provide two distinct representations of the solution θ_ε to the PDE (2.1.8). The connection between the two is given by the identities

$$h_t = \theta_\varepsilon(t, m_t),\ t \in [0,T]\,; \quad k_t = \nabla_x \theta_\varepsilon(t, m_t),\ t \in [0,T), \tag{2.2.18}$$

which holds true under both $\mathbb{P}$ and $\mathbb{P}^{\boldsymbol{h}/\varepsilon}$. By a standard application of the maximum principle (for PDEs), θ_ε is bounded in terms of d, T, $\|Q^\dagger f\|_{1,\infty}$ and $\|R^\dagger g\|_{1,\infty}$, and, by Lemma 2.2.8 right below, $\nabla_x \theta_\varepsilon$ is also bounded in terms of d, ε, T, $\|Q^\dagger f\|_{1,\infty}$ and $\|R^\dagger g\|_{1,\infty}$. The relationships stated in (2.2.18) show in fact how to construct easily a solution to (2.2.15) and (2.2.16) by solving first for $(m_t)_{0 \leqslant t \leqslant T}$ in the forward equation and then by expanding $(\theta_\varepsilon(t, m_t))_{0 \leqslant t \leqslant T}$. The hence constructed solution $(h_t, k_t)_{0 \leqslant t \leqslant T}$ to the backward equation in (2.2.16) (equivalently (2.2.15)) is bounded. In turn, uniqueness to the backward equation in (2.2.16) (equivalently (2.2.15)) is easily shown to hold true in the class of bounded processes $(h_t, k_t)_{0 \leqslant t \leqslant T}$. Similarly, (2.2.11) has a unique solution, which is then obtained by the change of variable (2.2.17).

Remark 2.2.5. *The reader will observe that, in the equation* (2.2.7) *for the optimal trajectory in the fictitious play, the intensity of the idiosyncratic noise is σ/ϖ. This is different from the intensity of the idiosyncratic noise underpinning the controlled trajectories in the MFG* (2.1.1)–(2.1.2)–(2.1.3)*, which is σ. In particular, the sequence of processes $(\boldsymbol{X}^{n,\varpi})_{n \geqslant 1}$, defined in Lemma 2.2.1 (see also* (2.2.7)*), cannot be 'a good approximation' of the process $\boldsymbol{X}$ that minimizes* (2.2.14)*. Although this looks paradoxal with the result stated in Theorem 2.2.4, it must be clear that, implicitly, the two bounds* (2.2.12) *and* (2.2.13) *rely on the fact that the dynamics for the MFG equilibrium $\boldsymbol{m}$ does not depend on σ.*

Remark 2.2.6. *Our construction of a fictitious play with a geometric learning rate, proportional to ϖ^{-n} (see* (2.2.6)*), as opposed to the harmonic learning rate $1/(n+1)$ that is used in the standard version of the fictitious play, could be easily extended to other (neither harmonic nor geometric) rates. There are two key principles that should be followed: the first one is to make appear, in the dynamics* (2.2.8)*, a difference between the component $\boldsymbol{h}^{n+1}$ of the scheme at iteration $n+1$ and the component $\boldsymbol{h}^n$ of the scheme at iteration n, and the second one is to define the tilted measure depending on the form of the finite difference (which is exactly the case in* (2.2.1)*). What is remarkable in this approach is that the tilted measure has a bias. Whereas it would be natural to take the tilted measure as $\mathbb{P}^{\boldsymbol{h}^n/\varepsilon}$, it is here taken as $\mathbb{P}^{\varpi \boldsymbol{h}^n/\varepsilon}$ with $\varpi \neq 1$. The reader will easily*

see that, in order to recover the 'non-biased' setting (which formally corresponds to $\varpi = 1$), one needs to replace ϖh_t^n by h_t^n in (2.2.7) *and* (2.2.8)*. By summing over n as in* (2.2.9)*, we would then obtain $1/(n+1)$ as learning rate in* (2.2.6)*. In other words, the 'non-biased' regime corresponds to the harmonic fictitious play.*

Although Theorem 2.2.4 does not cover the case $\varpi = 1$, the interested will easily check that the proof that is given below also works when $\varpi = 1$. Then, the bound in (2.2.12) *must be replaced by $\exp\big(C\varepsilon^{-2}\big)/n^2$. Similarly,* (2.2.13) *must be replaced by $\exp\big(C\varepsilon^{-2}\big)/n$. In fact, the same remark applies to the forthcoming Theorems 2.2.17 and 2.2.27.*

Remark 2.2.7. *We notice for later purposes that the backward equation in* (2.2.15) *can be 'solved' explicitly. Indeed, calling $(P_t)_{0\leqslant t\leqslant T}$ the solution of the linear differential equation $\dot{P}_t = -P_t\eta_t$, for $t \in [0,T]$ with $P_0 = I_d$ as initial condition, where I_d is the d-dimensional identity matrix, it holds*

$$\mathrm{d}\big[P_t h_t\big] = -P_t Q^\dagger f\big(m_t\big)\mathrm{d}t + \varepsilon P_t k_t \mathrm{d}W_t^{\boldsymbol{h}/\varepsilon},$$

or equivalently,

$$P_t h_t = P_T R^\dagger g\big(m_T\big) + \int_t^T P_s Q^\dagger f\big(m_s\big)\mathrm{d}s - \varepsilon \int_t^T P_s k_s \mathrm{d}W_s^{\boldsymbol{h}/\varepsilon}, \quad t \in [0,T].$$

2.2.1.3 Discussion about the rate of convergence

Beside any specific application to learning, this is another of our contributions to provide an explicit bound for the rate of convergence of our variant of the fictitious play for mean field games with common noise. We feel worth to point out that, to the best of our knowledge, there are very few results on the rate of convergence for the fictitious play in the absence of common noise, whether the mean field game be potential or monotone (as we already explained, no result is available without common noise outside the potential or monotone cases, except [96] which is for a time-continuous version of the fictitious play). In most of the existing references, the analysis indeed involves an additional compactness argument which complicates the computation of the rate. Still, the reader can find in [59] an explicit rate for the exploitability for a monotone and potential game set in discrete time; the bound is of order $O(1/\sqrt{n})$ (hence weaker than ours). In [96], a bound is shown, also for the exploitability, but for the time-continuous version of the fictitious play, when the game is monotone; it is of order $O(1/n)$, and is also weaker than ours in the geometric setting but is hence comparable to ours in the harmonic framework.

In comparison, the thrust of our analysis is to provide a scheme with a geometric decay. Obviously, this must be tempered, due to the presence of the multiplicative constant $\exp(C\varepsilon^{-2})$. When the intensity ε is away from zero, the effective decay is really good, but when ε gets close to 0 (which is the typical regime when we use the common noise as an exploration noise), the exponential factor really matters. Of course the bound for the error may be rewritten as follows:

$$\exp\big(-n\ln(\varpi) + \tfrac{C}{\varepsilon^2}\big),$$

which says that n should be chosen on a scale larger than ε^{-2}. We think that this is numerically affordable. In comparison (see Remark 2.2.6), if one had to work with the standard fictitious play

(i.e., with a harmonic learning rate), the same analysis would lead to a bound of order $\exp(C\varepsilon^{-2})/n$, which is obviously much worse. In the first arXiv version [48] of this work, dedicated to the analysis of the sole harmonic regime, we claimed that this was possible to remove the exponential factor, but there was a clear mistake in the computations: for convenience reasons, we decided not to indicate the dependence upon ε in the various tilted measures, as a consequence of which we forgot a factor $1/\varepsilon$ in some of the main estimates.

The presence of the exponential factor comes from the following lemma, whose proof is deferred to the end of the subsection:

Lemma 2.2.8. *There exists a constant C_1, only depending on d, T, $\|Q^\dagger f\|_{1,\infty}$ and $\|R^\dagger g\|_{1,\infty}$, such that*

$$|\nabla_x \theta_\varepsilon(t,x)| \leqslant \frac{C_1}{\varepsilon^2}, \quad t \in [0,T),\ x \in \mathbb{R}^d,$$

with θ_ε the solution of the PDE (2.1.8).

The above L^∞ bound for the gradient is known to be sharp, but it looks rather poor because the estimate is precisely given in L^∞. Also, one may hope for better estimates in different norms. In other words, θ_ε may indeed become very steep, but maybe only on some localized parts of the space. This is the point where things become highly subtle because the process $\overline{\boldsymbol{m}}^n$ becomes localized itself as the diffusion coefficient ε tends to 0. So, the challenging question is to decide whether it may stay or not in parts of the space where the gradient is high. As exemplified in the analysis performed in Delarue and Foguen [45], this may be a challenging question, even in dimension 1. Unless we make additional assumptions on the model (assuming for instance that θ_ε is smooth independently of ε, which is for example the case when T is small enough), we must confess that we are not able to provide a more relevant bound for the gradient of θ_ε (which bound gives in the end a bound for the process $\boldsymbol{k}$ in (2.2.15), see (2.2.18)).

It must be also stressed that (2.2.13) provides a bound for the so-called *weak* error of the scheme and is fully relevant from the practical point of view. In our context, the *strong* error, as addressed in (2.2.12), does not provide the same information. Indeed, because the two densities $\mathcal{E}(\frac{1}{\varepsilon}\boldsymbol{h})$ and $\mathcal{E}(\frac{\varpi}{\varepsilon}\boldsymbol{h}^n)$ become singular when ε tends to 0, the passage from the strong to the weak error is not direct.

Remark 2.2.9. *The reader may wonder about the scope of Theorem 2.2.4 in the higher dimensional framework. This question will be addressed from a purely numerical prospect in the forthcoming Section 2.3, see in particular Subsection 2.3.4. From a more theoretical point of view, the same question can be addressed by investigating the dependence of the constant C in* (2.2.13) *upon the dimension d. Whereas we cannot provide a sharp (or at least reasonable) bound in full generality, we show below that, even in simple cases when the matrices Q and R reduce to the $d \times d$ identity matrix and the coefficients f and g are diagonal, i.e. each coordinate i of f (respectively g) writes $f^i(x) = f_0(x_i)$ (respectively $g^i(x) = g_0(x_i)$) for a function $f_0 : \mathbb{R} \to \mathbb{R}$ (respectively $g_0 : \mathbb{R} \to \mathbb{R}$), with x_i denoting the i^{th} coordinate of x, then the exponential factor in* (2.2.12) *and* (2.2.13) *is typically $\exp(C\mathcal{O}(\sqrt{d})\varepsilon^{-2})$, for C independent of d.*

2.2.1.4 Proof of Theorem 2.2.4

The proof relies on the following lemma, whose proof is also deferred to the end of the subsection.

Lemma 2.2.10. *There exists a constant $C_2 > 0$, only depending on d, T, $\|Q^\dagger f\|_{1,\infty}$ and $\|R^\dagger g\|_{1,\infty}$, such that, $\mathbb{P}$ almost surely,*

$$|h_t| \leqslant C_2\,; \quad |h_t^n| \leqslant C_2, \quad t \in [0,T],\ n \geqslant 1.$$

Proof of Theorem 2.2.4. Throughout, C is a generic constant that is allowed to vary from line to line, as long as it only depends on d, T, $\|Q^\dagger f\|_{1,\infty}$ and $\|R^\dagger g\|_{1,\infty}$.

First Step. Invoking Lemma 2.2.10 and recalling that $\varpi \in (1,\sqrt{2}]$, we then have

$$\Big|\tfrac{(1-\varpi^{-1})\varpi^{-n}}{1-\varpi^{-(n+1)}} h_t^{n+1}\Big| \leqslant C\varpi^{-n}, \quad t \in [0,T], \tag{2.2.19}$$

from which we deduce (consider the difference between (2.2.9) and the forward equation in (2.2.11)) that

$$\sup_{0\leqslant t\leqslant T} \big|\overline{m}_t^n - m_t\big| \leqslant C\varpi^{-n}. \tag{2.2.20}$$

We now make the difference between the backward equations in (2.2.10) and (2.2.11). We obtain

$$\begin{aligned}
\mathrm{d}\big(h_t^{n+1} - h_t^\varpi\big) &= -\tfrac{1}{\varpi}\Big(Q^\dagger f\big(\overline{m}_t^n\big) - Q^\dagger f\big(m_t\big)\Big)\mathrm{d}t + \eta_t\big(h_t^{n+1} - h_t^\varpi\big)\mathrm{d}t + \varpi k_t^{n+1}\big(h_t^n - h_t^\varpi\big)\mathrm{d}t \\
&\quad + \varpi\big(k_t^{n+1} - k_t^\varpi\big)h_t^\varpi \mathrm{d}t + \varepsilon\big(k_t^{n+1} - k_t^\varpi\big)\mathrm{d}W_t, \\
h_T^{n+1} - h_T^\varpi &= \tfrac{1}{\varpi}\Big[R^\dagger g\big(\overline{m}_T^n\big) - R^\dagger g\big(m_T\big)\Big].
\end{aligned} \tag{2.2.21}$$

In particular, rewriting the above equation under $\boldsymbol{W}^{\boldsymbol{h}/\varepsilon} = \boldsymbol{W}^{\varpi \boldsymbol{h}^\varpi/\varepsilon}$, we get

$$\begin{aligned}
\mathrm{d}\big(h_t^{n+1} - h_t^\varpi\big) &= -\tfrac{1}{\varpi}\Big(Q^\dagger f\big(\overline{m}_t^n\big) - Q^\dagger f\big(m_t\big)\Big)\mathrm{d}t + \eta_t\big(h_t^{n+1} - h_t^\varpi\big)\mathrm{d}t + \varpi k_t^{n+1}\big(h_t^n - h_t^\varpi\big)\mathrm{d}t \\
&\quad + \varepsilon\big(k_t^{n+1} - k_t^\varpi\big)\mathrm{d}W_t^{\boldsymbol{h}/\varepsilon}, \quad t \in [0,T].
\end{aligned} \tag{2.2.22}$$

Then, taking the square, using (2.2.20) and Lemmas 2.2.8 and 2.2.10, expanding by Itô's formula and applying Young's inequality, we get

$$\begin{aligned}
\mathrm{d}\big|h_t^{n+1} - h_t^\varpi\big|^2 &\geqslant -C\varpi^{-2n}\mathrm{d}t - C\big|h_t^{n+1} - h_t^\varpi\big|^2\mathrm{d}t - C\big|k_t\big|\big|h_t^n - h_t^\varpi\big|\big|h_t^{n+1} - h_t^\varpi\big|\mathrm{d}t \\
&\quad - C\big|k_t^{n+1} - k_t^\varpi\big|\big|h_t^n - h_t^\varpi\big|\big|h_t^{n+1} - h_t^\varpi\big|\mathrm{d}t \\
&\quad + \varepsilon^2\big|k_t^{n+1} - k_t^\varpi\big|^2\mathrm{d}t + 2\varepsilon\big(h_t^{n+1} - h_t^\varpi\big)\cdot\big[\big(k_t^{n+1} - k_t^\varpi\big)\mathrm{d}W_t^{\boldsymbol{h}/\varepsilon}\big] \\
&\geqslant -C\varpi^{-2n}\mathrm{d}t - \tfrac{C}{\varepsilon^2}\big|h_t^n - h_t^\varpi\big|^2\mathrm{d}t - \tfrac{C}{\varepsilon^2}\big|h_t^{n+1} - h_t^\varpi\big|^2\mathrm{d}t \\
&\quad + 2\varepsilon\big(h_t^{n+1} - h_t^\varpi\big)\cdot\big[\big(k_t^{n+1} - k_t^\varpi\big)\mathrm{d}W_t^{\boldsymbol{h}/\varepsilon}\big],
\end{aligned} \tag{2.2.23}$$

for $t \in [0,T]$ and for a constant C as in the statement. For a free parameter $\lambda > 1$, we obtain

$$\mathrm{d}\Big[\exp\big(\tfrac{C}{\varepsilon^2}\lambda t\big)\big|h_t^{n+1} - h_t^\varpi\big|^2\Big] \geqslant \exp\big(\tfrac{C}{\varepsilon^2}\lambda t\big)\Big(\tfrac{C(\lambda-1)}{\varepsilon^2}\big|h_t^{n+1} - h_t^\varpi\big|^2 - C\varpi^{-2n} - \tfrac{C}{\varepsilon^2}\big|h_t^n - h_t^\varpi\big|^2\Big)\mathrm{d}t$$

$$+ 2\varepsilon \exp\big(\tfrac{C}{\varepsilon^2}\lambda t\big)\big(h_t^{n+1} - h_t^{\varpi}\big)\cdot\big[\big(k_t^{n+1} - k_t^{\varpi}\big)\mathrm{d}W_t^{\boldsymbol{h}/\varepsilon}\big]. \tag{2.2.24}$$

And then, integrating from t to T, taking conditional expectation under $\mathbb{P}^{\boldsymbol{h}/\varepsilon}$ given $\mathcal{F}_t$ and recalling that $h_T^{n+1} = R^\dagger g(\overline{m}_T^n)/\varpi$ and $h_T^{\varpi} = R^\dagger g(m_T)/\varpi$, we get, for any $t \in [0,T]$,

$$\begin{aligned}
&\exp\big(\tfrac{C}{\varepsilon^2}\lambda t\big)\big|h_t^{n+1} - h_t^{\varpi}\big|^2 + \tfrac{C(\lambda-1)}{\varepsilon^2}\mathbb{E}^{\boldsymbol{h}/\varepsilon}\bigg[\int_t^T \exp\big(\tfrac{C}{\varepsilon^2}\lambda s\big)\big|h_s^{n+1} - h_s^{\varpi}\big|^2\mathrm{d}s \,|\, \mathcal{F}_t\bigg] \qquad (2.2.25)\\
&\quad\leqslant C\varpi^{-2n}\Big(\exp\big(\tfrac{C}{\varepsilon^2}\lambda T\big) + \int_t^T \exp\big(\tfrac{C}{\varepsilon^2}\lambda s\big)\mathrm{d}s\Big) + \tfrac{C}{\varepsilon^2}\mathbb{E}^{\boldsymbol{h}/\varepsilon}\bigg[\int_t^T \exp\big(\tfrac{C}{\varepsilon^2}\lambda s\big)\big|h_s^{n} - h_s^{\varpi}\big|^2\mathrm{d}s \,|\, \mathcal{F}_t\bigg].
\end{aligned}$$

Next, we choose $\lambda = 4$. This yields

$$\begin{aligned}
&\tfrac{\varepsilon^2}{3C}\exp\big(\tfrac{4C}{\varepsilon^2}t\big)\big|h_t^{n+1} - h_t^{\varpi}\big|^2 + \mathbb{E}^{\boldsymbol{h}/\varepsilon}\bigg[\int_t^T \exp\big(\tfrac{4C}{\varepsilon^2}s\big)\big|h_s^{n+1} - h_s^{\varpi}\big|^2\mathrm{d}s \,|\, \mathcal{F}_t\bigg]\\
&\quad\leqslant \tfrac{\varepsilon^2(1+T)}{3}\varpi^{-2n}\exp\big(\tfrac{4C}{\varepsilon^2}T\big) + \tfrac{1}{3}\mathbb{E}^{\boldsymbol{h}/\varepsilon}\bigg[\int_t^T \exp\big(\tfrac{4C}{\varepsilon^2}s\big)\big|h_s^{n} - h_s^{\varpi}\big|^2\mathrm{d}s \,|\, \mathcal{F}_t\bigg].
\end{aligned} \tag{2.2.26}$$

The above holds true for almost every $\omega \in \Omega$ (under $\mathbb{P}$ and $\mathbb{P}^{\boldsymbol{h}/\varepsilon}$) and for every $t \in [0,T]$. By a standard induction argument, using in addition Lemma 2.2.10, we deduce that, for a deterministic constant C' depending on the same parameters as C, for almost every $\omega \in \Omega$, for every $t \in [0,T]$ and for every $n \geqslant 1$,

$$\begin{aligned}
\mathbb{E}^{\boldsymbol{h}/\varepsilon}\bigg[\int_t^T \exp\big(\tfrac{4C}{\varepsilon^2}s\big)\big|h_s^{n} - h_s^{\varpi}\big|^2\mathrm{d}s \,|\, \mathcal{F}_t\bigg] &\leqslant C'\exp\big(\tfrac{4C}{\varepsilon^2}T\big)\sum_{k=0}^{n} 3^{-k}\varpi^{-2(n-k)}\\
&= C'\exp\big(\tfrac{4C}{\varepsilon^2}T\big)\varpi^{-2n}\sum_{k=0}^{n}\big(\tfrac{\varpi^2}{3}\big)^k \qquad (2.2.27)\\
&\leqslant C'\exp\big(\tfrac{4C}{\varepsilon^2}T\big)\sum_{k=0}^{n}\big(\tfrac{2}{3}\big)^k \leqslant C'\exp\big(\tfrac{4C}{\varepsilon^2}T\big)\varpi^{-2n},
\end{aligned}$$

where we used, in the last line, the fact that $\varpi \in (1, \sqrt{2}]$. Because ε is less than 1, we can easily remove the constant C' by increasing the constant C. And then, by (2.2.25), we obtain (2.2.12).

Second Step. We now turn to the proof of (2.2.13). In fact, it is a direct consequence of Pinsker's inequality, which says that

$$d_{\mathrm{TV}}\big(\mathbb{P}^{\boldsymbol{h}/\varepsilon}, \mathbb{P}^{\varpi\boldsymbol{h}^n/\varepsilon}\big) \leqslant \sqrt{2}\mathbb{E}^{\boldsymbol{h}/\varepsilon}\bigg[\ln\Big(\frac{\mathcal{E}(\boldsymbol{h}/\varepsilon)}{\mathcal{E}(\varpi\boldsymbol{h}^n/\varepsilon)}\Big)\bigg]^{1/2},$$

where d_{TV} in the left-hand side is the total variation distance. Now,

$$\begin{aligned}
&\ln\Big(\frac{\mathcal{E}(\boldsymbol{h}/\varepsilon)}{\mathcal{E}(\varpi\boldsymbol{h}^n/\varepsilon)}\Big) = \ln\Big(\frac{\mathcal{E}(\varpi\boldsymbol{h}^\varpi/\varepsilon)}{\mathcal{E}(\varpi\boldsymbol{h}^n/\varepsilon)}\Big)\\
&= -\tfrac{\varpi}{\varepsilon}\int_0^T \big(h_s^\varpi - h_s^n\big)\cdot \mathrm{d}W_s - \tfrac{\varpi^2}{2\varepsilon^2}\int_0^T \big(|h_s^\varpi|^2 - |h_s^n|^2\big)\mathrm{d}s\\
&= -\tfrac{\varpi}{\varepsilon}\int_0^T \big(h_s^\varpi - h_s^n\big)\cdot \mathrm{d}W_s^{\boldsymbol{h}/\varepsilon} + \tfrac{\varpi^2}{\varepsilon^2}\int_0^T \big(h_s^\varpi - h_s^n\big)\cdot h_s^\varpi \mathrm{d}s - \tfrac{\varpi^2}{2\varepsilon^2}\int_0^T \big(|h_s^\varpi|^2 - |h_s^n|^2\big)\mathrm{d}s\\
&= -\tfrac{\varpi}{\varepsilon}\int_0^T \big(h_s^\varpi - h_s^n\big)\cdot \mathrm{d}W_s^{\boldsymbol{h}/\varepsilon} + \tfrac{\varpi^2}{2\varepsilon^2}\int_0^T \big|h_s^\varpi - h_s^n\big|^2\mathrm{d}s.
\end{aligned}$$

And then, by (2.2.12),

$$d_{\mathrm{TV}}\big(\mathbb{P}^{\boldsymbol{h}/\varepsilon}, \mathbb{P}^{\varpi\boldsymbol{h}^n/\varepsilon}\big) \leqslant \tfrac{C}{\varepsilon}\varpi^{-n}\exp\big(\tfrac{C}{2\varepsilon^2}\big). \tag{2.2.28}$$

Modifying the constant C, we can easily get rid of the multiplicative constant $1/\varepsilon$ in the right-hand side. By (2.2.12) again, for any function F that is 1-bounded and 1-Lipschitz on the space $\mathcal{C}([0,T];\mathbb{R}^d\times\mathbb{R}^d)$,

$$\Big|\mathbb{E}^{\boldsymbol{h}/\varepsilon}\Big[F\big(\overline{\boldsymbol{m}}^n, \boldsymbol{h}^n\big)\Big] - \mathbb{E}^{\boldsymbol{h}/\varepsilon}\Big[F\big(\boldsymbol{m},\boldsymbol{h}\big)\Big]\Big| \leqslant \varpi^{-n}\exp\big(\tfrac{C}{\varepsilon^2}\big),$$

and then, by (2.2.28), we get (2.2.13). □

Remark 2.2.11. *The reader can deduce from the display* (2.2.27) *the reason why we assumed* ϖ *to be less than* $\sqrt{2}$*. In fact, even though* ϖ *were larger, the geometric decay would remain less than* 3^{-n} *because of the factor* $1/3$ *in* (2.2.26)*. Here, the factor* $1/3$ *must be understood as* $1/(\lambda-1)$ *for* $\lambda = 4$*. For sure, it would be tempting to choose* $1/(\lambda-1) = 1/\varpi$*, but this would lead to* $\lambda = \varpi + 1$*. The resulting exponential factor* $\exp(C\varepsilon^{-2}\lambda s)$ *in* (2.2.25) *would become much too high with* ϖ*.*

2.2.1.5 Proof of the auxiliary statements

It now remains to prove Lemmas 2.2.1, 2.2.10 and 2.2.8 and Remark 2.2.9.

Proof of Lemma 2.2.1. The proof mainly follows from the stochastic Pontryagin principle, see for instance [119, chapter 3]. Here, the stochastic Pontryagin principle provides a necessary and sufficient condition on the dynamics of the optimal control because the cost coefficients are convex in the spatial variable. However, the very fist step of the proof is to get rid of the parameter ϖ in the cost functional $\mathcal{R}^{\varpi X_0}(\varpi\boldsymbol{\alpha};\overline{\boldsymbol{m}}^n;\varpi\varepsilon\boldsymbol{W}^{\varpi\boldsymbol{h}^n/\varepsilon})$ (see (2.2.1)).

For a control $\boldsymbol{\alpha}$, the controlled dynamics driven by ϖX_0, $\varpi\boldsymbol{\alpha}$ and the common noise $\varpi\varepsilon\boldsymbol{W}^{\varpi\boldsymbol{h}^n/\varepsilon}$ write

$$\mathrm{d}X_t^{\varpi,\boldsymbol{\alpha}} = \varpi\alpha_t\mathrm{d}t + \sigma\mathrm{d}B_t + \varpi\varepsilon\mathrm{d}W_t^{\varpi\boldsymbol{h}^n/\varepsilon}, \quad t\in[0,T]\ ; \quad X_0^{\varpi,\boldsymbol{\alpha}} = \varpi X_0.$$

Dividing by ϖ, we can write $X_t^{\varpi,\boldsymbol{\alpha}}$ in the form $X_t^{\varpi,\boldsymbol{\alpha}} = \varpi X_t^{\boldsymbol{\alpha}}$, with (the notation $X_t^{\boldsymbol{\alpha}}$ is in fact a bit abusive because the dynamics below still depend on ϖ)

$$\mathrm{d}X_t^{\boldsymbol{\alpha}} = \alpha_t\mathrm{d}t + \tfrac{1}{\varpi}\sigma\mathrm{d}B_t + \varepsilon\mathrm{d}W_t^{\varpi\boldsymbol{h}^n/\varepsilon}, \quad t\in[0,T]\ ; \quad X_0^{\boldsymbol{\alpha}} = X_0.$$

Accordingly, the cost in (2.2.1) can be rewritten in the form

$$\begin{aligned}
&\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\big[\mathcal{R}^{\varpi X_0}\big(\varpi\boldsymbol{\alpha};\overline{\boldsymbol{m}}^n;\varpi\varepsilon\boldsymbol{W}^{\varpi\boldsymbol{h}^n/\varepsilon}\big)\big]\\
&=\frac{1}{2}\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\left[\Big|\varpi X_T^{\boldsymbol{\alpha}}+g\big(\overline{m}_T^n\big)\Big|^2+\int_0^T\Big(\varpi^2|\alpha_t|^2+\big|\varpi X_t^{\boldsymbol{\alpha}}+f\big(\overline{m}_t^n\big)\big|^2\Big)\mathrm{d}t\right]\\
&=\frac{\varpi^2}{2}\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\left[\Big|X_T^{\boldsymbol{\alpha}}+\tfrac{1}{\varpi}g\big(\overline{m}_T^n\big)\Big|^2+\int_0^T\Big(|\alpha_t|^2+\big|X_t^{\boldsymbol{\alpha}}+\tfrac{1}{\varpi}f\big(\overline{m}_t^n\big)\big|^2\Big)\mathrm{d}t\right]
\end{aligned}$$

The leading parameter ϖ^2 can be easily removed (because it does not change the minimizer). Next, the stochastic Pontryagin principle says that the optimal trajectory to the cost functional in the above right-hand side –and thus the optimal trajectory to the cost functional (2.2.1) (but up to the rescaling factor ϖ)– is

$$\mathrm{d}X_t^{n+1}=-\big(\eta_t X_t^{n+1}+h_t^{n+1}\big)\mathrm{d}t+\tfrac{1}{\varpi}\sigma\mathrm{d}B_t+\varepsilon\mathrm{d}W_t^{\varpi\boldsymbol{h}^n/\varepsilon},\quad t\in[0,T]\,;\quad X_0^{n+1}=X_0,$$

with $(\eta_t)_{0\leqslant t\leqslant T}$ and $\boldsymbol{h}^{n+1}$ as in the statement. □

Proof of Lemma 2.2.10. The proof is a straightforward consequence of the two equations for $\boldsymbol{h}$ and $\boldsymbol{h}^{n+1}$ under (respectively) the probability measures $\mathbb{P}^{\boldsymbol{h}/\varepsilon}$ and $\mathbb{P}^{\varpi\boldsymbol{h}^n/\varepsilon}$, see (2.2.15) and (2.2.4). One can make the argument especially clear by using the formulation presented in Remark 2.2.7 together with the fact that the coefficients f and g are bounded. □

Proof of Lemma 2.2.8. In (2.1.8), we perform the change of variable

$$\theta_\varepsilon(t,x)=\varphi_\varepsilon\big(\tfrac{t}{\varepsilon^2},\tfrac{x}{\varepsilon^2}\big),\quad (t,x)\in[0,T]\times\mathbb{R}^d.$$

We get

$$\begin{aligned}
&\tfrac{1}{\varepsilon^2}\partial_t\varphi_\varepsilon\big(\tfrac{t}{\varepsilon^2},\tfrac{x}{\varepsilon^2}\big)+\tfrac{1}{2\varepsilon^2}\Delta_{xx}^2\varphi_\varepsilon\big(\tfrac{t}{\varepsilon^2},\tfrac{x}{\varepsilon^2}\big)-\tfrac{1}{\varepsilon^2}\big(\eta_t x+\theta_\varepsilon(t,x)\big)\cdot\nabla_x\varphi_\varepsilon\big(\tfrac{t}{\varepsilon^2},\tfrac{x}{\varepsilon^2}\big)\\
&\quad+Q^\dagger f(x)-\eta_t\theta_\varepsilon(t,x)=0.
\end{aligned}$$

Multiplying by ε^2 and changing (t,x) into $(\varepsilon^2 t,\varepsilon^2 x)$, we obtain

$$\begin{aligned}
&\partial_t\varphi_\varepsilon(t,x)+\tfrac{1}{2}\Delta_{xx}^2\varphi_\varepsilon(t,x)-\Big(\eta_t\varepsilon^2 x+\theta_\varepsilon\big(\varepsilon^2 t,\varepsilon^2 x\big)\Big)\cdot\nabla_x\varphi_\varepsilon(t,x)\\
&\quad+\varepsilon^2 Q^\dagger f\big(\varepsilon^2 x\big)-\varepsilon^2\eta_t\theta_\varepsilon\big(\varepsilon^2 t,\varepsilon^2 x\big)=0.
\end{aligned}$$

Above, (t,x) belongs to $[0,T/\varepsilon^2]\times\mathbb{R}^d$. The terminal condition is $\varphi_\varepsilon(T/\varepsilon^2,x)=g(\varepsilon^2 x)$. Moreover, by Lemma 2.2.10, the function θ_ε can be bounded independently of ε. Then, for t at distance less than 1 from T/ε^2, we get a bound for $\nabla_x\varphi_\varepsilon$ from standard estimates for systems of nonlinear parabolic PDEs, as used in [44]. When t is at distance greater than 1 from T/ε^2, we get a bound for $\nabla_x\varphi_\varepsilon$ from interior estimates for systems of nonlinear parabolic PDEs, see for instance [**MR2053051**]. The result follows. □

Proof of Remark 2.2.9. We now add a few lines in order to prove the complementary Remark 2.2.9. When R and Q are the identity matrices, the solution $\boldsymbol{\eta}$ of the Riccati equation (2.2.3) becomes an homothecy (which we identify with a real-valued function). In turn, if the functions f and g are 'diagonal', in the sense of Remark 2.2.9, then the solution θ_ε to the PDE (2.1.8) is also diagonal, namely $\theta^i_\varepsilon(t,x) = \theta_{0,\varepsilon}(t,x_i)$, with $\theta_0 : [0,T] \times \mathbb{R} \to \mathbb{R}$ being the solution of the $1d$-equation:

$$\partial_t \theta_{0,\varepsilon}(t,x_0) + \tfrac{\varepsilon^2}{2}\partial^2_{x_0x_0}\theta_{0,\varepsilon}(t,x_0) - \big(\eta_t x_0 + \theta_{0,\varepsilon}(t,x_0)\big)\partial_{x_0}\theta_{0,\varepsilon}(t,x_0) + f_0(x_0) - \eta_t\theta_{0,\varepsilon}(t,x_0) = 0,$$

for $(t,x_0) \in [0,T] \times \mathbb{R}$, and with the terminal boundary condition $\theta_{0,\varepsilon}(T,x_0) = g_0(x_0)$.

In particular,

$$\sup_{(t,x)\in[0,T]\times\mathbb{R}^d} \big|\nabla_x \theta_\varepsilon(t,x)\big|^2 = d \sup_{(t,x_0)\in[0,T]\times\mathbb{R}} \big|\partial_{x_0}\theta_{0,\varepsilon}(t,x_0)\big|^2.$$

So, if we take for granted the fact that we cannot get better than C_0/ε^2 for the right-hand side, then the bound for the left-hand side becomes $C_0 d/\varepsilon^4$. Returning back to (2.2.23), the constant C therein grows like $\sqrt{d}$ with d, which completes the proof of the claim. □

2.2.2 The common noise as an exploration noise

Following the agenda explained in the introduction, we now regard the common noise as an exploration noise for an MFG without common noise. In words, ε is set equal to 0 in the original MFG (2.1.1)–(2.1.2)–(2.1.3) and this is only in the choice of the controls that we restore the presence of the common noise, in the form of a randomization.

2.2.2.1 Presentation of the model

In the absence of common noise, the state dynamics merely write

$$\mathrm{d}X_t = \beta_t \mathrm{d}t + \sigma \mathrm{d}B_t, \quad t \in [0,T].$$

However, we want $\boldsymbol{\beta} = (\beta_t)_{0\leqslant t\leqslant T}$ to be subjected to a random exploration of the form

$$\beta_t = \alpha_t + \varepsilon \dot{W}_t, \quad t \in [0,T],$$

where $\boldsymbol{\alpha} = (\alpha_t)_{0\leqslant t\leqslant T}$ is the control effectively chosen by the agent (or by the 'controller') and ε is (strictly) positive (and is kept fixed throughout the subsection). In this expansion, $(\dot{W}_t)_{0\leqslant t\leqslant T}$ is formally understood as the time-derivative of $(W_t)_{0\leqslant t\leqslant T}$. Obviously, the latter does not exist as a function, which makes the above decomposition non tractable. However, it prompts us to introduce a variant based upon a mollification of the common noise. To make it clear, we introduce a family of regular processes $((W^p_t)_{0\leqslant t\leqslant T})_{p\geqslant 1}$ such that, almost surely (under $\mathbb{P}$),

$$\lim_{p\to\infty} \sup_{0\leqslant t\leqslant T} |W_t - W^p_t| = 0,$$

and, for any $p \geqslant 1$, the paths of $(W^p_t)_{0\leqslant t\leqslant T}$ are continuous and piecewise continuously differentiable. Throughout this subsection and the next one, we use the following piecewise affine interpolation:

$$W^p_t := W_{\tau_p(t)} + \frac{p(t-\tau_p(t))}{T}\big(W_{\tau_p(t)+T/p} - W_{\tau_p(t)}\big), \tag{2.2.29}$$

where, by definition, $\tau_p(t) := \lfloor pt/T \rfloor (T/p)$. In words $\tau_p(t)$ is the unique element of $(T/p)\cdot\mathbb{N}$ such that $\tau_p(t) \leqslant t < \tau_p(t) + T/p = \tau_p(t + T/p)$, namely $\tau_p(t) = \ell T/p$, for $t \in [\ell T/p, (\ell+1)T/p)$ and $\ell \in \{0, \cdots, p-1\}$. It is worth observing that the time derivative of $\boldsymbol{W}^p$ writes in the form of finite differences of $\boldsymbol{W}$, which explains our definition of $\boldsymbol{W}^p$ as a linear interpolation. We elaborate on this observation in Remark 2.2.12 below.

For an $\mathbb{F}^{\boldsymbol{W}}$-adapted and continuous environment $\boldsymbol{m} = (m_t)_{0\leqslant t\leqslant T}$, the cost functional, as originally defined in (2.1.2), is turned into the following discrete time version

$$\widetilde{J}^p(\boldsymbol{\alpha};\boldsymbol{m}) := \frac{1}{2}\mathbb{E}\bigg[\big|R^\dagger X_T + g(m_T)\big|^2 + \int_0^T \Big\{\big|Q^\dagger X_{\tau_p(t)} + f(m_{\tau_p(t)})\big|^2 + |\alpha_t + \varepsilon \dot{W}^p_t|^2\Big\}\mathrm{d}t\bigg],$$

where the expectation is taken over both the idiosyncratic and exploration (common) noises. Above, we require $\boldsymbol{\alpha}$ to be progressively-measurable with respect to the filtration $\mathbb{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}} := (\sigma(X_0, (B_{\tau_p(s)}, W_{\tau_p(s)})_{s\leqslant t}))_{0\leqslant t\leqslant T}$ and to be constant on any interval $[\ell T/p, (\ell+1)T/p)$, for $\ell \in \{0, \cdots, p-1\}$. The above minimization problem can be regarded as a discrete-time control problem. The need for a time discretization is twofold: (i) On the one hand, it permits to avoid any anticipativity problem, since the linear interpolation at a time $t \neq \tau_p(t)$ anticipates on the future realization of the exploration noise; (ii) On the other hand, it is more adapted to numerical purposes. To the best of our knowledge, the formulation of $\widetilde{J}^p(\boldsymbol{\alpha};\boldsymbol{m})$ in the form of a cost functional featuring an additive randomization of the control is new.

In the presence of the randomization, the cost functional might become very large as p tends to ∞, even for a control $\boldsymbol{\alpha}$ of finite energy. This prompts us to renormalize $\widetilde{J}^p$. As a result of the adaptability constraint, we indeed have

$$\mathbb{E}\int_0^T \alpha_t \cdot \dot{W}^p_t \mathrm{d}t = \mathbb{E}\sum_{\ell=0}^{p-1}\int_{\ell T/p}^{(\ell+1)T/p} \alpha_t \dot{W}^p_t \mathrm{d}t = \sum_{\ell=0}^{p-1}\mathbb{E}\Big[\alpha_{\ell T/p}\cdot\big(W_{(\ell+1)T/p} - W_{\ell T/p}\big)\Big] = 0,$$

with the last line following from the fact that $\alpha_{\ell T/p}$ is $\mathcal{F}^{\boldsymbol{W}}_{\ell T/p}$-measurable. Also, since $\boldsymbol{W}^p$ is piecewise linear, we have

$$\mathbb{E}\int_0^T |\dot{W}^p_t|^2 \mathrm{d}t = \sum_{\ell=0}^{p-1}\int_{\ell T/p}^{(\ell+1)T/p} \mathbb{E}\Big[\big|\frac{p}{T}\big(W_{(\ell+1)T/p} - W_{\ell T/p}\big)\big|^2\Big]\mathrm{d}t = d\,p\frac{T}{p}\frac{p^2}{T^2}\frac{T}{p} = d\,p.$$

So, we must subtract to the cost a diverging term to recover the original cost functional. To make it clear, the effective cost must be

$$\begin{aligned} J^p(\boldsymbol{\alpha};\boldsymbol{m}) &:= \widetilde{J}^p(\boldsymbol{\alpha};\boldsymbol{m}) - \tfrac{1}{2}\varepsilon^2 d\,p \\ &= \frac{1}{2}\mathbb{E}\bigg[\big|R^\dagger X_T + g(m_T)\big|^2 + \int_0^T \Big\{\big|Q^\dagger X_{\tau_p(t)} + f(m_{\tau_p(t)})\big|^2 + |\alpha_{\tau_p(t)}|^2\Big\}\mathrm{d}t\bigg]. \end{aligned} \tag{2.2.30}$$

Noticeably, we recover (up to the time discretization) the same cost functional J as in (2.1.2), which explains why we have removed the tilde over the symbol J^p. Anyway, $J^p(\cdot;\boldsymbol{m})$ and $\widetilde{J}^p(\cdot;\boldsymbol{m})$ have the same minimizers.

Before we provide the form of the corresponding fictitious play, we need to clarify the notion of tilted measure. We assume that we are given a process $\boldsymbol{h} = (h_t)_{0\leqslant t\leqslant T}$ that is piecewise constant

and $\mathbb{F}^{p,\boldsymbol{W}}$-adapted, with $\mathbb{F}^{p,\boldsymbol{W}} := \mathbb{F}^{\boldsymbol{W}^p}$: For the same p as before, the process $\boldsymbol{h}$ is constant on each interval $([\ell T/p, (\ell+1)T/p))_{\ell=0,\cdots,p-1}$ and each $h_{\ell T/p}$ is $\mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}$-measurable, with $\mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p} = \sigma(W^p_s, s \leqslant \ell T/p)$. Then, as before, we can let $\boldsymbol{W}^{\boldsymbol{h}/\varepsilon}$:

$$W^{\boldsymbol{h}/\varepsilon}_t = W_t + \frac{1}{\varepsilon}\int_0^t h_s \mathrm{d}s.$$

We compute the p-piecewise linear interpolation $\boldsymbol{W}^{p,\boldsymbol{h}/\varepsilon}$ of $\boldsymbol{W}^{\boldsymbol{h}/\varepsilon}$:

$$\begin{aligned}
W^{p,\boldsymbol{h}/\varepsilon}_t &= W^{p,\boldsymbol{h}/\varepsilon}_{\tau_p(t)} + \frac{p(t-\tau_p(t))}{T}\big(W^{\boldsymbol{h}/\varepsilon}_{\tau_p(t)+T/p} - W^{\boldsymbol{h}/\varepsilon}_{\tau_p(t)}\big) \\
&= W^{p,\boldsymbol{h}/\varepsilon}_{\tau_p(t)} + \frac{p(t-\tau_p(t))}{T}\big(W_{\tau_p(t)+T/p} - W_{\tau_p(t)}\big) + \frac{1}{\varepsilon}\frac{p(t-\tau_p(t))}{T}\frac{T}{p}h_{\tau_p(t)} \\
&= W^{p,\boldsymbol{h}/\varepsilon}_{\tau_p(t)} + \frac{p(t-\tau_p(t))}{T}\big(W_{\tau_p(t)+T/p} - W_{\tau_p(t)}\big) + \frac{1}{\varepsilon}\big(t-\tau_p(t)\big)h_{\tau_p(t)} \\
&= W^{p,\boldsymbol{h}/\varepsilon}_{\tau_p(t)} + \int_{\tau_p(t)}^t \mathrm{d}W^p_s + \frac{1}{\varepsilon}\int_{\tau_p(t)}^t h_s \mathrm{d}s,
\end{aligned}$$

from which we deduce that

$$W^{p,\boldsymbol{h}/\varepsilon}_t = W^p_t + \frac{1}{\varepsilon}\int_0^t h_s \mathrm{d}s, \quad t \in [0,T]. \tag{2.2.31}$$

For sure, under the probability $\mathcal{E}(\boldsymbol{h}/\varepsilon)\cdot\mathbb{P}$, the process $\boldsymbol{W}^{p,\boldsymbol{h}/\varepsilon}$ is the piecewise linear interpolation of $\boldsymbol{W}^{\boldsymbol{h}/\varepsilon}$ and the latter is a Brownian motion. Also,

$$\dot{W}^{p,\boldsymbol{h}/\varepsilon}_t = \dot{W}^p_t + \frac{1}{\varepsilon}h_t,$$

which permits to say that the trajectories controlled by $(\alpha_t + \varepsilon\dot{W}^p_t)_{0\leqslant t\leqslant T}$ satisfy

$$\mathrm{d}X_t = \big(\alpha_t - h_t\big)\mathrm{d}t + \sigma \mathrm{d}B_t + \varepsilon \mathrm{d}W^{p,\boldsymbol{h}/\varepsilon}_t,$$

which coincides with (2.1.1). Importantly, since $\boldsymbol{h}$ is piecewise constant, $\mathcal{E}(\boldsymbol{h}/\varepsilon)$ can be expressed in terms of the sole $\boldsymbol{h}$ and $\boldsymbol{W}^p$.

Remark 2.2.12. *We feel useful to comment more on our choice to work with the linear interpolation $(W^p_t)_{0\leqslant t\leqslant T}$ of $(W_t)_{0\leqslant t\leqslant T}$. While it may look somewhat arbitrary, this choice is in fact dictated by the structure of the model and the form of the final result that is provided next. Indeed, the main result of this subsection, see Theorem 2.2.17 below, yields a bound for the weak error (in a convenient sense) between the solution of a time-discrete version of the mean field game (with a common noise) and the corresponding time-discrete variant of the fictitious play, with the time mesh being given by $(kT/p)_{k=0,\cdots,p}$.*

Obviously, working with a discrete-time version of the game makes perfect sense from a practical point of view. Here, it is also especially adapted to our objective. As we already explained, we want to regard the instantaneous value of the control at time t as being corrupted, at least formally, by

the time derivative of the common noise. Because the latter derivative does not exist as a true function, our strategy is to work instead with finite differences of the common noise. This is one first reason for working with $(W_t^p)_{0\leqslant t\leqslant T}$, because the linear interpolation is completely characterized by the finite differences of $(W_t)_{0\leqslant t\leqslant T}$ at times 0, T/p, $2T/p$, $\cdots$, T. Another related reason is that, for $\boldsymbol{h}$ a piecewise constant process that is $\mathbb{F}^{p,\boldsymbol{W}}$-adapted, the relationship (2.2.31) *holds true because $\boldsymbol{W}^p$ and $\boldsymbol{W}^{p,\boldsymbol{h}}$ are piecewise linear. This is another strong case for working with the piecewise linear interpolation because* (2.2.31) *makes it possible to apply Girsanov theorem directly, with the latter being the key ingredient of the whole analysis.*

In a nutshell, given the class of time-discrete games we address for approximating the original game (using in particular finite differences to perturb the controls), the linear interpolation is the most natural one to reconstruct time-continuous dynamics from time-discrete observations. As such, the linear interpolation does not directly impact the bound on the weak error that is stated below, because the weak error is defined a priori within the given class of approximating games, regardless of the choice of the interpolation. However, working with the linear interpolation makes the proof much easier.

2.2.2.2 Fictitious play

The analysis from the previous paragraph leads to the following new scheme, which should be regarded as the discrete-time analogue of the scheme presented in Subsection 2.2.1.1. Fix a real $\varpi \in (1, \sqrt{2}]$ and an integer $p \geqslant 1$ and, for the same initialization $\boldsymbol{m}^0 = (m_t^0 = \mathbb{E}(X_0))_{0\leqslant t\leqslant T}$ and $\boldsymbol{h}^0 = (h_t^0 = 0)_{0\leqslant t\leqslant T}$ as in the continuous-time setting, assume that we have defined two families of proxies $(\boldsymbol{m}^1, \cdots, \boldsymbol{m}^n)$ and $(\boldsymbol{h}^1, \cdots, \boldsymbol{h}^n)$ with the additional assumption that each process $\boldsymbol{h}^k$ is constant on each interval $[\ell T/p, (\ell+1)T/p)$, for $\ell \in \{0, \cdots, p-1\}$ and $k \in \{1, \cdots, n\}$. And we assume each $(m^k_{\ell T/p}, h^k_{\ell T/p})$ to be measurable with respect to $\mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}$ if $\ell \in \{0, \cdots, p\}$. Then, we solve for

$$\begin{aligned}\boldsymbol{\alpha}^{(p),n+1,\varpi} &= \mathrm{argmin}_{\boldsymbol{\alpha}}\left(\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{p,\varpi X_0}\Big(\varpi\boldsymbol{\alpha} + \varpi\varepsilon\dot{\boldsymbol{W}}^{p,\varpi\boldsymbol{h}^n/\varepsilon}; \overline{\boldsymbol{m}}^n; 0\Big)\Big] - \tfrac{1}{2} d\,\varpi^2\varepsilon^2 p\right)\\ &= \mathrm{argmin}_{\boldsymbol{\alpha}}\left(\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{p,\varpi X_0}\Big(\varpi\boldsymbol{\alpha}; \overline{\boldsymbol{m}}^n; \varpi\varepsilon\boldsymbol{W}^{p,\varpi\boldsymbol{h}^n/\varepsilon}\Big)\Big]\right),\end{aligned} \tag{2.2.32}$$

with the same measurability rules as those explained above and where $\mathcal{R}^{p,x_0}$ is the obvious discrete-time version of $\mathcal{R}^{x_0}$, namely

$$\mathcal{R}^{p,x_0}(\boldsymbol{\alpha}; \boldsymbol{n}; \varepsilon\boldsymbol{w}) = \frac{1}{2}\left[\big|Rx_T + g(n_T)\big|^2 + \sum_{\ell=0}^{p-1}\Big\{\big|Qx_{\ell T/p} + f(n_{\ell T/p})\big|^2 + \big|\alpha_{\ell T/p}\big|^2\Big\}\right], \tag{2.2.33}$$

where $x_{(\ell+1)T/p} = x_{\ell T/p} + (T/p)\alpha_{\ell T/p} + \sigma(B_{(\ell+1)T/p} - B_{\ell T/p}) + \varepsilon(w_{(\ell+1)T/p} - w_{\ell T/p})$. As before, $\overline{m}_t^n$ in (2.2.32) is $\overline{m}_t^n = \frac{\varpi(1-\varpi^{-1})}{1-\varpi^{-n}}\sum_{k=1}^n \varpi^{-k} m_t^k$.

The first step here is to solve the optimization problem (2.2.32). Very similar to Lemma 2.2.1, we can provide an explicit form for the optimal feedback through the solution of the following

time-discrete Riccati equation:

$$\begin{aligned}
&\frac{p}{T}\Big(\eta^{(p)}_{(\ell+1)T/p}-\eta^{(p)}_{\ell T/p}\Big)-\eta^{(p)}_{(\ell+1)T/p}\eta^{(p)}_{\ell T/p}+Q^\dagger Q\Big(I_d-\frac{T}{p}\eta_{\ell T/p}\Big)=0,\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \ell\in\{0,\cdots,p-2\}\ ;\\
&\frac{p}{T}\Big(\eta^{(p)}_{T}-\eta^{(p)}_{(p-1)T/p}\Big)-\eta^{(p)}_{T}\eta^{(p)}_{(p-1)T/p}=0\ ;\quad \eta^{(p)}_T=R^\dagger R.
\end{aligned}\tag{2.2.34}$$

The analysis of the latter (see the earlier reference [51]) goes through the auxiliary Riccati equation

$$\frac{p}{T}\Big(P^{(p)}_{(\ell+1)T/p}-P^{(p)}_{\ell T/p}\Big)-P^{(p)}_{(\ell+1)T/p}\Big(I_d+\frac{T}{p}P^{(p)}_{(\ell+1)T/p}\Big)^{-1}P^{(p)}_{(\ell+1)T/p}+Q^\dagger Q=0,\tag{2.2.35}$$

for $\ell\in\{0,\cdots,p-1\}$, with $P^{(p)}_T=R^\dagger R$ as boundary condition. The Riccati equation (2.2.35) can be solved inductively: the solution is symmetric and non-negative[9] , which guarantees that the inverse right above is well-defined. Then,

$$\eta^{(p)}_{\ell T/p}=\Big(I_d+\frac{T}{p}P^{(p)}_{(\ell+1)T/p}\Big)^{-1}P^{(p)}_{(\ell+1)T/p},\quad \ell\in\{0,\cdots,p-1\},\tag{2.2.36}$$

solves (2.2.34). Notice that the above left-hand side is symmetric and non-negative because the two matrices in the right-hand side commute.

Lemma 2.2.13. *Under the above assumptions, the minimization problem* (2.2.32) *has a unique solution* $\boldsymbol{\alpha}^{(p),n+1,\varpi}$*, which writes*

$$\alpha^{(p),n+1,\varpi}_t=-\Big(\eta^{(p)}_{\tau_p(t)}X^{(p),n+1,\varpi}_{\tau_p(t)}+\widetilde{h}^{n+1}_{\tau_p(t)}\Big),\quad t\in[0,T],\tag{2.2.37}$$

where $\widetilde{\boldsymbol{h}}^{n+1}$ *solves the backward SDE:*

$$\begin{aligned}
\mathrm{d}\widetilde{h}^{n+1}_t=&\Big\{-\tfrac{1}{\varpi}\mathbb{E}\Big[Q^\dagger f\big(\overline{m}^n_{\tau_p(t)+T/p}\big)\,|\,\mathcal{F}^{p,\boldsymbol{W}}_{\tau_p(t)}\Big]\mathbf{1}_{\{t\leqslant(p-1)T/p\}}\\
&\qquad+\Big(\eta^{(p)}_{\tau_{p(t)}+T/p}+\frac{T}{p}Q^\dagger Q\mathbf{1}_{\{t\leqslant(p-1)T/p\}}\Big)\widetilde{h}^{n+1}_{\tau_p(t)}\Big\}\mathrm{d}t\\
&\quad+\varpi k^{n+1}_t h^n_t\mathrm{d}t+\varepsilon k^{n+1}_t\mathrm{d}W_t,\quad t\in[0,T),\\
\widetilde{h}^{n+1}_T=&\tfrac{1}{\varpi}R^\dagger g\big(\overline{m}^n_T\big).
\end{aligned}\tag{2.2.38}$$

Accordingly, the optimal path $\boldsymbol{X}^{(p),n+1,\varpi}$ *solves (up to a rescaling factor* ϖ*) the forward SDE:*

$$\mathrm{d}X^{(p),n+1,\varpi}_t=\alpha^{(p),n+1,\varpi}_t\mathrm{d}t+\tfrac{1}{\varpi}\sigma\mathrm{d}B_t+\varepsilon\mathrm{d}W^{p,\varpi\boldsymbol{h}^n/\varepsilon}_t,\quad t\in[0,T];\quad X^{n+1}_0=X_0.\tag{2.2.39}$$

[9]Symmetry is obvious. Non-negativity is a bit more demanding. Assuming that $P^{(p)}_{(\ell+1)T/p}$ is non-negative, non-negativity of $P^{(p)}_{\ell T/p}$ is proved as follows. For any vector $x\in\mathbb{R}^d$, we have

$$\begin{aligned}
(P^{(p)}_{\ell T/p}x)\cdot x&=(P^{(p)}_{(\ell+1)T/p}x)\cdot x-\tfrac{T}{p}(P^{(p)}_{(\ell+1)T/p}x)\cdot\Big[\big(I_d+\tfrac{T}{p}P^{(p)}_{(\ell+1)T/p}\big)^{-1}P^{(p)}_{(\ell+1)T/p}x\Big]+\tfrac{T}{p}|Qx|^2\\
&\geqslant(P^{(p)}_{(\ell+1)T/p}x)\cdot x-(P^{(p)}_{(\ell+1)T/p}x)\cdot x+(P^{(p)}_{(\ell+1)T/p}x)\cdot\Big[\big(I_d+\tfrac{T}{p}P^{(p)}_{(\ell+1)T/p}\big)^{-1}x\Big],
\end{aligned}$$

and the right-hand side is obviously non-negative.

Remark 2.2.14. *Although it is not indicated in the notations $\widetilde{\boldsymbol{h}}^n$, $\boldsymbol{h}^n$, $\boldsymbol{k}^n$ and $\overline{\boldsymbol{m}}^n$, it should be clear to the reader that the latter four processes depend on p, ϖ and ε.*

Remark 2.2.15. *The convergence of $(P^{(p)}_{\tau_p(t)})_{0\leqslant t\leqslant T}$ to $(\eta_t)_{0\leqslant t\leqslant T}$ in Lemma 2.2.1 is obvious.*

By the symmetric and non-negative structure of the matrices $(P^{(p)}_\ell)_{\ell=0,\cdots,p}$, we can easily get uniform bounds on the latter. In turn, we can regard the equation (2.2.35) *as an Euler scheme for a matricial differential equation driven by a Lipschitz vector field. The rate of convergence is linear in p. By* (2.2.36)*, we deduce that*

$$\sup_{0\leqslant t\leqslant T}|\eta^{(p)}_t-\eta_t|\leqslant\frac{C}{p}, \tag{2.2.40}$$

for a constant C independent of p, but possibly depending on the dimension d. Typically, the bound on $(P^{(p)}_\ell)_{\ell=0,\cdots,p}$ should depend on the norms of $Q^\dagger Q$ and $R^\dagger R$ and is thus expected to depend on d in a polynomial way. In turn, stability arguments for finite difference equations say that those bounds should propagate to the constant C in the above estimate for the distance between $(\eta^{(p)}_t)_{0\leqslant t\leqslant T}$ and $(\eta_t)_{0\leqslant t\leqslant T}$. However, this argument may not be sharp: when Q and R are the identity matrix, the Riccati equation (2.2.35) *reduces to a scalar equation and the constant C should just scale like $\sqrt{d}$.*

Remark 2.2.16. *The backward SDE* (2.2.38) *is in fact a mere discrete-time equation. Indeed,*

$$\begin{aligned}\widetilde{h}^{n+1}_{\ell T/p}&=\widetilde{h}^{n+1}_{(\ell+1)T/p}+\tfrac{T}{\varpi p}Q^\dagger f(\overline{m}^n_{(\ell+1)T/p})\mathbf{1}_{\{\ell\leqslant p-2\}}\\&\quad-\tfrac{T}{p}\Big(\eta^{(p)}_{(\ell+1)T/p}+\tfrac{T}{p}Q^\dagger Q\mathbf{1}_{\{\ell\leqslant p-2\}}\Big)\widetilde{h}^{n+1}_{\ell T/p}-\varepsilon\int_{\ell T/p}^{(\ell+1)T/p}k^{n+1}_s\mathrm{d}W^{\varpi\boldsymbol{h}^n/\varepsilon}_s,\end{aligned} \tag{2.2.41}$$

for $\ell\in\{0,\cdots,p-1\}$. Taking conditional expectation given $\mathcal{F}^{\boldsymbol{W}}_{\ell T/p}$, we can solve for

$$\Big(I_d+\tfrac{T}{p}\eta^{(p)}_{(\ell+1)T/p}+\tfrac{T^2}{p^2}Q^\dagger Q\mathbf{1}_{\{\ell\leqslant p-2\}}\Big)\widetilde{h}^{n+1}_{\ell T/p}.$$

It is easy to deduce $\widetilde{h}^{n+1}_{\ell T/p}$. We get

$$\begin{aligned}\widetilde{h}^{n+1}_{\ell T/p}&=\Big(I_d+\tfrac{T}{p}\eta^{(p)}_{(\ell+1)/p}+\tfrac{T^2}{p^2}Q^\dagger Q\mathbf{1}_{\{\ell\leqslant p-2\}}\Big)^{-1}\\&\quad\times\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\widetilde{h}^{n+1}_{(\ell+1)T/p}+\tfrac{T}{\varpi p}Q^\dagger f\big(\overline{m}^n_{(\ell+1)T/p}\big)\mathbf{1}_{\{\ell\leqslant p-2\}}\,|\,\mathcal{F}^{\boldsymbol{W}}_{\ell T/p}\Big],\end{aligned} \tag{2.2.42}$$

which proves, by induction, that $\widetilde{h}^{n+1}_{\ell T/p}$ is $\mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}$-measurable. It suffices to observe that, for Z an $\mathcal{F}^{p,\boldsymbol{W}}_{(\ell+1)T/p}$-measurable random variable, the conditional expectation of Z given $\mathcal{F}^{\boldsymbol{W}}_{\ell T/p}$ is in fact $\mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}$-measurable.

Taking for granted Lemma 2.2.13 (the proof is given at the end of the subsection), we put

$$h^{n+1}_t:=\widetilde{h}^{n+1}_{\tau_p(t)}, \tag{2.2.43}$$

which satisfies the required measurability constraints thanks to Remark 2.2.16. Then, we let

$$m^{n+1}_t:=\mathbb{E}\big[X^{(p),n+1,\varpi}_t\,|\,\sigma(\boldsymbol{W})\big]=\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\big[X^{(p),n+1,\varpi}_t\,|\,\sigma(\boldsymbol{W})\big],\quad t\in[0,T], \tag{2.2.44}$$

together with

$$\overline{m}_t^{n+1} = \tfrac{\varpi(1-\varpi^{-1})}{1-\varpi^{-(n+1)}} \sum_{k=1}^{n+1} \varpi^{-k} m_t^k = \tfrac{\varpi^{-n}(1-\varpi^{-1})}{1-\varpi^{-(n+1)}} m_t^{n+1} + \Big(1 - \tfrac{\varpi^{-n}(1-\varpi^{-1})}{1-\varpi^{-(n+1)}}\Big)\overline{m}_t^n. \tag{2.2.45}$$

Importantly, notice that

$$m_{\ell T/p}^{n+1} = \mathbb{E}\big[X_{\ell T/p}^{(p),n+1,\varpi} \,|\, \sigma(\boldsymbol{W}^p)\big] = \mathbb{E}\big[X_{\ell T/p}^{(p),n+1,\varpi} \,|\, \mathcal{F}_{\ell T/p}^{p,\boldsymbol{W}}\big], \quad \ell \in \{0,\cdots,p\},$$

which proves in particular that the left-hand side is $\mathcal{F}_{\ell T/p}^{p,\boldsymbol{W}}$-measurable.

The analogue of Theorem 2.2.4 becomes:

Theorem 2.2.17. *There exists a threshold $c > 0$, depending on d, T, the norms $\|f\|_{1,\infty}$ and $\|g\|_{1,\infty}$ and the norms $|Q|$ and $|R|$ of the matrices Q and R, such that, for $p\varepsilon^2 \geqslant c$, the scheme (2.2.43)–(2.2.45) converges to $(\boldsymbol{m}^{(p)}, \widetilde{\boldsymbol{h}}^{(p)}/\varpi, \boldsymbol{k}^{(p)}/\varpi)$, where $(\boldsymbol{m}^{(p)}, \widetilde{\boldsymbol{h}}^{(p)}, \boldsymbol{k}^{(p)})$ is the unique solution of the decoupled discrete-time FBSDE system:*

$$\begin{aligned}
m_{(\ell+1)T/p}^{(p)} &= m_{\ell T/p}^{(p)} - \tfrac{T}{p}\eta_{\ell T/p}^{(p)} m_{\ell T/p}^{(p)} + \varepsilon\big(W_{(\ell+1)T/p} - W_{\ell T/p}\big),\\
\widetilde{h}_{\ell T/p}^{(p)} &= \widetilde{h}_{(\ell+1)T/p}^{(p)} + \tfrac{T}{p} Q^\dagger f\big(m_{(\ell+1)T/p}^{(p)}\big)\mathbf{1}_{\{\ell\leqslant p-2\}}\\
&\quad - \tfrac{T}{p}\Big(\eta_{(\ell+1)T/p}^{(p)} + \tfrac{T}{p} Q^\dagger Q \mathbf{1}_{\{\ell\leqslant p-2\}}\Big)\widetilde{h}_{\ell T/p}^{(p)} - \Big(\int_{\ell T/p}^{(\ell+1)T/p} k_s^{(p)}\mathrm{d}s\Big)\widetilde{h}_{\ell T/p}^{(p)}\\
&\quad - \varepsilon\int_{\ell T/p}^{(\ell+1)T/p} k_s^{(p)}\mathrm{d}W_s, \quad \ell \in \{0,\cdots,p-1\},\\
m_0^{(p)} &= \mathbb{E}(X_0), \quad \widetilde{h}_T^{(p)} = R^\dagger g\big(m_T^{(p)}\big),
\end{aligned} \tag{2.2.46}$$

with an explicit bound on the rate of convergence, namely

$$\mathrm{essup}_{\omega\in\Omega}\Big[\sup_{\ell=0,\cdots,p}\Big(|m_{\ell T/p}^{(p)} - \overline{m}_{\ell T/p}^n|^2 + |\varpi^{-1}\widetilde{h}_{\ell T/p}^{(p)} - h_{\ell T/p}^n|^2\Big)\Big] \leqslant \varpi^{-2n}\exp\big(C\varepsilon^{-2}\big), \tag{2.2.47}$$

for a constant C that also depends on d, T, $\|f\|_{1,\infty}$, $\|g\|_{1,\infty}$, $|Q|$ and $|R|$.

Moreover, if we extend $\boldsymbol{m}^{(p)}$ by continuous interpolation to the entire $[0,T]$ and if we call $\boldsymbol{h}^{(p)}$ the piecewise constant extension of $\widetilde{\boldsymbol{h}}^{(p)}$ to the entire $[0,T]$, then, up to a modification of the constant C, the weak error of the scheme for the Fortet-Mourier distance satisfies

$$\sup_F\Big|\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[F\big(\overline{\boldsymbol{m}}^n, \boldsymbol{h}^n\big)\Big] - \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[F\big(\boldsymbol{m}^{(p)}, \varpi^{-1}\boldsymbol{h}^{(p)}\big)\Big]\Big| \leqslant \varpi^{-n}\exp\big(C\varepsilon^{-2}\big), \tag{2.2.48}$$

the supremum being taken over all the functions F on $\mathcal{C}([0,T];\mathbb{R}^d\times\mathbb{R}^d)$ that are bounded by 1 and 1-Lipschitz continuous.

The following comments are in order:

- In (2.2.46), $\boldsymbol{m}^{(p)}$ and $\widetilde{\boldsymbol{h}}^{(p)}$ are implicitly understood as $\boldsymbol{m}^{(p)} = (m_{\ell T/p}^{(p)})_{\ell=0,\cdots,p}$ and $\widetilde{\boldsymbol{h}}^{(p)} = (h_{\ell T/p}^{(p)})_{\ell=0,\cdots,p}$, with $m_{\ell T/p}^{(p)}$ and $h_{\ell T/p}^{(p)}$ being $\mathbb{R}^d$-valued and $\mathcal{F}_{\ell T/p}^{p,\boldsymbol{W}}$-measurable for each $\ell \in \{0,\cdots,p\}$.

- The process $\boldsymbol{k}^{(p)}$ is understood as $(k_s^{(p)})_{0\leqslant s\leqslant T}$. It is $\mathbb{R}^{d\times d}$-valued and $\mathbb{F}^{\boldsymbol{W}}$-progressively measurable.

Existence and uniqueness of a solution to (2.2.46) is in fact ensured by the following proposition, whose proof is deferred to §2.2.2.5. The statement also explains the need for a threshold on the product $p\varepsilon^2$ in the two bounds (2.2.47) and (2.2.48).

Proposition 2.2.18. *There exists a constant c, depending on d, T, $\|f\|_{1,\infty}$, $\|g\|_{1,\infty}$, $|Q|$ and $|R|$, such that for $p\varepsilon^2 \geqslant c$, the backward equation in (2.2.46) admits a unique solution $(\widetilde{h}^{(p)}_{\ell T/p})_{\ell=0,\cdots,p}$, with $\widetilde{h}^{(p)}_{\ell T/p} \in L^2(\Omega, \mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}, \mathbb{P}; \mathbb{R}^d)$ for each $\ell \in \{0,\cdots,p\}$.*

The process $(k_t^{(p)})_{0\leqslant t\leqslant T}$ is $\mathbb{F}^{\boldsymbol{W}}$-progressively measurable and is square integrable over $[0,T]\times\Omega$ (Ω being equipped with $\mathbb{P}$).

We also feel useful to clarify the meaning of the continuous interpolation of $\boldsymbol{m}^{(p)}$ in (2.2.48). In fact, the definition is similar to (2.2.29):

$$m_t^{(p)} := m^{(p)}_{\tau_p(t)} + \frac{p(t-\tau_p(t))}{T}\big(m^{(p)}_{\tau_p(t)+T/p} - m^{(p)}_{\tau_p(t)}\big), \quad t\in[0,T]. \tag{2.2.49}$$

Moreover, the definition of the piecewise constant extension of $\widetilde{\boldsymbol{h}}^{(p)}$ is similar to (2.2.43):

$$h_t^{(p)} := \widetilde{h}^{(p)}_{\tau_p(t)}, \quad t\in[0,T]. \tag{2.2.50}$$

It is worth noting that $(\boldsymbol{m}^{(p)}, \boldsymbol{h}^{(p)})$ (hence extended to the entire $[0,T]$ as above) is the unique solution of the following two-step fixed point problem, which is nothing but the discrete-time version of the MFG with common noise addressed in Theorem 2.2.4:

(i) The solution of

$$\mathrm{argmin}_{\boldsymbol{\alpha}}\left(\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\Big(\boldsymbol{\alpha}+\varepsilon\dot{\boldsymbol{W}}^{p,\boldsymbol{h}^{(p)}/\varepsilon};\boldsymbol{m}^{(p)};0\Big)\Big] - \tfrac{1}{2}d\,\varepsilon^2 p\right), \tag{2.2.51}$$

which is also

$$\mathrm{argmin}_{\boldsymbol{\alpha}}\left(\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\Big(\boldsymbol{\alpha};\boldsymbol{m}^{(p)};\varepsilon\boldsymbol{W}^{p,\boldsymbol{h}^{(p)}/\varepsilon}\Big)\Big]\right) \tag{2.2.52}$$

is given by

$$\alpha_t^{(p),\star} = -\Big(\eta^{(p)}_{\tau_p(t)}X^{(p),\star}_{\tau_p(t)} + h_t^{(p)}\Big), \quad t\in[0,T], \tag{2.2.53}$$

where

$$\begin{aligned}\mathrm{d}X_t^{(p),\star} &= -\eta^{(p)}_{\tau_p(t)}X^{(p),\star}_{\tau_p(t)}\mathrm{d}t + \sigma\mathrm{d}B_t + \varepsilon\mathrm{d}W_t^p \\ &= \alpha_t^{(p),\star}\mathrm{d}t + \sigma\mathrm{d}B_t + \varepsilon\mathrm{d}W_t^{p,\boldsymbol{h}^{(p)}/\varepsilon}, \quad t\in[0,T],\end{aligned} \tag{2.2.54}$$

with X_0 as initial condition.

