UMA ABORDAGEM VIA FUNÇÕES DE LIAPUNOV COM CONTROLE À DINÂMICA DE APRENDIZADO EM JOGOS DE DOIS JOGADORES

Rodrigo Brandolt Sodré de Macedo

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Elétrica.

Orientador: Amit Bhaya

Rio de Janeiro
Dezembro de 2010

UMA ABORDAGEM VIA FUNÇÕES DE LIAPUNOV COM CONTROLE À DINÂMICA DE APRENDIZADO EM JOGOS DE DOIS JOGADORES

Rodrigo Brandolt Sodré de Macedo

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Amit Bhaya, Ph.D.

Prof. Daniel Ratton Figueiredo, Ph.D.

Prof. Eduardo Camponogara , Ph.D.

Prof. Eugenius Kaszkurewicz, D. Sc.

Prof. João Carlos dos Santos Basilio, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
DEZEMBRO DE 2010

Macedo, Rodrigo Brandolt Sodré de

Uma abordagem via funções de Liapunov com controle à dinâmica de aprendizado em jogos de dois jogadores/Rodrigo Brandolt Sodré de Macedo. – Rio de Janeiro: UFRJ/COPPE, 2010.

XV, 99 p.: il.; 29, 7cm.

Orientador: Amit Bhaya

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2010.

Referências Bibliográficas: p. 92 – 99.

1. teoria de jogos. 2. controle chaveado. 3. aprendizado por reforço. I. Bhaya, Amit. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Aos meus pais Marli e Valmir pelo amparo e exemplo ao longo dos anos e à minha irmã Kátia pelo companheirismo. A minha mulher Márcia pelo amor, tenacidade e paciência e aos nossos filhos, Leonardo e Maria Beatriz, nossos maiores tesouros.

Creio:

- *no esforço construtivo continuado, com intensidade, frequência e duração;*
- *na lucidez permanente como oposição a qualquer tipo de fuga;*
- *na aprendizagem, mais ainda nas adversidades, para a construção natural evolutiva e racional;*
- *na capacidade humana de transmissão de informações por meio de pensamento, formando uma rede telepática;*
- *no trabalho e no estudo;*
- *na estrutura familiar forte como base para a educação;*
- *na caridade que não cause dependência ou preguiça;*
- *no ato amoroso;*
- *em Deus.*

Agradecimentos

Agradeço ao Exército Brasileiro em especial aos seguintes membros do Centro Tecnológico do Exército: Exmo Sr. General-de-Divisão João Edison Minnicelli M.Sc, Chefe do CTEx, Sr. Cel QEM Hildo Vieira Prado Filho M.Sc, Subchefe do CTEx, Sr. TC QEM Antônio Marcos Yuan, Chefe da Divisão Bélica, Sr. Maj QEM Marco Antonio Alvares dos Prazeres M.Sc, Chefe do Grupo de Mísseis e Foguetes e demais membros do Grupo de Mísseis e Foguetes que me apoiaram. Aos seguintes membros do Núcleo de Computação de Alto Desempenho da Universidade Federal do Rio de Janeiro: Helenice Henderson, Fernando Pazos D.Sc, Leonardo Valente Ferreira D.Sc, Mara Prata, Myrian Christina de Aragão Costa D.Sc, Omar Diene D.Sc, Orlando Caldas, Paula Aída Sesini D.Sc, Ricardo Padilha Pareto, Sandra Carneiro da Silva, Valeriana Roncero D.Sc, pela oportunidade de convívio e crescimento profissional e em especial ao professor Amit Bhaya Ph.D, que soube me conduzir durante todo o trabalho com paciência irrestrita. Agradeço também aos companheiros de transporte solidário até o CTEx: Clayton Scouper das Chagas M.Sc, Felipe Aurélio Caetano de Bastos D.Sc, Ricardo Andrade, Víctor Santoro Santiago D.Sc, que tornaram menores as distâncias. A Fernando Apolinário Pereira M.Sc companheiro do CTEx, que inicia o trabalho de doutorado enquanto termino, pela troca de informações em problemas em comum. Enfim, agradeço a Deus: por me confortar quando as dificuldades pareciam grandes e me impulsionar sempre em frente; pela oportunidade de exercitar a paciência, a tenacidade, a concentração, o equilíbrio e a humildade; por proporcionar o desenvolvimento de meus filhos ao mesmo tempo que o doutorado, que se, a primeira vista, me pareceu dividir as forças, logo a seguir me serviu de exemplo de força e de superação.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

UMA ABORDAGEM VIA FUNÇÕES DE LIAPUNOV COM CONTROLE À DINÂMICA DE APRENDIZADO EM JOGOS DE DOIS JOGADORES

Rodrigo Brandolt Sodré de Macedo

Dezembro/2010

Orientador: Amit Bhaya

Programa: Engenharia Elétrica

Esta tese foca na dinâmica de aprendizado de jogos, especificamente nas várias propostas recentes da literatura em jogos de duas ações, dois jogadores. Essas propostas, conhecidas por gradiente ascendente incremental (IGA) e "vença ou aprenda rápido" (WoLF-IGA), utilizam uma forma de gradiente ascendente ao longo de uma função de valor, chamada também de função recompensa, junto com algumas leis de chaveamento heurísticas. A tese propõe o uso de função de Liapunov de controle (FLC) para desenvolver projetos de aprendizado envolvendo chaveamento, que é típico quando FLC são usadas em sistemas não lineares, permitindo a unificação de todas as propostas recentes na área de aprendizado por gradiente ascendente em jogos, também provendo provas rigorosas, inexistentes em muitas dessas propostas, de convergência ao equilíbrio misto. Ademais, a perspectiva de controle também conduz a generalizações dessas propostas, primeiro propondo novas leis de chaveamento que podem levar a melhoria de desempenho. E mais, com exceção de uma das propostas existentes, chamada de heurística de aprendizado de política ponderada (WPL), todas as propostas anteriores assumem que o jogo subjacente, isto é, ambas matrizes de recompensa, são conhecidas de ambos jogadores. Nesse respeito, a tese examina WPL e mostra que, pela introdução do conceito de equilíbrio virtual, essa política também pode ser considerada como um projeto de controle chaveado usando FLC, chegando-se a uma prova rigorosa. Além disso, a técnica de estimação de mínimos quadrados é introduzida de modo a estimar a combinação requerida de parâmetros de matrizes de recompensa que surgem em outra proposta recente, chamada aprendizado probabilístico de Boltzmann, que mostra que uma FLC adequada permite o projeto de estratégias que convergem ao equilíbrio de Nash misto desejado, em jogos padrão tal como o jogo casar moedas.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

A CONTROL LIAPUNOV FUNCTION PERSPECTIVE ON THE DYNAMICS OF LEARNING IN TWO PLAYER GAMES

Rodrigo Brandolt Sodré de Macedo

December/2010

Advisor: Amit Bhaya

Department: Electrical Engineering

This thesis focuses on the dynamics of learning in games – more specifically on the various recent proposals in the literature on two player, two action games. These proposals, known by names such as incremental gradient ascent (IGA) and Win or Learn Fast IGA (WoLF-IGA), utilize some form of gradient ascent along a value function, also known as a payoff or reward function, together with some heuristic switching laws. The thesis proposes the use of a control Liapunov function to design learning schemes involving switching, which is typical when control Liapunov functions are used in nonlinear systems, allowing the unification of all recent proposals in the area of gradient ascent learning in games, also providing rigorous proofs, missing in many of these proposals, of convergence to mixed equilibria in two-player, two-action games. In addition, the control perspective also leads naturally to generalizations of these proposals, by proposing new switching laws that can lead to improved performance. Furthermore, with the exception of one of the existing proposals, called weighted policy learning (WPL) heuristic, all previous proposals assume that the underlying game, i.e., both payoff matrices, is known to both players. In this respect, this thesis examines WPL and shows that, by the introduction of the concept of virtual equilibria, this policy can also be considered as a control Liapunov design and given a rigorous convergence proof. In addition, a technique of least squares estimation is introduced in order to estimate the required combination of payoff matrix parameters that arise in another recent proposal, called Boltzmann probabilistic learning, and it is shown, once again, that a suitable control Liapunov function allows the design of strategies that converge to the desired mixed Nash equilibrium of the standard two player two action games, such as matching pennies, that are used as benchmarks in the literature.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AR	Aprendizado por Reforço
ARM	Aprendizado por Reforço Multiagente (ARM)
ARM-GA	Aprendizado por Reforço Multiagente por Gradiente Ascendente (ARM-GA)
BPL	*Boltzmann Probabilities Learning* - Aprendizado com Probabilidades de Boltzmann
DR	Dinâmica do Replicador
EN	Equilíbrio de Nash
FLC	Funções de Liapunov com Controle
GIGA	*Generalized Infinitesimal Gradient Ascent* - Gradiente Ascendente Infinitesimal Generalizado
HAL	*Heuristically Accelerated Q-Learning* - Aprendizado Acelerado por Heurísticas
HEGS	*Hyperbolic Elliptic Gradient Switching* - Gradiente Chaveamento Hiperbólico Elíptico
IA	Inteligência Artificial
IGA	*Infinitesimal Gradient Ascent* - Gradiente Ascendente Infinitesimal
LSE	*Least Square Error* - Erros Mínimos Quadrados
MAB	*Multi-Armed Bandit* - máquina caça-níquel tipo bandido com múltiplos braços
MAR-EC	Método Aprendizado por Reforço com Estimação Corrente
MAR-EP	Método de Aprendizado por Reforço com Estimação Preliminar
P2P	*Peer-to-Peer* - Ponto a Ponto
WoLF	*Win or Learn Fast* - Ganhar ou Aprender Rapidamente
WPL	*Weighted Policy Learning*- Algoritmo de Aprendizado com Ponderação das Políticas

Capítulo 1

Introdução

1.1 Introdução geral

A teoria do jogos formaliza a interação entre agentes de um sistema e o processo de tomada de decisão. Tais situações ocorrem em conflitos armados, por exemplo num embate aéreo, na economia em competições pela fixação de preços de produtos ou ainda nas áreas de biologia, política e engenharia. A dificuldade de cada agente em decidir reside no fato de cada ação dele influenciar a reação dos outros.

O primeiro modelo de jogo cooperativo para n pessoas também é apresentado por NEUMANN e MORGENSTERN (1944), onde a teoria de soluções é baseada na formulação da função característica. Vários outros conceitos de solução foram definidos para teoria de jogos cooperativos. Em 1950, o equilíbrio de Nash (NASH, 1950) foi enunciado e NASH (1951) define a estratégia posteriormente denominada de equilíbrio de Nash.

De acordo com MANNOR e SHAMMA (2007), mais recentemente, a teoria de jogos tem sido estudada em diversos contextos distintos da origem histórica. Especificamente, a área de aprendizado multiagente, dentro da qual esta tese se situa, tem sido estudada por pesquisadores da área de inteligência artificial, dentro da ciência da computação, bem como da teoria de controle. O problema de projetar sistemas de controle ótimos (ou ao menos razoáveis) é conhecidamente difícil. Exemplos de problemas complexos de engenharia incluem planejamento em sistemas de manufaturas (AKELLA e KUMAR, 1986; GERSHWIN, 1994), roteamento em redes de dados (ALTMAN *et al.*, 2005; ALTMAN e SHIMKIN, 1998; ROUGHGARDEN, 2005), e comando e controle de forças em rede em ambientes adversos (AHUJA *et al.*, 2003; BEARD *et al.*, 2002). Essas aplicações abrangem uma coleção de componentes dispersos interagindo que buscam otimizar um objetivo coletivo global por meio de uma tomada de decisão local. A tarefa é complicada por limitações na capacidade de comunicação, na informação local e dinâmica, componentes defeituosos, e no incerto, se não hostil, ambiente. Em geral, não é viável passar toda informação para um centro de comando que poderia processar essa informação e disseminar instruções. Além disso, mesmo que isso fosse possível, a complexidade do sistema como um todo faria o problema de construir uma política ótima centralizada intratável.

É a complexidade e a natureza distribuída desses problemas que motivam o uso de métodos teóricos de jogos. Sob o ponto de vista multiagente, permite olhar o sistema todo como um conjunto de componentes simples interagindo. Nota-se que isso significa impor uma estrutura multiagente como uma abordagem de projeto. O resultado é que o processo de decisão para cada componente individual é ditado por um problema de otimização que é bastante simplificado se comparado ao problema centralizado, mas acoplado às decisões de outros componentes interconectados. Dessa forma, um equilíbrio de Nash reflete uma condição de otimização do

ponto de vista de cada componente individual, mas não necessariamente reflete a operação ótima do ponto de vista coletivo. No entanto, a possibilidade do sistema de se organizar num equilíbrio de Nash quase ótimo é menos desanimadora do que a perspectiva de construir uma política ótima centralizada.

É seguro dizer que muito da pesquisa em aprendizado multiagente tem raízes na literatura de teoria de jogos econômica. Por conseguinte, também é certo dizer que essa literatura não pretende oferecer uma metodologia para sistemas de engenharia. Isso não significa que aquele material, por exemplo as numerosas monografias excelentes em aprendizagem de jogos (YOUNG, 2006) ou jogos evolucionários (WEIBULL, 1995), não podem ser fonte de métodos de sistemas de engenharia. Ao contrário, destacam a importância de reconhecer qual é o problema quando considerando esse material para uma abordagem multiagente para projetar sistemas de engenharia.

Mencionam-se alguns aspectos que distinguem o aprendizado no contexto da engenharia. Em particular, vê-se que várias noções que tomam um papel descritivo ou preditivo na literatura econômica se tornam considerações de projeto quando considerando uma abordagem de aprendizado multiagente. Sugere-se que o aprendizado multiagente pode ser uma abordagem efetiva para os problemas descritos anteriormente. E, ainda, deve-se definir um jogo específico para aplicar métodos de aprendizado multiagente. Considera-se que uma motivação importante para se tomar uma perspectiva multiagente é minorar a complexidade. Isso significa que os elementos básicos dos jogadores, espaço de estratégias, e utilidades de jogadores são todas considerações de projeto quando se impõe uma abordagem multiagente. As operações dos agentes num meio reativo que muda com o tempo é uma realidade importante que uma metodologia de projeto deve lidar. O tipo de jogo que é escolhido para capturar tais dinâmicas pode ser um jogo do tipo *one-shot*, um jogo repetido ou um jogo estocástico. Escolher um jogo tipo *one-shot* implica que se ignora a dinâmica do problema. Um jogo repetido oferece uma estrutura mais intensa refletindo as consequências de uma estratégia sob múltiplos passos. Isso sugere técnicas de minimização de perda (ou arrependimento) como em AUER *et al.* (2002). Jogos estocásticos são um modelo mais natural para modelar dinâmica. Aprender nesse contexto é bastante complexo, como mostrado em HU e WELLMAN (2003), no contexto de chegar ao equilíbrio de Nash e em MANNOR e SHIMKIN (2003), no contexto de minimização de perda.

Finalmente, uma vez que jogadores e espaços de estratégia estão especificados, outra consideração de projeto é definir funções de recompensa (utilidade) dos agentes. Assim como a maioria das especficações de desempenho, há uma considerável complexidade em especificar funções de recompensas adequadas. Mesmo no caso ideal de um objetivo centralizado admitido, há considerações importantes em

"distribuir" esse objetivo entre diferentes jogadores.

Muito do trabalho em aprendizado em jogos consiste em caracterizar o comportamento de vários mecanismos de aprendizado (YOUNG, 2006) em termos do conjunto de equilíbrio limitante, por exemplo, equilíbrio de Nash, equilíbrio correlacionado. Isso é razoável, particularmente na ausência de um domínio de aplicação específico para obter o desempenho em termos do critérios de domínio específicos.

Se um equilíbrio de Nash de fato reflete uma condição de operação desejável, então métodos de aprendizado que conduzem comportamento para este equilíbrio são também bem motivadores.

Uma importante consideração em sistemas multiagente e aprendizado multiagente em particular é a informação disponível para cada agente. Um exemplo é se os agentes têm acesso a ações de outros agentes ou apenas suas recompensas individuais, desse modo distinguindo algoritmos de aprendizado baseados em ações daqueles baseados em ação (YOUNG, 2006). Outra questão, um pouco mais sutil, é se agentes tem acesso as funções de utilidade do outro agente. A falta dessa informação resulta no chamado aprendizado descasado e pode ter importantes conseqüências, comportamento limitado resultante de certas classes de dinâmicas de aprendizado (HART e MAS-COLELL, 2003). Deve-se comentar, também, que uma expressão funcional de utilidade está frequentemente indisponível para sistemas econômicos ao passo que em sistemas de engenharia, realizando o aprendizado baseado em recompensas atualizadas uma necessidade. Para sistemas de engenharia, a informação disponível para cada agente é uma consideração de projeto que é influenciada pelo domínio específico. Isso está em flagrante contraste com o modelo social descritivo, onde o fluxo de infomação é ditado pelo cenário modelado particular. Um bom exemplo é o conhecimento da função de utilidade de outros agentes. Numa aplicação de engenharia, onde agentes são componentes programados, esse conhecimento pode simplesmente ser comunicado, ainda que com algum custo, entre agentes. Nos contextos de modelos sociais, onde a utilidade reflete uma intenção vaga de um agente, tal comunicação de utilidade não pode ser adotada.

Há substancial e crescente número de mecanismos de adaptação para aprendizado multiagente. Para modelos sociais descritivos, há um interesse em entender como agentes (humanos) aprendem em jogos (por exemplo, em CAMERER (2003)). Isso talvez explique porque a maioria das dinâmicas na literatura econômica requer pouca capacidade de processamento temporal atualizada. De outro modo, em oposição, em sistemas de engenharia, agentes são componentes programados e, então, a especificação da dinâmica de aprendizado se torna ainda outra escolha de projeto. De fato, se se vê sistemas de controle por realimentação (por exemplo, em BASAR (2000)) como uma forma de tomador de decisão sequencial, perguntando quando um controlador reflete a tomada de decisão humana a resposta é dificilmente. Usa-

se aqui a noção racionalidade limitada simplesmente para significar limitações na capacidade de processamento de agentes individuais. Nesse contexto, o domínio da aplicação específico dita as fronteiras da racionalidade limitada e, consequentemente o conjunto de dinâmicas de aprendizado possíveis. Dependendo da taxa real temporal do tomador de decisão, o resultado é um largo espectro de possibilidades, abrangendo desde processamento intensivo computacional até simples regras de decisão. A lista de considerações de projeto de engenharia em aprendizado multiagente está longe de ser abrangida, e continuará a crescer da mesma forma que os métodos econômicos continuem a ser explorados por problemas de engenharia. Um exemplo é o grande interesse, particularmente na literatura de rede de computadores, em aplicar conceitos para mecanismos de projeto. Começando por KELLY (1997), a idéia de prover incentivo por meio de preços tornou-se extremamente atrativa na comunidade teórica de redes de computadores. Resultados recentes (por exemplo, em ROUGHGARDEN (2005)) apresentam um mecanismo simples levando a uma perda de eficiência limitada com respeito à otimização social. Contudo, o projeto *ab initio* (preliminar) de um mecanismo parece uma tarefa formidável.

Como foi argumentado anteriormente, a agenda de engenharia em sistemas multiagente é bastante diferente daquela da agenda da teoria econômica de jogos. Descrevem-se, ainda, dois aspectos adicionais que particularmente são importantes para sistemas de engenharia: robustez e domínio do conhecimento. Robustez, no sentido clássico da teoria de controle, significa a resiliência do sistema à variação de parâmetros ou da entrada esperada. Para que o aprendizado multiagente seja usado em sistemas de engenharia reais, ele deve possuir algumas propriedades de robustez. Há vários aspectos de robustez que são importantes para sistemas de engenharia. Primeiro, robustez implica que há algumas garantias de desempenho que devem ser mantidas sob as piores circuntâncias. Isso pode ser feito, talvez, usando aprendizado atualizado (CESA-BIANCHI e LUGOSI, 2006) onde a perda é minimizada e alguma garantia de desempenho é provida. Segundo, robustez implica que se pode garantir desempenho com respeito a algum conjunto de especificações. Isso significa que se se restringe o conjunto de oponentes e o ambiente atual para pertencer a um certo conjunto de oponentes e ambientes, então é possível garantir alguns níveis de desempenho. Finalmente, robustez é talvez mais importante na situação dinâmica, onde possa haver distúrbios transitórios. Nesse caso, um objetivo de projeto pode ser garantir que o desempenho medido ao longo do tempo nunca se deteriore abaixo de um certo nível. O domínio do conhecimento é um elemento vital da especificação de engenharia de sistemas. Algoritmos de aprendizado multiagente atuais disponíveis são tipicamente simples e não levam em conta o domínio do conhecimento. Enquanto isso talvez seja apropriado para uma agenda prescritiva, alguns sistemas de engenharia permitem aos agentes usar consideráveis recursos. O uso eficiente do domínio

do conhecimento no processo de projeto de algoritmos de aprendizado multiagente é crítico para fazer esses algoritmos efetivos e, em particular, para estendê-los além de problemas acadêmicos ilustrativos. A incorporação eficiente do domínio do conhecimento em planejar e na inteligência artificial em geral é um longa aventura. Devido à complexidade da interação entre múltiplos tomadores de decisão, desenvolver métodos rigorosos para descrever especificações dos sistemas multiagentes é crucial para o entendimento, análise e simulação de sistemas multiagentes. Possíveis especifições de tecnologias variam de linguagens simbólicas até modelos matemáticos. A especificação do domínio do conhecimento deve representar um compromisso entre a quantidade de detalhes necessários para descrever o modelo e as interações necessárias e a complexidade da descrição.

Contextualizou-se acima de maneira mais ampla, o aprendizado para o caso multiagente. Nessa tese, deixa-se explícito no entanto, estuda-se de maneira particular o aprendizado de dois agentes sob a perspectiva de engenharia. O jogo usado como base ao longo da tese é o jogo de casar moedas, um jogo de dois jogadores chamados R e C, e duas ações em que cada jogador quer ganhar. Casar moedas (*Matching Pennies*) é um jogo de soma zero em que cada jogador possui uma moeda (chamada de *penny* em inglês). Cada jogador opta por escolher cara ou coroa, arremessando a sua moeda para o alto e escondendo a mesma ao cair na mão. Se ambas moedas exibirem duas caras ou duas coroas, o jogador R vence, caso contrário o jogador C vence. O motivo pelo qual este jogo foi escolhido é que ele apresenta um equilíbrio de Nash chamado misto. Este tipo de equilíbrio possui característica de convergência difícil, conforme visto no capítulo 3.

1.2 Revisão bibliográfica

A seguir, realiza-se uma revisão bibliográfica, comentando alguns resultados mais espercíficos nas áreas estudadas na tese.

SHAMMA e ARSLAN (2005) trata de um jogo repetido, contínuo no tempo, no qual os jogadores continuamente atualizam suas estratégias em resposta às observações das ações dos oponentes, mas sem o conhecimento das intenções do oponente. O objetivo básico é entender como jogadores interagindo podem convergir para um equilíbrio de Nash, ou seja, um conjunto de estratégias para as quais nenhum jogador tem incentivo para mudar. A motivação é a que se segue. Há dois jogadores, cada um com um conjunto finito de ações possíveis. A cada instante que ocorre o jogo, cada jogador seleciona uma ação de acordo com a distribuição de probabilidades que representa aquela estratégia do jogador. A recompensa para cada jogador, chamada de utilidade do jogador, depende das ações tomadas por ambos jogadores. Enquanto cada jogador conhece sua própria recompensa, elas não são compartilhadas entre os

jogadores.

Supondo que um jogador sempre usa a mesma distribuição de probabilidade para gerar a sua ação, isto é, o jogador mantém sua estratégia constante, então, o outro jogador pode, ao longo do tempo, por meio de jogos repetidos, aprender essa distribuição, mantendo as médias móveis das ações do oponente. Essas médias móveis são chamadas de frequências empíricas. Jogando a melhor resposta otimizada para as frequências empíricas, o jogador otimizador convergirá, possivelmente, para a sua melhor estratégia face à estratégia fixa do outro jogador. E, se ambos jogadores presumem que o outro jogador está usando uma estratégia constante, seus mecanismos de atualização ficam sintonizados. Tal processo é o chamado jogo fictício. Nesse ambiente, jogadores jogam a melhor resposta otimizada das frequências empíricas do oponente, presumindo (de maneira errada) que a frequência empírica é representativa da distribuição de probabilidade constante.

O procedimento de jogo fictício, uma regra de aprendizado, foi introduzido em 1951 (BROWN, 1951; ROBINSON, 1951) como um mecanismo de calcular o equilíbrio de Nash. Há bastante literatura sobre o tópico. Linhas de pesquisa relacionadas são discutidas em FUDENBERG e LEVINE (1969) e em WEIBULL (1995) e na recente revisão em HART (2005). A questão principal é se jogos repetidos convergem para o equilíbrio de Nash. Uma linha do tempo resumida de resultados que estabelecem convergência de jogo fictício é a seguinte: em 1951, jogos de dois jogadores soma zero (ROBINSON, 1951); em 1961, jogos de dois jogadores dois movimentos (MIYASAWA, 1961); em 1993, jogos de dois movimentos, dois jogadores, com ruído, com um único equilíbrio de Nash (FUDENBERG e KREPS, 1993); em 1996, jogos com múltiplos jogadores com funções de recompensas idênticas (MONDERER e SHAPLEY, 1996); em 1999 jogos com ruído de dois movimentos, dois jogadores, com múltiplos equilíbrios de Nash (BENAIM e HIRSCH, 1999); e em 2003, jogos de dois jogadores onde um jogador tem apenas dois movimentos (BERGER, 2003).

Há uma coleção de resultados negativos com relação à possibilidade de equilíbrio misto surgir como resultado de comportamento interativo. CRAWFORD (1985) mostra que uma grande classe de mecanismos de ajuste de estratégia (diferente de jogo fictício) não pode convergir para um equilíbrio misto. KRISHNA e SJÖSTRÖM (1998) mostra que "quase todos" os jogos nos quais os jogadores têm mais de dois movimentos não podem convergir para um equilíbrio misto na melhor resposta do jogo fictício. Questões de não convergência são também discutidas em (ELLISON e FUDENBERG, 2000; FOSTER e YOUNG, 1998; HART e MAS-COLELL, 2003; VASIN, 1999). Em particular, HART e MAS-COLELL (2003) mostra que uma versão generalizada do jogo de Jordan não exibirá convergência para nenhum mecanismo de ajuste de estratégia, não apenas um mecanismo de melhor resposta,

conquanto que os jogadores não compartilhem suas funções de utilidade e os mecanismos de atualização sejam funções estáticas da frequências empíricas.

Ao contrário do caso do equilíbrio de Nash, há métodos (AUGUST *et al.*, 1999; FOSTER e VOHRA, 1997; HART e MAS-COLELL, 2000) que são garantidos, para todos os jogos, de convergir para o conjunto maior do equilíbrio correlacionado, o qual é um conjunto convexo que contém o conjunto de equilíbrios de Nash. Esses são algoritmos baseados em arrependimento (significando o erro por não ter escolhido a melhor decisão) que observam decisões passadas num esforço para avaliar quais foram as ações mais produtivas. (HART, 2005) provê um discursão extensa sobre algoritmos baseados em arrependimento.

Trabalhos similares no contexto de jogos de múltiplos jogadores são (BASAR, 1987) e (CONLISK, 1993). (BASAR, 1987) considera dois processos dinâmicos. O primeiro é aquele que jogadores usam uma estratégia que é a melhor resposta para ação anterior do oponente. O segundo é uma relaxação na qual jogadores usam a melhor resposta apenas para ajustar sua estratégia corrente, introduzindo assim alguma inércia, possivelmente melhorando propriedades de convergência. CONLISK (1993) considera jogos de soma zero jogados em intervalos, onde os jogadores ajustam sua estratégia de acordo com a previsão aproximada da estratégia do oponente.

1.2.1 A teoria de jogos evolucionários e o aprendizado multiagente

Em 'Se o aprendizado multiagente é a resposta, qual é a pergunta?' de (Y. SHOHAM e GRENAGER, 2007), os autores fazem um esforço produtivo para analisar o estado da arte de aprendizado multiagente, para resumir os resultados que foram alcançados, para dicernir as direções principais que foram seguidas. Em resumo, eles concluem que a maior parte em aprendizado multiagente pode ser enquadrado em um de cinco blocos, cada uma deles associado a uma linha de pesquisa distinta (ver Y. SHOHAM e GRENAGER, 2007, seção 5):

Computacional: algoritmos de aprendizagem são uma maneira de computar as propriedades de um jogo.

Descritivo: algoritmos de aprendizagem descrevem como agentes naturais aprendem no contexto de outros agentes também aprendizes.

Normativo: algoritmos de aprendizagem dão um significado para determinar quais conjuntos de regras de aprendizado estão em equilíbrio com outras.

Prescritivo, cooperativo: algoritmos de aprendizagem descrevem como agentes podem aprender a alcançar o controle distribuído de sistemas dinâmicos.

Prescritivo, não cooperativo: pergunta-se como um agente pode atuar para obter alta recompensa no jogo repetitivo. Há várias possibilidades para o termo alta recompensa. O trabalho de BOWLING e VELOSO (2001) expõe dois critérios para qualquer algoritmo de aprendizagem num ambiente de aprendizado multiagente: (1) o aprendizado deve convergir para uma política estacionária e (2) se o oponente converge para uma política estacionária, o algoritmo deve convergir para a melhor resposta. esta tese foca o desenvolvimento neste bloco.

Não é todo trabalho da área que se enquadra num desses cinco blocos. Isso significa que ou se precisa de mais blocos ou que algum trabalho adicional precisa ser feito para talvez rever as definições destes blocos em bases mais sólidas.

1.2.2 Aprendizado por reforço (AR)

Em (CLAUS e BOUTILIER, 1998) várias possibilidades são aventadas para a utilização de aprendizado por reforço (AR) (BOWLING, 2003; KAELBLING *et al.*, 1996; LITTMAN, 1994; RUSSELL e NORVIG, 2003) quando um agente se depara num meio em que precisa aprender uma estratégia por meio de tentativa e erro num ambiente dinâmico, sem um modelo pré- definido:

a) Há diferenças entre o modo de aprendizagem dos agentes?

b) O aprendizado no ambiente de múltiplos jogadores converge e com qual taxa?

c) As taxas de convergência influem no equilíbrio?

d) O aprendizado de um agente irá influenciar no de outros agentes?

WATKINS (1992) mostra que o algoritmo de aprendizado por reforço *Q-learning* converge para o ponto ótimo das ações e JAAKKOLA *et al.* (1994) estabelece um novo teorema que abrange a prova de convergência do método *Q-learning* e do método Diferença Temporal (*TD*)- *Temporal Difference Method*.

BANERJEE *et al.* (2001) estuda a técnica de aprendizagem por reforço chamada de minimax-Q e sua taxa de convergência.

RIBEIRO (2002) apresenta um tutorial sobre AR, comentando suas características e limitações como a convergência das ações condicionada a um número grande de observações.

BIANCHI (2004) propõe a aceleração de aprendizado por meio do algoritmo Aprendizado Acelerado por Heurísticas (*Heuristically Accelerated Q-Learning*-HAL). Em HAL a exploração é realizada por meio de uma função heurística que pode ser modificada a cada iteração. Garante-se a convergência das ações, mantém-se o aprendizado não supervisionado e acelera-se o aprendizado de modo heurístico.

BAZZAN (2005, 2009) exibem resultados de sistemas multiagentes aplicados a coordenação de controle de tráfego, sendo o primeiro fazendo uso de um coordenador do processo.

TUYLS *et al.* (2006) constrói um modelo temporal contínuo do algoritmo de aprendizado por reforço *Q-learning* (SUTTON e BARTO, 1998) onde os valores Q são probabilidades de Boltzmann para seleção de ação, relacionando Dinâmica do Replicador *Replicator Dynamics*-(RD) e *Q-learning*. O trabalho de TUYLS *et al.* (2006) teve como base (BORGERS e SARIN, 1997) que relacionou matematicamente Q-Learning e dinâmica do replicador (DR). Mais precisamente, o modelo de aprendizado cruzado (*Cross Learning*) [1] converge para a versão temporal contínua da DR. O limite de tempo contínuo é construído de tal forma que a cada intervalo de tempo tem-se várias iterações do jogo, e que os ajustes que os jogadores fazem entre duas iterações do jogo são pequenos. E, no limite, o processo de aprendizado torna-se determinístico. O processo limite satisfaz o sistema de equações da DR.

O termo evolucionário adjetivando a abordagem de aprendizado não se limita ao sentido biológico da palavra, como algo determinado geneticamente, por meio da reprodução de genes. BORGERS e SARIN (1997) cita também o sentido de evolução cultural, um processo de aprendizagem aonde cada agente procurar imitar, convencer o outro e apresenta um modelo de aprendizagem (CROSS, 1973) que converge para um modelo temporal contínuo de equação replicadora (ver SIGMUND, 2009, capítulo 2).

O Método de Aprendizado por Reforço Estimação Preliminar (MAR-EP) desenvolvido nesta tese usou probabilidades de Boltzmann para escolha das ações dos jogadores, nos moldes de (TUYLS *et al.*, 2006). MAR-EP, assim como BIANCHI (2004), usa uma heurística para acelerar o aprendizado, mantém a convergência das ações e seu aprendizado é não supervisionado. Em MAR-EP, a heurística tratada é uma modelagem inicial de todos os agentes, mantendo uma taxa de aprendizado constante inicial comum, e não uma função heurística como em BIANCHI (2004). Outra diferença é que a coleta de informações para modelagem da heurística dá-se apenas na fase inicial e não durante toda a iteração.

1.3 Objetivos desta tese

No contexto esboçado acima, os objetivos desta tese são:

1. Revisão de métodos de aprendizado por reforço baseados em gradientes ascendentes, fornecendo uma explicação unificada de todos eles, por meio da técnica de função de Liapunov com controle.

[1] o modelo de aprendizado cruzado é um modelo que considera vários agentes jogando o mesmo jogo repetidamente no tempo discreto.

2. Fornecimento de uma prova rigorosa do algoritmo de Aprendizado por Política Ponderada (*Weighted Policy Learning*) - (WPL), considerado algoritmo do estado da arte, bem como fornecer das condições de contorno que garantem a convergência de WPL, não apresentadas em (ABDALLAH e LESSER, 2008), utilizando a perspectiva de funções de Liapunov e a idéia de equilíbrios virtuais.

3. Desenvolvimento de uma metodologia de aprendizado por reforço aplicado a uma classe de jogos de dois jogadores, duas ações, o Método de Aprendizado por Reforço Estimação Preliminar - (MAR-EP) em que cada jogador usa um controlador projetado com Funções de Liapunov com Controle (FLC) com parâmetros estimados apenas nos primeiros contatos entre os jogadores (isto é, com apenas conhecimento parcial do jogo.)

1.4 Organização geral da tese

A tese encontra-se dividida em cinco capítulos. O primeiro capítulo é uma introdução, onde se posiciona o trabalho em relação ao meio científico e expõe-se os seus objetivos. O segundo capítulo projeta-se uma nova estratégia de controle chaveado, denominada de Chaveamento Gradiente Hiperbólico-Elíptico, *Hyperbolic-Elliptic Gradient Switching* (HEGS), baseada em controle chaveado.

O terceiro capítulo é uma revisão de métodos de aprendizado por reforço baseados em gradientes ascendentes, fornecendo uma explicação unificada de todos eles, por meio da técnica de função de Liapunov com controle, incluindo HEGS. Também o algoritmo WPL, considerado algoritmo do estado da arte, é examinado utilizando a perspectiva de funções de Liapunov e a ideia de equilíbrios virtuais, esclarecendo o funcionamento do método e fornecendo uma prova rigorosa da convergência do método. E ainda serão elencadas as condições de contorno que garantem a convergência de WPL, não apresentadas em ABDALLAH e LESSER (2008).

O quarto capítulo consiste na apresentação do Método de Aprendizado por Reforço Estimação Preliminar (MAR-EP), um algoritmo de aceleração do aprendizado de fato que usa controladores projetados com FLC por cada jogador que usam parâmetros estimados apenas nos primeiros contatos entre os jogadores.

Na parte final da tese, que é o quinto capítulo, expõem-se conclusões, contribuições e perspectivas para futuros trabalhos.

