Mathematical models for large populations, behavioral economics, and targeted advertising

Médéric Motte

Université de Paris

École doctorale **de Sciences Mathématiques de Paris-Centre 386**

Laboratoire de Probabilités, Statistique et Modélisation (LPSM)

Mathematical models for large populations, behavioral economics and targeted advertising

Par Médéric Motte

Thèse de doctorat de Mathématiques Appliquées

Dirigée par Huyên Pham

Présentée et soutenue publiquement le 15 décembre 2021

Devant un jury composé de :

Huyên Pham, prof. des universités, Univ. de Paris, directeur de thèse
Romuald Elie, prof. des universités, Univ. Gustave Eiffel, rapporteur
Xin Guo, prof. des universités, UC Berkeley CA, rapporteur
Jean-François Chassagneux, prof. des universités, Univ. de Paris, examinateur
François Delarue, prof. des universités, Univ. Nice-Sophia Antipolis, examinateur
Ashkan Nikeghbali, prof. des universités, Univ. of Zürich, examinateur
Vianney Perchet, prof. des universités, ENSAE-CREST, examinateur

Remerciements

Avant d'entrer dans le coeur de ce manuscrit, je souhaite formuler quelques remerciements aux personnes qui ont fait partie de ce voyage qu'à été ma thèse.

Je voudrais particulièrement remercier mon directeur de thèse Huyên Pham, qui m'a permis d'explorer les idées qui m'étaient chères, même lorsqu'elles étaient non conventionnelles, risquées, à l'état embryonaire, et prenaient du temps à mûrir. Huyên, merci pour ta patience, ta confiance, et l'aide que tu m'as apportée durant cette thèse.

Merci aux rapporteurs Romuald Elie et Xin Guo d'avoir pris le temps de lire en détail ce manuscrit et de le rapporter. Merci aux membres du jury Jean-François Chassagneux, François Delarue, Ashkan Nikeghbali, et Vianney Perchet de m'honorer de leur présence.

J'aimerais aussi remercier l'équipe du LPSM, chercheurs et doctorants, qui contribuent tous à créer une atmosphère de recherche stimulante ayant joué un rôle important dans l'écriture de cette thèse. Un merci particulier à Côme, Enzo, Cyril et Maximilien, pour les quelques moments de réflexion collective passés devant le tableau du bureau des thésards. Merci à Guillaume, qui venait aussi du master M2MO, et qui a vécu l'expérience si particulière de la thèse en parallèle de la mienne.

Je remercie aussi les chercheurs de l'université ETH Zürich, qui m'ont accueilli pour une visite de recherche, et particulièrement Delia Coculescu avec qui j'ai eu l'opportunité de développer un travail profondément connecté aux travaux présents dans ce manuscrit.

Je remercie Jean-François Chassagneux et Bastien Fernandez, qui ont pris de leur temps en été pour m'aider dans des moments difficiles.

Merci aux mathématiciens et économistes de l'Histoire, qui n'hésitent pas à développer des théories nouvelles, révolutionnaires, et à les défendre assez longtemps pour qu'elles s'établissent dans le paysage mathématique. J'ai eu plaisir à explorer ce monde quatre années durant, et j'en ressors riche de connaissances, de méthodes, et de beauté mathématique.

Mon remerciement le plus important va à ma femme Imane, qui m'a permis de me consacrer pleinement à ma thèse. Imane, sans ton soutien et ta patience, cette thèse n'aurait pas été possible.

Résumé

Cette thèse étudie des modèles mathématiques pour des problématiques sociales et économiques en ligne, et plus précisément, des modèles de grandes populations en interactions, d'influence sociale, de choix risqués, et de publicité ciblée. L'intérêt de la communauté mathématique pour les sujets socio-économiques sur Internet est relativement nouveau comparé à son intérêt classique pour la physique, ce qui en fait un domaine dans lequel beaucoup reste encore à explorer. La création d'Internet a conduit à de nombreux changements de paradigme concernant la façon dont les gens intéragissent (réseaux sociaux), agissent (navigation sur Internet, décision de clic et d'achat), et dont les business fonctionnent (sites gratuits vivant de la publicité ciblée). Le but de cette thèse est de fournir des outils mathématiques contribuant à une meileure compréhension de ces problématiques, d'un point de vue à la fois théorique et pratique. L'objectif est double: sur le plan mathématique, nous souhaitons illustrer le fait qu'Internet et les réseaux sociaux conduisent naturellement à des mathématiques diverses et intéressantes, et sur le plan des applications, nous voulons fournir des outils mathémathiques potentiellement utiles pour les problèmes concrets se posant sur Internet.

La première partie de ce manuscrit est consacrée à des problèmes de grande population avec une approche relativement théorique, comparée aux parties suivantes plus appliquées. Nous étudions un processus de décision markovien (MDP) à N-agent et un processus de décision markovien de type McKean-Vlasov, avec bruit commun et contrôles open-loop, en horizion infini. Nous obtenons dans un premier temps l'équation de Bellman pour le MDP de type McKean-Vlasov, en exposant et contournant des problèmes de mesurabilité dans le cas d'un espace d'états continu en présence de bruit commun, puis nous établissons la réduction à des politiques stationnaires feedback randomisées, et montrons que l'optimisation sur des contrôles randomisés peut donner un gain strictement supérieur à l'optimisation sur des contrôles non-randomisés, démontrant ainsi la nécessité de randomiser. Nous obtenons dans un second temps l'équation de Bellman du MDP à N-agents, et établissons enfin des résultats de propagation du chaos, à savoir, la convergence des valeurs optimales quand $N \to \infty$ vers la valeur optimal du MDP de type McKean-Vlasov, avec taux de convergence $\mathcal{O}(M_N^\gamma)$, où γ est explicite et M_N est lié à la convergence de mesures empiriques vers la mesure théorique associée au sens de Wasserstein, et le fait qu'une politique feedback randomisée stationnaire ε-optimale pour le MDP de type McKean-Vlasov est une politique $(\varepsilon + \mathcal{O}(M_N^\gamma))$-optimale pour le MDP à N-agents. Finalement, nous appliquons la propagation du chaos pour approximativement résoudre le MDP à N-agent via la résolution du MDP de type McKean-Vlasov associé dans des exemples jouets motivés par la publicité ciblée.

La seconde partie est dédiée à l'étude de modèles d'économie comportementale. Notre premier travail analyse des jeux en grande population via le concept de solution d'Elimination Itérative de Stratégies Dominées, au moyen de méthodes d'estimation de champ moyen. Plus précisément, nous étudions un jeu dans lequel une population de N joueurs doit faire un choix binaire (dans un espace de choix $\mathcal{X} = \{0, 1\}$). Ces choix peuvent avoir de nombreuses interprétations: acheter ou non un produit, s'inscrire ou non à un service, publiquement soutenir ou non une opinion, etc. Chaque joueur n, pour $n \leq N$, est caractérisé par deux informations:

1. Ses utilités intrinsèques: $u_n = (u_{n,x})_{x\in\mathcal{X}} \in \mathbb{R}^2$: pour $x \in \mathcal{X}$, $u_{n,x}$ représente l'utilité que le joueur n aurait de faire le choix x, *en dehors de toute influence sociale.*

2. Sa classe $k_n \in \mathcal{K} = \{1, ..., K\}$: La classe d'un individu peut représenter toute façon pertinente de grouper les individus dans une population (age, classe sociale, genre, orientation politque, etc). Cette séparation permet de modéliser et étudier l'influence sociale asymétrique. L'influence sociale multi-class permet notamment l'étude de phénomènes qualitatifs comme la répulsion de classe. Cela se produit quand deux classes d'individus différentes ne veulent pas agir ou penser de la même manière. Par exemple, si le choix est de soutenir ou non une opinion, il y a un phénomène de répulsion entre les individus de droite et de gauche: une personne de gauche est réticente à publiquement soutenir la même opinion qu'une personne de droite, et vice versa, même si les deux personnes sont intrinsèquement d'accord sur le sujet en question.

Un choix est un élément de $\mathcal{X} = \{0, 1\}$. Un profil de choix est un élément de $\mathcal{X}^N$ décrivant l'ensemble des configurations possible de choix dans la population de N joueurs. Par exemple, un profil de choix $\boldsymbol{x} = (x_n)_{n\in[\![1,N]\!]} \in \mathcal{X}^N$ signifie que chaque joueur n fait le choix x_n. Le reward perçu par le joueur n étant donné un profil de choix $\boldsymbol{x}$ est défini par

$$R_n(x_n, \boldsymbol{x}_{-n}) = u_{n,x_n} + \mathrm{u}(k_n, x_n, \frac{1}{N}\sum_{i=1}^{N} \delta_{k_i,x_i}),$$

où $\mathrm{u} : \mathcal{K} \times \mathcal{X} \times \mathcal{P}(\mathcal{K} \times \mathcal{X}) \to \mathbb{R}$ est appelée *fonction d'utilité sociale.* u_{n,x_n} est la partie du reward que le joueur n obtient du fait de choisir x_n *en soi*, et $\mathrm{u}(k_n, x_n, \frac{1}{N}\sum_{i=1}^{N} \delta_{k_i,x_i})$ est interprétée comme l'utilité *sociale* du choix x_n pour le joueur n étant donnée la distribution $\frac{1}{N}\sum_{i=1}^{N} \delta_{k_i,x_i}$ des paires (classe, choix) de la population. Nous étudions alors deux jeux basés sur ce framework:

1. Le jeu statique en information complète: les joueurs ne jouent qu'une fois, et connaissent chacun les données de toute la population,

2. Le jeu répété sans information initiale: les joueurs ne connaissent initialement presque rien les uns sur les autres mais jouent plusieurs fois au même jeu et peuvent observer les actions des jeux passés.

Les deux types de jeu sont étudiés via le concept d'Elimination Itérative de Stratégies Strictement Dominées (IESDS). Le résultat principal est la prédiction des choix rationnels de presque toute la population au moyen d'outils de champ moyen. Dans un second temps, nous utilisons ce résultats pour étudier des phénomènes qualitatifs d'influence sociale comme l'effet *boule de neige* et l'effet de *répulsion de classe*.

Le second travail de cette partie est un modèle de choix risqué paramétrant la théorie des prospects cumulés (CPT) de Kahneman et Tversky, permettant une calibration flexible et donnant une formule explicite pour l'évaluation d'un choix risqué donnant une récompense gaussienne. Dans chacun des deux travaux, nous discutons des applications commerciales et politiques.

Dans la troisième et dernière partie de ce manuscrit, nous développons deux études spécifiquement motivées par la publicité ciblée. Notre premier travail propose des algorithms d'apprentissage en ligne pour la prédiction de clics, prenant la forme d'un problème de classification binaire.

Un produit est associé à un *prix* $p \in \mathbb{R}$ et à des *caractéristiques* $\mathbf{f} \in \mathcal{F} := [0,1]^d$. Les caractéristiques d'un produit $\mathbf{f} \in \mathcal{F}$ peuvent correspondre par exemple à sa qualité, la réputation de la marque, sa forme, sa durée de vie, etc. Ainsi, un produit est caractérisé par un couple prix-caractéristiques $(p, \mathbf{f}) \in \mathbb{R} \times \mathcal{F}$. Dans la suite, nous identifions une publicité au produit dont elle fait la publicité, et nous dirons donc aussi bien "le produit $(p, \mathbf{f})$" que "la publicité $(p, \mathbf{f})$". Une intention de clic est représentée par une variable binaire $c \in \{-1, 1\}$, 1 signifiant "intention de cliquer", et -1 "pas d'intention de cliquer".

Le problème de prédiction de clic est alors défini comme suit. Considérons une suite aléatoire $(p_k, \mathbf{f}_k, c_k)_{k\in\mathbb{N}}$ de publicités $(p_k, \mathbf{f}_k)$ et d'intentions de clic c_k associées, pour tout $k \in \mathbb{N}$. Plus précisément, à chaque instant $k \in \mathbb{N}$, un nouveau produit est créé par une entreprise, avec un prix $p_k \in \mathbb{R}$ et des caractéristiques $\mathbf{f}_k = (f_{k,i})_{i\in[\![1,d]\!]} \in \mathcal{F}$, correspondant donc au produit/publicité $(p_k, \mathbf{f}_k)$. La variable binaire c_k représente *l'intention de clic* de l'individu pour ce produit, c'est-à-dire, c_k est la réponse à la question "si la publicité $(p_k, \mathbf{f}_k)$ était affichée à l'individu, cliquerait-il dessus?". Si oui, alors $c_k = 1$, sinon, $c_k = -1$.

Nous supposons qu'il existe une *fonction de récompense* $R_\star$ de forme polynomiale, avec degré borné, telle que $\forall k \in \mathbb{N}$,

$$p_k < R_\star(\mathbf{f}_k) - \varepsilon \Rightarrow c_k = 1, \quad \text{and} \quad p_k > R_\star(\mathbf{f}_k) + \varepsilon \Rightarrow c_k = -1,$$

où $\varepsilon \geq 0$ est une marge tenant compte des cas où aucun classifieur polynomial ne peut parfaitement séparer les données. Décrivons à présent les règles du problème de publicité ciblée. A chaque temps $k \in \mathbb{N}$ les étapes suivantes se produisent:

1. Nouvelle publicité: une nouvelle publicité $(p_k, \mathbf{f}_k)$ est créée.
2. Décision d'affichage: le publicitaire décide d'afficher ou non la publicté à l'individu.
3. Réaction de clic: si (et seulement si) la publicité à été affichée à l'individu à la précédente étape, sa réaction de clic (ou non) est observée par le publicitaire.
4. Mise à jour de la mémoire: Le publicitaire peut mettre à jour les variables stoquées en mémoire en fonctions des données observées durant les précédentes étapes, pour améliorer l'algorithme de décision d'affichage.

Les principales propriétés de l'algorithme développé dans ce travail sont les suivantes.

- Pas de faux négatifs: L'algorithme évite complètement les faux négatifs, i.e. ne pas afficher une publicité qui aurait conduit à un clic.
- Retour asymétrique: l'algorithme ne nécessite d'observer a posteriori que les intentions de click c_k de l'individu sur les publicités qui ont été affichées par l'algorithme.
- Faux positifs logarithmiques: Malgré la contrainte forte d'éviter tout faux négatif, l'algorithme fait seulement un nombre logarithmique $\mathcal{O}(\ln(k) + 1)$ de faux positifs après k publicités générées.

Le second travail de cette partie modélise et résout explicitement des problèmes de contrôle optimal pour les enchères de publicité ciblée. Nous étudions quatre modèles s'appliquant aussi bien à la publicité commerciale qu'au marketing social, et impliquant de la publicité ciblée, non-ciblée, ainsi que des interactions sociales. Nous détaillons ici le modèle le plus simple parmi les quatre étudiés dans ce travail. Introduisons progressivement les concepts.

L'élément clé est la notion d'information. Nous supposons qu'il y a une information qu'initialement personne, excepté l'agent (l'entreprise), ne connait. Cette information peut representer par exemple l'existence d'un nouveau produit vendu par l'agent. Dans la suite, nous appelons cette information l'Information I.

Nous modélisons à présent un Individu et son comportement. L'Individu est caractérisé par deux processus de Poisson indépendants $(N^{\mathbf{I}}, N^{\mathbf{T}})$ tels que:

- $N^{\mathbf{I}}$ est un processus de Poisson avec intensité $\eta_{\mathbf{I}}$, comptant les instants quand l'Individu se connecte à un site web contenant l'information I.

- $N^{\mathbf{T}}$ est un processus de Poisson avec intensité $\eta_{\mathbf{T}}$, comptant les instants où l'Individu se connecte à un site web ne contenant, a priori, pas l'Information I mais affichant de la publicité ciblée sur ses pages web, donc susceptible d'afficher l'Information I via une publicité ciblée.

Nous considérons à présent une famille de variables aléatoires réelles i.i.d. $(B_k^{\mathbf{T}})_{k\in\mathbb{N}}$. For $k \in \mathbb{N}$, $B_k^{\mathbf{T}}$ représente le prix au-dessus duquel l'enchère de l'agent doit être pour gagner la k-ème enchère de publicité ciblée. Nous modélisons donc ce prix de façon exogène (i.e. sans explicitement modéliser les autres enchérisseurs).

Le control d'enchère de l'Agent est supposé non-anticipatif, i.e. il ne dépend pas des événements futurs. Nous modélisons cela, en considérant la filtration $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$ telle que $\mathcal{F}_t = \sigma(N_s^{\mathbf{I}}, N_s^{\mathbf{T}}, B_{N_s^{\mathbf{T}}}^{\mathbf{T}}, s \leq t)$. Une stratégie d'enchère open-loop pour l'Agent est alors un processus aléatoire β à valeur réelle et progressivement mesurable par rapport à $(\mathcal{F}_{t-})_{t\in\mathbb{R}_+}$ (propriété non-anticipative), tel que β_t correspond à l'enchère que l'Agent ferait s'il y avait une opportunité d'afficher une publicité à l'Individu à l'instant t (i.e. si $\Delta N_t^{\mathbf{T}} = 1$).

Notons X^β le processus à valeur dans $\{0, 1\}$ tel que $X_t^\beta = 1$ si et seulement si l'Individu a obtenu l'Information avant l'instant t, supposant que la stratégie d'enchère open-loop de l'Agent est β. X^β est modélisé comme la solution au système dynamique

$$\begin{aligned} X_0^\beta &= 0 \\ dX_t^\beta &= (1 - X_{t-}^\beta)(\mathbf{1}_{\beta_t \geq B_{N_t^{\mathbf{T}}}^{\mathbf{T}}} dN_t^{\mathbf{T}} + dN_t^{\mathbf{I}}) \end{aligned}$$

Cette dynamique signifie que l'Individu commence non-informé. Il acquiert l'information aussitôt que 1) il se connecte à un site web affichant des publicités ciblées, et l'Agent gagne l'enchère (partie "$\mathbf{1}_{\beta_t \geq B_{N_t^{\mathbf{T}}}^{\mathbf{T}}} dN_t^{\mathbf{T}}$"), ou 2) il se connecte à un site web contenant l'Information I (partie "$dN_t^{\mathbf{I}}$"). Ensuite, il reste informé indéfiniment (partie "$(1 - X_{t-}^\beta)$").

Le gain moyen de l'Agent, étant donnée une stratégie d'enchère β, est

$$V(\beta) = \mathbb{E}\Big[\int_0^\infty e^{-\rho t}(K dX_t^\beta - \mathbf{1}_{\beta_t > B_{N_t^{\mathbf{T}}}^{\mathbf{T}}} \mathbf{c}(\beta_t, B_{N_t^{\mathbf{T}}}^{\mathbf{T}}) dN_t^{\mathbf{T}})\Big]$$

où $K \in \mathbb{R}$ représente le profit marginal fait par l'Agent quand l'Individu acquiert l'information, i.e. quand $\Delta X_t^\beta = 1$, et où $\mathbf{c} : \mathbb{R}^2 \to \mathbb{R}$ est une a fonction representant ce que l'Agent paiera s'il gagne l'enchère. La seconde partie correspond au prix payé quand l'enchère est gagnée par l'Agent: une enchère arrive quand $\Delta N_t^{\mathbf{T}} = 1$, et elle est gagnée si $\mathbf{1}_{\beta_t > B_{N_t^{\mathbf{T}}}^{\mathbf{T}}}$. Le prix payé est $\mathbf{c}(\beta_t, B_{N_t^{\mathbf{T}}}^{\mathbf{T}})$:

- Si on considère des enchères de type first-price, on a $\mathbf{c}(b, B) = b$, i.e. si l'Agent gange l'enchère, il paie son enchère β_t.
- Si on considère des enchères de type second-price, on a $\mathbf{c}(b, B) = B$, i.e. si l'Agent gagne l'enchère, il paie la seconde enchère le plus élevé de l'enchère, i.e. $B^{\mathbf{T}}_{N^{\mathbf{T}}_t}$.

Le but de l'Agent est alors d'utiliser une stratégie d'enchère $\beta^\star$ telle que $V(\beta^\star) = \sup_\beta V(\beta) =: V^\star$. Notre résultat principal est le suivant. La valeur optimale est donnée par

$$V^\star \;=\; \sup_{b\in\mathbb{R}} \frac{\eta^{\mathbf{I}} K + \eta^{\mathbf{T}} \mathbb{E}[(K - \mathbf{c}(b, B^{\mathbf{T}}_1))\mathbf{1}_{b\geq B^{\mathbf{T}}_1}]}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}} \mathbb{P}(b \geq B^{\mathbf{T}}_1)},$$

et tout contrôle d'enchère optimal $\beta^\star$ tel que $\beta^\star_t = (1 - X^{\beta^\star}_t)b^\star$, où

$$b^\star \in \operatorname*{argmax}_{b\in\mathbb{R}} \frac{\eta^{\mathbf{I}} K + \eta^{\mathbf{T}} \mathbb{E}[(K - \mathbf{c}(b, B^{\mathbf{T}}_1))\mathbf{1}_{b\geq B^{\mathbf{T}}_1}]}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}} \mathbb{P}(b \geq B^{\mathbf{T}}_1)},$$

est optimal. En d'autres termes, une politique d'enchères optimales est de faire l'enchère constante $b^\star$ tant que l'Individu n'est pas informé, et ensuite d'arrêter d'enchérir (ce qui est clairement optimal une fois que l'Individu est informé).

Le modèle que nous avons détaillé dans ce résumé est le modèle de publicité commerciale avec récompense basée sur l'achat. Nous avons aussi étudié trois autres modèles avec différentes applications: la publicité commerciale avec récompense basée sur une inscription, le marketing social avec discount rate, et un modèle plus riche de marketing social incluant, en plus de la publicité et les connections à des sites web contenant l'Information, 1) la publicité non-ciblée, et 2) les interactions sociales entre les individus de la population. Dans tous ces modèles, nous obtenons une formule fermée pour la valeur et la politique optimales.

Abstract

This thesis studies mathematical models for online social and economic problems, namely, models of large connected populations, social influence and interactions, risky choices, and targeted advertising.

The first part focuses on large population problems, with a more general and theoretical approach than the other parts. We study N-agent and McKean-Vlasov Markov Decision Processes with common noise and open-loop controls in infinite horizon. We obtain the Bellman equations for both problems, the reduction to stationary randomized feedback policies for the McKean-Vlasov MDP, as well as the propagation of chaos for the N-agent MDP, with convergence rates. We expose and circumvent measurability issues in the case of continuous state spaces in the presence of common noise, show that optimizing over randomized controls can yield a strictly greater gain than over non-randomized controls, and finally use propagation of chaos to approximately solve the N-agent MDP with toy model for advertising.

In part II, we study two economic behavioral models. Our first work analyses repeated game in a large population where players initially have no information about each others, via the Iterated Elimination of Strictly Dominated Strategies solution concept, and using mean-field estimating methods. We finally use our results to study the *snowball* and *class repulsion* social influence effects. Our second work is a choice under risk model, proposing a parametrization of Kahneman and Tversky's Cumulative Prospect theory, allowing for flexible calibration and yielding explicit formulas for the Certainty Equivalence in the case of gaussian prospects. In both works, we propose commercial and political applications.

In Part III, we make two studies specifically designed for targeted advertising. Our first study proposes an online click prediction learning algorithm for targeted advertising, for which we obtain a logarithm error efficiency, with bounded memory usage and computational complexity. Our second work is the modeling and explicit resolution of optimal control problems for targeted advertising auctions. We study four models, with applications to both commercial advertising and social marketing involving targeted advertising, non-targeted advertising, and social interactions.

Keywords: Large population, mean-field approximation, McKean-Vlasov, stochastic control, propagation of chaos, game theory, Mean-field games, social influence, social interactions, choice under risk, Cumulative Prospect theory, Prospect theory, targeted advertising, auctions, machine learning, online classification learning

Abstract

Cette thèse étudie des modèles mathématiques pour des problèmes socio-économiques en ligne, à savoir, des modèles de grande population en interaction, d'influence sociale, de choix risqué, et de publicité ciblée.

La première partie se concentre sur des problèmes de grande population, avec une approche plus générale et théorique que les autres parties. Nous étudions des processus de décision markoviens (MDP) à N-agent et de type McKean-Vlasov avec bruit commun et contrôles open-loop en horizion infini. Nous obtenons les équations de Bellman pour chaque problème, la réduction à des politiques stationnaires feedback randomisées pour le MDP de type McKean-Vlasov, ainsi que la propagation du chaos pour le MDP à N-agents, avec taux de convergence. Nous exposons et contournons des problèmes de mesurabilité dans le cas d'un espace d'état continu en présence de bruit commun, montrons qu'optimiser sur des contrôles randomisés peut donner un gain strictement plus grand que d'optimiser sur des contrôles non-randomisés, et utilisons finalement la propagation du chaos pour résoudre le MDP à N-agents pour des modèles jouets appliqués à la publicité.

Dans la partie II, nous étudions deux modèles de comportement économique. Notre premier travail analyse un jeu répété en grande population où les joueurs ne connaissent initialement rien les uns sur les autres, via le concept d'Elimination Itérative de Stratégies Strictement Dominées, et au moyen d'outils d'estimation de champ moyen. Nous utilisons finalement nos résultats pour étudier deux phénomènes d'influence sociale qualitatifs: l' effet *boule de neige* et l'effet de *répulsion de classe*. Notre second travail est un modèle de choix risqué, proposant une paramétrisation de la théorie des prospects cumulés de Kahneman et Tversky, permettant une calibration flexible, avec formule explicite de valorisation dans le cas de prospects gaussiens. Dans les deux travaux, nous proposons des applications politiques et commerciales.

Dans la partie III, nous faisons deux études spécifiquement dédiées à la publicité ciblée. Notre première étude propose un algorithme d'apprentissage en ligne pour la prédiction de clic sur des publicités, pour lequel nous obtenons une erreur de prédiction logarithmique, avec complexités mémoire et computationelle bornées. Notre second travail est la modélisation et résolution explicite de problèmes de contrôle optimal pour les enchères de publicité ciblée. Nous étudions quatre modèles, avec applications aussi bien à la publicité commerciale et qu'au marketing social, impliquant de la publicité ciblée, non ciblée, et des interactions sociales.

Mots-clé: Grande population, approximation de champ moyen, McKean-Vlasov, contrôle stochastique, propagation du chaos, théorie des jeux, jeux à champ moyen, influence sociale, interactions sociales, choix risqués, théorie des prospects cumulés, théorie des prospects, publicité ciblée, enchères, apprentissage statistique, apprentissage en ligne, classification

Contents

Chapter 1

Introduction

1.1 General motivations of the thesis

This thesis studies mathematical models addressing thematics that are particularly important on Internet, namely, large populations, social influence, social games, risky choices, interactions and targeted advertising. This work is part of the growing scientific interest towards modern *online* sociologic and economic matters. The interest of the mathematical community toward these problematics is much younger than its classical interest toward physics, and there is still a lot of room for exploration in that field, especially since the sociologic and economic world is evolving so fast. The creation of Internet led to several changes of paradigm regarding the way people interact (social networks), act (internet navigation, clicking and buying decisions), and regarding the way businesses work (free websites living from targeted advertising).

Internet and more specifically social networks rely on three basic elements: Content, social interactions, and advertising. Providing contents and facilitating social interactions are the mains services they provide. On the other hand, advertising is the way the vast majority of websites make profit and are viable economic systems.

These three aspects are well suited for mathematical analysis:

1. The "content" component is strongly related to choice under risk theories: indeed, each time an individual decides whether or not to access or pay for a content, he is making a risky choice based on partial information.

2. The "social interactions" component is linked to several mathematical theories, like mean-field theories modeling large population behaviors, game theory modeling rational choices when one's reward depends upon other people's choices (social

influence), and information spreading models, studying how a given information can spread in a population via social interactions.

3. The "advertising" component is a strategic element for marketing, and it is thus naturally well modeled by optimal control theory (model-based approach) and machine learning algorithms (model-free approach).

The aim of this thesis is to provide mathematical tools contributing to a better understanding of Internet and social network problematics, from a both theoretical and practical viewpoint. Our goal is twofold: from the mathematical's viewpoint, to illustrate how Internet and social networks can naturally lead to interesting and diverse mathematics, and from the applications' viewpoint, to provide potentially useful mathematical tools for Internet problems. Depending upon the model we study, the solutions will take the form of correspondences, formulas, explicit formulas, explicit optimal policies, and learning algorithms.

1.2 Background of related mathematical topics

In this section, we provide succinct overviews of the various research fields and theories related to the present work. We organize these non-exhaustive overviews in parts and paragraphs roughly following the thesis' organization. We will not provide full and detailed overviews of each research field, instead focusing on the aspects relevant to understand its contributions.

1.2.1 Large populations models and mean-field approximation

Part I of the thesis focuses on the control of large populations under mean-field interactions, with natural motivations to advertising. Large populations have been studied in several ways. An important separation in the literature comes from the way that individuals are assumed to be connected to each other. One approach is to consider that even when we make number of individuals N go to infinity, the number of neighbors of each individual stays the same, i.e. does not scale with N. Another currently very popular approach is to scale the number of neighbors of an individual with N, with the extreme case where each individual is connected to everyone. Our study of controlled large populations adopts the latter approach, which most naturally allow to use the mean-field approximation principle.

Mean-field theory

The mean-field approximation is a principle used in the mean field theory, taking its source in the statistical physics community, back to P. Weiss' work in the 1900's [94] and earlier in the work of P. Curie. The motivation of mean-field theory is to simplify the study of large particle systems. The starting point is a formal manipulation consisting in replacing empirical distributions by theoretical ones in the transition's dynamic of the system. This formal manipulation turns the dynamic system into another one, where the particles are independent, and, under symmetry assumptions, identically distributed. The analysis of such *mean-field* system can thus be reduced to the study of a single particle, called the *representative* particle.

The belief behind this manipulation is twofold:

1. the *mean-field* system is expected to be more tractable than the original large system,

2. its study is expected to provide insightful information on the original large system.

The origin of this belief is empirical. Indeed, apart from some exceptions like the one dimensional nearest neighbor Ising model, wrongly predicted by a mean-field approximation, and the mean-field system of the Sherrington-Kirkpatrick model of spin glasses [79] has raised many challenges, only recently seeing significant progress ([3, 14, 35, 36, 82]), this belief is empirically satisfied in a wide majority of cases, as in the ferromagnetic Ising and Potts models [46, 95].

These empirical evidences motivated mathematical rigorous justifications, and point 2 of the belief was justified for large classes of systems by means of law of large number arguments. The convergence of the large system to the mean-field system was termed *propagation of chaos*. There is now a large body of research investigating the links between mean field and actual systems. The mathematical formulation of mean field theory has found several applications outside of statistical physics, in particular in game theory and in optimal control theory, via the recent theories of Mean-field games and of McKean-Vlasov optimal control, see, for more details about these developments, the two-volumes book of Rene Carmona and François Delarue [15].

Optimal control and McKean-Vlasov optimal control

In part I, the large population models that we study are optimal control problems. Optimal control is a fundamental tool with important applications in the industry, as it is the main theory proposing strategies to maximize an agent's profit. Recalling that the goal of this thesis is to study mathematical models for Internet problematics, optimal

control may be the most appealing tool for modeling online advertising problems, the other natural tool being machine learning algorithms, some of them being theoretically based on optimal control as well (e.g. reinforcement learning). Optimal control is also a vast theory, with several branches, namely, deterministic optimal control and stochastic optimal control, adding randomness to the controlled dynamic. When considered in a discrete time framework, a stochastic optimal control problem is called a Markov Decision Process.

Recently, the popularization of mean-field theory outside of the world of statistical physics spread to the optimal and stochastic control communities, leading to a modern theory called McKean-Vlasov optimal control theory. The idea of McKean-Vlasov optimal control is to consider a problem involving the control of a large population to maximize a profit, when the population is subject to mean-field interactions. As always in mean-field theories, the underlying motivation is to solve the a priori more complex optimal control problem on N interacting individuals. Notice that such problem seems particularly relevant for advertising applications.

A large literature has already emerged on continuous-time models for the optimal control of McKean-Vlasov dynamics, and dynamic programming principle (in other words time consistency) has been established for these types of problems in [53], [72], [7], [24]. Propagation of chaos also has been established in several frameworks, and we refer to [49], which was the first paper to rigorously connect mean-field control to large systems of controlled processes, see also the recent paper [29] and [23]. We refer to the books [8], [15] for an overview of the subject.

Compared to continuous-time models, discrete-time McKean-Vlasov control problems have been less studied in the literature, but there is a growing interest for this framework, see [71] [16], and [33] for applications to the context of reinforcement learning.

1.2.2 Behavioral economics models: games and risky choices

The second part of this thesis is devoted to economic behavioral models, the first one being a large population game and the second one a choice under risk model. Such theories indeed are natural candidates to understand online behaviors: clicking on any web link is a risky choice as the reward generated by its content is not fully known, and interactions with other people on social networks are susceptible to introduce dependencies of one's reward in other people's choices (e.g. via trend phenomenons).

Whether they are games or choice under risk theories, economic behavioral theories traditionally aim to understand and predict people's choices, generally based on an assumption of rationality, i.e. on the basic principle that people's choices generally aim to maximize their happiness.

Game theory with large populations

Game theory aims to predict people's behavior in a situation where their respective rewards depend upon each other's choices.

Game theory started with the study of 2-player games, as in a letter written by James Waldegrave in 1713 and Antoine Cournot's duopoly in 1838, before game theory was formalized.

The study of N-player games, present in the work of Von Neumann and Morgenstern ([91]), followed by the developments brought by Nash ([68, 43, 66, 67]) and later Aumann ([5]), naturally led to observe a phenomenon that had already long been known in statistical physics: They noticed that when studying N-player games, some predictions for the N-player game often "converged" to asymptotic predictions when N was sent to infinity. Robert Aumann later published a seminal paper on games with infinitely many players (see [5]), initiating a long list of studies of games with continuum of players, see for instance the large games literature ([47, 41, 13, 42]), and the prolific literature in Mean-field games theory, a theory initiated 15 years ago simultaneously in the engineering community by Peter Caines, Minyi Huang and Roland Malhamé [38, 37], and in the mathematical community by Pierre Louis Lions and Jean Michel Lasry [50, 51, 52, 34], aiming to understand the limiting behavior, for N large, of N-player differential games under symmetry assumption.

Besides the well-known Nash-equilibrium concept introduced by John Nash ([43]), other concepts have been studied in game theory. A concept that particularly interests us in this thesis is the concept of Iterated Elimination of Dominated Strategies (IESDS). This concept is for instance studied in [64]. More precisely, in [64], Milgrom defines the concept of *serially undominated strategies*, which are simply all the remaining strategies after performing an IESDS. The general idea of the IESDS is the following. We define an iterative mechanism consisting in progressively eliminating strategies from the set of all possible strategies. At each iteration, there is thus a set of *non-eliminated strategies*, i.e. strategies that have not been eliminated *yet*. Each iteration consists in eliminating the strategies that are strictly dominated over all the *currently non-eliminated strategies*. The idea is that, assuming that each player is intelligent enough and knows that the other players are intelligent as well, he assumes that no player would play a strictly dominated strategy. He can thus eliminate dominated strategies of the set of strategies because they will not be played. He knows that all the players will also eliminate these strategies. Then, on the sub-game restricted to the non-eliminated strategies, some strategy might now, in turn, be strictly dominated. All players will then eliminate them as well. Then, Milgrom calls *serially undominated strategies* the set of strategies that

are never eliminated after any number of iterations of the above mechanism.

For large population games, the IESDS concept has not been studied a lot. See however Dufwenberg and Stegeman [26] and Chen, Long, Luo [17] for studies of the IESDS concept for general games, with potentially infinitely many players and strategies.

Risky choices and Cumulative prospect theory

The motivation of Cumulative Prospect theory is to propose a more realistic alternative to Expected Utility Theory (EUT) to model human behaviors when facing risky choices. EUT was initially proposed by Daniel Bernoulli as a response to what should be the reasonable maximal price to pay to enter a gamble. At the time, the natural assumption was that it should be the expectation of the gamble's reward. However, Bernoulli convincingly argued (see St. Petersburg game) that it could not truthfully model people's choices, and introduced the concept of *utility* function: the value associated to a gamble R (i.e. a choice with random reward) was not $\mathbb{E}[R]$ anymore but $V_{EUT}(R) = \mathbb{E}[v(R)]$, where v is the utility function. Therefore, when facing a family of gambles $\mathcal{G}$ (i.e. real random variables), an individual would choose the gamble

$$\operatorname*{argmax}_{R\in\mathcal{G}} V_{EUT}(R) = \operatorname*{argmax}_{R\in\mathcal{G}} \mathbb{E}[v(R)].$$

This was the first formulation of EUT. Later, John Von Neumann ([91]) proved that EUT is implied by a set of very compelling *axioms of rationality*, bringing a lot of credibility to EUT. However, empirical studies soon revealed that people's behaviors were consistently violating EUT.

A promising alternative was Kahneman and Tversky's Nobel Prize awarded Prospect theory ([40]). Later, John Quiggin, inspired by Prospect theory, developed the rank-dependent expected utility theory ([74]), which in turn led Kahneman and Tversky to improve their own theory by developing the Kahneman and Tversky's Cumulative Prospect theory ([40]).

Since then, rank-dependent utility and Cumulative Prospect theories have sparked a great interest outside of the economic behavioral community, e.g. in finance, see [10, 96].

1.2.3 Targeted advertising: learning algorithms and optimal control

In the last part of the thesis, we focus on models specifically designed for targeted advertising strategies. There are two natural mathematical tools to design strategies for optimizing a gain or reaching a satisfying one: learning methods, coming from machine learning and statistical learning, and optimal control.

Classification learning

Probably the most studied problem in machine learning are classification problem. Classification problems are all based on a common framework: there is an input space X and an output, or label, space Y. In our case, X can encode every aspect of an ad, and Y can correspond to the binary clicking decision that a given individual would make for this ad. The goal essentially is to find a map $f : X \to Y$ making a small amount of prediction errors. The map $f : X \to Y$ is then referred to as a classifier. Classification is a type of *supervised learning*, because the labels from past observed data are observed, and the algorithm does not have to "create" labels.

In classical "offline" learning, one assumes that we have access to training data $(X_1, Y_1), ..., (X_n, Y_n) \in X \times Y$, and a classification algorithm is then a procedure receiving the training data and outputting a classifier f. Another branch of classification learning, called *online classification learning*, refers to the situation where no training data is initially accessible, and where inputs $(X_t, Y_t)_{t\in\mathbb{N}}$ comes as time goes by. An online classification algorithm builds a sequence $(f_t)_{t\in\mathbb{N}}$ of classifiers such that f_t corresponds to the update of the classifier before time t. At each time t, two actions are taken: 1) a prediction of the output of X_t using classifier f_t, and 2) an update of the classifier to f_{t+1} for future predictions, taking into account the data received at time t.

Several algorithms have been developed for this task, the most used ones being Vapnik and Chervonenkis's Support Vector Machine algorithm presented in the seminal paper [12], and Logistic Regression, invented by Berkson [9] and Cox [20].

For more details about classification learning, we refer to the many textbooks, surveys, and monographs on these topics: [4], [57], [22], [25], [30], [45], [48], [60], [62], [63], [69], and [85, 86, 87].

Optimal control for targeted advertising, auctions, and social interactions

The other important theoretical tool for situations with a strategic component allowing to increase one's profit is the optimal control theory. An important application of optimal control is advertising.

Optimal control for advertising. Several approaches have been proposed in the past to model advertising problems: mathematical programming, dynamic programming, simulation, and heuristic procedures ([56, 98]). An important addition is optimal control theory. In this approach, a dynamical system is modeled with controlled differential equations and optimized by means of the maximum principle [42]. We mention the important Nerlove-Arrow ([70]), and Vidale-Wolfe ([89]) models, and for an overview of this research field, see [78] and its sequel [27].

The main particularity of the optimal control advertising models studied in this thesis is that they are *population models*, therefore modeling each individual and their behaviors in a population of N people, while the existing literature about optimal control for advertising focuses on differential models, considering, from the start, controlled differential equations directly modeling the dynamics of sales as a continuous process affected by an advertising expenditures process.

Auction theory. Auctions are an inevitable component of today's advertising, and particularly of targeted advertising. Indeed, ad emplacements allocation are made via *targeted advertising auctions*. Each time an individual connects to a website displaying targeted ads, several agents (companies, influencers, etc) compete in an auction for the *ad emplacement*. Each agent makes a bid, the winner pays the price resulting from the auction's rule, and his ad is displayed to the individual.

The long history of auctions, and their omnipresence on the Internet, illustrate the crucial importance of auction theory. As a sub-field of game theory, auctions have been widely studied by game theorists such as John Nash ([66]), William S. Vickrey ([88]), and the 2020 Nobel prize in economics winners Milgrom and Wilson, for their contributions to auction theory.

Information spreading. Finally, advertising also relates to information spreading, as the goal of an advertiser is to efficiently spread an information. Information spreading is a sub-field of population dynamics theory. This research field is very vast, often termed as opinion dynamics theory, as it is one of its main applications. Models goes from Bayesian to non Bayesian models and from games to evolutionary dynamics. They apply to information propagation, opinions formation, and choices dynamics. For a detailed overview of populations dynamics, opinions dynamics, and learning in social networks, see [2].

1.3 Contributions of the thesis

Last section illustrates that studying social networks and Internet problematics from a mathematical point of view is a widely interdisciplinary task and can be done from the viewpoint of several mathematical disciplines. Social networks reunites them in a common playground, where large populations, risky choices, social games, social interactions and targeted advertising cohabite.

In this thesis, we studied models from each of these points of view and sometimes mixing them. To make the presentation as clear as possible, we choose to group our

works into three parts, each containing two chapters, from the most theoretical to the most applied works, and separated by main thematics:

1. We study in Part I large population models. Our study is part of the modern wide interest in mean-field models for populations. Although works outside of this part will also involve populations, the studies of Part I specifically study phenomenons related to large populations with mean-field interactions. The main goal of these studies is to provide methods and tools to deal with large population problems in general (propagation of chaos, mean-field Bellman equation, problem lifting methods, etc), and to expose and resolve theoretical challenges in these types of problem. Part I is the most theoretical part of the thesis, as it deals with a general large population framework.

2. In part II, we study economic behavioral models, that is, how to model human behaviors when facing choices in situations involving two types of unknowns: games (involving other players choices), and choices under risk (involving randomness). Although still theoretical in the sense the works in this part theorize human behaviors, this part is more applied than part I, and we shall discuss some commercial and political applications in each study.

3. In Part III, we focus on very concrete problematics from targeted advertising. While this is the most applied part, the studies therein also have interesting theoretical aspects, as part of the optimal control and online learning research fields.

1.3.1 Large population models with mean-field interactions

In Part I, we study a theoretical model of large population in a work separated in two chapters. This part is based on the paper [65], with improved results of propagation of chaos. The initial common framework involves two models. In one model, we consider a controlled population with N individuals, and in the other model, we consider a controlled single individual (called *representative individual*). In both models, the controller aims to maximize a profit depending upon the models dynamic and his gain function. We start with a probabilistic universe Ω on which are defined the following random variables:

- Idiosyncratic noises $(\varepsilon_t^i)_{t\in\mathbb{N},i\in\mathbb{N}}$, such that ε^i will, for $i \in \mathbb{N}$, represent the idiosyncratic noise of an individual i in the studied models,

- Common noise $(\varepsilon_t^0)_{t\in\mathbb{N},i\in\mathbb{N}}$, representing a noise affecting all the individuals,

- Randomization variables $(U_t^i)_{t\in\mathbb{N},i\in\mathbb{N}}$, such that U^i serves to randomize actions on individual i.

We consider a state space $\mathcal{X}$ and an action space A.

Given this basic setup, we consider the two control problems, formally similar:

1. **The N-individual control problem:** In this model, an open-loop control is a random process α valued in A^N and adapted to $\mathcal{F}_t^N := \sigma((\varepsilon_s^i)_{i\in[\![1,N]\!]}, \varepsilon_s^0, s \leq t)$. For $t \in \mathbb{N}$ and $i \in [\![1, N]\!]$, α_t^i represents the action to send to individual i at time t (to keep an intuitive image of the model, let us assume that the action represents sending a targeted ad α_t^i to individual i, although it could be a marketing offer or any related type of action). The fact that α is adapted to $\mathcal{F}_t^N$ simply means that any action to any individual can depend upon all past data (past common noise and past idiosyncratic noise of all the individuals). The set of open-loop controls is denoted $\Pi_{OL,N}$. The dynamic of the population with control α is defined by

$$\begin{aligned} X_0^{i,N,\alpha} &= x_i, \quad i \in [\![1, N]\!] \\ X_{t+1}^{i,N,\alpha} &= F\Big(X_t^{i,N,\alpha}, \alpha_t^i, \frac{1}{N}\sum_{n=1}^N \delta_{(X_t^{n,N,\alpha},\alpha_t^n)}, \varepsilon_{t+1}^i, \varepsilon_{t+1}^0\Big), \quad i \in [\![1, N]\!], t \in \mathbb{N}. \end{aligned}$$

where F is a measurable function. This is a general dynamic, F does not necessarily have to depend upon all these objects, but in this theoretical work, our goal is to keep a neutral and general framework. An importent element is the dependence in $\frac{1}{N}\sum_{n=1}^N \delta_{(X_t^n,\alpha_t^n)}$, which encodes a mean-field interaction with the rest of the population, which is a key aspect of the model. The state $X_t^{i,N,\alpha}$ corresponds to the state of individual i at time t, in a dynamic where the control α was used to influence the population. As α was interpreted as sending targeted ads to each individual, X_t^i can be interpreted, for instance, as the company to which individual i is client at time t. The expected gain of the external agent, in this model, with control α, takes the form

$$V_N(\alpha) = \mathbb{E}[\frac{1}{N}\sum_{i=1}^N \sum_{t=0}^{+\infty} \beta^t f(X_t^{i,N,\alpha}, \alpha_t^i, \frac{1}{N}\sum_{n=1}^N \delta_{(X_t^{n,N,\alpha},\alpha_t^n)})].$$

As for F, the function f is a general reward function and does not have to depend upon all its parameters. The main idea is that it represents the reward of the external agent generated by individual i at time t. If the agent is a company, it is clear that its reward coming from individual i at time t depends upon his state X_t^i (saying whether or not he is client of the company) and the action α_t^i (did the

company paid to send him a targeted ad?). The dependence in $\frac{1}{N}\sum_{n=1}^{N}\delta_{(X_t^{n,N,\alpha},\alpha_t^n)}$ is an additional degree of freedom, of particular theoretical interest since it is another place where mean-field aspects can play a role. The goal is then to study $V_N(\alpha)$ and in particular the optimal expected gain $V_N := \sup_{\alpha\in\Pi_{OL,N}} V_N(\alpha)$ as well as providing tools to compute it and design controls α_ϵ leading to ϵ-optimal expected gains, for $\epsilon \geq 0$.

2. **The McKean-Vlasov control problem:** In this model, we focus on a single individual, e.g. individual 1, and an open-loop control is a random process α valued in A and adapted to $\mathcal{F}_t := \sigma(\varepsilon_s^1, \varepsilon_s^0, s \leq t)$, that is, only depending upon this single individual's past data and the past common noise, and only representing the actions made for this single individual. The set of open-loop controls is denoted by Π_{OL}. Given a control α, we define the following dynamic:

$$\begin{aligned} X_0^\alpha &= \xi, \\ X_{t+1}^\alpha &= F(X_t^\alpha, \alpha_t, \mathbb{P}^0_{(X_t^\alpha,\alpha_t)}, \varepsilon_{t+1}^1, \varepsilon_{t+1}^0), \quad t \in \mathbb{N}. \end{aligned}$$

where $\mathbb{P}^0_{(X_t^\alpha,\alpha_t)}$ denotes the probability distribution of (X_t^α, α_t) conditionally to the past common noise. The gain of the external agent, in this problem, is

$$V(\alpha) = \mathbb{E}[\sum_{t=0}^{\infty} \beta^t f(X_t^{i\alpha}, \alpha_t, \mathbb{P}^0_{(X_t^\alpha,\alpha_t)})].$$

Notice the formal similarity of this problem with the N-individual control problem. Essentially, the empirical distributions $\frac{1}{N}\sum_{n=1}^{N}\delta_{(X_t^{n,N,\alpha},\alpha_t^n)}$ were replaced by the theoretical conditional probability distributions $\mathbb{P}^0_{(X_t^\alpha,\alpha_t)}$. The reason to make such substitution simply results from our knowledge that in many probabilistic situations, empirical distributions tend to be close to theoretical distributions. This knowledge starts from the strong law of large numbers, its various extensions, eventually turning into the phenomenon of propagation of chaos in uncontrolled stochastic dynamic systems. We now know that replacing empirical distributions by theoretical distributions often leads to two models that are close to each others in several ways, and closer and closer as the population size N increases.

Given these kind of framework, usual mathematical studies involve:

1. Studying the McKean-Vlasov problem,
2. studying the proximity of the N-agent problem with the McKean-Vlasov problem.

There is already a vast literature on this subject, studying various models (uncontrolled, controlled, mean-field games, etc). From a theoretical viewpoint, here are the key challenges and contributions of our study:

1. **Possible finite spaces and regularity in expectation:** our models allow $\mathcal{X}$ and A to be finite. Although in general, finite spaces are simpler than continuous spaces, this is not the case when it comes to chaos propagation results. Propagation of chaos usually requires regularity in the state transition function that cannot reasonably be assumed when the state space is discrete, essentially because F's domain contains a continuous space $\mathcal{P}(\mathcal{X})$ and has a finite codomain $\mathcal{X}$, which prevents continuity unless F is constant in is $\mathcal{P}(\mathcal{X})$ coordinate, which is equivalent to have no mean-field interactions. However, a weaker form of regularity can reasonably be assumed even with finite spaces: regularity in expectation w.r.t. the idiosyncratic noise coordinate. We shall use this weak regularity assumption to obtain propagation of chaos.

2. **Measurability issues due to common noise and continuous state space:** We shall see that the simultaneous presence of common noise and a continuous state space poses measurability issues when attempting to use standard methods working in the no common noise or finite state space setup. Theses issues are avoided by means of a more flexible lifting procedure.

3. **Necessity of randomization:** We show that in our discrete time framework, randomization is in general necessary to build ε-optimal controls. This result contrasts with the usual result of Markov Decision Processes that one can restrict to feedback (non randomized) controls without reducing the optimal gain.

4. **Measurable objects on measure spaces:** we introduce measurable functions on measure spaces: a measurable coupling function, and a measurable projection, useful to build more complex functions on measure spaces in a mesurable way.

The study is split in two chapters:

1. In Chapter 1, we study the McKean-Vlasov control problem, establish the fixed point Bellman equation for V: for all $\mu \in \mathcal{P}(\mathcal{X})$,

$$V(\mu) = \sup_{\mathbf{a}:\mathcal{X}\times[0,1]\to A} \mathbb{E}[f(\xi, a(\xi, U_1), \mathbb{P}^0_{\xi,a(\xi,U_1)}) + \beta V(\mathbb{P}^0_{F(\xi,a(\xi,U_1),\mathbb{P}^0_{\xi,a(\xi,U_1)},\varepsilon^1_1,\varepsilon^0_1)})].$$

   Furthermore, we prove a verification result, and that reduction to stationary randomized feedback policies is possible, and *necessary* in the sense that reducing to

non-randomized feedback policies is susceptible to strictly decrease the optimal gain.

2. In Chapter 2, we prove results of propagation of chaos of the N-individual control problem to the McKean-Vlasov problem, with rates of convergence. More precisely, with an explicit sequence $(M_N)_{N\in\mathbb{N}}$ coming from the convergence rate of empirical measures toward theoretical measures in the Wasserstein sense, and with en explicit $\gamma \leq 1$:

   - We prove the convergence of value functions: $\|V_N - V\| = \mathcal{O}(M_N^\gamma)$. This result can be seen as a discrete time version of the result in [23] for diffusions, as it links the N-individual and McKean-Vlasov MDPs when the optimization is performed on open-loop controls. Compared to our published work [65], which only linked the McKean-Vlasov MDP to the N-individual MDP over *individualized* open-loop controls, we were able to improve our result by linking the McKean-Vlasov MDP with the N-individual MDP over *fully* open-loop controls.
   - We prove that any ϵ-optimal stationary randomized feedback policy for the McKean-Vlasov is a $(\epsilon + \mathcal{O}(M_N^\gamma))$-optimal stationary randomized feedback policy for the N-individual MDP.
   - Conversely, we provide a simple way to turn any ϵ-optimal stationary feedback policy for the N-individual MDP into a $(\epsilon + \mathcal{O}(M_N^\gamma))$-optimal stationary randomized feedback policy for the McKean-Vlasov MDP.

1.3.2 Games and economic behavioral models

In Part II, we turn to studies that are much more applied than Part I in their motivations, and yet still theoretical in that they aim to theorize human behavior. Our study contributes to the effort of providing mathematical models to economics and in particular behavioral economics. It contains two chapters:

1. Chapter 3 fits into game theory (study of choice when the reward depends upon other people's choices). More precisely, we study a large population game with social rewards, via the concept of Iterated Elimination of Strictly Dominated Strategies, and by means of mean-field approximation tools.

2. Chapter 4 fits into decision under risk theory (study of choice when the rewards are random). More precisely, our work provides a parametrization of the well-established Cumulative Prospect theory in which gambles certainty equivalence

can be explicitly computed when rewards are gaussian, and we discussed important applications to targeted advertising.

Although game theory and choice under risk are different theories, they are probably the most successful mathematical theories so far to study human behaviors in complex choice situations. The fact that they each tackle a different kind of complexity (presence of other players, and randomness) make them greatly complementary, as evidenced by their joint presence in the groundbreaking book *Games and Economic Behaviors*, by John Von Neumann.

Iterated Elimination of Strictly Dominate Strategies in large population games

In this study, we consider a binary space $\mathcal{X} = \{0, 1\}$, representing two possible choices. These choices can represent many things: buying or not a product, subscribing or not to a service, publicly supporting or not an opinion, etc. We consider a population with N players. Each player n, for $n \leq N$, is characterized by two pieces of information:

1. **His intrinsic utilities** $u_n = (u_{n,x})_{x \in \mathcal{X}} \in \mathbb{R}^2$: for $x \in \mathcal{X}$, $u_{n,x}$ represents the utility that player n has for choice x, *outside of all social influence and interactions*. For instance, if the choice is to buy (choice 1) or not (choice 0) a product, one could consider that $u_{n,1}$ represents how much player n will like the product in itself, $u_{n,0} = 0$ could represent the null utility of not buying it. If the choice is to publicly support an opinion (choice 1) or reject it (choice 0), $u_{n,1}$ represents the intrinsic happiness of player n to support this opinion, and $u_{n,0}$ to reject it.

2. **His class** $k_n \in \mathcal{K} = \{1, ..., K\}$**:** The class of an individual can represent any relevant way to group individuals in a population (age, social class, gender, political orientation, etc). This separation allows to study asymmetric social influence. Multi-class social influence makes it possible to study qualitative phenomenons like class repulsion. This happens when two different classes of individuals don't want to act or think the same way. Let us provide two natural examples. In the example of supporting or not an opinion, there is a repulsion between the left-wing and right-wing politically oriented classes: a politically left-wing person is reluctant to support the same opinion as a politically right-wing person, even if they *intrinsically* agree on it. In the example of buying or not a product, a well known example is the case of diet coke. It is known that one of the main reasons why Coca-cola commercialized coke zero is that males were reluctant to buy diet coke because they associated it to a female product. Coke zero was designed to have a less female connotation. Thus, the simple fact that diet coke was seen as a female product was enough to dissuade some men from buying it.

For each $n \leq N$, we then denote by $d_n = (k_n, u_n)$ the *data* of player n. Let us now describe the unrolling of the game. First of all, each player n makes a choice $x_n \in \mathcal{X}$ (buying or not the product, publicly supporting or not the opinion). This simple choice guarantees each player n to receive the associated intrinsic utility u_{n,x_n}. Then, players socially interact with each others, e.g. on a social network, and this way observe the choices made by other people. From these observations, player n perceives a *social reward* given by $\mathrm{u}(k_n, x_n, \frac{1}{N}\sum_{i=1}^{N} \delta_{k_i,x_i})$ (where $\mathrm{u} : \mathcal{K} \times \mathcal{X} \times \mathcal{P}(\mathcal{K} \times \mathcal{X})$ is a measurable function), meaning that his social reward depends upon his class, choice, and the distribution of class and choices in the population. Therefore, player n's overall utility for this game is defined, for all $(d_n)_{n\leq N} \in (\mathcal{K} \times \mathbb{R}^2)^N$ and $(x_n)_{n\leq N} \in \mathcal{X}^N$, by

$$R(d_n, x_n, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}) = u_{n,x_n} + \mathrm{u}(k_n, x_n, \frac{1}{N}\sum_{i=1}^{N} \delta_{k_i,x_i}).$$

Our study consists in investigating two games based on the above core framework but differing in the following aspects:

- The first game is a *static* game, and players are assumed to know the population's data $\boldsymbol{d}$, i.e. the data d_n of any player n in the population.
- The second game is a *repeated* game, and players are assumed to initially have no information about each others.

Both games are studied via the Iterated Elimination of Strictly Dominated Strategies (IESDS) solution concept. The reason why we study the IESDS is that it provides a way to describe people's rationality that is more convincing than the Nash-equilibrium solution concept. Let us briefly explain why.

The Nash-equilibrium concept essentially claims that players would not play a strategy profile $\boldsymbol{x} = (x_n)_{n\in[\![1,N]\!]}$ if for some $n \in [\![1, N]\!]$, player n could have a strictly better response x'_n to the strategies $\boldsymbol{x}_{-n}$ of the other players, thus only leaving, by definition, Nash-equilibrias as potential strategy profiles.

There are three natural ways to understand such concept:

1. **Empirically:** it has been observed, in many cases, that people play a (close to) Nash-equilibrium. This can in itself justify to study it.
2. **With the implicit mechanism argument:** In this interpretation, we say that each player uses an internal mechanism to determine what strategy to play, and, without specifying it, we assume that this mechanism involves the computation

of best responses in an iterative way. It thus seems natural to assume that such mechanism converges to a fix point of the best response functions i.e. a Nash-equilibrium.

3. **As an intrinsically rational decision criterion:** In this argument, the strategy x_n of player n is justified by the fact that, knowing that the others play $\boldsymbol{x}_{-n}$, he should rationally play his best response x_n. However, to justify that the others play $\boldsymbol{x}_{-n}$, the same argument requires to know that player n play x_n. In other words, this way of justifying that players will play a Nash-equilibrium is logically *circular*, and therefore does not lead to an actual logical *deductive* argument, starting from an obviously true assertion and logically deducting what players will play.

The downside of the empirical approach is that it does not give an explanation of *why* people play a Nash-equilibrium. The downside of the implicit argument mechanism is that it assumes the existence of such mechanism without describing it precisely, which, again, does not really rigorously explain *why* they play a Nash-equilibrium, and finally, the downside of the rational decision criterion approach is that it is logically flawed because of its circularity.

Therefore, the Nash-equilibrium is a good concept from a descriptive and empirically predictive standpoint, but not from a logical and explanatory point of view, which can be disappointing given that the players are supposed to be rational and logical.

The IESDS solution concept is, on the other hand, a logical and deductive iterative mechanism, consisting in starting from all possible strategy profiles, and, at each iteration, removing the strategy profiles containing strictly dominated strategies, simply encoding the idea that 1) no player would play a dominated strategy, and 2) all the players know this fact, and thus, all the players can simply dismiss all the strategy profiles with a dominated strategy.

The main result of Chapter 3 is that, using the IESDS solution concept, we are able to predict with precision the rational choices of most players in the population. Furthermore, this prediction will be obtained by means of mean-field methods. The second result is that we will be able to use these predicted choices to study qualitative social phenomenons like the *snowball effect* and the *class repulsion effect*.

Gaussian Cumulative Prospect Theory

In Chapter 4, we propose a parametrization of Cumulative Prospect Theory, yielding an explicit gamble valuation formula for gaussian prospects. Cumulative Prospect Theory

models how, given a set of *prospects*, that is, a set of gambles $\mathcal{G}$ represented by random variables, an individual attributes a subjective value $V(R)$ to each gamble $R \in \mathcal{G}$, to choose the gamble with highest gamble valuation. Such model is interesting to predict an individual's choices, and thus very useful for targeted advertising.

Cumulative Prospect Theory defines the gamble valuation assigned by an individual to a given prospect as follows:

1. **Functions with constraints:** For this individual, there exists three functions $v : \mathbb{R} \to \mathbb{R}$ and $w^-, w^+ : [0,1] \to [0,1]$ satisfying 1) $v(0) = 0$, v concave on $\mathbb{R}_+$, convex with steeper curve on $\mathbb{R}_-$, 2) $w^+(0) = w^-(0) = 0$, $w^+(1) = w^-(1) = 1$, both increasing inverse S shape functions, such that the gamble valuation $V(R)$ given to any gamble R is given by:

2. **Gamble valuation formula:**

$$V(R) = \int_{-\infty}^{0} v_-(r) d(w^- \circ F_R)(r) + \int_{+\infty}^{0} v_+(r) d(w^+ \circ \bar{F}_R)(r)$$

   where $F_R : \mathbb{R} \to [0,1]$ is the cumulative distribution function of R, and $\bar{F}_R : \mathbb{R} \to [0,1]$ is its tail function.

We stress that both points are *crucial* to Cumulative prospect theory: the constraints imposed on v, w^- and w^+ are as important as the gamble valuation formula, because each constraint was deduced from many experiments performed by Kahneman and Tversky in their work.

Our contribution is to propose a parametric model for Cumulative Prospect theory, i.e. to propose parametrized classes of 1) reward distributions $\mathcal{R}$ (from which to draw the gamble R), 2) value functions $\mathcal{V}$ (from which to draw v), and 3) weighting functions $\mathcal{W}$ (from which to draw w^- and w^+), satisfying the required constraints in 1., flexible enough to approximate any function satisfying these constraints, and yielding an explicit valuation formula. The classes we propose are:

1. For $\mathcal{R}$: the class of gaussian reward distributions, parametrized by their mean and variance,

2. For $\mathcal{V}$: the utility functions

$$\begin{aligned} v_{m^-,V^-,a^-,m^+,V^+,a^+}(x) &= -(m^- x + V^-(1 - e^{-a^-(-x)}))\mathbf{1}_{x<0} \\ &\quad +(m^+ + V^+(1 - e^{-a^+ x}))\mathbf{1}_{z\geq 0} \end{aligned}$$

   with $m^- \geq m^+$, $V^- \geq V^+$, and $a^- \geq a^+$.

3. For $\mathcal{W}$: the weighting functions

$$w_{p_0,\gamma}(p) = \mathcal{N}(\gamma\mathcal{N}^{-1}(p) + (1-\gamma)\mathcal{N}^{-1}(p_0)), \quad \forall p \in [0,1]$$

where $\mathcal{N}$ denotes the cumulative distribution of the standard normal distribution.

Our main result is the following.

Theorem 1.3.1 *We have*

- **Validity:** *any value function $v \in \mathcal{V}$ and weighting function $w \in \mathcal{W}$ satisfies the constraints in 1..*
- **Density:** *$\mathcal{R}$, $\mathcal{V}$, $\mathcal{W}$ contains Gaussian reward distributions with any mean and variance, value functions with any asymptotes and rate of convergence to the asymptotes, and weighting functions with any crossover point and slope at the crossover point.*
- **Analytic valuation function:** *We have*

$$\begin{aligned} & V_{\mu,\sigma,p_0,\lambda,m^-,V^-,a^-,m^+,V^+,a^+} \\ = \; & -\mathbf{m}^- \left(\boldsymbol{x}\mathcal{N}(\boldsymbol{x}) - \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\boldsymbol{x}^2} \right) - V^- \left(\mathcal{N}(\boldsymbol{x}) - e^{-\mathbf{a}^+\boldsymbol{x} + \frac{(\mathbf{a}^+)^2}{2}} \mathcal{N}(\boldsymbol{x} - \mathbf{a}^+) \right) \\ + \; & \mathbf{m}^+ \left(\boldsymbol{x}\mathcal{N}(\boldsymbol{x}) - \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\boldsymbol{x}^2} \right) + V^+ \left(\mathcal{N}(\boldsymbol{x}) - e^{-\mathbf{a}^+\boldsymbol{x} + \frac{(\mathbf{a}^+)^2}{2}} \mathcal{N}(\boldsymbol{x} - \mathbf{a}^+) \right) \end{aligned}$$

where, $\boldsymbol{x} := \frac{\hat{\mu}}{\hat{\sigma}}$, $\bar{\boldsymbol{x}} := \frac{\bar{\hat{\mu}}}{\hat{\sigma}}$, $\mathbf{m}^+ := \hat{\sigma} m^+$, $\mathbf{m}^- := \hat{\sigma} m^-$, $\mathbf{a}^+ := \hat{\sigma} a^+$, $\mathbf{a}^- := \hat{\sigma} a^-$, where $\hat{\mu} = \mu - \sigma(\gamma^{-1} - 1)\mathcal{N}^{-1}(p_0)$, $\bar{\hat{\mu}} = \mu + \sigma(\gamma^{-1} - 1)\mathcal{N}^{-1}(p_0)$, $\hat{\sigma} = \sigma\gamma^{-1}$.

Our second contribution is to to discuss some applications of our results, in particular to large population's behavioral models when facing risky choices, with commercial and political applications.

1.3.3 Models for targeted advertising

In Part III, we study concrete models for solving targeted advertising problems. This is the most applied part of the thesis. It is split in two chapters, each studying important targeted advertising problematics.

1. In Chapter 5, we study an online learning algorithm for ad clicking prediction. This work ranges into the class of online binary classification algorithms, which is a major field of machine learning, widely used for targeted advertising and web recommendations.

2. In Chapter 6, we study models of optimal bidding strategies for targeted advertising auctions. This is a crucial problem for commercial companies or companies specialized in targeted advertising: every time an individual connects to a website using targeted advertising, an auction is automatically opened, where the bidders are all the companies interested in displaying their ad to this individual. Each bidding company then makes a bid for this auction, and the company winning the auction has its ad displayed to the individual and pays the website an amount depending upon the auction rules. This chapter is at the intersection of optimal control and auction theory.

Although both studies are designed for specific applications, both works involve interesting theoretical techniques to derive, on one hand, a logarithmic bound for the prediction efficiency of the click prediction algorithm, and on the other hand an explicit formula for optimal bidding strategies for targeted advertising.

Online click prediction learning algorithm

Our first work for targeted advertising defines and studies an online click prediction learning algorithm. A product is associated to a *price* $p \in \mathbb{R}$ and to features $\mathbf{f} \in \mathcal{F} := [0,1]^d$. A product's features $\mathbf{f} \in \mathcal{F}$ represent the characteristics of a product (quality, brand's reputation, shape, life duration, etc). Thus, a product is characterized by a price-features pair $(p, \mathbf{f}) \in \mathbb{R} \times \mathcal{F}$. By misuse of language, we identify the product with any advertisement of the product. We will thus indifferently say "the product $(p, \mathbf{f})$" and "the ad $(p, \mathbf{f})$".

A click will be represented by a binary variable $c \in \{-1, 1\}$, 1 meaning "click", and -1 "no click". Depending upon what happens, it will have slightly different interpretations: as long as an ad $(p, \mathbf{f}) \in \mathbb{R} \times \mathcal{F}$ has not been displayed to the individual, the associated $c \in \{-1, 1\}$ is a *click intention*, but once (and if) the ad $(p, \mathbf{f})$ is displayed to the individual, c will correspond to his *clicking decision*.

We denote by $\mathcal{R}_{d,D}$ the set of multi-dimensional polynomial functions from $\mathcal{F}$ to $\mathbb{R}$ with maximal degree D in each coordinate, i.e. taking the form

$$R(\mathbf{f}) = \sum_{\mathbf{i} \in [\![0,D]\!]^d} r_{\mathbf{i}} \prod_{k=1}^{d} f_k^{i_k}, \quad \forall \mathbf{f} = (f_k)_{k \in [\![1,d]\!]} \in \mathcal{F}.$$

where $\mathbf{r} = (r_{\mathbf{i}})_{\mathbf{i} \in [\![0,D]\!]^d} \in \mathbb{R}^{D^d}$ is a multi-index vector. A function $R \in \mathcal{R}$ will be interpreted as a *reward function*, associating to any features $\mathbf{f} = (f_i)_{i \in [\![1,d]\!]}$ the reward $R(\mathbf{f})$.

The framework for the online learning algorithm is the following. We consider a random sequence $(p_k, \mathbf{f}_k, c_k)_{k\in\mathbb{N}}$ of ads $(p_k, \mathbf{f}_k)$ and associated clicking intentions c_k for all $k \in \mathbb{N}$. More precisely, at each time $k \in \mathbb{N}$, a new product is created by a company, with price $p_k \in \mathbb{R}$ and features $\mathbf{f}_k = (f_{k,i})_{i\in[\![1,d]\!]} \in \mathcal{F}$, yielding the product/ad $(p_k, \mathbf{f}_k)$. For a given individual, the company wonders if it should display the ad $(p_k, \mathbf{f}_k)$ to a given individual. The binary value c_k represents the *clicking intention* of the individual for this product, that is, c_k is the answer to the question "if ad $(p_k, \mathbf{f}_k)$ was displayed to the individual, would he click on it?". If so, then, by definition, $c_k = 1$. Otherwise, $c_k = -1$.

We assume that there exists a *reward function* $R_\star \in \mathcal{R}$ such that, $\forall k \in \mathbb{N}$,

$$p_k < R_\star(\mathbf{f}_k) - \varepsilon \Rightarrow c_k = 1, \quad \text{and} \quad p_k > R_\star(\mathbf{f}_k) + \varepsilon \Rightarrow c_k = -1,$$

where $\varepsilon \geq 0$ is a margin, important for realism because it encompasses several natural phenomenons:

- **Non-polynomial reward functions:** the "real" reward function of the individual might not be polynomial, but only approximable with a polynomial reward function up to an error ε,
- **Hidden variables, or inconsistent clicking decisions:** there might be an unobservable noise in the individual's evaluation of the product's utility making him value slightly differently a same product at two different times.
- **Time varying reward function:** The underlying reward function of the individual might slightly evolve with time, and thus, for $n \leq N$, all the successive utility functions of the individual are close to the first one up to a margin error ε.

I.i.d. products with atomless distribution: We assume that $(p_k, \mathbf{f}_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. random variables with common distribution ν assumed atomless and such that $\frac{d\nu}{d\lambda} \leq C$ for some constant C.

Upper and lower bounded conditional density at the margin: We assume that there exists $\eta > \varepsilon$ such that

$$c < \frac{d\mathcal{L}(p_1 - R_\star(\mathbf{f}_1) \mid \mathbf{f}_1)}{d\lambda}(y) < C, \quad \forall y \in [-\eta, \eta], \quad a.s.$$

Let us now informally describe the rules of the targeted advertising problem. At each time $k \in \mathbb{N}$ the following steps occur:

1. **New ad event:** a new ad advertising a new product $(p_k, \mathbf{f}_k)$ is created. At this point, $(p_k, \mathbf{f}_k)$ is observable to the advertiser and can be used, along with data stored in memory from past times, for the subsequent steps.

2. **Displaying decision:** the advertiser executes a program processing $(p_k, \mathbf{f}_k)$ and data stored in memory from last times to decide whether or not to display ad $(p_k, \mathbf{f}_k)$ to the individual.

3. **Clicking reaction (this step happens only if ad $(p_k, \mathbf{f}_k)$ was displayed to the individual):** Once ad $(p_k, \mathbf{f}_k)$ is displayed, the individual sees $(p_k, \mathbf{f}_k)$ it and either clicks on it or not, according to the clicking intention c_k. In either case, the advertiser observes the reaction of the individual, which means that he observes c_k. We stress that if the advertiser chose to not display ad $(p_k, \mathbf{f}_k)$ in last step, this step does not happen and the advertiser does *not* observe c_k.

4. **Memory update:** The advertiser has the possibility to update the variables stored in memory, and in particular he can choose to remember $(p_k, \mathbf{f}_k)$, and, provided that he displayed the ad to the individual at step 2, the clicking reaction c_k observed ad step 3, for future use.

For the sake of conciseness, let us not detail the algorithm here, but let us instead describe its main characteristics and properties.

Main characteristics. The main elements of the algorithm are:

- **Feature space transformation:** We transform the data $(\mathbf{f}_k)_{k\in\mathbb{N}} \subset [0,1]^d$ into data $(\phi(\mathbf{f}_k))_{k\in\mathbb{N}} \subset [0,1]^{D^d}$ for some function $\phi : [0,1]^d \to [0,1]^{D^d}$, essentially allowing us to see $R_\star$ as a *linear* function, i.e. such that $R_\star(\mathbf{f}_k) = \mathbf{r}_\star \cdot \phi(\mathbf{f}_k)$ for all $k \in \mathbb{N}$, for some *reward* vector $\mathbf{r}_\star \in \mathbb{R}^{D^d}$. Notice that in this case we have $R_\star(\mathbf{f}) - p = \boldsymbol{u}_\star \cdot (p, \phi(\mathbf{f}))$, where $\boldsymbol{u}_\star = (-1, \mathbf{r}_\star)$.

- **Utility vector approximation:** The learning side of the algorithm then essentially consists in approximating $\boldsymbol{u}_\star$ with a sequence of vectors $(\boldsymbol{u}_n)_{n\in\mathbb{N}} \subset \mathbb{R}^{D^d}$, supposed to be closer and closer to $\boldsymbol{u}_\star$, allowing to make better and better click predictions.

Main properties. Our main results are:

- **No false negative:** The algorithm completely avoids false negatives, i.e. not displaying an ad that would have led to a click. In other words, it does not miss any click.

- **Asymmetric feedback:** As discussed above, the algorithm only observes the clicking intention c_k of the individual after he displayed the ad $(p_k, \mathbf{f}_k)$ to him. Thus, when $(p_k, \mathbf{f}_k)$ is not displayed, c_k is never accessed.

- **Logarithmic false positives:** Despite the strong constraint the avoid all false negative, the algorithm is able to only make
  - a logarithm number $C\ln(k)$ of false positives (displaying an ad that does not lead to a click) before time k, in the case where $\varepsilon = 0$ (i.e. when there exists a polynomial reward function $R_\star$ perfectly determining the clicking intention).
  - In the general case $\varepsilon \geq 0$, a number $C(\ln(k) + k\varepsilon)$ of false positives (where we stress that C does not depend upon ε).
- **Computational and memory efficiency:** The algorithm has a bounded average memory usage and computational efficiency per each time k.

Optimal bidding strategies for targeted advertising

In this work, we design several population optimal control problems for targeted advertising. We here detail the simplest one, and then informally mention the other models with more features. To make this summary clearer, let us progressively introduce the features of the models.

Information: In our model, the key element is the notion of information. We assume that there is an information that initially nobody, except the agent (e.g. a company), knows. This information can represent for instance the existence of a new product sold by the agent. In the sequel, we denote by I the information.

The individual: We now model an individual and his behavior. The individual is characterized by two independent Poisson processes $(N^{\mathbf{I}}, N^{\mathbf{T}})$ such that:

- $N^{\mathbf{I}}$ is a Poisson process with intensity $\eta_{\mathbf{I}}$, counting the times when the individual connects to a website containing information I. It can be a specialized website about products, or even the agent's commercial website itself, both providing information I when one connects to these.
- $N^{\mathbf{T}}$ is a Poisson process with intensity $\eta_{\mathbf{T}} =: 1$ (by convention), counting the times when the individual connects to a website not containing, a priori, information I but displaying targeted ads on his web pages, thus susceptible to display information I via a targeted ad provided that the agent pays for it. Many websites display targeted ads: search engines, social networks, and many standard websites.

The targeted advertising auctions: We consider a family of i.i.d. real random variables $(B_k^{\mathbf{T}})_{k\in\mathbb{N}}$. For $k \in \mathbb{N}$, $B_k^{\mathbf{T}}$ represents the price above which the agent's bid has to be to win the k-th targeted advertising auction. We thus model this price with exogenous random variables (as opposed to modeling it endogenously by considering each

bidder, which would turn the problem into a N-bidders game). The process $(B^{\mathbf{T}}_{N^{\mathbf{T}}_t})_{t\in\mathbb{R}_+}$ has at each time $t \in \mathbb{R}_+$ the value of the price of the last auction.

Past data: We have now introduced all the randomness generating the model. The bidding control of the Agent is assumed to be non-anticipative, i.e. to not depend upon future random events. To represent this, we consider the filtration $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$ such that $\mathcal{F}_t = \sigma(N^{\mathbf{I}}_s, N^{\mathbf{T}}_s, B^{\mathbf{T}}_{N^{\mathbf{T}}_s}, s \leq t)$.

Open-loop bidding strategy: An open-loop bidding strategy, for the agent willing to diffuse information I, is a real valued random process β, progressively measurable w.r.t. $(\mathcal{F}_{t-})_{t\in\mathbb{R}_+}$ (non-anticipative property), such that β_t corresponds to the bid that the agent would make if there is an opportunity to display a targeted ad to the individual at time t (i.e. if $\Delta N^{\mathbf{T}}_t = 1$).