(ii) The process obtained by taking the conditional expectation of $\boldsymbol{X}^{(p),\star}$ given $\sigma(\boldsymbol{W}^p)$, namely $(\mathbb{E}[X_t^{(p),\star} \,|\, \sigma(\boldsymbol{W}^p)])_{0\leqslant t\leqslant T}$, solves the forward equation in (2.2.46), i.e.,

$$\mathbb{E}\big[X^{(p),\star}_{\ell T/p} \,|\, \sigma(\boldsymbol{W}^p)\big] = m^{(p)}_{\ell T/p}, \quad \ell = 0, \cdots, p. \tag{2.2.55}$$

Indeed, by a straightforward adaptation of Lemma 2.2.13, the system (2.2.46) can be shown to characterize the solution to the fixed point problem associated with the two items (i) and (ii) right above. This proves in particular that the discrete-time MFG constructed on the top of the cost functional (2.2.51) has a unique solution when $p\varepsilon^2 \geqslant c$, which is given by Proposition 2.2.18.

Remark 2.2.19. *As made in clear in the forthcoming proof of Proposition 2.2.18, there is another conceivable extension for $\widetilde{\boldsymbol{h}}^{(p)}$ (in addition to the extension defined in* (2.2.50)*). Indeed, very similar to* (2.2.38)*, given $\boldsymbol{h}^{(p)}$ (from* (2.2.50)*), one can define the extension of $\widetilde{\boldsymbol{h}}^{(p)}$ to the entire $[0,T]$ as the solution of the BSDE:*

$$\begin{aligned}
\mathrm{d}\widetilde{h}^{(p)}_t &= \Big\{-\mathbb{E}\Big[Q^\dagger f\big(m^{(p)}_{\tau_p(t)+T/p}\big) \,|\, \mathcal{F}^{p,\boldsymbol{W}}_{\tau_p(t)}\Big]\mathbf{1}_{\{t\leqslant(p-1)T/p\}} \\
&\qquad\qquad + \Big(\eta^{(p)}_{\tau_{p(t)}+T/p} + \tfrac{T}{p}Q^\dagger Q\mathbf{1}_{\{t\leqslant(p-1)T/p\}}\Big)h^{(p)}_t\Big\}\mathrm{d}t \\
&\quad + k^{(p)}_t h^{(p)}_t \mathrm{d}t + \varepsilon k^{(p)}_t \mathrm{d}W_t, \quad t\in[0,T), \\
\widetilde{h}^{(p)}_T &= R^\dagger g\big(m^{(p)}_T\big).
\end{aligned} \tag{2.2.56}$$

Consistency of this extension is explained in the proof of Proposition 2.2.18. The time-continuous process $\widetilde{\boldsymbol{h}}^p$ defined as the solution of (2.2.56) *coincides at times $(\ell T/p)_{\ell=0,\cdots,p}$ with the solution $(\widetilde{h}^{(p)}_{\ell T/p})_{\ell=0,\cdots,p}$ of* (2.2.46)*. Moreover, the process $\boldsymbol{k}^{(p)}$ in* (2.2.56) *coincides with $\boldsymbol{k}^{(p)}$ in* (2.2.46)*.*

Remark 2.2.20. *Similar to the proof of Theorem 2.2.4, the proof of Theorem 2.2.17 also shows that*

$$d_{\mathrm{TV}}\big(\mathbb{P}^{\varpi \boldsymbol{h}^n/\varepsilon}, \mathbb{P}^{\boldsymbol{h}^{(p)}/\varepsilon}\big) \leqslant \varpi^{-n}\exp\big(C\varepsilon^{-2}\big).$$

Remark 2.2.21. *Following Remark 2.2.9 about the scope of Theorem 2.2.4 to the higher dimensional setting, we could think of tracking the dependence of the constant C (in* (2.2.11) *and* (2.2.12)*), which is here independent of p, upon the dimension d. Although we prefer not to address this question in full detail, we insist on the fact that the conclusion of Remark 2.2.9 also holds true here, meaning that the constant C cannot be better than $O(\exp(O(\sqrt{d})))$ in simple cases when Q and R are the identity matrices and f and g are diagonal. Intuitively, we cannot expect a constant that would be, asymptotically in d and uniformly in p, better than the constant appearing in the statement of Theorem 2.2.4, as otherwise we could take the limit $p\to\infty$ in the statement of Theorem 2.2.17 and then get a better estimate in Theorem 2.2.4.*

2.2.2.3 Analysis of the convergence

Conditioning on $\boldsymbol{W}$ in (2.2.39) and recalling the two notations (2.2.43) and (2.2.44), we obtain

$$\mathrm{d}m^{n+1}_t = -\Big(\eta^{(p)}_{\tau_p(t)} m^{n+1}_{\tau_p(t)} + h^{n+1}_t\Big)\mathrm{d}t + \varepsilon \mathrm{d}W^{p,\varpi\boldsymbol{h}^n/\varepsilon}_t, \quad t\in[0,T]; \quad m^{n+1}_0 = \mathbb{E}(X_0).$$

Thanks to (2.2.45), the identity (2.2.9) becomes

$$\mathrm{d}\overline{m}_t^{n+1} = -\big(\eta^{(p)}_{\tau_p(t)}\overline{m}^{n+1}_{\tau_p(t)} + \tfrac{(1-\varpi^{-1})\varpi^{-n}}{1-\varpi^{-(n+1)}}h_t^{n+1}\big)\mathrm{d}t + \varepsilon\mathrm{d}W_t^p, \quad t\in[0,T]. \tag{2.2.57}$$

As in the analysis of (2.2.10), the second term in the middle disappears asymptotically at a geometric rate. This follows from the next result, which is an improved version of Lemma 2.2.10 and which is also used next in place of Lemma 2.2.8 (the latter relying on a Markovian structure that is not satisfied here):

Lemma 2.2.22. *With $(\overline{\boldsymbol{m}}^n, \widetilde{\boldsymbol{h}}^n)$ being defined as in* (2.2.38), (2.2.44) *and* (2.2.45) *and $(\boldsymbol{m}^{(p)}, \widetilde{\boldsymbol{h}}^{(p)})$ as in* (2.2.46) *and* (2.2.56)*, there exists a constant C_1, only depending on d, T, $\|f\|_{1,\infty}$, $\|g\|_{1,\infty}$, $|Q|$ and $|R|$, such that, $\mathbb{P}$ almost surely,*

$$|h_t^n| \leqslant C_1, \quad t\in[0,T],\ n\geqslant 1\,; \quad |\widetilde{h}_t^{(p)}| \leqslant C_1, \quad t\in[0,T]. \tag{2.2.58}$$

Moreover, for the same constant C_1, there exists $\delta\in(0,1]$, depending on the same parameters as C_1, such that, with probability 1 under $\mathbb{P}$,

$$\forall t\in[0,T], \quad \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\bigg[\exp\bigg(\delta\varepsilon^2\int_t^T |k_s^{(p)}|^2\mathrm{d}s\bigg)\,|\,\mathcal{F}_t^{\boldsymbol{W}}\bigg] \leqslant C_1. \tag{2.2.59}$$

The proof of Lemma 2.2.22 is deferred to §2.2.2.5. Taking for granted the statement, we now complete the proof of Theorem 2.2.17.

Proof of Theorem 2.2.17. Taking for granted Proposition 2.2.18, we assume that the system (2.2.46) has a unique solution. Implicitly, this requires $p\varepsilon^2\geqslant c$, but this is indeed one of the assumption of Theorem 2.2.17.

First Step. For any integer $n\geqslant 1$, we thus compare $(\overline{\boldsymbol{m}}^{n+1}, \widetilde{\boldsymbol{h}}^{n+1})$ and $(\overline{\boldsymbol{m}}^{(p)}, \widetilde{\boldsymbol{h}}^{(p)}/\varpi)$, with the latter two ones being extended to the entire $[0,T]$ as explained in (2.2.49) and (2.2.56). To do so, we first notice that

$$\mathrm{d}m_t^{(p)} = -\eta^{(p)}_{\tau_p(t)}m^{(p)}_{\tau_p(t)}\mathrm{d}t + \varepsilon\mathrm{d}W_t^p, \quad t\in[0,T].$$

By (2.2.57) and (2.2.58),

$$\mathrm{d}\big(\overline{m}_t^{n+1} - m_t^{(p)}\big) = -\eta^{(p)}_{\tau_p(t)}\big(\overline{m}_t^{n+1} - m_t^{(p)}\big)\mathrm{d}t - O\big(\varpi^{-n}\big)\mathrm{d}t, \quad t\in[0,T],$$

where $|O(\varpi^{-n})|\leqslant C\varpi^{-n}$ for a constant C as in the statement of Theorem 2.2.17. We easily get

$$\sup_{0\leqslant t\leqslant T}|\overline{m}_t^{n+1} - m_t^{(p)}| \leqslant C\varpi^{-n}. \tag{2.2.60}$$

Second Step. We turn to the difference $\widetilde{\boldsymbol{h}}^{n+1} - \widetilde{\boldsymbol{h}}^{(p)}/\varpi$. We rewrite (2.2.38) as

$$\begin{aligned}\mathrm{d}\widetilde{h}_t^{n+1} = \Big\{&-\tfrac{1}{\varpi}\mathbb{E}\Big[Q^\dagger f\big(\overline{m}^n_{\tau_p(t)+T/p}\big)\,|\,\mathcal{F}^{p,\boldsymbol{W}}_{\tau_p(t)}\Big]\mathbf{1}_{\{t\leqslant(p-1)T/p\}}\\ &+\Big(\eta^{(p)}_{\tau_{p(t)}+T/p} + \tfrac{T}{p}Q^\dagger Q\mathbf{1}_{\{t\leqslant(p-1)T/p\}}\Big)h_t^{n+1}\Big\}\mathrm{d}t\\ &+\varpi k_t^{n+1}\big(h_t^n - \tfrac{1}{\varpi}h_t^{(p)}\big)\mathrm{d}t + \varepsilon k_t^{n+1}\mathrm{d}W_t^{\boldsymbol{h}^{(p)}/\varepsilon}, \quad t\in[0,T].\end{aligned}$$

Similarly, we rewrite (2.2.56) as

$$\begin{aligned}
\mathrm{d}\big(\tfrac{1}{\varpi}\widetilde{h}_t^{(p)}\big) &= \Big\{-\tfrac{1}{\varpi}\mathbb{E}\Big[Q^\dagger f\big(m^{(p)}_{\tau_p(t)+T/p}\big)\,|\,\mathcal{F}^{p,\boldsymbol{W}}_{\tau_p(t)}\Big]\mathbf{1}_{\{t\leqslant(p-1)T/p\}} \\
&\qquad + \Big(\eta^{(p)}_{\tau_{p(t)}+T/p} + \frac{T}{p}Q^\dagger Q\mathbf{1}_{\{t\leqslant(p-1)T/p\}}\Big)\big(\tfrac{1}{\varpi}h_t^{(p)}\big)\Big\}\mathrm{d}t \\
&\quad + \tfrac{\varepsilon}{\varpi}k_t^{(p)}\mathrm{d}W_t^{\varpi\boldsymbol{h}^{(p),\varpi/\varepsilon}}, \quad t\in[0,T), \\
\tfrac{1}{\varpi}\widetilde{h}_T^{(p)} &= \tfrac{1}{\varpi}R^\dagger g\big(m_T^{(p)}\big).
\end{aligned}$$

Forming and squaring the difference between the two equations and using (2.2.60) (together with Remark 2.2.15, which gives a bound for $\boldsymbol{\eta}^{(p)}$ independent of p), we deduce (very like as in (2.2.22) and (2.2.23))

$$\begin{aligned}
&\mathrm{d}\Big[\big|\widetilde{h}_t^{n+1} - \tfrac{1}{\varpi}\widetilde{h}_t^{(p)}\big|^2\Big] \\
&\geqslant -C\varpi^{-2n}\mathrm{d}t - C\big|h_t^{n+1} - \tfrac{1}{\varpi}h_t^{(p)}\big|^2\mathrm{d}t - C\big|\widetilde{h}_t^{n+1} - \tfrac{1}{\varpi}\widetilde{h}_t^{(p)}\big|^2\mathrm{d}t \\
&\quad - C|k_t^{(p)}|\,\big|\widetilde{h}_t^{n+1} - \tfrac{1}{\varpi}\widetilde{h}_t^{(p)}\big|\,\big|h_t^{n} - \tfrac{1}{\varpi}h_t^{(p)}\big|\mathrm{d}t - C\big|\widetilde{h}_t^{n+1} - \tfrac{1}{\varpi}\widetilde{h}_t^{(p)}\big|\big|k_t^{n+1} - \tfrac{1}{\varpi}k_t^{(p)}\big|\mathrm{d}t \\
&\quad + \varepsilon^2\big|k_t^{n+1} - \tfrac{1}{\varpi}k_t^{(p)}\big|^2\mathrm{d}t + \varepsilon\big(h_t^{n+1} - \tfrac{1}{\varpi}\widetilde{h}_t^{(p)}\big)\cdot\big[\big(k_t^{n+1} - \tfrac{1}{\varpi}k_t^{(p)}\big)\mathrm{d}W_t^{\boldsymbol{h}^{(p)/\varepsilon}}\big],
\end{aligned}$$

with the constant C being allowed to vary from line to line as long as it depends on the same parameters as those indicated in the statement of Theorem 2.2.17. With δ as in (2.2.59), this leads to

$$\begin{aligned}
&\mathrm{d}\Big[\big|\widetilde{h}_t^{n+1} - \tfrac{1}{\varpi}\widetilde{h}_t^{(p)}\big|^2\Big] \\
&\geqslant -C\varpi^{-2n}\mathrm{d}t - C\big|h_t^{n+1} - \tfrac{1}{\varpi}h_t^{(p)}\big|^2\mathrm{d}t - C\big|\widetilde{h}_t^{n+1} - \tfrac{1}{\varpi}\widetilde{h}_t^{(p)}\big|^2\mathrm{d}t \\
&\quad - \tfrac{C}{\delta\varepsilon^2}\big|h_t^{n} - \tfrac{1}{\varpi}h_t^{(p)}\big|^2\mathrm{d}t - \big(\delta\varepsilon^2|k_t^{(p)}|^2 + \tfrac{C}{\varepsilon^2}\big)\big|\widetilde{h}_t^{n+1} - \tfrac{1}{\varpi}\widetilde{h}_t^{(p)}\big|^2\mathrm{d}t \\
&\quad + \tfrac{\varepsilon^2}{2}\big|k_t^{n+1} - \tfrac{1}{\varpi}k_t^{(p)}\big|^2\mathrm{d}t + \varepsilon\big(h_t^{n+1} - \tfrac{1}{\varpi}\widetilde{h}_t^{(p)}\big)\cdot\big[\big(k_t^{n+1} - \tfrac{1}{\varpi}k_t^{(p)}\big)\mathrm{d}W_t^{\boldsymbol{h}^{(p)/\varepsilon}}\big].
\end{aligned}$$

Using the fact that, by convention, ε is taken in $(0,1)$, we can get rid of the last term on the first line (which is dominated by the last term on the second line). Then, letting

$$E_t^{(p)} := \exp\bigg(\int_0^t \big(\delta\varepsilon^2|k_s^{(p)}|^2 + \tfrac{C}{\delta\varepsilon^2}\big)\mathrm{d}s\bigg), \quad t\in[0,T], \tag{2.2.61}$$

and assuming without any loss of generality that $\delta\in(0,1)$, we obtain

$$\begin{aligned}
&\mathrm{d}\Big[E_t^{(p)}\big|\widetilde{h}_t^{n+1} - \tfrac{1}{\varpi}\widetilde{h}_t^{(p)}\big|^2\Big] \\
&\geqslant -CE_t^{(p)}\Big[\varpi^{-2n} + \big|h_t^{n+1} - \tfrac{1}{\varpi}h_t^{(p)}\big|^2 + \tfrac{1}{\delta\varepsilon^2}\big|h_t^{n} - \tfrac{1}{\varpi}h_t^{(p)}\big|^2\Big]\mathrm{d}t + \tfrac{\varepsilon^2}{2}E_t^{(p)}\big|k_t^{n+1} - \tfrac{1}{\varpi}k_t^{(p)}\big|^2\mathrm{d}t \\
&\quad + \varepsilon E_t^{(p)}\big(h_t^{n+1} - \tfrac{1}{\varpi}h_t^{(p)}\big)\cdot\Big[\big(k_t^{n+1} - \tfrac{1}{\varpi}k_t^{(p)}\big)\mathrm{d}W_t^{\boldsymbol{h}^{(p)/\varepsilon}}\Big].
\end{aligned} \tag{2.2.62}$$

Then, using in addition the fact that $\widetilde{h}_T^{n+1} - \widetilde{h}_T^{(p)}/\varpi = [g(\overline{m}_T^n) - g(m_T^{(p)})]/\varpi$, we get

$$
\begin{aligned}
&E_t^{(p)}\big|\widetilde{h}_t^{n+1} - \tfrac{1}{\varpi}\widetilde{h}_t^{(p)}\big|^2 + \tfrac{\varepsilon^2}{2}\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\bigg[\int_t^T E_s^{(p)}\big|k_s^{n+1} - \tfrac{1}{\varpi}k_s^{(p)}\big|^2\mathrm{d}s \,|\, \mathcal{F}_t^{\boldsymbol{W}}\bigg] \\
&\quad \leqslant C\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\bigg[E_T^{(p)}\varpi^{-2n} + \int_t^T E_s^{(p)}\Big[\varpi^{-2n} + \big|h_s^{n+1} - \tfrac{1}{\varpi}h_s^{(p)}\big|^2 + \tfrac{1}{\delta\varepsilon^2}\big|h_s^n - \tfrac{1}{\varpi}h_s^{(p)}\big|^2\Big]\mathrm{d}s \,|\, \mathcal{F}_t^{\boldsymbol{W}}\bigg].
\end{aligned}
$$

Take now $t = \ell T/p$, for some $\ell \in \{0, \cdots, p-1\}$. Then, with C and δ as above,

$$
\begin{aligned}
&E_{\ell T/p}^{(p)}\big|h_{\ell T/p}^{n+1} - \tfrac{1}{\varpi}h_{\ell T/p}^{(p)}\big|^2 \\
&\leqslant C\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\bigg[E_T^{(p)}\varpi^{-2n} + \int_{\ell T/p}^T E_s^{(p)}\Big[\big|h_s^{n+1} - \tfrac{1}{\varpi}h_s^{(p)}\big|^2 + \tfrac{1}{\delta\varepsilon^2}\big|h_s^n - \tfrac{1}{\varpi}h_s^{(p)}\big|^2\Big]\mathrm{d}s \,|\, \mathcal{F}_{\ell T/p}^{\boldsymbol{W}}\bigg] \\
&\leqslant C\exp\bigg(\delta\varepsilon^2\int_0^{\ell T/p}|k_s^{(p)}|^2\mathrm{d}s\bigg)\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\bigg[\exp\bigg(\delta\varepsilon^2\int_{\ell T/p}^T|k_s^{(p)}|^2\mathrm{d}s\bigg) \,|\, \mathcal{F}_{\ell T/p}^{\boldsymbol{W}}\bigg] \\
&\times\bigg(\exp\big(C\tfrac{T}{\delta\varepsilon^2}\big)\varpi^{-2n} + \tfrac{T}{p}\sum_{k=\ell}^{p-1}\exp\big(C\tfrac{kT}{p\delta\varepsilon^2}\big)\operatorname*{essup}_{\omega\in\Omega}\Big[\big|h_{kT/p}^{n+1} - \tfrac{1}{\varpi}h_{kT/p}^{(p)}\big|^2 + \tfrac{1}{\delta\varepsilon^2}\big|h_{kT/p}^n - \tfrac{1}{\varpi}h_{kT/p}^{(p)}\big|^2\Big]\bigg).
\end{aligned}
$$

By (2.2.59), we have a bound for the expectation on the penultimate line. Also, recalling the form of $E_{\ell T/p}^{(p)}$ in (2.2.61), we can divide both sides of the inequality by $\exp(\delta\varepsilon^2\int_0^{\ell T/p}|k_s^{(p)}|^2\mathrm{d}s)$. We get

$$
\begin{aligned}
&\exp\big(C\tfrac{\ell T}{p\delta\varepsilon^2}\big)\operatorname*{essup}_{\omega\in\Omega}\big|h_{\ell T/p}^{n+1} - \tfrac{1}{\varpi}h_{\ell T/p}^{(p)}\big|^2 \\
&\leqslant C\exp\big(C\tfrac{T}{\delta\varepsilon^2}\big)\varpi^{-2n} + \tfrac{CT}{p}\sum_{k=\ell}^{p-1}\exp\big(C\tfrac{kT}{p\delta\varepsilon^2}\big)\operatorname*{essup}_{\omega\in\Omega}\Big[\big|h_{kT/p}^{n+1} - \tfrac{1}{\varpi}h_{kT/p}^{(p)}\big|^2 + \tfrac{1}{\delta\varepsilon^2}\big|h_{kT/p}^n - \tfrac{1}{\varpi}h_{kT/p}^{(p)}\big|^2\Big].
\end{aligned}
$$

For $CT/p \leqslant 1/2$ (which assumption is consistent with the lower bound $p\varepsilon^2 \geqslant c$), we get (for a possibly new value of C)

$$
\exp\big(C\tfrac{\ell T}{p\delta\varepsilon^2}\big)\operatorname*{essup}_{\omega\in\Omega}\big|h_{\ell T/p}^{n+1} - \tfrac{1}{\varpi}h_{\ell T/p}^{(p)}\big|^2 \leqslant A_\ell + \tfrac{CT}{p}\sum_{k=\ell+1}^{p-1}\exp\big(C\tfrac{kT}{p\delta\varepsilon^2}\big)\operatorname*{essup}_{\omega\in\Omega}\big|h_{kT/p}^{n+1} - \tfrac{1}{\varpi}h_{kT/p}^{(p)}\big|^2,
$$

with the notation

$$
A_\ell := C\exp\big(C\tfrac{T}{\delta\varepsilon^2}\big) + \tfrac{CT}{p\delta\varepsilon^2}\sum_{k=\ell}^{p-1}\exp\big(C\tfrac{kT}{p\delta\varepsilon^2}\big)\operatorname{essup}_{\omega\in\Omega}\big|h_{kT/p}^n - \tfrac{1}{\varpi}h_{kT/p}^{(p)}\big|^2.
$$

By the discrete version of Gronwall's lemma, we obtain (for a possibly new value of C):

$$
\begin{aligned}
&\exp\big(C\tfrac{\ell T}{p\delta\varepsilon^2}\big)\operatorname{essup}_{\omega\in\Omega}\big|h_{\ell T/p}^{n+1} - \tfrac{1}{\varpi}h_{\ell T/p}^{(p)}\big|^2 \\
&\quad \leqslant CA_\ell \leqslant C\exp\big(C\tfrac{T}{\delta\varepsilon^2}\big) + \tfrac{CT}{p\delta\varepsilon^2}\sum_{k=\ell}^{p-1}\exp\big(C\tfrac{kT}{p\delta\varepsilon^2}\big)\operatorname{essup}_{\omega\in\Omega}\big|h_{kT/p}^n - \tfrac{1}{\varpi}h_{kT/p}^{(p)}\big|^2.
\end{aligned}
\tag{2.2.63}
$$

And then, for some real $\lambda > 0$,

$$\begin{aligned}
&\sum_{\ell=0}^{p-1} \exp\bigl(\lambda \tfrac{\ell T}{p\delta\varepsilon^2}\bigr) \exp\bigl(C \tfrac{\ell T}{p\delta\varepsilon^2}\bigr) \operatorname*{essup}_{\omega\in\Omega} \bigl|h^{n+1}_{\ell T/p} - \tfrac{1}{\varpi} h^{(p)}_{\ell T/p}\bigr|^2 \\
&\leqslant C \exp\bigl(C \tfrac{T}{\delta\varepsilon^2}\bigr) \sum_{\ell=0}^{p-1} \exp\bigl(\lambda \tfrac{\ell T}{p\delta\varepsilon^2}\bigr) \varpi^{-2n} \\
&\quad + \tfrac{CT}{p\delta\varepsilon^2} \sum_{k=0}^{p-1} \biggl(\sum_{\ell=0}^{k} \exp\bigl(\lambda \tfrac{\ell T}{p\delta\varepsilon^2}\bigr) \biggr) \exp\bigl(C \tfrac{kT}{p\delta\varepsilon^2}\bigr) \operatorname*{essup}_{\omega\in\Omega} \bigl|h^{n}_{kT/p} - \tfrac{1}{\varpi} h^{(p)}_{kT/p}\bigr|^2 \\
&\leqslant \tfrac{C}{\exp(\lambda T/(p\delta\varepsilon^2))-1} \exp\bigl((C+\lambda) \tfrac{T}{\delta\varepsilon^2}\bigr) \varpi^{-2n} \\
&\quad + \tfrac{CT}{p\delta\varepsilon^2} \tfrac{\exp(\lambda T/(p\delta\varepsilon^2))}{\exp(\lambda T/(p\delta\varepsilon^2))-1} \sum_{k=0}^{p-1} \exp\bigl((C+\lambda) \tfrac{kT}{p\delta\varepsilon^2}\bigr) \operatorname*{essup}_{\omega\in\Omega} \bigl|h^{n}_{kT/p} - \tfrac{1}{\varpi} h^{(p)}_{kT/p}\bigr|^2.
\end{aligned}$$

Now,

$$\tfrac{T}{p\delta\varepsilon^2} \tfrac{\exp(\lambda T/(p\delta\varepsilon^2))}{\exp(\lambda T/(p\delta\varepsilon^2))-1} = \tfrac{T}{p\delta\varepsilon^2} \tfrac{1}{1-\exp(-\lambda T/(p\delta\varepsilon^2))} = \tfrac{1}{\lambda} \varphi\bigl(\tfrac{\lambda T}{p\delta\varepsilon^2}\bigr),$$

where

$$\varphi(x) = \tfrac{x}{1-\exp(-x)}, \quad x > 0.$$

Clearly, $\varphi(x) \to 1$ as $x \to 0$. Take now $\lambda = 6C$ and $p\varepsilon^2$ large enough so that $\varphi(\tfrac{\lambda T}{p\delta\varepsilon^2}) = \varphi(\tfrac{6CT}{p\delta\varepsilon^2})$ is less than 2. We have

$$\begin{aligned}
&\sum_{\ell=0}^{p-1} \exp\bigl(7C \tfrac{\ell T}{p\delta\varepsilon^2}\bigr) \operatorname*{essup}_{\omega\in\Omega} \bigl|h^{n+1}_{\ell T/p} - \tfrac{1}{\varpi} h^{(p)}_{\ell T/p}\bigr|^2 \\
&\leqslant \tfrac{1}{3} \sum_{\ell=0}^{p-1} \exp\bigl(7C \tfrac{\ell T}{p\delta\varepsilon^2}\bigr) \operatorname*{essup}_{\omega\in\Omega} \bigl|h^{n}_{\ell T/p} - \tfrac{1}{\varpi} h^{(p)}_{\ell T/p}\bigr|^2 + \tfrac{C}{1-\exp(-6CT/(p\delta\varepsilon^2))} \exp\bigl(7C \tfrac{T}{\delta\varepsilon^2}\bigr) \varpi^{-2n}.
\end{aligned}$$

Here, we notice that (assuming without any loss of generality that $C \geqslant 1$)

$$\tfrac{C}{1-\exp(-6CT/(p\delta\varepsilon^2))} = \tfrac{p\delta\varepsilon^2}{6T} \varphi\Bigl(\tfrac{6CT}{p\delta\varepsilon^2}\Bigr) \leqslant \tfrac{p\delta\varepsilon^2}{3T}$$

and then,

$$\begin{aligned}
&\sum_{\ell=0}^{p-1} \exp\bigl(7C \tfrac{\ell T}{p\delta\varepsilon^2}\bigr) \operatorname*{essup}_{\omega\in\Omega} \bigl|h^{n+1}_{\ell T/p} - h^{(p),\varpi}_{\ell T/p}\bigr|^2 \\
&\leqslant \tfrac{1}{3} \sum_{\ell=0}^{p-1} \exp\bigl(7C \tfrac{\ell T}{p\delta\varepsilon^2}\bigr) \operatorname*{essup}_{\omega\in\Omega} \bigl|h^{n}_{\ell T/p} - h^{(p),\varpi}_{\ell T/p}\bigr|^2 + C' p \exp\bigl(7C \tfrac{T}{\delta\varepsilon^2}\bigr) \varpi^{-2n},
\end{aligned}$$

for a constant C' depending on the same parameters as C. The above inequality is very similar to (2.2.26). Similar to (2.2.27), we obtain

$$\sum_{\ell=0}^{p-1} \exp\bigl(7C \tfrac{\ell T}{p\delta\varepsilon^2}\bigr) \operatorname*{essup}_{\omega\in\Omega} \bigl|h^{n+1}_{\ell T/p} - \tfrac{1}{\varpi} h^{(p)}_{\ell T/p}\bigr|^2 \leqslant C' p \exp\bigl(7C \tfrac{T}{\delta\varepsilon^2}\bigr) \varpi^{-2n},$$

for a possibly new value of C'. Back to (2.2.63), we obtain (2.2.47). Inequality (2.2.48) is proven as (2.2.13). □

2.2.2.4 Convergence to the MFG

We now study the distance between $(\boldsymbol{m}^{(p)}, \boldsymbol{h}^{(p)})$ and $(\boldsymbol{m}, \boldsymbol{h})$ (see (2.2.11)) as p tends to ∞. While this looks a natural question, the following result, which is given for reader's interest only, has a secondary role in our study. The limit $p \to \infty$ will be addressed in the next subsection but from another point of view.

Proposition 2.2.23. *With $(\boldsymbol{m}^{(p)}, \widetilde{\boldsymbol{h}}^{(p)})$ as in the statement of Theorem 2.2.17 (with the same extension as in Remark 2.2.19) and with $(\boldsymbol{m}, \boldsymbol{h})$ as in the statement of Theorem 2.2.4 and* (2.2.14)*, there exists a constant C, only depending on d, T, $\|f\|_{1,\infty}$, $\|g\|_{1,\infty}$, $|Q|$ and $|R|$, such that*

$$\sup_{0\leqslant t\leqslant T} \mathbb{E}\Big[|h_t - \widetilde{h}_t^{(p)}|^2\Big] \leqslant \exp\Big(\tfrac{C}{\varepsilon^2}\Big)\tfrac{\ln(p)}{p}.$$

Proof. For an integer $p \geqslant 1$, we write

$$\mathrm{d}m_t^{(p)} = -\eta^{(p)}_{\tau_p(t)} m^{(p)}_{\tau_p(t)}\mathrm{d}t + \varepsilon \mathrm{d}W_t^p, \quad t \in [0,T].$$

We notice that

$$\begin{aligned} \big|m_t^{(p)} - m^{(p)}_{\tau_p(t)}\big| &\leqslant C\tfrac{T}{p}|m^{(p)}_{\tau_p(t)}| + \varepsilon \sup_{0\leqslant s-\tau_p(t)\leqslant T/p} |W_s - W_{\tau_p(t)}| \\ &\leqslant C\tfrac{T}{p}\sup_{0\leqslant s\leqslant T}|m_s^{(p)}| + \varepsilon \max_{k=0,\cdots,p-1}\sup_{kT/p\leqslant s\leqslant (k+1)T/p}|W_s - W_{kT/p}|, \quad t\in[0,T]. \end{aligned} \tag{2.2.64}$$

Moreover,

$$\begin{aligned} \mathrm{d}\big(m_t^{(p)} - m_t\big) &= -\big(\eta^{(p)}_{\tau_p(t)} m^{(p)}_{\tau_p(t)} - \eta_t m_t\big)\mathrm{d}t + \mathrm{d}(W_t^p - W_t) \\ &= -\big(\eta^{(p)}_{\tau_p(t)} m^{(p)}_{\tau_p(t)} - \eta_t^{(p)} m_t^{(p)}\big)\mathrm{d}t - \big(\eta_t^{(p)} - \eta_t\big)m_t^{(p)}\mathrm{d}t - \eta_t\big(m_t^{(p)} - m_t\big)\mathrm{d}t \\ &\quad + \mathrm{d}(W_t^p - W_t). \end{aligned}$$

By combining the last two displays with (2.2.40), we deduce from Gronwall's lemma that

$$\begin{aligned} |m_t^{(p)} - m_t| &\leqslant C\tfrac{T}{p}\sup_{0\leqslant s\leqslant T}|m_s^{(p)}| + \varepsilon \sup_{0\leqslant s\leqslant T}|W_s - W_s^p| \\ &\quad + \varepsilon \max_{k=0,\cdots,p-1}\sup_{kT/p\leqslant s\leqslant (k+1)T/p}|W_s - W_{kT/p}| \\ &\leqslant C\tfrac{T}{p}\sup_{0\leqslant s\leqslant T}|m_s^{(p)}| + C\varepsilon \max_{k=0,\cdots,p-1}\sup_{kT/p\leqslant s\leqslant (k+1)T/p}|W_s - W_{kT/p}|, \end{aligned} \tag{2.2.65}$$

for $t \in [0,T]$. By Lemma 2.2.8 (the proof of which also provides a bound for the time-derivative of

θ_ε) and (2.2.17)–(2.2.18) and by the analogue of (2.2.64) but for $\boldsymbol{m}$, we also have

$$\begin{aligned}
|h_{\tau_p(t)} - h_t| &= \big|\theta_\varepsilon\big(\tau_p(t), m_{\tau_p(t)}\big) - \theta_\varepsilon\big(t, m_t\big)\big| \\
&\leqslant \tfrac{C}{\varepsilon^2}\Big(\tfrac{T}{p} + |m_{\tau_p(t)} - m_t|\Big) \\
&\leqslant \tfrac{C}{\varepsilon^2}\Big(\tfrac{T}{p} + \tfrac{T}{p}\sup_{0\leqslant s\leqslant T}|m_s| + \varepsilon \max_{k=0,\cdots,p-1}\sup_{kT/p\leqslant s\leqslant (k+1)T/p}|W_s - W_{kT/p}|\Big) \\
&\leqslant \tfrac{C}{\varepsilon^2}\Big(\tfrac{T}{p} + \tfrac{T}{p}\sup_{0\leqslant s\leqslant T}|m^{(p)}_s| + \varepsilon \max_{k=0,\cdots,p-1}\sup_{kT/p\leqslant s\leqslant (k+1)T/p}|W_s - W_{kT/p}|\Big).
\end{aligned} \tag{2.2.66}$$

Next, we rewrite the first term in the right-hand side in (2.2.46) as (which follows from the last paragraph in Remark 2.2.16)

$$\mathbb{E}\Big[Q^\dagger f\big(m^{(p)}_{\tau_p(t)+T/p}\big)\,|\,\mathcal{F}^{p,\boldsymbol{W}}_{\tau_p(t)}\Big] = \mathbb{E}\Big[Q^\dagger f\big(m^{(p)}_{\tau_p(t)+T/p}\big)\,|\,\mathcal{F}^{\boldsymbol{W}}_{\tau_p(t)}\Big].$$

Meanwhile,

$$\begin{aligned}
&\Big|\mathbb{E}\Big[Q^\dagger f\big(m^{(p)}_{\tau_p(t)+T/p}\big)\,|\,\mathcal{F}^{\boldsymbol{W}}_{\tau_p(t)}\Big] - Q^\dagger f\big(m_t\big)\Big| \\
&\leqslant \Big|\mathbb{E}\Big[Q^\dagger f\big(m^{(p)}_{\tau_p(t)+T/p}\big)\,|\,\mathcal{F}^{\boldsymbol{W}}_{\tau_p(t)}\Big] - Q^\dagger f\big(m^{(p)}_{\tau_p(t)}\big)\Big| + \Big|Q^\dagger f\big(m^{(p)}_{\tau_p(t)}\big) - Q^\dagger f\big(m_t\big)\Big|.
\end{aligned}$$

By the first line in (2.2.64),

$$\Big|\mathbb{E}\Big[Q^\dagger f\big(m^{(p)}_{\tau_p(t)+T/p}\big)\,|\,\mathcal{F}^{\boldsymbol{W}}_{\tau_p(t)}\Big] - Q^\dagger f\big(m^{(p)}_{\tau_p(t)}\big)\Big| \leqslant C\Big(\tfrac{T}{p}\sup_{0\leqslant s\leqslant T}|m^{(p)}_s| + \varepsilon\sqrt{\tfrac{T}{p}}\Big).$$

Together with (2.2.65), this gives

$$\begin{aligned}
&\Big|\mathbb{E}\Big[Q^\dagger f\big(m^{(p)}_{\tau_p(t)+T/p}\big)\,|\,\mathcal{F}^{\boldsymbol{W}}_{\tau_p(t)}\Big] - Q^\dagger f\big(m_t\big)\Big| \\
&\leqslant C\Big(\tfrac{T}{p}\sup_{0\leqslant s\leqslant T}|m^{(p)}_s| + \varepsilon \max_{k=0,\cdots,p-1}\sup_{kT/p\leqslant s\leqslant (k+1)T/p}|W_s - W_{kT/p}| + \varepsilon\sqrt{\tfrac{T}{p}}\Big).
\end{aligned}$$

By the above display and by (2.2.66), we can rewrite the equation for $\boldsymbol{h}$ in (2.2.16) in the form:

$$\begin{aligned}
\mathrm{d}h_t = \Big\{&-\mathbb{E}\Big[Q^\dagger f\big(m^{(p)}_{\tau_p(t)+T/p}\big)\,|\,\mathcal{F}^{\boldsymbol{W}}_{\tau_p(t)}\Big]\mathbf{1}_{\{t\leqslant (p-1)T/p\}} \\
&+ \Big(\eta^{(p)}_{\tau_p(t)+T/p} + \tfrac{T}{p}Q^\dagger Q\mathbf{1}_{\{t\leqslant (p-1)T/p\}}\Big)h_{\tau_p(t)}\Big\}\mathrm{d}t + e_t\mathrm{d}t + k_t h_t\mathrm{d}t + \varepsilon k_t\mathrm{d}W_t,
\end{aligned}$$

for $t\in[0,T]$, where

$$\begin{aligned}
|e_t| &\leqslant C\mathbf{1}_{\{(p-1)T/p<t\leqslant T\}} + C\varepsilon\sqrt{\tfrac{T}{p}} \\
&\quad + C\Big(\tfrac{T}{p} + \tfrac{T}{p}\sup_{0\leqslant s\leqslant T}|m^{(p)}_s| + \varepsilon \max_{k=0,\cdots,p-1}\sup_{kT/p\leqslant s\leqslant (k+1)T/p}|W_s - W_{kT/p}|\Big).
\end{aligned} \tag{2.2.67}$$

Next, by (2.2.46),

$$\begin{aligned}\mathrm{d}\big[h_t-\widetilde{h}^{(p)}_t\big]&=\Big(\eta^{(p)}_{\tau_p(t)+T/p}+\tfrac{T}{p}Q^\dagger Q\mathbf{1}_{\{t\leqslant(p-1)T/p\}}\Big)\big(h_{\tau_p(t)}-\widetilde{h}^{(p)}_{\tau_p(t)}\big)\mathrm{d}t+e_t\mathrm{d}t\\&\quad+\big(k_th_t-k^{(p)}_t\widetilde{h}^{(p)}_t\big)\mathrm{d}t+\varepsilon\big(k_t-k^{(p)}_t\big)\mathrm{d}W_t,\quad t\in[0,T].\end{aligned}$$

Then, proceeding as in (2.2.22) and (2.2.23) (using Lemma 2.2.8),

$$\begin{aligned}\mathrm{d}|h_t-\widetilde{h}^{(p)}_t|^2&\geqslant-C|h_t-\widetilde{h}^{(p)}_t|^2\mathrm{d}t-C|h_{\tau_p(t)}-\widetilde{h}^{(p)}_{\tau_p(t)}|^2\mathrm{d}t-2|h_t-\widetilde{h}^{(p)}_t||e_t|\mathrm{d}t\\&\quad-C|h_t-\widetilde{h}^{(p)}_t|\,|k_t-k^{(p)}_t|-C|k_t||h_t-\widetilde{h}^{(p)}_t|^2\mathrm{d}t+\varepsilon^2|k_t-k^{(p)}_t|^2\mathrm{d}t\\&\quad+2\varepsilon\big(h_t-\widetilde{h}^{(p)}_t\big)\cdot\big[\big(k_t-k^{(p)}_t\big)\mathrm{d}W_t\big]\\&\geqslant-\tfrac{C}{\varepsilon^2}|h_t-\widetilde{h}^{(p)}_t|^2\mathrm{d}t-\varepsilon^2|e_t|^2\mathrm{d}t-C|h_{\tau_p(t)}-h^{(p)}_{\tau_p(t)}|^2\mathrm{d}t\\&\quad+2\varepsilon\big(h_t-\widetilde{h}^{(p)}_t\big)\cdot\big[\big(k_t-k^{(p)}_t\big)\mathrm{d}W_t\big].\end{aligned}$$

There is one small subtlety here to handle the last term on the penultimate line because it is indexed by time $\tau_p(t)$ (and not t). The strategy is to take expectation on both sides in the above inequality and then to integrate in time between $\tau_p(t)$ and t for a given $t\in[0,T)$. As for the dynamics of the expectation, we have

$$\mathrm{d}\mathbb{E}\big[|h_t-\widetilde{h}^{(p)}_t|^2\big]\geqslant-\tfrac{C}{\varepsilon^2}\mathbb{E}\big[|h_t-\widetilde{h}^{(p)}_t|^2\big]\mathrm{d}t-\mathbb{E}\big[|e_t|^2\big]\mathrm{d}t-C\mathbb{E}\big[|h_{\tau_p(t)}-h^{(p)}_{\tau_p(t)}|^2\big]\mathrm{d}t.\tag{2.2.68}$$

By (2.2.67), notice that

$$\mathbb{E}\big[|e_t|^2\big]\leqslant C\mathbf{1}_{\{(p-1)T/p<t\leqslant T\}}+\tfrac{C}{p^2}+C\mathbb{E}\Big[\max_{k=0,\cdots,p-1}\sup_{kT/p\leqslant s\leqslant(k+1)T/p}|W_s-W_{kT/p}|^2\Big].\tag{2.2.69}$$

We admit for a while that

$$\mathbb{E}\Big[\max_{k=0,\cdots,p-1}\sup_{kT/p\leqslant s\leqslant(k+1)T/p}|W_s-W_{kT/p}|^2\Big]\leqslant C\tfrac{\ln(p)}{p}.\tag{2.2.70}$$

And then, by integrating (2.2.68) between $\tau_p(t)$ and t and by invoking boundedness of the two processes $\boldsymbol{h}$ and $\widetilde{\boldsymbol{h}}^{(p)}$ (see Lemmas 2.2.10 and 2.2.22), we deduce that

$$\begin{aligned}\mathbb{E}\big[|h_{\tau_p(t)}-\widetilde{h}^{(p)}_{\tau_p(t)}|^2\big]&\leqslant C\mathbb{E}\big[|h_t-\widetilde{h}^{(p)}_t|^2\big]+\tfrac{C}{\varepsilon^2}\int_{\tau_p(t)}^t\mathbb{E}\big[|h_t-\widetilde{h}^{(p)}_t|^2\big]\mathrm{d}s+\tfrac{C\ln(p)}{p}\\&\leqslant C\mathbb{E}\big[|h_t-\widetilde{h}^{(p)}_t|^2\big]+\tfrac{C\ln(p)}{\varepsilon^2p}.\end{aligned}$$

Inserting the above estimate into (2.2.68) and then invoking (2.2.69), we obtain

$$\mathrm{d}\mathbb{E}\big[|h_t-\widetilde{h}^{(p)}_t|^2\big]\geqslant-\tfrac{C}{\varepsilon^2}\mathbb{E}\big[|h_t-\widetilde{h}^{(p)}_t|^2\big]\mathrm{d}t-C\Big(\mathbf{1}_{\{(p-1)T/p<t\leqslant T\}}+\tfrac{\ln(p)}{\varepsilon^2p}\Big)\mathrm{d}t.$$

By Gronwall's lemma (and by (2.2.65), which allows us to control the terminal condition), we get

$$\mathbb{E}\big[|h_t-\widetilde{h}^{(p)}_t|^2\big]\leqslant\exp\Big(\tfrac{C}{\varepsilon^2}\Big)\tfrac{C\ln(p)}{\varepsilon^2p},\quad t\in[0,T].$$

Changing the value of C, we easily complete the proof of the statement.

It remains to check (2.2.70). We have, for any $r > 0$,

$$\begin{aligned}
&\mathbb{P}\Big(\Big\{\max_{k=0,\cdots,p-1}\sup_{kT/p\leqslant s\leqslant(k+1)T/p}|W_s - W_{kT/p}|^2 \geqslant r\Big\}\Big)\\
&= 1 - \mathbb{P}\Big(\Big\{\max_{k=0,\cdots,p-1}\sup_{kT/p\leqslant s\leqslant(k+1)T/p}|W_s - W_{kT/p}|^2 < r\Big\}\Big)\\
&= 1 - \mathbb{P}\Big(\Big\{\sup_{0\leqslant s\leqslant T/p}|W_s|^2 < r\Big\}\Big)^p\\
&\leqslant 1 - \big(1 - \exp(-cpr)\big)^p \leqslant p\exp(-cpr),
\end{aligned}$$

for a constant c only depending on T. Replacing r by $\ln(p)r/p$, we easily deduce that

$$\mathbb{E}\Big[\max_{k=0,\cdots,p-1}\sup_{kT/p\leqslant s\leqslant(k+1)T/p}|W_s - W_{kT/p}|^2\Big] \leqslant C\frac{\ln(p)}{p},$$

which is (2.2.70). □

2.2.2.5 Proof of auxiliary results

Proof of Lemma 2.2.13.

First Step. We first notice that, for a given choice of $(\overline{\boldsymbol{m}}^n, \boldsymbol{h}^n)$, the problem (2.2.32) has at least one minimizer. Indeed, the cost $\boldsymbol{\alpha} \mapsto \mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}[\mathcal{R}^{p,\varpi X_0}(\varpi\boldsymbol{\alpha} + \varpi\varepsilon\dot{\boldsymbol{W}}^{p,\varpi\boldsymbol{h}^n/\varepsilon}; \overline{\boldsymbol{m}}^n; 0)]$ is lower semicontinuous with respect to $\boldsymbol{\alpha}$ when the latter is identified with a collection of random variables $(\alpha_0, \cdots, \alpha_{(p-1)T/p}) \in \times_{\ell=0}^{p-1} L^2(\Omega, \mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{\ell T/p}, \mathbb{P}^{\varpi\boldsymbol{h}^n/\varepsilon}; \mathbb{R}^d)$, with the product space being equipped with the weak topology. The identification is made possible by the fact that the controls $\boldsymbol{\alpha}$ are chosen to be piecewise constant. Moreover, the cost functional blows up when the L^2-norm of $\boldsymbol{\alpha}$ tends to ∞. The minimization problem can hence be restricted to a bounded subset of $\times_{\ell=0}^{p-1} L^2(\Omega, \mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{\ell T/p}, \mathbb{P}^{\varpi\boldsymbol{h}^n/\varepsilon}; \mathbb{R}^d)$, the latter being obviously compact for the weak topology. Existence of a minimizer easily follows.

We check below that any critical point of the cost functional is a solution of the forward-backward system (2.2.37)–(2.2.38). Since the latter is uniquely solvable, this indeed provides a characterization of the minimizer.

Second Step. In order to prove that critical points solve (2.2.37)–(2.2.38), we follow the usual lines of the Pontryagin principle. Although the result is certainly not new, we feel better to provide the complete proof, since the formulation we use is tailor-made to our needs. We start to notice that, under the probability $\mathbb{P}^{\varpi\boldsymbol{h}^n/\varepsilon}$, the path driven by the control $\varpi\boldsymbol{\alpha}$ and the initial condition ϖX_0 reads $(\varpi X_t)_{0\leqslant t\leqslant T}$ with

$$\mathrm{d}X_t = \alpha_t\mathrm{d}t + \tfrac{1}{\varpi}\sigma\mathrm{d}B_t + \varepsilon\mathrm{d}W_t^{p,\varpi\boldsymbol{h}^n/\varepsilon}, \quad t \in [0,T].$$

We then introduce an additive perturbation of the cost and hence replace $\boldsymbol{\alpha} = (\alpha_t)_{0\leqslant t\leqslant T}$ by $\boldsymbol{\alpha}+\delta\boldsymbol{\beta} = (\alpha_t+\delta\beta_t)_{0\leqslant t\leqslant T}$ in the above expansion, with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ being identified as before as elements of $\times_{\ell=0}^{p-1} L^2(\Omega, \mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{\ell T/p}, \mathbb{P}^{\varpi\boldsymbol{h}^n/\varepsilon}; \mathbb{R}^d)$. We then write $\boldsymbol{X}^\delta = (X_t^\delta)_{0\leqslant t\leqslant T}$ in order to emphasize the

dependence of $\boldsymbol{X}$ upon δ. When $\delta = 0$, we merely write $\boldsymbol{X}$ instead of $\boldsymbol{X}^0$. The formal derivative with respect to δ at $\delta = 0$ reads

$$\mathrm{d}\big(\partial X_t\big) = \beta_t \mathrm{d}t, \quad t \in [0,T]\,; \quad \partial X_0 = 0.$$

In turn, the derivative of the cost $\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}[\mathcal{R}^{p,\varpi X_0}(\varpi[\boldsymbol{\alpha} + \delta\boldsymbol{\beta}] + \varpi\varepsilon \dot{\boldsymbol{W}}^{p,\varpi \boldsymbol{h}^n/\varepsilon}; \overline{\boldsymbol{m}}^n; 0)]$, with respect to δ, is

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\delta}\Big\{\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{p,\varpi X_0}\Big(\varpi\boldsymbol{\alpha} + \varpi\varepsilon \dot{\boldsymbol{W}}^{p,\varpi \boldsymbol{h}^n/\varepsilon}; \overline{\boldsymbol{m}}^n; 0\Big)\Big]\Big\}_{|\delta=0} \\
&= \varpi^2 \mathbb{E}^{\boldsymbol{h}^n}\Bigg[\Big(RX_T + \tfrac{1}{\varpi} g\big(\overline{m}^n_T\big)\Big)\cdot R\partial X_T \\
&\qquad + \int_0^T \Big\{\Big(QX_{\tau_p(t)} + \tfrac{1}{\varpi} f\big(\overline{m}^n_{\tau_p(t)}\big)\Big)\cdot Q\partial X_{\tau_t(p)} + \alpha_{\tau_p(t)}\cdot\beta_{\tau_p(t)}\Big\}\mathrm{d}t\Bigg].
\end{aligned}$$

We now solve the BSDE (under $\mathbb{P}^{\varpi \boldsymbol{h}^n/\varepsilon}$)

$$\begin{aligned}
&\mathrm{d}Y_t = -\big(Q^\dagger Q X_{\tau_p(t)} + \tfrac{1}{\varpi} Q^\dagger f(\overline{m}^n_{\tau_p(t)})\big)\mathrm{d}t + Z^{\boldsymbol{B}}_t \mathrm{d}B_t + Z^{\boldsymbol{W}}_t \mathrm{d}W^{\varpi \boldsymbol{h}^n/\varepsilon}_t, \quad t \in [0,T], \\
&Y_T = R^\dagger R X_T + \tfrac{1}{\varpi} R^\dagger g\big(\overline{m}^n_T\big).
\end{aligned}$$

By discrete integration by parts, we have

$$\begin{aligned}
&\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\big[Y_T \cdot \partial X_T\big] \\
&= \sum_{\ell=0}^{p-1} \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\Big(Y_{(\ell+1)T/p} - Y_{\ell T/p}\Big)\cdot \partial X_{\ell T/p} + Y_{(\ell+1)T/p}\cdot\Big(\partial X_{(\ell+1)T/p} - \partial X_{\ell T/p}\Big)\Big] \\
&= \tfrac{T}{p}\sum_{\ell=0}^{p-1} \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[-\Big(Q^\dagger Q X_{\ell T/p} + \tfrac{1}{\varpi} Q^\dagger f(\overline{m}^n_{\ell T/p})\Big)\cdot \partial X_{\ell T/p} + Y_{(\ell+1)T/p}\cdot\beta_{\ell T/p}\Big].
\end{aligned}$$

Rewriting the above sum as an integral, we obtain

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\delta}\Big\{\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{p,\varpi X_0}\Big(\varpi\boldsymbol{\alpha} + \varpi\varepsilon \dot{\boldsymbol{W}}^{p,\varpi \boldsymbol{h}^n/\varepsilon}; \overline{\boldsymbol{m}}^n; 0\Big)\Big]\Big\}_{|\delta=0} \\
&= \varpi^2 \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\int_0^T \Big(Y_{\tau_p(t)+T/p} + \alpha_{\tau_p(t)}\Big)\cdot\beta_{\tau_p(t)}\mathrm{d}t\Big],
\end{aligned}$$

which means that the optimizer (which we merely write $\boldsymbol{\alpha}^{n+1}$ without specifying the indices p and ϖ) satisfies

$$\alpha^{n+1}_{\ell T/p} = -\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[Y^{n+1}_{(\ell+1)T/p} \,|\, \mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{\ell T/p}\Big] = -Y^{n+1}_{\ell T/p} + \tfrac{T}{p}\Big(Q^\dagger Q X^{n+1}_{\ell T/p} + \tfrac{1}{\varpi} Q^\dagger f(\overline{m}^n_{\ell T/p})\Big), \tag{2.2.71}$$

for $\ell \in \{0, \cdots, p-1\}$, where

$$\begin{aligned}
&\mathrm{d}X^{n+1}_t = \alpha^{n+1}_t \mathrm{d}t + \tfrac{1}{\varpi}\sigma \mathrm{d}B_t + \varepsilon \mathrm{d}W^{p,\varpi \boldsymbol{h}^n/\varepsilon}_t, \\
&\mathrm{d}Y^{n+1}_t = -\big(Q^\dagger Q X^{n+1}_{\tau_p(t)} + \tfrac{1}{\varpi} Q^\dagger f(\overline{m}^n_{\tau_p(t)})\big)\mathrm{d}t + Z^{n+1,\boldsymbol{B}}_t \mathrm{d}B_t + Z^{n+1,\boldsymbol{W}}_t \mathrm{d}W^{\varpi \boldsymbol{h}^n/\varepsilon}_t, \quad t \in [0,T], \\
&Y^{n+1}_T = R^\dagger R X^{n+1}_T + \tfrac{1}{\varpi} R^\dagger g(\overline{m}^n_T).
\end{aligned}$$

Importantly, we stress that, by construction, $\boldsymbol{X}^{n+1}_{\tau_p(\cdot)}$ is $(\mathbb{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}})_{0\leqslant t\leqslant T}$-adapted. We then let, for the same solution $\boldsymbol{\eta}^{(p)}$ as in (2.2.34)–(2.2.36),

$$\begin{aligned}\widetilde{h}^{n+1}_{\ell T/p} &= \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[Y^{n+1}_{(\ell+1)T/p} \,|\, \mathcal{F}^{X_0,\boldsymbol{B},\boldsymbol{W}}_{\ell T/p}\Big] - \eta^{(p)}_{\ell T/p} X^{n+1}_{\ell T/p} \\ &= Y^{n+1}_{\ell T/p} - \tfrac{T}{p}\Big(Q^\dagger Q X^{n+1}_{\ell T/p} + \tfrac{1}{\varpi} Q^\dagger f(\overline{m}^n_{\ell T/p})\Big) - \eta^{(p)}_{\ell T/p} X^{n+1}_{\ell T/p},\end{aligned} \tag{2.2.72}$$

for $\ell \in \{0, \cdots, p-1\}$. We obtain, for $\ell \in \{0, \cdots, p-2\}$,

$$\begin{aligned}\widetilde{h}^{n+1}_{(\ell+1)T/p} - \widetilde{h}^{n+1}_{\ell T/p} &= -\tfrac{T}{p}\Big(Q^\dagger Q X^{n+1}_{(\ell+1)T/p} + \tfrac{1}{\varpi} Q^\dagger f\big(\overline{m}^n_{(\ell+1)T/p}\big)\Big) - \tfrac{T}{p}\eta^{(p)}_{\ell T/p}\alpha^{n+1}_{\ell T/p} \\ &\quad - \Big(\eta^{(p)}_{(\ell+1)T/p} - \eta^{(p)}_{\ell T/p}\Big) X^{n+1}_{(\ell+1)T/p} + \int_{\ell T/p}^{(\ell+1)T/p} k^{n+1,\boldsymbol{B}}_s \mathrm{d}B_s + \int_{\ell T/p}^{(\ell+1)T/p} k^{n+1,\boldsymbol{W}}_s \mathrm{d}W^{\boldsymbol{h}^n}_s,\end{aligned}$$

for some square integrable and $\mathbb{F}^{X_0,\boldsymbol{B},\boldsymbol{W}}$-progressively measurable processes $\boldsymbol{k}^{n+1,\boldsymbol{B}}$ and $\boldsymbol{k}^{n+1,\boldsymbol{W}}$. Above, we used the fact that $W^{p,\varpi\boldsymbol{h}^n/\varepsilon}_{\tau_p(t)+T/p} - W^{p,\varpi\boldsymbol{h}^n/\varepsilon}_{\tau_p(t)} = W^{\varpi\boldsymbol{h}^n/\varepsilon}_{\tau_p(t)+T/p} - W^{\varpi\boldsymbol{h}^n/\varepsilon}_{\tau_p(t)}$. Now, using (2.2.71) and conditioning on $\mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{\tau_p(t)}$ in (2.2.72), we obtain

$$\alpha^{n+1}_{\ell T/p} = -\eta^{(p)}_{\ell T/p} X^{n+1}_{\ell T/p} - \mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\widetilde{h}^{n+1}_{\ell T/p} \,|\, \mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{\ell T/p}\Big], \quad \ell \in \{0, \cdots, p-1\},$$

which permits to write

$$\begin{aligned}X^{n+1}_{(\ell+1)T/p} &= X^{n+1}_{\ell T/p} + \tfrac{T}{p}\alpha^{n+1}_{\ell T/p} + \tfrac{1}{\varpi}\sigma\Big(B_{(\ell+1)T/p} - B_{\ell T/p}\Big) + \varepsilon\Big(W^{\varpi\boldsymbol{h}^n/\varepsilon}_{(\ell+1)T/p} - W^{\varpi\boldsymbol{h}^n/\varepsilon}_{\ell T}\Big) \\ &= \Big(I_d - \tfrac{T}{p}\eta^{(p)}_{\ell T/p}\Big) X^{n+1}_{\ell T/p} - \tfrac{T}{p}\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\widetilde{h}^{n+1}_{\ell T/p} \,|\, \mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{\ell T/p}\Big] \\ &\quad + \tfrac{1}{\varpi}\sigma\Big(B_{(\ell+1)T/p} - B_{\ell T/p}\Big) + \varepsilon\Big(W^{\varpi\boldsymbol{h}^n/\varepsilon}_{(\ell+1)T/p} - W^{\varpi\boldsymbol{h}^n/\varepsilon}_{\ell T}\Big).\end{aligned} \tag{2.2.73}$$

Modifying the processes $\boldsymbol{k}^{n+1,\boldsymbol{B}}$ and $\boldsymbol{h}^{n+1,\boldsymbol{W}}$ and using in addition the first equation for $\boldsymbol{\eta}^{(p)}$ in (2.2.34), we end-up with

$$\begin{aligned}\widetilde{h}^{n+1}_{(\ell+1)T/p} - \widetilde{h}^{n+1}_{\ell T/p} &= -\tfrac{T}{p\varpi} Q^\dagger f(\overline{m}^n_{(\ell+1)T/p}) + \tfrac{T}{p}\Big(\eta^{(p)}_{\ell T/p} + \rho^{(p)}_{\ell T/p}\Big)\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\widetilde{h}^{n+1}_{\ell T/p} \,|\, \mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{\ell T/p}\Big] \\ &\quad + \int_{\ell T/p}^{(\ell+1)T/p} k^{n+1,\boldsymbol{B}}_s \mathrm{d}B_s + \int_{\ell T/p}^{(\ell+1)T/p} k^{n+1,\boldsymbol{W}}_s \mathrm{d}W^{\boldsymbol{h}^n}_s,\end{aligned} \tag{2.2.74}$$

for $\ell \in \{0, \cdots, p-2\}$, where

$$\rho^{(p)}_{\ell T/p} = \tfrac{T}{p\varpi} Q^\dagger Q + \Big(\eta^{(p)}_{(\ell+1)T/p} - \eta^{(p)}_{\ell T/p}\Big).$$

By recalling that $\overline{m}^n_{(\ell+1)T/p}$ is $\mathcal{F}^{p,\boldsymbol{W}}_{(\ell+1)T/p}$-measurable and by modifying the process $\boldsymbol{k}^{n+1,\boldsymbol{W}}$ accordingly, we can rewrite (2.2.74) in the form

$$\begin{aligned}\widetilde{h}^{n+1}_{(\ell+1)T/p} - \widetilde{h}^{n+1}_{\ell T/p} &= -\tfrac{T}{p}\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\tfrac{1}{\varpi} Q^\dagger f(\overline{m}^n_{(\ell+1)T/p}) \,|\, \mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}\Big] \\ &\quad + \tfrac{T}{p}\Big(\eta^{(p)}_{\ell T/p} + \rho^{(p)}_{\ell T/p}\Big)\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\widetilde{h}^{n+1}_{\ell T/p} | \mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{\ell T/p}\Big] \\ &\quad + \int_{\ell T/p}^{(\ell+1)T/p} k^{n+1,\boldsymbol{B}}_s \mathrm{d}B_s + \int_{\ell T/p}^{(\ell+1)T/p} k^{n+1,\boldsymbol{W}}_s \mathrm{d}W^{\boldsymbol{h}^n}_s,\end{aligned} \tag{2.2.75}$$

for $\ell \in \{0, \cdots, p-2\}$. When $\ell = p-1$, we deduce from (2.2.72) and (2.2.73) that

$$\begin{aligned}
&\widetilde{h}^{n+1}_{(p-1)T/p} \\
&= R^\dagger R\, \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\big[X^{n+1}_T | \mathcal{F}^{X_0,\boldsymbol{B},\boldsymbol{W}}_{(p-1)T/p}\big] + \tfrac{1}{\varpi} R^\dagger \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\big[g\big(\overline{m}^n_T\big) \,|\, \mathcal{F}^{X_0,\boldsymbol{B},\boldsymbol{W}}_{(p-1)T/p}\big] - \eta^{(p)}_{(p-1)T/p} X^{n+1}_{(p-1)T/p} \\
&= \Big[R^\dagger R\Big(I_d - \frac{T}{p}\eta^{(p)}_{(p-1)T/p}\Big) - \eta^{(p)}_{(p-1)T/p}\Big] X^{n+1}_{(p-1)T/p} \\
&\quad + \tfrac{1}{\varpi} R^\dagger \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\big[g\big(\overline{m}^n_T\big) \,|\, \mathcal{F}^{X_0,\boldsymbol{B},\boldsymbol{W}}_{(p-1)T/p}\big] - \tfrac{T}{p} R^\dagger R\, \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\widetilde{h}^{n+1}_{(p-1)T/p} | \mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{(p-1)T/p}\Big],
\end{aligned}$$

which yields, using the boundary condition in (2.2.34) together with the fact that $\overline{m}^n_T$ is $\mathcal{F}^{\boldsymbol{W}}_T$-measurable

$$\widetilde{h}^{n+1}_{(p-1)T/p} = \tfrac{1}{\varpi} R^\dagger \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\big[g\big(\overline{m}^n_T\big) \,|\, \mathcal{F}^{\boldsymbol{W}}_{(p-1)T/p}\big] - \tfrac{T}{p} R^\dagger R\, \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\widetilde{h}^{n+1}_{(p-1)T/p} | \mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{(p-1)T/p}\Big]. \tag{2.2.76}$$

We now recall that, by construction, $\widetilde{h}^{n+1}_{\ell T/p}$ is $\mathcal{F}^{X_0,\boldsymbol{B},\boldsymbol{W}}_{\ell T/p}$-measurable. Arguing as in Remark 2.2.16, we can prove inductively (over ℓ) that $\widetilde{h}^{n+1}_{\ell T/p}$ is in fact $\mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}$-measurable. As a consequence, $\boldsymbol{k}^{n+1,\boldsymbol{B}}$ has to be zero. By combining (2.2.75) and (2.2.76), we hence recover BSDE (2.2.38).

□

Proof of Lemma 2.2.22. The bound (2.2.58) on $\boldsymbol{h}^n$ is a direct consequence of (2.2.42), which yields:

$$\mathrm{essup}_{\omega\in\Omega}|\widetilde{h}^{n+1}_{\ell T/p}| \leqslant \big(1 + \tfrac{C_1}{p}\big)\Big[\mathrm{essup}_{\omega\in\Omega}|\widetilde{h}^{n+1}_{(\ell+1)T/p}| + \tfrac{C_1}{p}\Big], \quad \ell \in \{0, \cdots, p-1\},$$

for C_1 as in the statement. Observing that $\mathrm{essup}_{\omega\in\Omega}|\widetilde{h}^{n+1}_T|$ is bounded by C_1, we get the result by means of a straightforward backward induction. The bound on $\widetilde{\boldsymbol{h}}^{(p)}$ is proven in a similar way.

In order to prove the second claim in the statement (see (2.2.59)), we invoke [76, Theorem 2.2]. Indeed, we notice that the $\mathbb{P}^{\boldsymbol{h}^{(p)}/\varepsilon}$-martingale

$$\Big(\varepsilon \int_0^t k^{(p)}_s \mathrm{d}W^{\boldsymbol{h}^{(p)}/\varepsilon}_s\Big)_{0\leqslant t\leqslant T}$$

has a finite BMO norm, see for instance (2.1) with $p = 2$ in the book [76] by Kazamaki. The proof is as follows. Rewriting (2.2.56) in the form

$$\begin{aligned}
\mathrm{d}\widetilde{h}^{(p)}_t &= \Big\{-\mathbb{E}\Big[Q^\dagger f\big(m^{(p)}_{\tau_p(t)+T/p}\big) \,|\, \mathcal{F}^{p,\boldsymbol{W}}_{\tau_p(t)}\Big]\mathbf{1}_{\{t\leqslant(p-1)T/p\}} \\
&\qquad + \Big(\eta^{(p)}_{\tau_{p(t)}+T/p} + \frac{T}{p} Q^\dagger Q \mathbf{1}_{\{t\leqslant(p-1)T/p\}}\Big) h^{(p)}_t\Big\}\mathrm{d}t + \varepsilon k^{(p)}_t \mathrm{d}W^{\boldsymbol{h}^{(p)}/\varepsilon}_t, \quad t \in [0,T), \\
\widetilde{h}^{(p)}_T &= R^\dagger g\big(m^{(p)}_T\big),
\end{aligned}$$

we easily deduce that, for any stopping τ,

$$\Big|\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\int_\tau^T \varepsilon k^{(p)}_s \mathrm{d}W^{\boldsymbol{h}^{(p)}/\varepsilon}_s \big| \mathcal{F}^{\boldsymbol{W}}_\tau\Big]\Big| \leqslant C_1,$$

for a possibly new value of C_1. We conclude by means of [76, Theorem 2.2]. □

Proof of Proposition 2.2.18. The equation for $(\widetilde{h}^{(p)}_{\ell T/p})_{\ell=0,\cdots,p}$ in (2.2.46) can be rewritten in the form:

$$\begin{aligned}&\Big[I_d + \tfrac{T}{p}\Big(\eta^{(p)}_{(\ell+1)T/p} + \tfrac{T}{p}Q^\dagger Q \mathbf{1}_{\{\ell\leqslant p-2\}}\Big)\Big]\widetilde{h}^{(p)}_{\ell T/p}\\ &= \mathbb{E}\Big[\mathcal{E}\big(\tfrac{1}{\varepsilon}\widetilde{h}^{(p)}_{\ell T/p}\big)\Big(\widetilde{h}^{(p)}_{(\ell+1)T/p} + \tfrac{T}{p}Q^\dagger f\big(m^{(p)}_{(\ell+1)T/p}\big)\mathbf{1}_{\{\ell\leqslant p-2\}}\Big) \,|\, \mathcal{F}^{\boldsymbol{W}}_{\ell T/p}\Big],\end{aligned} \tag{2.2.77}$$

for $\ell \in \{0,\cdots,p-1\}$, with the shorten notation

$$\mathcal{E}\big(\tfrac{1}{\varepsilon}\widetilde{h}^{(p)}_{\ell T/p}\big) := \exp\Big(-\tfrac{1}{\varepsilon}\widetilde{h}^{(p)}_{\ell T/p}\cdot\big(W_{(\ell+1)T/p} - W_{\ell T/p}\big) - \tfrac{T}{2\varepsilon^2 p}|\widetilde{h}^{(p)}_{\ell T/p}|^2\Big).$$

The challenge is thus to find, for each $\ell \in \{0,\cdots,p-1\}$, a variable $\widetilde{h}^{(p)}_{\ell T/p} \in L^2(\Omega, \mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}, \mathbb{P}; \mathbb{R}^d)$ that solves (2.2.77), when $\widetilde{h}^{(p)}_{(\ell+1)T/p}$ is given in $L^2(\Omega, \mathcal{F}^{p,\boldsymbol{W}}_{(\ell+1)T/p}, \mathbb{P}; \mathbb{R}^d)$.