Capítulo 2

Sistemas dinâmicos de Aprendizado por Reforço Multiagente por Gradiente Ascendente (ARM-GA) pela perspectiva de controle chaveado

2.1 Introdução

SINGH *et al.* (2000) analisaram o comportamento de agentes que adaptam incrementalmente sua estratégia por meio do gradiente ascendente das funções de recompensa esperada, num jogo de soma geral dois jogadores, duas ações, e mostraram que o resultado ou as estratégias convergem para um equilíbrio de Nash ou, senão, a média de suas recompensas converge para as recompensas do equilíbrio de Nash. Este método foi chamado de gradiente ascendente incremental (*Incremental Gradient Ascent* - IGA).

BOWLING e VELOSO (2002) mostraram, por meio de argumentos geométricos, que chavear entre duas taxas de aprendizado, apelidado de ganhar ou aprender rapidamente, WoLF-IGA (*Win or Learn Fast*-Ganhar ou Aprender Rapidamente - IGA), poderia levar à convergência das estratégias e das recompensas. Sucessivamente, BANERJEE e PENG (2002) e ZHANG e HUANG (2004) mostraram que os casos de não convergência da estratégia WoLF-IGA em jogos de soma geral podem ter modificada convenientemente sua estratégia, usando intuição geométrica advinda de um sistema de segunda ordem associado.

Nesse capítulo, são feitas as seguintes contribuições. Primeiro, observa-se que todos os métodos propostos de convergência acima citados são uma técnica de estrutura de chaveamento ou estrutura variável que envolve, usando terminologia de controle, uma realimentação de estado multiplicativa. Isso permite revisão rápida das propriedades geométricas básicas das trajetórias do sistema em IGA e WoLF-IGA.

Segundo, esse ponto de vista de controle por chaveamento permite (i) fornecer uma prova simples usando função de Liapunov das estratégias propostas em (BOWLING e VELOSO, 2002), (BANERJEE e PENG, 2002) e (ZHANG e HUANG, 2004) e, além, disso, (ii) projetar uma nova estratégia que foi chamada de Chaveamento Gradiente Hiperbólico-Elíptico, *Hyperbolic-Elliptic Gradient Switching* (HEGS), que é similar à estratégia de (BANERJEE e PENG, 2002), (ZHANG e HUANG, 2004), mas conduz a uma convergência mais rápida das estratégias e das recompensas. Enfatiza-se que a estratégia de chaveamento proposta é dependente de estado e que se acredita que este é o contexto correto para examinar os resultados anteriores citados acima, que enfatizaram a natureza de dependência do tempo das estratégias ao invés da natureza de dependência de estado.

2.1.1 Definições

O contexto para a política de gradiente ascendente é um cenário de dois jogadores e duas ações, com as seguintes matrizes de recompensa:

$$R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \qquad C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \tag{2.1}$$

Sejam x_1 e x_2, respectivamente, as probabilidades dos jogadores linha e coluna selecionarem as primeiras ações dos seus respectivos conjuntos de possíveis ações (OWEN, 1995). As recompensas dos jogadores linha e coluna são, $V_r(x_1, x_2)$ e $V_c(x_1, x_2)$, respectivamente:

$$\begin{aligned} V_r &= r_{11}(x_1x_2) + r_{22}((1-x_1)(1-x_2)) \\ &\quad + r_{12}(x_1(1-x_2)) + r_{21}((1-x_1)x_2) \\ V_c &= c_{11}(x_1x_2) + c_{22}((1-x_1)(1-x_2)) \\ &\quad + c_{12}(x_1(1-x_2)) + c_{21}((1-x_1)x_2) \end{aligned} \tag{2.2}$$

Definindo

$$\begin{aligned} \bar{r} &= r_{11} + r_{22} - (r_{12} + r_{21}) \\ \bar{c} &= c_{11} + c_{22} - (c_{12} + c_{21}) \end{aligned} \tag{2.3}$$

e, ainda,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, A = \begin{bmatrix} 0 & \bar{r} \\ \bar{c} & 0 \end{bmatrix}, b = \begin{bmatrix} -(r_{22} - r_{12}) \\ -(c_{22} - c_{21}) \end{bmatrix} \tag{2.4}$$

o sistema de dinâmica gradiente que atualiza as probabilidades, também conhecidas como as estratégias, x_1, x_2, pode ser escrito como:

$$\dot{\mathbf{x}} = \begin{bmatrix} \partial V_r(x_1, x_2)/\partial x_1 \\ \partial V_c(x_1, x_2)/\partial x_2 \end{bmatrix} = \begin{bmatrix} x_2\bar{r} - (r_{22} - r_{12}) \\ x_1\bar{c} - (c_{22} - c_{21}) \end{bmatrix} \tag{2.5}$$

que pode ser reescrito como:

$$\dot{\mathbf{x}} = A\mathbf{x} + b, \tag{2.6}$$

que define um sistema dinâmico afim no interior de um quadrado unitário.

Na verdade, como x_1 e x_2 representam probabilidades, o ponto (x_1, x_2) deve estar dentro do quadrado unitário $S = [0, 1] \times [0, 1]$. Assim, o sistema dinâmico que representa a atualização das probabilidades deve, de fato, ser escrito como:

$$\dot{\mathbf{x}} = \mathbf{P}(A\mathbf{x} + b), \tag{2.7}$$

onde **P** representa o operador projeção aplicado individualmente em cada componente do vetor, de modo que, se o componente $x_i(t)$ se torna zero em t_z e ao mesmo tempo $\dot{x}_i(t_z)$ é negativo, então $x_i(t)$ continua igual a 0 para $t > t_z$, até que $\dot{x}_i(t)$ se torna positivo ou até que a trajetória converge para um ponto de equilíbrio (ambos gradientes ou gradientes projetados iguais a zero). Esse sistema dinâmico gradiente projetado (2.7) é conhecido como sistema gradiente ascendente infinitesimal, *Infinitesimal Gradient Ascent* (IGA) e seu ponto de equilíbrio, onde seus gradientes são zero, é chamado de $\mathbf{x}^{\text{eq}} = ({x_1}^{\text{eq}}, {x_2}^{\text{eq}})$ e calculado como:

$$ {x_1}^{\text{eq}} = \frac{(c_{22} - c_{21})}{\bar{c}} \tag{2.8} $$

$$ {x_2}^{\text{eq}} = \frac{(r_{22} - r_{12})}{\bar{r}} \tag{2.9} $$

Dadas essas definições, o problema geral descrito nesse trabalho pode ser formulado da seguinte forma: *achar a lei que atualiza estratégias para jogos de soma geral de duas pessoas, duas ações, descritas acima de uma forma que as estratégias convergem para um equilíbrio de Nash.*

Esse problema corresponde à noção de *convergência forte*, na qual as estratégias convergem para um equilíbrio de Nash. *Convergência fraca* é dita quando as recompensas esperadas convergem para aquelas correspondentes ao equilíbrio de Nash, apesar de as estratégias talvez não convergirem.

2.2 Perspectiva de controle chaveado em WoLF-IGA

O sistema dinâmico gradiente (2.7) foi introduzido em (SINGH *et al.*, 2000) e o seguinte resultado fundamental de convergência fraca está exibido a seguir:

Teorema 2.2.1 *Em um jogo de soma geral, duas pessoas, duas ações, se ambos jogadores atualizam suas ações selecionando probabilidades x_i de acordo com a dinâmica IGA (2.7), então suas recompensas médias irão convergir para as recompensas esperadas para algum equilíbrio de Nash. Isso ocorre em uma de duas maneiras: (i) as trajetórias de (2.7) convergem para um equilíbrio de Nash, ou (ii) as trajetórias não convergem, mas as médias das recompensas convergem para as recompensas esperadas de algum par Nash.*

A observação crucial é que a dinâmica do sistema afim (2.6) e a localização do ponto de equilíbrio $\mathbf{x}^{\text{eq}}$ determinam que item (i) ocorre apenas quando $\bar{r}\bar{c} \geq 0$ e corresponde à convergência para pares Nash na borda do quadrado unitário S

(nesse caso, gradientes projetados se tornam zero), e item (ii) ocorre apenas quando $\bar{r}\bar{c} < 0$ e corresponde a $\mathbf{x}^{\text{eq}}$ no interior do quadrado unitário S.

Esse teorema foi mostrado por exaustivas análises gráficas de diferentes possibilidades que ocorrem com a chamada dinâmica irrestrita (2.6). Especificamente, item (i) do teorema 2.2.1, que apenas ocorre se $\bar{r}\bar{c} \geq 0$, corresponde a um ponto de equilíbrio sela da equação (2.6), enquanto item (ii) corresponde ao centro de equilíbrio de (2.6) e apenas ocorre quando $\bar{r}\bar{c} < 0$.

2.2.1 Ganhar ou aprender rapidamente- IGA, WoLF-IGA ("*Win or Learn Fast*"-IGA)

O próximo desenvolvimento, devido a (BOWLING e VELOSO, 2002), foi um resultado de convergência forte, que mostrou que a convergência de estratégias para o equilíbrio de Nash pode ocorrer mesmo no caso $\bar{r}\bar{c} < 0$, correspondendo ao item (ii) do teorema 2.2.1, com o centro $\mathbf{x}^{\text{eq}} \in S$, provando que o sistema dinâmico (2.7) pode ser modificado pela introdução, do ponto de vista de controle, de um ganho de realimentação ou matriz de aprendizagem na dinâmica gradiente, como a seguir:

$$\dot{\mathbf{x}} = \mathbf{K}(A\mathbf{x} + b), \tag{2.10}$$

onde

$$\mathbf{K} = \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix} \tag{2.11}$$

é uma matriz diagonal a ser especificada abaixo. Note, também, que a matriz de aprendizagem está sendo aplicada apenas na dinâmica afim, que é seguida pelas trajetórias do sistema que são confinadas no quadrado unitário.

O algoritmo da dinâmica de WoLF-IGA é dado por

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & \ell^r(t)\bar{r} \\ \ell^c(t)\bar{c} & 0 \end{bmatrix}\mathbf{x} + \begin{bmatrix} -\ell^r(t)(r_{22} - r_{12}) \\ -\ell^c(t)(c_{22} - c_{21}) \end{bmatrix}, \tag{2.12}$$

onde as taxas de aprendizagem ou ganhos $\ell^c(t)$ e $\ell^r(t)$ são determinados pelas seguintes regras:

$$\ell^r(t) = \begin{cases} \ell_{min} \text{ se } V_r(x_1, x_2) \geq V_r({x_1}^{\text{eq}}, x_2) \\ \ell_{max} \text{ senão} \end{cases} \tag{2.13}$$

$$\ell^c(t) = \begin{cases} \ell_{min} \text{ se } V_c(x_1, x_2) \geq V_c(x_1, {x_2}^{\text{eq}}) \\ \ell_{max} \text{ senão} \end{cases} \tag{2.14}$$

(BOWLING e VELOSO, 2002) referenciaram a técnica como WoLF-IGA, já que,

seguindo as regras acima, corresponde a aprender lentamente ou cuidadosamente quando ganhando (ganho = $\ell_{\min}$), e aprender rapidamente (ganho = $\ell_{\max}$) quando perdendo.

Essa regra foi investigada posteriormente por (BANERJEE e PENG, 2002), que também usaram

$$K = \begin{bmatrix} \ell^r(t) & 0 \\ 0 & \ell^c(t) \end{bmatrix} \tag{2.15}$$

e por (ZHANG e HUANG, 2004), que escolheram

$$K = \begin{bmatrix} (1+\delta_{x_1}(t))\gamma & 0 \\ 0 & (1+\delta_{x_2}(t))\gamma \end{bmatrix} \tag{2.16}$$

com $\ell^c(t)$, $\ell^r(t)$, $\delta_{x_2}(t)$, $\delta_{x_1}(t)$, chaveados entre valores máximo e mínimo, de acordo com algum critério descrito em detalhes em (ZHANG e HUANG, 2004). Nota-se que, em todos os casos acima, k_1 e k_2 são sempre escolhidos como valores positivos.

Antes de continuar, destaca-se que todas essas abordagens redescobriram um fato antigo da teoria de estrutura variável e sistemas chaveados, que é possível chavear entre dois osciladores harmônicos e atingir estabilidade assintótica (UTKIN, 1977). Mais especificamente, no presente contexto, o seguinte teorema, que contém os resultados de (BOWLING e VELOSO, 2002), (BANERJEE e PENG, 2002) e (ZHANG e HUANG, 2004) podem fornecer uma prova curta e rigorosa, baseada numa função de Liapunov.

Teorema 2.2.2 *O sistema dinâmico afim da equação (2.6) ($\dot{\boldsymbol{x}} = \boldsymbol{K}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$), com $\boldsymbol{A}$ tal que $\bar{r}\bar{c} < 0$ pode ser estabilizado por um controle chaveado dependente de estado do tipo (2.10):*

$$\dot{\boldsymbol{z}} = \begin{bmatrix} 0 & k_1(\boldsymbol{z})\bar{r} \\ k_2(\boldsymbol{z})\bar{c} & 0 \end{bmatrix} \boldsymbol{z}, \tag{2.17}$$

onde $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{x}^{\text{eq}}$. Para

$$\bar{r} < 0, \quad \bar{c} > 0, \tag{2.18}$$

as leis de chaveamento são definidas como

$$\begin{aligned} |k_1(\boldsymbol{z})\bar{r}| &> |k_2(\boldsymbol{z})\bar{c}|, \text{ se } z_1 z_2 > 0 \\ |k_1(\boldsymbol{z})\bar{r}| &< |k_2(\boldsymbol{z})\bar{c}|, \text{ se } z_1 z_2 \leq 0 \end{aligned} \tag{2.19}$$

Se as leis de chaveamento são escolhidas positivas e constantes por trecho, (2.19) simplificam para

$$\begin{aligned} |k_1\bar{r}| &> |k_2\bar{c}|, \text{ se } z_1 z_2 > 0 \\ |\hat{k}_1\bar{r}| &< |\hat{k}_2\bar{c}|, \text{ se } z_1 z_2 \leq 0, \end{aligned} \tag{2.20}$$

onde $k_i(\boldsymbol{z}) = k_i$, se $z_1 z_2 > 0$ e $k_i(\boldsymbol{z}) = \hat{k}_i$, se $z_1 z_2 \leq 0$. Mudanças apropriadas dos

sinais das desigualdades definem as leis de chaveamento no caso $\bar{r} > 0$ *e* $\bar{c} < 0$.

Prova: Primeiro, nota-se que, para o sistema dinâmico afim (2.6), após uma mudança de coordenadas para seu ponto de equilíbrio, isto é, definindo

$$\mathbf{z} = \mathbf{x} - \mathbf{x}^{\text{eq}}$$

pode ser reescrito como

$$\dot{\mathbf{z}} = \mathbf{A}\mathbf{z}$$

Assim, o controle de chaveamento pode ser desenvolvido para o sistema (2.17). Considere a função de Liapunov quadrática

$$V(\mathbf{z}) = (1/2)\mathbf{z}^T\mathbf{z} = (1/2)(z_1^2 + z_2^2). \tag{2.21}$$

As derivadas ao longo das trajetórias de (2.17) podem ser calculadas como:

$$\dot{V}(\mathbf{z}) = (k_1(\mathbf{z})\bar{r} + k_2(\mathbf{z})\bar{c})z_1 z_2, \tag{2.22}$$

que se verifica facilmente ser estritamente negativa fora dos eixos coordenados, sob condições (2.18) e (2.20). Para completar a prova, observa-se que o menor conjunto invariante contido na união dos eixos coordenados (onde $\dot{V} = 0$) é a origem. Então, a estabilidade assintótica global segue o princípio de invariância de LaSalle. ■

Três observações são adequadas nesse momento. Primeira, observa-se que o uso de uma lei dependente de estado, em oposição a uma lei dependente do tempo, torna possível o uso do princípio de invariância de LaSalle e isso é crucial para a prova de estabilidade. Segunda, a prova é dada apenas para dinâmica sem restrições. Isso porque as modificações requeridas para atingir a mesma conclusão para a dinâmica restrita

$$\dot{\mathbf{x}} = \mathbf{P}\mathbf{K}(\mathbf{A}\mathbf{x} + \mathbf{b}) \tag{2.23}$$

são descritas em detalhes em (BOWLING e VELOSO, 2002), (BANERJEE e PENG, 2002). Terceira, observa-se que, das condições $\bar{r}\bar{c} < 0$, $\bar{r}$ e $\bar{c}$ não podem ser identicamente iguais a zero. A matriz $\mathbf{A}$, nesse caso, não é inversível conforme a equação (2.24) e SINGH *et al.* (2000) já mostraram que o algoritmo IGA leva a trajetória do par de estratégias a convergir num ponto da borda.

$$\overbrace{\begin{bmatrix} 0 & \frac{1}{k_2(z)\bar{c}} \\ \frac{1}{k_1(z)\bar{c}z_1} & 0 \end{bmatrix}}^{A^{-1}} \cdot \begin{bmatrix} 0 & k_1(z)\bar{r} \\ k_2(z)\bar{c} & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I \tag{2.24}$$

2.2.2 Interpretação geométrica das leis de chaveamento

O sistema (2.17) pode ser reescrito como:

$$\begin{cases} \frac{dz_2}{dt} = k_2(z)\bar{c}z_1 \\ \frac{dz_1}{dt} = k_1(z)\bar{r}z_2 \end{cases} \tag{2.25}$$

e então:

$$\frac{dz_2}{dz_1} = \frac{k_2(\mathbf{z})\bar{c}z_1}{k_1(\mathbf{z})\bar{r}z_2}$$

Integração conduz aos lugares das trajetórias no plano de fase:

$$z_2^2 = \frac{k_2(\mathbf{z})\bar{c}}{k_1(\mathbf{z})\bar{r}}z_1^2 + m, \tag{2.26}$$

onde m é a constante de integração e pode ser determinada pela condição inicial . Nota-se que esses lugares são elipses se $\frac{k_2(\mathbf{z})\bar{c}}{k_1(\mathbf{z})\bar{r}} < 0$ e hipérboles se $\frac{k_2(\mathbf{z})\bar{c}}{k_1(\mathbf{z})\bar{r}} > 0$. No caso de elipses, os eixos são alinhados com os eixos coordenados e seus comprimentos são calculados como

$$d_1 = \sqrt{\left|\frac{mk_1(\mathbf{z})\bar{r}}{k_2(\mathbf{z})\bar{c}}\right|},\ d_2 = \sqrt{|m|}.$$

No caso das trajetórias que são hipérboles, suas assíntotas são linhas retas passando pela origem:

$$z_2 = \pm\left(\sqrt{\left|\frac{k_2\bar{c}}{k_1\bar{r}}\right|}\right)z_1 \tag{2.27}$$

Sob o ponto de vista geométrico, observa-se que a lei expressa na equação (2.20) assegura que k_1 e k_2 são chaveados de tal maneira que o eixo maior da elipse está na direção vertical quando a trajetória está no primeiro e terceiro quadrantes e na direção horizontal quando a trajetória está no segundo e quarto quadrantes, e não é difícil ver que a trajetória do sistema chaveado converge para o centro sob a lei de chaveamento, como anteriormente observado, num contexto geral, em (UTKIN, 1977).

2.3 Sistema de dinâmica Gradiente de Chaveamento Hiperbólico Elíptico (*Hyperbolic-Elliptic Gradient Switching* (HEGS))

Da prova do Teorema 2.2.2 e da descrição geométrica acima, é possível chegar aos refinamentos das regras de WoLF-IGA que conduzem a uma convergência mais rápida. Uma descrição geométrica dessa modificação, na figura 2.1, é seguida por

uma descrição matemática.

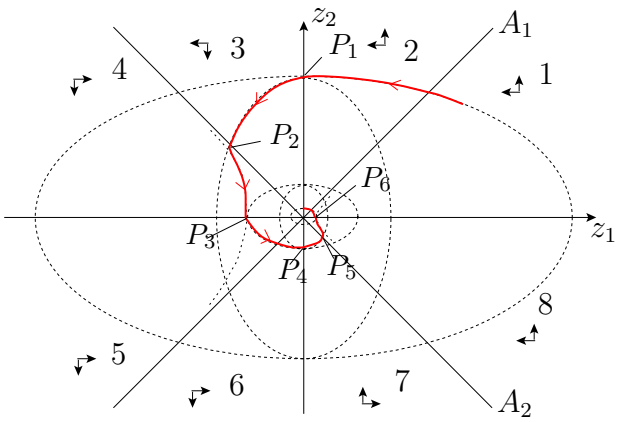


Figura 2.1: Divisão do plano de fase z_1-z_2 em setores chaveados de 1 até 8 para o caso $\bar{r} < 0$, $\bar{c} > 0$, o que implica movimento no sentido elíptico anti-horário, qualquer que seja $k_i(\mathbf{z}) > 0, i = 1, 2$. P_i são os pontos de chaveamento: os segmentos de trajetória P_2P_3 e P_5P_6 são hipérboles, todos os outros são elipses. As linhas A_1 e A_2 são, respectivamente, as assíntotas estáveis e instáveis para as trajetórias de hipérboles. Pequenas setas indicam as componentes do campo vetorial em cada setor para a dinâmica de HEGS definida no texto.

Na figura 2.1 é exibida a nova estratégia HEGS, que é idêntica a de WoLF-IGA em todos os setores, exceto 4 e 8, onde o sinal do ganho k_1 é escolhido negativo (implicando gradiente descendente para o jogador linha). Geometricamente falando, essa mudança conduz a uma convergência mais rápida, conduz a uma trajetória que percorre internamente a trajetória convergente de WoLF-IGA e mais próxima da origem.

O estabelecimento preciso da convergência sob a dinâmica de HEGS é listado como teorema principal desse capítulo:

Teorema 2.3.1 *O sistema dinâmico afim (2.17), com* $\boldsymbol{A}$ *tal que (2.18) são atendidas, pode ser estabilizado por um controle de chaveamento elíptico-hiperbólico, constante por partes, dependente de estado definido, com referência aos setores definidos na figura 2.1, como segue:*

$$\begin{aligned} k_1(\boldsymbol{z}) &> 0, \boldsymbol{z} \notin \text{setor 4 ou 8} \\ k_1(\boldsymbol{z}) &< 0, \boldsymbol{z} \in \text{setor 4 ou 8} \end{aligned} \tag{2.28}$$

$$k_2(\boldsymbol{z}) > 0, \text{para todos } \boldsymbol{z} \tag{2.29}$$

e, além disso, os valores constantes assumidos pelos ganhos chaveados são sujeitos às restrições:

$$\begin{aligned} |k_1\bar{r}| &> |k_2\bar{c}|, \text{ se } z_1z_2 > 0 \text{ e } \boldsymbol{z} \notin \text{setor 4 ou 8} \\ |\hat{k}_1\bar{r}| &< |\hat{k}_2\bar{c}|, \text{ se } z_1z_2 \leq 0 \text{ e } \boldsymbol{z} \notin \text{setor 4 ou 8} \end{aligned} \tag{2.30}$$

onde $k_i(\boldsymbol{z}) = k_i$, *se* $z_1 z_2 > 0$ *e* $\boldsymbol{z} \notin$ *setor 4 ou 8. De maneira similar,* $k_i(\boldsymbol{z}) = \hat{k}_i$, *se* $z_1 z_2 \leq 0$ *e* $\boldsymbol{z} \notin$ *setor 4 ou 8. A dinâmica definida por (2.6), (2.28)–(2.30) e* $\bar{r}\bar{c} < 0$ *é chamada de dinâmica gradiente de chaveamento hiperbólico elíptico (*hyperbolic-elliptic gradient switching (HEGS) dynamics*) e converge para um equilíbrio de Nash mais rapidamente que o sistema de dinâmica WoLF-IGA (2.12)–(2.14). Finalmente, mudanças apropriadas dos sinais de desigualdade definem as leis de chaveamento no caso* $\bar{r} > 0$ *e* $\bar{c} < 0$.

Prova: É necessário apenas examinar a dinâmica modificada nos setores 4 e 8 (aonde os segmentos de trajetória seguem a dinâmica de hipérbole), pois a dinâmica nos setores restantes é idêntica àquela das condições do teorema 2.2.2. Usando a mesma função quadrática de Liapunov (2.21) como acima, a derivada ao longo das trajetórias (2.22) é facilmente verificada como estritamente negativa sob as condições (2.18), (2.28)–(2.30), fora dos eixos coordenados. Nota-se, além disso, que o campo vetorial, em ambos os lados de cada assíntota ou eixo coordenado que define um setor, não satisfaz a condição de modo de deslizamento (*sliding*) (de apontar no sentido do segmento) e que o menor conjunto invariante contido na união dos eixos coordenados (onde $\dot{V} = 0$) é a origem, estabilidade assintótica advém do princípio de invariância de LaSalle. A condição sobre a convergência mais rápida da dinâmica de HEGS em relação à dinâmica de WoLF-IGA, é por argumento geométrico. ■

Observa-se que, de maneira rigorosa, o chaveamento no setor 4 e 8 pode acontecer depois da trajetória elíptica ter entrado no interior do setor, como não é prático requerer chaveamento exato em A_2. De fato, se isso fosse possível, a convergência para a origem poderia ocorrer ao longo de A_2. Chaveamento não ideal que ocorre na prática, ou na dinâmica discretizada que será usada em aplicações, garante que o chaveamento de fato ocorre no interior dos setores 4 e 8 e a prova de convergência não é afetada por essa mudança. Da mesma forma, é possível chavear de volta para uma trajetória de elipse em quaisquer pontos da trajetória de hipérbole antes do ponto P_3 na figura 2.1 sem afetar convergência, mas ainda resultando em convergência mais rápida que em WoLF-IGA. Nesses comentários ressalta-se que HEGS possui alguma robustez com respeito a linhas de chaveamento.

Assim como no teorema 2.2.2, o fato da lei de chaveamento ser dependente de estado é crucial para o uso do princípio de invariância de Krasovskii-LaSalle (LASALLE, 1960)[1]. A prova é dada apenas na dinâmica irrestrita. As modificações necessárias para atingir a mesma conclusão para dinâmica restrita (2.23) são descritas em detalhes em (BOWLING e VELOSO, 2002), (BANERJEE e PENG, 2002). Outra maneira de dizer isso, segundo (BOWLING e VELOSO, 2002), é dizer que, no caso da dinâmica restrita (2.23), qualquer par de estratégia inicial que seja

[1]LaSalle foi o primeiro autor do Ocidente a publicar o princípio em 1960, enquanto Nikolai N. Krasovskii teve sua primeira publicação em 1952 em russo.

suficientemente perto do centro (equilíbrio de Nash) convergirá para ele. Aqui, "suficientemente perto" significa que a trajetória elíptica da dinâmica irrestrita (2.6) por meio desse ponto inicial permanece inteiramente dentro do quadrado unitário S. Finalmente, observa-se que existe flexibilidade na escolha do ganho negativo k_1 nos setores 4 e 8. De maneira direta, quanto maior o ganho, maior o componente do movimento na direção da linha estável A_2, conduzindo a uma convergência mais rápida. Na prática, pode haver restrições em quão grande esse ganho pode ser.

2.3.1 Quão mais rápido a dinâmica HEGS pode convergir ?

Concentra-se o foco na condição $\bar{r} < 0, \bar{c} > 0$ e nos setores 4 e 8, onde as trajetórias de HEGS são hipérboles, já que, no setores restantes, HEGS usa os mesmos ganhos positivos de WoLF-IGA.

De acordo com a estratégia de WoLF-IGA, nos setores 4 e 8, os ganhos de realimentação devem ser $k_1 > 0$, k_{1min}, e $k_2 > 0$, k_{2max}, tais que

$$\dot{V}_{W-IGA}(\mathbf{z}) = (k_{1min}(\mathbf{z})\bar{r} + k_{2max}(\mathbf{z})\bar{c})z_1 z_2 \tag{2.31}$$

A estratégia HEGS, nos setores 4 e 8, requer que os ganhos de realimentação sejam $k_{1HEGS} < 0$, e $k_{2HEGS} > 0$, tais que

$$\dot{V}_{HEGS}(\mathbf{z}) = (k_{1HEGS}(\mathbf{z})\bar{r} + k_{2HEGS}(\mathbf{z})\bar{c})z_1 z_2 \tag{2.32}$$

Comparando (2.31) com (2.32), a condição suficiente para $\dot{V}_{HEGS}(\mathbf{z}) \leq \dot{V}_{W-IGA}(\mathbf{z})$ é

$$k_{2HEGS}(\mathbf{z}) \geq k_{2max}(\mathbf{z}) \tag{2.33}$$

Nota-se que a equação (2.33) implica na condução de estratégia de HEGS à convergência mais rápida que àquela de WoLF-IGA.

Para quantificar quanto mais rápido HEGS é em relação a WoLF-IGA nos setores 4,8, considera-se $\bar{r} = -\bar{c}$, $\bar{c} > 0$, $k_{1HEGS} = -1$, $k_{2HEGS} = 1$, $k_{1min} = 0,08$, $k_{2max} = 1$, que, de (2.31) e (2.32), pode ser escrito como:

$$\dot{V}_{W-IGA}(\mathbf{z}) = 0,92\bar{c}z_1 z_2 \tag{2.34}$$

$$\dot{V}_{HEGS}(\mathbf{z}) = 2\bar{c}z_1 z_2 \tag{2.35}$$

Como $z_1 z_2 < 0$ nos setores 4 e 8, $\dot{V}_{HEGS}(\mathbf{z})$ é cerca de duas vezes mais negativo que $\dot{V}_{W-IGA}(\mathbf{z})$.

2.4 Exemplo numérico 1

As seguintes matrizes de recompensa

$$R = \begin{bmatrix} 0 & 3 \\ 1 & 2 \end{bmatrix} \qquad C = \begin{bmatrix} 3 & 2 \\ 0 & 1 \end{bmatrix} \tag{2.36}$$

definem um jogo de soma geral de dois jogadores, duas ações, com um único equilíbrio de Nash em $(0,5, 0,5)$. Para esse jogo, BOWLING e VELOSO (2002) mostrou que WoLF-IGA não converge. Subsequentemente, BANERJEE e PENG (2002); ZHANG e HUANG (2004) introduziram modificações, mostradas na seção 2.2.1, que induzem convergência. Nessa seção, é feito um estudo comparativo da dinâmica de HEGS proposta e das duas modificações WoLF-IGA citadas.

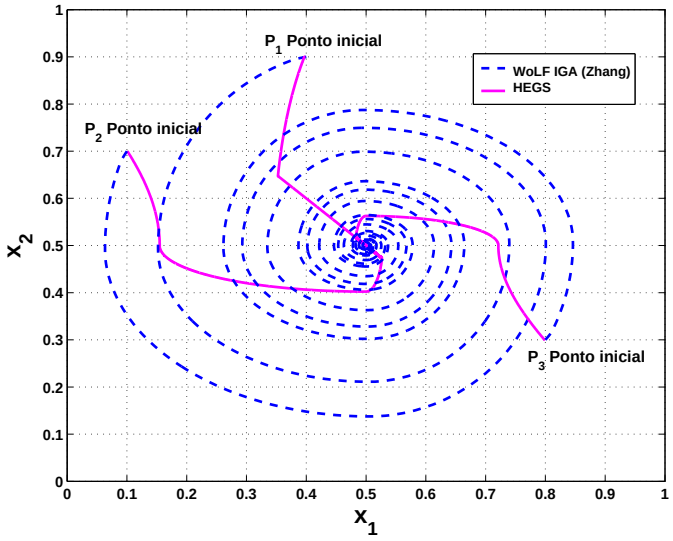
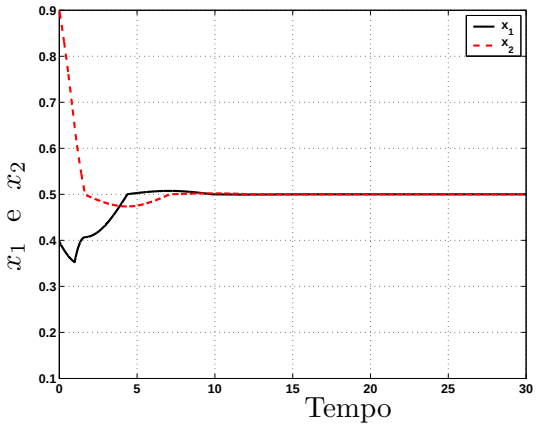


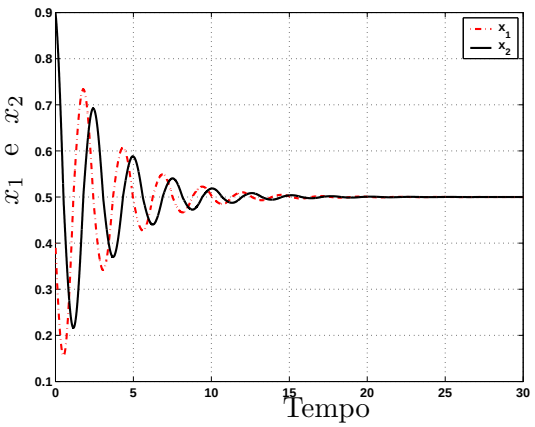
Figura 2.2: Comparação das trajetórias no diagrama de fase com HEGS em linha sólida e em Wolf-IGA em Zhang e Huang (2004) em linha tracejada no exemplo numérico 1. Foram traçadas três trajetórias com pontos iniciais em $P_1 = (0,4, 0,9)$, $P_2 = (0,1, 0,7)$, e $P_3 = (0,8, 0,3)$.

Na figura 2.2 mostram-se os resultados no diagrama de fase $x_1 \times x_2$, aplicando-se HEGS com $k_1 = -1$ nos setores 4 e 8, comparado com WoLF-IGA modificado por (ZHANG e HUANG, 2004). HEGS está identificado por linha tracejada enquanto WoLF-IGA modificado (ZHANG e HUANG, 2004) por linha sólida. Traçaram-se três trajetórias com ponto iniciais de $(0,1, 0,7)$, $(0,4, 0,9)$, $(0,8, 0,3)$. Já na figura 2.3 mostram-se os mesmo resultados, mas por meio do gráfico temporal, com ponto inicial escolhido de $(0,4, 0,9)$. Têm-se, na figura 2.3(a) x_1 em linha pontilhada e x_2 em linha sólida e, na figura 2.3(b), x_1 em linha pontilhada e x_2 em linha sólida de Wolf-IGA em Zhang e Huang (2004).

Na figura 2.4 mostram-se os resultados aplicando-se HEGS com $k_1 = -1$ nos



(a) Gráfico temporal de HEGS com x_1 em linha sólida e x_2 em linha tracejada no exemplo numérico 1, ponto inicial da trajetória $P_1 = (0, 4, 0, 9)$.



(b) Gráfico temporal de Wolf-IGA em Zhang e Huang (2004) com x_1 em linha sólida e x_2 em linha tracejada no exemplo numérico 1, ponto inicial da trajetória $P_1 = (0, 4, 0, 9)$.

Figura 2.3: Comparação temporal de HEGS e em Wolf-IGA em Zhang e Huang (2004) no exemplo numérico 1.

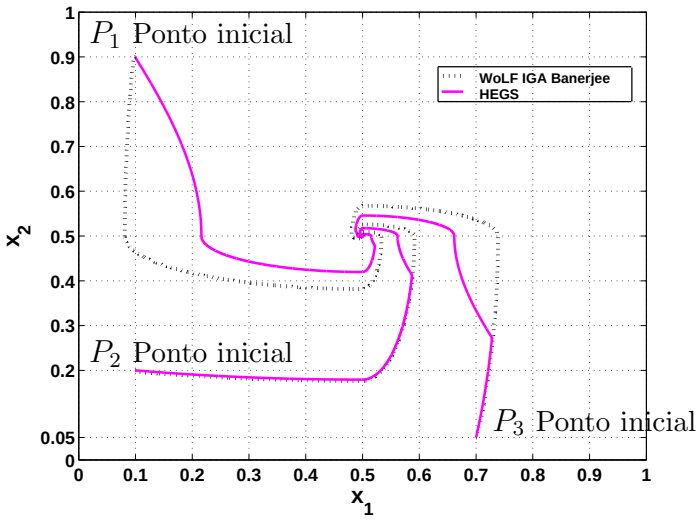
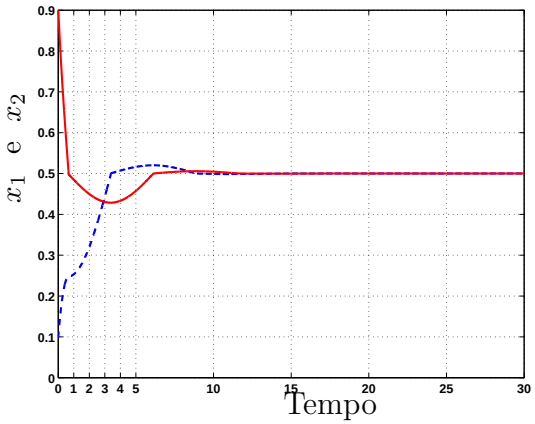


Figura 2.4: Comparação das trajetórias no diagrama de fase com HEGS em linha sólida e com Wolf-IGA em Banerjee e Peng (2002), em linha pontilhada no exemplo numérico 1. Foram traçadas três trajetórias com pontos iniciais $P_1 = (0, 1, 0, 9)$, $P_2 = (0, 1, 0, 2)$ e $P_3 = (0, 7, 0, 05)$.

