PROVABLE REINFORCEMENT LEARNING FOR
CONSTRAINED AND MULTI-AGENT CONTROL SYSTEMS

by

Dongsheng Ding

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ELECTRICAL ENGINEERING)

December 2022

Copyright 2022 Dongsheng Ding
Dedication

To my family.
Acknowledgements

First and foremost, I would like to express my deep and sincere gratitude to my advisor, Professor Mihailo R. Jovanović, for his generous guidance during my graduate study. Every time of meeting with him was a lesson, and I am grateful for all the knowledge and experience he has shared with me. His commitment to excellence and open-mindedness encouraged me to perfect my research and have big-picture thinking. I will always be grateful for all his advice and support, which always inspire me to be a better control engineer.

I was fortunate to have the opportunity to work with many excellent researchers, including Dr. Xiaohan Wei, Professor Zhuoran Yang, Professor Zhaoran Wang, Dr. Kaiqing Zhang, Dr. Chen-Yu Wei, and Professor Bin Hu. Their creativity and patience have made our collaboration joyful. They generously offered me an abundant amount of time for discussing and shaping several exciting research problems presented in this dissertation.

I am also privileged to collaborate with Professor Tamer Başar. I would like to thank him for his precious advice and generous support. His amazing enthusiasm and vision have always inspired me to work hard.

My gratitude extends to Professors Ashutosh Nayyar and Meisam Razaviyayn, for serving on my defense/qualifying exam committees; Professors Pierluigi Nuzzo and Mahdi Soltanolkotabi, for serving on my qualifying exam committee. I have benefitted from their invaluable comments.
and feedback. I am also grateful to many excellent teachers who have guided my graduate study, including Professors Zhi-Quan Luo, Andrew Lamperski, Yousef Saad, Peter Seiler, Georgios Giani- nakis, Arindam Banerjee, and Soheil Mohajer at the University of Minnesota Twin Cities; Profes- sors Bart Kosko, Meisam Razaviyayn, Mahdi Soltanolkotabi, Igor Kukavica, Jason D. Lee, Haipeng Luo, Steven M. Heilman, Sergey Lototsky, Andrew Manion, and Rahul Jain at the University of Southern California. Additionally, I would like to thank Anastassia Tzoytzyrakos, Lucienne Aarsen, and Elizabeth Fife, for teaching me English communication.

It is an amazing journey for me to start my doctoral program at the University of Minnesota Twin Cities and finish it at the University of Southern California. I am very grateful to have my past and current labmates, Dr. Xiaofan Wu, Professor Armin Zare, Dr. Neil Dhingra, Professor Yongxin Chen, Dr. Hamza Farooq, Dr. Sepideh Hassan-Moghaddam, Dr. Saakar Byahut, Dr. Wei Ran, Dr. Anubhav Dwivedi, Hesameddin Mohammadi, Samantha Samuelson, Emily Reed, Zalan Fabian, Jin Zhou, Milad Pooladsanj, Ibrahim Ozaslan, for making my graduate study more en- joyable. Moreover, I would like to thank many friends I met in this journey for the accompany, including Dr. Diqing Su, Dr. Jianjun Yuan, Dr. Huaijin Yao, Wen Zhou, Dr. Rui Ma, Dr. Shreyas Bhaban, and Dr. Bhaskar Sen in Minneapolis; Dr. Huimei Cheng, Dr. Sarah Cooney, Shushan Arakelyan, Nripsuta Ani Saxena, Qinyi Luo, Yifang Chen, Dr. Qian Yu, Dr. Yihang Zhang, Dr. Jiali Duan, Dr. Bin Wang, Dr. Nathan Dahlin, Dr. Yeji Shen, Dr. Buyun Chen, Dr. Seyed Moham- madreza Mousavi Kalan, Dr. Mukul Gagrani, Dr. Filipe Vital, Renyan Sun, Panagiotis Kyriakis, Yijun Liu, Yunsong Liu, Jayson Sia, Krishna Chaitanya Kalagarla, Tiancheng Jin, Fernando V. Monteiro, Mengxiao Zhang, Chung-Wei Lee, and Sagar Sudhakara in Los Angeles. Also, I would like to thank my roommates, Dr. Yaobin Qin, Dr. Jinfeng Yang, Jian Shi, Xuansheng Wang, Tianyi Chen, Gele Qin, and Dr. Xin Bai, for the help in my daily life.
Finally, my deepest gratitude goes to my granduncle, parents, sisters, and cousins, for their endless love and support. My granduncle and father have taught me to be independent and pursue my dreams; my mother has taught me to appreciate every moment in life; my sisters and cousins have always encouraged me to believe in myself. I am the most fortunate guy in the world to have you as my extended family, and I am forever in your debt.
Table of Contents

Dedication ........................................................................................................ ii
Acknowledgements ......................................................................................... iii
List of Figures ................................................................................................. x
Abstract ........................................................................................................... xiii

1 Introduction ................................................................................................. 1
  1.1 Policy gradient methods for constrained MDPs .................................. 4
  1.2 Provably efficient RL for constrained MDPs ...................................... 6
  1.3 Temporal-difference learning with linear function approximation ...... 7
  1.4 Markov potential games and independent learning ......................... 9
  1.5 Organization of the dissertation ....................................................... 10
  1.6 Contributions of the dissertation ....................................................... 12

I Reinforcement learning for constrained control systems ............... 16

2 Natural policy gradient primal-dual method for constrained MDPs .... 17
  2.1 Introduction ....................................................................................... 18
  2.2 Problem setup .................................................................................... 19
    2.2.1 Constrained Markov decision processes .................................... 20
    2.2.2 Method of Lagrange multipliers ................................................. 22
    2.2.3 Policy parametrization ............................................................... 24
  2.3 Natural policy gradient primal-dual method ...................................... 27
    2.3.1 Constrained policy optimization methods ................................ 27
    2.3.2 Natural policy gradient primal-dual method .............................. 32
  2.4 Tabular softmax policy case ............................................................. 34
    2.4.1 Non-asymptotic convergence analysis ...................................... 36
  2.5 Function approximation case ............................................................ 44
    2.5.1 Log-linear policy class ............................................................. 46
    2.5.2 Non-asymptotic convergence analysis ...................................... 51
    2.5.3 General smooth policy class ..................................................... 59
  2.6 Sample-based algorithms .................................................................. 63
5.4 Independent policy gradient methods ............................................. 163
  5.4.1 Policy gradient for Markov potential games ........................... 163
  5.4.2 Faster rates for Markov cooperative games ............................... 167
5.5 Nash regret analysis ................................................................. 168
  5.5.1 Setting up the analysis ....................................................... 168
  5.5.2 Nash regret analysis for Markov potential games .................... 170
  5.5.3 Nash regret analysis for Markov cooperative games ................. 177
5.6 Independent policy gradient with function approximation ............... 180
5.7 Game-agnostic convergence ...................................................... 185
5.8 Computational experiments ...................................................... 189
5.9 Concluding remarks .................................................................. 193

6 Discussion and future directions ................................................ 195
  6.1 Policy gradient primal-dual algorithms ...................................... 195
  6.2 Provably efficient RL for constrained MDPs ............................... 196
  6.3 Multi-agent policy evaluation in other settings ........................... 197
  6.4 Multi-agent policy gradient methods ........................................ 198

Bibliography .................................................................................... 199

Appendices ....................................................................................... 225

A Supporting proofs in Chapter 2 ...................................................... 225
  A.1 Proof of Lemma 1 ................................................................. 225
  A.2 Proof of Lemma 2 ................................................................. 225
  A.3 Proof of Lemma 3 ................................................................. 226
  A.4 Proof of Theorem 4 ............................................................... 227
  A.5 Proof of Lemma 5 ................................................................. 232
  A.6 Sample-based algorithm with function approximation .......... 234
  A.7 Proof of Theorem 13 ............................................................. 236
  A.8 Proof of Theorem 14 ............................................................. 242

B Supporting proofs in Chapter 3 ...................................................... 245
  B.1 Proof of Theorem 23 ............................................................. 245
  B.2 Proof of Formulas (3.9) and (3.11) ....................................... 248
  B.3 Proof of Formulas (3.14) and (3.15) .................................... 249
     B.3.1 Proof of Formula (3.16) .................................................. 250
     B.3.2 Proof of Formula (B.1) .................................................... 253
  B.4 Supporting lemmas from optimization .................................... 256
  B.5 Other supporting lemmas ...................................................... 258

C Supporting proofs in Chapter 4 ...................................................... 261
  C.1 Proof of Lemma 25 .............................................................. 261
  C.2 Martingale concentration bound ............................................ 262
  C.3 Proof of Lemma 27 .............................................................. 263

viii
C.4 Proof of Lemma 28 ................................................. 263

D Supporting proofs in Chapter 5 ............................................. 265
  D.1 Proof of Lemma 34 ....................................................... 265
  D.2 Proof of Lemma 35 ....................................................... 266
  D.3 Proof of Lemma 36 ....................................................... 267
  D.4 Proof of Lemma 37 ....................................................... 268
  D.5 Proofs for Section 5.6 .................................................... 272
    D.5.1 Unbiased estimate .................................................. 272
    D.5.2 Proof of Theorem 40 ............................................... 273
    D.5.3 Proof of Theorem 41 ............................................... 279
    D.5.4 Sample complexity ................................................. 282
  D.6 Proofs for Section 5.7 .................................................... 283
    D.6.1 Proof of Theorem 42 ............................................... 283
    D.6.2 Proof of Theorem 43 ............................................... 290
  D.7 Other auxiliary lemmas ................................................ 299
    D.7.1 Auxiliary lemma for potential functions ....................... 299
    D.7.2 Auxiliary lemma for single-player MDPs ....................... 300
    D.7.3 Auxiliary lemma for multi-player MDPs ....................... 301
    D.7.4 Auxiliary lemma for stochastic projected gradient descent .... 302
List of Figures

1.1 Empirical success of RL in two representative applications: (a) playing the game of Go [203] and (b) solving Rubik’s cube with a robot hand [10]. ........................................ 2

1.2 Diagram of controller-system interaction. In addition to reward, other learning objectives are critical as constraints, e.g., energy cost [94], traffic delay [11], and infections [253]. .............................................................. 3

1.3 Diagram of multiple controllers-system interactions. More than one controller participates in multi-agent systems, e.g., computer games [229], swarm robotics [105], and financial management [283]. .................................................. 3

2.1 An example of a constrained MDP for which the objective function $V_{r,g}^\pi(s)$ in Problem (2.3) is not concave and the constraint set $\{ \theta \in \Theta \mid V_{r,g}^\pi(s) \geq b \}$ is not convex. The pair $(r,g)$ associated with the directed arrow represents (reward, utility) received when an action at a certain state is taken. This example is utilized in the proof of Lemma 3. ............................................................. 25

2.2 Learning curves of NPG-PD (—) and FOCOPS [280] (—) for Hopper-v3, Swimmer-v3, and Half Cheetah-v3 robotic tasks with the respective speed limits 82.748, 24.516, and 151.989. The vertical axes represent the average reward and the average cost (i.e., average speed). The solid lines show the means of 1000 bootstrap samples obtained over 5 random seeds and the shaded regions display the bootstrap 95% confidence intervals. ................................................................. 70

2.3 Learning curves of NPG-PD (—) and FOCOPS [280] (—) for Walker2d-v3, Ant-v3, and Humanoid-v3 robotic tasks with the respective speed limits 81.886, 103.115, and 20.140. The vertical axes represent the average reward and the average cost (i.e., average speed). The solid lines show the means of 1000 bootstrap samples obtained over 5 random seeds and the shaded regions display the bootstrap 95% confidence intervals. ................................................................. 71
2.4 Learning curves of NPG-PD (—) and FOCOPS [280] (—) for Humanoid Circle-v0 and Ant Circle-v0 robotic task. The horizontal axis represents the number of dual updates. The average cost is constrained to go below 50. The vertical axes represent the average reward and the average cost (i.e., average speed). The solid lines show the means of 1000 bootstrap samples obtained over 5 random seeds and the shaded regions display the bootstrap 95% confidence intervals.

4.1 A system with six agents that communicate over a connected undirected network. Each agent interacts with the environment by receiving a private reward and taking a local action.

4.2 Performance comparison for the centralized problem with \( N = 1 \). Our algorithm with stepsize \( \eta = 0.1 \) and \( K = 4 \) (—) achieves a smaller optimality gap than stochastic primal-dual algorithm with stepsize: \( \eta = 0.1 \) (---), \( \eta = 0.05 \) (---), \( \eta = 0.025 \) (---), and \( \eta = 1/\sqrt{t} \) (—). It also provides a smaller optimality gap than the approach that utilizes pre-collected i.i.d. samples in a buffer (—).

4.3 Performance comparison for the distributed case with \( N = 10 \). Our algorithm with stepsize \( \eta = 0.05 \) and \( K = 4 \) (—) achieves a smaller optimality gap than stochastic primal-dual algorithm with stepsize: \( \eta = 0.05 \) (---), \( \eta = 0.025 \) (---), \( \eta = 0.0125 \) (---), and \( \eta = 1/\sqrt{t} \) (—). It also provides a smaller optimality gap than the approach that utilizes pre-collected iid samples in a buffer (—).

4.4 Performance comparison for the distributed case with \( N = 5 \). Our algorithm with stepsize \( \eta = 0.1 \) and \( K = 3 \) (—) achieves a smaller optimality gap than stochastic primal-dual algorithm with stepsize: \( \eta = 0.5 \) (---), \( \eta = 0.25 \) (---), \( \eta = 0.125 \) (---), and \( \eta = 1/\sqrt{t} \) (—). It also provides a smaller optimality gap than the approach that utilizes pre-collected iid samples in a buffer (—).

5.1 Convergence performance. (a) Learning curves for our independent policy gradient (—) with stepsize \( \eta = 0.001 \) and the projected stochastic gradient ascent (—) with \( \eta = 0.0001 [131] \). Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval. (b) Learning curves for six individual runs of our independent policy gradient (solid line) and the projected stochastic gradient ascent (dash line) three each. (c) Distribution of players in one of two states taking four actions. In (a) and (b), we measure the accuracy by the absolute distance of each iterate to the converged Nash policy, i.e., \( \frac{1}{N} \sum_{i=1}^{N} \| \pi_i^{(t)} - \pi_i^{Nash} \|_1 \). In our computational experiments, the initial distribution \( \rho \) is uniform.
5.2 Convergence performance. (a) Learning curves for our independent policy gradient (—) with stepsize $\eta = 0.002$ and the projected stochastic gradient ascent (—) with $\eta = 0.0001$ [131]. Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval. (b) Learning curves for six individual runs of our independent policy gradient (solid line) and the projected stochastic gradient ascent (dash line) three each. (c) Distribution of players in one of two states taking four actions. In (a) and (b), we measure the accuracy by the absolute distance of each iterate to the converged Nash policy, i.e., $\frac{1}{N} \sum_{i=1}^{N} \| \pi_i(t) - \pi_i^{Nash} \|_1$. In our computational experiments, the initial distribution $\rho$ is uniform.

5.3 Convergence performance. (a) Learning curves for our independent policy gradient (—) with stepsize $\eta = 0.001$ and the projected stochastic gradient ascent (—) with $\eta = 0.0001$ [131]. Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval. (b) Learning curves for six individual runs of our independent policy gradient (solid line) and the projected stochastic gradient ascent (dash line) three each. (c) Distribution of players in one of two states taking four actions. In (a) and (b), we measure the accuracy by the absolute distance of each iterate to the converged Nash policy, i.e., $\frac{1}{N} \sum_{i=1}^{N} \| \pi_i(t) - \pi_i^{Nash} \|_1$. In our computational experiments, the initial distribution is nearly degenerate $\rho = (0.9999, 0.0001)$.

5.4 Convergence performance. (a) Learning curves for our independent policy gradient (—) with stepsize $\eta = 0.001$ and the projected stochastic gradient ascent (—) with $\eta = 0.0001$ [131]. Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval. (b) Learning curves for six individual runs of our independent policy gradient (solid line) and the projected stochastic gradient ascent (dash line) three each. (c) Distribution of players in one of two states taking four actions. In (a) and (b), we measure the accuracy by the absolute distance of each iterate to the converged Nash policy, i.e., $\frac{1}{N} \sum_{i=1}^{N} \| \pi_i(t) - \pi_i^{Nash} \|_1$. In our computational experiments, the initial distribution is nearly degenerate $\rho = (0.0001, 0.9999)$. 

xii
Abstract

Reinforcement learning (RL) has proven its value through great empirical success in many artificial sequential decision-making control systems, but uptake in complex real-world systems has been much slower. A wide gap between standard RL setups and realities often results from constraints and multiple agents in real-world systems. To reduce this gap, we develop effective RL algorithms for two types of real-world stochastic control systems: constrained systems and multi-agent systems to search for better control policies, with theoretical performance guarantees.

Part I of the dissertation is devoted to RL for constrained control systems. We study two settings of sequential decision-making control problems described by constrained Markov decision processes (MDPs) in which a controller (or an agent) aims at satisfying a constraint in addition to maximizing the standard reward objective. In the simulation setting, we propose a direct policy search method for infinite-horizon constrained MDPs: natural policy gradient primal-dual method, which updates the primal policy via natural policy gradient ascent and the dual variable via projected sub-gradient descent. We establish a global convergence theory for our method using softmax, log-linear, and general smooth policy parametrizations, and demonstrate finite-sample complexity guarantees for two model-free extensions of our method. In the online episodic setting, we propose an online policy optimization method for episodic finite-horizon
constrained MDPs: optimistic primal-dual proximal policy optimization, where we effectuate safe exploration through the upper-confidence bound optimism and address constraints via the primal-dual optimization. We establish sublinear regret and constraint violation bounds that depend on the size of the state-action space only through the dimension of the feature mapping, making our results hold even when the number of states goes to infinity.

Part II of the dissertation is devoted to RL for multi-agent control systems. We study two setups of multi-agent sequential decision-making control problems modeled by multi-agent MDPs in which multiple agents aim at maximizing their reward objectives. In the cooperative setup, we propose an online distributed temporal-difference learning algorithm for solving the classical policy evaluation problem with networked agents. Our algorithm works as a true stochastic primal-dual update using online Markovian samples and homotopy-based adaptive stepsizes. We establish optimal finite-time error bound with a sharp dependence on the network size and topology. In the cooperative/competitive setup, we propose a new independent policy gradient method for learning a Nash policy of Markov potential games. We establish sublinear Nash regret bounds that are free of explicit dependence on the state space size, enabling our method to work for problems with large size of state space and a large number of players. We demonstrate finite-sample complexity guarantees for a model-free extension of our method in the function approximation setting. Moreover, we identify a class of independent policy gradient methods that enjoys last-iterate convergence and sublinear Nash regret bound for learning both zero-sum Markov games and Markov cooperative games.
Chapter 1

Introduction

Reinforcement Learning (RL) is an algorithmic paradigm for sequential decision-making in which a controller (or an agent) aims to maximize the task-associated long-term reward by interacting with an unknown system (or environment) over time to learn a good control policy. In recent years, RL has achieved remarkable empirical success in a large set of simulated systems such as playing computer games and manipulating robotics; see Figure 1.1. However, it is challenging to extend such success to real-world applications by directly applying existing RL methods. In this thesis, we grow our current RL progress for real-world applications by establishing algorithmic solutions for two types of stochastic control systems: constrained systems and multi-agent systems, from a mainly theoretical point of view.

In many real-world RL tasks, it is not sufficient for the agent to only maximize the long-term reward associated with the single learning objective. The control system is also subject to constraints on its utilities/costs in many safety-critical applications, e.g., in autonomous driving, robotics, cyber-security, and financial management. Application of standard RL techniques for such constrained systems stimulates an active line of research on constrained RL: in addition to maximizing the long-term reward, it is also critical to take into account the (safety) constraint on
Figure 1.1: Empirical success of RL in two representative applications: (a) playing the game of Go [203] and (b) solving Rubik’s cube with a robot hand [10].

the long-term utility/cost as an extra learning objective; see Figure 1.2. Along this line of research, we focus on a fundamental class of constrained control systems in a model of constrained Markov Decision Processes (MDPs). Part I of the dissertation is devoted to the direct policy search method for constrained MDPs in two fundamental RL setups: with or without policy simulators. We are interested in exploiting the structure of RL objective to design RL algorithms with provably performance guarantees. Two topics we investigate are given as follows.

- Natural policy gradient primal-dual methods for constrained MDPs.
- Provably efficient policy optimization for constrained MDPs.

Not just a controller (or an agent), many successful RL applications involve the participation of more than an agent, e.g., computer games, swarm robotics, and financial management. Leveraging standard RL techniques for such multi-agent systems encourages a rich line of research on multi-agent RL: multiple agents operate in a common system and each of them aims to maximize its long-term reward by interacting with the unknown system and other agents; see Figure 1.3.
In this line of research, we are interested in two multi-agent RL tasks: the policy evaluation problem of multi-agent MDPs and the independent direct policy search method for Markov potential games. Part II of the dissertation is devoted to the following two topics.

- Multi-agent temporal difference learning for multi-agent MDPs.
- Independent policy gradient methods for Markov potential games.
Having stated two lines of research above, a critical capability we aim to develop for our RL methods is that they can work in large state spaces with function approximation. With this attribute, we support our RL methods by theoretical convergence guarantees that result from a solid integration of control, optimization, statistics, and game theory.

The remainder of this introduction is organized as follows. We briefly overview policy gradient methods and their applications for constrained MDPs in Section 1.1. We discuss provably efficient RL algorithms for constrained MDPs in Section 1.2. We overview temporal-difference learning with linear function approximation in Section 1.3. We discuss Markov potential games and independent learning in Section 1.4. In Section 1.5, we provide an outline of the dissertation. Finally, we summarize the main contributions of the dissertation in Section 1.6.

1.1 Policy gradient methods for constrained MDPs

Policy gradient methods lie at the heart of the empirical success of RL, which motivates a rich line of global convergence results. In [87, 152, 170, 167, 168, 169], the authors provided global convergence guarantees and quantified sample complexity of (natural) policy gradient methods for nonconvex linear quadratic regulator problem of both discrete- and continuous-time systems. In [276], the authors showed that locally optimal policies for MDPs are achievable using policy gradient methods with reward reshaping. It was demonstrated in [233] that (natural) policy gradient methods converge to the globally optimal value when overparametrized neural networks are used. A variant of natural policy gradient, trust-region policy optimization (TRPO) [195], converges to the globally optimal policy with overparametrized neural networks [139] and for
regularized MDPs [199]. In [32, 33], the authors studied global optimality and convergence of policy gradient methods from a policy iteration perspective. In [8], the authors characterized global convergence properties of (natural) policy gradient methods and studied computational, approximation, and sample size issues. Additional recent advances along these lines include [159, 271, 50, 142, 75, 117, 247]. While all these references handle a lack of convexity in the objective function, we make additional effort to deal with nonconvex constraints that arise in optimal control of constrained MDPs.

In many constrained MDP algorithms [11, 2, 1, 37], Lagrangian-based policy gradient methods have been widely used to address constraints. However, convergence guarantees of these algorithms are either local (to stationary-point or locally optimal policies) [35, 56, 220] or asymptotic [37]. When function approximation is used for policy parametrization, [264] recognized the lack of convexity and showed asymptotic convergence (to a stationary point) of a method based on successive convex relaxations. In [179], the authors provided duality analysis for constrained MDPs in the policy space and proposed a provably convergent dual descent algorithm by assuming access to a nonconvex optimization oracle. However, it is not clear how to obtain the solution to a primal nonconvex optimization problem and the global convergence guarantees are not established. In [180], the authors proposed a primal-dual algorithm and provided computational experiments but did not offer any convergence analysis. This motivates us to establish non-asymptotic convergence of Lagrangian-based policy gradient methods to a globally optimal solution. Other related Lagrangian-based policy optimization methods include CPG [227], accelerated PDPO [135], CPO [7, 259], FOCOPS [280], IPPO [143], P3O [201], and CUP [257] but theoretical guarantees for these algorithms are still lacking.
1.2 Provably efficient RL for constrained MDPs

Provably efficient RL algorithms have shown the power of function approximation to achieve the
statistical efficiency through the tradeoff between exploration and exploration. Using the optim-
ism in the face of uncertainty [17, 45], [256, 255, 110, 47, 267] addressed the exploration and
exploitation trade-off by adding the Upper Confidence Bound (UCB) bonus, and proposed algo-
rijthms are provably sample-efficient. In [47], optimism has been combined with policy-based RL:
an optimistic proximal policy optimization with UCB exploration. However, all these references
only studied some particular MDPs in unconstrained RL. This motivates us to design an optimistic
variant of proximal policy optimization for constrained MDPs. For the large constrained MDPs
with unknown transition models, there is a line of literature that is related to the policy optimiza-
tion under constraints, e.g., [227, 7, 259, 220, 143, 280, 201, 257]. However, the exploration under
constraints is less studied and their theoretical guarantees are unknown. The present work fills
in this gap in the linear function approximation setting.

The study of provably efficient RL algorithms for constrained MDPs has received growing at-
tention, especially those on learning constrained MDPs with unknown transitions and rewards.
Most of them are model-based and only apply to finite state-action spaces. [204, 85] leveraged up-
per confidence bound (UCB) on fixed reward, utility, and transition probability to propose sample-
efficient algorithms for tabular constrained MDPs; [204] established an $\tilde{O}(\sqrt{|A|T^{1.5}\log T})$ regret
and constraint violation via linear program in the average-cost case in time $T$; [85] achieved an $\tilde{O}(|S|\sqrt{H^3T})$ regret and constraint violation in the episodic setting via linear program and
primal-dual policy optimization, where $|S|$ is the size of state space, $|A|$ is the size of action space,
and $H$ is the horizon of episode. In [187], the authors studied an adversarial stochastic shortest path problem under constraints and unknown transitions with $\tilde{O}(|S|\sqrt{|A|H^2T})$ regret and constraint violation. [21] extended Q-learning with optimism for finite state-action constrained MDPs with peak constraints. [42] proposed UCB-based convex planning for episodic tabular constrained MDPs in dealing with convex or hard constraints. [115, 98] established probably approximately correct (PAC) guarantees that enjoy better problem-dependent sample-complexity. In contrast, our proposed algorithm can potentially apply to scenarios with infinite state-space, and our provided sublinear regret and constraint violation bounds only depend on the implicit dimension instead of the true dimension of the state space. Compared to more recent references [73, 52, 271, 251, 201, 257], we attack the exploration directly and does not rely on any policy simulators (or generative models).

1.3 Temporal-difference learning with linear function approximation

Temporal-difference (TD) learning with linear function approximators is a popular approach for estimating the value function for an agent that follows a particular policy. The asymptotic convergence of the original TD method, which is known as TD(0), and its extension TD($\lambda$) was established in [225]. In spite of their wide-spread use, these algorithms can become unstable and convergence cannot be guaranteed in off-policy learning scenarios [26, 218]. To ensure stability, batch methods, e.g., Least-Squares Temporal-Difference (LSTD) learning [41], have been proposed at the expense of increased computational complexity. To achieve both stability and low computational cost, a class of gradient-based temporal-difference (GTD) algorithms [215, 216],
e.g., GTD, GTD2, and TDC, were proposed and their asymptotic convergence in off-policy settings was established by analyzing certain ordinary differential equations (ODEs) [38]. However, these are not true stochastic gradient methods with respect to the original objective functions [218], because the underlying TD objectives, e.g., mean square Bellman error (MSBE) in GTD or mean square projected Bellman error (MSPBE) in GTD2, involve products and inverses of expectations. As such, these cannot be sampled directly and it is difficult to analyze efficiency with respect to the TD objectives.