The controlled dynamic system: We denote by X^β the $\{0,1\}$-valued process such that $X^\beta_t = 1$ iff the individual has obtained the information before time t, given the bidding strategy β of the agent. X^β is modeled as the solution to the dynamic system

$$\begin{aligned} X^\beta_0 &= 0 \\ dX^\beta_t &= (1 - X^\beta_{t-})(\mathbf{1}_{\beta_t \geq B^{\mathbf{T}}_{N^{\mathbf{T}}_t}} dN^{\mathbf{T}}_t + dN^{\mathbf{I}}_t) \end{aligned}$$

Essentially, this dynamic means that the individual starts uninformed. He gets informed as soon as either 1) he connects to a website displaying targeted ads, and the agent wins the auction ("$\mathbf{1}_{\beta_t \geq B^{\mathbf{T}}_{N^{\mathbf{T}}_t}} dN^{\mathbf{T}}_t$" part), or 2) he connects to a website containing information I ("$dN^{\mathbf{I}}_t$" part). Then, he stays informed forever ("$(1 - X^\beta_{t-})$" part).

The agent's gain: The expected gain of the agent, given a bidding strategy β, is

$$V(\beta) = \mathbb{E}\Big[\int_0^\infty e^{-\rho t}(K dX^\beta_t - \mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N^{\mathbf{T}}_t}} \mathbf{c}(\beta_t, B^{\mathbf{T}}_{N^{\mathbf{T}}_t}) dN^{\mathbf{T}}_t)\Big]$$

where $K \in \mathbb{R}$ represents the margin profit made by the agent when the individual gets informed, i.e. when $\Delta X^\beta_t = 1$, and where $\mathbf{c} : \mathbb{R}^2 \to \mathbb{R}$ is a function representing what the Agent will pay if he wins the auction. More precisely, the second part corresponds to the price paid when auctions are won by the agent: an auction happens when $\Delta N^{\mathbf{T}}_t = 1$, the auction is won if $\mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N^{\mathbf{T}}_t}}$. The price paid is $\mathbf{c}(\beta_t, B^{\mathbf{T}}_{N^{\mathbf{T}}_t})$:

- If we consider a first-price auction rule, we have $\mathbf{c}(b, B) = b$, i.e. if the Agent wins the auction, he pays his bid β_t.
- If we consider a second-price auction rule, we have $\mathbf{c}(b, B) = B$, i.e. if the Agent wins the auction, he pays the second highest bid in the auction, i.e. $B^{\mathbf{T}}_{N^{\mathbf{T}}_t}$.

The goal of the agent is then to use a bidding strategy $\beta^\star$ such that $V(\beta^\star) = \sup_\beta V(\beta) =: V^\star$.

The optimal policy and value: The optimal value is given by

$$V^\star \;=\; \sup_{b\in\mathbb{R}} \frac{\eta^{\mathbf{I}} K + \eta^{\mathbf{T}} \mathbb{E}[(K - \mathbf{c}(b, B_1^{\mathbf{T}}))\mathbf{1}_{b\geq B_1^{\mathbf{T}}}]}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}} \mathbb{P}(b \geq B_1^{\mathbf{T}})},$$

and the optimal bidding control $\beta^\star$ is the unique open-loop bidding control such that $\beta_t^\star = (1 - X_t^{\beta^\star})b^\star$, where

$$b^\star := \operatorname*{argmax}_{b\in\mathbb{R}} \frac{\eta^{\mathbf{I}} K + \eta^{\mathbf{T}} \mathbb{E}[(K - \mathbf{c}(b, B_1^{\mathbf{T}}))\mathbf{1}_{b\geq B_1^{\mathbf{T}}}]}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}} \mathbb{P}(b \geq B_1^{\mathbf{T}})}$$

In other words, the optimal bidding policy is to make the constant bid $b^\star$ as long as the individual is not informed, and then to stop bidding (which is an obvious part of the strategy).

We now mention the other models in this work. We have designed models for two types of advertising:

1. **Commercial advertising**, modeling situations where informing an individual triggers a reward for the agent, which is generally the case in commercial advertising. We consider two types of reward: *purchase-based reward*, modeling the case where the information triggers a purchase and thus a punctual payment from the individual to the agent, and *subscription-based reward*, modeling cases where the information triggers a subscription of the individual to a service proposed by the agent, and thus pays a regular fee to the agent.

2. **Social marketing**, modeling situations where informing an individual cancels a cost continuously perceived by the agent. In this model, each individual, as long as he is not informed, incurs a continuous cost to the agent, which is particularly well suited for social marketing where the agent's goal is not to make profit but instead to change people's behaviors and promoting social change by sensitizing them about dangers (anti-drugs campaigns, road-safety campaigns, sexual-safety campaigns, low-fat diet campaigns, etc).

The model we detailed in this summary is the commercial advertising model with purchased-base reward. The model with subscription-based reward only differs in the gain function, because informing the individual triggers his subscription and thus a regular fee instead of a punctual payment. The social marketing model has much more features than the commercial marketing models, as it additionally involves 1) social interactions, and 2) non-targeted advertising. In all these models, we obtain a closed formula for the optimal value and the optimal policy, with a similar form.

1.4 Outline of the thesis

1. Part I: Chapter 2: Mean-field Markov decision processes with common noise and open-loop controls. Chapter 3: Chaos propagation of N-agent Markov decision processes with common noise and open-loop controls.

2. Part II: Behavioral economics models. Chapter 4: Large population games with the Iterative Elimination of Strictly Dominated Strategies concept. Chapter 5: Gaussian cumulative prospect theory.

3. Part III: Models for targeted advertising. Chapter 6: Online click prediction learning algorithm. Chapter 7: Optimal control for targeted advertising.

Part I

Large populations with mean-field interactions

Chapter 2

Mean-field Markov decision processes with common noise and open-loop controls

Abstract. In this chapter, we develop an exhaustive study of Markov decision process (MDP) under mean field interaction both on states and actions in the presence of common noise, and when optimization is performed over open-loop controls on infinite horizon. Such model, called CMKV-MDP for conditional McKean-Vlasov MDP, is formally obtained by substituting empirical distributions by theoretical ones in a N-agent Markov Decision Process. We highlight the crucial role of relaxed controls and randomization hypothesis for this class of models with respect to classical MDP theory. We prove the correspondence between CMKV-MDP and a general lifted MDP on the space of probability measures, and establish the dynamic programming Bellman fixed point equation satisfied by the value function, as well as the existence of ϵ-optimal randomized feedback controls. The arguments of proof involve an original measurable optimal coupling for the Wasserstein distance.

2.1 Introduction

Optimal control of McKean-Vlasov (MKV) systems, also known as mean-field control (MFC) problems, has sparked a great interest in the domain of applied probabilities during the last decade. In these optimization problems, the transition dynamics of the system and the reward/gain function depend not only on the state and action of the agent/controller, but also on their probability distributions. These problems are motivated from models of large population of interacting cooperative agents obeying to a social planner (center of decision), and are often justified heuristically as the asymptotic

regime with infinite number of agents under Pareto efficiency. Such problems have found numerous applications in distributed energy, herd behavior, finance, etc.

A large literature has already emerged on continuous-time models for the optimal control of McKean-Vlasov dynamics, and dynamic programming principle (in other words time consistency) has been established in this context in the papers [53], [72], [7], [24]. We refer to the books [8], [15] for an overview of the subject.

Our work and main contributions. In this paper, we introduce a general discrete time framework by providing an exhaustive study of Markov decision process (MDP) under mean-field interaction in the presence of *common noise*, and when optimization is performed over *open-loop controls* on infinite horizon. Such model is called conditional McKean-Vlasov MDP, shortly abbreviated in the sequel as CMKV-MDP, and the set-up is the mathematical framework for a theory of reinforcement learning with mean-field interaction. Let us first briefly describe and motivate the main features of our framework:

(i) The controls are open-loop, which is a natural assumption to study the problem with the richest possible set of controls adapted to the past. This is useful to prove that apparently more restrictive control sets are actually sufficient in the sense that allowing access to more information would not increase the optimal value.

(ii) The dynamics of individuals depend upon a common noise, emulating the fact that they are influenced by common information (public data) which may vary over time. We consider an i.i.d. common noise sequence $(\varepsilon_t^0)_{t\in\mathbb{N}}$, but we stress that this framework contains the apparently more realistic framework of a Markovian common noise sequence, up to a change of state space, as we shall discuss.

Compared to continuous-time models, discrete-time McKean-Vlasov control problems have been less studied in the literature. In [71], the authors consider a finite-horizon problem without common noise and state the dynamic programming (Bellman) equation for MFC with closed-loop (also called feedback) controls, that are restricted to depend on the state. Very recently, the works [16], [33] addressed Bellman equations for MFC problems in the context of reinforcement learning. The paper [33] considers relaxed controls in their MFC formulation but without common noise, and derives the Bellman equation for the Q-value function as a deterministic control problem that we obtain here as a particular case (see our Remark 2.4.10). The framework in [16] is closest to ours by considering also common noise, however with the following differences: these authors restrict their attention to stationary feedback policies, and reformulate their MFC control problem as a MDP on the space of probability measures by deriving formally (leaving aside the measurability issues and assuming the existence of a stationary feedback control) the associated Bellman equation, which is then used for the develop-

ment of Q-learning algorithms. Notice that [16], [33] do not consider dependence upon the probability distribution of the control in the state transition dynamics and reward function.

Our first contribution is to obtain the correspondence of our CMKV-MDP with a suitable lifted MDP on the space of probability measures. Starting from open-loop controls, this is achieved in general by introducing relaxed (i.e. measure-valued) controls in the enlarged state/action space, and by emphasizing the measurability issues arising in the presence of common noise and with continuous state space. In the special case without common noise or with finite state space, the relaxed control in the lifted MDP is reduced to the usual notion in control theory, also known as mixed or randomized strategies in game theory. While it is known in standard MDP that an optimal control (when it exists) is in pure form, relaxed control appears naturally in MFC where the social planner has to sample the distribution of actions instead of simply assigning the same pure strategy among the population in order to perform the best possible collective gain.

The reformulation of the original problem as a lifted MDP leads us to consider an associated dynamic programming equation written in terms of a Bellman fixed point equation in the space of probability measures. Our second contribution is to establish rigorously the Bellman equation satisfied by the state value function of the CMKV-MDP, and then by the state-action value function, called Q-function in the reinforcement learning terminology. This is obtained under the crucial assumption that the initial information filtration is generated by an atomless random variable, i.e., that it is rich enough, and calls upon original measurable optimal coupling results for the Wasserstein distance. Moreover, and this is our fourth contribution, the methodology of proof allows us to obtain as a by-product the existence of an ϵ-optimal control, which is constructed from randomized feedback policies under a randomization hypothesis. This shows in particular that the value function of CMKV-MDP over open-loop controls is equal to the value function over randomized feedback controls, and we highlight that it may be strictly larger than the value function of CMKV-MDP over "pure" feedback controls, i.e., without randomization. This is a notable difference with respect to the classical (without mean-field dependence) theory of MDP as studied e.g. in [11], [81]. We discuss and illustrate with a set of simple examples the difference of control strategies (open loop vs feedback), and the crucial role of the randomization hypothesis.

Finally, we discuss how to compute the value function and approximate optimal randomized feedback controls from the Bellman equation according to value or policy iteration methods and by discretization of the state space and of the space of probability measures.

Outline of the paper. The rest of the paper is organized as follows. In Section 2.3, we establish the correspondence of the CMKV-MDP with a lifted MDP on the space of probability measures with usual relaxed controls when there is no common noise or when the state space is finite. In the general case considered in Section 2.4, we show how to lift the CMKV-MDP by a suitable enlargement of the action space in order to get the correspondance with a MDP on the Wasserstein space. We then derive the associated Bellman fixed point equation satisfied by the value function, and obtain the existence of approximate randomized feedback controls. We also highlight the differences between open-loop vs feedback vs randomized controls. We conclude in Section 2.5 by indicating some questions for future research. Finally, we collect in the Appendix some useful and technical results including measurable coupling arguments used in the proofs of the paper.

Notations. Given two measurable spaces $(\mathcal{X}_1, \Sigma_1)$ and $(\mathcal{X}_2, \Sigma_2)$, we denote by pr_1 (resp. pr_2) the projection function $(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2 \mapsto x_1 \in \mathcal{X}_1$ (resp. $x_2 \in \mathcal{X}_2$). For a measurable function $\Phi : \mathcal{X}_1 \to \mathcal{X}_2$, and a positive measure μ_1 on $(\mathcal{X}_1, \Sigma_1)$, the pushforward measure $\Phi \star \mu_1$ is the measure on $(\mathcal{X}_2, \Sigma_2)$ defined by

$$\Phi \star \mu_1(B_2) = \mu_1\big(\Phi^{-1}(B_2)\big), \quad \forall B_2 \in \Sigma_2.$$

We denote by $\mathcal{P}(\mathcal{X}_1)$ the set of probability measures on $\mathcal{X}_1$, and $\mathcal{C}(\mathcal{X}_1)$ the cylinder (or weak) σ-algebra on $\mathcal{P}(\mathcal{X}_1)$, that is the smallest σ-algebra making all the functions $\mu \in \mathcal{P}(\mathcal{X}_1) \mapsto \mu(B_1) \in [0,1]$, measurable for all $B_1 \in \Sigma_1$.

A probability kernel ν on $\mathcal{X}_1 \times \mathcal{X}_2$, denoted $\nu \in \hat{\mathcal{X}}_2(\mathcal{X}_1)$, is a measurable mapping from $(\mathcal{X}_1, \Sigma_1)$ into $(\mathcal{P}(\mathcal{X}_2), \mathcal{C}(\mathcal{X}_2))$, and we shall write indifferently $\nu(x_1, B_2) = \nu(x_1)(B_2)$, for all $x_1 \in \mathcal{X}_1$, $B_2 \in \Sigma_2$. Given a probability measure μ_1 on $(\mathcal{X}_1, \Sigma_1)$, and a probability kernel $\nu \in \hat{\mathcal{X}}_2(\mathcal{X}_1)$, we denote by $\mu_1 \cdot \nu$ the probability measure on $(\mathcal{X}_1 \times \mathcal{X}_2, \Sigma_1 \otimes \Sigma_2)$ defined by

$$(\mu_1 \cdot \nu)(B_1 \times B_2) = \int_{B_1 \times B_2} \mu_1(\mathrm{d}x_1)\nu(x_1, \mathrm{d}x_2), \quad \forall B_1 \in \Sigma_1,\ B_2 \in \Sigma_2.$$

Let X_1 and X_2 be two random variables valued respectively on $\mathcal{X}_1$ and $\mathcal{X}_2$, denoted $X_i \in L^0(\Omega; \mathcal{X}_i)$. We denote by $\mathcal{L}(X_i)$ the probability distribution of X_i, and by $\mathcal{L}(X_2|X_1)$ the conditional probability distribution of X_2 given X_1. With these notations, when $X_2 = \Phi(X_1)$, then $\mathcal{L}(X_2) = \Phi \star \mathcal{L}(X_1)$.

When $(\mathcal{Y}, d)$ is a compact metric space, the set $\mathcal{P}(\mathcal{Y})$ of probability measures on $\mathcal{Y}$ is equipped with the Wasserstein distance

$$\mathcal{W}(\mu, \mu') = \inf\Big\{ \int_{\mathcal{Y}^2} d(y, y')\boldsymbol{\mu}(\mathrm{d}y, \mathrm{d}y') : \boldsymbol{\mu} \in \boldsymbol{\Pi}(\mu, \mu')\Big\},$$

where $\mathbf{\Pi}(\mu,\mu')$ is the set of probability measures on $\mathcal{Y}\times\mathcal{Y}$ with marginals μ and μ', i.e., $\mathrm{pr}_1\star\boldsymbol{\mu}=\mu$, and $\mathrm{pr}_2\star\boldsymbol{\mu}=\mu'$. Since $(\mathcal{Y},d)$ is compact, it is known (see e.g. Corollary 6.13 in [90]) that the Borel σ-algebra generated by the Wasserstein metric coincides with the cylinder σ-algebra on $\mathcal{P}(\mathcal{Y})$, i.e., Wasserstein distances metrize weak convergence. We also recall the dual Kantorovich-Rubinstein representation of the Wasserstein distance

$$\mathcal{W}(\mu,\mu')=\sup\Big\{\int_{\mathcal{Y}}\phi\,\mathrm{d}(\mu-\mu'):\phi\in L_{lip}(\mathcal{Y};\mathbb{R}),[\phi]_{lip}\leq 1\Big\},$$

where $L_{lip}(\mathcal{Y};\mathbb{R})$ is the set of Lipschitz continuous functions ϕ from $\mathcal{Y}$ into $\mathbb{R}$, and $[\phi]_{lip}$ $=\sup\{|\phi(y)-\phi(y')|/d(y,y'):y,y'\in\mathcal{Y},y\neq y'\}$.

2.2 The N-agent and McKean-Vlasov MDP

We formulate the mean-field Markov Decision Process (MDP) in a large population model with indistinguishable agents $i\in\mathbb{N}^*=\mathbb{N}\setminus\{0\}$.

Let $\mathcal{X}$ (the state space) and A (the action space) be two compact Polish spaces equipped respectively with their metric d and d_A. We denote by $\mathcal{P}(\mathcal{X})$ (resp. $\mathcal{P}(A)$) the space of probability measures on $\mathcal{X}$ (resp. A) equipped respectively with their Wasserstein distance $\mathcal{W}$ and $\mathcal{W}_A$. We also consider the product space $\mathcal{X}\times A$, equipped with the metric $\boldsymbol{d}((x,a),(x',a'))=d(x,x')+d_A(a,a')$, $x,x'\in\mathcal{X}$, $a,a'\in A$, and the associated space of probability measure $\mathcal{P}(\mathcal{X}\times A)$, equipped with its Wasserstein distance $\boldsymbol{W}$. Let G, E, and E^0 be three measurable spaces, representing respectively the initial information, idiosyncratic noise, and common noise spaces.

Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space on which are defined the following family of mutually i.i.d. random variables

- $(\Gamma^i,\xi^i)_{i\in\mathbb{N}^\star}$ (initial informations and initial states) valued in $G\times\mathcal{X}$
- $(\varepsilon^i_t)_{i\in\mathbb{N}^\star,t\in\mathbb{N}}$ (idiosyncratic noises) valued in E with probability distribution λ_ε
- $\varepsilon^0:=(\varepsilon^0_t)_{t\in\mathbb{N}}$ (common noise) valued in E^0.

We assume that $\mathcal{F}$ contains an atomless random variable, i.e., $\mathcal{F}$ is rich enough, so that any probability measure ν on $\mathcal{X}$ (resp. A or $\mathcal{X}\times A$) can be represented by the law of some random variable Y on $\mathcal{X}$ (resp. A or $\mathcal{X}\times A$), and we write $Y\sim\nu$, i.e., $\mathcal{L}(Y)=\nu$.

Given $N\in\mathbb{N}^*$, we denote by $\mathcal{A}_N$ the set of open-loop controls for the N-individual MDP, that is, the set of A^N-valued random sequences α, adapted to the filtration $(\mathcal{F}_{N,t})_{t\in\mathbb{N}}$ defined by $\mathcal{F}_{N,t}=\sigma(\Gamma^i,\xi^i,(\varepsilon^i_s)_{s\leq t},i\in[\![1,N]\!],(\varepsilon^0_s)_{s\leq t})$.

Given $\alpha \in \mathcal{A}_N$, the state process of agent $i = 1, \ldots, N$ in an N-agent MDP is given by the dynamical system

$$\left\{ \begin{array}{rcl} X_0^{i,N,\alpha} & = & \xi^i \\ X_{t+1}^{i,N,\pi} & = & F(X_t^{i,N,\alpha}, \alpha_t^i, \frac{1}{N}\sum_{j=1}^N \delta_{(X_t^{j,N,\alpha},\alpha_t^{j,\pi})}, \varepsilon_{t+1}^i, \varepsilon_{t+1}^0), \quad t \in \mathbb{N}, \end{array} \right.$$

where F is a measurable function from $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X} \times A) \times E \times E^0$ into $\mathcal{X}$, called state transition function. The i-th individual contribution to the influencer's gain over an infinite horizon is defined by

$$J_i^{N,\alpha} := \sum_{t=0}^{\infty} \beta^t f\Big(X_t^{i,N,\alpha}, \alpha_t^i, \frac{1}{N}\sum_{j=1}^N \delta_{(X_t^{j,N,\alpha},\alpha_t^j)}\Big), \quad i = 1, \ldots, N,$$

where the reward f is a measurable real-valued function on $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X} \times A)$, assumed to be bounded (recall that $\mathcal{X}$ and A are compact spaces), and β is a positive discount factor in $[0, 1)$. The influencer's renormalized and expected gains are

$$J^{N,\alpha} \; := \; \frac{1}{N}\sum_{i=1}^N J_i^{N,\alpha}, \qquad V^{N,\alpha} \; := \; \mathbb{E}\big[J^{N,\alpha}\big],$$

and the optimal value of the influencer is $V^N := \sup_{\alpha \in \mathcal{A}_N} V^{N,\alpha}$. Observe that the agents are indistinguishable in the sense that the initial pair of information/state $(\Gamma^i, \xi^i)_i$, and idiosyncratic noises are i.i.d., and the state transition function F, reward function f, and discount factor β do not depend on i.

Let us now consider the asymptotic problem when the number of agents N goes to infinity. In view of the propagation of chaos argument, we expect the N-individual MDP to converge in some sense to the following McKean-Vlasov MDP.

Let us rename Γ, ξ and $(\varepsilon_t)_{t\in\mathbb{N}}$ the random variables Γ^1, ξ^1, and $(\varepsilon_t^1)_{t\in\mathbb{N}}$. We also introduce $\mathcal{A}$, the set of open-loop controls for the McKean-Vlasov MDP, that is, the set of A-valued random sequences α adapted to the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ such that $\mathcal{F}_t := \sigma(\Gamma, \xi, (\varepsilon_s)_{s\le t}, (\varepsilon_s^0)_{s\le t})$. Given $\alpha \in \mathcal{A}$, we define the conditional McKean-Vlasov dynamic

$$\left\{ \begin{array}{rcl} X_0^\alpha & = & \xi \\ X_{t+1}^\alpha & = & F(X_t^\alpha, \alpha_t, \mathbb{P}^0_{(X_t^\alpha,\alpha_t)}, \varepsilon_{t+1}, \varepsilon_{t+1}^0), \quad t \in \mathbb{N}. \end{array} \right. \tag{2.2.1}$$

Here, we denote by $\mathbb{P}^0$ and $\mathbb{E}^0$ the conditional probability and expectation knowing the common noise ε^0, and then, given a random variable Y valued in $\mathcal{Y}$, we denote by $\mathbb{P}^0_Y$ or $\mathcal{L}^0(Y)$ its conditional law knowing ε^0, which is a random variable valued in $\mathcal{P}(\mathcal{Y})$ (see Lemma 2.6.2). The influencer's expected gain in the McKean-Vlasov model is

$$V^\alpha \; := \; \mathbb{E}\big[\sum_{t=0}^{\infty} \beta^t f\big(X_t^\alpha, \alpha_t, \mathbb{P}^0_{(X_t^\alpha,\alpha_t)}\big)\big], \qquad V := \sup_{\alpha\in\Pi_{OL}} V^\alpha. \tag{2.2.2}$$

Problem (3.2.1)-(3.2.2) is called conditional McKean-Vlasov Markov decision process, CMKV-MDP in short. The study of the CMKV-MDP is a priori justified by the empirical efficiency of mean-field approximations. It is often preferable to have a good understanding of the limit candidate before proving that it is indeed the limit of the N-individual MDPs, which is why we first study it and postpone rigorous convergence results to next chapter.

In the sequel, we make the following regularity assumptions on F and f:

($\mathbf{HF_{lip}}$) There exists $K_F > 0$, such that for all $a \in A$, $e^0 \in E^0$, $x, x' \in \mathcal{X}$, $\nu, \nu' \in \mathcal{P}(\mathcal{X}\times A)$,

$$\mathbb{E}\big[d\big(F(x,a,\nu,\varepsilon_1^1,e^0),F(x',a,\nu',\varepsilon_1^1,e^0)\big)\big] \leq K_F\big(d(x,x') + \boldsymbol{W}(\nu,\nu')\big)\big).$$

($\mathbf{Hf_{lip}}$) There exists $K_f > 0$, such that for all $a \in A$, $x, x' \in \mathcal{X}$, $\nu, \nu' \in \mathcal{P}(\mathcal{X} \times A)$,

$$d(f(x,a,\nu),f(x',a,\nu')) \leq K_f\big(d(x,x') + \boldsymbol{W}(\nu,\nu')\big)\big).$$

Remark 2.2.1 We stress the importance of making the regularity assumptions for F in *expectation* only. For the same argument as in Remark **??**, when $\mathcal{X}$ is finite, F cannot be, strictly speaking, Lipschitz. However, F can be Lipschitz *in expectation*, e.g. once integrated w.r.t. the idiosyncratic noise, which is a very natural assumption. □

2.3 Lifted MDP on $\mathcal{P}(\mathcal{X})$

In the sequel, we shall denote by $\mathcal{G} = \sigma(\Gamma)$ the σ-algebra generated by the random variable Γ, hence representing the initial information filtration, and by $L^0(\mathcal{G};\mathcal{X})$ the set of $\mathcal{G}$-measurable random variables valued in $\mathcal{X}$. We shall assume that the initial state $\xi \in L^0(\mathcal{G};\mathcal{X})$, which means that the policy has access to the agent's initial state through the initial information filtration $\mathcal{G}$.

From now on, we denote the expected gain of the agent associated to initial state ξ and open-loop control α $V^\alpha(\xi)$, equal to

$$V^\alpha(\xi) = \mathbb{E}\Big[\sum_{t\in\mathbb{N}} \beta^t f(X_t,\alpha_t,\mathbb{P}^0_{(X_t,\alpha_t)})\Big],$$

where we stress the dependence upon the initial state ξ. The value function to the CMKV-MDP is then defined by

$$V(\xi) = \sup_{\alpha\in\mathcal{A}} V^\alpha(\xi), \quad \xi \in L^0(\mathcal{G};\mathcal{X}).$$

Let us now show how one can lift the CMKV-MDP to a (classical) MDP on the space of probability measures $\mathcal{P}(\mathcal{X})$. We set $\mathbb{F}^0$ as the filtration generated by the common noise

ε^0. Given an open-loop control $\alpha \in \mathcal{A}$, and its state process $X = X^{\xi,\alpha}$, denote by $\{\mu_t = \mathbb{P}^0_{X_t},\ t \in \mathbb{N}\}$, the random $\mathcal{P}(\mathcal{X})$-valued process, and notice from Proposition 2.6.1 that $(\mu_t)_t$ is $\mathbb{F}^0$-adapted. From (**??**), and recalling the pushforward measure notation, we have

$$\mu_{t+1} = F\big(\cdot,\cdot,\mathbb{P}^0_{(X_t,\alpha_t)},\cdot,\varepsilon^0_{t+1}\big) \star \big(\mathbb{P}^0_{(X_t,\alpha_t)} \otimes \lambda_\varepsilon\big), \quad a.s. \tag{2.3.1}$$

As the probability distribution λ_ε of the idiosyncratic noise is a fixed parameter, the above relation means that μ_{t+1} only depends on $\mathbb{P}^0_{(X_t,\alpha_t)}$ and ε^0_{t+1}. Moreover, by introducing the so-called *relaxed control* associated to the open-loop control α as

$$\hat{\alpha}_t(x) = \mathcal{L}^0\big(\alpha_t | X_t = x\big), \quad t \in \mathbb{N},$$

which is valued in $\hat{A}(\mathcal{X})$, the set of probability kernels on $\mathcal{X} \times A$ (see Lemma 2.6.2), we see from Bayes formula that $\mathbb{P}^0_{(X_t,\alpha_t)} = \mu_t \cdot \hat{\alpha}_t$. The dynamics relation (2.3.1) is then written as

$$\mu_{t+1} = \hat{F}(\mu_t, \hat{\alpha}_t, \varepsilon^0_{t+1}), \quad t \in \mathbb{N},$$

where the function $\hat{F} : \mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X}) \times E^0 \to \mathcal{P}(\mathcal{X})$ is defined by

$$\hat{F}(\mu, \hat{a}, e^0) = F\big(\cdot,\cdot,\mu\cdot\hat{a},\cdot,e^0\big) \star \big((\mu\cdot\hat{a}) \otimes \lambda_\varepsilon\big). \tag{2.3.2}$$

On the other hand, by the law of iterated conditional expectation, the expected gain can be written as

$$V^\alpha(\xi) = \mathbb{E}\Big[\sum_{t\in\mathbb{N}} \beta^t \mathbb{E}^0\big[f(X_t, \alpha_t, \mathbb{P}^0_{(X_t,\alpha_t)})\big]\Big],$$

with the conditional expectation term equal to

$$\mathbb{E}^0\big[f(X_t, \alpha_t, \mathbb{P}^0_{(X_t,\alpha_t)})\big] = \hat{f}(\mu_t, \hat{\alpha}_t),$$

where the function $\hat{f} : \mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X}) \to \mathbb{R}$ is defined by

$$\hat{f}(\mu, \hat{a}) = \int_{\mathcal{X}\times A} f(x, a, \mu\cdot\hat{a})(\mu\cdot\hat{a})(\mathrm{d}x, \mathrm{d}a). \tag{2.3.3}$$

The above derivation suggests to consider a MDP with state space $\mathcal{P}(\mathcal{X})$, action space $\hat{A}(\mathcal{X})$, a state transition function $\hat{F}$ as in (2.3.2), a discount factor $\beta \in [0,1)$, and a reward function $\hat{f}$ as in (2.3.3). A key point is to endow $\hat{A}(\mathcal{X})$ with a suitable σ-algebra in order to have measurable functions $\hat{F}$, $\hat{f}$, and $\mathbb{F}^0$-adapted process $\hat{\alpha}$ valued in $\hat{A}(\mathcal{X})$, so that the MDP with characteristics $(\mathcal{P}(\mathcal{X}), \hat{A}(\mathcal{X}), \hat{F}, \hat{f}, \beta)$ is well-posed. This issue is investigated in the next sections, first in special cases, and then in general case by a suitable enlargement of the action space.

2.3.1 Case without common noise

When there is no common noise, the original state transition function F is defined from $\mathcal{X} \times A \times \mathcal{P}(\mathcal{X} \times A) \times E$ into $\mathcal{X}$, and the associated function $\hat{F}$ is then defined from $\mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X})$ into $\mathcal{P}(\mathcal{X})$ by

$$\hat{F}(\mu, \hat{a}) = F\big(\cdot, \cdot, \mu \cdot \hat{a}, \cdot\big) \star \big((\mu \cdot \hat{a}) \otimes \lambda_\varepsilon\big).$$

In this case, we are simply reduced to a deterministic control problem on the state space $\mathcal{P}(\mathcal{X})$ with dynamics

$$\mu_{t+1} = \hat{F}(\mu_t, \kappa_t), \quad t \in \mathbb{N}, \;\; \mu_0 \;=\; \mu \in \mathcal{P}(\mathcal{X}),$$

controlled by $\kappa = (\kappa_t)_{t \in \mathbb{N}} \in \widehat{\mathcal{A}}$, the set of deterministic sequences valued in $\hat{A}(\mathcal{X})$, and cumulated gain/value function:

$$\widehat{V}^\kappa(\mu) = \sum_{t=0}^{\infty} \beta^t \hat{f}(\mu_t, \kappa_t), \quad \widehat{V}(\mu) \;=\; \sup_{\kappa \in \widehat{\mathcal{A}}} \widehat{V}^\kappa(\mu), \quad \mu \in \mathcal{P}(\mathcal{X}),$$

where the bounded function $\hat{f} : \mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X}) \to \mathbb{R}$ is defined as in (2.3.3). Notice that there are no measurability issues for $\hat{F}$, $\hat{f}$, as the problem is deterministic and all the quantities defined above are well-defined.

We aim to prove the correspondence and equivalence between the MKV-MDP and the above deterministic control problem. From similar derivation as in (2.3.1)-(2.3.3) (by taking directly law under $\mathbb{P}$ instead of $\mathbb{P}^0$), we clearly see that for any $\alpha \in \mathcal{A}$, $V^\alpha(\xi) = \widehat{V}^{\hat{\alpha}}(\mu)$, with $\mu = \mathcal{L}(\xi)$, and $\hat{\alpha} = \mathcal{R}_\xi(\alpha)$ where $\mathcal{R}_\xi$ is the relaxed operator

$$\begin{array}{rcl} \mathcal{R}_\xi : \mathcal{A} & \longrightarrow & \widehat{\mathcal{A}} \\ \alpha = (\alpha_t)_t & \longmapsto & \hat{\alpha} = (\hat{\alpha}_t)_t : \; \hat{\alpha}_t(x) = \mathcal{L}\big(\alpha_t | X_t^{\xi,\alpha} = x\big), \;\; t \in \mathbb{N}, \; x \in \mathcal{X}. \end{array}$$

It follows that $V(\xi) \leq \widehat{V}(\mu)$. In order to get the reverse inequality, we have to show that $\mathcal{R}_\xi$ is surjective. Notice that this property is not always satisfied: for instance, when the σ-algebra generated by ξ is equal to $\mathcal{G}$, then for any $\alpha \in \mathcal{A}$, α_0 is $\sigma(\xi)$-measurable at time $t = 0$, and thus $\mathcal{L}(\alpha_0|\xi)$ is a Dirac distribution, hence cannot be equal to an arbitrary probability kernel $\kappa_0 = \hat{a} \in \hat{A}(\mathcal{X})$. We shall then make the following randomization hypothesis.

Rand$(\xi, \mathcal{G})$**:** There exists a uniform random variable $U \sim \mathcal{U}([0,1])$, which is $\mathcal{G}$-measurable and independent of $\xi \in L^0(\mathcal{G}; \mathcal{X})$.

Remark 2.3.1 The randomization hypothesis **Rand**$(\xi, \mathcal{G})$ implies in particular that Γ is atomless, i.e., $\mathcal{G}$ is rich enough, and thus $\mathcal{P}(\mathcal{X}) = \{\mathcal{L}(\zeta) : \zeta \in L^0(\mathcal{G}; \mathcal{X})\}$. Furthermore,

it means that there is extra randomness in $\mathcal{G}$ besides ξ, so that one can freely randomize via the uniform random variable U the first action given ξ according to any probability kernel $\hat{a}$. Moreover, one can extract from U, by standard separation of the decimals of U (see Lemma 2.21 in [44]), an i.i.d. sequence of uniform variables $(U_t)_{t\in\mathbb{N}}$, which are $\mathcal{G}$-measurable, independent of ξ, and can then be used to randomize the subsequent actions. □

Theorem 2.3.1 (Correspondence in the no common noise case) *Assume that* $\mathbf{Rand}(\xi, \mathcal{G})$ *holds true. Then* $\mathcal{R}_\xi$ *is surjective from* $\mathcal{A}$ *into* $\widehat{\mathcal{A}}$*, and we have* $V(\xi) = \widehat{V}(\mu)$*, for* $\mu = \mathcal{L}(\xi)$*. Moreover, for* $\epsilon \geq 0$*, if* $\alpha^\epsilon \in \mathcal{A}$ *is an* ϵ*-optimal control for* $V(\xi)$*, then* $\mathcal{R}_\xi(\alpha^\epsilon) \in \widehat{\mathcal{A}}$ *is an* ϵ*-optimal control for* $\widehat{V}(\mu)$*, and conversely, if* $\hat{\alpha}^\epsilon \in \widehat{\mathcal{A}}$ *is an* ϵ*-optimal control for* $\widehat{V}(\mu)$*, then any* $\alpha^\epsilon \in \mathcal{R}_\xi^{-1}(\hat{\alpha}^\epsilon)$ *is an* ϵ*-optimal control for* $V(\xi)$*. Consequently, an optimal control for* $V(\xi)$ *exists iff an optimal control for* $\widehat{V}(\mu)$ *exists.*

Proof. In view of the above discussion, we only need to prove the surjectivity of $\mathcal{R}_\xi$. Fix a control $\kappa \in \widehat{\mathcal{A}}$ for the MDP on $\mathcal{P}(\mathcal{X})$. By Lemma 2.22 in [44], for all $t \in \mathbb{N}$, there exists a measurable function $\mathrm{a}_t : \mathcal{X} \times [0,1] \to A$ such that $\mathbb{P}_{\mathrm{a}_t(x,U)} = \kappa_t(x)$, for all $x \in \mathcal{X}$. It is then clear that the control α defined recursively by $\alpha_t := \mathrm{a}_t(X_t^{\xi,\alpha}, U_t)$, where $(U_t)_t$ is an i.i.d. sequence of $\mathcal{G}$-measurable uniform variables independent of ξ under $\mathbf{Rand}(\xi, \Gamma)$, satisfies $\mathcal{L}(\alpha_t \mid X^{\xi,\alpha} = x) = \kappa_t(x)$ (observing that U_t is independent of $X_t^{\xi,\alpha}$), and thus $\hat{\alpha} = \kappa$, which proves the surjectivity of $\mathcal{R}_\xi$. □

Remark 2.3.2 The above correspondence result shows in particular that the value function V of the MKV-MDP is law invariant, in the sense that it depends on its initial state ξ only via its probability law $\mu = \mathcal{L}(\xi)$, for ξ satisfying the randomization hypothesis. □

2.3.2 Case with finite state space $\mathcal{X}$ and with common noise

We consider the case with common noise but when the state space $\mathcal{X}$ is finite, i.e., its cardinal $\#\mathcal{X}$ is finite, equal to n.

In this case, one can identify $\mathcal{P}(\mathcal{X})$ with the simplex $\mathbb{S}^{n-1} = \{p = (p_i)_{i=1,\dots,n} \in [0,1]^n : \sum_{i=1}^n p_i = 1\}$, by associating any probability distribution $\mu \in \mathcal{P}(\mathcal{X})$ to its weights $(\mu(\{x\}))_{x\in\mathcal{X}} \in \mathbb{S}^{n-1}$. We also identify the action space $\hat{A}(\mathcal{X})$ with $\mathcal{P}(A)^n$ by associating any probability kernel $\hat{a} \in \hat{A}(\mathcal{X})$ to $(\hat{a}(x))_{x\in\mathcal{X}} \in \mathcal{P}(A)^n$, and thus $\hat{A}(\mathcal{X})$ is naturally endowed with the product σ-algebra of the Wasserstein metric space $\mathcal{P}(A)$.

Lemma 2.3.1 *Suppose that* $\#\mathcal{X} = n < \infty$*. Then,* $\hat{F}$ *in* (2.3.2) *is a measurable function from* $\mathbb{S}^{n-1} \times \mathcal{P}(A)^n \times E^0$ *into* $\mathbb{S}^{n-1}$*,* $\hat{f}$ *in* (2.3.3) *is a real-valued measurable function on*

$\mathbb{S}^{n-1} \times \mathcal{P}(A)^n$. *Moreover, for any* $\xi \in L^0(\mathcal{G};\mathcal{X})$, *and* $\alpha \in \mathcal{A}$, *the* $\mathcal{P}(A)^n$*-valued process* $\hat{\alpha}$ *defined by* $\hat{\alpha}_t(x) = \mathcal{L}^0(\alpha_t | X_t^{\xi,\alpha} = x)$, $t \in \mathbb{N}$, $x \in \mathcal{X}$, *is* $\mathbb{F}^0$*-adapted.*

Proof. By Lemma 2.6.1, it is clear, by measurable composition, that we only need to prove that $\Psi : (\mu, \hat{a}) \in (\mathcal{P}(\mathcal{X}), \hat{A}(\mathcal{X})) \mapsto \mu \cdot \hat{a} \in \mathcal{P}(\mathcal{X} \times A)$ is measurable. However, in this finite state space case, $\mu \cdot \hat{a}$ is here simply equal to $\sum_{x \in \mathcal{X}} \mu(x)\hat{a}(x)$ and, thus Ψ is clearly measurable. □

In view of Lemma 2.3.1, the MDP with characteristics $(\mathcal{P}(\mathcal{X}) \equiv \mathbb{S}^{n-1}, \hat{A}(\mathcal{X}) \equiv \mathcal{P}(A)^n, \hat{F}, \hat{f}, \beta)$ is well-posed. Let us then denote by $\widehat{\mathcal{A}}$ the set of $\mathbb{F}^0$-adapted processes valued in $\mathcal{P}(A)^n$, and given $\kappa \in \widehat{\mathcal{A}}$, consider the controlled dynamics in $\mathbb{S}^{n-1}$

$$\mu_{t+1} = \hat{F}(\mu_t, \kappa_t, \varepsilon^0_{t+1}), \quad t \in \mathbb{N},\ \mu_0 = \mu \in \mathbb{S}^{n-1}, \tag{2.3.4}$$

the associated expected gain and value function

$$\widehat{V}^\kappa(\mu) \ = \ \mathbb{E}\Big[\sum_{t=0}^{\infty} \beta^t \hat{f}(\mu_t, \kappa_t)\Big], \quad \widehat{V}(\mu) \ = \ \sup_{\kappa \in \hat{\mathcal{A}}} \widehat{V}^\kappa(\mu). \tag{2.3.5}$$

We aim to prove the correspondence and equivalence between the CMKV-MDP and the MDP (2.3.4)-(2.3.5). From the derivation in (2.3.1)-(2.3.3) and by Lemma 2.3.1, we see that for any $\alpha \in \mathcal{A}$, $V^\alpha(\xi) = \widehat{V}^{\hat{\alpha}}(\mu)$, where $\mu = \mathcal{L}(\xi)$, and $\hat{\alpha} = \mathcal{R}^0_\xi(\alpha)$ where $\mathcal{R}^0_\xi$ is the relaxed operator

$$\begin{array}{rcl} \mathcal{R}^0_\xi : \mathcal{A} & \longrightarrow & \widehat{\mathcal{A}} \\ \alpha = (\alpha_t)_t & \longmapsto & \hat{\alpha} = (\hat{\alpha}_t)_t :\ \hat{\alpha}_t(x) = \mathcal{L}^0\big(\alpha_t | X_t^{\xi,\alpha} = x\big),\ \ t \in \mathbb{N},\ x \in \mathcal{X}. \end{array} \tag{2.3.6}$$

It follows that $V(\xi) \le \widehat{V}(\mu)$. In order to get the reverse inequality from the surjectivity of $\mathcal{R}^0_\xi$, we need again as in the no common noise case to make some randomization hypothesis. It turns out that when $\mathcal{X}$ is finite, this randomization hypothesis is simply reduced to the atomless property of Γ.

Lemma 2.3.2 *Assume that* Γ *is atomless, i.e.,* $\mathcal{G}$ *is rich enough. Then, any* $\xi \in L^0(\mathcal{G};\mathcal{X})$ *taking a countable number of values, satisfies* **Rand**(ξ, Γ).

Proof. Let S be a countable set s.t. $\xi \in S$ a.s., and $\mathbb{P}[\xi = x] > 0$ for all $x \in S$. Fix $x \in S$ and denote by $\mathbb{P}_x$ the probability "knowing $\xi = x$", i.e., $\mathbb{P}_x[B] := \frac{\mathbb{P}[B, \xi = x]}{\mathbb{P}[\xi = x]}$, for all $B \in \mathcal{F}$. It is clear that, endowing Ω with this probability, Γ is still atomless, and so there exists a $\mathcal{G}$-measurable random variable U_x that is uniform under $\mathbb{P}_x$. Then, the random variable $U := \sum_{x \in S} U_x \mathbf{1}_{\xi = x}$ is a $\mathcal{G}$-measurable uniform random variable under $\mathbb{P}_x$ for all $x \in S$, which implies that it is a uniform variable under $\mathbb{P}$, independent of ξ. □

Theorem 2.3.2 (Correspondance with the MDP on $\mathcal{P}(\mathcal{X})$ in the $\mathcal{X}$ finite case) *Assume that $\mathcal{G}$ is rich enough. Then $\mathcal{R}^0_\xi$ is surjective from $\mathcal{A}$ into $\widehat{\mathcal{A}}$, and $V(\xi) = \widehat{V}(\mu)$, for any $\mu \in \mathcal{P}(\mathcal{X})$, $\xi \in L^0(\mathcal{G};\mathcal{X})$ s.t. $\mu = \mathcal{L}(\xi)$. Moreover, for $\epsilon \geq 0$, if $\alpha^\epsilon \in \mathcal{A}$ is an ϵ-optimal control for $V(\xi)$, then $\mathcal{R}^0_\xi(\alpha^\epsilon) \in \widehat{\mathcal{A}}$ is an ϵ-optimal control for $\widehat{V}(\mu)$. Conversely, if $\hat{\alpha}^\epsilon \in \widehat{\mathcal{A}}$ is an ϵ-optimal control for $\widehat{V}(\mu)$, then any $\alpha^\epsilon \in (\mathcal{R}^0_\xi)^{-1}(\hat{\alpha}^\epsilon)$ is an ϵ-optimal control for $V(\xi)$. Consequently, an optimal control for $V(\xi)$ exists iff an optimal control for $\widehat{V}(\mu)$ exists.*

Proof. From the derivation in (2.3.4)-(2.3.6), we only need to prove the surjectivity of $\mathcal{R}^0_\xi$. Fix $\kappa \in \widehat{\mathcal{A}}$ and let $\boldsymbol{\pi}_t \in L^0((E^0)^t; \hat{A}(\mathcal{X}))$ be such that $\kappa_t = \boldsymbol{\pi}_t((\varepsilon^0_s)_{s\leq t})$. As $\mathcal{X}$ is finite, by definition of the σ-algebra on $\hat{A}(\mathcal{X})$, $\boldsymbol{\pi}_t$ can be seen as a measurable function in $L^0((E^0)^t \times \mathcal{X}; \mathcal{P}(A))$. Let $\phi \in L^0(A, \mathbb{R})$ be an embedding as in Lemma 2.6.2. By Lemma 2.6.1, we know that $\phi \star \boldsymbol{\pi}_t$ is in $L^0((E^0)^t \times \mathcal{X}; \mathcal{P}(\mathbb{R}))$. Given $m \in \mathcal{P}(\mathbb{R})$ we denote by F^{-1}_m the generalized inverse of its distribution function, and it is known that the mapping $m \in (\mathcal{P}(\mathbb{R}), \mathcal{W}) \mapsto F^{-1}_m \in (L^1_{caglad}(\mathbb{R}), \|\cdot\|_1)$ is an isometry and is thus measurable. Therefore, $F^{-1}_{\phi\star\boldsymbol{\pi}_t}$ is in $L^0((E^0)^t \times \mathcal{X}; (L^1_{caglad}(\mathbb{R}), \|\cdot\|_1))$. Finally, the mapping $(f, u) \in (L^1_{caglad}(\mathbb{R}), \|\cdot\|_1) \times ([0,1], \mathcal{B}([0,1])) \mapsto f(u) \in (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is measurable, since it is the limit of the sequence $n \sum_{i\in\mathbb{Z}} \mathbf{1}_{[\frac{i+1}{n}, \frac{i+2}{n})}(u) \int_{\frac{i}{n}}^{\frac{i+1}{n}} f(y)dy$ when $n \to \infty$. Therefore, the mapping

$$\begin{array}{rcl} \mathrm{a}_t : (E^0)^t \times \mathcal{X} \times [0,1] & \longrightarrow & A \\ ((e^0_s)_{s\leq t}, x, u) & \longmapsto & \phi^{-1} \circ F^{-1}_{\phi\star\boldsymbol{\pi}_t((e^0_s)_{s\leq t}, x)}(u) \end{array}$$

is measurable. We thus define, by induction, $\alpha_t := \mathrm{a}_t((\varepsilon^0_s)_{s\leq t}, X^{\xi,\alpha}_t, U_t)$. By construction and by the generalized inverse simulation method, it is clear that $\hat{\alpha}_t = \kappa_t$. □

Remark 2.3.3 We point out that when both state space $\mathcal{X}$ and action space A are finite, equipped with the metrics $d(x,x') := \mathbf{1}_{x\neq x'}$, $x, x' \in \mathcal{X}$ and $d_A(a,a') := \mathbf{1}_{a\neq a'}$, $a, a' \in A$, the transition function $\hat{F}$ and reward function $\hat{f}$ of the lifted MDP on $\mathcal{P}(\mathcal{X})$ inherits the Lipschitz condition ($\mathbf{HF_{lip}}$) and ($\mathbf{Hf_{lip}}$) used for the propagation of chaos. Indeed, it is known that the Wasserstein distance obtained from d (resp. d_A) coincides with twice the total variation distance, and thus to the $\mathbb{L}^1$ distance when naturally embedding $\mathcal{P}(\mathcal{X})$ (resp. $\mathcal{P}(A)$) in $[0,1]^{\#\mathcal{X}}$ (resp. $[0,1]^{\#A}$). Thus, embedding $\hat{A}(\mathcal{X})$ in $\mathcal{M}_{\#\mathcal{X},\#A}([0,1])$, the set of $\#\mathcal{X} \times \#A$ matrices with coefficients valued in $[0,1]$, we have

$$\|\hat{F}(\mu, \hat{a}, e^0), \hat{F}(\nu, \hat{a}', e^0)\|_1 \leq (1 + K_F)(2\|\mu - \mu'\|_1 + \sup_{x\in\mathcal{X}} \|\hat{a}_{x,\cdot} - \hat{a}'_{x,\cdot}\|_1).$$

We obtain a similar property for f. In other words, lifting the CMKV-MDP not only turns it into an MDP, but also its state and action spaces $[0,1]^{\#\mathcal{X}}$ and $[0,1]^{\#\mathcal{X}\times\#A}$ are

very standard, and its dynamic and reward are Lipschitz functions with factors of the order of K_F and K_f according to the norm $\|\cdot\|_1$. Thus, due to the standard nature of this MDP, most MDP algorithms can be applied and their speed will be simply expressed in terms of the original parameters of the CMKV-MDP, K_F and K_f. □

Remark 2.3.4 As in the no common noise case, the correspondence result in the finite state space case for $\mathcal{X}$ shows notably that the value function of the CMKV-MDP is law-invariant.

The general case (common noise and continuous state space $\mathcal{X}$) raises multiple issues for establishing the equivalence between CMKV-MDP and the lifted MDP on $\mathcal{P}(\mathcal{X})$. First, we have to endow the action space $\hat{A}(\mathcal{X})$ with a suitable σ-algebra for the lifted MDP to be well-posed: on the one hand, this σ-algebra has to be large enough to make the functions $\hat{F} : \mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X}) \times E^0 \to \mathcal{P}(\mathcal{X})$ and $\hat{f} : \mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X}) \to \mathbb{R}$ measurable, and on the other hand, it should be small enough to make the process $\hat{\alpha} = \mathcal{R}^0_\xi(\alpha)$ $\mathbb{F}^0$-adapted for any control $\alpha \in \mathcal{A}$ in the CMKV-MDP. Beyond the well-posedness issue of the lifted MDP, the second important concern is the surjectivity of the relaxed operator $\mathcal{R}^0_\xi$ from $\mathcal{A}$ into $\widehat{\mathcal{A}}$. Indeed, if we try to adapt the proof of Theorem 2.3.2 to the case of a continuous state space $\mathcal{X}$, the issue is that we cannot in general equip $\hat{A}(\mathcal{X})$ with a σ-algebra such that $L^0((E^0)^t; \hat{A}(\mathcal{X})) = L^0((E^0)^t \times \mathcal{X}; \mathcal{P}(A))$, and thus we cannot see $\boldsymbol{\pi}_t \in L^0((E^0)^t; \hat{A}(\mathcal{X}))$ as an element of $L^0((E^0)^t \times \mathcal{X}; \mathcal{P}(A))$, which is crucial because the control α (such that $\hat{\alpha} = \kappa$) is defined with α_t explicitly depending upon $\boldsymbol{\pi}_t((\varepsilon^0_s)_{s\le t}, X_t)$.

In the next section, we shall fix these measurability issues in the general case, and prove the correspondence between the CMKV-MDP and a general lifted MDP on $\mathcal{P}(\mathcal{X})$. □

2.4 General case and Bellman fixed point equation in $\mathcal{P}(\mathcal{X})$

We address the general case with common noise and possibly continuous state space $\mathcal{X}$, and our aim is to state the correspondence of the CMKV-MDP with a suitable lifted MDP on $\mathcal{P}(\mathcal{X})$ associated to a Bellman fixed point equation, characterizing the value function, and obtain as a by-product an ϵ-optimal control. We proceed as follows:

(i) We first introduce a well-posed lifted MDP on $\mathcal{P}(\mathcal{X})$ by enlarging the action space to $\mathcal{P}(\mathcal{X} \times A)$, and call $\tilde{V}$ the corresponding value function, which satisfies: $V(\xi) \le \tilde{V}(\mu)$, for $\mu = \mathcal{L}(\xi)$.

(ii) We then consider the Bellman equation associated to this well-posed lifted MDP on $\mathcal{P}(\mathcal{X})$, which admits a unique fixed point, called $V^\star$.

(iii) Under the randomization hypothesis for ξ, we show the existence of an ϵ-randomized feedback policy, which yields both an ϵ-randomized feedback control for the CMKV-MDP and an ϵ-optimal feedback control for $\tilde{V}$. This proves that $V(\xi) = \tilde{V}(\mu) = V^*(\mu)$, for $\mu = \mathcal{L}(\xi)$.

(iv) Under the condition that $\mathcal{G}$ is rich enough, we conclude that V is law-invariant and is equal to $\tilde{V} = V^\star$, hence satisfies the Bellman equation.

Finally, we show how to compute from the Bellman equation by value or policy iteration approximate optimal strategy and value function.

2.4.1 A general lifted MDP on $\mathcal{P}(\mathcal{X})$

We start again from the relation (2.3.1) describing the evolution of $\mu_t = \mathbb{P}^0_{X_t}$, $t \in \mathbb{N}$, for a state process $X_t = X_t^{\xi,\alpha}$ controlled by $\alpha \in \mathcal{A}$:

$$\mu_{t+1} = F(\cdot, \cdot, \mathbb{P}^0_{(X_t,\alpha_t)}, \cdot, \varepsilon^0_{t+1}) \star (\mathbb{P}^0_{(X_t,\alpha_t)} \otimes \lambda_\varepsilon), \quad a.s. \tag{2.4.1}$$

Now, instead of disintegrating as in Section 2.3, the conditional law of the pair (X_t, α_t), as $\mathbb{P}^0_{(X_t,\alpha_t)} = \mu_t \cdot \hat{\alpha}_t$ where $\hat{\alpha} = \mathcal{R}^0_\xi(\alpha)$ is the relaxed control in (2.3.6), we directly consider the control process $\boldsymbol{\alpha}_t = \mathbb{P}^0_{(X_t,\alpha_t)}$, $t \in \mathbb{N}$, which is $\mathbb{F}^0$-adapted (see Proposition 2.6.1), and valued in the space of probability measures $\boldsymbol{A} := \mathcal{P}(\mathcal{X} \times A)$, naturally endowed with the σ-algebra of its Wasserstein metric. Notice that this $\boldsymbol{A}$-valued control $\boldsymbol{\alpha}$ obtained from the CMKV-MDP has to satisfy by definition the marginal constraint $\mathrm{pr}_1 \star \boldsymbol{\alpha}_t = \mu_t$ at any time t. In order to tackle this marginal constraint, we shall rely on the following coupling results.

Lemma 2.4.1 (Measurable coupling)
There exists a measurable function $\zeta \in L^0(\mathcal{P}(\mathcal{X})^2 \times \mathcal{X} \times [0,1]; \mathcal{X})$ s.t. for any $(\mu, \mu') \in \mathcal{P}(\mathcal{X})$, and if $\xi \sim \mu$, then

- *$\zeta(\mu, \mu', \xi, U) \sim \mu'$, where U is an uniform random variable independent of ξ.*
- *(i) When $\mathcal{X} \subset \mathbb{R}$:*

$$\mathbb{E}\big[d(\xi, \zeta(\mu, \mu', \xi, U))\big] \;=\; \mathcal{W}(\mu, \mu').$$

(ii) In general when $\mathcal{X}$ Polish: $\forall\, \varepsilon > 0$, $\exists \eta > 0$ s.t.

$$\mathcal{W}(\mu, \mu') \,<\, \eta \;\Rightarrow\; \mathbb{E}\big[d(\xi, \zeta(\mu, \mu', \xi, U))\big] \,<\, \varepsilon.$$

Proof. See Appendix 2.6.2. □

Remark 2.4.1 Lemma 2.4.1 can be seen as a measurable version of the well-known coupling result in optimal transport, which states that given μ, $\mu' \in \mathcal{P}(\mathcal{X})$, there exists ξ and ξ' random variables with $\mathcal{L}(\xi) = \mu$, $\mathcal{L}(\xi') = \mu'$ such that $\mathcal{W}(\mu, \mu') = \mathbb{E}\big[d(\xi, \xi')\big]$. A similar measurable optimal coupling is proved in [28] under the assumption that there exists a transfer function realizing an optimal coupling between μ and μ'. However, such transfer function does not always exist, for instance when μ has atoms but not μ'. Lemma 2.4.1 builds a measurable coupling without making such assumption (essentially using the uniform variable U to randomize when μ has atoms). □

From the measurable coupling function ζ as in Lemma 2.4.1, we define the coupling projection $\mathbf{p} : \mathcal{P}(\mathcal{X}) \times \boldsymbol{A} \to \boldsymbol{A}$ by

$$\mathbf{p}(\mu, \mathbf{a}) = \mathcal{L}\big(\zeta(\mathrm{pr}_1 \star \mathbf{a}, \mu, \xi', U), \alpha_0\big), \quad \mu \in \mathcal{P}(\mathcal{X}), \mathbf{a} \in \boldsymbol{A},$$

where $(\xi', \alpha_0) \sim \mathbf{a}$, and U is a uniform random variable independent of ξ'.

Lemma 2.4.2 (Measurable coupling projection)
The coupling projection $\mathbf{p}$ *is a measurable function from* $\mathcal{P}(\mathcal{X}) \times \boldsymbol{A}$ *into* $\boldsymbol{A}$*, and for all* $(\mu, \mathbf{a}) \in \mathcal{P}(\mathcal{X}) \times \boldsymbol{A}$*:*

$$pr_1 \star \boldsymbol{p}(\mu, \mathbf{a}) = \mu, \quad \textit{and if} \quad pr_1 \star \mathbf{a} = \mu, \ \textit{then } \mathbf{p}(\mu, \mathbf{a}) = \mathbf{a}. \tag{2.4.2}$$

Proof. By construction, it is clear that $\zeta(\mu, \mu, \xi, U) = \xi$, and so relation (2.4.2) is obvious. The only result that is not trivial is the measurability of $\mathbf{p}$. Observe that $\mathbf{p}(\mu, \mathbf{a}) = g(\mu, \mathbf{a}, \cdot, \cdot, \cdot) \star (\mathbf{a} \otimes \mathcal{U}([0,1]))$ where g is the measurable function

$$\begin{aligned} g : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X} \times A) \times \mathcal{X} \times A \times [0,1] &\longrightarrow \mathcal{X} \times A \\ (\mu, \mathbf{a}, x, a, u) &\longmapsto (\zeta(\mathrm{pr}_1 \star \mathbf{a}, \mu, x, u), a) \end{aligned}$$

We thus conclude by Lemma 2.6.1. □

By using this coupling projection $\mathbf{p}$, we see that the dynamics (2.4.1) can be written as

$$\mu_{t+1} = \boldsymbol{F}(\mu_t, \boldsymbol{\alpha}_t, \varepsilon^0_{t+1}), \quad t \in \mathbb{N}, \tag{2.4.3}$$

where the function $\boldsymbol{F} : \mathcal{P}(\mathcal{X}) \times \boldsymbol{A} \times E^0 \to \mathcal{P}(\mathcal{X})$ defined by

$$\boldsymbol{F}(\mu, \mathbf{a}, e^0) = F(\cdot, \cdot, \mathbf{p}(\mu, \mathbf{a}), \cdot, e^0) \star \big(\mathbf{p}(\mu, \mathbf{a}) \otimes \lambda_\varepsilon\big),$$

is clearly measurable. Let us also define the measurable function $\tilde{f} : \mathcal{P}(\mathcal{X}) \times \boldsymbol{A} \to \mathbb{R}$ by

$$\tilde{f}(\mu, \mathbf{a}) = \int_{\mathcal{X}\times A} f(x, a, \mathbf{p}(\mu, \mathbf{a}))\mathbf{p}(\mu, \mathbf{a})(\mathrm{d}x, \mathrm{d}a).$$

The MDP with characteristics $(\mathcal{P}(\mathcal{X}), \boldsymbol{A} = \mathcal{P}(\mathcal{X} \times A), \tilde{F}, \tilde{f}, \beta)$ is then well-posed. Let us then denote by $\boldsymbol{\mathcal{A}}$ the set of $\mathbb{F}^0$-adapted processes valued in $\boldsymbol{A}$, and given an open-loop control $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$, consider the controlled dynamics

$$\mu_{t+1} = \boldsymbol{F}(\mu_t, \boldsymbol{\nu}_t, \varepsilon^0_{t+1}), \quad t \in \mathbb{N},\ \mu_0 = \mu \in \mathcal{P}(\mathcal{X}), \tag{2.4.4}$$

with associated expected gain/value function

$$\widetilde{V}^{\boldsymbol{\nu}}(\mu) = \mathbb{E}\Big[\sum_{t\in\mathbb{N}} \beta^t \tilde{f}(\mu_t, \boldsymbol{\nu}_t)\Big], \qquad \widetilde{V}(\mu) = \sup_{\boldsymbol{\nu}\in\boldsymbol{\mathcal{A}}} \widetilde{V}^{\boldsymbol{\nu}}(\mu). \tag{2.4.5}$$

Given $\xi \in L^0(\mathcal{G}; \mathcal{X})$, and $\alpha \in \mathcal{A}$, we set $\boldsymbol{\alpha} = \boldsymbol{\mathcal{L}}^0_\xi(\alpha)$, where $\boldsymbol{\mathcal{L}}^0_\xi$ is the lifted operator

$$\begin{array}{rcl} \boldsymbol{\mathcal{L}}^0_\xi : \mathcal{A} & \longrightarrow & \boldsymbol{\mathcal{A}} \\ \alpha = (\alpha_t)_t & \longmapsto & \boldsymbol{\alpha} = (\boldsymbol{\alpha}_t)_t :\ \boldsymbol{\alpha}_t = \mathbb{P}^0_{(X^{\xi,\alpha}_t, \alpha_t)},\ \ t \in \mathbb{N}. \end{array}$$

By construction from (2.4.3), we see that $\mu_t = \mathbb{P}^0_{X^{\xi,\alpha}_t}$, $t \in \mathbb{N}$, follows the dynamics (2.4.4) with the control $\boldsymbol{\nu} = \boldsymbol{\mathcal{L}}^0_\xi(\alpha) \in \boldsymbol{\mathcal{A}}$. Moreover, by the law of iterated conditional expectation, and the definition of $\tilde{f}$, the expected gain of the CMKV-MDP can be written as

$$\begin{aligned} V^\alpha(\xi) &= \mathbb{E}\Big[\sum_{t\in\mathbb{N}} \beta^t \mathbb{E}^0\big[f(X^{\xi,\alpha}_t, \alpha_t, \mathbb{P}^0_{(X^{\xi,\alpha}_t, \alpha_t)})\big]\Big] \\ &= \mathbb{E}\Big[\sum_{t\in\mathbb{N}} \beta^t \tilde{f}(\mathbb{P}^0_{X^{\xi,\alpha}_t}, \boldsymbol{\alpha}_t)\Big] = \widetilde{V}^{\boldsymbol{\alpha}}(\mu), \quad \text{with } \mu = \mathcal{L}(\xi). \end{aligned} \tag{2.4.6}$$

It follows that $V(\xi) \le \widetilde{V}(\mu)$, for $\mu = \mathcal{L}(\xi)$. Our goal is to prove the equality, which implies in particular that V is law-invariant, and to obtain as a by-product the corresponding Bellman fixed point equation that characterizes analytically the solution to the CMKV-MDP.

2.4.2 Bellman fixed point on $\mathcal{P}(\mathcal{X})$

We derive and study the Bellman equation corresponding to the general lifted MDP (2.4.4)-(2.4.5) on $\mathcal{P}(\mathcal{X})$.

By defining this MDP on the canonical space $(E^0)^{\mathbb{N}}$, we identify ε^0 with the canonical identity function in $(E^0)^{\mathbb{N}}$, and ε^0_t with the t-th projection in $(E^0)^{\mathbb{N}}$. We also denote by

$\theta : (E^0)^{\mathbb{N}} \to (E^0)^{\mathbb{N}}$ the shifting operator, defined by $\theta((e_t^0)_{t\in\mathbb{N}}) = (e_{t+1}^0)_{t\in\mathbb{N}}$. Via this identification, an open-loop control $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$ is a sequence $(\boldsymbol{\nu}_t)_t$ where $\boldsymbol{\nu}_t$ is a measurable function from $(E^0)^t$ into $\boldsymbol{A}$, with the convention that $\boldsymbol{\nu}_0$ is simply a constant in $\boldsymbol{A}$. Given $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$, and $e^0 \in E^0$, we define $\vec{\boldsymbol{\nu}}^{e^0} := (\vec{\boldsymbol{\nu}}_t^{e^0})_t \in \boldsymbol{\mathcal{A}}$, where $\vec{\boldsymbol{\nu}}_t^{e^0}(.) := \boldsymbol{\nu}_{t+1}(e^0, .)$, $t \in \mathbb{N}$. Given $\mu \in \mathcal{P}(\mathcal{X})$, and $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$, we denote by $(\mu_t^{\mu,\boldsymbol{\nu}})_t$ the solution to (2.4.4) on the canonical space, which satisfies the flow property

$$\big(\mu_{t+1}^{\mu,\boldsymbol{\nu}}, \boldsymbol{\nu}_{t+1}\big) \equiv \big(\mu_t^{\mu_1^{\mu,\boldsymbol{\nu}},\vec{\boldsymbol{\nu}}^{\varepsilon_1^0}(\theta(\varepsilon^0))}, \vec{\boldsymbol{\nu}}_t^{\varepsilon_1^0}(\theta(\varepsilon^0))\big), \quad t \in \mathbb{N}.$$

where $\equiv$ denotes the equality between functions on the canonical space. Given that $\varepsilon_1^0 \perp\!\!\!\perp \theta(\varepsilon^0) \overset{d}{=} \varepsilon^0$, we obtain that the expected gain of this MDP in (2.4.5) satisfies the relation

$$\widetilde{V}^{\boldsymbol{\nu}}(\mu) = \tilde{f}(\mu, \boldsymbol{\nu}_0) + \beta\mathbb{E}\Big[\widetilde{V}^{\vec{\boldsymbol{\nu}}^{\varepsilon_1^0}}(\mu_1^{\mu,\boldsymbol{\nu}})\Big]. \tag{2.4.7}$$

Let us denote by $L^\infty(\mathcal{P}(\mathcal{X}))$ the set of bounded real-valued functions on $\mathcal{P}(\mathcal{X})$, and by $L_m^\infty(\mathcal{P}(\mathcal{X}))$ the subset of measurable functions in $L^\infty(\mathcal{P}(\mathcal{X}))$. We then introduce the Bellman "operator" $\mathcal{T} : L_m^\infty(\mathcal{P}(\mathcal{X})) \to L^\infty(\mathcal{P}(\mathcal{X}))$ defined for any $W \in L_m^\infty(\mathcal{P}(\mathcal{X}))$ by:

$$[\mathcal{T}W](\mu) := \sup_{\mathbf{a}\in\boldsymbol{A}} \Big\{\tilde{f}(\mu, \mathbf{a}) + \beta\mathbb{E}\big[W\big(\tilde{F}(\mu, \mathbf{a}, \varepsilon_1^0)\big)\big]\Big\}, \quad \mu \in \mathcal{P}(\mathcal{X}). \tag{2.4.8}$$

Notice that the sup can a priori lead to a non measurable function $\mathcal{T}W$. This Bellman operator is consistent with the lifted MDP derived in Section 2.3, with characteristics $(\mathcal{P}(\mathcal{X}), \hat{A}(\mathcal{X}), \hat{F}, \hat{f}, \beta)$, although this MDP is not always well-posed. Indeed, its corresponding Bellman operator is well-defined as it only involves the random variable ε_1^0 at time 1, hence only requires the measurability of $e^0 \mapsto \hat{F}(\mu, \hat{a}, e^0)$, for any $(\mu, \hat{a}) \in \mathbb{P}(\mathcal{X}) \times \hat{A}(\mathcal{X})$ (which holds true), and it turns out that it coincides with $\mathcal{T}$.

Proposition 2.4.1 *For any $W \in L_m^\infty(\mathcal{P}(\mathcal{X}))$, and $\mu \in \mathcal{P}(\mathcal{X})$, we have*

$$[\mathcal{T}W](\mu) = \sup_{\hat{a}\in\hat{A}(\mathcal{X})} [\hat{\mathcal{T}}^{\hat{a}}W](\mu) = \sup_{\mathrm{a}\in L^0(\mathcal{X}\times[0,1];A)} [\mathbb{T}^{\mathrm{a}}W](\mu), \tag{2.4.9}$$

where $\hat{\mathcal{T}}^{\hat{a}}$ and $\mathbb{T}^{\mathrm{a}}$ are the operators defined on $L^\infty(\mathcal{P}(\mathcal{X}))$ by

$$\begin{aligned}
[\hat{\mathcal{T}}^{\hat{a}}W](\mu) &= \hat{f}(\mu, \hat{a}) + \beta\mathbb{E}\big[W\big(\hat{F}(\mu, \hat{a}, \varepsilon_1^0)\big)\big], \\
[\mathbb{T}^{\mathrm{a}}W](\mu) &= \mathbb{E}\Big[f(\xi, \mathrm{a}(\xi, U), \mathcal{L}(\xi, \mathrm{a}(\xi, U))) + \beta W\big(\mathbb{P}^0_{F(\xi,\mathrm{a}(\xi,U),\mathcal{L}(\xi,\mathrm{a}(\xi,U)),\varepsilon_1,\varepsilon_1^0)}\big)\Big]
\end{aligned} \tag{2.4.10}$$

for any $(\xi, U) \sim \mu \otimes \mathcal{U}([0,1])$ (it is clear that the right-hand side in (2.4.10) does not depend on the choice of such (ξ, U)). Moreover, we have

$$[\mathcal{T}W](\mu) = \sup_{\alpha_0\in L^0(\Omega;A)} \mathbb{E}\Big[f(\xi, \alpha_0, \mathcal{L}(\xi, \alpha_0)) + \beta W\big(\mathbb{P}^0_{F(\xi,\alpha_0,\mathcal{L}(\xi,\alpha_0),\varepsilon_1,\varepsilon_1^0)}\big)\Big]. \tag{2.4.11}$$

Proof. Fix $W \in L^\infty_m(\mathcal{P}(\mathcal{X}))$, and $\mu \in \mathcal{P}(\mathcal{X})$. Let $\mathbf{a}$ be arbitrary in $\boldsymbol{A}$. Since $\mathbf{p}(\mu,\mathbf{a})$ has first marginal equal to μ, there exists by assertion 3 in Lemma 2.6.2 a probability kernel $\hat{a} \in \hat{A}(\mathcal{X})$ such that $\mathbf{p}(\mu,\mathbf{a}) = \mu \cdot \hat{a}$. Therefore, $\tilde{F}(\mu,\mathbf{a},e^0) = \hat{F}(\mu,\hat{a},e^0)$, $\tilde{f}(\mu,\mathbf{a}) = \hat{f}(\mu,\hat{a})$, which implies that $[\mathcal{T}W](\mu) \le \sup_{\hat{a}\in\hat{A}(\mathcal{X})}[\hat{\mathcal{T}}^{\hat{a}}W](\mu) =: \mathbb{T}^1$.

Let us consider the operator $\mathcal{R}$ defined by

$$\begin{array}{rcl} \mathcal{R} : L^0(\mathcal{X}\times[0,1];A) & \longrightarrow & \hat{A}(\mathcal{X}) \\ \mathrm{a} & \longmapsto & \hat{a}: \ \hat{a}(x) = \mathcal{L}\big(\mathrm{a}(x,U)\big), \ \ x\in\mathcal{X},\ U\sim\mathcal{U}([0,1]), \end{array}$$

and notice that it is surjective from $L^0(\mathcal{X}\times[0,1];A)$ into $\hat{A}(\mathcal{X})$, by Lemma 2.22 in [44]. By noting that for any $\mathrm{a} \in L^0(\mathcal{X}\times[0,1];A)$, and $(\xi,U)\sim\mu\otimes\mathcal{U}([0,1])$, we have $\mathcal{L}\big(\xi,\mathrm{a}(\xi,U)\big) = \mu\cdot\mathcal{R}(\mathrm{a})$, it follows that $[\mathbb{T}^{\mathrm{a}}W](\mu) = [\hat{\mathcal{T}}^{\mathcal{R}(\mathrm{a})}W](\mu)$. Since $\mathcal{R}$ is surjective, this yields $\mathbb{T}^1 = \sup_{\mathrm{a}\in L^0(\mathcal{X}\times[0,1];A)}[\mathbb{T}^{\mathrm{a}}W](\mu) =: \mathbb{T}^2$.

Denote by $\mathbb{T}^3$ the right-hand-side in (2.4.11). It is clear that $\mathbb{T}^2 \le \mathbb{T}^3$. Conversely, let $\alpha_0 \in L^0(\Omega;A)$. We then set $\mathbf{a} = \mathcal{L}(\xi,\alpha_0) \in \mathcal{P}(\mathcal{X}\times A)$, and notice that the first marginal of $\mathbf{a}$ is μ. Thus, $\mathbf{p}(\mu,\mathbf{a}) = \mathcal{L}(\xi,\alpha_0)$, and so

$$\begin{array}{rcl} \tilde{f}(\mu,\mathbf{a}) = & \displaystyle\int_{\mathcal{X}\times A} f(x,a,\mathbf{p}(\mu,\mathbf{a}))\mathbf{p}(\mu,\mathbf{a})(\mathrm{d}x,\mathrm{d}a) & = \ \mathbb{E}\big[f(\xi,\alpha_0,\mathcal{L}(\xi,\alpha_0))\big] \\ \tilde{F}(\mu,\mathbf{a},\varepsilon_1^0) = & F(\cdot,\cdot,\mathbf{p}(\mu,\mathbf{a}),\cdot,\varepsilon_1^0)\star\big(\mathbf{p}(\mu,\mathbf{a})\otimes\lambda_\varepsilon\big) & = \ \mathbb{P}^0_{F(\xi,\alpha_0,\mathcal{L}(\xi,\alpha_0),\varepsilon_1,\varepsilon_1^0)}. \end{array}$$

We deduce that $\mathbb{T}^3 \le [\mathcal{T}W](\mu)$, which gives finally the equalities (2.4.9) and (2.4.11). □

We state the basic properties of the Bellman operator $\mathcal{T}$.

Proposition 2.4.2 *Assume that* $(\mathbf{H_{lip}})$ *holds true. (i) The operator* $\mathcal{T}$ *is monotone increasing: for* $W_1, W_2 \in L^\infty_m(\mathcal{P}(\mathcal{X}))$, *if* $W_1 \le W_2$, *then* $\mathcal{T}W_1 \le \mathcal{T}W_2$. *(ii) Furthermore, it is contracting on* $L^\infty_m(\mathcal{P}(\mathcal{X}))$ *with Lipschitz factor* β, *and admits a unique fixed point in* $L^\infty_m(\mathcal{P}(\mathcal{X}))$, *denoted by* $V^\star$, *hence solution to:*

$$V^\star = \ \mathcal{T}V^\star.$$

(iii) $V^\star$ *is* γ*-Hölder, with* $\gamma = \min\Big(1, \frac{|\ln\beta|}{\ln(2K_F)}\Big)$, *i.e. there exists some positive constant* $K_\star$ *(depending only on* K_F, K_f, β, *and explicit in the proof), such that*

$$\big|V^\star(\mu) - V^\star(\mu')\big| \le \ K_\star \mathcal{W}(\mu,\mu')^\gamma, \quad \forall \mu,\mu' \in \mathcal{P}(\mathcal{X}).$$

Proof. (i) The monotonicity of $\mathcal{T}$ is shown by standard arguments.