In order to do so, we consider the mapping

$$\begin{aligned}\Phi_\ell : L^2(\Omega, \mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}, \mathbb{P}; \mathbb{R}^d) \ni y \mapsto &\Big[I_d + \tfrac{T}{p}\Big(\eta^{(p)}_{(\ell+1)T/p} + \tfrac{T}{p}Q^\dagger Q \mathbf{1}_{\{\ell\leqslant p-2\}}\Big)\Big]^{-1}\\ &\times \mathbb{E}\Big[\mathcal{E}\big(\tfrac{1}{\varepsilon}y\big)\Big(\widetilde{h}^{(p)}_{(\ell+1)T/p} + \tfrac{T}{p}Q^\dagger f\big(m^{(p)}_{(\ell+1)T/p}\big)\mathbf{1}_{\{\ell\leqslant p-2\}}\Big) \,|\, \mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}\Big]\end{aligned}$$

and then look for a fixed point to it.

Following the second step of Lemma 2.2.10 (see in particular (2.2.28)), we have, for any two $y, y' \in L^2(\Omega, \mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}, \mathbb{P}; \mathbb{R}^d)$, with probability 1 under $\mathbb{P}$,

$$d_{\mathrm{TV}}\Big(\mathbb{P}_y\big(\cdot \,|\, \mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}\big), \mathbb{P}_{y'}\big(\cdot \,|\, \mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}\big)\Big) \leqslant \sqrt{2}\Big(\tfrac{1}{2\varepsilon^2}\tfrac{T}{p}|y-y'|^2\Big)^{1/2} \leqslant \big(\tfrac{T}{p\varepsilon^2}\big)^{1/2}|y-y'|,$$

where $\mathbb{P}_y = \mathcal{E}(y)\cdot\mathbb{P}$ (and similarly for y'). Above, $\mathbb{P}_y(\cdot \,|\, \mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p})$ should be understood as a regular conditional probability of $\mathbb{P}_y$ on $(\Omega, \mathcal{F}^{p,\boldsymbol{W}}_{(\ell+1)T/p})$ given $\mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}$ (and similarly for y').

Recalling that $\widetilde{h}^{(p)}_{(\ell+1)T/p} + [T/p]Q^\dagger f(m^{(p)}_{(\ell+1)T/p})\mathbf{1}_{\{\ell\leqslant p-2\}}$ is bounded (by a constant independent of ε, ℓ and p), we deduce that there exists a constant c (depending on the same parameters as those quoted in the statement) such that, with probability 1 under $\mathbb{P}$,

$$|\Phi_\ell(y') - \Phi_\ell(y)| \leqslant c\tfrac{1}{\sqrt{p}\varepsilon}|y'-y|.$$

And then,

$$\mathbb{E}\big[|\Phi_\ell(y') - \Phi_\ell(y)|^2\big] \leqslant \tfrac{c^2}{p\varepsilon^2}\mathbb{E}\big[|y'-y|^2\big],$$

which proves that Φ_ℓ is a contraction in $L^2(\Omega, \mathcal{F}^{p,\boldsymbol{W}}_{\ell T/p}, \mathbb{P}; \mathbb{R}^d)$ when $\sqrt{p}\varepsilon > c$. By (backward) induction on ℓ (following the proof of Lemma 2.2.22), we know that each $\widetilde{h}^{(p)}_{\ell T/p}$ is bounded in L^∞ by a constant C (depending on the same parameters as those quoted in the statement).

Once $\widetilde{h}^{(p)}_{\ell T/p}$ has been found in (2.2.46), it remains to find the process $(k^{(p)}_s)_{\ell T/p\leqslant s\leqslant(\ell+1)T/p}$. This can be done by solving the following standard BSDE (with the pair process $(\boldsymbol{Y},\boldsymbol{Z}) = (Y_s,Z_s)_{\ell T/p\leqslant s\leqslant(\ell+1)T/p}$ as unknown, taking values in $\mathbb{R}^d\times\mathbb{R}^{d\times d}$):

$$\begin{aligned}\mathrm{d}Y_s &= -Q^\dagger f\big(m^{(p)}_{(\ell+1)T/p}\big)\mathbf{1}_{\{\ell\leqslant p-2\}}\mathrm{d}s + \Big(\eta^{(p)}_{(\ell+1)T/p} + \tfrac{T}{p}Q^\dagger Q\mathbf{1}_{\{\ell\leqslant p-2\}}\Big)\widetilde{h}^{(p)}_{\ell T/p}\mathrm{d}s\\ &\quad + Z_s\widetilde{h}^{(p)}_{\ell T/p}\mathrm{d}s + \varepsilon Z_s\mathrm{d}W_s,\end{aligned}$$

for $s\in[\ell T/p,(\ell+1)T/p]$, with the boundary condition $Y_{(\ell+1)T/p} = \widetilde{h}^{(p)}_{(\ell+1)T/p}$ (at time $(\ell+1)T/p$). Existence and uniqueness of a solution is standard because $\widetilde{h}^{(p)}_{\ell T/p}$ is in L^∞. Following the last sentence in Remark 2.2.16 (about measurability of the conditioning of an $\mathcal{F}^{p,\boldsymbol{W}}_{(\ell+1)T/p}$-measurable random variable given $\mathcal{F}^{\boldsymbol{W}}_{\ell T/p}$), we deduce that $Y_{\ell T/p}$ coincides with $\widetilde{h}^{(p)}_{\ell T/p}$ and eventually $(Z_s)_{\ell T/p\leqslant s\leqslant(\ell+1)T/p}$ solves the equation with $(k^{(p)}_s)_{\ell T/p\leqslant s\leqslant(\ell+1)T/p}$ as unknown in (2.2.46). Uniqueness of this choice can be easily checked by observing (from Itô's isometry) that, for another solution, say $(k^{(p),\prime}_s)_{\ell T/p\leqslant s\leqslant(\ell+1)T/p}$,

$$\begin{aligned}\varepsilon^2\mathbb{E}\int_{\ell T/p}^{(\ell+1)T/p}\big|k^{(p)}_s - k^{(p),\prime}_s\big|^2\mathrm{d}s &\leqslant C\mathbb{E}\bigg[\bigg|\int_{\ell T/p}^{(\ell+1)T/p}\big(k^{(p)}_s - k^{(p),\prime}_s\big)\mathrm{d}s\bigg|^2\bigg]\\ &\leqslant \tfrac{CT}{p}\mathbb{E}\bigg[\int_{\ell T/p}^{(\ell+1)T/p}\big|k^{(p)}_s - k^{(p),\prime}_s\big|^2\mathrm{d}s\bigg],\end{aligned}$$

for a possibly new choice of the constant C. Uniqueness easily follows if $CT/p<\varepsilon^2$. □

2.2.3 Exploitation versus exploration

We now address the exploitability, when we play the equilibrium strategy given by the fictitious play.

2.2.3.1 Approximate Nash equilibrium formed by the solution of the MFG with common noise

We first prove (which is easier) that the solution of the p-discrete MFG with common noise (as defined in the statement of Theorem 2.2.17 and in (2.2.52)–(2.2.55)) forms an approximate Nash equilibrium of the game without common noise. In order to do so, we recall from (2.2.53) that $\boldsymbol{\alpha}^{(p),\star}$ is the equilibrium strategy of the p-discrete MFG associated with the cost functional (2.2.52) and that the optimal trajectory is given by (2.2.54).

We claim:

Proposition 2.2.24. *There exists a constant C, only depending on the parameters d, T, $\|f\|_{1,\infty}$, $\|g\|_{1,\infty}$, $|Q|$, $|R|$ and $\mathbb{E}[|X_0|^2]$, such that, for any integer $p\geqslant 1$,*

$$\inf_{\boldsymbol{\alpha}}\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\boldsymbol{\alpha};\boldsymbol{m}^{(p)};0\big)\Big]\geqslant\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\boldsymbol{\alpha}^{(p),\star};\boldsymbol{m}^{(p)};0\big)\Big]-C\varepsilon,$$

the infimum being taken over $\mathbb{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}} = (\sigma(X_0,(B_{\tau_p(s)},W_{\tau_p(s)})_{s\leqslant t}))_{0\leqslant t\leqslant T}$-progressively measurable and square-integrable process $(\boldsymbol{\alpha}_t)_{0\leqslant t\leqslant T}$ that is piecewise constant on any interval $[\ell T/p,(\ell+1)T/p)$, for $\ell\in\{0,\cdots,p-1\}$.

We start with the following lemma:

Lemma 2.2.25. *There exists a constant C, only depending on d, T, $\|f\|_{1,\infty}$, $\|g\|_{1,\infty}$, $|Q|$ and $|R|$, such that, for any integer $p \geqslant 1$ and any $\mathbb{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}$-progressively measurable and square-integrable process $\boldsymbol{\alpha} = (\alpha_t)_{0\leqslant t\leqslant T}$ that is piecewise constant on any interval $[\ell T/p, (\ell+1)T/p)$, for $\ell \in \{0, \cdots, p-1\}$,*

$$\begin{aligned}&\left|\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}; \boldsymbol{m}^{(p)}; \varepsilon \boldsymbol{W}^{p,\boldsymbol{h}^{(p)}/\varepsilon}\big)\Big] - \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}; \boldsymbol{m}^{(p)}; 0\big)\Big]\right| \\ &\qquad \leqslant C\left(1 + \mathbb{E}\big[|X_0|^2\big]^{1/2} + \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\left[\int_0^T |\alpha_t|^2 \mathrm{d}t\right]^{1/2}\right)\varepsilon.\end{aligned}$$

The symbol $\mathbb{E}$ instead of $\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}$ for the L^2 moment of the initial condition is fully justified by the fact that X_0 is $\mathcal{F}_0$-measurable.

Proof. For the same initial condition X_0 and the same $\boldsymbol{\alpha}$ as in the statement, we let

$$\mathrm{d}X_t^{(p),\varepsilon} = \alpha_t \mathrm{d}t + \sigma \mathrm{d}B_t + \varepsilon \mathrm{d}W_t^{p,\boldsymbol{h}^{(p)}/\varepsilon}, \quad t \in [0,T].$$

In particular, we write $\boldsymbol{X}^{(p),0} = (X_t^{(p),0})_{0\leqslant t\leqslant T}$ when there is no common noise. Then, we obviously have

$$\sup_{0\leqslant t\leqslant T} |X_t^\varepsilon - X_t^0| \leqslant C\varepsilon \sup_{0\leqslant t\leqslant T} |W_t^{p,\boldsymbol{h}^{(p)}/\varepsilon}| \leqslant C\varepsilon \sup_{0\leqslant t\leqslant T} |W_t^{\boldsymbol{h}^{(p)}/\varepsilon}|.$$

Meanwhile, we observe that

$$\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\sup_{0\leqslant t\leqslant T}\big(|X_t^\varepsilon|^2 + |X_t^0|^2\big)\Big] \leqslant C\left(1 + \mathbb{E}\big[|X_0|^2\big] + \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\left[\int_0^T |\alpha_t|^2 \mathrm{d}t\right]\right).$$

Recalling (2.2.33), we get the result. □

The next lemma says that we can easily restrict ourselves to controls with a finite energy.

Lemma 2.2.26. *There exists a constant A, only depending on the parameters d, T, $\|f\|_{1,\infty}$, $\|g\|_{1,\infty}$, $|Q|$, $|R|$ and $\mathbb{E}[|X_0|^2]$ such that, for any $\varepsilon \in (0,1]$, any integer $p \geqslant 1$ and any $\mathbb{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}$-progressively measurable and square-integrable process $\boldsymbol{\alpha} = (\alpha_t)_{0\leqslant t\leqslant T}$ that is piecewise constant on any interval $[\ell T/p, (\ell+1)T/p)$, for $\ell \in \{0, \cdots, p-1\}$,*

$$\begin{aligned}&\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\left[\int_0^T |\alpha_t|^2 \mathrm{d}t\right] \geqslant A \\ &\quad \Rightarrow \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\boldsymbol{\alpha}; \boldsymbol{m}^{(p)}; 0\big)\Big] \geqslant \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(0; \boldsymbol{m}^{(p)}; \varepsilon \boldsymbol{W}^{p,\boldsymbol{h}^{(p)}/\varepsilon}\big)\Big] + 1.\end{aligned}$$

Proof. It suffices to observe that

$$\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\boldsymbol{\alpha}; \boldsymbol{m}^{(p)}; 0\big)\Big] \geqslant \frac{1}{2}\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\left[\int_0^T |\alpha_t|^2 \mathrm{d}t\right].$$

Meanwhile,

$$\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(0; \boldsymbol{m}^{(p)}; \varepsilon \boldsymbol{W}^{p,\boldsymbol{h}^{(p)}/\varepsilon}\big)\Big] \leqslant C.$$

The proof is easily completed. □

We now complete the

Proof of Proposition 2.2.24. By Lemma 2.2.26, we can reduce ourselves to the set $\mathcal{E}_A$ of processes $\boldsymbol{\alpha}$ (as in the statement of Proposition 2.2.24) such that

$$\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\int_0^T |\alpha_t|^2 \mathrm{d}t\Big] \leqslant A.$$

We then apply Lemma 2.2.25, which says that

$$\sup_{\boldsymbol{\alpha}\in\mathcal{E}_A}\Big|\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\boldsymbol{\alpha};\boldsymbol{m}^{(p)};\varepsilon\boldsymbol{W}^{p,\boldsymbol{h}^{(p)}/\varepsilon}\big)\Big] - \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\boldsymbol{\alpha};\boldsymbol{m};0\big)\Big]\Big| \leqslant C\varepsilon.$$

The proof is easily completed, using (2.2.52)–(2.2.53), from which we get that, for any $\boldsymbol{\alpha}\in\mathcal{E}_A$,

$$\begin{aligned}\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\boldsymbol{\alpha}^{(p),\star};\boldsymbol{m}^{(p)};\varepsilon\boldsymbol{W}^{p,\boldsymbol{h}^{(p)}/\varepsilon}\big)\Big] &\leqslant \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\boldsymbol{\alpha};\boldsymbol{m}^{(p)};\varepsilon\boldsymbol{W}^{p,\boldsymbol{h}^{(p)}/\varepsilon}\big)\Big]\\ &\leqslant \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\boldsymbol{\alpha};\boldsymbol{m}^{(p)};0\big)\Big] + C\varepsilon.\end{aligned}$$

This completes the proof. □

2.2.3.2 Approximate Nash equilibrium formed by the return of the fictitious play

We now state a similar result, but for the approximation returned by the fictitious play subjected to the ε-randomization and to the p-discretization. Throughout this paragraph, $\boldsymbol{h}^{n+1}$ and $\boldsymbol{m}^{n+1}$ are as in (2.2.43) and (2.2.44). We recall that these two processes depend on p (although it is not indicated in the notation).

The point is then to consider the strategy $\boldsymbol{\alpha}^{(p),n,\diamond}$ defined by

$$\alpha_t^{(p),n,\diamond} := -\eta^{(p)}_{\tau_p(t)} X^{(p),\star}_{\tau_p(t)} - \varpi h_t^n, \tag{2.2.78}$$

where we recall (see (2.2.54))

$$\begin{aligned}\mathrm{d}X_t^{(p),\star} &= -\eta^{(p)}_{\tau_p(t)} X^{(p),\star}_{\tau_p(t)}\mathrm{d}t + \sigma\mathrm{d}B_t + \varepsilon\mathrm{d}W_t^p\\ &= \alpha_t^{(p),n,\diamond}\mathrm{d}t + \sigma\mathrm{d}B_t + \varepsilon\mathrm{d}W_t^{p,\varpi\boldsymbol{h}^n/\varepsilon}, \quad t\in[0,T]; \quad X_0^{(p),n,\diamond} = X_0.\end{aligned} \tag{2.2.79}$$

Following (2.2.44), we have

$$\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[X_t^{(p),\star}\,|\,\sigma(\boldsymbol{W})\Big] = \mathbb{E}\Big[X_t^{(p),\star}\,|\,\sigma(\boldsymbol{W})\Big] = m_t^{(p)}, \quad t\in[0,T]. \tag{2.2.80}$$

In other words, the conditional state under the candidate strategy is the environment itself, which is a pre-requisite in mean field games. Notice that $\boldsymbol{\alpha}^{(p),n,\diamond}$ and $\boldsymbol{h}^n$ and do depend on ε.

Theorem 2.2.27. *Assume that the law of the initial condition has sub-Gaussian tails, i.e.* $\mathbb{P}(\{|X_0|\geqslant r\}) \leqslant a^{-1}\exp(-ar^2)$, *for some* $a>0$ *and for any* $r>0$. *Then, there exist two positive constants*

c and C, only depending on the parameters a, d, T, $\|f\|_{1,\infty}$, $\|g\|_{1,\infty}$, $|Q|$ and $|R|$, such that, for any $\varepsilon \in (0,1]$, any integer $p \geqslant 1$ satisfying $p\varepsilon^2 \geqslant c$ and any integer $n \geqslant 1$,

$$\inf_{\boldsymbol{\alpha}} \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}; \boldsymbol{m}^{(p)}; 0\big)\Big] \geqslant \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}^{(p),n,\diamond}; \boldsymbol{m}^{(p)}; 0\big)\Big] - C\varepsilon - Cp^{-1} - \exp(C\varepsilon^{-2})\varpi^{-n}, \tag{2.2.81}$$

the infimum in the left-hand side being taken over $\mathbb{F}^{X_0,\boldsymbol{B},\boldsymbol{W}}$-progressively measurable and square-integrable processes $(\boldsymbol{\alpha}_t)_{0\leqslant t\leqslant T}$.

The difference between the term in the left-hand side and the first term in the right-hand side of (2.2.81) *is called the $\mathbb{P}^{\varpi \boldsymbol{h}^n/\varepsilon}$-mean exploitability of the policy $\boldsymbol{\alpha}^{(p),n,\diamond}$. (Obviously, this difference is non-positive.)*

We feel useful to comment on the meaning and scope of Theorem 2.2.27. We first observe that, in the functional $\mathcal{R}$ that appears in $\mathcal{R}^{X_0}(\boldsymbol{\alpha}; \boldsymbol{m}^{(p)}; 0)$ and $\mathcal{R}^{X_0}(\boldsymbol{\alpha}^{(p),n,\diamond}; \boldsymbol{m}^{(p)}; 0)$, on both sides of the main inequality (2.2.81), there is no time discretization and no common noise (since the third input is 0). In other words, both the costs to $(\boldsymbol{\alpha}, \boldsymbol{m}^{(p)})$ and $(\boldsymbol{\alpha}^{(p),n,\diamond}, \boldsymbol{m}^{(p)})$ are computed according to the time-continuous original model without common noise (even though the control $\boldsymbol{\alpha}^{(p),n,\diamond}$ is piecewise constant and random as an output of the scheme). In particular, for a given realization of the common noise $\boldsymbol{W}$ (which does manifest here because $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^{(p),n,\diamond}$, $\boldsymbol{m}^{(p)}$ and $\boldsymbol{h}^n$ depend on $\boldsymbol{W}$), the conditional expectations $\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}[\mathcal{R}^{X_0}(\boldsymbol{\alpha}; \boldsymbol{m}^{(p)}; 0) \,|\, \sigma(\boldsymbol{W})]$ and $\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}[\mathcal{R}^{X_0}(\boldsymbol{\alpha}^{(p),n,\diamond}; \boldsymbol{m}^{(p)}; 0) \,|\, \sigma(\boldsymbol{W})]$ coincide with the costs $J(\boldsymbol{\alpha}(\cdot,\cdot,\boldsymbol{W}); \boldsymbol{m}^{(p)}(\boldsymbol{W}))$ and $J(\boldsymbol{\alpha}^{(p),n,\diamond}(\cdot,\cdot,\boldsymbol{W}); \boldsymbol{m}^{(p)}(\boldsymbol{W}))$, for J as in (2.1.2) (with the expectation in the latter being just performed over $(X_0, \boldsymbol{B})$). In particular, the notations $\boldsymbol{\alpha}(\cdot,\cdot,\boldsymbol{W})$, $\boldsymbol{\alpha}^{(p),n,\diamond}(\cdot,\cdot,\boldsymbol{W})$ and $\boldsymbol{m}^{(p)}(\boldsymbol{W})$ are used here to emphasize that, in the computation of J, the realization of the common noise is kept frozen in the inputs $\boldsymbol{\alpha}(\cdot,\cdot,\boldsymbol{W})$, $\boldsymbol{\alpha}^{(p),n,\diamond}(\cdot,\cdot,\boldsymbol{W})$ and $\boldsymbol{m}^{(p)}(\boldsymbol{W})$. In turn, the two expectations in (2.2.81) can be rewritten as $\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}[J(\boldsymbol{\alpha}(\cdot,\cdot,\boldsymbol{W}); \boldsymbol{m}^{(p)}(\cdot,\boldsymbol{W}))]$ and $\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}[J(\boldsymbol{\alpha}^{(p),n,\diamond}(\cdot,\cdot,\boldsymbol{W}); \boldsymbol{m}^{(p)}(\cdot,\boldsymbol{W}))]$, with the expectation $\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}$ being now taken under the sole common noise. This can be rephrased quite simply: Theorem 2.2.27 provides a bound for the $\mathbb{P}^{\varpi \boldsymbol{h}^n/\varepsilon}$-mean exploitability associated with the original MFG, the mean being here taken with respect to the tilted law of the common noise (i.e., the law of $\boldsymbol{W}$ under $\mathbb{P}^{\varpi \boldsymbol{h}^n/\varepsilon}$). The bound that is given for the mean exploitability depends on the three parameters ε, n and p. Typically, we want to choose ε small and p large, which is well understood: the equilibrium $(\boldsymbol{m}^{(p)}, \boldsymbol{\alpha}^{(p),n,\diamond})$ is learnt for the p-discrete MFG with an ε-common noise; if ε is large or p is small, the equilibrium that is learnt cannot be a 'good' approximate equilibrium of the original MFG. This is exactly what the terms $-C\varepsilon$ and $-C/p$ say in (2.2.81). As for the last term in (2.2.81), it becomes smaller and smaller as n increases. This is also consistent with the intuition: the mean exploitability becomes small when the number n of iterations of the fictitious play is large. For sure, there is a price to pay: as ε tends 0, the impact of the common noise becomes smaller and n has to be chosen larger to attain the same accuracy for the mean exploitability.

The reader must also realize that, differently from what is done in (2.2.32), the environment used in the cost functional is the conditional state of the reference particle when using the strategy $\boldsymbol{\alpha}^{(p),n,\diamond}$. This looks subtle, but this makes a big difference in the analysis. Indeed, if we had to follow (2.2.32) and thus use $\overline{\boldsymbol{m}}^n$ as environment in the cost functional, then the resulting minimizing path would NOT be typical of the environment, meaning that its (theoretical) conditional expectation (under $\mathbb{P}^{\varpi \boldsymbol{h}^n/\varepsilon}$ or $\mathbb{P}$) given the common noise would NOT match $\overline{\boldsymbol{m}}^n$. Differently, our choice to

use $\boldsymbol{m}^{(p)}$ as environment in the statement of Theorem 2.2.27 allows for the following property: the dynamics (2.2.79) are typical of the environment, meaning that the environment is exactly given by the conditional state given the realization of the common noise, which is exactly what (2.2.80) says. In this framework, Theorem 2.2.27 asserts that, by deviating unilaterally from the rest of the population, a reference player could hardly increase her cost, at best by a remainder that is small when p and n are large and ε is fixed.

We end-up this discussion with the following two observations, which may echo possible questions of the reader. First, one may wonder about the practical implementation of (2.2.79), as it requires to know the value of $\boldsymbol{\eta}^{(p)}$. In fact, as it is clarified in the next section, any efficient method that would be used in the implementation of the fictitious play should learn, at the same time, the solution to the Riccati equation (2.2.34), even though the coefficients f and g are not known. This is somehow a pre-requisite in the implementation of the fictitious play. Second, the fact that the dynamics (2.2.79) are independent of n (under the historical measure) may be rather intriguing. Of course, one must recall in this regard that, on a statistical point of view, the dynamics that truly matter are in fact those computed under the tilted probability measure. Indeed, the two costs in (2.2.81) are averaged under the tilted measure. Should we represent those costs in terms of an infinite cloud of particles, then all the particles forming the approximate Nash equilibrium provided by Theorem 2.2.27 should be subjected to a common noise that would be sampled under the tilted probability measure. This is exactly the point where the dependence on n manifests.

We start with the following refinement of Theorem 2.2.17:

Lemma 2.2.28. *With the same notation as in the statement of Theorem 2.2.17, there exist two positive constants c and C, only depending on the parameters a, d, T, $\|f\|_{1,\infty}$, $\|g\|_{1,\infty}$, $|Q|$ and $|R|$, such that, for any $\varepsilon \in (0,1]$, any integer $p \geqslant 1$ satisfying $p\varepsilon^2 \geqslant c$ and any integer $n \geqslant 1$,*

$$\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\sup_{0\leqslant t\leqslant T} |\alpha_t^{(p),n,\diamond} - \alpha_t^{(p),\star}|^2\Big] \leqslant \varpi^{-2n}\exp(C\varepsilon^{-2}),$$

where we recall that $\boldsymbol{\alpha}^{(p),\star}$ is the equilibrium strategy of the p-discrete MFG with an ε common noise, as given by (2.2.53).

Proof of Lemma 2.2.28. We recall from (2.2.53) and (2.2.78) that

$$\alpha_t^{(p),n,\diamond} - \alpha_t^{(p),\star} = \varpi h_t^n - h_t^{(p)}, \quad t \in [0,T],$$

and the bound in the statement follows from Theorem 2.2.17. □

We now turn to

Proof of Theorem 2.2.27.

First Step. Recalling the shape of the optimizer $\boldsymbol{\alpha}^{(p),n+1,\diamond}$ in (2.2.78)–(2.2.79) , together with the fact that $\boldsymbol{h}^{n+1}$ is bounded, uniformly in $n \geqslant 1$, see Lemma 2.2.22 (recalling that C_1 therein is independent of n, as mentioned in the statement), we deduce that

$$\Big|\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}^{(p),n,\diamond};\boldsymbol{m}^{(p)};0\big)\Big]\Big| \leqslant C,$$

where the constant C is independent n. In the rest of the proof, the value of C may vary from line to line provided it remains independent of n. Then, we must invoke the fact that the problem

$$\inf_{\boldsymbol{\alpha}} \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}; \boldsymbol{m}^{(p)}; 0\big)\Big]$$

is a linear quadratic-problem in a random environment. Similar to the minimization problem studied in Lemma 2.2.1, its solution is given in feedback form. If we call $\widehat{\boldsymbol{\alpha}}^{n+1}$ the minimizer, it reads

$$\widehat{\alpha}_t^n = -\big(\eta_t \widehat{X}_t^n + \widehat{h}_t^n\big), \quad t \in [0,T], \tag{2.2.82}$$

with

$$\mathrm{d}\widehat{X}_t^n = -\big(\eta_t \widehat{X}_t^n + \widehat{h}_t^n\big)\mathrm{d}t + \sigma \mathrm{d}B_t, \quad t \in [0,T]; \quad \widehat{X}_0^n = X_0, \tag{2.2.83}$$

and where

$$\begin{aligned} \mathrm{d}\widehat{h}_t^n &= \Big(-Q^\dagger f\big(m_t^{(p)}\big) + \eta_t \widehat{h}_t^n\Big)\mathrm{d}t + \varepsilon \widehat{k}_t^n \mathrm{d}W_t^{\varpi \boldsymbol{h}^n/\varepsilon}, \quad t \in [0,T], \\ \widehat{h}_T^n &= R^\dagger g\big(m_T^{(p)}\big). \end{aligned} \tag{2.2.84}$$

We start with a series of results whose proof is given in the last step below. We first claim that, for each $k \in \{0, \cdots, p\}$, $\widehat{h}_{kT/p}^n$ is $\mathcal{F}_{kT/p}^{\boldsymbol{W}^p}$-measurable. Moreover,

$$\mathrm{essup}_{\omega \in \Omega} \sup_{0 \leqslant t \leqslant T} |\widehat{h}_t^n| \leqslant C. \tag{2.2.85}$$

As a result, with probability 1 under $\mathbb{P}^{\varpi \boldsymbol{h}^n/\varepsilon}$ (and also under $\mathbb{P}$),

$$\begin{aligned} \sup_{0 \leqslant t \leqslant T} |\widehat{X}_t^n| &\leqslant C\Big(1 + |X_0| + \sup_{0 \leqslant t \leqslant T} |B_t|\Big), \\ \sup_{0 \leqslant t \leqslant T} |\widehat{\alpha}_t^n| &\leqslant C\Big(1 + |X_0| + \sup_{0 \leqslant t \leqslant T} |B_t|\Big), \end{aligned} \tag{2.2.86}$$

which implies

$$\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\sup_{0 \leqslant t \leqslant T} \big|\widehat{X}_t^n\big|^2\Big] \leqslant C. \tag{2.2.87}$$

and

$$\mathcal{R}^{X_0}\big(\widehat{\boldsymbol{\alpha}}^n; \boldsymbol{m}^{(p)}; 0\big) \leqslant C\Big(1 + |X_0|^2 + \sup_{0 \leqslant t \leqslant T} |B_t|^2\Big). \tag{2.2.88}$$

Second Step. By the first step,

$$\inf_{\boldsymbol{\alpha}} \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}; \boldsymbol{m}^{(p)}; 0\big)\Big] = \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\widehat{\boldsymbol{\alpha}}^n; \boldsymbol{m}^{(p)}; 0\big)\Big].$$

Here,

$$\begin{aligned} &\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\widehat{\boldsymbol{\alpha}}^n; \boldsymbol{m}^{(p)}; 0\big)\Big] \\ &= \tfrac{1}{2}\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\bigg[|R^\dagger \widehat{X}_T^n + g(m_T^{(p)})|^2 + \int_0^T \Big\{|Q^\dagger \widehat{X}_t^n + f(m_t^{(p)})|^2 + |\widehat{\alpha}_t^n|^2\Big\}\mathrm{d}t\bigg], \end{aligned}$$

and the purpose is to lower bound the above cost by $\mathcal{R}^{p,X_0}(\widehat{\boldsymbol{\alpha}}^{(p),n};\boldsymbol{m}^{(p)};0)$ for some well-chosen $\mathbb{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}$-progressively measurable and piecewise constant control $\widehat{\boldsymbol{\alpha}}^{(p),n}$ and with $\mathcal{R}^{p,X_0}$ as in (2.2.33). In order to do so, we let

$$\widehat{h}^{(p),n}_{\tau_p(t)} := \tfrac{p}{T}\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\int_{\tau_p(t)}^{\tau_p(t)+T/p} \widehat{h}^n_s \mathrm{d}s \,|\, \mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{\tau_p(t)}\Big], \quad t\in[0,T], \tag{2.2.89}$$

which allows us to define by induction:

$$\widehat{X}^{(p),n}_{(k+1)T/p} := \widehat{X}^{(p),n}_{kT/p} - \tfrac{T}{p}\eta_{kT/p}\widehat{X}^{(p),n}_{kT/p} - \tfrac{T}{p}\widehat{h}^{(p),n}_{kT/p} + \sigma\big(B_{(k+1)T/p} - B_{kT/p}\big), \tag{2.2.90}$$

for $k\in\{0,\cdots,p-1\}$ (with X_0 as initial condition). We prove in the fourth step below that

$$\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\sup_{k=0,\cdots,p}\big|\widehat{X}^n_{kT/p} - \widehat{X}^{(p),n}_{kT/p}\big|^2\Big] \leqslant \tfrac{C}{p}, \tag{2.2.91}$$

and

$$\sup_{0\leqslant t\leqslant T}\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\big|\widehat{X}^n_t - \widehat{X}^{(p),n}_{\tau_p(t)}\big|^2\Big] \leqslant \tfrac{C}{p}, \tag{2.2.92}$$

for a constant C as in the statement of Theorem 2.2.27. By (2.2.40), we then also have

$$\sup_{0\leqslant t\leqslant T}\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\big|\eta_t\widehat{X}^n_t - \eta^{(p)}_{\tau_p(t)}\widehat{X}^{(p),n}_{\tau_p(t)}\big|^2\Big] \leqslant \tfrac{C}{p}, \tag{2.2.93}$$

Together with (2.2.87), we deduce from the latter displays that

$$\begin{aligned}
&\tfrac{1}{2}\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[|Q^\dagger\widehat{X}^n_t + f(m^{(p)}_t)|^2\Big]\\
&\geqslant \tfrac{1}{2}\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[|Q^\dagger\widehat{X}^{(p),n}_{\tau_p(t)} + f\big(m^{(p)}_{\tau_p(t)}\big)|^2\Big]\\
&\quad - \mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\Big(|Q^\dagger\big(\widehat{X}^n_t - \widehat{X}^{(p),n}_{\tau_p(t)}\big)| + |f(m^{(p)}_t) - f(m^{(p)}_{\tau_p(t)})|\Big)\,|Q^\dagger\widehat{X}^{(p),n}_{\tau_p(t)} + f(m^{(p)}_{\tau_p(t)})|\Big]\\
&\geqslant \tfrac{1}{2}\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[|Q^\dagger\widehat{X}^{(p),n}_{\tau_p(t)} + f(m^{(p)}_{\tau_p(t)})|^2\Big] - \tfrac{C}{\sqrt{p}}.
\end{aligned} \tag{2.2.94}$$

Proceeding similarly for the other terms in the cost functional, we obtain

$$\begin{aligned}
\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\widehat{\boldsymbol{\alpha}}^n;\boldsymbol{m}^{(p)};0\big)\Big] \geqslant \tfrac{1}{2}\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[&|R^\dagger\widehat{X}^{(p),n}_T + g(m^{(p)}_T)|^2\\
&+ \int_0^T\Big\{|Q^\dagger\widehat{X}^{(p),n}_{\tau_p(t)} + f(m^{(p)}_{\tau_p(t)})|^2 + |\eta^{(p)}_{\tau_p(t)}\widehat{X}^{(p),n}_{\tau_p(t)} + \widehat{h}^n_t|^2\Big\}\mathrm{d}t\Big] - \tfrac{C}{\sqrt{p}},
\end{aligned}$$

and by conditioning the last term in the integral by $\mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{\tau_p(t)}$, we deduce from (2.2.89) and Jensen's inequality that

$$\begin{aligned}
\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\widehat{\boldsymbol{\alpha}}^n;\boldsymbol{m}^{(p)};0\big)\Big] \geqslant \tfrac{1}{2}\mathbb{E}^{\varpi\boldsymbol{h}^n/\varepsilon}\Big[&|R^\dagger\widehat{X}^{(p),n}_T + g(m^{(p)}_T)|^2\\
&+ \int_0^T\Big\{|Q^\dagger\widehat{X}^{(p),n}_{\tau_p(t)} + f(m^{(p)}_{\tau_p(t)})|^2 + |\widehat{\alpha}^{(p),n}_t|^2\Big\}\mathrm{d}t\Big] - \tfrac{C}{\sqrt{p}},
\end{aligned} \tag{2.2.95}$$

where

$$\widehat{\alpha}_t^{(p),n} = -\eta^{(p)}_{\tau_p(t)} \widehat{X}^{(p),n}_{\tau_p(t)} - \widehat{h}^{(p),n}_{\tau_p(t)},$$

for $t \in [0,T]$. Clearly, the above conditioning is licit because the process $(\widehat{X}^{(p),n}_{kT/p})_{k=0,\cdots,p}$ is $(\mathcal{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}_{kT/p})_{k=0,\cdots,p}$-progressively measurable. By the way, we deduce from the latter that the process $(\widehat{\alpha}_t^{(p),n})_{0\leqslant t\leqslant T}$ is $\mathbb{F}^{p,X_0,\boldsymbol{B},\boldsymbol{W}}$-progressively measurable, square-integrable and piecewise constant on any interval $[\ell T/p, (\ell+1)T/p)$, for $\ell \in \{0,\cdots,p-1\}$. In turn, the right-hand side in (2.2.95) can be rewritten

$$\mathbb{E}^{\varpi \boldsymbol{h}^{n}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\widehat{\boldsymbol{\alpha}}^{(p),n};\boldsymbol{m}^{(p)};0\big)\Big] - \tfrac{C}{\sqrt{p}}.$$

We deduce that, for any $\varrho > 1$,

$$\begin{aligned}
&\inf_{\boldsymbol{\alpha}} \mathbb{E}^{\varpi \boldsymbol{h}^{n}/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha};\boldsymbol{m}^{(p)};0\big)\Big] \\
&\quad \geqslant \mathbb{E}^{\varpi \boldsymbol{h}^{n}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\widehat{\boldsymbol{\alpha}}^{(p),n};\boldsymbol{m}^{(p)};0\big)\mathbf{1}_{\{\mathcal{R}^{p,X_0}(\widehat{\boldsymbol{\alpha}}^{(p),n};\boldsymbol{m}^{(p)};0)\leqslant\varrho\}}\Big] - \tfrac{C}{\sqrt{p}}.
\end{aligned}$$

Using the upper bound on $d_{\mathrm{TV}}\big(\mathbb{P}^{\boldsymbol{h}^{(p)}/\varepsilon}, \mathbb{P}^{\varpi \boldsymbol{h}^{n}/\varepsilon}\big)$ (see Remark 2.2.20), we obtain

$$\begin{aligned}
&\inf_{\boldsymbol{\alpha}} \mathbb{E}^{\varpi \boldsymbol{h}^{n}/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha};\boldsymbol{m}^{(p)};0\big)\Big] \\
&\geqslant \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\widehat{\boldsymbol{\alpha}}^{(p),n};\boldsymbol{m}^{(p)};0\big)\mathbf{1}_{\{\mathcal{R}^{p,X_0}(\widehat{\boldsymbol{\alpha}}^{(p),n};\boldsymbol{m}^{(p)};0)\leqslant\varrho\}}\Big] - \varrho\exp(C\varepsilon^{-2})\varpi^{-n} - \tfrac{C}{\sqrt{p}}.
\end{aligned}$$

And then, by (2.2.88) and the sub-Gaussian property of the law of X_0, we obtain

$$\begin{aligned}
\inf_{\boldsymbol{\alpha}} \mathbb{E}^{\varpi \boldsymbol{h}^{n}/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha};\boldsymbol{m}^{(p)};0\big)\Big] &\geqslant \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\widehat{\boldsymbol{\alpha}}^{(p),n};\boldsymbol{m}^{(p)};0\big)\Big] \\
&\quad - C\exp\big(-\eta\varrho\big) - \varrho\exp(C\varepsilon^{-2})\varpi^{-n} - \tfrac{C}{\sqrt{p}},
\end{aligned} \tag{2.2.96}$$

for some $\eta > 0$. Finally, by Proposition 2.2.24,

$$\begin{aligned}
&\inf_{\boldsymbol{\alpha}} \mathbb{E}^{\varpi \boldsymbol{h}^{n}/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha};\boldsymbol{m}^{(p)};0\big)\Big] \\
&\quad \geqslant \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\boldsymbol{\alpha}^{(p),\star};\boldsymbol{m}^{(p)};0\big)\Big] - C\varepsilon - C\exp\big(-\eta\varrho\big) - \rho\exp(C\varepsilon^{-2})\varpi^{-n} - \tfrac{C}{\sqrt{p}}.
\end{aligned} \tag{2.2.97}$$

Third Step. We now revert the computations achieved in the second step. Following (2.2.95) and then (2.2.94) and the proof of Lemma 2.2.25, the cost on the last line can be lower bounded as follows:

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{p,X_0}\big(\boldsymbol{\alpha}^{(p),\star};\boldsymbol{m}^{(p)};0\big)\Big] \\
&= \tfrac{1}{2}\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[|R^\dagger X_T^{(p),\star,0} + g(m_T^{(p)})|^2 + \int_0^T \Big\{|Q^\dagger X^{(p),\star,0}_{\tau_p(t)} + f(m^{(p)}_{\tau_p(t)})|^2 + |\alpha_t^{(p),\star}|^2\Big\}\mathrm{d}t\Big] \\
&\geqslant \tfrac{1}{2}\mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[|R^\dagger X_T^{(p),\star} + g(m_T^{(p)})|^2 + \int_0^T \Big\{|Q^\dagger X_t^{(p),\star} + f(m_t^{(p)})|^2 + |\alpha_t^{(p),\star}|^2\Big\}\mathrm{d}t\Big] \\
&\quad - \tfrac{C}{\sqrt{p}} - C\varepsilon,
\end{aligned}$$

where $\boldsymbol{X}^{(p),\star,0}$ on the second line denotes the time-continuous continuous process defined by

$$X_t^{(p),\star,0} = X_{kT/p}^{(p),\star,0} + \big(t - \tfrac{kT}{p}\big)\alpha_{kT/p}^{(p),\star} + \sigma\big(B_t - B_{kT/p}\big), \quad t \in [\tfrac{kT}{p}, \tfrac{(k+1)T}{p}], \quad k \in \{0, \cdots, p-1\}.$$

In particular, $\boldsymbol{X}^{(p),\star,0}$ is the time-continuous process driven by $\boldsymbol{\alpha}^{(p),\star}$ when there is no common noise. Notice that $\boldsymbol{X}^{(p),\star,0}$ differs from the process $\boldsymbol{X}^{(p),\star}$ that appears on the third line and that is defined in (2.2.54).

In turn, by Proposition 2.2.28 and for ρ as in (2.2.97),

$$\begin{aligned}
&\inf_{\boldsymbol{\alpha}} \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}; \boldsymbol{m}^{(p)}; 0\big)\Big] \\
&\geqslant \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}^{(p),\star}; \boldsymbol{m}^{(p)}; 0\big)\Big] - C\varepsilon - C\exp\big(-\eta\varrho\big) - \rho\exp(C\varepsilon^{-2})\varpi^{-n} - \tfrac{C}{\sqrt{p}} \\
&\geqslant \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}^{(p),n,\diamond}; \boldsymbol{m}^{(p)}; 0\big)\Big] - C\varepsilon - C\exp\big(-\eta\varrho\big) - \rho\exp(C\varepsilon^{-2})\varpi^{-n} - \tfrac{C}{\sqrt{p}} \\
&\geqslant \mathbb{E}^{\boldsymbol{h}^{(p)}/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}^{(p),n,\diamond}; \boldsymbol{m}^{(p)}; 0\big)\mathbf{1}_{\{\mathcal{R}^{X_0}(\boldsymbol{\alpha}^{(p),n,\diamond}; \boldsymbol{m}^{(p)}; 0) \leqslant \varrho\}}\Big] - C\varepsilon - C\exp\big(-\eta\varrho\big) \\
&\quad - \rho\exp(C\varepsilon^{-2})\varpi^{-n} - \tfrac{C}{\sqrt{p}}.
\end{aligned}$$

And then, as in the derivation of (2.2.96),

$$\begin{aligned}
&\inf_{\boldsymbol{\alpha}} \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}; \boldsymbol{m}^{(p)}; 0\big)\Big] \\
&\geqslant \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}^{(p),n,\diamond}; \boldsymbol{m}^{(p)}; 0\big)\mathbf{1}_{\{\mathcal{R}^{X_0}(\boldsymbol{\alpha}^{(p),n,\diamond}; \boldsymbol{m}^{(p)}; 0) \leqslant \varrho\}}\Big] - C\varepsilon - C\exp\big(-\eta\varrho\big) \\
&\quad - \rho\exp(C\varepsilon^{-2})\varpi^{-n} - \tfrac{C}{\sqrt{p}} \\
&\geqslant \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\mathcal{R}^{X_0}\big(\boldsymbol{\alpha}^{(p),n,\diamond}; \boldsymbol{m}^{(p)}; 0\big)\Big] - C\varepsilon - C\exp\big(-\eta\varrho\big) - \rho\exp(C\varepsilon^{-2})\varpi^{-n} - \tfrac{C}{\sqrt{p}}.
\end{aligned}$$

Choosing $\eta\varrho = -\ln(\varepsilon)$ and modifying the value of C, we complete the proof.

Fourth Step. It now remains to prove some auxiliary results that are used in the previous steps.

We start with the analysis of $\widehat{\boldsymbol{h}}^n$ in (2.2.84). We recall from (2.2.43) that, that, for each $t \in [0,T]$, h_t^n is $(\sigma((W_k^p)_{k \leqslant \lfloor tp/T \rfloor T/p}))_{0 \leqslant t \leqslant T}$-measurable. Hence, from (2.2.78) and (2.2.79), each $m_t^{(p)}$ is $(\sigma((W_k^p)_{k \leqslant \lfloor tp/T \rfloor T/p}))_{0 \leqslant t \leqslant T}$-measurable (notice the use of the ceil part instead of the floor part in the time index of the filtration). Further, we notice from (2.2.84) and Remark 2.2.7 that

$$P_{kT/p}\widehat{h}_{kT/p}^n = \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\bigg[P_T R^\dagger g\big(m_T^{(p)}\big) + \int_{kT/p}^T P_s Q^\dagger f\big(m_t^{(p)}\big)\mathrm{d}t \,|\, \mathcal{F}_{kT/p}^{\boldsymbol{W}}\bigg], \quad k \in \{0, \cdots, p\}.$$

Following the analysis of (2.2.42), the left-hand side is $\mathcal{F}_{kT/p}^{\boldsymbol{W}^p}$-measurable. Noting that each P_t, for $t \in [0,T]$, is invertible, we deduce that $\widehat{h}_{kT/p}^n$ is $\mathcal{F}_{kT/p}^{\boldsymbol{W}^p}$-measurable, which property is used below.

And then, (2.2.85) follows from the above representation of $(P_{kT/p}\widehat{h}_{kT/p}^n)_{k=0,\cdots,p}$ (with a similar representation of $P_t\widehat{h}_t^n$ for any $t \in [0,T]$). The two bounds in (2.2.86) easily follow.

We now turn to the proof of (2.2.91) and (2.2.92). By (2.2.90) , we know that

$$\widehat{X}_{(k+1)T/p}^{(p),n} = \widehat{X}_{kT/p}^{(p),n} - \tfrac{T}{p}\eta_{kT/p}\widehat{X}_{kT/p}^{(p),n} - \tfrac{T}{p}\widehat{h}_{kT/p}^{(p),n} + \sigma\big(B_{(k+1)T/p} - B_{kT/p}\big), \quad k \in \{0, \cdots, p-1\}.$$

Meanwhile, recalling (2.2.83), we write

$$\begin{aligned}\widehat{X}^n_{(k+1)T/p} &= \widehat{X}^n_{kT/p} - \int_{kT/p}^{(k+1)T/p} \big[\eta_s \widehat{X}^n_s + \widehat{h}^n_s\big]\mathrm{d}s + \sigma\big(B_{(k+1)T/p} - B_{kT/p}\big) \\ &= \widehat{X}^n_{kT/p} - \bigg(\tfrac{T}{p}\eta_{kT/p}\widehat{X}^n_{kT/p} + \int_{kT/p}^{(k+1)T/p} \widehat{h}^n_s \mathrm{d}s\bigg) - \tfrac{T}{p}\widehat{A}^n_k + \sigma\big(B_{(k+1)T/p} - B_{kT/p}\big),\end{aligned}$$

with

$$\widehat{A}^n_k = \tfrac{p}{T}\int_{kT/p}^{(k+1)T/p} \big[\eta_s \widehat{X}^n_s - \eta_{kT/p}\widehat{X}^n_{kT/p}\big]\mathrm{d}s. \tag{2.2.98}$$

And then,

$$\begin{aligned}\widehat{X}^{(p),n}_{(k+1)T/p} - \widehat{X}^n_{(k+1)T/p} = \widehat{X}^{(p),n}_{kT/p} - \widehat{X}^n_{kT/p} - \tfrac{T}{p}\eta_{kT/p}\big(\widehat{X}^{(p),n}_{kT/p} - \widehat{X}^n_{kT/p}\big) \\ - \int_{kT/p}^{(k+1)T/p} \big[\widehat{h}^{(p),n}_{\tau_p(s)} - \widehat{h}^n_s\big]\mathrm{d}s + \tfrac{T}{p}\widehat{A}^n_k.\end{aligned}$$

By discrete Gronwall's lemma, there exists a constant C (depending on the same parameters as in the statement of Theorem 2.2.27) such that

$$\sup_{\ell=0,\cdots,p} \big|\widehat{X}^{(p),n}_{\ell T/p} - \widehat{X}^{n+1}_{\ell T/p}\big| \leqslant C\sum_{\ell=0}^{p-1}\bigg(\bigg|\int_{\ell T/p}^{(\ell+1)T/p} \big[\widehat{h}^{(p),n}_{\tau_p(s)} - \widehat{h}^n_s\big]\mathrm{d}s\bigg| + \tfrac{T}{p}\big|\widehat{A}^n_\ell\big|\bigg).$$

And then,

$$\begin{aligned}&\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\sup_{k=0,\cdots,p}\big|\widehat{X}^{(p),n}_{kT/p} - \widehat{X}^n_{kT/p}\big|^2\Big] \\ &\leqslant C\tfrac{T}{p}\sum_{\ell=0}^{p-1}\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\bigg[\tfrac{p^2}{T^2}\bigg|\int_{\ell T/p}^{(\ell+1)T/p}\big[\widehat{h}^{(p),n}_{\tau_p(s)} - \widehat{h}^n_s\big]\mathrm{d}s\bigg|^2 + \big|\widehat{A}^n_\ell\big|^2\bigg].\end{aligned} \tag{2.2.99}$$

Here, we observe from (2.2.89) that

$$\begin{aligned}&\tfrac{T}{p}\sum_{\ell=0}^{p-1}\tfrac{p^2}{T^2}\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\bigg[\bigg|\int_{\ell T/p}^{(\ell+1)T/p}\big[\widehat{h}^{(p),n}_{\tau_p(s)} - \widehat{h}^n_s\big]\mathrm{d}s\bigg|^2\bigg] \\ &= \tfrac{p}{T}\sum_{\ell=0}^{p-1}\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\bigg[\bigg|\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\bigg[\int_{\ell T/p}^{(\ell+1)T/p}\widehat{h}^n_s\mathrm{d}s \,|\, \mathcal{F}^{\boldsymbol{W}^p}_{\ell T/p}\bigg] - \int_{\ell T/p}^{(\ell+1)T/p}\widehat{h}^n_s\mathrm{d}s\bigg|^2\bigg] \\ &\leqslant \sum_{\ell=0}^{p-1}\int_{\ell T/p}^{(\ell+1)T/p}\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\big|\widehat{h}^n_s - \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\big[\widehat{h}^n_s \,|\, \mathcal{F}^{\boldsymbol{W}^p}_{\ell T/p}\big]\big|^2\Big]\mathrm{d}s.\end{aligned} \tag{2.2.100}$$

By (2.2.84), we have, for $\ell \in \{0,\cdots,p-1\}$ and for $s \in [\ell T/p, (\ell+1)T/p]$,

$$\widehat{h}^n_s = \widehat{h}^n_{\ell T/p} + \int_{\ell T/p}^s \big[-Q^\dagger f\big(m^{(p)}_r\big) + \eta_r \widehat{h}^n_r\big]\mathrm{d}r + \varepsilon\int_{\ell T/p}^s \widehat{k}^n_r \mathrm{d}W^{\varpi \boldsymbol{h}^n/\varepsilon}_r.$$

Taking the conditional expectation given $\mathcal{F}^{\boldsymbol{W}^p}_{\ell T/p}$ under $\mathbb{P}^{\varpi \boldsymbol{h}^n/\varepsilon}$, we obtain

$$\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\big[\widehat{h}^n_s \,|\, \mathcal{F}^{\boldsymbol{W}^p}_{\ell T/p}\big] = \widehat{h}^n_{\ell T/p} + \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\bigg[\int_{\ell T/p}^s \big[-Q^\dagger f\big(m^{(p)}_r\big) + \eta_r \widehat{h}^n_r\big]\mathrm{d}r \,|\, \mathcal{F}^{\boldsymbol{W}^p}_{\ell T/p}\bigg],$$

and then

$$\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\big|\widehat{h}^n_s - \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\big[\widehat{h}^n_s \,|\, \mathcal{F}^{\boldsymbol{W}^p}_{\ell T/p}\big]\big|^2\Big] \leqslant \tfrac{C}{p^2} + C\varepsilon^2 \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\bigg(\int_{\ell T/p}^s |\widehat{k}^n_r|^2 \mathrm{d}r\bigg).$$

Finally, by (2.2.100),

$$\begin{aligned}
&\tfrac{T}{p}\sum_{\ell=0}^{p-1} \tfrac{p^2}{T^2}\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\bigg[\bigg|\int_{\ell T/p}^{(\ell+1)T/p} \big[\widehat{h}^{(p),n}_{\tau_p(s)} - \widehat{h}^n_s\big]\mathrm{d}s\bigg|^2\bigg] \\
&\leqslant \tfrac{C}{p^2} + \sum_{\ell=0}^{p-1}\int_{\ell T/p}^{(\ell+1)T/p} \mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\bigg(\int_{\ell T/p}^s \varepsilon^2 |\widehat{k}^n_r|^2 \mathrm{d}r\bigg)\,\mathrm{d}s \\
&\leqslant \tfrac{C}{p^2} + \tfrac{C}{p}\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\bigg(\int_0^T \varepsilon^2 |\widehat{k}^n_r|^2 \mathrm{d}r\bigg).
\end{aligned}$$

Taking the square in (2.2.84) and using (2.2.85), we can prove that

$$\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\bigg(\int_0^T \varepsilon^2 |\widehat{k}^n_r|^2 \mathrm{d}r\bigg) \leqslant C.$$

Back to (2.2.98) and (2.2.99), we get

$$\mathbb{E}^{\varpi \boldsymbol{h}^n/\varepsilon}\Big[\sup_{k=0,\cdots,p} \big|\widehat{X}^{(p),n}_{kT/p} - \widehat{X}^n_{kT/p}\big|^2\Big] \leqslant \tfrac{C}{p},$$

which is (2.2.91).

By (2.2.83), we easily obtain (2.2.92). □

2.3 Numerical experiments

This section is devoted to several numerical experiments based on our learning algorithm. We first provide in Subsection 2.3.1 a theoretical analysis of a benchmark example that we use throughout the section. In Subsection 2.3.2, we present several variants of the implemented version of the algorithm. We also explain how to compute a reference solution. Then, we study in Subsection 2.3.3 the numerical behavior of the algorithm when the intensity ε of the common noise is equal to 1. In Subsection 2.3.4, we explain some of the difficulties that arise in the large dimension setting. Finally, in Subsection 2.3.5, we address the small viscosity regime. In both Subsections 2.3.3 and 2.3.5, we discuss the influence of the parameter ϖ. In particular, we compare the two harmonic ($\varpi = 1$) and geometric ($\varpi > 1$) versions of the algorithm. In this regard, it is worth recalling Remark 2.2.6: all our results hold for the harmonic version of the fictitious play with the geometric decay ϖ^{-n} being replaced by $1/n$.

2.3.1 A benchmark example

As a particular instance of the cost functional J in (2.1.2), we focus here on

$$J(\boldsymbol{\alpha};\boldsymbol{m}) = \frac{1}{2}\mathbb{E}\bigg[\big|X_T + g(m_T)\big|^2 + \int_0^T |\alpha_t|^2 \mathrm{d}t\bigg], \tag{2.3.1}$$

which implicitly means that $f = 0$. The examples that are addressed below are in dimension $d = 1$, $d = 2$, $d = 12$ and $d = 20$, with $T = 1$ in any cases. For simplicity, we also work with $X_0 = 0$.

In dimension $d = 1$, we choose

$$g(x) = \cos(\kappa x), \quad x \in \mathbb{R}, \tag{2.3.2}$$

for a free positive parameter κ that we tune from smaller to larger values. Obviously, the Lipschitz constant of g becomes larger with κ. Accordingly, the coupling between the two forward and backward equations in (2.1.7) becomes stronger as κ gets larger, which makes it more difficult to solve, especially when $\varepsilon = 0$. We illustrate this in Lemma 2.3.1 below: it says that, when $\varepsilon = 0$, (2.1.7) is uniquely solvable when $|\kappa| < 2$. Even more, the analysis performed in [38] shows that the more standard (and obviously simpler) Picard scheme (2.1.9)–(2.1.10) would converge when κ is small, whether there is a common noise or not. Our result is thus especially relevant when κ gets larger.

In dimension 2, we work with a similar terminal cost $g = (g_1, g_2)$:

$$g_1(x_1, x_2) = \cos(\kappa x_1)\cos(\kappa x_2), \quad g_2(x_1, x_2) = \sin(\kappa x_1)\sin(\kappa x_2), \quad x_1,\ x_2 \in \mathbb{R}, \tag{2.3.3}$$

where, as in dimension 1, κ is a free positive parameter that we tune from smaller to larger values.

As we just said, we also address higher dimensional examples, with $d = 12$ or $d = 20$. In order to encode the function g in a systematic manner, we then choose:

$$g_i(x) = \cos\bigg(\sum_{j=1}^{d} \Theta_{i,j} x_j\bigg) \quad x \in \mathbb{R}^d, \quad i = 1, \cdots, d, \tag{2.3.4}$$

for a fixed square matrix Θ of size $d \times d$, whose choice depends on the examples. For instance, we may take $\Theta_{i,j} = (i + j)/(2d)$. The impact of Θ is then quite similar to the impact of κ in the low dimensional case as it induces highly oscillatory phenomena, which make the coupling in (2.1.7) stronger.

Regardless of the choice of g, the Riccati equation (2.2.3) associated with (2.3.1) writes

$$\dot{\eta}_t - \eta_t^2 = 0, \quad t \in [0, 1]\ ; \quad \eta_1 = 1, \tag{2.3.5}$$

the solution of which is given by

$$\eta_t = \frac{1}{2 - t}, \quad t \in [0, 1]. \tag{2.3.6}$$

Obviously, the solution to (2.2.3) is here identified with a scalar-valued function while, formally, it takes values in the set of square $d \times d$ matrices. This follows from the fact that Q and R in (2.2.3) are just the identity matrix. Numerically, this makes both the code and the analysis slightly easier, but the resulting restriction on the scope of the results is in fact limited, the real challenge being to learn $\boldsymbol{h}^n$ in (2.2.32).

2.3.1.1 Solutions without common noise

Even though this is not needed for all the examples we handle below, we feel useful to have a preliminary discussion about the shape of the solutions when there is no common noise, keeping in mind that, originally, the true model of interest is the MFG without common noise. Recalling (2.1.7), we indeed have that, whenever there is no common noise (i.e., $\varepsilon = 0$), the equilibria are given as the solutions of the deterministic system

$$\dot{m}_t = -\big(\eta_t m_t + h_t\big), \quad \dot{h}_t = \eta_t h_t, \quad t \in [0,1]\ ; \quad h_1 = g(m_1), \tag{2.3.7}$$

with $m_0 = 0$, which prompts us to perform the changes of variable:

$$\widetilde{m}_t = \frac{m_t}{2-t}, \quad \widetilde{h}_t = (2-t)h_t, \quad t \in [0,1]. \tag{2.3.8}$$

We then have that $(m_t, h_t)_{0 \leqslant t \leqslant 1}$ solves (2.3.7) if and only if

$$\dot{\widetilde{m}}_t = -\frac{1}{2-t} h_t = -\frac{1}{(2-t)^2}\widetilde{h}_t, \quad \dot{\widetilde{h}}_t = 0, \quad t \in [0,1]\ ; \quad \widetilde{h}_1 = g\big(\widetilde{m}_1\big),$$

from which we deduce the following simpler characterization (recalling that $m_0 = 0$)

$$\widetilde{m}_1 = -g\big(\widetilde{m}_1\big) \int_0^1 \frac{\mathrm{d}t}{(2-t)^2} = -\frac{g(\widetilde{m}_1)}{2}. \tag{2.3.9}$$

There are as many equilibria as solutions to the equation $x = -g(x)/2$. We thus have the following obvious lemma:

Lemma 2.3.1. *When $\varepsilon = 0$ and regardless of the dimension, the solutions of the MFG associated with the cost functional* (2.3.1) *are given by the roots $\widetilde{m}_1$ of the equation $2x + g(x) = 0$ and then by the changes of variable* (2.3.8)*. In particular, if the Lipschitz constant of the function g is strictly less than 2, then the MFG has one and only one solution.*

2.3.1.2 Potential structure when $\varepsilon = 0$

Following (2.1.13) and (2.1.14), we may associate a mean field control problem with the MFG (with $\varepsilon = 0$) if the function g derives from a potential G. The cost functional is given by

$$\mathcal{J}(\boldsymbol{\alpha}) = \mathbb{E}\bigg[\frac{1}{2}|X_1|^2 + G\big(\mathbb{E}(X_1)\big) + \frac{1}{2}\int_0^1 |\alpha_t|^2 \mathrm{d}t\bigg],$$

where, in the right-hand side,

$$X_1 = \int_0^1 \alpha_t \mathrm{d}t.$$

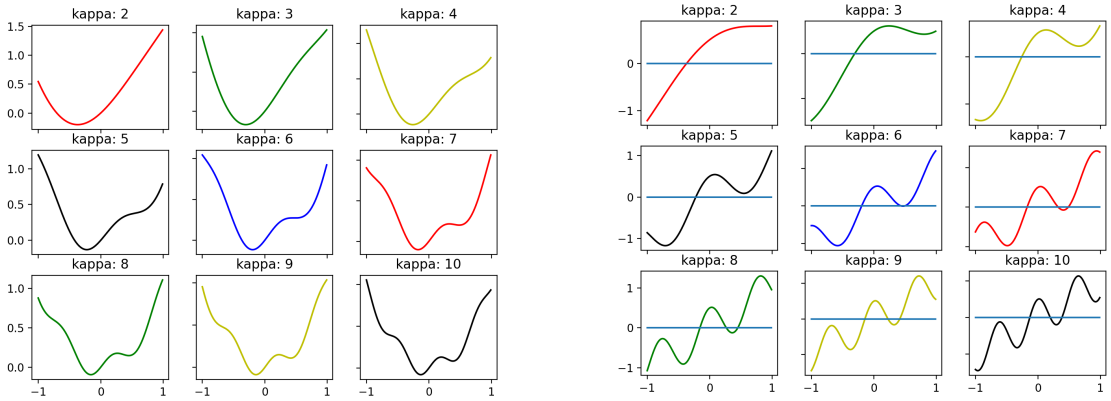
The analysis of the minimization problem $\inf_{\boldsymbol{\alpha}} \mathcal{J}(\boldsymbol{\alpha})$ is straightforward. Indeed, by an obvious convexity argument, we observe that

$$\mathcal{J}(\boldsymbol{\alpha}) \geqslant \mathcal{J}\bigg(\mathbb{E}\int_0^1 \alpha_t \mathrm{d}t\bigg),$$

where, in the right-hand side, it is obviously understood that the argument of $\mathcal{J}$ is a constant control. This shows that the minimizers are deterministic and constant in time. Using the generic notation β for $\int_0^1 \alpha_t dt$, the problem is thus to minimize

$$\mathscr{J}(\beta) = \beta^2 + G(\beta), \quad \beta \in \mathbb{R}.$$

Of course, we observe that any minimizer of $\mathscr{J}$ is also a solution of the equation $\beta = -g(\beta)/2$: we recover the fact that any solution to the mean field control problem is an MFG solution. Obviously, the converse may not be true. Accordingly, the set of solutions to the mean field control problem may be strictly included in the set of equilibria of the corresponding mean field game. This principle is illustrated by Figure 2.5 below with g as in (2.3.2): For all the values of $\kappa \in \{2, 3, \cdots, 10\}$, the potential $\mathscr{J}$ has a unique minimizer; but, for $\kappa \in \{7, 8, 9, 10\}$, the derivative of $\mathscr{J}$ has several zeros, hence proving that the MFG has several solutions.



(a) Evolution of $\mathscr{J}$ with κ.

(b) Evolution of $\frac{\mathrm{d}}{\mathrm{d}\beta}\mathscr{J}$ with κ.

Figure 2.5: Minimizers of the mean field control problem and equilibria of the mean field game for $\kappa \in \{2, \cdots, 10\}$ when $d = 1$ and $g(x) = \cos(\kappa x)$: Solutions of the mean field control problem are the minimizers of $\mathcal{J}$; Solutions of the mean field game are the roots of $\frac{\mathrm{d}}{\mathrm{d}\beta}\mathscr{J}$.

Interestingly, the recent contribution [37] says that, when the MFG is potential, the equilibria that are not associated with a minimizer of the potential should be ruled out. Equivalently, only the MFG solutions that minimize (globally) the potential should be selected. Even more, the guess (which has been rigorously established in [37, 45] but in different settings that the one addressed here) is that the minimizers of the potential are precisely those that appear by forcing uniqueness with a common noise and then by tuning down the intensity of the common noise. We check this prediction numerically in Subsection 2.3.5, by using our learning method.

2.3.2 Algorithms for a fixed intensity of the common noise

We here explain how to implement the fictitious play in the two benchmark examples (2.3.2) and (2.3.3) along the lines of Figure 2.3. Throughout the subsection, the intensity of the common noise

is fixed. For simplicity, the intensity of the common noise is chosen as $\varepsilon = 1$ and the intensity of the independent noise is chosen as $\sigma = 0$ or 1. Numerical experiments are discussed in the next subsection.

2.3.2.1 Numerical reference solution

We first explain how to compute a reference solution by solving numerically the decoupled forward-backward equation (2.2.11). Observe that there is no contradiction in using a numerical method for testing our learning algorithm: the numerical method relies on the explicit knowledge of the coefficient g, whereas the learning algorithm is intended to work on the sole observations of the costs.

The numerical method we use is tailor-made to our problem. Indeed, we observe that the solution to the backward equation in (2.2.15) writes

$$h_t = \mathbb{E}^{\boldsymbol{h}}\left[\exp\left(-\int_t^1 \eta_s \mathrm{d}s\right) g\big(m_1\big) \,|\, \mathcal{F}_t^{\boldsymbol{W}}\right], \quad t \in [0,1], \tag{2.3.10}$$

where $(m_t)_{0\leqslant t\leqslant 1}$ is an Ornstein-Uhlenbeck process (with independent coordinates)

$$\mathrm{d}m_t = -\eta_t m_t \mathrm{d}t + \mathrm{d}W_t, \quad t \in [0,1]\ ; \quad m_0 = 0. \tag{2.3.11}$$

We then employ a Picard scheme for solving the fixed point equation (2.3.10), by iterating

$$h_t^{\mathrm{ref},n+1} = \mathbb{E}^{\boldsymbol{h}^{\mathrm{ref},n}}\left[\exp\left(-\int_t^1 \eta_s \mathrm{d}s\right) g\big(m_1\big) \,|\, \mathcal{F}_t^{\boldsymbol{W}}\right], \quad t \in [0,1], \tag{2.3.12}$$

Numerically, we use a time grid $\{t_k = k/p\}_k$ with p uniform steps (for the same p as the one used in the learning method, which makes the comparison easier) and, following the seminal work of [60], we employ a regression method in order to approximate the conditional expectation appearing in the right-hand side. Precisely, we find, at each iteration n of the Picard sequence and at each node $t_j > 0$ of the time mesh, an approximation of the conditional expectation in the right-hand side of (2.3.12) in the form of a (deterministic) function of m_{t_j} (the forward component in (2.3.12)), with the deterministic function being chosen in a given class of functions from $\mathbb{R}^d$ into itself. The point is then to choose a convenient class within which the regression is achieved.

Here, we address two approaches, already reported in the literature. The first one consists in performing regression on linear combinations of Hermite polynomials. It is inspired from [20], which offers (in a somewhat more complicated framework) some theoretical bounds on the regression error. The second approach is taken from the work [53], which paved the way for a systematic use of neural networks in the computation of numerical solutions to nonlinear PDEs and related Markovian BSDEs. The point in this second approach is thus to approximate the conditional expectation in (2.3.12), at time $t = t_j$, by means of a neural network with m_{t_j} as entry.

For the sake of completeness, we provide of a short overview of the shape of the functions within these two classes.