The finite-time or finite-sample performance analysis of TD algorithms is critically important in applications with limited time budgets and finite amount of data. Early results are based on the stochastic approximation approach under i.i.d. sampling. For TD(0) and GTD, $O(1/T^\alpha)$ error bound with $\alpha \in (0, 0.5)$ was established in [61, 62] and an improved $O(1/T)$ bound was provided in [122], where $T$ is the total number of iterations. In the Markov setting, $O(1/T)$ bound was established for TD(0) that involves a projection step [34]; for a linear stochastic approximation algorithm driven by Markovian noise [207]; and for an on-policy TD algorithm known as SARSA [286]. Recent work [102] provided complementary analysis of TD algorithms using the Markov jump linear system theory. To enable both on- and off-policy implementation, an optimization-based approach [138] was used to cast MSBPE into a convex-concave objective that allows the use of stochastic gradient algorithms [175] with $O(1/\sqrt{T})$ bound for i.i.d. samples; an extension of this approach to the Markov setting was provided in [238]. In [222], the finite-time error bound of GTD was improved to $O(1/T)$ for i.i.d. sampled data but it remains unclear how to extend these results to multi-agent scenarios where data is not only Markovian but also distributed over a network. The present work fills in this gap by proposing a new online distributed TD learning algorithm that operates online Markovian samples.
1.4 Markov potential games and independent learning

In stochastic optimal control, the Markov potential games (MPGs) model dates back to \[66, 95\]. More recent studies include \[268, 157, 147, 160\] and all of these studies focus on systems with known dynamics. MPGs have also attracted attention in multi-agent RL. In the infinite-horizon setting, \[131, 277\] extended the policy gradient method \[8, 112\] for multiple players and established the iteration/sample complexity that scales with the size of state space; \[92\] generalized the natural policy gradient method \[112, 8\] and established the global asymptotic convergence. In the finite-horizon setting, \[205\] built on the single-agent Nash-VI \[140\] to propose a sample efficient turn-based algorithm and \[153\] studied the policy gradient method. Earlier, \[234, 144\] studied Markov cooperative games and \[118, 178, 57\] studied one-state MPGs; both of these are special cases of MPGs. We note that the term: Markov potential game has also been used to refer to state-based potential MDPs \[154, 161\], which are different from the MPGs that we study; see counterexamples in \[131\].

Despite recent advances on the theory of policy gradient \[32, 8\], the theory of policy gradient methods for multi-agent RL is relatively less studied. In the basic two-player zero-sum Markov games, \[274, 44, 64, 282\] established global convergence guarantees for policy gradient methods for learning an (approximate) Nash equilibrium. More recently, \[49, 239\] examined variants of policy gradient methods and provided last-iterate convergence guarantees. However, it is much harder for the policy gradient methods to work in general Markov games \[158, 97\]. The effectiveness of (natural) policy gradient methods for tabular MPGs was demonstrated in \[131, 277, 92, 278\]. Moreover, \[249, 235, 262, 181\] reported impressive empirical performance of multi-agent
policy gradient methods with function approximation in cooperative Markov games, but the theoretical foundation has not been provided, which motivates the present work.

Independent learning recently received attention in multi-agent RL [64, 273, 177, 193, 111, 205, 116], because it only requires local information for learning and naturally yields algorithms that scale to a large number of players. The algorithms in [131, 277, 92, 278] can also be generally categorized as independent learning algorithms for MPGs. In addition, we establish a new independent learning method for MPGs with large size of state space and number of players.

Being game-agnostic is a desirable property for independent learning in which players are oblivious to the types of games being played. In particular, classical fictitious-play warrants average-iterate convergence for several games [189, 172, 100]. Although online learning algorithms, e.g., the one based on multiplicative weight updates (MWU) [51], offer average-iterate convergence in zero-sum matrix games, they often do not provide last-iterate convergence guarantees [25], which motivates recent studies [65, 171, 241]. Interestingly, while MWU converges in last-iterate for potential games [178, 57], this is not the case for zero-sum matrix games [53]. Recently, [130, 129] established last-iterate convergence of $Q$-learning dynamics for both zero-sum and potential/cooperative matrix games. However, it is open question whether an algorithm can have last-iterate convergence for both zero-sum and potential/cooperative Markov games. We provide the first affirmative answer to this question.

1.5 Organization of the dissertation

The dissertation contains two parts. In each part, we investigate two problem setups in two chapters, respectively. Part I of the dissertation is devoted to reinforcement learning for constrained
control systems: Chapter 2 is based on the joint work with Kaiqing Zhang, Jiali Duan, and Tamer Başar [73, 69, 68], and Chapter 3 is based on the joint work with Xiaohan Wei, Zhuoran Yang, and Zhaoran Wang [74]. Part II of the dissertation is devoted to reinforcement learning for multi-agent control systems: Chapter 4 is based on the joint work with Xiaohan Wei, Zhuoran Yang, and Zhaoran Wang [70, 71], and Chapter 5 is based on the joint work with Chen-Yu Wei and Kaiqing Zhang [72].

Part I

In Chapter 2, we first formulate an optimal control problem for constrained MDPs and provide necessary background material. Then, we describe our natural policy gradient primal-dual method and provide convergence guarantees for our algorithm under the tabular softmax, log-linear, and general smooth policy parametrizations. Next, we establish convergence and finite-sample complexity guarantees for two model-free primal-dual algorithms and provide computational experiments. Finally, we close this chapter with concluding remarks.

In Chapter 3, we first introduce the episodic finite-horizon constrained MDPs, the metrics of learning performance, and the linear function approximation. We then propose an optimistic primal-dual policy optimization algorithm for learning constrained MDPs. Next, we establish a regret and constraint violation analysis for the proposed algorithm. We further present some improved results in the tabular setting. Finally, we close this chapter with concluding remarks.

Part II

In Chapter 4, we first introduce a class of multi-agent stochastic saddle point problems that contain, as a special instance, minimization of a mean square projected Bellman error via distributed temporal-difference learning. We then develop a homotopy-based online distributed
primal-dual algorithm to solve this problem and establish a finite-time performance bound for the proposed algorithm. We offer computational experiments to demonstrate the merits and the effectiveness of our theoretical findings. Finally, we close the chapter with concluding remarks.

In Chapter 5 we first introduce Markov potential games, Nash equilibrium, and provide necessary background material. Then, we present an independent learning protocol, and propose an independent policy gradient method for Markov potential games and establish a Nash regret analysis. We next establish a model-free extension of our method and analysis to the linear function approximation setting. Furthermore, we establish game-agnostic convergence of an optimistic independent policy gradient method for learning both Markov cooperative games and zero-sum Markov games. We also provide computational experiments to demonstrate the merits and the effectiveness of our theoretical findings. Finally, we close the chapter with concluding remarks.

1.6 Contributions of the dissertation

This section summarizes the most important contributions of the dissertation.

Part 1

Convergence and sample complexity of natural policy gradient primal-dual methods for constrained MDPs. We propose a simple but effective primal-dual algorithm for solving discounted infinite-horizon optimal control problems for constrained MDPs. Our Natural Policy Gradient Primal-Dual (NPG-PD) method employs natural policy gradient ascent to update the primal variable and projected sub-gradient descent to update the dual variable. We exploit the structure of softmax policy parametrization to establish global convergence guarantees in spite of the fact that the objective function in maximization problem is not concave and the constraint
set is not convex. In particular, we prove that our NPG-PD method achieves global convergence with rate $O(1/\sqrt{T})$ in both the optimality gap and the constraint violation, where $T$ is the total number of iterations. Our convergence guarantees are dimension-free, i.e., the rate is independent of the size of the state-action space. We further establish convergence with rate $O(1/\sqrt{T})$ in both the optimality gap and the constraint violation for log-linear and general smooth policy parametrizations, up to a function approximation error caused by restricted policy parametrization. We also provide convergence and finite-sample complexity guarantees for two sample-based NPG-PD algorithms. Finally, we utilize computational experiments to showcase the merits and the effectiveness of our approach.

**Provably efficient policy optimization for learning constrained MDPs with linear function approximation.** We propose a provably efficient safe RL algorithm for constrained MDPs with an unknown transition model in the linear episodic setting: an Optimistic Primal-Dual Proximal Policy OPtimization (OPDOP) algorithm, where the value function is estimated by combining the least-squares policy evaluation and an additional bonus term for the exploration under constraints (or safe exploration). Theoretically, we prove that the proposed algorithm achieves an $\tilde{O}(dH^{2.5}\sqrt{T})$ regret and the same $\tilde{O}(dH^{2.5}\sqrt{T})$ constraint violation, where $d$ is the dimension of the feature mapping, $H$ is the horizon of each episode, and $T$ is the total number of steps. We establish these bounds in the setting where the reward/utility functions are fixed but the feedback after each episode is bandit. Our bounds depend on the capacity of the state space only through the dimension of the feature mapping and thus hold even when the number of states goes to infinity. To the best of our knowledge, our result is the first provably
efficient online policy optimization for constrained MDPs in the function approximation setting, with safe exploration.

**Part II**

**Homotopy stochastic primal-dual optimization for multi-agent temporal-difference learning.** We formulate the multi-agent temporal-difference (TD) learning as the minimization problem of mean square projected Bellman error (MSPBE). We employ Fenchel duality to cast the MSPBE minimization as a stochastic saddle point problem where the primal-dual objective is convex and strongly-concave with respect to primal and dual variables, respectively. Since the primal-dual objective has a linear dependence on expectations, we can obtain unbiased estimates of gradients from state samples thereby overcoming a challenge that approaches based on naive TD objectives face [138]. This allows us to design a true stochastic primal-dual learning algorithm and perform the finite-time performance analysis in Markov setting [175, 83]. Our primal-dual formulation utilizes distributed dual averaging [82] and for our homotopy-based distributed learning algorithm we establish a sharp finite-time error bound in terms of network size and topology. This differentiates our work from the approaches and results in [127, 76, 77]. To the best of our knowledge, we are the first to utilize the homotopy-based approach for solving a class of distributed convex-concave saddle point programs with $O(1/T)$ finite-time performance bound.

**Independent policy gradient for large-scale Markov potential games (MPGs): sharper rates, function approximation, and game-agnostic convergence.** First, we propose an independent policy gradient algorithm for learning an $\epsilon$-Nash equilibrium of Markov potential games
(MPGs) with $O(1/\epsilon^2)$ iteration complexity. In contrast to the state of the art results [131, 277], such iteration complexity does not explicitly depend on the state space size. Second, we consider a linear function approximation setting and design an independent sample-based policy gradient algorithm that learns an $\epsilon$-Nash equilibrium with $O(1/\epsilon^5)$ sample complexity. This appears to be the first result for learning MPGs with function approximation. Third, we establish the convergence of an independent optimistic policy gradient algorithm (which has been proved to converge in learning zero-sum Markov games [239]) for learning a subclass of MPGs: Markov cooperative games. We show that the same type of optimistic policy learning algorithm provides an $\epsilon$-Nash equilibrium in both zero-sum Markov games and Markov cooperative games while the players are oblivious to the types of games being played. To the best of our knowledge, this appears to be the first game-agnostic convergence result in Markov games.
Part I

Reinforcement learning for constrained control systems
Chapter 2

Natural policy gradient primal-dual method for constrained MDPs

In this chapter, we study sequential decision making control problems aimed at maximizing the expected total reward while satisfying a constraint on the expected total utility. We employ the natural policy gradient method to solve the discounted infinite-horizon optimal control problem for constrained Markov decision processes (MDPs). Specifically, we propose a new Natural Policy Gradient Primal-Dual (NPG-PD) method that updates the primal variable via natural policy gradient ascent and the dual variable via projected sub-gradient descent. Although the underlying maximization involves a nonconcave objective function and a nonconvex constraint set, under the softmax policy parametrization we prove that our method achieves global convergence with sublinear rates regarding both the optimality gap and the constraint violation. Such convergence is independent of the size of the state-action space, i.e., it is dimension-free. Furthermore, for log-linear and general smooth policy parametrizations, we establish sublinear convergence rates up to a function approximation error caused by restricted policy parametrization. We also provide convergence and finite-sample complexity guarantees for two sample-based NPG-PD algorithms.
Finally, we use computational experiments to showcase the merits and the effectiveness of our approach.

2.1 Introduction

Policy gradient [217] and natural policy gradient [112] methods have enjoyed substantial empirical success in solving MDPs [195, 136, 164, 194, 214]. Policy gradient methods, or more generally direct policy search methods, have also been used to solve constrained MDPs [227, 37, 35, 56, 220, 135, 180, 7, 206], but most existing theoretical guarantees are asymptotic and/or only provide local convergence guarantees to stationary-point policies. On the other hand, it is desired to show that, for arbitrary initial condition, a solution that enjoys $\epsilon$-optimality gap and $\epsilon$-constraint violation is computed using a finite number of iterations and/or samples. It is thus imperative to establish global convergence guarantees for policy gradient methods when solving constrained MDPs.

In this chapter, we provide a theoretical foundation for non-asymptotic global convergence of the natural policy gradient method in solving optimal control problems for constrained MDPs and answer the following questions:

(i) Can we employ natural policy gradient methods to solve optimal control problems for constrained MDPs?

(ii) Do natural policy gradient methods converge to the globally optimal solution that satisfies constraints?
(iii) What is the convergence rate of natural policy gradient methods and the effect of the function approximation error caused by a restricted policy parametrization?

(iv) What is the sample complexity of model-free natural policy gradient methods?

In Section 2.2, we formulate an optimal control problem for constrained Markov decision processes and provide necessary background material. In Section 2.3, we describe our natural policy gradient primal-dual method. We provide convergence guarantees for our algorithm under the tabular softmax policy parametrization in Section 2.4 and under log-linear and general smooth policy parametrizations in Section 2.5. We establish convergence and finite-sample complexity guarantees for two model-free primal-dual algorithms in Section 2.6 and provide computational experiments in Section 2.7. We close the chapter with remarks in Section 2.8.

2.2 Problem setup

In Section 2.2.1, we introduce constrained Markov decision processes. In Section 2.2.2, we present the method of Lagrange multipliers, formulate a saddle-point problem for the constrained policy optimization, and exhibit several problem properties: strong duality, boundedness of the optimal dual variable, and constraint violation. In Section 2.2.3, we introduce a parametrized formulation of the constrained policy optimization problem, provide an example of a constrained MDP which is not convex, and present several useful policy parametrizations.
2.2.1 Constrained Markov decision processes

We consider a discounted constrained Markov decision process [11],

\[
\text{CMDP}(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, g, b, \gamma, \rho)
\]

where \( \mathcal{S} \) is a state space, \( \mathcal{A} \) is an action space, \( \mathbb{P} \) is a transition probability measure which specifies the transition probability \( \mathbb{P}(s' \mid s, a) \) from state \( s \) to the next state \( s' \) under action \( a \in \mathcal{A} \), \( r: \mathcal{S} \times \mathcal{A} \rightarrow [0, 1] \) is a reward function, \( g: \mathcal{S} \times \mathcal{A} \rightarrow [0, 1] \) is a utility function, \( b \) is a constraint offset, \( \gamma \in [0, 1) \) is a discount factor, and \( \rho \) is an initial distribution over \( \mathcal{S} \).

For any state \( s_t \), a stochastic policy \( \pi: \mathcal{S} \rightarrow \Delta(\mathcal{A}) \) is a function in the probability simplex \( \Delta(\mathcal{A}) \) over action space \( \mathcal{A} \), i.e., \( a_t \sim \pi(\cdot \mid s_t) \) at time \( t \). Let \( \Pi \) be a set of all possible policies. A policy \( \pi \in \Pi \), together with initial state distribution \( \rho \), induces a distribution over trajectories \( \tau = \{(s_t, a_t, r_t, g_t)\}_{t=0}^{\infty} \), where \( s_0 \sim \rho, a_t \sim \pi(\cdot \mid s_t) \) and \( s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t) \) for all \( t \geq 0 \).

Given a policy \( \pi \), the value functions \( V_r^\pi, V_g^\pi: \mathcal{S} \rightarrow \mathbb{R} \) associated with the reward \( r \) or the utility \( g \) are determined by the expected values of total discounted rewards or utilities received under policy \( \pi \),

\[
V_r^\pi(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid \pi, s_0 = s \right]
\]

\[
V_g^\pi(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t g(s_t, a_t) \mid \pi, s_0 = s \right]
\]
where the expectation $\mathbb{E}$ is taken over the randomness of the trajectory $\tau$ induced by $\pi$. Starting from an arbitrary state-action pair $(s, a)$ and following a policy $\pi$, we also introduce the state-action value functions $Q^\pi_r(s, a), Q^\pi_g(s, a): S \times A \to \mathbb{R}$ together with their advantage functions $A^\pi_r, A^\pi_g: S \times A \to \mathbb{R}$,

\[
\begin{align*}
Q^\pi_r(s, a) &:= \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t \diamond (s_t, a_t) \mid \pi, s_0 = s, a_0 = a \right] \\
A^\pi_r &:= Q^\pi_r(s, a) - V^\pi_r(s)
\end{align*}
\]

where the symbol $\diamond$ represents $r$ or $g$. Since $r, g \in [0, 1]$, we have $V^\pi_r(s) \in [0, 1/(1-\gamma)]$ and their expected values under the initial distribution $\rho$ are determined by $V^\pi_r(\rho) := \mathbb{E}_{s_0 \sim \rho} \left[ V^\pi_r(s_0) \right]$.

Having defined a policy as well as the state-action value functions for the discounted constrained MDP, the objective is to find a policy that maximizes the expected reward value over all policies subject to a constraint on the expected utility value,

\[
\begin{align*}
\text{maximize} & \quad V^\pi_r(\rho) \\
\text{subject to} & \quad V^\pi_g(\rho) \geq b.
\end{align*}
\]

In view of the aforementioned boundedness of $V^\pi_r(s)$ and $V^\pi_g(s)$, we set the constraint offset $b \in (0, 1/(1-\gamma)]$ to make Problem (2.1) meaningful.

**Remark 1** For notational convenience we consider a single constraint in Problem (2.1) but our convergence guarantees are readily generalizable to the problems with multiple constraints.
2.2.2 Method of Lagrange multipliers

By dualizing constraints [145,29], we cast Problem (2.1) into the following max-min problem,

\[
\begin{align*}
\text{maximize} & \quad \pi \\
\text{minimize} & \quad \lambda \\
\text{subject to} & \quad \lambda \geq 0
\end{align*}
\]

\[V^\pi_r(\rho) + \lambda (V^\pi_g(\rho) - b)
\]

(2.2)

where \(V^\pi,\lambda_L(\rho) := V^\pi_r(\rho) + \lambda (V^\pi_g(\rho) - b)\) is the Lagrangian of Problem (2.1), \(\pi\) is the primal variable, and \(\lambda\) is the nonnegative Lagrange multiplier or dual variable. The associated dual function is defined as

\[V^\lambda_D(\rho) := \max_{\pi \in \Pi} V^\pi,\lambda_L(\rho).\]

Instead of utilizing the linear program method [11], we employ direct policy search method to solve Problem (2.2). Direct methods are attractive for three reasons: (i) they allow us to directly optimize/monitor the value functions that we are interested in; (ii) they can deal with large state-action spaces via policy parameterization, e.g., neural nets; and (iii) they can utilize policy gradient estimates via simulations of the policy. Since Problem (2.1) is a nonconcave constrained maximization problem with the policy space \(\Pi\) that is often infinite-dimensional, Problems (2.1) and (2.2) are challenging.

In spite of these challenges, Problem (2.1) has nice properties in the policy space when it is strictly feasible. We adapt the standard Slater condition [29] and assume strict feasibility of Problem (2.1) throughout the chapter.

Assumption 1 (Slater condition) There exists \(\xi > 0\) and \(\bar{\pi} \in \Pi\) such that \(V^\bar{\pi}_g(\rho) - b \geq \xi\).
The Slater condition is mild in practice because we usually have \textit{a priori} knowledge on a strictly feasible policy, e.g., the minimal utility is achievable by a particular policy so that the constraint becomes loose.

Let \( \pi^* \) denote an optimal solution to Problem \((2.1)\), let \( \lambda^* \) be an optimal dual variable

\[
\lambda^* \in \arg\min_{\lambda \geq 0} V_{D}^\lambda(\rho)
\]

and let the set of all optimal dual variables be \( \Lambda^* \). We use the shorthand notation \( V_{r}^\pi(\rho) = V_{r}^*(\rho) \) and \( V_{D}^\lambda(\rho) = V_{D}^*(\rho) \) whenever it is clear from the context. We recall the strong duality for constrained MDPs and we prove boundedness of optimal dual variable \( \lambda^* \).

**Lemma 1 (Strong duality and boundedness of \( \lambda^* \))** Let Assumption \([1]\) hold. Then

(i) \( V_{r}^*(\rho) = V_{D}^*(\rho) \);

(ii) \( 0 \leq \lambda^* \leq (V_{r}^*(\rho) - V_{r}^\pi(\rho))/\xi \).

**Proof.** See Appendix \([A.1]\). \( \square \)

Let the value function associated with Problem \((2.1)\) be determined by

\[
v(\tau) := \maximize_{\pi \in \Pi} \{ V_{r}^\pi(\rho) \mid V_{g}^\pi(\rho) \geq b + \tau \}.
\]

Using the concavity of \( v(\tau) \) (e.g., see \([179, \text{Proposition 1}]\)), in Lemma \([2]\) we establish a bound on the constraint violation; see Appendix \([A.2]\) for proof.
Lemma 2 (Constraint violation) Let Assumption 1 hold. For any $C \geq 2\lambda^*$, if there exists a policy $\pi \in \Pi$ and $\delta > 0$ such that $V^\ast_r(\rho) - V^\pi_r(\rho) + C[b - V^\pi_g(\rho)]_+ \leq \delta$, then $[b - V^\pi_g(\rho)]_+ \leq 2\delta/C$, where $[x]_+ = \max(x, 0)$.

Proof. See Appendix A.2

Aided by the above properties implied by the Slater condition, we target the max-min Problem (2.2) in a primal-dual domain.

2.2.3 Policy parametrization

Introduction of a set of parametrized policies $\{\pi_\theta \mid \theta \in \Theta\}$ brings Problem (2.1) into a constrained optimization problem over the finite-dimensional parameter space $\Theta$,

$$\text{maximize} \quad \theta \in \Theta \quad V_{r}^{\pi_\theta}(\rho)$$

subject to $V_{g}^{\pi_\theta}(\rho) \geq b$. (2.3)

A parametric version of max-min Problem (2.2) is given by

$$\text{maximize} \quad \theta \in \Theta \quad \text{minimize} \quad \lambda \geq 0 \quad V_{r}^{\pi_\theta}(\rho) + \lambda(V_{g}^{\pi_\theta}(\rho) - b).$$

where $V_{L}^{\pi_\theta,\lambda}(\rho) := V_{r}^{\pi_\theta}(\rho) + \lambda(V_{g}^{\pi_\theta}(\rho) - b)$ is the associated Lagrangian and $\lambda$ is the Lagrange multiplier. The dual function is determined by $V_{D}^{\lambda}(\rho) := \text{maximize}_\theta V_{L}^{\pi_\theta,\lambda}(\rho)$. The primal maximization problem (2.3) is finite-dimensional but not concave even if in the absence of a constraint [8]. In Lemma 3 we prove that, in general, Problem (2.3) is not convex because it involves
maximization of a non concave objective function over non convex constraint set. The proof is
provided in Appendix A.3 and it utilizes an example of a constrained MDP in Figure 2.1.

Lemma 3 (Lack of convexity) There exists a constrained MDP for which the objective function
\( V_{\pi^\theta}(s) \) in Problem (2.3) is not concave and the constraint set \( \{ \theta \in \Theta \mid V_{g}^{\pi^\theta}(s) \geq b \} \) is not convex.

![Figure 2.1: An example of a constrained MDP for which the objective function \( V_{\pi^\theta}(s) \) in Problem (2.3) is not concave and the constraint set \( \{ \theta \in \Theta \mid V_{g}^{\pi^\theta}(s) \geq b \} \) is not convex. The pair \((r, g)\) associated with the directed arrow represents (reward, utility) received when an action at a certain state is taken. This example is utilized in the proof of Lemma 3.](image)

In general, the Lagrangian \( V^{\pi^\theta, \lambda}(\rho) \) in Problem (2.4) is convex in \( \lambda \) but not concave in \( \theta \).

While many algorithms for solving max-min optimization problems, e.g., those proposed in [137, 176, 254], require extra assumptions on the max-min structure or only guarantee convergence to a stationary point, we exploit problem structure and propose a new primal-dual method to compute globally optimal solution to Problem (2.4). Before doing that, we first introduce several useful classes of policies.

**Direct policy parametrization.** A direct parametrization of a policy is the probability distribution,

\[
\pi_\theta(a \mid s) = \theta_{s,a} \text{ for all } \theta \in \Delta(\mathcal{A})^{\mid S\mid}
\]


where $\theta_s \in \Delta(A)$ for any $s \in S$, i.e., $\theta_{s,a} \geq 0$ and $\sum_{a \in A} \theta_{s,a} = 1$. This policy class is complete since it directly represents any stochastic policy. Even though it is challenging to deal with from both theoretical and the computational viewpoints [159, 8], it offers a sanity check for many policy search methods.

**Softmax policy parametrization.** This class of policies is parametrized by the softmax function,

$$
\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in A} \exp(\theta_{s,a'})}
$$

for all $\theta \in \mathbb{R}^{|S||A|}$.

The softmax policy can be used to represent any stochastic policy and its closure contains all stationary policies. It has been utilized to study convergence properties of many RL algorithms [32, 8, 159, 50, 117] and it offers several algorithmic advantages: (i) it equips the policy with a rich structure so that the natural policy gradient update works like the classical multiplicative weights update in the online learning literature (e.g., see [51]); (ii) it can be used to interpret the function approximation error [8].

**Log-linear policy parametrization.** A log-linear policy is given by

$$
\pi_\theta(a \mid s) = \frac{\exp(\theta^\top \phi_{s,a})}{\sum_{a' \in A} \exp(\theta^\top \phi_{s,a'})}
$$

for all $\theta \in \mathbb{R}^d$, where $\phi_{s,a} \in \mathbb{R}^d$ is the feature map at a state-action pair $(s, a)$. The log-linear policy builds on the softmax policy by applying the softmax function to a set of linear functions in a given feature space. More importantly, it exactly characterizes the linear function approximation via policy parametrization [8]; see [162, 13] for linear constrained MDPs.
General policy parametrization. A general class of stochastic policies is given by \( \{ \pi_\theta \mid \theta \in \Theta \} \) with \( \Theta \subset \mathbb{R}^d \) without specifying the structure of \( \pi_\theta \). The parameter space has dimension \( d \) and this policy class covers a setting that utilizes nonlinear function approximation, e.g., (deep) neural networks \([139, 233]\).

When we choose \( d \ll |S||A| \) in either the log-linear policy or the general nonlinear policy, the policy class has a limited expressiveness and it may not contain all stochastic policies. Motivated by this observation, the theory that we develop in Section 2.5 establishes global convergence up to error caused by the restricted policy class.

2.3 Natural policy gradient primal-dual method

In Section 2.3.1, we provide a brief summary of three basic algorithms that have been used to solve constrained policy optimization problem (2.3). In Section 2.3.2, we propose a natural policy gradient primal-dual method which represents an extension of natural policy gradient method to constrained optimization problems.

2.3.1 Constrained policy optimization methods

We briefly summarize three basic algorithms that can be employed to solve the primal problem (2.3). We assume that the value function and the policy gradient can be evaluated exactly for any given policy.
We first introduce some useful definitions. The discounted visitation distribution \(d^\pi_{s_0}(s)\) of a policy \(\pi\) and its expectation over the initial distribution \(\rho\) are respectively given by,

\[
d^\pi_{s_0}(s) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr^\pi(s_t = s \mid s_0)
\]

\[
d^\pi_{\rho}(s) = \mathbb{E}_{s_0 \sim \rho} \left[ d^\pi_{s_0}(s) \right]
\]

where \(\Pr^\pi(s_t = s \mid s_0)\) is the probability of visiting state \(s\) at time \(t\) under the policy \(\pi\) with an initial state \(s_0\). When the use of parametrized policy \(\pi_\theta\) is clear from the context, we use \(V_\pi^\theta(\rho)\) to denote \(V_{r_\pi}^\theta(\rho)\). When \(\pi_\theta(\cdot \mid s)\) is differentiable and when it belongs to the probability simplex, i.e., \(\pi_\theta \in \Delta(A)^{|S|}\) for all \(\theta\), the policy gradient of the Lagrangian (2.4) is determined by,

\[
\nabla_\theta V_{L}^{\theta,\lambda}(s_0) = \nabla_\theta V_{r}^{\theta}(s_0) + \lambda \nabla_\theta V_{g}^{\theta}(s_0)
\]

\[
= \frac{1}{1 - \gamma} \mathbb{E}_{s_0 \sim d^\pi_{s_0}} \mathbb{E}_{a \sim \pi_\theta(\cdot \mid s)} \left[ A_{L}^{\theta,\lambda}(s, a) \nabla_\theta \log \pi_\theta(a \mid s) \right]
\]

where \(A_{L}^{\theta,\lambda}(s, a) = A_{r}^{\theta}(s, a) + \lambda A_{g}^{\theta}(s, a)\).