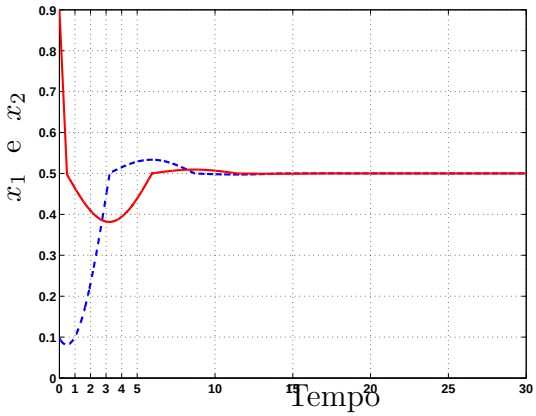
setores 4 e 8, comparado com WoLF-IGA modificado por (BANERJEE e PENG, 2002). Aqui HEGS está identificado por linha sólida enquanto WoLF-IGA modificado por (BANERJEE e PENG, 2002) por linha pontilhada.

Já na figura 2.5 mostram-se os mesmos resultados, mas por meio do gráfico temporal, com ponto inicial escolhido de $(0, 4, 0, 9)$. Tem-se na figura 2.5(a) x_1 em linha tracejada e x_2 em linha sólida em HEGS e na figura 2.5(b) x_1 em linha tracejada e x_2 em linha sólida de Wolf-IGA em Zhang e Huang (2004).

Verificam-se que os segmentos de trajetória de hipérbole são decisivos no cresci-



(a) Gráfico temporal de HEGS com x_1 em linha tracejada e x_2 em linha sólida no exemplo numérico 1, ponto inicial $P_1 = (0, 1, 0, 9)$.



(b) Gráfico temporal de Wolf-IGA em Banerjee e Peng (2002) com x_1 em linha tracejada e x_2 em linha sólida no exemplo numérico 1, ponto inicial $P_1 = (0, 1, 0, 9)$.

Figura 2.5: Comparação temporal de x_1 e x_2 em HEGS e em Wolf-IGA em Banerjee e Peng (2002).

mento da taxa de convergência, para o equilíbrio de Nash, das trajetórias de HEGS em relação às trajetórias de WoLF-IGA.

A escolha do ponto inicial nos setores 4 ou 8 enfatiza a diferença entre as estratégias HEGS e WoLF-IGA, mas qualquer escolha de condição inicial conduzirá a uma convergência mais rápida da estratégia HEGS, já que ela sempre vai por meio de segmentos de hipérbole para o equilíbrio e, esses segmentos, são escolhidos para serem mais rápidos que WoLF-IGA.

2.5 Exemplo numérico 2

Nas figuras 2.6 e 2.7 ilustram-se o diagrama de fase $x_1 \times x_2$ do exemplo numérico 2 com trajetórias de 9×9 pontos iniciais, usando Wolf-IGA e HEGS, respectivamente, num jogo com ponto de equilíbrio de Nash $(0, 9, 0, 9)$, citado em ABDALLAH e LESSER (2008), com as características dadas a seguir na equação (2.40).

De fato, para utilização de Wolf-IGA e HEGS, não é necessário conhecer todos os elementos das matrizes R e C, apenas dois parâmetros de cada matriz definidos como se seguem:

$$\begin{aligned} \bar{r} &= r_{11} + r_{22} - (r_{12} + r_{21}) \\ \check{r} &= -(r_{22} - r_{12}) \\ \bar{c} &= c_{11} + c_{22} - (c_{12} + c_{21}) \\ \check{c} &= -(c_{22} - c_{21}) \end{aligned} \tag{2.37}$$

O equilíbrio $\mathbf{x}^{\text{eq}} = ({x_1}^{\text{eq}}, {x_2}^{\text{eq}})$ é obtido como:

$$
\begin{aligned}
{x_1}^{\text{eq}} &= \frac{(c_{22} - c_{21})}{\bar{c}} = \frac{-\check{c}}{\bar{c}} = 0,9 && (2.38)\\
{x_2}^{\text{eq}} &= \frac{(r_{22} - r_{12})}{\bar{r}} = \frac{-\check{r}}{\bar{r}} = 0,9 && (2.39)
\end{aligned}
$$

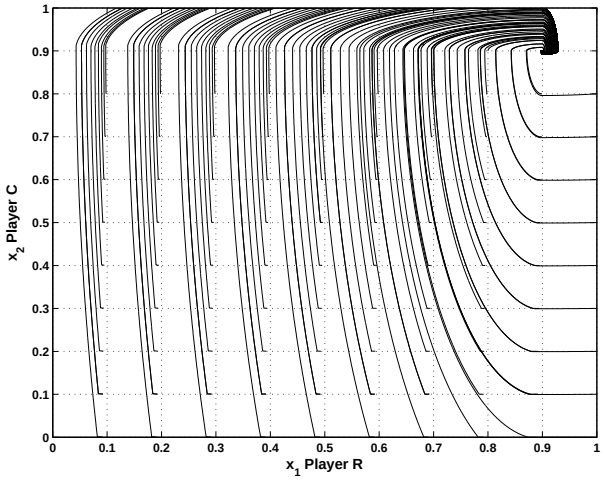


Figura 2.6: Diagrama de fase do exemplo numérico 2 com Wolf-IGA com trajetórias de 9 × 9 pontos iniciais.

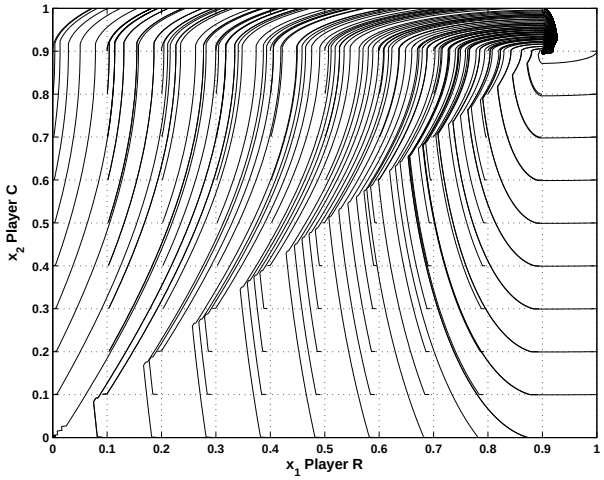


Figura 2.7: Diagrama de fase do exemplo numérico 2 com HEGS com trajetórias de 9 × 9 pontos iniciais.

$$
\begin{aligned}
\bar{r} &= 0,50\\
\check{r} &= 0,45\\
\bar{c} &= -0,50\\
\check{c} &= -0,45
\end{aligned}
\qquad (2.40)
$$

Esse é um exemplo em que $\bar{r} > 0, \bar{c} < 0$ e portanto, nos setores 1 e 5, as trajetórias

de HEGS são hipérboles, já que, no setores restantes, HEGS usa os mesmos ganhos positivos de WoLF-IGA.

Na figura 2.8 exibe-se o diagrama de fase de HEGS para o caso em que $\bar{r} > 0, \bar{c} < 0$, onde o chaveamento ocorre nos setores 1 e 5.

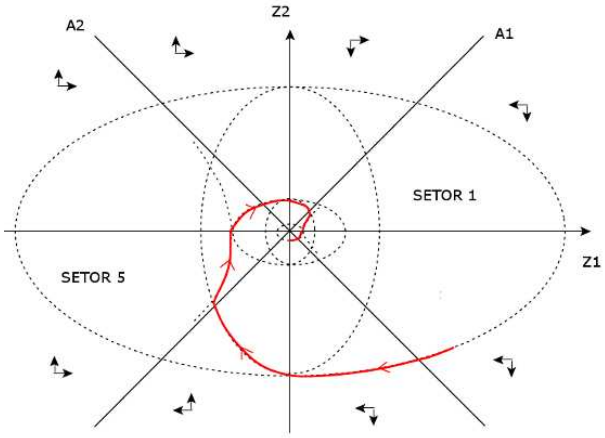


Figura 2.8: Divisão do plano de fase em setores chaveados de 1 até 8 para o caso $\bar{c} > 0$, $\bar{r} < 0$, o que implica movimento no sentido elíptico horário, qualquer que seja $k_i(\mathbf{z}) > 0, i = 1, 2$. Os segmentos de trajetória nos setores 1 e 5 são hipérboles, todos os outros são elipses. As linhas A_1 e A_2 são, respectivamente, as assíntotas estáveis e instáveis para as trajetórias de hipérboles. Pequenas setas indicam as componentes do campo vetorial em cada setor para a dinâmica de HEGS definida no texto.

De acordo com a estratégia de WoLF-IGA, nos setores 1 e 5, os ganhos de realimentação devem ser $k_1 > 0$, k_{1min}, e $k_2 > 0$, k_{2max}, tais que

$$\dot{V}_{W-IGA}(\mathbf{z}) = (k_{1min}(\mathbf{z})\bar{r} + k_{2max}(\mathbf{z})\bar{c})z_1 z_2 \tag{2.41}$$

A estratégia HEGS, nos setores 1,5, requer que os ganhos de realimentação sejam $k_{1HEGS} < 0$, e $k_{2HEGS} > 0$, tais que

$$\dot{V}_{HEGS}(\mathbf{z}) = (k_{1HEGS}(\mathbf{z})\bar{r} + k_{2HEGS}(\mathbf{z})\bar{c})z_1 z_2 \tag{2.42}$$

Comparando (2.41) com (2.32), a condição suficiente para $\dot{V}_{HEGS}(\mathbf{z}) \leq \dot{V}_{W-IGA}(\mathbf{z})$ é

$$k_{2HEGS}(\mathbf{z}) \geq k_{2max}(\mathbf{z}) \tag{2.43}$$

Nota-se que (2.43) implicará na condução de estratégia de HEGS a uma convergência mais rápida nos setores em referência que àquela de WoLF-IGA.

Arbitrou-se $k_{1HEGS} = -0,5$, $k_{2HEGS} = 1$, $k_{1min} = 0,08$, $k_{2max} = 1$, tal que (2.41)

e (2.42), podem ser escritas como:

$$\dot{V}_{W-IGA}(\mathbf{z}) = 0,42\bar{c}z_1z_2 \tag{2.44}$$

$$\dot{V}_{HEGS}(\mathbf{z}) = 1,5\bar{c}z_1z_2 \tag{2.45}$$

Como $z_1z_2 > 0$ nos setores 1 e 5, $\dot{V}_{HEGS}(\mathbf{z})$ é mais de três vezes e meia negativa do que $\dot{V}_{W-IGA}(\mathbf{z})$.

2.6 Uma escolha de ganhos suficiente para HEGS

Na seção anterior escolheu-se de maneira arbitrária os ganhos de modo a garantir $\dot{V}_{W-IGA}(\mathbf{z})$ negativo. Uma escolha de ganhos suficiente nos setores 4 e 8 para $\bar{c} > 0$ e $\bar{r} < 0$ ou nos setores 1 e 5 para $\bar{r} > 0$ e $\bar{c} < 0$ é dada a seguir:

$$k_{1HEGS} = -k_1\frac{sign(\bar{r})}{\bar{r}} \tag{2.46}$$

$$k_{2HEGS} = k_2\frac{sign(\bar{c})}{\bar{c}} \tag{2.47}$$

onde *sign* é a função sinal e $k_1 > 0$ e $k_2 > 0$. A escolha de (2.46) e (2.47) garante que V_{Liap} de (2.32) é negativa. A escolha de $k_1 = k_2 = 1$ torna as equações (2.46) e (2.47) como:

$$k_{1HEGS} = -\frac{sign(\bar{r})}{\bar{r}} \tag{2.48}$$

$$k_{2HEGS} = \frac{sign(\bar{c})}{\bar{c}} \tag{2.49}$$

2.7 Conclusões

Enquadrou-se os métodos de convergência WoLF-IGA e IGA e os seus desenvolvimentos em BANERJEE e PENG (2002) e em ZHANG e HUANG (2004) como uma técnica de estrutura variável, permitindo a revisão das propriedades geométricas das trajetórias de IGA e WoLF-IGA. Esse ponto de vista de controle chaveado permitiu que se provasse a convergência, via função de Liapunov, das propostas de (BOWLING e VELOSO, 2002), (BANERJEE e PENG, 2002) e (ZHANG e HUANG, 2004) de maneira unificada. E ainda uma nova estratégia foi implementada, chamada de chaveamento Gradiente Hiperbólico-Elíptico, *Hyperbolic-Elliptic Gradient Switching*

(HEGS), conduzindo a uma convergência mais rápida das estratégias e das recompensas que (ZHANG e HUANG, 2004).

Ainda que HEGS tenha a convergência no mesmo instante de (BANERJEE e PENG, 2002), a aceleração (nos setores 4 e 8, para $\bar{r} < 0$ e $\bar{c} > 0$ ou nos setores 1 e 5 para $\bar{r} > 0$ e $\bar{c} < 0$) leva à convergência mais rápida nos setores citados. Usando um termo de controle, o tempo de amortecimento de HEGS é menor que o de (BANERJEE e PENG, 2002), o que é vantajoso. O tempo de amortecimento do sinal de saída de um sistema é definido como o período de tempo que garante-se, após o sistema receber um degrau unitário, que o sinal da saída do sistema não varie mais que um determinado percentual do valor assintótico do sinal de saída.

Capítulo 3

Aprendizado por Reforço Multiagente por Gradiente Ascendente (ARM-GA)

O problema de decisão de um agente pode ser visto como o problema de selecionar uma determinada ação quando ele está em um determinado estado. Um exemplo bem conhecido de um problema de decisão envolvendo um único agente é o problema de um bandido com múltiplos braços (um tipo de máquina caça-níquel), mais conhecido pelo sigla MAB (*Multi-Armed Bandit*), no qual o agente deve escolher uma alavanca (ou braço) entre várias. A recompensa de uma determinada escolha segue uma distribuição aleatória (porém fixa). A meta do agente é escolher a alavanca (a ação) que resulta na maior recompensa esperada. Para alcançar esta meta, o agente "recolhe" amostras da distribuição de recompensas associadas a cada ação (simplesmente executando ações diferentes e observando a recompensa obtida). A meta de algoritmos de aprendizado por reforço em geral é estabilizar em (convergir a) uma estratégia que maximiza a recompensa média do agente ao longo de jogos repetidos. Os algoritmos tradicionais desta área, como Q-learning (SUTTON e BARTO, 1998) asseguram convergência a estratégia (também chamada política) ótima em ambiente estacionário, que significa que a distribuição da recompensa associada a cada ação fixa é invariante no tempo.

Em um sistema multiagente, a recompensa recebida quando cada agente executa uma determinada ação depende não somente da sua própria escolha de ação, como também das escolhas dos demais agentes. A título de um exemplo, considere a extensão do problema MAB ao caso multiagente. A recompensa recebida pelo agente A quando escolhe alavanca 1 depende de qual alavanca o agente B tenha escolhido. Se ambos os agentes A e B estão aprendendo e adaptando suas estratégias, a hipótese de estacionaridade é violada (pois a distribuição das recompensas está variando ao longo do tempo) e, portanto, estratégias de aprendizado por reforço concebidas para um único agente podem deixar de convergir. Ademais, no caso multiagente, o critério de otimização também fica menos claro do que no caso de um único agente. Idealmente, é desejável que todos os agentes atinjam um equilíbrio que maximize suas recompensas individuais. Porém, quando não há comunicação entre agentes ou agentes não cooperam, nem sempre é possível atingir um equilíbrio globalmente ótimo (CLAUS e BOUTILIER, 1998). Uma meta alternativa e consagrada é a convergência a um equilíbrio de Nash (EN) (BANERJEE e PENG, 2007; BOWLING, 2005; CONITZER e SANDHOLM, 2006; NASH, 1950, 1951), o qual, por definição, é um máximo local para todos os agentes (no sentido de que nenhum agente poderia obter recompensa maior ao desviar unilateralmente da estratégia correspondente ao EN).

Um aspecto importante para compreender um algoritmo de Aprendizado por Reforço Multiagente (ARM) consiste em analisar a sua dinâmica: isto é, como estratégias de múltiplos agentes evoluem ao longo do tempo enquanto interagem entre si. Tal análise revela não somente se agentes utilizando um determinado al-

goritmo ARM eventualmente convergirão, como também esclarece aspectos qualitativos da trajetória (transitórios, rapidez, suavidade, etc.). A análise da dinâmica de um algoritmo ARM, mesmo nos domínios mais simples (especificamente, jogos de dois jogadores com duas ações cada, que é o foco deste trabalho), é desafiadora e, até o momento, tem sido realizada para poucos algoritmos ARM, e para poucos tipos de jogos e com dinâmicas lineares ou lineares por pedaços (BOWLING e VELOSO, 2002; SINGH *et al.*, 2000). Mais recentemente, alguns outros algoritmos ARM foram propostos (ABDALLAH e LESSER, 2008; BANERJEE e PENG, 2007; BHAYA e MACEDO, 2006; BOWLING e VELOSO, 2002; BOWLING, 2005; CLAUS e BOUTILIER, 1998; CONITZER e SANDHOLM, 2006; HU e WELLMAN, 2003; LITTMAN, 2001; PESHKIN ET AL., 2000; SINGH *et al.*, 2000; ZHANG e HUANG, 2004). A maioria destes algoritmos parte da hipótese básica de que os agentes conhecem a estrutura do jogo (isto é, as respectivas matrizes de recompensa), ou então a localização do equilíbrio de Nash (BANERJEE e PENG, 2007; BOWLING e VELOSO, 2002). Alguns supõem conhecidos quais ações são executadas pelos outros agentes e quais recompensas recebem (CONITZER e SANDHOLM, 2006; HU e WELLMAN, 2003). Tais hipóteses são restritivas em algumas situações nas quais comunicação entre agentes é rara ou ausente (por exemplo, compartilhamento de arquivos no esquema P2P, ou eBay): nestes contextos, um agente raramente tem conhecimento da existência dos outros agentes, e menos conhecimento ainda das ações deles e recompensas associadas. Evidentemente, se os agentes não conhecem o jogo subjacente (especificamente o equilíbrio de Nash) e não se observam mutuamente, então mesmo um jogo simples com dois jogadores e duas ações pode ficar desafiador. Apesar disso, ABDALLAH e LESSER (2008) propuseram um algoritmo ARM no qual cada agente utiliza apenas a informação da recompensa associada e sua própria ação e não possui nenhuma informação sobre o outro agente no jogo. Mostraram, por meio de simulações utilizando alguns exemplos padrão, que o algoritmo proposto por eles (*Weighted Policy Learning* (WPL)) levava a convergência a um EN. Alguns pontos devem ser destacados: a dinâmica WPL é uma dinâmica derivada a partir da dinâmica de gradiente ascendente WoLF-IGA proposta anteriormente por BOWLING e VELOSO (2002), substituindo os ganhos fixos de WoLF-IGA por ganhos dependentes do estado, tornando a dinâmica WPL não-linear. Outro ponto importante, destacado pelos próprios proponentes do método, é que, por ter uma dinâmica não linear, a análise de convergência se torna mais difícil. De fato, a convergência do algoritmo não foi demonstrada no artigo original de ABDALLAH e LESSER (2008), tampouco as possíveis restrições.

Com esses prolegômenos, o plano deste capítulo pode ser descrito da seguinte maneira. Primeiramente, será feita uma revisão de métodos de aprendizado por reforço baseados em gradientes ascendentes, fornecendo uma explicação unificada de

todos eles, por meio da técnica de função de Liapunov com controle, utilizado nesta tese como a ferramenta unificadora. Esta formulação levará naturalmente a uma extensão do método WoLF-IGA, que, entretanto, utilizará as mesmas hipóteses, um tanto restritivas, feitas no trabalho original de BOWLING e VELOSO (2002). Em seguida, o algoritmo WPL, considerado algoritmo do estado da arte, será examinado utilizando a perspectiva de funções de Liapunov e a ideia de equilíbrios virtuais, esclarecendo o funcionamento do método e fornecendo uma prova rigorosa da convergência do método. E ainda serão elencadas as condições de contorno que garantem a convergência de WPL, não apresentadas em ABDALLAH e LESSER (2008).

3.1 Esboço da teoria dos jogos de dois jogadores, duas ações

Esta seção revisa, de forma extremamente resumida, as definições da teoria de jogos que serão utilizadas neste capítulo, com o intuito de torná-lo independente dos demais capítulos. Em seguida, revisam-se os conceitos básicos de algoritmos da classe ARM-GA.

A teoria do jogos clássica lida com conflitos: os indivíduos envolvidos chamados também de agentes ou jogadores, normalmente controlam apenas parte de sua função objetivo (BERTSEKAS, 1995): a outra parte é devida aos outros jogadores e a incertezas do modelo. Jogos são modelos de situações de conflito ou de cooperação: embates aéreos, competições econômicas para fixação de preços de produtos, jogos de salão como xadrez ou conjunto de agentes robóticos buscando em conjunto cumprir um objetivo comum, são alguns exemplos. Uma característica básica nas situações de conflito é que o resultado final depende essencialmente da combinação das estratégias adotadas por cada jogador ou agente [1].

A teoria dos jogos evolucionários lida com os mesmos objetos da teoria dos jogos clássica, a mudança está no enfoque. A teoria dos jogos clássica lida principalmente com jogos entre indivíduos racionais (jogadores). Cada jogador decide entre as opções de jogada (estratégias), visando aumentar a sua recompensa e toma tal decisão sabendo que outros jogadores também assim agirão. A teoria dos jogos evolucionários trata de populações de agentes que interagem entre si num ambiente (jogo) e se comportam ou estão definidos para usar determinada estratégia.

A teoria de jogos fornece um arcabouço para a modelagem da interação entre agentes (jogadores) e foi utilizada na literatura sobre ARM para formular e analisar

[1] As principais referências consultadas da teoria do jogos foram (BARTO *et al.*, 1989; BASAR e OLSDER, 1998; FUDENBERG e TIROLE, 1991; HOFBAUER e SIGMUND, 1998; OWEN, 1995; SIGMUND, 2009)

(a) Coordenação

Ações	a1	a2
a1	$2,1$	$0,0$
a2	$0,0$	$1,2$

(b) Casar moedas

Ações	H	T
H	$1,-1$	$-1,1$
T	$-1,1$	$1,-1$

(c) Ardiloso

Ações	a1	a2
a1	$0,3$	$3,2$
a2	$1,0$	$2,1$

Jogo geral

Ações	a1	a2
a1	r_{11}, c_{11}	r_{12}, c_{12}
a2	r_{21}, c_{21}	r_{22}, c_{22}

Tabela 3.1: Jogos padrão (*benchmark*) de dois jogadores, duas ações. O jogo de coordenação possui dois ENs puros: $((0,1)_r,\ (0,1)_c)$ e $((1,0)_r,\ (1,0)_c)$. Tanto o jogo casar moedas quanto o jogo ardiloso possuem um EN misto no qual todas as ações são escolhidas com a mesma probabilidade, $((0,5,0,5)_r,\ (0,5,0,5)_c)$, onde os subescritos r e c referenciam os jogadores linha (*row*) e coluna.

o problema de convergência de algoritmos ARM(ABDALLAH e LESSER, 2008; BOWLING e VELOSO, 2002; BOWLING, 2005; CLAUS e BOUTILIER, 1998; CONITZER e SANDHOLM, 2006; SINGH *et al.*, 2000; WANG e SANDHOLM, 2002).

Um jogo é representado por uma ênupla, com o número de jogadores, suas regras e quais estratégias possíveis a cada lance (movimento), além da lista de resultados (ou recompensas) para cada combinação de estratégias. Assumindo por hipótese a racionalidade dos jogadores, o objetivo de cada jogador é maximizar a sua recompensa. Um jogo define portanto, de forma compacta, como a recompensa de um agente depende da ação dos outros agentes participando do jogo. Um jogo em forma normal é definido pela ênupla $(n, A_1, \cdots, A_n, R_1, \cdots, R_n)$, sendo n o número de agentes, A_i o conjunto de ações disponíveis para o i-ésimo agente e $R_i : A_1 \times \cdots \times A_n \rightarrow \mathbb{R}$ é a recompensa que o i-ésimo agente receberá em função da ação conjunta de todos os agentes. Quando o jogo possui apenas dois jogadores, torna-se usual e conveniente definir suas ações como uma matriz (estritamente falando, um arranjo) de recompensas conforme mostrado na tabela 3.1. Cada célula (i, j) no arranjo contém um par ordenado de números reais, sendo os elementos do par as recompensas respectivas recebidas pelo jogador linha (jogador 1) e pelo jogador coluna (jogador 2), se o jogador linha escolhe ação i e o jogador coluna escolhe ação j. Na tabela 3.1 mostram-se exemplos padrão (*benchmark*) [2] que foram utilizados na avaliação de algoritmos propostos anteriormente na literatura.

[2]Casar moedas (*Matching Pennies*) é um jogo de soma zero em que cada jogador possui uma moeda (chamada de *penny* em inglês). Cada jogador opta por escolher cara ou coroa, arremessando a sua moeda para o alto e escondendo a mesma ao cair na mão. Se ambas moedas exibirem duas caras ou duas coroas, o jogador R vence, caso contrário o jogador C vence.

Uma *política* ou *estratégia* de um agente i é denotada $\pi_i \in PD(A_i)$, sendo $PD(A_i)$ o conjunto de distribuições de probabilidade sobre ações A_i. A probabilidade de escolher uma ação a_k de acordo com a política π_i é denotada $\pi_i(a_k)$. Uma política é denominada *determinística* ou *pura* se a probabilidade de escolher uma ação é 1, enquanto a probabilidade de escolher as demais ações é nula. Em notação matemática, π_i é pura se $\exists k : \pi_i(a_k) = 1$ e, ao mesmo tempo, $\forall j \neq k, \pi_i(a_j) = 0$. Caso contrário, a política é denominada *mista* ou *estocástica*.

Uma *política conjunta* π é a coleção das políticas dos agentes individuais, i.e., $\pi = \langle \pi_1, \pi_2, \ldots, \pi_n \rangle$, sendo n o número de agentes. Uma notação abreviada conveniente, muito utilizada na literatura, é $\pi = \langle \pi_i, \pi_{-i} \rangle$, sendo π_{-i} a coleção das políticas de todos os agentes, excluindo o i-ésimo.

Da mesma forma, seja $A_{-i} := \{\langle a_1, \ldots, a_n \rangle : i \neq j\}$. A recompensa esperada que o agente i obteria, se os demais agentes seguissem a política conjunta π, é $V_i(\langle \pi_i, \pi_{-i} \rangle) = \sum_{a_i \in A_i} \pi_i(a_i)\pi_{-i}(a_{-i})R_i(a_i, a_{-i})$; isto é, a recompensa média calculada ao longo da política conjunta. Uma política conjunta é denominada um *equilíbrio de Nash* (EN) se nenhum agente consegue uma recompensa esperada maior por meio de uma mudança unilateral de sua política. Utilizando a notação matemática estabelecida acima, $\langle \pi_i^*, \pi_{-i}^* \rangle$ é um equilíbrio de Nash se, para todo i, $V_i(\langle \pi_i^*, \pi_{-i}^* \rangle) \geq V_i(\langle \pi_i, \pi_{-i}^* \rangle)$. Um EN é puro se todas as políticas que o constituem sejam puras; caso contrário, é denominado misto ou estocástico. Todo jogo possui ao menos um EN, porém poderia não possuir qualquer equilíbrio puro.

A título de ilustração destes conceitos, considere novamente os jogos na tabela 3.1. O jogo de coordenação é um exemplo de jogo que possui ao menos um EN puro. O jogo casar moedas é exemplo de jogo que não possui nenhum EN puro e possui apenas um EN misto, no qual cada jogador escolhe ações a_1 e a_2 com probabilidades iguais $(0,5)$. A convergência de algoritmos ARM-GA em jogos com EN puro é mais fácil do que em jogos nos quais o único EN é misto (BOWLING e VELOSO, 2002; SINGH *et al.*, 2000; ZINKEVICH, 2003). O jogo ardiloso (*tricky*) é parecido com o jogo casar moedas no sentido de possuir apenas um EN misto (e nenhum puro); entretanto, foi mostrado em (ABDALLAH e LESSER, 2008; BOWLING e VELOSO, 2002) que alguns algoritmos do tipo ARM-GA que conseguem convergir no jogo casar moedas divergem no jogo ardiloso, portanto ele será considerado um exemplo modelo nesta tese.

Antes de discutir os algoritmos ARM na seção 3.2, vale ressaltar dois aspectos cruciais. O primeiro ponto é a hipótese geral no contexto ARM de que os agentes jogam o mesmo jogo repetidas vezes, para um número grande de vezes. Esta hipótese é necessária para que exista dinâmica de aprendizado. A segunda hipótese é que as recompensas são determinísticas, uma vez especificada a ação conjunta. Entretanto, do ponto de vista de cada agente, as recompensas são estocásticas, por causa da

aleatoriedade imposta pela ação dos outros agentes no jogo.

Definição 3.1.1 (subclasses de jogos dois jogadores, duas ações) *No cenário de dois jogadores, duas ações, sejam r_{ij} a recompensa para o jogador linha (jogador 1) recebe por escolher a estratégia pura s_i do conjunto S_1 quando o jogador coluna escolhe a estratégia s_j do conjunto S_2. c_{ij} é a recompensa que o jogador coluna (jogador 2) recebe por escolher a estratégia pura s_j do conjunto S_2 quando o jogador coluna escolhe a estratégia s_i do conjunto S_1. sendo os elementos das seguintes matrizes de recompensa 3.1:*

$$R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \qquad C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \tag{3.1}$$

Os jogos podem ser classificados em três subclasses (ver VEGA-REDONDO, 2003, página 403):

Subclasse 1: uma das equações a seguir vale:

$$\begin{aligned} (r_{11} - r_{21})(r_{12} - r_{22}) &> 0 \\ (c_{11} - c_{21})(c_{12} - c_{22}) &> 0 \end{aligned} \tag{3.2}$$

A subclasse 1 possui dois equilíbrios. Essa subclasse inclui aqueles jogos em que cada jogador tem uma estratégia dominante [3]*, e, portanto, uma dinâmica menos interessante.*

Exemplo:

$$R = \begin{bmatrix} 1 & 5 \\ 0 & 3 \end{bmatrix} \qquad C = \begin{bmatrix} 1 & 0 \\ 5 & 3 \end{bmatrix} \tag{3.3}$$

Nesse jogo, a estratégia linha 1 do jogador linha é dominante em relação à estratégia da linha 2, pois, observando a matriz R, a recompensa do jogador 1 será $1 > 0$ se o jogador escolher a estratégia da coluna 1, e $5 > 3$ se o jogador 2 escolher a estratégia da coluna 2. Esse jogo é subclasse 1 conforme equação (3.4).

$$\begin{aligned} (r_{11} - r_{21})(r_{12} - r_{22}) &= (1-0)(5-3) > 0 \\ (c_{11} - c_{21})(c_{12} - c_{22}) &= (1-5)(0-3) > 0 \end{aligned} \tag{3.4}$$

[3] A estratégia dominante de um jogador é aquela que é sempre melhor que as outras, não importando o que o outro jogador faça.

Subclasse 2: todas as equações a seguir valem:

$$\begin{aligned}(r_{11} - r_{21})(r_{12} - r_{22}) &< 0 \\ (c_{11} - c_{21})(c_{12} - c_{22}) &< 0 \\ (r_{11} - r_{21})(c_{11} - c_{12}) &> 0\end{aligned} \tag{3.5}$$

A subclasse 2 pode ser vista como uma extensão da subclasse 1, possuindo dois equilíbrios puros. Ambos equilíbrios podem ser atingidos dependendo das matrizes de recompensa e das condições iniciais. Há ainda nesta subclasse um terceiro equilíbrio instável, ou seja pequenas pertubações no ponto conduzem as trajetórias de um dos dois equilíbrios puros.

Subclasse 3: todas as equações a seguir valem:

$$\begin{aligned}(r_{11} - r_{21})(r_{12} - r_{22}) &< 0 \\ (c_{11} - c_{21})(c_{12} - c_{22}) &< 0 \\ (r_{11} - r_{21})(c_{11} - c_{12}) &< 0\end{aligned} \tag{3.6}$$

A subclasse 3 possui um equilíbrio misto. Para o problema de aprendizado em jogos, esta terceira subclasse é a mais desafiadora, pois vários algoritmos propostos na literatura deixam de convergir para ela. Será, portanto, a subclasse focada nesta tese.

Sejam as definidas as seguintes constantes:

$$\begin{aligned}\bar{r} &= r_{11} + r_{22} - (r_{12} + r_{21}) \\ \check{r} &= -(r_{22} - r_{12}) \\ \bar{c} &= c_{11} + c_{22} - (c_{12} + c_{21}) \\ \check{c} &= -(c_{22} - c_{21})\end{aligned} \tag{3.7}$$

Reescrevendo as equações da subclasse 3 em função das equações (3.7) tem-se:

$$\begin{aligned}(\bar{r} + \check{r})\check{r} &< 0 \\ (\bar{c} + \check{c})\check{c} &< 0 \\ (\bar{r} + \check{r})(\bar{c} + \check{c}) &< 0\end{aligned} \tag{3.8}$$

Na figura 3.1 ilustra-se uma trajetória de um jogo subclasse 3 que possui apenas um EN misto, com as matrizes de recompensa dadas na equação (3.9). Essa subclasse de jogo é a de interesse nesta tese.

$$R = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad C = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \tag{3.9}$$

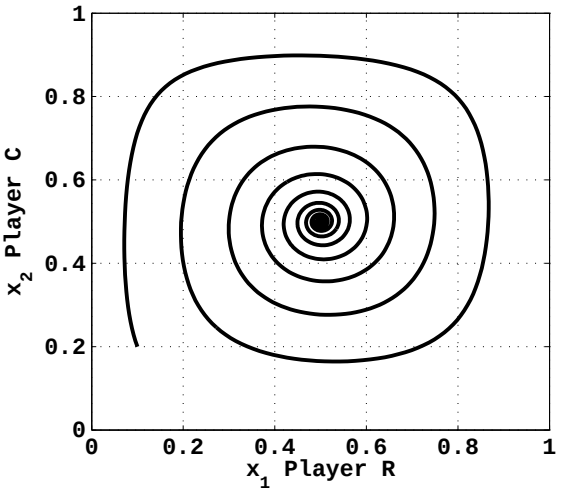


Figura 3.1: Trajetória de aprendizado por reforço num jogo subclasse 3 com um único EN misto.

3.2 Aprendizado por Reforço Multiagente (ARM)

Os primeiros algoritmos ARM eram baseados no algoritmo de Q-Aprendizado (SUTTON e BARTO, 1998) e portanto poderiam aprender apenas políticas determinísticas, limitando sua aplicabilidade. Uma outra classe de algoritmos ARM é a chamada classe de aprendizes de equilíbrio (*Equilibrium Learners*), como Nash-Q (HU e WELLMAN, 2003), e AWESOME (CONITZER e SANDHOLM, 2006). A maioria destes algoritmos supõe que cada agente observa as ações dos outros agentes, além de conhecer o jogo. Cada agente computa o EN e o objetivo do aprendizado é a convergência de todos os agentes a um determinado EN, supondo ainda que todos os agentes executam o mesmo algoritmo ARM. A observação de outros agentes ou o conhecimento da estrutura do jogo subjacente não são hipóteses razoáveis em todas as situações e isto motivou o desenvolvimento de aprendizes utilizando gradientes ascendentes. Algoritmos ARM-GA aprendem uma política estocástica seguindo diretamente o gradiente da recompensa esperada. A habilidade de aprender uma política estocástica é particularmente importante quando o mundo não está completamente observável ou quando existe competição. Considere, por exemplo, um robô cego em um labirinto: sendo cego, o robô não distingue entre locais distintos no labirinto. Nesta situação, qualquer política determinística (que sempre escolhe uma determinada ação em toda parte) poderia não conseguir sair do labirinto, ao passo que uma política estocástica que escolhe cada ação com uma determinada probabilidade eventualmente encontrará a saída. De forma análoga, em um domínio competitivo, uma política estocástica pode ser a única política estável (um exemplo

disso é o EN em um jogo competitivo). No restante desta seção, será feita uma revisão rápida da literatura sobre esta família de algoritmos, pois o foco desta tese está na análise e projeto de algoritmos desta classe.