Using Hermite polynomials. The real-valued components of the regression functions are taken in the linear span of $\{H_\ell^d(\cdot/\sigma_{t_j})\}_{|\ell|\leqslant D}$, where σ_{t_j} is the common standard deviation of the coordinates

of m_{t_j} (which is explicitly computable) and $\{H^d_\ell(\cdot)\}_{|\ell|\leqslant D}$ is the collection of Hermite polynomials of dimension d ($d = 1, 2$ in our case) and of degree less than D, for a given integer D (here ℓ is a d-tuple of integers and $|\ell|$ is the sum of its entries). We recall that, when $d = 1$,

$$H^1_\ell(x) = \frac{(-1)^\ell}{\sqrt{2^\ell \ell!}} e^{x^2} \frac{\mathrm{d}}{\mathrm{d}x^\ell}[e^{-x^2}], \quad x \in \mathbb{R}.$$

When $d = 2$,

$$H^2_{(\ell_1,\ell_2)}(x_1, x_2) = H^1_{\ell_1}(x_1) H^1_{\ell_2}(x_2), \quad x_1, x_2 \in \mathbb{R}.$$

Using a neural network. For a number of layers L, the functions used for the regression have the form $\psi^{L+1} \circ \varphi^L \circ \psi^L \cdots \circ \psi^2 \circ \varphi^1 \circ \psi^1$, with ψ^ℓ, for each $\ell = 1, \cdots, L+1$, being an affine function from $\mathbb{R}^{d_{\ell-1}}$ to $\mathbb{R}^{d_\ell}$, where $d_0 = d_{L+1} = d$. For each $\ell = 1, \cdots, L$, φ^ℓ is a function from $\mathbb{R}^{d_\ell}$ into itself that maps $(x_1, \cdots, x_{d_\ell})$ onto $(\varsigma(x_1), \cdots, \varsigma(x_{d_\ell}))$ for some activation function ς (which, for simplicity, is taken to be the same for any ℓ). The activation function is fixed *a priori*. Standard examples for it are the ReLu or Sigmoid functions. We then call neurons of the networks the coefficients encoding the linear mappings $(\psi^\ell)_{\ell=1,\cdots,L+1}$. In clear, the principle is to tune, for each ℓ, matrices $A^\ell \in \mathbb{R}^{d_{\ell-1} \times d_\ell}$ and vectors $B^\ell \in \mathbb{R}^{d_\ell}$ in the decomposition of ψ_ℓ as

$$\psi^\ell(x) = A^\ell x + B^\ell, \quad x \in \mathbb{R}^{d_{\ell-1}}.$$

Neural networks are notoriously known to (possibly) behave well in higher dimension. Below, we provide examples with $d = 12$ and $d = 20$.

Given a finite dimensional class $\mathcal{C}$ of regression functions (obtained by Hermite polynomials or neural networks), we use the following algorithm for the the computation of a reference solution.

Algorithm 1. [Reference solution]

Input: Introduce $\Delta_{t_k} w^{(j)}$ with $j \in \{1, \cdots, N\}$ and $k \in \{1, \cdots, p\}$, a collection of N realizations of independent Gaussian variables $\mathcal{N}(0, I_d)$.

Task: For each i, compute $(m^{(j)}_{t_k})_{j=0,\cdots,p}$ the realizations of the Euler scheme associated with (2.3.11) and driven by the simulations

$$\left(w^{(j)}_{t_k} = \frac{1}{\sqrt{p}} \Big[\Delta_{t_1} w^{(j)} + \cdots + \Delta_{t_k} w^{(j)} \Big] \right)_{k=1,\cdots,p}.$$

Loop: One iteration consists of one Picard iteration. The input at iteration number n is encoded in the form of an array $(h^{n,(j)}_{t_k})_{j,k}$ with entries in $\mathbb{R}^d$. The following procedure returns the next input for iteration number $n + 1$.

(a) Look for $(h^{n+1,(j)}_{t_k})_{j,k}$ in terms of $(h^{n,(j)}_{t_k})_{j,k}$ by solving, for each $k = 0, \cdots, p-1$, the minimization problem

$$\min_{\mathfrak{h} \in \mathcal{C}} \frac{1}{N} \sum_{j=1}^N \mathcal{E}^{n,(j)}_k \left| \exp\Big(- \sum_{s=k}^{p-1} \eta_{t_s} \Big) g(m^{(j)}_1) - \mathfrak{h}\big(m^{(j)}_{t_k}\big) \right|^2,$$

with

$$\mathcal{E}_k^{n,(j)} = \exp\Big(- \sum_{s=k}^{p-1} h_{t_s}^{n,(j)} \cdot \Delta_{t_s} w^{(j)} - \frac{1}{2p} \sum_{s=k}^{p-1} |h_{t_s}^{n,(j)}|^2 \Big).$$

(b) To enforce some stability, update each coordinate of $(h_{t_k}^{n+1,(j)})_{j,k}$ by projecting it onto $[-1, 1]$.

Remark 2.3.2. *For instance, when Hermite polynomials are used, the minimization step in Algorithm 1 can be rewritten as*

$$\min_{\boldsymbol{c}=(c_\ell)_{|\ell|\leqslant D}} \frac{1}{N} \sum_{j=1}^{N} \mathcal{E}_k^{n,(j)} \Big| \exp\Big(- \sum_{s=k}^{p-1} \eta_{t_s} \Big) g(m_1^{(j)}) - \sum_{|\ell|\leqslant D} c_\ell H_\ell^d\big(\tfrac{1}{\sigma_{t_k}} m_{t_k}^{(j)}\big) \Big|^2,$$

where each c_ℓ in $\boldsymbol{c} = (c_\ell)_{|\ell|\leqslant D}$ is in $\mathbb{R}^d$. (When $k = 0$, only $c_{\mathbf{0}}$, where $|\mathbf{0}| = 0$, matters and it is given by a mere empirical mean.)

Here, the coordinates in (2.3.11) *have the same marginal variances because $(\eta_t)_{0\leqslant t\leqslant T}$ is scalar-valued (or equivalent is a d-diagonal matrix). This is no longer true when R and Q in* (2.1.2) *are general matrices. In such a case (and more generally when the components are not centered), the components of $(m_t)_{0\leqslant t\leqslant T}$ are no longer independent and the 'normalized' Hermite polynomials $\{H_\ell^d(\cdot/\sigma_{t_k})\}_{|\ell|\leqslant D}$ used above should be instead replaced by $\{H_\ell^d(\mathbb{K}^{-1/2}(m_{t_k})(\cdot - \mathbb{E}(m_{t_k}))\}_{|\ell|\leqslant D}$, where $\mathbb{K}(m_{t_j})$ is the covariance matrix of m_{t_k} and $\mathbb{K}^{1/2}(m_{t_k})$ is any root of it (e.g., as given by the Cholesky decomposition).*

2.3.2.2 Numerical approximation by fictitious play

We now address the implementation of the fictitious play according to the principle stipulated by Figure 2.3. The black-box therein is discretized in a suitable manner. It is used to solve numerically the optimization problem (2.2.32), and then to implement the updating rule (2.2.45) for the environment. Since Lemma 2.2.13 says that the corresponding optimal law has an affine Markov feedback form, we can perform the optimization in (2.2.32) over closed loop controls that are affine in the state variable, with the linear coefficient being a scalar only depending on time and with the intercept possibly depending on the environment. From the practical point of view, this choice is fully justified if the model is expected to be linear-quadratic in the space/action variables (as it is in (2.1.1)–(2.1.2), with Q and R being scalars), but this does not require the coefficients Q, R, f and g to be known explicitly. Mathematically, this writes as follows. We can restrict ourselves to controls $\boldsymbol{\alpha} = (\alpha_{t_k})_{k=1,\cdots,p}$ of the form

$$\alpha_{t_k} = a_{t_{k-1}} X_{t_{k-1}} + C_{t_{k-1}}, \tag{2.3.13}$$

where $a_{t_{k-1}}$ is a scalar and $C_{t_{k-1}}$ is a d-dimensional random vector (which is typically adapted to the increments $(\Delta_{t_\ell} w)_{\ell\leqslant k-1}$, but the measurability properties of which are specified in a finer manner right below). In comparison with the formula (2.2.37), $a_{t_{k-1}}$ should be understood as a proxy for $-\eta_{t_{k-1}}$ and $C_{t_{k-1}}$ as a proxy for the intercept therein. Part of the numerical difficulty is thus to have a tractable regression method for capturing the randomness of $C_{t_{k-1}}$. Recalling that the intercept

$\boldsymbol{h}$ in the solution of the mean field game with common noise is a function of the environment $(m_t)_{0\leqslant t\leqslant 1}$, see (2.2.18), a key point in our numerical method is to look for $C_{t_{k-1}}$ as a $\sigma(\overline{m}_{t_{k-1}})$-measurable random variable, where $\overline{m}_{t_{k-1}}$ is the current proxy for the state of the environment at time t_{k-1}. Numerically, we use a regression method, very much in the spirit of Algorithm 1: either we use a regression based on a d-dimensional linear combination of Hermite polynomials of degree less than D, for a fixed value of D, or we use a neural network with L layers and with a prescribed numbers of neurons. Of course, the reader may wonder about another approach to (2.3.13), in which we optimize over general (possibly non-affine) controls in Markov feedback form. In fact, while this seems an attractive way to bypass any *a priori* knowledge about the shape of the model, this approach requires an extra step for updating the intercept $\boldsymbol{h}$ at the next iteration of the fictitious play. Intuitively, the new intercept value can be found by linearly regressing the controls on the states. At this point, however, we assert that the presence of this additional step in the numerical procedure can only be fully justified if the linear-quadratic structure of the model is known. This makes the advantage of not postulating (2.3.13) very limited.

We now make this construction explicit when $\sigma = \varepsilon = 1$ and the postulate (2.3.13) is in force, noticing that the algorithm must rely on simulations for the independent and common noises. Throughout, the learning parameter ϖ is arbitrary: it may be strictly greater than 1 (in which case we are using the geometric variant of the fictitious play) or equal to 1 (in which case we are using the harmonic variant). Below, we call M the number of particles (given a realization of the common noise) and N the number of simulations of the whole system (or equivalently of the common noise), with i denoting the generic label for a particle and j denoting the generic label for a realization. We are thus given a collection

$$\Delta_{t_k} b^{(i,j)},\ \Delta_{t_k} w^{(j)}, \quad i \in \{1,\cdots,M\},\ j \in \{1,\cdots,N\},\ k \in \{1,\cdots,p\}, \tag{2.3.14}$$

of realizations of independent $\mathcal{N}(0, I_d)$ Gaussian variables. At iteration number n of the fictitious play, the current proxy for the environment is thus given in the form of a collection $(\overline{m}_{t_k}^{n,(j)})_{j,k}$, with j running from 1 to N and k running from 0 to p. Similarly, the proxy for the intercept in the optimal law (2.2.37) is given in the form of a collection $(h_{t_k}^{n,(j)})_{j,k}$, also with j running from 1 to N and k running from 0 to $p-1$. The fact that the two proxies are independent of i is fully consistent with the fact that $\overline{\boldsymbol{m}}^n$ and $\boldsymbol{h}^n$ in Remark 2.2.16 are adapted to the realization of the common noise. Following our introductory discussion, we solve, as an approximation of the stochastic control problem (2.2.32), the minimization problem:

$$\min\left\{\frac{1}{N}\sum_{j=1}^{N}\left(\mathcal{E}^{n,(j)}\frac{1}{M}\sum_{i=1}^{M}\left[\frac{1}{2p}\sum_{k=1}^{p}\varpi^2|\alpha_{t_k}^{(i,j)}|^2+\frac{1}{2}\left|\varpi x_1^{(i,j)}+g\left(\overline{m}_1^{n,(j)}\right)\right|^2\right]\right)\right\}, \tag{2.3.15}$$

where

$$\mathcal{E}^{n,(j)} = \exp\left(-\varpi\sqrt{\frac{1}{p}}\sum_{k=0}^{p-1} h_{t_k}^{n,(j)}\cdot\Delta_{t_{k+1}} w^{(j)} - \frac{\varpi^2}{2p}\sum_{k=0}^{p-1}|h_{t_k}^{n,(j)}|^2\right).$$

Moreover, in the minimization problem, $\alpha_{t_k}^{(i,j)}$, $x_{t_k}^{(i,j)}$ are required to be of the form

$$\begin{aligned} x_{t_k}^{(i,j)} &= x_{t_{k-1}}^{(i,j)} + \frac{1}{p}\alpha_{t_k}^{(i,j)} + \frac{1}{\varpi\sqrt{p}}\Delta_{t_k} b^{(i,j)}, \quad \ell = 1,\cdots,p\ ; \quad x_0^{(i,j)} = x_0, \\ \alpha_{t_k}^{(i,j)} &= a_{t_{k-1}} x_{t_{k-1}}^{(i,j)} + C_{t_{k-1}}^{(j)} + \varpi h_{t_{k-1}}^{n,(j)} + \sqrt{p}\Delta_{t_k} w^{(j)}, \quad k = 1,\cdots,p. \end{aligned} \tag{2.3.16}$$

In particular, $(x_{t_k}^{(i,j)})_{k=0,\cdots,p}$ solves the following Euler scheme:

$$x_{t_k}^{(i,j)} = x_{t_{k-1}}^{(i,j)} + \frac{1}{p}\Big(a_{t_{k-1}} x_{t_{k-1}}^{(i,j)} + C_{t_{k-1}}^{(j)} + \varpi h_{t_{k-1}}^{n,(j)}\Big) + \frac{1}{\varpi\sqrt{p}}\Delta_{t_k} b^{(i,j)} + \frac{1}{\sqrt{p}}\Delta_{t_k} w^{(j)}, \quad k = 1, \cdots, p.$$

Moreover, $C_{t_k}^{(j)}$ is required to be of the form

$$C_{t_k}^{(j)} = \mathfrak{h}_{t_k}\big(\overline{m}_{t_k}^{n,(j)}\big), \tag{2.3.17}$$

for a function $\mathfrak{h}_{t_k}$ within one of the two classes $\mathcal{C}$ described in Subsection 2.3.2.1, depending on whether we use Hermite polynomials or neural networks. As already explained, neural networks may provide better results in higher dimension (or, to put it differently, Hermite polynomials may be of an intractable complexity in higher dimension).

As before, we feel useful to exemplify (2.3.17) within each of the two aforementioned case.

Using Hermite polynomials. In this case, we use a slightly modified form of (2.3.17), as we also center $\overline{m}_{t_k}^{n,(j)}$ by means of the empirical mean in the argument of the Hermite polynomials. Definition (2.3.17) thus becomes:

$$C_{t_k}^{(j)} = \sum_{|\ell| \leqslant D} c_{t_k}(\ell) H_\ell^d\bigg(\big(U_{t_k}^n\big)^{-1}\Big(\overline{m}_{t_k}^{n,(j)} - \frac{1}{N}\sum_{r=1}^N \overline{m}_k^{n,(r)}\Big)\bigg), \tag{2.3.18}$$

where $c_k(\ell) \in \mathbb{R}^d$ and $U_{t_k}^n$ is the diagonal matrix obtained by taking the roots of the diagonal of the empirical covariance matrix:

$$\Sigma_{t_k}^n = \frac{1}{N}\sum_{j=1}^N \Big(\overline{m}_{t_k}^{n,(j)} - \frac{1}{N}\sum_{r=1}^N \overline{m}_{t_k}^{n,(r)}\Big) \otimes \Big(\overline{m}_{t_k}^{n,(j)} - \frac{1}{N}\sum_{r=1}^N \overline{m}_{t_k}^{n,(r)}\Big).$$

Remark 2.3.3. *Following Remark 2.3.2, we can define $U_{t_k}^n$ (in a more general fashion) as the upper triangular matrix given by the Cholesky decomposition of $\Sigma_{t_k}^n$ in the form $\Sigma_{t_k}^n = (U_{t_k}^n)^\dagger U_{t_k}^n$.*

The minimization problem in (2.3.15) is thus defined (when using Hermite polynomials) over $\boldsymbol{a} = (a_{t_k})_k \in \mathbb{R}^p$ and $\boldsymbol{c} = (c_{t_k}(\ell))_{k,\ell} \in (\mathbb{R}^d)^{p\times L}$, where L is the number of distinct d-tuples of (non-negative) integers whose sum is less than D.

Using neural networks. With the same definition as in Subsection 2.3.2.1, (2.3.17) takes the form

$$C_{t_k}^{(j)} = \psi_{t_k}^{L+1} \circ \varphi_{t_k}^L \circ \psi_{t_k}^L \cdots \circ \psi_{t_k}^2 \circ \varphi_{t_k}^1 \circ \psi_{t_k}^1\big(\overline{m}_{t_k}^{n,(j)}\big). \tag{2.3.19}$$

Typically, the number of layers L, the dimensions of the various layers and the activation functions are the same for any time t_k, $k = 0, \cdots, N-1$. In other words, we can write $\varphi_{t_k}^L = \varphi^L$, $\varphi_{t_k}^{L-1} = \varphi^{L-1}$, $\cdots$. Each $\psi_{t_k}^\ell$, for $\ell = 1, \cdots, L+1$, writes as

$$\psi_{t_k}^\ell(x) = A_k^\ell x + B_k^\ell, \quad x \in \mathbb{R}^{d_{\ell-1}}.$$

The minimization problem in (2.3.15) is thus defined over $\boldsymbol{a} = (a_{t_k})_k \in \mathbb{R}^p$, $\boldsymbol{A} = (A_k^\ell)_{k,\ell}$ and $\boldsymbol{B} = (B_k^\ell)_{k,\ell}$, with $A_k^\ell \in \mathbb{R}^{d_{\ell-1}\times d_\ell}$ and $B_k^\ell \in \mathbb{R}^{d_\ell}$.

We can summarize the algorithm in the following form. It has the same form whether we work with the harmonic ($\varpi = 1$) or geometric ($\varpi > 1$) version of the fictitious play.

Algorithm 2. [Learning with two noises, $\sigma = \varepsilon = 1$]

Input: Take as input the realizations (2.3.14), which are the same at any iteration of the algorithm.

Loop: At rank $n+1$ of the fictitious play:

(a) Take as input the proxies $(\overline{m}_{t_k}^{n,j})_{j,k}$ and $(h_{t_k}^{n,j})_{j,k}$ for (respectively) the environment and the intercept.

(b) Solve the minimization problem (2.3.15) over $\boldsymbol{a} = (a_{t_k})_k$ in (2.3.16) and $\mathfrak{h} = (\mathfrak{h}_{t_k})_k$ in (2.3.17). Call $\boldsymbol{a}^{n+1} = (a_{t_k}^{n+1})_k$ and $\mathfrak{h}^{n+1} = (\mathfrak{h}_{t_k}^{n+1})_k$ the optimal points (returned by any optimization method).

(c) With $\boldsymbol{a}^{n+1} = (a_{t_k}^{n+1})_k$ and $\mathfrak{h}^{n+1} = (\mathfrak{h}_{t_k}^{n+1})_k$, associate $(x_{t_k}^{n+1,(i,j)})_{i,j,k}$ as in (2.3.16) and $(C_{t_k}^{n+1,(j)})_{k,j}$ as in (2.3.17).

(d) Update the proxies by letting

$$m_{t_k}^{n+1,(j)} = \frac{1}{M}\sum_{i=1}^{M} x_{t_k}^{n+1,(i,j)}, \quad \overline{m}_{t_k}^{n+1,(j)} = \pi_{n+1}(\varpi) \times m_{t_k}^{n+1,(j)} + \big(1 - \pi_{n+1}(\varpi)\big) \times \overline{m}_{t_k}^{n,(j)},$$
$$h_{t_k}^{n+1,(j)} = -C_{t_k}^{n+1,(j)},$$

where $\pi_n(\varpi) = \begin{cases} 1/n & \text{if } \varpi = 1, \\ \frac{\varpi^{-(n-1)}(1-\varpi^{-1})}{1-\varpi^{-n}} & \text{if } \varpi > 1. \end{cases}$

Remark 2.3.4. *The construction of Algorithm 2 relies on the various noises* (2.3.14)*. In fact, it is worth mentioning that the algorithm can be reformulated in a similar manner when the increments* $(\Delta_{t_k} b^{(i,j)})_{i,j,k}$ *defining the idiosyncratic noises are assumed to just depend on* (i,k) *(equivalently they are equal for the same value of* i *but for two different values of* j*). As before, the resulting random variables* $(\Delta_{t_k} b^{(i)})_{i=1,\cdots,M,k=1,\cdots,p}$ *are required to be independent and identically distributed.*

Obviously, the smallest family $(\Delta_{t_k} b^{(i)})_{i=1,\cdots,M,k=1,\cdots,p}$ *is of a cheapest numerical cost. However, it creates additional correlations between the particles: for two different indices* j *and* j'*, the corresponding two empirical means over* $i \in \{1,\cdots,N\}$ *in* (2.3.15) *are dependent, with the correlations vanishing asymptotically as* N *tends to* ∞*. Even though we do not provide any bound for the numerical error in this section, it is worth mentioning that the presence of additional correlations would make the Monte-Carlo error harder to estimate. As reported below, we tested the two approaches. For the range of parameters we used, we have not seen any major difference.*

In some of the numerical examples below, we will use variants of Algorithm 2. We first focus on the shape of the algorithm when $\sigma = 0$, recalling that our results do not require the presence of an idiosyncratic noise. Implicitly, we then have a single particle only, i.e. $M = 1$. Moreover, $\Delta_{t_k} b^{(i,j)}$ no longer appears in (2.3.16). We then have the following variant of Algorithm 2, which is of lower complexity:

Algorithm 3. [Learning with common noise only, $\sigma = 0$, $\varepsilon = 1$]

Input: Take as input the realizations $(\Delta_{t_k} w^{(j)})_{j,k}$ in (2.3.14), which are the same at any iteration of the algorithm.

Loop: At rank $n+1$ of the fictitious play, apply the same loop as in Algorithm 2, but with $M = 1$ and with $\Delta_{t_k} b^{(1,j)} = 0$ in (2.3.16).

We end up our presentation with the case when there is no common noise, i.e. $\sigma = 1$ and $\varepsilon = 0$. In that case, our strategy no longer applies. The point is thus to implement the standard version of the fictitious play. Accordingly, there is no Girsanov density in (2.3.15) and $\alpha_{t_k}^{(i,j)}$ in (2.3.16) merely writes

$$\alpha_{t_k}^{(i,j)} = a_{t_{k-1}} x_{t_{k-1}}^{(i,j)} + C_{t_{k-1}}, \quad k = 1, \cdots, p.$$

with C_{t_k} being a deterministic d-dimensional vector (which is independent of j). In particular, there is no need to use the proxy $(h_{t_k}^{n,j})_{j,k}$ for the intercept. Equivalently, we can assume that $N = 1$ and $\mathfrak{h}_{t_k}$ in (2.3.17) to be a mere constant function (i.e., the function $\mathfrak{h}_{t_k}$ is identified with $\mathfrak{h}_{t_k}(0)$). The rest of the algorithm is similar. It may be written as follows.

Algorithm 4. [Learning with idiosyncratic noise only, $\sigma = 1$, $\varepsilon = 0$]

Input: Take as input the realizations $(\Delta_{t_k} b^{(i,1)})_k$ in (2.3.14), which are the same at any iteration of the algorithm.

Loop: At rank $n+1$ of the fictitious play:

(a) Take as input the proxy $(\overline{m}_{t_k}^n)_k$ for the environment and the intercept.

(b) Solve the minimization problem (2.3.15) over $\boldsymbol{a} = (a_{t_k})_k$ in (2.3.16) and $\boldsymbol{\mathfrak{h}} = (\mathfrak{h}_{t_k})$ in (2.3.17), with $\mathcal{E}^{n,(j)} = 1$ in (2.3.15) and with the last line in (2.3.16) being replaced by

$$\alpha_{t_k}^{(i,1)} = a_{t_{k-1}} x_{t_{k-1}}^{(i,1)} + \mathfrak{h}_{t_{k-1}}(0), \quad k = 1, \cdots, p.$$

Call $\boldsymbol{a}^{n+1} = (a_{t_k}^{n+1})_k$ and $\boldsymbol{\mathfrak{h}}^{n+1}(0) = (\mathfrak{h}_{t_k}^{n+1}(0))_k$ the optimal points.

(c) With $\boldsymbol{a}^{n+1} = (a_{t_k}^{n+1})_k$ and $\boldsymbol{\mathfrak{h}}^{n+1} = (\mathfrak{h}_{t_k}^{n+1}(0))_k$, associate $(x_{t_k}^{n+1,(i,1)})_{i,k}$ as in (2.3.16) (with the same prescription as above).

(d) Update the proxy by letting

$$m_{t_k}^{n+1,(1)} = \frac{1}{M} \sum_{i=1}^{M} x_{t_k}^{n+1,(i,1)}, \quad \overline{m}_{t_k}^{n+1,(1)} = \pi_{n+1}(\varpi) m_{t_k}^{n+1,(1)} + \big(1 - \pi_{n+1}(\varpi)\big) \overline{m}_{t_k}^{n,(1)},$$

with $\pi_n(\varpi)$ being defined as in Algorithm 2.

In all the numerical experiments, we use ADAM optimizer (as implemented in `TensorFlow`) in order to solve the optimization step in Algorithms 2, 3 and 4. We recall from the discussion in Subsection 2.1.7 that the code in `TensorFlow` relies on some automatic differentiation and thus makes an explicit use of the linear-quadratic form of the coefficients. In this sense, our numerical implementation requires part of the model to be known, but, as we already explained in Subsection 2.1.7, this does not really affect the scope of our conclusions: Provided the optimization method used in Algorithms 2, 3 and 4 is sufficiently accurate, the whole works well and demonstrates the interest for exploring the state space by means of the common noise.

2.3.3 Numerical experiments in low dimension

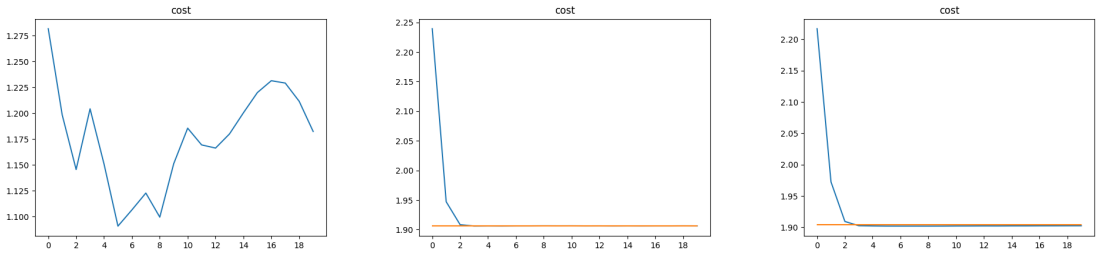
The numerical examples that we provide below are here to illustrate several features of our form of fictitious play. In all the examples treated in this subsection, we choose $d = 2$, $f \equiv 0$ and g as in (2.3.3) with $\kappa = 10$. The intercept $\boldsymbol{h}^n$ in (2.2.32) is approximated by means of Hermite polynomials, as explained in Subsection 2.3.2.1. Similar conclusions to the ones that are reported here could be drawn if the regression of $\boldsymbol{h}^{n+1}$ onto $\overline{\boldsymbol{m}}^n$ (see (2.3.17)) was done by using a neural network. For brevity, we feel however more appropriate to restrict the exposition of the numerical results obtained with neural networks to the higher dimensional setting only, which is done in the next section.

It is worth mentioning that, whatever the method that is used, our aim is to demonstrate, numerically, the positive impact of the common noise, but not to compare the numerical rates of convergence obtained in the numerical experiments with the bounds obtained in the theoretical analysis. The reason is that the algorithms we present below include additional features (like numerical optimization methods and regressions) that are not addressed in the theoretical analysis. This makes difficult any precise comparison. Also, it is fair to say that the time of execution of the full algorithm becomes quite long (several hours for the longest experiments), even for low values of n and p (n less than 50 and p less than 100).

As for the parameter ϖ, we address its influence in several manners. To do so, we implement both the harmonic and geometric variants of the fictitious play. Even in the harmonic regime (i.e., $\varpi = 1$), we can easily see the impact of the common noise, especially when the latter has intensity 1 (which is our choice in this subsection). Obviously, this is very good. A related observation is that, although our theoretical guarantees (given by the analogues of Theorems 2.2.4 and 2.2.17 in the harmonic regime, see Remark 2.2.6) just provide a harmonic decay when $\varpi = 1$, the observed rate of convergence may be better. In particular, the difference with the geometric version is rather tiny on the examples that are tested in this subsection, for a common noise with intensity 1 (as we will see in the forthcoming Subsection 2.3.5, the picture is different when the viscosity is small). There are even situations where, for numerical reasons, the harmonic variant has better results than the geometric one. For sure, this could be rather surprising for the reader (and even disappointing in some sense), but it is clear that our (theoretical) estimates rely on rather generic properties of the dynamics in hand. Obviously, there might be more subtle features of the equations that should be taken into account in order to explain the behavior of the algorithm as the number n of iterations of the fictitious play increases. Having a sharp understanding of the algorithm when n is large and, at the same time, ε is small is even more difficult. As we already explained, the constant $\exp(C\varepsilon^{-2})$ that appears in all of our estimates is certainly non-optimal in many situations.

2.3.3.1 Evolution of the learnt cost with one type of noise only

We first test the influence of each of the two types of noises onto the behavior of the algorithm. We thus compare the evolution of the cost learnt by the harmonic and geometric fictitious plays in three scenarios (Figure 2.6): In the first scenario, there is an independent noise, but no common noise (Algorithm 4); in the second scenario, there is a common noise but no independent noise and $\varpi = 1$ (Algorithm 3, harmonic fictitious play); in the third scenario, there is a common noise but no independent noise and $\varpi = 1.1$ (Algorithm 3, geometric fictitious play).



(a) With independent but no common noise

(b) With common but no independent noise, $\varpi = 1$

(c) With common but no independent noise, $\varpi = 1.1$

Figure 2.6: Harmonic and geometric fictitious plays. Comparison of the learnt cost depending on the type of noise and the value of ϖ. In x-axis: number of iterations; In y-axis; learnt cost.

The experiments are computed with: $n = 20$ learning iterations, $p = 30$ time steps, $M = 4 \times 10^5$, $N = 1$, $\sigma = 1$ and $\varepsilon = 0$ in case (A) and $n = p = 20$, $M = 1$, $N = 4 \times 10^5$, $\sigma = 0$, $\varepsilon = 1$ and $D = 4$ in cases (B) and (C). In ADAM method, the learning rate is 0.01, with 15 epochs and one batch.

In plots (B) and (C), the orange line is the theoretical equilibrium cost as computed by the BSDE method explained in Subsection 2.3.2.1. In plot (A), there is no computed reference cost. As we already explained, there might not be a unique equilibrium and the notion of reference cost no longer makes sense. Notice by the way that the equilibrium cost in cases (B) and (C) is not an equilibrium cost in case (A) because the problems are different (as being set over different forms of dynamics). Anyway, the conclusion is clear: the learnt cost exhibits an oscillatory behavior in case (A), whereas it does not in cases (B) and (C). This is an evidence of the numerical impact of the common noise onto the behavior of the fictitious play. Of course, it would be desirable to have more theoretical guarantees on the possible divergence of the fictitious play in absence of the common noise. Unfortunately, we are not aware of such results. In our numerical experiments (including some that are not reported here), we have been able to reproduce the oscillatory behavior characterizing the pane (A) in Figure 2.6 in other two-dimensional examples without common noise, meaning that, for a given batch (even of large size), the learnt (or training) cost may feature some non-trivial oscillations. However, we have not been able to reproduce[10] a similar phenomenon in dimension 1 (in which case the fictitious play is known to converge), even when the MFG is known to have multiple equilibria. As for panes (B) and (C), it is worth observing that the results are consistent. As announced, the convergence is fast, even in case (B), for which our theoretical guarantees just provide a harmonically decaying bound for the error.

[10]To be complete on this point, the experiments that we did are for a class of coefficients of the same nature as in the $2d$ example, in particular with a terminal boundary condition exhibiting oscillations of the same frequency, see (2.3.2) with κ between 1 and 10.

2.3.3.2 Error in the optimal path/intercept with common noise and no independent noise

We now compute the error achieved by the learning algorithm. Using the same notation as in the statement of Theorem 2.2.17, we focus on the following L^2 error:

$$\left[\mathbb{E}\int_0^1 \Big(|\overline{m}_t^n - m_t^{(p)}|^2 + |\varpi h_t^n - h_t^{(p)}|^2\Big)\mathrm{d}t\right]^{1/2}.$$

Focusing here on the time-averaged error is obviously more advantageous from the numerical point of view, but it is sufficient to do so in order to demonstrate the efficiency of the algorithm and also to identify some of its limitations. Notice also that the expectation in the above error is taken under the non-tilted expectation. This looks a reasonable choice because ε is here equal to 1. In particular, the Girsanov density $\mathcal{E}(\boldsymbol{h})$ and its inverse have bounded moments of any order (independently of n), as a consequence of which integrating under either measure should not make a big difference.

Numerically, the error is approximated by

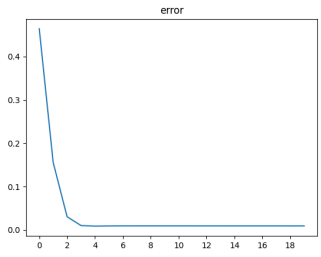
$$\left[\frac{1}{Np}\sum_{j=1}^{N}\sum_{k=0}^{p-1}\Big(|\overline{m}_{t_k}^{n,(j)} - m_{t_k}^{\star,(j)}|^2 + |h_{t_k}^{n,(j)} - \varpi h_{t_k}^{\star,(j)}|^2\Big)\right]^{1/2}, \tag{2.3.20}$$

where $(\overline{m}_{t_k}^{n,(j)})_{j,k}$ and $(h_{t_k}^{n,(j)})_{j,k}$ are the returns of Algorithm 2 or 3 (depending on the value of σ) and $(\overline{m}_{t_k}^{\star,(j)})_{j,k}$ and $(h_{t_k}^{\star,(j)})_{j,k}$ are the reference solutions computed with Algorithm 1 (with a sufficiently high number of Picard iterations: 10 in practice).

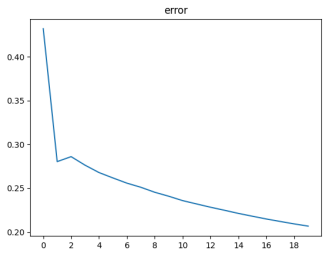
We perform the first experiment by running Algorithm 3 (no idiosyncratic noise), but assuming that the solution $(\eta_t)_{0\leqslant t\leqslant 1}$ to the Riccati equation (2.3.5) is known, see (2.3.6): In Algorithm 3, a_{t_k} is replaced by η_{t_k}. This amounts to say that the coefficients Q and R in (2.1.2) are explicitly known. The evolution of the error with the number of iterations is plotted in Figure 2.7, Plot (A) when $\varpi = 1$ (harmonic fictitious play) and Plot (C) when $\varpi = 1.1$ (geometric fictitious play). We perform the second experiment by running Algorithm 3, but without assuming any further knowledge of the solution of the Riccati equation (2.3.5). The evolution of the error with the number of iterations is plotted in Figure 2.7, Plot (B) when $\varpi = 1$ and Plot (D) when $\varpi = 1.1$.

The experiments are computed with: $n = 20$ learning iterations, $p = 30$ time steps, $M = 1$, $N = 4 \times 10^5$, $\sigma = 0$, $\varepsilon = 1$ and $D = 4$. On Plots (A) and (C), the error is less than 0.01 after iteration 3.

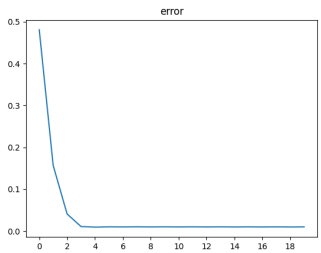
The main conclusion one may draw from the comparison of these plots is quite clear: when randomness only comes from the common noise, it is very difficult for the optimization method to distinguish between $a_{t_{k-1}}x_{t_{k-1}}^{(1,j)}$ and $C_{t_{k-1}}^{(j)}$ in (2.3.16). In this matter, there is also a difference between Plots (B) and (D), from which we deduce that the harmonic fictitious play behaves better than the geometric one in the absence of independent noise. Comparison of plots (A) and (C), which are very close, suggests that the difference between (B) and (D) mostly comes from the accuracy of the estimate $(a_{t_{k-1}})_{k=1,\cdots,p}$ of the solution to the Riccati equation (2.2.34). Our guess is that the presence of an additional biais in (2.3.16) (due to the fact that $\varpi > 1$) makes the estimation more difficult. On the contrary, when $\varpi = 1$, there is no bias and the term $C_{t_{k-1}}^{(j)} + \varpi h_{t_{k-1}}^{n,(j)}$ in (2.3.16) is very close to 0. We think that this should help for the identification of $(a_{t_{k-1}})_{k=1,\cdots,p}$.



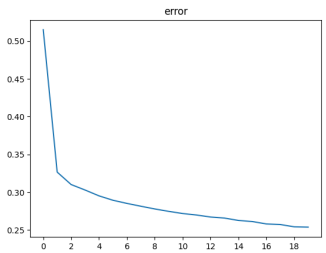
(a) Riccati known, $\varpi = 1$



(b) Riccati unknown, $\varpi = 1$



(c) Riccati known, $\varpi = 1.1$

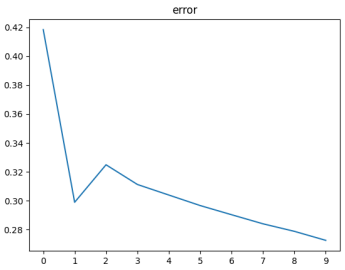


(d) Riccati unknown, $\varpi = 1.1$

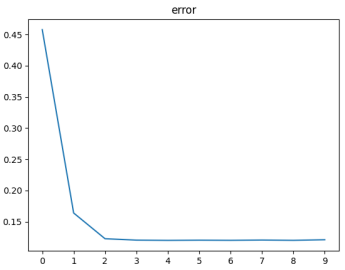
Figure 2.7: Comparison of the error returned by Algorithm 3 (common noise only), depending on whether the solution to the Riccati equation is known or not. On the top line, $\varpi = 1$. On the bottom line, $\varpi = 1.1$.

2.3.3.3 Error in the optimal path/intercept with both noises

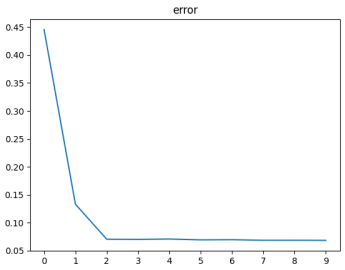
We now proceed with the same analysis but putting the two noises. As the total number of simulated paths is $N \times M$, this may be however rather costly. We proceed below by freezing the quantity $N \times M$.



(a) $M = 4 \times 10^4$, $N = 1$, time (cpu): 4867



(b) $M = 2 \times 10^4$, $N = 20$, time (cpu): 4096



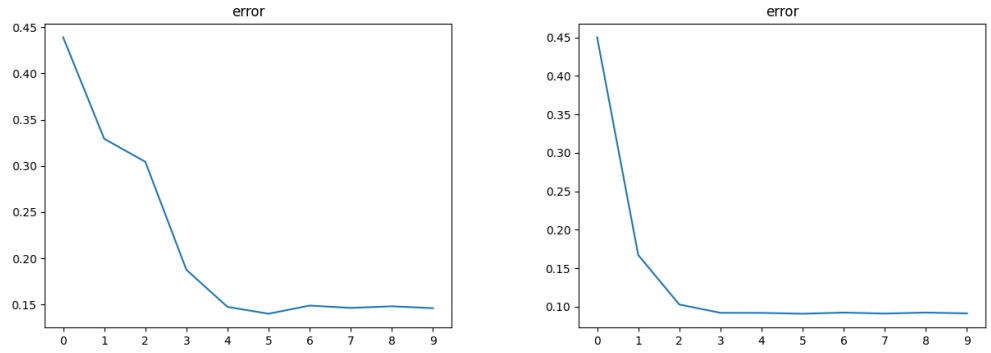
(c) $M = 10^4$, $N = 40$, time (cpu): 8409

Figure 2.8: Comparison of the error returned by Algorithm 2, depending on the number of particles/realizations, $\varpi = 1$.

The experiments are run with the harmonic version of the fictitious play (i.e., $\varpi = 1$), with:

$n = 10$ learning iterations, $p = 30$ time steps, $\sigma = 1$ and $\varepsilon = 1$; 4000 units in cpu time is around 1h. The conclusion is that idiosyncratic noises demonstrate to be useful from the numerical point of view, even though the results stated in Section 2.2 remain the same whatever the value of σ. A careful inspection of the numerical results shows that, in fact, idiosyncratic noises provide a better fit of the solution to the Riccati equation, which is fully consistent with the conclusion of the previous paragraph. In turn, we guess that a model-based method, in which Q and R would be learnt first, and then the solution to the Riccati equation would be learnt separately, would make sense in this specific setting.

When $\varpi = 1.1$, we observe a phenomenon similar to the one reported in Figure 2.7: the estimate is less accurate with the geometric version of the fictitious play. Again, we believe that this is due to the learning of the solution to the Riccati equation. In presence of a bias, the letter seems more difficult to catch. To wit, we have plotted in Figure 2.9, Plot (A), the error for the geometric version of the fictitious play (here, $\varpi = 1.1$) and for the same parameter as in Plot (C) in Figure 2.8: $n = 10$, $p = 30$, $\sigma = 1$, $\varepsilon = 1$, $M = 10^4$ and $N = 40$. In order to stress that the drawback observed on Plot (A) does not contradict our theoretical results, we have plotted in the same Figure, Plot (B), the results when the solution to the Riccati is known (which is a bit different from what is done in Figure 2.7 because there is here an additional idiosyncratic noise which requires additional Monte-Carlo approximations). It is clear that, in the latter case, the result is better. Possibly, this opens the door for refined procedures in which one first estimates the solution to the Riccati equation.



(a) $M = 10^4$, $N = 40$, Riccati unknown (b) $M = 10^4$, $N = 40$, Riccati known

Figure 2.9: Algorithm 2, $\varpi = 1.1$, depending on whether Riccati is known or not.

2.3.3.4 Validation

The reader could worry about a possible overfitting in our numerical experiments. Actually, in order to validate our results, we can learn the coefficients $\boldsymbol{a}^n$ and $\boldsymbol{c}^n$ in Algorithm 2 for a first series of data in (2.3.14). In brief, this first series of data is used to learn the equilibrium feedback. Then, we can use a second series of data (say, of the same size) in (2.3.14) in order to compute the error. In clear, given this second series of data, we can implement Algorithm 1 and then compute

the resulting error (2.3.20) when $(\overline{m}_{t_k}^{n,(j)})_{j,k}$ and $(h_{t_k}^{n,(j)})_{j,k}$ therein are obtained from the formulas (2.3.16) and (2.3.18) by using the coefficients $\boldsymbol{a}^n$ and $\boldsymbol{c}^n$ returned by the first series of data and then by implementing the Euler scheme with respect to the second series of data. The resulting plots are very similar to those represented in Figure 2.8 (e.g., when $\varpi = 1$). For this reason, we feel useless to insert them here. Intuitively, what happens is that there are sufficiently many realizations of the common noise in our experiments to guarantee, in the regression (2.3.15), a convenient form of averaging with respect to all these realizations.

2.3.4 Experiments in higher dimension

The challenge in higher dimension is twofold. The very main one is to provide an efficient regression of $\boldsymbol{h}^{n+1}$ onto $\overline{\boldsymbol{m}}^n$ (recall (2.3.17)). We already reported this difficulty and, as we already announced, we handle it here by using neural networks. The second issue is directly related with the computational cost. Intuitively, all the tensors that enter the code include an axis corresponding to all the possible coordinates of the noises and the states. When the dimension increases, the size of those tensors increase accordingly, which impacts the computational effort. In all the examples below, $\sigma = \varepsilon = 1$. For simplicity, we just present the result in the case $\varpi = 1$ because the difficulties that we report below are the same in the case $\varpi > 1$.

2.3.4.1 Results with Algorithm 2.

In the examples below, we reduce part of the complexity by labelling the increments of the independent noises $\Delta_{t_k} b^{(i,j)}$ in (2.3.14) by the sole i, which allows us to save memory. We refer to Remark 2.3.4 for more details about this. We also limit ourselves to batches carrying $N = 10^3$ realizations of the common noise and $M = 10^2$ realizations of the independent noise. These numbers are a bit less than those used in the lower dimensional setting (compare for instance with Figure 2.8). However, differently from what we did in the previous subsection, we now use several batches. In the experiments below, the batches are indexed by an index, called *batch*. For a given value i of this index *batch*, we simulate one stack, say $G_{\rm com}[i]$, of $N = 10^3$ realizations of the common noise and then a number $m_{\rm ind}$ of sub-batches (or sub-stacks) containing, each, $M = 10^2$ realizations of the independent noise. Those sub-batches are denoted $G_{\rm ind}[i, j]$, for $j = 1, \cdots, m_{\rm ind}$. The batches contain independent realizations. In the experiments below, $m_{\rm ind} = 10$.

When we start experiments with a new value i of the index *batch*, we perform several runs of ADAM. All these runs use the same batch $G_{\rm com}[i]$ of common noises. We then repeat several times the following episode: we do a first run with the first batch $G_{\rm ind}[i, 1]$ of independent noises, and then a second one with the second batch $G_{\rm ind}[i, 2]$, and so on and so forth up until the batch $G_{\rm ind}[i, m_{\rm ind}]$ of index $m_{\rm ind}$. We repeat those episodes several times: 4 times in the experiments ran below. Our choice to proceed in such a way is dictated by the fact that the number N is pretty low, hence the desire to have more batches for the independent noises. As for the choice $N \gg M$, it is dictated by our wish to have the lowest possible fluctuations with respect to the common noise, as the main object that we want to learn here is $\mathfrak{h}$ in (2.3.17), the input of which is measurable with respect to the common noise only. In our experiments, the index *batch* is running over $\{1, \cdots, 20\}$. We have chosen to pass once on the realizations of the common noise.

For some choices of the matrix Θ in (2.3.4), our results are bad. In order to appreciate this

observation, it is worth recalling that our strategy relies on a change of measure based on the Girsanov density $\mathcal{E}(\boldsymbol{h}^n)$, see (2.1.15). In our code, the training cost is precisely estimated under this new probability measure, as made clear in (2.3.15). Drawing a parallel with preferential sampling in Monte Carlo methods, one may then guess that the variance resulting from the empirical estimate of the loss in (2.3.15) may dramatically deteriorate as the dimension increases. Indeed, if $\boldsymbol{h}$ in (2.1.15) has all its coordinates constant equal to 1, then

$$\mathbb{E}\Big[\mathcal{E}(\boldsymbol{h})^2\Big] = \exp\Big(\int_0^T |h_s|^2 ds\Big) = \exp(d).$$

Even more, the fact that the control appearing in (2.3.15) contains a finite difference of the common noise (see (2.3.16)) says that the typical size of the variance may be of order $p^2 \exp(d)$.

In order to validate numerically this intuition, we consider the case when the matrix $(\Theta_{i,j})_{1\leqslant i,j\leqslant d}$ in (2.3.4) is given by $\Theta_{i,2i(\mathrm{mod})\mathrm{d}} = -10$, $\Theta_{i,3i(\mathrm{mod})\mathrm{d}} = 10$ and $\Theta_{i,j} = 0$ otherwise, where $2i(\mathrm{mod})\mathrm{d}$ is the rest of the Euclidean division of $2i$ by d (and similarly for $3i(\mathrm{mod})d$). We then represent in Figure 2.10 the evolution of the variance of the term

$$\mathcal{E}(\boldsymbol{h}) \int_0^T |\dot{W}_t^{p,\boldsymbol{h}}|^2 dt, \tag{2.3.21}$$

when $\boldsymbol{h}$ therein is computed numerically by the FBSDE solver described in Subsection 2.3.2.1. We expect this term to give the worst contribution to the global variance. As shown by the plot in Figure 2.10, the log variance becomes higher at multiples of 6 (because of the choice Θ) and the plot even suggests that the variance could blow up exponentially fast along the subsequence 6, 12, 18, 24...

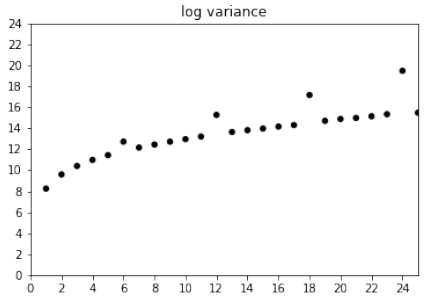
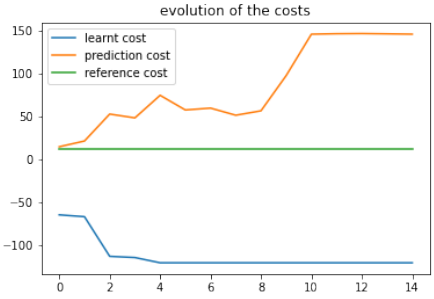


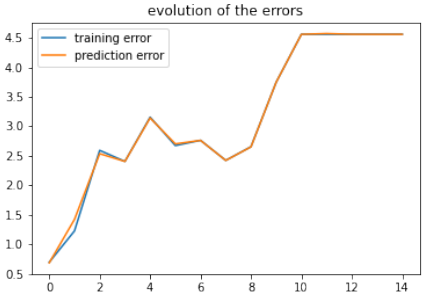
Figure 2.10: Evolution of the log variance of the loss with the dimension for $\Theta_{i,2i(\mathrm{mod})\mathrm{d}} = -10$, $\Theta_{i,3i(\mathrm{mod})\mathrm{d}} = 10$ in (2.3.4). Dimension is in x-axis. Log variance is in y-axis.

We thus chose to run Algorithm 2 with $d = 12$ with aforementioned values of M and N and with $p = 20$ time steps. The results are reported in Figure 2.11 below. We clearly observe that the learnt cost blows-up. Subsequently, the training error does not vanish, with the training error being here computed as in (2.3.20) but solely for the $\boldsymbol{h}$ part, that is

$$\left[\frac{1}{Np}\sum_{j=1}^{N}\sum_{k=0}^{p-1} |h_{t_k}^{(j)} - h_{t_k}^{\star,(j)}|^2\right]^{1/2}. \tag{2.3.22}$$



(a) Evolution of the costs



(b) Evolution of the errors

Figure 2.11: Results in dimension $d = 12$ for $\Theta_{i,2i(\mathrm{mod})\mathrm{d}} = -10$, $\Theta_{i,3i(\mathrm{mod})\mathrm{d}} = 10$ in (2.3.4). Harmonic variant of the fictitious play ($\varpi = 1$).

In the plot, we also include a prediction cost and a prediction error, which are computed on a batch different from the batches used for training (but of the same size). Given a new batch for the common and independent noises, we can indeed use the neural network returned by the training phase to compute, for this new batch, a prediction of the cost (together with a prediction of $\boldsymbol{h}$), but under the historical probability measure $\mathbb{P}$ (in order to avoid any variance issues). The prediction is thus constructed as follows. Given the returns $(a_{t_k})_{k=0,\cdots,p-1}$ and $(\mathfrak{h}_{t_k})_{k=0,\cdots,p-1}$ of the neural network for the affine feedback, see (2.3.13) and (2.3.17), we implement the following Euler scheme:

$$
\begin{aligned}
x_{t_k}^{(i,j)} &= x_{t_{k-1}}^{(i,j)} + \frac{1}{p}\alpha_{t_k}^{(i,j)} + \frac{1}{\sqrt{p}}\Delta_{t_k} b^{(i)} + \frac{1}{\sqrt{p}}\Delta_{t_k} w^{(j)}, \quad \ell = 1, \cdots, p\,; \quad x_0^{(i,j)} = x_0, \\
\alpha_{t_k}^{(i,j)} &= a_{t_{k-1}} x_{t_{k-1}}^{(i,j)} + \mathfrak{h}_{t_{k-1}}(\overline{m}_{t_{k-1}}^{(j)}), \quad k = 1, \cdots, p,
\end{aligned}
$$

with $\overline{m}_{t_{k-1}}^{(j)} = M^{-1}\sum_{i=1}^{M} x_{t_{k-1}}^{(i,j)}$.

The predicted cost is then

$$
\frac{1}{N}\sum_{j=1}^{N}\frac{1}{M}\sum_{i=1}^{M}\left[\frac{1}{2p}\sum_{k=1}^{p}|\alpha_{t_k}^{(i,j)}|^2 + \frac{1}{2}\left|x_1^{(i,j)} + g\left(\overline{m}_1^{n,(j)}\right)\right|^2\right],
$$

and the predicted error is (2.3.22).

While their role is rather limited at this stage, the two prediction cost and error play a more important role in the next paragraph.

2.3.4.2 Variance reduction.

The issue reported in the previous paragraph calls for the implementation of a variance reduction method. The method that is addressed in this new paragraph aims at reducing the variance associated with the dominant term (2.3.21). The key point is to return back to (2.2.30), to write the normalization factor $-d\,\varepsilon^2 p/2$ as the expectation

$$
\varepsilon\mathbb{E}^{\boldsymbol{h}_n}\int_0^T \alpha_t \dot{W}_t^{p,\boldsymbol{h}^n} dt - \frac{\varepsilon^2}{2}\mathbb{E}^{\boldsymbol{h}_n}\int_0^T |\dot{W}_t^{p,\boldsymbol{h}^n}|^2 dt, \tag{2.3.23}
$$

which is indeed equal to $-d\,\varepsilon^2 p/2$, and then to approximate the above two expectations by two empirical means. Intuitively (and this is the computation achieved in (2.2.30)), the resulting estimator of the cost coincides with the estimator that would be computed if we had to solve (the reader may compare the following cost with (2.2.32), with $\varpi = 1$ for simplicty)

$$\mathrm{argmin}_{\boldsymbol{\alpha}} \mathbb{E}^{\boldsymbol{h}^n}\big[\mathcal{R}^p\big(\boldsymbol{\alpha}; \overline{\boldsymbol{m}}^n; \varepsilon \boldsymbol{W}^p\big)\big], \tag{2.3.24}$$

which is implicitly set over the dynamics

$$\mathrm{d}X_t = \alpha_t \mathrm{d}t + \sigma \mathrm{d}B_t + \varepsilon \mathrm{d}W_t^{p,\boldsymbol{h}^n}, \quad t \in [0,T]. \tag{2.3.25}$$

Here, we have two formulations depending on the line that is used in (2.2.32): (i) We have the formulation (2.3.24)–(2.3.25), which is based on the second line in (2.2.32); (ii) And we have the formulation based on the top line in (2.2.32) when the corrector $-d\,\varepsilon^2 p/2$ therein is replaced by the empirical mean deriving from (2.3.23). Interestingly, the two formulations do not have the same practical interpretation. Indeed, the formulation (i) does not fit the principle of Figure 2.2 in introduction since the common randomization therein does not impact the control directly, but impacts the dynamics. Differently, the formulation (ii) based on the empirical corrector associated with (2.3.23) preserves the principle of Figure 2.2 since the corrector can be computed separately from the control. Of course, it has a practical price: implicitly, computing the correction empirically makes sense only if the dependence of the running cost upon the control is known. In other words, part of the model must be known.

Figure 2.12 below provides an estimate of the variance of the (random) cost in (2.3.24) when $\boldsymbol{h}^n$ is replaced by the numerical solution $\boldsymbol{h}$ of the FBSDE solver described in Subsection 2.3.2.1 and when Θ is computed as in Figure 2.10. As the reader can notice, the growth is slower (than in Figure 2.10).

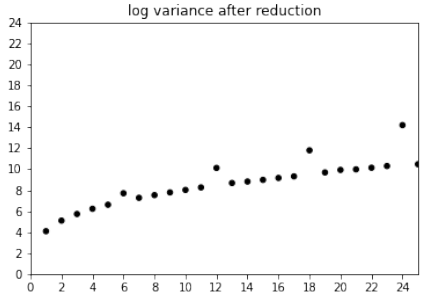
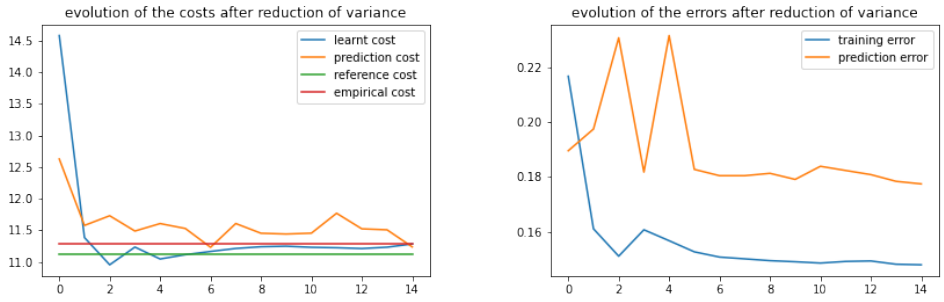


Figure 2.12: Evolution of the log variance of the corrected optimal cost (2.3.24). with the dimension for $\Theta_{i,2i(\mathrm{mod})\mathrm{d}} = -10$, $\Theta_{i,3i(\mathrm{mod})\mathrm{d}} = 10$ in (2.3.4). Dimension is in x-axis. Log variance is in y-axis. Harmonic variant of the fictitious play ($\varpi = 1$).

The estimator for (2.3.23) writes as in (2.3.15), namely

$$\frac{\varepsilon}{N}\sum_{j=1}^{N}\left(\mathcal{E}^{n,(j)}\frac{1}{M}\sum_{i=1}^{M}\left[\frac{1}{2p}\sum_{k=1}^{p}\alpha_{t_k}^{(i,j)}\cdot\left(h_{t_k}^{n,(j)}+p\Delta_{t_{k+1}}w^{(j)}\right)\right]\right)$$
$$-\frac{\varepsilon^2}{2N}\sum_{j=1}^{N}\left(\mathcal{E}^{n,(j)}\frac{1}{2p}\sum_{k=1}^{p}\left|h_{t_k}^{n,(j)}+p\Delta_{t_{k+1}}w^{(j)}\right|^2\right).$$

The results are represented in Figure 2.13 for the same choice of d and Θ as in Figure 2.11. Obviously, they are much more convincing than in Figure 2.7. On the left pane (A), the reference cost is computed by solving first the FBSDE described in Subsection 2.3.2.1 on a series of batches and then by averaging the corresponding costs over the batches. In contrast, the empirical cost is computed by solving the FBSDE and the associated cost on the batch on which the training loss is computed. The fact that the reference and empirical costs do not coincide shows that there is, in between, some additional Monte-Carlo error that is distinct from the learning procedure.



(a) Evolution of the losses

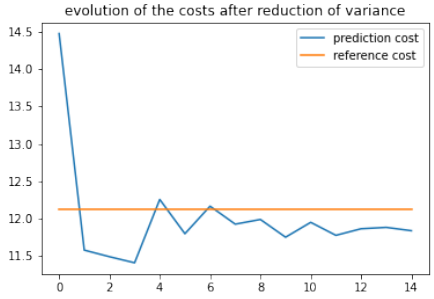
(b) Evolution of the errors

Figure 2.13: Results in dimension $d = 12$ for $\Theta_{i,2i(\mathrm{mod})\mathrm{d}} = -10$, $\Theta_{i,3i(\mathrm{mod})\mathrm{d}} = 10$ in (2.3.4). Harmonic variant of the fictitious play ($\varpi = 1$).

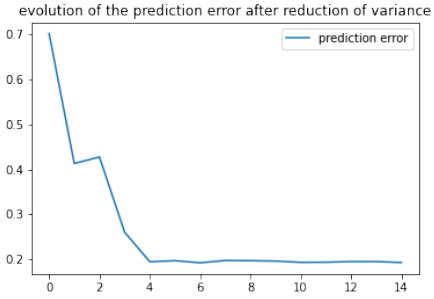
The prediction cost and error are computed on a series of 10 batches (of the same size as before). The cost and the error that are plotted are obtained by averaging out on the batches. We observe a small bias between the training and prediction errors that would deserve further experiments. Anyway the main message is clear: the plots after variance reduction are much better than before variance reduction.

We now provide another example, with a new choice of Θ, that exhibits a less trivial evolution of the prediction error. Namely, we choose $\Theta = (\theta_{i,j} = (i+j)/(2d))_{i,j}$. The results in Figure 2.14 may be summarized as follows: the relative error on the cost is less than 0.03 and the prediction error is less than 0.2. As for the latter, it must be recalled that the prediction error writes as a d-dimensional norm, see (2.3.22). In particular, the mean square error per coordinate is less than $0.2^2/d$. Here, $d = 12$ and the mean error per coordinate is less than 0.06. We address the same example in Figure 2.15 in dimension $d = 20$. The absolute error on the cost is good, except at iterations 11 and 13, but even for these two the relative error is less than 0.05. The prediction error

is between 0.22 and 0.25 after iteration 6, except at iterations 11 and 13, where the prediction error reaches 0.32. In any case, the mean error per coordinate is less than 0.07 after iteration 6.

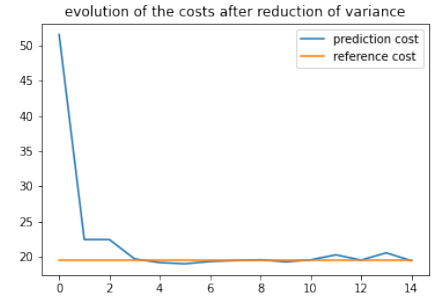


(a) Evolution of the losses

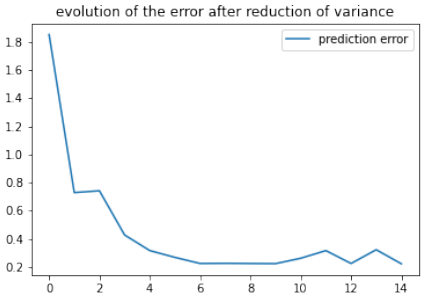


(b) Evolution of the errors

Figure 2.14: Results in dimension $d = 12$ for $\Theta = (\theta_{i,j} = (i + j)/(2d))_{i,j}$ in (2.3.4). The reference loss is around 12.17 and the prediction loss at iteration 14 is 11.84. The prediction error at iteration 14 is around 0.19. Harmonic variant of the fictitious play ($\varpi = 1$).



(a) Evolution of the losses



(b) Evolution of the errors

Figure 2.15: Results in dimension $d = 20$ for $\Theta = (\theta_{i,j} = (i + j)/(2d))_{i,j}$ in (2.3.4). The reference loss is around 19.47 and the prediction loss at iteration 14 is 19.45. The prediction error at iteration 14 is around 0.22. Harmonic variant of the fictitious play ($\varpi = 1$).

2.3.5 Small viscosity

The next question in our numerical experiments is to address the behavior of the algorithm as the viscosity tends to 0. As made clear in the analysis performed in Section 2.2, the influence of the small viscosity may manifest in an exponential manner and this was our original motivation to design a geometrically converging scheme.

Beyond the theoretical challenge raised by the possible occurrence of singularities in the vanishing viscosity limit (an example of which is given by Lemma 2.2.8), small viscosity may also create additional numerical instability phenomena. As a main example, high variances may occur in the

computations of empirical measures under the tilted probability measure. Obviously, this follows from the fact that the Girsanov density driving the change of reference measure in the algorithm becomes highly singular as the viscosity tends to 0. We reported a similar drawback in the previous subsection, but in the large dimensional framework. In the first arXiv version [48] of this work, the observation of this phenomenon prompted us to introduce a method at the intersection between annealed simulating and preferential sampling, which we illustrated on a specific example that is recalled below. While this method is performing well (we revisit it in the framework of the geometric fictitious play in this subsection), it requires to repeat the scheme for several values of the viscosity and it is thus rather costly.

Instead, we here show that the geometric fictitious play can quickly return a very accurate numerical solution to the vanishing viscosity property problem (at least in the example addressed in the first arXiv version [48]). This is a very striking exemplification of our approach. It clearly demonstrates the benefits of choosing a higher value of the rate ϖ and therefore of working with the geometric variant (instead of the harmonic variant) of the fictitious play. Our numerical experiment based on the aforementioned benchmark model. We choose a variant of g in (2.3.2), with $d = 1$:

$$g(x) = \cos\big(\kappa(x - x_0)\big) - 2x_0, \tag{2.3.26}$$

with $\kappa = 10$ and where x_0 is a root of the equation

$$\cos\big(\kappa x_0\big) = 2x_0.$$

Numerically, we find that a choice is $x_0 \approx -0.384$. The motivation for such an x_0 is that 0 is a solution of (2.3.9). In other words, 0 is an equilibrium. Even more, we observe that, if the viscosity is zero, then the iterative sequence defined through the two updating rules (2.1.9) and (2.1.10) remains in $(\boldsymbol{m}^n, \boldsymbol{h}^n) = (0, -2x_0)$ for any $n \geqslant 1$ if $\boldsymbol{m}^0$ is chosen as 0. In other words, the standard fictitious (without any exploration) play converges (as predicted by the theory since this 1-dimensional MFG is potential, see §2.3.1.2) and chooses the 0-equilibrium. In order to compare with the prediction method exposed in §2.3.1.2, we have plotted the corresponding potential (which is given by a primitive G of g). The plots are given in Figure 2.16. We observe that 0 is just a local minimizer of the potential and that the global minimizer is around -0.5.

Our geometric fictitious play is able to rule out the 0-equilibrium and to retain the other one. In order to check this numerically, we have used three values for ϖ: $\varpi = 4$, $\varpi = 1.5$ and $\varpi = 1$. As for the other parameters, we have chosen $\sigma = 0$ (no idiosyncratic noise, but Riccati is known), $M = 2 \times 10^4$ (number of Monte-Carlo simulations) and $p = 100$. In all the experiments, ε is set equal to 0.2. The algorithm is initialized from the local minimizer. Our regression on the Hermite polynomials goes up to polynomials of degree 6. The results are presented in Figures 2.17, 2.18 and 2.19. Therein, we have plotted the histogram, under the tilted probability measure and at the end of each episode, of the terminal conditional mean of the system. Even though the regime $\varpi = 4$ is outside the scope of Theorem 2.2.4 (because we assumed $\varpi \in (1, \sqrt{2}]$), the result with this choice is quite impressive: the right equilibrium is selected in five iterations only. When $\varpi = 1.5$, ten iterations are necessary. When $\varpi = 1$, the results are satisfactory with around fifteen iterations but it is clear that the histogram features a kind of residual variance, which makes the selection less obvious. We believe that these plots are a strong case for supporting our results.

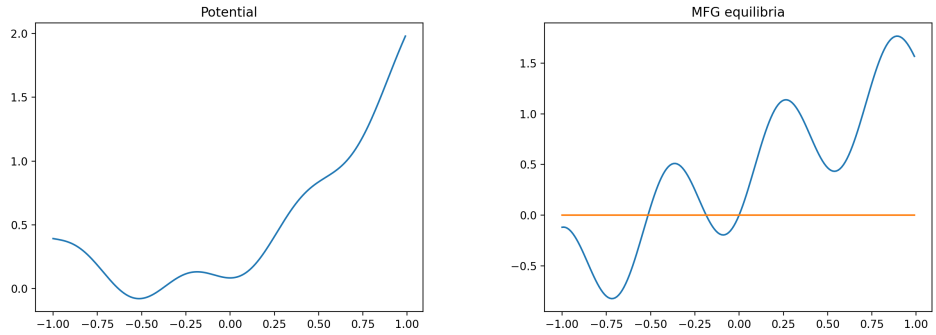


Figure 2.16: Equilibrium predicted by the potential rule: Potential G on the left pane; Zeros of the function g in the right pane.

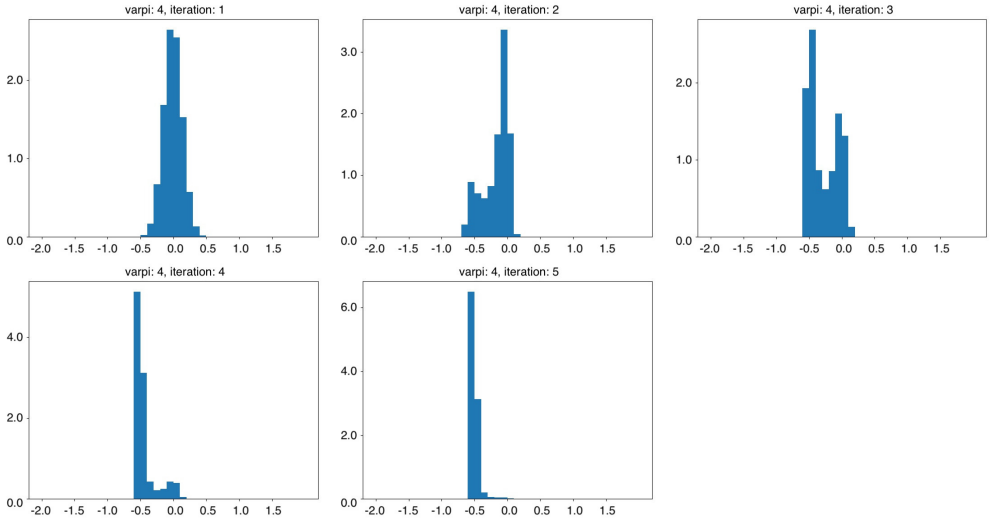


Figure 2.17: Selection of a solution by vanishing viscosity: histogram under the tilted measure of the terminal mean for $\varepsilon = 0.2$ and $\varpi = 4$. Equilibrium is clearly selected in 5 iterations.

For sure, the reader may wonder about the same plots if we decrease the value of ε (say $\varepsilon = 0.1$). It is fair to recognize that the results are not so good when $\varepsilon = 0.1$. To our mind, this is due to the variance issues that we reported above: the Girsanov change of reference may induce high variances. Numerical observations seem to indicate that those issues may even get worse when increasing the value of ϖ (which is not so easy to explain from a theoretical point of view because the integrand $\varpi \boldsymbol{h}^n$ in (2.2.1) is close to the true solution $\boldsymbol{h}$, regardless of the value of ϖ). One possible strategy to reduce the underlying variance would be to implement the same preferential sampling argument as in the first arXiv version [48], but it is fair to say that it is difficult to obtain a plot that is as good as the one obtained in Figure 2.17. For this reason, we feel better to address this possible extension in a future work.