**Dual method**

When strong duality in Lemma 1 holds, it is convenient to work with the dual formulation of the primal problem (2.3),

\[
\text{minimize}_{\lambda \geq 0} \; V_{D}^{\lambda}(\rho).
\]
While the dual function is convex regardless of concavity of the primal maximization problem, it is often non-differentiable \[^{[30]}\]. Thus, a projected dual subgradient descent can be used to solve the dual problem,

\[ \lambda^{(t+1)} = P_+ (\lambda^{(t)} - \eta \partial_\lambda V_{\lambda}^{\lambda(t)} (\rho)) \]

where \( P_+ (\cdot) \) is the projection to the non-negative real axis, \( \eta > 0 \) is the stepsize, and

\[ \partial_\lambda V_{\lambda}^{\lambda(t)} (\rho) := \partial_\lambda V_{\lambda}^{\lambda(t)} (\rho) \bigg|_{\lambda = \lambda^{(t)}} \]

is the subgradient of the dual function evaluated at \( \lambda = \lambda^{(t)} \).

The dual method works in the space of dual variables and it requires efficient evaluation of the subgradient of the dual function. We note that computing the dual function \( V_{\lambda}^{\lambda}(\rho) \) for a given \( \lambda = \lambda^{(t)} \) in each step amounts to a standard unconstrained RL problem \[^{[179]}\]. In spite of global convergence guarantees for several policy search methods in the tabular setting, it is often challenging to obtain the dual function and/or to compute its sub-gradient, e.g., when the problem dimension is high and/or when the state space is continuous. Although the primal problem can be approximated using the first-order Taylor series expansion \[^{[7, 259]}\], inverting Hessian matrices becomes the primary computational burden and it is costly to implement the dual method.

**Primal method**

In the primal method, a policy search strategy works directly on the primal problem \(^{(2.3)}\) by seeking an optimal policy in a feasible region. The key challenge is to ensure the feasibility of the
next iterate in the search direction, which is similar to the use of the primal method in nonlinear programming [145].

An intuitive approach is to check the feasibility of each iterate and determine whether the constraint is active [251]. If the iterate is feasible or the constraint is inactive, we move towards maximizing the single objective function; otherwise, we look for a feasible direction. For the softmax policy parametrization (2.5), this can be accomplished using a simple first-order gradient method,

\[
\theta_{s,a}^{(t+1)} = \theta_{s,a}^{(t)} + \eta G_{s,a}^{(t)}(\rho)
\]

where we use the \( A_r^{(t)}(s,a) \) and \( A_g^{(t)}(s,a) \) to denote \( A_{\pi}^{\theta^{(t)}}(s,a) \) and \( A_g^{\theta^{(t)}}(s,a) \), respectively, \( G_{s,a}^{(t)}(\rho) \) is the gradient ascent direction determined by the scaled version of advantage functions, and \( \epsilon_b > 0 \) is the relaxation parameter for the constraint \( V^\pi_{g}(\rho) \geq b \). When the iterate violates the relaxed constraint, \( V^\pi_{g}(\rho) > b - \epsilon_b \), it maximizes the constraint function to gain feasibility. More reliable evaluation of the feasibility often demands a more tractable characterization of the constraint, e.g., by utilizing Lyapunov function [54], Gaussian process modeling [211], backward value function [191], and logarithmic penalty function [143]. Hence, the primal method offers the adaptability of adjusting a policy to satisfy the constraint, which is desirable in safe training applications. However, global convergence theory is still lacking and recent progress [251] requires a careful relaxation of the constraint.
Primal-dual method

The primal-dual method simultaneously updates primal and dual variables \([16]\). A basic primal-dual method with the direct policy parametrization \(\pi_\theta(a \mid s) = \theta_{s,a}\) performs the following Policy Gradient Primal-Dual (PG-PD) update \([1]\),

\[
\theta^{(t+1)} = \mathcal{P}_\Theta\left(\theta^{(t)} + \eta_1 \nabla_\theta V_{\pi_\theta,L}^{\theta^{(t)},\lambda^{(t)}}(\rho)\right)
\]

\[
\lambda^{(t+1)} = \mathcal{P}_\Lambda\left(\lambda^{(t)} - \eta_2 \left(V_{g}^{\theta^{(t)}}(\rho) - b\right)\right)
\]

(2.10)

where \(\nabla_\theta V_{\pi_\theta,L}^{\theta^{(t)},\lambda^{(t)}}(\rho) := \nabla_\theta V_{r}^{\theta^{(t)}}(\rho) + \lambda^{(t)} \nabla_\theta V_{g}^{\theta^{(t)}}(\rho)\), \(\eta_1 > 0\) and \(\eta_2 > 0\) are the stepsizes, \(\mathcal{P}_\Theta\) is the projection onto probability simplex \(\Theta := \Delta(\mathcal{A})^{\mid S\mid}\), and \(\mathcal{P}_\Lambda\) is the projection that will be specified later. For the max-min formulation (2.4), PG-PD method (2.10) directly performs projected gradient ascent in the policy parameter \(\theta\) and descent in the dual variable \(\lambda\), both over the Lagrangian \(V_{\pi_\theta,L}^{\theta^{(t)},\lambda^{(t)}}(\rho)\). The primal-dual method overcomes disadvantages of the primal and dual methods either by relaxing the precise calculation of the subgradient of the dual function or by changing the descent direction via tuning of the dual variable. While this simple method provides a foundation for solving constrained MDPs \([56, 220]\), lack of convexity in (2.4) makes it challenging to establish convergence theory, which is the primary objective of this chapter.

We first leverage structure of constrained policy optimization problem (2.3) to provide a positive result in terms of optimality gap and constraint violation.

**Theorem 4 (Restrictive convergence: direct policy)** Let Assumption\([2]\) hold with a policy class \(\{\pi_\theta = \theta \mid \theta \in \Theta\}\) and let \(\Lambda = [0, 2/((1 - \gamma)\xi)]\), \(\rho > 0\), \(\lambda^{(0)} = 0\), and \(\theta^{(0)}\) be such that \(V_{r}^{\theta^{(0)}}(\rho) \geq \)]
If we choose \( \eta_1 = O(1) \) and \( \eta_2 = O(1/\sqrt{T}) \), then the iterates \( \theta^{(t)} \) generated by PG-PD method (2.10) satisfy

\[
\frac{1}{T} \sum_{t=0}^{T-1} (V_r^*(\rho) - V_r^{(t)}(\rho)) \leq C_1 \frac{|A||S|}{(1 - \gamma)^6 T^{1/4}} \left\| \frac{d^\pi_r*}{\rho} \right\|_\infty^2
\]

\[
\frac{1}{T} \sum_{t=0}^{T-1} (b - V_g^{(t)}(\rho)) \leq C_2 \frac{|A||S|}{(1 - \gamma)^6 T^{1/4}} \left\| \frac{d^\pi_r*}{\rho} \right\|_\infty^2
\]

where \( C_1 \) and \( C_2 \) are absolute constants that are independent of \( T \).

For the tabular constrained MDPs with direct policy parametrization, Theorem 4 guarantees that, on average, the optimality gap \( V_r^*(\rho) - V_r^{(t)}(\rho) \) and the constraint violation \( b - V_g^{(t)}(\rho) \) decay to zero with the sublinear rate \( 1/T^{1/4} \). However, this rate explicitly depends on the sizes of state/action spaces \( |S| \) and \( |A| \), and the distribution shift \( \left\| \frac{d^\pi_r*}{\rho} \right\|_\infty \) that specifies the exploration factor. A careful initialization \( \theta^{(0)} \) that satisfies \( V_r^{\theta^{(0)}}(\rho) \geq V_r^*(\rho) \) is also required.

The proof of Theorem 4 is provided in Appendix A.4 and it exploits the problem structure that casts the primal problem (2.3) as a linear program in the occupancy measure [11] and applies the convex optimization analysis. This method is not well-suited for large-scale problems, and projections onto the high-dimensional probability simplex are not desirable in practice. We next introduce a natural policy gradient primal-dual method to overcome these challenges and provide stronger convergence guarantees.

### 2.3.2 Natural policy gradient primal-dual method

The Fisher information matrix induced by \( \pi_\theta \),

\[
F_\rho(\theta) := \mathbb{E}_{s \sim \rho^*} \mathbb{E}_{a \sim \pi_\theta(s \mid \cdot)} \left[ \nabla_\theta \log \pi_\theta(a \mid s) \left( \nabla_\theta \log \pi_\theta(a \mid s) \right)^\top \right]
\]
is used in the update of the primal variable in our primal-dual algorithm. The expectations are taken over the randomness of the state-action trajectory induced by $\pi_\theta$ and Natural Policy Gradient Primal-Dual (NPG-PD) method for solving Problem (2.4) is given by,

$$
\theta^{(t+1)} = \theta^{(t)} + \eta_1 F^\dagger(\theta^{(t)}) \nabla_\theta V_L^{\theta^{(t)}, \lambda^{(t)}}(\rho)
$$

$$
\lambda^{(t+1)} = \mathcal{P}_\Lambda \left( \lambda^{(t)} - \eta_2 \left( V_{g}^{\theta^{(t)}}(\rho) - b \right) \right)
$$

where $\dagger$ denotes the Moore-Penrose inverse of a given matrix, $\mathcal{P}_\Lambda(\cdot)$ denotes the projection to the interval $\Lambda$ that will be specified later, and $(\eta_1, \eta_2)$ are constant positive stepsizes in the updates of primal and dual variables. The primal update $\theta^{(t+1)}$ is obtained using a pre-conditioned gradient ascent via the natural policy gradient $F^\dagger(\theta^{(t)}) \nabla_\theta V_L^{(t)}(\rho)$ and it represents the policy gradient of the Lagrangian $V_L^{(t)}(\rho)$ in the geometry induced by the Fisher information matrix $F_\rho(\theta^{(t)})$.

On the other hand, the dual update $\lambda^{(t+1)}$ is obtained using a projected sub-gradient descent by collecting the constraint violation $b - V_g^{(t)}(\rho)$, where, for brevity, we use $V_L^{\theta^{(t)}}(\rho)$ and $V_g^{\theta^{(t)}}(\rho)$ to denote $V_L^{\theta^{(t)}, \lambda^{(t)}}(\rho)$ and $V_g^{\theta^{(t)}}(\rho)$, respectively.

In Section 2.4, we establish global convergence of NPG-PD method (2.11) under the softmax policy parametrization. In Section 2.5, we examine the general policy parametrization and, in Section 2.6, we analyze sample complexity of two sample-based implementations of NPG-PD method (2.11).

**Remark 2** The performance difference lemma [113, 8], which quantifies the difference between two state value functions, $V_\pi(s_0)$ and $V_{\pi'}(s_0)$, for any two policies $\pi$ and $\pi'$ and any state $s_0$,

$$
V_\pi(s_0) - V_{\pi'}(s_0) = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{s_0}, a \sim \pi(\cdot|s)} \left[ A_{\pi'}(s, a) \right]
$$
is utilized in our analysis, where the symbol $\diamondsuit$ denotes $r$ or $g$.

### 2.4 Tabular softmax policy case

We first examine NPG-PD method (2.11) under softmax policy parametrization (2.5). Strong duality in Lemma 1 holds on the closure of the softmax policy class, because of completeness of the softmax policy class. Even though maximization problem (2.3) is not concave, we establish global convergence of our algorithm with dimension-independent convergence rates.

We first exploit the softmax policy structure to show that the primal update in (2.11) can be expressed in a more compact form; see Appendix A.5 for the proof.

**Lemma 5 (Primal update as MWU)** Let $\Lambda := [0, 2/(1 - \gamma)\xi]$ and let $A_L^{(t)}(s, a) := A_r^{(t)}(s, a) + \lambda^{(t)}A_g^{(t)}(s, a)$. Under softmax parametrized policy (2.5), NPG-PD algorithm (2.11) can be brought to the following form,

\[
\begin{align*}
\theta_{s,a}^{(t+1)} &= \theta_{s,a}^{(t)} + \frac{\eta_1}{1 - \gamma} A_L^{(t)}(s, a) \\
\lambda^{(t+1)} &= \mathcal{P}_\Lambda (\lambda^{(t)} - \eta_2 (V_r^{(t)}(\rho) - b)).
\end{align*}
\]

Furthermore, the primal update in (2.12a) can be equivalently expressed as

\[
\pi^{(t+1)}(a \mid s) = \pi^{(t)}(a \mid s) \frac{\exp \left( \frac{\eta_1}{1 - \gamma} A_L^{(t)}(s, a) \right)}{Z^{(t)}(s)}
\]

where $Z^{(t)}(s) := \sum_{a \in A} \pi^{(t)}(a \mid s) \exp \left( \frac{\eta_1}{1 - \gamma} A_L^{(t)}(s, a) \right)$.

The primal updates in (2.12a) do not depend on the state distribution $d_\rho^{(t)}$ that appears in NPG-PD algorithm (2.11) through the policy gradient. This is because of the Moore-Penrose inverse of the Fisher information matrix in (2.11). Furthermore, policy update (2.12b) is given by the
multiplicative weights update (MWU) which is commonly used in online linear optimization [51]. In contrast to the online linear optimization, an advantage function appears in the MWU policy update at each iteration in (2.12b).

In Theorem 6, we establish global convergence of NPG-PD algorithm (2.12a) with respect to both the optimality gap $V^*(\rho) - V^t(\rho)$ and the constraint violation $b - V^t(\rho)$. Even though we set $\theta_{s,a}(0) = 0$ and $\lambda(0) = 0$ in the proof of Theorem 6 in Section 2.4.1, global convergence can be established for arbitrary initial conditions.

**Theorem 6 (Dimension-free global convergence: softmax policy)** Let us fix $T > 0$ and $\rho \in \Delta_S$ and let Assumption 1 hold for $\xi > 0$. If we choose $\eta_1 = 2 \log |A|$ and $\eta_2 = 2(1 - \gamma)/\sqrt{T}$, then the iterates $\pi(t)$ generated by algorithm (2.12) satisfy,

$$(\text{Optimality gap}) \quad \frac{1}{T} \sum_{t=0}^{T-1} (V^*(\rho) - V^t(\rho)) \leq \frac{7}{(1 - \gamma)^2} \frac{1}{\sqrt{T}}$$

$$(\text{Constraint violation}) \quad \left[ \frac{1}{T} \sum_{t=0}^{T-1} (b - V^t(\rho)) \right]_+ \leq \frac{2/\xi + 4 \xi}{(1 - \gamma)^2} \frac{1}{\sqrt{T}}.$$ 

Theorem 6 demonstrates that, on average, the reward value function converges to its globally optimal value and that the constraint violation decays to zero. In other words, for a desired accuracy $\epsilon$, it takes $O(1/\epsilon^2)$ iterations to compute the solution which is $\epsilon$ away from the globally optimal one (with respect to both the optimality gap and the constraint violation). We note that the required number of iterations only depends on the desired accuracy $\epsilon$ and is independent of the sizes of the state and action spaces. Although maximization problem (2.3) is not concave, our rate $(1/\sqrt{T}, 1/\sqrt{T})$ for optimality/constraint violation gap outperforms the classical one $(1/\sqrt{T}, 1/T^{3/4})$ [149] and it matches the achievable rate for solving online convex minimization problems with convex constraint sets [263]. Moreover, in contrast to the bounds established for
PG-PD algorithm (2.10) in Theorem 4, the bounds in Theorem 6 for NPG-PD algorithm (2.11) under softmax policy parameterization do not depend on the initial distribution $\rho$.

As shown in Lemma 7 in Section 2.4.1, the reward and utility value functions are coupled and the natural policy gradient method in the unconstrained setting does not provide monotonic improvement to either of them [8, Section 5.3]. To address this challenge, we introduce a new line of analysis. To bound the optimality gap via a drift analysis of the dual update we first establish the bounded average performance in Lemma 8 in Section 2.4.1. Furthermore, instead of using methods from constrained convex optimization [149, 263, 242, 265], which either require extra assumptions or have slow convergence rate, under strong duality we establish that the constraint violation for nonconvex Problem (2.3) converges with the same rate as the optimality gap.

### 2.4.1 Non-asymptotic convergence analysis

We first show that the policy improvement is not monotonic in either the reward value function or the utility value function.

**Lemma 7 (Non-monotonic improvement)** For any distribution of the initial state $\mu$, iterates $(\pi^{(t)}, \lambda^{(t)})$ of algorithm (2.12) satisfy

$$V_r^{(t+1)}(\mu) - V_r^{(t)}(\mu) + \lambda^{(t)}(V_g^{(t+1)}(\mu) - V_g^{(t)}(\mu)) \geq \frac{1 - \gamma}{\eta_1} E_{s \sim \mu} \log Z^{(t)}(s) \geq 0. \quad (2.13)$$
PROOF. Let \( d_{\mu}^{(t+1)} := d_{\pi}^{(t+1)} \). The performance difference lemma in conjunction with the multiplicative weights update in (2.12b) yield,

\[
V_{r}^{(t+1)}(\mu) - V_{r}^{(t)}(\mu) = \left[ \frac{1}{1 - \gamma} - \lambda(t) \right] \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\mu}^{(t+1)}} \left[ \sum_{a \in \mathcal{A}} \pi^{(t+1)}(a \mid s) A_{r}^{(t)}(s, a) \right] + \frac{1}{\eta_{1}} \mathbb{E}_{s \sim d_{\mu}^{(t+1)}} \left[ D_{KL} \left( \pi^{(t+1)}(a \mid s) \mid \pi^{(t)}(a \mid s) \right) \right] + \frac{1}{\eta_{1}} \mathbb{E}_{s \sim d_{\mu}^{(t+1)}} \log Z^{(t)}(s) - \lambda(t) \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\mu}^{(t+1)}} \left[ \sum_{a \in \mathcal{A}} \pi^{(t+1)}(a \mid s) A_{g}^{(t)}(s, a) \right]
\]

where the last equality follows from the definition of the Kullback-Leibler divergence or relative entropy between distributions \( p \) and \( q \), \( D_{KL}(p \parallel q) := \mathbb{E}_{x \sim p} \log(p(x)/q(x)) \). Furthermore,

\[
\frac{1}{\eta_{1}} \mathbb{E}_{s \sim d_{\mu}^{(t+1)}} \left[ D_{KL} \left( \pi^{(t+1)}(a \mid s) \mid \pi^{(t)}(a \mid s) \right) \right] + \frac{1}{\eta_{1}} \mathbb{E}_{s \sim d_{\mu}^{(t+1)}} \log Z^{(t)}(s) - \lambda(t) \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\mu}^{(t+1)}} \left[ \sum_{a \in \mathcal{A}} \pi^{(t+1)}(a \mid s) A_{g}^{(t)}(s, a) \right] \geq \frac{1}{\eta_{1}} \mathbb{E}_{s \sim d_{\mu}^{(t+1)}} \log Z^{(t)}(s) - \lambda(t) \left( V_{g}^{(t+1)}(\mu) - V_{g}^{(t)}(\mu) \right)
\]

is a consequence of the performance difference lemma, where we drop a nonnegative term in (a) and (b). The first inequality in (2.13) follows from a componentwise inequality \( d_{\mu}^{(t+1)} \geq (1 - \gamma) \mu \), which is obtained using (2.7).
Now we prove that $\log Z(t)(s) \geq 0$. From the definition of $Z(t)(s)$ we have

$$
\log Z(t)(s) = \log \left( \sum_{a \in A} \pi(t)(a | s) \exp \left( \frac{\eta}{1 - \gamma} (A_r(t)(s, a) + \lambda(t) A_g(t)(s, a)) \right) \right),
$$

$$
\geq \sum_{a \in A} \pi(t)(a | s) \log \left( \exp \left( \frac{\eta}{1 - \gamma} (A_r(t)(s, a) + \lambda(t) A_g(t)(s, a)) \right) \right)
$$

$$
= \frac{\eta}{1 - \gamma} \sum_{a \in A} \pi(t)(a | s) (A_r(t)(s, a) + \lambda(t) A_g(t)(s, a))
$$

$$
= \frac{\eta}{1 - \gamma} \sum_{a \in A} \pi(t)(a | s) A_r(t)(s, a) + \frac{\eta}{1 - \gamma} \lambda(t) \sum_{a \in A} \pi(t)(a | s) A_g(t)(s, a)
$$

where in (a) we apply the Jensen’s inequality to the concave function $\log(x)$. On the other hand, the last equality is due to that

$$
\sum_{a \in A} \pi(t)(a | s) A_r(t)(s, a) = \sum_{a \in A} \pi(t)(a | s) (Q_r(t)(s, a) - V_r(t)(s)) = 0
$$

$$
\sum_{a \in A} \pi(t)(a | s) A_g(t)(s, a) = 0
$$

which follow from the definitions of $A_r(t)(s, a)$ and $A_g(t)(s, a)$. $\square$

We next compare the value functions of policy iterates generated by algorithm (2.12) with the ones that result from the use of optimal policy.

**Lemma 8 (Bounded average performance)** Let Assumption 2 hold and let us fix $T > 0$ and $\rho \in \Delta_S$. Then the iterates $(\pi(t), \lambda(t))$ generated by algorithm (2.12) satisfy

$$
\frac{1}{T} \sum_{t=0}^{T-1} \left( (V_r^*(\rho) - V_r(t)(\rho)) + \lambda(t) (V_g^*(\rho) - V_g(t)(\rho)) \right) \leq \frac{\log |A|}{\eta_1 T} + \frac{1}{(1 - \gamma)^2 T} + \frac{2\eta_2}{(1 - \gamma)^3}. \tag{2.14}
$$
Let $d^* := d^\mu$. The performance difference lemma in conjunction with the multiplicative weights update in (2.12b) yield,

$$V_r^*(\rho) - V_r^{(t)}(\rho) = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d^*} \left[ \sum_{a \in \mathcal{A}} \pi^*(a \mid s) A_r^{(t)}(s, a) \right]$$

$$= \frac{1}{\eta_1} \mathbb{E}_{s \sim d^*} \left[ \sum_{a \in \mathcal{A}} \pi^*(a \mid s) \log \left( \frac{\pi^{(t+1)}(a \mid s)}{\pi^{(t)}(a \mid s)} Z^{(t)}(s) \right) \right]$$

$$- \frac{\lambda^{(t)}}{1 - \gamma} \mathbb{E}_{s \sim d^*} \left[ \sum_{a \in \mathcal{A}} \pi^*(a \mid s) A_g^{(t)}(s, a) \right].$$

Application of the definition of the Kullback–Leibler divergence or relative entropy between distributions $p$ and $q$, $D_{KL}(p \parallel q) := \mathbb{E}_{x \sim p} \log(p(x)/q(x))$, and the performance difference lemma again yield,

$$V_r^*(\rho) - V_r^{(t)}(\rho) = \frac{1}{\eta_1} \mathbb{E}_{s \sim d^*} \left[ D_{KL}(\pi^*(a \mid s) \parallel \pi^{(t)}(a \mid s)) - D_{KL}(\pi^*(a \mid s) \parallel \pi^{(t+1)}(a \mid s)) \right]$$

$$+ \frac{1}{\eta_1} \mathbb{E}_{s \sim d^*} \log Z^{(t)}(s) - \frac{\lambda^{(t)}}{1 - \gamma} \mathbb{E}_{s \sim d^*} \left[ \sum_{a \in \mathcal{A}} \pi^*(a \mid s) A_g^{(t)}(s, a) \right]$$

$$= \frac{1}{\eta_1} \mathbb{E}_{s \sim d^*} \left[ D_{KL}(\pi^*(a \mid s) \parallel \pi^{(t)}(a \mid s)) - D_{KL}(\pi^*(a \mid s) \parallel \pi^{(t+1)}(a \mid s)) \right]$$

$$+ \frac{1}{\eta_1} \mathbb{E}_{s \sim d^*} \log Z^{(t)}(s) - \lambda^{(t)}(V_g^*(\rho) - V_g^{(t)}(\rho)).$$

(2.15)

On the other hand, the first inequality in (2.13) with $\mu = d^*$ becomes

$$V_r^{(t+1)}(d^*) - V_r^{(t)}(d^*) + \lambda^{(t)}(V_g^{(t+1)}(d^*) - V_g^{(t)}(d^*)) \geq \frac{1 - \gamma}{\eta_1} \mathbb{E}_{s \sim d^*} \log Z^{(t)}(s).$$

(2.16)
Hence, application of (2.16) to the average of (2.15) over \( t = 0, 1, \ldots, T - 1 \) leads to,

\[
\frac{1}{T} \sum_{t=0}^{T-1} (V^*_r(\rho) - V^r(t)(\rho))
\]

\[= \frac{1}{\eta_1 T} \sum_{t=0}^{T-1} \mathbb{E}_{s \sim d^*} \left[ D_{KL}(\pi^*(a | s) \| \pi^t(a | s)) - D_{KL}(\pi^*(a | s) \| \pi^{t+1}(a | s)) \right]
\]

\[+ \frac{1}{\eta_1 T} \sum_{t=0}^{T-1} \mathbb{E}_{s \sim d^*} \log Z^t(s) - \frac{1}{T} \sum_{t=0}^{T-1} \lambda^t(V^*_g(\rho) - V^g(t)(\rho))
\]

\[\leq \frac{1}{\eta_1 T} \sum_{t=0}^{T-1} \mathbb{E}_{s \sim d^*} \left[ D_{KL}(\pi^*(a | s) \| \pi^t(a | s)) - D_{KL}(\pi^*(a | s) \| \pi^{t+1}(a | s)) \right]
\]

\[+ \frac{1}{(1 - \gamma)T} \sum_{t=0}^{T-1} (V^t_{r+1}(d^*) - V^t_r(d^*))
\]

\[+ \frac{1}{(1 - \gamma)T} \sum_{t=0}^{T-1} \lambda^t(V^t_{g+1}(d^*) - V^t_g(d^*)) - \frac{1}{T} \sum_{t=0}^{T-1} \lambda^t(V^*_g(\rho) - V^g(t)(\rho)).
\]

From the dual update in (2.12a) we have

\[
\frac{1}{T} \sum_{t=0}^{T-1} \lambda^t(V^t_{g+1}(\mu) - V^t_g(\mu))
\]

\[= \frac{1}{T} \sum_{t=0}^{T-1} (\lambda^{t+1}V^t_g(\mu) - \lambda^tV^t_g(\mu)) + \frac{1}{T} \sum_{t=0}^{T-1} (\lambda^t - \lambda^{t+1})V^t_{g+1}(\mu)
\]

\[(a) \leq \frac{1}{T} \lambda^T V^T_g(\mu) + \frac{1}{T} \sum_{t=0}^{T-1} |\lambda^t - \lambda^{t+1}| V^t_{g+1}(\mu)
\]

\[(b) \leq \frac{2\eta_2}{(1 - \gamma)^2}
\]

where we take a telescoping sum for the first sum in (a) and drop a non-positive term, and in (b) we utilize \( |\lambda^T| \leq \eta_2 T/(1 - \gamma) \) and \( |\lambda^t - \lambda^{t+1}| \leq \eta_2/(1 - \gamma) \), which follows from the dual update in (2.12a), the non-expansiveness of projection \( P_A \), and boundedness of the value function.
\[ \text{Bounding the optimality gap. From the dual update in (2.12a) we have} \]

\[
0 \leq \left( \lambda^{(T)} \right)^2 = \sum_{t=0}^{T-1} \left( (\lambda^{(t+1)})^2 - (\lambda^{(t)})^2 \right) \\
= \sum_{t=0}^{T-1} \left( (\mathcal{P}_A (\lambda^{(t)}) - \eta_2 (V_g^{(t)}(\rho) - b)) \right)^2 - (\lambda^{(t)})^2 \\
\overset{(a)}{\leq} \sum_{t=0}^{T-1} \left( (\lambda^{(t)})^2 - \eta_2 (V_g^{(t)}(\rho) - b) \right)^2 - (\lambda^{(t)})^2 \\
= 2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)} (b - V_g^{(t)}(\rho)) + \eta_2^2 \sum_{t=0}^{T-1} (V_g^{(t)}(\rho) - b)^2 \\
\overset{(b)}{\leq} 2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)} (V_g^{(t)}(\rho) - V_g^{(t)}(\rho)) + \frac{\eta_2^2 T}{(1 - \gamma)^2}
\]

where (a) because of the projection \(\mathcal{P}_A\), (b) is because of the feasibility of the optimal policy \(\pi^*\):

\[ V_g^{(t)}(\rho) \geq b, \text{ and } |V_g^{(t)}(\rho) - b| \leq 1/(1 - \gamma). \text{ Hence,} \]

\[
-\frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)} (V_g^{(t)}(\rho) - V_g^{(t)}(\rho)) \leq \frac{\eta_2}{2(1 - \gamma)^2}. \quad (2.19b)
\]
To obtain the optimality gap bound, we now substitute (2.19b) into (2.14), apply $D_{KL}(p \parallel q) \leq \log |A|$ for $p \in \Delta(A)$ and $q = \text{Unif}_{A}$, and take $\eta_1 = 2 \log |A|$ and $\eta_2 = 2(1 - \gamma)/\sqrt{T}$.