Um dos primeiros algoritmos desta classe a ser proposto, o algoritmo chamado Infinitesimal Gradient Ascent (IGA) (SINGH *et al.*, 2000) e sua subsequente generalização (*generalized* IGA, chamado GIGA) (ZINKEVICH, 2003) teve sua convergência provada em jogos com ENs puros. Entretanto, ambos os algoritmos deixam de convergir em jogos com EN mistos e portanto deixam de ser adequados para aplicações que exigem políticas mistas.

Várias modificações a IGA e GIGA foram propostas para evitar divergência em jogos com EN mistos: IGA/PHC-WoLF (BOWLING e VELOSO, 2002), PHC-PDWoLF (BANERJEE e PENG, 2003) e GIGA-WoLF (BOWLING, 2005). Todos estes algoritmos usam alguma forma da *heurística Win or Learn Fast* (WoLF) (Ganhe ou Aprenda Rápido), proposta em BOWLING e VELOSO (2002) e cuja proposta é chavear entre um aprendizado rápido se o agente está aquém da sua política do EN (isto se define como perder) e um aprendizado mais lento caso o agente esteja além (recebendo recompensa maior) da política EN. Evidentemente, uma hipótese básica (e restritiva) desta política é que ela não pode ser implementada a menos que os agentes conheçam o jogo e possam computar a política EN correspondente. Portanto, uma implementação prática da heurística WoLF deverá utilizar algum tipo de aproximação ou estimação da recompensa associada a uma política EN desejada de cada agente. ABDALLAH e LESSER (2008) propuseram uma nova heurística, denominada *Weighted Policy Learner* (WPL) (Aprendizado com Ponderação das Políticas) que não requer conhecimento do EN e converge em classe mais ampla de jogos.

Neste contexto, a contribuição deste capítulo é uma análise rigorosa com prova de convergência do algoritmo heurístico WPL.

3.3 Algoritmos de Aprendizado por Reforço Multiagente baseados em Gradientes Ascendentes (ARM-GA)

O primeiro algoritmo proposto da classe ARM-GA foi o *Infinitesimal Gradient Ascent* (IGA) (SINGH *et al.*, 2000) analisado no capítulo 2 desta tese. Nesta seção revisitam-se os algoritmos do capítulo 2, utilizando uma notação mais geral, a fim de generalizar estes e chegar ao algoritmo WPL.

IGA é um algoritmo simples de gradiente ascendente no qual cada agente i atualiza sua política π_i para seguir o gradiente de recompensas esperadas (também

conhecido como função valor) V_i, de acordo com as seguintes equações:

$$\begin{aligned} \Delta\pi_i^{t+1} &\leftarrow \eta\frac{\partial V_i(\pi^t)}{\partial \pi_i} \\ \pi_i^{t+1} &\leftarrow \text{Proj}(\pi_i^t + \Delta\pi_i^{t+1}) \end{aligned} \tag{3.10}$$

O parâmetro η é denominado taxa de aprendizado da política e, convencionalmente, se aproxima a zero no limite ($\eta \to 0$), motivando a utilização da palavra infinitesimal no nome do método. A função Proj denota projeção da política atualizada ao espaço de políticas válidas: em termos simples, como uma política nada mais é do que uma probabilidade de escolha de uma determinada ação, esta função projeta ou limita políticas atualizadas, calculadas pelo termo $\pi_i^t + \Delta\pi_i^{t+1}$ ao intervalo $[0, 1]$ para que possam representar probabilidades. A definição mais geral desta função, dada em ZINKEVICH (2003), é $\text{Proj}(x) := \text{argmin}_{x' \in \{\text{política válida}\}}|x - x'|$, sendo $|x - x'|$ a distância Euclideana entre x e x'. Formalmente, uma política válida π sobre um conjunto de ações A deve satisfazer as seguintes condições: (i) para toda ação $a \in A$, $0 \leq \pi(a) \leq 1$ e (ii) $\sum_\pi := \sum_{a \in A} \pi(a) = 1$. Em linguagem geométrica, o espaço de políticas válidas é um simplex (segmento da reta entre os pontos $(1, 0), (0, 1)$ no caso de duas ações). Um política conjunta, π, é um ponto neste simplex. É bem possível, quando se está seguindo um gradiente aproximado, que a soma das políticas atualizadas ultrapasse (escape) deste simplex: a tarefa da função de projeção é justamente trazer estas políticas de volta ao simplex válido, substituindo a política inválida pela política válida mais próxima.

O algoritmo IGA não converge em todo jogo de dois jogadores, duas ações. Especificamente, deixa de convergir em jogos que não possuem EN puros, porém apenas EN mistos (SINGH *et al.*, 2000). Em seguida, BOWLING e VELOSO (2002) propuseram a modificação chamada WoLF-IGA para melhorar as propriedades do algoritmo IGA e propiciar convergência nos jogos que possuem EN mistos. A ideia (heurística), proposta sem prova formal em BOWLING e VELOSO (2002), é a seguinte: se o jogador está obtendo uma recompensa abaixo da recompensa que obteria se jogasse a estratégia EN, então a taxa de aprendizado deve ser maior; caso contrário, a taxa de aprendizado deve ser menor. Formalmente, sendo π_i^* a política EN para agente i, e $\eta_{\text{perd}}, \eta_{\text{ganh}}$ as taxas de aprendizado, o algoritmo WoLF-IGA é representado pela seguinte dinâmica chaveada:

$$\begin{aligned} \Delta\pi_i(a) &\leftarrow \frac{\partial V_i(\pi)}{\partial \pi_i} s_i, \text{sendo } s_i = \begin{cases} \eta_{\text{perd}}, \text{ se } V_i(\pi_i, \pi_{-i}) < V_i(\pi_i^*, \pi_{-i}) \\ \eta_{\text{ganh}}, \text{ senão} \end{cases} \\ \pi_i &\leftarrow \text{Proj}(\pi_i + \Delta\pi_i) \end{aligned} \tag{3.11}$$

Argumentos geométricos foram utilizados em BOWLING e VELOSO (2002) para afirmar a convergência da heurística WoLF-IGA em jogos com dois agentes e duas

ações: uma prova rigorosa será dada na seção seguinte, utilizando um função de Liapunov. Deve ser ressaltado que a heurística WoLF-IGA possui aplicabilidade limitada na prática, pois exige que cada agente conheça sua política de equilíbrio de Nash, o que implica no conhecimento do jogo subjacente (matrizes R, C). Uma aproximação a WoLF-IGA, chamada PHC-WoLF, proposta pelos mesmos autores (BOWLING e VELOSO, 2002), utilizou a ideia de estimar a estratégia EN utilizando a média da estratégia do próprio agente ao longo do tempo. Entretanto, essa aproximação apresenta problemas de convergência em determinados jogos (por exemplo, no jogo ardiloso (*tricky*).

O último algoritmo deste ciclo, GIGA-WoLF, estende o algoritmo GIGA, munindo-o da heurística WoLF. GIGA-WoLF armazena duas políticas π e ν ao longo do tempo. A política π é utilizada para escolher ações a serem executadas, ao passo que a política ν é utilizada para aproximar a política EN. As equações de GIGA-WoLF são as seguintes (BOWLING, 2005):

$$\begin{aligned} \hat{\pi}^{t+1} &= \text{Proj}(\pi^t + \delta r^t) \\ \nu^{t+1} &= \text{Proj}(\pi^t + \delta r^t/3) \\ \eta^{t+1} &= \min\left(1, \frac{\|\nu^{t+1} - \nu^t\|}{\nu^{t+1} - \hat{\pi}^t}\right) \\ \pi^{t+1} &= \hat{\pi}^{t+1} + \eta^{t+1}(\nu^{t+1} - \hat{\pi}^{t+1}) \end{aligned} \tag{3.12}$$

A ideia principal deste algoritmo é uma modificação da heurística WoLF: o agente i atualiza sua política π_i mais rápido se está tendo recompensa menor do que a da política ν, i.e., $V^{\pi_i} < V^{\nu}$. Como, por projeto, ν se movimenta mais lentamente que π, GIGA-WoLF utiliza ν para se dar conta de que necessita mudar a direção atual do gradiente. Isto proporciona uma melhora na convergência do algoritmo. Entretanto, ainda existem problemas de convergência para alguns jogos, além da falta de uma prova rigorosa.

3.4 Dinâmica discreta de algoritmos ARM-GA

No artigo pioneiro de SINGH *et al.* (2000), foi utilizada uma equação ordinária diferencial para modelar a dinâmica IGA. A análise, como em grande parte da literatura e nesta tese, é feita para o caso de dois jogadores, duas ações. A ação conjunta de dois jogadores no instante t será denotada pelo par (x_1^t, x_2^t), sendo x_1^t (resp. x_2^t) a probabilidade do primeiro [linha] (resp. segundo [coluna]) jogador escolher sua primeira ação. Em outras palavras, $\pi_1 = (x_1^t, 1 - x_1^t)$ é a política do jogador linha (primeiro jogador) e $\pi_2 = (x_2^t, 1 - x_2^t)$ é a política do jogador coluna (segundo jogador). O superescrito será omitido sempre que não haja risco de confusão.

Utilizando a notação padrão (r_{ij}, c_{ij} sendo recompensas dos jogadores linha e coluna e V_r, V_c sendo as funções valor esperado das respectivas recompensas), a dinâmica IGA, em tempo discreto, do jogador linha, pode ser expressa da seguinte maneira:

$$x_1^{t+1} = x_1^t + \eta \frac{\partial V_r(x_1^t, x_2^t)}{\partial x_1} = x_1^t + \eta(V_r(1, x_2^t) - V_r(0, x_2^t)) \tag{3.13}$$

sendo

$$\begin{aligned} V_r(1, x_2^t) - V_r(0, x_2^t) &= (r_{11}x_2^t + r_{12}(1 - x_2^t)) - (r_{21}x_2^t + r_{22}(1 - x_2^t)) \\ &= x_2^t(r_{11} - r_{12} - r_{21} + r_{22}) + (r_{12} - r_{22}) \end{aligned} \tag{3.14}$$

De forma análoga, obtém-se a dinâmica discreta do jogador coluna:

$$x_2^{t+1} = x_2^t + \eta \frac{\partial V_c(x_1^t, x_2^t)}{\partial x_2} = x_2^t + \eta(V_c(x_1^t, 1) - V_c(x_1^t, 0)) \tag{3.15}$$

sendo

$$\begin{aligned} V_c(x_1^t, 1) - V_c(x_1^t, 0) &= (c_{11}x_1^t + c_{21}(1 - x_1^t)) - (c_{12}x_1^t + c_{22}(1 - x_1^t)) \\ &= x_1^t(c_{11} - c_{12} - c_{21} + c_{22}) + (c_{21} - c_{22}) \end{aligned} \tag{3.16}$$

O significado completo destas equações, bem como a análise do sistema dinâmico contínuo associado ao sistema discreto descrito acima, serão abordados na seção seguinte.

3.5 WPL

Conforme mencionado acima, a limitação dos algoritmos WoLF-IGA e, de modo mais geral, a classe HEGS, é a necessidade de conhecer a política EN. Nesta seção, será apresentado o algoritmo *Weighted Policy Learner* (WPL), proposta sem prova de convergência em ABDALLAH e LESSER (2008), que não necessita desta hipótese. Será também demonstrado que o algoritmo WPL é suscetível ao mesmo tipo de análise via função de Liapunov, chegando-se a uma prova rigorosa de sua convergência.

A dinâmica discreta do algoritmo WPL é dada pelas seguintes equações:

$$\begin{aligned} \Delta\pi_i(a) &\leftarrow \tfrac{\partial V_i(\pi)}{\partial \pi_i(a)} \eta s_i, \text{sendo } s_i = \begin{cases} \pi_i(a) & \text{se } \frac{\partial V_i(\pi)}{\partial \pi_i(a)} < 0 \\ 1 - \pi_i(a) & \text{senão} \end{cases} \\ \pi_i &\leftarrow \text{Proj}(\pi_i + \Delta\pi_i) \end{aligned} \tag{3.17}$$

A função Proj é a mesma adotada pelo ZINKEVICH (2003), *epsilon*-modificada: isto é, para toda ação a, a probabilidade $\pi(a) \in [\epsilon, 1]$ (ao invés de $[0, 1]$), onde ϵ é um

incremental positivo menor que 1. A intuição por trás deste algoritmo é a seguinte. Se o gradiente para uma determinada ação é negativo, então ele é ponderado por $\pi_i(a)$; senão, o gradiente positivo é ponderado por $(1 - \pi_i(a))$. Isto significa que a probabilidade de escolher uma boa ação decresce por uma taxa que diminui na medida em que esta probabilidade se aproxima a um (fronteira do simplex viável); da mesma forma, a probabilidade de escolher uma ação ruim decresce por uma taxa que também decresce conforme esta probabilidade se aproxima a zero. Uma discussão maior da heurística que motiva o método pode ser encontrado em (ABDALLAH e LESSER, 2008).

No caso de um jogo de dois jogadores, duas ações, a dinâmica discreta WPL pode ser escrita da seguinte maneira, para $i, j = 1, 2$:

$$x_i^t \leftarrow x_i^{t-1} + \eta(\bar{g}x_j^{t-1} + \check{g})s_i, \text{sendo } s_i = \begin{cases} (1 - x_i^{t-1}) & \text{se } \bar{g}x_j^{t-1} + \check{g} > 0 \\ x_i^{t-1} & \text{senão} \end{cases} \tag{3.18}$$

sendo $g = r$, se $i = 1$ e $g = c$, se $i = 2$. Note, ainda, que, $\bar{r}x_2^{t-1} + \check{r} = V_r(1, x_2^{t-1}) - V_r(0, x_2^{t-1})$ e $\bar{c}x_1^{t-1} + \check{c} = V_c(x_1^{t-1}, 1) - V_c(x_1^{t-1}, 0)$. No limite, quando a taxa $\eta \to 0$, obtêm-se as equações diferenciais:[4]

$$\begin{aligned} \dot{x}_1 &= (\bar{r}x_2 + \check{r})s_i, \text{sendo } s_i = \begin{cases} (1 - x_1) & \text{se } \bar{r}x_2 + \check{r} > 0 \\ x_1 & \text{senão} \end{cases} \\ \dot{x}_2 &= (\bar{c}x_1 + \check{c})s_i, \text{sendo } s_i = \begin{cases} (1 - x_2) & \text{se } \bar{c}x_1 + \check{c} > 0 \\ x_2 & \text{senão} \end{cases} \end{aligned} \tag{3.19}$$

Nesta seção, será examinada a convergência das trajetórias do sistema WPL contínuo (3.19), para fornecer prova de convergência, ausente na proposta original em ABDALLAH e LESSER (2008), e também para obter um melhor entendimento do funcionamento do algoritmo. A primeira observação é que a interseção dos dois chaveamentos das equações (3.19) se dá no ponto indicado na equação (3.20) a seguir.

$$x_1{}^{\text{eq}} = -\frac{\check{c}}{\bar{c}} \tag{3.20}$$

$$x_2{}^{\text{eq}} = -\frac{\check{r}}{\bar{r}} \tag{3.21}$$

[4]Aproveita-se aqui para corrigir um erro tipográfico no artigo ABDALLAH e LESSER (2008), que originalmente propôs a dinâmica WPL. Na versão do artigo, aparecem versões atrasadas dos estados (x_1^{t-1}, x_2^{t-1}) do lado direito da equação (3.19). Estes atrasos, naturais no caso discreto que motivou a introdução da dinâmica contínua, criam uma complexidade muito grande na análise do sistema contínuo. Ademais, a julgar pelas simulações e análises parciais apresentadas pelos próprios autores (ABDALLAH e LESSER, 2008, ver página 532), não era essa a intenção dos autores.

A segunda observação é que a dinâmica WPL é composta de quatro subsistemas e que há um chaveamento entre estes que depende do estado (x_1, x_2) do sistema. Percebe-se, entretanto, que, diferente do HEGS, que chaveia apenas os ganhos em função dos estados, portanto não mudando o ponto de equilíbrio, no caso do WPL, há chaveamento entre dinâmicas não lineares distintas. Inicia-se, portanto, a análise do sistema WPL (3.19) com o cálculo dos pontos de equilíbrio e a determinação da natureza de cada um. Conforme está exibido na figura 3.2, cada subsistema possui

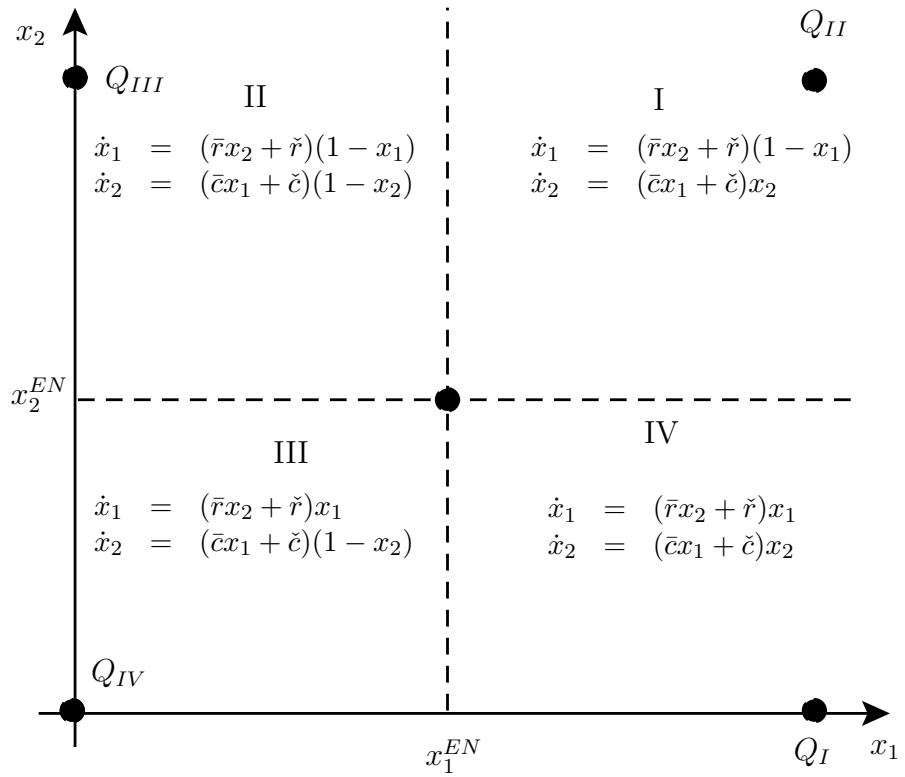


Figura 3.2: Diagrama de fase da dinâmica chaveada WPL, em que se mostra os pontos de equilíbrio de cada subsistema. Cada subsistema possui dois pontos de equilíbrio: um no ponto de interseção das retas pontilhadas que é o EN (x_1^{EN}, x_2^{EN}) e o outro no ponto denotado Q_i, subescrito i em letra romana.

um ponto de equilíbrio no ponto (x_1^{EN}, x_2^{EN}) e outro no ponto Q_i, i em letra romana. O ponto notável é que, enquanto o ponto (x_1^{EN}, x_2^{EN}) se localiza na fronteira da região correspondente à dinâmica (para cada uma delas), o outro ponto Q_i pertence a uma região onde a dinâmica já é outra (por ter chaveado). Este tipo de ponto de equilíbrio, que não pertence à região da sua dinâmica correspondente, é denominado *ponto de equilíbrio virtual*, citado em COSTA *et al.* (2000). Evidentemente, as trajetórias de um subsistema jamais poderiam convergir a um equilíbrio virtual, ainda que este fosse estável, pois na tentativa de se aproximar a este, que se localiza em outra região, entrariam em região contígua onde a dinâmica e, consequentemente, os equilíbrios, reais e virtuais, são diferentes. Sob esta perspectiva, fundamental no projeto de sistemas de controle à estrutura variável, pode dizer-se que o objetivo

do projeto é o arranjo engenhoso dos pontos de equilíbrio reais e virtuais de tal modo que todas as trajetórias de todos os subsistemas são levados a um ponto de equilíbrio ou ciclo limite estável. Para mais detalhes e exemplos do uso desta técnica, veja (MEZA *et al.*, 2005).

No caso específico da dinâmica WPL, será feita a seguinte análise: linearização da dinâmica de cada subsistema em torno de cada ponto de equilíbrio, para determinar o seu tipo; em seguida, uma análise, via função de Liapunov, daqueles pontos que poderiam se candidatar a equilíbrios globalmente estáveis.

3.5.1 Linearização dos subsistemas da dinâmica WPL

Os pontos de equilíbrio do subsistema I são $(1,0)$ e $(-\check{r}/\bar{r}, -\check{c}/\bar{c})$. O Jacobiano $J_I(x_1, x_2)$ é dada por

$$J_I(x_1, x_2) = \begin{bmatrix} -\bar{r}x_2 - \check{r} & \bar{r}(1-x_1) \\ \bar{c}x_2 & \bar{c}x_1 + \check{c} \end{bmatrix} \tag{3.22}$$

Avaliando o Jacobiano J_I nos pontos de equilíbrio $(1,0)$ (virtual) e $(-\check{r}/\bar{r}, -\check{c}/\bar{c})$ (real), obtém-se:

$J_I \downarrow$ e $x \rightarrow$	$(1,0)$	$(-\check{r}/\bar{r}, -\check{c}/\bar{c})$
$J_I(x_1, x_2)$	$\begin{bmatrix} -\check{r} & 0 \\ 0 & \bar{c}+\check{c} \end{bmatrix}$	$\begin{bmatrix} (\bar{r}\check{c} - \check{r}\bar{c})/\bar{c} & \bar{r}+\check{r} \\ -\check{c} & (\check{c}\bar{r} - \bar{c}\check{r})/\bar{r} \end{bmatrix}$
Tipo	Virtual	Real

Os pontos de equilíbrio do subsistema II são $(1,1)$ e $(-\check{r}/\bar{r}, -\check{c}/\bar{c})$. O Jacobiano $J_{II}(x_1, x_2)$ é dada por

$$J_{II}(x_1, x_2) = \begin{bmatrix} -\bar{r}x_2 - \check{r} & \bar{r}(1-x_1) \\ \bar{c}(1-x_2) & -\bar{c}x_1 - \check{c} \end{bmatrix} \tag{3.23}$$

Avaliando o Jacobiano J_{II} nos pontos de equilíbrio $(1,1)$ (virtual) e $(-\check{r}/\bar{r}, -\check{c}/\bar{c})$ (real), obtém-se:

$J_{II} \downarrow$ e $x \rightarrow$	$(1,1)$	$(-\check{r}/\bar{r}, -\check{c}/\bar{c})$
$J_{II}(x_1, x_2)$	$\begin{bmatrix} -(\bar{r}+\check{r}) & 0 \\ 0 & -(\bar{c}+\check{c}) \end{bmatrix}$	$\begin{bmatrix} (\bar{r}\check{c} - \check{r}\bar{c})/\bar{c} & \bar{r}+\check{r} \\ \bar{c}+\check{c} & (\bar{c}\check{r} - \check{c}\bar{r})/\bar{r} \end{bmatrix}$
Tipo	Virtual	Real

Os pontos de equilíbrio do subsistema III são $(1,1)$ e $(-\check{r}/\bar{r}, -\check{c}/\bar{c})$. O Jacobiano $J_{III}(x_1, x_2)$ é dada por

$$J_{III}(x_1, x_2) = \begin{bmatrix} (\bar{r}x_2 + \check{r}) & \bar{r}x_1 \\ \bar{c}(1-x_2) & -(\bar{c}x_1 + \check{c}) \end{bmatrix} \tag{3.24}$$

Avaliando o Jacobiano J_{III} nos pontos de equilíbrio $(0,1)$ (virtual) e $(-\check{r}/\bar{r}, -\check{c}/\bar{c})$ (real), obtém-se:

$J_{III} \downarrow$ e $x \rightarrow$	$(0,1)$	$(-\check{r}/\bar{r}, -\check{c}/\bar{c})$
$J_{III}(x_1, x_2)$	$\begin{bmatrix} \bar{r}+\check{r} & 0 \\ 0 & -\check{c} \end{bmatrix}$	$\begin{bmatrix} (\bar{c}\check{r}-\check{c}\bar{r})/\bar{c} & -\check{r} \\ \bar{c}+\check{c} & (\bar{c}\check{r}-\check{c}\bar{r})/\bar{r} \end{bmatrix}$
Tipo	Virtual	Real

Os pontos de equilíbrio do subsistema IV são $(0,0)$ e $(-\check{r}/\bar{r}, -\check{c}/\bar{c})$. O Jacobiano $J_{IV}(x_1, x_2)$ é dado por

$$J_{IV}(x_1, x_2) = \begin{bmatrix} (\bar{r}x_2 + \check{r}) & \bar{r}x_1 \\ \bar{c}x_2 & \bar{c}x_1 + \check{c} \end{bmatrix} \tag{3.25}$$

Avaliando o Jacobiano J_{IV} nos pontos de equilíbrio $(0,0)$ (virtual) e $(-\check{r}/\bar{r}, -\check{c}/\bar{c})$ (real), obtém-se:

$J_{IV} \downarrow$ e $x \rightarrow$	$(0,0)$	$(-\check{r}/\bar{r}, -\check{c}/\bar{c})$
$J_{IV}(x_1, x_2)$	$\begin{bmatrix} \check{r} & 0 \\ 0 & \check{c} \end{bmatrix}$	$\begin{bmatrix} (-\bar{r}\check{c}+\check{r}\bar{c})/\bar{c} & -\check{r} \\ -\check{c} & (-\bar{c}\check{r}+\check{c}\bar{r})/\bar{r} \end{bmatrix}$
Tipo	Virtual	Real

A partir das tabelas acima, fica evidente que, para os quatro subsistemas, o equilíbrio virtual pode ser uma sela ou um nó (estável ou instável), de acordo com os sinais dos termos na diagonal principal dos respectivos Jacobianos. Note que esta conclusão está garantida pelo teorema de Hartman–Grobman, (HOFBAUER e SIGMUND, 1998), pois percebe-se que todos os equilíbrios virtuais são hiperbólicos, desde que nenhum dos termos $\check{r}, \check{c}, \bar{r}+\check{r}, \bar{c}+\check{c}$ seja igual a zero.

Ao mesmo tempo, para os quatro subsistemas, observa-se que o equilíbrio de Nash é sempre um equilíbrio real, cujo tipo de estabilidade pode ser determinado pelos autovalores do Jacobiano em torno deste ponto, novamente pelo teorema de Hartman-Grobman, desde que nenhum autovalor tenha parte real nula. Será visto adiante, entretanto, que o termo $\bar{r}\check{c} - \check{r}\bar{c}$ (que aparece no numerador de todos os elementos diagonais dos Jacobianos dos equilíbrios reais) se anula nos casos de maior interesse (EN mistos). Com isso, a parte real dos autovalores destes Jacobianos fica nula, e o teorema de Hartman-Grobman não pode ser utilizado para inferir algo sobre a natureza (local) do ponto de equilíbrio real.

Para tornar concreta a análise utilizando estes conceitos, bem como a prova de estabilidade utilizando uma função de Liapunov[5], considere o jogo casar moedas

[5]Ao invés de usar linearização dos subsistemas da dinâmica WPL feita acima, uma prova da estabilidade de WPL utilizando uma função de Liapunov num caso mais geral é feita na subseção 3.5.2 a seguir e ainda elencadas as condições de contorno que garantem a convergência de WPL, não apresentadas em (ABDALLAH e LESSER, 2008).

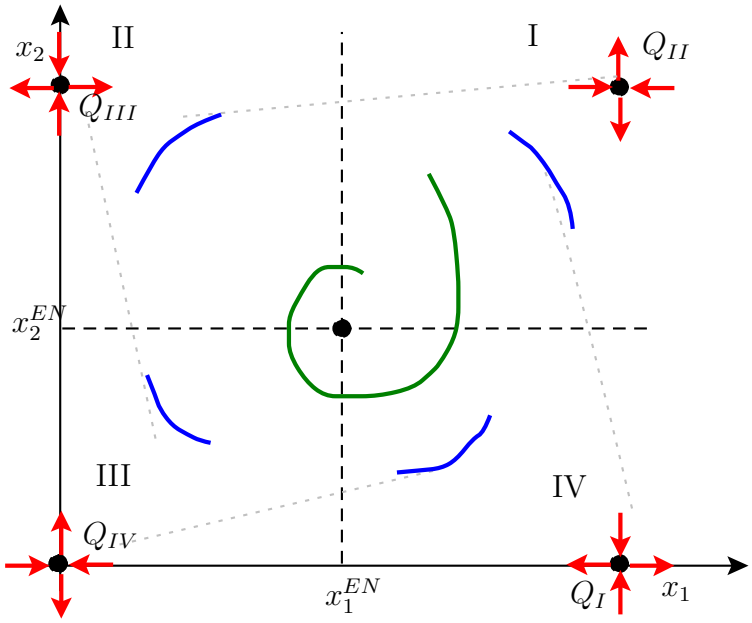


Figura 3.3: Diagrama de fase da dinâmica chaveada WPL, em que se mostra os pontos de equilíbrio de cada subsistema. Cada subsistema possui dois pontos de equilíbrio: um no ponto de interseção das retas pontilhadas que é o EN (x_1^{EN}, x_2^{EN}) e o outro no ponto denotado Q_i, subescrito i em letra romana. Os trechos de trajetórias em azul estão desenhados de acordo com a dinâmica da sela virtual que rege suas dinâmicas (indicada pela linha pontilhada). A trajetória sólida (verde) no meio da figura mostra uma concatenação possível destes trechos, motivando a conjectura da convergência assintótica de todas as trajetórias iniciadas dentro do quadrado unitário $[0,1] \times [0,1]$.

(veja tabela 3.1), para o qual define-se $R = [1, -1; -1, 1]$, $C = -R$, de modo que tem-se $\bar{r} = 4, \check{r} = -2, \bar{c} = -4, \check{c} = 2$, o que implica que o EN está localizado em $(0,5, 0,5)$.

Para este jogo, ao substituir os valores $\bar{r} = 4, \check{r} = -2, \bar{c} = -4, \check{c} = 2$ nas expressões calculadas acima para os Jacobianos, chega-se à conclusão que os quatro equilíbrios virtuais dos quatro subsistemas são selas, com as seguintes matrizes:

$$J_I = J_{III} = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix} \qquad J_{II} = J_{IV} = \begin{bmatrix} -2 & 0 \\ 0 & 2 \end{bmatrix} \tag{3.26}$$

Para os quatro equilíbrios reais que coincidem no EN para os quatro subsistemas, os quatro Jacobianos são iguais:

$$J_I = J_{II} = J_{III} = J_{IV} = \begin{bmatrix} 0 & 2 \\ -2 & 0 \end{bmatrix} \tag{3.27}$$

Esta configuração dos pontos de equilíbrio virtual e real, está mostrada na figura 3.3 que utiliza os autovalores dos Jacobianos calculados para indicar as variedades

estáveis e instáveis de cada sela virtual, o que permite traçar o retrato de fase aproximado do sistema enquanto chaveia entre os quatro subsistemas e conjecturar a convergência de todas as trajetórias válidas (isto é, iniciadas dentro do quadrado unitário no quadrante não-negativo) ao equilíbrio de Nash localizado no centro do quadrado.

Para finalizar a demonstração formal desta convergência, basta utilizar uma função de Liapunov. A forma geral dela é:

$$\begin{aligned} V_{\text{WLP}} &= 0,5(x_1 - x_1^{EN})^2 + 0,5(x_2 - x_2^{EN})^2 \\ &= 0,5(x_1 + \tfrac{\check{c}}{\bar{c}})^2 + 0,5(x_2 + \tfrac{\check{r}}{\bar{r}})^2 \end{aligned} \tag{3.28}$$

No caso específico do jogo casar moedas, a função de Liapunov fica:

$$V_{\text{WLP}-\text{MP}} = 0,5(x_1 - 0,5)^2 + 0,5(x_2 - 0,5)^2 \tag{3.29}$$

A figura 3.4 exibe o diagrama de fase da dinâmica chaveada WPL, para o jogo casar moedas.

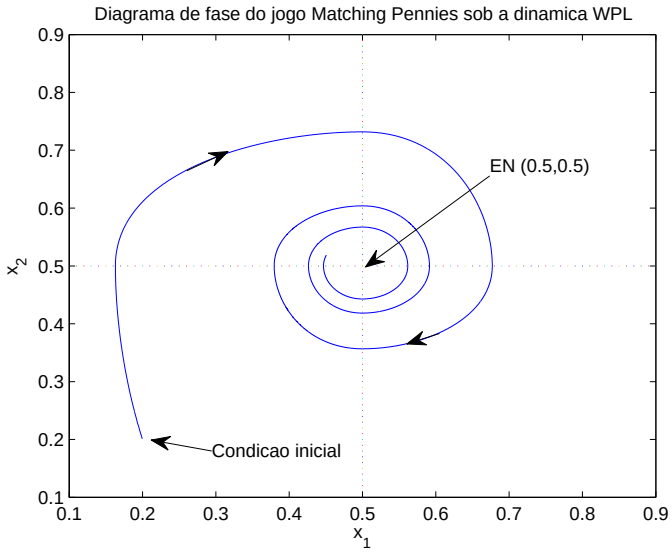


Figura 3.4: Diagrama de fase da dinâmica chaveada WPL, para o jogo casar moedas, em que se mostra a convergência oscilatória e levemente amortecida ao EN localizado no ponto $(0,5, 0,5)$, e, para clareza, mostra-se uma única trajetória iniciando-se em $(0,2, 0,2)$ (trajetórias a partir de outras condições iniciais são parecidas).

As figuras 3.5 e 3.6 exibem os gráficos da evolução da função de Liapunov e da sua derivada ao longo de uma trajetória do sistema.

Para a função de Liapunov da equação (3.29), é possível demonstrar, quadrante a quadrante, que a derivada dela, ao longo das trajetórias do sistema, fica negativa,

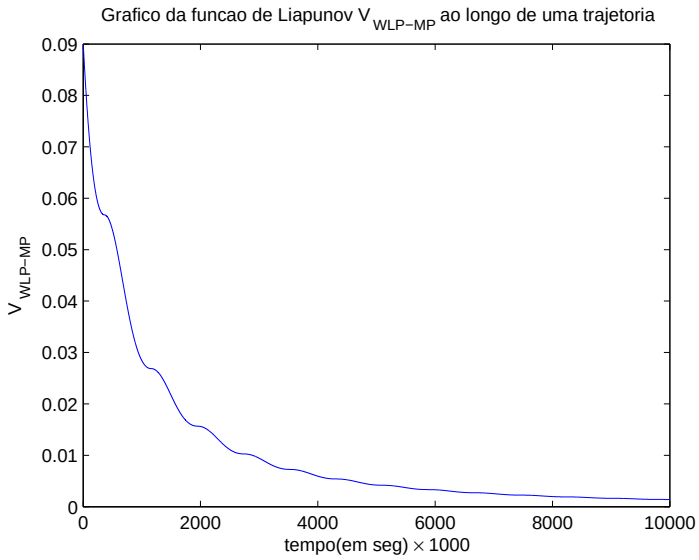


Figura 3.5: Para o jogo casar moedas, gráfico em que se mostra o decrescimento da função de Liapunov V_{WPL-MP}, para clareza, ao longo de uma única trajetória iniciando-se em $(0, 2, 0, 2)$ (gráficos a partir de outras condições iniciais são parecidos).