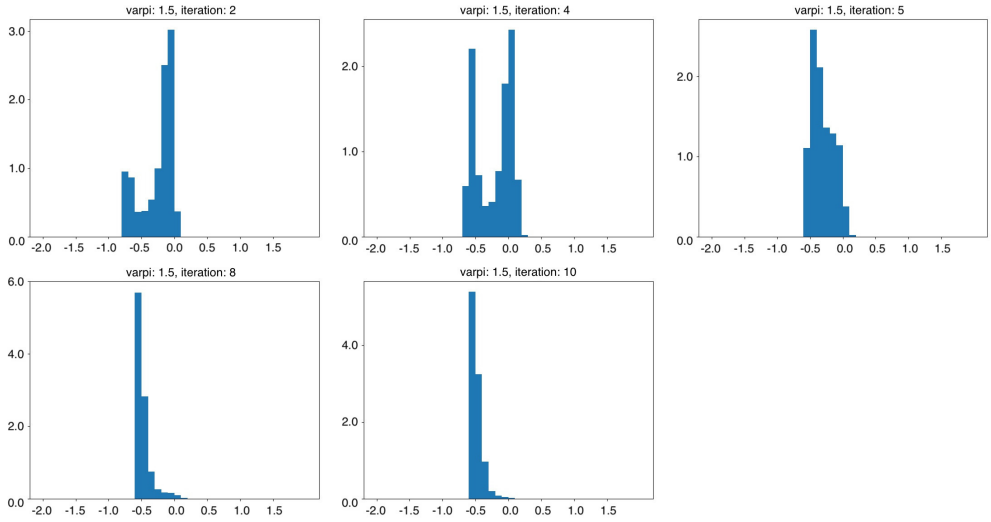


Figure 2.18: Selection of a solution by vanishing viscosity: histogram under the tilted measure of the terminal mean for $\varepsilon = 0.2$ and $\varpi = 1.5$. Equilibrium is selected in 10 iterations.

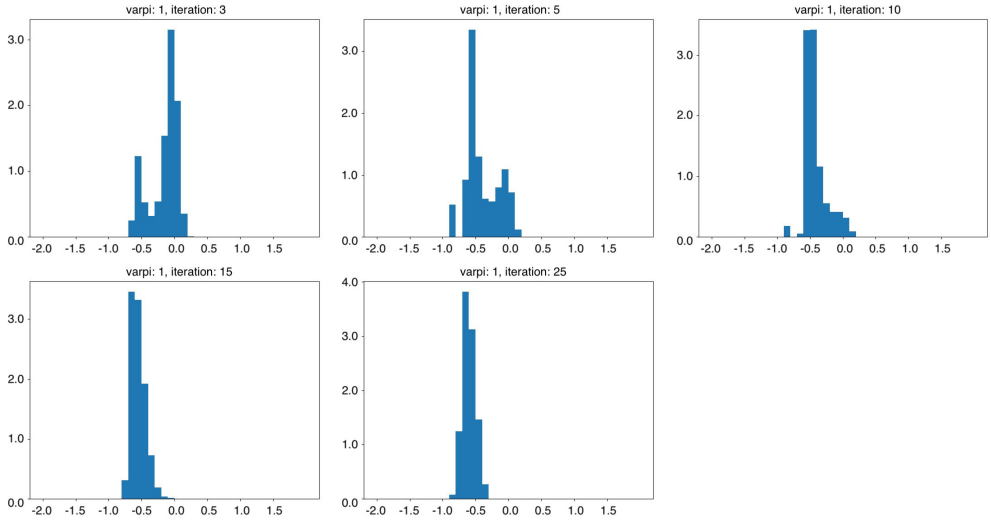


Figure 2.19: Selection of a solution by vanishing viscosity: histogram under the tilted measure of the terminal mean for $\varepsilon = 0.2$ and $\varpi = 1$. Equilibrium is selected in more than 10 iterations, with a non-trivial residual variance.

Chapter 3

Some quantitative bounds for Q-learning in continuous spaces and an application to Mean Field MDP with finite states

3.1 Introduction

Since its conception by Watkins in 1989 [114] , Q-learning has been one of the prominent algorithms for solving Reinforcement Learning (RL) problems. In the typical case of RL, an agent being in a given state, interacts with an environment performing actions that yield rewards and update her state. Usually (but not exclusively) the agent has some target that she wants to obtain, possibly in an optimal way.

The standard Q-learning is a model-free algorithm, meaning that it does not require to represent the underlying system dynamics in form of a model. Instead, the algorithm learns through interaction with the environment, progressively refining its decision-making strategy based on received feedback. Q-learning uses real or simulated data (sequences of states, actions and rewards) to approximate the value function of dynamic programming as a function of the initial state. Furthermore, the algorithm is recursive the sense that each new piece of information is used to update the current estimates.

Even though Q-learning figures in all majors Reinforcement Learning textbooks, still up until this day it attracts considerable attention, with various new variants appearing in the literature. Its convergence has been demonstrated early by Watkins [115] using a prototypical proof device and later by Tsitsiklis [109] using stochastic approximation on more general assumptions. However little has been known on an explicit rate of convergence in the general case of a continuous space, this fact is one of the main motivations of our work. One key idea in this regard is the notion of exploration: the true Q-function (which we also call the action value function) can be well approximated on the whole state and action spaces provided that the space-action observations visit sufficiently well the ambient spaces. Implicitly, this requires a form of non-degeneracy in the

transition kernels governing the occurrence of observations.

The first main contribution of this article is a new proof that follows globally the steps of the initial proof of Watkins [115] for discrete states and of Carden [27] for continuous spaces but instead provides an exact bound for the convergence in probability of the approximate value function to the optimal one. Our motivation for addressing models with continuous spaces comes from our original desire to study reinforcement learning for mean field control. In line with the analysis carried out in the previous chapter for linear-quadratic mean field, this is indeed here our objective to provide a sharp exploration-exploitation analysis for mean field learning, at least when learning is performed by tabulating the Q-function. In particular, we want here to interpret the non-degeneracy properties that are required on the transition kernels as properties that derive from a common noise that we regard in the end as an exploration noise. In this way, our work here complements the study from the previous chapter and offers a new example of 'exploration through common noise'.

In fact, one key point in the construction of Carden is to use a Nadaraya-Watson kernel in order to extend the empirical Q-function (or approximated Q-function) from the observations to the whole space. Similar to [27] (see also [115]), our analysis of the resulting approximated Q-function is indeed based on the so-called Action Replay Process (ARP) of Watkins (see Section 3.2.2 for a definition), which is an auxiliary Markov decision process whose Q-function coincides with the estimator returned by the Nadaraya-Watson regression. In this regard and in comparison with the aforementioned references [27, 115], there are several main differences in our work: (i) We show that we can get rates that remain tractable in the high-dimensional setting by assuming that the coefficients and the true Q-function are sufficiently smooth. This is consistent with standard methods from numerical analysis in which smoothness of the data is used to decrease the underlying complexity; (ii) We use tools from stochastic analysis and in particular from the theory of martingales to show that the ARP features averaging properties that we prove useful to get bounds (in a convenient topology) between the transition kernel of the ARP and the original transition kernel; (iii) We use thorough bounds for the coupon collector problem to derive maximal inequalities for the covering times of ARP, i.e. the time it needs to visit every partition of the state-action space.

In what follows we start by introducing some notation to clarify the objects that we are going to study and provide some context for our methodological choices. We end this introductory section with a discussion about the results of the article in contrast with related works and further organisation of the rest.

3.1.1 Set-up

In this subsection, we introduce basic notations that will be used throughout the chapter. We consider two bounded open subsets S and A of $\mathbb{R}^{d_S}$ and $\mathbb{R}^{d_A}$ respectively, for d_S and d_A two positive integers, with both S and A satisfying a uniform interior cone condition with Lipschitz boundary (see [6, Chapter 4] and [107] for various reformulations). In brief, $\overline{S}$ and $\overline{A}$ are (respectively) the state and action spaces of the Markov decision process under study. (We refer to [12] for a textbook on Markov decision processes and to [11] for a focus in the mean field framework). Below, s is frequently used to denote a generic element of $\overline{S}$ and a to denote a generic element of $\overline{A}$. Moreover, we use the notation D for $D := d_S + d_A$.

Moreover, we consider (on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$) a stochastic process $(\mathrm{s}_n, \mathrm{a}_n)_{n\geqslant 0}$ with

values in $\overline{S} \times \overline{A}$, adapted to some filtration $(\mathcal{F}_n)_{n\geqslant 0}$, which is understood as the sequence of states $((\mathrm{s}_n)_{n\geqslant 0})$ occupied and actions $((\mathrm{a}_n)_{n\geqslant 0})$ played by the observer. We assume the following two properties property:

Assumption (Kernel Bounds). *For any given $h > 0$, there exists $\eta, \eta' > 0$ such that, for any balls B_S of $\mathbb{R}^{d_S}$ and B_A of $\mathbb{R}^{d_A}$, of radius greater than h each and respectively included in $\overline{S}$ and $\overline{A}$,*

$$\eta h^D \leqslant \mathbb{P}\Big(\big\{\big(\mathrm{s}_{n+1}, \mathrm{a}_{n+1}\big) \in B_S \times B_A\big\} \,|\, \mathcal{F}_n\Big) \leqslant \eta' h^D. \tag{3.1.1}$$

Assumption (Markov Transition Kernel). *In line with the description of our objective, the process $(\mathrm{s}_n)_{n\geqslant 0}$ must coincide with the observations of a Markov Decision Process (MDP for short) driven by the sequence of actions $(\mathrm{a}_n)_{n\geqslant 0}$. The transition kernel of the MDP is denoted:*

$$\begin{aligned} \overline{\mathsf{P}} : \overline{S} \times \overline{A} &\to \mathcal{P}\big(\overline{S}\big) \\ (s,a) &\mapsto \Big(\overline{\mathsf{P}}\big((s,a),\cdot\big) : E \in \mathcal{B}(S) \mapsto \overline{\mathsf{P}}\big((s,a),E\big)\Big), \end{aligned} \tag{3.1.2}$$

where $\mathcal{B}(\overline{S})$ is the Borel σ-field on $\overline{S}$. As usual with Markov kernels, we require the mapping $\overline{\mathsf{P}}$ to be measurable (which means here that $(s,a) \mapsto \overline{\mathsf{P}}((s,a),E)$ is measurable for the standard Borel σ-fields and for any $E \in \mathcal{B}(S)$). In fact, more assumptions are made below on the regularity with respect to (s,a), using a convenient distance on $\mathcal{P}(S)$.

This notations makes it possible to reformulation part of the identity(3.1.1). For any $n \geqslant 0$, we have

$$\mathbb{P}\Big(\big\{\mathrm{s}_{n+1} \in E\big\} \,|\, \mathcal{F}_n\Big) = \overline{\mathsf{P}}\Big(\big(\mathrm{s}_n, \mathrm{a}_n\big), E\Big), \quad E \in \mathcal{B}(S). \tag{3.1.3}$$

Importantly, the combination of (3.1.1), (**??**) and (3.1.3) puts a constraint on the transition kernel $\overline{\mathsf{P}}$: from (3.1.1) and (**??**), the latter is required to feature some non-degeneracy properties; in words, mass must cover the whole space (this is (3.1.1)) and cannot be concentrated (this is (**??**)). This is a key assumption in our analysis. In the forthcoming Section 3.2, we will reinterpret these conditions as conditions put on an additional exploration noise inserted in the dynamics for the purpose of learning. As far the action process $(\mathrm{a}_n)_{n\geqslant 0}$, assumptions (3.1.1) and (**??**) are in fact less stringent on the model: whereas the dynamics of $(\mathrm{s}_n)_{n\geqslant 0}$ are prescribed by the transition kernel $\overline{\mathsf{P}}$ under study, the actions $(\mathrm{a}_n)_{n\geqslant 0}$ can be chosen exogenously by the 'observer'. In particular, the observer may be able to choose actions that satisfy the non-degeneracy constraints. We will come back to these important points next.

Here comes now the reward function. At each step n, we receive a random reward r_n depending on the state s_n and the action a_n. Whereas more general structures would be conceivable, we require (throughout the analysis) that r_n has the following structure:

Assumption (Reward). *There exists a bounded continuous function $R : \overline{S} \times \overline{A} \to \mathbb{R}$ such that $\mathrm{r}_n = R(\mathrm{s}_n, \mathrm{a}_n)$ for any $n \geqslant 0$. In fact, more conditions will be put next on the regularity of R. Very briefly, R is assumed to have bounded derivatives up to the order $5(\lfloor d_S/2 \rfloor + 1)$, which is the price to pay to obtain tractable convergence rates for the learning algorithm that is presented in the next subsection.*

As we already alluded to several times, we can wrap-up the previous in the definition of a *Markov Decision Process*:

Definition 3.1.1. *For a real $\gamma \in (0,1)$, the 6-tuple $(\overline{S}, \overline{A}, \overline{\mathsf{P}}, R, \gamma)$ forms a Markov decision process, with $\overline{S}$ as state space, $\overline{A}$ as action space, $\overline{\mathsf{P}}$ as transition kernel, R as reward and γ as discount factor.*

A policy π is a measurable function $\pi : \overline{S} \to \overline{A}$. Under a policy π, the value of a state s is

$$V^{\pi}(s) := \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k R\big(\mathrm{s}_k, \pi(\mathrm{s}_k)\big) \;\big|\; \mathrm{s}_0 = s\Big],$$

where $(\mathrm{s}_n)_{n\geqslant 0}$ is the Markov chain associated with the transition kernel $s \in \overline{S} \mapsto \overline{\mathsf{P}}((s,\pi(s)),\cdot) \in \mathcal{P}(\overline{S})$, and the action value function of a pair $(s,a) \in \overline{S} \times \overline{A}$ is

$$Q^{\pi}(s,a) := R(s,a) + \mathbb{E}\Big[\sum_{k=1}^{\infty} \gamma^k R\big(\mathrm{s}_k, \pi(\mathrm{s}_k)\big) \;\big|\; \mathrm{s}_0 = s\Big] = R(s,a) + \gamma \mathbb{E}\big[V^{\pi}(\mathrm{s}_1)|\mathrm{s}_0 = s\big].$$

Solving the MDP is to find a policy $\pi^\star$ such that its value function $V^\star := V^{\pi^\star}$ satisfies

$$V^\star(s) \geqslant V^{\pi}(s), \quad s \in \overline{S},$$

for any other policy π.

We recall that $V^\star$ solves the Bellman equation

$$V^\star(s) = \sup_{a \in \overline{A}} \Big[R(s,a) + \gamma \int_{\overline{S}} V^\star(s') \overline{\mathsf{P}}\big((s,a), ds'\big) \Big],$$

which equation is sometimes formalized as a fixed point for the Bellman optimality operator defined as follows

$$\begin{aligned} \big(\mathcal{T}^{\pi} U\big)(s) &= R\big(s, \pi(s)\big) + \gamma \int_{\overline{S}} U(s') \overline{\mathsf{P}}\big((s,\pi(s), ds'\big), \\ \big(\mathcal{T}^{\star} U\big)(s) &= \sup_{\pi} \big\{ \mathcal{T}^{\pi} U(s) \big\}, \end{aligned} \tag{3.1.4}$$

where U is a test function from $\overline{S}$ to $\mathbb{R}$. We refrain from detailing the functional space in which those test functions are taken. Below, we always assume that the value function $V^\star$ is sufficiently smooth (examples are given in [12]). What is clear is that if $\mathcal{T}^\star U$ maps continuous functions on continuous functions, then $\mathcal{T}^\star$ forms a contraction on the space of continuous functions on $\overline{S}$. Indeed, it is trivial to observe that, for any two continuous functions U and U' on $\overline{S}$,

$$\begin{aligned} \sup_{s \in \overline{S}} \big| \big(\mathcal{T}^\star U - \mathcal{T}^\star U'\big)(s) \big| &\leqslant \sup_{\pi} \sup_{s \in \overline{S}} \big| \big(\mathcal{T}^{\pi} U - \mathcal{T}^{\pi} U'\big)(s) \big| \\ &\leqslant \gamma \sup_{s \in \overline{S}} \big| \big(U - U'\big)(s) \big|. \end{aligned}$$

This argument is used several times in the sequel and is very powerful. The main question here is in fact connected with the choice of the functional space underpinning the contraction and this, in

turn, is connected with the regularity properties of the kernel $\overline{\mathsf{P}}$, which we feel better to present later on in the chapter.

Next, we associate with $V^\star$ the optimal action-value function:

$$Q^\star(s,a) = R(s,a) + \gamma \int_{\overline{S}} V^\star(s')\overline{\mathsf{P}}\big((s,a), ds'\big), \quad (s,a) \in \overline{S} \times \overline{A}. \tag{3.1.5}$$

And we make the following two observations:

1. Since $V^\star \geqslant V^\pi$ for any policy π, one must have

$$Q^\star(s,a) \geqslant Q^\pi(s,a) \quad \forall (s,a) \in \overline{S} \times \overline{A}.$$

2. From the optimal action-value function we can recover the optimal policies by

$$\pi^\star(s) = \operatorname*{argmax}_{a \in A} Q^\star(s,a),$$

provided that we can select in this manner a measurable maximizer. Again we refrain from entering this discussion. We refer for instance to [21] for general results in this direction.

3.1.2 Value iteration and motivation for a kernel based approach

We now address possible strategies for approximating $Q^\star$, which is the real objective in practice. In this regard, it is fair to say that, whenever we have access to full information for the trajectories, or the spaces involved are small and finite, we can directly implement the fixed point iterations underpinning to the Bellman optimality operator (3.1.4) (which we already explained to form a strict contraction in sup norm). For instance, when the problem is set in a finite time horizon $N \in \mathbb{N}$ (and is thus time dependent), the value iteration takes the form backward induction, initialized from the value function V_N at time N:

$$V_n^\star(s) = \sup_{a \in \overline{A}} \Big\{ \mathbb{E}\big[R(\mathrm{s}_n, \mathrm{a}_n) + \gamma \mathbb{E}[V_{n+1}^\star(\mathrm{s}_{n+1}) \,\big|\, (\mathrm{s}_n, \mathrm{a}_n) = (s,a)]\big] \Big\}, \quad \text{for } \; n = N-1, ..., 1, 0,$$

with the analogue of the aforementioned infinite time horizon problem being $V^\star = T^\star V^\star$.

In most reinforcement learning problems this strategy fails because it scales very badly with dimension, which phenomenon is usually referred to as 'curse of dimensionality'. A possibly efficient way to learn the optimal value function is the so called temporal difference (TD) learning. In short, updates of the value function are computed 'on the go', which results in a significantly faster procedure.

Central to this method is the TD-update or *target* that we use to update the value function. In the Q-learning method addressed in the article, the target is $R(s,a) + \gamma \sup_{a' \in \overline{A}} Q(\mathrm{s}'[s], a')$, where $\mathrm{s}'[s]$ is the random state that is reached from s, that is $\mathrm{s}'[s]$ is sampled from the distribution $\overline{\mathsf{P}}((s,a), \cdot)$. Then, the update rule takes the form

$$Q^{n+1}(s,a) = Q^n(s,a) + \alpha \bigg(R(s,a) + \gamma \sup_{a' \in \overline{A}} Q^n(\mathrm{s}'[s], a') - Q^n(s,a) \bigg), \tag{3.1.6}$$

where α is a learning rate in $(0,1)$, possibly depending on the rank of the iteration and on the pair (s,a) at which the approximation Q^{n+1} is computed. Intuitively, we then expect that, under some averaging properties and for a well chosen sequence of learning rates,

$$Q^{n+1}(s,a) \sim Q^n(s,a) + \alpha \int_{\overline{S}} \Big(R(s,a) + \gamma \sup_{a' \in \overline{A}} Q^n(s',a') - Q^n(s,a) \Big) \overline{\mathsf{P}}\big((s,a), ds'\big),$$

which is the very close to the Bellman equation for $Q^\star$.

However, the very first issue here is that the problem is a continuous one. Due to that, we need to apply some form of function approximation on top of Q-learning in order to handle the continuous inputs, but this might not be so easy. Indeed, it is well known from [104] when function approximation, bootstrapping (TD learning) and off policy TD control are combined, they form a 'deadly triad' that can possibly create divergences in the learning scheme, see also [108] for an example of divergence occurring under linear function approximation. In contrast, we follow [27] and make the choice for kernel approximation. The main motivation for this comes from the fact that kernel regression is non-expanding operation and thus does not cause any additional divergence in the Bellman operator. For a general discussion on kernel regression, we refer to [69].

The regression relies on the following:

Assumption (Regression Kernel). *We call $\mathscr{K} : \mathbb{R}^{d_S} \times \mathbb{R}^{d_A} \to \mathbb{R}$ a smooth non-negative compactly supported function that is (strictly) positive on the D-dimensional ball $B(0,1)$ of center 0 and of radius 1. We denote by $\|\mathscr{K}\|_\infty$ its L^∞ norm and we let $\lambda_{\mathscr{K}} = \inf_{B(0,1)} \mathscr{K}$. We let $\varrho > 1$ be the smallest real such that the support of $\mathscr{K}$ is in $B(o,\rho) = \rho B(0,1)$.*

For a bandwidth h that represents the radius at which estimation is performed, we use the notation $\mathscr{K}_h$ for

$$\mathscr{K}_h(s,a) = \mathscr{K}\Big(\frac{s}{h}, \frac{a}{h}\Big), \quad (s,a) \in \overline{S} \times \overline{A}. \tag{3.1.7}$$

We can refine the upper boundedness assumption of the transition kernel, assuming that there exists another constant $\eta' > 1$ such that, for any D-dimensional ball of radius $3\rho h$,

$$\mathbb{P}\Big(\big\{ \big(\mathrm{s}_{n+1}, \mathrm{a}_{n+1}\big) \in B_{2\rho h} \big\} \,|\, \mathcal{F}_n \Big) \leqslant \eta' h^D. \tag{3.1.8}$$

Here is now how the regression kernel is used in (3.1.6). For a given realization $(\mathrm{s}_n, \mathrm{a}_n)_{0 \leqslant n \leqslant N}$ as in the previous subsection, we then introduce the sequence of updating ratios $\alpha_n : \overline{S} \times \overline{A} \ni (s,a) \mapsto \alpha_n(s,a)$ and $n \in \mathbb{N}$, defined by

$$\alpha_n(s,a) = \begin{cases} \dfrac{\mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)}{\sum_{j=0}^n \mathscr{K}_h\big(s - \mathrm{s}_j, a - \mathrm{a}_j\big)} & \text{if} \quad \displaystyle\sum_{j=0}^n \mathscr{K}_h\big(s - \mathrm{s}_j, a - \mathrm{a}_j\big) > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{3.1.9}$$

Implicitly, α_n is a random field (as it depends on the realization of $(\mathrm{s}_n, \mathrm{a}_n)_{n \geqslant 0}$). Notice also that we feel better not to include the parameter h in the notation at this stage of the analysis since h is fixed.

The following definition is the basis for clarifying the rule (3.1.6):

Definition 3.1.2. *Given an integer $n \geqslant 0$ and given a realization $(\mathrm{s}_k, \mathrm{a}_k)_{0 \leqslant k \leqslant n}$ satisfying the properties highlighted in Subsection 3.1.1, the kernel approximation of a function $f : \overline{S} \times \overline{A} \to \mathbb{R}$ at a point (s, a) and at depth n is given by*

$$\mathcal{A}_{h,n} f(s,a) := \sum_{k=0}^{n} \frac{\mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) f(\mathrm{s}_k, \mathrm{a}_k)}{\sum_{k=0}^{n} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)}, \tag{3.1.10}$$

provided the denominator is not zero. If the denominator is zero, we return 0 for $\mathcal{A}_{h,n} f(s,a)$. The approximation satisfies the recursion property:

$$\mathcal{A}_{h,n} f(s,a) = \alpha_n(s,a) f(\mathrm{s}_n, \mathrm{a}_n) + \big(1 - \alpha_n(s,a)\big) \mathcal{A}_{h,n-1} f(s,a). \tag{3.1.11}$$

The recursion property is reminiscent of (3.1.6). It can be proven as follows:

$$\begin{aligned}
&\sum_{k=0}^{n} \frac{\mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) f(\mathrm{s}_k, \mathrm{a}_k)}{\sum_{k=0}^{n} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)} \\
&= \alpha_n(s,a) f(\mathrm{s}_n, \mathrm{a}_n) + \sum_{k=0}^{n-1} \frac{\mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) f(\mathrm{s}_k, \mathrm{a}_k)}{\sum_{k=0}^{n} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)} \\
&= \alpha_n(s,a) f(\mathrm{s}_n, \mathrm{a}_n) + \big(1 - \alpha_n(s,a)\big) \sum_{k=0}^{n-1} \frac{\mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) f(\mathrm{s}_k, \mathrm{a}_k)}{\sum_{k=0}^{n-1} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)},
\end{aligned}$$

at least when all the denominators are non-zero. When one of them is zero, the proof is straightforward.

A key property is that the weights entering the definition of $\mathcal{A}_{h,n} f$ is an element of the simplex of dimension $n-1$ (i.e., the weights form a convex combination), from which we deduce the following two properties:

1. Whenever $\sum_{k=0}^{n} \mathscr{K}_h(s - \mathrm{s}_k, a - \mathrm{a}_k) > 0$ and f is Lipschitz with Lipschitz constant L, the distance between f and $\mathcal{A}_{h,N} f$ is less than Lh, namely

$$|\mathcal{A}_{h,N} f(s,a) - f(s,a)| = \left| \sum_{k=0}^{n} \frac{\mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) \big(f(\mathrm{s}_n, \mathrm{a}_n) - f(s,a)\big)}{\sum_{k=0}^{n} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)} \right| \leqslant Lh. \tag{3.1.12}$$

2. The operator $\mathcal{A}_{h,N}$ preserves the sup-norm which guarantees that, in turn, the contraction property of the Bellman optimality operator under a kernel approximation is preserved. We write this in the form

$$\|\mathcal{A}_{h,N} \mathcal{T}^\star U - \mathcal{A}_{h,N} \mathcal{T}^\star U'\|_\infty \leqslant \gamma \|U - U'\|_\infty, \tag{3.1.13}$$

   where U and U' are functions on $\overline{S}$ with the required form of regularity.

As stated in [94] , the moral behind is that kernel regression approximation offers a convenient 'plug-in' estimate of the value functions. This motivates our choice for kernel based Q-learning. In

clear, we can define the Nadaraya-Watson estimator for the Q-function. First, we adapt the update rule (3.1.6) and let (for $n \geqslant 1$)

$$\widehat{Q}_n(s,a) = \widehat{Q}_{n-1}(s,a) + \alpha_{n-1}(s,a)\Big(R(\mathrm{s}_{n-1},\mathrm{a}_{n-1}) + \gamma \mathrm{sup}_{a' \in A}\widehat{Q}_{n-1}(\mathrm{s}_n,a') - \widehat{Q}_{n-1}(s,a)\Big). \quad (3.1.14)$$

Thus, by the recursion formula (3.1.11), we have

$$\widehat{Q}_n(s,a) = \begin{cases} \dfrac{\sum_{j=0}^{n-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big) y_j}{\sum_{0=1}^{n-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)} & \text{if} \quad \sum_{0=1}^{n-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big) > 0 \\ 0 & \text{otherwise} \end{cases}, \quad (3.1.15)$$

where

$$y_j = R(\mathrm{s}_j,\mathrm{a}_j) + \gamma \sup_{a \in \overline{A}} \widehat{Q}_j(\mathrm{s}_{j+1},a). \quad (3.1.16)$$

Notice that the computation of $\widehat{Q}_n$ requires $n+1$ observations from the Markov decision process (as the latter starts from time 0). At the initial time, we choose $\widehat{Q}_0 \equiv 0$.

Last but not least, to rephrase [27]: equation (3.1.14) is the right one for comparing with TD learning, but equation (3.1.15) is the useful one for numerical purposes.

To wrap-up everything, we present a pseudo code of the algorithm in an episodic fashion to be closer to existing methods of implementation in reinforcement learning community.

Algorithm 2: Kernel Based Q-Learning

input : Type of kernel $\mathscr{K}$, bandwidth h, discount γ, number of iteration n
output: Approximate Action Value Function $\widehat{Q}_n$

initialization:;
$\widehat{Q}_0(s,a) = 0 \ \forall (s,a)$;
set inital state s_0;
for k *in* n **do**
  Choose action a_k;
  Get: reward r_k, next state s';
  Compute $y_k = r_k + \gamma \underset{a \in \bar{A}}{max}\, \widehat{Q}_k(s',a)$ using (3.1.15) ;
  Store in memory (s_k, a_k, y_k);
  $k = k+1$;
  $s_k = s'$;

Given the nature of memory based approximations, essentially all we need is just successive iterations $\mathrm{s}_n, \mathrm{a}_n, \mathrm{y}_n, \mathrm{s}_{n+1}, \cdots$

For sure, a natural question concerns the choice of a_n at each iteration. In this regard, it should be stressed that the ε-greedy strategy (that consists in sampling a_n uniformly with probability ε and, otherwise, in choosing it as the minimizer of the current approximation of the Q-function does not satisfy the assumption **(Upper Bound Kernel)**, because it may induce accumulation of mass phenomena. This is clearly a drawback of our proof which entirely relies on the mixing properties of the kernels, as stated in the two assumptions **(Lower Bound Kernel)** and **(Upper Bound Kernel)**. Instead, one may just sample a_n: this satisfies the conditions prescribed in our analysis.

In our analysis below, there is only one episode. We refer to Section 3.5 for numerical aspects in which several episodes are run.

3.1.3 Summary of main results

It is well known that memory based techniques can approximate any function with arbitrary precision, at the expense of memory. However, memory costs may take a dramatic turn in high dimension. In the current framework, one wants in particular to return a relevant approximation of the function $Q^\star$, but with a rather reasonable number of observations. In contrast, the work of Carden [27] asserts that one can reach any desired accuracy provided that the number of observations is large enough. This result may be however intractable in practice if the required number of observations is much too large. Intuitively, such a picture may indeed happen when the dimensions d_S and d_A become larger and larger. Here, we are able to resolve the curse of dimensionality, from which these methods usually suffer by resorting to smoother functions in the appropriate Sobolev spaces.

This prompts us to let the following two assumptions:

Assumption (Regularity Cost and Transition Kernel). *We assume that the function R has bounded derivatives of any order up to the order $5(\lfloor d_S/2 \rfloor + 1)$. And, we assume that, for any function $\varphi : \overline{S} \to \mathbb{R}$, whose derivatives up to a certain order $k \in \{1, \cdots, 5(\lfloor d_S/2 \rfloor + 1)\}$ are bounded by a certain constant C, the function*

$$(s,a) \in \overline{S} \times \overline{A} \mapsto \int_{\overline{S}} \varphi(s') \overline{\mathsf{P}}\big((s,a), ds'\big)$$

also has bounded derivatives up to k, with bounds only depending on C.

The order $5(\lfloor d_S/2 \rfloor + 1)$ here, is found from Sobolev embedding theorems, which play a great role in the mathematical analysis that is provided next and which draw a clear connection between the dimension of the state and action spaces and the required regularity on the data. (As we already mentioned several times, this is our choice to impose (very) strong assumptions on the coefficients in order to get tractable rates of convergence with an affordable complexity.)

In parallel, we also assume:

Assumption (Regularity Value Function) *We assume that the function $V^\star$ has bounded derivatives of any order up to the order $5(\lfloor d_S/2 \rfloor + 1)$.*

Obviously, this assumption is implicit. In general, one cannot directly deduce that $V^\star$ is smooth from the simple fact that the data are smooth. However, we have additional structural conditions under which $V^\star$ is indeed smooth. We provide examples in Subsection 3.2.6.

In clear, the main objective of the article is to address the rate convergence of $\widehat{Q}_{h,n}(s,a)$ to $Q^\star(s,a)$ when n becomes large and h becomes small. Our main result in this regard can be summarized through the following meta-statement:

Main Meta-Statement. *There exists a constant C, depending on the various parameters underpinning the aforementioned assumptions, such that, for an error threshold $\varepsilon > 0$, we can find h_ε and n_ε such that the* sup *distance between $\widehat{Q}_{h_\varepsilon, n_\varepsilon}$ and $Q^\star$ is less than $C(1 + |\ln(\varepsilon)|)\varepsilon$ on an*

event of probability greater than $1 - C\varepsilon^D$. *The number* n_ε *that is necessary to do so is less than* $\exp(C|\ln(\varepsilon)|^3)$.

A more rigorous statement is given in Theorem 3.2.8 and Proposition 3.2.10 below (it is fair to say that the latter result is stated in the framework of mean field control, but the spirit is very much the same). Notice that Proposition 3.2.10 in fact includes the case when the original kernel $\overline{\mathsf{P}}$ does not satisfy the non-degeneracy properties states in Assumptions **(Kernel Lower Bounds)** and **(Kernel Upper Bounds)**. The idea in this case is to add an additional exploration noise. What Proposition 3.2.10 says is that the function $Q^\star$ (for the model that does not satisfy the non-degeneracy properties) can be approximated with the same rate before and for the same order of observations as in the meta-statement but on an event of probability greater than $1 - C\varepsilon$ (instead of $1 - C\varepsilon^D$). This difference quantifies the exploration-exploitation trade-off in this setting.

3.1.4 Literature review

Obviously our work inherits from the classical works on the field, see for instance [99, 104], both in the way the problem is formulated and in the way the solution strategy is implemented. At the same time we distinguish ourselves in a few important points. Here we avoid a general literature review since the literature on the subject is quite vast and mature. Instead we focus on few selected contributions that relate best to our work and put it into perspective.

In [94], the authors provide a complete analysis of kernel based reinforcement learning with respect to Approximate Value Iteration (AVI) methods. In a reinforcement learning problem defined on continuous states and discrete actions, they prove consistency of the estimates of the value function when a kernel regression operator similar to (3.1.10) is used. They decompose the approximation error in terms of a bias and a variance term (that is very often the case when kernel methods are used, see [69, Chapter 6]) and choose an optimal 'shrinkage rate' for the bandwidth of the kernels h to reduce bias and variance. In contrast, in our analysis we keep h fixed throughout the proof but adapted to the data of the problem. In particular, the final estimates reported in the statement of the Meta-Theorem is for an h that is chosen in terms of the regularity of the value function and the size of the spaces involved. In this regard, it is worth insisting on the fact that our proof gives a clear insight on the averaging effect underlying the learning procedure.

In the family of kernel based reinforcement learning and specifically kernel based Q-learning in continuous state and action spaces, we mention [77] where the authors use a Reproducing Kernel Hilbert Space (RKHS) approach. Instead of 'plug-in' approximation by kernel regression, they resort to a regularized Bellman equation whose fixed point is identified via a reformulation of a functional descent in a RKHS. The authors highlight the nested expectation nature of Q-learning that they treat via a two timescale stochastic approximation approach. The resulting algorithm is a variant of Q-learning in which the authors control the complexity by imposing a memory compression by sparse projection to lower dimensional subspaces. Similarly to their approach, we assume here that the Q-function belongs to a suitable (Hilbert) space of regular functions, as result of which we can control the complexity. As far as the algorithm is concerned, the update rule we use for the value function is quite different and allows us to analyse the error per iteration. In particular, we are here able to bound directly the distance (at each iteration) between the learnt and true Q-functions.

Last, we mention [50], which also addresses kernel based reinforcement learning. The authors use an episodic model-based optimistic algorithm called Kernel-UCBVI (Upper Confidence Bound Value Iteration) to solve continuous MDPs under relatively weak assumptions. Even though their algorithm and approach are completely different (model based vs model free), we feel that their mathematical analysis of the convergence is very much of the same flavor as our work. The authors derive concentration inequalities for the transition kernels applied to the value function, in their notation $|(\widehat{P}_n^k - P_n)V_n^*|$[1], which is reminiscent of the computations exposed in Subsection 3.2.5 below. However the form of Value Iteration is different because of the added exploration bonuses which relay on counts of visits for state action pairs and thus their final regret relies on the convering dimension instead of the physical one and is not directly comparable to ours.

3.1.5 Organisation

We present a strategy of proof together with some refined versions of the main meta-statement in the subsequent Section 3.2. Therein, we also expose the application to mean field control, in agreement with the objective of this manuscript, and we give some further directions of research. The core is of the proof is exposed in Sections 3.3 and 3.4. Section 3.5 is dedicated to some numerical illustration. In Section 3.6, we give the proofs of some auxiliary results.

3.2 General structure of the proof

The purpose of this section is to present the main steps of the proof towards the main result of this chapter. The main intuition follows from the original works by Watkins (and co-author) [114, 115].

3.2.1 Definition and transitions of the Action Replay Process

The analysis relies on the notion of 'Action Replay Process' (ARP), in a strategy similar to the original one conceived by Watkins (and Dayan) [114, 115] in the discrete setting and then extended by Carden [27] to continuous spaces.

The construction of the ARP is as follows. We associate with a realization $(\underline{\mathrm{s}}, \underline{\mathrm{a}}) = (\mathrm{s}_n, \mathrm{a}_n)_{n\in\mathbb{N}}$ of the Markov decision process introduced in Definition 3.1.1 a new Markov decision process whose transitions explicitly depend on the realization $(\underline{\mathrm{s}}, \underline{\mathrm{a}})$ (and are thus random) and whose Q-function coincides exactly with the function $\widehat{Q}_n(s,a)$ defined in (3.1.14). In the latter, the additional parameter n is seen as part of the state variable since the ARP lives on an extended space comprising the time variable. As we will see next, the ARP is in fact a Markov process with values in $\mathbb{N}\times\overline{S}\times\overline{A}$, with the key feature that the transitions only allow the time component to decrease. States with a time component that is equal to 0 are absorbing.

Definition 3.2.1. *Let $(\underline{\mathrm{s}}, \underline{\mathrm{a}}) = (\mathrm{s}_n, \mathrm{a}_n)_{n\in\mathbb{N}}$ be a realization of the Markov decision process satisfying Definition 3.1.1 and* (3.1.3) *Then, for an element $(n,s,a)\in\mathbb{N}\times\overline{S}\times\overline{A}$, we define a probability measure*

$$\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\big((n,s,a),\cdot\big)$$

[1] where $\widehat{P}_n^k$ is the transition kernel that is identified via a weighted sum of Dirac measures, for weights similar to (3.1.9)

on $\mathbb{N} \times \overline{S} \times \overline{A}$ *as follows. If* $n = 0$*, then* $\overline{\Pi}_{(\underline{s},\underline{a})}\big((0,s,a),\cdot\big)$ *is the Dirac point mass at* $(0,s,a)$*.*

If $n \geqslant 1$*, we consider a sequence* $(U_1, \cdots, U_n)$ *of independent random variables with uniform distribution on* $(0,1)$*. Then, for* $(\alpha_k(s,a))_{0\leqslant k\leqslant n-1}$ *being defined as in* (3.1.9)*, we call* τ *the following random time*

$$\tau = \max\Big\{k \in \{1, \cdots, n\} : \alpha_{k-1}(s,a) > U_k\Big\},$$

with τ *being equal to* 0 *if all the events in the maximum are empty, and we define the transition probability* $\overline{\Pi}_{(\underline{s},\underline{a})}((n,s,a),\cdot)$ *as the law of*

$$\begin{aligned}\Big(\tau - 1, \mathrm{s}_\tau, \mathrm{a}_\tau\Big) \quad &\textit{if} \quad \tau \geqslant 1,\\ \Big(0,s,a\Big) \quad &\textit{if} \quad \tau = 0.\end{aligned}$$

It should be noticed that, under the condition $\sum_{k=0}^{n-1} \mathscr{K}_h(s - \mathrm{s}_k, a - \mathrm{a}_k) > 0$, we necessarily have $\tau \geqslant 1$ since $\alpha_k(s,a) = 1$ for at least one $k \in \{0, \cdots, n-1\}$. Indeed, if we call k the smallest integer such that $\mathscr{K}_h(s - \mathrm{s}_k, a - \mathrm{a}_k) > 0$, then, necessarily, $\alpha_k(s,a) = 1$. On the opposite, $\tau = 0$ if $\sum_{k=0}^{n-1} \mathscr{K}_h(s - \mathrm{s}_k, a - \mathrm{a}_k) = 0$. In particular, it should be observed that $\overline{\Pi}_{(\underline{s},\underline{a})}((n,s,a),\cdot)$ is always well-defined, even though the trajectory $(\underline{s},\underline{a})$ has not visited the neighborhood of (s,a). Moreover, we emphasize that the states that the process reaches upon termination are arbitrary and might differ at each realisation.

Observe also that the transitions starting from a time component $n \geqslant 1$ only depend on the observations of $(\mathrm{s}_k, \mathrm{a}_k)$ at times $k = 0, \cdots, n-1$. Below, we denote by $\mathbf{P}_{(\underline{s},\underline{a})}$ a probability measure on an auxiliary probability space $(\Xi, \mathcal{G})$ together with a process $(\Lambda_k, \Sigma_k, B_k)_{k\in\mathbb{N}}$ also constructed on $(\Xi, \mathcal{G})$ with values in $\mathbb{N}\times\overline{S}\times\overline{A}$ such that, under $\mathbf{P}_{(\underline{s},\underline{a})}$, the process $(\Lambda_k, \Sigma_k, B_k)_{k\in\mathbb{N}}$ is a homogeneous Markov chain with transition probabilities $\big(\overline{\Pi}_{(\underline{s},\underline{a})}\big((n,s,a),\cdot\big)_{n\in\mathbb{N},s\in\overline{S},a\in\overline{A}}$, namely

$$\mathbf{P}_{(\underline{s},\underline{a})}\Big(\Big\{(\Lambda_{k+1}, \Sigma_{k+1}, B_{k+1}) \in E\Big\} \,|\, \mathcal{G}_k^{(\Lambda,\Sigma,B)}\Big) = \overline{\Pi}_{(\underline{s},\underline{a})}\big((\Lambda_k, \Sigma_k, B_k), E\big), \tag{3.2.1}$$

for E a Borel subset of $\mathbb{N} \times \overline{S} \times \overline{A}$, where $\mathbb{G}^{(\Lambda,\Sigma,B)} = (\mathcal{G}_k^{(\Lambda,\Sigma,B)})_{k\in\mathbb{N}}$ is the filtration generated by $(\Lambda_k, \Sigma_k, B_k)_{k\in\mathbb{N}}$.

The following lemma clarifies the semi-group induced by the ARP, with the relationship below playing a key role in the subsequent analysis.

Lemma 3.2.2. *Let* $\psi : \mathbb{N} \times \overline{S} \times \overline{A} \to \mathbb{R}$ *be a bounded and measurable function. Then, for any fixed realization* $(\underline{\mathrm{s}}, \underline{\mathrm{a}})$ *of the Markov process* $(\mathrm{s}_n, \mathrm{a}_n)_{n\geqslant 0}$*, for any* $n \in \mathbb{N}$*, for any* $(s,a) \in \bar{S} \times \bar{A}$*, such that*

$$\sum_{l=0}^{n-1} \mathscr{K}_h\big(s - \mathrm{s}_l, a - \mathrm{a}_l\big) > 0, \tag{3.2.2}$$

the following identity holds true:

$$\mathbf{E}_{(\underline{s},\underline{a})}\Big[\psi\big(\Lambda_1, \Sigma_1, B_1\big) \,|\, (\Lambda_0, \Sigma_0, B_0) = (n,s,a)\Big] = \sum_{k=0}^{n-1} \frac{\mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)\psi\big(k, \mathrm{s}_{k+1}, \mathrm{a}_{k+1}\big)}{\sum_{l=1}^{n} \mathscr{K}_h\big(s - \mathrm{s}_l, a - \mathrm{a}_l\big)}.$$

otherwise, whenever $\sum_{l=0}^{n-1} \mathscr{K}_h\big(s - \mathrm{s}_l, a - \mathrm{a}_l\big) = 0$*,*

$$\mathbf{E}_{(\underline{s},\underline{a})}\Big[\psi\big(\Lambda_1, \Sigma_1, B_1\big) \,|\, (\Lambda_0, \Sigma_0, B_0) = (n,s,a)\Big] = \psi(0,s,a).$$

Proof. We fix the initial condition $(n, s, a) \in \mathbb{N} \times \overline{S} \times \overline{A}$ of the ARP. Then, following Definition 3.2.1, we introduce the stopping time

$$\tau = \max\Big\{k \in \{1, \cdots, n\} : \alpha_{k-1}(s, a) > U_k\Big\},$$

for the same sequence $(U_1, \cdots, U_n)$ as therein. We recall that $\tau \geqslant 1$ under the condition (3.2.2). Moreover,

$$\mathbf{P}_{(\underline{s},\underline{a})}\Big(\big\{\tau = k\big\}\Big) = \Big[\prod_{j=k}^{n-1}\big(1 - \alpha_j(s, a)\big)\Big]\alpha_{k-1}(s, a), \quad k \in \{1, \cdots, n\},$$

with the convention that the product is equal to 1 if $k = n$. Then, for ψ as in the statement,

$$\mathbf{E}_{(\underline{s},\underline{a})}\Big[\psi\big(\Lambda_1, \Sigma_1, B_1\big) \,|\, (\Lambda_0, \Sigma_0, B_0) = (n, s, a)\Big] = \mathbf{E}_{(\underline{s},\underline{a})}\Big[\psi\big(\tau - 1, \mathrm{s}_\tau, \mathrm{a}_\tau\big)\Big],$$

and

$$\begin{aligned}\mathbf{E}_{(\underline{s},\underline{a})}\Big[\psi\big(\tau - 1, \mathrm{s}_\tau, \mathrm{a}_\tau\big)\Big] &= \sum_{k=1}^{n}\Big[\Big(\prod_{j=k}^{n-1}\big(1 - \alpha_j(s, a)\big)\Big)\alpha_{k-1}(s, a)\psi\big(k - 1, \mathrm{s}_k, \mathrm{a}_k\big)\Big] \\ &= \sum_{k=0}^{n-1}\Big[\Big(\prod_{j=k+1}^{n-1}\big(1 - \alpha_j(s, a)\big)\Big)\alpha_k(s, a)\psi\big(k, \mathrm{s}_{k+1}, \mathrm{a}_{k+1}\big)\Big].\end{aligned}$$

For a given j such that $\alpha_j(s, a) > 0$ (which implies in particular that $\sum_{l=0}^{j}\mathscr{K}_h(s - \mathrm{s}_l, a - \mathrm{a}_l) > 0$), we have

$$1 - \alpha_j(s, a) = \frac{\sum_{l=1}^{j-1}\mathscr{K}_h\big(s - \mathrm{s}_l, a - \mathrm{a}_l\big)}{\sum_{l=1}^{j}\mathscr{K}_h\big(s - \mathrm{s}_l, a - \mathrm{a}_l\big)},$$

and then, for a given k such that $\alpha_k(s, a) > 0$,

$$\prod_{j=k+1}^{n-1}\big(1 - \alpha_j(s, a)\big) = \frac{\sum_{l=0}^{k}\mathscr{K}_h\big(s - \mathrm{s}_l, a - \mathrm{a}_l\big)}{\sum_{l=0}^{n-1}\mathscr{K}_h\big(s - \mathrm{s}_l, a - \mathrm{a}_l\big)},$$

so that

$$\prod_{j=k+1}^{n-1}\big(1 - \alpha_j(s, a)\big)\alpha_k(s, a) = \frac{\mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)}{\sum_{l=0}^{n-1}\mathscr{K}_h\big(s - \mathrm{s}_l, a - \mathrm{a}_l\big)}.$$

If $\alpha_k(s, a) = 0$, the equality still holds true (recalling that the denominator in the right-hand side cannot be 0 in our setting). In the end, we can write

$$\mathbf{E}_{(\underline{s},\underline{a})}\Big[\psi\big(\tau - 1, \mathrm{s}_\tau, \mathrm{a}_\tau\big)\Big] = \sum_{k=0}^{n-1}\frac{\mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)\psi\big(k, \mathrm{s}_{k+1}, \mathrm{a}_{k+1}\big)}{\sum_{l=1}^{n}\mathscr{K}_h\big(s - \mathrm{s}_l, a - \mathrm{a}_l\big)},$$

which completes the proof when (3.2.2) holds true.

When (3.2.2) does not hold, the identity follows from the fact that $\overline{\Pi}_{(\underline{s},\underline{a})}((n, s, a), \cdot)$ is the delta mass at $(0, s, a)$. □

3.2.2 Q-function associated with the ARP

In order to complete our description of the ARP as an MDP, we now associate with it the following optimization problem:

$$V^{\mathrm{ARP},\star}(n,s) = \sup_{\pi} \mathbf{E}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\Big[\mathbf{1}_{\{\tau\geqslant 1\}} \sum_{k=0}^{\infty} \gamma^k R\big(\Sigma_k, \pi(\Lambda_k,\Sigma_k)\big) \,|\, (\Lambda_0,\Sigma_0) = (n,s)\Big],$$

with the supremum being taken over strategies π from $\mathbb{N}\times\overline{S}$ into $\overline{A}$ (be careful that strategies are time dependent), and where the variable τ in the first line is implicitly understood as in Definition 3.2.1.

Following (3.1.5), the optimal action-value function $Q^{\mathrm{ARP},\star}$ is the solution of

$$\begin{aligned} Q^{\mathrm{ARP},\star}\big(n,s,a\big) &= R(s,a)\mathbf{P}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\big(\{\tau\geqslant 1\}\,|\,(\Lambda_0,\Sigma_0,B_0) = (n,s,a)\big) \\ &\quad + \gamma\, \mathbf{E}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\Big[\sup_{a'\in\bar{A}} Q^{\mathrm{ARP},\star}\big(\Lambda_1,\Sigma_1,a'\big)\,|\,(\Lambda_0,\Sigma_0,B_0) = (n,s,a)\Big], \end{aligned} \tag{3.2.3}$$

In fact, the effective connection with $V^{\mathrm{ARP},\star}$ is not used in the rest of the analysis, and we can regard (3.2.3) as an autonomous equation. As shown in the proof of Lemma 3.2.3, any solution to (3.2.3) must satisfy $Q^{\mathrm{ARP},\star}\big(0,s,a\big) = 0$. The values of $Q^{\mathrm{ARP},\star}(n,s,a)$, for $n\geqslant 1$, are then given by the recursion formula (which follows from Lemma 3.2.2):

$$Q^{\mathrm{ARP},\star}(n,s,a) = R(s,a) + \gamma \sum_{k=0}^{n-1} \frac{\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)}{\sum_{l=1}^{n}\mathscr{K}_h\big(s-\mathrm{s}_l,a-\mathrm{a}_l\big)} \sup_{a'\in\overline{A}} Q^{\mathrm{ARP},\star}(k,s_{k+1},a'), \tag{3.2.4}$$

whenever (3.2.2) is satisfied (recalling that τ is necessarily greater than 1 under the standing condition (3.2.2)). When the latter does not hold, we have $Q^{\mathrm{ARP},\star}(n,s,a) = Q^{\mathrm{ARP},\star}(0,s,a) = 0$. This solves the equation. Measurability can be argued as in the first step of the proof of Lemma 3.2.3 below.

To conclude this introductory section we can show that the ARP and the kernel based algorithm of Section 2 have (almost) the same optimal Q-values (this should be not a big surprise in light of (3.2.4), which is very close to (3.1.15)). The following lemma is indeed a reformulation of [27, Lemma 1]:

Lemma 3.2.3. *We have*

$$\sup_{n\in\mathbb{N}} \sup_{(s,a)\in\bar{S}\times\bar{A}} \Big|Q^{\mathrm{ARP},*}(n,s,a) - \widehat{Q}_n(s,a)\Big| \leqslant Ch.$$

Proof. First Step. By a straightforward induction, we see that, for any $n\in\mathbb{N}$, the function $(s,a)\mapsto\widehat{Q}_n(s,a)$ is continuous, which proves in particular that the function $s\mapsto\sup_{a'\in\bar{A}}\widehat{Q}_n(s,a')$ is measurable.

Next, the very key point is to reformulate the right-hand side in (3.1.15) as an expectation with respect to the variable τ introduced in Definition 3.2.1. Indeed, if

$$\sum_{j=0}^{n-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big) > 0, \tag{3.2.5}$$

then

$$\widehat{Q}_n(s,a) = \frac{\sum_{j=0}^{n-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big) R(\mathrm{s}_j,\mathrm{a}_j)}{\sum_{j=1}^{n-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)} + \gamma \frac{\sum_{j=1}^{n-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big) \sup_{a'\in A} \widehat{Q}_{j-1}(\mathrm{s}_j,a')}{\sum_{j=0}^{n-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)}.$$

We now invoke Lemma 3.2.2. On the event $\big\{\sum_{l=0}^{n-1} \mathscr{K}_h\big(s-s_l, a-a_l\big) \geqslant 1\big\}$,

$$\begin{aligned}\widehat{Q}_n(s,a) = {} & \frac{\sum_{j=0}^{n-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big) R(\mathrm{s}_j,\mathrm{a}_j)}{\sum_{j=1}^{n-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)} \\ & + \gamma \mathbf{E}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\Big[\sup_{a'\in\bar{A}} \widehat{Q}_{\Lambda_1}\big(\Sigma_1,a'\big) \,|\, (\Lambda_0,\Sigma_0,B_0) = (n,s,a)\Big].\end{aligned}$$

Here, we observe from (3.1.12) and from the Lipschitz regularity of the function R that

$$\left|R(s,a) - \frac{\sum_{j=0}^{n-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big) R(\mathrm{s}_j,\mathrm{a}_j)}{\sum_{j=1}^{n-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)}\right| \leqslant Ch.$$

We easily deduce from (3.2.3) (recalling again that τ is necessarily greater than 1 under the standing condition (3.2.5)) that

$$\sup_{(s,a)\in\bar{S}\times\bar{A}} \Big|Q^{\mathrm{ARP},\star}(n,s,a) - \widehat{Q}_n(s,a)\Big| \leqslant Ch + \gamma \sup_{n\in\mathbb{N}} \sup_{(s,a)\in\bar{S}\times\bar{A}} \Big|Q^{\mathrm{ARP},\star}(n,s,a) - \widehat{Q}_n(s,a)\Big|, \tag{3.2.6}$$

under the condition (3.2.5), i.e. $\sum_{j=0}^{n-1} \mathscr{K}_h(s-\mathrm{s}_j, a-\mathrm{a}_j) > 0$.

Second Step. Now, if (3.2.5) fails, that is

$$\sum_{j=0}^{n-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big) = 0,$$

then, by construction,

$$\gamma \mathbf{E}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\Big[\sup_{a'\in\bar{A}} Q^{\mathrm{ARP},\star}\big(\Lambda_1,\Sigma_1,a'\big) \,|\, (\Lambda_0,\Sigma_0,B_0) = (n,s,a)\Big] = \gamma \sup_{a'\in\bar{A}} Q^{\mathrm{ARP},\star}\big(0,s,a'\big),$$

and

$$Q^{\mathrm{ARP},\star}\big(n,s,a\big) = \gamma \sup_{a'\in\bar{A}} Q^{\mathrm{ARP},\star}\big(0,s,a'\big).$$

In particular, when $n = 0$, we have that

$$Q^{\mathrm{ARP},\star}\big(0,s,a\big) = \gamma \sup_{a'\in\bar{A}} Q^{\mathrm{ARP},\star}\big(0,s,a'\big),$$

which yields

$$Q^{\mathrm{ARP},\star}\big(0,s,a\big) = 0,$$

and then

$$Q^{\mathrm{ARP},\star}\big(n,s,a\big) = 0,$$

(under the condition $\sum_{0=1}^{n-1} \mathscr{K}_h(s-\mathrm{s}_j, a-\mathrm{a}_j)=0$). By definition of $\widehat{Q}_n(s,a)$, we thus have (under the same condition as before) that

$$Q^{\mathrm{ARP},\star}\big(n,s,a\big)=\widehat{Q}_n(s,a).$$

Conclusion. Back to (3.2.6), we get

$$\sup_{(n,s,a)\in\mathbb{N}\times\bar{S}\times\bar{A}}\Big|Q^{\mathrm{ARP},\star}(n,s,a)-\widehat{Q}_n(s,a)\Big|\leqslant Ch+\gamma\sup_{n\in\mathbb{N}}\sup_{(s,a)\in\bar{S}\times\bar{A}}\Big|Q^{\mathrm{ARP},*}(n,s,a)-\widehat{Q}_n(s,a)\Big|,$$

from which the result easily follows. □

3.2.3 Repeated covering times

The purpose of this subsection is to provide two results that revisit some of the arguments introduced in [27] about covering times of the process $(\mathrm{s}_n,\mathrm{a}_n)_{n\in\mathbb{N}}$ and in particular to give quantitative bounds for those covering times. Intuitively, the covering time is the time duration that is needed for the MDP to visit all the cells that are attached to the kernel based algorithm. This is a very important concept in probability theory and this is exactly the point where we can quantify the impact of the mixing properties of the noise.

The results are rather technical, but they have the great advantage to be explicit. Also they depend on a series of parameters that may be interpreted as follows:

1. ℓ below is a time at which we start to study the covering times;
2. m is time duration over which we study the covering times;
3. h is the bandwidth in the kernel $\mathscr{K}_h$, see (3.1.7);
4. J counts the number of cells of radius h in the domain $\overline{S}\times\overline{A}$ and is order h^{-D};
5. η and η' are the mixing parameters in (3.1.1) and (**??**);
6. β and δ are two macroscopic free parameters whose values are fixed in the end only.

Using these parameters, the following statement provides a quantitative version of Lemma 2 in [27]:

Proposition 3.2.4. *Assume that $\eta' h^D\leqslant 1$ and $J\geqslant 9$. There exists a universal constant C, such that, for any $\delta\in(0,1/2)$ and $\beta>e^2$, and for any integers $\ell\geqslant 0$ and $m\geqslant 1$,*

$$\mathbb{P}\bigg(\Big\{\inf_{(s,a)\in\bar{S}\times\bar{A}}\sum_{j=\ell}^{\ell+m}\alpha_j\big(s,a\big)\geqslant A(\ell,m)\Big\}\bigg)\geqslant 1-\varepsilon(m),$$

$$A(\ell,m):=\frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{\eta}{2\beta\eta'\ln(J)}\ln\Big(\frac{\ell+m/2}{\ell+2m^{1/2+\delta}E_{\eta,J,h}^{1/2-\delta}}\Big),$$

$$\varepsilon(m):=\exp\big(C\beta^2\big)J\Big(\eta' h^D\Big)^{\beta/2-1}+\exp\Big(-\frac{(\ln(J))^{2(1-\delta)}(\eta h^D m)^{2\delta}}{C}\Big),$$

where η and η' are as in (3.1.1) *and* (**??**), *and*

$$E_{\eta,J,h} := \frac{\ln(J)+1}{\eta h^D}. \tag{3.2.7}$$

Remark 3.2.5. *The following remarks are in order:*

- *In the rest of the analysis, we want to have $A(\ell,m)$ large and $\varepsilon(m)$ small. This is possible to achieve by letting m tend to ∞ when all the other parameters are fixed. As m is here understood as the number of observations of the Markov chain $(\mathrm{s}_n,\mathrm{a}_n)_{n\in\mathbb{N}}$, choosing m large is however costly.*

  When m is fixed, but ℓ increases, the lower bound decreases. This comes from the fact that the denominator in the definition of the Nadaraya-Watson estimator increases with the number of observations.

- *In this respect, it is interesting to note that, when ℓ and m are fixed, the two terms $A(\ell,m)$ and $\varepsilon(m)$ depend in a somewhat dramatic way on the parameter η in the lower bound* (3.1.1) *and on the bandwith h in $\mathscr{K}_h$. Indeed, when we would like $A(\ell,m)$ to be large and $\varepsilon(m)$ to be small. This is difficult to achieve when h is small whilst the accuracy of the Nadaraya-Watson estimator gets better when h decreases. This makes the difficulty of the analysis.*

Proposition 3.2.4 allows us to control the dynamics of the time component $(\Lambda_n)_{n\geqslant 0}$ of the ARP. In short, we want $(\Lambda_n)_{n\geqslant 0}$ to remain as large as possible: intuitively, the larger Λ_n, the more accurate the mixing properties of the sequence $(\mathrm{s}_k,\mathrm{a}_k)_{0\leqslant k\leqslant\Lambda_n}$. This is the purpose of the following statement to clarify this fact:

Proposition 3.2.6. *Let ℓ_0 and n be two integers greater than 1. For J,h,β and δ as in the statement of Proposition 3.2.4 and for an additional real $\Gamma_0\geqslant 1$ such that $\Gamma_0^{(1-2\delta)/(1+2\delta)}E_{\eta,J,h}\geqslant 4$, let*

$$\ell_n = \Big\lfloor \ell_0\Big(1+\Gamma_0^{(1-2\delta)/(1+2\delta)}E_{\eta,J,h}\Big)^n\Big\rfloor. \tag{3.2.8}$$

Then, there exists an event $D(n)\in\mathcal{F}$, with

$$D(n)=\bigcap_{k=0}^{n-1}\Big\{\inf_{(s,a)\in\bar{S}\times\bar{A}}\sum_{j=\ell_k}^{\ell_{k+1}}\alpha_j(s,a)\geqslant A(\ell_k,\ell_{k+1}-\ell_k)\Big\}$$

for $A(\ell_k,\ell_{k+1}-\ell_k)$ as in Proposition 3.2.4, and

$$\mathbb{P}\big(D(n)\big)\geqslant 1-\exp\big(C\beta^2\big)nJ\Big(\eta' h^D\Big)^{\beta/2-1}-\min\big[n,\frac{C}{\delta}\big]\exp\Big(-\frac{(\ln(J))^4\ell_0^{2\delta}\Gamma_0^{2\delta(1-2\delta)/(1+2\delta)}}{C}\Big),$$

such that, for any realization $(\underline{\mathrm{s}},\underline{\mathrm{a}})=(\mathrm{s}_n,\mathrm{a}_n)_{n\in\mathbb{N}}$ (of the Markov process defined in the previous subsection) that belongs to $D(n)$, we have, for any integer $p\in\mathbb{N}$,

$$\mathbf{P}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\Big(\big\{\Lambda_{p+n}<\ell_0\big\}\,|\,\mathcal{G}_p^{(\Lambda,\Sigma,B)}\Big)\leqslant n\exp\Big(-\frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{\eta}{2\beta\eta'\ln(J)}\ln\Big(\frac{\Gamma_0^{(1-2\delta)^2/[2(1+2\delta)]}}{16}\Big)\Big),$$

on the event $\{\Lambda_p\geqslant\ell_n\}$.

Very intuitively, ℓ_n plays in the end the role of available observations of the sequence $(\underline{s}, \underline{a})$ and ℓ_0 is some threshold above which we want the time component of the ARP to stay. The parameter Γ_0 will be chosen in the end.

Proofs of Propositions 3.2.4 and 3.2.6 are given in Section 3.3.

3.2.4 Distance between the transition probabilities

We here introduce another key result. It is related to [27, Lemma 5], but it provides a quantitative bound of the phenomenon spotted therein. Meanwhile, it also clarifies the proof, as the results from [68] invoked in [27, Lemma 5] are just stated for a fixed choice of test functions (whilst the estimate of the distance below implicitly requires to address a supremum over a collection of test functions); moreover, the assumptions used in [68] involve additional stationarity properties, which we do not use here.

In short, the ARP provides a form of averaging, which, in the proof below, manifests in the form of diffusive bounds for martingales. The non-trivial point is to transform those diffusive bounds into a result for the distance between the transitions kernel of the original MDP and that of the ARP. In this regard, a key question is about the distance that should indeed equip the space of probability measures. While [27] makes use of the Wasserstein distance, already existing proofs for the rate of convergence, in the same Wasserstein distance, of the law of large numbers (see in particular [57] and the references therein) show that the resulting speed of convergence would dramatically suffer from the dimension. Instead, our approach here is to reduce the space of test functions from Lipschitz to much more regular functions. Although this requires to work with very strong assumptions on the coefficients, this allows us to preserve reasonable rate of convergence and related complexity, even in higher dimension. This approach is taken from the analysis of central limit theorems for particle systems, see for instance [89, 106].

Our space of test functions is defined as the Sobolev space $H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})$. In short, this is the space of functions with square-integrable Sobolev derivatives up to the order $5(\lfloor d_S/2\rfloor+1)$, see for instance the book [6] by Adams and Fournier. We denote by $\|\cdot\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})}$ the corresponding norm. We write the dual space as $H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})$. It is a Sobolev space of negative distributions equipped with the norm

$$\|q\|_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})} = \sup_{\varphi:\|\varphi\|_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\leqslant 1} \langle \varphi, q\rangle, \tag{3.2.9}$$

where the bracket in the right-hand side is seen as the usual duality bracket. One very key property is that, by Sobolev embedding (which holds true under the standing regularity of the domain, see [6]), any function in $H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})$) is bounded. In particular, any probability measure on $\overline{S}$ can be regarded as an element of the dual space $H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})$. The restriction of the norm $\|\cdot\|_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}$ to the space of probability measures induces a distance that is denoted by $\mathrm{d}_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}$.

Our objective is to address

$$\begin{aligned}&\mathrm{d}_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\Big(\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\big((n,s,a),\cdot\big),\overline{\mathsf{P}}\big((s,a),\cdot\big)\Big)\\&=\sup_{\|\varphi\|_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\leqslant 1}\Big\langle\varphi,\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\big((n,s,a),\cdot\big)-\overline{\mathsf{P}}\big((s,a),\cdot\big)\Big\rangle.\end{aligned}\tag{3.2.10}$$

We then notice from assumption **(Regularity Cost and Transition Kernel)** and from Sobolev embedding theorem that there exist two constants c and C such that

$$\begin{aligned}\mathrm{d}_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\Big(\overline{\mathsf{P}}\big((s,a),\cdot\big),\overline{\mathsf{P}}\big((s',a'),\cdot\big)\Big)&\leqslant\sup_{\|\varphi\|_{1,\infty}\leqslant c}\Big\langle\varphi,\overline{\mathsf{P}}\big((s,a),\cdot\big)-\overline{\mathsf{P}}\big((s',a'),\cdot\big)\Big\rangle\\&\leqslant C\big(|s-s'|+|a-a'|\big),\end{aligned}$$

where, for a continuously differentiable function φ on $\overline{S}\times\overline{A}$, $\|\varphi\|_{1,\infty}$ denotes the supremum norm of $|\varphi|+|\nabla\varphi|$. This plays a key role is the proof of the following result.

Proposition 3.2.7. *Under the standing assumption, for any real $\theta\in(0,1/2)$, we can find two constants C and C_θ, not depending on the discretization parameters (but possibly depending on the domains, on the dimensions, on the choice of $\mathscr{K}$) and with C_θ being allowed to depend on θ, with the following property: for any fixed realization $(\underline{\mathrm{s}},\underline{\mathrm{a}})$ of the process $(\mathrm{s}_\ell,\mathrm{a}_\ell)_{\ell\geqslant 0}$, for any integers $n\geqslant 1$ and $L\geqslant 2$, and any real $\varepsilon>0$ and $\beta\geqslant e^2$,*

$$\begin{aligned}&\mathbb{P}\Big(\bigcup_{j\geqslant n}\Big\{\mathrm{d}_{H^{-5(\lfloor d_S/2\rfloor+1)}}\Big(\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\big((j,s,a),\cdot\big),\overline{\mathsf{P}}\big((s,a),\cdot\big)\Big)\geqslant E_n(\theta)\Big\}\cap\Big\{\sum_{k=1}^n\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\leqslant L\Big\}\Big)\\&\leqslant C_\theta\big(\varepsilon h\big)^{-D}L^{-1/\theta+1}\quad+C\frac{1}{\beta\eta'\varepsilon^D h^{2D}}\exp\Big(-\beta\eta' h^D(n-1)\Big)+\frac{C}{\eta h^D}\exp\Big(-\frac{\eta h^D}{C}(n-1)\Big),\end{aligned}$$

where

$$E_n(\theta)=C\Big(\sum_{k=1}^n\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\Big)^{-(1-\theta)/2}+Ch+C\beta\varepsilon\frac{\eta'}{\eta}\big(1+\ln(J)\big).$$

Proof of the above result is given in Section 3.4.

3.2.5 Main statement

We now provide a rigorous version of the main Meta-Theorem.

Theorem 3.2.8. *Under the standing assumptions, we can find a constant C (not depending on the parameters of the algorithm but possibly depending on all the data entering the assumptions) such that, for any $\varepsilon\in(0,1)$ and for a number of observations*

$$n\geqslant\exp\big(C\frac{\eta'}{\eta}|\ln(\varepsilon)|^3\big),$$

we have

$$\sup_{(s,a)\in\overline{S}\times\overline{A}}\left|Q^{\star}(s,a)-Q^{\mathrm{ARP},\star}(n,s,a)\right|\leqslant C\Big[1-\Big(1+\frac{\eta'}{\eta}\Big)\ln(\varepsilon)\Big]\varepsilon+C\eta^{-3/8}\varepsilon^{1/8+7D/8},$$

on an event of probability greater than

$$1-C\big(\eta'\big)^{\beta/2-1}\varepsilon^{2D}-C\Big(1+\eta^{-4}\varepsilon^{6D}+(\eta')^{-4}\varepsilon^{4D}\Big)\varepsilon.$$

Proof. The proof is splitted into several steps.