**Bounding the constraint violation.** For any $\lambda \in \left[ 0, \frac{2}{2((1 - \gamma)\xi)} \right]$, from the dual update in (2.12a) we have

$$0 \leq \frac{1}{T}|\lambda^{(T)} - \lambda^0|^2 \leq \frac{1}{T}|\lambda^{(0)} - \lambda|^2 - \frac{2\eta_2}{T} \sum_{t=0}^{T-1} (V^*_g(\rho) - b)(\lambda^{(t)} - \lambda) + \frac{\eta_2^2}{(1 - \gamma)^2},$$

where (a) is because of the non-expansiveness of projection $P_{\Lambda}$ and (b) is due to $(V^*_g(\rho) - b)^2 \leq 1/(1 - \gamma)^2$. Averaging the above inequality over $t = 0, \ldots, T - 1$ yields

$$0 \leq \frac{1}{T}|\lambda^{(0)} - \lambda|^2 - \frac{2\eta_2}{T} \sum_{t=0}^{T-1} (V^*_g(\rho) - b)(\lambda^{(t)} - \lambda) + \frac{\eta_2^2}{(1 - \gamma)^2},$$

which implies,

$$\frac{1}{T} \sum_{t=0}^{T-1} (V^*_g(\rho) - b)(\lambda^{(t)} - \lambda) \leq \frac{1}{2\eta_2 T}|\lambda^{(0)} - \lambda|^2 + \frac{\eta_2}{2(1 - \gamma)^2}. \quad (2.20)$$

We now add (2.20) to (2.14) on both sides of the inequality, and utilize $V^*_g(\rho) \geq b$,

$$\frac{1}{T} \sum_{t=0}^{T-1} (V^*_r(\rho) - V^*_g(\rho)) + \frac{\lambda}{T} \sum_{t=0}^{T-1} (b - V^*_g(\rho)) \leq \frac{\log |A|}{\eta_1 T} + \frac{1}{(1 - \gamma)^2 T} + \frac{2\eta_2}{(1 - \gamma)^2} + \frac{1}{2\eta_2 T}|\lambda^{(0)} - \lambda|^2 + \frac{\eta_2}{2(1 - \gamma)^2}.$$
Taking $\lambda = \frac{2}{(1-\gamma)\xi}$ when $\sum_{t=0}^{T-1} (b - V_{g}^{(t)}(\rho)) \geq 0$ and $\lambda = 0$ otherwise, we obtain

$$V_{r}^{\star}(\rho) = \frac{1}{T} \sum_{t=0}^{T-1} V_{r}^{(t)}(\rho) + \frac{2}{(1-\gamma)\xi} \left[ b - \frac{1}{T} \sum_{t=0}^{T-1} V_{g}^{(t)}(\rho) \right] + \frac{\log |A|}{\eta_1 T} + \frac{1}{(1-\gamma)^2 T} + \frac{2\eta_2}{(1-\gamma)^3 T} + \frac{2}{\eta_2(1-\gamma)\xi^2 T} + \frac{\eta_2}{2(1-\gamma)^2}.$$ 

Note that both $V_{r}^{(t)}(\rho)$ and $V_{g}^{(t)}(\rho)$ can be expressed as linear functions in the same occupancy measure [11, Chapter 10] that is induced by policy $\pi^{(t)}$ and transition $P(s' | s, a)$. The convexity of the set of occupancy measures shows that the average of $T$ occupancy measures is an occupancy measure that produces a policy $\pi'$ with value $V_{r}^{\pi'}$ and $V_{g}^{\pi'}$. Hence, there exists a policy $\pi'$ such that $V_{r}^{\pi'}(\rho) = \frac{1}{T} \sum_{t=0}^{T-1} V_{r}^{(t)}(\rho)$ and $V_{g}^{\pi'}(\rho) = \frac{1}{T} \sum_{t=0}^{T-1} V_{g}^{(t)}(\rho)$. Thus,

$$V_{r}^{\star}(\rho) - V_{r}^{\pi'}(\rho) + \frac{2}{(1-\gamma)\xi} \left[ b - V_{g}^{\pi'}(\rho) \right] + \frac{\log |A|}{\eta_1 T} + \frac{1}{(1-\gamma)^2 T} + \frac{2\eta_2}{(1-\gamma)^3} + \frac{2}{\eta_2(1-\gamma)\xi^2 T} + \frac{\eta_2}{2(1-\gamma)^2}.$$ 

Application of Lemma 2 with $2/((1-\gamma)\xi) \geq 2\lambda^*$ yields

$$\left[ b - V_{g}^{\pi'}(\rho) \right]_{+} \leq \frac{\xi \log |A|}{\eta_1 T} + \frac{\xi}{(1-\gamma) T} + \frac{2\eta_2 \xi}{(1-\gamma)^2 T} + \frac{2}{\eta_2(1-\gamma)\xi T} + \frac{\eta_2 \xi}{2(1-\gamma)},$$ 

which leads to our constraint violation bound if we further utilize $\frac{1}{T} \sum_{t=0}^{T-1} (b - V_{g}^{(t)}(\rho)) = b - V_{g}^{\pi'}(\rho), \eta_1 = 2 \log |A|,$ and $\eta_2 = 2(1-\gamma)/\sqrt{T}$. \qed
2.5 Function approximation case

Let us consider a general form of NPG-PD algorithm (2.11),

\[
\begin{align*}
\theta^{(t+1)} &= \theta^{(t)} + \frac{\eta_1}{1-\gamma} w^{(t)} \\
\lambda^{(t+1)} &= \mathcal{P}_\lambda \left( \lambda^{(t)} - \eta_2 \left( V^{(t)}_\theta (\rho) - b \right) \right)
\end{align*}
\] (2.21)

where \( w^{(t)}/(1-\gamma) \) denotes either the exact natural policy gradient or its sample-based approximation. For a general policy class, \( \{ \pi_\theta \mid \theta \in \Theta \} \), with the parameter space \( \Theta \subset \mathbb{R}^d \), the strong duality in Lemma 1 does not necessarily hold and our analysis of Section 2.4 does not apply directly. Let the parametric dual function \( V^{\lambda_\theta}_\theta (\rho) := \max_{\theta \in \Theta} V^{\pi_\theta,\lambda(\rho)}_L (\rho) \) be minimized at the optimal dual variable \( \lambda^*_\theta \). Under the Slater condition in Assumption 1, the parametrization gap [179, Theorem 2] is determined by,

\[
V^{\pi^*}_r (\rho) = V^{\lambda^*}_D (\rho) \geq V^{\lambda_\theta}_D (\rho) \geq V^{\pi^*}_r (\rho) - M \epsilon
\]

where \( \epsilon := \max_s \| \pi(\cdot | s) - \pi_\theta(\cdot | s) \|_1 \) is the policy approximation error and \( M > 0 \) is a problem-dependent constant. Application of item (ii) in Lemma 1 to the set of all optimal dual variables \( \lambda^*_\theta \) yields \( \lambda^*_\theta \in [0, 2/((1-\gamma)\xi)] \) and, thus, \( \Lambda = [0, 2/((1-\gamma)\xi)] \).

To quantify errors caused by the restricted policy parametrization, let us first generalize NPG. For a distribution over state-action pair \( \nu \in \Delta(S \times A) \), we introduce the compatible function approximation error as the following regression objective [112],

\[
E^\nu (w; \theta, \lambda) := \mathbb{E}_{(s,a) \sim \nu} \left[ \left( A^{\theta,\lambda}_L (s,a) - w^T \nabla_\theta \log \pi_\theta(a | s) \right)^2 \right]
\]
where \( A^\theta_\lambda(s, a) := A^\theta_r(s, a) + \lambda A^\theta_g(s, a) \). We can view NPG in (2.11) as a minimizer of \( E^\nu(w; \theta, \lambda) \) for \( \nu(s, a) = d^\pi_\rho(s) \pi_\theta(a \mid s) \),

\[
(1 - \gamma) F^\dagger_\rho(\theta) \nabla_\theta V^\theta_\lambda(\rho) \in \arg \min_\omega E^\nu(w; \theta, \lambda). \tag{2.22}
\]

Expression (2.22) follows from the first-order optimality condition and the use of \( \nabla_\theta V^\theta_\lambda(\rho) := \nabla_\theta V^\theta_r(\rho) + \lambda \nabla_\theta V^\theta_g(\rho) \) allows us to rewrite it as

\[
(1 - \gamma) F^\dagger_\rho(\theta) \nabla_\theta V^\theta(\rho) \in \arg \min_\omega E^\nu(\omega; \theta) \tag{2.23}
\]

where \( \omega \) denotes \( r \) or \( g \).

Let the minimal error be \( E^\nu,\star := \minimize_\omega E^\nu(\omega; \theta) \), where the compatible function approximation error \( E^\nu(\omega; \theta) \) is given by

\[
E^\nu(\omega; \theta) := \mathbb{E}_{(s,a) \sim \nu} \left[ (A^\theta_\omega(s, a) - w^\top_\omega \nabla_\theta \log \pi_\theta(a \mid s))^2 \right]. \tag{2.24}
\]

When the compatible function approximation error is zero, the global convergence follows from Theorem 6. However, this is not the case for a general policy class because it may not include all possible policies (e.g., if we take \( d \ll |S||A| \) for the tabular constrained MDPs). The intuition behind compatibility is that any minimizer of \( E^\nu(\omega; \theta) \) can be used as the NPG direction without affecting convergence theory; also see discussions in [112, 217, 8].
Since the state-action measure \( \nu \) of some feasible comparison policy \( \pi \) is not known, we introduce an exploratory initial distribution \( \nu_0 \) over state-action pairs and define a state-action visitation distribution \( \nu_{\nu_0}^{\pi} \) of a policy \( \pi \) as

\[
\nu_{\nu_0}^{\pi}(s, a) = (1 - \gamma)E_{(s_0, a_0) \sim \nu_0} \left[ \sum_{t=0}^{\infty} \gamma^t \Pr_{\pi}(s_t = s, a_t = a \mid s_0, a_0) \right]
\]

where \( \Pr_{\pi}(s_t = s, a_t = a \mid s_0, a_0) \) is the probability of visiting a state-action pair \((s, a)\) under policy \( \pi \) for an initial state-action pair \((s_0, a_0)\). Whenever clear from context, we use \( \nu^{(t)} \) to denote \( \nu_{\nu_0}^{\pi^{(t)}} \) for notational convenience. It the minimizer is computed exactly, we can update \( w^{(t)} \) in (2.21) using \( w^{(t)} = w_r^{(t)} + \lambda^{(t)} w_g^{(t)} \), where \( w_r^{(t)} \) and \( w_g^{(t)} \) are given by

\[
w_r^{(t)} \in \argmin_{w_r} E_{\nu^{(t)}}(w_r; \theta^{(t)}).
\]

Even though the exact computation of the minimizer may not be feasible, we can use sample-based algorithms to approximately solve the empirical risk minimization problem. By characterizing errors that result from sample-based solutions and from function approximation, we next prove convergence of (2.21) for the log-linear and for the general smooth policy classes.

### 2.5.1 Log-linear policy class

We first consider policies \( \pi_\theta \) in the log-linear class (2.6), with linear feature maps \( \phi_{s,a} \in \mathbb{R}^d \). In this case, the gradient \( \nabla_{\theta} \log \pi_\theta(a \mid s) \) becomes a shifted version of feature \( \phi_{s,a} \),

\[
\nabla_{\theta} \log \pi_\theta(a \mid s) = \phi_{s,a} - \mathbb{E}_{a' \sim \pi_\theta(\cdot \mid s)}[\phi_{s,a'}] := \bar{\phi}_{s,a}.
\]
Thus, the compatible function approximation error (2.24) captures how well the linear function
\( \theta^T \bar{\phi}_{s,a} \) approximates the advantage functions \( A_\theta^r(s,a) \) or \( A_\theta^g(s,a) \) under the state-action distribution \( \nu \). We also introduce the compatible function approximation error with respect to the state-action value functions \( Q_\theta^\diamond(s,a) \),

\[
E_\nu^\diamond(w_\diamond; \theta) := \mathbb{E}_{(s,a) \sim \nu} \left[ (Q_\theta^\diamond(s,a) - w_\diamond^T \phi_{s,a})^2 \right].
\]

When there are no compatible function approximation errors, the policy update in (2.21) for \( w(t) \) that is calculated by (2.25) is given by \( w(t) = w^r(t) + \lambda(t) w^g(t), w^\diamond(t) \in \text{argmin}_{w_\diamond} E_{\nu}^{\diamond(t)}(w_\diamond; \theta^{(t)}) \) for \( \diamond = r \) or \( g \), where \( \nu^{(t)}(s,a) = d^{(t)}(s) \pi^{(t)}(a | s) \) is an on-policy state-action visitation distribution. This is because the softmax function is invariant to any terms that are independent of the action.

Let us consider an approximate solution,

\[
w_\diamond^{(t)} \approx \text{argmin}_{\|w_\diamond\|_2 \leq W} E_{\nu}^{\diamond(t)}(w_\diamond; \theta^{(t)})
\]

where the bounded domain \( W > 0 \) can be viewed as an \( \ell_2 \)-regularization and let the exact minimizer be \( w_\diamond^{(t)} \in \text{argmin}_{\|w_\diamond\|_2 \leq W} E_{\nu}^{\diamond(t)}(w_\diamond; \theta^{(t)}) \). Fixing a state-action distribution \( \nu^{(t)} \), the estimation error in \( w_\diamond^{(t)} \) arises from the discrepancy between \( w_\diamond^{(t)} \) and \( w_\diamond^{(t),\star} \), which comes from the randomness in a sample-based optimization algorithm and the mismatch between the linear function and the true state-action value function. We represent the estimation error as

\[
E_{\diamond,\text{est}}^{(t)} := \mathbb{E} \left[ E_{\nu}^{\diamond(t)}(w_\diamond^{(t)}; \theta^{(t)}) - E_{\nu}^{\diamond(t)}(w_\diamond^{(t),\star}; \theta^{(t)}) \right]
\]
where the expectation $\mathbb{E}$ is taken over the randomness of approximate algorithm that is used to solve (2.27).

Note that the state-action distribution $\nu^{(t)}$ is on-policy. To characterize the effect of distribution shift on $w_{\nu^*,\nu}^{(t)}$, let us introduce some notation. We represent a fixed distribution over state-action pairs $(s, a)$ by

$$\nu^*(s, a) := d_{\rho^*}^\pi(s) \circ \text{Unif}_A(a).$$

(2.28)

The fixed distribution $\nu^*$ samples a state from $d_{\rho^*}^\pi(s)$ and an action uniformly from $\text{Unif}_A(a)$. We characterize the error in $w_{\nu^*,\nu}^{(t)}$ that arises from the distribution shift using the transfer error,

$$\mathcal{E}_{\nu^*,\nu}^{(t)} := \mathbb{E}\left[\mathcal{E}_\nu^{(t)}\right].$$

Assumption 2 (Estimation error and transfer error) Both the estimation error and the transfer error are bounded, i.e., $\mathcal{E}_{\nu^*,\nu}^{(t)} \leq \epsilon_{\text{est}}$ and $\mathcal{E}_{\nu^*,\nu}^{(t)} \leq \epsilon_{\text{bias}}$, where $\odot$ denotes $r$ or $g$.

When we apply a sample-based algorithm to (2.27), it is standard to have $\epsilon_{\text{est}} = O(1/\sqrt{K})$, where $K$ is the number of samples; e.g., see [20, Theorem 1]. A special case is the exact tabular softmax policy parametrization for which $\epsilon_{\text{bias}} = \epsilon_{\text{est}} = 0$.

For any state-action distribution $\nu$, we define $\Sigma_{\nu} := \mathbb{E}_{(s, a)} \sim \nu \left[\phi_{s, a}^\top \phi_{s, a}\right]$ and, to compare $\nu$ with $\nu^*$, we introduce the relative condition number,

$$\kappa := \sup_{w \in \mathbb{R}^d} \frac{w^\top \Sigma_{\nu^*} w}{w^\top \Sigma_{\nu_0} w}.$$

Assumption 3 (Relative condition number) For an initial state-action distribution $\nu_0$ and $\nu^*$ determined by (2.28), the relative condition number $\kappa$ is finite.

48
With the estimation error $\epsilon_{\text{est}}$, the transfer error $\epsilon_{\text{bias}}$, and the relative condition number $\kappa$ in place, in Theorem 9 we provide convergence guarantees for algorithm (2.21) using the approximate update (2.27). Even though we set $\theta^{(0)} = 0$ and $\lambda^{(0)} = 0$ in the proof of Theorem 9, global convergence can be established for arbitrary initial conditions.

**Theorem 9 (Convergence and optimality: log-linear policy)** Let Assumption 1 hold for $\xi > 0$ and let us fix a state distribution $\rho$ and a state-action distribution $\nu_0$. If the iterates $(\theta^{(t)}, \lambda^{(t)})$ generated by algorithm (2.21) and (2.27) with $\|\phi_{s,a}\| \leq B$ and $\eta_1 = \eta_2 = 1/\sqrt{T}$ satisfy Assumptions 2 and 3 then,

\[
E \left[ \frac{1}{T} \sum_{t=0}^{T-1} (V^*_r(\rho) - V_r^{(t)}(\rho)) \right] \leq \frac{C_3}{(1 - \gamma)^5} \frac{1}{\sqrt{T}} + \frac{2 + 4/\xi}{(1 - \gamma)^2} \left( \sqrt{|A|} \epsilon_{\text{bias}} + \sqrt{\frac{\kappa |A| \epsilon_{\text{est}}}{1 - \gamma}} \right)
\]

\[
E \left[ \frac{1}{T} \sum_{t=0}^{T-1} (b - V_g^{(t)}(\rho)) \right] \leq \frac{C_4}{(1 - \gamma)^4} \frac{1}{\sqrt{T}} + \left( \frac{4 + 2\xi}{1 - \gamma} \right) \left( \sqrt{|A|} \epsilon_{\text{bias}} + \sqrt{\frac{\kappa |A| \epsilon_{\text{est}}}{1 - \gamma}} \right)
\]

where $C_3 = 1 + \log |A| + 5B^2W^2/\xi$ and $C_4 = (1 + \log |A| + B^2W^2)\xi + (2 + 4B^2W^2)/\xi$.

Theorem 9 shows that, on average, the reward value function converges to its globally optimal value and that the constraint violation decays to zero (up to an estimation error $\epsilon_{\text{est}}$ and a transfer error $\epsilon_{\text{bias}}$). When $\epsilon_{\text{bias}} = \epsilon_{\text{est}} = 0$, the rate $(1/\sqrt{T}, 1/\sqrt{T})$ matches the result in Theorem 6 for the exact tabular softmax case. In contrast to the optimality gap, the lower order of effective horizon $1/(1 - \gamma)$ in the constraint violation yields a tighter error bound.

**Remark 3** In the standard error decomposition,

\[
E_{\omega}^{\nu(t)}(w_{\omega}^{(t)}; \theta^{(t)}) = E_{\omega}^{\nu(t)}(w_{\omega}^{(t)}; \theta^{(t)}) - E_{\omega}^{\nu(t)}(w_{\omega,s}^{(t)}; \theta^{(t)}) + E_{\omega}^{\nu(t)}(w_{\omega,s}^{(t)}; \theta^{(t)})
\]
the difference term is the standard estimation error that result from the discrepancy between $w_{\diamond}^{(t)}$ and $w_{\diamond,*}^{(t)}$, and the last term characterizes the approximation error in $w_{\diamond,*}^{(t)}$. In Corollary 10, we repeat Theorem 9 in terms of an upper bound $\epsilon_{\text{approx}}$ on the approximation error,

$$E_{\diamond,\text{approx}}^{(t)} := \mathbb{E} \left[ \mathcal{E}_{\diamond}^{\nu(t)} \left( w_{\diamond,*}^{(t)} ; \theta^{(t)} \right) \right].$$

Since $E_{\diamond,\text{approx}}^{(t)}$ utilizes on-policy state-action distribution $\nu^{(t)}$, the error bounds in Corollary 10 depend on the worst-case distribution mismatch coefficient $\|\nu^*/\nu_0\|_{\infty}$. In contrast, application of estimation and transfer errors in Theorem 9 does not involve the distribution mismatch coefficient. Therefore, the error bounds in Theorem 9 are tighter than the ones in Corollary 10 that utilize the standard error decomposition.

**Corollary 10 (Convergence and optimality: log-linear policy)** Let Assumption 1 hold for $\xi > 0$ and let us fix a state distribution $\rho$ and a state-action distribution $\nu_0$. If the iterates $(\theta^{(t)}, \lambda^{(t)})$ generated by algorithm (2.21) and (2.27) with $\|\phi_{s,a}\| \leq B$ and $\eta_1 = \eta_2 = 1/\sqrt{T}$ satisfy Assumption 2 except for $E_{\diamond,\text{bias}}^{(t)}$, Assumption 3 and $E_{\diamond,\text{approx}}^{(t)} \leq \epsilon_{\text{approx}}$, $\diamond = r$ or $g$, then,

$$\mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} \left( V_r^*(\rho) - V_r^{(t)}(\rho) \right) \right] \leq \frac{C_3}{(1 - \gamma)^5} \frac{1}{\sqrt{T}} + C'_3 \left( \frac{|A| \epsilon_{\text{approx}}}{1 - \gamma} \frac{\|\nu^*\|}{\nu_0} + \sqrt{\frac{\kappa |A| \epsilon_{\text{est}}}{1 - \gamma}} \right)$$

$$\mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} (b - V_g^{(t)}(\rho)) \right] \leq \frac{C_4}{(1 - \gamma)^4} \frac{1}{\sqrt{T}} + C'_4 \left( \frac{|A| \epsilon_{\text{approx}}}{1 - \gamma} \frac{\|\nu^*\|}{\nu_0} + \sqrt{\frac{\kappa |A| \epsilon_{\text{est}}}{1 - \gamma}} \right)$$

where $C_3 = 1 + \log |A| + 5B^2W^2/\xi$, $C_4 = (1 + \log |A| + B^2W^2)\xi + (2 + 4B^2W^2)/\xi$, $C'_3 = (2 + 4/\xi)/(1 - \gamma)^2$, and $C'_4 = (4 + 2\xi)/(1 - \gamma)$.
PROOF. From the definitions of $\mathcal{E}_\nu^\nu$ and $\mathcal{E}_0^{\nu(t)}$ we have

$$\mathcal{E}_\nu^\nu\left(w_\star^{(t)}; \theta^{(t)}\right) \leq \left\| \nu^* \right\|_\infty \mathcal{E}_0^{\nu(t)}\left(w_\star^{(t)}; \theta^{(t)}\right) \leq \frac{1}{1 - \gamma} \left\| \nu_0 \right\|_\infty \mathcal{E}_0^{\nu(t)}\left(w_\star^{(t)}; \theta^{(t)}\right)$$

where the second inequality is because of $(1 - \gamma)\nu_0 \leq \nu^{(t)}$. Thus,

$$\mathcal{E}_\nu^{(t)} \leq \frac{1}{1 - \gamma} \left\| \nu^* \right\|_\infty \mathcal{E}_0^{(t)}$$

which allows us to replace $\mathcal{E}_\nu^{(t)}$ in the proof of Theorem 9 by $\mathcal{E}_0^{(t)}$. □

2.5.2 Non-asymptotic convergence analysis

We first provide a regret-type analysis for our primal-dual method.

Lemma 11 (Regret/violation lemma) Let Assumption 1 hold for $\xi > 0$, let us fix a state distribution $\rho$ and $T > 0$, and let $\log \pi(a | s)$ be $\beta$-smooth in $\theta$ for any $(s, a)$. If the iterates $(\theta^{(t)}, \lambda^{(t)})$ are generated by algorithm (2.21) with $\theta^{(0)} = 0$, $\lambda^{(0)} = 0$, $\eta_1 = \eta_2 = 1/\sqrt{T}$, and $\|w^{(t)}\| \leq W$, then,

$$\frac{1}{T} \sum_{t=0}^{T-1} \left( V^* - V^{(t)} \right) \leq \frac{C_3}{(1 - \gamma)^5} \frac{1}{\sqrt{T}} + \sum_{t=0}^{T-1} \frac{\text{err}_r^{(t)}(\pi^*)}{(1 - \gamma)T} + \sum_{t=0}^{T-1} \frac{2 \times \text{err}_g^{(t)}(\pi^*)}{(1 - \gamma)^2 \xi T}$$

$$\left[ \frac{1}{T} \sum_{t=0}^{T-1} (b - V_g^{(t)}(\rho)) \right]_+ \leq \frac{C_4}{(1 - \gamma)^4} \frac{1}{\sqrt{T}} + \sum_{t=0}^{T-1} \frac{\xi \times \text{err}_r^{(t)}(\pi^*)}{T} + \sum_{t=0}^{T-1} \frac{2 \times \text{err}_g^{(t)}(\pi^*)}{(1 - \gamma)T}$$

where $C_3 = 1 + \log |A| + 5\beta W^2/\xi$, $C_4 = (1 + \log |A| + \beta W^2)\xi + (2 + 4\beta W^2)/\xi$, and

$$\text{err}_\delta^{(t)}(\pi) := \left| \mathbb{E}_{s \sim d^\rho} \mathbb{E}_{a \sim \pi(\cdot | s)} \left[ A_\delta^{(t)}(s, a) - (w_\delta^{(t)})^T \nabla_{\pi} \log \pi^{(t)}(a | s) \right] \right|$$
where $\diamond = r$ or $g$.