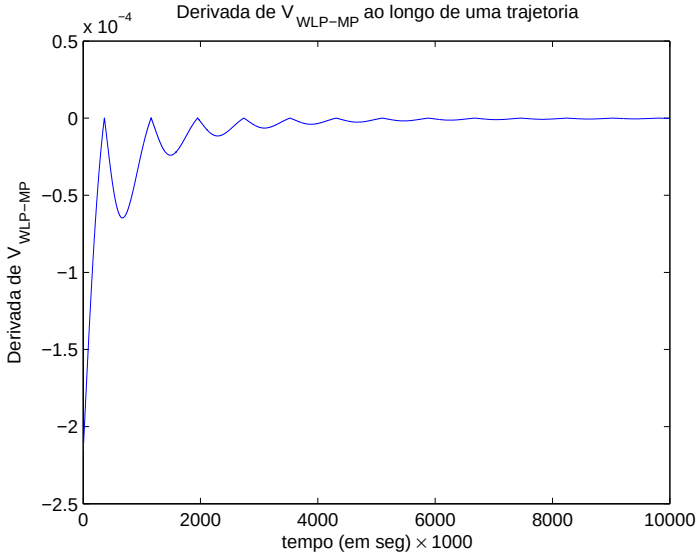


Figura 3.6: Para o jogo casar moedas, gráfico em que se mostra a negatividade da derivada da função de Liapunov $\dot{V}_{WPL-MP}$, para clareza, ao longo de uma única trajetória iniciando-se em $(0, 2, 0, 2)$ (gráficos a partir de outras condições iniciais são parecidos).

demonstrando sua estabilidade. Especificamente, calcula-se:

$$\dot{V}_{\text{WLP}-\text{MP}} = (x_1 - 0,5)(4x_2 - 2)s_1 + (x_2 - 0,5)(-4x_1 + 2)s_2 \tag{3.30}$$

Em seguida, avaliam-se as quatro possibilidades para os termos chaveados s_1, s_2, de acordo com o quadrante no qual o ponto da trajetória se encontra num jogo

genérico, com $\bar{r}\bar{c} < 0$. São cálculos exaustivos e estão listados na subseção 3.5.2, a seguir, para o caso de $\bar{r} < 0$ e $\bar{c} > 0$. Por meio deles, é possível mostrar a negatividade de $\dot{V}_{\mathrm{WLP-MP}}$.

3.5.2 Demonstração de estabilidade da dinâmica WPL usando Liapunov e definição das condições de contorno

De acordo com a equação (3.19) a dinâmica WPL pode ser escrita, destacando as hipóteses condicionais, da seguinte maneira:

$$\begin{aligned}
\dot{x}_1 &= (\bar{r}x_2 + \check{r})s_1, \text{sendo } s_1 = \begin{cases} (1-x_1) & \text{se } \bar{r}x_2 + \check{r} > 0 \quad 1^a \text{ hipótese} \\ x_1 & \text{senão} \quad\quad\quad\quad\quad\quad 3^a \text{ hipótese} \end{cases} \\
\dot{x}_2 &= (\bar{c}x_1 + \check{c})s_2, \text{sendo } s_2 = \begin{cases} (1-x_2) & \text{se } \bar{c}x_1 + \check{c} > 0 \quad 2^a \text{ hipótese} \\ x_2 & \text{senão} \quad\quad\quad\quad\quad\quad 4^a \text{ hipótese} \end{cases}
\end{aligned} \tag{3.31}$$

considere a função de Liapunov com controle V_{WLP} a seguir:

$$\begin{aligned}
V_{\mathrm{WLP}} &= 0,5(x_1 - x_1^{EN})^2 + 0,5(x_2 - x_2^{EN})^2 \\
&= 0,5(x_1 + \tfrac{\check{c}}{\bar{c}})^2 + 0,5(x_2 + \tfrac{\check{r}}{\bar{r}})^2
\end{aligned} \tag{3.32}$$

Derivando a FLC V_{WLP} ao longo das trajetórias de (3.31), tem-se:

$$\dot{V}_{\mathrm{WLP}} = (x_1 + \tfrac{\check{c}}{\bar{c}})\dot{x}_1 + (x_2 + \tfrac{\check{r}}{\bar{r}})\dot{x}_2 \tag{3.33}$$

$$\dot{V}_{\mathrm{WLP}} = (x_1 + \tfrac{\check{c}}{\bar{c}})(\bar{r}x_2 + \check{r})s_1 + (x_2 + \tfrac{\check{r}}{\bar{r}})(\bar{c}x_1 + \check{c})s_2 \tag{3.34}$$

$$\dot{V}_{\mathrm{WLP}} = (x_1 + \tfrac{\check{c}}{\bar{c}})(x_2 + \tfrac{\check{r}}{\bar{r}})\bar{r}s_1 + (x_2 + \tfrac{\check{r}}{\bar{r}})(x_1 + \tfrac{\check{c}}{\bar{c}})\bar{c}s_2 \tag{3.35}$$

$$\dot{V}_{\mathrm{WLP}} = (x_1 + \tfrac{\check{c}}{\bar{c}})(x_2 + \tfrac{\check{r}}{\bar{r}})\underbrace{(s_2\bar{c} + s_1\bar{r})}_{T_3} \tag{3.36}$$

As hipóteses da equação (3.31) dividem o espaço do quadrado unitário $[0,1] \times [0,1]$ em quatro quadrantes numerados em ordem crescente no sentido anti-horário de 1 a 4, com centro em (x_1^{EN}, x_2^{EN}) (veja figura 3.3). Deseja-se verificar se $\dot{V}_{\mathrm{WLP}}$ de (3.36) é negativa (ou semi-negativa definida), o que garantiria a convergência ao ponto (x_1^{EN}, x_2^{EN}). As hipóteses da equação (3.31) fornecem os sinais das duas primeiras parcelas de (3.36). Resta analisar o sinal da terceira e última parcela $(s_1\bar{c} + s_2\bar{r})$, denominada T_3. Sem perda de generalidade, para os quatro casos analisados a seguir, utiliza-se a condição de (BOWLING e VELOSO, 2002) e de (ABDALLAH

e LESSER, 2008), que é $\bar{r} \cdot \bar{c} < 0$, hipótese de subclasse 3. Em particular usa-se a hipótese de subclasse 3 com $\bar{c} > 0$, $\bar{r} < 0$, $\bar{c} = -e\bar{r}$, $\check{c} = -f\check{r}$, onde $e > 0$, $f > 0$. Para facilidade, chama-se nessa seção essa hipótese de subclasse 3 com $\bar{c} > 0$, $\bar{r} < 0$ de hipótese padrão. A combinação das hipóteses duas a duas dão origem a quatro casos, que estão analisados a seguir.

1^o caso: hipótese padrão e 1^a e 2^a hipóteses válidas

Da 1^a hipótese, tem-se:

$$\bar{r}x_2 + \check{r} > 0 \tag{3.37}$$

E usando da hipótese padrão que $\bar{r} < 0$, tem-se

$$x_2 < -\frac{\check{r}}{\bar{r}} = x_2^{EN} \tag{3.38}$$

Observando a figura 3.3 e da equação (3.38), tem-se x_2 no 3^o ou 4^o quadrantes. Da 2^a hipótese, tem-se:

$$\bar{c}x_1 + \check{c} > 0 \tag{3.39}$$

E

$$x_1 > -\frac{\check{c}}{\bar{c}} = x_1^{EN} \tag{3.40}$$

Observando a figura 3.3 e da equação (3.40), tem-se x_1 no 1^o ou 4^o quadrantes. Assim, com a 1^a e 2^a hipóteses elencadas, a região definida é do 4^o quadrante e deve-se ter $T_3 \geq 0$.

$$T_3 = ((1 - x_1)\bar{r} + (1 - x_2)\bar{c}) = \bar{r}\,(1 - x_1 - e(1 - x_2)) \tag{3.41}$$

Seja a reta oriunda de (3.41) a seguir:

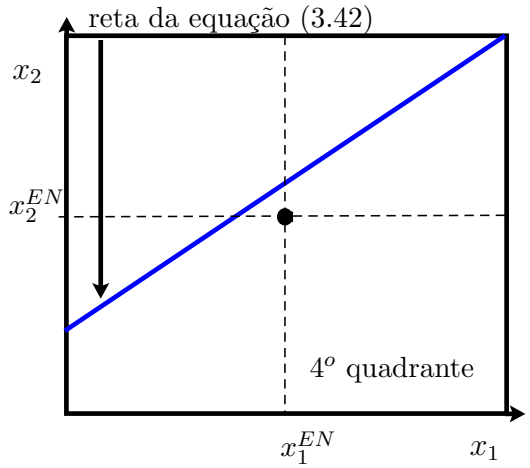
$$(x_1 - 1) = e(x_2 - 1) \tag{3.42}$$

A equação (3.42) é de uma reta que passa pelo ponto $(1, 1)$ com inclinação $1/e$. Para $1/e \leq \frac{x_2^{EN}}{x_1^{EN}}$, todo o 4^o quadrante está localizado abaixo da reta citada, conforme mostra a figura 13(a) e portanto $T_3 \geq 0$. A inequação $1/e \leq \frac{x_2^{EN}}{x_1^{EN}}$ é equivalente a:

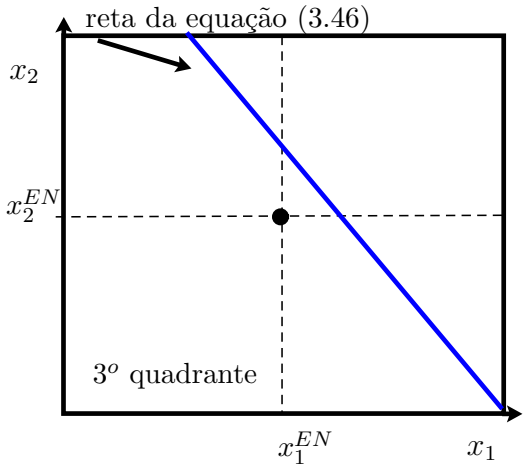
$$\frac{1}{e} \leq \frac{x_2^{EN}}{x_1^{EN}} = \frac{\check{r}}{\bar{r}}\frac{\bar{c}}{\check{c}} = \frac{e}{f} \tag{3.43}$$

e

$$f \leq e^2 \tag{3.44}$$



(a) Todos os pontos do 4^o quadrante estão abaixo da reta da equação (3.42), o que garante $T_3 \geq 0$.



(b) Todos os pontos do 3^o quadrante estão abaixo da reta da equação (3.46), o que garante $T_3 \leq 0$.

Figura 3.7: Ilustração do sinal do termo T_3 para os 3^o e 4^o quadrantes.

2^o caso: hipótese padrão e 1^a e 4^a hipóteses válidas

Da 1^a hipótese, tem-se x_2 no 3^o ou 4^o quadrantes. Da 4^a hipótese, tem-se x_1 no 2^o ou 3^o quadrantes. Assim, com a 1^a e 4^a hipóteses elencadas, a região definida é do 3^o quadrante e deve-se ter $T_3 \leq 0$.

$$T_3 = ((1 - x_1)\bar{r} + x_2\bar{c}) = \bar{r}\,(1 - x_1 - ex_2) \tag{3.45}$$

Seja a reta oriunda de (3.45) a seguir:

$$x_2 = -\frac{1}{e}(x_1 - 1) \tag{3.46}$$

A equação (3.46) é de uma reta que passa pelo ponto $(1, 0)$ com inclinação $-1/e$. Para $1/e \geq \frac{x_2^{EN}}{1-x_1^{EN}}$:

$$\frac{1}{e} \geq \frac{x_2^{EN}}{1 - x_1^{EN}} = \frac{-\frac{\check{r}}{\bar{r}}}{1 + \frac{\check{c}}{\bar{c}}} = -\frac{\bar{c}}{\bar{r}}\frac{\check{r}}{\bar{c} + \check{c}} \tag{3.47}$$

Substituindo $\bar{c} = -e\bar{r}$ em (3.47)

$$\frac{1}{e} \geq \frac{e\check{r}}{\bar{c} + \check{c}} \tag{3.48}$$

Substituindo $\check{r} = -\frac{\check{c}}{f}$ em (3.48)

$$\frac{f}{e^2} \geq -\frac{\check{c}}{(\bar{c} + \check{c})} \tag{3.49}$$

Da equação subclasse 3 (3.8) na equação (3.49), tem-se:

$$\frac{f}{e^2} \geq 0 \tag{3.50}$$

Para a condição de (3.49), todo o 3^o quadrante está localizado abaixo da reta citada, conforme mostra a figura 13(b) e portanto $T_3 \leq 0$.

3^o caso: hipótese padrão e 3^a e 2^a hipóteses válidas

Da 3^a hipótese, tem-se x_2 no 1^o ou 2^o quadrantes. Da 2^a hipótese, tem-se x_1 no 1^o ou 4^o quadrantes. Assim, com a 3^a e 2^a hipóteses elencadas, a região definida é do 1^o quadrante e deve-se ter $T_3 \leq 0$.

$$T_3 = (x_1\bar{r} + (1 - x_2)\bar{c}) = \bar{r}\,(x_1 - e(1 - x_2)) \tag{3.51}$$

Seja a reta oriunda de (3.51) a seguir:

$$-\frac{1}{e}x_1 = x_2 - 1 \tag{3.52}$$

A equação (3.52) é de uma reta que passa pelo ponto $(0, 1)$ com inclinação $-1/e$.

Para $1/e \geq \frac{1-x_2^{EN}}{x_1^{EN}}$:

$$\frac{1}{e} \geq \frac{1 - x_2^{EN}}{x_1^{EN}} = \frac{1 + \frac{\check{r}}{\bar{r}}}{-\frac{\check{c}}{\bar{c}}} = -\frac{\bar{c}}{\bar{r}}\frac{(\bar{r} + \check{r})}{\check{c}} \tag{3.53}$$

Substituindo $\bar{c} = -e\bar{r}$ em (3.53):

$$\frac{1}{e} \geq e\frac{\bar{r} + \check{r}}{\check{c}} \tag{3.54}$$

Substituindo $\check{c} = -f\check{r}$ em (3.54):

$$\frac{f}{e^2} \geq -\frac{(\bar{r} + \check{r})}{\check{r}} \tag{3.55}$$

Da equação subclasse 3 (3.8) na equação (3.55), tem-se:

$$\frac{f}{e^2} \geq 0 \tag{3.56}$$

Para a condição de (3.55), todo o 1^o quadrante está localizado acima da reta citada e portanto $T_3 \leq 0$.

4^o caso: hipótese padrão e 3^a e 4^a hipóteses válidas

Da 3^a hipótese, tem-se x_2 no 1^o ou 2^o quadrantes. Da 4^a hipótese, tem-se x_1 no 2^o ou 3^o quadrantes. Assim, com a 3^a e 4^a hipóteses elencadas, a região definida é do

2^o quadrante e deve-se ter $T_3 \geq 0$.

$$T_3 = x_1\bar{r} + x_2\bar{c} = \bar{r}\,(x_1 - ex_2) \tag{3.57}$$

Seja a reta oriunda de (3.59) a seguir:

$$x_2 = \frac{1}{e}x_1 \tag{3.58}$$

A equação (3.58) é de uma reta que passa pelo ponto $(0, 0)$ com inclinação $1/e$.

Para $1/e \leq \frac{x_2^{EN}}{x_1^{EN}}$:

$$\frac{1}{e} \leq \frac{x_2^{EN}}{x_1^{EN}} = \frac{-\frac{\check{r}}{\bar{r}}}{-\frac{\check{c}}{\bar{c}}} = \frac{\bar{c}}{\bar{r}}\frac{\check{r}}{\check{c}} = \frac{e}{f} \tag{3.59}$$

e tem-se

$$\frac{e^2}{f} \geq 1 \tag{3.60}$$

De posse das inequações (3.44), (3.50), (3.56) e (3.60), as condições de contorno para que WPL convirja estão resumidas nas inequações (3.61):

$$\begin{cases} \text{Se} \quad \bar{c} > 0 \quad \text{e} \quad \bar{r} < 0, \quad \text{então} \quad 0 < \frac{x_1^{EN}}{x_2^{EN}}\frac{1}{e} \leq 1 \\ \text{Se} \quad \bar{c} < 0 \quad \text{e} \quad \bar{r} > 0, \quad \text{então} \quad \frac{x_1^{EN}}{x_2^{EN}}\frac{1}{e} \geq 1 \end{cases} \tag{3.61}$$

onde $\check{c} = -f\check{r}$, $\bar{c} = -e\bar{r}$, $e > 0$, $f > 0$.

As inequações (3.61) que são as condições de contorno para que WPL convirja são de fato as condições do jogo de subclasse 3. Este fato, que é relevante, não é mencionado por (ABDALLAH e LESSER, 2008).

As inequações (3.62) a seguir são equivalentes:

$$\begin{cases} 0 < \frac{x_1^{EN}}{x_2^{EN}}\frac{1}{e} \leq 1 \\ 0 < \frac{f}{e^2} \leq 1 \\ 0 < -\frac{\check{c}}{\check{r}}\left(\frac{\bar{r}}{\bar{c}}\right)^2 \leq 1 \end{cases} \tag{3.62}$$

3.6 Conclusões

A abordagem de controle, especificamente, Funções de Liapunov com Controle (FLC), levou a uma unificação da teoria de algoritmos do tipo ARM-GA, permitindo análise de algoritmos desta classe, bem como projeto de algoritmos novos. Neste capítulo, limitou-se ao estudo de algoritmos que tivessem algum tipo de conhecimento do jogo subjacente, ou por conhecer o equilíbrio de Nash ou por conhecer as matrizes de recompensa. Uma exceção é o algoritmo WPL, que funciona mesmo sem conhecimento do EN, porém ainda exige (implicitamente) observação da ação

do outro agente. Para este algoritmo, a contribuição deste capítulo consistiu em colocá-lo no mesmo contexto geral, ainda utilizando uma função de Liapunov e o conceito de equilíbrios virtuais para fornecer uma prova de estabilidade, resolvendo uma questão aberta na literatura sobre algoritmos ARM-GA. No próximo capítulo, será investigada uma classe de algoritmos para a mesma classe de jogos considerada neste capítulo. Embora a abordagem adotada seja diferente, será mostrado que, novamente, a abordagem FLC permite o projeto de um algoritmo que supera os existentes em alguns aspectos e permite o relaxamento de algumas das hipóteses restritivas.

Capítulo 4

Método de Aprendizado por Reforço com Estimação Preliminar (MAR-EP) no contexto de jogo de dois agentes e duas ações

4.1 O Método de Aprendizado por Reforço com Estimação Preliminar (MAR-EP)

No capítulo 2 para um jogo de dois jogadores, duas ações, a partir da dinâmica de introduzida por (SINGH *et al.*, 2000), equação (2.7), desenvolveu-se a estratégia HEGS na qual é necessária a informação do setor onde se encontra o jogador oponente (de forma indireta, uma informação do estado do jogador oponente). No capítulo 3, por sua vez, a partir da dinâmica de WPL introduzida por (ABDALLAH e LESSER, 2008), equação (3.18), não é necessário o conhecimento da ação de outro jogador. No capítulo 4 buscou-se restringir ainda mais a informação de cada jogador. Cada jogador tem acesso a uma estimativa da ação do jogador oponente. E ainda, empregou-se uma outra dinâmica mais geral, dependente de um parâmetro τ chamado de temperatura, oriunda da DR, denominada aqui de Aprendizado por Reforço usando Probabilidades de Boltzmann (*Boltzmann Probabilities Learning*) - BPL, que é uma dinâmica híbrida. Para τ tendendo ao infinito a dinâmica BPL se comporta exatamente como IGA, HEGS, etc. Para τ tendendo a zero, a dinâmica BPL tem equilíbrio no centro, o que significa probabilidades das ações iguais a $0,5$. Para τ com valores intermediários tem-s uma solução que é híbrida da duas citadas. Desta forma desenvolve-se no capitulo 4 o Método de Aprendizado por Reforço com Estimação Preliminar - MAR-EP. O MAR-EP desenvolvido no capítulo 4 é uma alternativa ao problema da convergência lenta (BIANCHI, 2004; RIBEIRO, 2002) de AR. Tem como cenário o jogo de dois agentes, duas ações e usa uma heurística inspirada em FLC para acelerar a convergência das ações.

Segue uma breve descrição do Método de Aprendizado por Reforço com Estimação Preliminar. O MAR-EP possui três fases. Na primeira fase, que é curta e com poucas interações entre os agentes, a taxa de aprendizagem α é mantida constante e comum para os agentes e cada um faz uma modelagem resultante do contato com o agente oposto, estimando parâmetros (isto é, elementos da matriz de recompensa). Esta fase é importante no sentido de tornar desnecessário o conhecimento do jogo.

A segunda fase dos controladores traz como vantagem a redução do tempo de convergência. A taxa de aprendizagem α, baseada na modelagem da fase inicial, passa ser vista sob a ótica de um controle para cada agente, sendo cada uma delas projetada com auxílio de uma função de Liapunov, conduzindo à convergência das estratégias de forma mais rápida que aquela usada em TUYLS *et al.* (2006). Assume-se como característica do jogo que cada agente tem apenas a informação atrasada de uma unidade da ação do outro agente e que os agentes "jogam o mesmo jogo", ou seja obedecem intrinsecamente as fases do jogo.

Na terceira e última fase, retornam-se aos controladores de ganho constante

positivo, o que garante a convergência ao mesmo ponto que se empregasse durante todo o tempo o ganho constante positivo (JAAKKOLA *et al.*, 1994; WATKINS, 1992) [1] [2].

O MAR-EP procura por um lado a maior eficiência dos algoritmos baseados em modelagem em ambientes complexos reais (RUSSELL e NORVIG, 2003) quando usa modelagem na fase inicial. Está implícito, portanto, que o ponto de equilíbrio não se encontra próximo ao ponto inicial. Por outro lado, quando já se está suficientemente próximo do ponto de equilíbrio, usa-se *Q-Learning* para atingir esse. A dualidade do método, ao agir de duas formas diferentes, implica embora de maneira heurística, a vantagem sobre o algoritmo *Q-Learning*.

Em resumo, o MAR-EP possui uma heurística na exploração. Define-se um padrão de comportamento para o agentes, o que é na verdade um tipo de modelagem, ainda que superficial, e essa modelagem se dá na fase inicial do aprendizado, por meio de estimação de parâmetros.

Feito o resumo informal acima, descreve-se a organização do capítulo a seguir. Primeiramente, são vistas as definições usadas por TUYLS *et al.* (2006) e em seguida na seção 4.3, seguindo TUYLS *et al.* (2006), desenvolve-se a distribuição de probabilidade de Boltzmann até a obtenção da dinâmica BPL.

A dinâmica BPL é a dinâmica advinda do desenvolvimento de aprendizado por reforço *Q-Learning* (WATKINS, 1992), usando probabilidades de Boltzmann, feita por TUYLS *et al.* (2006). A partir daí, projetam-se, na seção 4.4, os controladores adequados para que as estratégias convirjam ao equilíbrio de Nash, quando se tem disponível a ação e dois parâmetros do jogador oponente. Em exemplos ilustram-se o potencial do uso dos controladores. Na seção 4.5, apresenta-se o método de estimação dos parâmetros do jogador oponente baseado em mínimos quadrados e sua aplicação nos controladores, aliado à informação atrasada da ação do jogador oponente. Um chaveamento adequado se faz necessário, já que os ganhos obtidos do controlador vão decrescendo, reduzindo a velocidade de convergência, e é apresentado na seção 4.6. Na seção 4.7 exibem-se alguns exemplos numéricos.

4.2 Definições preliminares

Nessa seção, exibe-se a convenção adotada por TUYLS *et al.* (2006) no desenvolvimento de aprendizado por reforço *Q-Learning* (WATKINS, 1992), usando Probabilidades de Boltzmann. A dinâmica advinda desse desenvolvimento é aqui denominada de Aprendizado por Reforço usando Probabilidades de Boltzmann (*Boltzmann*

[1] Detalhes do chaveamento entre a segunda e a terceira fase encontram-se na seção 4.6.

[2] JAAKKOLA *et al.* (1994) fornece a prova de convergência para o caso de α constante, resultado que é invocado na terceira fase de MAR-EP.

Probabilities Learning) - BPL. A convenção adotada por TUYLS *et al.* (2006) será denominada de convenção BPL.

Pela convenção BPL num jogo de múltiplos agentes, múltiplas ações, sejam x_1, x_2, x_3,$\cdots$, x_n ações do agente 1, y_1, y_2, y_3,$\cdots$, y_n ações do agente 2, e assim de modo sucessivo. Por facilidade de análise, o trabalho foca no jogo de duas ações, dois agentes, de forma que as matrizes de recompensa A e B dos agentes 1 e 2 são dadas por:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \tag{4.1}$$

A convenção BPL é válida para $n > 2$ agentes e $m > 2$ ações por agente. No entanto o que foi desenvolvido nesse capítulo é válido para jogos com $n = 2$ agentes e $m = 2$ ações por agente.

E pode-se dizer então que

$$x_2 = 1 - x_1 \tag{4.2}$$

$$y_2 = 1 - y_1 \tag{4.3}$$

4.3 Desenvolvendo a distribuição de Boltzmann até chegar em BPL

Nessa seção, parte-se de um modelo de Q-Learning temporal, onde os valores Q são interpretados como probabilidades de Boltzmann para seleção da ação, até atingir a dinâmica BPL, um tipo de dinâmica do replicador (DR), nos moldes de (TUYLS *et al.*, 2006).

A equação do replicador (ver HOFBAUER e SIGMUND, 1998, página 87) é o principal modelo determinista para descrição da evolução temporal das frequências das estratégias de uma população. Há uma ligação estreita entre a equação replicadora e o equilíbrio de Nash: todos os pontos interiores da solução da equação replicadora são pontos de equilíbrio de Nash.

A equação do replicador forma a DR, formalizada como um sistema de equações diferenciais como na equação (4.4):

$$\frac{dx_i}{dt} = [(A\mathbf{x})_i - \mathbf{x}.A\mathbf{x}]x_i \tag{4.4}$$

onde x_i representa a probabilidade da estratégia i, A, a matriz de recompensa que

descreve os diferentes valores de recompensa e $\mathbf{x}.A\mathbf{x}$ é $\mathbf{x}^T A\mathbf{x}$.

Formalmente, a distribuição de Boltzmann é dada por:

$$x_i(k) = \frac{e^{\tau Q_{a_i}(k)}}{\sum_{j=1}^{n} e^{\tau Q_{a_j}(k)}} \tag{4.5}$$

sendo $x_i(k)$ a probabilidade de jogar a estratégia i no instante de tempo k, $\tau > 0$ o parâmetro chamado temperatura e n o número de ações que o agente pode tomar.[3]

O parâmetro τ elevado torna as ações equiprováveis, incentivando a estratégia denominada explorar (*exploration*). Para valores de τ pequenos, têm-se grandes diferenças de probabilidade, incentivando a exploração intensiva (*exploitation*), se concentrando no ganho imediato. Tal estratégia é chamada de gulosa ou sugar. A heurística normalmente adotada é um balanceamento (balanço) entre sugar e explorar, quando se tem um ambiente desconhecido.

Calculando a derivada temporal de $x_i(t)$,

$$\frac{dx_i(t)}{dt} = \frac{d}{dt}\frac{e^{\tau Q_i(t)}}{\sum_j e^{\tau Q_j(t)}} \tag{4.6}$$

Fazendo $u = e^{\tau Q_i(t)}$, $v = \sum_j e^{\tau Q_j(t)}$, tem-se $u' = e^{\tau Q_i(t)} \tau \frac{dQ_i(t)}{dt}$, derivada de u temporal e $v' = \sum_j e^{\tau Q_j(t)} \cdot e^{\tau Q_i(t)} \cdot \tau$, derivada de v temporal.

Assim,

$$\frac{dx_i(t)}{dt} = \frac{u'}{v} - \frac{u \cdot v'}{v^2} \tag{4.7}$$

$$\frac{dx_i(t)}{dt} = \frac{u'}{v} - \frac{u \cdot v'}{v^2} \tag{4.8}$$

$$\frac{u'}{v} = \tau x_i(t) \frac{Q_i(t)}{dt} \tag{4.9}$$

$$\frac{uv'}{v^2} = \frac{\sum_j \tau \frac{Q_i(t)}{dt} e^{\tau Q_j(t)} e^{\tau Q_i(t)}}{\sum_j e^{\tau Q_j(t)} \cdot \sum_j e^{\tau Q_j(t)}} \tag{4.10}$$

$$\frac{uv'}{v^2} = \tau x_i(t) \sum_j x_j(t) \frac{dQ_j(t)}{dt} \tag{4.11}$$

$$\frac{dx_i(t)}{dt} = \tau x_i(t) \left[\frac{Q_i(t)}{dt} - \sum_j x_j(t) \frac{dQ_j(t)}{dt} \right], \tag{4.12}$$

Pode-se escrever $x_i(k+1)/x_i(k)$ da equação (4.5) como:

[3]Para tornar as expressões menos densas, os limites do somatórios a seguir, de 1 a n, serão omitidos.

$$\frac{x_i(k+1)}{x_i(k)} = \frac{e^{\tau Q_{a_i}(k+1)} \sum_j e^{\tau Q_{a_j}(k)}}{e^{\tau Q_{a_i}(k)} \sum_j e^{\tau Q_{a_j}(k+1)}} \tag{4.13}$$

$$\frac{x_i(k+1)}{x_i(k)} = \frac{e^{\tau Q_{a_i}(k+1)} e^{-\tau Q_{a_i}(k)} \sum_j e^{\tau Q_{a_j}(k)} \sum_j}{e^{\tau Q_{a_i}(k)} \sum_j e^{\tau Q_{a_j}(k+1)}} \tag{4.14}$$

$$\frac{x_i(k+1)}{x_i(k)} = \frac{e^{\tau \Delta Q_{a_i}(k)}}{\sum_j x_j e^{\tau \Delta Q_{a_j}(k)}} \tag{4.15}$$

onde $\Delta Q_{a_i}(k) = Q_{a_i}(k+1) - Q_{a_i}(k)$.

Pode-se escrever a equação (4.15) então:

$$x_i(k+1) = \frac{x_i(k) e^{\tau \Delta Q_{a_i}(k)}}{\sum_j x_j e^{\tau \Delta Q_{a_j}(k)}} \tag{4.16}$$

E, ainda, subtraindo $x_i(k)$ de cada lado da equação (4.16),

$$x_i(k+1) - x_i(k) = \frac{x_i(k) e^{\tau \Delta Q_{a_i}(k)}}{\sum_j x_j(k) e^{\tau \Delta Q_{a_j}(k)}} - x_i(k) \tag{4.17}$$

$$x_i(k+1) - x_i(k) = x_i(k) \left(\frac{e^{\tau \Delta Q_{a_i}(k)} - \sum_j x_j(k) e^{\tau \Delta Q_{a_j}(k)}}{\sum_j x_j(k) e^{\tau \Delta Q_{a_j}(k)}} \right) \tag{4.18}$$

Para a versão contínua no tempo, supõe-se que o tempo entre duas jogadas é dado por δ, $0 \leq \delta \leq 1$ e $x_i(k\delta)$ descreve os valores de x_i no tempo $k\delta = t$,

$$\frac{x_i(k\delta + \delta) - x_i(k\delta)}{\delta} = \frac{x_i(k\delta)}{\delta \sum_j x_j(k\delta) e^{\tau \Delta Q_{a_j}(k\delta)}} \cdot \left(e^{\tau \Delta Q_{a_j}(k\delta)} - \sum_j x_j(k\delta) e^{\tau \Delta Q_{a_j}(k\delta)} \right) \tag{4.19}$$

Fazendo δ tender a zero,

$$\lim_{\delta \to 0} \frac{\Delta x_i(k\delta)}{\delta} = \lim_{\delta \to 0} \frac{x_i(k\delta)}{\delta \sum_j x_j(k\delta) e^{\tau \Delta Q_{a_j}(k\delta)}} \cdot \left(e^{\tau \Delta Q_{a_i}(k\delta)} - \sum_j x_j(k\delta) e^{\tau \Delta Q_{a_j}(k\delta)} \right) \tag{4.20}$$

$$\lim_{\delta \to 0} \frac{\Delta x_i(k\delta)}{\delta} = \lim_{\delta \to 0} \frac{x_i(k\delta)}{\sum_j x_j(k\delta) e^{\tau \Delta Q_{a_j}(k\delta)}} \cdot \lim_{\delta \to 0} \left(\frac{e^{\tau \Delta Q_{a_i}(k\delta)}}{\delta} - \frac{\sum_j x_j(k\delta) e^{\tau \Delta Q_{a_j}(k\delta)}}{\delta} \right) \tag{4.21}$$

A primeira parcela da equação (4.38), a seguir, é x_i, que $\Delta Q_{a_j}(k\delta)$ se torna zero

$$\lim_{\delta \to 0} \frac{x_i(k\delta)}{\delta \sum_j x_j(k\delta) e^{\tau \Delta Q_{a_j}(k\delta)}} = x_i, \tag{4.22}$$

pois $\Delta Q_{a_j}(k\delta)$ se torna zero e $\sum_j x_j(k\delta) = 1$. Assim,

$$\lim_{\delta \to 0} \frac{\Delta x_i(k\delta)}{\delta} = x_i \cdot T_2 \ , \tag{4.23}$$

onde T_2 é igual a:

$$T_2 = \lim_{\delta \to 0} \left(\frac{e^{\tau \Delta Q_{a_i}(k\delta)}}{\delta} - \frac{\sum_j x_j e^{\tau \Delta Q_{a_j}(k\delta)}}{\delta} \right) \tag{4.24}$$

O limite é indefinido, e usa-se, então, a regra de l'hôpital.

$$T_2 = \lim_{\delta \to 0} \frac{\tau \Delta Q_{a_j}(k\delta) e^{\tau \Delta Q_{a_j}(k\delta)}}{\delta} - \sum_j x_j(k\delta) \cdot \lim_{\delta \to 0} \left(\tau \Delta Q_{a_j}(k\delta) \frac{e^{\tau \Delta Q_{a_j}(k\delta)}}{\delta} \right) \tag{4.25}$$

$$T_2 = \tau \frac{dQ_{a_i}}{dt} - \sum_j x_j \tau \frac{dQ_{a_j}}{dt} \tag{4.26}$$

O limite de (4.20) será então:

$$\frac{\dot{x}_i(t)}{x_i(t)} = \tau \left(\frac{dQ_{a_i}}{dt} - \sum_j \frac{dQ_{a_j}}{dt} x_j \right) \tag{4.27}$$

A equação (4.27) é o modelo contínuo no tempo do Q-Learning. A regra de atualização para o 1^o jogador é dada pela equação (4.28),

$$Q_{a_i}(k+1) = Q_{a_i}(k) + \alpha \left(r_{ai}(k+1) + \gamma \max_{a_i} Q - Q_{a_i}(k) \right), \tag{4.28}$$

onde r_{ai} é dado por:

$$r_{ai} = \sum_j a_{ij} y_j \tag{4.29}$$

Da equação (4.27) é necessário ter então $\frac{dQ_{a_i}(t)}{dt}$. Parte-se, então, da equação de Q-Learning escrita na equação (4.27), o que conduz a:

$$\Delta Q_{a_i}(k) = \alpha \left(r_{ai}(k+1) + \gamma \max_{a_i} Q - Q_{a_i}(k) \right) \tag{4.30}$$

(4.30) é a equação das diferenças para a função Q. Ao se fazer essa equação infini-

tesimal, tem-se:

$$\Delta Q_{a_i}(k\delta) = \alpha(r_{ai}((k+1)\delta) + \gamma \max_{a_i} Q - Q_{a_i}(k\delta)) \cdot ((k+1)\delta - k\delta) \tag{4.31}$$

Reescrevendo a equação (4.31):

$$\Delta Q_{a_i}(k\delta) = \alpha \left(r_{ai}((k+1)\delta) + \gamma \max_{a_i} Q - Q_{a_i}(k\delta) \right) \delta \tag{4.32}$$

Ao se fazer $\delta \to 0$ e $k\delta \to 0$, tem-se:

$$\frac{Q_{a_i}}{dt} = \alpha \left(r_{ai} + \gamma \max_{a_i} Q - Q_{a_i} \right) \tag{4.33}$$

Substituindo (4.33) em (4.27):

$$\frac{\dot{x}_i}{x_i(t)} = \tau \left(\alpha r_{ai} - \gamma \max_{a_i} Q_{a_i} - \alpha Q_{a_i} - \sum_j x_j \alpha \left(r_{a_i} + \gamma \max_{a_i} Q_{a_i} - Q_{a_j} \right) \right) \tag{4.34}$$

e

$$\frac{\dot{x}_i}{x_i(t)} = \tau\alpha \left(r_{ai} - \sum_j x_j r_{aj} - Q_{a_i} + \sum_j Q_{a_j} x_j \right) \tag{4.35}$$

Como $\sum_j x_j = 1$, a equação (4.35) se torna:

$$\frac{\dot{x}_i}{x_i(t)} = \tau\alpha \left(r_{ai} - \sum_j x_j r_{aj} - Q_{a_i} \sum_j x_j + \sum_j Q_{a_j} x_j \right) \tag{4.36}$$

$$\frac{\dot{x}_i}{x_i(t)} = \tau\alpha \left(r_{ai} - \sum_j x_j r_{aj} + \sum_j x_j (Q_{a_j} - Q_{a_i}) \right) \tag{4.37}$$

Como $\frac{x_j}{x_i(t)}$ é igual a $\frac{e^{\tau \Delta Q_{a_j}}}{e^{\tau \Delta Q_{a_i}}}$,

$$\alpha \sum_j x_j \log \left(\frac{x_j}{x_i(t)} \right) = \alpha\tau \sum_j x_j (Q_{a_j} - Q_{a_i}). \tag{4.38}$$

Isso conduz a:

$$\dot{x}_i = x_i(t) \left[\alpha\tau \overbrace{(r_{ai} - \sum_j x_j r_{aj})}^{1^a \text{ parcela}} + \alpha \overbrace{\left[\sum_j x_j \log \left(\frac{x_j}{x_i} \right) \right]}^{2^a \text{ parcela}} \right] \tag{4.39}$$

Para $i = 1$, faz-se o cálculo da primeira parcela de (4.39):
r_{a1} é dado por:

$$r_{a1} = \sum_j a_{1j} y_j = a_{11} y_1 + a_{12} y_2 = a_{11} y_1 + a_{12}(1 - y_1) \tag{4.40}$$

r_{a2} é dado por:

$$r_{a2} = \sum_j a_{2j} y_j = a_{21} y_1 + a_{22} y_2 = a_{21} y_1 + a_{22}(1 - y_1) \tag{4.41}$$

$$\sum_j x_j r_{aj} = x_1 r_{a1} + x_2 r_{a2} = (a_{11} - a_{12}) x_1 y_1 + a_{12} x_1 + (a_{21} - a_{22}) x_2 y_1 + a_{22} x_2 \tag{4.42}$$

$$\sum_j x_j r_{aj} = x_1 r_{a1} + x_2 r_{a2} = (a_{11} - a_{12}) x_1 y_1 + a_{12} x_1 + (a_{21} - a_{22})(1 - x_1) y_1 + a_{22}(1 - x_1) \tag{4.43}$$

e

$$\sum_j x_j r_{aj} = \bar{a} x_1 y_1 + (a_{12} - a_{22}) x_1 + (a_{21} - a_{22}) y_1 - a_{22} \tag{4.44}$$

Substituindo a equação (4.44) na primeira parcela da equação (4.39):

$$r_{a1} - \sum_j x_j r_{aj} = (a_{11} - a_{12}) y_1 + a_{12} - \bar{a} x_1 y_1 - (a_{21} - a_{22}) x_1 - (a_{21} - a_{22}) y_1 - a_{22} \tag{4.45}$$

$$r_{a1} - \sum_j x_j r_{aj} = \bar{a} y_1 + (a_{22} - a_{12}) x_1 - \bar{a} x_1 y_1 + a_{12} - a_{22} \tag{4.46}$$

E, então, a primeira parcela da equação (4.39) é dada por (4.47)

$$r_{a1} - \sum_j x_j r_{aj} = (1 - x_1)(\bar{a} y_1 + a_{12} - a_{22}) \tag{4.47}$$

Para $i = 1$, faz-se o cálculo da segunda parcela da equação (4.39):

$$\sum_j x_j \log\left(\frac{x_j}{x_i}\right) = x_1 \log\left(\frac{x_1}{x_1}\right) + x_2 \log\left(\frac{x_2}{x_1}\right) = (1 - x_1) \log\left(\frac{1 - x_1}{x_1}\right) \tag{4.48}$$

Com equações (4.47) e (4.48) em (4.39), têm-se para o jogador x, da dinâmica

da ação x_1:

$$\frac{x_1}{dt} = x_1 \left[\alpha\tau(1-x_1)(\bar{a}y_1 + a_{12} - a_{22}) + \alpha(1-x_1)\log\left(\frac{1-x_1}{x_1}\right)\right] \tag{4.49}$$

De forma análoga a equação (4.50) obtém-se a equação para o jogador y, da dinâmica da ação y_1:

$$\frac{y_1}{dt} = y_1 \left[\alpha\tau(1-y_1)(\bar{b}x_1 + b_{21} - b_{22}) + \alpha(1-y_1)\log\left(\frac{1-y_1}{y_1}\right)\right] \tag{4.50}$$

onde $\bar{b} = b_{11} + b_{22} - b_{12} - b_{21}$.