First Step. We recall the equation for the optimal action value function. For any $(n,s,a)\in\mathbb{N}\times\overline{S}\times\overline{A}$,

$$\begin{aligned}Q^{\mathrm{ARP},\star}\big(n,s,a\big)&=R(s,a)\mathbf{P}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\big(\{\tau\geqslant 1\}\,|\,(\Lambda_0,\Sigma_0,B_0)=(n,s,a)\big)\\&\quad+\gamma\,\mathbf{E}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\Big[\sup_{a'\in\bar{A}}Q^{\mathrm{ARP},\star}\big(\Lambda_1,\Sigma_1,a'\big)\,|\,(\Lambda_0,\Sigma_0,B_0)=(n,s,a)\Big].\end{aligned}$$

Similarly, the original MDP satisfies:

$$Q^{\star}(s,a)=R(s,a)+\gamma\int_S\sup_{a'\in A}Q^{\star}\big(s',a'\big)\overline{\mathsf{P}}\big((s,a),ds'\big),$$

which implies (using the fact that the function R is bounded)

$$\begin{aligned}\Big|Q^{\mathrm{ARP},\star}(n,s,a)-Q^{\star}(s,a)\Big|&\leqslant C\mathbf{P}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\big(\{\tau=0\}\,|\,(\Lambda_0,\Sigma_0,B_0)=(n,s,a)\big)\\&\quad+\gamma\int_S\sup_{a'\in A}\Big|Q^{\mathrm{ARP},\star}\big(n',s',a'\big)-Q^{\star}(s',a')\Big|\,\overline{\Pi}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\big((n,s),(dn',ds')\big)\\&\quad+\gamma\Big|\int_S\sup_{a'\in A}Q^{\star}\big(s',a'\big)\Big[\overline{\mathsf{P}}\big((s,a),ds'\big)-\overline{\Pi}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\big((n,s),(dn',ds')\big)\Big]\Big|.\end{aligned}$$

We show in the second step of the proof of Proposition 3.2.6 that

$$\mathbf{P}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\big(\{\tau=0\}\,|\,(\Lambda_0,\Sigma_0,B_0)=(n,s,a)\big)\leqslant\exp\Big(-\inf_{(s',a')\in\bar{S}\times\bar{A}}\sum_{j=0}^{n-1}\alpha_j\big(s',a'\big)\Big).$$

And then, seeing n as the initial value Λ_0, we reformulate the above inequality as

$$\begin{aligned}\sup_{(s,a)\in\overline{S}\times\overline{A}}\left|Q^{\star}(s,a)-Q^{\mathrm{ARP},\star}(\Lambda_0,s,a)\right|&\leqslant C\exp\Big(-\inf_{(s',a')\in\overline{S}\times\overline{A}}\sum_{j=0}^{\Lambda_0-1}\alpha_j\big(s',a'\big)\Big)\\&\quad+\gamma\mathbf{E}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\Big[\sup_{(s,a)\in\overline{S}\times\overline{A}}\left|Q^{\star}(s,a)-Q^{\mathrm{ARP},\star}(\Lambda_1,s,a)\right|\,|\,\mathcal{G}_0^{(\Lambda,\Sigma,B)}\Big]\\&\quad+\gamma\|Q^{\star}(\cdot)\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\sup_{(s,a)\in\overline{S}\times\overline{A}}\big\|\overline{\mathsf{P}}\big((s,a),\cdot\big)-\overline{\Pi}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\big((\Lambda_0,s,a),\cdot\big)\big\|_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}.\end{aligned}$$

By induction, we get, for any $\ell \in \{0, \cdots, m\}$,

$$\begin{aligned}
\sup_{(s,a)\in\overline{S}\times\overline{A}} \big|Q^\star(s,a) - Q^{\mathrm{ARP},\star}(\Lambda_0,s,a)\big| &\leqslant \gamma^{\ell+1}\mathbf{E}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\Big[\sup_{(s,a)\in\overline{S}\times\overline{A}} \big|Q^\star(s,a) - Q^{\mathrm{ARP},\star}(\Lambda_{\ell+1},s,a)\big| \,|\, \mathcal{G}_0^{(\Lambda,\Sigma,B)}\Big] \\
&+ C\sum_{k=0}^{\ell}\gamma^k \mathbf{E}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\Big[\exp\Big(-\inf_{(s',a')\in\bar{S}\times\bar{A}}\sum_{j=0}^{\Lambda_k-1}\alpha_j\big(s',a'\big)\Big)\,|\,\mathcal{G}_0^{(\Lambda,\Sigma,B)}\Big] \\
&+ \|Q^\star(\cdot)\|_{2,D}\sum_{k=0}^{\ell}\gamma^{k+1}\mathbf{E}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\Big[\sup_{(s,a)\in\overline{S}\times\overline{A}}\big\|\overline{\mathsf{P}}\big((s,a),\cdot\big) - \overline{\Pi}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\big(\Lambda_k,s,a\big),\cdot\big)\big\|_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\,|\,\mathcal{G}_0^{(\Lambda,\Sigma,B)}\Big].
\end{aligned}$$

In particular, for $\ell = m$,

$$\begin{aligned}
&\sup_{(s,a)\in\overline{S}\times\overline{A}} \big|Q^\star(s,a) - Q^{\mathrm{ARP},\star}(\Lambda_0,s,a)\big| \\
&\leqslant \gamma^{m}\mathbf{E}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\Big[\sup_{(s,a)\in\overline{S}\times\overline{A}} \big|Q^\star(s,a) - Q^{\mathrm{ARP},\star}(\Lambda_{\ell},s,a)\big| \,|\, \mathcal{G}_0^{(\Lambda,\Sigma,B)}\Big] \\
&\quad + C\sum_{k=0}^{m}\gamma^k \mathbf{E}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\Big[\exp\Big(-\inf_{(s',a')\in\bar{S}\times\bar{A}}\sum_{j=0}^{\Lambda_k-1}\alpha_j\big(s',a'\big)\Big)\,|\,\mathcal{G}_0^{(\Lambda,\Sigma,B)}\Big] \\
&\quad + \|Q^\star(\cdot)\|_{2,D}\sum_{k=0}^{m}\gamma^{k+1}\mathbf{E}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\Big[\sup_{(s,a)\in\overline{S}\times\overline{A}}\big\|\overline{\mathsf{P}}\big((s,a),\cdot\big) - \overline{\Pi}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\big(\Lambda_k,s,a\big),\cdot\big)\big\|_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\,|\,\mathcal{G}_0^{(\Lambda,\Sigma,B)}\Big] \\
&=: T_1 + T_2 + T_3.
\end{aligned}$$

Second Step. We now handle each of the three terms in conclusion of the first step. We start with T_1. From the boundedness of the action-value functions, we clearly have

$$T_1 \leqslant C\gamma^m =: e_1.$$

As far as T_2 is concerned, we deduce from the fact that $(\Lambda_k)_{k\geqslant 0}$ is non-increasing that, for $k \in \{0, \cdots, m\}$,

$$\begin{aligned}
&\mathbf{E}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\Big[\exp\Big(-\inf_{(s',a')\in\overline{S}\times\overline{A}}\sum_{j=0}^{\Lambda_k-1}\alpha_j\big(s',a'\big)\Big)\,|\,\mathcal{G}_0^{(\Lambda,\Sigma,B)}\Big] \\
&\leqslant \mathbf{E}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\Big[\exp\Big(-\inf_{(s',a')\in\overline{S}\times\overline{A}}\sum_{j=0}^{\Lambda_m-1}\alpha_j\big(s',a'\big)\Big)\,|\,\mathcal{G}_0^{(\Lambda,\Sigma,B)}\Big] \\
&\leqslant \mathbf{P}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\big(\{\Lambda_m < \ell_0\}\,|\,\mathcal{G}_0^{(\Lambda,\Sigma,B)}\big) \;+ \exp\Big(-\inf_{(s',a')\in\overline{S}\times\overline{A}}\sum_{j=0}^{\ell_0-1}\alpha_j\big(s',a'\big)\Big) \\
&\leqslant \mathbf{P}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\big(\{\Lambda_m < \ell_0\}\,|\,\mathcal{G}_0^{(\Lambda,\Sigma,B)}\big) \;+ e\exp\Big(-\inf_{(s',a')\in\overline{S}\times\overline{A}}\sum_{j=0}^{\ell_0}\alpha_j\big(s',a'\big)\Big),
\end{aligned}$$

where ℓ_0 is as in the statement of Proposition 3.2.6. Using the same notation as in the latter statement and combining with Proposition 3.2.4 (with $\ell = 0$ and $m = \ell_0$ therein), we have, under

the condition $\{\Lambda_0 \geqslant \ell_m\}$,

$$T_2 \leqslant e_2 := C(m+1)\exp\Big(-\frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{\eta}{2\beta\eta'\ln(J)}\ln\Big(\frac{\Gamma_0^{(1-2\delta)^2/[2(1+2\delta)]}}{16}\Big)\Big) + C\exp\Big(-\frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{\eta}{2\beta\eta'\ln(J)}\ln\Big(\frac{\ell_0^{1/2-\delta}}{4E_{\eta,J,h}^{1/2-\delta}}\Big)\Big),$$

on an event F_2 with probability

$$\begin{aligned}&\mathbb{P}(F_2)\\ &\geqslant 1-\exp\big(C\beta^2\big)(m+2)J\Big(\eta' h^D\Big)^{\beta/2-1} - \min\big[m+1,\frac{C}{\delta}\big]\exp\Big(-\frac{(\ln(J))^4\ell_0^{2\delta}\Gamma_0^{2\delta(1-2\delta)/(1+2\delta)}}{C}\Big)\\ &\quad-\exp\Big(-\frac{(\ln(J))^{2(1-\delta)}(\eta h^D\ell_0)^{2\delta}}{C}\Big).\end{aligned}$$

At last, we handle T_3. Observing that any probability measure has a universally bounded norm $\|\cdot\|_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}$ (as a consequence of Sobolev's embedding theorem), we deduce that

$$\begin{aligned}&\mathbf{E}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\Big[\sup_{(s,a)\in\overline{S}\times\overline{A}}\big\|\overline{\mathsf{P}}\big((s,a),\cdot\big)-\overline{\Pi}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\big(\Lambda_k,s,a\big),\cdot\big)\big\|_{2,-D}\,|\,\mathcal{G}_0^{(\Lambda,\Sigma,B)}\Big]\\ &\leqslant C\mathbf{P}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\big(\{\Lambda_m<\ell_0\}\,|\,\mathcal{G}_0^{(\Lambda,\Sigma,B)}\big) \;+C\sup_{j\geqslant\ell_0}\sup_{(s,a)\in\overline{S}\times\overline{A}}\big\|\overline{\mathsf{P}}\big((s,a),\cdot\big)-\overline{\Pi}_{\underline{\mathrm{s}},\underline{\mathrm{a}}}\big(j,s,a\big),\cdot\big)\big\|_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}.\end{aligned}$$

By changing the value of C in the definition of e_2, we obtain

$$T_3 \leqslant e_2 + e_3,$$

on $F_2\cap F_3$, where, by Proposition 3.2.7,

$$e_3 := E_{\ell_0}(\theta),$$

and

$$\begin{aligned}&\mathbb{P}(F_3)\\ &\geqslant 1-C_\theta\big(\varepsilon h\big)^{-D}L^{-1/\theta+1} \;-C\frac{1}{\beta\eta'\varepsilon^D h^{2D}}\exp\Big(-\beta\eta' h^D(\ell_0-1)\Big)-\frac{C}{\eta h^D}\exp\Big(-\frac{\eta h^D}{C}(\ell_0-1)\Big)\\ &\quad-\mathbb{P}\Big(\Big\{\sum_{k=1}^{\ell_0}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\leqslant L\Big\}\Big),\end{aligned}$$

where L is a free parameter.

Here, we claim that, for a universal constant $c>0$ (the proof is given next),

$$\mathbb{P}\Big(\Big\{\sum_{k=1}^{\ell_0}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\leqslant L\Big\}\Big)\leqslant\exp\big(-c\eta^2\ell_0 h^{2D}\big), \tag{3.2.11}$$

when

$$\frac{2\|\mathscr{K}\|_\infty}{\lambda_{\mathscr{K}}}\frac{L}{\ell_0} \leqslant \frac{1}{2}\eta h^D. \tag{3.2.12}$$

Choosing L to get equality in the above inequality, we obtain

$$\begin{aligned}\mathbb{P}(F_3) \geqslant 1 - C_\theta \eta^{-1/\theta+1}\varepsilon^{-D}h^{-D/\theta}\ell_0^{-1/\theta+1} \quad &- \exp\bigl(-c\eta^2\ell_0 h^{2D}\bigr) \\ - C\frac{1}{\beta\eta'\varepsilon^D h^{2D}}\exp\Bigl(-\beta\eta' h^D(\ell_0-1)\Bigr) &- \frac{C}{\eta h^D}\exp\Bigl(-\frac{\eta h^D}{C}(\ell_0-1)\Bigr).\end{aligned}$$

It remains to see that, on F_3, we have implicitly

$$\sum_{k=1}^{\ell_0}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\bigl(s-\mathrm{s}_k, a-\mathrm{a}_k\bigr) \geqslant L.$$

Therefore, recalling the statement of Proposition 3.2.7, we obtain

$$e_3 = E_{\ell_0}(\theta) \leqslant C\bigl(\eta h^D\ell_0\bigr)^{-(1-\theta)/2} + ch + C\beta\varepsilon\frac{\eta'}{\eta}\bigl(1+\ln(J)\bigr).$$

Collecting all the terms, we end up with

$$\begin{aligned}&\sup_{(s,a)\in\overline{S}\times\overline{A}}\bigl|Q^\star(s,a) - Q^{\mathrm{ARP},\star}(\Lambda_0,s,a)\bigr| \\ &\leqslant C\gamma^m + C\bigl(\eta h^D\ell_0\bigr)^{-(1-\theta)/2} + ch + C\beta\varepsilon\frac{\eta'}{\eta}\bigl(1+\ln(J)\bigr) \\ &\quad + C(m+1)\exp\biggl(-\frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{\eta}{2\beta\eta'\ln(J)}\ln\Bigl(\frac{\Gamma_0^{(1-2\delta)^2/[2(1+2\delta)]}}{16}\Bigr)\biggr) \\ &\quad + C\exp\biggl(-\frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{\eta}{2\beta\eta'\ln(J)}\ln\Bigl(\frac{\ell_0^{1/2-\delta}}{4E_{\eta,J,h}^{1/2-\delta}}\Bigr)\biggr),\end{aligned} \tag{3.2.13}$$

on $\{\Lambda_0 \geqslant \ell_m\}\cap F$, where

$$\begin{aligned}\mathbb{P}(F) \geqslant 1 - \exp\bigl(C\beta^2\bigr)(m+1)J\Bigl(\eta' h^D\Bigr)^{\beta/2-1} \\ - \min\bigl[m+1, \frac{C}{\delta}\bigr]\exp\Bigl(-\frac{(\ln(J))^4\ell_0^{2\delta}\delta_0^{2\delta(1-2\delta)/(1+2\delta)}}{C}\Bigr) \\ - \exp\Bigl(-\frac{(\ln(J))^{2(1-\delta)}(\eta h^D\ell_0)^{2\delta}}{C}\Bigr) \\ - C_\theta\eta^{-1/\theta+1}\varepsilon^{-D}h^{-D/\theta}\ell_0^{-1/\theta+1} \quad - \exp\bigl(-c\eta^2\ell_0 h^{2D}\bigr) \\ - \frac{1}{\beta\eta'\varepsilon^D h^{2D}}\exp\Bigl(-\beta\eta' h^D(\ell_0-1)\Bigr) - \frac{C}{\eta h^D}\exp\Bigl(-\frac{\eta h^D}{C}(\ell_0-1)\Bigr).\end{aligned} \tag{3.2.14}$$

Third Step. We now tune the various parameters.
We first recall the meaning of each of them:

1. The parameter m is an arbitrary integer that plays the role of a threshold in the various truncations appearing in the expansions of T_1, T_2 and T_3 in the first step. We may choose it freely. In the bound (3.2.13), m appears through the factor $C\gamma^m$ and in the form a linear factor in the penultimate term. In (3.2.14), m also plays the role of a linear factor. Last but not least, m appears in the choice of the initial condition ℓ_m, as we required Λ_0 to be greater than ℓ_m.

2. In connection with the role of m, the parameter ℓ_m represents the number of observations that are needed in the end. This is a quantity that can be tuned by the observer, but, for obvious practical reasons, its value must be chosen in a minimal way depending on the choices for the other parameters.

3. In fact, the parameter ℓ_m is related to ℓ_0, m and Γ_0 through the formula (3.2.8). The largest m, ℓ_0 and Γ_0, the largest ℓ_m, but at the cost of an increase of complexity. The parameter ℓ_0 appears repeatedly in the two inequalities (3.2.13) and (3.2.14), with rates that are polynomial or exponential. The parameter Γ_0 also appears in various exponential terms.

4. The polynomial rates in the previous item are dictated by the exponent θ. The latter comes from the statement of Proposition 3.2.7. It is required to be in $(0, 1/2)$. Similarly, the exponent δ comes from Propositions 3.2.4 and 3.2.6 and is required to belong to $(0, 1/2)$.

5. The parameter ε also appears in the statement Proposition 3.2.7. The smallest ε, the smallest its contribution in (3.2.13), but the largest its impact in (3.2.14).

6. The parameter β arises in the statement of Proposition 3.2.4. It must be greater than e^2. The highest β, the highest the error in (3.2.13). The impact on (3.2.14) is more subtle. When β is large, the factor $(h^D)^{\beta/2-1}$ is typically expected to be small (see below for the choice of h) and the factor $\exp(C\beta^2)$ becomes large. Clearly, β should not be too large.

7. As far as h and J are concerned, we recall that h is the bandwidth in the kernel $\mathscr{K}_h$. Typically, it is small. And J is connected with the numbers of balls of radius h that are need to cover the domain $\overline{S} \times \overline{A}$. We should think it as h^{-D} up to a scaling factor that depends on the diameter of the domain.

8. The parameter γ is directly connected with the optimization problem: it is the discount factor. The parameters η and η' reflects the non-degeneracy (or, simply, the presence) of the noise underpinning the dynamics.

In order to choose all these parameters together with proceed as follows. We fix two tolerance thresholds $\mathrm{Tol}_{\mathrm{error}}$ and $\mathrm{Tol}_{\mathrm{prob}}$ that we do not want to exceed in both the error (3.2.13) and the probability (3.2.13) (up to the operation $x \mapsto 1 - x$). Also choose $\beta = e^2 + 1$ and $\theta = \delta = 1/4$, consistently with the observations that we have just made in the above description .

By (3.2.13) , we then choose

$$m = \Big\lceil \frac{\ln\big(\mathrm{Tol}_{\mathrm{error}}\big)}{\ln(\gamma)} \Big\rceil, \quad h = \varepsilon = \mathrm{Tol}_{\mathrm{error}}, \quad \tfrac{1}{16}\Gamma_0^{1/12} = \exp\Big(-\frac{2\beta\eta'\|\mathscr{K}\|_\infty}{\eta\lambda_{\mathscr{K}}} \ln(J)\ \ln\big(\mathrm{Tol}_{\mathrm{error}}\big)\Big),$$

and ℓ_0 as the maximum between $\ell_{\rm a}$ and $\ell_{\rm b}$ where

$$\varepsilon^{-D}h^{-9D}\ell_{\rm a}^{-3} = {\rm Tol}_{\rm prob}, \quad \ell_{\rm b} = \Gamma_0^{(1-2\delta)/(1+2\delta)}E_{\eta,J,h}$$

We get

$$\ell_{\rm a} = {\rm Tol}_{\rm prob}^{-1/3}\varepsilon^{-D/3}h^{-3D}, \quad h^D\ell_{\rm a} = {\rm Tol}_{\rm prob}^{-1/3}\varepsilon^{-D/3}h^{-2D}.$$

We obtain

$$\begin{aligned}
&\exp\bigl(-c\eta^2\ell_0h^{2D}\bigr) \leqslant C\eta^{-3/2}\varepsilon^Dh^{3D}{\rm Tol}_{\rm prob},\\
&\exp\bigl(-\beta\eta' h^D(\ell_0-1)\bigr) \leqslant C\bigl(\eta'\bigr)^{-3}\varepsilon^Dh^{6D}{\rm Tol}_{\rm prob},\\
&\exp\bigl(-\frac{\eta h^D}{C}(\ell_0-1)\bigr) \leqslant C'\eta^{-3}\varepsilon^Dh^{6D}{\rm Tol}_{\rm prob}.
\end{aligned}$$

As far as the first four terms in (3.2.14) are concerned,

$$\begin{aligned}
&\exp\bigl(C\beta^2\bigr)(m+1)J\bigl(\eta' h^D\bigr)^{\beta/2-1} \leqslant C\ln\bigl({\rm Tol}_{\rm error}\bigr)\bigl(\eta'\bigr)^{\beta/2-1}\bigl(h^D\bigr)^{\beta/2-2},\\
&\min\bigl[m+1,\frac{C}{\delta}\bigr]\exp\Bigl(-\frac{(\ln(J))^4\ell_0^{2\delta}\Gamma_0^{2\delta(1-2\delta)/(1+2\delta)}}{C}\Bigr) \leqslant C\ell_{\rm a}^{-3} \leqslant C{\rm Tol}_{\rm prob}\varepsilon^Dh^{9D},\\
&\exp\Bigl(-\frac{(\ln(J))^{2(1-\delta)}(\eta h^D\ell_0)^{2\delta}}{C}\Bigr) \leqslant C\eta^{-3}\varepsilon^Dh^{6D}{\rm Tol}_{\rm prob}, \quad ,\\
&\eta^{-1/\theta+1}\varepsilon^{-D}h^{-D/\theta}\ell_0^{-1/\theta+1} \leqslant \eta^{-3}h^{5D}{\rm Tol}_{\rm prob}.
\end{aligned}$$

As for (3.2.13), we have

$$\begin{aligned}
&\bigl(\eta h^D\ell_0\bigr)^{-(1-\theta)/2} \leqslant C\eta^{-3/8}{\rm Tol}_{\rm prob}^{1/8}\varepsilon^{D/8}h^{3D/4},\\
&(m+1)\exp\Bigl(-\frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{\eta}{2\beta\eta'\ln(J)}\ln\Bigl(\frac{\Gamma_0^{(1-2\delta)^2/[2(1+2\delta)]}}{16}\Bigr)\Bigr) \leqslant C\ln\Bigl({\rm Tol}_{\rm error}\Bigr){\rm Tol}_{\rm error},\\
&\exp\Bigl(-\frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{\eta}{2\beta\eta'\ln(J)}\ln\Bigl(\frac{\ell_0^{1/2-\delta}}{4E_{\eta,J,h}^{1/2-\delta}}\Bigr)\Bigr) \leqslant C{\rm Tol}_{\rm error}.
\end{aligned}$$

We end up with

$$\begin{aligned}
&\sup_{(s,a)\in\overline{S}\times\overline{A}}\bigl|Q^\star(s,a) - Q^{{\rm ARP},\star}(\Lambda_0,s,a)\bigr|\\
&\leqslant C\Bigl[1-\Bigl(1+\frac{\eta'}{\eta}\Bigr)\ln\Bigl({\rm Tol}_{\rm error}\Bigr)\Bigr]{\rm Tol}_{\rm error} + C\eta^{-3/8}{\rm Tol}_{\rm prob}^{1/8}{\rm Tol}_{\rm error}^{7D/8},
\end{aligned} \tag{3.2.15}$$

on an event of probability greater than

$$1 - C\bigl(\eta'\bigr)^{\beta/2-1}\ln\Bigl({\rm Tol}_{\rm error}\Bigr){\rm Tol}_{\rm error}^{D(\beta/2-2)} - C\Bigl(1+\eta^{-4}{\rm Tol}_{\rm error}^{6D} + (\eta')^{-4}{\rm Tol}_{\rm error}^{4D}\Bigr){\rm Tol}_{\rm prob}. \tag{3.2.16}$$

Now,

$$\begin{aligned}\ell_m &\leqslant \ell_0\Big(1+\Gamma_0^{1/3}\frac{|\ln(\mathrm{Tol}_{\mathrm{error}})|}{\eta\mathrm{Tol}_{\mathrm{error}}^D}\Big)^{|\ln(\mathrm{Tol}_{\mathrm{error}})|}\\ &= \max(\ell_{\mathrm{a}},\ell_{\mathrm{b}})\Big(1+\Gamma_0^{1/3}\frac{|\ln(\mathrm{Tol}_{\mathrm{error}})|}{\eta\mathrm{Tol}_{\mathrm{error}}^D}\Big)^{|\ln(\mathrm{Tol}_{\mathrm{error}})|}\\ &\leqslant \mathrm{Tol}_{\mathrm{prob}}^{-1/3}\mathrm{Tol}_{\mathrm{error}}^{-10D/3}\Big(1+\exp\big(C\frac{\eta'}{\eta}|\ln(\mathrm{Tol}_{\mathrm{error}})|^2\big)\frac{|\ln(\mathrm{Tol}_{\mathrm{error}})|}{\eta\mathrm{Tol}_{\mathrm{error}}^D}\Big)^{|\ln(\mathrm{Tol}_{\mathrm{error}})|+1},\end{aligned} \tag{3.2.17}$$

which is enough to complete the proof.

Fourth Step. We now prove (3.2.11). We have

$$\mathbb{P}\Big(\Big\{\sum_{k=1}^{\ell_0}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\leqslant L\Big\}\Big)=\mathbb{P}\Big(\Big\{\sum_{k=1}^{\ell_0}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\leqslant 2\|\mathscr{K}\|_\infty L\Big\}\Big)$$

Here,

$$\Big\{\sum_{k=1}^{\ell_0}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\leqslant 2\|\mathscr{K}\|_\infty L\Big\}\subset\Big\{\sum_{j=1}^{\ell_0}\mathbf{1}_B\big(\mathrm{s}_j,\mathrm{a}_j\big)\leqslant\frac{2\|\mathscr{K}\|_\infty}{\lambda_{\mathscr{K}}}L\Big\},$$

where B is the ball of center (s,a) and of radius ρh. By the lower bound for the transition kernel,

$$\mathbb{P}\Big(\Big\{\big(\mathrm{s}_{j+1},\mathrm{a}_{j+1}\big)\in B\Big\}\,|\,\mathcal{F}_j\Big)\geqslant\eta h^D.$$

Therefore, the sum

$$\sum_{j=1}^{\ell_0}\mathbf{1}_B\big(\mathrm{s}_j,\mathrm{a}_j\big)$$

is stochastically lower bounded by

$$\sum_{i=1}^{\ell_0}\varepsilon_i,$$

where $(\varepsilon_1,\cdots,\varepsilon_{\ell_0})$ are independent and identically distributed Bernoulli random variables with parameter ηh^D, which implies that

$$\begin{aligned}\mathbb{P}\Big(\Big\{\sum_{j=1}^{\ell_0}\mathbf{1}_B\big(\mathrm{s}_j,\mathrm{a}_j\big)\leqslant\frac{2\|\mathscr{K}\|_\infty}{\lambda_{\mathscr{K}}}L\Big\}\Big)&\leqslant\mathbb{P}\Big(\Big\{\sum_{j=1}^{\ell_0}\varepsilon_j\leqslant\frac{2\|\mathscr{K}\|_\infty}{\lambda_{\mathscr{K}}}L\Big\}\Big)\\ &=\mathbb{P}\Big(\Big\{\frac{1}{\ell_0}\sum_{j=1}^{\ell_0}\varepsilon_j\leqslant\frac{2\|\mathscr{K}\|_\infty}{\lambda_{\mathscr{K}}}\frac{L}{\ell_0}\Big\}\Big).\end{aligned}$$

Assume now that

$$\frac{2\|\mathscr{K}\|_\infty}{\lambda_{\mathscr{K}}}\frac{L}{\ell_0}\leqslant\frac{1}{2}\eta h^D.$$

By Hoeffding inequality (see [49, Chapter 2, Section 4]), we obtain

$$\mathbb{P}\Big(\Big\{\sum_{j=1}^{\ell_0}\mathbf{1}_B\big(\mathrm{s}_j,\mathrm{a}_j\big)\leqslant\frac{2\|\mathscr{K}\|_\infty}{\lambda_{\mathscr{K}}}L\Big\}\Big)\leqslant\exp\big(-c\eta^2\ell_0h^{2D}\big),$$

for a universal constant $c>0$. □

3.2.6 Example from mean field control

Our main example is taken from mean field control (see for instance [11, 30]). In this situation, we assume that $\overline{S}$ is the space of probability measures over a finite set, denoted by $\mathscr{S}$. In this situation, the Markov decision process $(\mathrm{s}_n)_{n\geqslant0}$ is implicitly understood as describing the evolution of a large cloud of N agents who cooperate in order to maximise a common reward assigned to the collectivity (or to minimise a common energy).

A prototype for such a situation is the case when agent $i\in\{1,\cdots,N\}$ in the population obey dynamics of the form

$$X^i_{n+1}=F\big(X^i_n,\bar{\mu}^N_n,\alpha^i_n,U^i_{n+1},U^0_{n+1}\big),\tag{3.2.18}$$

where F is map (common to all the agents) from $\mathscr{S}\times\overline{S}\times\overline{A}\times[0,1]\times[0,1]$ into $\mathscr{S}$. Above, the notation $\bar{\mu}^N_n$ is the standard notation for the empirical measure of the cloud of agents at time n, namely

$$\bar{\mu}^N_n:=\frac{1}{N}\sum_{j=1}^N\delta_{X^j_n}.$$

Moreover, in (3.2.18), quantities $((U^k_\ell)_{k=0,\cdots,N})_{\ell\geqslant0}$ are independent random variables with uniform distribution on $[0,1]$. The interpretation of those samples is as follows: the variable U^i_{n+1} (indexed by the same label i as the player itself) should be thought as an idiosyncratic noise to which agent i is subjected. Importantly, this noise does not impact directly the rest of the collectivity (it does impact indirectly the other players through the mean field interaction term). As made clear below, idiosyncratic noises do not randomize the state of the population: we drew a similar observation in the chapter dedicated to mean field games. Differently, the variable U^0_{n+1} is the same in the dynamics of all the players. This noise is said to be common and should induce in the end a form of randomization of the population.

Last but not least, the variable α^i_n in (3.2.18) denotes the action chosen by player i at time n, based upon the observations made up until time n. Whereas this would be in fact very welcome for practical purposes, we refrain from entering a lengthy discussion on the quantities that are indeed observable in practice. This is somehow outside the scope of the chapter.

In mean field control, players cooperate to maximize a global reward. With the notations we introduced before, this common reward may be expressed as

$$\mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^N\sum_{k\geqslant0}\gamma^kR(X^i_k,\alpha^i_k)\Big],$$

expectation being taken with respect to all the noises. Typically, players are assumed to start from independent and identically distributed initial conditions.

Then, the key idea is to take the limit $N \to \infty$ and to address directly the same problem but when the population is infinite. Inspired by mean field theory in statistical physics, the ansatz is the following: one may directly optimize over the state of the population, and the latter should coincide with the theoretical distributional state of any agent in the population (which, in the optimal state, are expected to be exchangeable). Intuitively, the question is thus to derive the dynamics for the evolution of the population from the relation (3.2.18). Thinking of the action α_n^i as a function (through a policy π) of the private state X_n^i and of the state of the population $\bar{\mu}_n^N$, one postulates that, given the state s_n of the (now infinite) population at time n, the state s_{n+1} is given by

$$s_{n+1} = \Big(s_n \otimes \mathrm{Unif}([0,1])\Big) \circ \Big((x,u) \mapsto F\big(x, s_n, \pi(x,s_n), u, U_{n+1}^0\big)\Big)^{-1}. \tag{3.2.19}$$

Obviously, this identity may be rewritten in terms of transition probabilities on the space $\overline{S}$. For $(s,a) \in \overline{S} \times \overline{A}$ and for $v \in [0,1]$, we define

$$\overline{\mathsf{P}}(s,a)\big(E\big) = \mathbb{P}\Big(\Big\{\Big(s \otimes \mathrm{Unif}([0,1])\Big) \circ \Big((x,u) \mapsto F\big(x,s,a,u,U^0\big)\Big)^{-1} \in E\Big\}\Big),$$

with U^0 being uniformly distributed on $[0,1]$. The mean field control problem is nothing but the Markov decision process associated with the kernel $\overline{\mathsf{P}}$. It is also worth emphasizing that (3.2.19) corresponds to the dynamics $(s_n = \mathcal{L}(X_n|(U_k^0)_{1\leqslant k\leqslant n}))_{n\geqslant 0}$ of the conditional marginal laws (given the common noise) of a typical agent in the population obeying the following rule

$$X_{n+1} = F\big(X_n, \mathcal{L}(X_n|(\varepsilon_k^0, U_k^0)_{1\leqslant k\leqslant n}), \alpha_n, U_{n+1}^1, U_{n+1}^0\big), \quad n \geqslant 0. \tag{3.2.20}$$

As far as the control parameter is concerned, the set $\overline{A}$ may be itself the space of probability measures $\mathcal{P}(\mathscr{A})$ over some finite set $\mathscr{A}$ or it may a more general 'continuous' set. When $\overline{A}$ is understood as $\mathcal{P}(\mathscr{A})$, it means that players are allowed to play random strategies.

We now give several examples for the form of the common noise. These are important in order to obtain a model that satisfies our assumptions:

1. Call $G : [0,1] \to \overline{S} = \mathcal{P}(\mathscr{S})$ a random variable with uniform distribution (when $[0,1]$ is equipped with the Lebesgue measure) and then call $H : [0,1] \times \mathcal{P}(\mathscr{S}) \to \mathscr{S}$ a mapping such that, for any $\mu \in \mathcal{P}(\mathscr{S})$, the image of the Lebesgue measure on $[0,1]$ by $u \in [0,1] \mapsto H(u,\mu)$ is μ-distributed (see [30, Chapter 5] for the construction). Then, for $(U_n^0)_{n\geqslant 1}$ and $(U_n^1)_{n\geqslant 1}$, two independent sequences of independent and identically distributed random variables as above, with the first one representing the common noise and the second one representing the idiosyncratic noice, we may consider the dynamics

   $$X_{n+1} = H\big(U_{n+1}^1, G(U_{n+1}^0)\big), \quad n \geqslant 0.$$

   Then, any X_{n+1} is distributed (conditional on the common noise) according to a measure on $\mathscr{S}$ that is sampled randomly (uniformly) on $\overline{S}$.

2. For sure, the above example may be generalized. For instance, if $F : S \times \mathscr{S} \times A \times [0,1] \to S$ is given (notice that this one is independent of the common noise), one may consider in addition

a sequence $(\varepsilon_n^0)_{n\geqslant 1}$, independent of the two sequences $(U_n^0)_{n\geqslant 1}$ and $(U_n^1)_{n\geqslant 1}$, such that each ε_n^0 is Bernoulli(p) distributed for $p \in (0,1)$. Then, we may consider the dynamics:

$$X_{n+1} = \varepsilon_{n+1} H\big(U_{n+1}^1, G(U_{n+1}^0)\big) + (1-\varepsilon_{n+1}) F\big(X_n, \mathcal{L}(X_n|(\varepsilon_k^0, U_k^0)_{1\leqslant k\leqslant n}), \alpha_n, U_{n+1}^1\big), \quad n \geqslant 0.$$

This corresponds to the case where, with probability $1-\varepsilon_n$, the agent follows the 'original' dynamics (3.2.20) (without common noise), and with probability ε_0, it is resampled uniformly on the whole space.

3. There is another way to extend the first example. We consider a smooth density function φ from $\mathbb{R}^{d_S} \to \mathbb{R}$ with a support that is localized around 0. We call Z^0 a random variable distributed according to φ. We observe that, for a probability measure $s \in \overline{S}$ such that

$$\inf_{i=1,\cdots,|\mathscr{S}|} s_i \geqslant q,$$

for some $q > 0$ (and with $s_1, \cdots, s_{|\mathscr{S}|}$ denoting the weights of the probability measure s on the set $\mathscr{S}$), the random variable $s + Z^0$ remains a probability measure (if the support of φ is sufficiently concentrated around 0). In order to clarify this, denote by c the smallest radius of the ball containing the support of φ. Then, for any $s \in \overline{S}$ and for $q \geqslant c$, we have

$$\mathbb{P}\Big(\big\{q\mathbf{1} + (1-q)s + Z^0 \in \overline{S}\big\}\Big) = 1,$$

where $\mathbf{1}$ is (here) the vector $(1/|\mathscr{S}|, \cdots, 1/|\mathscr{S}|)$. This says that

$$X_{n+1} = H\Big(U_{n+1}, q\mathbf{1} + (1-q)\mathcal{L}(X_n|(Z_k^0)_{1\leqslant k\leqslant n}) + Z_{n+1}^0\Big),$$

is also a model with common noise, when $(Z^0)_{k\geqslant 1}$ are independent copies of Z^0.

For sure, all the examples may be reformulated in terms of the sole kernel $\overline{\mathsf{P}}$. In fact, what really matters is the practical meaning of the various types of transformation $H(\dots)$ presented right above:

1. In examples (1) and (2), the transformation H is used to lower bound the density of the law $\mathcal{L}(X_n|(U_k^0)_{1\leqslant k\leqslant n})$. In clear, with probability p, one picks up randomly (with uniform distribution) an element of $\overline{S}$. Therefore, for any Borel subset $E \subset \overline{S}$,

$$\overline{\mathsf{P}}\big((s,a),E\big) \geqslant p\frac{\mathrm{Leb}_{\overline{S}}(E)}{\mathrm{Leb}_{\overline{S}}(S)}, \tag{3.2.21}$$

where $\mathrm{Leb}_{\overline{S}}$ is the Lebesgue measure on $\overline{S}$. This fits the constraint (3.1.1) (at least for the marginal in space).

2. In example (3), one has

$$\overline{\mathsf{P}}\big((s,a),E\big) \leqslant \sup\varphi \times \mathrm{Leb}_{\overline{S}}(E), \tag{3.2.22}$$

which fits (**??**).

Now, for a measurable kernel $\overline{\mathsf{P}}^0 : \overline{S} \times \overline{A} \to \mathcal{P}(\overline{S})$ that would correspond to the dynamics of a mean field control problem (obtained for instance by means of (3.2.20)), we can construct the following (new) kernel $\overline{\mathsf{P}}$ (by means of the same principles as those underpinning the aforementioned examples 1 and 3):

$$\overline{\mathsf{P}}\big((s,a),E\big) = p\frac{\mathrm{Leb}_{\overline{S}}(E)}{\mathrm{Leb}_{\overline{S}}(S)} + (1-p)\int_{\mathbb{R}^{d_S}}\Big(\int_{\overline{S}} \mathbf{1}_E\big(q+(1-q)z+y\big)\overline{\mathsf{P}}^0\big((s,a),dz\big)\Big)\varphi(y)dy. \quad (3.2.23)$$

It satisfies (3.2.21) and (3.2.22) (with respect to possibly different multiplicative constants).

For this example, we now address the Bellman equation. It reads

$$\begin{aligned} V(s) &= \sup_{a\in\overline{A}}\Big[R(s,a) + \gamma\int_{\overline{S}} V(s')\overline{\mathsf{P}}\big((s,a),ds'\big)\Big] \\ &= \sup_{a\in\overline{A}}\Big[R(s,a) + \gamma\int_{\mathbb{R}^{|\mathscr{I}|}}(1-p)\int_{\mathbb{R}^{|\mathscr{I}|}}\Big(\int_{\overline{S}} V\big(q+(1-q)z+y\big)\overline{\mathsf{P}}^0\big((s,a),dz\big)\Big)\varphi(y)dy\Big] \\ &\quad + \frac{\gamma p}{\mathrm{Leb}_{\overline{S}}(S)}\int_S V(s')ds'. \end{aligned}$$

There may be several types of conditions under which the value function V is regular in s (which is a prerequisite in our main theorem). A very basic example is: R and $\overline{\mathsf{P}}^0$ are smooth and have a separated structure in (s,a). We exemplify the application of our main result to the above example in the next subsection.

3.2.7 The common noise as an exploration noise

In the previous subsection, we have given some examples of noises in the framework of mean field control that would induce a form of exploration consistent with the two conditions (3.1.1) and (**??**). In fact, in this discussion, we have just addressed the randomization of the state variable (i.e., the law of the state in a mean field control problem). In fact, both (3.1.1) and (**??**) also require to address the 'transitions' of the action variable. In this regard, it must be stressed that, in the algorithm (2), we can only choose a_n to be randomly distributed on the whole space. Whilst this may open some questions from the practical point view about the accuracy of such a strategy, this fits in fact the mathematical analysis that we have exposed in this section. Very briefly, the reader will notice from the reading of the forthcoming Section 3.3 that our strategy for proving Propositions 3.2.4 and 3.2.6 is greatly inspired from the so-called coupon collector problem (see for instance [13]) in which new coupons are discovered with a purely random strategy. In this sense, the ratio η'/η can be understood as the ratio (up to a multiplicative constant) $\|\varphi\|_\infty/p$ appearing in (3.2.21) and (3.2.22). This leads us to the following observation: in (3.2.23), one may start from $\overline{\mathsf{P}}^0$ and then regard $\overline{\mathsf{P}}$ as a randomized version of it just used for the purpose of learning. This point of view is very close to the one considered in the chapter dedicated to mean field games.

To make this clear, we have the following statement:

Lemma 3.2.9. *Within the mean field control framework (as described in the previous subsection) and with the same notation as in* (3.2.23)*, call $V^{\star,0}$ (resp. $Q^{\star,0}$) the value function (resp. the action*

value function) associated with the kernel $\overline{\mathsf{P}}^0$ *and with the cost* R*. Assume that* $V^{\star,0}$ *is Lipschitz continuous in* s*. Then,*

$$\sup_{(s,a)\in\overline{S}\times\overline{A}} |Q^{\star,0}(s,a) - Q^{\star}(s,a)| \leqslant C\big(p+q\big),$$

where q *denotes the radius of the support of* φ*.*

Before we give the proof of Lemma 3.2.9, we now provide the meaning of it. Typically, $\|\varphi\|_\infty$ is of order q^{-d_S}. We thus deduce that (3.2.15), (3.2.16) and (3.2.17) hold true with

$$\eta = p, \quad \eta' = q^{|\mathscr{I}|-1}.$$

In particular, we obtain

Proposition 3.2.10. *In the example* (3.2.23)*, choose*

$$p = q^{d_S} = \mathrm{Tol}_{\mathrm{error}}^{d_S},$$

then the bounds (3.2.15)*,* (3.2.16) *and* (3.2.17) *give*

$$\sup_{(s,a)\in\overline{S}\times\overline{A}} \big|Q^{\star}(s,a) - Q^{\mathrm{ARP},\star}(n,s,a)\big| \leqslant C\Big[1 - \ln\Big(\mathrm{Tol}_{\mathrm{error}}\Big)\Big]\mathrm{Tol}_{\mathrm{error}},$$

on an event of probability greater than

$$1 - C\mathrm{Tol}_{\mathrm{error}} - C\mathrm{Tol}_{\mathrm{prob}}.$$

This requires

$$n \geqslant \exp\big(C|\ln(\mathrm{Tol}_{\mathrm{error}})|^3\big)$$

observations and this yields an approximation of the original $Q^{\star,0}$ *action-value function at order* Tol_{error} *(when the value functions are Lipschitz continuous).*

Obviously, the thrust of this result is that the dimensions of the state and action spaces just come through the various constants.

We now prove:

Proof of Lemma 3.2.9. Given the expressions of $Q^{\star,0}$ and $Q^{\star}$, it suffices to prove the same bound but for $V^{\star,0}$ and $V^{\star}$.

The main point is to observe from (3.2.23) that

$$\mathcal{W}_1\Big(\overline{\mathsf{P}}\big((s,a),\cdot\big),\overline{\mathsf{P}}^0\big((s,a),\cdot\big)\Big) \leqslant C\big(p+q\big),$$

where $\mathcal{W}_1$ is the standard Wasserstein distance

$$\mathcal{W}_1(\mu,\nu) = \sup_{\varphi} \int_S \varphi d\big(\mu-\nu\big), \quad \mu,\nu\in\mathcal{P}(\overline{S}),$$

the supremum being taken over all the Lipschitz continuous (and hence bounded) functions on S with 0 mean. Back to the Bellman equation, we deduce that

$$\sup_{s\in\overline{S}} \big|V^{\star}(s) - V^{\star,0}(s)\big| \leqslant \gamma \sup_{s\in\overline{S}} \big|V^{\star}(s) - V^{\star,0}(s)\big| + C\big(p+q\big),$$

from which the result easily follows. □

3.2.8 Further prospects

Here are some possible directions of research:

1. It is clear that the various constants appearing in the rates of convergence depend on the geometries of the space and action domains. We think it possible to specify the explicit dependence upon the diameters of those domains. Intuitively, the rates should be better on smaller domains. As an application of this, we hope to be able to design some refinement of the method close to the optimal policy, in order to get better accuracy of the action value function in the neighborhood of the latter.

2. In connection with mean field control, one should come back to the original problem with a finite number of players and then see how the Q-function that is learnt for the limiting (over the size of the population) problem provides a relevant approximation of the Q-function for the particle system. This question looks very close to questions related to the rate of convergence of the value functions in mean field control problem, see for instance [**CecchinFinite**, 25, 26].

3. In connection with the previous item, one should also wonder about the practical implementation of the common noise when dealing with the particle system. Assuming for instance that the system represents a flock of bots, one understands that, in order to realize the common noise, one may need to reshuffle the locations of the bots. Whereas the mean field control problem features a lot of symmetries, there might be in fact a break of symmetry when reallocating new positions to the bots: one may guess that there should exist an additional cost that one should pay for reallocating the positions of the bots (in addition to the effective feasibility of the operation). Possibly, we could address this question by introducing additional transport costs (in connection with questions of optimal transportation).

3.3 Proofs of the estimates for the repeated covering times

The purpose of this section is to prove Propositions 3.2.4 and 3.2.6.

3.3.1 Proof of Proposition 3.2.4

The proof of Proposition 3.2.4 relies on the following lemma:

Lemma 3.3.1. *Let $d \geqslant 1$ be an integer and $E \subset \mathbb{R}^d$ be an open subset satisfying a uniform (interior) cone condition. Then, there exist two constants, c_0 and C_0, only depending on E, such that, for any $\delta > 0$, there exist an integer J, with $c_0\delta^{-d} \leqslant J \leqslant C_0\delta^{-d}$, J elements $(x_j)_{j=1,\cdots,J}$, J pairwise disjoint balls $(B_j(c_0\delta))_{j=1,\cdots,J}$, of radius $c_0\delta$ each and respectively of centers $(x_j)_{j=1,\cdots,J}$, and J balls $(B_j(C_0\delta))_{j=1,\cdots,J}$, of radius $C_0\delta$ each and respectively of centers $(x_j)_{j=1,\cdots,J}$, such that*

$$\bigcup_{j=1}^{J} B_j(c_0\delta) \subset E \subset \bigcup_{j=1}^{J} B_j(C_0\delta).$$

Intuitively, Lemma 3.3.1 says that we can partition $\bar{S} \times \bar{A}$ into subsets that have (up to a multiplicative to constant) the same volume as a ball of radius h. This property plays a key role in the proof of Proposition 3.2.4: when h is not too small, assumption 3.1.1 provides a uniform lower bound for the probability of reaching each of these subsets. If the interior cone condition were not satisfied, we could think, as counter-examples, of domains with arbitrarily thiner tubes of fixed length where the process $(\mathrm{s}_n, \mathrm{a}_n)_{n \geqslant 0}$ would go with arbitrarily low probability.

The proof of Lemma 3.3.1 is postponed to the end of the subsection. We now directly turn to:

Proof of Proposition 3.2.4. We recall the notation $D = d_S + d_A$, which we use throughout the proof.

By means of Lemma 3.3.1 in dimension d_S and d_A respectively (with $c_0\delta = h$ in both cases, but with c implicitly depending on the dimension), we can cover $\overline{S} \times \overline{A}$ by J subsets (called cells) $B_1, \cdots, B_J$, with the properties that (using obvious notation similar to the notation of the statement of Lemma 3.3.1) $c_0^{D+1} h^{-D} \leqslant J \leqslant C_0 c_0^D h^{-D}$ and that B_j contains, for each $j = 1, \cdots, J$, the product of two balls of radius h, one in dimension d_S and one in dimension d_A. For any $j = 1, \cdots, J$, we denote by B'_j by the D-dimensional ball with the same center as B_j and with radius $3\varrho h$. Without any loss of generality, we can assume $J \geqslant 3$.

We notice from (3.1.1) that, for an arbitrary $n \geqslant 0$,

$$1 = \mathbb{P}\Big(\big\{(\mathrm{s}_n, \mathrm{a}_n) \in \overline{S} \times \overline{A}\big\}\Big) \geqslant \sum_{j=1}^{J} \mathbb{P}\Big(\big\{(\mathrm{s}_n, \mathrm{a}_n) \in B_j\big\}\Big) \geqslant \sum_{j=1}^{J} \eta h^D = \eta J h^D \geqslant \eta c_0^{D+1}. \tag{3.3.1}$$

We use the two bounds $\eta J h^D \leqslant 1$ in the proof. The bound $\eta c_0^{(D+1)} \leqslant 1$ is just given for information but we don't use it: as a main drawback, the constant c_0 depends on the geometry of the domain $\overline{S} \times \overline{A}$.

Throughout the proof, we denote by ℓ a time from which the process $(\mathrm{s}_n, \mathrm{a}_n)_{n \geqslant 0}$ is considered. Moreover, for $(s, a) \in \overline{S} \times \overline{A}$, we call $\jmath(s, a)$ the unique index $j \in \{1, \cdots, J\}$ such that $(s, a) \in B_j$.

First Step. Let T_k be the first time after ℓ when each set has been visited at least k times by the process $(\mathrm{s}_n, \mathrm{a}_n)_{n \geqslant 0}$:

$$T_k = \min\left\{ n \geqslant \ell : \min_{1 \leqslant j \leqslant J} \sum_{i=\ell}^{n} \mathbf{1}_{B_j}\big(\mathrm{s}_i, \mathrm{a}_i\big) \geqslant k \right\}, \tag{3.3.2}$$

with the convention that $T_0 = \ell - 1$. We prove below that each T_k is almost surely finite.

Elaborating on the first steps of Carden's proof, we get, for any integer $n \geqslant 1$ and any $(s, a) \in \overline{S} \times \overline{A}$,

$$\sum_{i=\ell}^{T_n} \alpha_i\big(s, a\big) = \sum_{k=1}^{n} \sum_{i=T_{k-1}+1}^{T_k} \alpha_i\big(s, a\big) \geqslant \frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty} \sum_{k=1}^{n} \left(\max_{1 \leqslant j \leqslant J} \sum_{i=0}^{T_k - 1} \mathbf{1}_{B'_j}\big(\mathrm{s}_i, \mathrm{a}_i\big) \right)^{-1}, \tag{3.3.3}$$

where we used the fact that, between $T_{k-1} + 1$ and T_k, the process $(\mathrm{s}_i, \mathrm{a}_i)_{i \geqslant 1}$ passed at least once

by the cell containing (s,a). In words, there exists a least one $i \in \{T_{k-1}+1, \cdots, T_k\}$, such

$$\alpha_i\big(s,a\big) = \lambda_{\mathscr{K}}\bigg(\sum_{j=0}^{i-1} \mathscr{K}\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)\bigg)^{-1} \geqslant \lambda_{\mathscr{K}}\bigg(\sum_{j=0}^{T_k-1} \|\mathscr{K}\|_\infty \mathbf{1}_{B_{\varrho h}(s,a)}\big(\mathrm{s}_j,\mathrm{a}_j\big)\bigg)^{-1}$$
$$= \frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\bigg(\sum_{j=0}^{T_k-1} \mathbf{1}_{B'}\big(\mathrm{s}_j,\mathrm{a}_j\big)\bigg)^{-1},$$

where, in the first line, we denoted by $B_{\varrho h}(s,a)$ the D-dimensional ball of center (s,a) and of radius ϱh and where, in the second line, we denoted by B' the ball $B'_{\mathrm{J}(s,a)}$. The fact that we can pass from the ball $B_{\varrho h}(s,a)$ in the first line to the ball B' in the second line is justified as follows: Since $(s,a) \in B_{\mathrm{J}(s,a)}$, it holds that

$$B_{\varrho h}(s,a) \subset \Big\{(s',a') \in \bar{S}\times\bar{A} : \mathrm{dist}\Big((s',a'), B_{\mathrm{J}(s,a)}\Big) \leqslant \varrho h\Big\} \subset B'_{\mathrm{J}(s,a)} = B'.$$

In order to get the second inclusion, we used the triangular inequality together with the fact that $B_{\mathrm{J}(s,a)}$ is included in the D-dimensional ball of center $(\mathrm{s}_{\mathrm{J}(s,a)}, \mathrm{a}_{\mathrm{J}(s,a)})$ and of radius $\sqrt{2}h < 2\varrho h$.

For comparison, the reader should observe that, in Carden's paper, the weaker lower bound is used in lieu of (3.3.3):

$$\alpha_i\big(s,a\big) \geqslant \lambda_{\mathscr{K}}\bigg(\sum_{j=0}^{T_k-1} \|\mathscr{K}\|_\infty\bigg)^{-1}.$$

Although the above is sufficient to prove the convergence of the learning algorithm introduced in Carden, it leads in fact to very poor quantitative bounds, whence the need for the refined version (3.3.3).

Second Step. While most of the proof is dedicated to the obtention of an upper bound for the random variables $(T_k)_{k\geqslant 1}$, we start with a lower bound for T_1. The motivation is the following: in order to handle (3.3.3), we need an averaging result for the sum $\sum_{i=0}^{T_k-1} \mathbf{1}_{B'_j}(\mathrm{s}_i,\mathrm{a}_i)$, which requires T_k (and thus T_1 when $k=1$) to be large enough.

In fact, both the methods to get a lower bound for T_1 and an upper bound for the variables $(T_k)_{k\geqslant 1}$ are very much inspired by the coupon collector problem. Here, the coupon collector problem we use to lower bound T_1 is associated with iterated discoveries of new cells of the form $(B'_j)_{j=1,\cdots,J}$. We let $\tau'_1 := \ell$ (this is understood as the first discovery time of a new cell). By induction and as long as $i \in \{2,\cdots,J\}$, we define τ'_i as the discovery time of the ith new cell:

$$\tau'_i := \min\left\{ n > \tau'_{i-1} : (\mathrm{s}_n,\mathrm{a}_n) \notin \bigcup_{k=1}^{\tau'_{i-1}} B'_{\mathrm{J}(\mathrm{s}_k,\mathrm{a}_k)} \right\}. \tag{3.3.4}$$

For a given $i \in \{2,\cdots,J\}$ and for $k \geqslant \ell$, we have, on the event $\{\tau'_{i-1} \leqslant k < \tau'_i\}$ (which belongs to

$\mathcal{F}_k$),

$$\begin{aligned}\mathbb{P}\Big(\{\tau'_i = k+1\}|\mathcal{F}_k\Big) &\leqslant \mathbb{P}\Big(\bigcup_{j\notin(\mathrm{J}(\mathrm{s}_i,\mathrm{a}_i))_{1\leqslant i\leqslant k}} B'_j \,|\, \mathcal{F}_k\Big)\\ &\leqslant \sum_{j\notin\{\mathrm{J}(\mathrm{s}_{\tau_l},\mathrm{a}_{\tau_l}),1\leqslant l\leqslant i-1\}} \mathbb{P}\Big(B'_j \,|\, \mathcal{F}_k\Big) \qquad (3.3.5)\\ &\leqslant \eta'\Big(J-(i-1)\Big)h^D = \eta' J h^D\Big(1-\frac{i-1}{J}\Big).\end{aligned}$$

So, conditional on $\mathcal{F}_{\tau'_{i-1}}$, $t'_i := \tau'_i - \tau'_{i-1}$ is stochastically bounded from below by a geometric random variable of parameter $\min(1, \eta' J h^D[1-(i-1)/J])$. We deduce that T_1 is stochastically lower bounded by $\sum_{i=1}^J \widetilde{t}'_i$, where $\widetilde{t}'_1, \cdots, \widetilde{t}'_J$ are independent random variables of geometric distributions, with $(\min(1, \eta' J h^D[1-(i-1)/J]))_{1\leqslant i\leqslant J}$ as parameters of success. We then apply Lemma 3.6.2 in the appendix. We get that there exists a universal constant C such that, for any $r > 0$,

$$\mathbb{P}\Big(\sum_{i=1}^J \widetilde{t}'_i < \frac{J\ln(J/[2\widetilde{\eta}'])}{2\widetilde{\eta}'}\Big) \leqslant \exp\big(1+Cr^2\big)\Big(\frac{2\widetilde{\eta}'}{J}\Big)^{r/2},$$

with $\widetilde{\eta}' = \eta' J h^D$, which is less than J since $\eta' h^D \leqslant 1$, and then

$$\mathbb{P}\Big(T_1 \leqslant \Big\lceil -\frac{\ln(2\eta' h^D)}{2\eta' h^D}\Big\rceil\Big) = \mathbb{P}\Big(T_1 \leqslant \Big\lceil \frac{\ln(1/[2\eta' h^D])}{2\eta' h^D}\Big\rceil\Big) \leqslant \exp\big(1+Cr^2\big)\Big(2\eta' h^D\Big)^{r/2}. \qquad (3.3.6)$$

Back to (3.3.3), we now provide an upper bound for the sequence

$$\Big(\frac{1}{n}\sum_{i=0}^{n-1} \mathbf{1}_{B'}\big(\mathrm{s}_i,\mathrm{a}_i\big)\Big)_{n\geqslant 1},$$

for a given cell B' of the type B'_j for $j \in \{1, \cdots, J\}$. Thanks to (**??**), we know that the above mean can be dominated by

$$\Big(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\Big)_{n\geqslant 1},$$

where $(\varepsilon_n)_{n\geqslant 1}$ is a sequence of independent and identically distributed Bernoulli random variables with parameter $\eta' h^D$, in the sense that

$$\mathbf{1}_{B'}\big(\mathrm{s}_i,\mathrm{a}_i\big) \leqslant \varepsilon_{i+1}, \quad i \geqslant 0.$$

We then deduce from Lemma 3.6.1 that, for $\beta > e^2$,

$$\mathbb{P}\Big(\bigcup_{n\geqslant\lceil -\ln(\eta' h^D/2)/\{2\eta' h^D\}\rceil} \Big\{\frac{1}{n}\sum_{k=1}^{n}\varepsilon_k \geqslant \beta\eta' h^D\Big\}\Big) \leqslant \big(\eta' h^D\big)^{\beta/2-1}. \qquad (3.3.7)$$

Therefore,

$$\mathbb{P}\Big(\bigcup_{n\geqslant\lceil -\ln(\eta' h^D/2)\rceil/\{2\eta' h^D\}} \Big\{\frac{1}{n}\sum_{k=1}^{n}\mathbf{1}_{B'}\big(\mathrm{s}_i,\mathrm{a}_i\big) \geqslant \beta\eta' h^D\Big\}\Big) \leqslant \big(\eta' h^D\big)^{\beta/2-1},$$

and then,

$$\mathbb{P}\Big(\bigcup_{j=1,\cdots,J}\bigcup_{n\geqslant[-\ln(\eta' h^D/2)]/\{2\eta' h^D\}}\Big\{\frac{1}{n}\sum_{k=0}^{n-1}\mathbf{1}_{B'_j}(\mathrm{s}_i,\mathrm{a}_i)\geqslant\beta\eta' h^D\Big\}\Big)\leqslant J\big(\eta' h^D\big)^{\beta/2-1}.$$

Combining with (3.3.6) and recalling that $\eta'\geqslant 1$, we get (for a new value of C)

$$\mathbb{P}\Big(\bigcup_{j=1,\cdots,J}\bigcup_{n\geqslant T_1}\Big\{\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{1}_{B'_j}(\mathrm{s}_i,\mathrm{a}_i)\geqslant\beta\eta' h^D\Big\}\Big)\leqslant\exp\big(Ca\beta^2\big)J\Big(\eta' h^D\Big)^{\beta/2-1}.$$

Back to (3.3.3), we get, for $n\geqslant 1$,

$$\mathbb{P}\Big(\Big\{\sum_{i=\ell}^{T_n}\alpha_i(\mathrm{s},\mathrm{a})\geqslant\frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{1}{\beta\eta' h^D}\ \sum_{k=1}^{n}\frac{1}{T_k}\Big\}\Big)\geqslant 1-\exp\big(C\beta^2\big)J\Big(\eta' h^D\Big)^{\beta/2-1}. \tag{3.3.8}$$

Third Step. As made clear in (3.3.8), the next step is to provide a lower bound $1/T_k$ and thus to upper bound T_k. Once again, the argument is very much inspired from the coupon collector problem, but there are some subtle differences from the second step, which requires some care.

Following the previous step, we first explain how to upper bound T_1. To do so, we define the analogue of τ'_1 in (3.3.4) but with respect to the balls $(B_j)_{j=1,\cdots,J}$ instead of $(B'_j)_{j=1,\cdots,J}$. We thus let $\tau_1=\ell$ and, as as long as $i\in\{2,\cdots,J\}$, we define by induction τ_i as the discovery time of the ith new cell:

$$\tau_i=\min\left\{n>\tau_{i-1}:(\mathrm{s}_n,\mathrm{a}_n)\notin\bigcup_{k=1}^{\tau_{i-1}}B_{\mathrm{J}(\mathrm{s}_k,\mathrm{a}_k)}\right\}.$$

For $i\geqslant 1$ and $k\geqslant 0$, we have, on the event $\{\tau_{i-1}\leqslant k<\tau_i\}$,

$$\begin{aligned}\mathbb{P}\Big(\{\tau_i=k+1\}|\mathcal{F}_k\Big)&\geqslant\mathbb{P}\Big(\bigcup_{j\notin(\ell(\mathrm{s}_i,\mathrm{a}_i))_{1\leqslant i\leqslant k}}B_j|\mathcal{F}_k\Big)\\&=\sum_{j\notin\{\ell(\mathrm{s}_{\tau_j},\mathrm{a}_{\tau_j}),1\leqslant j\leqslant i-1\}}\mathbb{P}\Big(B_j|\mathcal{F}_k\Big)\\&\geqslant\eta\Big(J-(i-1)\Big)h^D=\eta Jh^D\Big(1-\frac{i-1}{J}\Big).\end{aligned}$$

Here, we recall from (3.3.1) that $\eta Jh^D\leqslant 1$. So, conditional on $\mathcal{F}_{\tau_{i-1}}$, $t_i:=\tau_i-\tau_{i-1}$ is stochastically bounded from above by a geometric random variable of parameter $\eta Jh^D(1-(i-1)/J)$. We deduce that $T_1-\ell$ is stochastically dominated by

$$\sum_{i=1}^{J}\widetilde{t}_i,$$

where $\widetilde{t}_1,\cdots,\widetilde{t}_J$ are independent random variables of geometric distributions, with $\eta Jh^D(1-(i-1)/J)_{1\leqslant i\leqslant J}$ as parameters of success. We now let

$$W_1=\sum_{i=1}^{J}\widetilde{t}_i,$$

and use a concentration inequality on W_1 stated in Lemma 3.6.2 in the appendix. We deduce from item (i) therein that there exists a universal constant C such that, for any $r \in (0,1)$,

$$\mathbb{E}\Big[\exp\Big(r\eta h^D\big[W_1 - \mathbb{E}(W_1)\big]\Big)\Big] = \mathbb{E}\Big[\exp\Big(r\frac{\eta J h^D}{J}\big[W_1 - \mathbb{E}(W_1)\big]\Big)\Big] \leqslant \exp\big(C\frac{r^2}{1-r}\big), \tag{3.3.9}$$

where $\widetilde{W}_1$ is obtained from W_1 from a constructive manner and satisfies $W_1 - 1/(\eta h^D) \leqslant \widetilde{W}_1 \leqslant W_1$.

Back to the first step, we can now (stochastically) bound the conditional law of the increment $T_k - T_{k-1}$ given $\mathcal{F}_{T_{k-1}}$ by the law of a new coupon collector problem as defined in the second and third step. We deduce that

$$T_k \leqslant \ell + \sum_{j=1}^{k} W_j, \quad k \geqslant 1, \tag{3.3.10}$$

where $(W_k)_{k\geqslant 1}$ is a collection of independent random variables of the same law as W_1. And then,

$$\sum_{k=1}^{n} \frac{1}{T_k} \geqslant \sum_{k=1}^{n} \left(\ell + \sum_{j=1}^{k} W_j\right)^{-1},$$

which prompts us to define, for any $\kappa > 0$, the event:

$$A_n(\kappa) = \left\{ \max_{1\leqslant k\leqslant n} \left[\sum_{j=1}^{k} (\widetilde{W}_j - \mathbb{E}[\widetilde{W}_1]) \right] < \kappa \right\}, \tag{3.3.11}$$

where $(\widetilde{W}_1, \cdots, \widetilde{W}_n)$ are independent and identically distributed random variables obtained by means of Lemma 3.6.2 and satisfy $W_i - 1/(\eta h^D) \leqslant \widetilde{W}_i \leqslant W_i$.

We notice that

$$\mathbb{E}[W_1] = \sum_{i=1}^{J} \mathbb{E}[\tilde{t}_i] = \frac{J}{\eta J h^D} \sum_{i=1}^{J} \frac{1}{J-(i-1)} = \frac{1}{\eta h^D} \sum_{i=1}^{J} \frac{1}{i}.$$

It is well-known that

$$\ln(J) = \int_1^J \frac{1}{x} dx \leqslant \sum_{i=1}^{J} \frac{1}{i} \leqslant \int_1^J \frac{1}{x} dx + 1 = \ln(J) + 1,$$

and then,

$$\frac{\ln(J)}{\eta h^D} \leqslant \mathbb{E}[W_1] \leqslant \frac{\ln(J)+1}{\eta h^D} = E_{\eta,J,h}. \tag{3.3.12}$$

In the end, we obtain, on the event $A_n(\kappa)$,

$$\forall k \in \{1, \cdots, n\}, \quad \sum_{j=1}^{k} W_j \leqslant \sum_{j=1}^{k} \widetilde{W}_j + \frac{k}{\eta h^D} \leqslant \kappa + k\widetilde{E}_{\eta,J,h}, \quad \widetilde{E}_{\eta,J,h} := \frac{\ln(J)+2}{\eta h^D}. \tag{3.3.13}$$

Fourth Step. (**Good event analysis**.) We now use (3.3.10). We deduce that, on $A_n(\kappa)$, for all $k \in \{1, \cdots, n\}$,

$$T_k \leqslant \ell + \sum_{j=1}^{k} W_j \leqslant \ell + \kappa + k\widetilde{E}_{\eta,J,h}. \tag{3.3.14}$$

Then,

$$\sum_{k=1}^{n} T_k^{-1} \geqslant \sum_{k=1}^{n} \Big(\ell + \sum_{j=1}^{k} W_j\Big)^{-1} \geqslant \sum_{k=1}^{n} \Big(\ell + \kappa + k\widetilde{E}_{\eta,J,h}\Big)^{-1} \geqslant \int_1^n \frac{1}{\ell + \kappa + x\widetilde{E}_{\eta,J,h}} dx$$
$$= \frac{1}{\widetilde{E}_{\eta,J,h}} \ln\Big(\frac{\ell + \kappa + n\widetilde{E}_{\eta,J,h}}{\ell + \kappa + \widetilde{E}_{\eta,J,h}}\Big).$$

Therefore, intersecting with the event introduced in (3.3.8), we obtain

$$\inf_{(\mathrm{s},\mathrm{a}) \in \bar{S} \times \bar{A}} \sum_{j=\ell}^{T_n} \alpha_j(\mathrm{s},\mathrm{a}) \geqslant \frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty} \frac{1}{\beta\eta' h^D \widetilde{E}_{\eta,J,h}} \ln\Big(\frac{\ell + \kappa + n\widetilde{E}_{\eta,J,h}}{\ell + \kappa + \widetilde{E}_{\eta,J,h}}\Big),$$

on the event

$$\widetilde{A}_n(\kappa) := A_n(\kappa) \cap \Big\{\sum_{i=\ell}^{T_n} \alpha_i(\mathrm{s},\mathrm{a}) \geqslant \frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty} \frac{1}{\beta\eta' h^D} \sum_{k=1}^{n} \frac{1}{T_k}\Big\}.$$

Recalling the definition of $\widetilde{E}_{\eta,J,h}$ in (3.3.13) and using (3.3.14), we have, on $\widetilde{A}_n(\kappa)$, for all $(s,a) \in \overline{S} \times \overline{A}$,

$$\begin{aligned} \sum_{j=\ell}^{\ell + \lfloor \kappa + n\widetilde{E}_{\eta,J} \rfloor} \alpha_j(s,a) &\geqslant \frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty} \frac{\eta}{\beta\eta'(\ln(J)+1)} \ln\Big(\frac{\ell + \kappa + n\widetilde{E}_{\eta,J,h}}{\ell + \kappa + \widetilde{E}_{\eta,J,h}}\Big) \\ &\geqslant \frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty} \frac{\eta}{2\beta\eta' \ln(J)} \ln\Big(\frac{\ell + \kappa + n\widetilde{E}_{\eta,J,h}}{\ell + \kappa + \widetilde{E}_{\eta,J,h}}\Big) \end{aligned} \tag{3.3.15}$$

where we used the fact that $J \geqslant 3$, we have $\ln(J) \geqslant 1$ and $2\ln(J) \geqslant 2$, which implies $2\ln(J) \geqslant \ln(J) + 1$.

Fifth Step. (**Bad event analysis**.) We bound $\mathbb{P}(A_n(\kappa)^{\complement})$ from above. Let

$$V_k = \sum_{j=1}^{k} \widetilde{W}_j, \quad M_k = V_k - k\mathbb{E}[\widetilde{W}_1], \quad k \geqslant 1\ ; \quad \tau = \min\big\{k \geqslant 1 : M_k \geqslant \kappa\big\}.$$

Note that $(M_k)_{1 \leqslant k \leqslant n}$ is a martingale and, thus, for any $r \in (0,1)$, $(\exp(r\eta h^D M_k/2))_{1 \leqslant k \leqslant n}$ is a submartingale. Hence, by Doob's martingale inequality and by (3.3.9),

$$\begin{aligned} \mathbb{P}\big(A_n(\kappa)^{\complement}\big) = \mathbb{P}\big(\{\tau \leqslant n\}\big) &\leqslant \inf_{r \in (0,1)} \Big[\exp\Big(-nr\kappa\eta \frac{h^D}{2}\Big) \mathbb{E}\Big[\exp\Big(r\eta \frac{h^D}{2} M_n\Big)\Big]\Big] \\ &\leqslant \inf_{r \in (0,1/2)} \Big[\exp\Big(-nr\kappa\eta \frac{h^D}{2} + Cn\frac{r^2}{2}\Big)\Big] \\ &= \inf_{r \in (0,1/2)} \Big[\exp\Big(\frac{Cn}{2}\Big[r - \frac{\kappa\eta h^D}{2Cn}\Big]^2 - \frac{\kappa^2\eta^2 h^{2D}}{8Cn}\Big)\Big]. \end{aligned} \tag{3.3.16}$$

In particular, if $\kappa\eta h^D < Cn$, then

$$\mathbb{P}\big(A_n(\kappa)^{\complement}\big) \leqslant \exp\Big(-\frac{\kappa^2\eta^2 h^{2D}}{4Cn}\Big). \tag{3.3.17}$$

If $\kappa\eta h^D \geqslant Cn$, then, by choosing $r = 1/2$ in the second line of (3.3.17), we get

$$\mathbb{P}\big(A_n(\kappa)^{\complement}\big) \leqslant \exp\Big(-\kappa\eta\frac{h^D}{4} + \frac{Cn}{8}\Big) \leqslant \exp\Big(-\kappa\eta\frac{h^D}{4} + \kappa\eta\frac{h^D}{8}\Big) = \exp\Big(-\kappa\eta\frac{h^D}{8}\Big). \tag{3.3.18}$$

And then, we can combine (3.3.17) into

$$\mathbb{P}\big(A_n(\kappa)^{\complement}\big) \leqslant \exp\Big(-\min\Big[\frac{\kappa\eta h^D}{8}, \frac{\kappa^2\eta^2 h^{2D}}{4Cn}\Big]\Big), \tag{3.3.19}$$

which holds true whatever the value of κ.