**Proof.** The smoothness of the log-linear policy in conjunction with an application of Taylor series expansion to $\log \pi_\theta^{(t)}(a \mid s)$ yield

$$
\log \frac{\pi_\theta^{(t)}(a \mid s)}{\pi_\theta^{(t+1)}(a \mid s)} + \left(\theta^{(t+1)} - \theta^{(t)}\right)^T \nabla_\theta \log \pi_\theta^{(t)}(a \mid s) \leq \frac{\beta}{2} \|\theta^{(t+1)} - \theta^{(t)}\|^2 \tag{2.29}
$$

where $\theta^{(t+1)} - \theta^{(t)} = \eta_t w^{(t)}/(1 - \gamma)$. Fixing $\pi$ and $\rho$, we use $d$ to denote $d_\rho^\pi$ to obtain,

$$
\mathbb{E}_{s \sim d} \left( D_{KL}(\pi(\cdot \mid s) \| \pi_\theta^{(t)}(\cdot \mid s)) - D_{KL}(\pi(\cdot \mid s) \| \pi_\theta^{(t+1)}(\cdot \mid s)) \right) \\
= - \mathbb{E}_{s \sim d} \mathbb{E}_{a \sim \pi(\cdot \mid s)} \log \frac{\pi_\theta^{(t)}(a \mid s)}{\pi_\theta^{(t+1)}(a \mid s)} \\
\geq \eta_1 \mathbb{E}_{s \sim d} \mathbb{E}_{a \sim \pi(\cdot \mid s)} \left[ \nabla_\theta \log \pi_\theta^{(t)}(a \mid s) w^{(t)} \right] - \beta \frac{\eta_1^2}{2(1 - \gamma)^2} \|w^{(t)}\|^2 \\
\overset{(b)}{=} \eta_1 \mathbb{E}_{s \sim d} \mathbb{E}_{a \sim \pi(\cdot \mid s)} \left[ \nabla_\theta \log \pi_\theta^{(t)}(a \mid s) w^{(t)} \right] \\
+ \eta_1 \lambda^{(t)} \mathbb{E}_{s \sim d} \mathbb{E}_{a \sim \pi(\cdot \mid s)} \left[ \nabla_\theta \log \pi_\theta^{(t)}(a \mid s) \frac{w^{(t)}(a \mid s)}{w^{(t)}(s,a)} \right] - \beta \frac{\eta_1^2}{2(1 - \gamma)^2} \|w^{(t)}\|^2 \\
= \eta_1 \mathbb{E}_{s \sim d} \mathbb{E}_{a \sim \pi(\cdot \mid s)} A_{r}^{(t)}(s,a) + \eta_1 \lambda^{(t)} \mathbb{E}_{s \sim d} \mathbb{E}_{a \sim \pi(\cdot \mid s)} A_{g}^{(t)}(s,a) \\
+ \eta_1 \mathbb{E}_{s \sim d} \mathbb{E}_{a \sim \pi(\cdot \mid s)} \left[ \nabla_\theta \log \pi_\theta^{(t)}(a \mid s) \frac{w^{(t)}(a \mid s)}{w^{(t)}(s,a)} - \frac{\lambda^{(t)} A_{r}^{(t)}(s,a) + \lambda^{(t)} A_{g}^{(t)}(s,a)}{w^{(t)}(s,a)} \right] \\
- \beta \frac{\eta_1^2}{(1 - \gamma)^2} \left( \|w^{(t)}\|^2 + \|\lambda^{(t)}\|^2 \|w^{(t)}\|^2 \right) \\
\overset{(c)}{=} \eta_1 (1 - \gamma) \left( V_{r}^{\pi}(\rho) - V_{r}^{(t)}(\rho) \right) + \eta_1 (1 - \gamma) \lambda^{(t)} \left( V_{g}^{\pi}(\rho) - V_{g}^{(t)}(\rho) \right) \\
- \eta_1 \text{err}^{(t)}_{r}(\pi) - \eta_1 \lambda^{(t)} \text{err}^{(t)}_{g}(\pi) - \beta \frac{\eta_1^2 W^2}{(1 - \gamma)^2} - \beta \frac{\eta_1^2 W^2}{(1 - \gamma)^2} \left( \lambda^{(t)} \right)^2
$$
where (a) is because of (2.29). On the other hand, we use the update $w^{(t)} = w_r^{(t)} + \lambda^{(t)} w_g^{(t)}$ for a given $\lambda^{(t)}$ in (b) and in (c) we apply the performance difference lemma, definitions of $\text{err}_r^{(t)}(\pi)$ and $\text{err}_g^{(t)}(\pi)$, and $\|w^{(t)}\| \leq W$. Rearrangement of the above inequality yields

$$V_r^\pi(\rho) - V_r^{(t)}(\rho)$$

$$\leq \frac{1}{1 - \gamma} \left( \frac{1}{\eta_1} \mathbb{E}_{s \sim d} \left( D_{KL}(\pi(\cdot | s) \| \pi^{(t)}(\cdot | s)) - D_{KL}(\pi(\cdot | s) \| \pi^{(t+1)}(\cdot | s)) \right) \right)$$

$$+ \frac{1}{1 - \gamma} \text{err}_r^{(t)}(\pi) + \frac{2}{(1 - \gamma)^2 \xi} \text{err}_g^{(t)}(\pi) + \frac{\eta_1 W^2}{(1 - \gamma)^3} + \frac{4 \eta_1 W^2}{(1 - \gamma)^5 \xi^2}$$

$$- \lambda^{(t)} (V_g^{(t)}(\rho) - V_g^{(t)}(\rho)).$$

where we utilize $0 \leq \lambda^{(t)} \leq 2/((1 - \gamma)\xi)$ from the dual update in (2.21).

Averaging the above inequality above over $t = 0, 1, \ldots, T - 1$ yields

$$\frac{1}{T} \sum_{t=0}^{T-1} (V_r^\pi(\rho) - V_r^{(t)}(\rho))$$

$$\leq \frac{1}{(1 - \gamma)\eta_1 T} \sum_{t=0}^{T-1} \left( \mathbb{E}_{s \sim d} \left( D_{KL}(\pi(\cdot | s) \| \pi^{(t)}(\cdot | s)) - D_{KL}(\pi(\cdot | s) \| \pi^{(t+1)}(\cdot | s)) \right) \right)$$

$$+ \frac{1}{(1 - \gamma) T} \sum_{t=0}^{T-1} \text{err}_r^{(t)}(\pi) + \frac{2}{(1 - \gamma)^2 \xi T} \sum_{t=0}^{T-1} \text{err}_g^{(t)}(\pi) + \frac{\eta_1 W^2}{(1 - \gamma)^3} + \frac{4 \eta_1 W^2}{(1 - \gamma)^5 \xi^2}$$

$$- \frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)} (V_g^{(t)}(\rho) - V_g^{(t)}(\rho))$$

which implies that,

$$\frac{1}{T} \sum_{t=0}^{T-1} (V_r^\pi(\rho) - V_r^{(t)}(\rho))$$

$$\leq \frac{\log |A|}{(1 - \gamma)\eta_1 T} + \frac{1}{(1 - \gamma) T} \sum_{t=0}^{T-1} \text{err}_r^{(t)}(\pi) + \frac{2}{(1 - \gamma)^2 \xi T} \sum_{t=0}^{T-1} \text{err}_g^{(t)}(\pi)$$

$$+ \frac{\eta_1 W^2}{(1 - \gamma)^3} + \frac{4 \eta_1 W^2}{(1 - \gamma)^5 \xi^2} + \frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)} (V_g^{(t)}(\rho) - V_g^{(t)}(\rho)).$$
If we choose the comparison policy $\pi = \pi^*$, then we have

$$\frac{1}{T} \sum_{t=0}^{T-1} (V^*_r(\rho) - V^{(t)}_r(\rho)) + \frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)}(V^*_g(\rho) - V^{(t)}_g(\rho))$$

$$\leq \frac{\log |A|}{(1 - \gamma)T} + \frac{1}{(1 - \gamma)T} \sum_{t=0}^{T-1} \text{err}^{(t)}(\pi^*) + \frac{2}{(1 - \gamma)^2} \sum_{t=0}^{T-1} \text{err}^{(t)}(\pi^*)$$

$$+ \beta \frac{\eta_1 W^2}{(1 - \gamma)^3} + \beta \frac{4\eta_1 W^2}{(1 - \gamma)^5}.$$  \hspace{1cm} (2.30)

**Proving the first inequality.** By the same reasoning as in (2.19a),

$$0 \leq \left( \lambda^{(T)} \right)^2 = \sum_{t=0}^{T-1} \left( (\lambda^{(t+1)})^2 - (\lambda^{(t)})^2 \right)$$

$$\leq 2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)} \left( b - V^{(t)}_r(\rho) \right) + \eta_2^2 \sum_{t=0}^{T-1} \left( V^{(t)}_g(\rho) - b \right)^2$$

$$\leq 2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)} \left( V^*_g(\rho) - V^{(t)}_g(\rho) \right) + \frac{\eta_2^2 T}{1 - \gamma}.$$ \hspace{1cm} (2.31a)

where \((a)\) is because of feasibility of $\pi^*$: $V^*_g(\rho) \geq b$, and $|V^{(t)}_g(\rho) - b| \leq 1/(1 - \gamma)$. Hence,

$$- \frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)} \left( V^*_g(\rho) - V^{(t)}_g(\rho) \right) \leq \frac{\eta_2}{2(1 - \gamma)^2}.$$ \hspace{1cm} (2.31b)

By adding the inequality \((2.31b)\) to \((2.30)\) on both sides and taking $\eta_1 = \eta_2 = 1/\sqrt{T}$, we obtain the first inequality.
Proving the second inequality. Since the dual update in (2.21) is the same as the one in (2.12a), we can use the same reasoning to conclude (2.20). Adding the inequality (2.20) to (2.30) on both sides and using \( V_g^*(\rho) \geq b \) yield

\[
\begin{align*}
\frac{1}{T} \sum_{t=0}^{T-1} (V_r^*(\rho) - V_r^{(t)}(\rho)) + \frac{\lambda}{T} \sum_{t=0}^{T-1} (b - V_g^{(t)}(\rho)) &
\leq \frac{\log |\mathcal{A}|}{(1 - \gamma) \eta_1 T} + \frac{1}{(1 - \gamma) T} \sum_{t=0}^{T-1} \text{err}_r^{(t)}(\pi^*) + \frac{2}{(1 - \gamma)^2 \xi T} \sum_{t=0}^{T-1} \text{err}_g^{(t)}(\pi^*) \\
&\quad + \beta \frac{\eta_1 W^2}{(1 - \gamma)^3} + \beta \frac{4 \eta_1 W^2}{(1 - \gamma)^3 \xi^2 T} + \frac{1}{2 \eta_2 T} \left| \lambda^{(0)} - \lambda^* \right|^2 + \frac{\eta_2}{2(1 - \gamma)^2}.
\end{align*}
\]

Taking \( \lambda = \frac{2}{(1 - \gamma) \xi} \) when \( \sum_{t=0}^{T-1} (b - V_g^{(t)}(\rho)) \geq 0 \) and \( \lambda = 0 \) otherwise, we obtain

\[
\begin{align*}
V_r^*(\rho) - \frac{1}{T} \sum_{t=0}^{T-1} V_r^{(t)}(\rho) + \frac{2}{(1 - \gamma) \xi} \left[ b - \frac{1}{T} \sum_{t=0}^{T-1} V_g^{(t)}(\rho) \right] &
\leq \frac{\log |\mathcal{A}|}{(1 - \gamma) \eta_1 T} + \frac{1}{(1 - \gamma) T} \sum_{t=0}^{T-1} \text{err}_r^{(t)}(\pi^*) + \frac{2}{(1 - \gamma)^2 \xi T} \sum_{t=0}^{T-1} \text{err}_g^{(t)}(\pi^*) \\
&\quad + \beta \frac{\eta_1 W^2}{(1 - \gamma)^3} + \beta \frac{4 \eta_1 W^2}{(1 - \gamma)^3 \xi^2 T} + \frac{2}{\eta_2(1 - \gamma)^2 \xi^2} + \frac{\eta_2}{2(1 - \gamma)^2}.
\end{align*}
\]

Since \( V_r^{(t)}(\rho) \) and \( V_g^{(t)}(\rho) \) are linear functions in the occupancy measure [11, Chapter 10], there exists a policy \( \pi' \) such that \( V_r^{\pi'}(\rho) = \frac{1}{T} \sum_{t=0}^{T-1} V_r^{(t)}(\rho) \) and \( V_g^{\pi'}(\rho) = \frac{1}{T} \sum_{t=0}^{T-1} V_g^{(t)}(\rho) \). Hence,

\[
\begin{align*}
V_r^*(\rho) - V_r^{\pi'}(\rho) + \frac{2}{(1 - \gamma) \xi} \left[ b - V_g^{\pi'}(\rho) \right] &
\leq \frac{\log |\mathcal{A}|}{(1 - \gamma) \eta_1 T} + \frac{1}{(1 - \gamma) T} \sum_{t=0}^{T-1} \text{err}_r^{(t)}(\pi^*) + \frac{2}{(1 - \gamma)^2 \xi T} \sum_{t=0}^{T-1} \text{err}_g^{(t)}(\pi^*) \\
&\quad + \beta \frac{\eta_1 W^2}{(1 - \gamma)^3} + \beta \frac{4 \eta_1 W^2}{(1 - \gamma)^3 \xi^2 T} + \frac{2}{\eta_2(1 - \gamma)^2 \xi^2} + \frac{\eta_2}{2(1 - \gamma)^2}.
\end{align*}
\]
Application of Lemma 2 with $\frac{2}{((1-\gamma)\xi)} \geq 2\lambda^*$ yields

$$[b - V_g^\pi'(\rho)]_+ \leq \frac{\xi \log |\mathcal{A}|}{\eta_1 T} + \frac{\xi}{T} \sum_{t=0}^{T-1} \text{err}_r^{(t)}(\pi^*) + \frac{2}{(1-\gamma)T} \sum_{t=0}^{T-1} \text{err}_g^{(t)}(\pi^*)$$

$$+ \beta \frac{\eta_1 \xi W^2}{(1-\gamma)^2} + \beta \frac{4\eta_1 W^2}{(1-\gamma)^4 \xi} + \frac{2}{\eta_2 (1-\gamma) \xi T} + \frac{\eta_2 \xi}{2(1-\gamma)}.$$

which leads to our constraint violation bound if we further utilize $\frac{1}{T} \sum_{t=0}^{T-1} (b - V_g^{(t)}(\rho)) = b - V_g^\pi'(\rho)$ and $\eta_1 = \eta_2 = 1/\sqrt{T}$.

□

PROOF. [Proof of Theorem 9]

When $\|\phi_{s,a}\| \leq B$, for the log-linear policy class, $\log \pi_\theta(a \mid s)$ is $\beta$-smooth with $\beta = B^2$. By Lemma 11, it remains to consider the randomness in sequences of $w^{(t)}$ and the error bounds for $\text{err}_r^{(t)}(\pi^*)$. Application of the triangle inequality yields

$$\text{err}_r^{(t)}(\pi^*) \leq \left| \mathbb{E}_{s \sim d^*_\rho} \mathbb{E}_{a \sim \pi^*(\cdot \mid s)} \left[ A^{(t)}_r (s, a) - (w^{(t)}_{r,\star})^\top \nabla_\theta \log \pi^{(t)}(a \mid s) \right] \right|$$

$$+ \left| \mathbb{E}_{s \sim d^*_\rho} \mathbb{E}_{a \sim \pi^*(\cdot \mid s)} \left[ (w^{(t)}_{r,\star} - w^{(t)}_r)^\top \nabla_\theta \log \pi^{(t)}(a \mid s) \right] \right|. \quad (2.32)$$
Application of (2.26) and $A_r(t)(s, a) = Q_r(t)(s, a) - \mathbb{E}_{a' \sim \pi^t(\cdot | s)} Q_r(t)(s, a')$ yields

$$
\mathbb{E}_{s \sim d^\nu_s} \mathbb{E}_{a \sim \pi^t(\cdot | s)} \left[ A_r(t)(s, a) - (w_{r,s}^{(t)})^\top \nabla_\theta \log \pi^t(a | s) \right] = \mathbb{E}_{s \sim d^\nu_s} \mathbb{E}_{a \sim \pi^t(\cdot | s)} \left[ Q_r(t)(s, a) - \phi_{s,a} w_{r,s}^{(t)} \right] - \mathbb{E}_{s \sim d^\nu_s} \mathbb{E}_{a' \sim \pi^t(\cdot | s)} \left[ Q_r(t)(s, a') - \phi_{s,a'} w_{r,s}^{(t)} \right] \leq \sqrt{\mathbb{E}_{s \sim d^\nu_s} \mathbb{E}_{a \sim \pi^t(\cdot | s)} \left( Q_r(t)(s, a) - \phi_{s,a} w_{r,s}^{(t)} \right)^2} + \sqrt{\mathbb{E}_{s \sim d^\nu_s} \mathbb{E}_{a' \sim \pi^t(\cdot | s)} \left( Q_r(t)(s, a') - \phi_{s,a'} w_{r,s}^{(t)} \right)^2} \leq 2 \sqrt{|A|} \mathbb{E}_{s \sim d^\nu_s} \mathbb{E}_{a \sim \text{Unif}_A} \left[ \left( Q_r(t)(s, a) - \phi_{s,a} w_{r,s}^{(t)} \right)^2 \right] = 2 \sqrt{|A|} \mathbb{E}_{\nu^*}(w_{r,s}^{(t)}; \theta(t)).
$$

Similarly,

$$
\mathbb{E}_{s \sim d^\nu_s} \mathbb{E}_{a \sim \pi^t(\cdot | s)} \left[ (w_{r,s}^{(t)} - w_{r}^{(t)})^\top \nabla_\theta \log \pi^t(a | s) \right] = \mathbb{E}_{s \sim d^\nu_s} \mathbb{E}_{a \sim \pi^t(\cdot | s)} \left[ (w_{r,s}^{(t)} - w_{r}^{(t)})^\top \phi_{s,a} \right] - \mathbb{E}_{s \sim d^\nu_s} \mathbb{E}_{a' \sim \pi^t(\cdot | s)} \left[ (w_{r,s}^{(t)} - w_{r}^{(t)})^\top \phi_{s,a'} \right] \leq 2 \sqrt{|A|} \mathbb{E}_{s \sim d^\nu_s} \mathbb{E}_{a \sim \text{Unif}_A} \left[ \left( (w_{r,s}^{(t)} - w_{r}^{(t)})^\top \phi_{s,a} \right)^2 \right] = 2 \sqrt{|A|} \|w_{r,s}^{(t)} - w_{r}^{(t)}\|^2_{\Sigma^*_{\nu^*}},
$$

where $\Sigma^*_{\nu^*} := \mathbb{E}_{(s,a) \sim \nu^*} \left[ \phi_{s,a} \phi_{s,a}^\top \right]$. From the definition of $\kappa$ we have

$$
\|w_{r,s}^{(t)} - w_{r}^{(t)}\|^2_{\Sigma^*_{\nu^*}} \leq \kappa \|w_{r,s}^{(t)} - w_{r}^{(t)}\|^2_{\Sigma_{\nu^*}} \leq \kappa \|w_{r,s}^{(t)} - w_{r}^{(t)}\|^2_{\Sigma_{\nu^*}} \leq \frac{\kappa}{1 - \gamma} \|w_{r,s}^{(t)} - w_{r}^{(t)}\|^2_{\Sigma_{\nu^*}}.
$$

(2.35)
where we use $(1 - \gamma)\nu_0 \leq \nu_{\nu_0}^{(t)} := \nu^{(t)}$ in the second inequality. Evaluation of the first-order optimality condition of $w^{(t)}_{r,\ast} \in \arg\min_{\|w_r\|_2 \leq \mathcal{W}} \mathcal{E}_r^{\nu^{(t)}}(w_r; \theta^{(t)})$ yields

\[
(w_r - w^{(t)}_{r,\ast})^\top \nabla_\theta \mathcal{E}_r^{\nu^{(t)}}(w^{(t)}_{r,\ast}, \theta^{(t)}) \geq 0, \text{ for any } w_r \text{ satisfying } \|w_r\| \leq \mathcal{W}.
\]

Thus,

\[
\mathcal{E}_r^{\nu^{(t)}}(w_r; \theta^{(t)}) - \mathcal{E}_r^{\nu^{(t)}}(w^{(t)}_{r,\ast}; \theta^{(t)}) = \mathbb{E}_{s,a \sim \nu^{(t)}} \left[ \left( Q_r^{(t)}(s, a) - \phi_{s,a}^\top w^{(t)}_{r,\ast} + \phi_{s,a}^\top w^{(t)}_{r} - \phi_{s,a}^\top w_r \right)^2 \right] - \mathcal{E}_r^{\nu^{(t)}}(w^{(t)}_{r,\ast}; \theta^{(t)})
\]

\[
= 2 \left( w^{(t)}_{r,\ast} - w_r \right)^\top \mathbb{E}_{s,a \sim \nu^{(t)}} \left[ \left( Q_r^{(t)}(s, a) - \phi_{s,a}^\top w^{(t)}_{r,\ast} \right) \phi_{s,a} \right]
\]

\[
+ \mathbb{E}_{s,a \sim \nu^{(t)}} \left[ \left( \phi_{s,a}^\top w^{(t)}_{r,\ast} - \phi_{s,a}^\top w_r \right)^2 \right]
\]

\[
= \left( w_r - w^{(t)}_{r,\ast} \right)^\top \nabla_\theta \mathcal{E}_r^{\nu^{(t)}}(w^{(t)}_{r,\ast}; \theta^{(t)}) + \|w_r - w^{(t)}_{r,\ast}\|_2^2.
\]

Taking $w_r = w^{(t)}_{r,\ast}$ in the above inequality and combining it with (2.34) and (2.35), yield

\[
\mathbb{E}_{s \sim d^*_r} \mathbb{E}_{a \sim \pi^{(t)}(\cdot | s)} \left[ \left( w^{(t)}_{r,\ast} - w^{(t)}_r \right)^\top \nabla_\theta \log \pi^{(t)}(a | s) \right] \leq 2 \sqrt{\frac{\kappa |A|}{1 - \gamma}} \left( \mathcal{E}_r^{\nu^{(t)}}(w^{(t)}_r; \theta^{(t)}) - \mathcal{E}_r^{\nu^{(t)}}(w^{(t)}_{r,\ast}; \theta^{(t)}) \right).
\]

(2.36)

Substitution of (2.33) and (2.36) into the right-hand side of (2.32) yields

\[
\mathbb{E} \left[ err_r^{(t)}(\pi^\ast) \right] \leq 2 \sqrt{|A|} \mathbb{E} \left[ \mathcal{E}_r^{\nu^{(t)}(w^{(t)}_{r,\ast}; \theta^{(t)})} \right] + 2 \sqrt{\frac{\kappa |A|}{1 - \gamma}} \mathbb{E} \left[ \mathcal{E}_r^{\nu^{(t)}}(w^{(t)}_r; \theta^{(t)}) - \mathcal{E}_r^{\nu^{(t)}}(w^{(t)}_{r,\ast}; \theta^{(t)}) \right].
\]
By the same reasoning, we can establish a similar bound on $E[\text{err}^{(t)}_\nu(\pi^*)]$. Finally, our desired results follow by applying Assumption 2 and Lemma 11.

2.5.3 General smooth policy class

For a general class of smooth policies [276, 8], we now establish convergence of algorithm (2.21) with approximate gradient update,

$$w^{(t)} = w^{(t)}_r + \lambda^{(t)} w^{(t)}_g$$

$$w^{(t)}_o \approx \arg\min_{\|w_o\|_2 \leq W} E^{(t)}_o(w_o; \theta^{(t)})$$

(2.37)

where $\circ$ denotes $r$ or $g$ and the exact minimizer is given by $w^{(t)}_{o,*} \in \arg\min_{\|w_o\|_2 \leq W} E^{(t)}_o(w_{o,*}; \theta^{(t)})$.

Assumption 4 (Policy smoothness) For all $s \in S$ and $a \in A$, log $\pi_\theta(a \mid s)$ is a $\beta$-smooth function of $\theta$,

$$\|\nabla_\theta \log \pi_\theta(a \mid s) - \nabla_{\theta'} \log \pi_{\theta'}(a \mid s)\| \leq \beta \|\theta - \theta'\| \text{ for all } \theta, \theta' \in \mathbb{R}^d.$$ 

Since both tabular softmax and log-linear policies satisfy Assumption 4 [8], Assumption 4 covers a broader function class relative to softmax policy parametrization (2.5).

Given a state-action distribution $\nu^{(t)}$, we introduce the estimation error as

$$E^{(t)}_{o,\text{est}} := \mathbb{E} \left[ E^{(t)}_o(w^{(t)}_{o,*}; \theta^{(t)}) - E^{(t)}_o(w^{(t)}_{o}; \theta^{(t)}) \mid \theta^{(t)} \right].$$
Furthermore, given a state distribution $\rho$ and an optimal policy $\pi^*$, we define a state-action distribution $\nu^*(s, a) := d^*_{\rho}(s)\pi^*(a \mid s)$ as a comparator and introduce the transfer error,

$$E_{\phi, \text{bias}}^{(t)} := \mathbb{E} \left[ E_{\phi}^{\nu^*}(w_{\phi,a}^{(t)}, \theta^{(t)}) \right].$$

Finally, for any state-action distribution $\nu$, we define

$$\Sigma_{\nu}^\theta = \mathbb{E}_{(s,a) \sim \nu} \left[ \nabla_{\theta} \log \pi_{\theta}(a \mid s) \left( \nabla_{\theta} \log \pi_{\theta}(a \mid s) \right)^\top \right]$$

and use $\Sigma_{\nu}^{(t)}$ to denote $\Sigma_{\nu}^{\theta^{(t)}}$.

**Assumption 5 (Estimation/transfer errors and relative condition number)** The above estimation and transfer errors as well as the expected relative condition number are bounded, i.e.,

$$E_{\phi, \text{est}}^{(t)} \leq \epsilon_{\text{est}} \text{ and } E_{\phi, \text{bias}}^{(t)} \leq \epsilon_{\text{bias}}, \text{ for } \phi = r \text{ or } g,$$

and

$$\mathbb{E} \left[ \sup_{w \in \mathbb{R}^d} \frac{w^\top \Sigma_{\nu}^{(t)} w}{w^\top \Sigma_{\nu_0} w} \right] \leq \kappa.$$

We next provide convergence guarantees for algorithm (2.21) in Theorem 12 using the approximate update (2.37). Even though we set $\theta^{(0)} = 0$ and $\lambda^{(0)} = 0$ in the proof of Theorem 12, convergence can be established for arbitrary initial conditions.

**Theorem 12 (Convergence and optimality: general smooth policy)** Let us fix a state distribution $\rho$, a state-action distribution $\nu_0$, and $T > 0$, and let Assumptions 1 and 4 hold. If the iterates
\((\theta^{(t)}, \lambda^{(t)})\) generated by algorithm (2.21) and (2.37) with \(\eta_1 = \eta_2 = 1/\sqrt{T}\) satisfy Assumption 5 and \(\|w_\circ^{(t)}\| \leq W\), then,

\[
\mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} (V_\circ^*(\rho) - V_\circ^{(t)}(\rho)) \right] \leq \frac{C_3}{(1-\gamma)^5} \frac{1}{\sqrt{T}} + \frac{1 + 2/\xi}{(1-\gamma)^2} \left( \sqrt{\epsilon_{\text{bias}}} + \sqrt{\kappa \epsilon_{\text{est}}} \right)
\]

\[
\mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} \left( b - V_g^{(t)}(\rho) \right) \right] \leq \frac{C_4}{(1-\gamma)^4} \frac{1}{\sqrt{T}} + \frac{2 + \frac{2}{\xi}}{1-\gamma} \left( \sqrt{\epsilon_{\text{bias}}} + \sqrt{\kappa \epsilon_{\text{est}}} \right)
\]

where \(C_3 = 1 + \log |A| + 5\beta W^2 / \xi\) and \(C_4 = (1 + \log |A| + \beta W^2)\xi + (2 + 4\beta W^2) / \xi\).