(ii) The β-contraction property of $\mathcal{T}$ is also obtained by standard arguments. Let us now prove by induction that the iterative sequence $V_{n+1} = \mathcal{T}V_n$, with $V_0 \equiv 0$ is well defined and such that

$$|V_n(\mu) - V_n(\mu')| \leq 2K_f \sum_{t=0}^{\infty} \beta^t \min((2K_F)^t \mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}}) \tag{2.4.12}$$

for all $n \in \mathbb{N}$. The property is obviously satisfied for $n = 0$. Assume that the property holds true for a fixed $n \in \mathbb{N}$, and let us prove it for $n + 1$. First of all, the inequality (3.3.3) implies that V_n is continuous, and thus $V_n \in L_m^\infty(\mathcal{P}(\mathcal{X})$. Therefore, $V_{n+1} = \mathcal{T}V_n$ is well defined. Fix $\mu, \mu' \in \mathcal{P}(\mathcal{X})$. In order to use the expression (2.4.11) of the Bellman operator $\mathcal{T}$, we consider an optimal coupling (ξ, ξ') of μ and μ', i.e. $\xi \sim \mu$, $\xi' \sim \mu'$, and $\mathbb{E}[d(\xi, \xi')] = \mathcal{W}(\mu, \mu')$, and fix an A-valued random variable α_0. Let us start with two preliminary estimations: under ($\mathbf{H_{lip}}$), we have

$$\begin{aligned}
\mathbb{E}\left[|f(\xi, \alpha_0, \mathcal{L}(\xi, \alpha_0)) - f(\xi', \alpha_0, \mathcal{L}(\xi', \alpha_0))|\right] &\leq K_f(\mathbb{E}[d(\xi, \xi')] + \mathcal{W}(\mathcal{L}(\xi, \alpha_0), \mathcal{L}(\xi', \alpha_0))) \\
&\leq K_f(\mathbb{E}[d(\xi, \xi')] + \mathbb{E}[d((\xi, \alpha_0), (\xi', \alpha_0))]) \\
&\leq 2K_f \mathbb{E}[d(\xi, \xi')] = 2K_f \mathcal{W}(\mu, \mu'). \quad (2.4.13)
\end{aligned}$$

Similarly, for $e^0 \in E^0$, we have

$$\mathbb{E}[d(F(\xi, \alpha_0, \mathcal{L}(\xi, \alpha_0), \varepsilon_1^1, e^0), F(\xi', \alpha_0, \mathcal{L}(\xi, \alpha_0), \varepsilon_1^1, e^0))] \leq 2K_F \mathcal{W}(\mu, \mu'). \tag{2.4.14}$$

Now, we prove the hereditary property. The definition of $\mathcal{T}$ and V_{n+1} combined with (2.4.13) and the induction hypothesis, imply that

$$|V_{n+1}(\mu) - V_{n+1}(\mu')| \quad \leq \quad 2K_f \mathcal{W}(\mu, \mu') + \beta \mathbb{E}[2K_f \sum \beta^t \min((2K_F^t \mathcal{W}(\mu_1, \mu_1'), \Delta_{\mathcal{X}})]$$

where $\mu_1 = \mathcal{L}^0(F(\xi, \alpha_0, \mathcal{L}(\xi, \alpha_0), \varepsilon_1^1, \varepsilon_1^0))$ and $\mu_1' = \mathcal{L}^0(F(\xi', \alpha_0, \mathcal{L}(\xi', \alpha_0), \varepsilon_1^1, \varepsilon_1^0))$. By Jensen's inequality and (2.4.14), we have

$$\begin{aligned}
& |V_{n+1}(\mu) - V_{n+1}(\mu')| \\
\leq \quad & 2K_f \min(\mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}}) + \beta 2K_f \sum \beta^t \min((2K_F^t \mathbb{E}\mathcal{W}(\mu_1, \mu_1'), \Delta_{\mathcal{X}}) \\
\leq \quad & 2K_f \min(\mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}}) + \beta 2K_f \sum \beta^t \min((2K_F^t 2K_F \mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}}) \\
\leq \quad & 2K_f \sum \beta^t \min((2K_F^t \mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}}).
\end{aligned}$$

This concludes the induction and proves that V_n is well defined and satisfies the inequality (3.3.3) for all $n \in \mathbb{N}$. As $\mathcal{T}$ is β-contracting, a standard argument from the proof of the Banach fixed point theorem shows that $(V_n)_n$ is a Cauchy sequence in the complete metric space $L_m^\infty(\mathcal{P}(\mathcal{X}))$, and therefore admits a limit $V^\star \in L_m^\infty(\mathcal{P}(\mathcal{X}))$. Notice that

$$V^\star(\mu) = \lim_n V_{n+1}(\mu) = \lim_n \mathcal{T}V_n(\mu) = \mathcal{T}V^\star$$

by continuity of the contracting operator $\mathcal{T}$.
(iii) By sending n to infinity in (3.3.3), we obtain

$$|V(\mu) - V(\mu')| \leq 2K_f \sum_{t=0}^{\infty} \beta^t \min((2K_F)^t \mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}}) =: S(\mathcal{W}(\mu, \mu')).$$

where $S(m) = 2K_f \sum_{t=0}^{\infty} \beta^t \min((2K_F)^t m, \Delta_{\mathcal{X}})$. If $2\beta K_F < 1$, we clearly have

$$S(m) \leq m \sum_{t=0}^{\infty} (\beta 2K_F)^t = \frac{m}{1 - \beta 2K_F},$$

and so V is 1-Hölder. Let us now study the case $2\beta K_F > 1$. In this case, in particular, $2K_F > 1$, thus $t \mapsto s_t(m)$ is nondecreasing, and so

$$\begin{aligned} S(m) \leq & \sum_{t=0}^{\infty} \int_t^{t+1} \beta^t \min\big[s_t(m); \Delta_{\mathcal{X}}\big] ds \\ \leq & \frac{1}{\beta} \sum_{t=0}^{\infty} \int_t^{t+1} \beta^s \min\big[m(2K_F)^s; \Delta_{\mathcal{X}}\big] ds \\ \leq & \frac{1}{\beta} \int_0^{\infty} e^{-|\ln \beta| s} \min\big[m e^{\ln(2K_F)s}; \Delta_{\mathcal{X}}\big] ds. \end{aligned}$$

Let $t_\star$ be such that $m e^{\ln(2K_F)t_\star} = \Delta_{\mathcal{X}}$, i.e. $t_\star = \frac{\ln(\Delta_{\mathcal{X}}/m)}{\ln(2K_F)}$. Then,

$$\begin{aligned} \int_0^{\infty} e^{-|\ln \beta| s} \min\big[m e^{\ln(2K_F)s}; \Delta_{\mathcal{X}}\big] ds \quad \leq & \quad m \int_0^{t^\star} e^{\ln(2K_F\beta)s} ds + \Delta_{\mathcal{X}} \int_{t^\star}^{\infty} e^{\ln(\beta)s} ds \\ \leq & \quad \frac{m}{\ln(2K_F\beta)} \Big[e^{\ln(2K_F\beta)t^\star} - 1\Big] - \frac{\Delta_{\mathcal{X}}}{\ln \beta} e^{\ln(\beta)t^\star}. \end{aligned}$$

After substituting $t_\star$ by its explicit value, we then obtain

$$\begin{aligned} & \int_0^{\infty} e^{-|\ln \beta| s} \min\big[m e^{\ln(2K_F)s}; \Delta_{\mathcal{X}}\big] ds \\ \leq & \quad \frac{m}{\ln(2K_F\beta)} \Big[\Big(\frac{\Delta_{\mathcal{X}}}{m}\Big)^{\frac{\ln(2K_F\beta)}{\ln(2K_F)}} - 1\Big] - \frac{\Delta_{\mathcal{X}}}{\ln \beta} \left(\frac{\Delta_{\mathcal{X}}}{m}\right)^{\frac{\ln(\beta)}{\ln(2K_F)}} \\ \leq & \quad \Delta_{\mathcal{X}} \Big(\frac{1}{\ln(2K_F\beta)} - \frac{1}{\ln \beta}\Big) \Big(\frac{\Delta_{\mathcal{X}}}{m}\Big)^{\frac{\ln(\beta)}{\ln(2K_F)}} - \frac{m}{\ln(2K_F\beta)} \\ \leq & \quad \mathcal{O}\Big(m^{\min\left[1, \frac{|\ln \beta|}{\ln(2K_F)}\right]}\Big). \end{aligned}$$

This implies that V is γ-Hölder and concludes the proof. □

Remark 2.4.2 In the proof of Proposition 2.4.2, one could also have proved that the set $\mathcal{S}$ of functions $W : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ such that

$$|W(\mu) - W(\mu')| \leq 2K_f \sum_{t=0}^{\infty} \beta^t \min((2K_F)^t \mathcal{W}(\mu, \mu'), \Delta_{\mathcal{X}})$$

for all $\mu, \mu' \in \mathcal{P}(\mathcal{X})$ is a complete metric space, as it is a closed set of the complete metric space $L_m^\infty(\mathcal{P}(\mathcal{X}))$, and is stabilized by the contracting operator $\mathcal{T}$ (which is essentially proved by replacing V_n by W in the proof). One could then have invoked the Banach fixed point theorem on this set $\mathcal{S}$, implying the existence and uniqueness of the fixed point $V^\star$. Notice that this argument would not work if we considered, instead of $\mathcal{S}$, the set of γ-Hölder continuous functions. Indeed, while it is true that such set is stabilized by $\mathcal{T}$ (it essentially follows from (2.4.13) and (2.4.14)), the set of γ-Hölder continuous functions is not closed in $L_m^\infty(\mathcal{P}(\mathcal{X}))$ (and thus not a complete metric space): there might indeed exist a converging sequence of γ-Hölder continuous functions with multiplicative factors (in the Hölder property) tending toward infinity, such that the limit function is not γ-Hölder anymore. □

As a consequence of Proposition 2.4.2, we can easily show the following relation between the value function $\tilde{V}$ of the general lifted MDP, and the fixed point $V^\star$ of the Bellman operator.

Lemma 2.4.3 *For all $\mu \in \mathcal{P}(\mathcal{X})$, we have $\tilde{V}(\mu) \leq V^\star(\mu)$.*

Proof. From (2.4.7), we have

$$\begin{aligned}
&\inf_{\mu \in \mathcal{P}(\mathcal{X})} \left\{ V^\star(\mu) - \widetilde{V}^{\boldsymbol{\nu}}(\mu) \right\} \\
\geq\ &\inf_{\mu \in \mathcal{P}(\mathcal{X})} \left\{ \mathcal{T}V^\star(\mu) - \left(\tilde{f}(\mu, \boldsymbol{\nu}_0) + \beta \mathbb{E}\Big[V^\star(\mu_1^{\mu, \boldsymbol{\nu}}) \Big] \right) + \beta \mathbb{E}\Big[V^\star(\mu_1^{\mu,\boldsymbol{\nu}}) - \widetilde{V}^{\vec{\boldsymbol{\nu}}^{\varepsilon^0_1}}(\mu_1^{\mu,\boldsymbol{\nu}}) \Big] \right\} \\
\geq\ &\beta \mathbb{E}\Big[V^\star(\mu_1^{\mu,\boldsymbol{\nu}}) - \widetilde{V}^{\vec{\boldsymbol{\nu}}^{\varepsilon^0_1}}(\mu_1^{\mu,\boldsymbol{\nu}}) \Big] \geq \beta \inf_{\mu \in \mathcal{P}(\mathcal{X})} \left\{ V^\star(\mu) - \widetilde{V}^{\boldsymbol{\nu}}(\mu) \right\}.
\end{aligned}$$

This shows that $\inf_{\mu \in \mathcal{P}(\mathcal{X})}(V^\star(\mu) - \widetilde{V}^{\boldsymbol{\nu}}(\mu)) \geq 0$, hence

$$\widetilde{V}^{\boldsymbol{\nu}}(\mu) \leq V^\star(\mu) \quad \forall \mu \in \mathcal{P}(\mathcal{X}).$$

Taking the sup over $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$, we obtain the required result. □

We aim to prove rigorously the equality $\tilde{V} = V^\star$, i.e., the value function $\tilde{V}$ of the general lifted MDP satisfies the Bellman fixed point equation: $\tilde{V} = \mathcal{T}\tilde{V}$, and also to show

the existence of an ϵ-optimal control for $\tilde{V}$. Notice that it cannot be obtained directly from classical theory of MDP as we consider here open-loop controls $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$ while MDP usually deals with feedback controls on finite-dimensional spaces. Anyway, following the standard notation in MDP theory with state space $\mathcal{P}(\mathcal{X})$ and action space $\boldsymbol{A}$, and in connection with the Bellman operator in (3.3.2), we introduce, for $\boldsymbol{\pi} \in L^0(\mathcal{P}(\mathcal{X}); \boldsymbol{A})$ (the set of measurable functions from $\mathcal{P}(\mathcal{X})$ into $\boldsymbol{A}$) called (measurable) feedback policy, the so-called $\boldsymbol{\pi}$-Bellman operator $\mathcal{T}^{\boldsymbol{\pi}}$ on $L^\infty(\mathcal{P}(\mathcal{X}))$, defined for $W \in L^\infty(\mathcal{P}(\mathcal{X}))$ by

$$[\mathcal{T}^{\boldsymbol{\pi}} W](\mu) = \tilde{f}(\mu, \boldsymbol{\pi}(\mu)) + \beta \mathbb{E}\big[W\big(\tilde{F}(\mu, \boldsymbol{\pi}(\mu), \varepsilon_1^0)\big)\big], \quad \mu \in \mathcal{P}(\mathcal{X}). \tag{2.4.15}$$

As for the Bellman operator $\mathcal{T}$, we have the basic properties on the operator $\mathcal{T}^{\boldsymbol{\pi}}$.

Lemma 2.4.4 *Fix $\boldsymbol{\pi} \in L^0(\mathcal{P}(\mathcal{X}); \boldsymbol{A})$.*

(i) The operator $\mathcal{T}^{\boldsymbol{\pi}}$ is contracting on $L^\infty(\mathcal{P}(\mathcal{X}))$ with Lipschitz factor β, and admits a unique fixed point denoted $\tilde{V}^{\boldsymbol{\pi}}$.

(ii) Furthermore, it is monotone increasing: for $W_1, W_2 \in L^\infty(\mathcal{P}(\mathcal{X}))$, if $W_1 \leq W_2$, then $\mathcal{T}^{\boldsymbol{\pi}} W_1 \leq \mathcal{T}^{\boldsymbol{\pi}} W_2$.

Remark 2.4.3 It is well-known from MDP theory that the fixed point $\tilde{V}^{\boldsymbol{\pi}}$ to the operator $\mathcal{T}^{\boldsymbol{\pi}}$ is equal to

$$\tilde{V}^{\boldsymbol{\pi}}(\mu) = \mathbb{E}\Big[\sum_{t \in \mathbb{N}} \tilde{f}(\mu_t, \boldsymbol{\pi}(\mu_t))\Big],$$

where (μ_t) is the MDP in (2.4.4) with the feedback and stationary control $\boldsymbol{\nu}^{\boldsymbol{\pi}} = (\boldsymbol{\nu}_t^{\boldsymbol{\pi}})_t \in \boldsymbol{\mathcal{A}}$ defined by $\boldsymbol{\nu}_t^{\boldsymbol{\pi}} = \boldsymbol{\pi}(\mu_t)$, $t \in \mathbb{N}$. In the sequel, we shall then identify by misuse of notation $\tilde{V}^{\boldsymbol{\pi}}$ and $\tilde{V}^{\boldsymbol{\nu}^{\boldsymbol{\pi}}}$ as defined in (2.4.5). □

Our ultimate goal being to solve the CMKV-MDP, we introduce a subclass of feedback policies for the lifted MDP.

Definition 2.4.1 (Lifted randomized feedback policy)
A feedback policy $\boldsymbol{\pi} \in L^0(\mathcal{P}(\mathcal{X}); \boldsymbol{A})$ is a lifted randomized feedback policy if there exists a measurable function $\mathfrak{a} \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)$, called randomized feedback policy, such that $\big(\xi, \mathfrak{a}(\mu, \xi, U)\big) \sim \boldsymbol{\pi}(\mu)$, for all $\mu \in \mathcal{P}(\mathcal{X})$, with $(\xi, U) \sim \mu \otimes \mathcal{U}([0,1])$.

Remark 2.4.4 Given $\mathfrak{a} \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)$, denote by $\boldsymbol{\pi}^{\mathfrak{a}} \in L^0(\mathcal{P}(\mathcal{X}); \boldsymbol{A})$ the associated lifted randomized feedback policy, i.e., $\boldsymbol{\pi}^{\mathfrak{a}}(\mu) = \mathcal{L}\big(\xi, \mathfrak{a}(\mu, \xi, U)\big)$, for $\mu \in \mathcal{P}(\mathcal{X})$, and $(\xi, U) \sim \mu \otimes \mathcal{U}([0,1])$. By definition of the π-Bellman operator $\mathcal{T}^{\boldsymbol{\pi}^{\mathfrak{a}}}$ in (2.4.15), and

observing that $\mathbf{p}(\mu, \boldsymbol{\pi}^{\mathfrak{a}}(\mu)) = \boldsymbol{\pi}^{\mathfrak{a}}(\mu) = \mathcal{L}\big(\xi, \mathrm{a}^{\mu}(\xi, U)\big)$, where we set $\mathrm{a}^{\mu} = \mathfrak{a}(\mu, \cdot, \cdot) \in L^0(\mathcal{X} \times [0,1] : A)$, we see (recalling the notation in (2.4.10)) that for all $W \in L^{\infty}(\mathcal{P}(\mathcal{X}))$,

$$[\mathcal{T}^{\boldsymbol{\pi}^{\mathfrak{a}}} W](\mu) = [\mathbb{T}^{\mathrm{a}^{\mu}} W](\mu), \quad \mu \in \mathcal{P}(\mathcal{X}). \tag{2.4.16}$$

On the other hand, let $\xi \in L^0(\mathcal{G}; \mathcal{X})$ be some initial state satisfying the randomization hypothesis $\mathbf{Rand}(\xi, \mathcal{G})$, and denote by $\alpha^{\mathfrak{a}} \in \mathcal{A}$ the randomized feedback stationary control defined by $\alpha_t^{\mathfrak{a}} = \mathfrak{a}(\mathbb{P}^0_{X_t}, X_t, U_t)$, where $X = X^{\xi, \alpha^{\mathfrak{a}}}$ is the state process in (**??**) of the CMKV-MDP, and $(U_t)_t$ is an i.i.d. sequence of uniform $\mathcal{G}$-measurable random variables independent of ξ. By construction, the associated lifted control $\boldsymbol{\alpha}^{\mathfrak{a}} = \boldsymbol{\mathcal{L}}^0_{\xi}(\alpha^{\mathfrak{a}})$ satisfies $\boldsymbol{\alpha}_{\boldsymbol{t}}^{\mathfrak{a}} = \mathbb{P}^0_{(X_t, \alpha_t^{\mathfrak{a}})} = \boldsymbol{\pi}^{\mathfrak{a}}(\mu_t)$, where $\mu_t = \mathbb{P}^0_{X_t}$, $t \in \mathbb{N}$. Denoting by $V^{\mathfrak{a}} := V^{\alpha^{\mathfrak{a}}}$ the associated expected gain of the CMKV-MDP, and recalling Remark 2.4.3, we see from (2.4.6) that $V^{\mathfrak{a}}(\xi) = \tilde{V}^{\boldsymbol{\nu}^{\boldsymbol{\pi}^{\mathfrak{a}}}}(\mu) = \tilde{V}^{\boldsymbol{\pi}^{\mathfrak{a}}}(\mu)$, where $\mu = \mathcal{L}(\xi)$. □

We show a verification type result for the general lifted MDP, and as a byproduct for the CMKV-MDP, by means of the Bellman operator.

Proposition 2.4.3 (Verification result)
Fix $\epsilon \geq 0$, and suppose that there exists an ϵ-optimal feedback policy $\boldsymbol{\pi}_{\epsilon} \in L^0(\mathcal{P}(\mathcal{X}); \boldsymbol{A})$ for $V^{\star}$ in the sense that

$$V^{\star} \leq \mathcal{T}^{\boldsymbol{\pi}_{\epsilon}} V^{\star} + \epsilon.$$

Then, $\boldsymbol{\nu}^{\pi_{\epsilon}} \in \boldsymbol{\mathcal{A}}$ is $\frac{\epsilon}{1-\beta}$-optimal for $\tilde{V}$, i.e., $\tilde{V}^{\boldsymbol{\pi}_{\epsilon}} \geq \tilde{V} - \frac{\epsilon}{1-\beta}$, and we have $\tilde{V} \geq V^{\star} - \frac{\epsilon}{1-\beta}$. Furthermore, if $\boldsymbol{\pi}_{\epsilon}$ is a lifted randomized feedback policy, i.e., $\boldsymbol{\pi}_{\varepsilon} = \boldsymbol{\pi}^{\mathfrak{a}_{\epsilon}}$, for some $\mathfrak{a}_{\varepsilon} \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)$, then under $\mathbf{Rand}(\xi, \mathcal{G})$, $\alpha^{\mathfrak{a}_{\epsilon}} \in \mathcal{A}$ is an $\frac{\epsilon}{1-\beta}$-optimal control for $V(\xi)$, i.e., $V^{\mathfrak{a}_{\epsilon}}(\xi) \geq V(\xi) - \frac{\epsilon}{1-\beta}$, and we have $V(\xi) \geq V^{\star}(\mu) - \frac{\epsilon}{1-\beta}$, for $\mu = \mathcal{L}(\xi)$.

Proof. Since $\tilde{V}^{\boldsymbol{\pi}_{\epsilon}} = \mathcal{T}^{\boldsymbol{\pi}_{\epsilon}} \tilde{V}^{\boldsymbol{\pi}_{\epsilon}}$, and recalling from Lemma 2.4.3 that $V^{\star} \geq \tilde{V} \geq \tilde{V}^{\boldsymbol{\pi}_{\epsilon}}$, we have for all $\mu \in \mathcal{P}(\mathcal{X})$,

$$\Big|(V^{\star} - \tilde{V}^{\boldsymbol{\pi}_{\epsilon}})(\mu)\Big| \leq \Big|\mathcal{T}^{\boldsymbol{\pi}_{\epsilon}}(V^{\star} - \tilde{V}^{\boldsymbol{\pi}_{\epsilon}})(\mu) + \epsilon\Big| \leq \beta \|V^{\star} - \tilde{V}^{\boldsymbol{\pi}_{\epsilon}}\| + \epsilon,$$

where we used the β-contraction property of $\mathcal{T}^{\boldsymbol{\pi}_{\epsilon}}$ in Lemma 2.4.4. We deduce that $\|V^{\star} - \tilde{V}^{\boldsymbol{\pi}_{\epsilon}}\| \leq \frac{\epsilon}{1-\beta}$, and then, $\tilde{V} \geq \tilde{V}^{\boldsymbol{\pi}_{\epsilon}} \geq V^{\star} - \frac{\epsilon}{1-\beta}$, which combined with $V^{\star} \geq \tilde{V}$, shows the first assertion. Moreover, if $\boldsymbol{\pi}_{\varepsilon} = \boldsymbol{\pi}^{\mathfrak{a}_{\epsilon}}$ is a lifted randomized feedback policy, then by Remark 2.4.4, and under $\mathbf{Rand}(\xi, \mathcal{G})$, we have $V^{\mathfrak{a}_{\epsilon}}(\xi) = \tilde{V}^{\boldsymbol{\pi}_{\epsilon}}(\mu)$. Recalling that $V(\xi) \leq \tilde{V}(\mu)$, and together with the first assertion, this proves the required result. □

Remark 2.4.5 If we can find for any $\epsilon > 0$, an ϵ-optimal lifted randomized feedback policy for $V^{\star}$, then according to Proposition 2.4.3, and under $\mathbf{Rand}(\xi, \mathcal{G})$, one could

restrict to randomized feedback policies in the computation of the optimal value $V(\xi)$ of the CMKV-MDP, i.e., $V(\xi) = \sup_{\mathfrak{a} \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)} V^{\mathfrak{a}}(\xi)$. Moreover, this would prove that $V(\xi) = \tilde{V}(\mu) = V^\star(\mu)$, hence V is law-invariant, and satisfies the Bellman fixed equation.

Notice that instead of proving directly the dynamic programming Bellman equation for V, we start from the fixed point solution $V^\star$ to the Bellman equation, and show via a verification result that V is indeed equal to $V^\star$, hence satisfies the Bellman equation.

By the formulation (2.4.9) of the Bellman operator in Proposition 2.4.1, and the fixed point equation satisfied by $V^\star$, we know that for all $\epsilon > 0$, and $\mu \in \mathcal{P}(\mathcal{X})$, there exists $\mathrm{a}_\epsilon^\mu \in L^0(\mathcal{X} \times [0,1]; A)$ such that

$$V^\star(\mu) \leq [\mathbb{T}^{\mathrm{a}_\epsilon^\mu} V^\star](\mu) + \epsilon. \tag{2.4.17}$$

The crucial issue is to prove that the mapping $(\mu, x, u) \mapsto \mathfrak{a}_\epsilon(\mu, x, u) := \mathrm{a}_\epsilon^\mu(x, u)$ is measurable so that it defines a randomized feedback policy $\mathfrak{a}_\epsilon \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)$, and an associated lifted randomized feedback policy $\boldsymbol{\pi}^{\mathfrak{a}_\epsilon}$. Recalling the relation (2.4.16), this would then show that $\boldsymbol{\pi}^{\mathfrak{a}_\epsilon}$ is a ϵ-optimal lifted randomized feedback policy for $V^\star$, and we could apply the verification result. □

We now address the measurability issue for proving the existence of an ϵ-optimal randomized feedback policy for $V^\star$. The basic idea is to construct as in (2.4.17) an ϵ-optimal $\mathrm{a}_\epsilon^\mu \in L^0(\mathcal{X} \times [0,1]; A)$ for $V^\star(\mu)$ when μ lies in a suitable finite grid of $\mathcal{P}(\mathcal{X})$, and then "patchs" things together to obtain an ε-optimal randomized feedback policy. This is made possible under some uniform continuity property of $V^\star$.

The next result provides a suitable discretization of the set of probability measures.

Lemma 2.4.5 (Quantization of $\mathcal{P}(\mathcal{X})$)
Fix $\eta > 0$. Then for each finite $\eta/2$-covering $\mathcal{X}_\eta$ of $\mathcal{X}$, one can construct a finite subset $\mathcal{M}_\eta$ of $\mathcal{P}(\mathcal{X})$, of size $N_\eta = n_\eta^{\#\mathcal{X}_\eta - 1}$, where n_η is a grid size of $[0,1]$, that is an η-covering of $\mathcal{P}(\mathcal{X})$.

Proof. As $\mathcal{X}$ is compact, there exists a finite subset $\mathcal{X}_\eta \subset \mathcal{X}$ such that $d(x, x_\eta) \leq \eta/2$ for all $x \in \mathcal{X}$, where x_η denotes the projection of x on $\mathcal{X}_\eta$. Given $\mu \in \mathcal{P}(\mathcal{X})$, and $\xi \sim \mu$, we denote by ξ_η the quantization, i.e., the projection of ξ on $\mathcal{X}_\eta$, and by μ_η the discrete law of ξ_η. Thus, $\mathbb{E}[d(\xi, \xi_\eta)] \leq \eta/2$, and therefore $\mathcal{W}(\mu, \mu_\eta) \leq \eta/2$. The probability measure μ_η lies in $\mathcal{P}(\mathcal{X}_\eta)$, which is identified with the simplex of $[0,1]^{\#\mathcal{X}_\eta}$. We then use another grid $G_\eta = \{\frac{i}{n_\eta} : i = 0, \ldots, n_\eta\}$ of $[0,1]$, and project its weights $\mu_\eta(y) \in [0,1]$, $y \in \mathcal{X}_\eta$, on G_η, in order to obtain another discrete probability measure μ_{η, n_η}. From the dual

Kantorovich representation of Wasserstein distance, it is easy to see that for n_η large enough, $\mathcal{W}(\mu_\eta, \mu_{\eta,n_\eta}) \leq \eta/2$, and so $\mathcal{W}(\mu, \mu_{\eta,n_\eta}) \leq \eta$. We conclude the proof by noting that μ_{η,n_η} belongs to the set $\mathcal{M}_\eta$ of probability measures on $\mathcal{X}_\eta$ with weights valued in the finite grid G_η, hence $\mathcal{M}_\eta$ is a finite set of $\mathcal{P}(\mathcal{X}_\eta)$, of cardinal $N_\eta = n_\eta^{\#\mathcal{X}_\eta - 1}$. □

Remark 2.4.6 Lemma 2.4.5 is actually a simple consequence of Prokhorov's theorem, but the above proof has some advantages:

- it provides an explicit construction of the quantization grid $\mathcal{M}_\eta$,
- it explicitly gives the size of the grid as a function of η, which is particularly useful for computing the time/space complexity of algorithms,
- this special grid simultaneously quantizes $\mathcal{P}(\mathcal{X})$ and $\mathcal{X}$, in the sense that the measures from $\mathcal{M}_\eta$ are all supported on the finite set $\mathcal{X}_\eta$. This is also useful for algorithms because for $\mu \in \mathcal{M}_\eta$, $\hat{a} \in \hat{A}(\mathcal{X})$, and $W \in L^\infty(\mathbb{R})$, the expression $\hat{\mathcal{T}}^{\hat{a}} W(\mu)$ only depends upon $(\hat{a}(x))_{x \in \mathcal{X}_\eta}$. Therefore, in the Bellman fixed point equation, the computation of an ϵ-argmax over the set $L^0(\mathcal{X}, \mathcal{P}(A))$ is reduced to a computation of an ϵ-argmax over the set $\mathcal{P}(A)^{\mathcal{X}_\eta}$, which is more tractable for a computer.

We can conclude this paragraph by showing the existence of an ϵ-optimal lifted randomized feedback policy for the general lifted MDP on $\mathcal{P}(\mathcal{X})$, and obtain as a by-product the corresponding Bellman fixed point equation for its value function and for the optimal value of the CMKV-MDP under randomization hypothesis.

Theorem 2.4.1 *Assume that* $(\mathbf{H_{lip}})$ *holds true. Then, for all* $\epsilon > 0$*, there exists a lifted randomized feedback policy* $\boldsymbol{\pi}^{\mathfrak{a}_\epsilon}$*, for some* $\mathfrak{a}_\epsilon \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)$*, that is* ϵ*-optimal for* $V^\star$*. Consequently, under* $\mathbf{Rand}(\xi, \mathcal{G})$*, the randomized feedback stationary control* $\alpha^{\mathfrak{a}_\epsilon} \in \mathcal{A}$ *is* $\frac{\epsilon}{1-\beta}$*-optimal for* $V(\xi)$*, and we have* $V(\xi) = \tilde{V}(\mu) = V^\star(\mu)$*, for* $\mu = \mathcal{L}(\xi)$*, which thus satisfies the Bellman fixed point equation.*

Proof. Fix $\epsilon > 0$, and given $\eta > 0$, consider a quantizing grid $\mathcal{M}_\eta = \{\mu^1, \ldots, \mu^{N_\eta}\} \subset \mathcal{P}(\mathcal{X})$ as in Lemma 2.4.5, and an associated partition C_η^i, $i = 1, \ldots, N_\eta$, of $\mathcal{P}(\mathcal{X})$, satisfying

$$C_\eta^i \subset B_\eta(\mu^i) := \Big\{\mu \in \mathcal{P}(\mathcal{X}) : \mathcal{W}(\mu, \mu^i) \leq \eta\Big\}, \quad i = 1, \ldots, N_\eta.$$

For any μ^i, $i = 1, \ldots, N_\eta$, and by (2.4.17), there exists $\mathrm{a}_\epsilon^i \in L^0(\mathcal{X} \times [0,1]; A)$ such that

$$V^\star(\mu^i) \leq [\mathbb{T}^{\mathrm{a}_\epsilon^i} V^\star](\mu^i) + \frac{\epsilon}{3}. \tag{2.4.18}$$

From the partition C_η^i, $i = 1, \ldots, N_\eta$ of $\mathcal{P}(\mathcal{X})$, associated to $\mathcal{M}_\eta$, we construct the function $\mathfrak{a}_\epsilon : \mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1] \to A$ as follows. Let h_1, h_2 be two measurable functions from $[0,1]$ into $[0,1]$, such that if $U \sim \mathcal{U}([0,1])$, then $(h_1(U), h_2(U)) \sim \mathcal{U}([0,1])^{\otimes 2}$. We then define, for all $\mu \in \mathcal{P}(\mathcal{X}), x \in \mathcal{X}, u \in [0,1]$,

$$\mathfrak{a}_\epsilon(\mu, x, u) \;=\; \mathrm{a}_\epsilon^i\big(\zeta(\mu, \mu^i, x, h_1(u)), h_2(u)\big), \quad \text{when } \mu \in C_\eta^i,\ i = 1, \ldots, N_\eta,$$

where ζ is the measurable coupling function defined in Lemma 2.4.1. Such function $\mathfrak{a}_\varepsilon$ is clearly measurable, i.e., $\mathfrak{a}_\varepsilon \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)$, and we denote by $\boldsymbol{\pi}_\varepsilon = \boldsymbol{\pi}^{\mathfrak{a}_\varepsilon}$ the associated lifted randomized feedback policy, which satisfies

$$[\mathcal{T}^{\boldsymbol{\pi}_\varepsilon} V^\star](\mu^i) = [\mathbb{T}^{\mathrm{a}_\epsilon^i} V^\star](\mu^i), \quad i = 1, \ldots, N_\eta, \tag{2.4.19}$$

by (2.4.16). Let us now check that such $\boldsymbol{\pi}_\epsilon$ yields an ϵ-optimal randomized feedback policy for η small enough. For $\mu \in \mathcal{P}(\mathcal{X})$, with $(\xi, U) \sim \mu \otimes \mathcal{U}([0,1])$, we set $U_1 := h_1(U)$, $U_2 := h_2(U)$, and define $\mu_\eta = \mu^i$, when $\mu \in C_\eta^i$, $i = 1, \ldots, N_\eta$, and $\xi_\eta := \zeta(\mu, \mu_\eta, \xi, U_1)$. Observe by Lemma 2.4.5 that $\mathcal{W}(\mu, \mu_\eta) \le \eta$, and by Lemma 2.4.1 that $(\xi_\eta, U_2) \sim \mu_\eta \otimes \mathcal{U}([0,1])$. We then write for any $\mu \in \mathcal{P}(\mathcal{X})$,

$$\begin{aligned}
[\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu) - V^\star(\mu) &= \Big([\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu) - [\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu_\eta)\Big) + \Big([\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu_\eta) - V^\star(\mu_\eta)\Big) \\
&\qquad + \big(V^\star(\mu_\eta) - V^\star(\mu)\big) \\
&\ge \Big([\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu) - [\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu_\eta)\Big) - \frac{\epsilon}{3} - \frac{\epsilon}{3},
\end{aligned} \tag{2.4.20}$$

where we used (3.3.6)-(2.4.19) and the fact that $|V^\star(\mu_\eta) - V^\star(\mu)| \le \epsilon/3$ for η small enough by uniform continuity of $V^\star$ in Proposition 2.4.2. Moreover, by observing that $\mathfrak{a}_\epsilon(\mu, \xi, U) = \mathfrak{a}_\epsilon(\mu_\eta, \xi_\eta, U_2) =: \alpha_0$, so that $\boldsymbol{\pi}_\epsilon(\mu) = \mathcal{L}(\xi, \alpha_0)$, $\boldsymbol{\pi}_\epsilon(\mu_\eta) = \mathcal{L}(\xi_\eta, \alpha_0)$, we have

$$\begin{aligned}
[\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu) &= \mathbb{E}\Big[f(Y) + \beta V^\star(\mathbb{P}^0_{F(Y, \varepsilon_1, \varepsilon_1^0)})\Big], \\
[\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu_\eta) &= \mathbb{E}\Big[f(Y_\eta) + \beta V^\star(\mathbb{P}^0_{F(Y_\eta, \varepsilon_1, \varepsilon_1^0)})\Big],
\end{aligned}$$

where $Y = (\xi, \alpha_0, \boldsymbol{\pi}_\epsilon(\mu))$, and $Y_\eta = (\xi_\eta, \alpha_0, \boldsymbol{\pi}_\epsilon(\mu_\eta))$. Under $(\mathbf{H_{lip}})$, by using the γ-Hölder property of $V^\star$ with constant $K_\star$ in Proposition 2.4.2, and by definition of the Wasserstein distance (recall that $\xi \sim \mu$, $\xi_\eta \sim \mu_\eta$), we then get

$$\begin{aligned}
&\big|[\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu) - [\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](\mu_\eta)\big| \\
\le\;& 2K\mathbb{E}\big[d(\xi, \xi_\eta)\big] + \beta K_\star \mathbb{E}\Big[\mathbb{E}\big[d\big(F(\xi, \alpha_0, \boldsymbol{\pi}_\epsilon(\mu), \varepsilon_1, e), F(\xi_\eta, \alpha_0, \boldsymbol{\pi}_\epsilon(\mu_\eta), \varepsilon_1, e)\big)^\gamma\big]_{e := \varepsilon_1^0}\Big] \\
\le\;& 2K\mathbb{E}\big[d(\xi, \xi_\eta)\big] + \beta K_\star \mathbb{E}\Big[\mathbb{E}\big[d\big(F(\xi, \alpha_0, \boldsymbol{\pi}_\epsilon(\mu), \varepsilon_1, e), F(\xi_\eta, \alpha_0, \boldsymbol{\pi}_\epsilon(\mu_\eta), \varepsilon_1, e)\big)\big]_{e := \varepsilon_1^0}\Big]^\gamma \\
\le\;& C\mathbb{E}\big[d(\xi, \xi_\eta)\big]^\gamma.
\end{aligned}$$

for some constant C independent from $\mathbb{E}\big[d(\xi,\xi_\eta)\big]$. Now, by the coupling Lemma 2.4.1, one can choose η small enough so that $C\mathbb{E}\big[d(\xi,\xi_\eta)\big]^\gamma \leq \frac{\epsilon}{3}$. Therefore, $\big|[\mathcal{T}^{\boldsymbol{\pi}_\epsilon}V^\star](\mu) - [\mathcal{T}^{\boldsymbol{\pi}_\epsilon}V^\star](\mu_\eta)\big| \leq \epsilon/3$, and, plugging into (2.4.20), we obtain $\mathcal{T}^{\boldsymbol{\pi}_\epsilon}V^\star(\mu) - V^\star(\mu) \geq -\epsilon$, for all $\mu \in \mathcal{P}(\mathcal{X})$, which means that $\boldsymbol{\pi}_\epsilon$ is ϵ-optimal for $V^\star$. The rest of the assertions in the Theorem follows from the verification result in Proposition 2.4.3. □

Remark 2.4.7 We stress the importance of the coupling Lemma in the construction of ϵ-optimal control in Theorem 3.3.1. Indeed, as we do not make any regularity assumption on F and f with respect to the "control arguments", the only way to make $[\mathcal{T}^{\boldsymbol{\pi}_\epsilon}V^\star](\mu)$ and $[\mathcal{T}^{\boldsymbol{\pi}_\epsilon}V^\star](\mu_\eta)$ close to each other is to couple terms to have the same control in F and f. This is achieved by turning μ into μ_η, ξ into ξ_η and set $\alpha_0 = \mathfrak{a}_\epsilon(\mu,\xi,U)$ $= \mathfrak{a}_\epsilon(\mu_\eta,\xi_\eta,U_2)$. Turning μ into μ_η is a simple quantization, but turning ξ into ξ_η is obtained thanks to the coupling Lemma. □

Remark 2.4.8 Theorem 3.3.1, although applying to a more general case than the results from Section 2.3, provides a weaker result. Indeed, it does not state that any control $\boldsymbol{\nu}$ for the lifted MDP $\tilde{V}(\mu)$ can be represented, i.e., associated to a control α for $V(\xi)$ such that $\boldsymbol{\alpha}_t := \mathbb{P}^0_{(X_t,\alpha_t)} = \boldsymbol{\nu}_t$ for all $t \in \mathbb{N}$. This theorem only implies that one can *restrict* the optimization to representable controls without changing the optimal value. Consequently, contrarily to Theorem 2.3.1 and Theorem 2.3.2, here one cannot conclude that an optimal control for $V(\xi)$ exists iff an optimal control for $\tilde{V}(\mu)$ exists. More precisely, it is possible that an optimal control $\boldsymbol{\nu}$ for $\tilde{V}(\mu)$ exists but cannot be associated to a control α for $V(\xi)$ such that $\boldsymbol{\alpha} = \boldsymbol{\nu}$, and thus the existence of an optimal control for $\tilde{V}(\mu)$ does not guarantee the existence of an optimal control for $V(\xi)$. □

Remark 2.4.9 From Theorems 3.3.1 under the condition that ($\mathbf{H_{lip}}$) holds true, the value function V of the CMKV-MDP is law-invariant, and the supremum in the Bellman fixed point equation for $V \equiv \tilde{V}$ with the operator $\mathcal{T}$ can be restricted to lifted randomized feedback policies, i.e.,

$$V = \mathcal{T}V = \sup_{\mathfrak{a}\in L^0(\mathcal{P}(\mathcal{X})\times\mathcal{X}\times[0,1];A)} \mathcal{T}^{\mathfrak{a}}V$$

where we set $\mathcal{T}^{\mathfrak{a}} := \mathcal{T}^{\boldsymbol{\pi}^{\mathfrak{a}}}$ equal to

$$[\mathcal{T}^{\mathfrak{a}}W](\mu) = \mathbb{E}\Big[f(Y^{\mathfrak{a}}(\mu,\xi,U)) + \beta W(\mathbb{P}^0_{F(Y^{\mathfrak{a}}(\mu,\xi,U),\varepsilon_1,\varepsilon_1^0)})\Big],$$

with $Y^{\mathfrak{a}}(\mu,x,u) := (x,\mathfrak{a}(\mu,x,u),\boldsymbol{\pi}^{\mathfrak{a}}(\mu))$, and $(\xi,U) \sim \mu\otimes\mathcal{U}([0,1])$. Notice that this Bellman fixed point equation is not the same as the Bellman fixed point equation obtained

by optimizing over feedback controls only (not randomized nor open-loop). Let us call V_f the associated optimal value. Then it is known that

$$V_f = \mathcal{T}_f V_f = \sup_{\mathfrak{a}_f \in L^0(\mathcal{P}(\mathcal{X})\times\mathcal{X};A)} \mathcal{T}^{\mathfrak{a}_f} V.$$

In other words, in the feedback case, the sup in the Bellman fixed point equation is only taken over (non-randomized) feedback policies. □

2.4.3 Open-loop vs feedback vs randomized controls

In this paper, we have mentioned different types of controls: open-loop controls, feedback controls, and randomized feedback controls. To fix ideas, let us address three problems:

- **Feedback problem**: Optimizing over stationary feedback controls. We note V_f the corresponding optimal value.
- **Open-loop problem**: Optimizing over open-loop controls. We note V_{ol} the corresponding optimal value.
- **Randomized feedback problem**: Optimizing over stationary randomized feedback controls. We note V_r the corresponding optimal value.

When do the optimal values coincide? Theorem 3.3.1 shows that V_r is the same as the optimal value when the optimization is performed over open-loop controls. Also, we clearly have $V_{ol}(\xi) \geq V_f(\xi)$. The problem is now to figure out if the inequalities can be strict. Examples 2.4.1 and 2.4.2 below illustrate that one can have $V_r(\xi) > V_f(\xi)$.

Example 2.4.1 (Feedback problem) *Let us take an example similar to Example 3.1 in [33]. Consider $\mathcal{X} = \{-1, 1\} = A$, $\varepsilon_1 \sim \mathcal{B}(1/2)$, $F(x, a, \nu, e, e^0) = ax$, $f(x, a, \nu) = -\mathcal{W}(pr_1 \star \nu, \mathcal{B}(1/2))$. In other words, the reward is maximal and equal to 0 when the law of the state is a Bernoulli$(1/2)$ on $\mathcal{X}$, and minimal equal to $-1/2$ when the law of the state is a Dirac (δ_{-1} or δ_1). Assume that $\Gamma \sim \mathcal{U}([0, 1])$ a.s.. Fix $\xi =: x$ to be deterministic. We perform the optimization over feedback controls. It is clear that the law of X_t will always be a Dirac, and thus the gain will be $V_f(\xi) = \sum_{t=0}^{\infty} \beta^t(-\frac{1}{2}) = -\frac{1}{2(1-\beta)}$ which is the worst possible gain.*

Example 2.4.2 (Randomized feedback problem) *Let us consider the same problem as in Example 2.4.1, and then optimize over stationary randomized feedback controls. The randomization allows to set $\alpha_0 = sgn(U - \frac{1}{2})$ and $\alpha_t = 1$ for $t \in \mathbb{N}_\star$. It is clear that the strategy is optimal and leads to a gain $V_r(\xi) = -\frac{1}{2}$.*

Put the big example with explicit resolution

2.4.4 Computing value function and ϵ-optimal strategies in CMKV-MDP

Having established the correspondence of our CMKV-MDP with lifted MDP on $\mathcal{P}(\mathcal{X})$, and the associated Bellman fixed point equation, we can (up to a simple discretization of the state space in the Bellman fixed point equation) design two methods for computing the value function and optimal strategies:

(a) Value iteration. We approximate the value function $V = \tilde{V} = V^\star$ by iteration from the Bellman operator: $V_{n+1} = \mathcal{T}V_n$, and at iteration N, we compute an approximate optimal randomized feedback policy $\mathfrak{a}_N$ by (recall Remark 2.4.9)

$$\mathfrak{a}_N \in \arg\max_{\mathfrak{a} \in L^0(\mathcal{P}(\mathcal{X})\times\mathcal{X}\times[0,1];A)} \mathcal{T}^{\mathfrak{a}} V_N.$$

From $\mathfrak{a}_N$, we then construct an approximate randomized feedback stationary control $\alpha^{\mathfrak{a}_N}$ according to the procedure described in Remark 2.4.4.

(b) Policy iteration. Starting from some initial randomized feedback policy $\mathfrak{a}_0 \in L^0(\mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1]; A)$, we iterate according to:

- Policy evaluation: we compute the expected gain $\tilde{V}^{\boldsymbol{\pi}^{\mathfrak{a}_0}}$ of the lifted MDP
- Greedy strategy: we compute

$$\mathfrak{a}_{k+1} \in \arg\max_{\mathfrak{a} \in L^0(\mathcal{P}(\mathcal{X})\times\mathcal{X}\times[0,1];A)} \mathcal{T}^{\mathfrak{a}} \tilde{V}^{\boldsymbol{\pi}^{\mathfrak{a}_k}}.$$

We stop at iteration K to obtain $\mathfrak{a}_K$, and then construct an approximate randomized feedback control $\alpha^{\mathfrak{a}_K}$ according to the procedure described in Remark 2.4.4.

Practical computation. Since a randomized feedback control α is a measurable function $\mathfrak{a}$ of $(\mathbb{P}^0_{X_t^{\xi,\alpha}}, X_t^{\xi,\alpha}, U_t)$, we would need to compute and store the (conditional) law of the state process, which is infeasible in practice when $\mathcal{X}$ is a continuous space. In this case, to circumvent this issue, a natural idea is to discretize the compact space $\mathcal{X}$ by considering a finite subset $\mathcal{X}_\eta = \{x^1, \ldots, x^{N_\eta}\} \subset \mathcal{X}$ associated with a partition B^i_η, $i = 1, \ldots, N_\eta$, of $\mathcal{X}$, satisfying: $B^i_\eta \subset \big\{x \in \mathcal{X} : d(x, x^i) \leq \eta\big\}$, $i = 1, \ldots, N_\eta$, with $\eta > 0$. For any $x \in \mathcal{X}$, we denote by $[x]_\eta$ (or simply x_η) its projection on $\mathcal{X}_\eta$, defined by: $x_\eta = x^i$, for $x \in B^i_\eta$, $i = 1, \ldots, N_\eta$.

Definition 2.4.2 (Discretized CMKV-MDP) *Fix $\eta > 0$. Given $\xi \in L^0(\mathcal{G}; \mathcal{X}_\eta)$, and a control $\alpha \in \mathcal{A}$, we denote by $X^{\eta,\xi,\alpha}$ the McKean-Vlasov MDP on $\mathcal{X}_\eta$ given by*

$$X^{\eta,\xi,\alpha}_{t+1} = \big[F(X^{\eta,\xi,\alpha}_t, \alpha_t, \mathbb{P}^0_{(X^{\eta,\xi,\alpha}_t,\alpha_t)}, \varepsilon_{t+1}, \varepsilon^0_{t+1})\big]_\eta, \quad t \in \mathbb{N}, \;\; X^{\eta,\xi,\alpha}_0 = \xi,$$

i.e., obtained by projecting the state on $\mathcal{X}_\eta$ after each application of the transition function F. The associated expected gain V_η^α is defined by

$$V_\eta^\alpha(\xi) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \beta^t f\big(X_t^{\eta,\xi,\alpha}, \alpha_t, \mathbb{P}^0_{(X_t^{\eta,\xi,\alpha},\alpha_t)}\big)\Big].$$

Notice that the (conditional) law of the discretized CMKV-MDP on $\mathcal{X}_\eta$ is now valued in a finite-dimensional space (the simplex of $[0,1]^{N_\eta}$), which makes the computation of the associated randomized feedback control accessible, although computationally challenging due to the high-dimensionality (and beyond the scope of this paper). The next result states that an ϵ-optimal randomized feedback control in the initial CMKV-MDP can be approximated by a randomized feedback control in the discretized CMKV-MDP.

Proposition 2.4.4 *Assume that $\mathcal{G}$ is rich enough and $(\mathbf{H_{lip}})$ holds true. Fix $\xi \in L^0(\mathcal{G};\mathcal{X})$. Given $\eta > 0$, let us define ξ_η the projection of ξ on $\mathcal{X}_\eta$. As $\mathbf{Rand}(\xi_\eta, \mathcal{G})$ holds true, let us consider an i.i.d. sequence $(U_{\eta,t})_{t\in\mathbb{N}}$ of $\mathcal{G}$-measurable uniform variables independent of ξ_η. For $\epsilon > 0$, let $\mathfrak{a}_\epsilon$ be a randomized feedback policy that is ϵ-optimal for the Bellman fixed point equation satisfied by V. Finally, let $\alpha^{\eta,\epsilon}$ be the randomized feedback control in the discretized CMKV-MDP recursively defined by $\alpha_t^{\eta,\epsilon} = \mathfrak{a}_\epsilon(\mathbb{P}^0_{X_t^{\eta,\epsilon}}, X_t^{\eta,\epsilon}, U_{\eta,t})$, $t \in \mathbb{N}$, where we set $X_t^{\eta,\epsilon} := X_t^{\eta,\xi_\eta,\alpha^{\epsilon,\eta}}$. Then the control $\alpha^{\eta,\epsilon}$ is $\mathcal{O}(\eta^\gamma + \epsilon)$-optimal for the CMKV-MDP X with initial state ξ, where $\gamma = \min\big(1, \frac{|\ln\beta|}{(\ln 2K)_+}\big)$.*

Proof. *Step 1.* Let us show that

$$\sup_{\alpha\in\mathcal{A}} \sum_{t=0}^{\infty} \beta^t \mathbb{E}\big[d(X_t^{\xi,\alpha}, X_t^{\eta,\xi_\eta,\alpha})\big] \le C\eta^\gamma, \tag{2.4.21}$$

for some constant C that depends only on K, β and γ. Indeed, notice by definition of the projection on $\mathcal{X}_\eta$, and by a simple conditioning argument that for all $\alpha \in \mathcal{A}$, and $t \in \mathbb{N}$,

$$\mathbb{E}\big[d(X_{t+1}^{\xi,\alpha}, X_{t+1}^{\eta,\xi_\eta,\alpha})\big] \le \eta + \mathbb{E}\big[\Delta\big(X_t^{\xi,\alpha}, X_t^{\eta,\xi_\eta,\alpha}, \alpha_t, \mathbb{P}^0_{(X_t^{\xi,\alpha},\alpha_t)}, \mathbb{P}^0_{(X_t^{\eta,\xi_\eta,\alpha},\alpha_t)}, \varepsilon^0_{t+1}\big)\big],$$

where

$$\Delta(x,x',a,\nu,\nu',e^0) = \mathbb{E}[d(F(x,a,\nu,\varepsilon_{t+1},e^0), F(x',a,\nu',\varepsilon_{t+1},e^0))].$$

Under $(\mathbf{H_{lip}})$, we then get

$$\begin{aligned}\mathbb{E}\big[d(X_{t+1}^{\xi,\alpha}, X_{t+1}^{\eta,\xi_\eta,\alpha})\big] &\le \eta + K\mathbb{E}\Big[d(X_t^{\xi,\alpha}, X_t^{\eta,\xi_\eta,\alpha}) + \mathcal{W}(\mathbb{P}^0_{X_t^{\xi,\alpha}}, \mathbb{P}^0_{X_t^{\eta,\xi_\eta,\alpha}})\Big] \\ &\le \eta + 2K\mathbb{E}\big[d(X_t^{\xi,\alpha}, X_t^{\eta,\xi_\eta,\alpha})\big],\end{aligned}$$

by the same argument as in (**??**). Hence, the sequence $(\mathbb{E}\big[d(X_t^{\xi,\alpha}, X_t^{\eta,\xi_\eta,\alpha})\big])_{t\in\mathbb{N}}$ satisfies the same type of induction inequality as in (**??**) in Theorem **??** with η instead of M_N, and thus the same derivation leads to the required result (2.4.21). From the Lipschitz condition on f, we deduce by the same arguments as in (**??**) in Lemma **??** that

$$\sup_{\alpha\in\mathcal{A}} \big|V^\alpha(\xi_\eta) - V_\eta^\alpha(\xi_\eta)\big| = \mathcal{O}(\eta^\gamma). \tag{2.4.22}$$

Step 2. Denote by $\mu = \mathcal{L}(\xi)$, and $\mu_\eta = \mathcal{L}(\xi_\eta)$, and observe that $\mathcal{W}(\mu,\mu_\eta) \le \mathbb{E}[d(\xi,\xi_\eta)] \le \eta$. We write

$$\begin{aligned} V^{\alpha^{\eta,\epsilon}}(\xi) - V(\xi) &= \big[V^{\alpha^{\eta,\epsilon}}(\xi) - V^{\alpha^{\eta,\epsilon}}(\xi_\eta)\big] + \big[V^{\alpha^{\eta,\epsilon}}(\xi_\eta) - V_\eta^{\alpha^{\eta,\epsilon}}(\xi_\eta)\big] \\ &\quad + \big[V_\eta^{\alpha^{\eta,\epsilon}}(\xi_\eta) - V(\xi_\eta)\big] + \big[V(\xi_\eta) - V(\xi)\big] =: I_1 + I_2 + I_3 + I_4. \end{aligned}$$

The first and last terms I_1 and I_4 are smaller than $\mathcal{O}(\eta^\gamma)$ by the γ-Hölder property of V^α and V in Lemma **??**. By (2.4.22), the second term I_2 is of order $\mathcal{O}(\eta^\gamma)$ as well for η small enough. Regarding the third term I_3, notice that by definition, $V_\eta^{\alpha^{\eta,\epsilon}}(\xi_\eta)$ corresponds to the gain associated to the randomized feedback policy $\mathfrak{a}_\epsilon$ for the discretized CMKV-MDP. Denote by $\boldsymbol{\pi}_\epsilon$ the lifted randomized feedback policy associated to $\mathfrak{a}_\epsilon$, and recall by Remark 2.4.4 the identification with the lifted MDP: $V_\eta^{\alpha^{\eta,\epsilon}}(\xi') = \tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon}(\mu')$, $\mu' = \mathcal{L}(\xi')$, where $\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon}$ is the expected gain of the lifted MDP associated to the discretized CMKV-MDP, hence fixed point of the operator

$$[\mathcal{T}_\eta^{\mathfrak{a}_\epsilon} W](\mu') = \mathbb{E}\Big[f(Y^{\mathfrak{a}_\epsilon}(\mu',\xi',U)) + \beta W\big(\mathbb{P}^0_{\left[F(Y^{\mathfrak{a}_\epsilon}(\mu',\xi',U),\varepsilon_1,\varepsilon_1^0)\right]_\eta}\big)\Big],$$

$Y^{\mathfrak{a}}(\mu,x,u) = (x,\mathfrak{a}(\mu,x,u),\boldsymbol{\pi}^{\mathfrak{a}}(\mu))$ and $(\xi',U) \sim \mu'\otimes\mathcal{U}([0,1])$. Recalling that $V(\xi') = \tilde{V}(\mu')$, $\mu' = \mathcal{L}(\xi')$, with $\tilde{V}$ fixed point to the Bellman operator $\mathcal{T}$, it follows that

$$\begin{aligned} I_3 = \tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon}(\mu_\eta) - \tilde{V}(\mu_\eta) &= \Big([\mathcal{T}_\eta^{\mathfrak{a}_\epsilon}\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon}](\mu_\eta) - [\mathcal{T}_\eta^{\mathfrak{a}_\epsilon}\tilde{V}](\mu_\eta)\Big) + \Big([\mathcal{T}_\eta^{\mathfrak{a}_\epsilon}\tilde{V}](\mu_\eta) - [\mathcal{T}^{\mathfrak{a}_\epsilon}\tilde{V}](\mu_\eta)\Big) \\ &\quad + \Big([\mathcal{T}^{\mathfrak{a}_\epsilon}\tilde{V}](\mu_\eta) - \tilde{V}(\mu_\eta)\Big) =: I_3^1 + I_3^2 + I_3^3. \end{aligned}$$

By definition of $\mathfrak{a}_\epsilon$, we have $|I_3^3| \le \epsilon$. For I_3^2 notice that the only difference between the operators $\mathcal{T}_\eta^{\mathfrak{a}_\epsilon}$ and $\mathcal{T}^{\mathfrak{a}_\epsilon}$ is that F is projected on $\mathcal{X}_\eta$. Thus,

$$\Big|[\mathcal{T}_\eta^{\mathfrak{a}_\epsilon}\tilde{V}](\mu_\eta) - [\mathcal{T}^{\mathfrak{a}_\epsilon}\tilde{V}](\mu_\eta)\Big| \le \beta\mathbb{E}\Big[\big|\tilde{V}\big(\mathbb{P}^0_{[F(Y_\eta,\varepsilon_1,\varepsilon_1^0)]_\eta}\big) - \tilde{V}\big(\mathbb{P}^0_{F(Y_\eta,\varepsilon_1,\varepsilon_1^0)}\big)\big|\Big],$$

where $Y_\eta = (\xi_\eta, \mathfrak{a}_\epsilon(\mu,\xi_\eta,U), \boldsymbol{\pi}_\epsilon(\mu_\eta))$. It is clear by definition of the Wasserstein distance and the projection on $\mathcal{X}_\eta$ that

$$\mathcal{W}\big(\mathbb{P}^0_{[F(Y_\eta,\varepsilon_1,\varepsilon_1^0)]_\eta}, \mathbb{P}^0_{F(Y_\eta,\varepsilon_1,\varepsilon_1^0)}\big) \le \mathbb{E}^0[d(F(Y_\eta,\varepsilon_1,\varepsilon_1^0),[F(Y_\eta,\varepsilon_1,\varepsilon_1^0)]_\eta)] \le \eta.$$

From the γ-Hölder property of $\tilde{V}$ in Proposition 2.4.2, we deduce that $I_3^2 = \mathcal{O}(\eta^\gamma)$. Finally, for I_3^1, since $\mathcal{T}_\eta^{\mathfrak{a}_\epsilon}$ is a β-contracting operator on $(L^\infty(\mathcal{M}_\eta), \|\cdot\|_{\eta,\infty})$, we have

$$\big|[\mathcal{T}_\eta^{\mathfrak{a}_\epsilon}\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon}](\mu_\eta) - [\mathcal{T}_\eta^{\mathfrak{a}_\epsilon}\tilde{V}](\mu_\eta)\big| \leq \beta\|\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon} - \tilde{V}\|_{\eta,\infty},$$

and thus $|\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon}(\mu_\eta) - \tilde{V}(\mu_\eta)| = |I_3| \leq |I_3^1| + |I_3^2| + |I_3^3| \leq \beta\|\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon} - \tilde{V}\|_{\eta,\infty} + \mathcal{O}(\eta^\gamma + \epsilon)$. Taking the sup over $\mu_\eta \in \mathcal{M}_\eta$ on the left, we obtain that $\|\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon} - \tilde{V}\|_{\eta,\infty} \leq \frac{1}{1-\beta}\mathcal{O}(\eta^\gamma + \epsilon) = \mathcal{O}(\eta^\gamma + \epsilon)$, and we conclude that $|I_3| \leq \|\tilde{V}_\eta^{\boldsymbol{\pi}_\epsilon} - \tilde{V}\|_{\eta,\infty} \leq \mathcal{O}(\eta^\gamma + \epsilon)$, which ends the proof. □

Remark 2.4.10 (Q function) In view of the Bellman fixed point equation satisfied by the value function V of the CMKV-MDP in terms of randomized feedback policies, let us introduce the corresponding state-action value function Q defined on $\mathcal{P}(\mathcal{X}) \times \hat{A}(\mathcal{X})$ by

$$Q(\mu, \hat{a}) = [\hat{\mathcal{T}}^{\hat{a}}V](\mu) \;=\; \hat{f}(\mu, \hat{a}) + \beta\mathbb{E}\big[V\big(\hat{F}(\mu, \hat{a}, \varepsilon_1^0)\big)\big],$$

From Proposition 2.4.1, and since $V = \mathcal{T}V$, we recover the standard connection between the value function and the state-action value function, namely $V(\mu) = \sup_{\hat{a}\in\hat{A}(\mathcal{X})} Q(\mu, \hat{a})$, from which we obtain the Bellman equation for the Q function:

$$Q(\mu, \hat{a}) = \hat{f}(\mu, \hat{a}) + \beta\mathbb{E}\Big[\sup_{\hat{a}'\in\hat{A}(\mathcal{X})} Q\big(\mu_1^{\hat{a}}, \hat{a}'\big)\Big], \tag{2.4.23}$$

where we set $\mu_1^{\hat{a}} = \hat{F}(\mu, \hat{a}, \varepsilon_1^0)$. Notice that this Q-Bellman equation extends the equation in [33] (see their Theorem 3.1) derived in the no common noise case and when there is no mean-field dependence with respect to the law of the control. The Bellman equation (2.4.23) is the starting point in a model-free framework when the state transition function is unknown (in other words in the context of reinforcement learning) for the design of Q-learning algorithms in order to estimate the Q-value function by Q_n, and then to compute a relaxed control by

$$\hat{a}_n^\mu \in \arg\max_{\hat{a}\in\hat{A}(\mathcal{X})} Q_n(\mu, \hat{a}), \quad \mu \in \mathcal{P}(\mathcal{X}).$$

From Lemma 2.22 [44], one can associate to such probability kernel $\hat{a}_n^\mu$, a function $\mathfrak{a}_n : \mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1] \to A$, such that $\mathcal{L}(\mathfrak{a}_n(\mu, x, U)) = \hat{a}_n^\mu(x)$, $\mu \in \mathcal{P}(\mathcal{X})$, $x \in \mathcal{X}$, where U is an uniform random variable. In practice, one has to discretize the state space $\mathcal{X}$ as in Definition 2.4.2, and then to quantize the space $\mathcal{P}(\mathcal{X})$ as in Lemma 2.4.5 in order to reduce the learning problem to a finite-dimensional problem for the computation of an approximate optimal randomized feedback policy $\mathfrak{a}_n$ for the CMKV-MDP. □

2.5 Conclusion

We have developed a theory for mean-field Markov decision processes with common noise and open-loop controls, called CMKV-MDP, for general state space and action space. Such problem is motivated and shown to be the asymptotic problem of a large population of cooperative agents under mean-field interaction controlled by a social planner/influencer, and we provide a rate of convergence of the N-agent model to the CMKV-MDP. We prove the correspondence of CMKV-MDP with a general lifted MDP on the space of probability measures, and emphasize the role of relaxed control, which is crucial to characterize the solution via the Bellman fixed point equation. Approximate randomized feedback controls are obtained from the Bellman equation in a model-based framework, and future work under investigation will develop algorithms in a model-free framework, in other words in the context of reinforcement learning with many interacting and cooperative agents.

2.6 Appendix

2.6.1 Some useful results on conditional law

Lemma 2.6.1 *Let $(S,\mathcal{S})$, $(T,\mathcal{T})$, and $(U,\mathcal{U})$ be three measurable spaces, and $F \in L^0((S,\mathcal{S})\times(T,\mathcal{T});(U,\mathcal{U}))$ be a measurable function, then the function $\hat{F} : (\mathcal{P}(S),\mathcal{C}(S))\times(T,\mathcal{T}) \to (\mathcal{P}(U),\mathcal{C}(U))$ given by $\hat{F}(\mu,x) := F(\cdot,x)\star\mu$ is measurable.*

Proof. This follows from the measurability of the maps:

- $x \in (S,\mathcal{S}) \mapsto \delta_x \in (\mathcal{P}(S),\mathcal{C}(S))$,
- $(\mu,\nu) \in (\mathcal{P}(S),\mathcal{C}(S)) \times (\mathcal{P}(T),\mathcal{C}(T)) \mapsto \mu\otimes\nu \in (\mathcal{P}(S\times T),\mathcal{C}(S\times T))$,
- $\mu \in (\mathcal{P}(S),\mathcal{C}(S)) \mapsto F\star\mu \in (\mathcal{P}(T),\mathcal{C}(T))$,

and the measurability of the composition $(\mu,x) \mapsto (\mu,\delta_x) \mapsto \mu\otimes\delta_x \mapsto F\star(\mu\otimes\delta_x) = \hat{F}(\mu,x)$. □

Lemma 2.6.2 (Conditional law) *Let $(S,\mathcal{S})$ and $(T,\mathcal{T})$ be two measurable spaces.*

1. *If $(S,\mathcal{S})$ is a Borel space, there exists a conditional law of Y knowing X.*
2. *If $Y = \varphi(X,Z)$ where $Z \perp\!\!\!\perp X$ is a random variable valued in a measurable space V and $\varphi : S\times V \to T$ is a measurable function, then $\mathcal{L}(\varphi(x,Z))\,|_{x=X}$ is a conditional law of Y knowing X. In the case $S = S_1\times S_2$, $X = (X_1,X_2)$, and $Y = \varphi(X_1,Z)$, then $\mathbb{P}_Y^X = \mathcal{L}(\varphi(x_1,Z))\,|_{x_1=X_1}$, and thus $\mathbb{P}_Y^X$ is $\sigma(X_1)$-measurable in $(\mathcal{P}(T),\mathcal{C}(T))$.*

3. *For any probability kernel ν from S to $\mathcal{P}(T)$, there exists a measurable function $\phi : S \times [0,1] \to T$ s.t. $\nu(s) = \mathcal{L}(\phi(s,U))$, for all $s \in S$, where U is a uniform random variable.*

Proof. The first assertion is stated in Theorem 6.3 in [44], and the second one follows from Fubini's theorem. The third assertion is a consequence of the two others. □

Proposition 2.6.1 *Given an open-loop control $\alpha \in \mathcal{A}$, and an initial condition $\xi \in L^0(\mathcal{X};\mathcal{G})$, the solution $X^{\xi,\alpha}$ to the conditional McKean-Vlasov equation is such that: for all $t \in \mathbb{N}$, $X_t^{\xi,\alpha}$ is $\sigma(\xi,\Gamma,(\varepsilon)_{s\le t},(\varepsilon_s^0)_{s\le t})$-measurable, and $\mathbb{P}^0_{(X_t^{\xi,\alpha},\alpha_t)}$ is $\mathcal{F}_t^0$-measurable.*

Proof. We prove the result by induction on t. It is clear for $t = 0$. Assuming that it holds true for some $t \in \mathbb{N}$, we write

$$X_{t+1}^{\xi,\alpha} = F(X_t^{\xi,\alpha}, \alpha_t, \mathbb{P}^0_{(X_t^{\xi,\alpha},\alpha_t)}, \varepsilon_{t+1}, \varepsilon_{t+1}^0), \quad t \in \mathbb{N}.$$

By induction hypothesis, there is a measurable function $f_{t+1} : \mathcal{X}\times G\times E^{t+1}\times(E^0)^{t+1} \to \mathcal{X}$ s.t. $X_{t+1}^{\xi,\alpha} = f_{t+1}(\xi,\Gamma,(\varepsilon_s)_{s\le t+1},(\varepsilon_s^0)_{s\le t+1})$, and thus $X_{t+1}^{\xi,\alpha}$ is $\sigma(\xi,\Gamma,(\varepsilon)_{s\le t+1},(\varepsilon_s^0)_{s\le t+1})$-measurable and $\mathbb{P}^0_{(X_{t+1}^{\xi,\alpha},\alpha_{t+1})}$ is $\sigma(\varepsilon_s^0, s \le t+1)$-measurable by Lemma 2.6.2. □

2.6.2 Proof of coupling results

Lemma 2.6.1 *Let U, V be two independent uniform variables, and F a distribution function on $\mathbb{R}$. We have*

$$\Big(F^{-1}(U), F(F^{-1}(U)) - U\Big) \overset{d}{=} (F^{-1}(U), V\Delta F(F^{-1}(U)),$$

where we denote $\Delta F := F - F_-$.

Proof. Notice that $F(F^{-1}(U)) - U$ is the position (from top to bottom) of U in the set $\{u \in [0,1], F^{-1}(u) = F^{-1}(U)\}$ and is thus smaller than $\Delta F(F^{-1}(U))$. Now, given a measurable function $f \in L^0(A \times [0,1];\mathbb{R})$, we have

$$\begin{aligned}&\mathbb{E}\Big[f\big(F^{-1}(U), F(F^{-1}(U)) - U\big)\Big] \qquad (2.6.1)\\ &= \mathbb{E}\big[f(F^{-1}(U),0)\mathbf{1}_{\Delta F(F^{-1}(U))=0}\big] + \mathbb{E}\big[f(F^{-1}(U), F(F^{-1}(U)) - U))\mathbf{1}_{\Delta F(F^{-1}(U))>0}\big].\end{aligned}$$

The second term can be decomposed as

$$\sum_{\Delta F(c)>0} \mathbb{E}\Big[f\big(c, F(c) - U\big)\mathbf{1}_{F^{-1}(U)=c}\Big] \;=\; \sum_{\Delta F(c)>0} \int_0^1 f\left(c, \Delta F(c)u\right)\Delta F(c)du.$$

where the equality comes from a change of variable. Summing over $\Delta F(c) > 0$, we obtain $\mathbb{E}\big[f\big(F^{-1}(U), V\Delta F(F^{-1}(U))\big)\mathbf{1}_{\Delta F(F^{-1}(U))>0}\big]$, and combined with (2.6.1), we get

$$\mathbb{E}\Big[f\big(F^{-1}(U), F(F^{-1}(U)) - U\big)\Big] = \mathbb{E}\big[f\big(F^{-1}(U), V\Delta F(F^{-1}(U))\big)\big],$$

which proves the result. □

Lemma 2.6.2 *Let $\mathcal{X}$ be a compact Polish space, then there exists an embedding $\phi \in L^0(\mathcal{X}, \mathbb{R})$ such that*

1. *ϕ and ϕ^{-1} are uniformly continuous,*
2. *for any probability measure $\mu \in \mathcal{P}(\mathcal{X})$, we have $Im\left(F^{-1}_{\phi\star\mu}\right) \subset Im(\phi)$. In particular, $\phi^{-1} \circ F^{-1}_{\phi\star\mu}$ is well posed.*

Proof. 1. Without loss of generality, we assume that $\mathcal{X}$ is bounded by 1. Fix a countable dense family $(x_n)_{n\in\mathbb{N}}$ in $\mathcal{X}$. We define the map $\phi_1 : x \in \mathcal{X} \mapsto (d(x, x_n))_{n\in\mathbb{N}} \in [0,1]^{\mathbb{N}}$. Let us endow $[0,1]^{\mathbb{N}}$ with the metric $\mathrm{d}((u_n)_{n\in\mathbb{N}}, (v_n)_{n\in\mathbb{N}}) := \sum_{n\geq 0} \frac{1}{2^n}|u_n - v_n|$. ϕ_1 is clearly injective and uniformly continuous (even Lipschitz). The compactness of $\mathcal{X}$ implies that its inverse ϕ_1^{-1} is uniformly continuous as well. Let us now consider $\phi_2 : ([0,1]^{\mathbb{N}}, \mathrm{d}) \mapsto [0,1]$ where $\phi_2((u_n)_{n\in\mathbb{N}})$ essentially groups the decimals of the real numbers u_n, $n \in \mathbb{N}$, in a single real number. More precisely, let $\iota : \mathbb{N} \to \mathbb{N}^2$ be a surjection, then we define the k-th decimal of $\phi_2((u_n)_{n\in\mathbb{N}})$ as the $(\iota(k))_2$-th decimal of $u_{(\iota(k))_1}$ (with the convention that for a number with two possible decimal representations, we choose the one that ends with 000...). ϕ_2 is clearly injective, uniformly continuous, as well as its inverse ϕ_2^{-1}. Thus, $\phi := \phi_2 \circ \phi_1$ defines an embedding of $\mathcal{X}$ into $\mathbb{R}$, such that ϕ and ϕ^{-1} are uniformly continuous.
2. $F^{-1}_{\phi\star\mu}$ being caglad, and $\mathrm{Im}(\phi)$ being closed (by compactness of $\mathcal{X}$), it is enough to prove that $F^{-1}_{\phi\star\mu}(u) \in \mathrm{Im}\phi$ for almost every $u \in [0,1]$ (in the Lebesgue sense). However, given a uniform variable U, we have $F^{-1}_{\phi\star\mu}(U) \sim \phi \star \mu$, and thus

$$\mathbb{P}(F^{-1}_{\phi\star\mu}(U) \in \mathrm{Im}(\phi)) = \mathbb{P}_{Y\sim\mu}(\phi(Y) \in \mathrm{Im}(\phi)) = 1.$$

□

Proof of Lemma 2.4.1

(1) We first consider the case where $\mathcal{X} \subset \mathbb{R}$. Let us call F_μ the distribution function of $\mu \in \mathcal{P}(\mathcal{X})$, and F_μ^{-1} its generalized inverse. Let us define the function $\zeta : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \times \mathcal{X} \times [0,1] \to \mathcal{X}$ by

$$\zeta(\mu, \mu', x, u) \;:=\; F_{\mu'}^{-1}\big(F_\mu(x) - u\Delta F_\mu(x)\big),$$

which is measurable by noting that the measurability in μ, μ' comes from the continuity of

$$\begin{array}{rcl} \mathcal{P}(\mathcal{X}) & \to & L^1_{caglad}([0,1], \mathcal{X}) \\ \mu & \mapsto & F_\mu^{-1}. \end{array}$$

By construction, we then have for any $\xi \sim \mu$, and U, V two independent uniform variables, independent of ξ

$$\begin{array}{rcl} (\xi, \zeta(\mu, \mu', \xi, V)) & = & (\xi, F_{\mu'}^{-1}\big(F_\mu(\xi) - V\Delta F_\mu(\xi)\big)) \\ & \stackrel{d}{=} & (F_\mu^{-1}(U), F_{\mu'}^{-1}\big(F_\mu(F_\mu^{-1}(U)) - V\Delta F_\mu(F_\mu^{-1}(U))\big)) \\ & = & (F_\mu^{-1}(U), F_{\mu'}^{-1}\big(F_\mu(F_\mu^{-1}(U)) - V\Delta F_\mu(F_\mu^{-1}(U))\big)) \\ & \stackrel{d}{=} & \big(F_\mu^{-1}(U), F_{\mu'}^{-1}(U)\big), \end{array}$$

where the last equality holds by Lemma 2.6.1. It is well-known (see e.g. Theorem 3.1.2 in [75]) that $\big(F_\mu^{-1}(U), F_{\mu'}^{-1}(U)\big)$ is an optimal coupling for (μ, μ'), and so $\mathcal{W}(\mu, \mu') = \mathbb{E}\big[d(\xi, \zeta(\mu, \mu', \xi, V))\big]$.