Conclusion. We take an integer $m \geqslant 12\widetilde{E}_{\eta,J,h}$ and, for a fixed $\delta \in (0, 1/2)$, we let $n \geqslant 1$ (n being an integer) such that

$$\big(n^{1/2+\delta} + n\big)\widetilde{E}_{\eta,J,h} \leqslant m < \Big[\big(n+1\big)^{1/2+\delta} + n + 1\Big]\widetilde{E}_{\eta,J,h}.$$

We observe that

$$m < 2(n+1)\widetilde{E}_{\eta,J,h} \leqslant 4n\widetilde{E}_{\eta,J,h}.$$

Then,

$$5 \leqslant \frac{m}{2\widetilde{E}_{\eta,J,h}} - 1 \leqslant n \leqslant \frac{m}{\widetilde{E}_{\eta,J,h}}.$$

We then take $\kappa = n^{1/2+\delta}\widetilde{E}_{\eta,J,h}$ in (3.3.11). Using the fact that $\widetilde{E}_{\eta,J,h} \leqslant \kappa/2$ and $\widetilde{E}_{\eta,J,h} \leqslant (4/3)E_{\eta,J,h}$ (because $J \geqslant 9$ implies $\ln(J) \geqslant 2$ and then $(4/3)[\ln(J)+1] = \ln(J)+1+(\ln(J)/3+1/3) \geqslant \ln(J)+2$) together with (3.3.8), (3.3.12), (3.3.15) and (3.3.19), we end-up with

$$\begin{aligned}
&\mathbb{P}\bigg(\Big\{\inf_{(\mathrm{s},\mathrm{a})\in\overline{S}\times\overline{A}}\sum_{j=\ell}^{\ell+m}\alpha_j(\mathrm{s},\mathrm{a}) \geqslant \frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{\eta}{2\beta\eta'\ln(J)}\ln\Big(\frac{\ell+m/2}{\ell+2m^{1/2+\delta}E_{\eta,J,h}^{1/2-\delta}}\Big)\Big\}\bigg)\\
&\geqslant 1 - \exp\big(C\beta^2\big)J\Big(\eta' h^D\Big)^{\beta/2-1} - \exp\Big(-\min\Big[\frac{(\ln(J)+1)n^{1/2+\delta}}{8}, \frac{(\ln(J)+1)^2 n^{2\delta}}{4C}\Big]\Big)\\
&\geqslant 1 - \exp\big(C\beta^2\big)J\Big(\eta' h^D\Big)^{\beta/2-1} - \exp\Big(-\frac{(\ln(J)+1)n^{2\delta}}{\max(8,4C)}\Big).
\end{aligned}$$

For a new choice of C, we obtain

$$\begin{aligned}
&\mathbb{P}\bigg(\Big\{\inf_{(\mathrm{s},\mathrm{a})\in\overline{S}\times\overline{A}}\sum_{j=\ell}^{\ell+m}\alpha_j(\mathrm{s},\mathrm{a}) \geqslant \frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{\eta}{2\beta\eta'\ln(J)}\ln\Big(\frac{\ell+m/2}{\ell+2m^{1/2+\delta}E_{\eta,J,h}^{1/2-\delta}}\Big)\Big\}\bigg)\\
&\geqslant 1 - \exp\big(C\beta^2\big)J\Big(\eta' h^D\Big)^{\beta/2-1} - \exp\Big(-\frac{(\ln(J))^{2(1-\delta)}(\eta h^D m)^{2\delta}}{C}\Big).
\end{aligned}$$

The bound remains trivially true when $m < 12\widetilde{E}_{\eta,J,h} \leqslant 16E_{\eta,J,h}$: it suffices to observe that the logarithm inside the probability becomes negative. □

We now end-up the subsection with the proof of Lemma 3.3.1.

Proof of Lemma 3.3.1. We start with the following obvious remark: it suffices to make the proof for δ small enough. We then recall the following two results:

1. From [116], the set E can be decomposed into the countable union of closed dyadic cubes $(Q_j)_{j\in\mathbb{N}}$, with pairwise disjoint interiors, such that

$$\sqrt{d}\,\mathrm{diam}(Q_j) \leqslant \mathrm{dist}\big(Q_j, \partial E\big) \leqslant 4\sqrt{d}\,\mathrm{diam}(Q_j), \quad j \in \mathbb{N},$$

   where ∂E denotes the boundary of E and dist is here used to denote the (Euclidean) distance between two subsets of $\mathbb{R}^d$. Below, we call x_j the center of each cube Q_j.

2. From [6], there exist $\delta > 0$ and a finite collection of J cones $(\mathscr{C}_j)_{1\leqslant j\leqslant J}$ such that, for any $x \in E$ with $\mathrm{dist}(x, \partial E) < \delta$, $x + \mathscr{C}_j \subset E$. In particular, assuming without loss of generality that the heights of the cones are greater than δ, we can find a (small) constant $c > 0$, only depending on E, such that, for any $x \in E$ with $\mathrm{dist}(x, \partial E) < \delta$, there exists a cube Q of radius $c\delta$ such that $Q \subset E$, $\mathrm{dist}(x, Q) < \delta$ and $\mathrm{dist}(Q, \partial E) > 18\sqrt{d}c\delta$.

We now combine 1 and 2. For a fixed $x \in E$ with $\mathrm{dist}(x, \partial E) < \delta$, we call j the index such that the center x_Q of the cube Q (as in item 2) belongs to the interior of Q_j (as in item 1). If the radius of Q_j is less than $c\delta$, then

$$\mathrm{dist}\big(Q_j, \partial E\big) \geqslant \mathrm{dist}\big(x_Q, \partial E\big) - \mathrm{diam}\big(Q_j\big) = \mathrm{dist}\big(Q, \partial E\big) - 2c\delta \geqslant 16\sqrt{d}c\delta,$$

which implies (from the first item) that the diameter of Q_j is greater than $4\sqrt{d}c\delta \geqslant 4\delta$, hence obtaining a contradiction.

Moreover, again by the first item,

$$\begin{aligned}\mathrm{diam}(Q_j) \leqslant \mathrm{dist}\big(Q_j, \partial E\big) &\leqslant \mathrm{dist}\big(x_Q, \partial Q\big) + \mathrm{dist}\big(Q, \partial E\big)\\ &\leqslant \mathrm{dist}\big(x_Q, \partial Q\big) + \mathrm{dist}\big(x, \partial E\big) + \mathrm{dist}\big(x, Q\big)\\ &\leqslant (2 + c)\delta,\end{aligned}$$

from which we deduce that

$$|x - x_j| \leqslant \mathrm{dist}(x, Q) + c\delta + |x_Q - x_j| \leqslant (3 + 2c)\delta.$$

For a well-chosen C, we get that x belongs to the ball $B_j(C\delta)$ of center x_j and of radius $C\delta$, Also, $\mathrm{diam}(Q_j) \geqslant 2c\delta$

Observe now that if $x \in E$ but $\mathrm{dist}(x, \partial E) \geqslant \delta$, then $x \in Q_j$ for some j. If $\mathrm{diam}(Q_j) < 2c\delta$, then the first item yields $\mathrm{dist}(Q_j, \partial E) \leqslant 8\sqrt{d}c\delta$ and, in turn, $\mathrm{dist}(x, \partial E) \leqslant 10\sqrt{d}c\delta$. Assuming without any loss of generality that $10\sqrt{d}c < 1$, we get a contradiction. So, we deduce that $\mathrm{diam}(Q_j) \geqslant 2c\delta$.

This prompts us to call $\mathcal{J} = \{j : \mathrm{diam}(Q_j) \geqslant 2c\delta\}$. Our analysis shows that

$$\bigcup_{j\in\mathcal{J}} B_j(c\delta) \subset E \subset \bigcup_{j\in\mathcal{J}} B_j(C\delta).$$

For those $j \in \mathcal{J}$ for which $\mathrm{diam}(Q_j) \geqslant 4c\delta$, we can use the dyadic structure to divide them into new dyadic cubes diameters less than or equal to $4c\delta$. For simplicity, we still denote the resulting cubes by Q_j. We then clearly have

$$|J|c^d\delta^d|B_d(0,1)| \leqslant |E| \leqslant |J|C^d\delta^d|B_d(0,1)|,$$

where $|B_d(0,1)|$ is the volume of the d-dimensional ball. □

3.3.2 Proof of Proposition 3.2.6

Proof. Since the Markov process $(\Lambda_n, \Sigma_n, B_n)_{n\in\mathbb{N}}$ is homogeneous, we can assume that $p = 0$.

First Step. We first explain the choice of ℓ_n in the statement. In order to do so, we let $\Gamma := \Gamma_0 E_{\eta,J,h}^{(1+2\gamma)/(1-2\gamma)}$. Then, using the fact that $\ell_0 \geqslant 1$ and $\Gamma^{(1-2\gamma)/(1+2\gamma)} \geqslant 4$, we have, for any $k \geqslant 0$,

$$\begin{aligned}
\ell_{k+1} - \ell_k &\geqslant \ell_0\Big(1 + \Gamma^{(1-2\gamma)/(1+2\gamma)}\Big)^{k+1} - \ell_0\Big(1 + \Gamma^{(1-2\gamma)/(1+2\gamma)}\Big)^{k} - 2 \\
&\geqslant \ell_0\Big(1 + \Gamma^{(1-2\gamma)/(1+2\gamma)}\Big)^{k}\Big[1 + \Gamma^{(1-2\gamma)/(1+2\gamma)} - 3\Big] \\
&\geqslant \frac{\Gamma^{(1-2\gamma)/(1+2\gamma)}}{2}\ell_0\Big(1 + \Gamma^{(1-2\gamma)/(1+2\gamma)}\Big)^{k} \geqslant \frac{\Gamma^{(1-2\gamma)/(1+2\gamma)}}{2}\ell_k.
\end{aligned}$$

By the same argument,

$$\begin{aligned}
\ell_{k+1} - \ell_k &\leqslant \ell_0\Big(1 + \Gamma^{(1-2\gamma)/(1+2\gamma)}\Big)^{k+1} - \ell_0\Big(1 + \Gamma^{(1-2\gamma)/(1+2\gamma)}\Big)^{k} + 1 \\
&\leqslant \big(\ell_k + \mathbf{1}_{\{k\geqslant 1\}}\big)\Gamma^{(1-2\gamma)/(1+2\gamma)} + 1 \leqslant 3\frac{\Gamma^{(1-2\gamma)/(1+2\gamma)}}{2}\ell_k
\end{aligned}$$

Second Step. For an integer $k \geqslant 0$, we deduce from Definition 3.2.1 that, for any two integers $\ell \geqslant 1$ and $L \geqslant 0$, on the event $\{\Lambda_k \geqslant \ell + L\}$,

$$\begin{aligned}
\mathbf{P}_{(\underline{s},\underline{a})}\Big(\big\{\Lambda_{k+1} < \ell\big\} \,|\, \mathcal{G}_k^{(\Lambda,\Sigma,B)}\Big) &\leqslant \prod_{j=\ell+1}^{\Lambda_k}\Big(1 - \alpha_j\big(\Sigma_k, B_k\big)\Big) \\
&\leqslant \exp\Bigg(-\inf_{(s,a)\in\overline{S}\times\overline{A}}\sum_{j=\ell+1}^{\ell+L}\alpha_j\big(s,a\big)\Bigg),
\end{aligned}$$

with the convention that $\prod_{j=1}^{0}[\cdots] = 1$ and, equivalently, that $\sum_{j=1}^{0}[\cdots] = 0$. Therefore, for a

fixed integer $n \geqslant 1$ (as in the statement) and for $k \in \{1, \cdots, n\}$,

$$\begin{aligned}
&\mathbf{P}_{(\underline{s},\underline{a})}\Big(\big\{\Lambda_k < \ell_{n-k}\big\} \,|\, \mathcal{G}_{k-1}^{(\Lambda,\Sigma,B)}\Big) \\
&\leqslant \mathbf{P}_{(\underline{s},\underline{a})}\Big(\big\{\Lambda_{k-1} < \ell_{n+1-k}\big\} \,|\, \mathcal{G}_{k-1}^{(\Lambda,\Sigma,B)}\Big) \\
&\quad + \mathbf{P}_{(\underline{s},\underline{a})}\Big(\big\{\Lambda_k < \ell_{n-k}, \Lambda_{k-1} \geqslant \ell_{n+1-k}\big\} \,|\, \mathcal{G}_{k-1}^{(\Lambda,\Sigma,B)}\Big) \\
&\leqslant \mathbf{P}_{(\underline{s},\underline{a})}\Big(\big\{\Lambda_{k-1} < \ell_{n+1-k}\big\} \,|\, \mathcal{G}_{k-1}^{(\Lambda,\Sigma,B)}\Big) \\
&\quad + \mathbf{E}_{(\underline{s},\underline{a})}\Bigg[\exp\Bigg(-\inf_{(s,a)\in\overline{S}\times\overline{A}} \sum_{j=\ell_{n-k}+1}^{\ell_{n+1-k}} \alpha_j(s,a)\Bigg)\mathbf{1}_{\{\Lambda_{k-1}\geqslant\ell_{n+1-k}\}} \,|\, \mathcal{G}_{k-1}^{(\Lambda,\Sigma,B)}\Bigg].
\end{aligned}$$

By induction, we obtain that, on the event $\{\Lambda_0 \geqslant \ell_n\}$,

$$\begin{aligned}
\mathbf{P}_{(\underline{s},\underline{a})}\Big(\big\{\Lambda_n < \ell_0\big\} \,|\, \mathcal{G}_0^{(\Lambda,\Sigma,B)}\Big) &\leqslant \sum_{k=0}^{n-1} \exp\Bigg(-\inf_{(s,a)\in\overline{S}\times\overline{A}} \sum_{j=\ell_{n-k}+1}^{\ell_{n+1-k}} \alpha_j(s,a)\Bigg) \\
&\leqslant e\sum_{k=0}^{n-1} \exp\Bigg(-\inf_{(s,a)\in\overline{S}\times\overline{A}} \sum_{j=\ell_{n-k}}^{\ell_{n+1-k}} \alpha_j(s,a)\Bigg).
\end{aligned}$$

Third Step. Using the same notation as in Proposition 3.2.4, we now define the collection of events

$$\Bigg\{\inf_{(s,a)\in\overline{S}\times\overline{A}} \sum_{j=\ell_k}^{\ell_{k+1}} \alpha_j(s,a) \geqslant A(\ell_k, \ell_{k+1}-\ell_k)\Bigg\}, \quad k = 0, \cdots, n-1.$$

We then make use of the first step. We know that

$$\frac{1}{2}\Gamma^{(1-2\delta)/(1+2\delta)}\ell_k \leqslant \ell_{k+1} - \ell_k \leqslant \frac{3}{2}\Gamma^{(1-2\delta)/(1+2\delta)}\ell_k.$$

Recalling that

$$A(\ell_k, \ell_{k+1}-\ell_k) = \frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{\eta}{2\beta\eta'\ln(J)}\ln\Big(\frac{\ell_k + (\ell_{k+1}-\ell_k)/2}{\ell_k + 2(\ell_{k+1}-\ell_k)^{1/2+\delta}E_{\eta,J,h}^{1/2-\delta}}\Big),$$

we get

$$A(\ell_k, \ell_{k+1}-\ell_k) \geqslant \frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{\eta}{2\beta\eta'\ln(J)}\ln\Big(\frac{\Gamma^{(1-2\delta)/(1+2\delta)}\ell_k}{4\ell_k + 12\,\Gamma^{1/2-\delta}\ell_k^{1/2+\delta}E_{\eta,J,h}^{1/2-\delta}}\Big).$$

Recall now that $\Gamma = \Gamma_0 E_{\eta,J,h}^{(1+2\delta)/(1-2\delta)} \geqslant 4 \geqslant 1$, with $\Gamma_0 \geqslant 1$. Therefore, since $\ell_k \geqslant 1$,

$$4\ell_k + 12\Gamma^{1/2-\delta}\ell_k^{1/2+\delta}E_{\eta,J,h}^{1/2-\delta} \leqslant 16\Gamma_0^{1/2-\delta}E_{\eta,J,h}\ell_k.$$

We deduce that

$$A(\ell_k, \ell_{k+1}-\ell_k) \geqslant \frac{\lambda_{\mathscr{K}}}{\|\mathscr{K}\|_\infty}\frac{\eta}{2\beta\eta'\ln(J)}\ln\Big(\frac{\Gamma_0^{(1-2\delta)^2/[2(1+2\delta)]}}{16}\Big).$$

It remains to see from Proposition 3.2.4 and from the conclusion of the first step that

$$\begin{aligned}
&\mathbb{P}\Big(\Big\{\inf_{(s,a)\in\overline{S}\times\overline{A}}\sum_{j=\ell_k}^{\ell_{k+1}}\alpha_j(s,a)\geqslant A(\ell_k,\ell_{k+1}-\ell_k)\Big\}\Big)\\
&\geqslant 1-\exp\big(C\beta^2\big)J\Big(\eta' h^D\Big)^{\beta/2-1}-\exp\Big(-\frac{(\ln(J))^{2(1-\delta)}(\eta h^D[\ell_{k+1}-\ell_k])^{2\delta}}{C}\Big)\\
&\geqslant 1-\exp\big(C\beta^2\big)J\Big(\eta' h^D\Big)^{\beta/2-1}-\exp\Big(-\frac{(\ln(J))^{2(1-\delta)}(\eta h^D\ell_k)^{2\delta}\Gamma^{2\delta(1-2\delta)/(1+2\delta)}}{2^{2\delta}C}\Big)\\
&\geqslant 1-\exp\big(C\beta^2\big)J\Big(\eta' h^D\Big)^{\beta/2-1}-\exp\Big(-\frac{(\ln(J))^{2(2+k\delta)}(\eta h^D)^{-2k\delta}\ell_0^{2\delta}\Gamma_0^{2(k+1)\delta(1-2\delta)/(1+2\delta)}}{2^{2\delta}C}\Big).
\end{aligned}$$

We observe that, in the exponential right above, $\ln(J)\geqslant 1$, $\eta h^D\leqslant 1$, $\ell_0\geqslant 1$ and $\Gamma_0\geqslant 1$.

Now we introduce the event

$$D(n)=\bigcap_{k=0}^{n-1}\Big\{\inf_{(s,a)\in\overline{S}\times\overline{A}}\sum_{j=\ell_k}^{\ell_{k+1}}\alpha_j(s,a)\geqslant A(\ell_k,\ell_{k+1}-\ell_k)\Big\}$$

and using the inequality $ab\geqslant(a+b)/2$ for $a,b\geqslant 1$, we deduce that (choosing a new value of the constant C):

$$\begin{aligned}
\mathbb{P}\Big(D(n)\Big)&\geqslant 1-\exp\big(C\beta^2\big)nJ\Big(\eta' h^D\Big)^{\beta/2-1}\\
&\quad-\exp\Big(-\frac{(\ln(J))^4\ell_0^{2\delta}\Gamma_0^{2\delta(1-2\delta)/(1+2\delta)}}{C}\Big)\sum_{k=0}^{n-1}\exp\Big(-\frac{(\ln(J))^{2k\delta}(\eta h^D)^{-2k\delta}\Gamma_0^{2k\delta(1-2\delta)/(1+2\delta)}}{C}\Big)
\end{aligned}$$

We then use the following inequality, that holds true for $a>1$,

$$\sum_{k=0}^{n-1}\exp\big(-a^k\big)\leqslant 1+\int_0^{+\infty}\exp\big(-a^x\big)dx=1+\frac{1}{\ln(a)}\int_0^{+\infty}\exp\big(-e^y\big)dy,$$

where we performed the change of variable $a^x=e^y\Leftrightarrow x=y/\ln(a)$. Since $\ln(J)\geqslant\ln(3)>1$, we deduce that, for a new value of C,

$$\mathbb{P}\Big(D(n)\Big)\geqslant 1-\exp\big(C\beta^2\big)nJ\Big(\eta' h^D\Big)^{\beta/2-1}-\min\big[n,\frac{C}{\delta}\big]\exp\Big(-\frac{(\ln(J))^4\ell_0^{2\delta}\Gamma_0^{2\delta(1-2\delta)/(1+2\delta)}}{C}\Big).$$

□

3.4 Distance Between MDPs: Proof of Proposition 3.2.7

The proof of Proposition 3.2.7 is split into two parts. In the first one, we address the distance between $\overline{\Pi}_{(\underline{s},\underline{a})}((n,s,a),\cdot)$ and $\overline{\mathsf{P}}((s,a),\cdot)$ for fixed s and a. In the second part, we take the supremum over (s,a).

Throughout the section, we use repeatedly the notations (3.2.9) and (3.2.10).

3.4.1 Distance between transition kernels for fixed s and a

We first address the distance between $\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}((n,s,a),\cdot)$ and $\overline{\mathsf{P}}((s,a),\cdot)$ for fixed s and a

Lemma 3.4.1. *For any real $\theta \in (0,1/2)$, we can find two constants C and C_θ, with the second one allowed to depend on θ, with the following property: for any fixed realization $(\underline{\mathrm{s}},\underline{\mathrm{a}})$ of the process $(\mathrm{s}_n,\mathrm{a}_n)_{n\geqslant 0}$, for any integer $n \geqslant 1$, any integer $L \geqslant 2$ and any $(s,a) \in \overline{S}\times\overline{A}$,*

$$\mathbb{P}\Big(\bigcup_{j\geqslant n}\big(F_j(\theta)\big)^{\complement} \cap \Big\{\sum_{k=0}^{n-1}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\leqslant L\Big\}\Big) \leqslant C_\theta L^{-1/\theta+1},$$

where

$$\begin{aligned}F_j(\theta) &= \Big\{\mathrm{d}_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\Big(\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}((n,s,a),\cdot),\overline{\mathsf{P}}((s,a),\cdot)\Big)\\ &\qquad\leqslant C\Big[\Big(\sum_{k=0}^{n-1}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\Big)^{-(1-\theta)/2}+h\Big]\Big\}.\end{aligned}$$

In fact, it must be noted that the set $F_j(\theta)$ can be reformulated in a more explicit way. Indeed, we know from Lemma 3.2.2 that, under the condition (3.2.2),

$$\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}((n,s,a),\cdot) = \sum_{k=0}^{n-1}\frac{\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\delta_{k,\mathrm{s}_{k+1}},\mathrm{a}_{k+1}}{\sum_{l=0}^{n-1}\mathscr{K}_h\big(s-\mathrm{s}_l,a-\mathrm{a}_l\big)}, \tag{3.4.1}$$

which yields

$$\begin{aligned}F_j(\theta) &= \Big\{\sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\leqslant 1}\Big|\sum_{k=0}^{n-1}\frac{\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\big[\varphi\big(\mathrm{s}_{k+1}\big)\big]}{\sum_{k=1}^{n}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)}-\big[\overline{\mathsf{P}}\varphi\big](s,a)\Big|\\ &\qquad\leqslant C\Big[\Big(\sum_{k=0}^{n-1}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\Big)^{-(1-\theta)/2}+ch\Big]\Big\}.\end{aligned}$$

When condition (3.2.2) is not satisfied, the right-hand side in the definition of $F_j(\theta)$ is infinite and the inequality is trivially satisfied.

Proof. The spirit of the proof is to show a form of averaging, which we do here by using martingale techniques. This is completely different from the approach used in [27].

First Step. For a given bounded and measurable test function $\varphi : \overline{S}\to\mathbb{R}$, we define the following process:

$$M_n[\varphi] = \sum_{l=0}^{n-1}\mathscr{K}_h\big(s-\mathrm{s}_l,a-\mathrm{a}_l\big)\Big[\varphi\big(\mathrm{s}_{l+1}\big)-\mathbb{E}\Big(\varphi\big(\mathrm{s}_{l+1}\big)\,|\,\mathcal{F}_l\Big)\Big],$$

where the pair (s,a) is fixed throughout the proof.

Obviously, the process $(M_n[\varphi])_{n\geqslant 0}$ is a martingale with respect to the filtration $(\mathcal{F}_n)_{n\geqslant 0}$. We observe that

$$\mathbb{E}\Big[\varphi\big(\mathrm{s}_{l+1}\big)\,|\,\mathcal{F}_l\Big] = \big[\overline{\mathsf{P}}\varphi\big](\mathrm{s}_l,\mathrm{a}_l),$$

where $\overline{\mathsf{P}}\varphi$ is a shorten notation for the semi group induced by the transition kernel $\overline{\mathsf{P}}$, namely

$$\big[\overline{\mathsf{P}}\varphi\big](s',a') = \int_{\overline{S}} \varphi\big(\sigma\big)\overline{\mathsf{P}}\big((s',a'),d\sigma\big) = \mathbb{E}\Big[\varphi(\mathrm{s}_1)\,|\,(\mathrm{s}_0,\mathrm{a}_0)=(s',a')\Big].$$

Assume now on that $\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})} \leqslant 1$. Then, by Sobolev embedding theorem, φ has a bounded derivative and, by assumption **(Regularity Cost and Transition Kernel)**, the function $(s,a)\mapsto [\overline{\mathsf{P}}\varphi](s,a)$ is also Lipschitz continuous (with a know Lipschitz constant).

By (3.1.12), we deduce that, under the condition (3.2.2),

$$\begin{aligned}
&\sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\leqslant 1}\left|\sum_{l=0}^{n-1}\frac{\mathscr{K}_h\big(s-\mathrm{s}_l,a-\mathrm{a}_l\big)\mathbb{E}\Big(\varphi\big(\mathrm{s}_{l+1}\big)\,|\,\mathcal{F}_l\Big)}{\sum_{l=0}^{n-1}\mathscr{K}_h\big(s-\mathrm{s}_l,a-\mathrm{a}_l\big)} - \big[\overline{\mathsf{P}}\varphi\big](s,a)\right| \\
&= \sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\leqslant 1}\left|\sum_{l=0}^{n-1}\frac{\mathscr{K}_h\big(s-\mathrm{s}_l,a-\mathrm{a}_l\big)\big[\overline{\mathsf{P}}\varphi\big](\mathrm{s}_l,\mathrm{a}_l)}{\sum_{l=0}^{n-1}\mathscr{K}_h\big(s-\mathrm{s}_l,a-\mathrm{a}_l\big)} - \big[\overline{\mathsf{P}}\varphi\big](s,a)\right| \leqslant Ch.
\end{aligned} \tag{3.4.2}$$

By combining (3.4.2) with (3.4.1), we understand that it now remains to study

$$\sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\leqslant 1} |M_n[\varphi]|.$$

Obviously, this is the core of the proof.

Second Step. Our first step in the analysis of the above quantity is to study $(M_n[\varphi])_{n\geqslant 1}$ for a given choice of φ. Part of the difficulty will be to collect all the φ's in a single estimate.

In this step, φ is a mere bounded and measurable function. It is not required to be smooth. This point is important for the rest of the proof. We then introduce the following new sequence of stopping times:

$$\begin{aligned}
&S_0 := 0, \\
&S_\ell := \inf\left\{k \geqslant S_{\ell-1}+1 : \sum_{j=S_{\ell-1}}^{k}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_j,a-\mathrm{a}_j\big)\geqslant 1\right\}, \quad \ell\geqslant 1.
\end{aligned} \tag{3.4.3}$$

By Proposition 3.2.4, it is quite easy to see that, almost surely, $S_\ell<\infty$ for any $\ell\in\mathbb{N}$. We come back to this point in the fourth step below. At this stage, it suffices for the reader to know that there is no issue with the definition of the sequence $(S_\ell)_{\ell\geqslant 0}$.

Using these notations, we have

$$M_{S_\ell-1}[\varphi] = \sum_{l=1}^{\ell}\sum_{j=S_{l-1}}^{S_l-1}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_j,a-\mathrm{a}_j\big)\Big[\varphi\big(\mathrm{s}_{j+1}\big)-\mathbb{E}\Big(\varphi\big(\mathrm{s}_{j+1}\big)\,|\,\mathcal{F}_j\Big)\Big].$$

We claim that $(M_{S_\ell-1}[\varphi])_{\ell\geqslant 0}$ (with the convention that $M_{-1}[\varphi]=0$) is a martingale with respect to the filtration $(\mathcal{F}_{S_\ell})_{\ell\geqslant 0}$. Next, we estimate the martingale by means of Burkolder-Davis-Gundy's

inequality. This requires to address first the form of the quadratic variation. The key point in this regard is to use the fact that

$$\sum_{j=S_{l-1}}^{S_l-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big) \leqslant 1, \tag{3.4.4}$$

which allows us to get a simple bound for the jumps $(M_{S_{\ell+1}-1}[\varphi] - M_{S_\ell-1}[\varphi])_{\ell\geqslant 0}$:

$$\begin{aligned}
&\Big(\sum_{j=S_{l-1}}^{S_l-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)\Big[\varphi\big(\mathrm{s}_{j+1}\big) - \mathbb{E}\Big(\varphi\big(\mathrm{s}_{j+1}\big)\,|\,\mathcal{F}_j\Big)\Big]\Big)^2 \\
&\leqslant \sum_{j=S_{l-1}}^{S_l-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)\Big[\varphi\big(\mathrm{s}_{j+1}\big) - \mathbb{E}\Big(\varphi\big(\mathrm{s}_{j+1}\big)\,|\,\mathcal{F}_j\Big)\Big]^2 \\
&\leqslant 2\|\varphi\|_\infty \sum_{j=S_{l-1}}^{S_l-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)\Big[|\varphi|\big(\mathrm{s}_{j+1}\big) + \mathbb{E}\Big(|\varphi|\big(\mathrm{s}_{j+1}\big)\,|\,\mathcal{F}_j\Big)\Big].
\end{aligned}$$

We are now given an exponent $q > 1$. By discrete-time Burkolder-Davis-Gundy's inequality ([22]), we can find a universal constant C_q depending on q such that

$$\begin{aligned}
&\mathbb{E}\Big[\max_{1\leqslant l\leqslant \ell}\Big|M_{S_l-1}[\varphi]\Big|^{2q}\Big] \\
&\leqslant C_q\|\varphi\|_\infty^q \mathbb{E}\Big[\Big|\sum_{l=1}^{\ell}\sum_{j=S_{l-1}}^{S_l-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)\Big[|\varphi|\big(\mathrm{s}_{j+1}\big) + \mathbb{E}\Big(|\varphi|\big(\mathrm{s}_{j+1}\big)\,|\,\mathcal{F}_j\Big)\Big]\Big|^q\Big] \\
&\leqslant C_q\|\varphi\|_\infty^q \ell^q \mathbb{E}\Big[\Big(\ell^{-1}\sum_{j=0}^{S_\ell-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)\Big[|\varphi|\big(\mathrm{s}_{j+1}\big) + \mathbb{E}\Big(|\varphi|\big(\mathrm{s}_{j+1}\big)\,|\,\mathcal{F}_j\Big)\Big]\Big)^q\Big].
\end{aligned}$$

Using the fact that

$$\sum_{j=0}^{S_\ell-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big) = \sum_{l=1}^{\ell}\sum_{j=S_{l-1}}^{S_l-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big) \leqslant \ell, \tag{3.4.5}$$

we obtain (for a new value of C_q)

$$\begin{aligned}
&\mathbb{E}\Big[\max_{1\leqslant l\leqslant \ell}\Big|M_{S_l-1}[\varphi]\Big|^{2q}\Big] \\
&\leqslant C_q\|\varphi\|_\infty^{2q-1} \ell^q \mathbb{E}\Big[\ell^{-1}\sum_{j=1}^{S_\ell-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)\Big[|\varphi|\big(\mathrm{s}_{j+1}\big) + \mathbb{E}\Big(|\varphi|\big(\mathrm{s}_{j+1}\big)\,|\,\mathcal{F}_j\Big)\Big]\Big].
\end{aligned}$$

And then, for a new choice of C_q,

$$\mathbb{E}\Big[\max_{1\leqslant l\leqslant \ell}\Big|M_{S_l-1}[\varphi]\Big|^{2q}\Big] \leqslant C_q\|\varphi\|_\infty^{2q-1} \ell^q \mathbb{E}\Big[\ell^{-1}\sum_{j=1}^{S_\ell-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)|\varphi|\big(\mathrm{s}_{j+1}\big)\Big]. \tag{3.4.6}$$

Third Step. In order to clarify, we introduce, for $\ell \geqslant 2$, the following two (random) probability measures:

$$\mu_\ell(E) = \Big(\sum_{j=0}^{S_\ell-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)\Big)^{-1} \sum_{j=0}^{S_\ell-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)\mathbf{1}_E\big(\mathrm{s}_{j+1}\big),$$
$$\nu_\ell(E) = \Big(\sum_{j=0}^{S_\ell-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)\Big)^{-1} \sum_{j=0}^{S_\ell-1} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)\mathbb{E}\Big[\mathbf{1}_E\big(\mathrm{s}_{j+1}\big) \,|\, \mathcal{F}_j\Big],$$

where E is a generic Borel subset of $\overline{S}$.

Back to (3.4.1), we notice that

$$\mu_\ell = \overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}\big((S_\ell-1, s, a), \cdot\big).$$

Back to (3.4.2), we notice the first sum therein (at time $n = S_\ell$) is equal to

$$\sum_{l=0}^{S_\ell-1} \frac{\mathscr{K}_h\big(s-\mathrm{s}_l, a-\mathrm{a}_l\big)\mathbb{E}\Big(\varphi\big(\mathrm{s}_{l+1}\big) \,|\, \mathcal{F}_l\Big)}{\sum_{l=0}^{S_\ell-1} \mathscr{K}_h\big(s-\mathrm{s}_l, a-\mathrm{a}_l\big)} = \int_{\overline{S}} \varphi d\nu_\ell.$$

Back to the definition of the stopping time S_l in (3.4.3) of the second step, we notice that

$$\begin{aligned}
&\sum_{j=0}^{S_\ell-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big) \\
&\geqslant \sum_{l=0}^{\ell} \sum_{j=S_{l-1}}^{S_l} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big) - \sum_{l=0}^{\ell} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_{S_l}, a-\mathrm{a}_{S_l}\big) \\
&\geqslant \ell - \frac{1}{2}(\ell+1) = \frac{1}{2}(\ell-1).
\end{aligned} \tag{3.4.7}$$

Therefore, for a test function φ as in the second step,

$$\Big|\int_{\overline{S}} \varphi d\mu_\ell - \int_{\overline{S}} \varphi d\nu_\ell\Big| \leqslant 2(\ell-1)^{-1}\Big|M_{S_\ell-1}[\varphi]\Big|.$$

For $\ell \geqslant 2$, $\ell - 1 \geqslant \ell/2$ and then, by (3.4.6) and then (3.4.7),

$$\begin{aligned}
\mathbb{E}\Big[\Big|\int_{\overline{S}} \varphi d\mu_\ell - \int_{\overline{S}} \varphi d\nu_\ell\Big|^{2q}\Big] &\leqslant C_q \ell^{-2q} \mathbb{E}\Big[\max_{1\leqslant l\leqslant \ell}\Big|M_{S_l-1}[\varphi]\Big|^{2q}\Big] \\
&\leqslant C_q \|\varphi\|_\infty^{2q-1} \ell^{-q} \mathbb{E}\Big[\ell^{-1} \sum_{j=1}^{S_\ell-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_j, a-\mathrm{a}_j\big)|\varphi|\big(\mathrm{s}_{j+1}\big)\Big] \\
&\leqslant C_q \ell^{-q} \|\varphi\|_\infty^{2q-1} \mathbb{E}\Big[\int_{\overline{S}} |\varphi| d\mu_\ell\Big].
\end{aligned} \tag{3.4.8}$$

Next, we take φ in the Sobolev space $H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})$ with $\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})} \leqslant 1$ and we call $(e_i)_{i\geqslant 0}$ an orthonormal basis of $H^{3(\lfloor d_S/2\rfloor+1)}(\overline{S})$. We take for granted the following two properties:

$$\begin{aligned}&\sum_{i\geqslant 0}\big|\big(\varphi,e_i\big)_{H^{3(\lfloor d_S/2\rfloor+1)}(\overline{S})}\big|\leqslant C,\\&\sum_{i\geqslant 0}\big\|e_i\big\|_{L^\infty(\overline{S})}\leqslant C,\end{aligned}\tag{3.4.9}$$

for a constant C depending on the geometry of S. The demonstration of the above two bounds is left for the appendix 3.6.3. Then,

$$\left|\int_{\overline{S}}\varphi d\mu_\ell-\int_{\overline{S}}\varphi d\nu_\ell\right|\leqslant\sum_{i\geqslant 0}\left(\big|\big(\varphi,e_i\big)_{H^{3(\lfloor d_S/2\rfloor+1)}(\overline{S})}\big|\left|\int_{\overline{S}}e_i d\mu_\ell-\int_{\overline{S}}e_i d\nu_\ell\right|\right),$$

and then, by Hölder inequality and for φ satisfying $\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\leqslant 1$,

$$\begin{aligned}&\left|\int_{\overline{S}}\varphi d\mu_\ell-\int_{\overline{S}}\varphi d\nu_\ell\right|^{2q}\\&\leqslant\left(\sum_{i\geqslant 0}\big|\big(\varphi,e_i\big)_{H^{3(\lfloor d_S/2\rfloor+1)}(\overline{S})}\big|^{2p/(p+1)}\right)^{q(p+1)/p}\sum_{i\geqslant 0}\left|\int_{\overline{S}}e_i d\mu_\ell-\int_{\overline{S}}e_i d\nu_\ell\right|^{2q}\\&\leqslant\left(\sum_{i\geqslant 0}\big|\big(\varphi,e_i\big)_{H^{3(\lfloor d/2\rfloor+1)}(\overline{S})}\big|\right)^{q(p+1)/p}\sum_{i\geqslant 0}\left|\int_{\overline{S}}e_i d\mu_\ell-\int_{\overline{S}}e_i d\nu_\ell\right|^{2q}\end{aligned}$$

with $1/p+1/q=1$ (and so $(p+1)/(2p)+1/(2q)=1$). Above, we used the bound $\|\varphi\|_{H^{3(\lfloor d_S/2\rfloor+1)}(\overline{S})}\leqslant\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\leqslant 1$. Taking the supremum over φ in the unit ball of $\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})}$ and then invoking (3.4.9) first and (3.4.8) next, we deduce that

$$\begin{aligned}\mathbb{E}\left[\sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\leqslant 1}\left|\int_{\overline{S}}\varphi d\mu_\ell-\int_{\overline{S}}\varphi d\nu_\ell\right|^{2q}\right]&\leqslant C^{q(p+1)/p}\sum_{i\geqslant 0}\mathbb{E}\left[\left(\left|\int_{\overline{S}}e_i d\mu_\ell-\int_{\overline{S}}e_i d\nu_\ell\right|^2\right)^q\right]\\&\leqslant C_q\ell^{-q}\sum_{i\geqslant 0}\mathbb{E}\left[\int_{\overline{S}}|e_i|d\mu_\ell\right]\\&\leqslant C_q\ell^{-q},\end{aligned}$$

for a constant C_q depending on q.

Fourth Step. By Markov inequality, the conclusion of the third step yields, for $\ell\in\mathbb{N}$, for any $\theta\in(0,1/2)$,

$$\mathbb{P}\left[\left\{\sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\leqslant 1}\left|\int_{\overline{S}}\varphi d\mu_\ell-\int_{\overline{S}}\varphi d\nu_\ell\right|\geqslant\ell^{-(1-\theta)/2}\right\}\right]\leqslant C_q\ell^{-q\theta/2}.\tag{3.4.10}$$

And then, for $q\theta>2$ and another integer $L\in\mathbb{N}$,

$$\mathbb{P}\left[\bigcup_{\ell\geqslant L}\left\{\sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\leqslant 1}\left|\int_{\overline{S}}\varphi d\mu_\ell-\int_{\overline{S}}\varphi d\nu_\ell\right|\geqslant\ell^{-(1-\theta)/2}\right\}\right]\leqslant C_{q,\theta}L^{-q\theta/2+1},$$

the constant $C_{q,\theta}$ being also allowed to depend on θ. This prompts us to introduce the collection of events

$$E_\ell(\theta) = \Big\{ \sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})\leqslant 1}} \Big| \int_{\overline{S}} \varphi d\mu_\ell - \int_{\overline{S}} \varphi d\nu_\ell \Big| < \ell^{-(1-\theta)/2} \Big\}, \quad \ell \in \mathbb{N}.$$

On the event $\bigcap_{k\geqslant L} E_k(\theta)$ and for $\ell > L \geqslant 2$, we have, for any $n \in \{S_{\ell-1}, \cdots, S_\ell - 1\}$ and for any bounded (measurable) test function $\varphi : \overline{S} \to \mathbb{R}$,

$$\begin{aligned}
&\Big| \sum_{k=0}^{n-1} \frac{\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)\varphi\big(\mathrm{s}_{k+1}\big)}{\sum_{k=1}^{n}\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)} - \int_{\overline{S}} \varphi d\mu_\ell \Big| \\
&= \Big| \sum_{k=0}^{n-1} \frac{\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)\varphi\big(\mathrm{s}_{k+1}\big)}{\sum_{k=0}^{n-1}\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)} - \sum_{k=0}^{S_\ell-1} \frac{\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)\varphi\big(\mathrm{s}_{k+1}\big)}{\sum_{k=0}^{S_\ell-1}\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)} \Big| \\
&\leqslant 2\|\varphi\|_\infty \sum_{k=S_{\ell-1}}^{S_\ell-1} \frac{\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)}{\sum_{k=0}^{S_\ell-1}\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)} \\
&\leqslant 2\|\varphi\|_\infty (\ell-1)^{-1},
\end{aligned}$$

where we used (3.4.4) and (3.4.7) in the last line. Proceeding similarly with $\mathbb{E}[\varphi(\mathrm{s}_{k+1})|\mathcal{F}_k]$ instead of $\varphi(\mathrm{s}_{k+1})$, we obtain, for any $n \in \{S_{\ell-1}, \cdots, S_\ell - 1\}$,

$$\Big| \sum_{k=0}^{n-1} \frac{\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)\big[\varphi\big(\mathrm{s}_{k+1}\big) - \mathbb{E}\big(\varphi\big(\mathrm{s}_{k+1}\big)|\mathcal{F}_k\big)\big]}{\sum_{k=1}^{n}\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)} - \int_{\bar{S}} \varphi d\big(\mu_\ell - \nu_\ell\big) \Big| \leqslant 4\|\varphi\|_\infty (\ell-1)^{-1}.$$

And then, using (3.4.9), we get

$$\begin{aligned}
&\sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})\leqslant 1}} \Big| \sum_{k=0}^{n-1} \frac{\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)\big[\varphi\big(\mathrm{s}_{k+1}\big) - \mathbb{E}\big(\varphi\big(\mathrm{s}_{k+1}\big)|\mathcal{F}_k\big)\big]}{\sum_{k=0}^{n-1}\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)} \Big| \\
&\leqslant \sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})\leqslant 1}} \Big| \int_{\overline{S}} \varphi d\big(\mu_\ell - \nu_\ell\big) \Big| + C(\ell-1)^{-1},
\end{aligned}$$

for a possibly new constant C (the value of which is allowed to vary from line to line), independent of q.

And, since we are on $\bigcap_{\ell\geqslant L} E_\ell(\theta)$,

$$\sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})\leqslant 1}} \Big| \sum_{k=0}^{n-1} \frac{\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)\big[\varphi\big(\mathrm{s}_{k+1}\big) - \mathbb{E}\big(\varphi\big(\mathrm{s}_{k+1}\big)|\mathcal{F}_k\big)\big]}{\sum_{k=0}^{n-1}\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)} \Big| \leqslant \ell^{-(1-\theta)/2} + C\ell^{-1}.$$

And then, by (3.4.2), we get, for $n \in \{S_{\ell-1}, \cdots, S_\ell - 1\}$

$$\begin{aligned}
&\sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}(\overline{S})\leqslant 1}} \Big| \sum_{k=0}^{n-1} \frac{\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)\big[\varphi\big(\mathrm{s}_{k+1}\big)\big]}{\sum_{k=0}^{n-1}\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)} - \big[\overline{\mathsf{P}}\varphi\big](s,a) \Big| \\
&\quad \leqslant \ell^{-(1-\theta)/2} + C\ell^{-1} + Ch,
\end{aligned} \tag{3.4.11}$$

where we used, in the derivation of the very last term in the right-hand side, (3.4.9) in order to upper bound $\|\varphi\|_{1,\infty}$ by C when $\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}} \leqslant 1$. By (3.4.5), we know that, for $n \in \{S_{\ell-1}, \cdots, S_\ell - 1\}$,

$$\sum_{j=1}^{n} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s - \mathrm{s}_j, a - \mathrm{a}_j\big) \leqslant \ell,$$

which yields, by combination with (3.4.11),

$$\begin{aligned} &\sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}}(\overline{S})\leqslant 1} \left| \sum_{k=1}^{n} \frac{\mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)\big[\varphi\big(\mathrm{s}_{k+1}\big)\big]}{\sum_{k=1}^{n} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)} - \big[\overline{\mathsf{P}}\varphi\big](s,a) \right| \\ &\quad \leqslant \left(\sum_{k=0}^{n-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) \right)^{-(1-\theta)/2} \\ &\quad\quad + C\left(\sum_{k=0}^{n-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) \right)^{-1} + Ch. \end{aligned} \tag{3.4.12}$$

We now observe that, if

$$\sum_{k=0}^{n-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) > L,$$

then, by (3.4.5), $n > S_L - 1$, since

$$\sum_{k=1}^{S_L-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) \leqslant L.$$

Therefore, if we are on the event

$$\bigcap_{\ell \geqslant L} E_\ell(\theta) \cap \left\{ \sum_{k=1}^{n} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) > L \right\},$$

then (3.4.12) holds true. By using the asumption $L \geqslant 1$, we even have the simpler bound

$$\begin{aligned} &\sup_{\|\varphi\|_{H^{5(\lfloor d_S/2\rfloor+1)}}(\overline{S})\leqslant 1} \left| \sum_{k=1}^{n} \frac{\mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)\big[\varphi\big(\mathrm{s}_{k+1}\big)\big]}{\sum_{k=1}^{n} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)} - \big[\overline{\mathsf{P}}\varphi\big](s,a) \right| \\ &\quad \leqslant C\left(\sum_{k=1}^{n} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) \right)^{-(1-\theta)/2} + Ch, \end{aligned} \tag{3.4.13}$$

for a possibly new value of the constant C. We thus recover the definition of $F_n(\theta)$ in the statement by choosing $q = 2/\theta^2$ in (3.4.10). Our analysis says that

$$\bigcap_{\ell \geqslant L} E_\ell(\theta) \cap \left\{ \sum_{k=1}^{n} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) > L \right\} \subset \bigcap_{j \geqslant n} F_j(\theta).$$

Alternatively,

$$\bigcup_{j\geqslant n}\Big(F_j(\theta)^{\complement}\Big)\subset\bigcup_{\ell\geqslant L}\Big(E_\ell(\theta)^{\complement}\Big)\cup\Big\{\sum_{k=1}^{n}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\leqslant L\Big\},$$

and, then,

$$\bigcup_{j\geqslant n}\Big(F_j(\theta)^{\complement}\Big)\cap\Big\{\sum_{k=1}^{n}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)> L\Big\}\subset\bigcup_{\ell\geqslant L}\Big(E_\ell(\theta)^{\complement}\Big).$$

We then end-up the proof by using (3.4.10).

□

3.4.2 Taking the supremum over s and a

As the reader can notice in Lemma 3.4.1, the value of (s,a) is fixed. We here want to take the supremum over (s,a). This is done thanks to the following lemma, which allows us to reduce the analysis to points (s,a) in a finite lattice. The reader will easily deduce Proposition 3.2.7 from Lemmas 3.4.1 and 3.4.2.

Lemma 3.4.2. *For $\varepsilon>0$, consider a lattice $\mathscr{N}$ of $\overline{S}\times\overline{A}$ that is εh-dense in the sense that any $(s,a)\in\overline{S}\times\overline{A}$ in distance at most εh from $\mathscr{N}$. Then, with η' as in (**??**), for any $\beta\geqslant 1$ and $\theta>0$, there exists a constant C as in the statement of Proposition 3.2.7 (in particular C is independent of the lattice) such that*

$$\begin{aligned}\mathbb{P}\Bigg[\bigcap_{j\geqslant n}\Bigg\{&\sup_{(s,a)\in\overline{S}\times\overline{A}}\sup_{(s',a')\in\mathscr{N}:|(s',a')-(s,a)|\leqslant\varepsilon h}\mathrm{d}_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\Big(\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}((n,s,a),\cdot),\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}((n,s',a'),\cdot)\Big)\\&\leqslant C\beta\varepsilon\frac{\eta'}{\eta}\big(1+\ln(J)\big)\Bigg\}\Bigg]\\&\geqslant 1-\frac{1}{\beta\eta'\varepsilon^D h^{2D}}\exp\Big(-\beta\eta' h^D(n-1)\Big)-\frac{C}{\eta h^D}\exp\Big(-\frac{\eta h^D}{C}(n-1)\Big).\end{aligned}$$

Proof. First Step. The first step is to address the following probability:

$$\mathbb{P}\Bigg(\Bigg\{\inf_{(s,a)\in\overline{S}\times\overline{A}}\sum_{k=0}^{n-1}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)>L\Bigg\}\Bigg).$$

It is worth observing that the event inside the probability symbol appears in the statement of Lemma 3.4.1, but without the infimum. Back to the proof of Proposition 3.2.4 and using the same notation as therein (with $\ell=1$), we observe that

$$\begin{aligned}&\Bigg\{\inf_{(s,a)\in\overline{S}\times\overline{A}}\sum_{k=0}^{n-1}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)>L\Bigg\}\\&\supset\Bigg\{\min_{1\leqslant j\leqslant J}\sum_{k=0}^{n-1}\frac{\lambda_{\mathscr{K}}}{2\|\mathscr{K}\|_\infty}\mathbf{1}_{B_j}\big(\mathrm{s}_k,\mathrm{a}_k\big)\geqslant L+1\Bigg\}\supset\big\{\,T_{\lceil 2\|\mathscr{K}\|_\infty(L+1)/\lambda_{\mathscr{K}}\rceil}\leqslant n\big\},\end{aligned}$$

with $T_{\lceil 2\|\mathscr{K}\|_\infty(L+1)/\lambda_{\mathscr{K}}\rceil}$ being defined in (3.3.2). Now, by (3.3.14) in the proof of Proposition 3.2.4 (using the same notation as therein),

$$T_{\lceil 2\|\mathscr{K}\|_\infty(L+1)/\lambda_{\mathscr{K}}\rceil} \leqslant 1 + \kappa + \Big\lceil \frac{2\|\mathscr{K}\|_\infty(L+1)}{\lambda_{\mathscr{K}}} \Big\rceil \widetilde{E}_{\eta,J,h}, \tag{3.4.14}$$

on the event $A_{\lceil 2\|\mathscr{K}\|_\infty(L+1)/\lambda_{\mathscr{K}}\rceil}(\kappa)$ defined in (3.3.11). What really matters is the bound we have for $\mathbb{P}(A_{\lceil 2\|\mathscr{K}\|_\infty(L+1)/\lambda_{\mathscr{K}}\rceil}(\kappa)^{\complement})$, see (3.3.19). We deduce that

$$\mathbb{P}\big(A_{\lceil 2\|\mathscr{K}\|_\infty(L+1)/\lambda_{\mathscr{K}}\rceil}(\kappa)^{\complement}\big) \leqslant \exp\Big(-\min\Big[\frac{\kappa\eta h^D}{8}, \frac{\kappa^2\eta^2h^{2D}}{4C\lceil 2\|\mathscr{K}\|_\infty(L+1)/\lambda_{\mathscr{K}}\rceil}\Big]\Big).$$

We now choose $\kappa = \lceil (2\|\mathscr{K}\|_\infty(L+1)/\lambda_{\mathscr{K}})\rceil \widetilde{E}_{\eta,J,h}$, recalling from (3.3.13) that

$$\widetilde{E}_{\eta,J,h} = \frac{\ln(J)+2}{\eta h^D}.$$

We obtain

$$\mathbb{P}\big(A_{\lceil 2\|\mathscr{K}\|_\infty(L+1)/\lambda_{\mathscr{K}}\rceil}(\kappa)\big) \geqslant 1 - \exp\Big(-\frac{\lceil 2\|\mathscr{K}\|_\infty(L+1)/\lambda_{\mathscr{K}}\rceil(\ln(J)+2)}{C}\Big),$$

for a new value of the constant C. So, if we can choose L such that

$$L+1 = \Big\lfloor \frac{\lambda_{\mathscr{K}}}{2\|\mathscr{K}\|_\infty} \Big\lfloor \frac{n-1}{2\widetilde{E}_{\eta,J,h}} \Big\rfloor \Big\rfloor, \tag{3.4.15}$$

then

$$2\Big\lceil \frac{2\|\mathscr{K}\|_\infty(L+1)}{\lambda_{\mathscr{K}}} \Big\rceil \widetilde{E}_{\eta,J,h} \leqslant n-1.$$

Since $L+1 \geqslant 1$, then $L+2 \leqslant 2(L+1)$ and

$$L+1 \geqslant \frac{\lambda_{\mathscr{K}}}{4\|\mathscr{K}\|_\infty} \Big\lfloor \frac{n-1}{2\widetilde{E}_{\eta,J,h}} \Big\rfloor.$$

Also, we necessarily have $n-1 \geqslant 2E_{\eta,J,h}$ and thus

$$L+1 \geqslant \frac{\lambda_{\mathscr{K}}}{8\|\mathscr{K}\|_\infty} \frac{n-1}{\widetilde{E}_{\eta,J,h}}.$$

Therefore, by (3.4.14), and for $n \geqslant 2$,

$$\mathbb{P}\Big(\Big\{ \inf_{(s,a)\in\overline{S}\times\overline{A}} \sum_{k=0}^{n-1} \frac{1}{2\|\mathscr{K}\|_\infty} \mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big) > L \Big\}\Big) \geqslant 1 - \exp\Big(-\frac{\eta n h^D}{C}\Big), \tag{3.4.16}$$

for a new value of the constant C. Inserting into the statement of Lemma 3.4.1, we obtain, for $L \geqslant 1$ and $n \geqslant 2$,

$$\mathbb{P}\Big(\bigcup_{j\geqslant n} \big(F_j(\theta)\big)^{\complement}\Big) \leqslant C_\theta \Big(\frac{E_{\eta,J,h}}{n}\Big)^{1/\theta-1} + \exp\Big(-\frac{\eta n h^D}{C}\Big),$$

for any $(s,a) \in \overline{S} \times \overline{A}$, where C_θ. Above, we replaced $\widetilde{E}_{\eta,J,h}$ by $CE_{\eta,J,h}$, which is licit if $J \geqslant 3$. In fact, if $L = 1$ or if there is no way to define L according to (3.4.15), then, necessarily, $n - 1 \leqslant C' \widetilde{E}_{\eta,J,h}$, for a new constant C'. If $n \geqslant 2$, this implies $n \leqslant 2C' \widetilde{E}_{\eta,J,h}$ and we can increase the value of the constant C right above so that the right-hand side becomes (in that regime of parameters) greater than 1, in which case the inequality is also satisfied. In other words, we can take the above inequality for granted for any value of L,

And then,

$$\mathbb{P}\Big(\bigcup_{j \geqslant n} \big(\widetilde{F}_j(\theta)\big)^{\complement} \Big) \leqslant C_\theta \Big(\frac{E_{\eta,J,h}}{n} \Big)^{1/\theta - 1} + \exp\Big(- \frac{\eta n h^D}{C} \Big),$$

where

$$\widetilde{F}_j(\theta) := \Big\{ \sup_{\|\varphi\|_{H^{5(\lfloor d_S/2 \rfloor + 1)}(\overline{S})} \leqslant 1} \Big| \sum_{k=1}^n \frac{\mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)\big[\varphi\big(\mathrm{s}_{k+1}\big)\big]}{\sum_{k=1}^n \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big)} - \big[\overline{\mathsf{P}}\varphi\big](s,a) \Big| \leqslant C \Big(\frac{E_{\eta,J,h}}{n} \Big)^{(1-\theta)/2} + Ch \Big\},$$

The reader may compare the above definition with the definition of $F_j(\theta)$ in the statement. In short, we have inserted $E_{\eta,J,h}/n$ in the right-hand side. Here as well, the proof directly follows from (3.4.16) if $L \geqslant 2$. If $L = 1$ or if L cannot be defined as in (3.4.15), then we call easily increase the value of C_S by recalling that the elements of $H^{5(\lfloor d/2 \rfloor + 1)}(\overline{S})$ are necessarily bounded by a constant only depending on the geometry of S.

Second Step. The next step is to address

$$\mathrm{d}_{H^{-5(\lfloor d_S/2 \rfloor + 1)}(\overline{S})} \Big(\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}((n,s,a),\cdot), \overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}((n,s',a'),\cdot) \Big)$$

for two points (s,a), (s',a') in $\overline{S} \times \overline{A}$.

We consider $\mathscr{N}$ an εh-net covering $\overline{S} \times \overline{A}$: for any point of $\overline{S} \times \overline{A}$, we can find a point of the net at distance less than εh. Then, for $(s,a) \in \overline{S} \times \overline{A}$, we find $(s',a') \in \mathscr{N}$ such that

$$\big|(s',a') - (s,a)\big| \leqslant \varepsilon h.$$

Then, for any $k \in \{1, \cdots, n\}$, we get

$$\Big| \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) - \mathscr{K}_h\big(s' - \mathrm{s}_k, a' - \mathrm{a}_k\big) \Big| \leqslant \|\mathscr{K}\|_{1,\infty} \varepsilon.$$

Notice that this is $\|\mathscr{K}\|_{1,\infty}$ in the above right-hand side and not $\|\mathscr{K}_h\|_{1,\infty}$. The estimate follows from the fact that $\|\mathscr{K}_h\|_{1,\infty} = \|\mathscr{K}\|_{1,\infty}/h$. If $|(s' - \mathrm{s}_k, a' - \mathrm{a}_k\big)| \geqslant \varepsilon h + \varrho h$, then the left-hand side is zero since the support of $\mathscr{K}_h$ is included in the ball $B(0, \varrho h)$. Therefore, we can write

$$\Big| \mathscr{K}_h\big(s - \mathrm{s}_k, a - \mathrm{a}_k\big) - \mathscr{K}_h\big(s' - \mathrm{s}_k, a' - \mathrm{a}_k\big) \Big| \leqslant \|\mathscr{K}\|_{1,\infty} \mathbf{1}_{\{|(s' - \mathrm{s}_k, a' - \mathrm{a}_k\big)| \leqslant \varepsilon h + \varrho h\}} \varepsilon.$$

And then for any bounded test function $\varphi : \overline{S} \to \mathbb{R}$,

$$\begin{aligned}&\Big|\sum_{k=0}^{n-1} \frac{\mathscr{K}_h\big(s-\mathrm{s}_k, a-\mathrm{a}_k\big)\varphi\big(\mathrm{s}_k\big)}{\sum_{k=0}^{n-1}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)} - \sum_{k=0}^{n-1}\frac{\mathscr{K}_h\big(s'-\mathrm{s}_k,a'-\mathrm{a}_k\big)\varphi\big(\mathrm{s}_k\big)}{\sum_{k=0}^{n-1}\mathscr{K}_h\big(s'-\mathrm{s}_k,a'-\mathrm{a}_k\big)}\Big| \\ &\leqslant 2\varepsilon\|\varphi\|_\infty\|\mathscr{K}\|_{1,\infty}\frac{\sum_{k=0}^{n-1}\mathbf{1}_{\{|(s'-\mathrm{s}_k,a'-\mathrm{a}_k)|\leqslant\varepsilon h+\varrho h\}}}{\sum_{k=0}^{n-1}\mathscr{K}_h\big(s'-\mathrm{s}_k,a'-\mathrm{a}_k\big)},\end{aligned} \tag{3.4.17}$$

from which we deduce that

$$\begin{aligned}&\sup_{(s,a)\in\overline{S}\times\overline{A}}\sup_{(s',a')\in\mathscr{N}:|(s',a')-(s,a)|\leqslant\varepsilon h} \mathrm{d}_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\Big(\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}((n,s,a),\cdot),\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}((n,s',a'),\cdot)\Big) \\ &\leqslant C\varepsilon\sup_{(s',a')\in\mathscr{N}}\frac{\sum_{k=0}^{n-1}\mathbf{1}_{\{|(s'-\mathrm{s}_k,a'-\mathrm{a}_k)|\leqslant\varepsilon h+\varrho h\}}}{\sum_{k=0}^{n-1}(2\|\mathscr{K}\|_\infty)^{-1}\mathscr{K}_h\big(s'-\mathrm{s}_k,a'-\mathrm{a}_k\big)},\end{aligned}$$

for a constant independent of n (but depending on the details of $\mathscr{K}$).

From the first step (see (3.4.16)), we already know that, if we can define L as in (3.4.15), then

$$\mathbb{P}\bigg(\Big\{\inf_{(s',a')\in\overline{S}\times\overline{A}}\sum_{k=0}^{n-1}\frac{1}{2\|\mathscr{K}\|_\infty}\mathscr{K}_h\big(s-\mathrm{s}_k,a-\mathrm{a}_k\big)\geqslant\frac{\lambda_{\mathscr{K}}}{32\|\mathscr{K}\|_\infty}\frac{n}{E_{\eta,J,h}}\Big\}\bigg)\geqslant 1-\exp\Big(-\frac{\eta n h^D}{C}\Big).$$

We now follow the derivation of (3.3.7) in the proof of Proposition 3.2.4 (which relies on Lemma 3.6.1, see in particular (3.6.1)). With η' as in (**??**), for any $\beta\geqslant e^2$,

$$\mathbb{P}\left(\Big\{\sum_{k=0}^{n-1}\mathbf{1}_{\{|(s-\mathrm{s}_k,a-\mathrm{a}_k)|\leqslant 3\varrho h\}}\geqslant 3^D\beta\eta' n\varrho^D h^D\Big\}\right)\leqslant\exp\big(-\beta\eta' n h^D\big).$$

Then,

$$\mathbb{P}\bigg(\Big\{\sup_{(s',a')\in\mathscr{N}}\sum_{k=0}^{n-1}\mathbf{1}_{\{|(s'-\mathrm{s}_k,a'-\mathrm{a}_k)|\leqslant 3\varrho h\}}\leqslant 3^D\beta\eta' n\varrho^D h^D\Big\}\bigg)\geqslant 1-Ch^{-D}\varepsilon^{-D}\exp\big(-\beta\eta' n h^D\big).$$

We deduce that, for $\varepsilon\leqslant 2\varrho$,

$$\begin{aligned}&\mathbb{P}\bigg[\Big\{\sup_{(s,a)\in\overline{S}\times\overline{A}}\sup_{(s',a')\in\mathscr{N}:|(s',a')-(s,a)|\leqslant\varepsilon h}\mathrm{d}_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\Big(\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}((n,s,a),\cdot),\overline{\Pi}_{(\underline{\mathrm{s}},\underline{\mathrm{a}})}((n,s',a'),\cdot)\Big) \\ &\qquad\leqslant C\beta\varepsilon\eta' h^D E_{\eta,J,h}\Big\}\bigg] \\ &\geqslant 1-Ch^{-D}\varepsilon^{-D}\exp\big(-\beta\eta' n h^D\big)-\exp\Big(-\frac{\eta n h^D}{C}\Big).\end{aligned}$$

We deduce that

$$\begin{aligned}
&\mathbb{P}\left[\bigcap_{j\geqslant n}\Big\{\sup_{(s,a)\in\overline{S}\times\overline{A}}\sup_{(s',a')\in\mathscr{N}:|(s',a')-(s,a)|\leqslant\varepsilon h}\mathrm{d}_{H^{-5(\lfloor d_S/2\rfloor+1)}(\overline{S})}\Big(\overline{\Pi}_{(\underline{s},\underline{a})}((n,s,a),\cdot),\overline{\Pi}_{(\underline{s},\underline{a})}((n,s',a'),\cdot)\Big)\right.\\
&\qquad\left.\leqslant C\beta\varepsilon\frac{\eta'}{\eta}\big(1+\ln(J)\big)\Big\}\right]\\
&\geqslant 1-Ch^{-D}\varepsilon^{-D}\sum_{j\geqslant n}\exp\big(-\beta\eta' h^D j\big)-\sum_{j\geqslant n}\exp\Big(-\frac{\eta h^D j}{C}\Big)\\
&\geqslant 1-\frac{1}{\beta\eta'\varepsilon^D h^{2D}}\exp\Big(-\beta\eta' h^D(n-1)\Big)-\frac{C}{\eta h^D}\exp\Big(-\frac{\eta h^D}{C}(n-1)\Big).
\end{aligned}$$

□

3.5 Numerical Results

As numerical example to illustrate our results, we present a linear model set on the 1-d torus $\overline{S}=\mathbb{T}^1\big([0,\pi]\big)$ for the states and actions $\overline{A}=[-\frac{\pi}{2},\frac{\pi}{2}]$. The main motivation behind this particular example comes from the fact that it has an analytical solution that we can compute explicitly and thus have strict comparisons for our algorithm. Because our main theoretical result concerns an infinite MDP, we made the choice here for a periodic example whose structure is the closest analog to an infinite MDP. Let us define the dynamics

$$x_{n+1}=(x_n+a_n+\sigma\varepsilon_{n+1})\mod\pi,\quad x_0=x,$$

where $\varepsilon_n\sim\mathcal{N}(0,1)$.

The task we give to our agent is to minimize the associated running cost

$$c(x,a)=\cos(2x)+1+\gamma C_\sigma+cos(2x+2a)$$

where $C_\sigma=e^{2\sigma^2}$ and γ the usual discount factor.

The value function is

$$V(x)=\mathbb{E}[\sum_{n\geqslant 0}\gamma^n c(x,a)\,\big|\,x_0=x]$$

and the Bellman operator

$$V(x)=\min_a\Big\{c(x,a)+\gamma\mathbb{E}\big[V(x')\big]\Big\}$$

Now to compute the optimal value function analytically, we suppose that $V(x)$ is even, $V(-x)=-V(x)$ and belongs to the Banach space of continuous periodic functions of the interval $[0,\pi]$. Let

$$V(x)=\frac{1}{2}\alpha_0+\sum_{n\neq 0}\alpha_n\cos(nx),\tag{3.5.1}$$

be the standard development of $V(x)$ in Fourier series with coefficients $\{\alpha_n\}_{n\geqslant 0}$ given by the known formulas. We plug (3.5.1) in the fixed point equation

$$\begin{aligned}\frac{1}{2}\alpha_0 + \sum_{n\neq 0}\alpha_n\cos(nx) &= \cos(2x) + 1 + \gamma C_\sigma + \min_a\Big\{\cos(2x+2a) + \gamma\mathbb{E}\big[\frac{1}{2}\alpha_0 + \sum_{n\neq 0}\alpha_n\cos(nx')\big]\Big\}\\ \frac{1}{2}\alpha_0 + \sum_{n\neq 0}\alpha_n\cos(nx) &= \cos(2x) + 1 + \gamma C_\sigma + \gamma\frac{1}{2}\alpha_0 + \min_a\Big\{\cos(2x+2a) + \gamma\mathbb{E}\big[\sum_{n\neq 0}\alpha_n\cos\big(n(x+a+\sigma\varepsilon)\big)\big]\Big\}.\end{aligned} \tag{3.5.2}$$

We work the term inside the expectation

$$\begin{aligned}&\mathbb{E}\big[\sum_{n\neq 0} a_n\cos\big(n(x+a+\sigma\varepsilon)\big)\big] = \sum_{n\neq 0}\alpha_n\mathbb{E}\big[\cos\big(n(x+a)\big)\cos(\sigma\varepsilon) - \sin(2x+2a)\sin(\sigma\varepsilon)\big]\\ &= \sum_{n\neq 0}\alpha_n\cos\big(n(x+a)\big)C_\sigma,\end{aligned}$$

where in the second equality we used the fact that $\mathbb{E}[\sin(\sigma\varepsilon)] \approx 2\sigma\int_{-\infty}^{\infty}\varepsilon\frac{1}{\sqrt{2\pi}}e^{-\frac{\varepsilon^2}{2}}d\varepsilon$ (given by a Taylor expansion of $\sin(\sigma\varepsilon)$) and in the final the definition of C_σ. Now turning to the computation of *min* we have

$$\min_a\Big\{\cos(2x+2a) + \sum_{n\neq 0}\alpha_n C_\sigma\cos\big(n(x+a)\big)\Big\},$$

which yields $a^* = \frac{\pi}{n} - x$.