**Proof.** Since Lemma 11 holds for any smooth policy class that satisfies Assumption 4, it remains to bound \(\text{err}_\circ^{(t)}(\pi^*)\) for \(\circ = r\) or \(g\). We next separately bound each term on the right-hand side of (2.32). For the first term,

\[
\mathbb{E}_{s \sim d_\rho^t} \mathbb{E}_{a \sim \pi_r^*(\cdot | s)} \left[ A_r^{(t)}(s, a) - (w_r^{(t)})^\top \nabla_\theta \log \pi_\theta^{(t)}(a | s) \right] \leq \sqrt{ \mathbb{E}_{s \sim d_\rho^t} \mathbb{E}_{a \sim \pi_r^*(\cdot | s)} \left( A_r^{(t)}(s, a) - (w_r^{(t)})^\top \nabla_\theta \log \pi_\theta^{(t)}(a | s) \right)^2 } \tag{2.38}
\]

Similarly,

\[
\mathbb{E}_{s \sim d_\rho^t} \mathbb{E}_{a \sim \pi_r^*(\cdot | s)} \left[ (w_r^{(t)} - w_r^{(t)})^\top \nabla_\theta \log \pi_\theta^{(t)}(a | s) \right] \leq \sqrt{ \mathbb{E}_{s \sim d_\rho^t} \mathbb{E}_{a \sim \pi_r^*(\cdot | s)} \left[ (w_r^{(t)} - w_r^{(t)})^\top \nabla_\theta \log \pi_\theta^{(t)}(a | s) \right]^2 } \tag{2.39a}
\]
Let \( \kappa(t) := \left\| (\Sigma^{(t)}_{\nu_0})^{-1/2} \Sigma^{(t)}_{\nu^*} (\Sigma^{(t)}_{\nu_0})^{-1/2} \right\|_2 \) be the relative condition number at time \( t \). Thus,

\[
\| w_{r,*}^{(t)} - w_r^{(t)} \|_{\Sigma^*_{\nu^*}}^2 \leq \| (\Sigma^{(t)}_{\nu_0})^{-1/2} \Sigma^{(t)}_{\nu^*} (\Sigma^{(t)}_{\nu_0})^{-1/2} \| \| w_{r,*}^{(t)} - w_r^{(t)} \|_{\Sigma^*_{\nu_0}}^2
\]

\[
\leq \left( \frac{\kappa(t)}{1 - \gamma} \right) \| w_{r,*}^{(t)} - w_r^{(t)} \|_{\Sigma^*_{\nu(t)}}^2
\]

\[
\leq \left( \frac{\kappa(t)}{1 - \gamma} \right) \left( E_{\nu}^{(t)}(w_r^{(t)}; \theta(t)) - E_{\nu}^{(t)}(w_{r,*}^{(t)}; \theta(t)) \right)
\]

where we use \((1 - \gamma)\nu_0 \leq \nu^{(t)}_{\nu_0} := \nu^{(t)}\) in (a), and we get (b) by the same reasoning as bounding (2.35). Taking an expectation over the inequality above from both sides yields

\[
\mathbb{E}\left[ \| w_{r,*}^{(t)} - w_r^{(t)} \|_{\Sigma^*_{\nu^*}}^2 \right] \leq \mathbb{E}\left[ \frac{\kappa(t)}{1 - \gamma} \mathbb{E}\left[ E_{\nu}^{(t)}(w_r^{(t)}; \theta(t)) - E_{\nu}^{(t)}(w_{r,*}^{(t)}; \theta(t)) | \theta(t) \right] \right]
\]

\[
\leq \mathbb{E}\left[ \frac{\kappa(t)}{1 - \gamma} \right] \epsilon_{est}
\]

\[
\leq \frac{\kappa \epsilon_{est}}{1 - \gamma}
\]

where the last two inequalities are because of Assumption [5]

Substitution of (2.38) and (2.39) to the right-hand side of (2.32) yields an upper bound on \( \mathbb{E}\left[ \text{err}_r^{(t)}(\pi^*) \right] \). By the same reasoning, we can establish a similar bound on \( \mathbb{E}\left[ \text{err}_r^{(t)}(\pi^*) \right] \). Finally, application of these upper bounds to Lemma [11] yields the desired result. \(\square\)
2.6 Sample-based algorithms

We now leverage convergence results established in Theorems 9 and 12 to design two model-free algorithms that utilize sample-based estimates. In particular, we propose a sample-based extension of NPG-PD algorithm (2.21) with function approximation and 

\[ \Lambda = \left[ 0, \frac{2}{((1 - \gamma) \xi)} \right] \]

\[
\begin{align*}
\theta(t+1) &= \theta(t) + \frac{\eta}{1 - \gamma} \hat{w}(t) \\
\lambda(t+1) &= \mathcal{P}_\Lambda \left( \lambda(t) - \eta_2 \left( \hat{V}_g(t)(\rho) - b \right) \right)
\end{align*}
\]

(2.40)

where \( \hat{w}(t) \) and \( \hat{V}_g(t)(\rho) \) are the sample-based estimates of the gradient and the value function. At each time \( t \), we can access constrained MDP environment by executing a policy \( \pi \) with terminating probability \( 1 - \gamma \). For the minimization problem in (2.37), we can run stochastic gradient descent (SGD) for \( K \) rounds, \( w_{o,k+1} = \mathcal{P}_{\|w_{o,k}\| \leq W} (w_{o,k} - \alpha G_{o,k}) \). Here, \( G_{o,k} \) is a sample-based estimate of the population gradient \( \nabla_\theta E_{o}^{(t)}(w_o; \theta(t)) \),

\[
G_{o,k} = 2 \left( (w_{o,k})^T \nabla_\theta \log \pi_{o}^{(t)}(a \mid s) - \hat{A}_{o}^{(t)}(s, a) \right) \nabla_\theta \log \pi_{o}^{(t)}(a \mid s)
\]

\( \hat{A}_{o}^{(t)}(s, a) := \hat{Q}_{o}^{(t)}(s, a) - \hat{V}_{o}^{(t)}(s) \), \( \hat{Q}_{o}^{(t)}(s, a) \) and \( \hat{V}_{o}^{(t)}(s) \) are undiscounted sums that are collected in Algorithm 2. In addition, we estimate \( \hat{V}_g(t)(\rho) \) using an undiscounted sum in Algorithm 3. As shown in Appendix A.6, \( G_{o,k}, \hat{A}_{o}^{(t)}(s, a), \) and \( \hat{V}_g(t)(\rho) \) are unbiased estimates and we approximate gradient using the average of the SGD iterates \( \hat{w}(t) = K^{-1} \sum_{k=1}^{K} (w_{r,k} + \lambda(t)w_{g,k}) \), which is an approximate solution for least-squares regression [20, Theorem 1].
Algorithm 1 Sample-based NPG-PD algorithm with general policy parametrization

1: **Initialization:** Learning rates $\eta_1$ and $\eta_2$, number of SGD iterations $K$, SGD learning rate $\alpha$.
2: Initialize $\theta^{(0)} = 0, \lambda^{(0)} = 0$.
3: for $t = 0, \ldots, T - 1$ do
4: Initialize $w_{r,0} = w_{g,0} = 0$.
5: for $k = 0, 1, \ldots, K - 1$ do
6: Estimate $\hat{A}_r(s, a)$ and $\hat{A}_g(s, a)$ for some $(s, a) \sim \nu^{(t)}$, using Algorithm 2 with policy $\pi^{(t)}_\theta$.
7: Take a step of SGD,
   \[
   w_{r,k+1} = P_{\|w_r\| \leq W} \left( w_{r,k} - 2\alpha \left( (w_{r,k})^T \nabla_\theta \log \pi^{(t)}_\theta(s, a) - \hat{A}_r^{(t)}(s, a) \right) \nabla_\theta \log \pi^{(t)}_\theta(s, a) \right),
   \]
   \[
   w_{g,k+1} = P_{\|w_g\| \leq W} \left( w_{g,k} - 2\alpha \left( (w_{g,k})^T \nabla_\theta \log \pi^{(t)}_\theta(s, a) - \hat{A}_g^{(t)}(s, a) \right) \nabla_\theta \log \pi^{(t)}_\theta(s, a) \right).
   \]
8: end for
9: Set $\hat{w}^{(t)} = \frac{1}{K} \sum_{k=0}^{K-1} w_{r,k}$ and $\hat{w}_g^{(t)} = \frac{1}{K} \sum_{k=0}^{K-1} w_{g,k}$.
10: Estimate $\hat{V}_g^{(t)}(\rho)$ using Algorithm 3 with policy $\pi^{(t)}_\theta$.
11: Natural policy gradient primal-dual update
   \[
   \theta^{(t+1)} = \theta^{(t)} + \eta_1 \hat{w}^{(t)}
   \]
   \[
   \lambda^{(t+1)} = P_{[0, 2/((1-\gamma)\xi)]} \left( \lambda^{(t)} - \eta_2 \left( \hat{V}_g^{(t)}(\rho) - b \right) \right).
   \]
12: end for

Algorithm 2 $A$-Unbiased estimate ($A_{est}^g$, $\diamond = r$ or $g$)

1: **Input:** Initial state-action distribution $\nu_0$, policy $\pi$, discount factor $\gamma$.
2: Sample $(s_0, a_0) \sim \nu_0$, execute the policy $\pi$ with probability $\gamma$ at each step $h$; otherwise, accept $(s_h, a_h)$ as the sample.
3: Start with $(s_h, a_h)$, execute the policy $\pi$ with the termination probability $1 - \gamma$. Once terminated, add all rewards/utilities from step $h$ onwards as $\hat{Q}^g_\diamond(s_h, a_h)$ for $\diamond = r$ or $g$.
4: Start with $s_h$, sample $a_h' \sim \pi(\cdot | s_h)$, and execute the policy $\pi$ with the termination probability $1 - \gamma$. Once terminated, add all rewards/utilities from step $h$ onwards as $\hat{V}^\pi_\diamond(s_h)$ for $\diamond = r$ or $g$.
5: **Output:** $(s_h, a_h)$ and $\hat{A}^g_\diamond(s_h, a_h) := \hat{Q}^g_\diamond(s_h, a_h) - \hat{V}^\pi_\diamond(s_h), \diamond = r$ or $g$.

Algorithm 3 $V$-Unbiased estimate ($\hat{V}^{\pi}_{est}$)

1: **Input:** Initial state distribution $\rho$, policy $\pi$, discount factor $\gamma$.
2: Sample $s_0 \sim \rho$, execute the policy $\pi$ with the termination probability $1 - \gamma$. Once terminated, add all utilities up as $\hat{V}^\pi_\diamond(\rho)$.
3: **Output:** $\hat{V}^\pi_\diamond(\rho)$.
2.6.1 Sample complexity

To establish sample complexity of Algorithm 1, we assume the score function $\nabla_\theta \log \pi(a \mid s)$ has bounded norm $[276, 8]$. 

Assumption 6 (Lipschitz policy) For $0 \leq t < T$, the policy $\pi^{(t)}$ satisfies

$$\|\nabla_\theta \log \pi^{(t)}(a \mid s)\| \leq L_\pi,$$ where $L_\pi > 0$.

Under Assumption 6, sample-based estimate of SGD gradient is bounded by $G := 2L_\pi(WL_\pi + 1/(1 - \gamma))$ and, in Theorem 13, we establish sample complexity of Algorithm 1.

Theorem 13 (Sample complexity: general smooth policy) Let Assumptions 1, 4, and 6 hold and let us fix a state distribution $\rho$, a state-action distribution $\nu_0$, and $T > 0$. If the iterates $(\theta^{(t)}, \lambda^{(t)})$ are generated by the sample-based NPG-PD method described in Algorithm 1 with $\eta_1 = \eta_2 = 1/\sqrt{T}$ and $\alpha = W/(G\sqrt{K})$, in which $K$ rounds of trajectory samples are used at each time $t$, then,

$$\mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} \left( V^*_r(\rho) - V^{(t)}_r(\rho) \right) \right] \leq \frac{C_5}{(1 - \gamma)^5} \frac{1}{\sqrt{T}} + \frac{1 + 2/\xi}{(1 - \gamma)^2} \left( \sqrt{\epsilon_{\text{bias}}} + \sqrt{\frac{\kappa GW}{(1 - \gamma)\sqrt{K}}} \right)$$

$$\mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} \left( b - V^{(t)}_g(\rho) \right) \right] \leq \frac{C_6}{(1 - \gamma)^4} \frac{1}{\sqrt{T}} + \frac{2 + \xi}{1 - \gamma} \left( \sqrt{\epsilon_{\text{bias}}} + \sqrt{\frac{\kappa GW}{(1 - \gamma)\sqrt{K}}} \right)$$

where $C_5 = 2 + \log |A| + 5\beta W^2/\xi$ and $C_6 = (2 + \log |A| + \beta W^2)\xi + (2 + 4\beta W^2)/\xi$.

In Theorem 13 the sampling effect appears as an error rate $1/K^{1/4}$, where $K$ is the size of sampled trajectories. This rate follows the standard SGD result [197, Theorem 14.8] and it can be improved to $1/\sqrt{K}$ under additional restrictions on the dataset [101, 58]. The proof of Theorem 13
in Appendix A.7 follows the proof of Theorem 12 except that we use sample-based estimates of
gradients in the primal update and sample-based value functions in the dual update.

Algorithm 4 Sample-based NPG-PD algorithm with log-linear policy parametrization

1: **Input**: Learning rates $\eta_1$ and $\eta_2$, number of SGD iterations $K$, SGD learning rate $\alpha$.
2: Initialize $\theta^{(0)} = 0$, $\lambda^{(0)} = 0$.
3: **for** $t = 0, \ldots, T - 1$ **do**
4: Initialize $w_{r,0} = w_{g,0} = 0$.
5: **for** $k = 0, 1, \ldots, K - 1$ **do**
6: Estimate $\hat{Q}_r^{(t)}(s, a)$ and $\hat{Q}_g^{(t)}(s, a)$ for some $(s, a) \sim \nu^{(t)}$, using Algorithm 5 with log-linear policy $\pi^{(t)}$.
7: Take a step of SGD,
   
   
   $$
   w_{r,k+1} = \mathcal{P}_{\|w_r\| \leq W} \left( w_{r,k} - 2\alpha \left( \phi_{s,a}^\top w_{r,k} - \hat{Q}_r^{(t)}(s, a) \right) \phi_{s,a} \right)
   $$

   $$
   w_{g,k+1} = \mathcal{P}_{\|w_g\| \leq W} \left( w_{g,k} - 2\alpha \left( \phi_{s,a}^\top w_{g,k} - \hat{Q}_g^{(t)}(s, a) \right) \phi_{s,a} \right).
   $$

8: **end for**
9: Set $\hat{w}^{(t)} = \hat{w}_r^{(t)} + \lambda^{(t)} \hat{w}_g^{(t)}$, where $\hat{w}_r^{(t)} = \frac{1}{K} \sum_{k=0}^{K-1} w_{r,k}$ and $\hat{w}_g^{(t)} = \frac{1}{K} \sum_{k=0}^{K-1} w_{g,k}$.
10: Estimate $\hat{V}_g^{(t)}(\rho)$ using Algorithm 3 with log-linear policy $\pi_g^{(t)}$.
11: Natural policy gradient primal-dual update
   
   $$
   \theta^{(t+1)} = \theta^{(t)} + \frac{\eta_1}{1 - \gamma} \hat{w}^{(t)}
   $$

   $$
   \lambda^{(t+1)} = \mathcal{P}_{[0, 1/(1-\gamma))] \left( \lambda^{(t)} - \eta_2 \left( \hat{V}_g^{(t)}(\rho) - b \right) \right).
   $$

(2.41)

12: **end for**

Algorithm 5 $Q$-Unbiased estimate ($\hat{Q}_{\phi}^{\text{est}}, \diamond = r$ or $g$)

1: **Input**: Initial state-action distribution $\nu_0$, policy $\pi$, discount factor $\gamma$.
2: Sample $(s_0, a_0) \sim \nu_0$, execute the policy $\pi$ with probability $\gamma$ at each step $h$; otherwise, accept $(s_h, a_h)$ as the sample.
3: Start with $(s_h, a_h)$, execute the policy $\pi$ with the termination probability $1 - \gamma$. Once terminated, add all rewards/utilities from step $h$ onwards as $\hat{Q}_\phi^{\pi}(s_h, a_h)$ for $\diamond = r$ or $g$, respectively.
4: **Output**: $(s_h, a_h)$ and $\hat{Q}_\phi^{\pi}(s_h, a_h), \diamond = r$ or $g$.

Algorithm 4 is utilized for log-linear policy parametrization. For the feature $\phi_{s,a}$ that has
bounded norm $\|\phi_{s,a}\| \leq B$, the sample-based gradient in SGD is bounded by $G := 2B(WB +
1/(1 − γ)). In Theorem 14 we establish sample complexity of Algorithm 4; see Appendix A.8 for proof.

**Theorem 14 (Sample complexity: log-linear policy)** Let Assumption 1 hold and let us fix a state distribution \( \rho \) and a state-action distribution \( \nu_0 \). If the iterates \((\theta(t), \lambda(t))\) generated by the sample-based NPG-PD method described in Algorithm 4 with \( \| \phi_{s,a} \| \leq B \), \( \eta_1 = \eta_2 = 1/\sqrt{T} \), and \( \alpha = W/(G \sqrt{K}) \), in which \( K \) rounds of trajectory samples are used at each time \( t \), then,

\[
\begin{align*}
\mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} (V^*_r(\rho) - V^{(t)}_r(\rho))\right] & \leq \frac{C_5}{(1-\gamma)^5 \sqrt{T}} + \frac{2 + 4/\xi}{(1-\gamma)^2} \left( \sqrt{\|A\|} \epsilon_{\text{bias}} + \frac{\kappa \|A\|GW}{(1-\gamma)\sqrt{K}} \right) \\
\mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} (b - V^{(t)}_g(\rho))\right] & \leq \frac{C_6}{(1-\gamma)^4 \sqrt{T}} + \frac{4 + 2\xi}{1-\gamma} \left( \sqrt{\|A\|} \epsilon_{\text{bias}} + \frac{\kappa \|A\|GW}{(1-\gamma)\sqrt{K}} \right)
\end{align*}
\]

where \( C_5 = 2 + \log |A| + 5\beta W^2/\xi \) and \( C_6 = (2 + \log |A| + \beta W^2)\xi + (2 + 4\beta W^2)/\xi \).

When we specialize the log-linear policy to be the softmax policy, Algorithm 4 becomes a sample-based implementation of NPG-PD method (2.12) that utilizes the state-action value functions. In this case, \( \epsilon_{\text{bias}} = 0 \) and \( B = 1 \) in Theorem 14. When there are no sampling effects, i.e., as \( K \rightarrow \infty \), our rate \((1/\sqrt{T}, 1/\sqrt{T})\) matches the rate in Theorem 6.

### 2.7 Computational experiments

We use two examples of robotic tasks with constraints to demonstrate the merits and the effectiveness of our sample-based NPG-PD method described in Algorithm 1. The first example involves robots with speed limit tasks [280] and the second example rewards robots for running in a circle while staying in a safe region [7]. The robotic environments are implemented using the OpenAI Gym [43] for the MuJoCo physical simulators [221].
We compare performance of our NPG-PD algorithm with First Order Constrained Optimization in Policy Space (FOCOPS) algorithm [280], an approach that provides the state-of-the-art performance for constrained robotic tasks. The policy is represented as a Gaussian distribution, where the mean action is parametrized by a two-layer neural network with the tanh activation and the state-independent logarithmic standard deviation is computed separately from the mean action. To have a fair comparison, we instantiate subroutines of Algorithms 2 and 3 by fitting a two-layer neural network to estimate the value function and we implement lines 5–8 of Algorithm 1 with $K = 8192$ by solving a regularized empirical risk minimization problem to reduce variance. We also use the FOCOPS’ hyperparameters [280, Table 3] as our default hyperparameters.

In the first example, the goal of a robot is to move along either a line or in a plane while satisfying a speed limit constraint [280]. We train six MuJoCo robotic agents to walk: Hopper-v3, Swimmer-v3, HalfCheetah-v3, Walker2d-v3, Ant-v3, and Humanoid-v3, while we constrain the moving speed to be under a given threshold. Figure 2.2 shows that in the first three tasks, our NPG-PD algorithm achieves higher rewards than the baseline FOCOPS algorithm while achieving similar constraint satisfaction cost. In the second and third tasks, we observe oscillatory response that arises from dual updates in NPG-PG algorithm [210]. For the last three tasks, Figure 2.3 shows a competitive performance of NPG-PD with FOCOPS. In Humanoid-v3, even though oscillations slow down the convergence of NPG-PD, it achieves higher rewards than FOCOPS in spite of early oscillatory behavior.

In the second example, the robot aims to move along a circular trajectory while remaining within a safe region [7, 280]. For Humanoid Circle-v0, Figure 2.4 shows slow initial response of NPG-PD compared to FOCOPS. We suspect that this is because of incremental update
of the dual variable which does not produce sufficient penalty for reducing constraint violation. As the dual variable (or the average cost) approaches a stationary point, the average reward converges quickly. In contrast, for Ant Circle-v0, NPG-PD achieves a much higher average reward than FOCOPS.

2.8 Concluding remarks

We have proposed a Natural Policy Gradient Primal-Dual (NPG-PD) algorithm for solving optimal control problems for constrained MDPs. Our algorithm utilizes natural policy gradient ascent to update the primal variable and projected sub-gradient descent to update the dual variable. Although the underlying maximization involves a nonconcave objective function and a nonconvex constraint set, we have established global convergence for softmax, log-linear, and general smooth policy parametrizations and have provided finite-sample complexity guarantees for two model-free extensions of the NPG-PD algorithm. To the best of our knowledge, our work is the first to offer finite-time performance guarantees for policy-based primal-dual methods in the context of discounted infinite-horizon constrained MDPs.
Figure 2.2: Learning curves of NPG-PD (-) and FOCOPS [280] (-) for Hopper-v3, Swimmer-v3, and Half Cheetah-v3 robotic tasks with the respective speed limits 82.748, 24.516, and 151.989. The vertical axes represent the average reward and the average cost (i.e., average speed). The solid lines show the means of 1000 bootstrap samples obtained over 5 random seeds and the shaded regions display the bootstrap 95% confidence intervals.
Figure 2.3: Learning curves of NPG-PD (—) and FOCOPS [280] (—) for Walker2d-v3, Ant-v3, and Humanoid-v3 robotic tasks with the respective speed limits 81.886, 103.115, and 20.140. The vertical axes represent the average reward and the average cost (i.e., average speed). The solid lines show the means of 1000 bootstrap samples obtained over 5 random seeds and the shaded regions display the bootstrap 95% confidence intervals.
Figure 2.4: Learning curves of NPG-PD (—) and FOCOPS [280] (—) for Humanoid Circle-v0 and Ant Circle-v0 robotic task. The horizontal axis represents the number of dual updates. The average cost is constrained to go below 50. The vertical axes represent the average reward and the average cost (i.e., average speed). The solid lines show the means of 1000 bootstrap samples obtained over 5 random seeds and the shaded regions display the bootstrap 95% confidence intervals.
Chapter 3

Provably efficient policy optimization for constrained MDPs

In this chapter, we focus on an episodic constrained Markov decision processes (MDPs) with the function approximation where the Markov transition kernels have a linear structure but do not impose any additional assumptions on the sampling model. Designing safe reinforcement learning algorithms with provable computational and statistical efficiency is particularly challenging under this setting because of the need to incorporate both the constraint and the function approximation into the fundamental exploitation/exploration tradeoff. To this end, we propose an Optimistic Primal-Dual Proximal Policy Optimization (OPDOP) algorithm where the value function is estimated by combining the least-squares policy evaluation and an additional bonus term for the exploration under constraints (or safe exploration). We prove that the proposed algorithm achieves an \( \tilde{O}(dH^{2.5}\sqrt{T}) \) regret and an \( \tilde{O}(dH^{2.5}\sqrt{T}) \) constraint violation, where \( d \) is the dimension of the feature mapping, \( H \) is the horizon of each episode, and \( T \) is the total number of steps. These bounds hold when the reward/utility functions are fixed but the feedback after each episode is bandit. Our bounds depend on the capacity of the state-action space only through the dimension of the feature mapping and thus our results hold even when the number of states goes to infinity.
3.1 Introduction

Safe Reinforcement Learning (safe RL) augments RL with a practical consideration of safety to deal with restrictions/constraints arising from real-world problems, e.g., collision-avoidance in autonomous robots, cost limitations in medical applications, and legal and business restrictions in financial management. There is considerable growth in safe RL, especially those studies on constrained MDPs, showing the successful integration of the constrained optimization and the policy-based RL for addressing constraints. However, these safe RL algorithms either do not have a convergence theory or are limited to asymptotic convergence. In practice, only a finite amount of data is available. Hence, it is imperative to design safe RL algorithms with computational and statistical efficiency guarantees. For this purpose, we must address the exploration/exploitation trade-off under constraints.

In this chapter, we look at the challenging problem of finding a sequence of policies in response to online streaming samples of transition, reward functions, and utility functions. We attempt to provide theoretical guarantees on the regret of an algorithm approaching the best policy in hindsight, and feasibility region determined by constraints. The task of safe exploration is to explore the unknown environment and learn to adapt the policy to the constraint set. Our problem setting deviates from existing scenarios, where good priors on constraints or transition models are more focused, e.g.,. Recent policy-based safe RL algorithms for constrained MDPs, e.g., constrained policy optimization and primal-dual policy optimization, seek a single safe policy via the constrained policy optimization whose sample efficiency guarantees do not have a theory.
In this chapter, we present our answer the following theoretical question.

Can we design a provably sample efficient online policy optimization algorithm for constrained MDPs in the function approximation setting?

In Section 3.2, we introduce an episodic control problem of constrained MDPs, the metrics of learning performance, and the linear function approximation. In Section 3.3, we propose an optimistic primal-dual policy optimization algorithm for constrained MDPs. In Section 3.4, we establish the regret and constraint violation analysis for the proposed algorithm. In Section 3.5, we present some improved results in the tabular setting. We close this chapter with concluding remarks in Section 3.6.

### 3.2 Problem setup

We consider an episodic constrained Markov decision process,

\[
\text{CMDP}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, g, b)
\]

where \(\mathcal{S}\) is a state space, \(\mathcal{A}\) is an action space, \(H\) is a fixed length of each episode, \(\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^{H}\) is a collection of transition probability measures, \(r = \{r_h\}_{h=1}^{H}\) is a collection of reward functions, \(g = \{g_h\}_{h=1}^{H}\) is a collection of utility functions, and \(b\) is a constraint offset. We assume that \(\mathcal{S}\) is a measurable space with possibly infinite number of elements. Moreover, for each step \(h \in [H]\), \(\mathbb{P}_h(\cdot | s, a)\) is the transition kernel over next state if action \(a\) is taken for state \(s\) and \(r_h: \mathcal{S} \times \mathcal{A} \to [0, 1]\) is a reward function. We assume that reward/utility functions are deterministic. Our analysis readily generalizes to the setting where reward/utility are random.
Let the policy space $\Delta(A \mid S, H)$ be $\{\{\pi_h(\cdot \mid \cdot)\}_{h=1}^H : \pi_h(\cdot \mid s) \in \Delta(A), \forall s \in S \text{ and } h \in [H]\}$, where $\Delta(A)$ denotes a probability simplex over the action space. Let $\pi^k \in \Delta(A \mid S, H)$ be the policy taken by the agent at episode $k$, where $\pi^k_h(\cdot \mid s^k_h) : S \rightarrow A$ is the action that the agent takes at state $s^k_h$. For simplicity, we assume the initial state $s^k_1$ to be fixed as $s_1$ in different episodes for brevity. The agent interacts with the environment in the $k$th episode as follows. At the beginning, the agent determines a policy $\pi^k$. Then, at each step $h \in [H]$, the agent observes the state $s^k_h \in S$, determines an action $a^k_h$ following the policy $\pi^k_h(\cdot \mid s^k_h)$, and receives a reward $r_h(s^k_h, a^k_h)$ together with a utility $g_h(s^k_h, a^k_h)$. Meanwhile, the MDP evolves into next state $s^k_{h+1}$ drawing from the probability $P_h(\cdot \mid s^k_h, a^k_h)$. The episode terminates at state $s^k_H$ in which no control action is taken and both reward and utility functions are equal to zero. Our focus is the bandit setting where the agent only observes the values of reward/utility functions, $r_h(s^k_h, a^k_h), g_h(s^k_h, a^k_h)$, at visited state-action pair $(s^k_h, a^k_h)$. We assume that reward/utility functions are fixed over episodes.

Given a policy $\pi \in \Delta(A \mid S, H)$, the value function $V^\pi_{r,h}$ associated with the reward function $r$ at each step $h$ are the expected values of total rewards,

$$V^\pi_{r,h}(s) = \mathbb{E}_\pi \left[ \sum_{i=h}^H r_i(s_i, a_i) \mid s_h = s \right]$$

for all $s \in S$, $h \in [H]$, where the expectation $\mathbb{E}_\pi$ is taken over the random state-action sequence $\{(s_h, a_h)\}_{h=i}^H$; the action $a_h$ follows the policy $\pi_h(\cdot \mid s_h)$ at the state $s_h$ and the next state $s_{h+1}$ follows the transition dynamics $P_h(\cdot \mid s_h, a_h)$. Thus, the state-action function $Q^\pi_{r,h}(s, a) : S \times A \rightarrow \mathbb{R}$
$\mathbb{R}$ associated with the reward function $r$ is the expected value of total rewards when the agent starts from state-action pair $(s, a)$ at step $h$ and follows policy $\pi$,

$$Q_{r,h}^\pi(s, a) = \mathbb{E}_\pi \left[ \sum_{i=h}^{H} r_i(s_i, a_i) \mid s_h = s, a_h = a \right]$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$. Similarly, we define the value function $V_{g,h}^\pi: \mathcal{S} \to \mathbb{R}$ and the state-action function $Q_{g,h}^\pi(s, a): \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ associated with the utility function $g$. Denote symbol $\diamond = r$ or $g$. For brevity, we take the shorthand $\mathbb{P}_h V_{\diamond,h+1}^\pi(s, a) := \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot \mid s, a)} V_{\diamond,h+1}^\pi(s')$.