(2) Let us now consider the case of a general compact Polish space $\mathcal{X}$. Denoting by $\zeta_\mathbb{R}$ the "ζ" from the case "$\mathcal{X} \subset \mathbb{R}$", and considering an embedding $\phi \in L^0(\mathcal{X}, \mathbb{R})$ as in Lemma 2.6.2, let us define

$$\zeta(\mu, \mu', x, u) := \phi^{-1}(\zeta_\mathbb{R}(\phi \star \mu, \phi \star \mu', \phi(x), u)),$$

which is well posed by definition of $\zeta_\mathbb{R}$ and Lemma 2.6.2. Now, fix $\xi \sim \mu$, U a uniform variable independent of ξ, and define $\xi' := \zeta(\mu, \mu', \xi, U)$. By definition of ζ, its clear that $\xi' \sim \mu'$, and

$$\mathbb{E}\big[d(\phi(\xi), \phi(\xi'))\big] = \mathcal{W}(\phi \star \mu, \phi \star \mu'). \tag{2.6.2}$$

Fix $\epsilon > 0$. We are looking for $\eta, \delta > 0$ such that

$$\mathcal{W}(\mu, \mu') < \eta \;\Rightarrow\; \mathcal{W}(\phi \star \mu, \phi \star \mu') < \delta \;\Leftrightarrow\; \mathbb{E}\big[d(\phi(\xi), \phi(\xi'))\big] < \delta \;\Rightarrow\; \mathbb{E}[d(\xi, \xi')] < \epsilon.$$

Let us first show that there exists $\delta > 0$ such that $\mathbb{E}[d(\phi(\xi), \phi(\xi'))] < \delta \Rightarrow \mathbb{E}[d(\xi, \xi')] < \epsilon$. Fix $\gamma > 0$ such that $d(x, x') < \gamma \Rightarrow d(\phi^{-1}(x), \phi^{-1}(x')) < \frac{\epsilon}{2}$. Denoting by $\Delta_{\mathcal{X}}$ the diameter of $\mathcal{X}$, we then have

$$\mathbb{E}[d(\xi, \xi')] \leq \mathbb{E}[d(\xi, \xi')\mathbf{1}_{d(\phi(\xi),\phi(\xi'))<\gamma}] + \frac{\Delta_{\mathcal{X}}}{\gamma}\mathbb{E}\big[d(\phi(\xi), \phi(\xi'))\big] \leq \frac{\epsilon}{2} + \frac{\Delta_{\mathcal{X}}}{\gamma}\mathbb{E}\big[d(\phi(\xi), \phi(\xi'))\big],$$

so that we can choose $\delta = \frac{\gamma}{\Delta_{\mathcal{X}}}\frac{\epsilon}{2}$. On the other hand, by uniform continuity of ϕ and by definition of the Wasserstein metric, there exists $\eta > 0$ such that $d(\mu, \mu') < \eta \Rightarrow \mathcal{W}(\phi \star \mu, \phi \star \mu') < \delta$. From (2.6.2), we thus conclude that $d(\mu, \mu') < \eta \Rightarrow \mathbb{E}[d(\xi, \xi')] < \epsilon$. □

Chapter 3

Chaos propagation of N-agent Markov decision processes with common noise and open-loop controls

Abstract. In this chapter, we study a N-agent Markov Decision Process (MDP) with common noise, infinite horizon, and where the optimization is performed open-loop controls. We first obtain the Bellman fixed point equation for this problem and the reduction to feedback controls. Then, by comparing it with the fixed point Bellman equation of the associated mean-field approximation studied in previous chapter (CMKVMDP), we obtain the propagation of chaos of the optimal value functions of the N-agent MDP to the CMKVMDP when $N \to +\infty$, with some convergence rate, denoted by $\mathcal{O}(M_N^\gamma)$. We also provide ways to build $(\varepsilon + \mathcal{O}(M_N^\gamma))$-optimal policies for the N-agent MDP from ε-optimal policies for the CMKVMDP, and vice-versa. We finally provide a concrete application of the propagation of chaos result, by approximately solving an N-agent advertising problem under social influence via the resolution of the associated CMKVMDP.

3.1 Introduction

This chapter is a companion work to the previous chapter. In the previous chapter, we formulated a N-agent Markov Decision Process, simply to naturally introduce the associated McKean-Vlasov Markov Decision Process, via the so-called *mean-field approximation*. The mean-field approximation is a procedure consisting in replacing empirical

distributions by theoretical ones in a dynamic system. This allows to formally obtain a different problem called the mean-field, or McKean-Vlasov, "limit" of the problem. This procedure comes from statistical physics in the study of large dynamic particle systems. Besides allowing to define another problem, the underlying belief is that this McKean-Vlasov problem must be, in some ways, "close" to the original N-agent problem it was formally derived from. This belief initially comes from experience, in the sense that such procedure has been applied many times, for many problems, and that one has experimentally observe the proximity of the N-agent and the McKean-Vlasov problems in *most* cases. Besides the empirical evidences, mathematical justifications of this phenomenon have been rigorously obtained for large classes of problems. It is thus now acknowledged that 1) the mean-field approximation is likely to yield a Mckean-Vlasov problem that is close to the N-agent base model, and 2) that, when it is indeed the case, the mathematical arguments to prove it rely on extensions of the law of large numbers. It is now very standard to first formally derive the McKean-Vlasov MDP, study it in detail, and then only, prove that the N-agent MDPs "converge" to it as $N \to \infty$.

A detailed study of the McKean-Vlasov MDP limit candidate was made in previous chapter. We shall now use our knowledge about this McKean-Vlasov MDP to establish the propagation of chaos of the N-agent MDPs, i.e. their convergence to the McKean-Vlasov MDP from previous chapter as $N \to \infty$.

We point out the work [49], which is the first paper to rigorously connect mean-field control to large systems of controlled processes, see also the recent paper [29], and refer to the books [8], [15] for an overview of the subject.

Main contributions. In this paper, we introduce a general N-individual Markov Decision Process in discrete time framework in the presence of *common noise*, and when optimization is performed over *open-loop controls* on infinite horizon, and link it to the associated McKean-Vlavov Markov Decision Process obtained by means of a mean-field approximation procedure.

Our first contribution is to provide a detailed study of the N-agent MDP, in particular yielding the Bellman fixed point equation and the reduction to stationary feedback policies.

Our second and main contribution is to rigorously connect the N-agent MDPs to the McKean-Vlasov MDP formally obtained via a mean-field approximation procedure. We prove the following results:

- Propagation of chaos of the value functions, i.e. the uniform convergence, as the number of interacting agents N tends to infinity, of the optimal value function of the N-individual MDP towards the optimal value function of the CMKV-MDP.

- We provide simple ways to turn any close-to-optimal stationary randomized feedback policies of the McKean-Vlasov MDP into a close-to-optimal randomized feedback policy for the N-individual MDP, and vice-versa.

Furthermore, by relying on rate of convergence in Wasserstein distance of the empirical measure, we give a rate of convergence for the limiting CMKV-MDP under suitable Lipschitz assumptions on the state transition and reward functions, which is new to the best of our knowledge.

Finally, we illustrate the usefulness of such chaos propagation result with an example that can be explicitly solved in the mean-field approximation.

Outline of the paper. The rest of the paper is organized as follows. Section 3.2 carefully formulates both the N-individual model and the CMKV-MDP. In Section 3.3, we establish the Bellman fixed point equation for the N-individual MDP, as well as the reduction to stationary feedback controls. Then, in Section 3.4, we show the several connections between the N-individual and the McKean-Vlasov MDP, with rates of convergence, when N goes to infinity. We then in Section 3.5 describe and study a concrete example. We finally conclude in Section 3.6 by discussing the concrete implications of these results for the N-individual MDP.

Notation. When $(\mathcal{Y}, d)$ is a compact metric space, the set $\mathcal{P}(\mathcal{Y})$ of probability measures on $\mathcal{Y}$ is equipped with the Wasserstein distance

$$\mathcal{W}(\mu,\mu') = \inf\Big\{ \int_{\mathcal{Y}^2} d(y,y')\boldsymbol{\mu}(\mathrm{d}y,\mathrm{d}y') : \boldsymbol{\mu} \in \boldsymbol{\Pi}(\mu,\mu')\Big\},$$

where $\boldsymbol{\Pi}(\mu,\mu')$ is the set of probability measures on $\mathcal{Y}\times\mathcal{Y}$ with marginals μ and μ', i.e., $\mathrm{pr}_1 \star \boldsymbol{\mu} = \mu$, and $\mathrm{pr}_2 \star \boldsymbol{\mu} = \mu'$.

3.2 The N-agent Markov Decision Process

We formulate the mean-field Markov Decision Process (MDP) in a large population model with indistinguishable agents $i \in \mathbb{N}^* = \mathbb{N} \setminus \{0\}$.

Let $\mathcal{X}$ (the state space) and A (the action space) be two compact Polish spaces equipped respectively with their metric d and d_A. We denote by $\mathcal{P}(\mathcal{X})$ (resp. $\mathcal{P}(A)$) the space of probability measures on $\mathcal{X}$ (resp. A) equipped respectively with their Wasserstein distance $\mathcal{W}$ and $\mathcal{W}_A$. We also consider the product space $\mathcal{X}\times A$, equipped with the metric $\boldsymbol{d}((x,a),(x',a')) = d(x,x') + d_A(a,a')$, $x,x' \in \mathcal{X}$, $a,a' \in A$, and the associated space of probability measure $\mathcal{P}(\mathcal{X}\times A)$, equipped with its Wasserstein distance $\boldsymbol{W}$. Let G, E, and E^0 be three measurable spaces, representing respectively the initial information, idiosyncratic noise, and common noise spaces.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which are defined the following family of mutually i.i.d. random variables

- $(\Gamma^i, \xi^i)_{i\in\mathbb{N}^\star}$ (initial informations and initial states) valued in $G \times \mathcal{X}$
- $(\varepsilon_t^i)_{i\in\mathbb{N}^\star, t\in\mathbb{N}}$ (idiosyncratic noises) valued in E with probability distribution λ_ε
- $\varepsilon^0 := (\varepsilon_t^0)_{t\in\mathbb{N}}$ (common noise) valued in E^0.

We assume that $\mathcal{F}$ contains an atomless random variable, i.e., $\mathcal{F}$ is rich enough, so that any probability measure ν on $\mathcal{X}$ (resp. A or $\mathcal{X}\times A$) can be represented by the law of some random variable Y on $\mathcal{X}$ (resp. A or $\mathcal{X}\times A$), and we write $Y \sim \nu$, i.e., $\mathcal{L}(Y) = \nu$.

Given $N \in \mathbb{N}^*$, we denote by $\mathcal{A}_N$ the set of open-loop controls for the N-individual MDP, that is, the set of A^N-valued random sequences α, adapted to the filtration $(\mathcal{F}_{N,t})_{t\in\mathbb{N}}$ defined by $\mathcal{F}_{N,t} = \sigma(\Gamma^i, \xi^i, (\varepsilon_s^i)_{s\le t}, i \le N, (\varepsilon_s^0)_{s\le t})$.

Given $\alpha \in \mathcal{A}_N$, the state process of agent $i = 1, \ldots, N$ in an N-agent MDP is given by the dynamical system

$$\left\{ \begin{array}{rcl} X_0^{i,N,\alpha} & = & \xi^i \\ X_{t+1}^{i,N,\pi} & = & F(X_t^{i,N,\alpha}, \alpha_t^i, \frac{1}{N}\sum_{j=1}^N \delta_{(X_t^{j,N,\alpha}, \alpha_t^{j,\pi})}, \varepsilon_{t+1}^i, \varepsilon_{t+1}^0), \quad t \in \mathbb{N}, \end{array} \right.$$

where F is a measurable function from $\mathcal{X}\times A \times \mathcal{P}(\mathcal{X}\times A) \times E \times E^0$ into $\mathcal{X}$, called state transition function. The i-th individual contribution to the influencer's gain over an infinite horizon is defined by

$$J_i^{N,\alpha} := \sum_{t=0}^{\infty} \beta^t f\Big(X_t^{i,N,\alpha}, \alpha_t^i, \frac{1}{N}\sum_{j=1}^N \delta_{(X_t^{j,N,\alpha}, \alpha_t^j)}\Big), \quad i = 1, \ldots, N,$$

where the reward f is a measurable real-valued function on $\mathcal{X}\times A \times \mathcal{P}(\mathcal{X}\times A)$, assumed to be bounded (recall that $\mathcal{X}$ and A are compact spaces), and β is a positive discount factor in $[0, 1)$. The influencer's renormalized and expected gains are

$$J^{N,\alpha} \;:=\; \frac{1}{N}\sum_{i=1}^N J_i^{N,\alpha}, \qquad V^{N,\alpha} \;:=\; \mathbb{E}\big[J^{N,\alpha}\big],$$

and the optimal value of the influencer is $V^N := \sup_{\alpha\in\mathcal{A}_N} V^{N,\alpha}$. Observe that the agents are indistinguishable in the sense that the initial pair of information/state $(\Gamma^i, \xi^i)_i$, and idiosyncratic noises are i.i.d., and the state transition function F, reward function f, and discount factor β do not depend on i.

Let us now consider the asymptotic problem when the number of agents N goes to infinity. In view of the propagation of chaos argument, we expect the N-individual MDP to converge in some sense to the following McKean-Vlasov MDP.

Let us rename Γ, ξ and $(\varepsilon_t)_{t\in\mathbb{N}}$ the random variables Γ^1, ξ^1, and $(\varepsilon^1_t)_{t\in\mathbb{N}}$. We also introduce $\mathcal{A}$, the set of open-loop controls for the McKean-Vlasov MDP, that is, the set of A-valued random sequences α adapted to the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ such that $\mathcal{F}_t := \sigma(\Gamma, \xi, (\varepsilon_s)_{s\le t}, (\varepsilon^0_s)_{s\le t})$. Given $\alpha \in \mathcal{A}$, we define the conditional McKean-Vlasov dynamic

$$\left\{ \begin{array}{rcl} X^\alpha_0 &=& \xi \\ X^\alpha_{t+1} &=& F(X^\alpha_t, \alpha_t, \mathbb{P}^0_{(X^\alpha_t,\alpha_t)}, \varepsilon_{t+1}, \varepsilon^0_{t+1}), \quad t \in \mathbb{N}. \end{array} \right. \tag{3.2.1}$$

Here, we denote by $\mathbb{P}^0$ and $\mathbb{E}^0$ the conditional probability and expectation knowing the common noise ε^0, and then, given a random variable Y valued in $\mathcal{Y}$, we denote by $\mathbb{P}^0_Y$ or $\mathcal{L}^0(Y)$ its conditional law knowing ε^0, which is a random variable valued in $\mathcal{P}(\mathcal{Y})$ (see Lemma 2.6.2). The influencer's expected gain in the McKean-Vlasov model is

$$V^\alpha \;:=\; \mathbb{E}\big[\sum_{t=0}^{\infty} \beta^t f\big(X^\alpha_t, \alpha_t, \mathbb{P}^0_{(X^\alpha_t,\alpha_t)}\big)\big], \qquad V := \sup_{\alpha\in\Pi_{OL}} V^\alpha. \tag{3.2.2}$$

In the sequel, we make the following regularity assumptions on F and f:

($\mathbf{HF_{lip}}$) There exists $K_F > 0$, such that for all $a \in A$, $e^0 \in E^0$, $x, x' \in \mathcal{X}$, $\nu, \nu' \in \mathcal{P}(\mathcal{X}\times A)$,

$$\mathbb{E}\big[d\big(F(x,a,\nu,\varepsilon^1_1,e^0), F(x',a,\nu',\varepsilon^1_1,e^0)\big)\big] \le\; K_F\big(d(x,x') + \boldsymbol{W}(\nu,\nu')\big)\big).$$

($\mathbf{Hf_{lip}}$) There exists $K_f > 0$, such that for all $a \in A$, $x, x' \in \mathcal{X}$, $\nu, \nu' \in \mathcal{P}(\mathcal{X} \times A)$,

$$d(f(x,a,\nu), f(x',a,\nu')) \le\; K_f\big(d(x,x') + \boldsymbol{W}(\nu,\nu')\big)\big).$$

Remark 3.2.1 We stress the importance of making the regularity assumptions for F in *expectation* only. For the same argument as in Remark **??**, when $\mathcal{X}$ is finite, F cannot be, strictly speaking, Lipschitz. However, F can be Lipschitz *in expectation*, e.g. once integrated w.r.t. the idiosyncratic noise, which is a very natural assumption. □

3.3 Bellman fixed point equation for the N-agent MDP

We derive and study the Bellman equation corresponding to N-agent MDP, seen as a Markov Decision Process with state space $\mathcal{X}^N$, action space $\mathcal{A}^N$, state transition function

$$\boldsymbol{F}(x, \mathbf{a}, (e^i)_{i\le N}, e^0) = \Big(F(x_i, \mathbf{a}_i, \frac{1}{N}\sum_{n=1}^N \delta_{x_n,a_n}, e^i, e^0)\Big)_{i\le N}$$

and reward function

$$\mathbf{f}(x, \boldsymbol{\nu}_0) = \frac{1}{N}\sum_{i=1}^N f\Big(x_i, \mathbf{a}_i, \frac{1}{N}\sum_{n=1}^N \delta_{x_n,a_n}\Big)$$

By defining this MDP on the canonical space $(E^N, E^0)^{\mathbb{N}}$, we identify $((\varepsilon^i)_{i\le N}, \varepsilon^0)$ with the canonical identity function in $(E^N, E^0)^{\mathbb{N}}$, and ε^i_t (resp. ε^0_t) with the projection $((e^i)_{i\le N}, e^0) \mapsto e^i_t$ for all $((e^i)_{i\le N}, e^0) \in (E^N, E^0)^{\mathbb{N}}$. We also denote by $\theta : (E^N, E^0)^{\mathbb{N}} \to (E^N, E^0)^{\mathbb{N}}$ the shifting operator, defined by $\theta((e^i_t)_{i\le N}, e^0_t)_{t\in\mathbb{N}} = ((e^i_{t+1})_{i\le N}, e^0_{t+1})_{t\in\mathbb{N}}$. Via this identification, an open-loop control $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$ is a sequence $(\boldsymbol{\nu}_t)_t$ where $\boldsymbol{\nu}_t$ is a measurable function from $(E^N, E^0)^t$ into $\boldsymbol{A}$, with the convention that $\boldsymbol{\nu}_0$ is simply a constant in $\boldsymbol{A}$. Given $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$, and $((e^i)_{i\le N}, e^0) \in (E^N, E^0)^{\mathbb{N}}$, we define $\vec{\boldsymbol{\nu}}^{(e^i_1)_{i\le N}, e^0_1} := (\vec{\boldsymbol{\nu}}_t^{(e^i_1)_{i\le N}, e^0_1})_t \in \boldsymbol{\mathcal{A}}$, where $\vec{\boldsymbol{\nu}}_t^{(e^i_1)_{i\le N}, e^0_1}(.) := \boldsymbol{\nu}_{t+1}((e^i_1)_{i\le N}, e^0_1, .)$, $t \in \mathbb{N}$. Given $x \in \mathcal{X}^N$, and $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$, we denote by $(x^{x,\boldsymbol{\nu}}_t)_t$ the solution to (2.4.4) on the canonical space, which satisfies the flow property

$$\big(x^{x,\boldsymbol{\nu}}_{t+1}, \boldsymbol{\nu}_{t+1}\big) \equiv \big(x_t^{x^{x,\boldsymbol{\nu}}_1, \vec{\boldsymbol{\nu}}^{(\varepsilon^i_1)_{i\le N}, \varepsilon^0_1}(\theta((\varepsilon^i)_{i\le N}, \varepsilon^0))}, \vec{\boldsymbol{\nu}}^{(\varepsilon^i_1)_{i\le N}, \varepsilon^0_1}(\theta((\varepsilon^i)_{i\le N}, \varepsilon^0))\big), \quad t \in \mathbb{N}.$$

where $\equiv$ denotes the equality between functions on the canonical space. Given that $((\varepsilon^i_1)_{i\le N}, \varepsilon^0_1) \perp\!\!\!\perp \theta((\varepsilon^i)_{i\le N}, \varepsilon^0)) \overset{d}{=} (\varepsilon^i)_{i\le N}, \varepsilon^0)$, we obtain that the expected gain of this MDP in (2.4.5) satisfies the relation

$$V^{\boldsymbol{\nu}}(x) = \mathbf{f}(x, \boldsymbol{\nu}_0) + \beta\mathbb{E}\Big[V^{\vec{\boldsymbol{\nu}}^{(\varepsilon^i_1)_{i\le N}, \varepsilon^0_1}}(x^{x,\boldsymbol{\nu}}_1)\Big]. \tag{3.3.1}$$

Let us denote by $L^\infty(\mathcal{X}^N)$ the set of bounded real-valued functions on $\mathcal{X}^N$, and by $L^\infty_m(\mathcal{X}^N)$ the subset of measurable functions in $L^\infty(\mathcal{X}^N)$. We then introduce the Bellman "operator" $\mathcal{T} : L^\infty_m(\mathcal{X}^N) \to L^\infty(\mathcal{X}^N)$ defined for any $W \in L^\infty_m(\mathcal{X}^N)$ by:

$$[\mathcal{T}W](x) := \sup_{\mathbf{a}\in\boldsymbol{A}} \Big\{\mathbf{f}(x, \mathbf{a}) + \beta\mathbb{E}\big[W\big(\boldsymbol{F}(x, \mathbf{a}, (\varepsilon^i_1)_{i\le N}, \varepsilon^0_1)\big)\big]\Big\}, \quad x \in \mathcal{X}^N. \tag{3.3.2}$$

Notice that the sup can a priori lead to a non measurable function $\mathcal{T}W$.

We state the basic properties of the Bellman operator $\mathcal{T}$.

Proposition 3.3.1 *Assume that* $(\mathbf{H_{lip}})$ *holds true. (i) The operator* $\mathcal{T}$ *is monotone increasing: for* $W_1, W_2 \in L^\infty_m(\mathcal{X}^N)$*, if* $W_1 \le W_2$*, then* $\mathcal{T}W_1 \le \mathcal{T}W_2$*. (ii) Furthermore, it is contracting on* $L^\infty_m(\mathcal{X}^N)$ *with Lipschitz factor* β*, and admits a unique fixed point in* $L^\infty_m(\mathcal{X}^N)$*, denoted by* $V^\star$*, hence solution to:*

$$V^\star = \mathcal{T}V^\star.$$

(iii) $V^\star$ *is* γ*-Hölder, with* $\gamma = \min\Big(1, \frac{|\ln\beta|}{\ln(2K_F)}\Big)$*, i.e. there exists some positive constant* $K_\star$ *(depending only on* K_F*,* K_f*,* β*, and explicit in the proof), such that*

$$\big|V^\star(x) - V^\star(x')\big| \le K_\star d_N(x, x')^\gamma, \quad \forall x, x' \in \mathcal{X}^N.$$

Proof. (i) The monotonicity of $\mathcal{T}$ is shown by standard arguments.
(ii) The β-contraction property of $\mathcal{T}$ is also obtained by standard arguments. Let us now prove by induction that the iterative sequence $V_{n+1} = \mathcal{T}V_n$, with $V_0 \equiv 0$ is well defined and such that

$$|V_n(x) - V_n(x')| \leq 2K_f \sum_{t=0}^{\infty} \beta^t \min((2K_F)^t d_N(x, x'), \Delta_{\mathcal{X}}) \tag{3.3.3}$$

for all $n \in \mathbb{N}$. The property is obviously satisfied for $n = 0$. Assume that the property holds true for a fixed $n \in \mathbb{N}$, and let us prove it for $n + 1$. First of all, the inequality (3.3.3) implies that V_n is continuous, and thus $V_n \in L_m^\infty(\mathcal{X}^N$. Therefore, $V_{n+1} = \mathcal{T}V_n$ is well defined. Fix $x, x' \in \mathcal{X}^N$. Fix an A-valued random variable α_0. Let us start with two preliminary estimations: under $(\mathbf{H_{lip}})$, we clearly have

$$\mathbb{E}\left[|\mathbf{f}(x, \alpha_0) - \mathbf{f}(x', \alpha_0)|\right] \leq 2K_f d_N(x, x'). \tag{3.3.4}$$

Similarly, for $e^0 \in E^0$, we have

$$\mathbb{E}[d(\boldsymbol{F}(x, \alpha_0, (\varepsilon_1^i)_{i\leq N}, e^0), \boldsymbol{F}(x', \alpha_0, (\varepsilon_1^i)_{i\leq N}, e^0))] \leq 2K_F d_N(x, x'). \tag{3.3.5}$$

Now, we prove the hereditary property. The definition of $\mathcal{T}$ and V_{n+1} combined with (3.3.4) and the induction hypothesis, imply that

$$|V_{n+1}(x) - V_{n+1}(x')| \quad \leq \quad 2K_f d_N(x, x') + \beta\mathbb{E}[2K_f \sum \beta^t \min((2K_F^t d_N(x_1, x_1'), \Delta_{\mathcal{X}})]$$

where $x_1 = \boldsymbol{F}(x, \alpha_0, (\varepsilon_1^i)_{i\leq N}, e^0)$ and $x_1' = \boldsymbol{F}(x', \alpha_0, (\varepsilon_1^i)_{i\leq N}, e^0)$. By Jensen's inequality and (3.3.5), we have

$$\begin{aligned}
& |V_{n+1}(x) - V_{n+1}(x')| \\
\leq \quad & 2K_f \min(d_N(x, x'), \Delta_{\mathcal{X}}) + \beta 2K_f \sum \beta^t \min((2K_F^t \mathbb{E} d_N(x_1, x_1'), \Delta_{\mathcal{X}}) \\
\leq \quad & 2K_f \min(d_N(x, x'), \Delta_{\mathcal{X}}) + \beta 2K_f \sum \beta^t \min((2K_F^t 2K_F d_N(x, x'), \Delta_{\mathcal{X}}) \\
\leq \quad & 2K_f \sum \beta^t \min((2K_F^t d_N(x, x'), \Delta_{\mathcal{X}}).
\end{aligned}$$

This concludes the induction and proves that V_n is well defined and satisfies the inequality (3.3.3) for all $n \in \mathbb{N}$. As $\mathcal{T}$ is β-contracting, a standard argument from the proof of the Banach fixed point theorem shows that $(V_n)_n$ is a Cauchy sequence in the complete metric space $L_m^\infty(\mathcal{X}^N)$, and therefore admits a limit $V^\star \in L_m^\infty(\mathcal{X}^N)$. Notice that

$$V^\star(x) = \lim_n V_{n+1}(x) = \lim_n \mathcal{T}V_n(x) = \mathcal{T}V^\star$$

by continuity of the contracting operator $\mathcal{T}$.

(iii) By sending n to infinity in (3.3.3), we obtain

$$|V^\star(x) - V^\star(x')| \leq 2K_f \sum_{t=0}^{\infty} \beta^t \min((2K_F)^t d_N(x,x'), \Delta_{\mathcal{X}}) =: S(d_N(x,x')).$$

where $S(m) = 2K_f \sum_{t=0}^{\infty} \beta^t \min((2K_F)^t m, \Delta_{\mathcal{X}})$. If $2\beta K_F < 1$, we clearly have

$$S(m) \leq m \sum_{t=0}^{\infty} (\beta 2K_F)^t = \frac{m}{1 - \beta 2K_F},$$

and so V is 1-Hölder. Let us now study the case $2\beta K_F > 1$. In this case, in particular, $2K_F > 1$, thus $t \mapsto s_t(m)$ is nondecreasing, and so

$$\begin{aligned} S(m) \leq & \sum_{t=0}^{\infty} \int_t^{t+1} \beta^t \min\big[s_t(m); \Delta_{\mathcal{X}}\big] ds \\ \leq & \frac{1}{\beta} \sum_{t=0}^{\infty} \int_t^{t+1} \beta^s \min\big[m(2K_F)^s; \Delta_{\mathcal{X}}\big] ds \\ \leq & \frac{1}{\beta} \int_0^{\infty} e^{-|\ln\beta| s} \min\big[m e^{\ln(2K_F)s}; \Delta_{\mathcal{X}}\big] ds. \end{aligned}$$

Let $t_\star$ be such that $m e^{\ln(2K_F)t_\star} = \Delta_{\mathcal{X}}$, i.e. $t_\star = \frac{\ln(\Delta_{\mathcal{X}}/m)}{\ln(2K_F)}$. Then,

$$\begin{aligned} \int_0^{\infty} e^{-|\ln\beta| s} \min\big[m e^{\ln(2K_F)s}; \Delta_{\mathcal{X}}\big] ds & \leq m \int_0^{t^\star} e^{\ln(2K_F\beta)s} ds + \Delta_{\mathcal{X}} \int_{t^\star}^{\infty} e^{\ln(\beta)s} ds \\ & \leq \frac{m}{\ln(2K_F\beta)} \Big[e^{\ln(2K_F\beta)t^\star} - 1\Big] - \frac{\Delta_{\mathcal{X}}}{\ln\beta} e^{\ln(\beta)t^\star}. \end{aligned}$$

After substituting $t_\star$ by its explicit value, we then obtain

$$\begin{aligned} & \int_0^{\infty} e^{-|\ln\beta| s} \min\big[m e^{\ln(2K_F)s}; \Delta_{\mathcal{X}}\big] ds \\ \leq & \frac{m}{\ln(2K_F\beta)} \Big[\Big(\frac{\Delta_{\mathcal{X}}}{m}\Big)^{\frac{\ln(2K_F\beta)}{\ln(2K_F)}} - 1\Big] - \frac{\Delta_{\mathcal{X}}}{\ln\beta} \left(\frac{\Delta_{\mathcal{X}}}{m}\right)^{\frac{\ln(\beta)}{\ln(2K_F)}} \\ \leq & \Delta_{\mathcal{X}} \Big(\frac{1}{\ln(2K_F\beta)} - \frac{1}{\ln\beta}\Big)\Big(\frac{\Delta_{\mathcal{X}}}{m}\Big)^{\frac{\ln(\beta)}{\ln(2K_F)}} - \frac{m}{\ln(2K_F\beta)} \\ \leq & \mathcal{O}\Big(m^{\min\left[1, \frac{|\ln\beta|}{\ln(2K_F)}\right]}\Big). \end{aligned}$$

This implies that V is γ-Hölder and concludes the proof. □

Remark 3.3.1 In the proof of Proposition 2.4.2, one could also have proved that the set $\mathcal{S}$ of functions $W : \mathcal{X}^N \to \mathbb{R}$ such that

$$|W(x) - W(x')| \leq 2K_f \sum_{t=0}^{\infty} \beta^t \min((2K_F)^t d_N(x,x'), \Delta_{\mathcal{X}})$$

for all $x, x' \in \mathcal{X}^N$ is a complete metric space, as it is a closed set of the complete metric space $L^\infty_m(\mathcal{X}^N)$, and is stabilized by the contracting operator $\mathcal{T}$ (which is essentially proved by replacing V_n by W in the proof). One could then have invoked the Banach fixed point theorem on this set $\mathcal{S}$, implying the existence and uniqueness of the fixed point $V^\star$. Notice that this argument would not work if we considered, instead of $\mathcal{S}$, the set of γ-Hölder continuous functions. Indeed, while it is true that such set is stabilized by $\mathcal{T}$ (it essentially follows from (3.3.4) and (3.3.5)), the set of γ-Hölder continuous functions is not closed in $L^\infty_m(\mathcal{X}^N)$ (and thus not a complete metric space): there might indeed exist a converging sequence of γ-Hölder continuous functions with multiplicative factors (in the Hölder property) tending toward infinity, such that the limit function is not γ-Hölder anymore. □

As a consequence of Proposition 3.3.1, we can easily show the following relation between the value function V of the general lifted MDP, and the fixed point $V^\star$ of the Bellman operator.

Lemma 3.3.1 *For all $x \in \mathcal{X}^N$, we have $V(x) \leq V^\star(x)$.*

Proof. From (3.3.1), we have

$$\begin{aligned}
&\inf_{x\in\mathcal{X}^N} \left\{V^\star(x) - V^{\boldsymbol{\nu}}(x)\right\} \\
\geq\ &\inf_{x\in\mathcal{X}^N} \left\{\mathcal{T}V^\star(x) - \left(\mathbf{f}(x,\boldsymbol{\nu}_0) + \beta\mathbb{E}\Big[V^\star(x_1^{x,\boldsymbol{\nu}})\Big]\right) + \beta\mathbb{E}\Big[V^\star(x_1^{x,\boldsymbol{\nu}}) - V^{\vec{\boldsymbol{\nu}}^{\varepsilon_1^0}}(x_1^{x,\boldsymbol{\nu}})\Big]\right\} \\
\geq\ &\beta\mathbb{E}\Big[V^\star(x_1^{x,\boldsymbol{\nu}}) - V^{\vec{\boldsymbol{\nu}}^{\varepsilon_1^0}}(x_1^{x,\boldsymbol{\nu}})\Big] \geq \beta \inf_{x\in\mathcal{X}^N} \left\{V^\star(x) - V^{\boldsymbol{\nu}}(x)\right\}.
\end{aligned}$$

This shows that $\inf_{x\in\mathcal{X}^N}(V^\star(x) - V^{\boldsymbol{\nu}}(x)) \geq 0$, hence

$$V^{\boldsymbol{\nu}}(x) \leq V^\star(x) \quad \forall x \in \mathcal{X}^N.$$

Taking the sup over $\boldsymbol{\nu} \in \boldsymbol{\mathcal{A}}$, we obtain the required result. □

We aim to prove rigorously the equality $V = V^\star$, i.e., the value function V of the general lifted MDP satisfies the Bellman fixed point equation: $V = \mathcal{T}V$, and also to show the existence of ϵ-optimal stationary feedback control for V.

A stationary feedback policy is a measurable function $\boldsymbol{\pi} \in L^0(\mathcal{X}^N; \boldsymbol{A})$ (the set of measurable functions from $\mathcal{X}^N$ into $\boldsymbol{A}$). The associated stationary feedback control is the unique control $\boldsymbol{\nu}^\pi$ satisfying the constraint $\boldsymbol{\nu}_t = \pi(X_t^{x,\boldsymbol{\nu}})$ for all $t \in \mathbb{N}$. The flow property applied to this control clearly implies that $V^{\boldsymbol{\nu}^\pi}$ is a fixed point of the operator $\mathcal{T}^{\boldsymbol{\pi}}$ on $L^\infty(\mathcal{X}^N)$, defined for $W \in L^\infty(\mathcal{X}^N)$ by

$$[\mathcal{T}^{\boldsymbol{\pi}}W](x) = \mathbf{f}(x, \boldsymbol{\pi}(x)) + \beta\mathbb{E}\big[W\big(\boldsymbol{F}(x, \boldsymbol{\pi}(x), (\varepsilon_1^i)_{i\leq N}, \varepsilon_1^0)\big)\big], \quad x \in \mathcal{X}^N.$$

By misuse of notation, we shall identify $V^{\boldsymbol{\pi}}$ and $V^{\boldsymbol{\nu}^{\boldsymbol{\pi}}}$. We have the basic properties on the operator $\mathcal{T}^{\boldsymbol{\pi}}$.

Lemma 3.3.2 *Fix $\boldsymbol{\pi} \in L^0(\mathcal{X}^N; \boldsymbol{A})$.*

(i) The operator $\mathcal{T}^{\boldsymbol{\pi}}$ is contracting on $L^\infty(\mathcal{X}^N)$ with Lipschitz factor β, and $V^{\boldsymbol{\pi}}$ is its unique fixed point.

(ii) Furthermore, it is monotone increasing: for $W_1, W_2 \in L^\infty(\mathcal{X}^N)$, if $W_1 \leq W_2$, then $\mathcal{T}^{\boldsymbol{\pi}} W_1 \leq \mathcal{T}^{\boldsymbol{\pi}} W_2$.

We show a verification type result for the N-individual MDP, by means of the Bellman operator.

Proposition 3.3.2 (Verification result)
Fix $\epsilon \geq 0$, and suppose that there exists an ϵ-optimal feedback policy $\boldsymbol{\pi}_\epsilon \in L^0(\mathcal{X}^N; \boldsymbol{A})$ for $V^\star$ in the sense that

$$V^\star \leq \mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star + \epsilon.$$

Then, $\boldsymbol{\nu}^{\pi_\epsilon} \in \boldsymbol{\mathcal{A}}$ is $\frac{\epsilon}{1-\beta}$-optimal for V, i.e., $V^{\boldsymbol{\pi}_\epsilon} \geq V - \frac{\epsilon}{1-\beta}$, and we have $V \geq V^\star - \frac{\epsilon}{1-\beta}$.

Proof. Since $V^{\boldsymbol{\pi}_\epsilon} = \mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^{\boldsymbol{\pi}_\epsilon}$, and recalling from Lemma 3.3.1 that $V^\star \geq V \geq V^{\boldsymbol{\pi}_\epsilon}$, we have for all $x \in \mathcal{X}^N$,

$$\left|(V^\star - V^{\boldsymbol{\pi}_\epsilon})(x)\right| \leq \left|\mathcal{T}^{\boldsymbol{\pi}_\epsilon}(V^\star - V^{\boldsymbol{\pi}_\epsilon})(x) + \epsilon\right| \leq \beta \|V^\star - V^{\boldsymbol{\pi}_\epsilon}\| + \epsilon,$$

where we used the β-contraction property of $\mathcal{T}^{\boldsymbol{\pi}_\epsilon}$ in Lemma 3.3.2. We deduce that $\|V^\star - V^{\boldsymbol{\pi}_\epsilon}\| \leq \frac{\epsilon}{1-\beta}$, and then, $V \geq \tilde{V}^{\boldsymbol{\pi}_\epsilon} \geq V^\star - \frac{\epsilon}{1-\beta}$, which combined with $V^\star \geq V$, concludes the proof. □

We can conclude this paragraph by showing the existence of an ϵ-optimal lifted randomized feedback policy for the general lifted MDP on $\mathcal{X}^N$, and obtain as a by-product the corresponding Bellman fixed point equation for its value function and for the optimal value of the CMKV-MDP under randomization hypothesis.

Theorem 3.3.1 *Assume that $(\mathbf{H_{lip}})$ holds true. Then, for all $\epsilon > 0$, there exists feedback policy $\boldsymbol{\pi}$ that is ϵ-optimal for $V^\star$. Consequently, the feedback stationary control $\boldsymbol{\nu}^\pi \in \mathcal{A}$ is $\frac{\epsilon}{1-\beta}$-optimal for $V(x)$, and we have $V(x) = V^\star(x)$, which thus satisfies the Bellman fixed point equation.*

Proof. Fix $\epsilon > 0$, and given $\eta > 0$, consider a quantizing grid $\mathcal{M}_\eta = \{x^1, \dots, x^{N_\eta}\} \subset \mathcal{X}^N$, and an associated partition C^i_η, $i = 1, \dots, N_\eta$, of $\mathcal{X}^N$, satisfying

$$C^i_\eta \subset \; B_\eta(x^i) := \Big\{ x \in \mathcal{X}^N : d_N(x, x^i) \le \eta \Big\}, \quad i = 1, \dots, N_\eta.$$

For any x^i, $i = 1, \dots, N_\eta$, and by (2.4.17), there exists $\pi^i_\varepsilon \in L^0(\mathcal{X}^N; A^N)$ such that

$$V^\star(x^i) \le \mathcal{T}^{\pi^i_\varepsilon} V^\star(x^i) + \frac{\epsilon}{3}. \tag{3.3.6}$$

From the partition C^i_η, $i = 1, \dots, N_\eta$ of $\mathcal{X}^N$, associated to $\mathcal{M}_\eta$, we construct the function $\pi : \mathcal{X}^N \to A^N$ as follows: we define, for all $x \in \mathcal{X}^N$,

$$\pi_\varepsilon(x) \;\; = \;\; \pi^i, \quad \text{when } x \in C^i_\eta, \; i = 1, \dots, N_\eta,$$

Such function π_ε is clearly measurable. Let us now check that such $\boldsymbol{\pi}_\epsilon$ yields an ϵ-optimal feedback policy for η small enough. For $x \in \mathcal{X}^N$, we define $x_\eta = x^i$, when $x \in C^i_\eta$, $i = 1, \dots, N_\eta$. Observe that $d_N(x, x_\eta) \le \eta$. We then write for any $x \in \mathcal{X}^N$,

$$\begin{aligned}
[\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](x) - V^\star(x) = \; & \Big([\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](x) - [\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](x_\eta) \Big) + \Big([\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](x_\eta) - V^\star(x_\eta) \Big) \\
& + \big(V^\star(x_\eta) - V^\star(x) \big) \\
\ge \; & \Big([\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](x) - [\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](x_\eta) \Big) - \frac{\epsilon}{3} - \frac{\epsilon}{3},
\end{aligned} \tag{3.3.7}$$

where we used (3.3.6)-(2.4.19) and the fact that $|V^\star(x_\eta) - V^\star(x)| \le \epsilon/3$ for η small enough by uniform continuity of $V^\star$ in Proposition 3.3.1. Moreover, by observing that $\pi_\varepsilon(x) = \pi_\varepsilon(x_\eta) =: \alpha_0$, we have

$$\begin{aligned}
[\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](x) = \; & \mathbb{E}\Big[\mathbf{f}(x, \alpha_0) + \beta V^\star(\boldsymbol{F}(x, \alpha_0, (\varepsilon^i_1)_{i \le N}, \varepsilon^0_1)) \Big], \\
[\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](x_\eta) = \; & \mathbb{E}\Big[\mathbf{f}(x_\eta, \alpha_0) + \beta V^\star(\boldsymbol{F}(x_\eta, \alpha_0, (\varepsilon^i_1)_{i \le N}, \varepsilon^0_1)) \Big],
\end{aligned}$$

Under $(\mathbf{H_{lip}})$, by using the γ-Hölder property of $V^\star$ with constant $K_\star$ in Proposition 3.3.1, we then get

$$\begin{aligned}
& \big| [\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](x) - [\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](x_\eta) \big| \\
\le \; & 2Kd(x, x_\eta) + \beta K_\star \mathbb{E}\Big[\mathbb{E}\big[d\big(\boldsymbol{F}(x, \alpha_0, (\varepsilon^i_1)_{i \le N}, e), \boldsymbol{F}(x_\eta, \alpha_0, (\varepsilon^i_1)_{i \le N}, e) \big)^\gamma \big]_{e := \varepsilon^0_1} \Big] \\
\le \; & 2Kd(x, x_\eta) + \beta K_\star \mathbb{E}\Big[\mathbb{E}\big[d\big(\boldsymbol{F}(x, \alpha_0, (\varepsilon^i_1)_{i \le N}, e), \boldsymbol{F}(x_\eta, \alpha_0, (\varepsilon^i_1)_{i \le N}, e) \big) \big]_{e := \varepsilon^0_1} \Big]^\gamma \\
\le \; & C d_N(x, x_\eta)^\gamma \le C \eta^\gamma.
\end{aligned}$$

for some constant C. Therefore, $\big| [\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](x) - [\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star](x_\eta) \big| \le \epsilon/3$, and, plugging into (3.3.7), we obtain $\mathcal{T}^{\boldsymbol{\pi}_\epsilon} V^\star(x) - V^\star(x) \ge -\epsilon$, for all $x \in \mathcal{X}^N$, which means that $\boldsymbol{\pi}_\epsilon$ is ϵ-optimal for $V^\star$. The rest of the assertions in the Theorem follows from the verification result in Proposition 3.3.2. $\square$

3.4 Propagation of chaos results

In this section, we establish propagation of chaos results between the N-individual MDP and the limiting McKean-Vlasov MDP.

We compare the optimal values of each problem. The first step is to compare the Bellman operators. Let $a(\xi, U)$ be a randomized feedback policy, and let $\mathbf{a}$ be a random action valued in A^N. Let us denote

$$\epsilon := \mathbb{E}[\mathcal{W}(\mathbb{P}_{\xi,a(\xi,U)}, \frac{1}{N}\sum_{n=1}^N \delta_{x^n,\mathbf{a}^n})]$$

Our goal is to compare $\mathcal{T}^a V^\star$ to $\mathcal{T}_N^{\mathbf{a}} V^\star$.

Lemma 3.4.1 *We have*

$$|\mathbb{E}[\mathcal{T}^a V^\star - \mathcal{T}_N^{\mathbf{a}} V^\star]| \;\leq\; 2(K_f + \beta K_F)\mathbb{E}[\mathcal{W}(\mathbb{P}_{\xi,a(\xi,U)}, \frac{1}{N}\sum_{n=1}^N \delta_{x^n,\mathbf{a}^n})] + \beta K_F M_N$$

Proof. We have

$$\begin{aligned}
& \mathbb{E}[\mathcal{T}^a V^\star - \mathcal{T}_N^{\mathbf{a}} V^\star] \\
= \;& \mathbb{E}\Big[\mathop{\mathbb{E}}_{\xi\sim\frac{1}{N}\sum\delta_{x_n}}\Big[f(\xi,a(\xi,U),\mathbb{P}_{\xi,a(\xi,U)}) + \beta V^\star\big(\mathbb{P}^0_{F(\xi,a(\xi,U),\mathcal{L}(\xi,a(\xi,U)),\varepsilon_1^n,\varepsilon_1^0}\big)\Big] \\
- \;& \Big(\frac{1}{N}\sum_{i=1}^N f(x^i,\mathbf{a}^i,\frac{1}{N}\sum_{n=1}^N \delta_{x^n,\mathbf{a}^n}) + \beta V^\star\big(\frac{1}{N}\sum_{i=1}^N \delta_{F(x^i,\mathbf{a}^i,\frac{1}{N}\sum_{n=1}^N \delta_{x^i,\mathbf{a}^i},\varepsilon_1^i,\varepsilon_1^0)}\big)\Big)\Big] \\
= \;& \mathbb{E}\Big[\mathop{\mathbb{E}}_{\xi\sim\frac{1}{N}\sum\delta_{x_n}}\Big[f(\xi,a(\xi,U),\mathbb{P}_{\xi,a(\xi,U)})\Big] - \frac{1}{N}\sum_{i=1}^N f(x^i,\mathbf{a}^i,\frac{1}{N}\sum_{n=1}^N \delta_{x^n,\mathbf{a}^n})\Big] \\
+ \;& \beta\mathbb{E}\Big[\mathop{\mathbb{E}}_{\xi\sim\frac{1}{N}\sum\delta_{x_n}}\Big[V^\star\big(\mathbb{P}^0_{F(\xi,a(\xi,U),\mathcal{L}(\xi,a(\xi,U)),\varepsilon_1^n,\varepsilon_1^0}\big)\Big] - V^\star\big(\frac{1}{N}\sum_{i=1}^N \delta_{F(x^i,\mathbf{a}^i,\frac{1}{N}\sum_{n=1}^N \delta_{x^i,\mathbf{a}^i},\varepsilon_1^i,\varepsilon_1^0)}\big)\Big)\Big]
\end{aligned}$$

It is easy to show that we have

$$\mathop{\mathbb{E}}_{\xi\sim\frac{1}{N}\sum\delta_{x_n}}\Big[f(\xi,a(\xi,U),\mathbb{P}_{\xi,a(\xi,U)})\Big] - \frac{1}{N}\sum_{i=1}^N f(x^i,\mathbf{a}^i,\frac{1}{N}\sum_{n=1}^N \delta_{x^n,\mathbf{a}^n}) \leq 2K_f\mathcal{W}(\mathbb{P}_{\xi,a(\xi,U)}, \frac{1}{N}\sum_{n=1}^N \delta_{x^n,\mathbf{a}^n})$$

as it is a difference of a Lipschitz function applied to $\mathbb{P}_{\xi,a(\xi,U)}$ and $\frac{1}{N}\sum_{n=1}^N \delta_{x^n,\mathbf{a}^n}$. Let

us focus on the second term:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\xi\sim\frac{1}{N}\sum\delta_{x_n}}\Big[V^\star\big(\mathbb{P}^0_{F(\xi,a(\xi,U),\mathcal{L}(\xi,a(\xi,U)),\varepsilon_1^n,\varepsilon_1^0}\big)\Big]-V^\star\Big(\frac{1}{N}\sum_{i=1}^N\delta_{F(x^i,\mathbf{a}^i,\frac{1}{N}\sum_{n=1}^N\delta_{x^i,\mathbf{a}^i},\varepsilon_1^i,\varepsilon_1^0)}\Big)\Big]\\
\leq\;& C\mathop{\mathbb{E}}_{\xi\sim\frac{1}{N}\sum\delta_{x_n}}\Big[\mathcal{W}(\mathbb{P}^0_{F(\xi,a(\xi,U),\mathcal{L}(\xi,a(\xi,U)),\varepsilon_1^n,\varepsilon_1^0},\frac{1}{N}\sum_{i=1}^N\delta_{F(x^i,\mathbf{a}^i,\frac{1}{N}\sum_{n=1}^N\delta_{x^i,\mathbf{a}^i},\varepsilon_1^i,\varepsilon_1^0)})^\gamma\Big]\\
\leq\;& C\mathop{\mathbb{E}}_{\xi\sim\frac{1}{N}\sum\delta_{x_n}}\Big[\mathcal{W}(\mathbb{P}^0_{F(\xi,a(\xi,U),\mathcal{L}(\xi,a(\xi,U)),\varepsilon_1^n,\varepsilon_1^0},\frac{1}{N}\sum_{i=1}^N\delta_{F(x^i,\mathbf{a}^i,\frac{1}{N}\sum_{n=1}^N\delta_{x^i,\mathbf{a}^i},\varepsilon_1^i,\varepsilon_1^0)})\Big]^\gamma
\end{aligned}$$

For any i.i.d. random variables $(U^n,\tilde{\varepsilon}_1^n)_{n\leq N}$ such that $(U^n,\tilde{\varepsilon}_1^n)\overset{d}{=}(U,\varepsilon_1^n)$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{W}\Big(\mathbb{P}^0_{F(\xi,a(\xi,U),\mathcal{L}(\xi,a(\xi,U)),\varepsilon_1^n,\varepsilon_1^0},\frac{1}{N}\sum\delta_{F(x^i,\mathbf{a}^i,\frac{1}{N}\sum_{n=1}^N\delta_{x^i,\mathbf{a}^i},\varepsilon_1^i,\varepsilon_1^0)}\Big)\Big]\\
\leq\;&\mathbb{E}\Big[\mathcal{W}\Big(\mathbb{P}^0_{F(\xi,a(\xi,U),\mathcal{L}(\xi,a(\xi,U)),\varepsilon_1^n,\varepsilon_1^0},\frac{1}{N}\sum\delta_{F(\xi_n,a(\xi_n,U_n),\mathcal{L}(\xi,a(\xi,U)),\tilde{\varepsilon}_1^n,\varepsilon_1^0)}\Big)\Big]\\
+\;&\mathbb{E}\Big[\mathcal{W}\Big(\frac{1}{N}\sum\delta_{F(\xi_n,a(\xi_n,U_n),\mathcal{L}(\xi,a(\xi,U)),\tilde{\varepsilon}_1^n,\varepsilon_1^0)},\frac{1}{N}\sum\delta_{F(x^i,\mathbf{a}^i,\frac{1}{N}\sum_{n=1}^N\delta_{x^i,\mathbf{a}^i},\varepsilon_1^i,\varepsilon_1^0}\Big)\Big]\\
\leq\;&M_N+\mathbb{E}\Big[\mathcal{W}\Big(\frac{1}{N}\sum\delta_{F(\xi_n,a(\xi_n,U_n),\mathcal{L}(\xi,a(\xi,U)),\tilde{\varepsilon}_1^n,\varepsilon_1^0)},\frac{1}{N}\sum\delta_{F(x^i,\mathbf{a}^i,\frac{1}{N}\sum_{n=1}^N\delta_{x^i,\mathbf{a}^i},\varepsilon_1^i,\varepsilon_1^0}\Big)\Big]
\end{aligned}$$

Let us now focus on

$$\mathbb{E}\Big[\mathcal{W}\Big(\frac{1}{N}\sum\delta_{F(\xi_n,a(\xi_n,U_n),\mathcal{L}(\xi,a(\xi,U)),\tilde{\varepsilon}_1^n,\varepsilon_1^0)},\frac{1}{N}\sum\delta_{F(x^i,\mathbf{a}^i,\frac{1}{N}\sum_{n=1}^N\delta_{x^i,\mathbf{a}^i},\varepsilon_1^i,\varepsilon_1^0}\Big)\Big].$$

The reason why we allowed to take a random family $(\tilde{\varepsilon}_1^n)_{n\leq N}$ with same distribution as $(\varepsilon_1^n)_{n\leq N}$ instead of just taking $(\varepsilon_1^n)_{n\leq N}$ was to allow us to *couple* things nicely before using the formula $\mathcal{W}(\frac{1}{N}\sum_{n=1}^N\delta_{y_n},\frac{1}{N}\sum_{n=1}^N\delta_{z_n})\leq\frac{1}{N}\sum_{n=1}^N d(y_n,z_n)$ in order to obtain a good estimation. It is known that there always exists a transport map realizing an optimal coupling between two measures of the form $\frac{1}{N}\sum_{n=1}^N\delta_{y_n}$ and $\frac{1}{N}\sum_{n=1}^N\delta_{z_n}$, that is, a function T such that $\frac{1}{N}\sum_{n=1}^N\delta_{z_n}=\frac{1}{N}\sum_{n=1}^N\delta_{T(y_n)}$. and such that $\mathcal{W}(\frac{1}{N}\sum_{n=1}^N\delta_{y_n},\frac{1}{N}\sum_{n=1}^N\delta_{z_n})=\frac{1}{N}\sum_n d(y_n,T(y_n))$. In this finite support framework, notice that there necessarily exists a permutation $\sigma\in\mathfrak{S}_n$ such that $T(y_n)=z_{\sigma_n}$. In other words, there always exists such permutation σ such that $\mathcal{W}(\frac{1}{N}\sum_{n=1}^N\delta_{y_n},\frac{1}{N}\sum_{n=1}^N\delta_{z_n})=\frac{1}{N}\sum_n d(y_n,z_{\sigma_n})$. This permutation of course depends upon y and z, so let us denote it $\sigma^{y,z}$. Because the number of permutations is finite, it is clear that $(y,z)\mapsto\sigma^{y,z}$ is a measurable function. Let us thus consider the random variable $\sigma^{(\xi_n,a(\xi_n,U_n))_{n\leq N},(x_n,\mathbf{a}_n)_{n\leq N}}$ that we shall, to simplify notation, simply note σ. Notice that as $(\xi_n,\xi_n,a(\xi_n,U_n))_{n\leq N}\perp\!\!\!\perp(\varepsilon_1^n)_{n\leq N}$, we clearly have that $(\varepsilon_1^{\sigma_n})_{n\leq N}$ satisfies the required condition for $(\tilde{\varepsilon}_1^n)_{n\leq N}$, i.e. $(\varepsilon_1^{\sigma_n})_{n\leq N}\overset{d}{=}(\varepsilon_1^n)_{n\leq N}$ and

$(\varepsilon_1^{\sigma_n})_{n\le N} \perp\!\!\!\perp (\xi_n, \xi_n, a(\xi_n, U_n))_{n\le N}$. Therefore the above relation applies to $(\tilde{\varepsilon}_1^n)_{n\le N} := (\varepsilon_1^{\sigma_n})_{n\le N}$. We are thus reduced to study

$$
\begin{aligned}
& \mathbb{E}\Big[\mathcal{W}\Big(\frac{1}{N}\sum \delta_{F(\xi_n, a(\xi_n,U_n), \mathcal{L}(\xi, a(\xi,U)), \varepsilon_1^{\sigma_n}, \varepsilon_1^0)}, \frac{1}{N}\sum \delta_{F(x_n, \mathbf{a}_n, \frac{1}{N}\sum_n \delta_{x_n,\mathbf{a}_n}, \varepsilon_1^n, \varepsilon_1^0)}\Big)\Big] \\
& \mathbb{E}\Big[\mathcal{W}\Big(\frac{1}{N}\sum \delta_{F(\xi_{\sigma_n^{-1}}, a(\xi_{\sigma_n^{-1}}, U_{\sigma_n^{-1}}), \mathcal{L}(\xi, a(\xi,U)), \varepsilon_1^n, \varepsilon_1^0)}, \frac{1}{N}\sum \delta_{F(x_n, \mathbf{a}_n, \frac{1}{N}\sum_n \delta_{x_n,\mathbf{a}_n}, \varepsilon_1^n, \varepsilon_1^0)}\Big)\Big] \\
\le\ & \frac{1}{N}\sum \mathbb{E}\Big[d(F(\xi_{\sigma_n^{-1}}, a(\xi_{\sigma_n^{-1}}, U_{\sigma_n^{-1}}), \mathcal{L}(\xi, a(\xi,U)), \varepsilon_1^n, \varepsilon_1^0), F(x_n, \mathbf{a}_n, \frac{1}{N}\sum_n \delta_{x_n,\mathbf{a}_n}, \varepsilon_1^n, \varepsilon_1^0))\Big]
\end{aligned}
$$

By condition w.r.t. $(\xi_n, U_n)_{n\le N}$ and using the regularity in expectation of F given by $(\mathbf{HF_{lip}})$, we obtain

$$
\begin{aligned}
& \frac{1}{N}\sum \mathbb{E}\Big[d(F(\xi_{\sigma_n^{-1}}, a(\xi_{\sigma_n^{-1}}, U_{\sigma_n^{-1}}), \mathcal{L}(\xi, a(\xi,U)), \varepsilon_1^n, \varepsilon_1^0), F(x_n, \mathbf{a}_n, \frac{1}{N}\sum_n \delta_{x_n,\mathbf{a}_n}, \varepsilon_1^n, \varepsilon_1^0))\Big] \\
\le\ & K_F \frac{1}{N}\sum \mathbb{E}\Big[d((\xi_{\sigma_n^{-1}}, a(\xi_{\sigma_n^{-1}}, U_{\sigma_n^{-1}}), (x_n, a_n))) + \mathcal{W}\Big(\mathcal{L}(\xi, a(\xi,U)), \frac{1}{N}\sum_n \delta_{x_n,\mathbf{a}_n}\Big)\Big] \\
=\ & K_F \mathbb{E}[\mathcal{W}(\frac{1}{N}\sum \delta_{(\xi_n, a(\xi_n,U_n))}, \frac{1}{N}\sum_n \delta_{x_n,\mathbf{a}_n}) + \mathcal{W}\Big(\mathcal{L}(\xi, a(\xi,U)), \frac{1}{N}\sum_n \delta_{x_n,\mathbf{a}_n}\Big)\Big] \\
\le\ & K_F(M_N + 2\epsilon)
\end{aligned}
$$

Combining all the above computations, we obtain

$$
|\mathbb{E}[\mathcal{T}^a V^\star - \mathcal{T}_N^{\mathbf{a}} V^\star]| \ \le\ 2K_f\epsilon + \beta K_F(M_N + 2\epsilon) = 2(K_f + \beta K_F)\epsilon + \beta K_F M_N
$$

which concludes the proof. □

We are thus clearly reduced to study how well one can couple randomized feedback policies of the form $a(\xi, U)$ and A^N-valued random variables $\mathbf{a}$ to have a small term

$$
\mathbb{E}[\mathcal{W}(\mathbb{P}_{\xi,a(\xi,U)}, \frac{1}{N}\sum_{n=1}^N \delta_{x^n,\mathbf{a}^n})]
$$

Given any randomized feedback policies of the form $a(\xi, U)$, one possibility is to define

$\mathbf{a}^a$ such that $\mathbf{a}^{a,n} = a(\xi^{\sigma_n^{\xi,x}}, U^n)$ for all $n \leq N$. Indeed, we have

$$\begin{aligned}
& \mathbb{E}[\mathcal{W}(\mathbb{P}_{\xi,a(\xi,U)}, \frac{1}{N}\sum_{n=1}^N \delta_{x^n,\mathbf{a}^{a,n}})] \\
\leq\ & \mathbb{E}[\mathcal{W}(\mathbb{P}_{\xi,a(\xi,U)}, \frac{1}{N}\sum \delta_{\xi^{\sigma_n^{\xi,x}},a(\xi^{\sigma_n^{\xi,x}},U^n)}) + \mathcal{W}(\frac{1}{N}\sum \delta_{\xi^{\sigma_n^{\xi,x}},a(\xi^{\sigma_n^{\xi,x}},U^n)}, \frac{1}{N}\sum_{n=1}^N \delta_{x^n,\mathbf{a}^{a,n}})] \\
\leq\ & M_N + \mathbb{E}[\mathcal{W}(\frac{1}{N}\sum \delta_{\xi^{\sigma_n^{\xi,x}},a(\xi^{\sigma_n^{\xi,x}},U^n)}, \frac{1}{N}\sum_{n=1}^N \delta_{x^n,\mathbf{a}^{a,n}})] \\
\leq\ & M_N + \mathbb{E}[\frac{1}{N}\sum_{n=1}^N d(\xi^{\sigma_n^{\xi,x}}, x^n)] \leq 2M_N
\end{aligned}$$

On the other hand, given a deterministic A^N-valued variables $\mathbf{a}$, we can clearly define $a^{\mathbf{a}}(x,U)$ such that $\mathcal{L}(a^{\mathbf{a}}(\xi,U)) = \frac{1}{N}\sum_n \delta_{x^n,\mathbf{a}^n}$.

Theorem 3.4.1 *Let* $\gamma = \min\Big(1, \frac{|\ln\beta|}{\ln(2K_F)_+}\Big)$. *We have*

$$\|V_N(x) - V\Big(\frac{1}{N}\sum_{n\leq N}\delta_{x_n}\Big)\|_{x\in\mathcal{X}^N} \leq \mathcal{O}(M_N^\gamma)$$

Proof. We have

$$\begin{aligned}
V^\star(x) &= \mathcal{T}V^\star(x) \\
&= \sup_{a(x,u)} \mathcal{T}^a V^\star(x) \leq \sup_{a(x,u)} \mathcal{T}_N^{\mathbf{a}^a} V^\star(x) + 2(K_f + \beta K_F)2M_N + \beta K_F M_N \\
&\leq \mathcal{T}_N V^\star(x) + (2K_f + \beta 3K_F)M_N
\end{aligned}$$

Likewise, we have

$$\begin{aligned}
V^\star(x) &= \mathcal{T}V^\star(x) = \sup_{a(x,u)} \mathcal{T}^a V^\star(x) \geq \sup_{\mathbf{a}} \mathcal{T}_N^{a^{\mathbf{a}}} V^\star(x) \\
&\geq \sup_{\mathbf{a}} \mathcal{T}_N^{\mathbf{a}} V^\star(x) - \beta K_F M_N \geq \mathcal{T}_N V^\star(x) - \beta K_F M_N^\gamma
\end{aligned}$$

Now, recalling that $V_N(x) = \mathcal{T}_N V_N(x)$, we have:

$$(V_N - V)(x) \leq (\mathcal{T}_N V_N - \mathcal{T}_N V(x)) + ((1+K_F)M_N)^\gamma$$

and thus

$$(V_N - V)(x) \leq \beta \sup_{x\in\mathcal{X}^N} (V_N - V)(x) + ((1+K_F)M_N)^\gamma$$

which implies

$$\sup_{x\in\mathcal{X}^N}(V - V^\star(x)) \quad \leq \quad \frac{((1+K_F)M_N)^\gamma}{1-\beta}$$

and thus $V_N(x) \leq V(x) + \frac{((1+K_F)M_N)^\gamma}{1-\beta}$ for all $x \in \mathcal{X}^N$. By a similar argument, one can prove that $V_N(x) \geq V(x) + \mathcal{O}(M_N^\gamma)$, and thus, $\|V - V_N\| = \mathcal{O}(M_N^\gamma)$, which concludes the proof. □

It is now possible to link policies of both problems to each others, using the comparison of operators, of the value functions, and the verification results of both problems.

Let a be an ε-optimal randomized feedback policy for the McKean-Vlasov MDP. We thus have

$$V(x) \geq \mathcal{T}^a V(x) - \beta\varepsilon, \quad \forall x \in \mathcal{X}^N.$$

Thus, we have

$$V(x) \geq \mathcal{T}_N^{\mathbf{a}^a} V(x) - \beta(\varepsilon + \mathcal{O}(M_N)^\gamma), \quad \forall x \in \mathcal{X}^N.$$

and thus

$$V_N(x) \geq \mathcal{T}_N^{\mathbf{a}^a} V_N(x) - \beta\varepsilon - \mathcal{O}(M_N)^\gamma), \quad \forall x \in \mathcal{X}^N.$$

which, by the verification result, implies that $\mathbf{a}^a$ is $\frac{\beta\varepsilon+\mathcal{O}(M_N)^\gamma)}{1-\beta}$-optimal for V_N. However, we can improve this policy by simply considering $(\hat{\mathbf{a}}^a)^n = a(x^n, U^n)$. Indeed, we have

$$\begin{aligned}
& \mathcal{T}_N^a V_N(x) \geq \mathbb{E}[\mathcal{T}_N^a V_N(\xi)] - \mathbb{E}[\mathcal{W}(\frac{1}{N}\sum \delta_{\xi^n}, \frac{1}{N}\sum \delta_{x^n})]^\gamma \\
= \quad & \mathbb{E}[\mathcal{T}_N^{\mathbf{a}^a} V_N(\xi^{\sigma^{x,\xi}})] - \mathcal{O}(M_N^\gamma) \geq \mathbb{E}[\mathcal{T}_N^{\mathbf{a}^a} V_N(x)] - 2\mathcal{O}(M_N^\gamma) \\
\geq \quad & V_N(x) - \mathcal{O}(\varepsilon + M_N^\gamma)
\end{aligned}$$

and we conclude by the verification result that a is $\mathcal{O}(\varepsilon + M_N^\gamma)$-optimal for V_N.

Conversely, the comparison of operators and the verification result for the MKV-MDP imply that given an ε-optimal feedback policy $\mathbf{a}$ for the N-individual MDP, the randomized feedback policy $a^{\mathbf{a}}$ is $\mathcal{O}(\varepsilon + M_N^\gamma)$-optimal for V.

3.5 Toy example for advertising

In this section, we provide an example illustrating the utility of the results of this chapter. A careful look at the proofs of these results shows that actually, all that we have done

was to compare the Bellman fixed point operator of the N-individual MDP to another operator. This operator happens to be the Bellman operator of the MKV-MDP from previous chapter, but we have not really used this fact. In other words, it is possible to see the previous section as simply performing a mean-field approximation of the Bellman operator of the N-individual MDP, independently of the fact that the resulting operator is in turn the Bellman operator of the MKV-MDP.

Therefore, one question could be: given that often, one solves an MDP via the study of its Bellman operator, is it really useful to link the N-individual MDP to the MKV-MDP, rather than simply linking its Bellman-operator to its mean-field approximation?

The answer is yes, and the following example illustrates it. The above question relies on the assumption that the MKV-MDP would always be solved via its Bellman operator. However, we will solve the MKV-MDP of the next example not simply by using the Bellman operator, but also with other tools directly related to the MKV-MDP itself. Solving it analytically entirely with the Bellman operator would be very hard, and thus, having the possibility to perform various analysis on the MKV-MDP rather than being limited to a mean-field Bellman-operator is in practice very useful.

The model is specified as follows: We consider a targeted advertising situation. At each time $t \in \mathbb{N}$, each individual connects to a website and can receive an ad from a given Company. The Company's goal is to use targeted ads to attract people as quickly as possible while minimizing its advertising cost. To fix ideas, let us say that the Company is selling phones.

- State space $\mathcal{X} = \{0, 1\}$ ($x = 0$ means "not being a customer of C", and $x = 1$ means being one of company C).
- Action space $A := \{0, 1\}$ ($a = 0$ means "SN does not display an ad to the user", and $a = 1$ means displaying one).
- Idiosyncratic noises $(\varepsilon_t^i)_{i \in \mathbb{N}_\star, t \in \mathbb{N}}$ where $\varepsilon_t^i \sim \mathcal{U}([0, 1])$ represents the time spent by the i-th individual on a forum about phones during day t.
- No common noise.
- State transition function: for $(x, a) \in \mathcal{X} \times A$, $e \in [0, 1]$, $\mu \in \mathcal{P}(\mathcal{X})$,

$$F(x, \mu, a, e) = \begin{cases} \mathbf{1}_{e > \mu(\{0\}) - \eta a} & \text{if} \quad x = 0 \\ \mathbf{1}_{e < \mu(\{1\}) + \eta a} & \text{if} \quad x = 1, \end{cases}$$

  for some parameter $\eta > 0$, measuring the efficiency of an ad for incentive to become a customer of C. The interpretation is the following: if a user is not a customer of

C ($x = 0$), then she will be more likely to become a customer of C if the proportion of people that are not customers of C is small (i.e. $\mu(\{0\})$ is small) and if an ad has been sent to him (i.e. $a = 1$), while spending enough time e in forum. On the other hand, if a user is already a customer of C ($x = 1$), she will be more likely to stay a customer of C if the proportion $\mu(\{1\})$ of customers of C is large, and if an ad has been sent to him ($a = 1$). Here $\eta > 0$ is a an efficiency parameter of ad for incentive to become a customer of C.

- Reward function: for $(x, a) \in \mathcal{X} \times A$,

$$f(x, a) \;=\; x - ca,$$

for some $c > 0$ representing an ad cost. This means that if the user is a customer of C ($x = 1$), she contributes to the revenue of the company C, but if C had to make SN send him an ad ($a = 1$), it costs c to the company.

One can easily verify that ($\mathbf{HF_{lip}}$) and ($\mathbf{Hf_{lip}}$) are satisfied. From Theorem 3.4.1, we thus know that propagation of chaos holds true. A useful reformulation of F is given by

$$F(x, a, \mu, e) = \mathbf{1}_{\epsilon(e,x)<p+\eta a}$$

where $\epsilon(e, x) := (1 - e)(1 - x) + ex$ is e when $x = 1$ and $1 - e$ when $x = 0$. Fix an initial state variable ξ and a control α. Let p_t (resp. q_t) denote the Bernoulli parameter of $\mathbb{P}_{X_t^{\xi,\alpha}}$ (resp. α_t) for $t \in \mathbb{N}$. Let $\varepsilon_{t+1}^X = \epsilon(\varepsilon_{t+1}, X_t)$. Then $(p_t)_{t\in\mathbb{N}}$ follows the dynamics:

$$\begin{aligned} p_{t+1} &= \mathbb{P}[F(X_t^{\xi,\alpha}, \mathbb{P}_{X_t^{\xi,\pi}}, \alpha_t, \varepsilon_{t+1}) = 1] \;=\; \mathbb{P}\big[\varepsilon_{t+1}^X < p_t + \eta\alpha_t\big] \\ &= \mathbb{P}\big[\varepsilon_{t+1}^X < p_t\big] + \mathbb{P}\big[p_t \le \varepsilon_{t+1}^X < p_t + \eta, \alpha_t = 1\big]. \end{aligned}$$

The conditional law of ε_{t+1}^X knowing $(\Gamma, (\varepsilon_s)_{s\le t})$ is constant equal to $\mathcal{U}([0, 1])$, thus ε_{t+1}^X is uniform and independent of $(\Gamma, (\varepsilon_s)_{s\le t})$, thus

$$p_{t+1} = p_t + q_t \min(\eta, 1 - p_t).$$

On the other hand, notice that the gain functional can be rewritten as

$$\begin{aligned} V^\pi(\xi) &:= \mathbb{E}\big[\sum_{t\in\mathbb{N}} \beta^t f(X_t^{\xi,\pi}, \alpha_t)\big] \;=\; \mathbb{E}\big[\sum_{t\in\mathbb{N}} \beta^t \big(X_t^{\xi,\pi} - c\alpha_t\big)\big] \\ &= \sum_{t\in\mathbb{N}} \beta^t (p_t - cq_t). \end{aligned}$$

This derivation leads us to consider the deterministic control problem on $[0,1]$ with dynamics:

$$p_{t+1} \;=\; \Phi_\eta(p_t,q_t) \;:=\; p_t+\min(\eta,1-p_t)q_t, \qquad t\in\mathbb{N}, \quad p_0 \;=\; p\in[0,1],$$

controlled by the deterministic sequence $q=(q_t)_t$ valued in $[0,1]$, and with value function:

$$\mathcal{V}(p) \;=\; \sup_{q\in[0,1]^{\mathbb{N}}} \mathcal{V}^q(p), \qquad \mathcal{V}^q(p) \;:=\; \sum_{t\in\mathbb{N}} \beta^t(p_t-cq_t).$$

Notice that the corresponding dynamic programming equation takes the form of the fixed point Bellman equation:

$$V(p) \;=\; \sup_{q\in[0,1]} \big[p-cq+\beta V(\Phi_\eta(p,q))\big], \quad p\in[0,1],$$

The above arguments show the equivalence between the MKV-MDP and the deterministic problem (3.5.4)-(3.5.5): fix some arbitrary initial state function ξ with Bernoulli parameter $p=\mathbb{P}[\xi(\Gamma)=1]$. Then,

- For any control α with policy π of the MKV-MDP, by defining $q=(q_t)$ with $q_t=\mathbb{P}[\alpha_t=1]$, we have $V^\pi(\xi)=\mathcal{V}^q(p)$
- Conversely, for any $q=(q_t)_t\in[0,1]^{\mathbb{N}}$, by defining $\alpha=(\alpha_t)_t$ with $\alpha_t=1_{U_t\le q_t}$, we have $V^\pi(\xi)=\mathcal{V}^q(p)$.

Besides reducing the MkV-MDP to a simpler problem, this correspondence shows that in the MkV-MDP, one can restrict to purely randomized controls of the form $\alpha_t:=\mathbf{1}_{U_t<q_t}$, i.e. not depending upon the state X_t.

Proposition 3.5.1 *Let us define the function $\hat{q}:[0,1]\to[0,1]$, depending on the position of $\frac{c}{\eta}$ relative to $[\beta,\frac{\beta}{1-\beta}]$:*

- *If $\frac{c}{\eta}<\beta$,*

$$\hat{q}(p) \;:=\; \begin{cases} 1, & \text{for} \quad p<1-c\frac{1-\beta}{\beta} \\ 0, & \text{for} \quad 1-c\frac{1-\beta}{\beta}\le p\le 1. \end{cases}$$

- *If $\beta\le\frac{c}{\eta}<\frac{\beta}{1-\beta}$,*

$$\hat{q}(p) \;:=\; \begin{cases} 1, & \text{for} \quad p<1-2\eta \\ \frac{1-\eta-p}{\eta}, & \text{for} \quad 1-2\eta\le p<1-(2-\beta)\eta \\ 1, & \text{for} \quad 1-(2-\beta)\eta\le p<1-c\frac{1-\beta}{\beta} \\ 0, & \text{for} \quad 1-c\frac{1-\beta}{\beta}\le p\le 1. \end{cases}$$

- *If* $\frac{\beta}{1-\beta} \leq \frac{c}{\eta}$,

$$\hat{q}(p) \ := \ 0, \qquad \textit{for all } p \in [0,1].$$

Then, the feedback control $q^\star = (q_t^\star)_t$ *with* $q_t^\star = \hat{q}(p_t)$, $t \in \mathbb{N}$, *is an optimal control for problem* (3.5.4)-(3.5.5), *and thus the stationary randomized control* $\alpha^\star = (\alpha_t^\star)_{t\in\mathbb{N}}$ *with* $\alpha_t^\star = \mathbf{1}_{U_t<\hat{q}(p_t)}$ *is optimal for the MkV-MDP.*

Proof. • *Step 1:* Tt will be useful to see this control problem completely in terms of "increments" from p_t to p_{t+1}, rather than in terms of the control q_t. In other words, it will be easier to consider that we directly choose the increment $p_{t+1} - p_t$, instead of a control q_t determining this increment. We thus rewrite the gain functional as

$$\begin{aligned}
\mathcal{V}^q(p_0) &= \sum_{t\in\mathbb{N}} \beta^t (p_t - cq_t) \\
&= \sum_{t\in\mathbb{N}} \beta^t \left(p_0 + \sum_{0\leq s<t} (p_{s+1} - p_s) - c\frac{p_{t+1}-p_t}{\min(\eta, 1-p_t)} \right) \\
&= \frac{p_0}{1-\beta} + \sum_{t\in\mathbb{N}} \beta^t (p_{t+1}-p_t) \left(\frac{\beta}{1-\beta} - \frac{c}{\min(\eta,1-p_t)} \right) \\
&= \frac{p_0}{1-\beta} + \sum_{t\in\mathbb{N}} \beta^t (p_{t+1}-p_t) r(p_t),
\end{aligned}$$

where we rearranged the sums in the third equality, and where $r(p) := \frac{\beta}{1-\beta} - \frac{c}{\min(\eta,1-p)}$. Notice that r is constant on $[0, 1-\eta]$, then decreases. So if $r(0)$ is negative, $r(p_t)$ will always be negative, thus the sum is negative, and thus the best thing to do is nothing: $p_t = p_0 \ \forall t \in \mathbb{N}$, corresponding to the control $q_t := 0 \ \forall t \in \mathbb{N}$. This trivial case is obtained under the assumption that $r(0) \leq 0$, which is equivalent to $\frac{c}{\eta} \geq \frac{\beta}{1-\beta}$, corresponding to the third case disjunction.

In the rest of the proof, we shall then focus on the case where $r(0) > 0$, i.e., $\frac{c}{\eta} < \frac{\beta}{1-\beta}$.

• *Step 2:* The nonincreasing function r starts from a positive value $r(0)$, and only becomes negative after the solution to $r(p) = 0$, given by $\bar{p} := 1 - c\frac{1-\beta}{\beta} > 1 - \eta$. Thus, for the same reason as in *Step 1*, as soon as $p_t \geq \bar{p}$, (say from $t = \bar{t}$), the optimal strategy is to do nothing, because $\sum_{t=\bar{t}}^{\infty} \beta^t (p_{t+1} - p_t) r(p_t) \leq 0$. Consequently, the optimal trajectory will remain constant after we get in the interval $[\bar{p}, 1]$ (and thus the optimal control will be $q_t = 0$ from that point).

We now analyze different situations:

(i) Assume that there is some point $p_{t_0} \in [1-\eta, \bar{p})$. Then $r(p_{t_0}) > 0$, and a possible strategy is to jump to $p_{t_0+1} = 1$ by taking $q_{t_0} = 1$. Moreover, for any strategy,

we have $\sum_{t=t_0}^{\infty} \beta^t (p_{t+1} - p_t) r(p_t) \leq \beta^{t_0}(1 - p_{t_0}) r(p_{t_0})$. Therefore in this case, the optimal strategy is indeed to jump directly to 1 from t_0.