Plugging everything in (3.5.2) we can identify the coefficients of the Fourier representation of the Value function as follows

$$\begin{aligned}\frac{1}{2}\alpha_0 &= \gamma C_\sigma + 1 - \cos(\frac{2\pi}{n}) + \gamma\frac{1}{2}\alpha_0 - \gamma C_\sigma\sum_{n\neq 0}\alpha_n\\ \sum_{n\neq 0}\alpha_n\cos(nx) &= \cos(2x),\end{aligned}$$

where from the second we get $n = 2$, $\alpha_2 = 1$, then $\alpha_0 = 0$. So finally the state value function

$$V(x) = \cos(2x),$$

and the action value function

$$Q(x,a) = c(x,a) + \mathbb{E}[V(x')] = \cos(2x) + 1 + (1+\gamma C_\sigma)\cos(2x+2a) + \gamma C_\sigma,$$

and

$$a^*(x) = \frac{\pi}{2} - x.$$

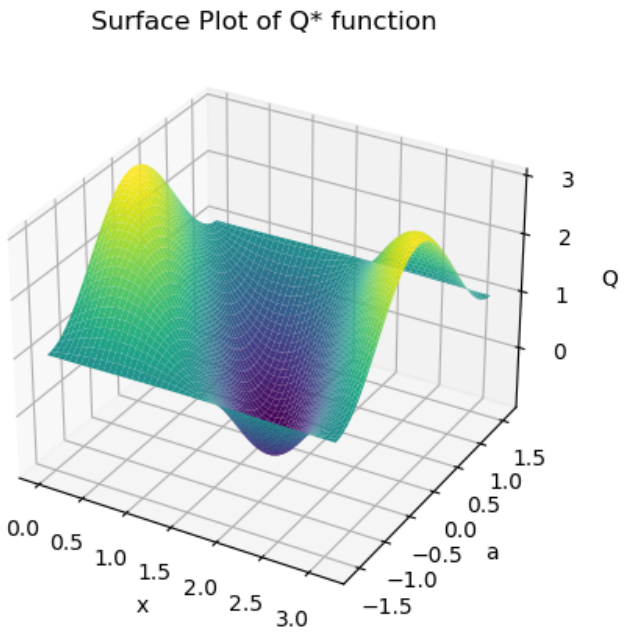


Figure 3.1: Theoretical Q function

For a visualisation of Q function see Figure 3.1

For our simulations to learn the optimal value function, we use a discretisation of the actions using $N_a = 100$ points for estimating min and $argmin$ using grid search and 10^6 iterations. At each iteration we choose uniformly a random action and observe the next state and the cost. Then we update our estimator of the value function according to Algorithm 2 and Equation 3.1.16 for the TD target. We emphasise that we just consider the data stored in memory as sequences $s_n, a_n, y_n, s_{n+1}, a_{n+1}, y_{n+1}$ where the total number of data points is the total number of iterations of our RL loop (i.e. 10^6). The choice of our kernel is Epanechnikov Kernel function

$$\mathscr{K}(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{if } x \in [-1,1] \\ 0 & \text{otherwise,} \end{cases}$$

and for the rest of our parameters, the size of the bandwidth $h = 0.2$, $\sigma = 1$ and $\gamma = 0.1$. First we compare the results for the convergence of the $\widehat{Q}_h$ to Q

The supnorm and L^2 errors respectively for the plot at 10^6 iterations over a 100×100 discretisation grid are as follows:

1. $\|\widehat{Q}_h - Q\|_\infty = 1.7678$
2. $\left(\int_0^1 \int_0^1 |\widehat{Q}_h(x,a) - Q(x,a)|^2 dxda\right)^{\frac{1}{2}} = 0.2986$

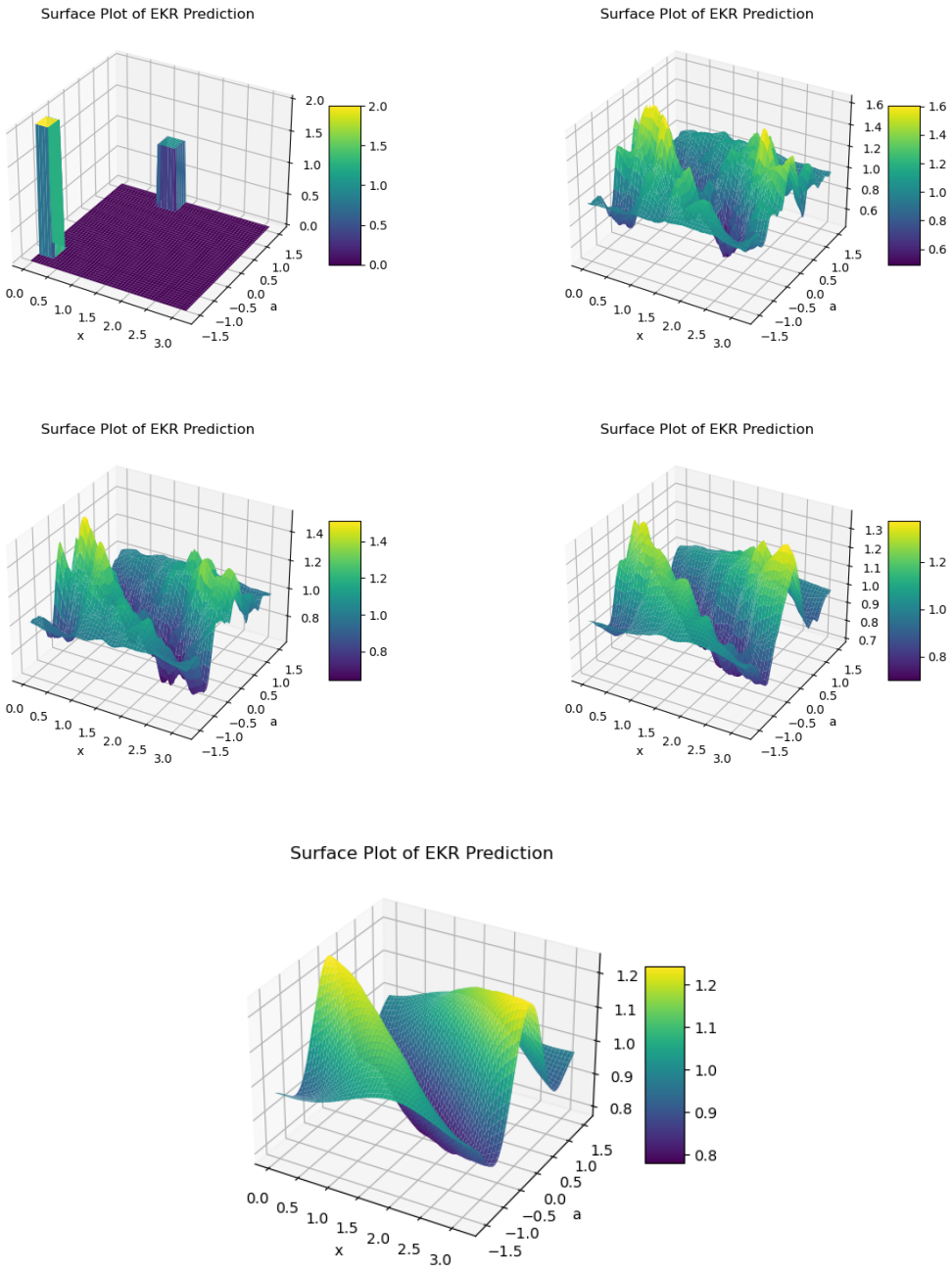


Figure 3.2: Approximated Q function at different iterations. From top left, clockwise: initial, 10^4, $2 * 10^4$, $9 * 10^4$ and 10^6

The main issues of the approximation that increase the supnorm error is the scale of $\widehat{Q}_h$ and

some areas around the boundary that is a common symptom of Kernel regression, see also the remarks that are following.

For the optimal control the issue with the boundary becomes even more evident

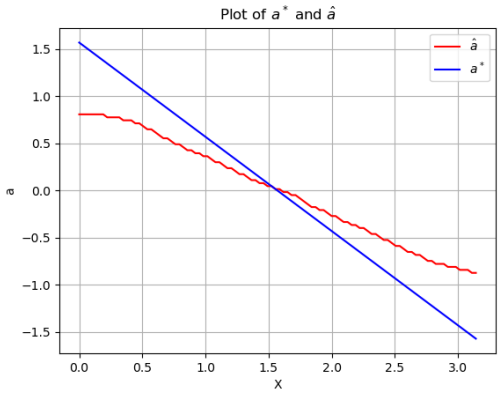


Figure 3.3: Approximated optimal policy

The L^2 error for the optimal policy is 0.037

Remark 3.5.1. *h should be proportional to σ^2 in a sense that if $h > \sigma^2$ the estimator will be flatter since the bandwidth of data for the kernel will be too wide, this corresponds to higher bias. On the other hand, for N_a to have a positive effect on the error as N_a goes up also h needs to go down but not too fast, since setting h too small makes the estimator's variance going up and bias going down.*

Remark 3.5.2. *As $h \downarrow$ we need increasingly more observations to make up for the lost information in the kernel matrix where zeros start to appear more often. Notice how the final plot of Figure 3.2 is much smoother than the previous with much less observations.*

Remark 3.5.3. *It is well known that kernel methods have issues close to the boundaries and the estimation can be well off, one standard way in the literature to correct boundary issues is local regression see [69, Chapter 6].*

3.6 Appendix

We here collect a series of results that are useful for our analysis.

3.6.1 Deviation inequalities for Bernoulli random variables

The following inequality is a direct consequence of Cramer inequality for Bernoulli variables:

Lemma 3.6.1. *Let $(\varepsilon_n)_{n\geqslant 1}$ be a sequence of independent and identically distributed random variables on $\{0,1\}$ of parameter $p \in (0,1]$. Then, for any $a \geqslant e^2$,*

$$\mathbb{P}\Big(\bigcup_{n\geqslant N}\Big\{\frac{1}{n}\sum_{k=1}^{n}\varepsilon_k \geqslant ap\Big\}\Big) \leqslant \Big(\frac{p}{2}\Big)^{a/2-1},$$

with $N = \lceil -\ln(p/2)/(2p) \rceil$.

Proof. We first notice that, without any loss of generality, we can assume $ap \leqslant 1$ as otherwise the bound is trivial.

Then, we use the following standard Cramer inequality for Bernoulli random variables (see for instance [49]). For any $a \in (0, 1/p]$, we get

$$\mathbb{P}\Big(\bigcup_{n \geqslant N} \Big\{ \frac{1}{n} \sum_{k=1}^{n} \varepsilon_k \geqslant ap \Big\} \Big) \leqslant \sum_{n \geqslant N} \exp\Big\{ -n \Big[ap \ln(a) + (1-ap) \ln\Big(\frac{1-ap}{1-p} \Big) \Big] \Big\}.$$

Here $(1-ap)/(1-p) \geqslant 1-ap$, and

$$\big(1-ap\big) \ln\Big(\frac{1-ap}{1-p} \Big) \geqslant \big(1-ap\big) \ln\big(1-ap\big) \geqslant -ap.$$

where we used the fact that, for $u \in [0,1]$,

$$(1-u)\ln\big(1-u\big) + u = \big[(1-u)\ln\big(1-u\big) - (1-u)\big] - [-1] = -\int_{1-u}^{1} \ln(x) dx \geqslant 0.$$

Therefore,

$$\mathbb{P}\Big(\bigcup_{n \geqslant N} \Big\{ \frac{1}{n} \sum_{k=1}^{n} \varepsilon_k \geqslant ap \Big\} \Big) \leqslant \sum_{n \geqslant N} \exp\Big\{ n \Big[-ap \ln(a) + ap \Big] \Big\}. \tag{3.6.1}$$

Then, for $a \geqslant e^2$,

$$\mathbb{P}\Big(\bigcup_{n \geqslant N} \Big\{ \frac{1}{n} \sum_{k=1}^{n} \varepsilon_k \geqslant ap \Big\} \Big) \leqslant \sum_{n \geqslant N} \exp\big(-nap\big) \leqslant \int_{N-1}^{+\infty} \exp(-apx) dx = \frac{1}{ap} \exp\Big(-ap(N-1) \Big).$$

Since $a \geqslant 1$, we get, for the value of N chosen in the statement,

$$\mathbb{P}\Big(\bigcup_{n \geqslant N} \Big\{ \frac{1}{n} \sum_{k=1}^{n} \varepsilon_k \geqslant ap \Big\} \Big) \leqslant \Big(\frac{p}{2} \Big)^{a/2-1}.$$

This completes the proof. □

3.6.2 Coupon collector

Lemma 3.6.2. *For an integer $J \geqslant 1$, let*

$$W_1 = \sum_{i=1}^{J} \widetilde{t}_i,$$

with $(\widetilde{t}_i)_{1 \leqslant i \leqslant J}$ being a sequence of independent variables, with each $\widetilde{t}_i$ following a geometric distribution of parameter $\min(1, \eta'(1-(i-1)/J))$, for some $\eta' \leqslant J$.

Then, there exists a universal constant C such that
(i) for any $\varepsilon \in (0,1)$,

$$\mathbb{E}\Big[\exp\Big((1-\varepsilon)\frac{\eta'}{J}\big[W_1 - \mathbb{E}(W_1)\big]\Big)\Big] \leqslant \exp\Big(1 + C\frac{(1-\varepsilon)^2}{\varepsilon}\Big),$$

and we can find a measurable map F from $\mathbb{R}\times[0,1]$ into $\mathbb{R}$ such that, for any $[0,1]$-valued uniformly distributed random variable U, the random variable $\widetilde{W}_1 := F(W_1, U)$ satisfies $W_1 - J/\eta' \leqslant \widetilde{W}_1 \leqslant W_1$ and

$$\mathbb{E}\Big[\exp\Big((1-\varepsilon)\frac{\eta'}{J}\big[\widetilde{W}_1 - \mathbb{E}(\widetilde{W}_1)\big]\Big)\Big] \leqslant \exp\Big(C\frac{(1-\varepsilon)^2}{\varepsilon}\Big);$$

(ii) for any $r > 0$,

$$\mathbb{E}\Big[\exp\Big(r\frac{\eta'}{J}\big[\mathbb{E}(W_1) - W_1\big]\Big)\Big] \leqslant \exp\big(1 + Cr^2\big);$$

(iii) for any $r > 0$,

$$\mathbb{P}\Big(W_1 \leqslant \frac{J\ln(J/[2\eta'])}{2\eta'}\Big) \leqslant \exp\big(1 + Cr^2\big)\Big(\frac{2\eta'}{J}\Big)^{r/2}.$$

Proof. First Step. We first notice that the parameter of $\widetilde{t}_i$ becomes (strictly) less than 1 for i satisfying $\eta'[1-(i-1)/J] < 1$, namely $J(1-1/\eta') < i-1 \Leftrightarrow i > J(1-1/\eta') + 1$. Below, we let $J_{\eta'} = \lceil J(1-1/\eta')\rceil + 1$. Since $\eta' \leqslant J$, we have $J(1-1/\eta') \leqslant J-1$ and thus $J_{\eta'} \leqslant J$. Clearly,

$$\frac{\eta'}{J}\sum_{i=1}^{J}\big[\widetilde{t}_i - \mathbb{E}[\widetilde{t}_i]\big] = \frac{\eta'}{J}\sum_{i=J_{\eta'}}^{J}\big[\widetilde{t}_i - \mathbb{E}[\widetilde{t}_i]\big].$$

Second Step. In this step, we prove the claim (i) in the statement. For $i \in \{J_{\eta'}, \cdots, J\}$, we recall that $\widetilde{t}_i = \lfloor Y_i\rfloor + 1$, where Y_i is an exponential random variable of parameter λ_i such that $1 - \exp(-\lambda_i) = \eta'(1-(i-1)/J) \Leftrightarrow \lambda_i = -\ln(1-\eta'+\eta'(i-1)/J)$. Therefore,

$$\begin{aligned}\mathbb{E}\Big[\exp\Big((1-\varepsilon)\frac{\eta'}{J}\sum_{i=J_{\eta'}}^{J}\big[\widetilde{t}_i - \mathbb{E}[\widetilde{t}_i]\big]\Big)\Big] &\leqslant \mathbb{E}\Big[\exp\Big(1 + (1-\varepsilon)\frac{\eta'}{J}\sum_{i=J_{\eta'}}^{J}\big[Y_i - \mathbb{E}[Y_i]\big]\Big)\Big]\\ &= \exp(1)\prod_{i=J_{\eta'}}^{J}\mathbb{E}\Big[\exp\Big((1-\varepsilon)\frac{\eta'}{J}\big[Y_i - \mathbb{E}[Y_i]\big]\Big)\Big],\end{aligned}$$

where we used the fact that $J - J_{\eta'} + 1 < J/\eta'$. Now,

$$\begin{aligned}\mathbb{E}\Big[\exp\Big((1-\varepsilon)\frac{\eta'}{J}\big[Y_i - \mathbb{E}[Y_i]\big]\Big)\Big] &= \lambda_i\exp\Big(-(1-\varepsilon)\frac{\eta'}{J\lambda_i}\Big)\int_0^\infty \exp\Big((1-\varepsilon)\frac{\eta'}{J}x - \lambda_i x\Big)dx\\ &= \lambda_i\exp\Big(-(1-\varepsilon)\frac{\eta'}{J\lambda_i}\Big)\Big(\lambda_i - (1-\varepsilon)\frac{\eta'}{J}\Big)^{-1}\\ &= \exp\Big(-(1-\varepsilon)\frac{\eta'}{J\lambda_i} - \ln\Big[1-(1-\varepsilon)\frac{\eta'}{J\lambda_i}\Big]\Big).\end{aligned}$$

Using the inequality $-\ln(1-u) \geqslant u$, for $u \in (0,1)$, we emphasize that $\lambda_i \geqslant \eta'[1-(i-1)/J]$, which gives $\lambda_i \geqslant \eta'/J$ and then $(1-\varepsilon)\eta'/[J\lambda_i] \leqslant 1-\varepsilon$. Now, we notice that, for $u \in [0, 1-\varepsilon]$,

$$\ln\big(1-u\big) = -\int_{1-u}^{1} \frac{1}{x} dx = -u + \int_{1-u}^{1} \frac{x-1}{x} dx \geqslant -u - \frac{1}{\varepsilon}\int_{1-u}^{1} (1-x)dx \geqslant -u - \frac{1}{2\varepsilon}u^2. \quad (3.6.2)$$

And then,

$$\mathbb{E}\Big[\exp\Big((1-\varepsilon)\frac{\eta'}{J}\big[Y_i - \mathbb{E}[Y_i]\big]\Big)\Big] \leqslant \exp\Big(\frac{(\eta')^2(1-\varepsilon)^2}{2J^2\lambda_i^2\varepsilon}\Big) \leqslant \exp\Big(\frac{(1-\varepsilon)^2}{2(J-(i-1))^2\varepsilon}\Big).$$

Taking the product over $i \in \{J_{\eta'}, \cdots, J\}$, we easily complete the proof of the two inequalities in the first claim. In this regard, we observe that, for each i, Y_i may be constructed in a canonical way from $\widetilde{t}_i$. Indeed, one has, for any integer $k \geqslant 1$ and any $x \in [0,1]$,

$$\begin{aligned}\mathbb{P}\Big((k+1) - Y_i \geqslant x \,\big|\, \lfloor Y_i \rfloor = k\Big) &= \mathbb{P}\Big(k \leqslant Y_i \leqslant k+1-x \,\big|\, k \leqslant Y_i \leqslant k+1\Big)\\ &= \frac{1-\exp(-\lambda_i x)}{1-\exp(-\lambda_i)}.\end{aligned}$$

The second line right above gives the cumulative distribution function of the law of $\widetilde{t}_i - Y_i$ given $\widetilde{t}_i$. In particular, calling F_i the corresponding quantile function, we may let $Y_i = \widetilde{t}_i - F_i(U_i)$ for U_i any $[0,1]$-valued uniformly distributed random variable. Letting $\widetilde{W}_1 := W_1 - \sum_{i=J_{\eta'}}^{J} F_i(U_i)$ for any independent $[0,1]$-valued uniformly distributed random variables $(U_{J_{\eta'}}, \cdots, U_J)$, this provides the form of F in item (i).

Third Step. We now prove the claim (ii). The procedure is very similar to the proof of (i) except that, due to the reversed sign in the exponential, the integrability properties are stronger. In clear, for any $r > 0$,

$$\begin{aligned}\mathbb{E}\Big[\exp\Big(r\frac{\eta'}{J}\big[\mathbb{E}[Y_i] - Y_i\big]\Big)\Big] &= \lambda_i \exp\Big(r\frac{\eta'}{J\lambda_i}\Big)\int_0^\infty \exp\Big(-r\frac{\eta'}{J}x - \lambda_i x\Big)dx\\ &= \lambda_i \exp\Big(r\frac{\eta'}{J\lambda_i}\Big)\Big(\lambda_i + r\frac{\eta'}{J}\Big)^{-1}\\ &= \exp\Big(r\frac{\eta'}{J\lambda_i} - \ln\Big[1 + r\frac{\eta'}{J\lambda_i}\Big]\Big).\end{aligned}$$

We now use the following analogue of (3.6.2). Recall indeed that, for $u > 0$,

$$\ln(1+u) = \int_0^u \frac{1}{1+x}dx = u - \int_0^u \frac{x}{1+x}dx \geqslant u - \frac{u^2}{2}.$$

Then,

$$\mathbb{E}\Big[\exp\Big(r\frac{\eta'}{J}\big[\mathbb{E}[Y_i] - Y_i\big]\Big)\Big] \leqslant \exp\Big(\frac{(\eta')^2 r^2}{2J^2\lambda_i^2}\Big) \leqslant \exp\Big(\frac{r^2}{2(J-(i-1))^2}\Big).$$

Taking the product over $i \in \{J_{\eta'}, \cdots, J\}$, we can easily follow the argument of the first claim. We get

$$\mathbb{E}\Big[\exp\Big(r\frac{\eta'}{J}\big[\mathbb{E}(W_1) - W_1\big]\Big)\Big] \leqslant \exp\big(r + Cr^2\big);$$

Using Young's equality, we upper bound r by $1/2 + r^2/2$ and then complete the proof.

Third Step. We observe that

$$\mathbb{E}\sum_{i=1}^{J}\widetilde{t}_i = J_{\eta'} - 1 + \sum_{i=J_{\eta'}}^{J}\frac{J}{\eta'(J-(i-1))} = J_{\eta'} - 1 + \frac{J}{\eta'}\sum_{i=1}^{J-J_{\eta'}+1}\frac{1}{i} \geqslant \frac{J\ln(J - J_{\eta'} + 1)}{\eta'}.$$

We know from the first step that $J_{\eta'} \leqslant J(1-1/\eta') + 2$ and then $J - J_{\eta'} + 1 \geqslant J/\eta' - 1 \geqslant J/(2\eta')$ since $J \geqslant \eta'$. Since $J_{\eta'} \geqslant 1$, we get

$$\mathbb{E}(W_1) = \mathbb{E}\sum_{i=1}^{J}\widetilde{t}_i \geqslant \frac{J\ln(J/[2\eta'])}{\eta'} \geqslant \frac{J\ln(J/[2\eta'])}{\eta'}.$$

From claim (ii), we deduce that, for any $r > 0$,

$$\begin{aligned}\mathbb{P}\Big(\mathbb{E}(W_1) - W_1 \geqslant \frac{J\ln(J/[2\eta'])}{2\eta'}\Big) &= \mathbb{P}\Big(r\Big[\mathbb{E}(W_1) - W_1\Big] \geqslant r\frac{J\ln(J/[2\eta'])}{2\eta'}\Big)\\ &\leqslant \exp\big(1 + Cr^2\big)\Big(\frac{2\eta'}{J}\Big)^{r/2}.\end{aligned}$$

Therefore,

$$\mathbb{P}\Big(W_1 \leqslant \frac{J\ln(J/[2\eta'])}{2\eta'}\Big) \leqslant \exp\big(1 + Cr^2\big)\Big(\frac{2\eta'}{J}\Big)^{r/2}.$$

□

3.6.3 Sobolev embeddings and bases

We establish the Sobolev embedding like properties (3.4.9). For two reals $k \geqslant 0$ and $m \geqslant \lfloor d_S/2 \rfloor + 1$, the embedding $i_{m+k,k}$ from $H^{m+k}(\overline{S})$ into $H^k(\overline{S})$ is Hilbert-Schmidt (and thus compact). In particular, by composition of Hilbert-Schmidt operators, the embedding $H^{2m+k}(\overline{S})$ into $H^k(\overline{S})$ is trace class (or 1-Schatten). As such, there exists an orthonormal basis $(e_i)_{i\geqslant 1}$ of $H^{2m+k}(\overline{S})$, it holds

$$\sum_{i\geqslant 0}\|e_i\|_{H^k(\overline{S})} < \infty. \tag{3.6.3}$$

If we now choose $k = m$, then classical Sobolev embedding theorem implies that

$$\sum_{i\geqslant 0}\|e_i\|_{L^\infty(\overline{S})} < \infty.$$

In fact, by duality, the adjoint mapping $i^*_{m+k,k}$ is also an Hilbert-Schmidt from $H^k(\overline{S})$ into $H^{m+k}(\overline{S})$. We thus have that the mapping

$$\begin{aligned}\big(i_{2m+k,m+k}, i^*_{2m+k,3m+k}\big) : H^{2m+k}(\overline{S}) &\to H^{m+k}(\overline{S}) \times H^{3m+k}(\overline{S})\\ h &\mapsto \big(i_{2m+k,m+k}(h), i^*_{2m+k,3m+k}(h)\big)\end{aligned}$$

is Hilbert-Schmidt. Similarly,

$$\begin{aligned}\big(i_{m+k,k}, i^*_{3m+k,4m+k}\big) : H^{m+k}(\overline{S}) \times H^{3m+k}(\overline{S}) &\to H^k(\overline{S}) \times H^{4m+k}(\overline{S}) \\ (h_1, h_2) &\mapsto \big(i_{m+k,k}(h_1), i^*_{3m+k,4m+k}(h_2)\big)\end{aligned}$$

is Hilbert-Schmidt. Therefore, by composition

$$\begin{aligned}\big(i_{2m+k,k}, i^*_{2m+k,4m+k}\big) : H^{2m+k}(\overline{S}) &\to H^k(\overline{\mathcal{S}}) \times H^{4m+k}(\overline{S}) \\ h &\mapsto \big(i_{2m+k,k}(h), i^*_{2m+k,4m+k}(h)\big)\end{aligned}$$

is trace class. We deduce that there exists an orthonormal basis $(e_i)_{i\geqslant 1}$ of $H^{2m+k}(\overline{S})$ such that

$$\sum_{i\geqslant 0} \sqrt{\|i_{2m+k,k}(e_i)\|^2_{H^k(\overline{S})} + \|i^*_{2m+k,4m+k}(e_i)\|^2_{H^{4m+k}(\overline{S})}} < \infty.$$

Equivalently,

$$\sum_{i\geqslant 0} \|i_{2m+k,k}(e_i)\|_{H^k(\overline{S})} + \sum_{i\geqslant 0} \|i^*_{2m+k,4m+k}(e_i)\|_{H^{4m+k}(\overline{S})} < \infty. \tag{3.6.4}$$

We now choose $k = m$. For $\varphi \in H^{5m}(\overline{S})$, with $\|\varphi\|_{H^{5m}(\overline{S})} \leqslant 1$, we have

$$\begin{aligned}\sum_{i\geqslant 0} \Big|\big(i_{5m,3m}\varphi, e_i\big)_{H^{3m}(\overline{S})}\Big| &= C_m \sum_{i\geqslant 0} \big|\big(\varphi, i^*_{5m,3m}e_i\big)_{H^{5m}(\overline{S})}\big| \\ &\leqslant C_m \sum_{i\geqslant 0} \big\|i^*_{5m,3m}(e_i)\big\|_{H^{5m}(\overline{S})} \leqslant C_m,\end{aligned}$$

for a new value of the constant C_m in the last line. This is the first line in (3.6.3) The second line in (3.6.3) (for the same basis) follows from (3.6.3) and (3.6.3).

Chapter 4

Perspectives for future research

In this final part of the thesis we will present two open problems that have not been addressed so far in the literature, along with some methodology and tools to potentially solve them. Both of them are connected and stem from practical applications related to collaborative MARL and in our opinion demonstrate how AI can fuel further research in mathematics.

4.1 How to make an implementable MFMDP

It is common practice in order to solve collaborative MARL, to define a Mean Field Markov Decision Process (MFMDP) to model the interactions of the representative player of the population with the environment and then try to learn the optimal value function of the MFMDP. Value Iteration is such an iterative algorithm that computes the optimal state value function by iteratively improving its estimate until convergence. The optimal policy can then be derived from the optimal value function.

In the case of single agent classical RL in Euclidian spaces, the algorithm requires either function approximation or states discretisation, which results to the well known Approximate Value Iteration (AVI). For the MFMDP on continuous spaces this is not enough because we are dealing with a problem that is set on the space of probability measures.

The standard single agent MDP can be implemented as a computer environment that anyone can interact with and thus solve by classical RL algorithms. Such an implementation can be found for example in `gymnasium` [64], and is very valuable for the development of the respective community. This is also part of our motivation for the research question we are proposing.

In a nutshell, our intention is to construct an approximate, finite dimensional MFMDP that we can simulate autonomously and then learn via classical algorithms of RL. The error between the theoretical model and the implementable one, is one part due to distribution approximation and one part due to construction of a finite-dimensional simulator for the dynamics. Our main result would be a quantification of this approximation error. We believe that by designing a methodology to create implementable MFMDPs we bring value to communities of RL and MFGs since we facilitate their interaction and collaboration.

4.1.1 Literature Review

So far examples of MFMDPs come from either exclusively finite state models, [36, 61], or continuous state ones with space discretisation [36, 14] relying on a theoretical model to provide the mean field transitions.

For the sake of brevity we focus our attention on [14] which is the closest to what we have in mind. The authors rely on several layers of approximation to construct a finite model that gives the optimal policy which is also optimal for the infinite population game. In one approach, direct aggregation of perfectly observable measures, at each step the best response is computed for all discretised states and approximated mean field states (empirical measure and nearest neighbor aggregation). Then, implementing the best response, the new mean field state is observed and accordingly approximated to introduce a new iteration. In a second approach, instead of direct observation of the whole mean field they have access to the distribution of n-players from the population and construct the best response given this observation. In both cases, the algorithm relies on the infinite population model for each iteration and thus it cannot be directly implemented is a form of a simulated MFMDP.

4.1.2 Tools

In order to construct our finite MFMDP we need two types of approximations, one for the distributions and one for the dynamics.

One possible choice for the first type is minimization of the Cramér distance

$$d_{C_p}(\mu,\nu) = \left(\int_X |F_\mu(x) - F_\nu(x)|^p dx\right)^{\frac{1}{p}}, \tag{4.1.1}$$

F_μ, F_ν the distribution of μ, ν. for the $p = 1$ in $d = 1$ it is the same as the 1-Wasserstein distance. There exists as well a useful dual formula,

$$d_{C_p}(\nu,\mu) = \left(\int |F_\mu(x) - F_\nu(x)|^p dx\right)^{\frac{1}{p}} = \sup_\varphi \Big\{\int \varphi(x) dF_\mu - \int \varphi(x) dF_\nu\Big\},$$

where the *sup* is taken over all absolutely continuous functions $\varphi(x) = \int_a^x \varphi'(t)dt$ with $\|\varphi'\|_{L^q} \leqslant 1$ and $\frac{1}{p} + \frac{1}{q} = 1$.

The Cramér distance has the benefit that can be computed efficiently numerically and preserves all the nice topological properties of Wasserstein. This minimization is quite flexible since it doesn't require an a priori fixed discretisation lattice. For some examples and algorithms see [10].

Another possible choice could be the method developed by Pham and Warin in [98]. There the authors approximate a functional $\mathcal{V} : \mu \in \mathcal{P}_2(\mathbb{R}^d) \to V(\cdot,\mu) \in L^2(\mu)$ using neural networks that get either a piecewise-constant density called bin approximation or another neural network on which acts the measure μ. The authors provide also details about data generation, training and examples of mean field models, which except from inspiration offer a valuable testbed for benchmarking.

We would like to comment that we don't necessarily see these tools as competing but rather as complementary where for example we could try to use Cramér distance as a form of preprocessing

for the discretisation of distributions that enter the network. After all neural networks are known for their flexibility and modularity.

Other tools that we could combine are cubature methods, Voronoi diagram projections and particle filters, all of them requiring further investigation.

4.1.3 Strategy/Methodology

The main idea is to define the infinite-dimensional MFMDP and an approximated finite-dimensional counterpart that we will learn afterwards by conventional methods. In practice and in accordance with what has been investigated so far in the thesis, we would like to have a first layer of approximation of type

$$\left|V^*(\mu) - \widehat{V}^*(\mu^K)\right| \leqslant \left|V^*(\mu) - V^*(\mu^K)\right| + \left|V^*(\mu^K) - \widehat{V}^*(\mu^K)\right|$$

where V^* is the value optimal value function of the infinite model and $\widehat{V}^*$ the one from the finite, with μ and μ^K accordingly. Under standard assumptions V^* can be proven to be continuous (especially due to regularity of rewards) and thus the first term is going to be easy to control, while in the second we expect to see the influence of approximating the transitions.

For a second layer of approximation, we learn the finite model by some RL algorithm whether is would be model based or model free.

$$\left|\widehat{V}^*(\mu^K) - \widehat{V}_n^*(\mu^K)\right|$$

4.1.4 Difficulties

From the two layers of approximation, second one is more standard since the literature both theoretically and numerically is very well developed however the problem could possibly be high dimensional. **The main difficulty and the novelty of the work would be to approximate effectively the transitions.**

In MFRL this problem is well known and usually falls under the umbrella of *optimization with a mean field oracle* where the oracle gives the transitions of mean field state. Instead, our intension is the constructed MFMDP to be *oracle free* where we don't assume anymore access to the model which gives the transitions[1].

This problem, in the case of an MFG is treated the recent work [120] where in a finite state model the occupancy measure is used to approximate the consistency condition for the distribution of the population. Of course the use of occupancy measure create a dependency of the derived policies on the history of visits for the states and thus non-Markovianity. However for standard RL with finite spaces it has been shown that this fact doesn't create problems since there always exists a Markovian policy that has the same occupancy measure as the non-Markovian. Very recently, [79] this property has been shown to hold also for continuous state MDPs. We would like to explore this opportunity and adapt their arguments in our set-up.

[1] not to be mistaken with model free

4.2 How to solve directly a continuous MFMDP, without state discretisation

The main motivation behind this research question is similar to the previous chapter of the thesis we would like to approximate the continuous MDP without passing from state discretisation and in addition have an efficient numerical implementation. On the one hand we have seen that using a kernel approximation for the value we respect the contraction property for the Bellman operator, on the other if we want to have an efficient implementation we necessarily have to constrain complexity since the memory requirements can grow pretty fast.

In our opinion combining, a stochastic approximation point of view with kernels would be beneficial for a MFMDP since off policy stochastic approximation (like Q-learning see [109]) requires to sample a large number of trajectories and then learn the optimal value. This combines very well with recent advancements like [98] and Langevin [40] so we can have efficient ways to sample probability measures that we then use in our kernels for an optimal representation. Note, that this strategy is different than what we proposed in the first research question where the interaction with the implementable MFMDP will provide the data for learning.

Since the approach is quite novel for mean field problems we refer to the literature review on the previous chapter of the thesis for kernel based approximation on single agent RL. We feel it is better first to detail our strategy and present our tools as we go since this time the idea is more complex and requires several steps to be complete.

4.2.1 Literature review

The only relevant references to our research question we were able to find while preparing this chapter was [39] and [113].

In [113] which appeared first chronologically, the authors consider a problem with continuous states, centralised actions for a finite space and instead of the whole mean field state they access only N agents distributed according to the mean field state μ. They assume to have access to a data set with uniform coverage. To represent the measures that enter into Q function they use a mean embedding into a Reproducing Kernel Hilbert Space (RKHS) and then a Mean Field Fitted-Q-Iteration (FQI) for updating the Q function.

In [39] the authors consider a multi-agent MDP where each agent has a local state and take local actions and the transitions depend upon these local quantities as well as the mean field state. In contrast with [113] they assume weak coverage over their dataset. They use as well mean embedding for the distributions and an algorithm based on the pessimism principle.

Both publications provide interesting implementations that should be taken into account for possible benchmarks of our work.

4.2.2 Strategy/Methodology and Tools

We start by assuming that the action value function belongs in some Reproducing Kernel Hilbert Space $\mathcal{H}$, so Q can be represented by an infinite sum of kernels[2], then revise a gradient descent on

[2]which we later turn into a finite via the representer's theorem

the RKHS to identify the fixed point of the Bellman operator. It's not hard to prove existence of such a space of functionals on the space of probability measures for $d = 1$. For a general introduction to RKHS see [103] and for an implementation of the strategy on single agent continuous MDP [77]

The first step is to rewrite the Bellman operator as an equation that we want to solve for all pairs μ, α in the respective domain

$$l(Q) = R(\mu, \alpha) + \mathbb{E}[\max_{\alpha'} Q(\mu_1, \alpha')] + Q(\mu, \alpha)$$

We take the square and integrate with respect to an arbitrary everywhere dense density over $\mathcal{P}(X) \times \mathcal{P}(A)$ to account for all initial conditions.

$$L(Q) = \frac{1}{2}\mathbb{E}_{\mu,\alpha}[l^2(Q)] = \int_{\mathcal{P}(X)\times\mathcal{P}(A)} \frac{1}{2} l^2(Q) \mathbb{P}(d\mu, d\alpha). \tag{4.2.1}$$

For technical reasons we need to add a penalisation so the final cost of the descent would be

$$J(Q) = L(Q) + \frac{\rho}{2}\|Q\|^2_{\mathcal{H}}$$

Now, it is pretty clear that in order to differentiate the cost (in Fréchet sense) we need it to be smooth, which could be done by replacing the max for a $softmax$, i.e. considering the Regularized Bellman operator. The major problem is that this procedure is not standard for measures and we have to give meaning to an operation like that.

There is one more remark to make before we deal with implementing the functional descent. Equation (4.2.1) involves a nested expectation problem that can be translated as follows "**in order to obtain samples of the gradient $\nabla_Q J$ we require two different queries to a simulation oracle: one to approximate the inner expectation over the transition dynamics defined by $\mu_1 = F(\mu, \alpha, \varepsilon^0)$, and one for each initial pair μ, α which defines the outer expectation**". As proposed in [77] we can resort to a two timescales stochastic approximation.

Following with the implementation of the descent, we apply some form of discretisation to probability measures to obtain finite-dimensional objects that we can store in memory and deal with the sampling issues. Here again we think that some Cramér distance, quantization bins or Langevin dynamics [40] could give us the solution to the problem.

Last, comes the question of constraining the complexity. This can be in form of decreasing the memory requirements or compressing the memory representations of the functionals. In the first category we mention the classical Nyström method [117] while in the second we refer to techniques of representation learning [19] for a general review and [110] for a method adapted to kernels.

4.2.3 Difficulties

On the theoretical side, in order to design the learning scheme we need a proper definition of a regularized Bellman operator for the Mean Field Value function that for the moment remains largely open.

On the practical side, we need efficient simulation of samples μ, α, μ_1 and to balance the memory requirements. This balance is by no means easy since the error will depend on the discretisation of the distributions that will be one of the crucial factors for the memory requirements. Nevertheless we are optimistic that a form of representation learning or compression will be able bring the scale into balance.

Bibliography

[1] Yves Achdou, Fabio Camilli, and Italo Capuzzo-Dolcetta. "Mean field games: convergence of a finite difference method". In: *SIAM J. Numer. Anal.* 51.5 (2013), pp. 2585–2612. ISSN: 0036-1429. DOI: 10.1137/120882421. URL: https://doi.org/10.1137/120882421.

[2] Yves Achdou, Fabio Camilli, and Italo Capuzzo-Dolcetta. "Mean field games: numerical methods for the planning problem". In: *SIAM J. Control Optim.* 50.1 (2012), pp. 77–109. ISSN: 0363-0129. DOI: 10.1137/100790069. URL: https://doi.org/10.1137/100790069.

[3] Yves Achdou and Italo Capuzzo-Dolcetta. "Mean field games: numerical methods". In: *SIAM J. Numer. Anal.* 48.3 (2010), pp. 1136–1162. ISSN: 0036-1429. DOI: 10.1137/090758477. URL: https://doi.org/10.1137/090758477.

[4] Yves Achdou and Mathieu Laurière. "Mean Field Games and Applications: Numerical Aspects". In: *Mean Field Games, Cetraro, Italy 2019, Cardaliaguet, Pierre, Porretta, Alessio (Eds.)* LNM 2281. Springer, 2021, pp. 203–248.

[5] Yves Achdou and Mathieu Laurière. "Mean field type control with congestion (II): An augmented Lagrangian method". In: *Appl. Math. Optim.* 74.3 (2016), pp. 535–578. ISSN: 0095-4616. DOI: 10.1007/s00245-016-9391-z. URL: https://doi.org/10.1007/s00245-016-9391-z.

[6] Robert A. Adams and John J. F. Fournier. *Sobolev spaces*. Second. Vol. 140. Pure and Applied Mathematics (Amsterdam). Elsevier/Academic Press, Amsterdam, 2003, pp. xiv+305. ISBN: 0-12-044143-8.

[7] Clémence Alasseur, Imen Ben Taher, and Anis Matoussi. "An extended mean field game for storage in smart grids". In: *J. Optim. Theory Appl.* 184.2 (2020), pp. 644–670. ISSN: 0022-3239. DOI: 10.1007/s10957-019-01619-3. URL: https://doi.org/10.1007/s10957-019-01619-3.

[8] L. Briceño Arias et al. "On the implementation of a primal-dual algorithm for second order time-dependent mean field games with local couplings". In: *CEMRACS 2017—numerical methods for stochastic models: control, uncertainty quantification, mean-field*. Vol. 65. ESAIM Proc. Surveys. EDP Sci., Les Ulis, 2019, pp. 330–348. DOI: 10.1051/proc/201965330. URL: https://doi.org/10.1051/proc/201965330.

[9] L. M. Briceño Arias, D. Kalise, and F. J. Silva. "Proximal methods for stationary mean field games with local couplings". In: *SIAM J. Control Optim.* 56.2 (2018), pp. 801–836. ISSN: 0363-0129. DOI: 10.1137/16M1095615. URL: https://doi.org/10.1137/16M1095615.

[10] Alessandro Barbiero and Asmerilda Hitaj. "Discrete approximations of continuous probability distributions obtained by minimizing Cramér-von Mises-type distances". In: *Statistical Papers* (2022). DOI: 10.1007/s00362-022-01356-2. URL: https://doi.org/10.1007/s00362-022-01356-2.

[11] Nicole Bäuerle. "Mean Field Markov Decision Processes". In: *Applied Mathematics & Optimization* 88.1 (2023), p. 12. DOI: 10.1007/s00245-023-09985-1. URL: https://doi.org/10.1007/s00245-023-09985-1.

[12] Nicole Bäuerle and Ulrich Rieder. *Markov decision processes with applications to finance*. Universitext. Springer, Heidelberg, 2011, pp. xvi+388. ISBN: 978-3-642-18323-2. DOI: 10.1007/978-3-642-18324-9. URL: https://doi.org/10.1007/978-3-642-18324-9.

[13] Leonard E. Baum and Patrick Billingsley. "Asymptotic distributions for the coupon collector's problem". In: *Ann. Math. Statist.* 36 (1965), pp. 1835–1839. ISSN: 0003-4851. DOI: 10.1214/aoms/1177699813. URL: https://doi.org/10.1214/aoms/1177699813.

[14] Erhan Bayraktar, Nicole Bauerle, and Ali Devran Kara. *Finite Approximations for Mean Field Type Multi-Agent Control and Their Near Optimality*. 2023. arXiv: 2211.09633 [math.OC].

[15] Erhan Bayraktar et al. "Finite state mean field games with Wright-Fisher common noise". In: *J. Math. Pures Appl.* 147 (2021), pp. 98–162.

[16] Christian Beck, Weinan E, and Arnulf Jentzen. "Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations". In: *J. Nonlinear Sci.* 29.4 (2019), pp. 1563–1619. ISSN: 0938-8974. DOI: 10.1007/s00332-018-9525-3. URL: https://doi.org/10.1007/s00332-018-9525-3.

[17] Jean-David Benamou and Guillaume Carlier. "Augmented Lagrangian methods for transport optimization, mean field games and degenerate elliptic equations". In: *J. Optim. Theory Appl.* 167.1 (2015), pp. 1–26. ISSN: 0022-3239. DOI: 10.1007/s10957-015-0725-9. URL: https://doi.org/10.1007/s10957-015-0725-9.

[18] Christian Bender and Jianfeng Zhang. "Time discretization and Markovian iteration for coupled FBSDEs". In: *Ann. Appl. Probab.* 18.1 (2008), pp. 143–177. ISSN: 1050-5164. DOI: 10.1214/07-AAP448. URL: https://doi.org/10.1214/07-AAP448.

[19] Yoshua Bengio, Aaron Courville, and Pascal Vincent. *Representation Learning: A Review and New Perspectives*. 2014. arXiv: 1206.5538 [cs.LG].

[20] Philippe Briand and Céline Labart. "Simulation of BSDEs by Wiener chaos expansion". In: *The Annals of Applied Probability* 24.3 (2014), pp. 1129 –1171. DOI: 10.1214/13-AAP943. URL: https://doi.org/10.1214/13-AAP943.

[21] L. D. Brown and R. Purves. "Measurable Selections of Extrema". In: *The Annals of Statistics* 1.5 (1973), pp. 902 –912. DOI: 10.1214/aos/1176342510. URL: https://doi.org/10.1214/aos/1176342510.

[22] D. L. Burkholder, B. J. Davis, and R. F. Gundy. "Integral inequalities for convex functions of operators on martingales". In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*. 1972, pp. 223–240.

[23] Pierre Cardaliaguet. *Notes on mean field games*. https://www.ceremade.dauphine.fr/ cardalia/MFG20130420.pdf/. 2013.

[24] Pierre Cardaliaguet and Saeed Hadikhanloo. "Learning in mean field games: the fictitious play". In: *ESAIM Control Optim. Calc. Var.* 23.2 (2017), pp. 569–591. ISSN: 1292-8119. DOI: 10.1051/cocv/2016004. URL: https://doi.org/10.1051/cocv/2016004.

[25] Pierre Cardaliaguet et al. "An algebraic convergence rate for the optimal control of McKean-Vlasov dynamics". In: *arXiv* 2203.14554 (2023).

[26] Pierre Cardaliaguet et al. *The master equation and the convergence problem in mean field games*. Vol. 201. Annals of Mathematics Studies. Princeton University Press, Princeton, NJ, 2019, pp. x+212. ISBN: 978-0-691-19071-6; 978-0-691-19070-9. DOI: 10.2307/j.ctvckq7qf. URL: https://doi.org/10.2307/j.ctvckq7qf.

[27] Stephen Carden. "Convergence of a Q-learning variant for continuous states and actions". In: *J. Artificial Intelligence Res.* 49 (2014), pp. 705–731. ISSN: 1076-9757. DOI: 10.1613/jair.4271. URL: https://doi.org/10.1613/jair.4271.

[28] E. Carlini and F. J. Silva. "A fully discrete semi-Lagrangian scheme for a first order mean field game problem". In: *SIAM J. Numer. Anal.* 52.1 (2014), pp. 45–67. ISSN: 0036-1429. DOI: 10.1137/120902987. URL: https://doi.org/10.1137/120902987.

[29] Elisabetta Carlini and Francisco J. Silva. "A semi-Lagrangian scheme for a degenerate second order mean field game system". In: *Discrete Contin. Dyn. Syst.* 35.9 (2015), pp. 4269–4292. ISSN: 1078-0947. DOI: 10.3934/dcds.2015.35.4269. URL: https://doi.org/10.3934/dcds.2015.35.4269.

[30] René Carmona and François Delarue. *Probabilistic theory of mean field games with applications. I*. Vol. 83. Probability Theory and Stochastic Modelling. Mean field FBSDEs, control, and games. Springer, Cham, 2018, pp. xxv+713. ISBN: 978-3-319-56437-1; 978-3-319-58920-6.

[31] René Carmona and François Delarue. *Probabilistic theory of mean field games with applications. II*. Vol. 84. Probability Theory and Stochastic Modelling. Mean field games with common noise and master equations. Springer, Cham, 2018, pp. xxiv+697. ISBN: 978-3-319-56435-7; 978-3-319-56436-4.

[32] René Carmona and Daniel Lacker. "A probabilistic weak formulation of mean field games and applications". In: *Ann. Appl. Probab.* 25.3 (2015), pp. 1189–1231. ISSN: 1050-5164. DOI: 10.1214/14-AAP1020. URL: https://doi.org/10.1214/14-AAP1020.

[33] René Carmona and Mathieu Laurière. "Convergence analysis of machine learning algorithms for the numerical solution of mean-field control and games: I-the ergodic case." In: *arXiv e-prints* arXiv:1907.05980 (2019).

[34] René Carmona and Mathieu Laurière. "Convergence analysis of machine learning algorithms for the numerical solution of mean-field control and games: II-the finite horizon case." In: *arXiv e-prints* arXiv:1908.01613 (2019).

[35] René Carmona, Mathieu Laurière, and Zongjun Tan. "Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods". In: *arXiv e-prints* arXiv:1910.04295 (2019).

[36] René Carmona, Mathieu Laurière, and Zongjun Tan. *Model-Free Mean-Field Reinforcement Learning: Mean-Field MDP and Mean-Field Q-Learning*. 2021. arXiv: 1910.12802 [math.OC].

[37] Alekos Cecchin and François Delarue. "Selection by vanishing common noise for potential finite state mean field games". In: *arXiv e-prints*, arXiv:2005.12153 (2020).

[38] Jean-François Chassagneux, Dan Crisan, and François Delarue. "Numerical method for FBSDEs of McKean-Vlasov type". In: *Ann. Appl. Probab.* 29.3 (2019), pp. 1640–1684. ISSN: 1050-5164. DOI: 10.1214/18-AAP1429. URL: https://doi.org/10.1214/18-AAP1429.

[39] Minshuo Chen et al. "Pessimism Meets Invariance: Provably Efficient Offline Mean-Field Multi-Agent RL". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021. URL: https://openreview.net/forum?id=Ww1e07fy9fC.

[40] Giovanni Conforti, Daniel Lacker, and Soumik Pal. *Projected Langevin dynamics and a gradient flow for entropic optimal transport*. 2023. arXiv: 2309.08598 [math.PR].

[41] Dan Crisan and Eamon McMurray. "Cubature on Wiener space for McKean-Vlasov SDEs with smooth scalar interaction". In: *Ann. Appl. Probab.* 29.1 (2019), pp. 130–177. ISSN: 1050-5164. DOI: 10.1214/18-AAP1407. URL: https://doi.org/10.1214/18-AAP1407.

[42] Jakša Cvitanić and Jianfeng Zhang. "The steepest descent method for forward-backward SDEs". In: *Electron. J. Probab.* 10 (2005), pp. 1468–1495. ISSN: 1083-6489. DOI: 10.1214/EJP.v10-295. URL: https://doi.org/10.1214/EJP.v10-295.

[43] Constantinos Daskalakis. *Lecture Notes*. URL: http://people.csail.mit.edu/costis/6853fa2011/lec4.pdf.

[44] François Delarue. "On the existence and uniqueness of solutions to FBSDEs in a non-degenerate case". In: *Stochastic Processes and their Applications* 99.2 (2002), pp. 209 – 286. ISSN: 0304-4149. DOI: https://doi.org/10.1016/S0304-4149(02)00085-6. URL: http://www.sciencedirect.com/science/article/pii/S0304414902000856.

[45] François Delarue and Rinel Foguen Tchuendom. "Selection of equilibria in a linear quadratic mean field game". In: *Stochastic Processes and their Applications* 130.2 (2020), pp. 1000–1040. DOI: 10.1016/j.spa.2019.04.005.

[46] François Delarue and Stéphane Menozzi. "A forward-backward stochastic algorithm for quasi-linear PDEs". In: *Ann. Appl. Probab.* 16.1 (2006), pp. 140–184. ISSN: 1050-5164. DOI: 10.1214/105051605000000674. URL: https://doi.org/10.1214/105051605000000674.

[47] François Delarue. "Restoring uniqueness to mean-field games by randomizing the equilibria". In: *Stochastics and Partial Differential Equations: Analysis and Computations* 7 (2019), pp. 598–678.

[48] François Delarue and Athanasios Vasileiadis. *Exploration noise for learning linear-quadratic mean field games*. 2023. arXiv: 2107.00839 [math.OC].

[49] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*. Second. Vol. 38. Applications of Mathematics (New York). Springer-Verlag, New York, 1998, pp. xvi+396. ISBN: 0-387-98406-2. DOI: 10.1007/978-1-4612-5320-4. URL: https://doi.org/10.1007/978-1-4612-5320-4.

[50] Omar Darwiche Domingues et al. *Kernel-Based Reinforcement Learning: A Finite-Time Analysis*. 2022. arXiv: 2004.05599 [cs.LG].

[51] P. Dorato and A. Levis. "Optimal linear regulators: The discrete-time case". In: *IEEE Transactions on Automatic Control* 16 (1971), pp. 613–620.

[52] Kenji Doya. "Reinforcement Learning in Continuous Time and Space". In: *Neural Comput.* 12.1 (2000), pp. 219–245. DOI: 10.1162/089976600300015961. URL: https://doi.org/10.1162/089976600300015961.

[53] Weinan E, Jiequn Han, and Arnulf Jentzen. "Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations". In: *Commun. Math. Stat.* 5.4 (2017), pp. 349–380. ISSN: 2194-6701. DOI: 10.1007/s40304-017-0117-6. URL: https://doi.org/10.1007/s40304-017-0117-6.

[54] Romuald Elie et al. "On the convergence of model free learning in mean field games". In: *AAAI* (2020).

[55] Dena Firoozi and Sebastian Jaimungal. "Exploratory LQG Mean Field Games with Entropy Regularization". In: *arXiv:2011.12946* (2020).

[56] Rinel Foguen Tchuendom. "Uniqueness for linear-quadratic mean field games with common noise". In: *Dyn. Games Appl.* 8.1 (2018), pp. 199–210. ISSN: 2153-0785. DOI: 10.1007/s13235-016-0200-8. URL: https://doi.org/10.1007/s13235-016-0200-8.

[57] Nicolas Fournier and Arnaud Guillin. "On the rate of convergence in Wasserstein distance of the empirical measure". In: *Probab. Theory Related Fields* 162.3-4 (2015), pp. 707–738. ISSN: 0178-8051. DOI: 10.1007/s00440-014-0583-7. URL: https://doi.org/10.1007/s00440-014-0583-7.

[58] D. Fudenberg and D.K. Levine. *The Theory of Learning in Games*. Economics Learning and Social Evolution Series. MIT Press, 1998. ISBN: 9780262061940. URL: https://books.google.gr/books?id=G6vTQFluxuEC.

[59] Matthieu Geist et al. "Concave Utility Reinforcement Learning: the Mean-field Game viewpoint". In: *arXiv:2106.03787* (2021).

[60] Emmanuel Gobet, Jean-Philippe Lemor, and Xavier Warin. "A regression-based Monte Carlo method to solve backward stochastic differential equations". In: *Ann. Appl. Probab.* 15.3 (2005), pp. 2172–2202. ISSN: 1050-5164. DOI: 10.1214/105051605000000412. URL: https://doi.org/10.1214/105051605000000412.

[61] Haotian Gu et al. *Mean-Field Controls with Q-learning for Cooperative MARL: Convergence and Complexity Analysis*. 2021. arXiv: 2002.04131 [cs.LG].

[62] Xin Guo, Renyuan Xu, and Thaleia Zariphopoulou. *Entropy Regularization for Mean Field Games with Learning*. 2021. arXiv: 2010.00145 [math.OC].

[63] Xin Guo et al. "Learning mean-field games". In: *Proceedings of NeurIPS*. 2019.

[64] *Gymnasium*. URL: https://gymnasium.farama.org/index.html.

[65] Saeed Hadikhanloo. "Learning in anonymous nonatomic games with applications to first-order mean field games". In: *arXiv e-prints* arXiv:1704.00378 (2017).

[66] Saeed Hadikhanloo and Francisco J. Silva. "Finite mean field games: fictitious play and convergence to a first order continuous mean field game". In: *J. Math. Pures Appl. (9)* 132 (2019), pp. 369–397. ISSN: 0021-7824. DOI: 10.1016/j.matpur.2019.02.006. URL: https://doi.org/10.1016/j.matpur.2019.02.006.

[67] Jiequn Han and Weinan E. *Deep Learning Approximation for Stochastic Control Problems*. 2016. arXiv: 1611.07422 [cs.LG].

[68] Bruce E. Hansen. "Uniform convergence rates for kernel estimation with dependent data". In: *Econometric Theory* 24.3 (2008), pp. 726–748. ISSN: 0266-4666. DOI: 10.1017/S0266466608080304. URL: https://doi.org/10.1017/S0266466608080304.

[69] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[70] Ruimeng Hu and Mathieu Laurière. "Recent Developments in Machine Learning Methods for Stochastic Control and Games". In: *HAL archives* hal-03656245 (May 2022). working paper or preprint. URL: https://hal.archives-ouvertes.fr/hal-03656245.

[71] Kuang Huang et al. "A game-theoretic framework for autonomous vehicles velocity control: bridging microscopic differential games and macroscopic mean field games". In: *Discrete Contin. Dyn. Syst. Ser. B* 25.12 (2020), pp. 4869–4903. ISSN: 1531-3492. DOI: 10.3934/dcdsb.2020131. URL: https://doi.org/10.3934/dcdsb.2020131.

[72] M. Huang, P.E. Caines, and R.P. Malhamé. "Individual and mass behavior in large population stochastic wireless power control problems: centralized and Nash equilibrium solutions". In: (2003), pp. 98 –103.

[73] Minyi Huang, Peter E. Caines, and Roland P. Malhamé. "The Nash certainty equivalence principle and McKean-Vlasov systems: An invariance principle and entry adaptation". In: *Decision and Control, 2007 46th IEEE Conference on*. IEEE. 2007, pp. 121–126.

[74] Minyi Huang, Roland P. Malhamé, and Peter E. Caines. "Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle". In: *Commun. Inf. Syst.* 6.3 (2006), pp. 221–251. ISSN: 1526-7555. URL: http://projecteuclid.org.proxy.lib.umich.edu/euclid.cis/1183728987.

[75] T. P. Huijskens, M. J. Ruijter, and C. W. Oosterlee. "Efficient numerical Fourier methods for coupled forward-backward SDEs". In: *J. Comput. Appl. Math.* 296 (2016), pp. 593–612. ISSN: 0377-0427. DOI: 10.1016/j.cam.2015.10.019. URL: https://doi.org/10.1016/j.cam.2015.10.019.

[76] Norihiko Kazamaki. In: *Continuous Exponential Martingales and BMO*. Berlin, Heidelberg. ISBN: 978-3-540-48421-9. DOI: 10.1007/BFb0073586.

[77] Alec Koppel et al. *Nonparametric Stochastic Compositional Gradient Descent for Q-Learning in Continuous Markov Decision Problems*. 2018. arXiv: 1804.07323 [cs.LG].

[78] Daniel Lacker. "Mean field games via controlled martingale problems: existence of Markovian equilibria". In: *Stochastic Process. Appl.* 125.7 (2015), pp. 2856–2894. ISSN: 0304-4149. DOI: 10.1016/j.spa.2015.02.006. URL: http://dx.doi.org/10.1016/j.spa.2015.02.006.

[79] Romain Laroche, Remi Tachet des Combes, and Jacob Buckman. *Non-Markovian policies occupancy measures*. 2022. arXiv: 2205.13950 [cs.LG].

[80] Jean-Michel Lasry and Pierre-Louis Lions. "Jeux à champ moyen. I. Le cas stationnaire". In: *C. R. Math. Acad. Sci. Paris* 343.9 (2006), pp. 619–625. ISSN: 1631-073X. DOI: 10.1016/j.crma.2006.09.019. URL: http://dx.doi.org.proxy.lib.umich.edu/10.1016/j.crma.2006.09.019.

[81] Jean-Michel Lasry and Pierre-Louis Lions. "Jeux à champ moyen. II. Horizon fini et contrôle optimal". In: *C. R. Math. Acad. Sci. Paris* 343.10 (2006), pp. 679–684. ISSN: 1631-073X. DOI: 10.1016/j.crma.2006.09.018. URL: http://dx.doi.org.proxy.lib.umich.edu/10.1016/j.crma.2006.09.018.

[82] Jean-Michel Lasry and Pierre-Louis Lions. "Mean field games". In: *Jpn. J. Math.* 2.1 (2007), pp. 229–260. ISSN: 0289-2316. DOI: 10.1007/s11537-007-0657-8. URL: http://dx.doi.org.proxy.lib.umich.edu/10.1007/s11537-007-0657-8.

[83] Mathieu Laurière et al. *Learning Mean Field Games: A Survey*. 2022. arXiv: 2205.12944 [cs.LG].

[84] Pierre-Louis Lions. "Cours au Collège de France, Equations aux dérivées partielles et applications". https://www.college-de-france.fr/site/pierre-louis-lions/course-2010-2011.htm, 2010-11.

[85] Pierre-Louis Lions. "Cours au Collège de France, Equations aux dérivées partielles et applications". https://www.college-de-france.fr/site/pierre-louis-lions/course.htm.

[86] Jin Ma, Philip Protter, and Jiongmin Yong. "Solving forward-backward stochastic differential equations explicitly – a four step scheme". In: *Probability Theory and Related Fields* 98.3 (1994), pp. 339–359. ISSN: 1432-2064. DOI: 10.1007/BF01192258. URL: https://doi.org/10.1007/BF01192258.

[87] Jin Ma, Jie Shen, and Yanhong Zhao. "On numerical approximations of forward-backward stochastic differential equations". In: *SIAM J. Numer. Anal.* 46.5 (2008), pp. 2636–2661. ISSN: 0036-1429. DOI: 10.1137/06067393X. URL: https://doi.org/10.1137/06067393X.

[88] H. P. McKean. "A Class of Markov Processes Associated with Nonlinear Parabolic Equations". In: *Proceedings of the National Academy of Sciences of the United States of America* 56.6 (1966), pp. 1907–1911. ISSN: 00278424. URL: http://www.jstor.org/stable/57643 (visited on 11/19/2023).

[89] S. Méléard. "Asymptotic behaviour of some interacting particle systems; McKean-Vlasov and Boltzmann models". In: *Probabilistic models for nonlinear partial differential equations (Montecatini Terme, 1995)*. Vol. 1627. Lecture Notes in Math. Springer, 1996, pp. 42–95.

[90] G. N. Milstein and M. V. Tretyakov. "Numerical algorithms for semilinear parabolic equations with small parameter based on approximation of stochastic equations". In: *Math. Comp.* 69.229 (2000), pp. 237–267. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-99-01134-5. URL: https://doi.org/10.1090/S0025-5718-99-01134-5.

[91] Médéric Motte and Huyên Pham. *Mean-field Markov decision processes with common noise and open-loop controls*. 2021. arXiv: 1912.07883 [math.OC].

[92] Rémi Munos. "A Study of Reinforcement Learning in the Continuous Case by the Means of Viscosity Solutions". In: *Machine Learning* 40 (2000), pp. 265–299.

[93] Ryan Murray and Michele Palladino. *A model for system uncertainty in reinforcement learning*. 2018. arXiv: 1802.07668 [math.OC].

[94] Dirk Ormoneit and Śaunak Sen. "Kernel-Based Reinforcement Learning". In: *Machine Learning* 49.2 (2002), pp. 161–178. DOI: 10.1023/A:1017928328829. URL: https://doi.org/10.1023/A:1017928328829.

[95] Shige Peng. "Stochastic Hamilton–Jacobi–Bellman Equations". In: *SIAM Journal on Control and Optimization* 30.2 (1992), pp. 284–304. DOI: 10.1137/0330018. eprint: https://doi.org/10.1137/0330018. URL: https://doi.org/10.1137/0330018.

[96] Sarah Perrin et al. "Fictitious play for mean field games: Continuous time analysis and applications". In: *Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS 2020)* (2020). URL: https://arxiv.org/abs/2007.03458.

[97] Sarah Perrin et al. "Mean Field Games Flock! The Reinforcement Learning Way". In: *arXiv:2105.07933* (2021).

[98] Huyên Pham and Xavier Warin. *Mean-field neural networks: learning mappings on Wasserstein space*. 2023. arXiv: 2210.15179 [math.OC].

[99] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[100] P. E. Chaudru de Raynal and C. A. Garcia Trillos. "A cubature based algorithm to solve decoupled McKean-Vlasov forward-backward stochastic differential equations". In: *Stochastic Process. Appl.* 125.6 (2015), pp. 2206–2255. ISSN: 0304-4149. DOI: 10.1016/j.spa.2014.11.018. URL: https://doi.org/10.1016/j.spa.2014.11.018.

[101] Gonçalo dos Reis and Vadim Platonov. "Itô-Wentzell-Lions Formula for Measure Dependent Random Fields under Full and Conditional Measure Flows". In: *Potential Analysis* 59.3 (2023), pp. 1313–1344. DOI: 10.1007/s11118-022-10012-1. URL: https://doi.org/10.1007/s11118-022-10012-1.

[102] Mohit Sewak. *Deep Reinforcement Learning*. Springer, 2019.

[103] Ingo Steinwart and Andreas Christmann. "Support Vector Machines". In: *Information Science and Statistics*. 2008.

[104] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Second. The MIT Press, 2018. URL: http://incompleteideas.net/book/the-book-2nd.html.

[105] A.S. Sznitman. "A fluctuation result for nonlinear diffusions". In: *Infinite Dimensional Analysis and Stochastic Processes* (1985), pp. 145–160.

[106] H. Tanaka and M. Hitsuda. "Central limit theorem for a simple diffusion model of interacting particles". In: *Hiroshima Mathematical Journal* 11.2 (1981), pp. 415–423.

[107] Marita Thomas. "Uniform Poincaré-Sobolev and isoperimetric inequalities for classes of domains". In: *Discrete Contin. Dyn. Syst.* 35.6 (2015), pp. 2741–2761. ISSN: 1078-0947. DOI: 10.3934/dcds.2015.35.2741. URL: https://doi.org/10.3934/dcds.2015.35.2741.

[108] John Tsitsiklis and Benjamin Van Roy. "Analysis of Temporal-Diffference Learning with Function Approximation". In: *Advances in Neural Information Processing Systems*. Ed. by M.C. Mozer, M. Jordan, and T. Petsche. Vol. 9. MIT Press, 1996. URL: https://proceedings.neurips.cc/paper_files/paper/1996/file/e00406144c1e7e35240afed70f34166a-Paper.pdf.

[109] John N. Tsitsiklis. "Asynchronous Stochastic Approximation and Q-Learning". In: *Machine Learning* 16.3 (1994), pp. 185–202. DOI: 10.1023/A:1022689125041. URL: https://doi.org/10.1023/A:1022689125041.

[110] Pascal Vincent and Y. Bengio. "Kernel Matching Pursuit". In: *Machine Learning* 48 (July 2002), pp. 165–187. DOI: 10.1023/A:1013955821559.

[111] Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. "Reinforcement Learning in Continuous Time and Space: A Stochastic Control Approach". In: *Journal of Machine Learning Research* 21.198 (2020), pp. 1–34. URL: http://jmlr.org/papers/v21/19-144.html.

[112] Haoran Wang, Thaleia Zariphopoulou, and Xunyu Zhou. "Exploration Versus Exploitation in Reinforcement Learning: A Stochastic Control Approach". In: *SSRN Electronic Journal* (Jan. 2019). DOI: 10.2139/ssrn.3316387.

[113] Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. *Breaking the Curse of Many Agents: Provable Mean Embedding Q-Iteration for Mean-Field Reinforcement Learning*. 2020. arXiv: 2006.11917 [cs.LG].

[114] C. J. C. H. Watkins. "Learning from Delayed Rewards". PhD thesis. King's College, Oxford, 1989.

[115] Christopher J. C. H. Watkins and Peter Dayan. "Q-learning". In: *Machine Learning* 8.3 (1992), pp. 279–292. DOI: 10.1007/BF00992698. URL: https://doi.org/10.1007/BF00992698.

[116] Hassler Whitney. "Analytic extensions of differentiable functions defined in closed sets". In: *Trans. Amer. Math. Soc.* 36.1 (1934), pp. 63–89. ISSN: 0002-9947. DOI: 10.2307/1989708. URL: https://doi.org/10.2307/1989708.

[117] Christopher Williams and Matthias Seeger. "Using the Nyström Method to Speed Up Kernel Machines". In: *Advances in Neural Information Processing Systems*. Ed. by T. Leen, T. Dietterich, and V. Tresp. Vol. 13. MIT Press, 2000. URL: https://proceedings.neurips.cc/paper_files/paper/2000/file/19de10adbaa1b2ee13f77f679fa1483a-Paper.pdf.

[118] Thaleia Zariphopoulou Xin Guo Renyuan Xu. "Entropy Regularization for Mean Field Games with Learning". In: *arXiv:2010.00145* (2020).

[119] Jiongmin Yong and Xun Yu Zhou. *Stochastic Controls*. Springer New York, NY, 22 June 1999.

[120] Muhammad Aneeq uz Zaman et al. *Oracle-free Reinforcement Learning in Mean-Field Games along a Single Sample Path*. 2023. arXiv: 2208.11639 [cs.LG].

[121] Eberhard Zeidler. *Applied functional analysis*. Vol. 108. Applied Mathematical Sciences. Applications to mathematical physics. Springer-Verlag, New York, 1995, pp. xxx+479. ISBN: 0-387-94442-7.

[122] Chiyuan Zhang et al. *Understanding deep learning requires rethinking generalization*. 2017. arXiv: 1611.03530 [cs.LG].

[123] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms*. 2021. arXiv: 1911.10635 [cs.LG].