The Bellman equations associated with a policy $\pi$ are given by

$$Q_{\diamond,h}^\pi(s, a) = (\diamond_h + \mathbb{P}_h V_{\diamond,h+1}^\pi(s, a))$$

(3.1)

where $V_{\diamond,h}^\pi(s) = \langle Q_{\diamond,h}^\pi(s, \cdot), \pi_h(\cdot \mid s) \rangle_{\mathcal{A}}$, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Here, the inner product of a function $f: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with $\pi(\cdot \mid s) \in \Delta(\mathcal{A})$ at fixed $s \in \mathcal{S}$ represents

$$\langle f(s, \cdot), \pi(\cdot \mid s) \rangle_{\mathcal{A}} := \sum_{a \in \mathcal{A}} \langle f(s, a), \pi(a \mid s) \rangle.$$ 

3.2.1 Learning performance

The design of optimal control policy reduces to finding a solution of a constrained problem in which the objective function is the expected total rewards and the constraint is on the expected total utilities,

$$\max_{\pi \in \Delta(\mathcal{A} \mid \mathcal{S}, H)} V_{r,1}^\pi(s_1)$$

subject to $V_{g,1}^\pi(s_1) \geq b$

(3.2)
where we take $b \in (0, H]$ to avoid triviality. It is readily generalized to the problem with multiple constraints. Let $\pi^* \in \Delta(A \mid S, H)$ be a solution to problem (3.2). Since the policy $\pi^*$ is computed from knowing the transition model and all reward and utility functions, we refer it as an optimal policy in-hindsight.

The associated Lagrangian of problem (3.2) is given by

$$V_L^\pi Y(s_1) := V_\pi r_1(s_1) + Y (V_\pi g_1(s_1) - b)$$

where $\pi$ is the primal variable and $Y \geq 0$ is the dual variable. We can cast (3.2) into a saddle-point problem,

$$\maximize_{\pi \in \Delta(A \mid S, H)} \minimize_{Y \geq 0} V_L^\pi Y(s_1)$$

where $V_L^\pi Y(s_1)$ is convex in $Y$ and is non-concave in $\pi$ in general. To address the non-concavity, we exploit the structure of value functions to propose a variant of Lagrange multiplier method for constrained RL problems in Section 3.3 which warrants a new line of primal-dual mirror descent type analysis in sequel. This distinguishes from unconstrained RL, e.g., [8, 47].

Another key feature of constrained RL is the safe exploration under constraints {93}. Without any constraint information \textit{a priori}, it is infeasible for each policy to satisfy the constraint since utility information on constraints is only revealed after a policy is decided. Instead, we allow each policy to violate the constraint in each episode and minimize regret while minimizing total constraint violations for safe exploration over $K$ episodes. We define the regret as the difference between the total reward value of policy $\pi^*$ in hindsight and that of the agent’s policy $\pi^k$ over $K$
episodes, and the constraint violation as a difference between the offset $Kb$ and the total utility value of the agent’s policy $\pi^k$ over $K$ episodes,

\[
\text{Regret}(K) = \sum_{k=1}^{K} \left( V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{\pi_k}(s_1) \right) \\
\text{Violation}(K) = \sum_{k=1}^{K} \left( b - V_{g,1}^{\pi_k}(s_1) \right).
\] (3.3)

In this chapter, we design algorithms, taking bandit feedback of the reward/utility functions, with both regret and constraint violation being sublinear in the total number of steps $T := HK$. Put differently, the algorithm should ensure that given $\epsilon > 0$, if $T = O(1/\epsilon^2)$, then both $\text{Regret}(K) = O(\epsilon)$ and $\text{Violation}(K) = O(\epsilon)$ hold with high probability.

Let $V_D^Y(s_1) := \max \pi V_L^{\pi,Y}(s_1)$ be the dual function and $Y^* := \arg\min_{Y \geq 0} V_D^Y(s_1)$ be the optimal dual variable. We assume feasibility for Problem (3.2) in Assumption 7 that is known as the Slater condition [179, 85, 187]. It is convenient to establish the strong duality [179] and the boundedness of the optimal dual variable $Y^*$ that can be found in Appendix B.4.

**Assumption 7 (Feasibility)** There exists $\gamma > 0$ and $\bar{\pi} \in \Delta(A|S,H)$ such that $V_{g,1}^{\bar{\pi}}(s_1) \geq b + \gamma$.

**Lemma 15 (Strong Duality and Boundedness of $Y^*$)** Let Assumption 7 hold. Then

(i) $V_{r,1}^{\pi^*}(s_1) = V_D^{Y^*}(s_1)$;

(ii) $0 \leq Y^* \leq (V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{\pi}(s_1))/\gamma$.

Lemma 15 provides useful optimization properties of (3.2) for our algorithm design and analysis.
3.2.2 Linear function approximation

We focus on a class of constrained MDPs, where transition kernels are linear in feature maps.

**Assumption 8** The CMDP($S, A, H, P, r, g$) is a linear MDP with a kernel feature map $\psi: S \times A \times S \rightarrow \mathbb{R}^{d_1}$, if for any $h \in [H]$, there exists a vector $\theta_h \in \mathbb{R}^{d_1}$ with $\|\theta_h\|_2 \leq \sqrt{d_1}$ such that for any $(s, a, s') \in S \times A \times S$,

$$P_h(s' | s, a) = \langle \psi(s, a, s'), \theta_h \rangle;$$

there exists a feature map $\varphi: S \times A \rightarrow \mathbb{R}^{d_2}$ and vectors $\theta_{r,h}, \theta_{g,h} \in \mathbb{R}^{d_2}$ such that for any $(s, a) \in S \times A$,

$$r_h(s, a) = \langle \varphi(s, a), \theta_{r,h} \rangle \text{ and } g_h(s, a) = \langle \varphi(s, a), \theta_{g,h} \rangle$$

where $\max(\|\theta_{r,h}\|_2, \|\theta_{g,h}\|_2) \leq \sqrt{d_2}$. Moreover, we assume that for any function $V: S \rightarrow [0, H]$,

$$\left\| \int_{S} \psi(s, a, s')V(s')ds' \right\|_2 \leq \sqrt{d_1} \cdot H \text{ for all } (s, a) \in S \times A$$

and $\max(d_1, d_2) \leq d$.

Assumption 8 adapts the definition of linear kernel MDP [18, 284, 47] for constrained MDPs. Linear kernel MDP examples include tabular MDPs [284], feature embedded transition models [255], and linear combinations of base models [166]. We can construct related examples of constrained MDPs with linear structure by adding proper constraints. For usefulness of linear structure, see discussions in the literature [80, 228, 124]. For more general transition dynamics, see factored MDPs [184].
Although our definition in Assumption 8 and linear MDPs [256, 110] all contain tabular MDPs as special cases, they define transition dynamics using different feature maps. They are not comparable since one cannot be implied by the other [284]. We provide more details on the tabular case of Assumption 8 in Section 3.5.

### 3.3 Optimistic primal-dual proximal policy optimization

In Algorithm 6, we present a new variant of proximal policy optimization [194]: an Optimistic Primal-Dual Proximal Policy OPtimization (OPDOP) algorithm. Specifically, we effectuate the optimism through the Upper-Confidence Bounds (UCB) [256, 255, 110], and address the constraints via the union of the Lagrange multipliers method with the value function structure that is captured by the performance difference lemma.

**Remark 4** For any two policies $\pi, \pi' \in \Delta(A | S, H)$, $\circ = r$ or $g$, the performance difference lemma [47] quantifies the value function difference by

$$V_{\circ, 1}^\pi(s_1) - V_{\circ, 1}^\pi'(s_1) = \mathbb{E}_{\pi'} \left[ \sum_{h=1}^{H} \left\{ Q_{\circ, h}^\pi(s_h, \cdot), \pi_h' (\cdot | s_h) - \pi_h (\cdot | s_h) \right\} \bigg | s_1 \right].$$

In each episode, our algorithm consists of three main stages. The first stage (lines 4–8) is **Policy Improvement**: we receive a new policy $\pi^k$ by improving previous $\pi^{k-1}$ via a mirror descent type optimization; The second stage (line 9) is **Dual Update**: we update dual variable $Y^k$ based on the constraint violation induced by previous policy $\pi^k$; The third stage (line 10) is **Policy Evaluation**: we optimistically evaluate newly obtained policy via the least-squares policy evaluation with an additional UCB bonus term for exploration.
3.3.1 Policy improvement

In the $k$-th episode, a natural attempt of obtaining a policy $\pi^k$ is to solve a Lagrangian-based policy optimization problem,

$$\max_{\pi \in \Delta(A|S,H)} V_L^{\pi,Y_{k-1}}(s_1) := V_{\pi,Y_{k-1}}(s_1) - V_{\pi,Y_{k-1}}^k(b - V_{\pi,Y_{k-1}}^k(s_1))$$

where $V_L^{\pi,Y}(s_1)$ is the Lagrangian and the dual variable $Y_{k-1} \geq 0$ is from the last episode; we show that $Y_{k-1}$ can be updated efficiently in Section 3.3.2. This type update also finds in [126, 179, 180, 220]. They rely on an oracle solver, e.g., Q-learning [86], proximal policy optimization [194], or trust region policy optimization [195], to deliver a near-optimal policy, making overall algorithmic complexity expensive. Hence, they are not suitable for online use. In contrast, we utilize the RL problem structure and show that only an easily-computable proximal step is sufficient for efficiently achieving near-optimal performance.

Recall symbol $\diamond = r$ or $g$. Via the performance difference lemma, we can expand value functions $V_{\phi,1}^\pi(s_1)$ at the previously known policy $\pi^{k-1}$,

$$V_{\phi,1}^\pi(s_1) = V_{\phi,1}^{\pi^{k-1}}(s_1^k) + \mathbb{E}_{\pi^{k-1}} \left[ \sum_{h=1}^H \langle Q_{\phi,h}(s_h, \cdot), (\pi_h - \pi^{k-1}_h)(\cdot | s_h) \rangle \right]$$

where $\mathbb{E}_{\pi^{k-1}}$ is taken over the random state-action sequence $\{(s_h, a_h)\}_{h=1}^H$. Thus, we introduce an approximation of $V_{\phi,1}^\pi(s_1)$ for any state-action sequence $\{(s_h, a_h)\}_{h=1}^H$ induced by $\pi$,

$$L_{\phi}^{k-1}(\pi) = V_{\phi,1}^{\pi^{k-1}}(s_1) + \sum_{h=1}^H \langle Q_{\phi,h}^{k-1}(s_h, \cdot), (\pi_h - \pi^{k-1}_h)(\cdot | s_h) \rangle$$
where $V^{k-1}$ and $Q^{k-1}$ can be estimated from an optimistic policy evaluation that will be discussed in Section 3.3.3. With this notion, in each episode, instead of solving a Lagrangian-based policy optimization, we perform a simple policy update in online mirror descent fashion,

$$
\max_{\pi \in \Delta(A|S,H)} L^{k-1}(\pi) - Y^{k-1}(b - L^{k-1}(\pi)) - \frac{1}{\alpha} \sum_{h=1}^{H} D_{KL}(\pi_h(\cdot|s_h) | \tilde{\pi}_h^{k-1}(\cdot|s_h))
$$

where $\tilde{\pi}_h^{k-1}(\cdot|s_h) = (1 - \theta) \pi_h^{k-1}(\cdot|s_h) + \theta \text{Unif}_A(\cdot)$ is a mixed policy of the previous one and the uniform distribution Unif$_A$ with $\theta \in (0, 1]$. The constant $\alpha > 0$ is a trade-off parameter, $D_{KL}(\pi | \tilde{\pi}^{k-1})$ is the Kullback-Leibler (KL) divergence between $\pi$ and $\tilde{\pi}^{k-1}$ in which $\pi$ is absolutely continuous in $\tilde{\pi}^{k-1}$. The policy mixing step ensures such absolute continuity and implies uniformly bounded KL divergence; see Lemma 54 in Appendix B.5. Ignoring other $\pi$-irrelevant terms, we update $\pi^k$ in terms of previous policy $\tilde{\pi}^{k-1}$ by

$$
\arg\max_{\pi \in \Delta(A|S,H)} \sum_{h=1}^{H} \left\langle (Q^{k-1}_{r,h} + Y^{k-1}Q^{k-1}_{g,h})(s_h, \cdot), \pi_h(\cdot|s_h) \right\rangle - \frac{1}{\alpha} \sum_{h=1}^{H} D_{KL}(\pi_h(\cdot|s_h) | \tilde{\pi}_h^{k-1}(\cdot|s_h)).
$$

Since the above update is separable over $H$ steps, we can update the policy $\pi^k$ as line 6 in Algorithm 6 a closed-form solution for any step $h \in [H]$. If we set $Y^{k-1} = 0$ and $\theta = 0$, the above update reduces to one step in an optimistic proximal policy optimization [47]. The idea of KL-divergence regularization in policy optimization has been widely used in many unconstrained scenarios [112, 194, 195, 233, 139]. Our method is distinct in that it is based on the performance difference lemma and the optimistically estimated value functions.
Algorithm 6 Optimistic Primal-Dual Proximal Policy Optimization (OPDOP)

1: **Initialization**: Let \( \{Q^0_{r,h}, Q^0_{g,h}\}_{h=1}^H \) be zero functions, \( \{\pi^0_h\}_{h \in [H]} \) be uniform distributions on \( \mathcal{A} \), \( V^0_{g,1} \) be \( b, Y^0 \) be 0, \( \chi = 2H/\gamma \), \( \alpha, \eta > 0, \theta \in (0, 1] \).

2: **for** episode \( k = 1, \ldots, K + 1 \) **do**
3:     Set the initial state \( s^1_k = s_1 \).
4:     **for** step \( h = 1, 2, \ldots, H \) **do**
5:         Mix the previous policy by \( \tilde{\pi}^{k-1}_h(\cdot | \cdot) = (1 - \theta)\pi^{k-1}_h(\cdot | \cdot) + \theta \text{Unif}_A \).
6:         Update the current policy by \( \pi^k_h(\cdot | \cdot) \propto \tilde{\pi}^{k-1}_h(\cdot | \cdot) e^{\alpha (Q^{k-1}_r + Y^{k-1}_g Q^{k-1}_{g,h}) (\cdot, \cdot)} \).
7:         Take an action \( a^k_h \sim \pi^k_h(\cdot | s^k_h) \) and recieve reward/utility \( r^k_h(s^k_h, a^k_h), g^k_h(s^k_h, a^k_h) \).
8:         Observe the next state \( s^{k+1}_h \).
9:     **end for**
10:    Update the dual variable \( Y^k \) by \( Y^k \leftarrow \mathcal{P}_{[0, \chi]}(Y^{k-1} + \eta (b - V^k_{g,1}(s^k_1))) \).
11:    Estimate the state-action or value functions \( \{Q^{k}_{r,h}, Q^{k}_{g,h}, V^{k}_{g,h}\}_{h=1}^H \) via \( \text{LSTD} \left( \{s^*_h, a^*_h, r_h(s^*_h, a^*_h), g_h(s^*_h, a^*_h)\}_{h, \tau = 1}^{H,k} \right) \).
12: **end for**

### 3.3.2 Dual update

To infer the constraint violation for the dual update, we estimate \( V^k_{g,1}(s^1_1) \) via an optimistic policy evaluation by \( V^{k-1}_{g,1}(s^1_1) \) that is discussed in Section 3.3.3. We update the Lagrange multiplier \( Y \) by moving \( Y^k \) to the direction of minimizing the Lagrangian \( V^k_{\pi,Y} \) over \( Y \geq 0 \) in line 10 of Algorithm 6, where \( \eta > 0 \) is a stepsize and \( \mathcal{P}_{[0, \chi]} \) is a projection onto \([0, \chi]\) with an upper bound \( \chi \) on \( Y^k \). By Lemma 15 we choose \( \chi = 2H/\gamma \geq 2Y^* \) so that projection interval \([0, \chi]\) includes the optimal dual variable \( Y^* \). This type design also finds in [85, 174].

The dual update works as a trade-off between the reward maximization and the constraint violation reduction. If the current policy \( \pi^k \) satisfies the approximated constraint, i.e., \( b - L^k_{g,1}(\pi^k) \leq 0 \), we put less weight on the state-action function associated with the utility and maximize the reward; otherwise, we sacrifice the reward a bit to satisfy the constraint. The dual update has a similar use in dealing with constraints in constrained MDPs, e.g., Lagrangian-based actor-critic [56].
and online constrained optimization \[263\,242\,265\]. In contrast, we handle the dual update via the optimistic policy evaluation, yielding a simple, but efficient estimation on the constraint violation.

**Algorithm 7** Least-squares temporal difference (LSTD) with UCB exploration

1. **Input:** \(\{s^\tau_h, a^\tau_h, r^\tau_h(s^\tau_h, a^\tau_h), g^\tau_h(s^\tau_h, a^\tau_h)\}_{h,\tau=1}^{H,k}\).
2. **Initialization:** Set \(\{V_{k}^h, V_{k}^g\}_{h=1}^{H,k}\) be zero functions and \(\lambda = 1, \beta = O(\sqrt{dH^2 \log (dT/p)})\).

3. for step \(h = H, H-1, \cdots, 1\) do
4. \[\Lambda^{k}_{h} = \sum_{\tau=1}^{k-1} \phi^\tau_{o,h}(s^\tau_h, a^\tau_h)\phi^\tau_{o,h}(s^\tau_h, a^\tau_h)^\top + \lambda I.\]
5. \[w^{k}_{o,h} = (\Lambda^{k}_{o,h})^{-1} \sum_{\tau=1}^{k-1} \phi^\tau_{o,h}(s^\tau_h, a^\tau_h)V^\tau_{o,h+1}(s^\tau_{h+1}).\]
6. \[\phi^{k}_{o,h}(\cdot, \cdot) = \int_{S} \psi(\cdot, \cdot, s')V_{k, h+1}^k(s')ds'.\]
7. \[\Gamma^{k}_{o,h}(\cdot, \cdot) = \beta(\phi^{k}_{o,h}(\cdot, \cdot)^\top (\Lambda^{k}_{o,h})^{-1}\phi^{k}_{o,h}(\cdot, \cdot))^{1/2}.\]
8. \[\Lambda^{k}_{h} = \sum_{\tau=1}^{k-1} \varphi(s^\tau_h, a^\tau_h)\varphi(s^\tau_h, a^\tau_h)^\top + \lambda I.\]
9. \[u^{k}_{o,h} = (\Lambda^{k}_{h})^{-1} \sum_{\tau=1}^{k-1} \varphi(s^\tau_h, a^\tau_h)\phi_h(s^\tau_h, a^\tau_h).\]
10. \[\Gamma^{k}_{h}(\cdot, \cdot) = \beta(\varphi(\cdot, \cdot)^\top (\Lambda^{k}_{h})^{-1}\varphi(\cdot, \cdot))^{1/2}.\]
11. \[Q^{k}_{o,h}(\cdot, \cdot) = \min \left( H - h + 1, \varphi(\cdot, \cdot)^\top u^{k}_{o,h} + \phi^{k}_{o,h}(\cdot, \cdot)^\top w^{k}_{o,h} + (\Gamma^{k}_{h} + \Gamma^{k}_{o,h})(\cdot, \cdot)\right)^+.\]
12. \[V^{k}_{o,h}(\cdot) = \langle Q^{k}_{o,h}(\cdot, \cdot), \pi^{k}_{\cdot}(\cdot | \cdot) \rangle_{A}.\]
13. end for
14. Return: \(\{Q^{k}_{o,h}(\cdot, \cdot), V^{k}_{o,h}(\cdot)\}_{h=1}^{H}\). 

**3.3.3 Policy evaluation**

The last stage of the \(k\)th episode takes the Least-Squares Temporal Difference (LSTD) \[41\,39\,125\,121\] to evaluate the policy \(\pi^k\) based on previous \(k - 1\) historical trajectories. For each step
\( h \in [H] \), instead of \( \mathbb{P}_h V_{r,h+1}^k \) in the Bellman equations (3.1), we estimate \( \mathbb{P}_h V_{r,h+1}^k \) by \( (\phi_{r,h}^k)^\top w_{r,h}^k \) where \( w_{r,h}^k \) is updated by the minimizer of the regularized least-squares problem over \( w \),

\[
\sum_{\tau=1}^{k-1} \left( V_{r,h+1}^\tau (s_{h+1}^\tau) - \phi_{r,h}^\tau (s_{h}^\tau, a_{h}^\tau)^\top w \right)^2 + \lambda \| w \|_2^2
\]  

(3.4)

where \( \phi_{r,h}^\tau (\cdot, \cdot) := \int_{S} \psi (\cdot, s') V_{r,h+1}^\tau (s') ds' \), \( V_{r,h+1}^\tau (\cdot) = \langle Q_{r,h+1}^\tau (\cdot, \cdot), \pi_{h+1}^\tau (\cdot | \cdot) \rangle_A \) for \( h \in [H-1] \) and \( V_{H+1}^\tau = 0 \), and \( \lambda > 0 \) is the regularization parameter. Similarly, we estimate \( \mathbb{P}_h V_{g,h+1}^k \) by \( (\phi_{g,h}^k)^\top w_{g,h}^k \). We display the least-squares solution in line 4–6 of Algorithm 7, where symbol \( \diamond = r \) or \( g \). We also estimate \( r_h (\cdot, \cdot) \) by \( \varphi^\top u_{r,h}^k \), where \( u_{r,h}^k \) is updated by the minimizer of another regularized least-squares problem,

\[
\sum_{\tau=1}^{k-1} \left( r_h (s_{h}^\tau, a_{h}^\tau) - \varphi (s_{h}^\tau, a_{h}^\tau)^\top u \right)^2 + \lambda \| u \|_2^2
\]  

(3.5)

where \( \lambda > 0 \) is the regularization parameter. Similarly, we estimate \( g_h (\cdot, \cdot) \) by \( \varphi^\top u_{g,h}^k \). The least-squares solutions lead to line 8–9 of Algorithm 7.

After obtaining estimates of \( \mathbb{P}_h V_{\diamond,h+1}^k \) and \( \phi_h (\cdot, \cdot) \) for \( \diamond = r \) or \( g \), we update the estimated state-action function \( \{ Q_{\diamond,h}^k \}_{h=1}^H \) iteratively in line 11 of Algorithm 7, where \( \varphi^\top u_{\diamond,h}^k \) is an estimate of \( \phi_h \) and \( (\phi_{\diamond,h}^k)^\top w_{\diamond,h}^k \) is an estimate of \( \mathbb{P}_h V_{\diamond,h+1}^k \); we add UCB bonus terms \( \Gamma_h^k (\cdot, \cdot), \Gamma_{\diamond,h}^k (\cdot, \cdot) : S \times A \to \mathbb{R}^+ \) so that

\[
\varphi^\top u_{\diamond,h}^k + \Gamma_h^k \quad \text{and} \quad (\phi_{\diamond,h}^k)^\top w_{\diamond,h}^k + \Gamma_{\diamond,h}^k
\]

all become their upper confidence bounds. Here, the bonus terms take \( \Gamma_h^k = \beta (\varphi^\top (\Lambda_h^k)^{-1} \varphi)^{1/2} \) and \( \Gamma_{\diamond,h}^k = \beta ((\phi_{\diamond,h}^k)^\top (\Lambda_{\diamond,h}^k)^{-1} \phi_{\diamond,h}^k)^{1/2} \) and we leave the parameter \( \beta > 0 \) to be tuned later. Moreover, the bounded reward/utility \( \diamond_h \in [0, 1] \) implies \( Q_{\diamond,h}^k \in [0, H-h+1] \).
We remark the computational efficiency of Algorithm 6. For the time complexity, since line 6 is a scalar update, they need $O(d|A|T)$ time. A dominating calculation is from lines 5/9 in Algorithm 7. If we use the Sherman–Morrison formula for computing $(\Lambda_k^h)^{-1}$, it takes $O(d^2T)$ time. Another important calculation is the integration from line 6 in Algorithm 7. We can either compute it analytically if it is tractable or approximate it via the Monte Carlo integration that assumes polynomial time. Therefore, the time complexity is $O(\text{poly}(d)|A|T)$ in total. For the space complexity, we don’t need to store policy since it is recursively calculated via line 6 of Algorithm 6. By updating $Y^k, \Lambda_h^k, \Lambda_{o,h}^k, w_{o,h}^k, \Lambda_{\cdot,h}^k, \Phi_h(s_h^k, a_h^k)$ recursively, it takes $O((d^2 + |A|)H)$ space.

### 3.4 Regret and constraint violation analysis

We now prove that the regret and the constraint violation for Algorithm 6 are sublinear in $T := KH$, the total number of steps taken by the algorithm, where $K$ is the total number of episodes and $H$ is the episode horizon. We recall that $|A|$ is the cardinality of action space $A$ and $d$ is the dimension of the feature map.

**Theorem 16 (Linear kernal MDP: regret and constraint violation)** Let Assumptions 7 and 8 hold. Fix $p \in (0, 1)$. We set $\alpha = \sqrt{\log |A|/H^2K}$, $\beta = C_1\sqrt{dH^2 \log (dT/p)}$, $\eta = 1/\sqrt{K}$, $\theta = 1/K$, and $\lambda = 1$ in Algorithm 6, where $C_1$ is an absolute constant. Suppose $\log |A| = O(d^2 \log^2 (dT/p))$. Then, with probability $1 - p$, the regret and the constraint violation in (3.3) satisfy

\[
\text{Regret}(K) \leq C dH^{2.5} \sqrt{T} \log \left( \frac{dT}{p} \right)
\]

\[
[\text{Violation}(K)]_+ \leq C' dH^{2.5} \sqrt{T} \log \left( \frac{dT}{p} \right)
\]
where $C$ and $C'$ are absolute constants.

Theorem 16 establishes that Algorithm 6 enjoys an $\tilde{O}(dH^{2.5}\sqrt{T})$ regret and an $\tilde{O}(dH^{2.5}\sqrt{T})$ constraint violation if we set algorithm parameters $\{\alpha, \beta, \eta, \theta, \lambda\}$ properly. Our results have the optimal dependence on the total number of steps $T$ up to some logarithmic factors. The $d$ dependence occurs due to the uniform concentration for controlling the fluctuations in the least-squares policy evaluation. This matches the existing bounds in the linear MDP setting without any constraints [47, 18, 284]. Our bounds differ from them only by $H$ dependence, which is a price introduced by the uniform bound on the constraint violation. It is noticed that our algorithm works for bandit feedback of reward/utility functions after each episode.

Regarding safe exploration, our violation bound provides finite-time convergence to the feasibility region defined by constraints. In the interaction with an unknown environment, the UCB exploration in the utility value function adds optimism towards constraint satisfaction. The dual update regularizes the policy improvement for governing actual constraint violation. Our regret and violation bounds readily lead to PAC guarantees [108]. Compared to most recent references [73, 251, 52, 271], our algorithm is sample-efficient in exploration and does not take any simulations of policy.

We remark the tabular setting for Algorithm 6. The tabular CMDP$(S, A, H, P, r, g, b)$ is a special case of Assumption 8 with $|S| < \infty$ and $|A| < \infty$. Let $d_1 = |S|^2|A|$ and $d_2 = |S||A|$. We take the following feature maps $\psi(s, a, s') \in \mathbb{R}^{d_1}$, $\varphi(s, a) \in \mathbb{R}^{d_2}$, and parameter vectors,

$$
\psi(s, a, s') = e_{(s,a,s')} , \quad \theta_h = P_h(\cdot, \cdot, \cdot) \\
\varphi(s, a) = e_{(s,a)} , \quad \theta_{r,h} = r_h(\cdot, \cdot), \quad \theta_{g,h} = g_h(\cdot, \cdot)
$$

(3.6)
where \( e_{(s,a,s')} \) is a canonical basis of \( \mathbb{R}^{d_1} \) associated with \((s, a, s')\) and \( \theta_h = \mathbb{P}_h(\cdot, \cdot, \cdot) \) reads that for any \((s, a, s')\) \( \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \), the \((s, a, s')\)th entry of \( \theta_h \) is \( \mathbb{P}(s' | s, a) \); similarly we define \( e_{(s,a)} \), \( \theta_{r,h} \), and \( \theta_{g,h} \). We can verify that \( \|\theta_h\| \leq \sqrt{d_1} \), \( \|\theta_{r,h}\| \leq \sqrt{d_2} \), \( \|\theta_{g,h}\| \leq \sqrt{d_2} \), and for any \( V: \mathcal{S} \to [0,H] \) and any \((s, a)\) \( \in \mathcal{S} \times \mathcal{A} \), we have \( \|\sum_{s' \in \mathcal{S}} \psi(s, a, s') V(s')\| \leq \sqrt{|\mathcal{S}|H} \leq \sqrt{d_1}H \).