(ii) Assume that for some t_0 we have $p_{t_0} \leq 1 - \eta$ and $p_{t_0+1} \leq 1 - \eta$. Notice that the gain function between p_{t_0} and p_{t_0+2} is given by

$$g(p_{t_0+1}) \;=\; p_{t_0} - c\frac{p_{t_0+1} - p_{t_0}}{\eta} + \beta\big(p_{t_0+1} - c\frac{p_{t_0+2} - p_{t_0+1}}{\eta}\big),$$

and its derivative is equal to

$$g'(p_{t_0+1}) \;=\; (1-\beta)\big(\frac{\beta}{1-\beta} - \frac{c}{\eta}\big) \;>\; 0,$$

which is negative, as $\frac{c}{\eta} < \frac{\beta}{1-\beta}$. This means that for an optimal strategy $p_{t_0+1} \leq 1 - \eta$, it cannot be moved to the right since otherwise it would increase the gain contradicting its optimality. In other words, an optimal $p_{t_0+1} \leq 1 - \eta$ should be associated to a control $q_{t_0} = 1$, and this can only occur when $p_{t_0} \leq 1 - 2\eta$ leading to $p_{t_0+1} = \Phi_\eta(p_{t_0}, 1) = p_{t_0} + \eta$.

To sum up *Step 2*, we have dealt with the optimal strategy in the areas $[0, 1 - 2\eta]$ and $[1 - \eta, 1]$: when $p_t \leq 1 - 2\eta$, it is optimal to jump to $p_{t+1} = p_t + \eta$ with a control $q_t = 1$; when $p_t \in [1 - \eta, \bar{p})$, we jump optimally to 1 (with a control $q_t = 1$), and when $p_t \in [\bar{p}, 1]$, we do not act anymore ($q_s = 0$ for $s \geq t$), hence keeping constant $p_s = p_t$ for $s \geq t$.

• *Step 3:* It remains to deal with the case when there is some point $p_{t_0} \in (1 - 2\eta, 1 - \eta)$, for which we only know from Step 2(ii) that p_{t_0+1} should lie in $[1 - \eta, p_{t_0} + \eta]$. Let us consider the gain function from t_0 as a function of $p_{t_0} + 1$:

$$G(p_{t_0+1}) \;:=\; \sum_{t=t_0}^{\infty} \beta^t (p_{t+1} - p_t) r(p_t).$$

From Step 2(i), we know that if $p_{t_0+1} \in [1 - \eta, \bar{p})$, then $p_t = 1$ for $t > t_0 + 1$, and so

$$\begin{aligned} G(p_{t_0+1}) &= (p_{t_0+1} - p_{t_0}) r(p_{t_0}) + \beta(1 - p_{t_0+1}) r(p_{t_0+1}) \\ &= (p_{t_0+1} - p_{t_0}) r(0) + \beta\big((1 - p_{t_0+1})\frac{\beta}{1-\beta} - c\big), \quad p_{t_0+1} \in [1 - \eta, \bar{p}), \end{aligned}$$

with derivative equal to $G'(p_{t_0+1}) = r(0) - \frac{\beta^2}{1-\beta} = \beta - \frac{c}{\eta}$. If $p_{t_0+1} \in [\bar{p}, 1]$, then we also know from Step 2 that $p_t = p_{t_0+1}$, for $t > t_0 + 1$, and so

$$G(p_{t_0+1}) \;=\; (p_{t_0+1} - p_{t_0}) r(p_{t_0}) = (p_{t_0+1} - p_{t_0}) r(0), \quad p_{t_0+1} \in [\bar{p}, 1],$$

which is increasing on $[\bar{p}, 1)$ as $r(0) = \frac{\beta}{1-\beta} - \frac{c}{\eta} > 0$.

We then make a second case disjunction:

(i) If $\frac{c}{\eta} < \beta$, then G is increasing w.r.t. $p_{t_0+1} \in [1-\eta, 1]$, and thus the optimal strategy is to take p_{t_0+1} as high as possible, i.e., $p_{t_0+1} = p_{t_0} + \eta$ (corresponding to $q_{t_0} = 1$).

(ii) If $\frac{c}{\eta} \geq \beta$, then G is first decreasing on $[1-\eta, \bar{p})$ and then increasing on $[\bar{p}, 1]$. Its maximum on $[1-\eta, p_{t_0}+\eta]$ is then reached either at $p_{t_0+1} = 1-\eta$ or at $p_{t_0+1} = p_{t_0} + \eta$. This situation corresponds to the case when two different phenomenons are fighting against each other, namely a "small" jump to $1-\eta$ vs a big jump to $p_{t_0} + \eta$ (with control $q_{t_0} = 1$). We shall then distinguish the subcases depending on the position of p_{t_0} in $(1-2\eta, 1-\eta)$,:

- If $p_{t_0} \in (1-2\eta, \bar{p}-\eta]$. Then $p_{t_0}+\eta \leq \bar{p}$, and so G is decreasing on $[1-\eta, p_{t_0}+\eta]$, and the maximum is reached at $p_{t_0+1} = 1-\eta$ corresponding to a control $q_{t_0} = (1-\eta-p_{t_0})/\eta$.
- If $p_{t_0} \in (\bar{p}-\eta, 1-\eta)$. We then compare $G(p_{t_0+1})$ at $p_{t_0+1} = 1-\eta$ and $p_{t_0}+\eta$. We have from (3.5.6)-(3.5.7)

$$\begin{aligned} G(1-\eta) &= (1-\eta-p_{t_0}+\beta\eta)r(0), \\ G(p_{t_0}+\eta) &= \eta r(0), \end{aligned}$$

  and then see that $G(1-\eta) > G(p_{t_0}+\eta)$ iff $p_{t_0} < 1-\eta(2-\beta)$. In this case, the optimal strategy is to go to $p_{t_0+1} = 1-\eta$, corresponding to a control $q_{t_0} = (1-\eta-p_{t_0})/\eta$. Otherwise, when $p_{t_0} \in [1-\eta(2-\beta), 1-\eta)$, it is better to jump to $p_{t_0+1} = p_{t_0}+\eta$, corresponding to a control $q_{t_0} = 1$.

□

By the propagation of chaos results in Theorem 3.4.1, we thus know that $a(p,x,u) = \mathbf{1}_{u<\hat{q}(p)}$ yields an $\mathcal{O}(\frac{1}{\sqrt{N}})$-optimal policy for the associated N-agent MDP. Notice that, from the above proof, it is clear that we did not simply use the Bellman equation of the MKV-MDP to solve this problem. We in particularly used variational arguments based on the trajectories of the mean-field proportion of clients. This illustrates that linking the N-agent MDP with the MKV-MDP can be really useful to solve the N-agent MDP.

3.6 Conclusion

We have developed a theory of mean-field Markov decision processes with common noise and open-loop controls, called CMKV-MDP, for general state space and action space,

and rigorously connected it to the N-individual Markov Decision Process it was originally formally derived from. We have provided a rate of convergence of the N-agent model to the CMKV-MDP. We have finally provided an example illustrating the usefulness of this result, by approximately solving a N-agent MDP via its limiting McKean-Vlasov MDP. Interesting developments could be to find other N-agent problems that are hard to solve (or unsolvable) in the N-agent framework but that yet can be approximately solved via the associated McKean-Vlasov MDP. We believe that although the example we provide is a toy model for advertising under social influence, using this general framework to model advertising problems with social influence is one of its natural applications, and we believe that it could be interesting to build and study other models of advertising and social influence within this general framework.

Part II

Behavioral economics models

Chapter 4

Population games in social networks with IESDS solution concepts

Abstract. In this work, we study a game with large population of players, by means of the so-called Iterative Elimination of Strictly Dominated Strategies (IESDS) solution concept. This concept has the advantage to describe a strongly rational iterative mechanism, as opposed to the Nash-equilibrium based on a circular, or fixed point, justification. Our game will also assume that players initially know nothing about each other, but can observe the result of past games as the game repeats itself. Our main result is to show that assuming that players strategies are consistent with the IESDS mechanism, almost all the choices of the population can be predicted with certainty after a few stage games. Furthermore, the distribution of the predicted choices is characterized as the fixed point of an analytical operator, thanks to arguments of mean-field approximation, and our second contribution is to use this characterization to analyze the population's choices, allowing us, in particular, to study social influence phenomenons like the snowball effect and the class repulsion effect.

4.1 Introduction

In this work, we study a game with a population of N player with N large by means of a solution concept that is not the Nash-equilibrium concept, but the stronger rational concept of *Iterated Elimination of Strictly Dominated Strategies (IESDS)*.

To motivate this study, let us briefly overview how game theory evolved towards large population game theories.

The first games that were studied, before game theory was even formalized, were

2-player games (e.g. a game studied in a letter by James Waldegrave in 1713, and the duopoly game by Antoine Cournot in 1838. The study of N-player games, present in the work of Von Neumann and Morgenstern ([91]), followed by the developments of Nash ([68, 43, 66, 67]) and later Aumann ([5]), led to observe propagation of chaos phenomenons: they noticed that predictions for N-player games often "converged" to asymptotic predictions when $N \to \infty$. Robert Aumann later published a seminal paper on games with infinitely many players ([5], [6]), initiating a long list of studies of games with continuum of players, see for instance the large games literature ([47, 41, 13, 42]), and the prolific Mean-field games literature initiated by the seminal paper of Lions and Lasry ([50, 51, 52, 34]), and independently in the engineering community by Caines, Huang and Malhamé [38, 37]. For a detailed exposition of the theory, we refer to the two-volume monograph by Rene Carmona and François Delarue [15].

However, the richer and richer structures of considered games made it harder and harder to study the more complex solution concepts. Therefore, research in large population games seems to have lately entirely focused on the concept of Nash-equilibria.

Despite its mathematical simplicity, the Nash-equilibrium has, as a standalone solution concept, been subject to criticism, because it relies on a circular rational justification. Indeed, given a Nash-equilibrium $(x_n)_{n\leq N}$, the justification that player n will indeed play strategy x_n is that it is the best response to the strategies x_{-n} of the other players. However, the strategies x_{-n} of the other players are themselves justified by assuming that player n plays x_n. There is thus clearly a circular (or fixed point) justification here, which raises some inconsistency issues regarding the non-cooperative aspect among players in Nash-equilibrium. The real strength of Nash-equilibrium is, besides its mathematical simplicity, its experimental validity. It has been observed, both experimentally and in theoretical games with a different solution concept, that players often end up playing a Nash-equilibrium or an ε-Nash-equilibrium. The reason is that most rational mechanisms are based on computing best responses in an iterative manner. It is thus not surprising that if the mechanism converges to a unique strategy for each player, the strategies obtained must form a fixed point for the best response functions. However, we stress that by directly focusing on the study of such fixed points, Nash-equilibria ignores the study of the convergence of the underlying rational mechanism (and thus assumes it). It is however easy to build games who have a unique Nash-equilibrium and such that, yet, rational solution concept mechanisms do not converge to a unique strategy profile, in which case nothing prevents the players to play something else than this Nash-equilibrium.

The IESDS solution concept, studied in this paper, breaks the circularity of Nash-equilibrium by proposing a strategic iterative rational mechanism starting from a universal set of strategies (as opposed to Nash which starts directly from the limiting fixed

point strategy). Let us informally illustrate how it works by modifying the justification of a Nash-equilibrium. The IESDS solution concepts encodes the idea that a solution strategy $(x_n)_{n\leq N}$ is played by all players when

1. some player i has a best response x_i regardless what the others do.

2. Then, any other player, knowing that player i is *rational*, knows that player i will play x_i with certainty. Given this knowledge, some other player $j \neq i$ may have a best response x_j regardless what the others do, *provided that player i indeed plays x_i*.

3. Then again, everyone knows that player j is rational and will thus necessarily play x_j. Likewise, some player $k \notin \{i, j\}$ may have a best response x_k regardless what the others do provided that players i and j indeed play x_i and x_j,

4. etc.

Notice that in this argument, the action of player i is intrinsically justified, the action of player j is justified by the fact that he knows that player i is rational, and the action of player k is justified by the fact that he knows that players i and j are rational. Notice that there is thus no circularity in this mechanism, but instead, a "hierarchy" of rational justifications among players.

In this paper, we shall thus focus on the IESDS. This concept is well known in the game theoretic community, see for instance the work of Milgrom on super-modular games [64]. The Iterated Elimination of Dominated Strategies has not been studied a lot for large population games. See Dufwenberg and Stegeman [26] and Chen, Long, Luo [17] for studies of the IESDS concept for general games, with potentially infinitely many players and strategies.

Here are the main other aspects of our study:

1. We assume virtually no information of players about each other at the beginning. The players are not assumed either to have a Bayesian information, i.e. they don't initially have a statistical representation of the population. Generally when players know nothing about each other, it is difficult for them to rationally eliminate strategies, however:

2. We make the game repeat itself, so that players are able to learn from past games, which will allow most of them to rationally eliminate strategies after a finite number of stage games.

Our first contribution is to prove that even by relying on this complex structure and solution concept, we are able to precisely predict the rational choices of almost all the

players. In particular, we characterize the class-wise proportion of rational choices 1 in the population as an approximate fixed point $p^\star \in [0,1]^K$ of a contracting operator.

Our second contribution is to provide different ways to analyze this class-wise choice distribution $p^\star$ using fixed point analysis methods: numerical methods, example with explicit formula for $p^\star$, and first order expansions for small social influence. Finally, we use these tools to study two qualitative social influence phenomena:

1. **The snowball (or amplification) effect**, occurring in situations where people like to make the same choices as other people, generally resulting in an accentuation of the popularity of the intrinsically more popular choice.

2. **The class repulsion effect:** corresponding to situations where there are two classes that don't like to make the same choices (e.g. left-wing people don't like to support the same reform as right-wing people, and vice versa). This generally results in the polarization of classes on the two choices, i.e. one class appropriates choice 1, and the other one, choice 0.

Outline of the chapter: In Section 7.2, we provide the core framework common to the games studied in this chapter. In Section 4.3, we study a static game based on this framework, with fully informed players. In Section 4.4, we study a repeated version of this game, with initially uninformed players. In Section 4.5, we provide tools and methods for analyzing the prediction made for both games, and we use them to study social phenomenons.

4.2 Core framework

We consider a binary space $\mathcal{X} = \{0,1\}$, representing two possible choices (buying or not a product, subscribing or not to a service, publicly supporting or not an opinion, etc). We consider a population with N players. Each player n, for $n \leq N$, is characterized by:

1. **His intrinsic utilities** $(u_{n,x})_{x\in\mathcal{X}} \in \mathbb{R}^2$**:** for $x \in \mathcal{X}$, $u_{n,x}$ represents the utility that player n has for choice x, *outside of all social influence and interactions*. For instance, if the choice is to buy (choice 1) or not (choice 0) a product, $u_{n,1}$ represents how much player n will like the product in itself, and for instance $u_{n,0} = 0$ represents the neutral utility from not buying it. If the choice is to publicly support an opinion (choice 1) or reject it (choice 0), $u_{n,1}$ represents the intrinsic happiness of player n to support this opinion, in itself, and $u_{n,0}$ to reject it, reflecting how his choice is consistent with his true opinion.

2. **His class $k_n \in \mathcal{K} = \{1, ..., K\}$:** The class of an individual can represent any relevant way to group individuals in a population (age, social class, gender, political orientation, etc). This separation allows to study asymmetric social influence. Multi-class social influence makes it possible to study qualitative phenomenons like class repulsion. This happens when two different classes of individuals don't want to act or think the same way. Let us provide two examples. In the case of supporting or not an opinion, there is a repulsion between the left-wing and right-wing politically oriented classes: a politically left-wing person is reluctant to support the same opinion as a politically right-wing person, and vice versa, even when they *intrinsically* agree on it. In the case of buying or not a product, a well known example is the diet coke. It is known that the main reason why Coca-cola commercialized coke zero is that males were reluctant to buy diet coke because they associated it to a female product. Coke zero, with a less female connotation, was invented for this reason. Thus, males were willing to drink sugar-free coke, but the simple fact that diet coke was seen as a female product dissuaded some of them to do so.

For each $n \leq N$, we denote by $d_n = (k_n, u_n)$ the *data* of player n, and $\boldsymbol{d} = (d_n)_{n \leq N}$.

Game theoretic notations: In the sequel, we shall adopt the standard index notation from game theory, that is:

- given a vector $\boldsymbol{y} \in \prod_{i=1}^{N} \mathcal{Y}_i$, for some spaces $(\mathcal{Y}_i)_{i \leq N}$, and given $n \leq N$, we denote $y_{-n} = (y_i)_{i \leq N, i \neq n}$.
- given a family of spaces $(\Pi^y)_{y \in \mathcal{Y}}$ parametrized by $y \in \mathcal{Y}$, and given a vector $\boldsymbol{y} = (y_i)_{i \leq N} \in \mathcal{Y}^N$, we introduce the the notation $\Pi^{\boldsymbol{y}} = \prod_{i=1}^{N} \Pi^{y_i}$.

Let us now describe the core structure of the game. First of all, each player n makes a choice $x_n \in \mathcal{X}$ (buying or not the product, publicly supporting or not the opinion). This simple choice guarantees each player n to receive the associated intrinsic utility u_{n,x_n}. Then, players socially interact with each others, e.g. on a social network, and this way observe the choices made by other people. From these observations, player n perceives a *social reward* given by $U(k_n, x_n, \frac{1}{N-1} \sum \delta_{k_i, x_i})$ (where $\mathrm{u} : \mathcal{K} \times \mathcal{X} \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ is assumed to be Lipschitz in its $\mathcal{P}(\mathcal{X})$ coordinate), meaning that his social reward depends upon his class, choice, and the distribution of class and choices in the population. Therefore, player n's overall utility for this game is defined, for all $(d_n)_{n \leq N} \in (\mathcal{K} \times \mathbb{R}^2)^N$ and

$(x_n)_{n\leq N} \in \mathcal{X}^N$, by

$$R(d_n, x_n, \boldsymbol{d}_{-n}, x_{-n}) = u_{n,x_n} + \mathrm{u}(k_n, x_n, \frac{1}{N}\sum_{i=1}^{N} \delta_{k_i,x_i}).$$

4.3 Static game under full information

4.3.1 The game

Let us start with the static game. In the static framework, each player n's strategy is a choice $x_n \in \mathcal{X}$, and the gain function V of player n coincides with his reward function R defined in previous section, i.e. we define $V \equiv R$ in this game. More precisely, a strategy profile is here given by a family $(x_n)_{n\leq N} \in \mathcal{X}^N$. Given data $\boldsymbol{d} = (d_n)_{n\leq N} \in (\mathcal{K} \times \mathbb{R}^2)^N$ and a strategy profile $(x_n)_{n\leq N}$, the gain of player n is defined by

$$V(d_n, x_n, \boldsymbol{d}_{-n}, x_{-n}) = u_{n,x_n} + \mathrm{u}\Big(k_n, x_n, \frac{1}{N}\sum_{i=1}^{N} \delta_{k_i,x_i}\Big)$$

The solution concept that we are interested in is the notion of Iterated Elimination of Strictly Dominated Strategies (IESDS). Let us start by describing the underlying principle of IESDS. The IESDS consists in defining, for each player n, a sequence $(\Pi_k^{d_n,\boldsymbol{d}})_{k\in\mathbb{N}} \in \mathcal{X}^{\mathbb{N}}$, starting with $\Pi_0^{d_n,\boldsymbol{d}} = \mathcal{X}$, and decreasing for the inclusion.

The interpretation of $\Pi_k^{d_n,\boldsymbol{d}}$ is as follows. To find his rational strategy, each player must think about what the other players could do. The idea is that each player i tries to remove as many irrational strategies as possible for other players, hoping that he may himself find a best strategy for him in response to the remaining set of strategies. The game thus forces each player to put himself in the shoes of all the other players and think about their interests.

Each player i wonders, for every player $n \leq N$: "What strategy would be irrational for player n to play?". *Because he knows that everyone is rational*, player i considers that each player n will not play such strategies, and denotes $\Pi_1^{d_n,\boldsymbol{d}}$ the remaining strategies for each player n. *Because he knows that everyone is as intelligent as him*, player i then asks himself: "Given that everyone knows that each player j will play in $\Pi_1^{d_j,\boldsymbol{d}}$, what strategy would be irrational for player n to play?".

As we shall see, this question will make new strategies appear as irrational for each player n. Again, player i then assumes that player n, being *rational*, will not play them, and he denotes $\Pi_2^{d_n,\boldsymbol{d}}$ the remaining strategies for each player n. This iterative process can then clearly be repeated over and over.

Because this is an elimination process, clearly, it must be stationary. Thus, after asking themselves these questions enough times, each player i will have computed a set $\Pi_\infty^{d_n,\boldsymbol{d}}$ for every player $n \leq N$, representing the strategies that could not be eliminated for player n, because he did not find a reason to consider them as irrational.

Notice that player i asked these questions for all player $n \leq N$, including himself. Therefore, $\Pi_\infty^{d_i,\boldsymbol{d}}$ corresponds to the set of strategies that he has not eliminated for himself.

The result of this process is that we consider that each player i will necessarily play a strategy in $\Pi_\infty^{d_i,\boldsymbol{d}}$, simply because the other strategies have, at some point in the process, be proved to be irrational.

Given the above description, the only thing that remains to be specified is: what do we call an "irrational strategy"? And why does it depend upon what we assume that the other players could do?

Perhaps the most convincing criterion to consider a strategy as irrational is that it is *dominated*. A strategy is said to be dominated if there exists another strategy performing strictly better than it regardless what the other players do. More generally, a strategy x_i is *dominated given that other players n play in* $\Pi_k^{d_n,\boldsymbol{d}}$ if there exists a strategy x_i' bringing to player i a gain strictly better than x_i regardless what the other players n do in $\Pi_k^{d_n,\boldsymbol{d}}$. Therefore, assuming that it was established before that each player n would play in $\Pi_k^{d_n,\boldsymbol{d}}$, player i has no reason to play x_i, as he would then strictly increase his gain with x_i'.

$\Pi_k^{d_n,\boldsymbol{d}}$ is the set of remaining strategies for a player with data d_n observing the whole data $\boldsymbol{d}$ after k iterations of dominated strategy eliminations.

Let us now translate the concept of IESDS in a rigorous mathematical definition: For all $n \leq N$, we define

$$\begin{aligned}
\Pi_0^{d_n,\boldsymbol{d}} &= \mathcal{X} \\
\Pi_{k+1}^{d_n,\boldsymbol{d}} &= \{x_n \in \Pi_k^{d_n,\boldsymbol{d}} : \nexists x \in \mathcal{X} \\
&\qquad \text{s.t. } V(d_n, x, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}) > V(d_n, x_n, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}), \forall \boldsymbol{x}_{-n} \in \Pi_k^{\boldsymbol{d}_{-n},\boldsymbol{d}}\}, \quad \forall k \in \mathbb{N}
\end{aligned}$$

Let us describe this elimination process:

1. Initially, player n assumes that for all $i \neq n$, player i, with data d_i, could choose any action. Player n thus defines $\Pi_0^{d_i,\boldsymbol{d}} = \mathcal{X}$, for all $i \leq N$.

2. Player n assumes that any player i with a strictly dominated strategy would never play it. As he knows the population's data $(d_i)_{i \leq N}$, he knows who has a strictly

dominated strategy. For these players i, player n eliminates their dominated strategy from their set of actions. In other words, from each strategy set $\Pi_0^{d_i}$, $i \leq N$, player n obtain new strategy sets $\Pi_1^{d_i}$, $i \leq N$, such that $\Pi_1^{d_i}$ only keeps the actions x_i that are not dominated, i.e. the actions x_i s.t.

$$\nexists x \in \mathcal{X} \text{ s.t. } V(d_i, x, \boldsymbol{d}_{-i}, \boldsymbol{x}_{-i}) > V(d_i, x_i, \boldsymbol{d}_{-i}, \boldsymbol{x}_{-i}), \forall \boldsymbol{x}_{-i} \in \Pi_0^{\boldsymbol{d}_{-n}, \boldsymbol{d}}.$$

3. The elimination step is then repeated over and over.

In this first game that we study, each player knows the data d_n of each player n and is thus not only able to compute his own reduction of strategies but also the reduction of strategies of the other players. As $(\Pi_k^{d_n, \boldsymbol{d}})_{k \in \mathbb{N}}$ is, for all $n \leq N$, a decreasing sequence of sets, it converges to a set $\Pi_\infty^{d_n, \boldsymbol{d}}$ that cannot be reduced more with this iterative procedure. The interpretation of $\Pi_\infty^{d_n, \boldsymbol{d}}$ is that any rational strategy of player n should thus belong to $\Pi_\infty^{d_n, \boldsymbol{d}}$, or, equivalently, any strategy outside of $\Pi_\infty^{d_n, \boldsymbol{d}}$ would be irrational for player n.

For a given $n \leq N$, $\Pi_\infty^{d_n, \boldsymbol{d}}$ might be reduced to a single remaining possible strategy, which is thus the strategy that they shall rationally play, but $\Pi_\infty^{d_n, \boldsymbol{d}}$ might also still contain more than one strategy, and in this case the above iterative procedure is not enough to reduce the set of strategies to a single one for this player. In this case, such player should use an additional criterion of reduction to keep reducing the possibilities, but we shall see in this study that modeling this additional criterion will not be necessary, as the above iterated elimination of dominated strategies will be enough to predict the behavior of the population with high precision.

Our goal is to study the properties of the set of remaining strategies $(\Pi_\infty^{d_n, \boldsymbol{d}})_{n \leq N}$ for each player in the population, to understand and predict as well as possible how the game will unroll.

We will use mean-field methods by introducing a family of K differentiable distribution functions $(F_k)_{k \in \mathcal{K}}$ and $\varepsilon > 0$ such that

$$\| \frac{1}{N_k} \sum_{n=1}^{N_k} \mathbf{1}_{u_n < v} - F_k(v) \| \leq \varepsilon, \quad \forall k \in \mathcal{K}$$

where $u_n := u_{n,1} - u_{n,0}$ is player n *differential intrinsic utility* for all $n \leq N$. Such family of distribution functions can essentially be obtained via statistical fitting methods and can be chosen in any class of distribution functions, and ε simply represents the associated fitting error.

To the distribution functions F_k we associate a random variable (K, U) such that $\mathbb{P}(U \leq v \mid K = k) = F_k(v)$, such that we have by definition

$$\|\frac{1}{N_k}\sum_{n=1}^{N_k} \mathbf{1}_{u_n<v} - \mathbb{P}(U \leq v \mid K = k)\| \leq \varepsilon, \quad \forall k \in \mathcal{K}$$

In other words, the distribution function of U uniformly approximate the empirical distributions of $(u_n)_{n\leq N}$ conditionally to each class. Another way to see it is to say that given a uniform random variable $V \sim \mathcal{U}(\{1, ..., N\})$, the random variable (K, U), is, in distribution, uniformly close to the random variable (k_V, u_V) selecting a player at random in the population. The upside of (K, U) is that its distribution is regular, which will allow us to use analytical tools coming from differentiation, contracting properties, etc, while keeping track of the approximation error ε.

We finally assume that the operator

$$p \mapsto \mathcal{L}(\underset{x\in\mathcal{X}}{\text{argmax}}\,(U_{1,x} + \mathrm{u}(k_1, x, p)),$$

is contracting. This assumption simply means (assuming that ε is small, i.e. that the mean-field approximation of the population is good) that when the distribution of class-choices pairs in the population changes, it is not able to change the best response of more people than the number of people who changed their choices. In other words, we assume that social influence affects people's utility in a small enough way that a small group of people is not able to trigger a huge change in the population (we thus exclude uncontrollable herd behavior effects). Notice that it does not mean that there is no social influence, but simply that it is not too strong.

The following result shows that if the population's data is the result of a stochastic genetic mechanism, then the data should satisfy the propagation of chaos, and one should thus, with high probability, be able to fit them with regular distribution functions $(F_k)_{k\in\mathcal{K}}$ with $\varepsilon = \mathcal{O}(\frac{1}{\sqrt{N}})$.

Theorem 4.3.1 (Fitting error and propagation of chaos) *Let us endow* $(\Omega, \mathcal{F}) = ((\mathcal{K} \times \mathbb{R}^2)^N, \mathcal{B}((\mathcal{K} \times \mathbb{R}^2)^N))$ *with the probability distribution* $\mathbb{P}$ *of a family of* N *i.i.d. random variables with common distribution* $\mathcal{L}(K, U)$*, i.e.* $\mathbb{P} = \mathcal{L}(K, U)^{\otimes N}$*. Then, for any* $\alpha > 0$*, there exists an event* $E_\alpha \in \mathcal{F}$ *and a constant* $C_\alpha \in /R_+$*, such that* $\mathbb{P}(E_\alpha) \geq 1-\alpha$*, and such that* $\forall (k_n, u_n)_{n\leq N} \in E_\alpha$*, we have*

$$\|\frac{1}{N_k}\sum_{n=1}^{N_k} \mathbf{1}_{u_{k,n}<v} - F_k(v)\| \leq \frac{C_\alpha}{\sqrt{N}}, \quad \forall k \in \mathcal{K}$$

Consequently, if the population's data $(k_n, u_n)_{n\leq N}$ was generated with N i.i.d. random variables with common distribution $\mathcal{L}(K,U)$ (e.g. via a stochastic genetic algorithm), then, with probability greater than α, the functions $(F_k)_{k\in\mathcal{K}}$ will fit the population's data with error at most $\frac{C_\alpha}{\sqrt{N}}$, and thus Theorem holds true with $\varepsilon = \frac{C_\alpha}{\sqrt{N}}$.

Proof. This directly follows from Dvoretzky–Kiefer–Wolfowitz's inequality. □

4.3.2 Game's analysis

We shall formalize our analysis by 1) introducing predictions (and motivate them), and 2) estimating their accuracy.

Heuristically guess a prediction by mean-field approximation: We derive a prediction for the game's unrolling under the assumption that each player n plays a strategy in $\Pi_\infty^{d_n,\boldsymbol{d}}$, with an heuristic. Let us assume that a prediction is "possible", that is, that there actually exists a vector of choices $(x_n^\star)_{n\leq N}$ such that $\Pi_\infty^{d_n,\boldsymbol{d}} \simeq \{x_n^\star\}$ for $n \leq N$. This implies that each player was able to restrict his possible strategies and other players' strategies such that each player n has a unique remaining possible choice $x_n^\star$ left. By definition of the iterated elimination process, and as $\Pi_\infty^{d_n,\boldsymbol{d}}$ is its limit, there should not exist $x \in \mathcal{X}$ such that $V(d_n, x, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}) > V(d_n, \tilde{x}_n, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}), \forall \boldsymbol{x}_{-n} \in \Pi_\infty^{\boldsymbol{d}_{-n}}$. However, as we assumed that $\Pi_\infty^{d_i} \simeq \{x_i^\star\}$, the part " $\forall \boldsymbol{x}_{-n} \in \Pi_\infty^{\boldsymbol{d}_{-n}}$ " essentially means " $\forall \boldsymbol{x}_{-n}$: $x_{-n} = x_{-n}^\star$ ", and thus we have $\nexists x \in \mathcal{X} : V(d_n, x, \boldsymbol{d}_{-n}, x_{-n}^\star) > V(d_n, x_n^\star, \boldsymbol{d}_{-n}, x_{-n}^\star)$. This means that the choices $x_n^\star$ should approximately satisfy

$$x_n^\star = \underset{\mathcal{X}}{\operatorname{argmax}} \left(u_{n,x} + \mathrm{u}(k_n, x, \frac{1}{N}\sum_{i=1}^{N} \delta_{k_i, x_i^\star})\right)$$

and thus

$$\frac{1}{N}\sum_n \delta_{k_n, x_n^\star} \quad = \quad \frac{1}{N}\sum_n \delta_{k_n, \underset{\mathcal{X}}{\operatorname{argmax}} \left(u_{n,x} + \mathrm{u}(k_n, x, \frac{1}{N}\sum_{i=1}^{N} \delta_{k_i, x_i^\star})\right)}$$

Which suggests that the distribution of class-choices $\frac{1}{N}\sum_n x_n^\star$ should approximately be the unique fixed point $p^\star$ of the contracting operator

$$p \mapsto \mathcal{L}(\underset{x\in\mathcal{X}}{\operatorname{argmax}} \left(U_{1,x} + \mathrm{u}(k_1, x, p)\right),$$

and that $x_n^\star$ should thus approximately satisfy

$$x_n^\star = \underset{\mathcal{X}}{\operatorname{argmax}} \left(u_{n,x} + \mathrm{u}(k_n, x, p^\star)\right)$$

Of course, this heuristic derivation is by essence not rigorous, but it has the advantage to make very quick the formal derivation of a prediction. Let us now make this prediction a rigorous mathematical object, so that we can then study our games with the goal to estimate how much such prediction is validated in each game.

Definition 4.3.1 (Predictions) *We define the following predictions:*

- **Predicted class-wise distribution of choices:** *The Predicted class-wise choices distribution $p^\star$ is defined as the unique fixed point of the contracting operator*

$$p \mapsto \mathcal{L}(\underset{x\in\mathcal{X}}{\operatorname{argmax}} \left(U_{1,x} + \mathrm{u}(k_1, x, p)\right)$$

- **Predicted choices:** *The predicted choice of player n is defined by*

$$x_n^\star = \underset{x}{\operatorname{argmax}} \left(u_{n,x} + \mathrm{u}(k_1, x, p^\star)\right)$$

- **Predicted rewards:** *The predicted reward of player n is defined by*

$$V_n^\star = u_{n,x_n^\star} + \mathrm{u}(k_1, x, p^\star))$$

- **Predicted differential reward:** *The predicted differential reward of player n is defined by*

$$\begin{aligned} \Delta V_n^\star &= \left[u_{n,x_n^\star} + \mathrm{u}(k_1, x_n^\star, p^\star))\right] \\ &\quad - \left[u_{n,1-x_n^\star} + \mathrm{u}(k_1, 1 - x_n^\star, p^\star))\right] \end{aligned}$$

We start by establishing the quality of this prediction in the current basic game in the next result.

Theorem 4.3.2 *For all $x \in \Pi_\infty^{d,d}$, we have the following properties:*

- **Distribution of choices prediction accuracy:**

$$\frac{1}{N}\sum \delta_{k_n,x_n} = p^\star + \mathcal{O}(\varepsilon)$$

i.e. $p^\star$ is an approximation with error $\mathcal{O}(\varepsilon)$ of the distribution, class by class, of the choices of a rational population in this game.

- **Choices prediction accuracy:**

$$x_n \quad = \quad x_n^\star \quad \forall n : \Delta V_n^\star > \mathcal{O}(\varepsilon)$$

i.e. $x^\star$ predicts correctly the rational choice of all players except the ones such that $\Delta V_n^\star = \mathcal{O}(\varepsilon)$. Furthermore, these wrongly predicted choices represent a proportion at most $\mathcal{O}(\varepsilon)$ of the population.

- **Reward prediction accuracy:** *For all n, including for the players with wrongly predicted choices, we have*

$$V(k_n, u_n, x_n, k_{-n}, x_{-n}) = V_n^\star + \mathcal{O}(\varepsilon)$$

i.e. $V_n^\star$ predicts the reward of each player up to an error $\mathcal{O}(\varepsilon)$.

The above result essentially means that the above predictions are able to predict most of the aspects of the game's unrolling from the computation of one single object, that is, $p^\star$, simply defined as the unique fixed point of the operator

$$p \mapsto \mathcal{L}(\underset{x \in \mathcal{X}}{\text{argmax}} \, (U_{1,x} + \mathrm{u}(k_1, x, p)),$$

which, depending upon the choice of the distributions F_k used to approximate the empirical distribution of the players data, can be an infinitely differentiable or an analytic function, for which the unique fixed point can be analytically estimated with a k-th order extension for small $\|\mathrm{u}\|$ using the implicit function theorem. Alternatively, one or two steps of Newton's method can yield a very precise analytical estimation of such fixed point. More generally, such analytical and regular object allows us to use many estimation techniques that could not directly apply to the irregular empirical distribution of (u). Another upside of using an analytical approximation is that it allows to express results in terms of a few meaningful parameters like the mean and variance of the approximate distribution, assuming that it is selected in a class of distribution that is parametrized by their mean and variance (e.g. gaussian distributions), which is interesting for deriving meaningful interpretations. The above results provides an estimation of the error that we have to agree to make in counterpart. Notice that the better the fit with the approximate distribution function F_k is (i.e. the smaller ε is), the better will be the predictions associated to this approximation.

4.4 Repeated game with no initial information

4.4.1 The game

The game we have studied up to now is a static game, where each player fully knows the data of each other player in the population. Although this is an interesting framework,

and although one could easily generalize it to a more realistic one where players only know an approximation of the class-utility distribution in the population, perhaps a more interesting framework is to assume that players do not know anything about each other.

If we do not provide any knowledge to the players in the static game, a simple study shows that, except for the players with a dominant choice from the start, player will not be able to eliminate any choice.

However, if we make the game repeats itself several times, and if we assume that players can observe the class-choices distribution from past games, there is hope that via these observations, players will be able to "learn" information about each other, and that the game will be precisely predictable after a few stage games.

Notice that observing the class-choices distributions does not require to observe people's intrinsic utilities, which is a realistic feature in many situations.

In the repeated game framework, a strategy for player n is represented as a sequence of function $\boldsymbol{x}_n := (\boldsymbol{x}_{n,t})_{t\leq T}$, where $T \in \mathbb{N}$ is the total number of stage games, such that $\boldsymbol{x}_{n,t} : ((\mathcal{K} \times \mathcal{X})^N)^t \to \{0,1\}$ what choice player n will make at the t-th game given the distributions of (class, choice) in the population in past games. The choices associated to data $d := (d_n)_{n\leq N}$ and strategy profile $\boldsymbol{x} := (\boldsymbol{x}_n)_{n\leq N}$ are described by the choice processes $(X_t^{n,d,\boldsymbol{x}})_{t\in\mathbb{N}}$ defined by induction by

$$\begin{aligned} X_0^{n,d,\boldsymbol{x}} &= x_{n,0} \\ X_{t+1}^{n,d,\boldsymbol{x}} &= x_{n,t+1}\Big(k_i, X_s^{i,d,\boldsymbol{x}}, i \leq N, s \leq t\Big) \end{aligned}$$

This is simply a repeated version of the game from previous chapter, such that each action at time t can depend upon the class and choices made in past games in the population. The reward of player n is

$$V(d_n, \boldsymbol{x}_n, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}) = \sum_{t\leq T} \beta^t \Big(u_{n,X_t^{n,d,\boldsymbol{x}}} + \mathrm{u}(k_n, X_t^{n,d,\boldsymbol{x}}, \frac{1}{N}\sum_{i=1}^{N} \delta_{k_i, X_t^{i,d,\boldsymbol{x}}}) \Big)$$

We introduce the following notation. For $x \in \Pi$ and $t \in \mathbb{N}$, we denote $\mathcal{X}_t^x := \{x' \in \Pi : x'_s \equiv x_s, \forall s \leq t\}$. This is simply the set of all strategy coinciding with x before time t. Intuitively, the meaning of this notation is that from time t, a player with strategy x can always "change" his strategy as long as he selects a new strategy in $\mathcal{X}_t^x$, because the games before time t have already happened and thus it is too late to change his strategy for these games, but games happening after time t have not happened yet and thus the individual can, at time t, readjust his strategy for these future games.

Although we said that players would not know anything about each other, we shall assume that they have a very minimal belief about the population. To describe this belief, let us introduce the following data set: for $\varepsilon \geq 0$, we consider

$$\mathcal{D}_\varepsilon := \{d \in \mathcal{D}^N : \exists (F_k)_{k\in\mathcal{K}} : F_k \text{ is 1 Lipschitz and } \| \frac{1}{N_k} \sum_{n=1}^{N_k} \mathbf{1}_{u_{k,n}<v} - F_k(v) \| \leq \varepsilon, \quad \forall k \in \mathcal{K}\}$$

This is simply the set of population data d such that there exists regular functions fitting the population data with error smaller than ε.

Essentially, we will assume that players know that $d \in \mathcal{D}_\varepsilon$, that is, they know (or assume) that the population data is regular enough to be well fitted by some Lipschitz distribution functions. We stress that players are not assumed to know such $(F_k)_{k\in\mathcal{K}}$, but only that it exists. This is, as we shall see in our probabilistic study, a belief justified by mean-field theory and propagation of chaos. To summarize, the actual population data d is, as we assumed earlier, in $\mathcal{D}_\varepsilon$, and players don't know d anymore, but simply that $d \in \mathcal{D}_\varepsilon$.

Let us know provide our solution concept for the repeated game with no initial information:

$$\begin{aligned}
\Pi_{k+1}^{d_n} \quad = \quad & \{x_n \in \Pi_k^{d_n} : \forall \boldsymbol{d}_{-n}, \forall \boldsymbol{x}_{-n} \in \Pi_k^{\boldsymbol{d}_{-n}}, \forall t \in \mathbb{N}, \\
& \nexists x'_n \in \mathcal{X}_{t-1}^{x_n} : V(d_n, x'_n, d'_{-n}, x'_{-n}) > V(d_n, \mathcal{X}_t^{x_n}, d'_{-n}, x'_{-n}), \\
& \forall d'_{-n} \in \mathcal{D}_\varepsilon, \forall x'_{-n} \in \Pi_k^{d'_{-n}} : \sum_{i=1}^N \delta_{d'_i, X_s^{i,x'}} = \sum_{i=1}^N \delta_{d_i, X_s^{i,x}}, s \leq t\}
\end{aligned}$$

The interpretation of this solution concept is as follows. If we compare it to the solution concept of the static game, the main differences are the *time* and *information* components. The general idea is that at each time $t \leq T$, there should not be a strategy *from time* t, i.e. a strategy $x'_n \in \mathcal{X}_t^{x_n}$, that is strictly better than x_n for player n, over all strategies of other players *that are consistent with player n's information*, i.e. such that $d'_{-n} \in \mathcal{D}_\varepsilon$, and such that

$$\frac{1}{N} \sum_{i=1}^N \delta_{d'_i, X_s^{i,x'}} = \frac{1}{N} \sum_{i=1}^N \delta_{d_i, X_s^{i,x}}, \quad \forall s \leq t$$

The underlying idea is thus that each player does not simply eliminate strategies that are strictly dominated from time 0, but also the strategies that are strictly dominated from any time t, with the interpretation that they would never play such strategy after time t since another strategy, that they can still adopt, is strictly better regardless the strategies of the other players.

4.4.2 Game's analysis

Before stating the result in the repeated game framework, let us, as for the static game, heuristically derive how should the game rationally unroll.

The idea is that if we in mind the idea that the game's repetition is supposed to allow players to learn about each other via their observations of past games, we can assume that after many stages, players should have a behavior that is close to a fully informed population's behavior. The class-wise choice distribution should thus eventually get close to $p^\star$ after enough stage games. If this is really the case, we should be able to predict that $\mathcal{W}(\frac{1}{N}\sum_{n=1}^{N}\delta_{k_n,X_t^{n,d,\boldsymbol{x}}}, p^\star) \leq \varepsilon_t$, where $(\varepsilon_t)_{t\in\mathbb{N}}$ is a decreasing sequence. The best way to then obtain the right sequence $(\varepsilon_t)_{t\in\mathbb{N}}$ is to try to prove that this prediction holds true and finally see the conditions required on $(\varepsilon_t)_{t\in\mathbb{N}}$ to make the proof work. We obtain the sequence characterized by

$$\begin{aligned} \varepsilon_0 &= 1 \\ \varepsilon_{t+1} &= 2\frac{c\varepsilon_t + \varepsilon}{1-\beta}, \quad t \in \mathbb{N} \end{aligned}$$

Notice that by explicitly writing ε_t, we can show that

$$\varepsilon_t = \left(\frac{2c}{1-\beta}\right)^t + \mathcal{O}(\varepsilon)$$

We now state our main result.

Theorem 4.4.1 *For all $x \in \Pi_\infty(\omega)$, we have the following properties:*

- **Distribution of choices prediction accuracy:**

$$\frac{1}{N}\sum_{n=1}^{N}\delta_{k_n,X_t^{n,d,\boldsymbol{x}}} = p^\star + \mathcal{O}(\varepsilon_t)$$

  i.e. $p^\star$ is an approximation with error $\mathcal{O}(\varepsilon)$ of the distribution, class by class, of the choices of a rational population in this game.

- **Choices prediction accuracy:**

$$X_t^{n,d,\boldsymbol{x}} = x_n^\star \quad \forall n : \Delta V_n^\star > \mathcal{O}(\varepsilon_t)$$

  i.e. $x^\star$ predicts correctly the rational choice of all players except the ones such that $\Delta V_n^\star = \mathcal{O}(\varepsilon_t)$. Furthermore, these wrongly predicted choices represent a proportion at most $\mathcal{O}(\varepsilon_t)$ of the population.

- **Reward prediction accuracy:** *For all n, including for the players with wrongly predicted choices, we have*

$$R(k_n, u_n, X_t^{n,d,\boldsymbol{x}}, k_{-n}, X_t^{-n,d,\boldsymbol{x}}) = V_n^\star + \mathcal{O}(\varepsilon_t)$$

i.e. $V_n^\star$ predicts the t-th reward of each player up to an error $\mathcal{O}(\varepsilon_t)$.

4.5 Analysis of the class-wise choice distribution $p^\star$

4.5.1 Iterative methods

Notice that the computation of the fixed point $p^\star$ is a straightforward numerical task as soon as

$$\phi(p) = \mathcal{L}(\underset{x}{\text{argmax}}\,(U_x + \mathrm{u}(k,x,p)), k) \sim (\mathbb{P}(U_k < \mathrm{u}(k,p)))_{k\in\mathcal{K}} = (F_k(\mathrm{u}(k,p)))_{k\in\mathcal{K}}$$

is fast to compute, which is the case for many classes of distribution functions.

4.5.2 Parametric fitting functions and explicit fixed point

Let us assume that the fitting functions F_k have been chosen in a parametric class of functions, then:

$$\phi(p) = (F(\theta_k, \mathrm{u}(k,p)))_{k\in\mathcal{K}}$$

If the class is parametrized by means and variances, we have

$$\phi(p) = (F(\frac{\mathrm{u}(k,p) - \mu_k}{\sigma_k}))_{k\in\mathcal{K}}$$

Many parametric class of distribution functions can be used in this context: gaussians, Logistics, etc. One can choose any of them, the prediction will hold true, but the error ε will be larger if the fitting is not good. Nonetheless, if we choose to use the class of uniform distributions, assuming that $\frac{\mathrm{u}(k,p)-\mu_k}{\sigma_k} \in [-1,1]$ for all $p \in [0,1]^K$, we have

$$\phi(p) = \frac{1}{2} + \frac{1}{2}(\frac{\mathrm{u}(k,p) - \mu_k}{\sigma_k})_{k\in\mathcal{K}}$$

Let us now assume that the social reward function u is affine in p.

$$\phi(p) = \frac{1}{2} + \frac{1}{2}(\frac{A_k p + b_k - \mu_k}{\sigma_k})_{k\in\mathcal{K}} = (\frac{A_k}{2\sigma_k})_{k\in\mathcal{K}} p + \frac{1}{2}(\frac{1 + b_k - \mu_k}{\sigma_k})_{k\in\mathcal{K}}$$

And thus the limit is

$$p^\star = (1 - (\frac{A_k}{2\sigma_k})_{k\in\mathcal{K}})^{-1}(\frac{1 + b_k - \mu_k}{2\sigma_k})_{k\in\mathcal{K}}$$

if it goes above $u(f) + \varepsilon$ or under $u(f) - \varepsilon$

4.5.3 Small social influence

In situations where social influence is very small, i.e. when $\mathrm{u} = c\tilde{\mathrm{u}}$ with c small, we can also make interesting analysis. In this case, the operator writes

$$\phi(p) = (F_k(c\tilde{\mathrm{u}}(k,p)))_{k\in\mathcal{K}}$$

Let us make explicit the dependence of its fixed point in c by writing $p^\star = p^\star(c)$. We have

$$p^\star(c) = \phi(p^\star(c)) = (F_k(c\tilde{\mathrm{u}}(k,p^\star(c))))_{k\in\mathcal{K}}$$

A standard argument of implicit function theorem implies that $p^\star$ is derivable for c small enough, and we have

$$(p^\star)'(c) = (\partial_x F(\theta_k, c\tilde{\mathrm{u}}(k,p^\star(c)))\tilde{\mathrm{u}}(k,p^\star(c)))_{k\in\mathcal{K}}$$

which, in $c = 0$, yields

$$(p^\star)'(0) = (\partial_x F_k(0)\mathrm{u}(k,p^\star(0)))_{k\in\mathcal{K}}$$

We thus have

$$(p^\star)'(0) = (f_k(0)\mathrm{u}(k,p^\star(0)))_{k\in\mathcal{K}}$$

And we can finally write the first order expansion of $p^\star(c)$:

$$p^\star(c) = p^\star(0) + (f_k(0)c\mathrm{u}(k,p^\star(0)))_{k\in\mathcal{K}} + \mathcal{O}(c^2)$$

which means that the effect of social influence on class k's choices is $f_k(0)\mathrm{u}(k,p(0))$, that is, the density of indecisive people from class k (i.e. $f_k(0)$) multiplied by the influence exercised on class k by the intrinsic class-wise choice distributions (i.e. $c\mathrm{u}(k,p(0))$). Let us consider two simple examples.

Snowball effect

First, we address the social phenomenon of amplification. This happens when social influence increases the largest proportions and decreases the smallest ones. We conjecture that this is a consequence of feeling happiness to share the same choice with someone else. To illustrate this, let us consider a single class, i.e. $K = 1$ and $\mathcal{K} = \{1\}$. Happiness to share the same choice would naturally lead to a social reward function $c\mathrm{u}(k,p) = c(p - \frac{1}{2})$. This thus leads to

$$p(c) = p(0) + f(0)c(p(0) - \frac{1}{2}) + \mathcal{O}(c^2)$$

and thus

$$p(c) - \frac{1}{2} = (p(0) - \frac{1}{2})(1 + f(0)c) + \mathcal{O}(c^2)$$

Class repulsion effect

In the second example, we address the social phenomenon of class repulsion. This happens when social influence prevents two opposite classes to mix together, i.e. to both concentrate on a same choice. We conjecture that this is a consequence of feeling unhappiness to share the same choice with someone in the other class. To illustrate it, we consider two classes, i.e. $K = 2$ and $\mathcal{K} = \{-1, 1\}$. Unhappiness to share the same choice as people in the opposite class would naturally lead to the social reward function $\mathrm{cu}(k, p) = -cp_{-k}$. This thus leads to

$$p(c) = p(0) + (-f_k(0)cp(0)_{-k})_{k\in\{-1,1\}} + \mathcal{O}(c^2)$$

and thus

$$p(c) = p(0) - c(f_k(0)p(0)_{-k})_{k\in\{-1,1\}} + \mathcal{O}(c^2)$$

i.e.

$$p(c)_k = p(0)_k - cf_k(0)p(0)_{-k} + \mathcal{O}(c^2)$$

The result is that

$$P(c) = P(0) - c(P_1 f_1(0)p(0)_{-1} + P_{-1}f_{-1}(0)p(0)_1) + \mathcal{O}(c^2)$$

Notice that the intrinsic choice distribution is indeed decreased because of social class repulsion.

4.6 Proofs

4.6.1 Proof of Theorem 4.3.2

We introduce a concept useful to study $(\Pi_k^{d_n})_{k\in\mathbb{N}}$, that we call "strategy elimination sup-process".

Definition 4.6.1 (strategy elimination sup-process) *A family $(\tilde{\Pi}_k^d)_{d\in\mathcal{D},k\in\mathbb{N}}$ of strategy sets is a strategy elimination sup-process if*

- $\tilde{\Pi}_0^d = \Pi$ *for all* $d_n \in \mathcal{D}$,
- $(\tilde{\Pi}^d)_{k\in\mathbb{N}}$ *is decreasing for the inclusion, for all* $d_n \in \mathcal{D}$,
- *we have,* $\forall k \in \mathbb{N}$,

$$\tilde{\Pi}_{k+1}^{d_n} \supset \{x_n \in \tilde{\Pi}_k^{d_n} : \nexists x \in \mathcal{X} \setminus \{x_n\} \\ s.t.\ V(d_n, x, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}) > V(d_n, x_n, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}), \forall \boldsymbol{x}_{-n} \in \tilde{\Pi}_k^{d_{-n}}\}, \quad \forall d_n \in \mathcal{D}$$

The idea of a strategy elimination sup-process is that it is a strategy elimination sequence such that each iteration does not require to eliminate *all* the strategies that can be eliminated: it can in particular eliminate only a subset of these strategies. This is an interesting tool because the sequence $(\Pi_k^{d_n})_{k\in\mathbb{N}}$, requiring to eliminate all the strategies that can be eliminated can be complex to express, because some strategies might involve a difficult argument to prove their irrationality, and it might be easier to eliminate them later in the process.

We can use this tool to study a well chosen strategy elimination sup-process $(\tilde{\Pi}_k^{d_n})_{k\in\mathbb{N}}$ keeping a simple form as k grows. The following result explains how strategy elimination sup-processes can be used in practice for the game's analysis.

Lemma 4.6.1 *Let $(\tilde{\Pi}_k)_{k\in\mathbb{N}}$ be a strategy elimination sup-process. We have*

$$\Pi_k^{d_n} \subset \tilde{\Pi}_k^{d_n}, \quad \forall n \leq N, \forall k \in \mathbb{N}$$

and thus $\Pi_\infty^{d_n} \subset \tilde{\Pi}_\infty^{d_n}$.

In other words, a strategy elimination sup-process is a good tool to analyze the solution concept: if well designed, $\tilde{\Pi}_k$ will keep a simple and tractable form, for instance allowing us to simply study properties shared by all its elements. As we have $\Pi_k \subset \tilde{\Pi}_k$, any property satisfied by all the elements of $\tilde{\Pi}_k$ will be satisified by all the elements of Π_k.

Proof of the Lemma. We prove this property by induction. For $k = 0$, we have $\Pi_0^{d_n} = \Pi = \tilde{\Pi}_0^{d_n}$ by definition. Let us assume that the property holds true for some $k \in \mathbb{N}$ and let us prove it for $k+1$. We have

$$\begin{aligned}
\Pi_{k+1}^{d_n} &= \{x_n \in \Pi_k^{d_n} : \nexists x \in \mathcal{X} \setminus \{x_n\} \\
&\qquad \text{s.t. } V(d_n, x, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}) > V(d_n, x_n, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}), \forall \boldsymbol{x}_{-n} \in \Pi_k^{\boldsymbol{d}_{-n}}\}, \\
&\subset \{x_n \in \tilde{\Pi}_k^{d_n} : \nexists x \in \mathcal{X} \setminus \{x_n\} \\
&\qquad \text{s.t. } V(d_n, x, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}) > V(d_n, x_n, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}), \forall \boldsymbol{x}_{-n} \in \tilde{\Pi}_k^{d_{-n}}\} \\
&\subset \tilde{\Pi}_{k+1}^{d_n}
\end{aligned}$$

where the first inclusion comes from the induction hypothesis. This concludes the induction and the proof. □

Let us now prove Theorem 4.3.2. We define the following candidate for our strategy elimination sup-process:

$$\tilde{\Pi}_k^{d_n} = \begin{cases} \{x_n^\star\} \text{ if } : x_n^\star = \underset{x}{\operatorname{argmax}}\, V(d_n, x, \mu), \forall \mu \in \mathcal{B}(p^\star, \varepsilon_k) \\ \mathcal{X} \text{ else.} \end{cases}$$

This sequence simply claims that the first players to find their strategy $x_n^\star$ are the players for whom it is dominant over the largest sets of possible choices distributions of other players. More precisely, it says that k elimination's iterations are sufficient for the players with strategy $x_n^\star$ dominant over choice distributions in $\mathcal{B}(p^\star, \varepsilon_k)$ to isolate their strategy $x_n^\star$. This is consistent with the intuition that the first players to find their rational choice are the players with the most dominant choice.

Lemma 4.6.2 *The sequence $(\tilde{\Pi}_k)_{k\in\mathbb{N}}$ is a strategy elimination sup-process.*

Proof. It is clear that $\tilde{\Pi}_0^{d_n} = \Pi$, and that $(\tilde{\Pi}_k^{d_n})_{k\in\mathbb{N}}$ is a decreasing sequence for the inclusion, for all $d_n \in \mathcal{D}$. To show that $\tilde{\Pi}$ is a strategy elimination sup-process, it only remains to prove that

$$\begin{aligned}\tilde{\Pi}_{k+1}^{d_n} \quad \supset \quad & \{x_n' \in \tilde{\Pi}_k^{d_n} : \nexists x'' \in \mathcal{X} \setminus \{x_n'\} \\ & \text{s.t. } V(d_n, x'', \boldsymbol{d}_{-n}, x'_{-n}) > V(d_n, x_n', \boldsymbol{d}_{-n}, x'_{-n}), \forall x'_{-n} \in \tilde{\Pi}_k^{d_{-n}}\}\end{aligned}$$

If $\tilde{\Pi}_{k+1}^{d_n} = \tilde{\Pi}_k^{d_n}$, i.e. if there has not been any elimination between stage k and $k+1$ for player n, this is obviously true. The only other case is when $\tilde{\Pi}_k^{d_n} = \mathcal{X}$ and then $\tilde{\Pi}_{k+1}^{d_n} = \{x_n^\star\}$. In this case, the eliminated strategy is $1 - x_n^\star$. Let us show that it was indeed right to eliminate it, i.e. that

$$\exists x' \in \mathcal{X} \setminus \{1 - x_n^\star\} \text{ s.t. } V(d_n, x', \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}) > V(d_n, 1 - x_n^\star, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}), \forall \boldsymbol{x}_{-n} \in \tilde{\Pi}_k^{d_{-n}}$$

As there are only two strategies, it reduces to prove that

$$V(d_n, x_n^\star, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}) > V(d_n, 1 - x_n^\star, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}), \forall \boldsymbol{x}_{-n} \in \tilde{\Pi}_k^{d_{-n}}$$

We have

$$\begin{aligned} V(d_n, 1 - x_n^\star, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}) &< V(d_n, 1 - x_n^\star, p^\star) + c\varepsilon_k + \varepsilon \\ &< V(d_n, x_n^\star, p^\star) - \varepsilon_{k+1} + c\varepsilon_k + \varepsilon \\ &< V(d_n, x_n^\star, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}) - \varepsilon_{k+1} + 2(c\varepsilon_k + \varepsilon) \\ &< V(d_n, x_n^\star, \boldsymbol{d}_{-n}, \boldsymbol{x}_{-n}) \end{aligned}$$

This concludes the proof. □

We can now establish the precision of our prediction for this game.

Corollary 4.6.1 *We have, for all $x \in \Pi_\infty$:*

$$x_n^\star = \operatorname*{argmax}_x V(d_n, x, \mu), \forall \mu \in \mathcal{B}(p^\star, \varepsilon_\infty) \Rightarrow x_n = x^\star$$

and thus, Theorem 4.3.2 holds true.

Proof. This is a direct implication of Lemma 4.6.1 and Lemma 4.6.2. □

4.6.2 Proof of Theorem 4.4.1

We adapt our proof for the static framework under full information for this framework. We start by transposing the notion of strategy elimination sup-process.

Definition 4.6.2 *A consistent strategy profile reduction process is a decreasing sequence* $\tilde{\Pi}_k$ *such that:*

1. $\tilde{\Pi}_0 = \Pi$,
2. *We have*

$$\begin{aligned}\tilde{\Pi}_{k+1}^{d_n} &= \{x_n \in \tilde{\Pi}_k^{d_n} : \forall \boldsymbol{d}_{-n}, \forall \boldsymbol{x}_{-n} \in \tilde{\Pi}_k^{\boldsymbol{d}_{-n}}, \forall t \in \mathbb{N}, \\ &\quad \nexists x'_n \in \mathcal{X}_{t-1}^{x_n} : V(d_n, x'_n, d'_{-n}, x'_{-n}) > V(d_n, \mathcal{X}_t^{x_n}, d'_{-n}, x'_{-n}), \\ &\quad \forall d'_{-n} \in \mathcal{D}_\varepsilon, \forall x'_{-n} \in \tilde{\Pi}_k^{d'_{-n}} : \sum_{i=1}^N \delta_{d'_i, X_s^{i,x'}} = \sum_{i=1}^N \delta_{d_i, X_s^{i,x}}, s \le t\}\end{aligned}$$

We have the following result.

Lemma 4.6.3 *For any consistent strategy profile reduction process* $\tilde{\Pi}^{d_n}$*, we have* $\Pi_\infty^{d_n} \subset \tilde{\Pi}_\infty^{d_n}$.

Proof. Let us show by induction that for all k we have $\Pi_k^{d_n} \subset \tilde{\Pi}_k^{d_n}$. For $k = 0$, it is obvious. Let us assume that it holds true for some $k \in \mathbb{N}$, then we have

$$\begin{aligned}\Pi_{k+1}^{d_n} &= \{x_n \in \Pi_k^{d_n} : \forall \boldsymbol{d}_{-n}, \forall \boldsymbol{x}_{-n} \in \Pi_k^{\boldsymbol{d}_{-n}}, \forall t \in \mathbb{N}, \\ &\quad \nexists x'_n \in \mathcal{X}_{t-1}^{x_n} : V(d_n, x'_n, d'_{-n}, x'_{-n}) > V(d_n, \mathcal{X}_t^{x_n}, d'_{-n}, x'_{-n}), \\ &\quad \forall d'_{-n} \in \mathcal{D}_\varepsilon, \forall x'_{-n} \in \Pi_k^{d'_{-n}} : \sum_{i=1}^N \delta_{d'_i, X_s^{i,x'}} = \sum_{i=1}^N \delta_{d_i, X_s^{i,x}}, s \le t\} \\ &\subset \{x_n \in \tilde{\Pi}_k^{d_n} : \forall \boldsymbol{d}_{-n}, \forall \boldsymbol{x}_{-n} \in \tilde{\Pi}_k^{\boldsymbol{d}_{-n}}, \forall t \in \mathbb{N}, \\ &\quad \nexists x'_n \in \mathcal{X}_{t-1}^{x_n} : V(d_n, x'_n, d'_{-n}, x'_{-n}) > V(d_n, \mathcal{X}_t^{x_n}, d'_{-n}, x'_{-n}), \\ &\quad \forall d'_{-n} \in \mathcal{D}_\varepsilon, \forall x'_{-n} \in \tilde{\Pi}_k^{d'_{-n}} : \sum_{i=1}^N \delta_{d'_i, X_s^{i,x'}} = \sum_{i=1}^N \delta_{d_i, X_s^{i,x}}, s \le t\} \\ &\subset \tilde{\Pi}_{k+1}\end{aligned}$$

which concludes the induction. We conclude by taking the limit in k. □

We define the following sequence of strategy profiles.

$$\tilde{\Pi}_{k+1}^{d_n} = \begin{cases} \{\boldsymbol{x}_n \in \tilde{\Pi}_k^{d_n} : X_t^{n,d,\boldsymbol{x}} = x_n^\star \ \forall t \ge k \ \forall \boldsymbol{x}_{-n} \in \tilde{\Pi}_k^{d_{-n}}\} \text{ if } x_n^\star = \underset{x}{\operatorname{argmax}}\, R(d_n, x, B(p^\star, \varepsilon_k)) \\ \Pi \text{ else.} \end{cases}$$

Lemma 4.6.4 *$\tilde{\Pi}$ is a consistent strategy profile reduction process.*

Proof. We only have to prove that

$$\begin{aligned}\tilde{\Pi}_{k+1}^{d_n} \quad \supset \quad & \{x_n \in \tilde{\Pi}_k^{d_n} : \forall \boldsymbol{d}_{-n}, \forall \boldsymbol{x}_{-n} \in \tilde{\Pi}_k^{d_{-n}}, \forall t \leq T, \\ & \nexists x'_n \in \mathcal{X}_t^{x_n} : V_t(d_n, x'_n, d'_{-n}, x'_{-n}) > V_t(d_n, x_n, d'_{-n}, x'_{-n}), \\ & \forall d'_{-n}, \forall x'_{-n} \in \tilde{\Pi}_k^{d'_{-n}} : \frac{1}{N_n} \sum_{i \in I_n} \delta_{d'_i, X_s^{i,x'}} = \frac{1}{N_n} \sum_{i \in I_n} \delta_{d_i, X_s^{i,x}}, s \leq t\}\end{aligned}$$

We prove it as in the static game case. First, when $\tilde{\Pi}_{k+1}^{d_n} = \tilde{\Pi}_k^{d_n}$, this is obvious. The only remaining case is when $\tilde{\Pi}_k^{d_n} = \Pi$ and

$$\tilde{\Pi}_{k+1}^{d_n} = \{\boldsymbol{x}_n \in \Pi : X_t^{n,d,\boldsymbol{x}} = x_n^\star \; \forall t \geq k \; \forall \boldsymbol{x}_{-n} \in \tilde{\Pi}_k^{d_{-n}}\}$$

which happens only if $x_n^\star = \underset{x}{\operatorname{argmax}}\, R(d_n, x, B(p^\star, \varepsilon_k))$. The strategies in $\tilde{\Pi}_k^{d_n}$ which have been eliminated in $\tilde{\Pi}_{k+1}^{d_n}$ are, in this case, the strategies $\boldsymbol{x}_n$ such that there exists $\boldsymbol{x}_{-n} \in \tilde{\Pi}_k^{d_{-n}}$ such that $X_t^{n,d,\boldsymbol{x}} \neq x_n^\star$. Let us show that it was right to eliminate this strategy, i.e. that indeed, there exists $x'_n \in \mathcal{X}_t^{\boldsymbol{x}_n}$ such that

$$\begin{aligned}& V_t(d_n, x'_n, d'_{-n}, x'_{-n}) > V_t(d_n, \boldsymbol{x}_n, d'_{-n}, x'_{-n}), \\ & \forall d'_{-n}, \forall x'_{-n} \in \tilde{\Pi}_k^{d'_{-n}} : \frac{1}{N_n} \sum_{i \in I_n} \delta_{d'_i, X_s^{i,\boldsymbol{x}_n, x'_{-n}}} = \frac{1}{N_n} \sum_{i \in I_n} \delta_{d_i, X_s^{i,\boldsymbol{x}_n, x_{-n}}}, s \leq t\}\end{aligned}$$

Let us naturally consider the strategy $(\boldsymbol{x}'_{n,t})_t \in \mathcal{X}_t^{\boldsymbol{x}_n}$ such that $\boldsymbol{x}'_{n,s} \equiv x_n^\star$ for all $s \geq t$, that is, the strategy stationary at $x_n^\star$ after time t. We study each reward for $s \geq t$.

$$\begin{aligned}& R(d_n, X_s^{n,d_n,\boldsymbol{x}_n,d'_{-n},\boldsymbol{x}'_{-n}}, d'_{-n}, X_s^{-n,d_n,\boldsymbol{x}_n,d'_{-n},\boldsymbol{x}'_{-n}}) \\ \leq \; & R(d_n, X_s^{n,d_n,\boldsymbol{x}_n,d'_{-n},\boldsymbol{x}'_{-n}}, p^\star) + c\varepsilon_k + \varepsilon \\ \leq \; & R(d_n, x_n^\star, p^\star) + c\varepsilon_k + \varepsilon \\ \leq \; & R(d_n, x_n^\star, X_s^{n,d_n,\boldsymbol{x}'_n,d'_{-n},\boldsymbol{x}'_{-n}}) + 2(c\varepsilon_k + \varepsilon)\end{aligned}$$

However, at time t, as we have assumed that $X_t^{n,d_n,\boldsymbol{x}_n,d'_{-n},\boldsymbol{x}'_{-n}} \neq x_n^\star$, we have the more refined estimation

$$\begin{aligned}& R(d_n, X_t^{n,d_n,\boldsymbol{x}_n,d'_{-n},\boldsymbol{x}'_{-n}}, d'_{-n}, X_t^{-n,d_n,\boldsymbol{x}_n,d'_{-n},\boldsymbol{x}'_{-n}}) \\ \leq \; & R(d_n, x_n^\star, X_s^{n,d_n,\boldsymbol{x}'_n,d'_{-n},\boldsymbol{x}'_{-n}}) - \varepsilon_{k+1} + 2(c\varepsilon_k + \varepsilon)\end{aligned}$$

As we have, by definition, $\varepsilon_{k+1} = 2\frac{c\varepsilon_k + \varepsilon}{1-\beta}$, we thus have

$$V_t(d_n, \boldsymbol{x}'_n, d'_{-n}, x'_{-n}) > V_t(d_n, \boldsymbol{x}_n, d'_{-n}, x'_{-n})$$

□

Proof of Theorem 4.4.1. This directly follows from Lemmas 4.6.3 and 4.6.4. □

4.7 Conclusion

In this work, we have studied games with a large amount of players with the strongly rational Iterative Elimination of Strictly Dominated Strategies solution concept. We have been able to study this concept in a repeated game where players initially do not have any information about each other, and we have proved that, although the remaining strategy profiles resulting from this process are not necessarily reduced to a unique one, they all share common properties allowing us to make precise predictions about people's rational choices after a few stage games. We believe that the ability to make precise predictions in such a complex game, with a large population, no initial information, repeated games, and the IESDS rational concept, with techniques that are mathematically not too heavy thanks to mean-field tools, is appealing and promising. Although the IESDS with finitely or infinitely many players is not new, to the best of our knowledge, using mean-field tools for the study of IESDS has not been done before. From this work, several directions can be taken: adapt the problem to non-binary space of choices, adding other types of players (e.g. advertisers), etc. In these possible extensions, making precise predictions is, in itself, not fundamentally harder, but the main challenge in such extensions is to keep the notations and proofs not too heavy. We believe that interesting work can be done in the direction of searching for the most elegant and powerful way to deal with more complex problems. Adding features to the game is susceptible to lead to heavy notations and formulas in two places: 1) the definition of the solution concept, which already takes three lines in our work, and 2) the proof of the prediction. Adding complexity to the game would indeed make the formulation of the solution concept a lot heavier in notations or require to introduce intermediary objects. On the other hand, the 3-lines solution concept introduced in this work for the repeated game, with no initial information but with past observations, and with the IESDS principle, seems interestingly concise considering the amount of features encoded in it. Likewise, the concept of elimination sup-process that we introduced for the proof made the argument simple and concise enough. It could be interesting to take advantage of these two aspects to make the game more complex while maintaining its tractability.

Chapter 5

Gaussian Cumulative Prospect theory

Abstract. In this work, we propose a parametrization of Daniel Kahneman and Amos Tversky's Cumulative Prospect Theory (CPT) leading to an explicit gamble valuation formula for gaussian rewards. More precisely, we define parametric functions v_θ, w_θ^- and w_θ^+, with the three following properties. We refer to the first property as "validity": for all θ, v_θ, w_θ^- and w_θ^+ satisfy all the properties stated by CPT (v_θ concave on $\mathbb{R}_+$, convex on $\mathbb{R}_-$ with steeper curve, w_θ^- and w_θ^+ increasing, inverse S shapes, mapping 0 to 0 and 1 to 1). We refer to the second property as "Density": the parametrization is flexible enough so that the choice of θ allows to generate a function v_θ with any asymptotes and convergence rate to asymptotes, and w_θ^- and w_θ^+ with any crossover points and slopes at crossover points. We call the third property "Explicit valuation": for any θ, the functions v_θ, w_θ^- and w_θ^+ lead to explicit gamble valuation for any gaussian reward, i.e. with any mean and variance (and therefore provides an explicit approximate valuation for any bell-shaped rewards). The motivation is to propose a CPT framework well suited for fast computations, for instance on large scale population problems. We illustrate such use with two examples of problems involving CPT and large populations.

5.1 Introduction

In this chapter, we propose a parametric model for Daniel Kahneman and Amos Tversky's Cumulative Prospect Theory ([40]), and for one of its major inspirations, John Quiggin's rank-dependent expected utility theory ([74]). Our parametric model has the advantage to yield a gamble valuation formula that is analytic when the gamble's reward is gaussian, which is an appealing property since most random variables naturally encountered in reality are close to gaussian random variables, as they result of the com-

bination of many of small causes, which by Central Limit Theorem arguments naturally lead to normal bell-shaped distributions.

An analytical gamble valuation formula is useful for many reasons. It can drastically improve the computational cost of the gamble valuation, which, when studying choice problems in large populations, can dramatically speed up computations that would be much costlier, longer, or simply impracticable with non-analytical gamble valuation formulas. Another advantage is the possibility to derive the formula w.r.t. any parameter of the problem, therefore obtaining analytical formulas for them, and use them for optimization problems involving risky choices (e.g. marketing problems) by means of optimization methods using the gradient and/or hessian of the gamble valuation function (e.g. gradient or stochastic gradient descent, Newton's method) without any instability issue. Finally, analytical formulas allow to study the gamble valuation qualitatively, by drawing relations between parameters, studying monotony, convexity, inflexion points, etc.

5.1.1 Origins and motivations

The choice under risk theories that we address in this work, namely, Kahneman and Tversky's Cumulative Prospect Theory ([40]), and John Quiggin's rank-dependent expected utility theory ([74]), are, to this day, considered to be the two most compelling non-expected utility theories for modeling human behavior when facing risky choices.

Both theories present themselves as alternatives to the so-called Expected Utility Theory (EUT), although they can also be seen as generalizations of the EUT. EUT was proposed by Daniel Bernoulli, modelling the reasonable maximal price $V(R)$ one should be willing to pay to enter a gamble R (a real random variable). At the time, the natural assumption was that such price should be the expectation of the gamble's reward, i.e. $V(R) = \mathbb{E}[R]$. However, Bernoulli presented the famous St. Petersburg game, convincingly arguing against this intuition.

Bernoulli then introduced the concept of *utility function*, with the idea that the price someone would pay is not the expectation $\mathbb{E}[R]$ of the monetary reward, but instead of its subjective *utility*, i.e. $V(R) = \mathbb{E}[v(R)]$. This was the first formulation of EUT. EUT started sparking a lot of attention from the 1950s, when John Von Neumann, in Games and Economic Behaviors ([91]), proved that EUT is implied by a set of very compelling *axioms of rationality*. However, empirical studies then revealed several patterns of choice behavior violating EUT ([54, 55]), and there is now a large number of evidences proving that actual choice behaviors systematically violate EUT.

To circumvent EUT's empirical inconsistencies, two opposite research branches emerged, that are now referred to as *conventional* and *non-conventional* choice theories (although

mixing approaches were proposed ([76, 21, 84]).

The non-conventional branch led to Kahneman and Tversky's Nobel Prize awarded Prospect theory ([40]). The conventional branch led to John Quiggin's rank-dependent expected utility theory ([74]), which, paradoxically, was inspired form Kahneman and Tversky's non-conventional Prospect theory ([40]).

Rank-dependent expected utility theory sparked a great interest in the research community ([61]). Axiomatizations were presented ([77, 92, 1, 97, 93]). Generalizations have also been proposed ([18, 32]), and extensions were discussed ([80, 59, 84]).

Later, Kahneman and Tversky themselves developped an improvement of Prospect theory, in turn inspired from Quiggin's rank-dependent expected utility theory ([74]), called Cumulative Prospect theory ([40]).

A conventional choice under risk theory (like Bernoulli's Expected Utility Theory, John Quiggin's rank-dependent expected utility theory [74], and Kahneman and Tversky's Cumulative Prospect theory [40]) models how, given a set of *prospects*, that is, a set of gambles $\mathcal{G}$, i.e. a set of real valued random variables, an individual attributes a deterministic value $V(R)$ to each gamble $R \in \mathcal{G}$ to choose the gamble with highest value, i.e. to choose the gamble

$$R_\star = \operatorname*{argmax}_R V(R).$$

Such model is obviously interesting to predict an individual's choices, and thus very useful for commercial or political applications.

5.1.2 Cumulative Prospect Theory

Cumulative Prospect Theory defines the gamble valuation assigned by an individual to a given prospect as follows:

1. **Functions with constraints:** For this individual, there exists three functions $v : \mathbb{R} \to \mathbb{R}$ and $w^-, w^+ : [0,1] \to [0,1]$ satisfying 1) $v(0) = 0$, v concave on $\mathbb{R}_+$, convex with steeper curve on $\mathbb{R}_-$, 2) $w^+(0) = w^-(0) = 0$, $w^+(1) = w^-(1) = 1$, both increasing inverse S shape functions, such that the gamble valuation $V(R)$ given to any gamble R is given by:

2. **Gamble valuation formula:**

$$V(R) = \int_{-\infty}^{0} v_-(r) d(w^- \circ F_R)(r) + \int_{+\infty}^{0} v_+(r) d(w^+ \circ \bar{F}_R)(r) \qquad (5.1.1)$$

   where $F_R : \mathbb{R} \to [0,1]$ is the cumulative distribution function of R, and $\bar{F}_R : \mathbb{R} \to [0,1]$ is its tail function.

It is possible to define the gamble valuation function V in a more probabilistic way: $V(R)$ is computed with the following steps.

1. The random reward R is split in a gain reward $R^+ := \max(R, 0)$ and a loss reward $R^- := \min(R, 0)$, such that $R = R^+ + R^-$ a.s.,

2. The cumulative distribution of R^- is distorted by applying the weighting w^- to it, yielding a distorted cumulative distribution function. Let us denote $\tilde{R}^-$ a random variable with such distorted CDF,

3. Likewise, the *tail* function of R^+ is distorted by applying the weighting w^+ to it, yielding a distorted tail function. Let us denote $\tilde{R}^+$ a random variable with such tail function,

4. The deterministic value attributed to gamble R is then $V(R) = \mathbb{E}[v(\tilde{R}^-)] + \mathbb{E}[v(\tilde{R}^+)]$.