A seguir, as equações (4.49) e (4.50) são expressas na notação adotada no tratamento de HEGS do capítulo 2 e em (OWEN, 1995). Sejam x_1 e x_2 as probabilidades dos dois agentes selecionarem as suas primeiras ações (OWEN, 1995). Como se trata de um jogo de duas ações, as segundas ações de cada agente são expressas por $1-x_1$ e $1-x_2$, respectivamente.

Nesse jogo de dois agentes, duas ações, têm-se as seguintes matrizes de recompensa:

$$R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \qquad C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \tag{4.51}$$

O sistema dinâmico dado pela equação (4.52) é denominado aqui de sistema de Aprendizado com Probabilidades de Boltzmann, (*Boltzmann Probabilities Learning*) - (BPL). Comparando a equação (4.52) com a equação (4.4) da DR, nota-se que o primeiro termo de ambas é igual.

$$\begin{bmatrix} \dot{x}_1 \\ \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} K_R x_1(1-x_1) \overbrace{\left[\tau(\bar{r}x_2 + \check{r}) + \log\left(\frac{1-x_1}{x_1}\right)\right]}^{f_1(x_1,x_2)} \\ \\ K_C x_2(1-x_2) \overbrace{\left[\tau(\bar{c}x_1 + \check{c}) + \log\left(\frac{1-x_2}{x_2}\right)\right]}^{f_2(x_1,x_2)} \end{bmatrix} \tag{4.52}$$

K_R e K_C são, respectivamente, os ganhos dos agentes 1 e 2 que são empregados em substituição à α nas equações (4.49) e (4.50). Reescreve-se a equação (2.37) para

facilitar a leitura do texto na equação (4.53) a seguir.

$$\begin{aligned} \bar{r} &= r_{11} + r_{22} - (r_{12} + r_{21}) \\ \check{r} &= -(r_{22} - r_{12}) \\ \bar{c} &= c_{11} + c_{22} - (c_{12} + c_{21}) \\ \check{c} &= -(c_{22} - c_{21}) \end{aligned} \tag{4.53}$$

Reescreve-se a equação (4.52) numa forma mais compacta como:

$$\begin{aligned} \dot{x}_1 &= K_R x_1(1 - x_1) f_1(x_1, x_2) \\ \dot{x}_2 &= K_C x_2(1 - x_2) f_2(x_1, x_2) \end{aligned} \tag{4.54}$$

4.3.1 Equilíbrios do sistema dinâmico BPL

Os equilíbrios do sistema dinâmico BPL no interior do quadrado unitário são soluções do seguinte sistema de equações:

$$\begin{aligned} \tau(\bar{r}x_2 + \check{r}) + \log\left(\tfrac{1-x_1}{x_1}\right) &= 0 \\ \tau(\bar{c}x_1 + \check{c}) + \log\left(\tfrac{1-x_2}{x_2}\right) &= 0 \end{aligned} \tag{4.55}$$

Examinando este sistema de equações é evidente que se o parâmetro τ tende a 0, o equilíbrio é determinado pelos segundos termos (logarítmicos) de cada equação, o que significa imediatamente que ele se localiza no ponto em que o argumento de cada logaritmo seja igual a um. Em outras palavras,

$$\frac{1 - x_1}{x_1} = \frac{1 - x_2}{x_2} = 1, \tag{4.56}$$

o que significa $x_1 = x_2 = 0,5$.

Dividindo ambos os lados das equações (4.58) por τ, analisa-se o que acontece quando $\tau \to \infty$; fica evidente que o equilíbrio é determinado pela solução das equações desacopladas (primeiros termos da equação (4.58)); isto é:

$$\begin{aligned} x_1 &= -\tfrac{\check{c}}{\bar{c}} \\ x_2 &= -\tfrac{\check{r}}{\bar{r}} \end{aligned} \tag{4.57}$$

Evidentemente, este equilíbrio é exatamente o equilíbrio dos sistemas dinâmicos da classe IGA, HEGS, etc. Para valores de τ intermediários, isto é entre 0 e ∞, o equilíbrio é determinado pela solução simultânea do seguinte sistema de equações:

$$\begin{aligned} x_1 &= -\tfrac{\check{c}}{\bar{c}} - \tfrac{1}{\tau\bar{c}} \log\left(\tfrac{1-x_2}{x_2}\right) \\ x_2 &= -\tfrac{\check{r}}{\bar{r}} - \tfrac{1}{\tau\bar{r}} \log\left(\tfrac{1-x_1}{x_1}\right) \end{aligned} \tag{4.58}$$

Este sistema pode ser escrito abreviadamente como:

$$\begin{aligned} x_1 &= h_1(x_2) \\ x_2 &= h_2(x_1) \end{aligned} \tag{4.59}$$

sendo

$$\begin{aligned} h_1(x_2) &:= -\tfrac{\check{c}}{\bar{c}} - \tfrac{1}{\tau\bar{c}} \log\left(\tfrac{1-x_2}{x_2}\right) \\ h_2(x_1) &:= -\tfrac{\check{r}}{\bar{r}} - \tfrac{1}{\tau\bar{r}} \log\left(\tfrac{1-x_1}{x_1}\right) \end{aligned} \tag{4.60}$$

Numericamente, é possível achar a solução do sistema por meio do método iterativo de ponto fixo, isto é, a partir de um ponto inicial $(x_1^{(0)}, x_2^{(0)})$:

$$\begin{aligned} x_1^{(k+1)} &= h_1(x_2^{(k)}) \\ x_2^{(k+1)} &= h_2(x_1^{(k)}) \end{aligned} \tag{4.61}$$

O ponto de equilíbrio de BPL, onde seus gradientes são zero, é chamado de $\mathbf{x}^{\text{eq}} = ({x_1}^{\text{eq}}, {x_2}^{\text{eq}})$ e pode ser calculado para a condição $x_i \neq 0$ e $x_i \neq 1$, $i \in \{1,2\}$ (ou seja $x_i \in (0,1)$, $i \in \{1,2\}$) como:

$$\begin{aligned} {x_1}^{\text{eq}} &= -\tfrac{\check{c}}{\bar{c}} + \frac{\log\left(\frac{x_2}{1-x_2}\right)}{\bar{c}\tau} \\ {x_2}^{\text{eq}} &= -\tfrac{\check{r}}{\bar{r}} + \frac{\log\left(\frac{x_1}{1-x_1}\right)}{\bar{r}\tau} \end{aligned} \tag{4.62}$$

4.4 Projeto de controladores de estimação preliminar

A solução de $f_1(x_1, x_2) = 0$ de (4.52) parametrizada em x_1 é:

$$\left(x_1, \underbrace{\frac{-\tau\check{r} + \log\left(\frac{x_1}{1-x_1}\right)}{\tau\bar{r}}}_{g_1(x_1)} \right) \tag{4.63}$$

A solução de $f_2(x_1, x_2) = 0$ de (4.52) parametrizada em x_2 é:

$$\left(\underbrace{\frac{-\tau\check{c} + \log\left(\frac{x_2}{1-x_2}\right)}{\tau\bar{c}}}_{g_2(x_2)}, x_2 \right) \tag{4.64}$$

Ressalta-se que, por construção, $0 \leq g_1(x_1) \leq 1$ e $0 \leq g_2(x_2) \leq 1$, já que $g_1(x_1)$ é o valor que x_2 deve assumir para $f_1(x_1, x_2) = 0$ e, de forma análoga, $g_2(x_2)$ é o valor que x_1 deve assumir para $f_2(x_1, x_2) = 0$.

Portanto, de (4.63) e de (4.64), destaca-se a condição $\bar{r} \cdot \bar{c} \neq 0$ para a aplicação

de MAR-EP. Essa restrição é também imposta em HEGS no teorema 2.2.2 e em BOWLING e VELOSO (2002). O trabalho em situações com $\bar{r} \cdot \bar{c} > 0$ e principalmente $\bar{r} \cdot \bar{c} = 0$ é um campo pouco explorado que merece investigações que aproveitem as características de divergência da dinâmica.

Seja a função de Liapunov escolhida a seguir:

$$V_{Liap} = 0,5 \left((x_1 - g_2(x_2))^2 + (x_2 - g_1(x_1))^2 \right) \tag{4.65}$$

É imediato com a escolha de V_{Liap}, conforme (4.65), que a dinâmica do sistema atinja o equilíbrio na interseção das curvas $f_1(x_1, x_2) = 0$, e $f_2(x_1, x_2) = 0$, $i \in \{1, 2\}$.

Calcula-se a $\dot{V}_{Liap}$ a seguir:

$$\dot{V}_{Liap} = (x_1 - g_2(x_2)) \left[\dot{x}_1 - \frac{\partial g_2(x_2)}{\partial x_2} \dot{x}_2 \right] + (x_2 - g_1(x_1)) \left[\dot{x}_2 - \frac{\partial g_1(x_1)}{\partial x_1} \dot{x}_1 \right] \tag{4.66}$$

Substituindo a equação (4.54) na equação (4.66),

$$\begin{aligned} \dot{V}_{Liap} &= (x_1 - g_2(x_2)) \left[x_1(1 - x_1) K_R f_1(x_1, x_2) - \tfrac{K_C f_2(x_1, x_2)}{\tau \bar{c}} \right] \\ &+ (x_2 - g_1(x_1)) \left[x_2(1 - x_2) K_C f_2(x_1, x_2) - \tfrac{K_R f_1(x_1, x_2)}{\tau \bar{r}} \right] \end{aligned} \tag{4.67}$$

Tem-se, então, que

$$\begin{aligned} \dot{V}_{Liap} &= K_R f_1(x_1, x_2) \left[x_1(1 - x_1)(x_1 - g_2(x_2)) - \tfrac{x_2 - g_1(x_1)}{\tau \bar{r}} \right] \\ &+ K_C f_2(x_1, x_2) \left[x_2(1 - x_2)(x_2 - g_1(x_1)) - \tfrac{x_1 - g_2(x_2)}{\tau \bar{c}} \right] \end{aligned} \tag{4.68}$$

Escolhem-se K_R e K_C, respectivamente, como nas equações (4.69) e (4.70):

$$K_R = -\frac{\kappa_R}{f_1(x_1, x_2)} \left[x_1(1 - x_1)(x_1 - g_2(x_2)) - \frac{x_2 - g_1(x_1)}{\tau \bar{r}} \right] \tag{4.69}$$

$$K_C = -\frac{\kappa_C}{f_2(x_1, x_2)} \left[x_2(1 - x_2)(x_2 - g_1(x_1)) - \frac{x_1 - g_2(x_2)}{\tau \bar{c}} \right] \tag{4.70}$$

onde $\kappa_R > 0$, $\kappa_C > 0$ são arbitradas pelos agentes. É imediato verificar que a escolha de K_R e K_C das equações (4.69) e (4.70) conduz à

$$\begin{aligned} \dot{V}_{Liap} &= -\kappa_R \left[x_1(1 - x_1)(x_1 - g_2(x_2)) - \tfrac{x_2 - g_1(x_1)}{\tau \bar{r}} \right]^2 \\ &- \kappa_C \left[x_2(1 - x_2)(x_2 - g_1(x_1)) - \tfrac{x_1 - g_2(x_2)}{\tau \bar{c}} \right]^2 < 0 \end{aligned} \tag{4.71}$$

Das equações (4.69) e (4.70), nota-se que o agente 1 precisa de $g_2(x_2)$ para obter K_R da mesma forma que o agente 2 precisa de $g_1(x_1)$ para obter K_C. Na seção 4.4.1 a seguir, registram-se os resultados com os controladores $g_2(x_2)$ e $g_1(x_1)$ teóricos, para que se tenha a noção da potencial superioridade de MAR-EP perante BPL.

Mesmo não sendo a condição de contorno adotada $\bar{r}\bar{c} \neq 0$, analisa-se a seguir, a título de comentário, o que acontece se $\bar{r} = 0$. Da primeira expressão da equação (4.52), observa-se que a dinâmica de x_1 independe de x_2. Nesse caso, portanto, o jogador 1 pode interferir no equilíbrio da ação do jogador 2 e o oposto não ocorre. De modo análogo se $\bar{c} = 0$, da segunda expressão da equação (4.52), observa-se que a dinâmica de x_2 independe de x_1. Portanto nesse caso, o jogador 2 pode interferir no equilíbrio da ação do jogador 1 e o inverso não ocorre. Fica claro a presença de um jogador dominante, líder no jogo, que soluciona a equação de sua dinâmica, escolhe o melhor ponto de equilíbrio da forma que melhor lhe convier, seja essa decisão egoísta (o melhor para si, se for "medianamente inteligente") ou cooperativa (o melhor equilíbrio para ambos, se for "superiormente inteligente"). Quanto à aceleração da convergência, pode-se dizer que, se $\bar{r} = 0$, apenas o jogador 1 pode acelerar a convergência e, se $\bar{c} = 0$, apenas o jogador 2 pode acelerar a convergência.

4.4.1 Exemplos teóricos comparando MAR-EP com BPL

Primeiro exemplo teórico com jogo casar moedas

Seja o exemplo de subclasse 3, de acordo com as matrizes a seguir.

$$R = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \quad C = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \tag{4.72}$$

Compara-se, nesse exemplo, BPL com MAR-EP com os controladores dados pelas equações (4.69) e (4.70). Foram definidos o ponto inicial em $(0,2\ ,\ 0,9)$, $\kappa_R = \kappa_C = 60$, e empregado em BPL definido como $\alpha = 0,1$. Para melhorar o desempenho de MAR-EP foi definido um limite inferior de módulo 0,1 para K_R e K_C, ou seja, controladores com ganhos menores que $0,1$ tem seus valores substituídos por 0,1.

Nas figuras do diagrama de fase, temporal agente 1, temporal agente 2 e ganhos K_R e K_C, respectivamente, figuras 4.1, 4.2, 4.3 e 4.4, são exibidos os resultados da proposta de controladores de MAR-EP, definidos por meio das equações (4.69) e (4.70) com $\kappa_1 = \kappa_2 = 60$, comparados com controladores $k_1 = k_2 = \alpha = 0,1$ fixos de BPL. BPL está representado com linha sólida, enquanto MAR-EP em linha tracejada.

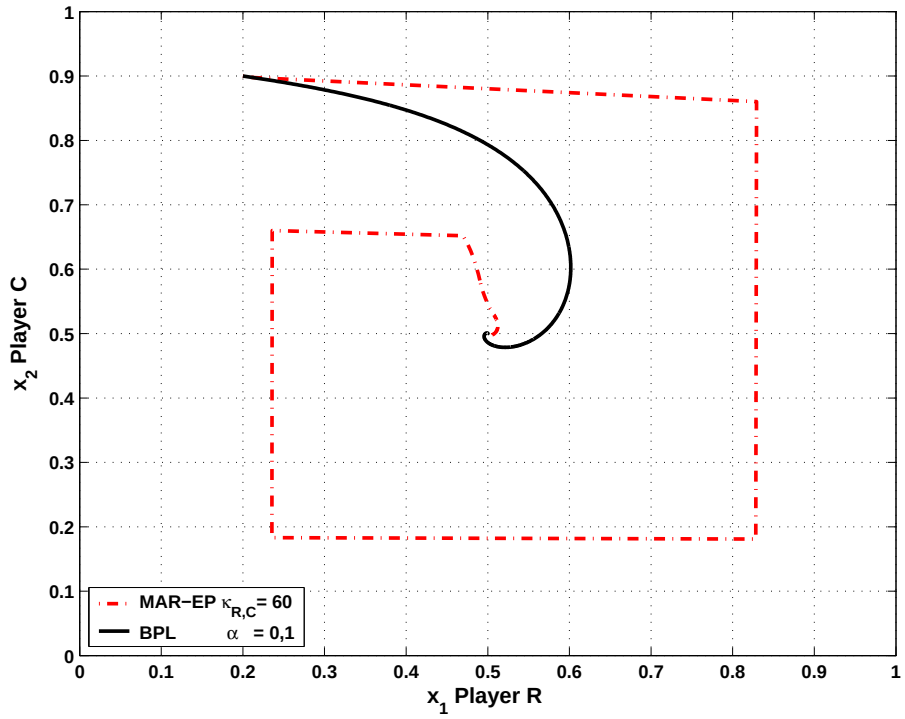


Figura 4.1: Diagrama de fase com MAR-EP em linha traço ponto com $g_1(x_1)$ e $g_2(x_2)$ teóricos e $\kappa_R = \kappa_C = 60$ e BPL em linha sólida com $\alpha = 0,1$, ponto inicial $(0,2, 0,9)$, do jogo casar moedas.

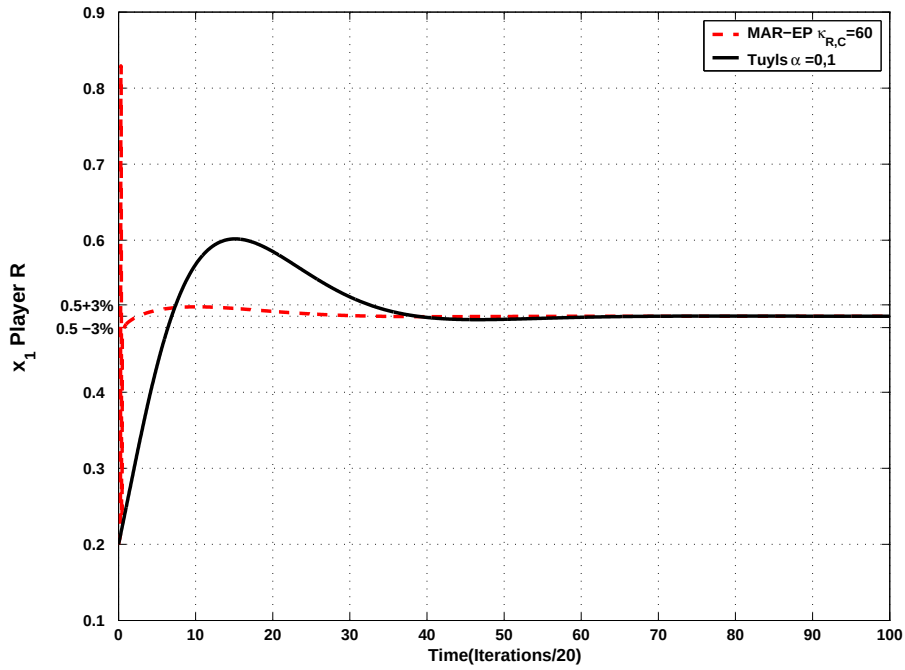


Figura 4.2: Gráfico temporal x_1 do agente 1 com MAR-EP em linha traço ponto com $g_1(x_1)$ e $g_2(x_2)$ teóricos e $\kappa_R = \kappa_C = 60$ e BPL em linha sólida com $\alpha = 0,1$, ponto inicial $(0,2, 0,9)$, do jogo casar moedas.

O gráfico de MAR-EP no diagrama de fase não é suave (é o compromisso que se paga ao se exigir rapidez de convergência) quando comparado a BPL.

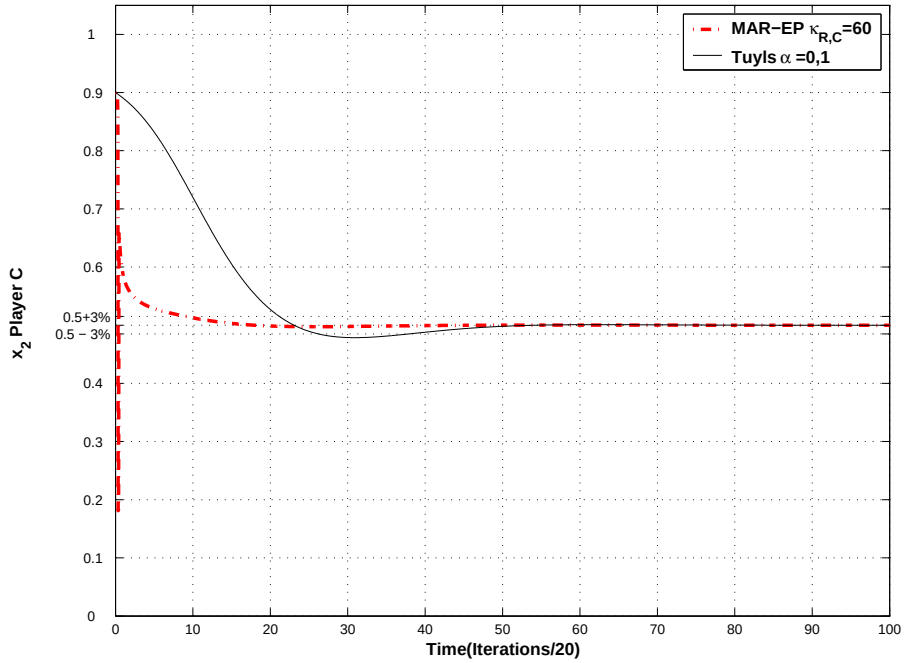


Figura 4.3: Gráfico temporal x_2 do agente 2 com MAR-EP em linha traço ponto com $g_1(x_1)$ e $g_2(x_2)$ teóricos e $\kappa_R = \kappa_C = 60$ e BPL em linha sólida com $\alpha = 0,1$, ponto inicial $(0,2, 0,9)$, do jogo casar moedas.

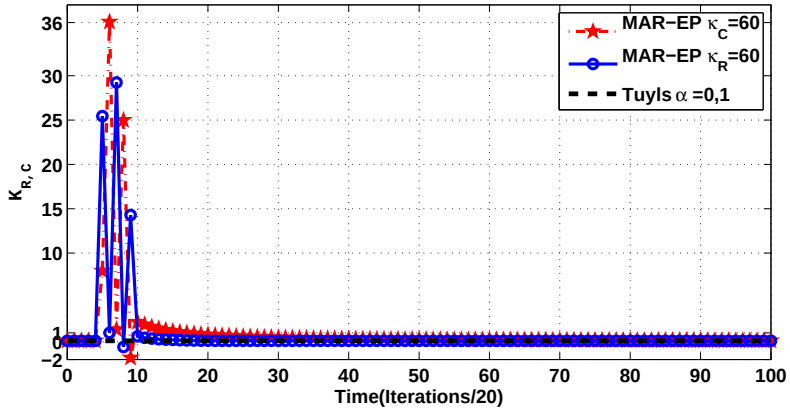


Figura 4.4: Amplitudes de K_R e K_C dos agentes 1 e 2, respectivamente, com MAR-EP com $g_1(x_1)$ e $g_2(x_2)$ teóricos e $\kappa_R = \kappa_C = 60$ e com BPL com $\alpha = 0,1$, ponto inicial $(0,2, 0,9)$, do jogo casar moedas.

Tanto do gráfico temporal de x_1, quanto o gráfico temporal de x_2, figuras 4.2 e 4.3, nota-se que a convergência das ações é mais rápida com MAR-EP do que com BPL. Em ambas figuras o jogo é o de casar moedas, o ponto inicial das trajetórias arbitrado foi $(0,2, 0,9)$, MAR-EP está implementado com $\kappa_R = \kappa_C = 60$ e está representado com linha traço ponto, enquanto BPL com $\alpha = 0,1$ em linha sólida.

As amplitudes dos ganhos dos controladores K_R e K_C de MAR-EP, representadas na figura 4.4, têm valores elevados se comparados a α, mas apenas no início do tempo. K_R está representado com linha sólida com círculo, K_C em linha traço

ponto com estrela e α em linha tracejada.

Segundo exemplo teórico com jogo casar moedas

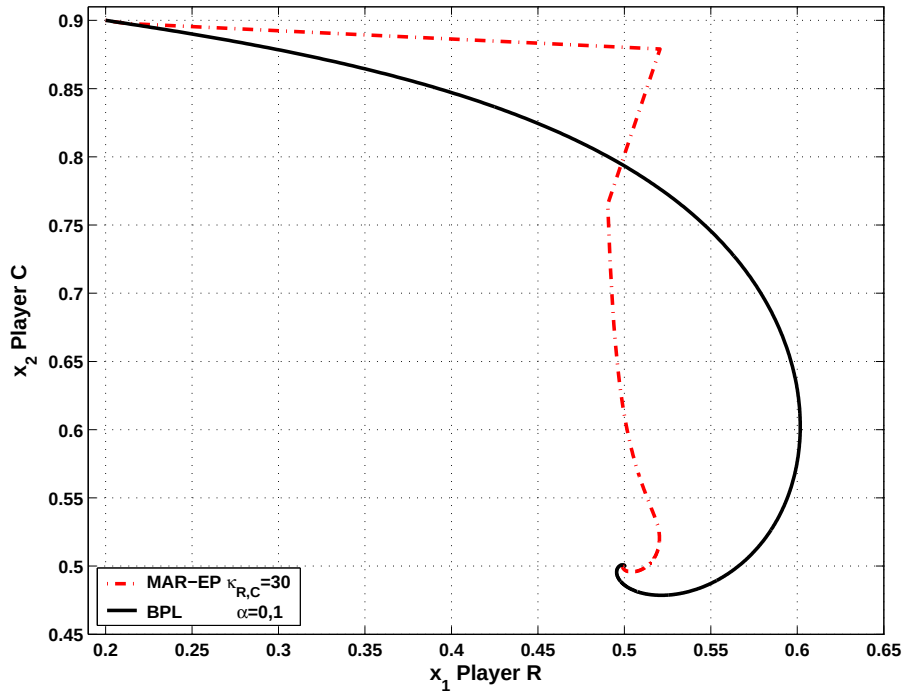


Figura 4.5: Diagrama de fase $x_1 \times x_2$ com MAR-EP com $\kappa_{R,C} = 30$ em linha tracejada e com BPL de ganho $\alpha = 0,1$ em linha sólida, ponto inicial $(0,2, 0,9)$, $\tau = 1$, no jogo de casar moedas.

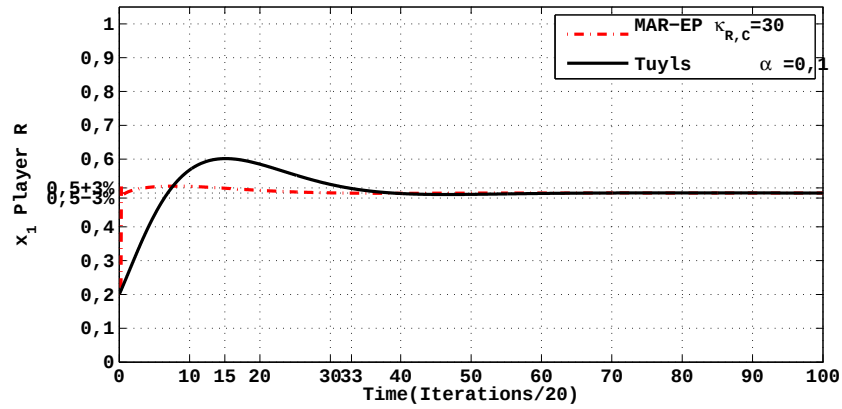


Figura 4.6: Gráfico temporal de x_1 do agente 1 com MAR-EP em linha tracejada com $\kappa_{R,C} = 30$ e com BPL em linha sólida, de ganho $\alpha = 0,1$, $\tau = 1$, ponto inicial $(0,2, 0,9)$, no jogo de casar moedas.

O segundo exemplo é feito comparando BPL com MAR-EP com ganho $\kappa_R = \kappa_C = 30$ para MAR-EP, $\alpha = 0,1$ fixo de BPL, obtendo-se as figuras do diagrama de

fase, temporal agente 1, temporal agente 2 e amplitudes de K_R e K_C, respectivamente, figuras 4.5, 4.6, 4.7 e 4.8. BPL está representado com linha sólida, enquanto MAR-EP em linha tracejada.

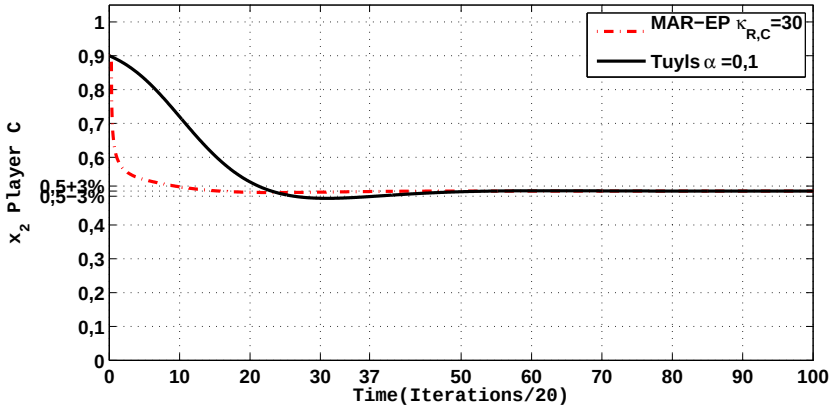


Figura 4.7: Gráfico temporal de x_2 do agente 2 em linha tracejada em MAR-EP com $\kappa_{R,C} = 30$ e em linha sólida em BPL de ganho $\alpha = 0,1$, $\tau = 1$, ponto inicial $(0,2, 0,9)$, no jogo de casar moedas.

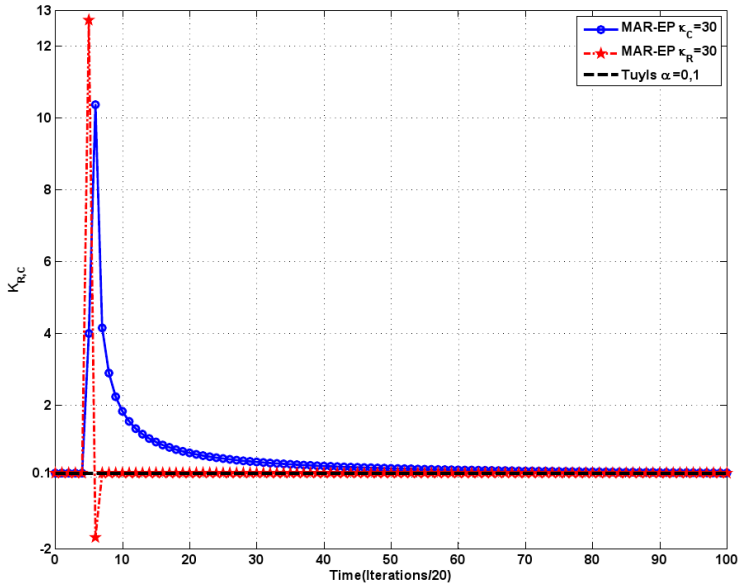


Figura 4.8: Amplitude de K_R em linha sólida com círculo em MAR-EP, amplitude de K_C em linha traço ponto com estrela em MAR-EP para $\kappa_R = \kappa_C = 5$ e BPL em linha tracejada, $\alpha = 0,1$, $\tau = 1$, ponto inicial $(0,2, 0,9)$, no jogo de casar moedas.

4.5 Determinação dos parâmetros $\hat{g}_1(\hat{x}_1)$ e $\hat{g}_2(\hat{x}_2)$ estimados

As equações (4.69) e (4.70), como comentado na seção anterior, têm alguns óbices: está implícito que o agente 1 tem as informações de x_2 e da matriz do agente 2, assim como que o agente 2 tem as informações de x_1 e da matriz do agente 1.

Para contornar os óbices comentados, supõe-se que o agente 1 possa obter uma aproximação de x_2 e dos parâmetros $\bar{c}$, $\check{c}$ do agente 2, assim como o agente 2 possa obter uma aproximação de x_1 e dos parâmetros $\bar{r}$, $\check{r}$ do agente 1.