We stress that the two aspects of CPT, i.e. the **constraints on the functions v, w^- and w^+**, and the **gamble valuation formula**, are both *crucial* to CPT: in particular, the constraints imposed on v, w^- and w^+ are as important as the gamble valuation formula, because each constraint was deduced from many experiments performed by Kahneman and Tversky in their work.

Besides the empirical exposition of v, w^- and w^+'s general shapes, mathematically encoded by these constraints, several parametric classes of analytical functions have been proposed for the value and weight functions v, w^- and w^+. For instance, for the weighting functions, some of the proposed parametrizations are the following.

1. In [84], the class of inverse S shape functions used to fit the empirical data is parametrized by $\gamma \in [0, 1]$, such that to all $\gamma \in [0, 1]$, we associate a function $w_\gamma : [0, 1] \to [0, 1]$ defined by

$$w_\gamma(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{\frac{1}{\gamma}}}, \quad \forall p \in [0, 1]$$

2. In [73], the following function is used, also with $\gamma \in [0, 1]$:

$$w_\gamma(p) = e^{-(-\ln(p))^\gamma}, \quad \forall p \in [0, 1]$$

3. In [83], the inverse S shape function used is the log-odds probability distortion function, $w_{p_0,\gamma} : [0, 1] \to [0, 1]$, where $p_0, \gamma \in [0, 1]$, characterized by the following identity:

$$Lo(w_{p_0,\gamma}(p)) = \gamma Lo(p) + (1-\gamma) Lo(p_0), \quad \forall p \in [0, 1] \tag{5.1.2}$$

where $Lo :]0,1[\to \mathbb{R}$ is the log-odds function defined by $Lo(p) = \ln\left(\frac{p}{1-p}\right)$, $\forall p \in]0,1[$.

It is possible to observe graphically that these are, indeed, inverse S shape functions, although proving it analytically is challenging.

All these functions are hard to distinguish according to [58]. Therefore, it seems that the parametric class of inverse S shape functions used is not really important as long as 1) its functions all are inverse S shape functions and 2) it is rich enough to reproduce the general shape of any inverse S shape function.

As Cumulative Prospect Theory was, in its origin, drawn from experiments and empirical evidences, it appears that even though these various parametrizations were proposed, they were mainly used for experimental purposes, i.e. to fit empirical data with a few parameters and analyze the psychology and behavior of a small group of individual. For these applications, it is not really necessary to have analytical formulas from end to end: numerical and discretization techniques are sufficient to run algorithms optimally fitting the empirical data observed in experiments with the right parameters.

However, we believe that interesting analytical problems involving CPT can be studied, e.g. commercial or political problems with large populations. In such problems, the goal is not to fit empirical data, but instead make a prediction, e.g. of the proportion of people who will make a given choice, or to optimize some gain, e.g. designing a commercial product or political program to optimize the proportion of people who will buy it or vote for it. The large number of computations of gamble valuations in such large population problem is susceptible to make the computation of integrals in (5.1.1) too costly. In these cases, being able to analytically compute the gamble valuations without computing integrals would make computations dramatically faster.

5.1.3 Contributions of this work

Our main contribution is to propose a parametrization of Cumulative Prospect theory, i.e. to propose parametrized classes of 1) reward distributions $\mathcal{R}$ (from which to draw the gamble R), 2) value functions $\mathcal{V}$ (from which to draw v), and 3) weighting functions $\mathcal{W}$ (from which to draw w^- and w^+), such that each $v \in \mathcal{V}$, $w \in \mathcal{W}$ satisfy the constraints in point 1. above, flexible enough to approximate any function satisfying these constraints, and yielding an explicit valuation formula, well suited for large number of gamble valuations in large population problems. The classes we propose are:

1. For $\mathcal{R}$: the class of gaussian reward distributions, parametrized by their mean and variance,

2. For $\mathcal{V}$: the utility functions

$$\begin{aligned} v_{m^-,V^-,a^-,m^+,V^+,a^+}(x) &= -(m^- x + V^-(1 - e^{-a^-(-x)}))\mathbf{1}_{x<0} \\ &\quad +(m^+ + V^+(1 - e^{-a^+ x}))\mathbf{1}_{z\geq 0} \end{aligned}$$

with $m^- \geq m^+$, $V^- \geq V^+$, and $a^- \geq a^+$.

3. For $\mathcal{W}$: the weighting functions

$$w_{p_0,\gamma}(p) = \mathcal{N}(\gamma \mathcal{N}^{-1}(p) + (1-\gamma)\mathcal{N}^{-1}(p_0)), \quad \forall p \in [0,1]$$

where $\mathcal{N}$ denotes the cumulative distribution of the standard normal distribution. This class of weighting functions is similar to the log-odds probability distortion function (5.1.2), except that the function $\mathcal{N}^{-1}$ is used instead of *Lo*. We shall see that this modification still yields inverse S functions (which can be proved analytically), and that they allow us to make explicit valuation computations.

Our main result is the following.

Theorem 5.1.1 *We have*

- **Validity:** *any value function $v \in \mathcal{V}$ and weighting function $w \in \mathcal{W}$ satisfies the constraints in 1..*
- **Density:** *$\mathcal{R}$, $\mathcal{V}$, $\mathcal{W}$ contains Gaussian reward distributions with any mean and variance, value functions with any asymptotes and rate of convergence to the asymptotes, and weighting functions with any crossover point and slope at the crossover point.*
- **Analytic valuation function:** *We have*

$$\begin{aligned} & V_{\mu,\sigma,p_0,\lambda,m^-,V^-,a^-,m^+,V^+,a^+} \\ = \; & -\mathbf{m}^- \left(\boldsymbol{x}\mathcal{N}(\boldsymbol{x}) - \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\boldsymbol{x}^2} \right) - V^- \left(\mathcal{N}(\boldsymbol{x}) - e^{-\mathbf{a}^+\boldsymbol{x} + \frac{(\mathbf{a}^+)^2}{2}} \mathcal{N}(\boldsymbol{x} - \mathbf{a}^+) \right) \\ + \; & \mathbf{m}^+ \left(\boldsymbol{x}\mathcal{N}(\boldsymbol{x}) - \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\boldsymbol{x}^2} \right) + V^+ \left(\mathcal{N}(\boldsymbol{x}) - e^{-\mathbf{a}^+\boldsymbol{x} + \frac{(\mathbf{a}^+)^2}{2}} \mathcal{N}(\boldsymbol{x} - \mathbf{a}^+) \right) \end{aligned}$$

where, $\boldsymbol{x} := \frac{\hat{\mu}}{\hat{\sigma}}$, $\bar{\boldsymbol{x}} := \frac{\bar{\hat{\mu}}}{\hat{\sigma}}$, $\mathbf{m}^+ := \hat{\sigma} m^+$, $\mathbf{m}^- := \hat{\sigma} m^-$, $\mathbf{a}^+ := \hat{\sigma} a^+$, $\mathbf{a}^- := \hat{\sigma} a^-$, where $\hat{\mu} = \mu - \sigma(\gamma^{-1} - 1)\mathcal{N}^{-1}(p_0)$, $\bar{\hat{\mu}} = \mu + \sigma(\gamma^{-1} - 1)\mathcal{N}^{-1}(p_0)$, $\hat{\sigma} = \sigma\gamma^{-1}$.

Our second contribution is to use the analytical expression of V to compute its explicit partial derivatives, and to discuss some applications of our results, in particular to commercial and political problems with large populations.

5.2 The model

In this section, we introduce and motivate each parametric class of functions for our parametric CPT model.

5.2.1 The class $\mathcal{R}$ of reward probability distribution

The probability distribution that is both ubiquitously met in Nature and mathematically convenient to deal with is the gaussian probability distribution. Let us thus simply assume that the class of rewards $\mathcal{R}$ is the class of gaussian variables with any mean and variance (μ, σ).

More realistically, we could also assume that it takes the form of a probability distribution that is only well approximated by a gaussian distribution with same mean and variance, which is essentially often the case in Nature and is mathematically explained by the Central Limit Theorem.

Nonetheless, to simplify, we shall only consider rewards R perfectly following a gaussian probability distribution with mean μ and standard deviation σ, i.e.

$$\mathcal{R} := \{\mathcal{N}(\mu, \sigma), \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+\}.$$

5.2.2 The class $\mathcal{W}$ of weighting functions

Next, we shall design the weighting function w in a way that 1) w has an increasing inverse S shape, and $w(0) = 0$ and $w(1) = 1$, and 2) w combines nicely with a reward following a gaussian probability distribution, as assumed in previous paragraph.

First of all, we write in detail the constraints required for a function $w : [0, 1] \to [0, 1]$ to be a valid wieghting function for CPT.

Definition 5.2.1 (Valid weighting function for CPT) *A valid weighting function is a derivable function $\phi : [0, 1] \to [0, 1]$ such that:*

- $\phi(0) = 0$, $\phi(1) = 1$,
- $\phi'(p) > 0$, $\forall p \in [0, 1]$ *(increasing property),*
- ϕ *has a unique inflexion point* $p^\star \in [0, 1]$, *such that* $\phi''(p^\star) = 0$, $\phi''(p) < 0$ *for* $p \in [0, p^\star[$ *and* $\phi''(p) > 0$ *for* $p \in]p^\star, 0]$ *(inverse S shape property).*

We shall now define a candidate of parametrized class of probability distortion functions for our model. If we take a look at the so-called log-odds probability distortion

function used in [83], $w_{p_0,\gamma} : [0,1] \to [0,1]$, characterized by the following identity:

$$Lo(w_{p_0,\gamma}(p)) = \gamma Lo(p) + (1-\gamma)Lo(p_0), \quad \forall p \in [0,1]$$

we see that it is built from the log-odds function, which can also be seen as the quantile function of the Logistic probability distribution. The Logistic probability distribution looks closely like the Normal probability distribution, in that they both have bell-shape density functions. Therefore, there distribution and quantile functions are both S shaped. The Logistic quantile function Lo would not combine very well with our normal distributed rewards $\mathcal{R}$, so a natural and convenient modification is to replace it by the Normal quantile function, i.e., essentially, to define $w^N_{p_0,\gamma} : [0,1] \to [0,1]$ characterized by the identity

$$\mathcal{N}^{-1}(w^N_{p_0,\gamma}(p)) = \gamma \mathcal{N}^{-1}(p) + (1-\gamma)\mathcal{N}^{-1}(p_0), \quad \forall p \in [0,1]$$

where $\mathcal{N}$ denotes the normal quantile function, which is equivalent to define it as follows.

Definition 5.2.2 (Normal (probability) distortion function) *The Normal probability distortion function $w^N_{p_0,\gamma} : [0,1] \to [0,1]$ associated to parameters $p_0 \in [0,1]$, $\gamma \in [0,1]$, is defined by*

$$w^N_{p_0,\gamma}(p) = \mathcal{N}(\gamma \mathcal{N}^{-1}(p) + (1-\gamma)\mathcal{N}^{-1}(p_0)), \quad \forall p \in [0,1].$$

Validity of $w^N_{p_0,\gamma}$ for $p_0 \in [0,1]$ and $\gamma \in [0,1]$

Let us now prove that for all $p_0 \in [0,1]$ and $\gamma \in [0,1]$, $w^N_{p_0,\gamma}$ is indeed a valid weighting function.

Proposition 5.2.1 *For all $p_0, \gamma \in [0,1]$, the normal probability distortion function $w_{p_0,\gamma}$ is a valid weighting function for the CPT.*

Proof. We only have to compute the successive derivatives and study them. We have

$$w_{p_0,\gamma}(p) = N(\gamma N^{-1}(p) + (1-\gamma)N^{-1}(p_0)), \quad \forall p \in [0,1]$$

For the sake of readability, let $u_0 := (1-\gamma)N^{-1}(p_0)$. We have

$$w_{p_0,\gamma}(0) = N(\gamma N^{-1}(0) + u_0) = N(-\infty) = 0$$

and

$$w_{p_0,\gamma}(1) = N(\gamma N^{-1}(1) + u_0) = N(+\infty) = 1$$

Let us now compute $w'_{p_0,\gamma}$: for all $p \in [0,1]$, we have

$$w'_{p_0,\gamma}(p) = \gamma(N^{-1})'(p)N'(\gamma N^{-1}(p) + u_0) = \gamma \frac{n(\gamma N^{-1}(p) + u_0)}{n(N^{-1}(p))} > 0,$$

where $n : \mathbb{R} \to \mathbb{R}$ denotes the gaussian density defined by $n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $\forall x \in \mathbb{R}$, and where the inequality simply comes from the positivity of the gaussian density. We now check that $w_{p_0,\gamma}$ has a unique inflexion point on $[0,1]$. Let $f : [0,1] \to \mathbb{R}$ be defined by $f(p) = \ln(w'_{p_0,\gamma}(p))$, for all $p \in [0,1]$. Notice that $f' = \frac{w''_{p_0,\gamma}}{w'_{p_0,\gamma}}$, and thus, $w''_{p_0,\gamma}(p)$ has only one zero if and only if f' has only one zero. We have

$$f(p) = \ln(\gamma) + \frac{1}{2}\Big(N^{-1}(p)^2 - (\gamma N^{-1}(p) + u_0)^2\Big)$$

Let us take the derivative:

$$\begin{aligned} f'(p) &= \frac{1}{2}\Big(2(N^{-1})'(p)N^{-1}(p) - 2\gamma(N^{-1})'(p)(\gamma N^{-1}(p) + u_0)\Big) \\ &= (N^{-1})'(p)\Big(N^{-1}(p) - \gamma(\gamma N^{-1}(p) + u_0)\Big) \end{aligned}$$

Notice that $(N^{-1})'(p) = \frac{1}{n(N^{-1}(p))} > 0$ for all $p \in [0,1]$ by the positivity of n. Therefore, we only need to study the sign of $N^{-1}(p) - \gamma(\gamma N^{-1}(p) + u_0)$. We have:

$$\begin{aligned} & N^{-1}(p) - \gamma(\gamma N^{-1}(p) + u_0) \geq 0 \\ \Leftrightarrow\ & N^{-1}(p)(1-\gamma^2) - \gamma u_0 \geq 0 \\ \Leftrightarrow\ & N^{-1}(p) \geq \frac{\gamma u_0}{1-\gamma^2} \Leftrightarrow p \geq N\Big(\frac{\gamma N^{-1}(p_0)}{1+\gamma}\Big) \end{aligned}$$

which means that $w''_{p_0,\gamma}$ is negative on $[0, N\Big(\frac{\gamma N^{-1}(p_0)}{1+\gamma}\Big)]$ and positive on $[N\Big(\frac{\gamma N^{-1}(p_0)}{1+\gamma}\Big), 1]$. □

Gaussian stability property

The class of weighting functions $\mathcal{W}$ has another particularity: it stabilizes gaussian distributions. More precisely, we have the following result.

Lemma 5.2.1 *Let $\mathcal{N}_{\mu,\sigma}$ be the distribution function of a gaussian variable with mean μ and standard deviation σ, and let $w^N_{p_0,\lambda}$ be a normal probability distortion function with crossover point p_0 and slope at the crossover point λ. Then, the function $w^N_{p_0,\lambda} \circ \mathcal{N}_{\mu,\sigma}$ is the distribution function of a gaussian variable with mean $\hat{\mu} := \mu - \sigma(\gamma^{-1} - 1)\mathcal{N}^{-1}(p_0)$ and standard deviation $\hat{\sigma} := \sigma\gamma^{-1}$, and $w^N_{p_0,\lambda} \circ \bar{\mathcal{N}}_{\mu,\sigma}$ is the tail function of a gaussian variable with mean $\bar{\hat{\mu}} := \mu + \sigma(\gamma^{-1} - 1)\mathcal{N}^{-1}(p_0)$ and standard deviation $\bar{\hat{\sigma}} := \sigma\gamma^{-1} = \hat{\sigma}$.*

Proof. We simply have

$$\begin{aligned} & w^N_{p_0,\lambda} \circ \mathcal{N}_{\mu,\sigma}(x) = \mathcal{N}(\gamma \mathcal{N}^{-1}(\mathcal{N}(\frac{x-\mu}{\sigma})) + (1-\gamma)\mathcal{N}^{-1}(p_0)) \\ = \; & \mathcal{N}\left(\gamma \frac{x-\mu}{\sigma} + (1-\gamma)\mathcal{N}^{-1}(p_0)\right) = \mathcal{N}\left(\frac{x-(\mu-(\gamma^{-1}-1)\sigma\mathcal{N}^{-1}(p_0))}{\sigma\gamma^{-1}}\right) \\ = \; & \mathcal{N}_{\hat{\mu},\hat{\sigma}}(x), \end{aligned}$$

where

$$\hat{\mu} = \mu - \sigma(\gamma^{-1}-1)\mathcal{N}^{-1}(p_0), \quad \hat{\sigma} = \sigma\gamma^{-1}$$

Likewise,

$$\begin{aligned} & w^N_{p_0,\lambda} \circ \bar{\mathcal{N}}_{\mu,\sigma}(x) = w^N_{p_0,\lambda} \circ \mathcal{N}_{-\mu,\sigma}(-x) \\ = \; & \mathcal{N}_{\widehat{-\mu},\hat{\sigma}}(-x) = \bar{\mathcal{N}}_{\widehat{--\mu},\hat{\sigma}}(x) = \bar{\mathcal{N}}_{\tilde{\mu},\hat{\sigma}}(x) \end{aligned}$$

where

$$\tilde{\mu} = \mu + \sigma(\gamma^{-1}-1)\mathcal{N}^{-1}(p_0)$$

□

Therefore, the class of normal weighting functions $\mathcal{W}$ "stabilizes" the set of gaussian distributions.

5.2.3 The value function

Let us finally naturally introduce the class $\mathcal{V}$ of value functions. In our model, up to now, we have fixed 1) the class of reward probability distribution to be gaussian distributions, and 2) the class of weighting functions to be the normal weighting functions. Therefore typically, when modeling a situation of choice under risk, one would choose μ, σ, the parameters of the gaussian reward distribution for this risky choice, and λ, p_0, the parameters of the weighting function for this individual. By the definition of the gamble valuation in CPT, and by Lemma 5.2.1, clearly, the class $\mathcal{V}$ of value functions should be a class of functions that are explicitly, or close to explicitly, integrable w.r.t. any gaussian distribution. The only thing left to do now is to build such function satisfying the requirements of prospect theory, that is, having a concave form on $\mathbb{R}_+$ and a steeper convex form on $\mathbb{R}_-$, and cancelling in 0.

An important class of functions that integrate well w.r.t. the gaussian distribution is the class of exponential functions $x \to e^{-ax}$. The function

$$v : x \to (-m^- x - V^-(1 - e^{-a^-(-x)}))\mathbf{1}_{x<0} + (m^+ x + V^+(1 - e^{-a^+ x})\mathbf{1}_{x \geq 0}$$

is concave on $\mathbb{R}_+$, convex on $\mathbb{R}_-$, and steeper on $\mathbb{R}_-$ if $V^- > V^+$, $a^- > a^+$, and $m^- > m^+$. Let us graphically interpreted the parameters of this utility function. First of all, it is separately designed on $\mathbb{R}_-$ and $\mathbb{R}_+$. We focus on the $\mathbb{R}_+$ part. To natural properties that one can imagine to represent a large variety of increasing concave functions are 1) an asymptote, to describe the behavior in $+\infty$, and 2) a rate of convergence toward the asymptote, to describe how quickly the function approaches this asymptote.

The function v defined above has an asymptote $x \mapsto V^+ + m^+ x$ in $+\infty$. and we have, for $x \geq 0$,

$$V^+ + m^+ x - v(x) = V^+ e^{-a^+ x}.$$

Therefore, the exponential rate of convergence of $v(x)$ to $V^+ + m^+ x$ is a^+ (it naturally starts from V^+ at $x = 0$ and then exponentially decays toward 0 with rate a^+).

5.3 The gamble valuation function

In this section, we provide an explicit expression of the gamble valuation $V(R)$ given $R \sim \mathcal{N}(\mu, \sigma)$, and given $w^-, w^+ \in \mathcal{W}$, $v \in \mathcal{V}$. Let us stress that, by "explicit", we mean any expression containing elementary operations, but we also allow the use of the exponential function and the normal cumulative distribution function, two functions who are of course very precisely implemented and optimized (i.e. with fast computation) in any statistical library.

Theorem 5.3.1 *We have*

$$\begin{aligned} & V_{\mu,\sigma,p_0,\lambda,m^-,V^-,a^-,m^+,V^+,a^+} \\ = \; & -\mathbf{m}^- \left(\boldsymbol{x}\mathcal{N}(\boldsymbol{x}) - \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\boldsymbol{x}^2} \right) - V^- \left(\mathcal{N}(\boldsymbol{x}) - e^{-\mathbf{a}^+ \boldsymbol{x} + \frac{(\mathbf{a}^+)^2}{2}} \mathcal{N}(\boldsymbol{x} - \mathbf{a}^+) \right) \\ + \; & \mathbf{m}^+ \left(\boldsymbol{x}\mathcal{N}(\boldsymbol{x}) - \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\boldsymbol{x}^2} \right) + V^+ \left(\mathcal{N}(\boldsymbol{x}) - e^{-\mathbf{a}^+ \boldsymbol{x} + \frac{(\mathbf{a}^+)^2}{2}} \mathcal{N}(\boldsymbol{x} - \mathbf{a}^+) \right) \end{aligned}$$

where, $\boldsymbol{x} := \frac{\hat{\mu}}{\hat{\sigma}}$, $\bar{\boldsymbol{x}} := \frac{\bar{\hat{\mu}}}{\hat{\sigma}}$, $\mathbf{m}^+ := \hat{\sigma} m^+$, $\mathbf{m}^- := \hat{\sigma} m^-$, $\mathbf{a}^+ := \hat{\sigma} a^+$, $\mathbf{a}^- := \hat{\sigma} a^-$, *where* $\hat{\mu} = \mu - \sigma(\gamma^{-1} - 1)\mathcal{N}^{-1}(p_0)$, $\bar{\hat{\mu}} = \mu + \sigma(\gamma^{-1} - 1)\mathcal{N}^{-1}(p_0)$, $\hat{\sigma} = \sigma\gamma^{-1}$.

Proof. The idea is to split the computation into several components. First of all, notice that the fact that gains and losses are processed separately (using different parameters for the weighting function and value function), we can clearly focus on computing the part corresponding to the gains: the part a corresponding to losses will take the same form but with different parameters. On $\mathbb{R}_+$, notice that the value function is

$$v(x) = m^+ x + V^+(1 - e^{-a^+ x}), \quad \forall x \in \mathbb{R}_+.$$

We can thus separately study the terms m^+x, V^+, and $-V^+e^{-a^+x}$. The "gain" part of gamble R, i.e. R_+ will have its tail function distorted by w^+. By Lemma 5.2.1, this will yield another gaussian variable with parameters $\bar{\hat{\mu}}$ and $\hat{\sigma}$ given by Lemma 5.2.1. All we are left to do is, given $Z \sim \mathcal{N}(\bar{\hat{\mu}}, \hat{\sigma})$ to compute separately

$$\mathbb{E}[m^+ Z \mathbf{1}_{Z\geq 0}], \quad \mathbb{E}[V^+ \mathbf{1}_{Z\geq 0}], \quad \mathbb{E}[-V^+ e^{-a^+x} \mathbf{1}_{Z\geq 0}].$$

The second term is simply given by $\mathbb{E}[V^+\mathbf{1}_{Z\geq 0}] = V^+\mathcal{N}(\frac{\bar{\hat{\mu}}}{\hat{\sigma}})$. Let us compute the third term. We have

$$\begin{aligned}
& -V^+\mathbb{E}[e^{-a^+Z}\mathbf{1}_{Z\geq 0}] = -V^+e^{-a^+\hat{\mu}}\mathbb{E}[e^{-a^+\hat{\sigma}N}\mathbf{1}_{N\geq -\frac{\hat{\mu}}{\hat{\sigma}}}] = -V^+e^{-a^+\hat{\mu}}\mathbb{E}[e^{a^+\hat{\sigma}N}\mathbf{1}_{N<\frac{\hat{\mu}}{\hat{\sigma}}}] \\
= & -V^+e^{-a^+\hat{\mu}}\int_{-\infty}^{\frac{\hat{\mu}}{\hat{\sigma}}} e^{a^+\hat{\sigma}x}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx = -V^+e^{-a^+\hat{\mu}}e^{\frac{(a^+\hat{\sigma})^2}{2}}\int_{-\infty}^{\frac{\hat{\mu}}{\hat{\sigma}}}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-a^+\hat{\sigma})^2}{2}}dx \\
= & -V^+e^{-a^+\hat{\mu}+\frac{(a^+\hat{\sigma})^2}{2}}\mathbb{P}(N+a^+\hat{\sigma}<\frac{\hat{\mu}}{\hat{\sigma}}) = -V^+e^{-a^+\hat{\mu}+\frac{(a^+\hat{\sigma})^2}{2}}\mathbb{P}(N<\frac{\hat{\mu}-a^+\hat{\sigma}^2}{\hat{\sigma}}) \\
= & -V^+(e^{-a^+\hat{\mu}+\frac{(a^+\hat{\sigma})^2}{2}}\mathcal{N}(\frac{\hat{\mu}-a^+\hat{\sigma}^2}{\hat{\sigma}}))
\end{aligned}$$

Finally, the first term is obtained as follows. Notice that

$$\mathbb{E}[Z\mathbf{1}_{Z\geq 0}] \quad = \quad -\partial_{a^+=0}\mathbb{E}[e^{-a^+Z}\mathbf{1}_{Z\geq 0}] = \hat{\mu}\mathcal{N}(\frac{\hat{\mu}}{\hat{\sigma}}) - \frac{\hat{\sigma}}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\hat{\mu}}{\hat{\sigma}}\right)^2}$$

□

Notice that the explicit expression of $V(R)$ makes it easily derivable in each of the model's parameters.

5.4 Applications to large population problems

In this section, we provide some ideas about how one can take advantage of the explicit formulas of this gaussian Cumulative Prospect model.

The reality is that there are an endless amount of possibilities to use analytical formulas to one's advantage: quick computation, interpretations, linking parameters together, studying sensitivities of the gamble valuation to parameters, using explicit gradient and hessian for optimization algorithms (gradient or stochastic gradient descent, Newton's method, etc), analytically solving problems, real-time plotting, high-frequency decision making (e.g. high-frequency trading) etc. Our goal here is not to investigate all these possibilities, but instead to detail two possible usages of the formula, in the context of large population problems.

5.4.1 Optimal product/program design for a large population

Although numerically approximating an integral, nowadays, does not take any noticeable computational time, it is still, at the microscopic level, orders of magnitude longer than computing an explicit formula. When computing a single gamble valuation, it might thus not make a big difference, but when repeatedly computing a large number of them, it makes a great one. The way we will illustrate a possible advantage of our analytical formula is thus by providing natural situations where one would want to repeatedly compute a large number of gamble valuations.

Consider an agent designing some program. By agent, we mean a company, a politician, or any influencer in general. By program, we mean a marketing program (designing a new product, a new show), or an electoral program (designing a presidential program, a reform, etc). Let $\mathcal{P}$ denote the set of all possible programs.

We now consider a population of N individuals. Each individual n, for $n \leq N$, has a *personality* e_n in some personality space E. We assume that the personality $e \in E$ of an individual fixes his utility and weighting functions, i.e. that there are functions

$$\begin{aligned} &p_0^+, \lambda^+, V^+, m^+, a^+, p_0^-, \lambda^-, V^-, m^-, a^- : e \in E \\ &\mapsto p_0^+(e), \lambda^+(e), V^+(e), m^+(e), a^+(e), p_0^-(e), \lambda^-(e), V^-(e), m^-(e), a^-(e), \end{aligned}$$

fixing the crossover point $p_0^+(e)$ and slope at the crossover point $\lambda^+(e)$ for the individual's weighting function on gains, the asymptote parameters $V^+(e)$ and $m^+(e)$ and the convergence rate to the asymptote $a^+(e)$ of the utility function on $\mathbb{R}_+$, and likewise on $\mathbb{R}_-$. Furthermore, we assume that there are functions $\mu, \sigma : E \times \mathcal{P} \to \mathbb{R}, \mathbb{R}_+$ such that any individual with personality $e \in E$ receives, from a given program $P \in \mathcal{P}$, a random reward with gaussian distribution with mean $\mu(e, P)$ and variance $\sigma(e, P)$. This can all be summarized by considering that there exists a function $\theta : E \times \mathcal{P} \to \Theta$ associating to a personality $e \in E$ and a program $P \in \mathcal{P}$ the parameters $\theta(e, P)$ associated to this individual in the model, for the program P, such that the value he attributes to the program (seen as a gamble) is $CE_{\theta(e,P)}$, and is thus an analytical function of the parameters $\theta(e, P)$. Provided that $\mu(e, P)$ and $\sigma(e, P)$ are analytic in P, $CE_{\theta(e,P)}$ is analytic in P as well.

Let us assume that the population has to make an action: in the case of a company, it can be "buying or not the new product", "watching or not the new show", etc. In the case of a politician, it can be "voting or not for the politician", or for instance "voting for or against the reform in a referendum". The action of an individual with personality $e \in E$, given program P, can be assumed to be $\mathbf{1}_{CE_{\theta(e,P)}>0}$, i.e. the individual sees the

choice as a binary choice with choice 1 representing "buying the product" or "voting for the candidate", which has value $CE_{\theta(e,P)}$, and the other choice, essentially representing "not buying it" or "not voting for him", can be, to fix ideas, seen as a neutral choice with null utility. Therefore, the action $\mathbf{1}_{CE_{\theta(e,P)}>0}$ simply means that we assume that the individual acts according to his preference.

Therefore, if program $P \in \mathcal{P}$ is presented to the population, the distribution of choices 1 in the population should be $\frac{1}{N}\sum_n \mathbf{1}_{CE_{\theta(\varepsilon_n,P)}>0}$. Notice that this is natural situation where a large number of certainty equivalences have to be computed, N typically representing millions, or even billions of people. Simply computing this expression with N integral numerical approximations versus N analytical expressions, is enough to see a big difference in computational time.

However, the complexity of the problem can naturally be pushed further away: indeed, the goal of the agent (the company, or the politician), is to design the right program $P \in \mathcal{P}$. We can assume that there exists a gain function for the agent, taking the form

$$G_N(P) = g(P, \frac{1}{N}\sum_n \mathbf{1}_{CE_{\theta(e_n,P)}>0})$$

that is, depending upon the program and the proportion of people who will choose this program if the agent proposes it. The goal of the agent is then to compute

$$P_N^\star = \operatorname*{argmax}_{P\in\mathcal{P}} G_N(P)$$

Notice how such optimization problem would require to compute $G_N(P)$ for many different $P \in \mathcal{P}$, each computation itself involving the estimation of N certainty equivalences. Here, again, we would see a great computational difference between certainty equivalences computed by integral numerical approximation or with an analytical formula.

Furthermore, by a mean-field approximation argument, one could approximate the problem with its mean-field version: if we approximate $\frac{1}{N}\sum_n \delta_{e_n}$ with a probability distribution ν, and denote ε a random variable with distribution ν, we consider the gain function:

$$G(P) = g(P, \mathbb{P}(CE_{\theta(\varepsilon,P)} > 0))$$

and the optimization problem

$$P^\star = \operatorname*{argmax}_{P\in\mathcal{P}} G(P)$$

The probability $\mathbb{P}(CE_{\theta(\varepsilon,P)} > 0)$ can be approximated with many methods, all involving the computation of several $CE_{\theta(e,P)}$, which, again, will be faster with analytical formulas. Finally, the optimization of such function is susceptible to require the computation of $\partial P \mathbb{P}(CE_{\theta(\varepsilon,P)} > 0) = \mathbb{E}[\partial_P CE_{\theta(\varepsilon,P)} \mid CE_{\theta(\varepsilon,P)} = 0]$: Here, notice that being able to explicitly compute the derivatives of the certainty equivalence w.r.t. the models parameters would be very useful.

5.4.2 Equilibrium computation in a social game with large population

Another situation where an analytical certainty equivalence formula can be useful is if we consider a social game in a large population.

In the framework introduced in previous section, let us assume that the agent is a company and that it is selling a product P. Let us fix the product P here. We however assume that the reward perceived by the individuals in population from making a choice 0 or 1 does not only come from the product P but also from subsequent social interactions they might have with other people: for instance, if two people who have both bought the product interact, they receive an additional positive reward.

Concretely, this means that if $(x_n)_{n\leq N} \in \{0,1\}^N$ represents the choices made by each individual n, then the reward received by individual n is not simply the product's reward $R_{P,\varepsilon_n} \sim \mathcal{N}(\mu(\varepsilon_n, P), \sigma(\varepsilon_n, P))$, but it instead

$$R_{P,\varepsilon_n} + \mathrm{u}(\frac{1}{N}\sum_i x_i) \sim \mathcal{N}\Big(\mu(\varepsilon_n, P) + \mathrm{u}(\frac{1}{N}\sum_i x_i), \sigma(\varepsilon_n, P)\Big)$$

where $\mathrm{u}(\frac{1}{N}\sum_i x_i)$ is a social reward depending upon other people's choices. This is equivalent to say that the mean of the reward depends upon the choice distribution of the other individuals. Thus, the parameter function θ now takes the form $\theta(\varepsilon_n, P, \mathrm{u}(\frac{1}{N}\sum_i x_i))$. The goal of individual n is to make the choice maximizing his certainty equivalence, i.e. he would like to make the choice x_n such that

$$x_n = \mathbf{1}_{CE_{\theta(\varepsilon_n,P,\mathrm{u}(\frac{1}{N}\sum_i x_i)}>0},$$

but to do so, he would have to know what choice the other players will make. This turns the problem into a game in large population. Notice that, as in this case, $CE_{\theta(\varepsilon_n,P,u)}$ is clearly strictly increasing in u , maximizing the certainty equivalence is equivalent to make the choice

$$x_n = \mathbf{1}_{\tau(\varepsilon_n,P)<\mathrm{u}(\frac{1}{N}\sum_i x_i)},$$

where $\tau(\varepsilon_n, P) = CE^{-1}_{\theta(\varepsilon_n,P,\cdot)}(0)$. We obtain the type of game that we studied in last chapter, and we know that rational players will end up essentially playing the unique fixed point of the operator

$$p \mapsto \mathbb{P}(\tau(\varepsilon_n, P) < \mathrm{u}(p))$$

that is, the unique fixed point of

$$p \mapsto \mathbb{P}(CE_{\theta(\varepsilon_n,P,\mathrm{u}(p))} > 0)$$

Such fixed point can be computed:

- either by applying several iterations of this contracting operator: in this case, it means computing several times a large amount of $CE_{\theta(\varepsilon_n,P,\mathrm{u}(p))}$, and thus, using an analytical expression for it is a great gain of time,
- or, in the case where $\mathrm{u}(p) = c\tilde{\mathrm{u}}(p)$ for c small, by using the fact that the fixed point $p^\star(c)$ of

  $$p \mapsto \mathbb{P}(CE_{\theta(\varepsilon_n,P,c\tilde{\mathrm{u}}(p))} > 0)$$

  is, for $c = 0$, simply $p^\star(0) = \mathbb{P}(CE_{\theta(\varepsilon_n,P)} > 0)$ (the distribution of choices without social influence), and thus, for c small, one can make a first order extension of $p^\star(c)$ using the implicit function theorem. In this case, clearly, such first order extension would require to derive $CE_{\theta(\varepsilon_n,P)}$ w.r.t. its parameters, which is again a lot more practical when ones has an analytical formula for it.

5.5 Conclusion

In this chapter, we have provided a parametric model for Cumulative Prospect Theory. More precisely, we have defined a set of parameters Θ and a map $\theta \in \Theta \mapsto (v_\theta, w^-_\theta, w^+_\theta)$ such that for all $\theta \in \Theta$, the functions v_θ, w^-_θ, and w^+_θ, are *valid* value and weighting functions (validity result), also such that any general shape of valid functions can be reproduced with the right parameter θ (density result), and yielding an explicit valuation formula for gaussian rewards. We have shown that such formula could easily be derived w.r.t. any coordinates of the parameter θ, and we also have provided two examples involving large populations, where such analytical valuation function (and its derivatives) can be very useful for speeding up computations associated to these problems by several orders of magnitude. We believe that interesting work can be made in this direction to develop these large population problems, only briefly discussed in this paper. A natural

application of any choice theory being to predict the choice of individuals, it indeed seems relevant to apply it to large populations in commercial or political problems, since generally, what one is really interested in, in these cases, is the aggregation of individuals' choices rather than the choice of a single individual. As mentioned in our brief discussion, such problem would naturally apply to the problem of designing the right product to sell to maximize the number of sales, or designing the right political program to reach a given proportion of votes.

Part III

Models for targeted advertising

Chapter 6

Online click learning algorithm for targeted advertising

Abstract. In this chapter, we introduce and study an online click prediction learning algorithm for targeted advertising. The algorithm is based on a polynomial classifier with soft or hard margin, and, to learn, only requires to observe clicks on displayed ads. We show that all the ads that would lead to a click will be displayed, and that the expected number of displayed ads that are not clicked on is logarithmic in the number of past ads. In classification terminology, this is to say that our algorithm makes no false negative and only makes a logarithmic amount of false positives in the number of past stages. We finally prove the boundedness of the average memory usage and of the computational complexity.

6.1 Introduction

In this chapter, we define and study an online click prediction learning algorithm specifically designed for targeted advertising. We prove that its learning efficiency is logarithmic. Furthermore, the memory and time complexity of the algorithm at each time $n \in \mathbb{N}$ is uniformly bounded over all times n. Finally, the particularities of our learning algorithm are:

- It makes no false negatives, i.e. the prediction errors consisting in not displaying an ad that would have led to a click never occurs. This is an important feature when the errors cost are very asymmetric between error types. In advertising, the social network will generally only be paid by the company if the individual has clicked on the ad (CPC advertising) or made a purchase/subscription (CPA advertising). Therefore, not displaying an ad that would have led to a click is a real loss. On the

other hand, the other prediction error type, consisting in displaying an ad that does not lead to a click, only generates a small "inconvenience" to the user's experience. Therefore, it is much more problematic for the social network to not display an ad that would have led to a click (and a profit) than displaying and ad which is not clicked on. This motivates the constraint to completely avoid the first error type while trying to minimize the other one.

- The algorithm does only need to access the clicking decisions for ads that it displays. In other words, for each ad that was not displayed by the algorithm, the clicking decision that *would have occurred* if the ad had been displayed is never revealed to the algorithm. In classification terminology, this is called *partial, or asymmetric feedback*. This feature is obviously necessary in targeted advertising, since, clearly, if an ad was not even displayed to the individual, there is no way to access the reaction he would have had, seeing this ad.

- Finally, our algorithm relies on polynomial classifiers to predict the clicks of the individual, with a soft margin in the case where the clicks can only be approximately well predicted with polynomial classifiers. The better the individual's clicks can be predicted with polynomial classifiers, the better will the algorithm's efficiency be. Eventually, if there exists a polynomial perfectly predicting the individual's clicks, the algorithm has a logarithmic number of prediction errors. This property is opposite to many classification algorithms relying on gradient descent or stochastic gradient descent, for which some regularity assumptions generally have to be made, making the algorithm non-efficient for perfectly separable data (hard margin), and thus generally requiring only a soft margin, see for instance Logistic Regression.

The use of classification algorithms for click prediction in targeted advertising and web recommendation is not new. Probably the most studied problems in machine learning are classification problems. Many algorithms have been studied to learn binary classification. Among the most famous are Support Vector Machines, first introduced by Vapnik and Cortes ([19]), Logistic Regression, invented by Berkson [9] and Cox [20], probit models, decision trees, and neural networks. Classification problems are all based on a common framework: there is an input space X and an output, or label, space Y. In our case, X can encode every aspect of an ad, and Y can correspond to the binary clicking decision that a given individual would make on this ad. The goal essentially is to find a map $f : X \to Y$ making a small amount of prediction errors. The map $f : X \to Y$ is then referred to as a classifier.

Classification is a type of *supervised learning*. In classical "offline" learning, the

setup is that one assumes that we have access to training data $(X_1, Y_1), ..., (X_n, Y_n) \in X \times Y$, and a classification algorithm is then a procedure receiving the training data and outputting a classifier f. An assumption made in general is that $(X_i, Y_i)_{i \leq n}$ are i.i.d. random variables (random framework), but not always (see adversarial framework).

For more details about classification learning, we refer to the many textbooks, surveys, and monographs on these topics, like [4], [57], [22], [25], [30], [45], [48], [60], [62], [63], [69], and [85, 86, 87].

Another branch of classification learning, called *online classification learning*, refers to the situation where no training data is initially accessible, and where inputs $(X_t)_{t \in \mathbb{N}}$ comes as time goes by. An online classification algorithm is more complex than an offline classification algorithm as it has to compute a sequence $(f_t)_{t \in \mathbb{N}}$ of classifiers such that f_t is the update of the classifier at time t. The idea is that at each time t, two actions are taken: 1) a prediction of the output of X_t using classifier f_t, and 2) an update of the classifier to f_{t+1} for future predictions, taking into account the data received at time t and the data stored in memory.

The particularity of online learning is to mix the learning and the prediction stages. This is susceptible to cause problems where one has to choose between learning (exploration) and predicting well (exploitation). A second challenge is the management of memory and computational time: the idea is, rather than making at each time $t+1$ an offline learning from all past data to compute the classifier f_{t+1}, to take advantage of the previous classifier f_t, and simply "surgically incorporate" the data processed since this last update in the classifier.

Organization of the chapter: In Section 6.2, we introduce the framework and problem. In Section 6.3, we directly and quickly provide the algorithm in a form that is as close as possible to its concrete implementation, in order to illustrate its formal simplicity and implementability. Then, in Section 6.4, we introduce the mathematical objects related to this algorithm that are important to both state *and* study its efficiency (prediction, memory, computation) in a rigorous way, and in Section 6.5 we state our main results, estimating this efficiency. Once these two steps are passed, in Section 6.6, we finally prove the results, and conclude in Section 6.7.

6.2 The problem

We start by modeling commercial products. A product is associated to a *price* $p \in \mathbb{R}$ and to features $\mathbf{f} \in \mathcal{F} := [0,1]^d$. A product's features $\mathbf{f} \in \mathcal{F}$ represent the characteristics of a product (quality, brand's reputation, shape, life duration, etc). Thus, a product is characterized by a price-features pair $(p, \mathbf{f}) \in \mathbb{R} \times \mathcal{F}$. By misuse of language, we identify

the product with any advertisement of the product. We will thus indifferently say "the product $(p, \mathbf{f})$" and "the ad $(p, \mathbf{f})$".

A click will be represented by a binary variable $c \in \{-1, 1\}$, 1 meaning "click", and -1 "no click". Depending upon the context, it will have slightly different interpretations: as long as an ad $(p, \mathbf{f}) \in \mathbb{R} \times \mathcal{F}$ has not been displayed to the individual, the associated $c \in \{-1, 1\}$ is a *clicking intention*, but once (and if) the ad $(p, \mathbf{f})$ is displayed to the individual, c will correspond to his *clicking decision*.

We denote by $\mathcal{R}_{d,D}$ the set of multi-dimensional polynomial functions from $\mathcal{F}$ to $\mathbb{R}$ with maximal degree D in each coordinate, i.e. taking the form

$$R(\mathbf{f}) = \sum_{\mathbf{i} \in [\![0,D]\!]^d} r_{\mathbf{i}} \prod_{k=1}^{d} f_k^{i_k}, \quad \forall \mathbf{f} = (f_k)_{k \in [\![1,d]\!]} \in \mathcal{F}.$$

where $\mathbf{r} = (r_{\mathbf{i}})_{\mathbf{i} \in [\![0,D]\!]^d} \in \mathbb{R}^{D^d}$ is a multi-index vector. A function $R \in \mathcal{R}$ will be interpreted as an (approximate) *reward function*, associating to any features $\mathbf{f} = (f_i)_{i \in [\![1,d]\!]}$ the reward $R(\mathbf{f}) \in \mathbb{R}$. The higher the degree D is, the better the class $\mathcal{R}$ is at approximating any regular function. Notice that as soon as $D \geq 1$, $\mathcal{F}$ contains all the affine functions from $\mathcal{F}$ to $\mathbb{R}$.

The framework for the online learning algorithm is the following. We consider a random sequence $(p_k, \mathbf{f}_k, c_k)_{k \in \mathbb{N}}$ of ads $(p_k, \mathbf{f}_k)$ and clicking intentions c_k for all $k \in \mathbb{N}$. More precisely, at each time $k \in \mathbb{N}$, a new product is created by a company, with price $p_k \in \mathbb{R}$ and features $\mathbf{f}_k = (f_{k,i})_{i \in [\![1,d]\!]} \in \mathcal{F}$, yielding the product/ad $(p_k, \mathbf{f}_k)$. For a given individual, the company wonders if it should display the ad $(p_k, \mathbf{f}_k)$ to a given individual. The binary value c_k represents the *clicking intention* of the individual for this product, that is, c_k is the answer to the question "if ad $(p_k, \mathbf{f}_k)$ were to be displayed to the individual, would he click on it?". If so, then, by definition, $c_k = 1$. Otherwise, $c_k = -1$.

We make the following assumption:

Existence of approximate polynomial reward function: We assume that there exists a *reward function* $R_\star \in \mathcal{R}$ such that, $\forall k \in \mathbb{N}$,

$$p_k < R_\star(\mathbf{f}_k) - \varepsilon \Rightarrow c_k = 1, \quad \text{and} \quad p_k > R_\star(\mathbf{f}_k) + \varepsilon \Rightarrow c_k = -1, \tag{6.2.1}$$

where $\varepsilon \geq 0$ is a margin. This is equivalent to

$$|R_\star(\mathbf{f}_k) - p_k| > \varepsilon \Rightarrow c_k = sgn(R_\star(\mathbf{f}_k) - p_k).$$

This simply means that except when the product's price p_k is too close to the theoretical reward $R_\star(\mathbf{f}_k)$ (or equivalently, when the theoretical net reward $R_\star(\mathbf{f}_k) - p_k$ is too close to

0), the clicking decision of the individual can be inconsistent with the natural prediction associated to the reward function $R_\star$, that is, one might have $c_k \neq sgn(R_\star(\mathbf{f}_k) - p_k)$, but outside of this case, $R_\star$ well predicts the individual's clicks, i.e. we have $c_k = sgn(R_\star(\mathbf{f}_k) - p_k)$.

The margin ε is important for realism because it encompasses several natural phenomenons:

- **Non-polynomial reward functions:** the "real" reward function of the individual might not be polynomial, but only approximable with a polynomial reward function up to an error ε,
- **Hidden variables, or inconsistent clicking decisions:** there might be an unobservable noise in the individual's evaluation of the product's utility making him value slightly differently a same product at two different times.
- **Time varying reward function:** The underlying reward function of the individual might slightly evolve with time, and thus, for $n \leq N$, all the successive reward functions of the individual are close to the first one up to a margin error ε.

We also make the following probabilistic assumptions:

I.i.d. products with atomless distribution: We assume that $(p_k, \mathbf{f}_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. random variables with common distribution ν assumed atomless and such that $\frac{d\nu}{d\lambda} \leq C$ for some constant C. We stress that the i.i.d. assumption is only made on $(p_k, \mathbf{f}_k)_{k\in\mathbb{N}}$, not $(p_k, \mathbf{f}_k, c_k)_{k\in\mathbb{N}}$. Actually, the only assumption involving $(c_k)_{k\in\mathbb{N}}$ is the previous one.

Upper and lower bounded conditional density at the margin: We assume that there exists $\eta > \varepsilon$ such that

$$c < \frac{d\mathcal{L}(p_1 - R_\star(\mathbf{f}_1) \mid \mathbf{f}_1)}{d\lambda}(y) < C, \quad \forall y \in [-\eta, \eta], \quad a.s.$$

Let us now informally describe the rules of the targeted advertising problem. At each time $k \in \mathbb{N}$ the following steps occur:

1. **New ad event:** a new ad advertising a new product $(p_k, \mathbf{f}_k)$ is created. At this point, $(p_k, \mathbf{f}_k)$ is observable to the advertiser and can be used, along with data stored in memory from past times, for the subsequent steps.
2. **Displaying decision:** the advertiser executes a program processing $(p_k, \mathbf{f}_k)$ and data stored in memory from last times to decide whether or not to display ad $(p_k, \mathbf{f}_k)$ to the individual.

3. **Clicking reaction (this step happens only if ad $(p_k, \mathbf{f}_k)$ was displayed to the individual):** Once ad $(p_k, \mathbf{f}_k)$ is displayed, the individual sees $(p_k, \mathbf{f}_k)$ it and either clicks on it or not, according to the clicking intention c_k. In either case, the advertiser observes the reaction of the individual, which means that he observes c_k. We stress that if the advertiser chose to not display ad $(p_k, \mathbf{f}_k)$ in last step, this step does not happen and the advertiser does *not* observe c_k.

4. **Memory update:** The advertiser has the possibility to update the variables stored in memory, and in particular he can choose to remember $(p_k, \mathbf{f}_k)$, and, provided that he displayed the ad to the individual at step 2, the clicking reaction c_k observed ad step 3, for future use.

6.3 The algorithm

The goal of this section is to write the online algorithm in a form that is as close as possible to its concrete implementation.

6.3.1 Feature space transformation ϕ

We introduce a function $\phi : [0,1]^d \to [0,1]^{D^d}$, defined by $\phi(\mathbf{f}) = (\phi(\mathbf{f})_\mathbf{i})_{\mathbf{i} \in [\![1,D]\!]^d}$, where, for any multi-index $\mathbf{i} = (i_1, ..., i_d) \in [\![1, D]\!]^d$, we have

$$\phi(\mathbf{f})_\mathbf{i} = \prod_{k=1}^{d} f_k^{\mathbf{i}_k}, \quad \forall \mathbf{f} = (f_k)_{k \in [\![1,d]\!]} \in [0,1]^d.$$

This is simply the function associating to a vector $\mathbf{f} \in [0,1]^d$ the vector of evaluations in $\mathbf{f}$ of each multi-dimensional monomials with degree smaller than D in each coordinate. The purpose of ϕ is to *linearize* the classification problem, i.e. turning the search of the *polynomial* reward function $R_\star$ into the search of a *linear* function. This is an important step for computational efficiency, because it will allow us to implement the core function of the algorithm, the *update* function, with standard linear programs.

6.3.2 The *update* function

In this section, we define the *update* function, a core function of our online algorithm. Its definition is here provided in a close to implemented way using linear programs, which are very standard optimization problems solved by several methods (Simplex method, interior point method, etc). A more theoretical and mathematically meaningful definition of the *update* function would require some intermediary definitions, that we postpone to Section 6.6.

Input. $(u, e, D) \in \mathbb{R}^{D^d+1} \times \mathbb{R}_+ \times \mathcal{P}(\mathbb{R}^{D^d+1} \times \{-1, 1\})$.

For any $v \in \{0\} \times \{0,1\}^{D^d}$, we denote $p^{\min}(v, u, e, D)$ (resp. $p^{\max}(v, u, e, D)$) the minimum (resp. maximum) reached by the following linear programs

$$\begin{aligned}
\text{Minimize/Maximize} \quad & v \cdot u', \forall u' \in \mathbb{R}^{D^d+1} \\
\text{subject to:} \quad & u'_1 = -1, \\
& a \cdot u' > -2\varepsilon, \quad \forall a : (a, 1) \in D \\
& a \cdot u' < 2\varepsilon, \quad \forall a : (a, 0) \in D \\
& v' \cdot u' > v' \cdot u - e, \quad \forall v \in \{0\} \times \{0,1\}^{D^d} \\
& v' \cdot u' < v' \cdot u + e, \quad \forall v \in \{0\} \times \{0,1\}^{D^d}
\end{aligned}$$

where $e_v(u') = u'(v)$ is the evaluation function.

Then, we denote by $update(u, e, D)$ the minimizer of the following linear program:

$$\begin{aligned}
\text{Minimize} \quad & e', \forall (u', e') \in \mathcal{U} \times \mathbb{R} \\
\text{subject to:} \quad & u'_1 = -1, \\
& u'(v) - e' < p^{\min}(v, u, e, D), \quad \forall v \in \{0,1\}^d \times \{0\} \\
& u'(v) + e' > p^{\max}(v, u, e, D), \quad \forall v \in \{0,1\}^d \times \{0\}
\end{aligned}$$

Output. $update(u, e, D) \in \mathbb{R}^{D^d+1} \times \mathbb{R}_+$.

The *update* function will be used to update our classifier during the online algorithm, and it will thus be called at different times to make the click predictions more and more efficient. A better understanding of what the *update* function truly does would require theoretical intermediary definitions, not needed for the implementation, and thus, again, postponed to Section 6.6. For now, let us just say that at each call of the *update* function, the argument $(u, e) \in \mathbb{R}^{1+D^d} \times \bar{\mathbb{R}}_+$ will correspond to the current classifier, and the set $D \subset \mathbb{R}^{D^d+1} \times \{-1, 1\}$ will correspond to a set of transformed labelled data (displayed ads and associated clicks), and the output $update(u, e, D) \in \mathbb{R}^{1+D^d} \times \bar{\mathbb{R}}_+$ will correspond to the new classifier. The *update* function will thus use the transformed labelled data set D to update the classifier (u, e) to the more accurate one $update(u, e, D)$. In this sense, the *update* function is how the algorithm learns and gets better and better at predicting clicks.

Essentially, given this *update* function, all that the online algorithm consists in is to define *how* the transformed labelled data set D is obtained before each update, and *when* each update is performed.

6.3.3 The online algorithm

In this section, we write our online algorithm in pseudo-language.

Before the problem starts, we run the Setup Algorithm 1 to initialize variables. Then, at

Algorithm 1 Setup.

$N \leftarrow 0$, $n \leftarrow 0$, $u \leftarrow (-1, 0_{\mathbb{R}^{D^d}}) \in \mathbb{R} \times [0,1]^{D^d}$, $e \leftarrow +\infty$, $D \leftarrow \emptyset$, $c \leftarrow -1$

each time $k \in \mathbb{N}$, we receive an ad (f_k, p_k) and process it with the Online Algorithm 2, where the *display* function is a function such that the call $display(f, p)$ has the following effect:

- it displays the ad (f, p) to the individual,
- it outputs the individual's reaction, i.e. $display(f, p) = 1$ if the individual clicked on the ad, and $display(f, p) = -1$ otherwise.

Algorithm 2 Online algorithm.

Input: $p \in \mathbb{R}, f \in [0,1]^d$
$a \leftarrow (p, \phi(f)) \in \mathbb{R} \times [0,1]^{D^d}$
if $u \cdot a > -e$ **then**
  $c \leftarrow display(p, f)$
  if $u \cdot a < e$ **then**
    $D \leftarrow D \cup \{(a, c)\}$
  end if
end if
$N \leftarrow N + 1$
if $N = 2^n$ **then**
  $(u, e) \leftarrow update(u, e, D)$
  $D \leftarrow \emptyset$, $N \leftarrow 0$, $n \leftarrow n + 1$
end if

The algorithm is here defined in a close to implemented way. Again, its rigorous theoretical study is postponed to Section 6.4 for the definition of the associated mathematical objects and to Section 6.6 for their analysis. Let us however here give the general idea of the algorithm. At each time, the variables "u" and "e" (with value e_k) are used to predict the next click, i.e. to decide whether or not to display the ads, with the test "$u \cdot a > -e$". Let us here admit that such test implies that there will be not click on the ad. This thus justifies that we don't display it. If, however, the first test does not fail, we

display the ad. Likewise, we admit that if the second test "$u \cdot a < e$" fails, there will be a click with certainty, and this means that it is not necessary to remember the observed individual's clicking reaction for improving our predictor since it was already able to predict it with certainty. If both tests are positive, we say that the click is non-predictable with "(u, e)", and thus, the ad is said to be non-predicted, and as we have displayed it, we can store the observe clicking reaction in the variable "D", which will be later used to update our predictor "(u, e)". Finally, the updates (calls "$(u, e) \leftarrow update(u, e, D)$") occur at the times $(2^n)_{n\in\mathbb{N}}$, i.e. the time between two successive updates doubles at each update. This is intuitively justified by the idea that the more we update the predictor "(u, e)", the more precise it is, and thus the more clicks it will be able to predict with certainty, and, therefore, the slower the set "D" of non-predicted clicks will fill itself, which is why more time is needed before using it to update our predictor.

6.3.4 Algorithm with tracking variables

In this section, we re-write the algorithm by adding *tracking variables*, that is, variables not really affecting how the algorithm operates, but simply recording the values of some of the variables at key moments, which will help us to analyze the algorithm.

The tracking variables are the lists "$\boldsymbol{d}$", "$\mathbf{D}^{\boldsymbol{M}}$", "$\mathbf{N}^{\boldsymbol{U}}$", "$\mathbf{D}^{\boldsymbol{U}}$", "$\mathbf{a}$", "$\boldsymbol{u}$", "$\mathbf{e}$",

Algorithm 3 Setup with tracking variables.

$N \leftarrow 0$, $n \leftarrow 0$, $u \leftarrow (-1, 0_{\mathbb{R}^{Dd}}) \in \mathbb{R} \times [0, 1]^{Dd}$, $e \leftarrow +\infty$, $D \leftarrow \emptyset$, $c \leftarrow -1$
$k \leftarrow 0$, $\boldsymbol{d} \leftarrow []$, $\mathbf{D}^{\boldsymbol{M}} \leftarrow []$, $\mathbf{N}^{\boldsymbol{U}} \leftarrow []$ $\mathbf{D}^{\boldsymbol{U}} \leftarrow []$, $\mathbf{a} \leftarrow []$, $\boldsymbol{u} \leftarrow []$, $\mathbf{e} \leftarrow []$

and the values that they will record are characterized by Algorithm 4, corresponding to the online Algorithm 2, with additional assignment instructions recording the values that we are interested in for the analysis. Let us analyze how the tracking variables are filled:

At each iteration k:

- The value of the variable "k" clearly is k, the number of past iterations.
- The k-th coordinate of "$\mathbf{D}^{\boldsymbol{M}}$", i.e. "$\mathbf{D}^{\boldsymbol{M}}[k]$", is assigned the value of "D" at the start of the iteration. Thus, $\mathbf{D}$ stores the sequence of values of "D" at the start of each iteration.
- The k-th coordinate of "$\mathbf{a}$", i.e. "$\mathbf{a}[k]$", is assigned the value $(p_k, \phi(\mathbf{f}_k))$, i.e. the price p_k and transformed features $\phi(\mathbf{f}_k)$ associated to the ad $(p_k, \mathbf{f}_k)$. Thus, $\mathbf{a}$ stores the sequence of price-(transformed features) pairs for each ad.

Algorithm 4 Online algorithm with tracking variables.

$\mathbf{D}^{\boldsymbol{M}}[k] \leftarrow D$, $\boldsymbol{d}[k] \leftarrow -1$, $\mathbf{N}^{\boldsymbol{U}}[k] \leftarrow \mathbf{N}^{\boldsymbol{U}}[k-1]$
$a \leftarrow (p, \phi(f))$
$\mathbf{a}[k] \leftarrow a$
if $u \cdot a > -e$ **then**
  $c \leftarrow display(p, f)$
  $\boldsymbol{d}[k] \leftarrow 1$
  if $u \cdot a < e$ **then**
    $D \leftarrow D \cup \{(a, c)\}$
  end if
end if
$N \leftarrow N + 1$
if $N = 2^n$ **then**
  $\mathbf{D}^{\boldsymbol{U}}[n] \leftarrow D$
  $(u, e) \leftarrow update(u, e, D)$,
  $\boldsymbol{u}[n] \leftarrow u$, $\mathbf{e}[n] \leftarrow e$, $\mathbf{N}^{\boldsymbol{U}}[k] \leftarrow \mathbf{N}^{\boldsymbol{U}}[k-1] + 1$
  $D \leftarrow \emptyset$, $N \leftarrow 0$, $n \leftarrow n + 1$
end if
$k \leftarrow k + 1$

- The k-th coordinate of "$\boldsymbol{d}$", i.e. "$\boldsymbol{d}[k]$", is by default -1, but is overwritten with the value 1 when "$display(p, f)$" is called. Thus, $\boldsymbol{d}$ stores the sequence of displaying decisions for each ad.

For $n \in \mathbb{N}$, at the n-th iteration where the "$update(u, e, D)$" is called:

- The n-th coordinate of "$\mathbf{D}^{\boldsymbol{U}}$", i.e. "$\mathbf{D}^{\boldsymbol{U}}[n]$", is assigned the value of "D" just before the call "$update(u, e, D)$". Thus, $\mathbf{D}^{\boldsymbol{U}}$ stores the sequence of values of "D" in the successive calls "$update(u, e, D)$".

- The n-th coordinate of "$\boldsymbol{u}$", i.e. "$\boldsymbol{u}[n]$", is assigned the value of "u" just after the call "$update(u, e, D)$". Thus, $\boldsymbol{u}$ stores the sequence of values of "u" after each "$update(u, e, D)$" call.

- The n-th coordinate of "$\mathbf{e}$", i.e. "$\mathbf{e}[n]$", is assigned the value of "e" just after the call "$update(u, e, D)$". Thus, $\mathbf{e}$ stores the sequence of values of "e" after each "$update(u, e, D)$" call.

- The n-th coordinate of "$\mathbf{N}^{\boldsymbol{U}}$", i.e. "$\mathbf{N}^{\boldsymbol{U}}[n]$", is incremented at each "$update(u, e, D)$" call. Thus, $\mathbf{N}^{\boldsymbol{U}}$ stores the number of past "$update(u, e, D)$" calls.

These tracking variables do not change the actual behavior of the algorithm, but will be very useful to its study. Let us now mathematically characterize the content of these tracking variables.

6.4 Preparing the mathematical analysis

6.4.1 Mathematical characterization of the. tracking variables

In this section, we define sequences $(d_k)_{k\in\mathbb{N}}$, $(D_k^{\boldsymbol{M}})_{k\in\mathbb{N}}$, $(N_k^{\boldsymbol{U}})_{k\in\mathbb{N}}$, $(D_k^{\boldsymbol{U}})_{k\in\mathbb{N}}$, $(a_k)_{k\in\mathbb{N}}$, $(u_k)_{k\in\mathbb{N}}$, and $(e_k)_{k\in\mathbb{N}}$ representing the content of the lists "$\boldsymbol{d}$", "$\mathbf{D}^{\boldsymbol{M}}$", "$\mathbf{N}^{\boldsymbol{U}}$", "$\mathbf{D}^{\boldsymbol{U}}$", "$\mathbf{a}$", "$\boldsymbol{u}$", "$\mathbf{e}$" once they are filled.

A careful look at Algorithm 4 allows to see that, by definition:

- $N_k^{\boldsymbol{U}} = \lfloor \log_2(k) \rfloor$ and $a_k = (p_k, \phi(f_k))$ for all $k \in \mathbb{N}$.
- We have

$$\begin{aligned} (u_0, e_0) &= (0, +\infty), \quad D_0^{\boldsymbol{U}} = \emptyset \\ D_n^{\boldsymbol{U}} &= \{(a_k, c_k) : |u_n \cdot a_k| < e_n, k \in [\![2^n, 2^{n+1}[\![\}, n \in \mathbb{N} \\ (u_{n+1}, e_{n+1}) &= update(u_n, e_n, D_n^{\boldsymbol{U}}) \end{aligned}$$

  Notice that the above relations fully determine $(u_n)_{n\in\mathbb{N}}$, $(D_n^{\boldsymbol{U}})_{n\in\mathbb{N}}$.

- We have, for all $k \in \mathbb{N}$,

$$\begin{aligned} d_k &= \begin{cases} 1 \text{ if } u_{N_k^{\boldsymbol{U}}} \cdot a_k > -e_{N_k^{\boldsymbol{U}}} \\ -1 \text{ else.} \end{cases}, \\ D_k^{\boldsymbol{M}} &= \{(a_N, c_N) : |u_{N_k^{\boldsymbol{U}}} \cdot a_N| < e_{N_k^{\boldsymbol{U}}}, N \in [\![2^{N_k^{\boldsymbol{U}}}, k[\![\}. \end{aligned}$$

These objects being fully determined by the above relations, we shall now take them as the mathematical definitions of $(N_k^{\boldsymbol{U}})_{k\in\mathbb{N}}$, $(a_k)_{k\in\mathbb{N}}$, $(u_n)_{n\in\mathbb{N}}$, $(D_n^{\boldsymbol{U}})_{n\in\mathbb{N}}$, $(d_k)_{k\in\mathbb{N}}$ and $(D_n^{\boldsymbol{M}})_{n\in\mathbb{N}}$.

6.4.2 Measures of efficiency

In this section, we define the measures of efficiency to analyze the algorithm.

Prediction's efficiency

The main principle of the prediction efficiency measure is to count, for all time $k \in \mathbb{N}$, the number of errors made before time k. As, in binary classification, there are two types of errors (false positives and false negatives), we introduce two measures of prediction efficiency.

Definition 6.4.1 (False negative efficiency measure) *The false negative efficiency measure at time $k \in \mathbb{N}$ is defined by*

$$E_k^- = \mathbb{E}\Big[\sum_{i=1}^{k} \mathbf{1}_{d_i=-1, c_i=1}\Big]$$

It is the expected number of false negatives before time k, i.e. the expected number of times when the algorithm did not display an ad that would have led to a click.

Definition 6.4.2 (False positive efficiency measure) *The false positive efficiency measure at time $k \in \mathbb{N}$ is defined by*

$$E_k^+ = \mathbb{E}\Big[\sum_{i=1}^{k} \mathbf{1}_{d_i=1, c_i=-1}\Big]$$

It is the expected number of false positives before time k, i.e. the expected number of times when the algorithm displayed an ad that did not lead to a click.

Memory efficiency

The memory efficiency at time $k \in \mathbb{N}$ represents the expected memory size taken by the algorithm at the k-th iteration. Most of the variables of the algorithm take a constant memory space at each iteration. The only variable that has a variable size is "D". Therefore, we will define the memory efficiency by

$$M_k = \mathbb{E}[\#D_k^{\boldsymbol{M}}], \quad \forall k \in \mathbb{N}.$$

Computational efficiency

The computational efficiency at time k corresponds to the total number of operations (additions, multiplications, comparisons, etc) performed before time k. Most of the operations of the algorithm during an iteration have a constant and very light complexity (it simply consists in a few comparisons, affectations, and incrementations). The operation that is susceptible to require a large number of operations is the call the the *update*

function. The instruction $update(u, e, D)$ performs $2d + 1$ linear programs with a set of $\#D + 2d = \mathcal{O}(\#D)$ linear constraints. It is known that there exists algorithms solving a linear program in *linear* time in the number of *constraints*. We thus consider that the complexity of the call $update(u, e, D)$ is linear in $\#D$. We define the computational complexity before time k by

$$C_k = \mathbb{E}\Big[\sum_{i=1}^{N_k^{\boldsymbol{U}}} \#D_i^{\boldsymbol{U}}\Big]$$

6.5 Main results and interpretations

In this section, we state our main results.

Theorem 6.5.1 (Prediction efficiency) *There exists two constants C_ℓ^p and C_f^p, independent from $\varepsilon \in \mathbb{R}_+$, such that*

$$E_k^- = 0, \quad E_k^+ \leq C_\ell^p \ln(k) + C_f^p \varepsilon k, \quad \forall k \in \mathbb{N}$$

Theorem 6.5.2 (Computational complexity estimation) *There exists a constant C^c such that we have*

$$C_k \leq C_\ell^c \ln(k) + ck, \quad \forall k \in \mathbb{N}$$

Theorem 6.5.3 (Memory efficiency estimation) *There exists a constant C_c, independent from $k \in \mathbb{N}$, such that we have*

$$M_k \leq C_c, \quad \forall k \in \mathbb{N}$$

Let us provide a few interpretations of these results:

- **Prediction efficiency:** an important qualitative result is that the constants do not depend upon ε. This is important because, at fixed ε, the estimation only claims that the prediction error is at most linear in k, but this is true of any classification algorithm: even one constantly making errors will have a linear prediction error. What gives qualitative meaning to our result is that, as C_ℓ^p and C_f^p do not depend upon ε, we have a bound over all ε, even $\varepsilon = 0$, where the error becomes perfectly logarithmic. This means that, provided that we make ε smaller and smaller, the prediction error will actually look more and more logarithmic (which is not the case for an algorithm constantly making errors).

- **Computational complexity:** In this result, the interpretation is that the long term cost will essentially corresponds to the fixed cost of the algorithm, but the cost associated to the function *update* is only logarithmic.
- **Memory efficiency:** The result simply means that a constant averaged memory space is necessary to run the algorithm.

6.6 Proofs

6.6.1 Basic definitions

In this section, we introduce some mathematical definitions relevant to our analysis.

- **Ball in $\mathcal{U}$:** we endow $\mathcal{U}$ with the distance

$$d(u,u') := \sup_{a\in\hat{\mathcal{F}}} |u\cdot a - u'\cdot a|, \quad \forall u,u'\in\mathcal{U}$$

  Given $u\in\mathcal{U}$ and $e\in\bar{\mathbb{R}}_+$, we denote $B(u,e)$ the ball with center u and radius e for this distance, that is

$$B(u,e) = \{u'\in\mathcal{U} : d(u',u) < e\}$$

  Given any subset $\mathcal{U}'\subset\mathcal{U}$, we denote $\mathcal{B}(\mathcal{U}')$ the smallest ball containing $\mathcal{U}'$. We then denote $rad(\mathcal{U}')$ the radius of $\mathcal{B}(\mathcal{U}')$.

- **ε-separating hyperplanes in $\mathcal{U}$:** Given a labelled data set D, the set of hyperplanes ε-separating D, denoted by $\mathcal{H}_\varepsilon(D)$, is the set of elements $u\in\mathcal{U}$ such that $sgn(u\cdot a + c\varepsilon) = c$ for all $(a,c)\in D$.

6.6.2 Mathematical characterization of the *update* function

In this section, we study what the *update* function does from a mathematical viewpoint.

Lemma 6.6.1 (Mathematical characterization of *update*) *For all $u\in\mathcal{U}$, $e\in\bar{\mathbb{R}}_+$, and $D\subset\mathcal{D}$, update(u,e,D) yields the center and radius of the smallest ball containing $B(u,e)\cap\mathcal{H}_{2\varepsilon}(D)$.*

Proof. Let $u'\in\mathcal{U}$ and $e'\in\bar{R}_+$ be such that $B(u,e)\cap\mathcal{H}_{2\varepsilon}(D)\subset B(u',e')$. This, by definition means that for all $u''\in B(u,e)\cap\mathcal{H}_{2\varepsilon}(D)$, we have $u''\subset B(u',e')$. As u'' and u'

are linear, their maximal difference is reached on the vertices of $[0,1]^{D^d}$, i.e. on $\{0,1\}^{D^d}$. Therefore, the property $u'' \in B(u', e')$ is equivalent to

$$\begin{aligned} u''(v) &< u'(v) + e', \quad \forall v \in \{0\} \times \{0,1\}^{D^d} \\ u''(v) &> u'(v) - e', \quad \forall v \in \{0\} \times \{0,1\}^{D^d} \end{aligned}$$

The fact that this must be satisfied for all $u'' \in B(u, e) \cap \mathcal{H}_{2\varepsilon}(D)$ is thus equivalent to have

$$\begin{aligned} u'(v) - e' &< p_v^-(u, e, D), \quad \forall v \in \{0\} \times \{0,1\}^{D^d} \\ u'(v) + e' &> p_v^+(u, e, D), \quad \forall v \in \{0\} \times \{0,1\}^{D^d} \end{aligned}$$

where

$$\begin{aligned} p_v^-(u, e, D) &= \min\{u''(v) : u'' \in B(u, e) \cap \mathcal{H}_{2\varepsilon}(D)\} \\ p_v^+(u, e, D) &= \max\{u''(v) : u'' \in B(u, e) \cap \mathcal{H}_{2\varepsilon}(D)\} \end{aligned}$$

or in other words, by definition of $B(u, e)$ and $\mathcal{H}_{2\varepsilon}(D)$, where $p^{\min}(v, u, e, D)$ (resp. $p^{\max}(v, u, e, D)$) are the minimum (resp. maximum) reached by the linear programs

$$\begin{aligned} \text{Minimize/Maximize} \quad & v \cdot u', \forall u' \in \mathcal{U} \\ \text{subject to:} \quad & u'_1 = -1 \\ & u' \cdot a > -2\varepsilon, \quad \forall a : (a, 1) \in \mathcal{D} \\ & u' \cdot a < 2\varepsilon, \quad \forall a : (a, 0) \in \mathcal{D} \\ & u'(v) > u(v) - e, \quad \forall v \in \{0\} \times \{0,1\}^d \\ & u'(v) < u(v) + e, \quad \forall v \in \{0\} \times \{0,1\}^d \end{aligned}$$

The fact that the function U minimizes e' over all (u', e') such that

$$\begin{aligned} u'_1 &= -1 \\ u'(v) - e' &< p_v^-(u, e, D), \quad \forall v \in \{0\} \times \{0,1\}^d \\ u'(v) + e' &> p_v^+(u, e, D), \quad \forall v \in \{0\} \times \{0,1\}^d \end{aligned}$$

thus exactly corresponds to finding the center and radius of the smallest ball containing $B(u, e) \cap \mathcal{H}_{2\varepsilon}(D)$. □

6.6.3 Study of u_n and e_n

Lemma 6.6.2 *We have $B(\hat{u_\star}, \varepsilon) \subset B(u_n, e_n)$ for all $n \in \mathbb{N}$.*

Proof. We prove this by induction on $n \in \mathbb{N}$. For $n = 0$, as $e_0 = +\infty$, it is clearly true. Assume that the property holds true for some $n \in \mathbb{N}$, and let us prove it for $n+1$. Notice that by definition, we have

$$(u_{n+1}, e_{n+1}) = update(u_n, e_n, D_n^{\boldsymbol{U}})$$

Notice that for all k, the fact that

$$u_\star(a_k) > \varepsilon \Rightarrow c_k = 1 \text{ and } u_\star(a_k) < -\varepsilon \Rightarrow c_k = -1$$

implies that

$$\hat{u_\star} \cdot \hat{a}_k > \varepsilon \Rightarrow c_k = 1 \text{ and } \hat{u_\star} \cdot \hat{a}_k < -\varepsilon \Rightarrow c_k = -1$$

which implies that $B(\hat{u_\star}, \varepsilon) \subset \mathcal{H}_{2\varepsilon}(\{(\hat{a}_k, c_k)\})$. Given that

$$D_n^{\boldsymbol{U}} = \{(\hat{a}_k, c_k) : |u_n(\hat{a}_k)| < e_n, k \in [\![2^n, 2^{n+1}[\![\}, n \in \mathbb{N}$$

we clearly have $B(\hat{u_\star}, \varepsilon) \subset \mathcal{H}_{2\varepsilon}(D_n^{\boldsymbol{U}})$. By induction hypothesis, we have $B(\hat{u_\star}, \varepsilon) \in B(u_n, e_n)$. Thus we have $B(\hat{u_\star}, \varepsilon) \in B(u_n, e_n) \cap \mathcal{H}_{2\varepsilon}(D_n^{\boldsymbol{U}})$, and thus, by the mathematical characterization of $update(u, e, D)$, we have

$$B(\hat{u_\star}, \varepsilon) \in B(u_n, e_n) \cap \mathcal{H}_{2\varepsilon}(D_n^{\boldsymbol{U}}) \subset B(update(u_n, e_n, D_n^{\boldsymbol{U}})) = B(u_{n+1}, e_{n+1}),$$

where we used the mathematical characterization of the *update* function proved in Lemma 6.6.1 which concludes the proof. □

6.6.4 Proof that $E_k^- = 0$

We have

$$\begin{aligned} E_k^- &= \mathbb{E}\Big[\sum_{i=1}^{k} \mathbf{1}_{d_i=-1, c_i=1}\Big] = \mathbb{E}\Big[\sum_{i=1}^{k} \mathbf{1}_{u_{N_i^{\boldsymbol{U}}} \cdot \hat{a}_i < -e_{N_i^{\boldsymbol{U}}}, c_i=1}\Big] \\ &\leq \mathbb{E}\Big[\sum_{i=1}^{k} \mathbf{1}_{\hat{u_\star} \cdot \hat{a}_i < -\varepsilon, c_i=1}\Big] \leq \mathbb{E}\Big[\sum_{i=1}^{k} \mathbf{1}_{u_\star(a_i) < -\varepsilon, c_i=1}\Big] \\ &\leq \mathbb{E}\Big[\sum_{i=1}^{k} \mathbf{1}_{c_i=-1, c_i=1}\Big] = 0 \end{aligned}$$

where the first inequality comes from the fact that $B(u_\star, \varepsilon) \in B(u_n, e_n)$ and the second inequality comes from Assumption (6.2.1). This concludes the proof. □

6.6.5 Proof of the other results

From E_k^+, M_k, and C_k to e_n

In this section, we show how the analysis of E_k^+, M_k, and C_k reduces to the study of e_n.

Lemma 6.6.3 *We have*

$$E_k^+ \leq C \sum_{n \leq \lceil \log_2(k) \rceil} 2^n \mathbb{E}[e_n], \quad M_k \leq k \mathbb{E}[e_{\lceil \log_2(k) \rceil}], \quad C_k \leq C \sum_{n \leq \lceil \log_2(k) \rceil} 2^n \mathbb{E}[e_n]$$

Proof. $\mathbf{E_k^+}$: We have

$$\mathbb{E}[E_k^+] \;=\; \mathbb{E}\Big[\sum_{i=1}^{k} \mathbf{1}_{d_i=1, c_i=-1}\Big]$$

By definition, $d_i = 1$ is equivalent to $u_{N_i^U} \cdot \hat{a}_i > -e_{N_i^U}$. Furthermore, $c_i = -1$ implies that $\hat{u_\star} \cdot \hat{a}_i < \varepsilon$, and thus, as $B(u_\star, \varepsilon) \in B(u_{N_i^U}, e_{N_i^U})$, it implies that $u_{N_i^U} \cdot \hat{a}_i < e_{N_i^U}$. Consequently, we have

$$\mathbf{1}_{d_i=1, c_i=-1} \leq \mathbf{1}_{|u_{N_i^U} \cdot \hat{a}_i| < e_{N_i^U}}$$

By a simple conditioning, we have

$$\mathbb{E}[\mathbf{1}_{|u_{N_i^U} \cdot \hat{a}_i| < e_{N_i^U}}] \;=\; \mathbb{E}[\mathbb{P}(\mathbf{1}_{|u \cdot \hat{a}_i| < e})_{u = u_{N_i^U}, e = e_{N_i^U}}].$$

As we assumed that (f_1, p_1) is a diffuse probability distribution with density bounded from above by C, we have

$$\mathbb{P}(\mathbf{1}_{|u \cdot \hat{a}_i| < e}) = \mathbb{P}(\mathbf{1}_{|\check{u}(a_i)| < e}) \leq C \lambda(\{\check{u}^{-1}([-e, e])\}) \leq Ce$$

and thus

$$\mathbb{E}[\mathbb{P}(\mathbf{1}_{|u \cdot a_i| < e})_{u = u_{N_i^U}, e = e_{N_i^U}}] \;\leq\; C \mathbb{E}[e_{N_i^U}]$$

Thus, we have

$$\mathbb{E}[E_k^+] \leq C \sum_{i=1}^{k} \mathbb{E}[e_{N_i^U}] \leq C \sum_{i=1}^{k} \mathbb{E}[e_{\lfloor \log_2(i) \rfloor}] \leq C \sum_{n \leq \lceil \log_2(k) \rceil} 2^n \mathbb{E}[e_n]$$

which concludes the proof.

$\mathbf{M_k}$: We have

$$\begin{aligned} M_k &= \mathbb{E}[\#D_k^{\boldsymbol{M}}] \leq \mathbb{E}[\#D^{\boldsymbol{U}}_{\lfloor \log_2(k) \rfloor}] = \mathbb{E}\Big[\sum_{N=2^{\lfloor \log_2(k) \rfloor}}^{2^{\lfloor \log_2(k) \rfloor+1}-1} \mathbf{1}_{|u_{\lfloor \log_2(k) \rfloor} \cdot a_N| < e_{\lfloor \log_2(k) \rfloor}} \Big] \\ &= 2^{\lfloor \log_2(k) \rfloor} \mathbb{P}(|u_{\lfloor \log_2(k) \rfloor} \cdot a_k| < e_{\lfloor \log_2(k) \rfloor}) \leq k \mathbb{P}(|u_{\lfloor \log_2(k) \rfloor} \cdot a_k| < e_{\lfloor \log_2(k) \rfloor}) \end{aligned}$$

We conclude by the same conditioning argument as in the first step.
$\mathbf{C_k}$: We have $C_k = \mathbb{E}\Big[\sum_{i=1}^{N_k^{\boldsymbol{U}}} \#D_i^{\boldsymbol{U}} \Big]$, where, by definition of $D_i^{\boldsymbol{U}}$, we have $\#D_i^{\boldsymbol{U}} = \sum_{k=2^i}^{2^{i+1}-1} \mathbf{1}_{|u_i(f_k)-p_k|<e_i}$. We again conclude by the same conditioning argument as in the first step. This concludes the proof. □

From e_k to r_k

For the sequel, we define the labelled data set sequence $(\mathcal{D}_k)_{k\in\mathbb{N}}$ by

$$\mathcal{D}_k = \{(a_i, sgn(\hat{u}_\star \cdot a_i) : |\hat{u}_\star \cdot a_i| \geq \varepsilon, i \leq k\}, \quad \forall k \in \mathbb{N}$$

and the sequence $(r_k)_{k\in\mathbb{N}}$ by

$$r_k = rad(\mathcal{H}_{2\varepsilon}(\mathcal{D}_k)), \quad \forall k \in \mathbb{N}.$$

The following result illustrate the utility of $(r_k)_{k\in\mathbb{N}}$.

Lemma 6.6.4 *We have*

$$\mathbb{E}[e_n] \leq \mathbb{E}[r_{2^n}]$$

Proof. For all $n \in \mathbb{N}$, as we have $B(u_\star, \varepsilon) \subset B(u_n, e_n)$, it is clear that for all $k \in [2^n, 2^{n+1}]$, if $|u_n \cdot \hat{a}_k| > e_n$ then $|u_\star(a_k)| > \varepsilon$, which implies that

$$sgn(u_n \cdot \hat{a}_k) = sgn(u_\star(a_k)) = c_k,$$

where the last identity comes from Assumption (6.2.1). In this case, we thus have $B(u_n, e_n) \subset \mathcal{H}(\{(\hat{a}_k, c_k)\})$ and thus also $B(u_n, e_n) \subset \mathcal{H}_{2\varepsilon}(\{(\hat{a}_k, c_k)\})$. Thus we have

$$B(u_n, e_n) \subset \mathcal{H}_{2\varepsilon}(\{(\hat{a}_k, c_k) : |u_n \cdot \hat{a}_k| > e_n, k \in [\![2^n, 2^{n+1}[\![\})$$

Recall that by definition, we have

$$D_n^{\boldsymbol{U}} = \{(\hat{a}_k, c_k) : |u_n \cdot \hat{a}_k| \leq e_n, k \in [\![2^n, 2^{n+1}[\![\}$$

and thus we have

$$B(u_n, e_n) \cap \mathcal{H}_{2\varepsilon}(D_n^{\boldsymbol{U}}) \subset \mathcal{H}_{2\varepsilon}(\{(\hat{a}_k, c_k) : k \in [\![2^n, 2^{n+1}[\![\})$$

Notice that for any $k \in [\![2^n, 2^{n+1}[\![$ such that $|u_\star(a_k)| > \varepsilon$, as, by Assumption (6.2.1), we have $c_k = sgn(u_\star(a_k))$, we clearly have

$$\tilde{\mathcal{D}}_n \subset \{(\hat{a}_k, c_k) : k \in [\![2^n, 2^{n+1}[\![\}$$

where

$$\tilde{\mathcal{D}}_n := \{(\hat{a}_k, sgn(u_\star(a_k))) : |u_\star(a_k)| > \varepsilon, k \in [\![2^n, 2^{n+1}[\![\}$$

Therefore, we clearly have

$$\mathcal{H}_{2\varepsilon}(\{(\hat{a}_k, c_k) : k \in [\![2^n, 2^{n+1}[\![\}) \subset \mathcal{H}_{2\varepsilon}(\tilde{\mathcal{D}}_n)$$

and thus

$$B(u_n, e_n) \cap \mathcal{H}_{2\varepsilon}(D_n^{\boldsymbol{U}}) \subset \mathcal{H}_{2\varepsilon}(\tilde{\mathcal{D}}_n)$$

Which implies, by definition of e_{n+1} and by the mathematical characterization of the *update* function in Lemma 6.6.1, that

$$e_{n+1} \leq rad(\mathcal{H}_{2\varepsilon}(\tilde{\mathcal{D}}_n))$$

Because the sequence $(a_k)_{k\in\mathbb{N}}$ is assumed to be i.i.d., we thus have

$$\mathbb{E}[e_{n+1}] \leq \mathbb{E}[rad(\mathcal{H}_{2\varepsilon}(\tilde{\mathcal{D}}_n))] = \mathbb{E}[rad(\mathcal{H}_{2\varepsilon}(\mathcal{D}_{2^n}))] = \mathbb{E}[r_{2^n}].$$

This concludes the proof. □

Preparing the probabilistic analysis

We have the following result.

Lemma 6.6.5 *There exists constants C, C', such that for all $u, u' \in \mathcal{U}$, we have*

$$d(u, u') \leq C(\|\check{u} - \check{u'}\| + \|\nabla(\check{u} - \check{u'})\|) \leq C' \max_{v \in \frac{1}{D}\{0,D\}^d} |(\check{u} - \check{u'})(v)|$$

Proof. This comes from the fact that $d(u, u') = \|u - u'\|_{\hat{\mathcal{F}}}$ is the norm of $u - u' : \hat{\mathcal{F}} \to \mathbb{R}$, which yields a pull-back norm of $\check{u} - \check{u'} : \mathcal{F} \to \mathbb{R}$, which is a finite-dimensional space, and thus in which all norms are equivalent. The middle term is clearly a norm, and

the third term is a norm: the only not completely trivial part is the definite positive aspect. It is easy to build D^d multi-dimensional Lagrange polynomials forming a basis of $\mathcal{U}$, indexed by $v \in \frac{1}{D}\{0,D\}^d$, each cancelling on all $v \in \frac{1}{D}\{0,D\}^d$ but one. In such basis, the coordinate of any polynomial is its values in $v \in \frac{1}{D}\{0,D\}^d$. Therefore, if they all cancel, then the polynomial is null. This proves the definite positive aspect. □

Lemma 6.6.6 *We have*

$$\|u\| \leq \frac{C'_{d,D}}{1-\delta C_{d,D}} \max_{v\in\frac{1}{D}\{0,D\}^d} \max(\min_{\|f-v\|\leq\delta} \check{u}(f), \min_{\|f-v\|\leq\delta}(-\check{u}(f)))$$

Proof. We can write $\check{u}(f) \geq \check{u}(v) - \|\nabla\check{u}\|\|f-v\|$, which implies $\check{u}(v) \leq \min_{\|f-v\|\leq\delta} \check{u}(f) + \|\nabla\check{u}\|\delta$. Likewise, we have

$$\check{u}(f) \leq \check{u}(v) + \|\nabla\check{u}\|\|f-v\|$$

and thus $\check{u}(v) \geq \max_{\|f-v\|\leq\delta} \check{u}(f) - \|\nabla\check{u}\|\delta$ which can be rewritten

$$-\check{u}(v) \leq \min_{\|f-v\|\leq\delta}(-\check{u}(f)) + \|\nabla\check{u}\|\delta$$

By combining these formulas, we obtain

$$|\check{u}(v)| \leq \max(\min_{\|f-v\|\leq\delta} \check{u}(f), \min_{\|f-v\|\leq\delta}(-\check{u}(f))) + \|\nabla\check{u}\|\delta$$

Taking the sup, we get

$$\max_{v\in\frac{1}{D}\{0,D\}^d} |\check{u}(v)| \leq \max_{v\in\frac{1}{D}\{0,D\}^d} \max(\min_{\|f-v\|\leq\delta} \check{u}(f), \min_{\|f-v\|\leq\delta}(-\check{u}(f))) + \delta C_{d,D} \max_{v\in\frac{1}{D}\{0,D\}^d} |\check{u}(v)|$$

and thus

$$\max_{v\in\frac{1}{D}\{0,D\}^d} |\check{u}(v)| \leq \frac{1}{1-\delta C_{d,D}} \max_{v\in\frac{1}{D}\{0,D\}^d} \max(\min_{\|f-v\|\leq\delta} \check{u}(f), \min_{\|f-v\|\leq\delta}(-\check{u}(f)))$$

Finally, we obtain that

$$\|u\| \leq \frac{C'_{d,D}}{1-\delta C_{d,D}} \max_{v\in\frac{1}{D}\{0,D\}^d} \max(\min_{\|f-v\|\leq\delta} \check{u}(f), \min_{\|f-v\|\leq\delta}(-\check{u}(f)))$$

which concludes the proof. □

Therefore, for any $u \in \mathcal{H}_{2\varepsilon}(\mathcal{D}_k)$, we have

$$\begin{aligned}
\|u-u_\star\| &\leq \frac{C'_{d,D}}{1-\delta C_{d,D}} \max_{v\in\frac{1}{D}\{0,D\}^d} \max\Big(\min_{\|f-v\|\leq\delta}((\hat{u}-u_\star)(f)), \min_{\|f-v\|\leq\delta}((u_\star-\hat{u})(f))\Big) \\
&\leq \frac{C'_{d,D}}{1-\delta C_{d,D}} \max_{v\in\frac{1}{D}\{0,D\}^d} \max\Big(\min_{(a,-1)\in\mathcal{D}_k,\|f-v\|\leq\delta}((\hat{u}-u_\star)(f)), \min_{(a,1)\in\mathcal{D}_k,\|f-v\|\leq\delta}((u_\star-\hat{u})(f))\Big) \\
&\leq \frac{C'_{d,D}}{1-\delta C_{d,D}} \max_{v\in\frac{1}{D}\{0,D\}^d} \max\Big(\min_{u_\star(a_k)<-\varepsilon,\|f-v\|\leq\delta}(2\varepsilon-u_\star(a)), \min_{u_\star(a_k)>\varepsilon,\|f-v\|\leq\delta}(u_\star(a)+2\varepsilon)\Big) \\
&\leq K_{d,D}\varepsilon + K_{d,D} \max_{v\in\frac{1}{D}\{0,D\}^d} \max\Big(\min_{u_\star(a_k)<-\varepsilon,\|f-v\|\leq\delta}|u_\star(a)|, \min_{u_\star(a_k)>\varepsilon,\|f-v\|\leq\delta}|u_\star(a)|\Big) \\
&\leq K_{d,D}\varepsilon + K'_{d,D} \sum_{v\in\frac{1}{D}\{0,D\}^d} \left(\min_{u_\star(a_k)<-\varepsilon,\|f_k-v\|\leq\delta}|u_\star(a)| + \min_{u_\star(a_k)>\varepsilon,\|f_k-v\|\leq\delta}|u_\star(a)|\right)
\end{aligned}$$

Probabilistic study

The probabilistic study of the problem relies on arguments of extreme values and records. This is natural since, in the deterministic study, we essentially made estimations based on minimums of families of real numbers, and once we put back probability in the framework, it natural turns into studying the probability distribution of a record. A core computation of record theory is the following:

$$\mathbb{E}[\min_{m\leq N} U_m] = \int_0^1 \mathbb{P}(\min_{m\leq N} U_m > x)dx = \int_0^1 (1-x)^N dx = \int_0^1 x^N dx = \frac{1}{N+1}$$

if $(U_m)_{m\leq N}$ are i.i.d. uniform random variables.

We shall now do the step 2) of our approach, to prove that falling into this band happens with small probability. Notice that the width of the band from previous Lemma, expressed with minimums and maximums, is clearly designed in a way that should facilitate the use of extreme value theory and records arguments to estimate the probability to fall into this band. However, as the situation is more complex than the introductory example, we have to design a more powerful way to do such type of estimations for our needs.

Lemma 6.6.7 *Let $(X_i, Y_i)_{i\leq n}$ be a family of i.i.d. random variables such that $\frac{d\mathcal{L}(Y_1|X_1)}{d\lambda}(y) \geq c > 0$ for all $0 \leq y \leq \frac{1}{C}$, a.s.. Then we have, for all measurable set A,*

$$\mathbb{E}[\min(1, C\min_{i:X_i\in A} Y_i)] \leq \frac{C}{c\mathbb{P}(X_1\in A)(n+1)}$$

Proof. measurable function. We have

$$\begin{aligned}
\mathbb{E}[\min(1, C\min_{i:X_i\in A} Y_i)] &\leq \mathbb{E}[\min(1, C\min_{i:X_i\in A}\min(Y_i, \frac{1}{C}))] \leq C\mathbb{E}[\min_{i:X_i\in A}\min(Y_i, \frac{1}{C})] \\
&\leq C\mathbb{E}[\min_{i\leq n}(\min(Y_i, \frac{1}{C})\mathbf{1}_{X_i\in A} + \frac{1}{C}\mathbf{1}_{X_i\notin A})] \\
&\leq C\int_0^\infty \mathbb{P}(\min(Y_i, \frac{1}{C})\mathbf{1}_{X_i\in A} + \frac{1}{C}\mathbf{1}_{X_i\notin A} \geq x)^n dx \\
&\leq C\int_0^{\frac{1}{C}} \mathbb{P}(\min(Y_i, \frac{1}{C})\mathbf{1}_{X_i\in A} + \frac{1}{C}\mathbf{1}_{X_i\notin A} \geq x)^n dx \\
&\leq C\int_0^{\frac{1}{C}} (1 - \mathbb{P}(X_1\in A) + \mathbb{P}(X_1\in A)(1-cx))^n dx \\
&\leq C\int_0^{\frac{1}{C}} (1 - cx\mathbb{P}(X_1\in A))^n dx \\
&\leq \frac{C}{c\mathbb{P}(X_1\in A)(n+1)}
\end{aligned}$$

□

Theorem 6.6.1 *We have*

$$\mathbb{P}(\mathcal{C}_{\mathcal{U},D_n}(f_{n+1}, p_{n+1}) = 0) \leq \frac{1}{N}$$

Proof. We have

$$\begin{aligned}
&\mathbb{P}(\mathcal{C}_{\mathcal{U},D_n}(f_{n+1}, p_{n+1}) = 0) \leq \mathbb{P}(|p_{n+1} - u(f_{n+1}| \leq \max_{w\in\mathcal{U}_{D_n}}\max_{v'\in[0,1]^d} |(u-w)(v)|) \\
\leq\ & \mathbb{P}(|p_{n+1} - u(f_{n+1})| \\
\leq\ & 6\max_{v\in\{0,1\}^d}\max(\min_{f_n\in\mathcal{B}(v,\frac{1}{3}),p_n<u(f_n)}(u(f_n)-p_n), \min_{f_n\in\mathcal{B}(v,\frac{1}{3}):p_n>u(f_n)}(p_n - u(f_n)))) \\
\leq\ & \mathbb{E}[C6\max_{v\in\{0,1\}^d}\max(\min_{f_n\in\mathcal{B}(v,\frac{1}{3}),p_n<u(f_n)}(u(f_n)-p_n), \min_{f_n\in\mathcal{B}(v,\frac{1}{3}):p_n>u(f_n)}(p_n - u(f_n)))] \\
\leq\ & 6C\sum_{v\in\{0,1\}^d}(\mathbb{E}[\min_{f_n\in\mathcal{B}(v,\frac{1}{3}),p_n<u(f_n)}(u(f_n)-p_n)] + \mathbb{E}[\min_{f_n\in\mathcal{B}(v,\frac{1}{3}):p_n>u(f_n)}(p_n - u(f_n)))] \\
\leq\ & 6C\sum_{v\in\{0,1\}^d}(\varepsilon + \mathbb{E}[\min_{f_n\in\mathcal{B}(v,\frac{1}{3}),p_n<u(f_n)-\varepsilon}(u(f_n)-p_n-\varepsilon)] \\
&+\mathbb{E}[\min_{f_n\in\mathcal{B}(v,\frac{1}{3}):p_n>u(f_n)+\varepsilon}(p_n - u(f_n) - \varepsilon))]
\end{aligned}$$

We can now apply the Lemma with $Y_i = u(f_i) - p_i$ and $X_i = (f_i, sgn(u(f_i) - p_i))$. □

Tracing back the estimations to prove the results

Before rigorously proving and adapting this argument in next sections, let us see how such estimation of height of the thinnest predicting band would allow us to conclude: provided that this estimation is rigorously proven, we would indeed have

$$\mathbb{E}[E_k] \leq C \sum_{n \leq n_k} 2^n \mathbb{E}[H_{2^n}] \leq C \sum_{n \leq n_k} 2^n (\frac{1}{2^n} + \varepsilon) = n_k + 2^{n_k+1}\varepsilon$$

However, notice that

$$k \geq \sum_{m=1}^{2^{n_k-1}} m = 2^{n_k} - 1$$

and thus $k + 1 \geq 2^{n_k}$, and $n_k \leq \ln(k+1)$, hence the (approximate) conclusion:

$$\mathbb{E}[E_k] \leq \ln(k+1) + (k+1)\varepsilon$$

Likewise, we would also have

$$\mathbb{E}[(\#(D_k))^m] \leq C_m(2^{mn_k}\mathbb{E}[\mathbf{e}^m_{2^{n_k-1}}]) \leq C_m K(1 + 2^{mn_k}\varepsilon^m)$$

After last update, such that $2^{n_k}\varepsilon = 1$, the labeled data set will be empty.

6.7 Conclusion

In this chapter, we have studied an online classification algorithm specifically designed for targeted advertising. The particularities of the problem were 1) Error asymmetry, in the sense that one error type had to be completely avoided, and the other one only minimized, and 2) Feedback asymmetry, i.e. the clicking intention of an individual is, afterward, only revealed when the ad has been displayed to him. Despite these constraints, we have proved that our online algorithm performs as efficiently as benchmark algorithms without these constraints, i.e. we obtain a number of prediction errors given k past predictions that has a logarithmic component and a linear component proportional to the problem's error margin ε. An important qualitative aspect is that the logarithmic component is independent from the linear component and applies even to the case of sharp separability, i.e. $\varepsilon = 0$ (where the linear component completely cancels out). In this work, we have also provided a description of our algorithm that is as close as possible to a concrete implementation, via the use of pseudo-language and linear programming. Besides the prediction efficiency, we have precisely studied both the memory usage and

computational complexity and proved that the algorithm was also efficient in these regards. Although the algorithm was designed with targeted advertising in mind, other standard applications of classification can clearly be found, e.g. in finance, to predict the evolution of a stock's price or to predict the trading decisions of an individual, or in medicine, e.g. to design medical tests.

Chapter 7

Optimal bidding strategies for advertising auctions

Abstract. In this work, we introduce and study several optimal control models of targeted advertising with auctions. Each model focuses on a different type of advertising, namely, commercial advertising for triggering purchases or subscriptions, and social marketing for sensitizing people about unhealthy behaviors (anti-drug, road-safety campaigns). All our models are based on a common framework encoding people's online behaviors and the targeted advertising auction mechanism widely used on Internet. Our main result is to provide semi-explicit formulas for each problem's optimal value and optimal bidding policy. Thanks to these formulas, we are able to draw interpretations about how phenomenons like people's online behaviors and social interactions affect the optimal bid to make for targeted advertising auctions. We also study how to efficiently combine targeted advertising and non-targeted advertising mechanism. We conclude by providing some classes of examples with fully explicit formulas.

7.1 Introduction

Through the emergence of new online channels and information technology, targeted advertising plays a growing role in our society and progressively replaces traditional forms of advertising like newspapers, billboards, etc. Indeed, companies can minimize wasted advertising costs by targeting directly individuals that are potentially interested by the product the advertiser is promoting. Modern targeted media use historical data on internet (cookies) such as tracking online or mobile web activities of consumers.

Optimal control is a suitable mathematical tool for studying advertising problems. There is already a vast literature on optimal control for advertising. Several approaches have been proposed in the past: mathematical programming, dynamic programming,

simulation, and heuristic procedures ([56, 98]). Optimal control theory was an important addition. In the classical approach, a dynamical system is modeled with controlled differential equations and optimized by means of the maximum principle. We mention the important Nerlove-Arrow ([70]), and Vidale-Wolfe ([89]) models, and for an overview of this research field up to the 90s, see [78] and its sequel [27]. We also mention two more recent works, optimal advertising with delay, studied by Gozzi and Marinelli ([31]), and Jack, Johnson and Zervos ([39]) on control applied to the goodwill problem. This existing literature about optimal control for advertising are differential models, considering, from the start, controlled differential equations directly modeling the dynamics of sales as a continuous process affected by an advertising expenditures process.

In this work, we study optimal control models for advertising strategies, and the main novelty is to consider individual or *individual or population models* aiming to model the behavior of an individual in a finite population, and describing how advertising will affect their states. The important upside of explicit population modeling is the integration of individual's behaviors, which is more natural and compelling than modeling an abstract object like the sales process, as the solution to a given differential equation. Indeed, while the latter approach requires to make a leap of faith to admit that such differential equation well describes the sales process dynamic, therefore making abstraction of the complex underlying mechanisms generating such sales process, the former approach defines the model at a more "atomic" level, that we can easily understand and be convinced by without too much effort or abstraction. Another important consequence of population models is that they also allow to model the world in which individuals live, and its rules, in a more explicit and concrete way. This, in particular, allows us to encode the feature of auctions for targeted advertising in our models. As they are a crucial component of online advertising, it would seem unreasonable to ignore them.

Auctions, in targeted advertising, are used to determine which company will have its ad displayed to a given individual. Each time an individual connects to a website using targeted advertising, several agents (companies, influencers, etc) compete in an auction for the *ad emplacement*. The agents make bids, the winner pays a price depending upon the auction mechanism, and his ad is displayed to the individual. The long history of auctions, starting from the groundbreaking works of John Nash ([66]) and later William S. Vickrey ([88]), and their omnipresence on the Internet, illustrate the crucial importance of auction theory, also evidenced by the 2020 Nobel prize in economics, Milgrom and Wilson, for their contribution to auction theory.

The output of auctions can be quite challenging to predict even in simple frameworks, and as the overall framework of our models is complex, we won't model each bidding company (which would turn our optimal control models into games) and instead assume that at each targeted advertising auction, the maximal bid from companies other than

our agent is a random variable independent from the past, and identically distributed across auctions. This assumption has the practical advantage to keep the control problem tractable.

Finally, one of our models will also encode social interactions allowing individuals who saw the ad to become themselves vectors of information. Again, our modeling of social interactions will be quite simple and symmetric, to keep the problem tractable. For an interesting and detailed overview of information spreading models in populations, we refer to [2].

We shall study two types of models, based on two types of usage of advertising:

1. **Commercial advertising**, modeling situations where informing an individual triggers a reward for the agent, which is thus particularly well suited for commercial advertising. We shall consider two types of rewards: *purchase-based reward*, modeling situations where the information triggers a purchase and thus a punctual payment from the individual to the agent, and *subscription-based reward*, modeling cases where the information triggers a subscription of the individual to a service proposed by the agent, and thus pays a regular fee to the agent.

2. **Social marketing**, modeling situations where informing an individual cancels a cost continuously perceived by the agent. In this model, each individual, as long as he is not informed, is considered to incur a continuous cost to the agent, and only when such individual gets informed, the cost stops. This model is particularly well suited for social marketing where the agent's goal is not to make profit but instead to change people's behaviors and promoting social change by sensitizing them about dangers. Classical social marketing campaigns are anti-drugs campaigns, road-safety campaigns, sexual-safety campaigns, low-fat diet campaigns, etc. From the agent's viewpoint, in such campaign, any individual who is not behaving safely is considered to represent a continuous cost to him.

Besides the different nature of their applications, the both aforementioned studies also importantly differ in their goals. On one hand, commercial targeted advertising is already widely spread on the Internet, and in this case, our study simply proposes a model that could potentially improve a company's bidding strategies. On the other hand, social marketing does currently not seem to use targeted advertising a lot, instead relying more on classical non-targeted advertising, and in this case, our model proposes a way to combine non-targeted advertising with targeted advertising for any organization or association using social marketing.

Our main contributions. Our first and main contribution, in this work, is to propose four advertising models, based on a common core framework explicitly modeling

individuals online behaviors and advertising auctions, each designed for various types of advertising, and to obtain for every model a semi-explicit form of the optimal value function and optimal bidding policies.

Our second contribution is to propose in one of these models a rich population model, involving individuals spontaneously finding an information, targeted advertising auctions, non-targeted advertising auctions, and social interactions, while keeping a problem tractable with semi-explicit optimal value and bidding policies.

Our third contribution is to provide classes of examples where the solutions (optimal value and bidding policy) are fully explicit.

By observing the form of the models solutions, we are able to clearly understand how 1) the optimal bid to make in a given targeted advertising auction depends not only upon the distribution of other bidders' maximal bids, but also upon the online behavior of the individual (rates at which he connects to various types of websites), and 2) In the fourth model, involving a population, and adding non-targeted advertising and social interactions in the population, we are able to understand a) how the presence of social interactions impact the optimal bid to make, and b) how the optimal bid to make for non-targeted advertising auctions relates to the optimal bid for targeted advertising auctions and the proportion of already informed people. More generally, this work allows to see how each way an individual can learn an information combine together and affect the optimal bid to make in advertising auctions, and this is our fourth contribution.

The mathematical method used to solve these problems is based on martingale tools, in particular, on techniques involving Poisson processes and their compensators. By means of these tools, we essentially prove the results in two steps: 1) bounding from above (resp. from below) the optimal value when it is a gain (resp. a cost), and then 2) providing a well chosen policy such that the inequalities in 1) become equalities, thus simultaneously proving that the optimal value is equal to its bound, and obtaining an optimal policy reaching it.

Outline of the chapter. We introduce in Section 7.2 the core framework on which each of the four models is based. In Section 7.3, we study two targeted advertising models designed for applications to commercial advertising, the first one modeling advertising to trigger a purchase, the second one modeling advertising to trigger a subscription. In Section 7.4, we study two advertising models applied to social marketing (anti-drug campaigns, road-safety campaigns, etc), the first one with an arbitrary discount factor, the second one with no discounting, but with extra features of non-targeted advertising and social interactions.

7.2 Basic framework

In this section, we introduce the framework on which all the subsequent models are based, simply enriching it in various ways.

The core framework essentially consists in modeling 1) the concept of information for this work, 2) an individual's online behavior, 3) the targeted advertising auction mechanism, 4) a targeted advertising bidding strategy, and finally describe how these four features combine together to determine the dynamic of an individual.

7.2.1 The Information

In this work, all our models will be about some *Information*. We shall denote it with a capital "I" to emphasize that it is a specific piece of information. It could a priori be any information. Let us give a few examples, further discussed in this work. The Information can be:

- the existence of a service (Netflix, Amazon, etc),
- the existence of a product (smartphone, computer),
- the unhealthiness or healthiness of a behavior (drug/alcohol consuming, road safety, sexual safety, etc).

In the various models studied in this paper, each model will naturally correspond to one of these three types of information, but for now, let us simply consider a generic Information.

The main characteristic of the Information is that any individual can either *not know it* or *know it*. In other words, the Information is naturally associated to a binary state for any individual: an individual in state 0 means that he does not know the Information, and an individual in state 1 means that he knows the Information.

7.2.2 The Agent

In our work, the Agent will represent any entity (company, association, etc) desiring to spread the Information.

- In the case of a new service or product, it will naturally be the company proposing this service or selling this product.
- In the case of the unhealthiness or healthiness of a given behavior, it will naturally be an philanthropic association or a governmental entity aiming to work for social good.

The main characteristics of the Agent is that 1) he wants to spread the Information, 2) he has a gain or cost function depending upon how the information spreads, and 3) he will use a targeted advertising strategy as a mean to diffuse the Information.

7.2.3 The Action

The Agent wants to give the Information to individuals to trigger an *Action*. The *Action* depends upon the type of the Information:

- If the Information is about the existence of a service, the expected Action is a *subscription*.
- If the Information is about the existence of a product, the expected Action is a *purchase*.
- If the Information is about an unhealthy behavior, the expected Action is a *healthier behavior*.

In this work, we assume that the Agent knows the individuals well enough to be aware of who would do the Action if they had the Information (who would subscribe to the service if he learns that it exists, buy the product if he learns that it exists, or stop some behavior if he learns that it is unhealthy).

The individuals who would not perform the Action, even informed, are dismissed: the Agent does not try to send them an ad. Therefore, we can assume that the individuals considered in this work are all such that

$$\text{Getting the Information} \Rightarrow \text{Doing the Action}$$

7.2.4 The Individual

Let us start by modelling the general behavior of an individual. Our model is in continuous time. An individual is associated to some random times when he does the following possible actions:

- Spontaneously connect to a website providing the Information. Websites intrinsically providing the information are numerous, depending upon the kind of information: specialized websites relaying the Information, company/association's own website, etc. Essentially, any website such that the Information is in the actual website's content, as opposed to the other case:
- Visit a website not a priori providing the information, but displaying targeted ads, and thus susceptible to display the Information in a targeted ad, provided that

the agent (company, association, etc) pays for it. Important websites displaying targeted ads typically are social networks and search engines.

An Individual is associated to independent Poisson processes $(N^{\mathbf{I}}, N^{\mathbf{T}})$ with respective intensities $\eta^{\mathbf{I}}$, $\eta^{\mathbf{T}}$. $N^{\mathbf{I}}$ counts the times when the Individual connects to websites intrinsically providing the Information. $N^{\mathbf{T}}$ counts the times when the Individual connects to websites displaying targeted ads.

We shall, in our fourth model, introduce a population with several individuals modeled on this basis, each with their own Poisson processes, independent across individuals.

7.2.5 The targeted advertising auctions

When the Individual connects to websites displaying targeted ads, in reality, many influencers are competing to win the right to display their ads to him. The mechanism used by the website to choose which influencer will display his ad is to make them bid for it. Each influencer has the possibility to propose a bid associated to the Individual's characteristics (intensities of his Poisson processes). This ad emplacement allocation mechanism is what we call *targeted advertising auctions*.

Auctions are complicated to study. They involve several bidders, and are thus part of game theory. The current framework is even more complicated since it is dynamic: an auction is opened each time the Individual connects to a website displaying targeted ads. Our goal is to focus on providing a *strategic* tool to the Agent, and keeping the problem tractable is important in this work.

A good compromise to both take targeted advertising auctions into account while having a strategically solvable problem is to model the maximal bid make by the other bidders (i.e. other than the Agent) as random variables, i.i.d. among auctions. We thus introduce a sequence of i.i.d. real random variables $(B_k^{\mathbf{T}})_{k\in\mathbb{N}}$, such that for $k \in \mathbb{N}$, $B_k^{\mathbf{T}}$ represents the maximal bid of other bidders during the k-th targeted advertising auction of the problem.

7.2.6 The targeted advertising bidding strategies

We now introduce the notion of targeted advertising bidding strategies. In essence, a targeted advertising bidding strategy is simply a real valued process β which depends *at most* from the past, i.e. which cannot depend upon the future, such that at each time $t \in \mathbb{R}_+$, β_t represents the bid that the Agent would make if the Individual connects to a website displaying targeted ads.

To rigorously formalize this, let us introduce the filtration $\mathbb{F} = (\mathcal{F}_t)_{t\in\mathbb{R}_+}$ generated by the processes

$$((N^{\mathbf{I}}, N^{\mathbf{T}}, B^{\mathbf{T}}_{N^{\mathbf{T}}}),$$

i.e. we have, for all $t \in \mathbb{R}_+$,

$$\mathcal{F}_t = \sigma((N^{\mathbf{I}}_s, N^{\mathbf{T}}_s, B^{\mathbf{T}}_{N^{\mathbf{T}}_s})_{s\leq t})$$

$\mathcal{F}_t$ thus represents all the information about event triggered before time t.

The set of open-loop bidding controls, denoted by Π_{OL}, is then the set of real-valued processes β predictable and progressively measurable w.r.t. the filtration $\mathbb{F}$.

7.2.7 Information dynamic, constant bidding, and advertising cost

We can now combine all the pieces of modeling previously introduced to define the *information dynamic* of the Individual, the notion of constant efficient bidding policy, and the advertising cost. Given an open-loop bidding control $\beta \in \Pi_{OL}$, the information dynamic of the Individual is the $\{0,1\}$-valued process X^β satisfying the relation

$$\begin{aligned} X^\beta_0 &= 0 \\ dX^\beta_t &= (1 - X^\beta_{t-})(dN^{\mathbf{I}}_t + \mathbf{1}_{\beta_t \geq B^{\mathbf{T}}_{N^{\mathbf{T}}_t}} dN^{\mathbf{T}}_t) \end{aligned}$$

Let us interpret this dynamic. The individual starts uninformed ($X^\beta_0 = 0$). Once he is informed ($X^\beta_t = 1$), he stays informed (hence the $(1 - X^\beta_{t-})$ part). As long as he is not informed, the remaining part of the dynamic is effective: when the individual connects to a website intrinsically providing the Information, he becomes informed ($dN^{\mathbf{I}}_t$ part). When he connects to a website displaying targeted ads ($dN^{\mathbf{T}}_t$ part), he becomes informed if and only if the Agent's ad is displayed to him, which happens if and only if the Agent wins the auction ($\mathbf{1}_{\beta_t \geq B^{\mathbf{T}}_{N^{\mathbf{T}}_t}}$ part).

Advertising cost: in the subsequent models, the gain or cost function of the agent will be the combination of 1) a component depending upon the information dynamic of the Individual, and 2) an advertising cost component. The component 1) will depend upon the model, but the advertising cost will always have the same form, namely:

$$C(\beta) = \mathbb{E}\Big[\int_0^\infty e^{-\rho t} \mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N^{\mathbf{T}}_t}} \mathbf{c}(\beta_t, B^{\mathbf{T}}_{N^{\mathbf{T}}_t}) dN^{\mathbf{T}}_t)\Big].$$

The interpretation is the following:

- $\rho \in \mathbb{R}_+$ is a discount rate. Usually, discount rate is chosen to be strictly positive in order to avoid infinite rewards or costs. However, in one of our models (the last

one), we will specifically assume $\rho = 0$, and it will be an important assumption to make the problem solvable. We shall see that in this model, infinite rewards/costs will never occur despite this assumption.

- When the Individual connects to a website displaying targeted advertising ($dN_t^{\mathbf{T}}$ part), if the targeted advertising auction is won by the agent ($\mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N_t^{\mathbf{T}}}}$ part), the agent has to pay a price $\mathbf{c}(\beta_t, B^{\mathbf{T}}_{N_t^{\mathbf{T}}})$, where $\mathbf{c} : \mathbb{R}^2 \to \mathbb{R}$ is a function depending upon the paying rule defined by the auction. Let us provide two important examples of such auction rule:

  1. **First-price auctions:** In first-price auctions, the winner of the auction pays his bid, and thus, we have $\mathbf{c}(b, B) = b$.
  2. **Second-price auctions:** In second-price auctions, the winner of the auction pays the *second winning bid*, i.e. the bid that he beat. In this case, we have $\mathbf{c}(b, B) = B$.

Constant efficient bidding policy: A constant efficient bidding policy is a constant $b \in \mathbb{R}$. The efficient constant bidding control $\beta^b \in \Pi_{OL}$ associated to a constant efficient bidding policy is defined by the feedback form constraint $\beta_t^b = (1 - X_{t-}^{\beta^b})b$. It simply models a policy where the Agent makes a constant bid b as long as the Individual is not informed (notice that it would be useless to make a positive bid once he is informed).

We have now introduced all the elements of the core framework. In the sequel, we shall study several advertising problems based on this framework. modeling different types of advertising:

- In Section 7.3, we model commercial advertising problems, i.e. problems where the Agent is a company either trying to sell a service or a product. The common property of both situations is that informing the Individual triggers an Action bringing a *reward* to the company (subscription regular fee, purchase punctual fee).

- In Section 7.4, we model social marketing problems, i.e. problems where the Agent is an association or government trying to alert people about unhealthy behaviors (anti-drug/alcohol campaigns, road-safety campaigns, etc). The particularity of such type of advertising is that informing people does not bring a reward to the Agent, but instead, it *cancels a cost*: as long as an individual has an unhealthy behavior, he incurs a continuous cost to the philanthropic association. Once informed, he behaves healthier and stops incurring such cost.

7.3 Commercial advertising model

In this section, we study models for commercial advertising. The Agent is thus a company trying to maximize its gain. We will study two types of commercial gains: the subscription-based gain, and the purchase-based gain.

7.3.1 Purchase-based gain function

In this section, we study the situation where the Information is the existence of a product, where the Agent is a company selling this product, and where the Action of the Individual, once informed, is to purchase the product. To this end, we consider the following *purchase based* gain function:

$$V(\beta) = \mathbb{E}\Big[\int_0^\infty e^{-\rho t} K dX_t^\beta\Big] - C(\beta), \quad \text{for } \beta \in \Pi_{OL}$$

Let us interpret this gain function. The part $C(\beta)$ is just the advertising cost from the core framework. ρ is still the discount rate. The part $\int_0^\infty e^{-\rho t} K dX_t^\beta$ simply represents a punctual payment K from the Individual to the Agent when he becomes informed (dX_t^β part). This thus naturally models the reward obtained by the Agent when the individual buys the product. Therefore, $V(\beta)$ represents the net profit of the Agent in the situation of selling a product.

We now state the result of this section.

Theorem 7.3.1 *We have*

$$V^\star := \sup_{\beta \in \Pi_{OL}} V(\beta) \;=\; \sup_{b \in \mathbb{R}} V(\beta^b),$$

with

$$V(\beta^b) = \frac{\eta^{\mathbf{I}} K + \eta^{\mathbf{T}} \mathbb{E}[(K - \mathbf{c}(b, B_1^{\mathbf{T}}))\mathbf{1}_{b \geq B_1^{\mathbf{T}}}]}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}} \mathbb{P}(b \geq B_1^{\mathbf{T}})}, \quad \forall b \in \mathbb{R}.$$

Furthermore, any $b_\star \in \mathbb{R}$ such that $b_\star = \operatorname{argmax}_{b \in \mathbb{R}} V(\beta^b)$ yields an optimal constant bid, i.e. an optimal open-loop bid taking the form of a constant efficient bid.

Interpretation. The simplest way to interpret this result is by first understanding the role of ρ. It is well known that a discount rate is mathematically equivalent to a random termination date of the problem following an exponential distribution with parameter ρ. In the above formulas, this is how ρ should be understood. Up to adding this random

termination time, we can thus consider that the problem has no discount rate. Given this interpretation, and assuming that the Agent plays a constant efficient bidding policy b, notice that the inner fraction can be seen as

$$p_{\mathbf{I}}K + p_\rho \times 0 + p_{\mathbf{T}}\mathbb{E}[K - \mathbf{c}(\mathbf{b}, B_1^{\mathbf{T}}) \mid \mathbf{b} \geq B_1^{\mathbf{T}}]$$

where $(p_{\mathbf{I}}, p_\rho, p_{\mathbf{T}})$ are probability weights proportional to $(\eta^{\mathbf{I}}, \rho, \eta^{\mathbf{T}}\mathbb{P}(b \geq B_1^{\mathbf{T}}))$. This expression should be seen as the expected reward of the Agent computed in terms of how the problem terminates:

- The problem terminates with the Individual finding the Information by himself with probability $p_{\mathbf{I}}$, and in this case, the Agent only perceives the reward K.
- The problem terminates with the random termination time we just introduced with probability ρ, and in this case, the Individual has not had the time to be informed: the Agent perceives nothing.
- The problem terminates with the Individual getting informed by seeing the Agent's targeted ad with probability $p_{\mathbf{T}}$, and in this case, the Agent perceives K and pays $\mathbf{c}(b, B_1^{\mathbf{T}})$ because he had to pay the auction's price.

Given that we removed the discount rate of the problem by introducing the random termination time, the Agent's expected reward indeed only depend upon *how* the game terminates, rather than *when*. It is thus natural that the optimal value consists in maximizing the expected reward at termination.

Besides the quantitative aspect of this result, an important qualitative property is that a constant bidding policy is enough to reach the optimal value over all open-loop bidding controls. This is particularly interesting from a model-free viewpoint (reinforcement learning) as it means that one can restrict the search for an optimal strategy to the set of constant bidding policies, which is a reasonably "small" set.

Cost dual viewpoint. Another interesting way to formulate the optimal value and bid is from a *cost viewpoint* (and this is actually how we prove this formula in this work): the idea is to consider the best possible scenario for the Agent, which is that the Individual directly connects to a website containing the information from the very beginning, and then look at the real scenario *relatively* to this best scenario. The real scenario necessarily brings a smaller gain than the best scenario, and thus, it is *as if* the Agent won the best scenario gain but then perceives a cost corresponding to the

difference. From this viewpoint, the goal is to minimize this cost. The best scenario gain clearly is K. This yields the following formulas: we have

$$\sup_{\beta\in\Pi_{OL}} V(\beta) \quad = \quad K - \inf_{b\in\mathbb{R}} \frac{\rho K + \eta^{\mathbf{T}}\mathbb{E}[\mathbf{c}(b, B_1^{\mathbf{T}})\mathbf{1}_{b\geq B_1^{\mathbf{T}}}]}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}}\mathbb{P}(b \geq B_1^{\mathbf{T}})},$$

and any $b_\star \in \mathbb{R}$ such that

$$b_\star \quad = \quad \operatorname*{argmin}_{b\in\mathbb{R}} \frac{\rho K + \eta^{\mathbf{T}}\mathbb{E}[\mathbf{c}(b, B_1^{\mathbf{T}})\mathbf{1}_{b\geq B_1^{\mathbf{T}}}]}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}}\mathbb{P}(b \geq B_1^{\mathbf{T}})}$$

yields an optimal constant bid.

Sensitivity of optimal bid to parameters. We start with a useful equivalence.

Lemma 7.3.1 *We have for all $\lambda_1, \lambda_2, \lambda_2', \lambda_3, \lambda_4, \lambda_4' \in \mathbb{R}_+$ such that $\lambda_3 > 0$ and $\lambda_4 < \lambda_4'$, we have*

$$\frac{\lambda_1 + \lambda_2}{\lambda_3 + \lambda_4} < \frac{\lambda_1 + \lambda_2'}{\lambda_3 + \lambda_4'} \Leftrightarrow \frac{\lambda_2' - \lambda_2}{\lambda_4' - \lambda_4} < \frac{\lambda_1}{\lambda_3} \tag{7.3.1}$$

Proof. We have

$$\begin{aligned}\frac{\lambda_1 + \lambda_2}{\lambda_3 + \lambda_4} > \frac{\lambda_1 + \lambda_2'}{\lambda_3 + \lambda_4'} \quad &\Leftrightarrow \quad (\lambda_3 + \lambda_4')(\lambda_1 + \lambda_2) > (\lambda_3 + \lambda_4)(\lambda_1 + \lambda_2') \Leftrightarrow \lambda_3\lambda_2 + \lambda_4'\lambda_1 > \lambda_3\lambda_2' + \lambda_4\lambda_1 \\ &\Leftrightarrow \quad (\lambda_4' - \lambda_4)\lambda_1 > \lambda_3(\lambda_2' - \lambda_2) \Leftrightarrow \frac{\lambda_1}{\lambda_3} > \frac{\lambda_2' - \lambda_2}{\lambda_4' - \lambda_4}\end{aligned}$$

□

Let us now consider constant bids $b < b'$ so that

$$\lambda_4 := \eta^{\mathbf{T}}\mathbb{P}(b \geq B_1^{\mathbf{T}}) \leq \eta^{\mathbf{T}}\mathbb{P}(b' \geq B_1^{\mathbf{T}}) =: \lambda_4'.$$

Let us denote

$$\begin{aligned}&\lambda_1 := \eta^{\mathbf{I}}K, \quad \lambda_3 := \eta^{\mathbf{I}} + \rho \\ &\lambda_2 := \eta^{\mathbf{T}}\mathbb{E}[(K - \mathbf{c}(b, B_1^{\mathbf{T}}))\mathbf{1}_{b\geq B_1^{\mathbf{T}}}], \quad \lambda_2' := \eta^{\mathbf{T}}\mathbb{E}[(K - \mathbf{c}(b', B_1^{\mathbf{T}}))\mathbf{1}_{b'\geq B_1^{\mathbf{T}}}].\end{aligned}$$

Applying the equivalence (7.3.1), we have

$$V(\beta^b) < V(\beta^{b'}) \Leftrightarrow \frac{\mathbb{E}[(K - \mathbf{c}(b', B_1^{\mathbf{T}}))\mathbf{1}_{b'\geq B_1^{\mathbf{T}}}] - \mathbb{E}[(K - \mathbf{c}(b, B_1^{\mathbf{T}}))\mathbf{1}_{b\geq B_1^{\mathbf{T}}}]}{\mathbb{P}(b' < B_1^{\mathbf{T}}) - \mathbb{P}(b < B_1^{\mathbf{T}})} < \frac{\eta^{\mathbf{I}}K}{\eta^{\mathbf{I}} + \rho}.$$

What is interesting in the right-hand side term is that it decouples the dynamic parameters of the problem, i.e. $\eta^{\mathbf{I}}$ and ρ, and the static parameters, i.e. K, $\mathbf{c}$, and $\mathcal{L}(B_1^{\mathbf{T}})$. In particular, this implies the following result: let $b_\star^{\eta^{\mathbf{I}},\rho}$ be the smallest optimal bid for parameters $\eta^{\mathbf{I}}, \rho$, then by definition, $V_{\eta^{\mathbf{I}},\rho}(\beta^{b_\star^{\eta^{\mathbf{I}},\rho}}) > V_{\eta^{\mathbf{I}},\rho}(\beta^b)$ for all $b < b_\star^{\eta^{\mathbf{I}},\rho}$ (where we stress the dependence of $V(\beta^b)$ in the parameters $\eta^{\mathbf{I}}, \rho$, and thus we have

$$\frac{\mathbb{E}[(K - \mathbf{c}(b_\star^{\eta^{\mathbf{I}},\rho}, B_1^{\mathbf{T}}))\mathbf{1}_{b_\star^{\eta^{\mathbf{I}},\rho} \geq B_1^{\mathbf{T}}}] - \mathbb{E}[(K - \mathbf{c}(b, B_1^{\mathbf{T}}))\mathbf{1}_{b \geq B_1^{\mathbf{T}}}]}{\mathbb{P}(b_\star^{\eta^{\mathbf{I}},\rho} < B_1^{\mathbf{T}}) - \mathbb{P}(b < B_1^{\mathbf{T}})} < \frac{\eta^{\mathbf{I}} K}{\eta^{\mathbf{I}} + \rho}.$$

Notice that this property will still be true if one decreases ρ (yielding $\tilde{\rho}$) or increases $\eta^{\mathbf{I}}$ (yielding $\tilde{\eta^{\mathbf{I}}}$), and thus we will still have $V_{\tilde{\eta^{\mathbf{I}}},\tilde{\rho}}(\beta^{b_\star^{\eta^{\mathbf{I}},\rho}}) > V_{\tilde{\eta^{\mathbf{I}}},\tilde{\rho}}(\beta^b)$ with these new parameters $\tilde{\rho}$ and $\tilde{\eta^{\mathbf{I}}}$, for all $b < b_\star^{\eta^{\mathbf{I}},\rho}$. This clearly implies that $b_\star^{\eta^{\mathbf{I}},\rho} \leq b_\star^{\tilde{\eta^{\mathbf{I}}},\tilde{\rho}}$.

In other words:

- The (smallest) optimal bid is increasing in ρ. This is consistent with the idea that when the Individual connects to a website displaying targeted ads, if ρ is small, then the Agent is more patient and thus takes into account the fact that there will be other opportunities for the Individual to get the Information, which is why he will bid less than if ρ is large, in which case the Agent ignores the future opportunities and will thus bid as if this was his only chance to display the ad.

- The (smallest) optimal bid is decreasing in $\eta^{\mathbf{I}}$. This is consistent with the fact that when the Individual connects to a website displaying targeted ads, if $\eta^{\mathbf{I}}$ is large, then the Agent knows that the Individual is susceptible to learn the Information by himself very soon anyway, which gives the Agent less incentive to bid high compared to the case where $\eta^{\mathbf{I}}$ is small, in which case the Individual has very little chance to learn the Information by himself.

7.3.2 Subscription-based gain function

In this section, we model the situation where the Information is the existence of a service, where the Agent is the company proposing this service, and where the Action of the Individual, once informed, is to subscribe to the service. To that aim, we simply consider the following *subscription-based* gain function:

$$V(\beta) = \mathbb{E}\Big[\sum_{n\in\mathbb{N}} e^{-(\tau^\beta + n)\rho} K\Big] - C(\beta), \quad \text{for } \beta \in \Pi_{OL},$$

where $\tau^\beta := \inf\{t \in \mathbb{R}_+ : X_t^\beta = 1\}$ is the time of information of the individual.

Let us interpret this gain function. The part $C(\beta)$ is simply the advertising cost described in the core framework. ρ is still the discount rate. The other part, $\mathbb{E}\Big[\sum_{n\in\mathbb{N}}\beta^{\tau^\beta+n}K\Big]$, represents the gain coming from the Individual's information dynamic. It simply corresponds to a regular payment of K every period 1 from the time of information τ^β (and thus the time of subscription) of the Individual.

We can now state the result of this section.

Theorem 7.3.2 *We have*

$$\sup_{\beta\in\Pi_{OL}} V(\beta) \;=\; \sup_{b\in\mathbb{R}} V(\beta^b),$$

with

$$V(\beta^b) \;=\; \frac{\eta^{\mathbf{I}}\frac{K}{1-\beta}+\mathbb{E}[\Big(\frac{K}{1-\beta}-\mathbf{c}(b,B_1^{\mathbf{T}})\Big)\mathbf{1}_{b\geq B_1^{\mathbf{T}}}]}{\eta^{\mathbf{I}}+\rho+\mathbb{P}(b\geq B_1^{\mathbf{T}})}$$

and any $b_\star\in\operatorname*{argmax}_{b\in\mathbb{R}} V(\beta^b)$ *yields an optimal constant bid, i.e. an optimal open-loop bid taking the form of a constant bid.*

Interpretation: Notice that the regular payment of K every period of duration 1 from the time of information is, from the Agent's viewpoint, equivalent to a unique payment of $\frac{K}{1-\beta}$ at the time of information. We are thus reduced to the previous case of purchase-based gain.

7.4 Social marketing model

We now model a very different kind of advertising, called *social marketing*. Social marketing is the activity of making advertising campaigns not to make profit but to sensitize people, in particular about unhealthy behaviors (anti-drug campaigns, road-safety campaigns, sexual-safety campaigns, etc). The common point of these campaigns is that they spread an Information about an unhealthy behavior. The Agent, here, is either an association or a governmental entity working for social good. As opposed to commercial advertising from previous section, informing an Individual here does not bring a reward to the Agent, but instead, cancels a cost.

The idea simply is that the Agent is philanthropic and considers that each Individual not behaving healthily incurs a cost to him. As long as the Individual is not informed, he thus keeps behaving unhealthily and incurring a continuous cost to the Agent. Once informed, the Individual does the Action to stop behaving unhealthily and the cost stops.

For this application, our study will be split in two sub-cases:

1. The case with a discount rate β, based on the same framework as previous models but with a cost function, and

2. the important case with no discounting (i.e. $\beta = 1$), where we will be able to make the framework much richer by introducing a population of N individuals as well as a non-targeted advertising mechanism, therefore turning the model into a population control problem.

In both cases the Agent's goal will be to *minimize* his cost.

7.4.1 Case with a discount rate

We start by the simpler case, with no social interaction nor non-targeted advertising, but with an arbitrary discount rate ρ. Besides the processes $N^{\mathbf{I}}$ and $N^{\mathbf{T}}$, we consider a third Poisson process N^E, independent from the others, with intensity η_E, counting the times when the Individual behaves unsafely.

In this social marketing problem, the cost function of the Agent is defined by

$$V(\beta) = \mathbb{E}\Big[\int_0^\infty e^{-\rho t}(K(1 - X_{t-}^\beta)dN_t^E\Big] + C(\beta), \quad \text{for } \beta \in \Pi_{OL}.$$

The part $C(\beta)$ is the advertising cost, and the part $\mathbb{E}\Big[\int_0^\infty e^{-\rho t}(K(1-X_{t-}^\beta)dN_t^E\Big]$ simply measures the cost perceived in the period before the Individual was informed, assuming that the Individual incurs a cost K to the Agent every time he behaves unsafely, discounted with the factor β.

We have the following result.

Theorem 7.4.1 *We have*

$$\inf_{\beta\in\Pi_{OL}} V(\beta) \;=\; \inf_{b\in\mathbb{R}} V(\beta^b),$$

with

$$V(\beta^b) \;=\; \frac{K + \eta^{\mathbf{T}}\mathbb{E}[\mathbf{c}(b, B_1^{\mathbf{T}})\mathbf{1}_{b\geq B_1^{\mathbf{T}}}]}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}}\mathbb{P}(b \geq B_1^{\mathbf{T}})}$$

and any $b_\star \in \operatorname{argmin}_{b\in\mathbb{R}} V(\beta^b)$ *yields an optimal constant bid, i.e. an optimal open-loop bid taking the form of a constant bid.*

Interpretation: Here again, we interpret ρ as the parameter of a random terminal time with exponential distribution. Notice that in the case of social marketing, where the Agent perceives a cost as long as the Individual is not informed, and where the cost stops as soon as he gets the information, there is already a random terminal time: the time when the Individual connects on the website intrinsically containing the information. Indeed, in such case, the cost stops and the problem stops as well. Both terminal times are exponential random variables with respective parameters $\eta^{\mathbf{I}}$ and ρ. It is known that they can be compressed in a unique terminal time (the minimum of both) with parameter $\eta^{\mathbf{I}}+\rho$. In other words, up to replacing the original intensity $\eta^{\mathbf{I}}$ of connection to a website containing the Information by $\eta^{\mathbf{I}}+\rho$, we are reduced to a problem with no discount rate ($\rho = 0$). The inner fraction can be split as follows:

$$\frac{K+\eta^{\mathbf{T}}\mathbb{E}[\mathbf{c}(b,B_1^{\mathbf{T}})\mathbf{1}_{b\geq B_1^{\mathbf{T}}}]}{\eta^{\mathbf{I}}+\rho+\eta^{\mathbf{T}}\mathbb{P}(b\geq B_1^{\mathbf{T}})} = \frac{K}{\eta^{\mathbf{I}}+\rho+\eta^{\mathbf{T}}\mathbb{P}(b\geq B_1^{\mathbf{T}})} + \frac{\eta^{\mathbf{T}}\mathbb{E}[\mathbf{c}(b,B_1^{\mathbf{T}})\mathbf{1}_{b\geq B_1^{\mathbf{T}}}]}{\eta^{\mathbf{I}}+\rho+\eta^{\mathbf{T}}\mathbb{P}(b\geq B_1^{\mathbf{T}})}$$

We first interpret the first term: $\eta^{\mathbf{I}}+\rho+\eta^{\mathbf{T}}\mathbb{P}(b\geq B_1^{\mathbf{T}})$ is the intensity of the time of information of the Individual, and thus $\frac{1}{\eta^{\mathbf{I}}+\rho+\eta^{\mathbf{T}}\mathbb{P}(b\geq B_1^{\mathbf{T}})}$ is the expected time before information. During this time, a continuous cost K is essentially perceived, which explains the term $\frac{K}{\eta^{\mathbf{I}}+\rho+\eta^{\mathbf{T}}\mathbb{P}(b\geq B_1^{\mathbf{T}})}$. Let us now interpret the second term. It has the same interpretation as for previous models: it essentially is the expected cost perceived at the time of termination of the problem, given that in this case, no reward, and only the ad cost, is perceived.

7.4.2 Case with no discount case, with social interactions and non-targeted advertising

In this section, we consider a social marketing model with no discounting, but with much more features than previous models. Although the model we study here is still based on our core framework, it is so much richer, and then for the sake of clarity, We redefine it from scratch.

In this model, we do not simply model websites intrinsically containing the Information and websites displaying targeted ads, but also:

- social interactions, and
- non-targeted advertising.

Essentially, rather than simply connecting to websites either containing the Information or displaying targeted ads, people will also connect on website displaying non-targeted

ads, and they will also socially interact. The reason why we introduce these two extra features is twofold:

1. **For relevance in terms of applications:** Social marketing nowadays still widely happens via non-targeted advertising (TV awareness campaigns, etc). Although our model proposes to use targeted advertising, it thus seems important to not completely dismiss the current method, and instead propose a way to combine both mechanisms.

2. **Mathematical reason:** The absence of discount factor allows the problem to still be tractable even by adding these features. In previous models, with a $\rho > 0$, one could not solve the problem with these extra features.

Let us reintroduce each component of the framework, one by one, with these additional features.

The population

Instead of a single Individual, we here consider a population with M individuals. Each individual $m \in [\![1, M]\!]$ is modeled with all the features of the Individual from previous models, and also extra features. The population's online behavior is characterized by:

- a family of M i.i.d. triplets $(N^{m,\mathbf{I}}, N^{m,\mathbf{T}}, N^{m,\bar{\mathbf{T}}}, N^{m,E})_{m \leq N}$ where, for $m \in [\![1, M]\!]$, $N^{m,\mathbf{I}}$, $N^{m,\mathbf{T}}$, $N^{m,\bar{\mathbf{T}}}$, and $N^{m,E}$ are four independent Poisson processes with respective intensities $\eta_{\mathbf{I}}$, $\eta_{\mathbf{T}}$, $\eta_{\bar{\mathbf{T}}}$, and η_E, Notice that it implies that each individual is assumed to share the same intensities. Up to this simplification, we shall see that the problem stays tractable.

- and a family $(N^{m,i,\mathbf{S}})_{m,i \in [\![1,M]\!]}$ of i.i.d. Poisson processes with intensity $\eta_{\mathbf{S}}$, independent from the other Poisson processes.

For all $m \in [\![1, M]\!]$, the processes $N^{m,\mathbf{I}}$, $N^{m,\mathbf{T}}$, and $N^{m,E}$, have the same interpretation as in previous model: $N^{m,\mathbf{I}}$ counts the times when individual n visits a website intrinsically containing the Information (in this case, it would be an association's website, the website specialized in health, etc). $N^{m,\mathbf{T}}$ counts the times when individual n connects to a website displaying targeted ads, and $N^{m,E}$ counts the time when he behaves unsafely. The new features are: $N^{m,\bar{\mathbf{T}}}$, simply counting the times when individual n visits a website displaying *non*-targeted ads, and for $n, i \in [\![1, M]\!]$, $N^{m,i,\mathbf{S}}$ counting the social interactions between individuals n and i in the population.

Targeted and non-targeted advertising auctions

Instead of only considering targeted advertising auctions, we also introduce a *non*-targeted advertising auction mechanism. Let us start by adapting the targeted advertising framework to the M-individual population.

Targeted advertising auctions: For each individual n, for $m \in [\![1, M]\!]$, each time individual n connects to a website displaying targeted ads, an auction is automatically opened where several agents bid to win the right to display their ads to the individual. As in previous models, to keep the problem tractable, we model the maximal bid from other bidders (other than our Agent), as random variables, i.i.d. across auctions and across individuals. We thus introduce an i.i.d. family of real valued random variables $(B_k^{m,\mathbf{T}})_{k\in\mathbb{N},m\in[\![1,M]\!]}$, where, for $m \in [\![1, M]\!]$ and $k \in \mathbb{N}$, $B_k^{m,\mathbf{T}}$ represents the maximal bid from other bidders at the k-th targeted advertising auction concerning individual n.

Non-targeted advertising auctions: In this model, we also consider non-targeted advertising. Each time when an individual (regardless of his index) connects to a website displaying non-targeted ads, here again, agents will compete to display their ads (with the only difference that they cannot make their bid depending upon the individual who connects to the website, hence the name "non-targeted advertising"). An auction is thus also opened at each such connection. As before, we model the maximum bid from other bidders (i.e. not the Agent) by random variables i.i.d. across non-targeted advertising auctions. We thus consider an i.i.d. family of real valued random variables $(B_k^{\bar{\mathbf{T}}})_{k\in\mathbb{N}}$, where for each $k \in \mathbb{N}$, $B_k^{\bar{\mathbf{T}}}$ represents the maximal bid of other bidders during the k-non-targeted advertising auction (in all the population). Furthermore, given an Agent's bid b and the other bidders maximal bid B, the price that the Agent has to pay if he wins the auction (i.e. if $b \geq B$) is defined by $\mathbf{c}^{\bar{\mathbf{T}}}(b, B)$, where $\mathbf{c}^{\bar{\mathbf{T}}} : \mathbb{R}^2 \to \mathbb{R}$ is a measurable function defining the auction rule.

The advertising bidding strategies

Given that there are now M individuals, targeted advertising, and non-targeted advertising, a general bidding map control will take a more complex form as in previous models.

Informally, a bidding map control, in this model, is a random process, depending only upon past events (i.e. not upon future events), valued in $\mathbb{R}^{M+1}$. The idea is that this vector will store the M bids that the Agent would like to make for each individual $m \in [\![1, M]\!]$ if he were to connect to a website displaying targeted ads, and the remaining coordinate corresponds to the bid that the Agent would like to make if someone (anonymous) connects to a website using non-targeted advertising. This is why $M+1$ potential

bids are required at any time, hence the term *bidding map*.

To make this intuition rigorous, let us first introduce the filtration $\mathbb{F} = (\mathcal{F}_t)_{t\in\mathbb{R}_+}$ generated by the processes

$$((N^{m,\mathbf{I}}, N^{m,\mathbf{T}}, N^{m,\bar{\mathbf{T}}}, N^{m,E}, B^{m,\mathbf{T}}_{N^{m,\mathbf{T}}}, N^{m,S})_{m\leq N}, B^{\bar{\mathbf{T}}}_{N^{\bar{\mathbf{T}}}}, ((N^{\{m,i\},\mathbf{S}})_{m,i\in[\![1,M]\!]})$$

where $N^{\bar{\mathbf{T}}} := \sum_{m=1}^{M} N^{m,\bar{\mathbf{T}}}$ globally counts the connections to a website displaying non-targeted ads. i.e. we have, for all $t \in \mathbb{R}_+$,

$$\mathcal{F}_t = \sigma(N^{m,\mathbf{I}}_s, N^{m,\mathbf{T}}_s, N^{m,\bar{\mathbf{T}}}_s, N^{m,E}_s, B^{m,\mathbf{T}}_{N^{m,\mathbf{T}}_s}, N^{m,S}_s, m \in [\![1,M]\!], B^{\bar{\mathbf{T}}}_{N^{\bar{\mathbf{T}}}_s}, N^{\{m,i\},\mathbf{S}}_s, m,i \in [\![1,M]\!], s \leq t)$$

Let us now define the notion of open-loop bidding map control. An open-loop bidding map is simply a process $\beta = (\beta^m)_{m=0,...,M}$, valued in $\mathbb{R}^{M+1}$, predictable and progressively measurable w.r.t. the filtration $\mathbb{F}$. It thus indeed means that for all t, the value of the open-loop bidding control β at time t, i.e. β_t, can only depend upon past events.

Let us now explain the meaning of the bidding map $\beta_t = (\beta^m_t)_{m=0,...,M} \in \mathbb{R}^{M+1}$. For $n = 1, ..., M$, β^m_t is the bid that the Agent would make if a targeted advertising auction for individual n happened at time t. The remaining coordinate, β^0_t is the bid that the Agent would make if a non-targeted advertising auction happens at time t.

In other words, the idea is that if an individual connects to a website displaying targeted ads, the website will open the targeted advertising auction for this individual, look at the bidding map $\beta_t = (\beta^m_t)_{m=0,...,M}$, and automatically use the bid β^m_t inscribed in this bidding map as the bid of the Agent for this auction. This allows the agent to specify a different bid for each individual, which encodes the idea of *targeted*-advertising. On the other hand, if an individual connects to a website displaying non-targeted ads, the website will open the non-targeted advertising auction for this connection, look at the Agent's bidding map $\beta_t = (\beta^m_t)_{m=0,...,M}$, and automatically use the bid β^0_t inscribed in this bidding map as the bid of the Agent for this auction. Notice that in this case, the bid of the Agent can thus not depend upon who connects, which encodes the idea of *non-targeted* advertising.

The information dynamic, cost function, and minimal bidding policies

Given an open-loop bidding map control β, we define the information dynamic of the population as follows. For all $m \in [\![1,M]\!]$, the information process X^n of individual n is a càd-làg process valued in $\{0,1\}$ satisfying the dynamic relation

$$\begin{aligned} X^{m,\beta}_0 &= 0 \\ dX^{m,\beta}_t &= (1 - X^{m,\beta}_{t-})(dN^{m,\mathbf{I}}_t + \mathbf{1}_{\beta^m_t \geq B^{m,\mathbf{T}}_{N^{m,\mathbf{T}}}} dN^{m,\mathbf{T}}_t + \mathbf{1}_{\beta^0_t \geq B^{\bar{\mathbf{T}}}_{N^{\bar{\mathbf{T}}}_t}} dN^{m,\bar{\mathbf{T}}}_t + \sum_{i=1}^{N} X^{i,\beta}_{t-} dN^{m,i,\mathbf{S}}_t). \end{aligned}$$

The interpretation of this dynamic is similar to previous sections, except that there are several individuals $m \in [\![1, M]\!]$, and that there is an additional term

$$\mathbf{1}_{\beta_t^0 \geq B^{\bar{\mathbf{T}}}_{N^{\bar{\mathbf{T}}}_t}} dN_t^{m,\bar{\mathbf{T}}} + \sum_{i=1}^{M} X_{t-}^{i,\beta} dN_t^{m,i,\mathbf{S}}.$$

Let us thus explain this extra term. It is essentially related to the new features of non-targeted advertising and social interactions. Nonetheless, to make the explanation clear, let us re-interpret the whole dynamic.

As in previous models, each individual n starts uninformed ($X_0^{m,\beta} = 0$). Once individual n is informed ($X_t^{m,\beta} = 1$), he stays informed ($(1 - X_{t-}^{m,\beta})$ part). As long as he is not informed, the rest of the dynamic is effective:

- As in previous models, when individual n connects to a website intrinsically containing the Information, individual becomes informed ($dN_t^{m,\mathbf{I}}$ part).
- Likewise, when individual n connects to a website displaying targeted ads ($dN_t^{m,\mathbf{T}}$ part), he gets informed if and only if the Agent's ad is displayed to him, i.e. iff the Agent wins the targeted advertising auction ($\mathbf{1}_{\beta_t^m \geq B^{m,\mathbf{T}}_{N^{m,\mathbf{T}}}}$ part).
- Additionally to previous models, individual n can also connect to websites displaying non-targeted ads ($dN_t^{m,\bar{\mathbf{T}}}$ part), in which case he will get informed if and only if the Agent's ad is displayed to him, i.e. iff the Agent wins the non-targeted advertising auction ($\mathbf{1}_{\beta_t^0 \geq B^{\bar{\mathbf{T}}}_{N^{\bar{\mathbf{T}}}_t}}$ part).
- Finally, if individual n socially interacts with individual i ($dN_t^{m,i,\mathbf{S}}$ part), he will get informed if and only if individual i is informed ($X_t^{i,\beta}$ part).

Given a bidding map control β, the expected cost incurred to the Agent is defined by

$$\begin{aligned} V(\beta) &= \mathbb{E}\Big[\sum_{i=1}^{M} \Big(\int_0^\infty K(1 - X_{t-}^{i,\beta}) dN_t^{E,i} + \int_0^\infty \mathbf{1}_{\beta_t^i > B^{i,\mathbf{T}}_{N^{i,\mathbf{T}}}} \mathbf{c}^{\mathbf{T}}(b_t^{i,\mathbf{T}}, B^{i,\mathbf{T}}_{N^{i,\mathbf{T}}}) dN^{i,\mathbf{T}} \\ &\quad + \int_0^\infty \mathbf{1}_{\beta_t^0 > B^{\bar{\mathbf{T}}}_{N_t^{\bar{\mathbf{T}}}}} \mathbf{c}^{\bar{\mathbf{T}}}(\beta_t^0, B^{\bar{\mathbf{T}}}_{N_t^{\bar{\mathbf{T}}}}) dN_t^{i,\bar{\mathbf{T}}} \Big) \Big]. \end{aligned}$$

This cost function is similar to previous model, except that there is a cost for each individual $i \in [\![1, M]\!]$ in the population, ($\sum_{i=1}^M$ part), and that there is an additional term $\int_0^\infty \mathbf{1}_{\beta_t^0 > B^{\bar{\mathbf{T}}}_{N_t^{\bar{\mathbf{T}}}}} \mathbf{c}(\beta_t^0, B^{\bar{\mathbf{T}}}_{N_t^{\bar{\mathbf{T}}}}) dN_t^{i,\bar{\mathbf{T}}}$. This new term is similar to the term $\int \mathbf{1}_{\beta_t^i > B^{i,\mathbf{T}}_{N^{i,\mathbf{T}}}} \mathbf{c}^{\mathbf{T}}(b_t^{i,\mathbf{T}}, B^{i,\mathbf{T}}_{N^{i,\mathbf{T}}}) dN^{i,\mathbf{T}}$

measuring the targeted adverting cost of the strategy. It instead clearly measures the non-targeted advertising cost of the strategy.

Minimal policy dynamic: A minimal policy is given by a pair of functions $\mathfrak{b} = (\mathfrak{b}^{\mathbf{T}}, \mathfrak{b}^{\bar{\mathbf{T}}})$ where $\mathfrak{b}^{\bar{\mathbf{T}}} : [0,1] \to \mathbb{R}$ and $\mathfrak{b}^{\mathbf{T}} : [0,1] \to \mathbb{R}$. To any minimal policy we associate the open-loop bidding map control $\beta^{\mathfrak{b}}$ satisfying the feedback form constraint

$$\beta_t^{m,\mathfrak{b}} = \mathfrak{b}^{\mathbf{T}}\Big(\frac{1}{M}\sum_{i=1}^{M} X_{t-}^{i,\beta^{\mathfrak{b}}}\Big)(1 - X_{t-}^{m,\beta^{\mathfrak{b}}}), \quad m \in [\![1, M]\!], \quad \beta_t^{0,\mathfrak{b}} = \mathfrak{b}^{\bar{\mathbf{T}}}\Big(\frac{1}{M}\sum_{i=1}^{M} X_{t-}^{i,\beta^{\mathfrak{b}}}\Big),$$

We now state the result for this model.

Theorem 7.4.2 *The optimal cost is given by*

$$\inf_{\beta\in\Pi_{OL}} V(\beta) = \sum_{p\in\frac{[\![0,M]\!]}{M}} \inf_{b\in\mathbb{R}} \frac{K + \eta^{\mathbf{T}}\mathbb{E}[\mathbf{c}(b, B_1^{\mathbf{T}})\mathbf{1}_{b>B_1^{1,\mathbf{T}}}] + \eta^{\bar{\mathbf{T}}}\mathbb{E}[\frac{\mathbf{c}^{\bar{\mathbf{T}}}(b,B_1^{\bar{\mathbf{T}}})}{1-p}\mathbf{1}_{b>\frac{B_1^{1,\bar{\mathbf{T}}}}{1-p}}]}{\eta^{\mathbf{I}} + \eta^{\mathbf{T}}\mathbb{P}(b \geq B_1^{1,\mathbf{T}}) + \eta^{\bar{\mathbf{T}}}\mathbb{P}(b \geq B_1^{\bar{\mathbf{T}}}) + p\eta^{\mathbf{S}}}$$

The minimal policy defined by $\mathfrak{b}_\star^{\bar{\mathbf{T}}}(p) = (1-p)\mathrm{b}_\star(p)$ *and* $\mathfrak{b}_\star^{\mathbf{T}}(p) = \mathrm{b}_\star(p)$*, where*

$$\mathrm{b}_\star(p) = \operatorname*{argmin}_{b\in\mathbb{R}} \frac{K + \eta^{\mathbf{T}}\mathbb{E}[\mathbf{c}(b, B_1^{\mathbf{T}})\mathbf{1}_{b>B_1^{1,\mathbf{T}}}] + \eta^{\bar{\mathbf{T}}}\mathbb{E}[\frac{\mathbf{c}^{\bar{\mathbf{T}}}(b,B_1^{\bar{\mathbf{T}}})}{1-p}\mathbf{1}_{b>\frac{B_1^{\bar{\mathbf{T}}}}{1-p}}]}{\eta^{\mathbf{I}} + \eta^{\mathbf{T}}\mathbb{P}(b \geq B_1^{1,\mathbf{T}}) + \eta^{\bar{\mathbf{T}}}\mathbb{P}(b \geq \frac{B_1^{\bar{\mathbf{T}}}}{1-p}) + p\eta^{\mathbf{S}}},$$

yields an optimal control $\beta^{\mathfrak{b}_\star}$*.*

Interpretations: Let us provide a few interpretations of this formula.

- **Interpretation of the part "$\sum_{p\in\frac{[\![0,M]\!]}{M}}$ ...":** We can split the problem in several successive problems each consisting in optimally going from a proportion $\frac{k}{M}$ of informed people to a proportion $\frac{k+1}{M}$, for $k \in \{0, ..., M-1\}$. The fact that there is no discount rate implies that the time when each problem starts does not matter, which implies that these successive problems can be optimized independently, i.e. one by one.

- **Interpretation of the term in the sum:** The justification of the form of the terms in the sum is similar to the justification given for the previous model: the fraction can be split into two fractions, one corresponding to the expected cost perceived during this period, and the other one corresponding to the expected cost perceived at the termination time of this period.