Seja $\hat{x}_1$ é uma estimativa do valor de x_1 obtida pelo agente 2 usando o valor atrasado de x_1 e $\hat{x}_2$ é uma estimativa do valor de x_2 obtida pelo agente 1 usando o valor atrasado de x_2. Sejam, ainda, $g_1(x_1)$ e $g_2(x_2)$ dados por (4.63) e (4.64) e reescritos a seguir em (4.73) e (4.74):

$$g_1(x_1) = \frac{-\tau\check{r} + \log\left(\frac{x_1}{1-x_1}\right)}{\tau\bar{r}} \tag{4.73}$$

$$g_2(x_2) = \frac{-\tau\check{c} + \log\left(\frac{x_2}{1-x_2}\right)}{\tau\bar{c}} \tag{4.74}$$

Sejam $\hat{g}_1(\hat{x}_1)$ e $\hat{g}_2(\hat{x}_2)$ dadas, respectivamente, por (4.75) e (4.76) a seguir, onde $\hat{g}_1(\hat{x}_1) = g_1(\hat{x}_1)$ e $\hat{g}_2(\hat{x}_2) = g_2(\hat{x}_2)$, aonde o termo $\hat{g}(\cdot)$ foi empregado para explicitar que se trata de uma estimativa da função $g(\cdot)$:

$$\hat{g}_1(\hat{x}_1) = \frac{\tau(r_{22} - r_{12}) + \log\left(\frac{\hat{x}_1}{1-\hat{x}_1}\right)}{\tau\bar{r}} = -\frac{\check{r}}{\bar{r}} + \frac{\log\left(\frac{\hat{x}_1}{1-\hat{x}_1}\right)}{\tau\bar{r}} \tag{4.75}$$

$$\hat{g}_2(\hat{x}_2) = \frac{\tau(c_{22} - c_{21}) + \log\left(\frac{\hat{x}_2}{1-\hat{x}_2}\right)}{\tau\bar{c}} = -\frac{\check{c}}{\bar{c}} + \frac{\log\left(\frac{\hat{x}_2}{1-\hat{x}_2}\right)}{\tau\bar{c}} \tag{4.76}$$

De forma análoga, $\hat{f}_1(\hat{x}_1, x_2)$ e $\hat{f}_2(x_1, \hat{x}_2)$ são dados a seguir, aonde o termo $\hat{f}(\cdot)$ foi empregado para explicitar que se trata de uma estimativa da função $f(\cdot)$ ao usar um termo estimado:

$$\hat{f}_1(\hat{x}_1, x_2) = \left[\tau(\bar{r}x_2 + \check{r}) + \log\left(\frac{1-\hat{x}_1}{\hat{x}_1}\right)\right] \tag{4.77}$$

$$\hat{f}_2(x_1, \hat{x}_2) = \left[\tau(\bar{c}x_1 + \check{c}) + \log\left(\frac{1-\hat{x}_2}{\hat{x}_2}\right)\right] \tag{4.78}$$

Projetam-se, assim, K_R e K_C da seguinte forma:

$$K_R = -\frac{\kappa_R}{\hat{f}_1(x_1, \hat{x}_2)} \left[x_1(1-x_1)(x_1 - \hat{g}_2(\hat{x}_2)) - \frac{\hat{x}_2 - g_1(x_1)}{\tau \bar{r}} \right] \tag{4.79}$$

$$K_C = -\frac{\kappa_C}{\hat{f}_2(\hat{x}_1, x_2)} \left[x_2(1-x_2)(x_2 - \hat{g}_1(\hat{x}_1)) - \frac{\hat{x}_1 - g_2(x_2)}{\tau \bar{c}} \right] \tag{4.80}$$

com $\kappa_R > 0$, $\kappa_C > 0$.

Seja

$$\hat{g}_1(\hat{x}_1) \approx g_1(x_1) \tag{4.81}$$

$$\hat{g}_2(\hat{x}_2) \approx g_2(x_2) \tag{4.82}$$

Então, com (4.81) e (4.77) em (4.79) e de (4.82) e (4.78) em (4.80) a seguir, e têm-se $\dot{V}_{Liap}$ de (4.68) como:

$$\begin{aligned} \dot{V}_{Liap} = \quad & -\kappa_R \left(\tfrac{\hat{f}_1(x_1, \hat{x}_2)}{f_1(x_1, x_2)} \right) \left[x_1(1-x_1)(x_1 - \hat{g}_2(\hat{x}_2)) - \tfrac{\hat{x}_2 - g_1(x_1)}{\tau \bar{r}} \right]^2 \\ & -\kappa_C \left(\tfrac{\hat{f}_2(\hat{x}_1, x_2)}{f_2(x_1, x_2)} \right) \left[x_2(1-x_2)(x_2 - \hat{g}_1(\hat{x}_1)) - \tfrac{\hat{x}_1 - g_2(x_2)}{\tau \bar{c}} \right]^2 \end{aligned} \tag{4.83}$$

A ideia do projeto dos controladores de estimação preliminar é que o agente 1 siga a curva $f_2(x_1, x_2) = 0$ e o agente 2 siga a curva $f_1(x_1, x_2) = 0$.

O agente 1 terá de fato uma estimativa de $f_2(x_1, x_2) = 0$, ao usar $\hat{g}_2(\hat{x}_2)$, assim como o agente 2 terá uma estimativa de $f_1(x_1, x_2) = 0$, ao usar $\hat{g}_1(\hat{x}_1)$.

Os agentes usam as respectivas estimativas de x_1 e de x_2 atrasados de uma unidade de tempo, quando do cálculo de seu controlador. Para tanto, as aproximações a seguir são usadas:

$$\hat{f}_1(\hat{x}_1, x_2) \approx f_1(x_1, x_2) \tag{4.84}$$

$$\hat{f}_2(x_1, \hat{x}_2) \approx f_2(x_1, x_2) \tag{4.85}$$

Então, tem-se com (4.84) e (4.85) em (4.83):

$$\begin{aligned} \hat{\dot{V}}_{Liap} = \quad & -\kappa_R \left[x_1(1-x_1)(x_1 - \hat{g}_2(\hat{x}_2)) - \tfrac{\hat{x}_2 - g_1(x_1)}{\tau \bar{r}} \right]^2 \\ & -\kappa_C \left[x_2(1-x_2)(x_2 - \hat{g}_1(\hat{x}_1)) - \tfrac{\hat{x}_1 - g_2(x_2)}{\tau \bar{c}} \right]^2 \end{aligned} \tag{4.86}$$

para $\kappa_R > 0$ e $\kappa_C > 0$.

Na realidade, não há garantia de $\dot{V}_{Liap} < 0$. O que se tem em (4.86) é apenas

uma aproximação, já que foram feitas considerações de aproximação em (4.81), (4.82), (4.84) e em (4.85). A escolha dos controladores feitas em (4.79) e em (4.80) conduzem a $\dot{V}_{Liap}$ que tende a ser negativo, mas que pode ser positivo pontualmente.

A medida que as ações se aproximam do ponto de equilíbrio, os ganhos definidos nas equações (4.79) e (4.80) decrescem. A solução adotada, para que os mesmos não se tornem pequenos e a convergência se torne demasiadamente lenta, foi a utilização de um critério para chaveamento do controlador da segunda fase para a terceira fase, como será abordado na seção 4.6, baseado na equação (4.87):

$$\begin{aligned} \dot{V}_{Liap} = \quad & -\kappa_R \left[x_1(1-x_1)(x_1 - \hat{g}_2(\hat{x}_2)) - \tfrac{\hat{x}_2 - g_1(x_1)}{\tau \bar{r}} \right]^2 \\ & -\kappa_C \left[x_2(1-x_2)(x_2 - \hat{g}_1(\hat{x}_1)) - \tfrac{\hat{x}_1 - g_2(x_2)}{\tau \bar{c}} \right]^2 \end{aligned} \tag{4.87}$$

Retornando ao cálculo dos parâmetros estimados $\hat{g}_1(\hat{x}_1)$ e $\hat{g}_2(\hat{x}_2)$, o controlador projetado pelo agente 1 usa, conforme equação (4.79), $\hat{g}_2(\hat{x}_2)$, enquanto o controlador projetado pelo agente 2 usa, conforme equação (4.80), $\hat{g}_1(\hat{x}_1)$. Cada agente i, $i \in \{1, 2\}$, usa a primeira fase do controlador para estimar dois parâmetros, que são informações não conhecidas do outro agente j, onde $i \neq j$ e $i, j \in \{1, 2\}$, e então estimar $\hat{g}_j(\hat{x}_j)$.

Repete-se aqui $\hat{g}_2(\hat{x}_2)$ para facilidade de entendimento:

$$\hat{g}_2(\hat{x}_2) = \frac{-\tau \check{c} + \log(\frac{\hat{x}_2}{1-\hat{x}_2})}{\tau \bar{c}} \tag{4.88}$$

Sejam b_0, b_1, definidos por (4.89), (4.90):

$$b_0 = -\tau \check{c} \tag{4.89}$$

$$b_1 = \tau \bar{c} \tag{4.90}$$

Escrevendo $\hat{g}_2(\hat{x}_2)$ em função dos parâmetros b_0, b_1:

$$\hat{g}_2(\hat{x}_2) = \frac{b_0 + \log(\frac{\hat{x}_2}{1-\hat{x}_2})}{b_1} \tag{4.91}$$

O agente 1, portanto, precisa estimar os parâmetros b_0, b_1 e substituí-los em (4.91) para obter $\hat{g}_2(\hat{x}_2)$, que é usada no seu controlador k_1.

De forma análoga, dado $\hat{g}_1(\hat{x}_1)$:

$$\hat{g}_1(\hat{x}_1) = \frac{-\tau \check{r} + \log(\frac{\hat{x}_1}{1-\hat{x}_1})}{\tau \bar{r}} \tag{4.92}$$

Sejam definidos a_0, a_1, respectivamente, por (4.93), (4.94):

$$a_0 = -\tau\check{r} \tag{4.93}$$

$$a_1 = \tau\bar{r} \tag{4.94}$$

Escrevendo $\hat{g}_1(\hat{x}_1)$ em função dos parâmetros a_0, a_1:

$$\hat{g}_1(\hat{x}_1) = \frac{a_0 + \log(\frac{\hat{x}_1}{1-\hat{x}_1})}{a_1} \tag{4.95}$$

Então, o agente 2 precisa estimar os parâmetros a_0, a_1, que são parâmetros do agente 1, e substituí-los em (4.95) para obter $\hat{g}_1(\hat{x}_1)$ que é usada no seu controlador k_2.

Os controladores dos agentes 1 e 2 foram definidos constantes iguais a $0 < \alpha < 1$ na primeira fase do controlador, as primeiras cinco iterações. Ressalta-se aqui que a primeira fase do controlador, pela hipótese de projeto, permite a estimação dos parâmetros a_0, a_1, b_0, b_1. As definições destes conduz $\hat{f}_1(x_1, \hat{x}_2)$ de (4.77) e $\hat{f}_2(\hat{x}_1, x_2)$ de (4.78) serem dados, respectivamente, por:

$$\hat{f}_1(\hat{x}_1, x_2) = \frac{\dot{\hat{x}}_1}{(1-\hat{x}_1)\hat{x}_1\alpha} = \left[\underbrace{\tau\bar{r}}_{a_1} x_2 \underbrace{-\tau\check{r}}_{a_0} + \log\left(\frac{1-\hat{x}_1}{\hat{x}_1}\right) \right] \tag{4.96}$$

E por:

$$\hat{f}_2(x_1, \hat{x}_2) = \frac{\dot{\hat{x}}_2}{(1-\hat{x}_2)\hat{x}_2\alpha} = \left[\underbrace{\tau\bar{c}}_{b_1} x_1 + \underbrace{-\tau\check{c}}_{b_0} + \log\left(\frac{1-\hat{x}_2}{\hat{x}_2}\right) \right] \tag{4.97}$$

Reescrevendo a equação (4.96), tem-se:

$$-a_0 + a_1 x_2 = \underbrace{\hat{f}_1(\hat{x}_1, x_2) - \log\left(\frac{1-\hat{x}_1}{\hat{x}_1}\right)}_{\text{termo independente}} \tag{4.98}$$

Reescrevendo a equação (4.97), tem-se:

$$-b_0 + b_1 x_1 = \underbrace{\hat{f}_2(x_1, \hat{x}_2) - \log\left(\frac{1-\hat{x}_2}{\hat{x}_2}\right)}_{\text{termo independente}} \tag{4.99}$$

Para a obtenção dos parâmetros a_0, a_1 estimados para o agente 2, monta-se um sistema do tipo:

$$A_2 \cdot X_2 = B_2 \tag{4.100}$$

Constrói-se a matriz A_2 com cinco linhas e duas colunas. Cada uma das cinco iterações corresponde a uma linha com os valores 1 e $\hat{x}_2$. O vetor B_2, com cinco linhas e uma coluna, é formado pelo termo independente dos parâmetros a_0, a_1, assinalado em (4.98), um para cada uma das cinco iterações. X_2 é o vetor coluna dos parâmetros a_0, a_1.

Pelo método de mínimos quadrados, obtém-se X_2 conforme a equação (4.101).

$$X_2 = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \left({A_2}^T \cdot A_2\right)^{-1} \cdot {A_2}^T \cdot B_2 \tag{4.101}$$

De modo análogo, para a obtenção dos parâmetros b_0, b_1 estimados para o agente 1, monta-se um sistema do tipo:

$$A_1 \cdot X_1 = B_1 \tag{4.102}$$

Constrõe-se a matriz A_1 com cinco linhas e duas colunas. Cada uma das cinco iterações corresponde a uma linha com os valores 1 e $\hat{x}_1$. O vetor B_1, com cinco linhas e uma coluna é formado pelo termo independente dos parâmetros b_0, b_1, assinalado em (4.99), um para cada uma das cinco iterações. X_1 é o vetor coluna dos parâmetros b_0, b_1.

Pelo método de mínimos quadrados, obtém-se X_1 conforme a equação (4.103).

$$X_1 = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \left({A_1}^T \cdot A_1\right)^{-1} \cdot {A_1}^T \cdot B_1 \tag{4.103}$$

Por construção do controlador de MAR-EP, quando (x_1, x_2) se aproxima do ponto de equilíbrio, os ganhos K_R, K_C são reduzidos até que $\dot{V}$ se tornam menores que um valor determinado, ficando a progressão de MAR-EP mais lenta que BPL. Para superar essa condição, foi definido um limitante inferior para $\hat{V}_{Liap_{\text{agente } i}}$, $i \in \{R, C\}$, uma aproximação de V_{Liap} definida na equação (4.105) da seção 4.6 a seguir, que serve de critério de chaveamento para a terceira fase de MAR-EP, aonde os ganhos K_R e K_C passam a ser constantes iguais a α como em (TUYLS *et al.*, 2006).

4.6 As fases dos controladores

Os controladores foram divididos em três fases:

1. Na primeira fase, que é composta de 5 iterações por definição, os agentes usam o controlador de ganho constante igual a α. Durante essa fase, cada agente

estima dois parâmetros: o agente 1 estima b_0, b_1 e o agente 2 estima a_0, a_1.

2. Na segunda fase, os agentes usam os parâmetros estimados na primeira fase para o projeto do controlador da segunda fase. O uso do controlador da segunda fase leva a dinâmica a um ponto de equilíbrio próximo ao ponto de equilíbrio obtido com controladores positivos. A segunda fase dos controladores é o diferencial do método, trazendo vantagens consideráveis como a redução do tempo de convergência, além da menor variação da amplitude das estratégias em relação ao ponto de convergência.

3. Na terceira fase os controladores adotados são ganhos constantes. No final da segunda fase os valores dos ganhos K_R e K_C tornam-se pequenos, assim como $\hat{V}_{Liap_{\text{agente }i}}$, $i \in \{R, C\}$. Quando $\hat{V}_{Liap_{\text{agente }i}}$ atinge um valor abaixo de limite determinado, os agentes chaveiam para o ganho constante igual a α. A escolha deste limite é uma variável de projeto a ser ajustada.

   Na segunda fase os ganhos dos controladores tornam-se pequenos (não necessariamente de modo simultâneo) e cada agente faz seu controlador ter ganho constante igual a α. Supõe-se que a estimação dos parâmetros é razoável tal que os ganhos dos controladores ao longo do tempo se tornem pequenos, menores que o limite inferior definido. Quanto maior o limite inferior, menor a segunda fase.

Os ganhos e o critério de chaveamento da segunda para terceira fase são dados por:

$$\begin{cases} 2^{\text{a}} \text{ fase} \quad k_i = K_i, \quad i \in \{R, C\} & \text{se } \hat{V}_{Liap_{\text{agente }i}} > -0,1 \\ 3^{\text{a}} \text{ fase} \quad k_i = \alpha & \text{se } \hat{V}_{Liap_{\text{agente }i}} \leq -0,1 \end{cases}, \tag{4.104}$$

onde k_i é ganho do agente i, $i \in \{R, C\}$.

Seja $\hat{V}_{Liap_{\text{agente }R}}(x_1, \hat{x}_2)$ dado por:

$$\hat{V}_{Liap_{\text{agente }R}}(x_1, \hat{x}_2) = \left((x_1 - \hat{g}_2(\hat{x}_2))^2 + (\hat{x}_2 - g_1(x_1))^2\right) \tag{4.105}$$

Seja $\hat{V}_{Liap_{\text{agente }C}}(\hat{x}_1, x_2)$ dado por:

$$\hat{V}_{Liap_{\text{agente }C}}(\hat{x}_1, x_2) = \left((\hat{x}_1 - g_2(x_2))^2 + (x_2 - \hat{g}_1(\hat{x}_1))^2\right) \tag{4.106}$$

Por construção, $\hat{V}_{Liap_{\text{agente }R}}(x_1, \hat{x}_2)$ e $\hat{V}_{Liap_{\text{agente }C}}(\hat{x}_1, x_2)$ são estimações de V_{Liap}, equação (4.86), e têm valores próximos um do outro. Servem de critério de chaveamento da segunda fase para a terceira fase para os dois agentes. O critério de chaveamento da segunda para terceira fase é absoluto, ou seja, quando $\hat{V}_{Liap_{\text{agente }i}}(x_i, \hat{x}_j)$ for menor que um determinado valor, chaveia-se para a terceira fase. O ajuste do valor de chaveamento é uma desvantagem do método.

4.7 Exemplos numéricos de BPL e de MAR-EP

Exemplo 1 de BPL: Na figura 4.9 mostra-se o diagrama de fase $x_1 \times x_2$ de BPL para o jogo casar moedas (*matching pennies*), com ponto inicial $(0,1,0,2)$, $\tau = 10$ e discretização $d = 0,05$ (o jogo é simulado em ambiente computacional com a dinâmica discretizada). Traçaram-se as trajetórias de BPL com linha sólida para $\alpha = 0,1$, de BPL com linha tracejada com quadrado para $\alpha = 0,5$ e de BPL com linha sólida com 'x' para $\alpha = 0,95$. Fica claro que a taxa de aprendizagem α deve ser baixa portanto, caso contrário BPL não converge para o d escolhido.

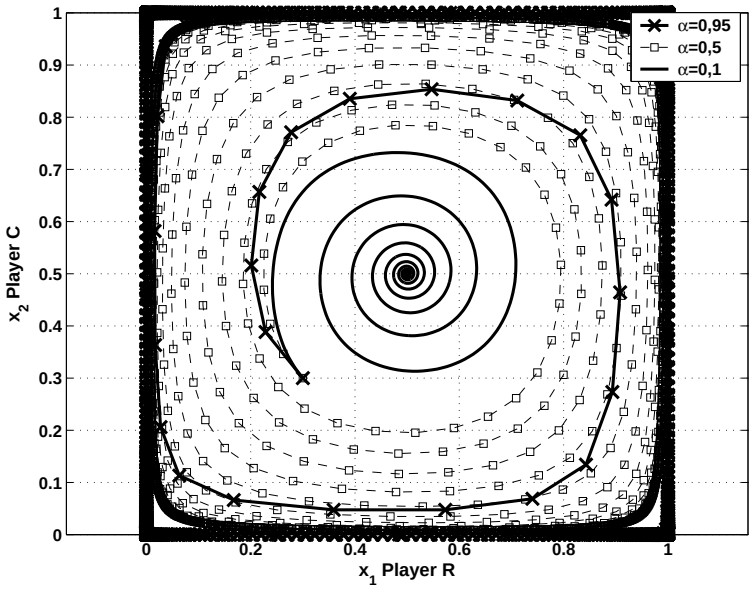


Figura 4.9: Diagrama de fase $x_1 \times x_2$ do jogo casar moedas, ponto inicial $(0,3,0,3)$ com três BPLs: BPL com $\tau = 0,1$ em linha sólida, BPL com $\tau = 0,5$ em linha tracejada com quadrado e BPL com $\tau = 0,95$ em linha sólida com 'x'.

Exemplo 2 de BPL: A figura 4.10 exibe no diagrama de fase $x_1 \times x_2$ a trajetória de BPL em linha sólida que não converge nem diverge, forma um ciclo limite. Foram usados $\alpha = 0,423$ e discretização $d = 0,05$. O uso de valor de α maior que $0,423$ em BPL diverge e, para valores inferiores a $0,423$ BPL converge conforme a trajetória tracejada, que usou $\alpha = 0,1$. O jogo exibido na figura 4.10 foi mais uma vez o jogo de casar moedas com ponto inicial $(0,1,0,2)$, $\tau = 10$ e discretização $d = 0,05$.

Exemplo 3 de BPL: Para exemplificar o efeito do valor de τ na dinâmica BPL, usou-se o jogo de casar moedas e ponto inicial $(0,3,0,3)$. Na figura 4.11 mostra-se no diagrama de fase $x_1 \times x_2$ o efeito do valor de τ na trajetória de BPL. Com $\alpha = 0,1$ e discretização $d = 0,05$, para $\tau = 0,1$ a trajetória foi uma curva tensa

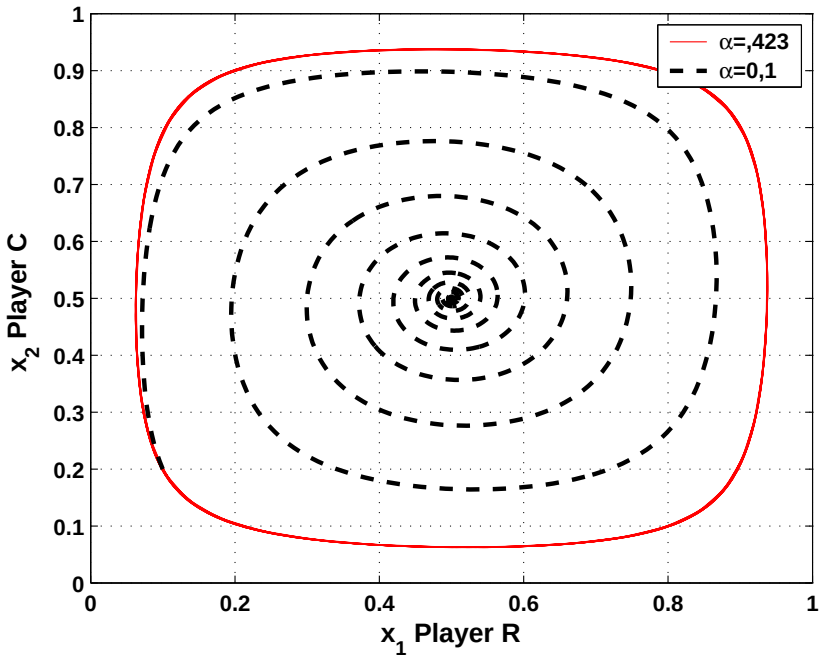


Figura 4.10: Diagrama de fase $x_1 \times x_2$ com ponto inicial $(0,1, 0,2)$ com BPL com $\alpha = 0,1$ em linha tracejada e BPL com $\alpha = 0,423$ em linha sólida.

linha tracejada no sentido de $(0,5, 0,5)$, enquanto para $\tau = 10$ a trajetória foi uma lenta espiral linha sólida até atingir $(0,5, 0,5)$. Já na figura 4.12 mostra-se o efeito do valor de τ escolhidos no gráfico temporal de x_1 do jogador R. Com $\tau = 10$ há oscilação grande em torno do ponto de convergência $0,5$ enquanto com $\tau = 0,1$ não há sobressinal (ultrapassagem do valor de convergência).

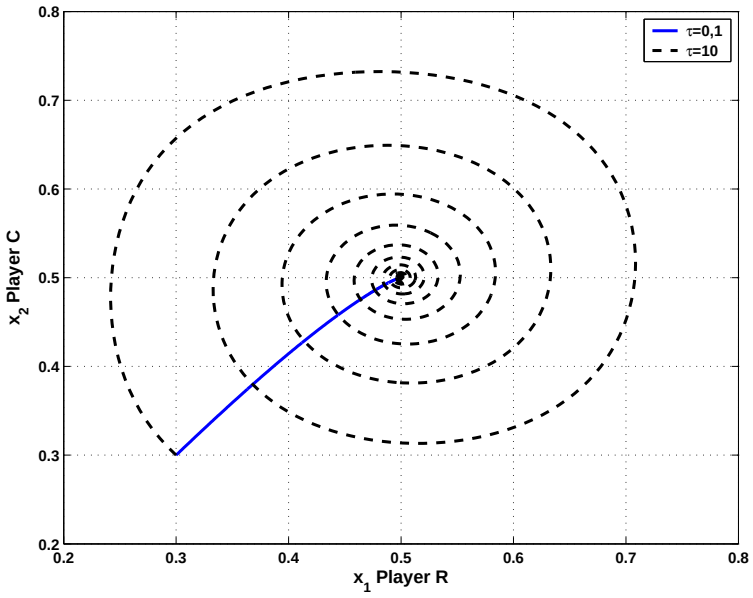


Figura 4.11: Diagrama de fase $x_1 \times x_2$ de BPL com parâmetro $\tau = 0,1$ em linha tracejada e com parâmetro $\tau = 10$ em linha sólida.

Exemplo 4 de MAR-EP $\times$ BPL: Comparou-se MAR-EP com BPL no exemplo numérico 4, por meio do jogo conhecido como jogo ardiloso [4]. Ilustram-se a comparação citada na figura 4.13 do diagrama de fase $x_1 \times x_2$, na figura 4.14 do gráfico temporal de x_1 do jogador R, na figura 4.15 do gráfico temporal de x_2 do jogador C e na figura 4.16 das amplitudes de K_R e K_C ao longo do tempo. Definiram-se o ponto inicial do jogo em $(0,1,0,2)$ e a taxa de aprendizagem $\alpha = 0,1$ para BPL e $\kappa_R = \kappa_C = 5$ e limite inferior dos ganhos os jogadores R e C iguais a $K_R = K_C = 0,1$ para MAR-EP. As matrizes de recompensa do jogo ardiloso estão dadas a seguir na equação (4.107).

$$R = \begin{bmatrix} 0 & 3 \\ 1 & 2 \end{bmatrix} \qquad C = \begin{bmatrix} 3 & 2 \\ 0 & 1 \end{bmatrix} \tag{4.107}$$

Nas figuras 4.13, 4.14 e 4.15 do exemplo numérico 4, BPL está representado em linha sólida e MAR-EP em linha traço ponto.

Exemplo numérico 5 de MAR-EP $\times$ BPL: Repetiu-se a comparação de MAR-EP com BPL no jogo ardiloso, mantendo o ponto inicial em $(0,1,0,2)$, mas dessa vez definindo $\alpha = 0,5$ e $\kappa_R = \kappa_C = 18,88125$. Na figura 4.17 do diagrama de fase $x_1 \times x_2$, na figura 4.18 do gráfico temporal de x_1 do jogador R, na figura 4.19

[4] O jogo ardiloso (*tricky game*) foi usado como exemplo por BOWLING e VELOSO (2002) e por ABDALLAH e LESSER (2008).

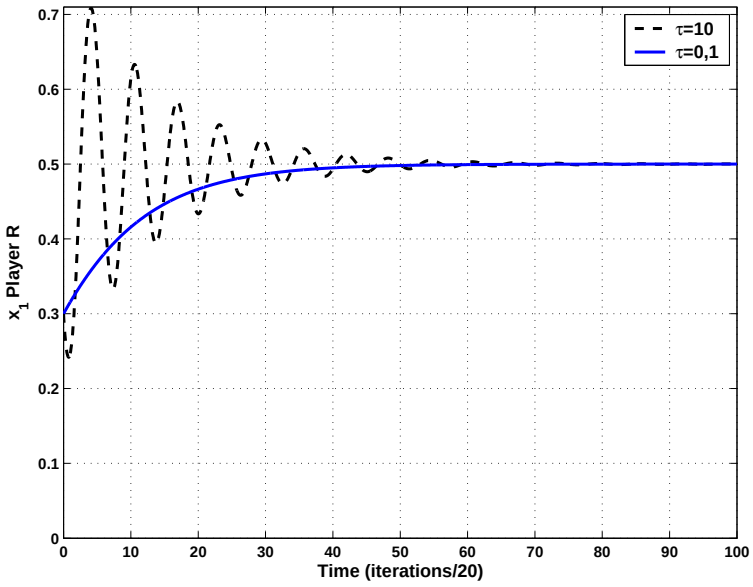


Figura 4.12: Gráfico temporal de x_1 do agente 1 de BPL com parâmetro $\tau = 0,1$ em linha tracejada e com parâmetro $\tau = 10$ em linha sólida.

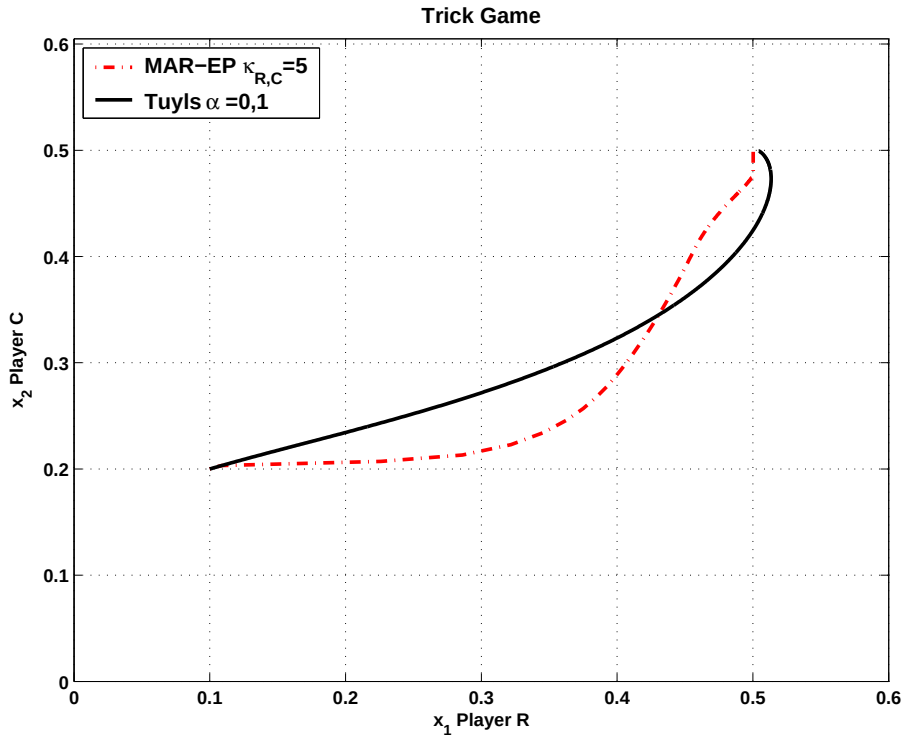


Figura 4.13: Diagrama de fase $x_1 \times x_2$ do exemplo numérico 4, jogo ardiloso, com MAR-EP em linha traço ponto com $\kappa_R = \kappa_C = 5$ e com BPL em linha sólida, parâmetro $\alpha = 0,1$.

do gráfico temporal de x_2 do jogador C e na figura 4.20 das amplitudes de K_R e K_C ao longo do tempo, ilustram-se a comparação citada.

Nas figuras 4.17, 4.18 e 4.19 do exemplo numérico 5, BPL está representado em linha sólida e MAR-EP em linha tracejada.

Na figura 4.20 do exemplo numérico 5, o eixo das ordenadas é a amplitude K_C que está representada em linha traço ponto e a amplitude K_R, que está representada em linha contínua para MAR-EP, enquanto o eixo das abcissas é o número de iterações dividido por 20. O valor de α para BPL está representado em linha tracejada grossa.

Os controladores do agentes adotados na fase 2 possuem cada um dois parâmetros estimados. Os erros de estimação inerentes levam a um ponto estimado P_E, numa vizinhança do ponto de equilíbrio, P_F, que é o ponto de equilíbrio da dinâmica. Daí, a necessidade de ambos agentes chavearem novamente para controladores fixos e assim poderem atingir o ponto P_F.

Na figura 4.21 exibe-se, no espaço $(x_1, x_2) \in [0, 1] \times [0, 1]$, o diagrama das primeiras estratégias dos agentes 1 e 2 (por simplicidade de notação é chamado ao longo do texto de diagrama de fase), respectivamente x_1 e x_2, um exemplo de jogo com dois agentes, duas ações de subclasse 3, conforme (TUYLS *et al.*, 2006), com matrizes R e C dadas pela equação (4.108), em que o ponto inicial (PI) foi definido em PI $= (0, 2, 0, 9)$. As curvas $f_1(x_1, x_2) = 0$ e $f_2(x_1, x_2) = 0$ se intercep-

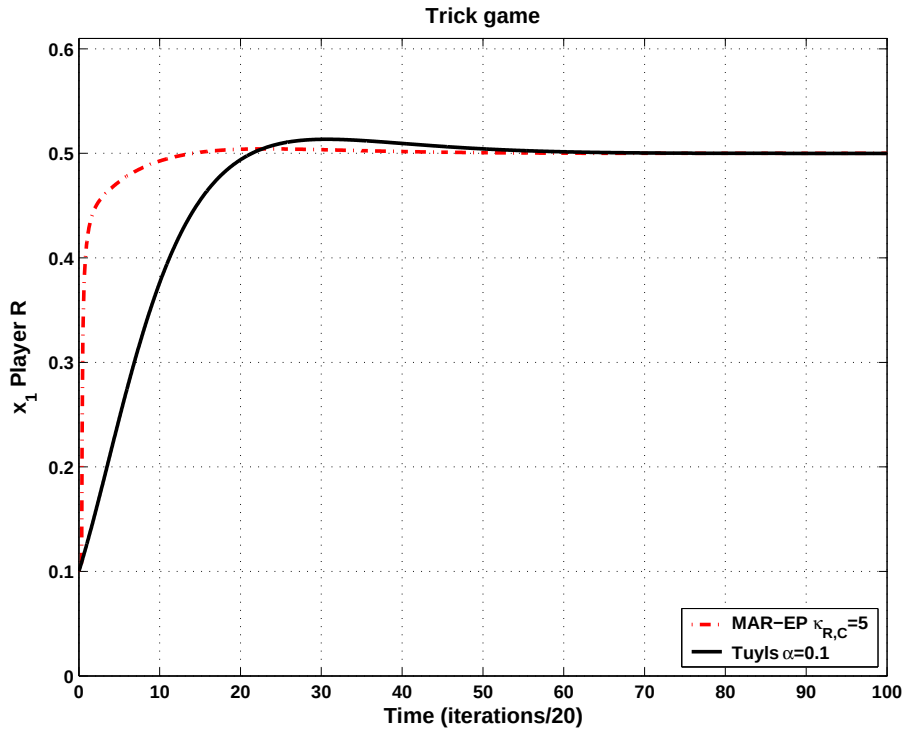


Figura 4.14: Gráfico temporal de x_1 do exemplo numérico 4, jogo ardiloso, com $\kappa_R = \kappa_C = 5$ com MAR-EP em linha traço ponto e com BPL em linha sólida, parâmetro $\alpha = 0,1$.

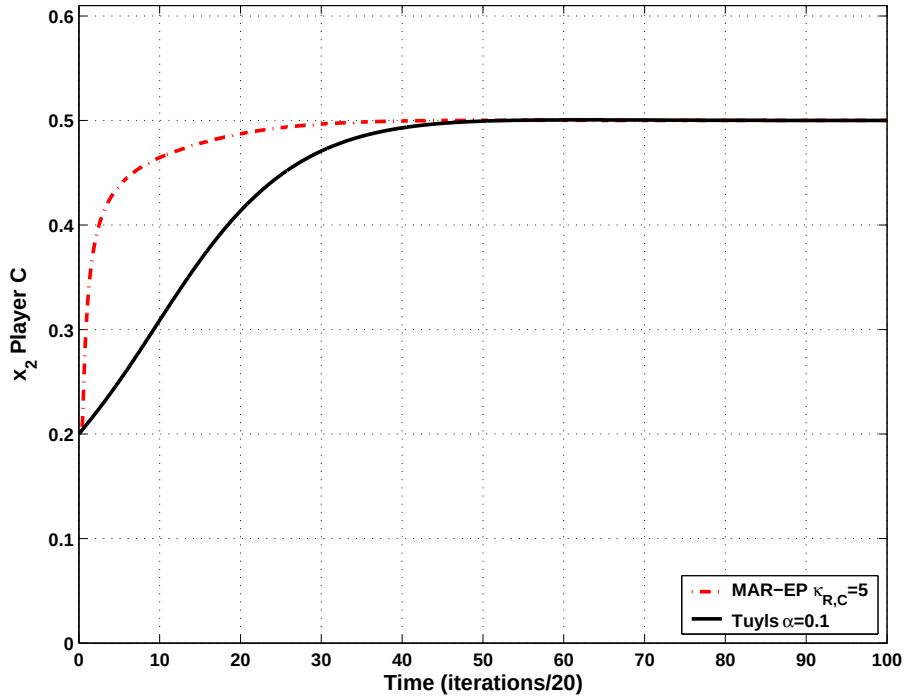


Figura 4.15: Gráfico temporal de x_2 do exemplo numérico 4, jogo ardiloso, com $\kappa_R = \kappa_C = 5$ com MAR-EP em linha traço ponto e com BPL em linha sólida, parâmetro $\alpha = 0,1$.