- **Interpretation of the term $\frac{\mathbf{c}^{\bar{\mathbf{T}}}(b,B_1^{\bar{\mathbf{T}}})}{1-p}$:** Notice that in the formula, $B_1^{1,\mathbf{T}}$ and $\frac{B_1^{\bar{\mathbf{T}}}}{1-p}$ play symmetric roles. It is *as if* the non-targeted advertising mechanism with price $B_1^{\bar{\mathbf{T}}}$ was equivalent a targeted advertising mechanism with price $\frac{B_1^{\bar{\mathbf{T}}}}{1-p}$. In other words, making the advertising mechanism not targeted essentially is equivalent to multiply the ad cost by $\frac{1}{1-p}$. This is natural since when the ad mechanism is not targeted, there is a probability p that it displays the ad to an already informed individual. Thus, statistically, for each ad displayed to an uninformed individual, $\frac{p}{1-p}$ ads will be displayed to already informed individuals (and be useless). This is thus equivalent to pay the price of $1+\frac{p}{1-p}=\frac{1}{1-p}$ ads to display an ad to an uninformed individual.

- **Interpretation of the term $p\eta^{\mathbf{S}}$:** Notice that in the formula, $p\eta^{\mathbf{S}}$ plays the same role as $\eta^{\mathbf{I}}$. This is consistent with the intuition that socially interacting with an informed individual has the same effect as visiting a website containing the information: it will inform the individual and not cost anything to the Agent. The more individuals are informed, the more likely such interaction is to occur. More precisely, each informed individual "plays the role" of a website containing the information, such that an individual has intensity $\frac{1}{M}\eta^{\mathbf{S}}$ to "visit" it, and thus, with a k informed individuals, it yields an intensity $\frac{k}{M}\eta^{\mathbf{S}}=p\eta^{\mathbf{S}}$.

Remark 7.4.1 (Mean-field approximation) *As in any population models with enough symmetry, it is expected that when M gets large, the model's result converge to a mean-field limit. Let us verify this, and see to what our result converges. Notice that the Agent's average optimal value per individual is thus*

$$\frac{1}{M}\inf_{\beta\in\Pi_{OL}}V(\beta)=\frac{1}{M}\sum_{p\in\frac{[\![0,M]\!]}{M}}\inf_{b\in\mathbb{R}}\frac{K+\eta^{\mathbf{T}}\mathbb{E}[\mathbf{c}^{\mathbf{T}}(b,B_1^{1,\mathbf{T}})\mathbf{1}_{b>B_1^{1,\mathbf{T}}}]+\eta^{\bar{\mathbf{T}}}\mathbb{E}[\frac{\mathbf{c}^{\bar{\mathbf{T}}}(b,B_1^{1,\bar{\mathbf{T}}})}{1-p}\mathbf{1}_{b>\frac{B_1^{1,\bar{\mathbf{T}}}}{1-p}}]}{\eta^{\mathbf{I}}+\eta^{\mathbf{T}}\mathbb{P}(b\geq B_1^{1,\mathbf{T}})+\eta^{\bar{\mathbf{T}}}\mathbb{P}(b\geq B_1^{\bar{\mathbf{T}}})+p\eta^{\mathbf{S}}}$$

which thus takes the form of a Riemann sum, implying that we have

$$\frac{1}{M}\inf_{b\in\Pi_{OL}}V(\beta)\simeq\int_0^1\inf_{b\in\mathbb{R}}\frac{K+\eta^{\mathbf{T}}\mathbb{E}[\mathbf{c}^{\mathbf{T}}(b,B_1^{1,\mathbf{T}})\mathbf{1}_{b>B_1^{1,\mathbf{T}}}]+\eta^{\bar{\mathbf{T}}}\mathbb{E}[\frac{B_1^{1,\mathbf{T}}}{1-p}\mathbf{1}_{b>\frac{\mathbf{c}^{\bar{\mathbf{T}}}(b,B_1^{1,\bar{\mathbf{T}}})}{1-p}}]}{\eta^{\mathbf{I}}+\eta^{\mathbf{T}}\mathbb{P}(b\geq B_1^{1,\mathbf{T}})+\eta^{\bar{\mathbf{T}}}\mathbb{P}(b\geq B_1^{\bar{\mathbf{T}}})+p\eta^{\mathbf{S}}}dp$$

Such result can be interesting for two important reasons:

1. *To obtain an analytical approximation of the optimal value in some cases where the integral can be explicitly computed,*

2. *and to provide a way to numerically approximate the optimal value, by discretizing the integral with a suitable discretization step. This can be useful with very large populations, where one might want to speed up the computation.*

7.5 Examples with explicit optimal bidding policies

7.5.1 Constant maximal bid from other bidders

In this section, we assume that the maximal bids from other bidders, i.e. $(B_k^{\mathbf{T}})_{k\in\mathbb{N}}$ for the targeted advertising auctions, and $(B_k^{\bar{\mathbf{T}}})_{k\in\mathbb{N}}$ for the non-targeted advertising auctions, are constant, i.e. $B_k^{\mathbf{T}} = B^{\mathbf{T}} \in \mathbb{R}$ and $B_k^{\bar{\mathbf{T}}} = B^{\bar{\mathbf{T}}} \in \mathbb{R}$. Given that they are constant, the first-rice auction or second-price auction cases essentially become equivalent, so let us focus on the second price type of auction, i.e. the auction payment rule $\mathbf{c}(b, B) = B$. Let us study two cases:

1. The commercial advertising problem with purchase-based gain function, and
2. The social marketing problem with no discount factor and with social interactions and non-targeted advertising.

Commercial advertising problem with purchase-based gain function

In this case, we have

$$V(\beta^b) \quad = \quad \frac{\eta^{\mathbf{I}} K + \eta^{\mathbf{T}}(K - B^{\mathbf{T}})\mathbf{1}_{b\geq B^{\mathbf{T}}}}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}}\mathbf{1}_{b\geq B^{\mathbf{T}}}},$$

and any $b_\star \in \underset{b\in\mathbb{R}}{\operatorname{argmax}}\ V(\beta^b)$ yields an optimal constant bid. Notice that $V(\beta^b)$ only takes two possible values, one for $b < B^{\mathbf{T}}$ and one for $b \geq B^{\mathbf{T}}$. The optimization thus reduces to choose either $b < B^{\mathbf{T}}$ (for instance $b = 0$), either $b \geq B^{\mathbf{T}}$ (for instance $b = B^{\mathbf{T}}$). Let us thus simply analyze the case where $b \geq B^{\mathbf{T}}$ is a better option: the optimal bid is $b_\star \geq B^{\mathbf{T}}$ iff

$$\frac{\eta^{\mathbf{I}} K + \eta^{\mathbf{T}}(K - B^{\mathbf{T}})}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}}} > \frac{\eta^{\mathbf{I}} K}{\eta^{\mathbf{I}} + \rho}$$

One straightforwardly proves that for $a, b, c, d > 0$, we have

$$\frac{a}{b} < \frac{a+c}{b+d} \Leftrightarrow \frac{a}{b} < \frac{c}{d} \tag{7.5.1}$$

and thus the optimal bid is $b_\star \geq B^{\mathbf{T}}$ iff

$$K - B^{\mathbf{T}} > \frac{\eta^{\mathbf{I}} K}{\eta^{\mathbf{I}} + \rho}$$

i.e. iff

$$B^{\mathbf{T}} < \frac{\rho}{\eta^{\mathbf{I}} + \rho} K.$$

We clearly see what we had already established in the general case: The optimal bids are "decreasing" in $\eta^{\mathbf{I}}$ and "increasing" in ρ, for instance in the sense that the smallest optimal bid is $B^{\mathbf{T}} \mathbf{1}_{B^{\mathbf{T}} \leq \frac{\rho}{\eta^{\mathbf{I}} + \rho}}$, and that this is clearly a decreasing function of $\eta^{\mathbf{I}}$ and an increasing function of ρ.

There is another interesting optimal bid, that is, the bid $\frac{\rho}{\eta^{\mathbf{I}} + \rho} K$. Indeed, this bid is the only one to be optimal *regardless* $B^{\mathbf{T}}$. In other words, simply knowing (or assuming) that other bidders' maximal bid is constant is enough to have a dominant bidding strategy $\frac{\rho}{\eta^{\mathbf{I}} + \rho} K$, optimal regardless the other bidders' maximal bid.

Social marketing problem with no discount factor and with social interactions and non-targeted advertising

In this case, the optimal bidding strategy is given by a minimal bidding policy $\mathfrak{b}_\star = (\mathfrak{b}^{\mathbf{T}}, \mathfrak{b}^{\bar{\mathbf{T}}})$ with $\mathfrak{b}^{\mathbf{T}}_\star(p) = \mathrm{b}_\star(p)$ and $\mathfrak{b}^{\bar{\mathbf{T}}}_\star(p) = (1-p)\mathrm{b}_\star(p)$, where

$$\mathrm{b}_\star(p) = \operatorname*{argmin}_{b \in \mathbb{R}} \frac{K + \eta^{\mathbf{T}} B^{1,\mathbf{T}} \mathbf{1}_{b > B^{1,\mathbf{T}}} + \eta^{\bar{\mathbf{T}}} \frac{B^{\bar{\mathbf{T}}}}{1-p} \mathbf{1}_{b > \frac{B^{\bar{\mathbf{T}}}}{1-p}}}{\eta^{\mathbf{I}} + \eta^{\mathbf{T}} \mathbf{1}_{b \geq B^{1,\mathbf{T}}} + \eta^{\bar{\mathbf{T}}} \mathbf{1}_{b \geq \frac{B^{\bar{\mathbf{T}}}}{1-p}} + p\eta^{\mathbf{S}}}.$$

In order to obtain simple and intepretable formulas let us assume that there is only one type of advertising.

Only targeted advertising. If there is only targeted advertising, i.e. if $\eta^{\bar{\mathbf{T}}} = 0$, we have

$$\mathfrak{b}^{\mathbf{T}}_\star(p) = \operatorname*{argmin}_{b \in \mathbb{R}} \frac{K + \eta^{\mathbf{T}} B^{1,\mathbf{T}} \mathbf{1}_{b > B^{1,\mathbf{T}}}}{\eta^{\mathbf{I}} + \eta^{\mathbf{T}} \mathbf{1}_{b \geq B^{1,\mathbf{T}}} + p\eta^{\mathbf{S}}},$$

Here again, we are reduced to compare two costs:

$$\frac{K}{\eta^{\mathbf{I}} + p\eta^{\mathbf{S}}} \text{ and } \frac{K + \eta^{\mathbf{T}} B^{1,\mathbf{T}}}{\eta^{\mathbf{I}} + \eta^{\mathbf{T}} + p\eta^{\mathbf{S}}},$$

the first one being obtained for $b < B^{1,\mathbf{T}}$, and the second one for $b \geq B^{1,\mathbf{T}}$. The best option will be $\mathfrak{b}_\star(p) \geq B^{1,\mathbf{T}}$ if and only if

$$\frac{K}{\eta^{\mathbf{I}} + p\eta^{\mathbf{S}}} > \frac{K + \eta^{\mathbf{T}} B^{1,\mathbf{T}}}{\eta^{\mathbf{I}} + \eta^{\mathbf{T}} + p\eta^{\mathbf{S}}},$$

Again using (7.5.1), this is equivalent to

$$B^{1,\mathbf{T}} < \frac{K}{\eta^{\mathbf{I}} + p\eta^{\mathbf{S}}}$$

which is equivalent to

$$p < \frac{\frac{K}{B^{1,\mathbf{T}}} - \eta^{\mathbf{I}}}{\eta^{\mathbf{S}}}.$$

This means that below the informed proportion $\frac{\frac{K}{B^{1,\mathbf{T}}} - \eta^{\mathbf{I}}}{\eta^{\mathbf{S}}}$, one should bid higher than $B^{\mathbf{T}}$ (and thus display ads), and after the informed proportion $\frac{\frac{K}{B^{1,\mathbf{T}}} - \eta^{\mathbf{I}}}{\eta^{\mathbf{S}}}$, one should bid lower than $B^{\mathbf{T}}$ (and thus stop displaying ads). Notice, in particular, that the threshold of informed proportion from which one has to stop displaying ads is decreasing in $\eta^{\mathbf{I}}$ and in $\eta^{\mathbf{S}}$. Let us interpret this.

- First of all, the fact that there is an informed proportion below which the Agent should display ads and above which he should not display ads necessarily comes from the social interactions, since the no-social interaction case ($\eta^{\mathbf{S}} = 0$) implies that $\mathfrak{b}_\star(p) \geq B^{1,\mathbf{T}}$ is optimal iff $B^{1,\mathbf{T}} < \frac{K}{\eta^{\mathbf{I}}}$. It is however interesting to see that $\eta^{\mathbf{I}}$ affects the threshold proportion $\frac{\frac{K}{B^{1,\mathbf{T}}} - \eta^{\mathbf{I}}}{\eta^{\mathbf{S}}}$. The fact that the presence of social interactions is susceptible to introduce a threshold proportion after which the Agent should stop displaying ads is the following: Let us assume that an individual connects to a website displaying targeted ads. With social interactions, the more people are informed, the sooner this individual will learn the Information anyway, by interacting with an informed individual. Therefore, the incentive of the Agent to display the ad to him is weaker as the proportion of informed individuals increases, which justifies that the bid he is willing to make is smaller, and once it is small enough to fall below $B^{1,\mathbf{T}}$, the Agent will stop displaying ads.

- The interpretation of the decreasing nature of the threshold proportion in $\eta^{\mathbf{I}}$ and $\eta^{\mathbf{S}}$ is the following. For a fixed proportion of informed individuals p, increasing the intensity of social interactions $\eta^{\mathbf{S}}$ will also make more probable a soon interaction with an informed people, thus weakening the Agent's incentive to display

an ad, such that this incentive will be fully compensated after a smaller informed proportion. Likewise, increasing the intensity $\eta^{\mathbf{I}}$ of connections to a website containing the information will make people inform themselves faster, thus catalyzing the increase of the informed proportion, in turn decreasing the Agent's incentive to display an ad.

Only non-targeted advertising. If there is only non-targeted advertising, i.e. if $\eta^{\mathbf{T}} = 0$, we have

$$\mathfrak{b}_\star^{\bar{\mathbf{T}}}(p) = \operatorname*{argmin}_{b\in\mathbb{R}} \frac{K + \eta^{\bar{\mathbf{T}}} \frac{B^{\bar{\mathbf{T}}}}{1-p} \mathbf{1}_{b>B^{\bar{\mathbf{T}}}}}{\eta^{\mathbf{I}} + \eta^{\bar{\mathbf{T}}} \mathbf{1}_{b\geq B^{\bar{\mathbf{T}}}} + p\eta^{\mathbf{S}}}.$$

Here again, we are reduced to compare two costs:

$$\frac{K}{\eta^{\mathbf{I}} + p\eta^{\mathbf{S}}} \quad \text{and} \quad \frac{K + \eta^{\bar{\mathbf{T}}} \frac{B^{\bar{\mathbf{T}}}}{1-p}}{\eta^{\mathbf{I}} + \eta^{\bar{\mathbf{T}}} + p\eta^{\mathbf{S}}},$$

the first one being obtained for $b < B^{\bar{\mathbf{T}}}$, and the second one for $b \geq B^{\bar{\mathbf{T}}}$. The best option will be $\mathfrak{b}_\star^{\bar{\mathbf{T}}}(p) \geq B^{\bar{\mathbf{T}}}$ if and only if

$$\frac{K}{\eta^{\mathbf{I}} + p\eta^{\mathbf{S}}} > \frac{K + \eta^{\bar{\mathbf{T}}} \frac{B^{\bar{\mathbf{T}}}}{1-p}}{\eta^{\mathbf{I}} + \eta^{\bar{\mathbf{T}}} + p\eta^{\mathbf{S}}},$$

Again using (7.5.1), this is equivalent to

$$\frac{B^{\bar{\mathbf{T}}}}{1-p} < \frac{K}{\eta^{\mathbf{I}} + p\eta^{\mathbf{S}}}$$

which is equivalent to

$$p < \frac{K - \eta^{\mathbf{I}} B^{\bar{\mathbf{T}}}}{K + \eta^{\mathbf{S}} B^{\bar{\mathbf{T}}}}$$

This means that before the informed proportion $\frac{K-\eta^{\mathbf{I}}B^{\bar{\mathbf{T}}}}{K+\eta^{\mathbf{S}}B^{\bar{\mathbf{T}}}}$, one should bid higher than $B^{\bar{\mathbf{T}}}$ (and thus display ads), and after the informed proportion $\frac{K-\eta^{\mathbf{I}}B^{\bar{\mathbf{T}}}}{K+\eta^{\mathbf{S}}B^{\bar{\mathbf{T}}}}$, one should bid lower than $B^{\bar{\mathbf{T}}}$ (and thus stop displaying ads). Notice, in particular, that, as in the "only targeted advertising" case, the informed proportion when one has to stop displaying ads is decreasing in $\eta^{\mathbf{I}}$ and in $\eta^{\mathbf{S}}$. Let us interpret this. All the interpretations given in the "only targeted advertising" case still apply here, but there is an additional justification of the fact that there is a threshold informed proportion above which the Agent should stop

displaying ads. Recall that in "only targeted advertising" case, we said that the presence of such threshold came from the presence of social interactions, and that when they are absent ($\eta^{\mathbf{S}} = 0$), or more generally when $\eta^{\mathbf{S}}$ is small enough, there is no threshold (the optimal bidding strategy is a constant bid). Here, notice that we have $\frac{K-\eta^{\mathbf{I}}B^{\bar{\mathbf{T}}}}{K+\eta^{\mathbf{S}}B^{\bar{\mathbf{T}}}} < 1$, even if $\eta^{\mathbf{S}} = 0$ (recall that we assumed that $\eta^{\mathbf{I}} > 0$). Thus, as opposed to the previous "only targeted advertising" case, the existence of such threshold does not only come from social interactions. To emphasize that this threshold also comes from the non-targeted nature of the advertising, let us set $\eta^{\mathbf{S}} = 0$, and see that even in this case, we still have a threshold. Indeed, with $\eta^{\mathbf{S}} = 0$, we obtain the threshold $\frac{K-\eta^{\mathbf{I}}B^{\bar{\mathbf{T}}}}{K} < 1$. Let us interpret this result. Displaying non-targeted ads always comes with the risk to display ads to already informed people, and thus paying for a useless ad. The more people are informed, the higher the risk. This is why after some proportion, it is simply not worth paying for displaying an ad, and thus the Agent has to stop doing so.

7.5.2 Uniform maximal bid from other bidders

Notice that the solutions of all the problems studied in this chapter have similar forms. We called them "semi-explicit" because they still required to compute an inf, sup, argmin, or argmax, although it was always an optimization on $\mathbb{R}$ of a rather simple fraction expressed in terms of the problem's parameters.

Also notice that what essentially prevented the formulas to be fully explicit was 1) that the probability distribution of $B_1^{\mathbf{T}}$ (and $B_1^{\bar{\mathbf{T}}}$ in the fourth model) had not been specified, and 2) that the auction payment rule $\mathbf{c}$ was not fixed.

In this section, we shall see that by assuming that the other bidders' maximal bid distribution is in the class of uniform distribution (with any mean and variance, in particular including constant bids), we are able to derive fully explicit formulas. Regarding the auction payment rule, we shall focus on the *first-price auction rule*, but the same argument applies to the *second-price auction rule*.

In this section, we study the case where the auction's payment rule is given by $\mathbf{c}(b, B) = B$, that is, if the Agent wins the auction ($b \geq B$), he pays his own bid, i.e. the maximal bid in the auction.

We shall study two classes of distributions for other bidders' maximal bid that will lead to fully explicit formulas. We focus on the example of the purchase-based commercial advertising model, but the same argument can be adapted to the other models, and we fix the first-price auction rule $\mathbf{c}(b, B) = b$, we get the formula,

$$V(\beta^b) = \frac{\eta^{\mathbf{I}}K + \eta^{\mathbf{T}}\mathbb{E}[(K-b)\mathbf{1}_{b\geq B_1^{\mathbf{T}}}]}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}}\mathbb{P}(b \geq B_1^{\mathbf{T}})} = \frac{\eta^{\mathbf{I}}K + \eta^{\mathbf{T}}(K-b)\mathbb{P}(b \geq B_1^{\mathbf{T}})}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}}\mathbb{P}(b \geq B_1^{\mathbf{T}})}.$$

Assuming that $B_1^{\mathbf{T}}$ follows a uniform distribution on $[b^-, b^+]$ where $B^- < B^+$, we can restrict the search for the argmax to the interval $[b^-, b^+]$, i.e.

$$b_\star = \underset{b\in[b^-,b^+]}{\operatorname{argmax}} \frac{\eta^{\mathbf{I}} K + \eta^{\mathbf{T}}(K-b)\mathbb{P}(b \geq B_1^{\mathbf{T}})}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}}\mathbb{P}(b \geq B_1^{\mathbf{T}})}$$

b being limited to this support, the term $\mathbb{P}(b \geq B_1^{\mathbf{T}}) = \frac{b-b^-}{b^+-b^-}$ becomes linear in b, and we have

$$b_\star = \underset{b\in[b^-,b^+]}{\operatorname{argmax}} \frac{\eta^{\mathbf{I}} K + \eta^{\mathbf{T}}(K-b)\frac{b-b^-}{b^+-b^-}}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}}\frac{b-b^-}{b^+-b^-}}$$

We can then clearly make a change of variable

$$b' = \eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}} \frac{b-b^-}{b^+-b^-}$$

i.e.

$$b = \lambda_1 + \lambda_2 b'$$

where

$$\lambda_1 = b^- - (b^+ - b^-)\frac{\eta^{\mathbf{I}} + \rho}{\eta^{\mathbf{T}}}, \quad \lambda_2 = \frac{b^+ - b^-}{\eta^{\mathbf{T}}}$$

such that

$$\begin{aligned}
\underset{b\in[b^-,b^+]}{\operatorname{argmax}} \frac{\eta^{\mathbf{I}} K + \eta^{\mathbf{T}}(K-b)\mathbb{P}(b \geq B_1^{\mathbf{T}})}{\eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}}\mathbb{P}(b \geq B_1^{\mathbf{T}})} &= \underset{b\in[b^-,b^+]}{\operatorname{argmax}} \frac{\eta^{\mathbf{I}} K + \eta^{\mathbf{T}}(K-\lambda_1-\lambda_2 b')\frac{b'-(\eta^{\mathbf{I}}+\rho)}{\eta^{\mathbf{T}}}}{b'} \\
&= \underset{b\in[b^-,b^+]}{\operatorname{argmax}} \frac{\eta^{\mathbf{I}} K + (K-\lambda_1-\lambda_2 b')(b'-(\eta^{\mathbf{I}}+\rho))}{b'} \\
&= \underset{b'\in[b'^-,b'^+]}{\operatorname{argmax}} \frac{a_0 + a_1 b' + a_2 b'^2}{b'}
\end{aligned}$$

where

$$a_0 = \lambda_1(\eta^{\mathbf{I}} + \rho) - K\rho, \quad a_1 = K - \lambda_1 + \lambda_2(\eta^{\mathbf{I}} + \rho), \quad a_2 = -\lambda_2 < 0,$$

and

$$b'^- = \eta^{\mathbf{I}} + \rho, \quad b'^+ = \eta^{\mathbf{I}} + \rho + \eta^{\mathbf{T}}.$$

We thus have

$$\underset{b\in[b^-,b^+]}{\operatorname{argmax}} \frac{\eta^{\mathbf{I}}K+\eta^{\mathbf{T}}(K-b)\mathbb{P}(b\geq B_1^{\mathbf{T}})}{\eta^{\mathbf{I}}+\rho+\eta^{\mathbf{T}}\mathbb{P}(b\geq B_1^{\mathbf{T}})} \quad = \quad \underset{b'\in[b'^-,b'^+]}{\operatorname{argmax}} \left(\frac{a_0}{b'}+a_1+a_2b'\right).$$

By deriving the last expression in b', we obtain $a_2-\frac{a_0}{b'^2}$ which is negative, for $b'\in[b'^-,b'^+]\subset\mathbb{R}_+$, if and only if $b'^2\geq\frac{a_0}{a_2}$, and thus if and only if $b'\geq\sqrt{\left(\frac{a_0}{a_2}\right)_+}$. The optimal b' is thus given by

$$b'_\star=\max\left(b'^-,\min\left(b'^+,\sqrt{\left(\frac{a_0}{a_2}\right)_+}\right)\right)$$

and thus

$$\begin{aligned} b_\star &= \lambda_1+\lambda_2\max\left(b'^-,\min\left(b'^+,\sqrt{\left(\frac{a_0}{a_2}\right)_+}\right)\right)\\ &= \max\left(b^-,\min\left(b^+,x\right)\right)\end{aligned}$$

where

$$\begin{aligned} x &= \lambda_1+\lambda_2\sqrt{\left(\frac{a_0}{a_2}\right)_+}=\lambda_1+\lambda_2\sqrt{\left(\frac{K\rho-\lambda_1(\eta^{\mathbf{I}}+\rho)}{\lambda_2}\right)_+}=\lambda_1+\sqrt{\lambda_2(K\rho-\lambda_1(\eta^{\mathbf{I}}+\rho))_+}\\ &= b^--(b^+-b^-)\frac{\eta^{\mathbf{I}}+\rho}{\eta^{\mathbf{T}}}+\sqrt{\frac{b^+-b^-}{\eta^{\mathbf{T}}}(K\rho-(b^--(b^+-b^-)\frac{\eta^{\mathbf{I}}+\rho}{\eta^{\mathbf{T}}})(\eta^{\mathbf{I}}+\rho))_+}\\ &= b^--(b^+-b^-)\frac{\eta^{\mathbf{I}}+\rho}{\eta^{\mathbf{T}}}+\sqrt{\frac{b^+-b^-}{\eta^{\mathbf{T}}}(K\rho-b^-(\eta^{\mathbf{I}}+\rho)+\frac{b^+-b^-}{\eta^{\mathbf{T}}}(\eta^{\mathbf{I}}+\rho)^2))_+}.\end{aligned}$$

7.6 Proofs

To simplify notations, in the proofs, let us focus on the second-price auction rule, i.e. $\mathbf{c}(b,B)=B$. The case of first-price auction rule is proved similarly.

7.6.1 Proof for social marketing with no discount factor

Fix an arbitrary open-loop bidding map control β. Let us denote

$$p_t^\beta=\frac{1}{M}\sum_{i=1}^M X_t^{i,\beta},\quad \forall t\in\mathbb{R}_+$$

the proportion of informed individuals at each time $t \in \mathbb{R}_+$. The underlying idea of this proof is a change of variable from the Poisson processes of the problem to the proportion p_t^β in the cost function. The motivation is that, as we saw in the interpretation of Theorem 7.4.2, intuitively, the problem is like a sequence of independent problems consisting in going from proportion p to $p + \frac{1}{M}$ with minimal cost, for all $p \in \frac{[\![0,M[\![}{M}$, and thus, the right way to look at the problem should be in terms of optimizing the cost over the proportions from $p = 0$ to $p = 1$ rather than over the times of jumps of the numerous Poisson processes defined in our model, the idea being that it should make clearer that one can simply optimize *locally*, i.e. for each transition from p to $p + \frac{1}{M}$ (point-wise optimization). Of course, the cost function is expressed in terms of the problem's Poisson processes, and we want to express it in terms of p^β, and all these processes are piece-wise constant processes. Thus, we have to be careful in the way we change of variable. The idea is to use the compensated processes of the Poisson processes with martingale arguments to apply the principles of continuous time change of variables to our discrete random jump processes.

More precisely, changing the variable to p_t^β in $V(\beta)$ essentially means to replace the Poisson processes dN^E, $dN^{\mathbf{T}}$, and $dN^{\bar{\mathbf{T}}}$ by dp^β, and to that end, we shall express $V(\beta)$ first with dt thanks to the intensity processes, then make the change of variable to obtain another intensity process, and then move back to the world of jump processes to obtain dp^β. Of course, if we want to perfectly obtain dp^β, we need to know what intensity process will fall back on it.

We are looking for an $\mathbb{F}$-predictable process G such that for all H positive and $\mathbb{F}$ predictable, we have

$$\mathbb{E}[\int_0^\infty H_t W_t dt] = \mathbb{E}[\int_0^\infty H_t dp_t^\beta]$$

Notice that we have

$$\begin{aligned}
dp_t^\beta &= \frac{1}{M}\sum_{i=1}^M dX_t^{i,\beta} \\
&= \frac{1}{M}\sum_{i=1}^M (1 - X_{t-}^{i,\beta})\Big(dN_t^{i,\mathbf{I}} + \mathbf{1}_{\beta_t^i \geq B^{i,\mathbf{T}}_{N_t^{i,\mathbf{T}}}} dN_t^{i,\mathbf{T}} + \mathbf{1}_{\beta_t^0 \geq B^{\bar{\mathbf{T}}}_{N_t^{\bar{\mathbf{T}}}}} dN_t^{i,\bar{\mathbf{T}}} + \sum_{m \in [\![1,M]\!]} X_{t-}^{m,\beta} dN_t^{i,n,\mathbf{S}}\Big) \\
&= \frac{1}{M}\sum_{i=1}^M (1 - X_{t-}^{i,\beta})\Big(dN_t^{i,\mathbf{I}} + \mathbf{1}_{\beta_t^i \geq B^{i,\mathbf{T}}_{N_{t-}^{i,\mathbf{T}}+1}} dN_t^{i,\mathbf{T}} + \mathbf{1}_{\beta_t^0 \geq B^{\bar{\mathbf{T}}}_{N_{t-}^{\bar{\mathbf{T}}}+1}} dN_t^{i,\bar{\mathbf{T}}} + \sum_{m \in [\![1,M]\!]} X_{t-}^{m,\beta} dN_t^{i,n,\mathbf{S}}\Big)
\end{aligned}$$

The process

$$t \mapsto p_t^\beta - \int_0^t \frac{1}{M}\sum_i (1-X_{s-}^{i,\beta})\Big(\eta^{\mathbf{I}} + \mathbf{1}_{\beta_s^i \geq B^{i,\mathbf{T}}_{N_{s-}^{i,\mathbf{T}}+1}}\eta^{\mathbf{T}} + \mathbf{1}_{\beta_s^0 \geq B^{\bar{\mathbf{T}}}_{N_{s-}^{\bar{\mathbf{T}}}+1}}\eta^{\bar{\mathbf{T}}} + \sum_{m\in[\![1,M]\!]} X_{s-}^{m,\beta}\eta^{\mathbf{S}}\Big)ds \tag{7.6.1}$$

is thus a martingale. Let us detail why. A classical result of Poisson processes and martingale theory is that for any Poisson process N with intensity η, the *compensated process of* N, defined by $dN_t - \eta dt$ (i.e. $(N_t - \eta t)_{t\in\mathbb{R}_+}$) is a martingale w.r.t. the filtration generated by N, but also, clearly, w.r.t. any filtration generated by N and any process Y independent of N. This implies that all the Poisson processes considered in this model are martingales w.r.t. the filtration $\tilde{\mathbb{F}} = (\tilde{\mathcal{F}}_t)_{t\in\mathbb{R}_+}$ defined by

$$\tilde{\mathcal{F}}_t = \sigma((B_k^{i,\mathbf{T}})_{i\in[\![1,M]\!],k\in\mathbb{N}_\star}, (B_k^{\bar{\mathbf{T}}})_{k\in\mathbb{N}_\star}, (N_s^{\mathbf{I}}, N_s^{\mathbf{T}}, N_s^{\bar{\mathbf{T}}}, N_s^{\mathbf{S}}, N_s^E)_{s\leq t}), \quad t\in\mathbb{R}_+,$$

that is, the filtration corresponding to the knowledge, *from the start*, of all the other bidders' maximal bids $(B_k^{i,\mathbf{T}})_{i\in[\![1,M]\!],k\in\mathbb{N}_\star}, (B_k^{\bar{\mathbf{T}}})_{k\in\mathbb{N}_\star}$, and the knowledge, revealed as time goes by, of the Poisson processes of the problem. Notice that, then, the processes in the integrand in (7.6.1) is $\tilde{\mathbb{F}}$-predictable, which thus implies that the process in (7.6.1) is a $\tilde{\mathbb{F}}$-martingale. Notice that $\mathcal{F}_t \subset \tilde{\mathcal{F}}_t$ for all $t\in\mathbb{R}_+$, and thus, for any bounded positive $\mathbb{F}$-predictable process H, we have

$$\begin{aligned}
&\mathbb{E}[\int_0^\infty H_t dp_t^\beta] \\
=\ &\mathbb{E}[\int_0^\infty H_t(\frac{1}{M}\sum_i (1-X_t^{i,\beta})\Big(\eta^{\mathbf{I}} + \mathbf{1}_{\beta_t^i \geq B^{i,\mathbf{T}}_{N_{t-}^{i,\mathbf{T}}+1}}\eta^{\mathbf{T}} + \mathbf{1}_{\beta_t^0 \geq B^{\bar{\mathbf{T}}}_{N_{t-}^{\bar{\mathbf{T}}}+1}}\eta^{\bar{\mathbf{T}}} + \sum_{m\in[\![1,M]\!]} X_{t-}^{m,\beta}\eta^{\mathbf{S}}\Big)dt)],
\end{aligned}$$

but as H is assumed $\mathbb{F}$-predictable, we clearly have

$$\begin{aligned}
&\mathbb{E}[\int_0^\infty H_t dp_t^\beta] \\
=\ &\mathbb{E}[\int_0^\infty H_t \frac{1}{M}\sum_i (1-X_t^{i,\beta})\Big(\eta^{\mathbf{I}} + \mathbb{P}(b\geq B_1^{i,\mathbf{T}})_{b:=\beta_t^i}\eta^{\mathbf{T}} + \mathbb{P}(b\geq B_1^{\bar{\mathbf{T}}})_{b:=\beta_t^0}\eta^{\bar{\mathbf{T}}} + \sum_{m\in[\![1,M]\!]} X_t^m\eta^{\mathbf{S}}\Big))dt].
\end{aligned}$$

We can simplify this expression as follows:

$$\begin{aligned}
&\mathbb{E}[\int_0^\infty H_t dp_t^\beta] \\
=\ &\mathbb{E}[\int_0^\infty H_t(1-p_t^\beta)(\eta^{\mathbf{I}} + \alpha_t^\beta\eta^{\mathbf{T}} + \mathbb{P}(b\geq B_1^{\bar{\mathbf{T}}})_{b:=\beta_t^0}\eta^{\bar{\mathbf{T}}} + p_t^\beta\eta^{\mathbf{S}})dt].
\end{aligned} \tag{7.6.2}$$

where $\alpha_t^\beta := \frac{\sum_{m=1}^M (1-X_{t-}^{m,\beta})\mathbb{P}(b\geq B_1^{\mathbf{T}})_{b:=\beta_t^i}}{M(1-p_{t-}^\beta)}$. The process

$$G_t := (1-p_t^\beta)(\eta^{\mathbf{I}} + \alpha_t^\beta \eta^{\mathbf{T}} + \mathbb{P}(b \geq B_1^{\bar{\mathbf{T}}})_{b:=\beta_t^0}\eta^{\bar{\mathbf{T}}} + p_t^\beta \eta^{\mathbf{S}}), \quad \forall t \in \mathbb{R}_+$$

will thus play the role of the intensity process we want to obtain in $V(\beta)$ to make our change of variable to p_t^β. Let us now express the Agent's cost in terms of dt in order to make this change of variable. First, we work a little bit on the cost function. We have

$$\begin{aligned}
V(\beta) &= \mathbb{E}\Big[\sum_{i=1}^M \Big(\int_0^\infty K(1-X_{t-}^{i,\beta})dN_t^{i,E} + \int \mathbf{1}_{\beta_t^i > B^{i,\mathbf{T}}_{N^{i,\mathbf{T}}}} B^{i,\mathbf{T}}_{N^{i,\mathbf{T}}} dN^{i,\mathbf{T}} + \int_0^\infty \mathbf{1}_{\beta_t^0 > B^{\bar{\mathbf{T}}}_{N_t^{\bar{\mathbf{T}}}}} B^{\bar{\mathbf{T}}}_{N_t^{\bar{\mathbf{T}}}} dN^{i,\bar{\mathbf{T}}}\Big)\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^M \int_0^\infty \Big(K(1-X_{t-}^{i,\beta}) + \mathbf{1}_{\beta_t^i > B^{i,\mathbf{T}}_{N^{i,\mathbf{T}}}} B^{i,\mathbf{T}}_{N^{i,\mathbf{T}}}\eta^{\mathbf{T}} + \mathbf{1}_{\beta_t^0 > B^{\bar{\mathbf{T}}}_{N_t^{\bar{\mathbf{T}}}}} B^{\bar{\mathbf{T}}}_{N_t^{\bar{\mathbf{T}}}}\eta^{\bar{\mathbf{T}}}\Big)dt\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^M \int_0^\infty \Big(K(1-X_{t-}^{i,\beta}) + \mathbb{E}[B_1^{1,\mathbf{T}}\mathbf{1}_{b>B_1^{1,\mathbf{T}}}]_{b:=\beta_t^i}\eta^{\mathbf{T}} + \mathbb{E}[B_1^{\bar{\mathbf{T}}}\mathbf{1}_{b>B_1^{\bar{\mathbf{T}}}}]_{b:=\beta_t^0}\eta^{\bar{\mathbf{T}}}\Big)dt\Big]
\end{aligned}$$

$$\begin{aligned}
V(\beta) &= \mathbb{E}\Big[\sum_{i=1}^M \Big(\int_0^\infty K(1-X_{t-}^{i,\beta})dN_t^{i,E} + \int \mathbf{1}_{\beta_t^i > B^{i,\mathbf{T}}_{N^{i,\mathbf{T}}}} B^{i,\mathbf{T}}_{N^{i,\mathbf{T}}} dN^{i,\mathbf{T}} + \int_0^\infty \mathbf{1}_{\beta_t^0 > B^{\bar{\mathbf{T}}}_{N_t^{\bar{\mathbf{T}}}}} B^{\bar{\mathbf{T}}}_{N_t^{\bar{\mathbf{T}}}} dN^{i,\bar{\mathbf{T}}}\Big)\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^M \int_0^\infty \Big(K(1-X_{t-}^{i,\beta}) + \mathbf{1}_{\beta_t^i > B^{i,\mathbf{T}}_{N^{i,\mathbf{T}}}} B^{i,\mathbf{T}}_{N^{i,\mathbf{T}}}\eta^{\mathbf{T}} + \mathbf{1}_{\beta_t^0 > B^{\bar{\mathbf{T}}}_{N_t^{\bar{\mathbf{T}}}}} B^{\bar{\mathbf{T}}}_{N_t^{\bar{\mathbf{T}}}}\eta^{\bar{\mathbf{T}}}\Big)dt\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^M \int_0^\infty \Big(K(1-X_{t-}^{i,\beta}) + \mathbb{E}[B_1^{1,\mathbf{T}}\mathbf{1}_{b>B_1^{1,\mathbf{T}}}]_{b:=\beta_t^i}\eta^{\mathbf{T}} + \mathbb{E}[B_1^{\bar{\mathbf{T}}}\mathbf{1}_{b>B_1^{\bar{\mathbf{T}}}}]_{b:=\beta_t^0}\eta^{\bar{\mathbf{T}}}\Big)dt\Big]
\end{aligned}$$

We can bound from below the part $\mathbb{E}[B_1^{1,\mathbf{T}}\mathbf{1}_{b>B_1^{1,\mathbf{T}}}]_{b:=\beta_t^i}$ by $(1-X_t^{\beta,i})\mathbb{E}[B_1^{1,\mathbf{T}}\mathbf{1}_{b>B_1^{1,\mathbf{T}}}]_{b:=\beta_t^i}$ and the part $\mathbb{E}[B_1^{\bar{\mathbf{T}}}\mathbf{1}_{b>B_1^{\bar{\mathbf{T}}}}]_{b:=\beta_t^0}$ by $\mathbf{1}_{p_t^\beta<1}\mathbb{E}[B_1^{\bar{\mathbf{T}}}\mathbf{1}_{b>B_1^{\bar{\mathbf{T}}}}]_{b:=\beta_t^0}$:

$$\begin{aligned}
&V(\beta) \\
&\geq \mathbb{E}\Big[\sum_{i=1}^M \int_0^\infty \Big(K(1-X_{t-}^{i,\beta}) + (1-X_{t-}^{i,\beta})\mathbb{E}[B_1^{1,\mathbf{T}}\mathbf{1}_{b>B_1^{1,\mathbf{T}}}]_{b:=\beta_t^i}\eta^{\mathbf{T}} + \mathbf{1}_{p_t^\beta<1}\mathbb{E}[B_1^{\bar{\mathbf{T}}}\mathbf{1}_{b>B_1^{\bar{\mathbf{T}}}}]_{b:=\beta_t^0}\eta^{\bar{\mathbf{T}}}\Big)dt\Big] \\
&= M\mathbb{E}\Big[\int_0^\infty \Big(K(1-p_t^\beta) + \frac{1}{M}\sum_{i=1}^M (1-X_{t-}^{i,\beta})\mathbb{E}[B_1^{1,\mathbf{T}}\mathbf{1}_{b>B_1^{1,\mathbf{T}}}]_{b:=\beta_t^i}\eta^{\mathbf{T}} + \mathbf{1}_{p_t^\beta<1}\mathbb{E}[B_1^{\bar{\mathbf{T}}}\mathbf{1}_{b>B_1^{\bar{\mathbf{T}}}}]_{b:=\beta_t^0}\eta^{\bar{\mathbf{T}}}\Big)dt\Big]
\end{aligned}$$

This is the first inequality of the proof. Afterward, we shall turn all the inequalities into equalities with a well chosen control. It is thus important to make sure that each inequality could be turned into an equality for some controls. For instance, here, the inequality

will clearly be an equality if the bidding map control makes null targeted advertising bids for individuals who are already informed, and null non-targeted advertising bids when the population is fully informed, which is obviously an efficient property. We can now use (7.6.2) with

$$H_t := \frac{K(1-p_t^\beta) + \frac{1}{M}\sum_{i=1}^M (1-X_{t-}^{i,\beta})\mathbb{E}[B_1^{1,\mathbf{T}}\mathbf{1}_{b>B_1^{1,\mathbf{T}}}]_{b:=\beta_t^i}\eta^{\mathbf{T}} + \mathbf{1}_{p_t^\beta<1}\mathbb{E}[B_1^{\bar{\mathbf{T}}}\mathbf{1}_{b>B_1^{\bar{\mathbf{T}}}}]_{b:=\beta_t^0}\eta^{\bar{\mathbf{T}}}}{(1-p_t^\beta)(\eta^{\mathbf{I}} + \alpha_t^\beta\eta^{\mathbf{T}} + \mathbb{P}(b\geq B_1^{\bar{\mathbf{T}}})_{b:=\beta_t^0}\eta^{\bar{\mathbf{T}}} + p_t^\beta\eta^{\mathbf{S}})}$$

with the convention that $\frac{0}{0}=0$. The process H is clearly predictible, positive, and bounded. We thus have

$$\begin{aligned} V(\beta) &\geq M\mathbb{E}\Big[\int_0^\infty H_t(1-p_t^\beta)(\eta^{\mathbf{I}} + \alpha_t^\beta\eta^{\mathbf{T}} + \mathbb{P}(b\geq B_1^{\bar{\mathbf{T}}})_{b:=\beta_t^0}\eta^{\bar{\mathbf{T}}} + p_t^\beta\eta^{\mathbf{S}})dt\Big] = M\mathbb{E}[\int_0^\infty H_t dp_t^\beta] \\ &= M\mathbb{E}\Big[\int_0^\infty \frac{K(1-p_t^\beta) + \frac{1}{M}\sum_{i=1}^M (1-X_{t-}^{i,\beta})\mathbb{E}[B_1^{1,\mathbf{T}}\mathbf{1}_{b>B_1^{1,\mathbf{T}}}]_{b:=\beta_t^i}\eta^{\mathbf{T}} + \mathbf{1}_{p_t^\beta<1}\mathbb{E}[B_1^{\bar{\mathbf{T}}}\mathbf{1}_{b>B_1^{\bar{\mathbf{T}}}}]_{b:=\beta_t^0}\eta^{\bar{\mathbf{T}}}}{(1-p_t^\beta)(\eta^{\mathbf{I}} + \alpha_t^\beta\eta^{\mathbf{T}} + \mathbb{P}(b\geq B_1^{\bar{\mathbf{T}}})_{b:=\beta_t^0}\eta^{\bar{\mathbf{T}}} + p_t^\beta\eta^{\mathbf{S}})} dp_t^\beta\Big] \end{aligned}$$

Now we can turn this cost into a sum over successive values of p_t^β:

$$V(\beta) \geq \mathbb{E}\Big[\sum_{p\in\frac{[\![0,M]\!]}{M}} \frac{K(1-p) + \frac{1}{M}\sum_{i=1}^M (1-X_{\tau_p^\beta}^{i,\beta})\mathbb{E}[B_1^{1,\mathbf{T}}\mathbf{1}_{b>B_1^{1,\mathbf{T}}}]_{b:=\beta^i_{\tau_p^\beta}}\eta^{\mathbf{T}} + \mathbb{E}[B_1^{\bar{\mathbf{T}}}\mathbf{1}_{b>B_1^{\bar{\mathbf{T}}}}]_{b:=\beta^0_{\tau_p^\beta}}\eta^{\bar{\mathbf{T}}}}{(1-p)(\eta^{\mathbf{I}} + \alpha^\beta_{\tau_p^\beta}\eta^{\mathbf{T}} + \mathbb{P}(b\geq B_1^{\bar{\mathbf{T}}})_{b:=\beta^0_{\tau_p^\beta}}\eta^{\bar{\mathbf{T}}} + p\eta^{\mathbf{S}})}\Big]$$

Where, for all $p\in\frac{[\![0,M]\!]}{M}$, τ_p^β is the time at which p^β reaches to p, i.e.,

$$\tau_p^\beta := \inf\{t\in\mathbb{R}_+ : p_t^\beta = p\}.$$

We now bound from below the inner fraction:

$$V(\beta) \geq \sum_{p\in\frac{[\![0,M]\!]}{M}} \inf_{\substack{b^{i,\mathbf{T}},b^{\bar{\mathbf{T}}}\in\mathbb{R} \\ i\in[\![1,M]\!]}} \frac{K(1-p) + \frac{1}{M}\sum_{i=1}^{M(1-p)}\mathbb{E}[B_1^{1,\mathbf{T}}\mathbf{1}_{b^{i,\mathbf{T}}>B_1^{1,\mathbf{T}}}]\eta^{\mathbf{T}} + \mathbb{E}[B_1^{\bar{\mathbf{T}}}\mathbf{1}_{b^{\bar{T}}>B_1^{\bar{\mathbf{T}}}}]\eta^{\bar{\mathbf{T}}}}{(1-p)\Big(\eta^{\mathbf{I}} + \frac{\sum_{i=1}^{M(1-p)}\mathbb{P}(b^{i,T}\geq B_1^{\mathbf{T}})}{M(1-p)}\eta^{\mathbf{T}} + \mathbb{P}(b^{\bar{T}}\geq B_1^{\bar{\mathbf{T}}})\eta^{\bar{\mathbf{T}}} + p\eta^{\mathbf{S}}\Big)} \tag{7.6.3}$$

Notice that this is the part explicitly suggesting that the problem reduces to a sum of local optimizations for all $p\in\frac{[\![0,M]\!]}{M}$. Now, notice that if we denote by B_p a random variable such that

$$B_p = ZB_1^{\mathbf{T}} + (1-Z)\frac{B_1^{\bar{\mathbf{T}}}}{1-p}$$

where $Z \perp\!\!\!\perp (B_1^{\mathbf{T}}, B_1^{\bar{\mathbf{T}}})$ and $Z\sim Bernoulli\left(\frac{\eta^{\mathbf{T}}}{\eta^{\mathbf{T}}+\eta^{\bar{\mathbf{T}}}}\right)$, we clearly have

$$V(\beta) \geq \sum_{p\in\frac{[\![0,M]\!]}{M}} \inf_{B\in L(\Omega,\mathbb{R})} \frac{\frac{K}{\eta^{\mathbf{T}}+\eta^{\bar{\mathbf{T}}}} + \mathbb{E}[B_p\mathbf{1}_{B>B_p}]}{\frac{\eta^{\mathbf{I}}+p\eta^{\mathbf{S}}}{\eta^{\mathbf{T}}+\eta^{\bar{\mathbf{T}}}} + \mathbb{P}(B\geq B_p)}$$

Where $L(\Omega, \mathbb{R})$ simply denotes the set of real random variables. Indeed, one retrieves the formula in (7.6.3) when

$$B = Z \sum_{i \in [\![1,M]\!]} b^{i,\mathbf{T}} \mathbf{1}_{U=i} + (1-Z) b^{\bar{\mathbf{T}}}$$

where $U \sim \mathcal{U}([\![1:M]\!])$ is a random uniform variable on $[\![1:M]\!]$ independent from the rest of the variables. Notice that a natural way to improve the choice of a given $B \in L(\Omega, \mathbb{R})$ in the above infimum is the following: under the constraint that $\mathbb{P}(\tilde{B} \geq B_p) = \mathbb{P}(B \geq B_p) =: P$, we can minimize $\mathbb{E}[B_p \mathbf{1}_{\tilde{B} > B_p}]$. It is clear that such $\tilde{B}$ is given by

$$\tilde{B} = \mathbf{1}_{U < \frac{P - F(F^{-1}(P))}{F(F^{-1}(P+)) - F(F^{-1}(P))}} F^{-1}(P) + \mathbf{1}_{U > \frac{P - F(F^{-1}(P))}{F(F^{-1}(P+)) - F(F^{-1}(P))}} F^{-1}(P+)$$

where $U \sim \mathcal{U}([0,1])$ and $F(x) = \mathbb{P}(c_p \leq x)$ is the distribution function of c_p. We thus have

$$\begin{aligned} V(\beta) &\geq \sum_{p \in \frac{[\![0,M]\!]}{M}} \inf_{P,\theta} \frac{\frac{K}{\eta^{\mathbf{T}} + \eta^{\mathbf{T}}} + \mathbb{E}[c_p \mathbf{1}_{\mathbf{1}_{U<\theta} F^{-1}(P) + \mathbf{1}_{U>\theta} F^{-1}(P_+) > c_p}]}{\frac{\eta^{\mathbf{I}} + p\eta^{\mathbf{S}}}{\eta^{\mathbf{T}} + \eta^{\bar{\mathbf{T}}}} + \mathbb{P}(\mathbf{1}_{U<\theta} F^{-1}(P) + \mathbf{1}_{U>\theta} F^{-1}(P_+) \geq c_p)} \\ &\geq \sum_{p \in \frac{[\![0,M]\!]}{M}} \inf_{P,\theta} \frac{\frac{K}{\eta^{\bar{\mathbf{T}}} + \eta^{\mathbf{T}}} + \theta \mathbb{E}[c_p \mathbf{1}_{F^{-1}(P) > c_p}] + (1-\theta)\mathbb{E}[c_p \mathbf{1}_{F^{-1}(P_+) > c_p}]}{\frac{\eta^{\mathbf{I}} + p\eta^{\mathbf{S}}}{\eta^{\mathbf{T}} + \eta^{\bar{\mathbf{T}}}} + \theta \mathbb{P}(F^{-1}(P) \geq c_p) + (1-\theta)\mathbb{P}(F^{-1}(P_+) \geq c_p)} \end{aligned}$$

However, notice that for a, b, c, d, we have

$$\frac{a + \theta b}{c + \theta d} = \frac{a - \frac{cb}{d}}{c + \theta d} + \frac{b}{d}$$

and thus the minimum in $\theta \in [0,1]$ is necessarily reached for either $\theta = 0$ or $\theta = 1$. Thus we have

$$V(\beta) \geq \sum_{p \in \frac{[\![0,M]\!]}{M}} \inf_{b \in \mathbb{R}} \frac{\frac{K}{\eta^{\bar{\mathbf{T}}} + \eta^{\mathbf{T}}} + \mathbb{E}[c_p \mathbf{1}_{b > c_p}]}{\frac{\eta^{\mathbf{I}} + p\eta^{\mathbf{S}}}{\eta^{\mathbf{T}} + \eta^{\bar{\mathbf{T}}}} + \mathbb{P}(b \geq c_p)}$$

We can express it back with the problem's original variables $B_1^{1,\mathbf{T}}$ and $B_1^{\bar{\mathbf{T}}}$:

$$V(\beta) \geq \sum_{p \in \frac{[\![0,M]\!]}{M}} \inf_{b \in \mathbb{R}} \frac{K + \eta^{\mathbf{T}} \mathbb{E}[B_1^{1,\mathbf{T}} \mathbf{1}_{b > B_1^{1,\mathbf{T}}}] + \eta^{\bar{\mathbf{T}}} \mathbb{E}[\frac{B_1^{1,\mathbf{T}}}{1-p} \mathbf{1}_{b > \frac{B_1^{1,\mathbf{T}}}{1-p}}]}{\eta^{\mathbf{I}} + \eta^{\mathbf{T}} \mathbb{P}(b \geq B_1^{1,\mathbf{T}}) + \eta^{\bar{\mathbf{T}}} \mathbb{P}(b \geq B_1^{\bar{\mathbf{T}}}) + p\eta^{\mathbf{S}}}$$

We have now determined a lower bound of $\inf_b V(\beta)$. It is then simple to retrace this above derivation with the control $\beta^{\mathfrak{b}}$ associated to the minimal policy defined by $\mathfrak{b}^T(p) = B_\star(p)$ and $\mathfrak{b}^{\bar{T}}(p) = (1-p)B_\star(p)$, where

$$B_\star(p) = \underset{b\in\mathbb{R}}{\operatorname{argmin}} \frac{K + \mathbb{E}[\frac{B_1^{\bar{\mathbf{T}}}}{1-p}\mathbf{1}_{b>\frac{B_1^{\bar{\mathbf{T}}}}{1-p}}]\eta^{\bar{\mathbf{T}}} + \mathbb{E}[B_1^{1,\mathbf{T}}\mathbf{1}_{b>B_1^{1,\mathbf{T}}}]\eta^{\mathbf{T}}}{\eta^{\mathbf{I}} + \eta^{\mathbf{T}}\mathbb{P}(b \geq B_1^{1,\mathbf{T}}) + \eta^{\bar{\mathbf{T}}}\mathbb{P}(b \geq \frac{B_1^{\bar{\mathbf{T}}}}{1-p}) + p\eta^{\mathbf{S}}},$$

and notice that in this case, all the inequalities turn into identities. □

7.6.2 Proof for social marketing with discount factor

Let us fix an open-loop bidding control β. We have

$$\begin{aligned}
V(\beta) &= \mathbb{E}\Big[\int_0^\infty e^{-\rho t}(K(1-X_{t^-}^b)dN_t^E + \mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N_t^{\mathbf{T}}}} B^{\mathbf{T}}_{N_t^{\mathbf{T}}} dN^{\mathbf{T}})\Big] \\
&\geq \mathbb{E}\Big[\int_0^\infty e^{-\rho t}(K(1-X_{t-}^\beta) + (1-X_{t-}^\beta)\mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N_t^{\mathbf{T}}}} B^{\mathbf{T}}_{N_t^{\mathbf{T}}}\eta^{\mathbf{T}})dt\Big] \\
&= \mathbb{E}\Big[\int_0^\infty e^{-\rho t}(1-X_{t-}^\beta)(K + \mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N_t^{\mathbf{T}}}} B^{\mathbf{T}}_{N_t^{\mathbf{T}}}\eta^{\mathbf{T}})dt\Big],
\end{aligned}$$

This first inequality will become an equality if the bidding control b makes null bids once the individual is informed.

Notice that, up to removing the discount factor, this problem is a particular case of previous one. The remaining part of the proof consists in getting rid of this discount factor and then referring to the result in the non discounted case. A discount factor can always be interpreted as a devaluation due to the possibility of an unpredictable termination event happening at an exponential time with parameter $\ln(\beta)$. Indeed, given such random exponential time τ independent from the existing random variables, we have

$$\begin{aligned}
V(\beta) &\geq \mathbb{E}\Big[\int_0^\infty e^{-\rho t}(1-X_{t-}^\beta)(K + \mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N_t^{\mathbf{T}}}} B^{\mathbf{T}}_{N_t^{\mathbf{T}}}\eta^{\mathbf{T}})dt\Big] \\
&= \mathbb{E}\Big[\int_0^\infty \mathbb{P}(\tau > t)(1-X_{t-}^\beta)(K + \mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N_t^{\mathbf{T}}}} B^{\mathbf{T}}_{N_t^{\mathbf{T}}}\eta^{\mathbf{T}})dt\Big] \\
&= \mathbb{E}\Big[\int_0^\infty \mathbf{1}_{\tau>t}(1-X_{t-}^\beta)(K + \mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N_t^{\mathbf{T}}}} B^{\mathbf{T}}_{N_t^{\mathbf{T}}}\eta^{\mathbf{T}})dt\Big] \\
&= \mathbb{E}\Big[\int_0^\tau (1-X_{t-}^\beta)(K + \mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N_t^{\mathbf{T}}}} B^{\mathbf{T}}_{N_t^{\mathbf{T}}}\eta^{\mathbf{T}})dt\Big]
\end{aligned}$$

We introduce a Poisson process M with intensity $\ln(\beta)$, first time of jump given by τ, and independent of the other random variables, and we denote the process $\tilde{X}^\beta$ satisfying

the dynamic

$$\begin{aligned}\tilde{X}_0^\beta &= 0\\ d\tilde{X}_t^\beta &= (1-\tilde{X}_{t^-}^\beta)(dN_t^{\mathbf{I}} + dN_t + \mathbf{1}_{\beta_t \geq B^{\mathbf{T}}_{N_t^{\mathbf{T}}}} dN_t^{\mathbf{T}})\end{aligned}$$

Notice that $\tilde{X}^\beta$ has exactly the same dynamic as X^β except that there is an additional cause of transition to state 1 given by the term dN. It is then clear that we have

$$\begin{aligned}\mathbb{E}\Big[\int_0^\tau (1-X_{t-}^\beta)(K+\mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N_t^{\mathbf{T}}}} B^{\mathbf{T}}_{N_t^{\mathbf{T}}}\eta^{\mathbf{T}})dt\Big] &= \mathbb{E}\Big[\int_0^\infty (1-\tilde{X}_t^\beta)(K+\mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N_t^{\mathbf{T}}}} B^{\mathbf{T}}_{N_t^{\mathbf{T}}}\eta^{\mathbf{T}})dt\Big]\\ &= \mathbb{E}\Big[\int_0^\infty (K(1-\tilde{X}_{t^-}^\beta)dN_t^E + \mathbf{1}_{\tilde{\beta}_t > B^{\mathbf{T}}_{N_t^{\mathbf{T}}}} B^{\mathbf{T}}_{N_t^{\mathbf{T}}} dN_t^{\mathbf{T}})\Big]\end{aligned}$$

where $\tilde{\beta}_t = (1-\tilde{X}_{t^-}^\beta)\beta_t$. By noting $\tilde{N}^{\mathbf{I}} = N^{\mathbf{I}} + N$, we obtain a Poisson process $\tilde{N}^{\mathbf{I}}$ with intensity $\eta+\rho$, and we have

$$\begin{aligned}\tilde{X}_0^\beta &= 0\\ d\tilde{X}_t^\beta &= (1-\tilde{X}_{t^-}^\beta)(d\tilde{N}_t^{\mathbf{I}} + \mathbf{1}_{\beta_t \geq B^{\mathbf{T}}_{N^{\mathbf{T}}}} dN_t^{\mathbf{T}})\end{aligned}$$

The cost $V(\beta)$ is thus expressed as the cost associated to the bidding map control $\beta = (0,\tilde{b})$ in the previous problem with a population with $M=1$, i.e. a single individual, and where $\eta^{\bar{\mathbf{T}}} = \eta^{\mathbf{S}} = 0$, i.e. the individual never connects to a website displaying non-targeted ads, and individuals do not socially interact. We know that $V(\beta)$ is thus bounded from below as follows:

$$V(\beta) \geq \inf_{b\in\mathbb{R}} \frac{K+\eta^{\mathbf{T}}\mathbb{E}[B_1^{\mathbf{T}}\mathbf{1}_{b>B_1^{1,\mathbf{T}}}]}{\eta^{\mathbf{I}}+\rho+\eta^{\mathbf{T}}\mathbb{P}(b\geq B_1^{\mathbf{T}})}$$

It is then simple to retrace this derivation with the particular bidding control $\beta^{b_\star}$ associated to the constant bidding policy $b_\star$ such that

$$b_\star = \operatorname*{argmin}_{b\in\mathbb{R}} \frac{K+\eta^{\mathbf{T}}\mathbb{E}[B_1^{\mathbf{T}}\mathbf{1}_{n>B_1^{\mathbf{T}}}]}{\eta^{\mathbf{I}}+\rho+\eta^{\mathbf{T}}\mathbb{P}(b\geq B_1^{\mathbf{T}})},$$

and to turn inequalities into equalities. This concludes the result for this case. □

7.6.3 Proof for commercial advertising with purchase-based reward

The idea is to reduce to the previous case. Given an open-loop bidding control β, we have

$$\begin{aligned} V(\beta) &= \mathbb{E}\Big[e^{-\rho\tau^\beta}K - \int_0^\infty e^{-\rho t}\mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N^{\mathbf{T}}_t}} B^{\mathbf{T}}_{N^{\mathbf{T}}_t} dN^{\mathbf{T}}_t)\Big] \\ &= \mathbb{E}\Big[\int_{\tau^\beta}^\infty \rho e^{-\rho t}K - \int_0^\infty e^{-\rho t}\mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N^{\mathbf{T}}_t}} B^{\mathbf{T}}_{N^{\mathbf{T}}_t} dN^{\mathbf{T}}_t)\Big] \\ &= \mathbb{E}\Big[\int_0^\infty e^{-\rho t}\rho K X^\beta_{t-}dt - \int_0^\infty e^{-\rho t}\mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N^{\mathbf{T}}_t}} B^{\mathbf{T}}_{N^{\mathbf{T}}_t} dN^{\mathbf{T}}_t)\Big] \end{aligned}$$

The problem is thus reduced to a continuous gain problem, with continuous reward ρK from the time of information. We shall now turn the continuous gain problem into a continuous cost problem as follows:

$$\begin{aligned} V(\beta) &= \mathbb{E}\Big[\int_0^\infty e^{-\rho t}(\rho K X^\beta_{t-}dt - \mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N^{\mathbf{T}}_t}} B^{\mathbf{T}}_{N^{\mathbf{T}}_t} dN^{\mathbf{T}}_t)\Big] \\ &= K - \mathbb{E}\Big[\int_0^\infty e^{-\rho t}(\rho K(1 - X^\beta_{t-})dt + \mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N^{\mathbf{T}}_t}} B^{\mathbf{T}}_{N^{\mathbf{T}}_t} dN^{\mathbf{T}}_t)\Big] \\ &= K - \mathbb{E}\Big[\int_0^\infty e^{-\rho t}(\rho K(1 - X^\beta_{t-})dN^E_t + \mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N^{\mathbf{T}}_t}} B^{\mathbf{T}}_{N^{\mathbf{T}}_t} dN^{\mathbf{T}}_t)\Big] \end{aligned}$$

We are reduced to the previous case (social marketing with discount factor). This concludes the proof. □

7.6.4 Proof for commercial advertising with subscription-based reward

Given an open-loop bidding control, we have

$$\begin{aligned} V(\beta) &= \mathbb{E}\Big[\sum_{k\in\mathbb{N}} e^{-\rho(\tau_b + k)}K + \int_0^\infty e^{-\rho t}\mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N^{\mathbf{T}}_t}} B^{\mathbf{T}}_{N^{\mathbf{T}}_t} dN^{\mathbf{T}}_t)\Big] \\ &= \mathbb{E}\Big[e^{-\rho\tau_b}\frac{K}{1 - e^{-\rho}} + \int_0^\infty e^{-\rho t}\mathbf{1}_{\beta_t > B^{\mathbf{T}}_{N^{\mathbf{T}}_t}} B^{\mathbf{T}}_{N^{\mathbf{T}}_t} dN^{\mathbf{T}}_t)\Big] \end{aligned}$$

This reduces the problem to the previous case and concludes the proof. □

7.7 Conclusion

In this work, we have developed several targeted advertising models with semi-explicit solutions. The advantage of these models is that they describe the advertising situation in a very concrete way: one or more individuals are really modeled, and their behaviors

are concretely described, involving connections to various types of websites at random times as well as social interactions. The advertising auctions are also precisely defined, even allowing to consider various auction rules (second-price auctions, first-price auctions). There is however still room for exploration to enrich the models while keeping them tractable with semi-explicit solutions: in the fourth model with an interacting population, it might be possible to add a bit of heterogeneity in the population connections and social interactions. Another possible development, regarding the auctions, could be to model the maximal bid from other bidders more realistically than with an i.i.d. sequence of random variables. A possible generalization could be, for instance, to model other bidders' maximal bid as a Markov process, but another approach could also be to explicitly model several bidding agents, for instance playing according to the so-called fictitious play principle, instead of modeling the other bidders' maximal bid in an exogenous way.

Bibliography

[1] M. Abdellaoui. Parameter-free elicitation of utility and probability weighting functions. *Management science*, 46(11):1497–1512, 2000.

[2] D. Acemoglu and A. Ozdaglar. Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, 1(1):3–49, 2011.

[3] M. Aizenman, R. Sims, and S. L. Starr. Extended variational principle for the sherrington-kirkpatrick spin-glass model. *Physical Review B*, 68(21):214403, 2003.

[4] . Anthony and P. L. Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[5] R. J. Aumann. Markets with a continuum of traders. *Econometrica: Journal of the Econometric Society*, pages 39–50, 1964.

[6] R. J Aumann. Values of markets with a continuum of traders. *Econometrica: Journal of the Econometric Society*, pages 611–646, 1975.

[7] E. Bayraktar, A. Cosso, and H. Pham. Randomized dynamic programming principle and feynman-kac representation for optimal control of mckean-vlasov dynamics. *Transactions of the American Mathematical Society*, 370(3):2115–2160, 2018.

[8] A. Bensoussan, J. Frehse, P. Yam, et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

[9] J. Berkson. Application of the logistic function to bio-assay. *Journal of the American statistical association*, 39(227):357–365, 1944.

[10] C. Bernard, X. He, J.-A. Yan, and X. Y. Zhou. Optimal insurance design under rank-dependent expected utility. *Mathematical Finance*, 25(1):154–186, 2015.

[11] D. P. Bertsekas. *Dynamic programming and optimal control 3rd edition, volume II*. 2011.

[12] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[13] G. Carmona and K. Podczeck. Ex-post stability of bayes–nash equilibria of large games. *Games and Economic Behavior*, 74(1):418–430, 2012.

[14] P. Carmona and Y. Hu. Universality in sherrington–kirkpatrick's spin glass model. 42(2):215–222, 2006.

[15] R. Carmona and F. Delarue. *Probabilistic Theory of Mean Field Games with Applications II: Mean Field Games with Common Noise and Master Equations*, volume 84. Springer, 2018.

[16] R. Carmona, M. Laurière, and Z. Tan. Model-free mean-field reinforcement learning: mean-field mdp and mean-field q-learning. *arXiv preprint arXiv:1910.12802*, 2019.

[17] Yi-Chun Chen, Ngo Van Long, and Xiao Luo. Iterated strict dominance in general games. *Games and Economic Behavior*, 61(2):299–315, 2007.

[18] S. H. Chew and L. G. Epstein. A unifying approach to axiomatic non-expected utility theories. *Journal of Economic Theory*, 49(2):207–240, 1989.

[19] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[20] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

[21] W. Darity. Keynes' principle of effective demand. by edward j. amadeo. aldershot: Edward elgar publishing limited, 1989. pp. 189. 42.75. *The Journal of Economic History*, 52(1):257–258, 1992.

[22] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[23] M. F. Djete. Extended mean field control problem: a propagation of chaos result. *arXiv preprint arXiv:2006.12996*, 2020.

[24] M. F. Djete, D. Possamai, and X. Tan. Mckean-vlasov optimal control: the dynamic programming principle. *arXiv preprint arXiv:1907.08860*, 2019.

[25] R. O. Duda, P. E. Hart, et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

[26] M. Dufwenberg and M. Stegeman. Existence and uniqueness of maximal reductions under iterated strict dominance. *Econometrica*, 70(5):2007–2023, 2002.

[27] G. Feichtinger, R. F. Hartl, and S. P. Sethi. Dynamic optimal control models in advertising: recent developments. *Management Science*, 40(2):195–226, 1994.

[28] J. Fontbona, H. Guérin, and S. Méléard. Measurability of optimal transportation and strong coupling of martingale measures. *Electronic communications in probability*, 15:124–133, 2010.

[29] M. Fornasier, S. Lisini, C. Orrieri, and G. Savaré. Mean-field optimal control as gamma-limit of finite agent controls. *European Journal of Applied Mathematics*, 30(6):1153–1186, 2019.

[30] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

[31] F. Gozzi, C. Marinelli, and S. Savin. On controlled linear diffusions with delay in a model of optimal advertising under uncertainty with memory effects. *Journal of optimization theory and applications*, 142(2):291–321, 2009.

[32] J. R. Green and B. Jullien. Ordinal independence in nonlinear utility theory. *Journal of risk and uncertainty*, 1(4):355–387, 1988.

[33] H. Gu, X. Guo, X. Wei, and R. Xu. Dynamic programming principles for learning mfcs. *arXiv preprint arXiv:1911.07314*, 2019.

[34] O. Guéant, J.-M. Lasry, and P.-L. Lions. Mean field games and applications. In *Paris-Princeton lectures on mathematical finance 2010*, pages 205–266. Springer, 2011.

[35] F. Guerra. Broken replica symmetry bounds in the mean field spin glass model. *Communications in mathematical physics*, 233(1):1–12, 2003.

[36] Francesco Guerra and Fabio Lucio Toninelli. Infinite volume limit and spontaneous replica symmetry breaking in mean field spin glass models. In *International Conference on Theoretical Physics*, pages 441–444. Springer, 2003.

[37] M. Huang, P. E. Caines, and R. P. Malhamé. Large-population cost-coupled lqg problems with nonuniform agents: individual-mass behavior and decentralized *varepsilon*-nash equilibria. *IEEE transactions on automatic control*, 52(9):1560–1571, 2007.

[38] M. Huang, R. P. Malhamé, P. E. Caines, et al. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. *Communications in Information & Systems*, 6(3):221–252, 2006.

[39] A. Jack, T. C. Johnson, and M. Zervos. A singular control model with application to the goodwill problem. *Stochastic processes and their applications*, 118(11):2098–2124, 2008.

[40] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[41] A. T. Kalai and E. Kalai. Cooperation and competition in strategic games with private information. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 345–346, 2010.

[42] E. Kalai. Large robust games. *Econometrica*, 72(6):1631–1665, 2004.

[43] E. Kalai and M. Smorodinsky. Other solutions to nash's bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 513–518, 1975.

[44] O. Kallenberg. *Foundations of Modern Probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

[45] M. J. Kearns, U. V. Vazirani, and U. Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

[46] H. Kesten and R. H. Schonmann. Behavior in large dimensions of the potts and heisenberg models. *Reviews in Mathematical Physics*, 1(02n03):147–182, 1989.

[47] M. A. Khan and Y. Sun. Non-cooperative games with many players. *Handbook of game theory with economic applications*, 3:1761–1808, 2002.

[48] S. R. Kulkarni, G. Lugosi, and S. S. Venkatesh. Learning pattern classification-a survey. *IEEE Transactions on Information Theory*, 44(6):2178–2206, 1998.

[49] D. Lacker. Limit theory for controlled mckean–vlasov dynamics. *SIAM Journal on Control and Optimization*, 55(3):1641–1672, 2017.

[50] J.-M. Lasry and P.-L. Lions. Jeux à champ moyen. i–le cas stationnaire. *Comptes Rendus Mathématique*, 343(9):619–625, 2006.

[51] J.-M. Lasry and P.-L. Lions. Jeux à champ moyen. ii–horizon fini et contrôle optimal. *Comptes Rendus Mathématique*, 343(10):679–684, 2006.

[52] J.-M. Lasry and P.-L. Lions. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

[53] M. Laurière and O. Pironneau. Dynamic programming for mean-field type control. *Comptes Rendus Mathematique*, 352(9):707–713, 2014.

[54] S. Lichtenstein and P. Slovic. Reversals of preference between bids and choices in gambling decisions. *Journal of experimental psychology*, 89(1):46, 1971.

[55] H. R Lindman. Inconsistent preferences among gambles. *Journal of Experimental Psychology*, 89(2):390, 1971.

[56] J. D. C. Little and L. M. Lodish. A media planning calculus. *Operations Research*, 17(1):1–35, 1969.

[57] W.-Y. Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.

[58] R. D. Luce. *Utility of gains and losses: Measurement-theoretical and experimental approaches*. Psychology Press, 2014.

[59] R. D. Luce and P. C. Fishburn. Rank-and sign-dependent linear utility models for finite first-order gambles. *Journal of risk and Uncertainty*, 4(1):29–59, 1991.

[60] G. Lugosi. Pattern classification and learning theory. In *Principles of nonparametric learning*, pages 1–56. Springer, 2002.

[61] M. J. Machina. " expected utility" analysis without the independence axiom. *Econometrica: Journal of the Econometric Society*, pages 277–323, 1982.

[62] G. J. McLachlan. Discriminant analysis and statistical pattern recognition. 544, 2004.

[63] S. Mendelson. A few notes on statistical learning theory. pages 1–40, 2003.

[64] P. Milgrom and J. Roberts. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica: Journal of the Econometric Society*, pages 1255–1277, 1990.

[65] M. Motte and H. Pham. Mean-field markov decision processes with common noise and open-loop controls. *To appear in Annals of Applied Probability*, 2019.

[66] J. Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.

[67] J. F. Nash. *8. Two-Person Cooperative Games*. Princeton University Press, 2016.

[68] J. F. Nash et al. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

[69] B. K. Natarajan. Machine learning: A theoretical approach. 2014.

[70] M. Nerlove and K. J. Arrow. Optimal advertising policy under dynamic conditions. *Economica*, pages 129–142, 1962.

[71] H. Pham and X. Wei. Discrete time mckean–vlasov control problem: a dynamic programming approach. *Applied Mathematics & Optimization*, 74(3):487–506, 2016.

[72] H. Pham and X. Wei. Dynamic programming for optimal control of stochastic mckean–vlasov dynamics. *SIAM Journal on Control and Optimization*, 55(2):1069–1101, 2017.

[73] D. Prelec. The probability weighting function. *Econometrica*, pages 497–527, 1998.

[74] J. Quiggin. A theory of anticipated utility. *Journal of Economic Behavior & Organization*, 3(4):323–343, 1982.

[75] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media, 1998.

[76] A. Rubinstein. Similarity and decision-making under risk (is there a utility theory resolution to the allais paradox?). *Journal of economic theory*, 46(1):145–153, 1988.

[77] U. Segal and A. Spivak. First order versus second order risk aversion. *Journal of Economic Theory*, 51(1):111–125, 1990.

[78] S. P. Sethi. Dynamic optimal control models in advertising: a survey. *SIAM review*, 19(4):685–725, 1977.

[79] D. Sherrington and S. Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975.

[80] C. Starmer and R. Sugden. Violations of the independence axion in common ratio problems: An experimental test of some competing hypotheses. *Annals of Operations Research*, 19(1):79–102, 1989.

[81] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[82] M. Talagrand. The parisi formula. *Annals of mathematics*, pages 221–263, 2006.

[83] A. Tversky and C. R Fox. Weighing risk and uncertainty. *Psychological review*, 102(2):269, 1995.

[84] A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

[85] V. Vapnik. Estimation of dependences based on empirical data. 2006.

[86] V. Vapnik. The nature of statistical learning theory. 2013.

[87] V. N. Vapnik. Direct methods in statistical learning theory. pages 225–265, 2000.

[88] W. Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

[89] M. L. Vidale and H. B. Wolfe. An operations-research study of sales response to advertising. *Operations research*, 5(3):370–381, 1957.

[90] C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[91] J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton university press, 2007.

[92] P. Wakker. Separating marginal utility and probabilistic risk aversion. *Theory and decision*, 36(1):1–44, 1994.

[93] P. Wakker, I. Erev, and E. U. Weber. Comonotonic independence: The critical test between classical and rank-dependent utility theories. *Journal of Risk and Uncertainty*, 9(3):195–230, 1994.

[94] P. Weiss. L'hypothèse du champ moléculaire et la propriété ferromagnétique. *J. Phys. Theor. Appl.*, 6(1):661–690, 1907.

[95] F.-Y. Wu. The potts model. *Reviews of modern physics*, 54(1):235, 1982.

[96] Z. Q. Xu, X. Y. Zhou, and S. C. Zhuang. Optimal insurance under rank-dependent utility and incentive compatibility. *Mathematical Finance*, 29(2):659–692, 2019.

[97] M. E. Yaari. The dual theory of choice under risk. *Econometrica: Journal of the Econometric Society*, pages 95–115, 1987.

[98] F. S. Zufryden. Optimal multi-period advertising budget allocation within a competitive environment. *Journal of the Operational Research Society*, 26(4):743–754, 1975.