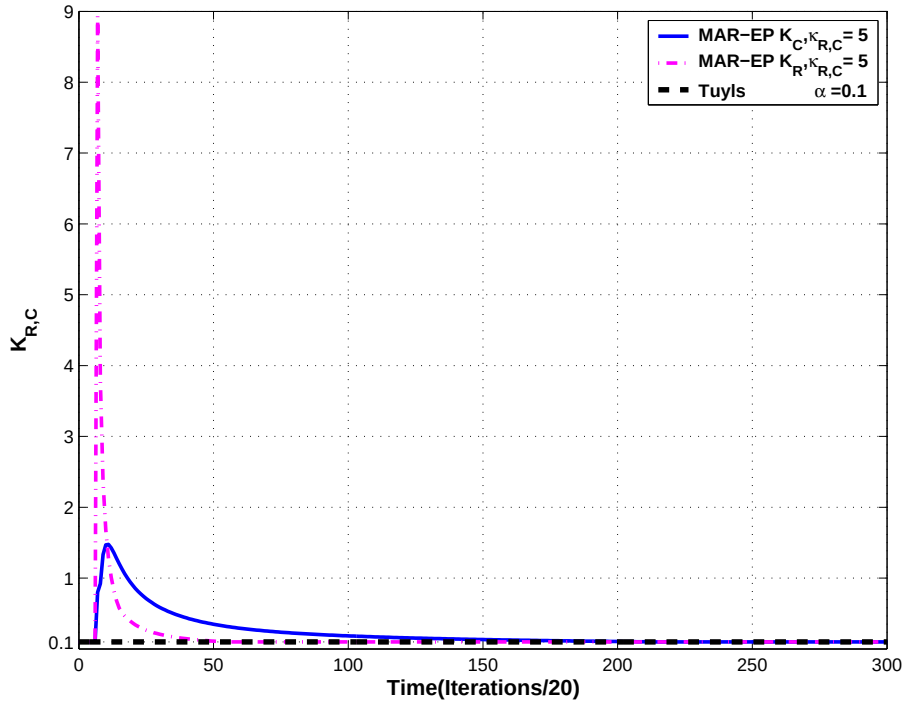


Figura 4.16: Amplitudes de K_R e de K_C do exemplo numérico 4, jogo ardiloso, com MAR-EP com $\kappa_R = \kappa_C = 5$ em linha traço ponto e com BPL em linha sólida, parâmetro $\alpha = 0,1$.

tam dando origem ao ponto de equilíbrio P_F, onde $P_F = (0,7187, 0,66606)$. Em pontilhado grosso tem-se a $f_2(x_1, x_2) = 0$ e em pontilhado fino tem-se a estimativa $\hat{f}_2(x_1, x_2) = 0$. Exibem-se em tracejado grosso a $f_1(x_1, x_2) = 0$ e em tracejado fino a estimativa $\hat{f}_1(x_1, x_2) = 0$. A curva da dinâmica de Tuyls foi representada em linha sólida fina e a dinâmica usando os controladores propostos em linha traço ponto média.

$$R = \begin{bmatrix} 2 & 0 \\ 4 & -3 \end{bmatrix} \qquad C = \begin{bmatrix} 3 & 0 \\ 2 & 4 \end{bmatrix} \tag{4.108}$$

4.8 Conclusões

Um método de aprendizado por reforço foi desenvolvido, MAR-EP, para o jogo de dois jogadores, duas ações. Não há a garantia de MAR-EP ser sempre melhor que o método adotado por TUYLS *et al.* (2006) com ganho constante, ainda que uma estimativa adequada da ação do jogador oponente leve a esse resultado. Porém, a natureza de MAR-EP, especificamente sua terceira fase, garante a convergência do método.

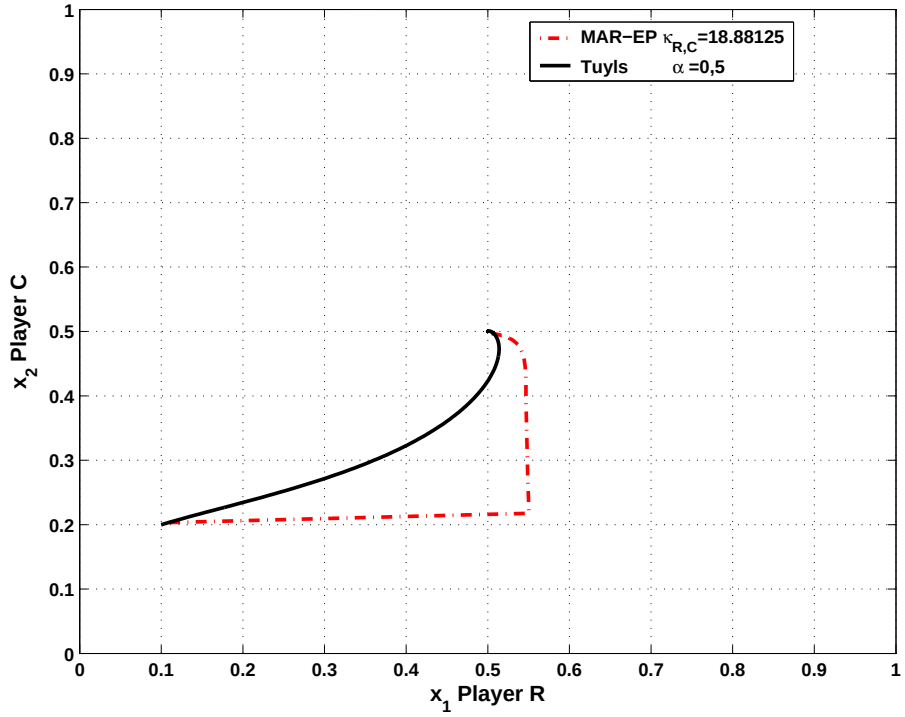


Figura 4.17: Diagrama de fase $x_1 \times x_2$ do exemplo numérico 5, jogo ardiloso, com MAR-EP em linha traço ponto e com BPL em linha sólida, com ponto inicial em $(0,1,0,2)$.

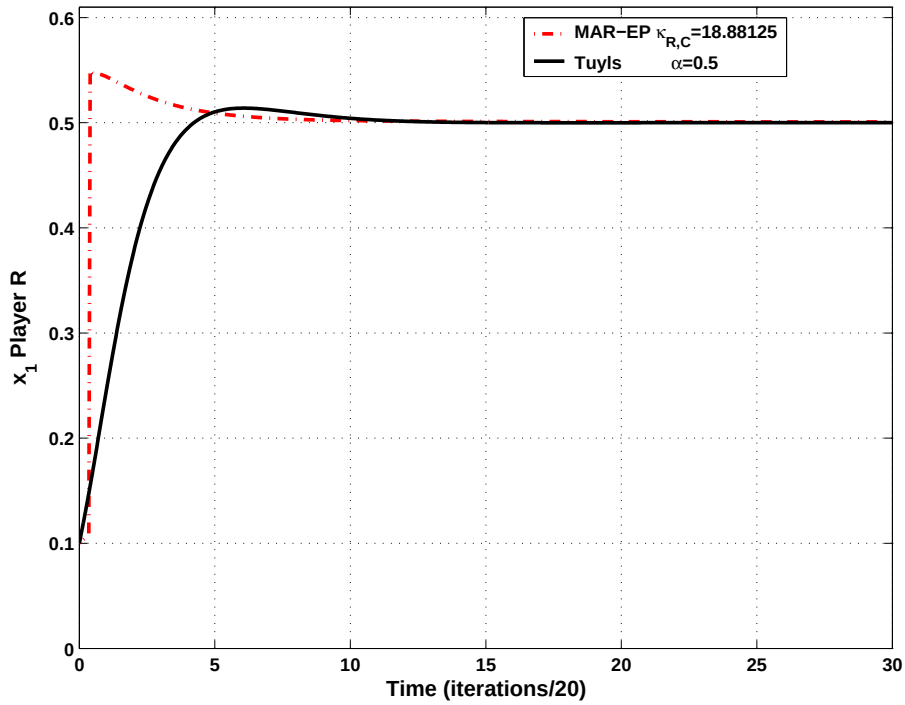


Figura 4.18: Comparação temporal de x_1 do exemplo numérico 5, jogo ardiloso, com MAR-EP em linha traço ponto e com BPL em linha sólida, com ponto inicial em $(0,1,0,2)$.

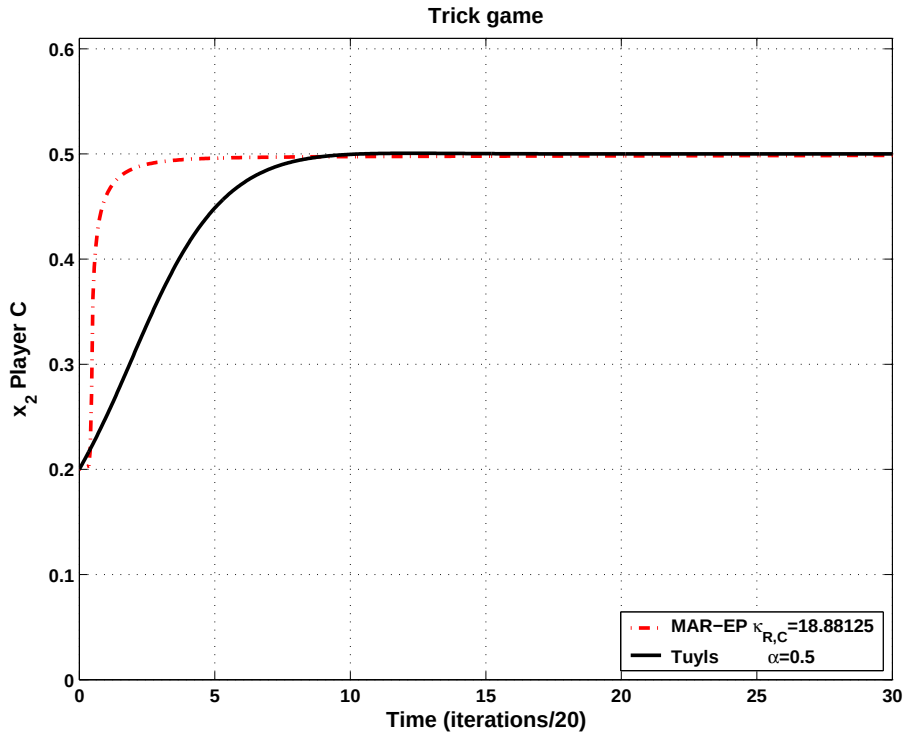


Figura 4.19: Comparação temporal de x_2 do exemplo numérico 5, jogo ardiloso, com MAR-EP em linha traço ponto e com BPL em linha sólida, com ponto inicial em $(0,1,0,2)$.

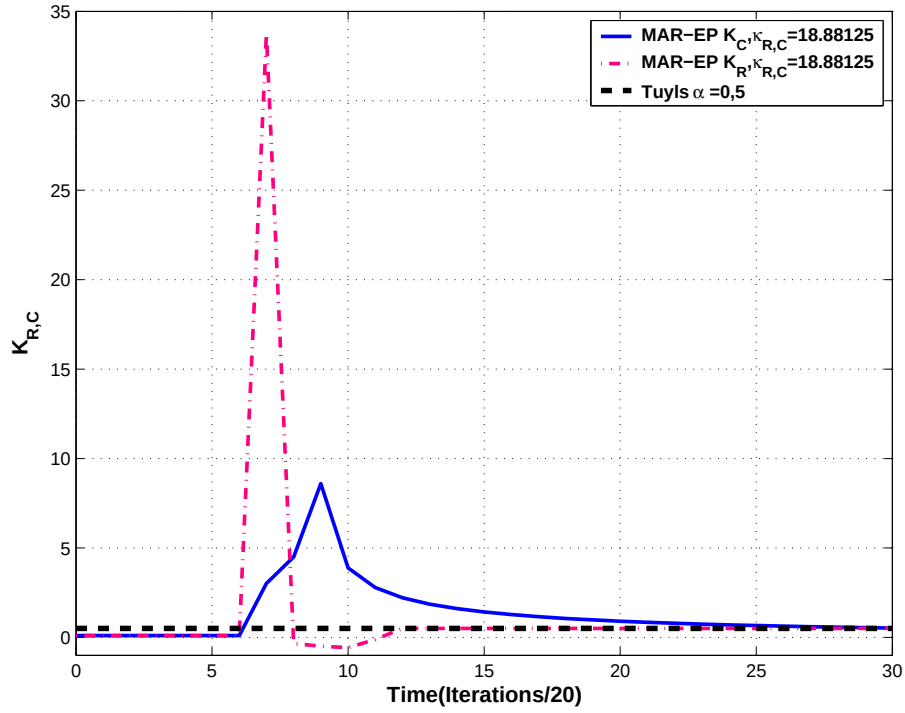


Figura 4.20: Amplitudes de K_R em linha traço ponto, K_C em linha sólida com MAR-EP e de $\alpha = 0,1$ com BPL em linha tracejada, do exemplo numérico 5, jogo ardiloso, com ponto inicial em $(0,1,0,2)$.

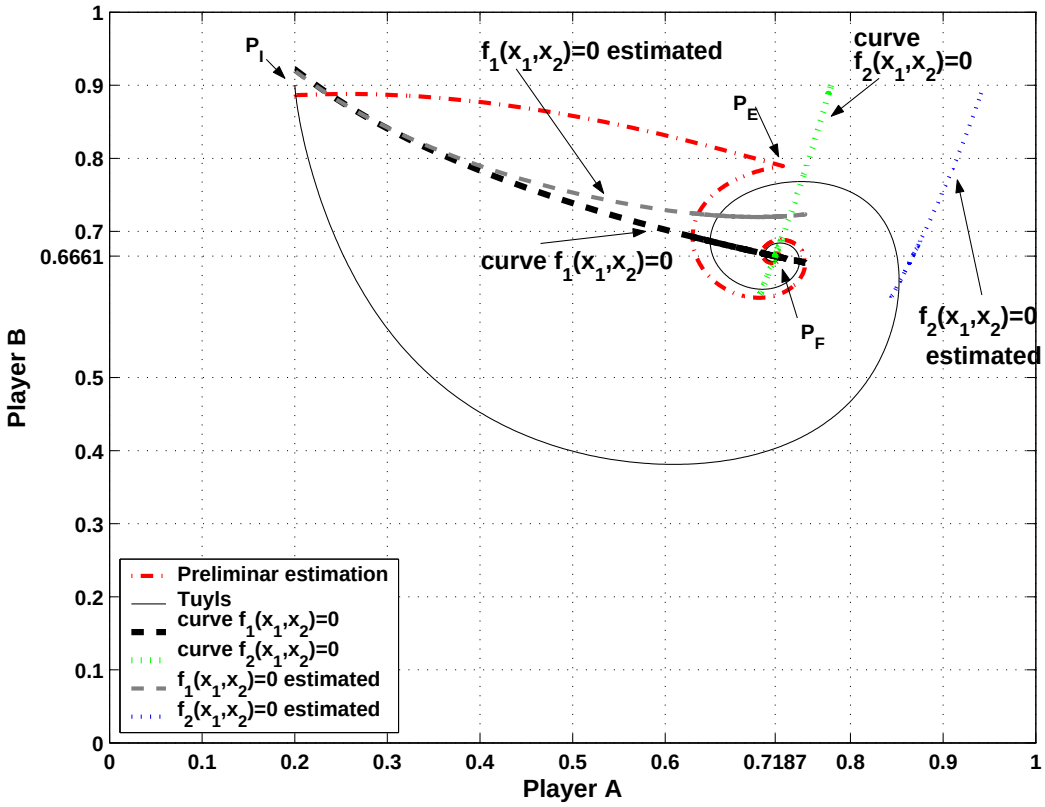


Figura 4.21: Comparação do diagrama de fase de um exemplo de matriz subclasse 3, para a dinâmica com ganho constante α de Tuyls e do MAR-EP. As curvas $f_1(x_1, x_2) = 0$ e $f_2(x_1, x_2) = 0$ se interceptam dando origem ao ponto de equilíbrio P_F.

Capítulo 5

Conclusões: contribuições e trabalhos futuros

5.1 Contribuições desta tese

No mesmo contexto geral de teoria de jogos (SHAMMA e ARSLAN, 2005), a resolução de problemas de teoria de jogos com dois jogadores e duas ações foi proposta e analisada. Mostra-se que o controle chaveado, dependente do estado, pode levar à estabilidade assintótica de um equilíbrio misto. Esta perspectiva de controle chaveado e uma função de Liapunov adequada:

a) explicou de forma sintética e unificada alguns resultados anteriores (BANERJEE *et al.*, 2001; BOWLING e VELOSO, 2002; SINGH *et al.*, 2000; ZHANG *et al.*, 2004), bem como permitiu uma generalização (HEGS), (BHAYA e MACEDO, 2006). Uma nova estratégia chamada de Chaveamento Gradiente Hiperbólico-Elíptico, *Hyperbolic-Elliptic Gradient Switching* (HEGS) foi projetada, similar à estratégia de BANERJEE e PENG (2002), ZHANG e HUANG (2004), mas que conduziu a uma convergência mais rápida das estratégias e das recompensas ao não se valer de apenas ganhos positivos.

b) mostrou que uma outra abordagem, chamada algoritmo de Aprendizado por Política Ponderada (*Weighted Policy Learning*)- WPL proposta na literatura, sem prova de convergência, como superior aos métodos de Aprendizado por Reforço Multiagente Gradiente Ascendente - ARM-GA mencionados no capítulo 2, também é suscetível a uma análise utilizando controle chaveado e funções de Liapunov, após a introdução do conceito de equilíbrios virtuais. Observa-se que a prova de convergência da dinâmica WPL era questão aberta na literatura, anterior a esta tese.

c) além disso, em relação à estimação da frequência empírica (isto é a estratégia do outro jogador), introduz-se técnica de estimação baseada em mínimos quadrados com controle chaveado para propiciar convergência de forma mais rápida sem conhecimento do jogo. É o Método de Aprendizado por Reforço com Estimação Preliminar - MAR-EP. Observa-se que este problema é abordado em SHAMMA e ARSLAN (2005) utilizando derivadores aproximados para as dinâmicas GA e jogo fictício, ao passo que, nesta tese, foi examinada a dinâmica BPL.

5.2 Propostas para continuação de trabalhos futuros

Seguem as propostas para continuação de futuros trabalhos nos seguintes itens abordados na tese.

5.2.1 WPL

a) Cabe posicionar a dinâmica de WPL quanto a sua generalização a jogos de múltiplos jogadores. Outros exemplos ajudariam a corroborar a aplicabilidade do método.

b) Propor novo algoritmo ARM-GA tal que não exista a restrição da condição de contorno apontada na equação (3.61), ou que torne mais rápida a convergência de WPL.

c) Aplicar o conceito de equilíbrios virtuais para projetar outras dinâmicas mais eficientes que o WPL.

5.2.2 MAR-EP

a) Enquadrar MAR-EP como um método gradiente com erro BERTSEKAS e TSITSIKLIS (2000). Esse ajuste deve ser tal que a estimação convirja para um erro menor, tornando mais breve a fase final de α constante positivo de MAR-EP.

b) Propor Método Aprendizado por Reforço com estimação Corrente -(MAR-EC) aonde os parâmetros do método sejam ajustados a cada instante.

c) Aplicar MAR-EP em sistemas com mais de dois jogadores. Investigar a escalabilidade, que é um problema sensível em aprendizado com múltiplos agentes, já que cresce o espaço de busca com o número de agentes, com a complexidade de seus objetivos e com a iteração entre os agentes.

d) Investigar uma versão cooperativa para MAR-EP, inserindo característica de robustez à perda de informação.

Referências Bibliográficas

ABDALLAH, S., LESSER, V. "A Multiagent Reinforcement Learning Algorithm with Non-linear Dynamics", *Journal of Artificial Intelligence Research*, v. 33, n. 1, pp. 521–549, 2008.

AHUJA, R., KUMAR, A., JHA, K. "Exact and heuristic methods for the weapon-target assignment problem", *Technical Report 4464-03, MIT, Sloan School of Management Working Papers*, 2003.

AKELLA, R., KUMAR, P. R. "Optimal control of production rate in a failure-prone manufacturing systems", *IEEE Transactions on Automatic Control*, 1986.

ALTMAN, E., BOULOGNE, T., AZOUZI, R. E. "A survey on networking games in telecommunications", *Computers and Operations Research*, v. 2, n. 33, pp. 286–311, 2005.

ALTMAN, E., SHIMKIN, N. "Individual equilibrium and learning in processor sharing systems", *Operations Research*, , n. 46, pp. 776–784, 1998.

AUER, P., CESA-BIANCHI, N., FREUND, Y. "The nonstochastic multiarmed bandit problem", *SIAM Journal of Computing*, v. 32, n. 1, pp. 48–77, 2002.

AUGUST, J. R., FUDENBERG, D., LEVINE, D. K. "Conditional universal consistency", *Games Econ. Behav.*, v. 29, pp. 104–130, 1999.

BANERJEE, B., SEN, S., PENG, J. "Fast concurrent reinforcement learners". In: *IJCAI'01: Proceedings of the 17th international joint conference on Artificial intelligence*, pp. 825–830, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN: 1-55860-812-5, 978-1-558-60812-2.

BANERJEE, B., PENG, J. "Convergent Gradient Ascent in General-Sum Games", *Proceedings of the Thirteenth European Conference on Machine Learning, Helsinki, Finland*, v. 1, pp. 1–9, 2002.

BANERJEE, B., PENG, J. "Adaptive policy gradient in multiagent learning", *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, v. 1, n. 1, pp. 686–692, 2003. http:/citeseer.ist.psu.edu/banerjee03adaptive.html.

BANERJEE, B., PENG, J. "Generalized multiagent learning with performance bound", *Autonomous Agents and Multi-Agent Systems*, v. 15, n. 3, pp. 281–312, 2007.

BARTO, A. G., SUTTON, R. S., WATKINS, C. J. *Learning and Sequential Decision Making*. Relatório técnico, University of Massachusetts, Amherst, MA, USA, 1989.

BASAR, T. "Relaxation Techniques and asynchronous algorithms for on-line computation of non-cooperative equilibria", *Journal of Economic Dynamics and Control*, v. 11, pp. 531–549, 1987.

BASAR, T. *Control Theory: Twenty-Five Seminal Papers*. New York, Wiley, 2000.

BASAR, T., OLSDER, G. J. *Dynamic Noncooperative Game Theory*. Classics in Applied Mathematics. 2o. ed. New York, SIAM, 1998. ISBN: 0-89871-429-X.

BAZZAN, A. L. C. "A distributed approach for coordination of traffic Signal Agents", *Autonomous Agents and Multi-Agent Systems*, v. 10, n. 3, pp. 131–164, 2005.

BAZZAN, A. L. C. "Opportunities for multiagent systems and multiagent reinforcement learning in traffic control", *Autonomous Agents and Multi-Agent Systems*, v. 18, n. 3, pp. 342–375, 2009.

BEARD, R. W., MCLAIN, T. W., GOODRICH, M. A. "Coordinated target assignment and intercept for unmanned air vehicles", *IEEE Transactions on Robotics and Automation*, v. 18, n. 6, pp. 911–922, 2002.

BENAIM, M., HIRSCH, M. W. "Mixed equilibria and dynamical systems arising from fictitious play in perturbed games", *Games Econ. Behav.*, v. 29, pp. 36–72, 1999.

BERGER, U. "Fictitious play in 2 x n games", *WUSTL, Economics Working Paper Archive*, 2003.

BERTSEKAS, D. P. *Nonlinear programming*. Optimization and computation series. 2o. ed. USA, Athenas Scientific, 1995. ISBN: 1-886529-00-0.

BERTSEKAS, D. P., TSITSIKLIS, J. N. "Gradient Convergence In Gradient Methods With Errors", *SIAM J. Optimization*, v. 10, n. 3, pp. 627–642, 2000.

BHAYA, A., MACEDO, R. B. S. "A Switching Control Approach to Convergence of Gradient Dynamics in General-Sum Games", *Congresso Brasileiro de Automática*, v. 1, pp. 3278–3283, Oct. 2006.

BIANCHI, R. A. C. *Uso de Heurísticas para a Aceleração do Aprendizado por Reforço*. Tese de Mestrado, Escola Politécnica da Universidade de São Paulo, - 2004.

BORGERS, T., SARIN, R. "Learning through reinforcement and replicator dynamics", *Journal of Economic Theory*, v. 77, n. 1, pp. 1–14, 1997.

BOWLING, M., VELOSO, M. "Convergence of Gradient Dynamics with a Variable Learning Rate". In: *Proc. 18th International Conf. on Machine Learning*, pp. 27–34. Morgan Kaufmann, San Francisco, CA, 2001. http:/citeseer.ist.psu.edu/article/bowling01convergence.html.

BOWLING, M., VELOSO, M. "Multiagent learning using a variable learning rate", *Artificial Intelligence*, v. 136, pp. 215–250, 2002.

BOWLING, M. *Multiagent Learning in the Presence of Agents with Limitations*. Tese de Mestrado, Carnegie Mellon University, Maio 2003.

BOWLING, M. "Convergence and no-regret in multiagent learning". In: *In Advances in Neural Information Processing Systems 17*, pp. 209–216. MIT Press, 2005.

BOYD, S., VANDENBERGHE, L. *Convex Optimization*. Cambridge, U.K., Cambridge University Press, 2004.

BROWN, G. *Iterative solutions of games by fictitious play, in Activity Analysis of Production and Allocation*. T. C. Koopmans, Ed. New York: Wiley, 1951.

CAMERER, C. F. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ, Princeton University Press, 2003.

CESA-BIANCHI, N., LUGOSI, G. *Prediction, Learning, and Games*. New York, Cambridge University Press, 2006.

CLAUS, C., BOUTILIER, C. "The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems". In: *In Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pp. 746–752. AAAI Press, 1998.

CONITZER, V., SANDHOLM, T. "AWESOME: A General Multiagent Learning Algorithm that Converges in Self-Play and Learns a Best Response against Stationary Opponents". In: *In Proceedings of the 20th International Conference On Machine Learning*, pp. 83–90, 2006.

CONLISK, J. "Adaptation in games: two solutions to the Crawford puzzle", *J. Econ. Behav. Organ.*, v. 22, pp. 25–50, 1993.

COSTA, M. I. S., KASZKUREWICZ, E., BHAYA, A. "Achieving global convergence to an equilibrium population in predator-prey systems by the use of discontinuous harvesting policy", *Ecological Modelling*, v. 128, pp. 89–99, 2000.

CRAWFORD, V. P. "Learning behavior and mixed strategy Nash equilibria", *J. Econ. Behav. Organ.*, v. 6, pp. 69–78, 1985.

CROSS, J. G. "A stochastic learning model of economic behavior", *Quarterly Journal of Economics*, v. 87, n. -, pp. 239–266, 1973.

ELLISON, G., FUDENBERG, D. "Learning purified mixed equilibria", *J. Econ. Theory*, v. 90, pp. 83–115, 2000.

FOSTER, D. P., VOHRA, R. V. "Calibrated learning and correlated equilibrium", *Games Econ. Behav.*, v. 21, pp. 40–55, 1997.

FOSTER, D. P., YOUNG, H. P. "On the nonconvergence of fictitious play in coordination games", *Games Econ. Behav.*, v. 25, pp. 79–96, 1998.

FUDENBERG, D., KREPS, D. "Learning mixed equilibria", *Games Econ. Behav.*, v. 5, pp. 320–367, 1993.

FUDENBERG, D., LEVINE, D. K. *The Theory of Learning in Games.* Cambridge, MA, MIT Press, 1969.

FUDENBERG, D., TIROLE, J. *Game Theory.* Cambridge MA, The MIT Press, 1991.

GERSHWIN, S. B. *Manufacturing Systems Engineering.* Englewood Cliffs, NJ, Prentice-Hall, 1994.

GREENWALD, A. R. *From Learning to Play Network Games: Does Rationality Yield Nash Equilibrium?* Tese de Mestrado, Department of Computer Science Graduate School of Arts and Science, New York University, may 1999.

HART, S., MAS-COLELL, A. "Uncoupled dynamics do not lead to Nash equilibrium", *Amer. Econ. Rev.*, v. 93, n. 5, pp. 1830–1836, 2003.

HART, S., MAS-COLELL, A. "A simple adaptive procedure leading to correlated equilibrium", *Econometrica*, v. 68, pp. 1127–1150, 2000.

HOFBAUER, J., SIGMUND, K. *Evolutionary Games and Population Dynamics*. United Kingdon, Cambridge University Press, 1998. ISBN: 9780521625708.

HART, S. "Adaptive heuristics", *IEEE Transactions on Circuits and Systems*, v. 73, n. 5, pp. 1401–1430, 2005.

HU, J., WELLMAN, M. P. "Nash Q-Learning for General-Sum Stochastic Games", *Journal Of Machine Learning Research*, v. 4, pp. 1039–1069, 2003.

JAAKKOLA, T., JORDAN, M. I., SINGH, S. P. "On the convergence of stochastic iterative dynamic programming algorithms", *Neural Computation*, v. 6, pp. 1185–1201, 1994.

KAELBLING, L. P., LITTMAN, M. L., MOORE, A. W. "Reinforcement Learning: A Survey", *Journal of Artificial Intelligence Research*, v. 4, n. 4, pp. 237–285, 1996.

KASZKUREWICZ, E., BHAYA, A. "Congestion Control using an improved variant of the AIMD scheme". In: *Proc. of the* 45^{th} *IEEE Conference on Decision and Control*, San Diego, CA, Dec. 2006.

KELLY, F. P. "Charging and rate control for elastic traffic", *European Transactions on Telecommunications*, v. 8, n. 1, pp. 33–37, 1997.

KRISHNA, V., SJÖSTRÖM, T. "On the convergence of fictitious play", *Math. Oper. Res.*, v. 23, n. 2, pp. 479–511, 1998.

LASALLE, J. "Some extensions of Liapunov's second method", *IRE Transactions on Circuit Theory*, v. CT-7, 1960.

LITTMAN, M. L. "Markov games as a framework for multi-agent reinforcement learning". In: *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 157–163, San Francisco, CA,USA, 1994. 2009.

LITTMAN, M. L. "Value-function reinforcement learning in Markov games", *Cognitive Systems Research*, v. 2, n. 1, pp. 55–66, 2001.

MANNOR, S., SHAMMA, J. S. "Multi-agent learning for engineers". In: *Special Issue on Foundations of Multi-Agent Learning*, pp. 417–422, 2007.

MANNOR, S., SHIMKIN, N. "The empirical Bayes envelope and regret minimization in competitive Markov decision processes", *Mathematics of Operations Research*, v. 28, n. 2, pp. 327–345, 2003.

MEZA, M. E. M., BHAYA, A., KASZKUREWICZ, E. "Controller design techniques for the Lotka-Volterra nonlinear system", *Sba: Controle & Automação Sociedade Brasileira de Automática*, v. 16, pp. 124–135, 06 2005. ISSN: 0103-1759.

MIYASAWA, K. *On the convergence of learning processes in a 2 x 2 non-zero-sum two person game*. Relatório Técnico 33, Princeton Univ., Economic Research Program, Princeton, NJ, 1961.

MONDERER, D., SHAPLEY, L. S. "Fictitious play property for games with identical interests", *J. Econ. Theory*, v. 68, pp. 258–265, 1996.

NASH, J. "Equilibrium points in n-person games", *Proceedings of the National Academy of Sciences*, v. 36, n. 1, pp. 48–49, 1950.

NASH, J. "Non-Cooperative Games", *The Annals of Mathematics*, v. 54, n. 2, pp. 286–295, 1951.

NEUMANN, J. V., MORGENSTERN, O. *Theory of Games and Economic Behavior*. Princeton University Press, 1944. ISBN: 0691119937. Disponível em: <`http://jmvidal.cse.sc.edu/library/neumann44a.pdf`>.

OWEN, G. *Game Theory*. United Kingdon, Academic Press, UK, 1995. ISBN: 0-72167-028-8.

PESHKIN, L., KIM, K. E., MEULEAU, N., KAELBLING , L. P. "Learning to Cooperate via Policy Search". pp. 307–314, 2000.

RIBEIRO, C. "Reinforcement Learning Agents", *Artificial Intelligence Review*, v. 17, n. 3, pp. 223–250, 2002.

ROBINSON, J. "An iterative method of solving a game", *Ann. Math.*, v. 54, pp. 296–301, 1951.

ROUGHGARDEN, T. *Selfish Routing and the Price of Anarchy*. Cambridge, MA, MIT Press, 2005.

RUSSELL, S., NORVIG, P. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 2003.

SHAMMA, J., ARSLAN, G. "Dynamic fictitious play, dynamic gradient play, and distributed convergence to Nash equilibria", *Automatic Control, IEEE Transactions on*, v. 50, n. 3, pp. 312 – 327, 2005.

SIGMUND, K. *The Calculus of Selfishness*. New Jersey - USA, Princeton University Press, 2009. ISBN: 978-0691-14275-3.

SINGH, S., KEARNS, M., MANSOUR, Y. "Nash Convergence of Gradient Dynamics in General-Sum Games", *In proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, v. 1, pp. 541–548, 2000.

SUTTON, R. S., BARTO, A. G. *Reinforcement Learning: an Introduction*. Html version ed. Cambridge, MA, MIT Press, 1998. ISBN: 0-262-19398-1.

TUYLS, K., HOEN, P. J., VANSCHOENWINKEL, B. "An Evolutionary Dynamical Analysis of Multi-Agent Learning in Iterated Games", *Autonomous Agents and Multi-agents Systems*, v. 12, n. 1, pp. 115–153, 2006. ISSN: 1387-2532. doi: http://dx.doi.org/10.1007/s10458-005-3783-9.

UTKIN, V. I. "Variable Structure Systems with Sliding Modes", *IEEE Trans. Automat. Control*, v. AC-22, n. 2, pp. 212–222, April 1977.

VASIN, A. "On stability of mixed equilibria", *Nonlinear Anal*, v. 38, pp. 793–802, 1999.

VEGA-REDONDO, F. *Economics and the theory of games*. United Kingdon, Cambridge University Press, 2003. ISBN: 0-52177590-6.

WANG, X., SANDHOLM, T. "Reinforcement Learning to Play an Optimal Nash Equilibrium in Team Markov Games". In: *in Advances in Neural Information Processing Systems*, pp. 1571–1578. MIT Press, 2002.

WATKINS, C. J. C. H. "Q-Learning", *Machine Learning*, v. 8, n. 3-4, pp. 279–292, 1992.

WEIBULL, J. W. *Evolutionary Game Theory*. Cambridge, MA, MIT Press, 1995.

ZHANG, H., HUANG, S. "Convergent gradient ascent with momentum in general-sum games", *Neurocomputing*, v. 61, pp. 449–454, 2004.

ZHANG, Y., KANG, S.-R., LOGUINOV, D. "Delayed stability and performance of distributed congestion control". In: *SIGCOMM'04*, pp. 307–318, Portland, OR, 2004. ACM.

ZINKEVICH, M. "Online Convex Programming and Generalized Infinitesimal Gradient Ascent". pp. 928–936, 2003.

YOUNG, H. P. *Strategic Learning and its Limits*. Oxford, Oxford University Press, 2006.

Y. SHOHAM, R. P., GRENAGER, T. "If multi-agent learning is the answer, what is the question?" *Artificial Intelligence*, 2007.