Therefore, we take \( d := \max(d_1, d_2) = |\mathcal{S}|^2 |\mathcal{A}| \) in Assumption 8 for the tabular case.

We now detail Algorithm 1 for the tabular case as follows. Our policy evaluation works with regression feature \( \phi_{\circ,h}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_2} \),

\[
\phi_{\circ,h}(s, a) = \sum_{s'} \psi(s, a, s') V_{\circ,h+1}^\tau(s'), \text{ for any } (s, a) \in \mathcal{S} \times \mathcal{A}
\]

where \( \circ = r \text{ or } g \). Thus, for any \((\bar{s}, \bar{a}, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \), the \((\bar{s}, \bar{a}, s')\)th entry of \( \phi_{\circ,h}(s, a) \) is given by

\[
\left[ \phi_{\circ,h}(s, a) \right]_{(\bar{s}, \bar{a}, s')} = \mathbf{1}\{(s, a) = (\bar{s}, \bar{a})\} V_{\circ,h+1}^\tau(\bar{s'})
\]

which shows that \( \phi_{\circ,h}(s, a) \) is a sparse vector with \(|\mathcal{S}|\) nonzero elements at \( \{(s, a, s'), s' \in \mathcal{S}\} \) and the \((s, a, s')\)th entry of \( \phi_{\circ,h}(s, a) \) is \( V_{\circ,h+1}^\tau(s') \). For instance of \( \circ = r \), the regularized least-squares problem (3.4) becomes

\[
\sum_{\tau=1}^{k-1} \left( V_{r,h+1}^\tau(s_{h+1}^\tau) - \sum_{(s,a,s')} \mathbf{1}\{(s, a) = (s_h^\tau, a_h^\tau)\} V_{r,h+1}^\tau(s')[w]_{(s,a,s')} \right)^2 + \lambda \|w\|_2^2
\]
where \([w]_{(s,a,s')}\) is the \((s, a, s')\)th entry of \(w\), and the solution \(w^k_{r,h}\) serves as an estimator of the transition kernel \(P_h(\cdot | \cdot, \cdot)\). On the other hand, since \(\varphi(s_h^\tau, a_h^\tau) = e_{(s_h^\tau, a_h^\tau)}\), the regularized least-squares problem (3.5) becomes

\[
\sum_{\tau = 1}^{k-1} (r_h(s_h^\tau, a_h^\tau) - [u]_{(s_h^\tau, a_h^\tau)})^2 + \lambda \| u \|_2^2
\]

where \([u]_{(s,a)}\) is the \((s,a)\)th entry of \(u\), the solution \(u^k_{r,h}\) is an estimate of \(r_h(s,a)\) as \(\varphi(s,a)^\top u^k_{r,h}\).

By adding similar UCB bonus terms \(\Gamma^k_h, \Gamma^k_{r,h} : S \times A \to \mathbb{R}\) given in Algorithm 7 we estimate the state-action function as follows,

\[
Q^k_{r,h}(s, a) = \min \left( [u^k_{r,h}]_{(s,a)} + \phi^k_{r,h}(s, a)^\top w^k_{r,h} + (\Gamma^k_h + \Gamma^k_{r,h})(s, a), H - h + 1 \right)^+
\]

\[
= \min \left( [u^k_{r,h}]_{(s,a)} + \sum_{s' \in S} V^k_{r,h+1}(s') [u^k_{r,h}]_{(s,a,s')} + (\Gamma^k_h + \Gamma^k_{r,h})(s, a), H - h + 1 \right)^+
\]

for any \((s, a) \in S \times A\). Thus, \(V^k_{r,h}(s) = \langle Q^k_{r,h}(s, \cdot), \pi_h^k(\cdot | s) \rangle_A\). Similarly, we estimate \(g_h(s, a)\) and thus \(Q^k_{g,h}(s, a)\) and \(V^k_{g,h}(s)\). Using already estimated \(\{Q^k_{r,h}(\cdot, \cdot), Q^k_{g,h}(\cdot, \cdot), V^k_{r,h}(\cdot), V^k_{g,h}(\cdot)\}_{h=1}^H\), we execute the policy improvement and the dual update in Algorithm 6.

We restate the result of Theorem 16 for the tabular case as follows.

**Corollary 17 (Regret and constraint violation)** For the tabular constrained MDP with feature maps (3.6), let Assumption 7 hold. Fix \(p \in (0, 1)\). In Algorithm 6 we set \(\alpha = \sqrt{\log |A| / (H^2 K)}\),
\[ \beta = C_1 \sqrt{|S|^2|A|H^2 \log (|S||A|T/p)}, \eta = 1/\sqrt{K}, \theta = 1/K, \text{ and } \lambda = 1, \text{ where } C_1 \text{ is an absolute constant. Then, the regret and the constraint violation in (3.3) satisfy} \]

\[
\begin{align*}
\text{Regret}(K) & \leq C|S|^2|A|H^{2.5}T^{\frac{1}{2}}\log \left( \frac{|S||A|T}{p} \right) \\
\text{Violation}(K) & \leq C'|S|^2|A|H^{2.5}T^{\frac{1}{2}}\log \left( \frac{|S||A|T}{p} \right)
\end{align*}
\]

with probability \( 1 - p \) where \( C \) and \( C' \) are absolute constants.

**Proof.** It follows the proof of Theorem 16 by noting that the tabular constrained MDP is a special linear MDP in Assumption 8, with \( d = |S|^2|A| \), and we have \( \log |A| \leq O(d^2 \log (dT/p)) \) automatically. \( \square \)

**Algorithm 8** Optimistic policy evaluation (OPE)

1: **Input:** \( \{s^r_h, a^r_h, r_h(s^r_h, a^r_h), g_h(s^r_h, a^r_h)\}_{h=1}^{H} \).
2: **Initialization:** Let \( \lambda = 1, \beta = C_1H \sqrt{|S|\log(|S||A|T/p)} \), and set \( \{V_{r,H+1}^k, V_{g,H+1}^k\} \) be zero functions.
3: **for** step \( h = H, H-1, \cdots, 1 \) **do**
4: \( \diamond = r, g \)
5: Compute counters \( n_h^k(s,a) \) and \( \hat{n}_h^k(s,a) \) via (3.36) for all \( (s,a,s') \in S \times A \times S \) and \( (s,a) \in S \times A \).
6: Estimate reward/utility functions \( \hat{r}_h^k, \hat{g}_h^k \) via (3.37) for all \( (s,a) \in S \times A \).
7: Estimate transition \( \hat{P}^k_h \) via (3.38) for all \( (s,a,s') \in S \times A \times S \), and take bonus \( \Gamma^k_h = \beta \left( \hat{n}_h^k(s,a) + \lambda \right)^{-1/2} \) for all \( (s,a) \in S \times A \).
8: \( Q_{o,h}^k(\cdot,\cdot) = \min \left( H-h+1, \hat{\phi}_h^k(\cdot,\cdot) + \sum_{s' \in S} \hat{P}_h(s'|\cdot,\cdot) \hat{V}_h^k(s')+2 \Gamma_h^k(\cdot,\cdot) \right) \)
9: \( V_h^k(\cdot) = \langle Q_h^k(\cdot,\cdot), \pi_h^k(\cdot|\cdot) \rangle_{\mathcal{A}} \).
10: **end for**
11: **Return:** \( \{Q_{r,h}^k(\cdot,\cdot), Q_{g,h}^k(\cdot,\cdot)\}_{h=1}^{H} \).
3.4.1 Setting up the analysis

Our analysis begins with decomposition of the regret given in (3.3).

\[
\text{Regret}(K) = \sum_{k=1}^{K} \left( V^*_{\pi,k}(s_1) - V^k_{r,1}(s_1) \right) + \sum_{k=1}^{K} \left( V^k_{r,1}(s_1) - V^\pi_{r,1}(s_1) \right)
\]

where we add and subtract the value \( V^k_{r,1}(s_1) \) estimated from an optimistic policy evaluation by Algorithm 7, the policy \( \pi^* \) in hindsight is the best policy in hindsight for problem (3.2). To bound the total regret (3.7), we would like to analyze (R.I) and (R.II) separately.

First, we define the model prediction error for the reward as

\[
\iota^k_{r,h} := r_h + \mathbb{P}_h V^k_{r,h+1} - Q^k_{r,h}
\]

for all \((k, h) \in [K] \times [H]\), which describes the prediction error in the Bellman equations (3.1) using \( V^k_{r,h+1} \) instead of \( V^\pi_{r,h+1} \). With this notation, we expand (R.I) into

\[
\text{(R.I)} = \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ \langle Q^k_{r,h}(s_h, \cdot), \pi^*_h(\cdot | s_h) - \pi^k_h(\cdot | s_h) \rangle \right] + \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ \iota^k_{r,h}(s_h, a_h) \right]
\]

where the first double sum is linear in terms of the policy difference and the second one describes the total model prediction error. The above expansion is based on the standard performance difference lemma (see Remark 4). Meanwhile, if we define the model prediction error for the utility as

\[
\iota^k_{g,h} := g_h + \mathbb{P}_h V^k_{g,h+1} - Q^k_{g,h}
\]
then, similarly, we can expand $\sum_{k=1}^{K} (V_{g,1}^*(s_1) - V_{g,1}^k(s_1))$ into

$$
\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ \langle Q_{g,h}^k(s_h, \cdot), \pi_h^* (\cdot | s_h) - \pi_h^k (\cdot | s_h) \rangle \right] + \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ \iota_{g,h}^k(s_h, a_h) \right].
$$

(3.11)

To analyze the constraint violation, we also introduce a useful decomposition,

$$
\text{Violation}(K) = \sum_{k=1}^{K} (b - V_{g,1}^k(s_1)) + \sum_{k=1}^{K} (V_{g,1}^k(s_1) - V_{g,1}^\pi(s_1))
$$

(VII)

(3.12)

which the inserted value $V_{g,1}^k(s_1)$ is estimated from an optimistic policy evaluation by Algorithm 7.

For notational simplicity, we introduce the underlying probability structure as follows. For any $(k, h) \in [K] \times [H]$, we define $\mathcal{F}_{h,1}^k$ as a $\sigma$-algebra generated by state-action sequences, reward and utility functions,

$$
\{(s^T_i, a^T_i)\} \cap [k-1] \times [H] \cup \{(s^k_i, a^k_i)\} \cup [h].
$$

Similarly, we define $\mathcal{F}_{h,2}^k$ as an $\sigma$-algebra generated by

$$
\{(s^T_i, a^T_i)\} \cap [k-1] \times [H] \cup \{(s^k_i, a^k_i)\} \cup [h] \cup \{s_{h+1}^k\}.
$$

Here, $s_{H+1}^k$ is a null state for any $k \in [K]$. A sequence of $\sigma$-algebras $\{\mathcal{F}_{h,m}^k\}_{(k,h,m) \in [K] \times [H] \times [2]}$ is a filtration in terms of time index

$$
t(k, h, m) := 2(k-1)H + 2(h-1) + m
$$

(3.13)
which holds that $\mathcal{F}_{h,m}^k \subset \mathcal{F}_{h',m}',$ for any $t \leq t'$. The estimated reward/utility value functions, $V_{r,h}^k, V_{g,h}^k$, and the associated state-action functions, $Q_{r,h}^k, Q_{g,h}^k$ are $\mathcal{F}_{1,1}^k$-measurable since they are obtained from previous $k - 1$ historical trajectories. With these notations, we can expand (R.II) in (3.7) into

$$(R.II) = - \sum_{k=1}^{K} \sum_{h=1}^{H} \iota_{k}(s_{h},a_{h}^{k}) + M_{r,H,2}^{K}$$

where $\{M_{r,h,m}^k\} \in [K] \times [H] \times [2]$ is a martingale adapted to the filtration $\{\mathcal{F}_{h,m}^k\} \in [K] \times [H] \times [2]$ in terms of time index $t$. Similarly, we have it for (V.II),

$$(V.II) = - \sum_{k=1}^{K} \sum_{h=1}^{H} \iota_{g}(s_{h},a_{h}^{k}) + M_{g,H,2}^{K}$$

where $\{M_{g,h,m}^k\} \in [K] \times [H] \times [2]$ is a martingale adapted to the filtration $\{\mathcal{F}_{h,m}^k\} \in [K] \times [H] \times [2]$ in terms of time index $t$. We prove (3.14) in Appendix B.3 (also see [47, Lemma 4.2]); (3.15) is similar.

We recall two UCB bonus terms in the state-action function estimation of Algorithm 7,

$$\Gamma_{o,h}^k := \beta((\phi_{o,h}^k)^\top (\Lambda_{o,h}^k)^{-1} \phi_{o,h}^k)^{1/2} \quad \text{and} \quad \Gamma_{h}^k := \beta((\varphi)^\top (\Lambda_{h}^k)^{-1} \varphi)^{1/2}$$

By the UCB argument, if we set $\lambda = 1$ and $\beta = C_1 \sqrt{dH^2 \log(dT/p)}$ where $C_1$ is an absolute constant, then for any $(k, h) \in [K] \times [H]$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$-2(\Gamma_{h}^k + \Gamma_{o,h}^k)(s, a) \leq \iota_{o,h}^k(x, a) \leq 0$$

(3.16)
with probability \(1 - p/2\) where the symbol \(\diamond = r\) or \(g\). We prove (3.16) in Appendix B.3.1 for completeness.

In what follows we delve into the analysis of the regret and the constraint violation.

### 3.4.2 Proof of regret bound

Our analysis begins with a primal-dual mirror descent type analysis for the policy update in line 6 of Algorithm 6. In Lemma 18, we present a key upper bound on the total differences of estimated values \(V^k_{\pi^*(r)}(s_1)\) and \(V^k_{\pi^*(g)}(s_1)\) given by Algorithm 7 to the optimal ones.

**Lemma 18 (Policy improvement: primal-dual mirror descent)** Let Assumptions 7-8 hold. In Algorithm 6, if we set \(\alpha = \sqrt{\log |A|/(H^2 \sqrt{K})}\) and \(\theta = 1/K\), then

\[
\sum_{k=1}^{K} \left( V^k_{\pi^*(r)}(s_1) - V^k_{\pi^*(r)}(s_1) \right) + \sum_{k=1}^{K} Y^k \left( V^k_{\pi^*(g)}(s_1) - V^k_{\pi^*(g)}(s_1) \right) \\
\leq C_2 H^{2.5} \sqrt{T \log |A|} + \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ \iota^k_{r,h}(s_h, a_h) \right] + \sum_{k=1}^{K} \sum_{h=1}^{H} Y^k \mathbb{E}_{\pi^*} \left[ \iota^k_{g,h}(s_h, a_h) \right] \tag{3.17}
\]

where \(C_2\) is an absolute constant and \(T = HK\).

**Proof.** We recall that line 6 of Algorithm 6 follows a solution \(\pi^k\) to the following subproblem,

\[
\max_{\pi \in \Delta(A|S,H)} \sum_{h=1}^{H} \left\langle Q^{k-1}_{r,h} + Y^{k-1}Q^{k-1}_{g,h}, \pi_h \right\rangle - \frac{1}{\alpha} \sum_{h=1}^{H} D_{KL}(\pi_h | \tilde{\pi}^{k-1}_h) \tag{3.18}
\]

where \(\left\langle Q^{k-1}_{r,h} + Y^{k-1}Q^{k-1}_{g,h}, \pi_h \right\rangle\) is a shorthand for \(\left\langle (Q^{k-1}_{r,h} + Y^{k-1}Q^{k-1}_{g,h})(s_h, \cdot), \pi_h(\cdot | s_h) \right\rangle\) and the shorthand \(D_{KL}(\pi_h | \tilde{\pi}^{k-1}_h)\) for \(D_{KL}(\pi_h(\cdot | s_h) | \tilde{\pi}^{k-1}_h(\cdot | s_h))\) if dependence on the state-action sequence \(\{s_h, a_h\}_{h=1}^{H}\) is clear from context. We note that (3.18) is in form of a mirror descent
We can apply the pushback property with \( x^* = \pi_h^k, y = \tilde{\pi}_h^{k-1} \) and 
\( z = \pi_h^* \),

\[
\sum_{h=1}^{H} \langle Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k}, \pi_h^k \rangle - \frac{1}{\alpha} \sum_{h=1}^{H} D_{KL}(\pi_h^k \| \tilde{\pi}_h^{k-1}) \\
\geq \sum_{h=1}^{H} \langle Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k}, \pi_h^* \rangle - \frac{1}{\alpha} \sum_{h=1}^{H} D_{KL}(\pi_h^* \| \tilde{\pi}_h^{k-1}) + \frac{1}{\alpha} \sum_{h=1}^{H} D_{KL}(\pi_h^* \| \pi_h^k).
\]

Equivalently, we write the above inequality as follows,

\[
\sum_{h=1}^{H} \langle Q_{r,h}^{k-1}, \pi_h^* - \pi_h^{k-1} \rangle + Y^{k-1} \sum_{h=1}^{H} \langle Q_{g,h}^{k-1}, \pi_h^* - \pi_h^{k-1} \rangle \\
\leq \sum_{h=1}^{H} \langle Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k}, \pi_h^k - \pi_h^{k-1} \rangle - \frac{1}{\alpha} \sum_{h=1}^{H} D_{KL}(\pi_h^k \| \tilde{\pi}_h^{k-1}) \\
+ \frac{1}{\alpha} \sum_{h=1}^{H} D_{KL}(\pi_h^* \| \tilde{\pi}_h^{k-1}) - \frac{1}{\alpha} \sum_{h=1}^{H} D_{KL}(\pi_h^* \| \pi_h^k).
\]

By taking expectation \( \mathbb{E}_{\pi^*} \) on both sides of (3.19) over the state-action sequence \( \{(s_h, a_h)\}_{h=1}^{H} \), starting from \( s_1 \), and applying decompositions (3.9) and (3.11), we have

\[
(V_{r,1}^*(s_1) - V_{r,1}^{k-1}(s_1)) + Y^{k-1}(V_{g,1}^*(s_1) - V_{g,1}^{k-1}(s_1)) \\
\leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ \langle Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k}, \pi_h^k - \pi_h^{k-1} \rangle \right] - \frac{1}{\alpha} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ D_{KL}(\pi_h^k \| \tilde{\pi}_h^{k-1}) \right] \\
+ \frac{1}{\alpha} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ D_{KL}(\pi_h^* \| \tilde{\pi}_h^{k-1}) \right] - D_{KL}(\pi_h^* \| \pi_h^k) \\
+ \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ \ell_{r,h}(s_h, a_h) \right] + Y^{k-1} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ \ell_{g,h}(s_h, a_h) \right]
\]

(3.20)
The rest is to bound the right-hand side of the above inequality (3.20). By the Hölder’s inequality and the Pinsker’s inequality, we first have

\[
\sum_{h=1}^{H} \langle Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}, \tilde{\pi}_{h}^{k-1} - \pi_{h}^{k-1} \rangle = \frac{1}{\alpha} \sum_{h=1}^{H} KL(\pi_{h}^{k} \mid \tilde{\pi}_{h}^{k-1})
\]

\[
\sum_{h=1}^{H} \langle Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}, \tilde{\pi}_{h}^{k-1} - \pi_{h}^{k-1} \rangle \cap \sum_{h=1}^{H} \langle Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}, \tilde{\pi}_{h}^{k-1} - \pi_{h}^{k-1} \rangle \\
\leq \sum_{h=1}^{H} \left( \|Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}\|_{\infty} \|\tilde{\pi}_{h}^{k} - \pi_{h}^{k-1}\|_{1} - \frac{1}{2\alpha} \|\tilde{\pi}_{h}^{k} - \pi_{h}^{k-1}\|_{1}^{2} \right)
\]

Then, using the square completion,

\[
\|Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}\|_{\infty} \|\tilde{\pi}_{h}^{k} - \pi_{h}^{k-1}\|_{1} - \frac{1}{2\alpha} \|\tilde{\pi}_{h}^{k} - \pi_{h}^{k-1}\|_{1}^{2} = -\frac{1}{2\alpha} (\alpha \|Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}\|_{\infty} - \|\tilde{\pi}_{h}^{k} - \pi_{h}^{k-1}\|_{1})^{2} + \frac{\alpha}{2} \|Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}\|_{\infty}^{2}
\]

\[
\leq \frac{\alpha}{2} \|Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}\|_{\infty}^{2}
\]

where we drop off the first quadratic term for the inequality, and \(\|\tilde{\pi}_{h}^{k-1} - \pi_{h}^{k-1}\|_{1} \leq \theta\), we have

\[
\sum_{h=1}^{H} \langle Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}, \tilde{\pi}_{h}^{k-1} - \pi_{h}^{k-1} \rangle = \frac{1}{\alpha} \sum_{h=1}^{H} KL(\pi_{h}^{k} \mid \tilde{\pi}_{h}^{k-1})
\]

\[
\sum_{h=1}^{H} \langle Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}, \tilde{\pi}_{h}^{k-1} - \pi_{h}^{k-1} \rangle \cap \sum_{h=1}^{H} \langle Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}, \tilde{\pi}_{h}^{k-1} - \pi_{h}^{k-1} \rangle \\
\leq \frac{\alpha}{2} \sum_{h=1}^{H} \|Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}\|_{\infty}^{2} + \theta \sum_{h=1}^{H} \|Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}\|_{\infty}
\]

\[
\leq \frac{\alpha(1 + \chi)^{2}H^{3}}{2} + \theta (1 + \chi) H^{2}
\]
where the last inequality is due to \( \|Q_{r,h}^{k-1}\|_\infty \leq H \), a fact from line 12 in Algorithm 7, and \( 0 \leq Y^{k-1} \leq \chi \). Taking the same expectation \( \mathbb{E}_{\pi^*} \) as previously on both sides of (3.21) and substituting it into the left-hand side of (3.20) yield,

\[
(V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{k-1}(s_1)) + Y^{k-1}(V_{g,1}^{\pi^*}(s_1) - V_{g,1}^{k-1}(s_1)) \leq \frac{\alpha (1 + \chi)^2 H^3}{2} + \theta (1 + \chi) H^2 + \frac{1}{\alpha} \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[D_{KL}(\pi_h^* \| \tilde{\pi}_h^{k-1}) - D_{KL}(\pi_h^* \| \pi_h^k)] \\
+ \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[l_{r,h}^{k-1}(s_h, a_h)] + Y^{k-1} \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[l_{g,h}^{k-1}(s_h, a_h)] \leq \frac{\alpha (1 + \chi)^2 H^3}{2} + \theta (1 + \chi) H^2 + \frac{1}{\alpha} \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[D_{KL}(\pi_h^* \| \tilde{\pi}_h^{k-1}) - D_{KL}(\pi_h^* \| \pi_h^k)] \\
+ \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[l_{r,h}^{k-1}(s_h, a_h)] + Y^{k-1} \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[l_{g,h}^{k-1}(s_h, a_h)].
\]  

(3.22)

where in the second inequality we note the fact that \( D_{KL}(\pi_h^* \| \tilde{\pi}_h^{k-1}) - D_{KL}(\pi_h^* \| \pi_h^k) \leq \theta \log |A| \) from Lemma 54.

We note that \( Y^0 \) is initialized to be zero. By taking a telescoping sum of both sides of (3.22) from \( k = 1 \) to \( k = K + 1 \) and shifting the index \( k \) by one, we have

\[
\sum_{k=1}^{K} (V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{k}(s_1)) + \sum_{k=1}^{K} Y^{k}(V_{g,1}^{\pi^*}(s_1) - V_{g,1}^{k}(s_1)) \leq \frac{\alpha (1 + \chi)^2 H^3(K + 1)}{2} + \theta (1 + \chi) H^2(K + 1) + \frac{\theta H(K + 1) \log |A|}{\alpha} + \frac{H \log |A|}{\alpha} \\
+ \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[l_{r,h}^{k}(s_h, a_h)] + \sum_{k=1}^{K} \sum_{h=1}^{H} Y^{k}\mathbb{E}_{\pi^*}[l_{g,h}^{k}(s_h, a_h)].
\]  

(3.23)
where we ignore $-\alpha^{-1} \sum_{h=1}^{H} \mathbb{E}_{\pi^{*}}[D(\pi^{*}_h \mid \pi^{K+1}_h)]$ and utilize

$$D_{KL}(\pi^{*}_h \mid \pi^{0}_h) = \sum_{a \in A} \pi^{*}_h(a \mid s_h) \log (|A| \pi^{*}_h(a \mid s_h)) \leq \log |A|$$

where $\pi^{0}_h$ is uniform over $A$ and we ignore $\sum_{a \in A} \pi^{*}_h(a \mid s_h) \log (\pi^{*}_h(a \mid s_h))$ that is nonpositive.

Finally, we take $\chi := H/\gamma$ and $\alpha, \theta$ in the lemma to complete the proof. □

By the dual update of Algorithm 6, we can simplify the result in Lemma 18 and return back to the regret (3.7).

**Lemma 19** Let Assumptions 7 and 8 hold. In Algorithm 6, if we set $\alpha = \sqrt{\log |A|/(H^2 \sqrt{K})}$, $\eta = 1/\sqrt{K}$, and $\theta = 1/K$, then with probability $1 - p/2$,

$$\text{Regret}(K) = C_3 H^{2.5} \sqrt{T \log |A|} + \sum_{k=1}^{K} \sum_{h=1}^{H} \left( \mathbb{E}_{\pi^{*}} \left[ \iota_{r,h}^{k}(s_h, a_h) \right] - \iota_{r,h}^{k}(s_h, a_h) \right) + M_{r,H,2}^{K} (3.24)$$

where $C_3$ is an absolute constant.

**Proof.** By the dual update in line 9 in Algorithm 6, we have

$$0 \leq \left( Y^{K+1} \right)^2$$

$$= \sum_{k=1}^{K+1} \left( \left( Y^{k} \right)^2 - \left( Y^{k-1} \right)^2 \right)$$

$$= \sum_{k=1}^{K+1} \left( \mathbb{P}_{\pi^{0},\chi} \left( Y^{k-1} + \eta (b - V_{g,1}^{k-1}(s_1)) \right)^2 - \left( Y^{k-1} \right)^2 \right)$$

$$\leq \sum_{k=1}^{K+1} \left( Y^{k-1} + \eta (b - V_{g,1}^{k-1}(s_1)) \right)^2 - \left( Y^{k-1} \right)^2$$

$$\leq \sum_{k=1}^{K+1} 2\eta Y^{k-1} \left( V_{g,1}^{\pi^{*}}(s_1) - V_{g,1}^{k-1}(s_1) \right) + \eta^2 (b - V_{g,1}^{k-1}(s_1))^2.$$
where we use the feasibility of $\pi^*$ in the last inequality. Since $Y^0 = 0$ and $|b - V_{g,1}^{k-1}(s_1)| \leq H$, the above inequality implies that

$$- \sum_{k=1}^{K} Y^k (V_{g,1}^\pi(s_1) - V_{g,1}^k(s_1)) \leq \sum_{k=1}^{K+1} \frac{\eta}{2} (b - V_{g,1}^{k-1}(s_1))^2 \leq \frac{\eta H^2 (K + 1)}{2}. \quad (3.25)$$

By noting the UCB result (3.16) and $Y^k \geq 0$, the inequality (3.17) implies that

$$\sum_{k=1}^{K} (V_{r,1}^\pi(s_1) - V_{r,1}^k(s_1)) + \sum_{k=1}^{K} Y^k (V_{g,1}^\pi(s_1) - V_{g,1}^k(s_1)) \leq C_2 H^{2.5} \sqrt{T \log |A|} + \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ l_{r,h}^k(s_h, a_h) \right]. \quad (3.26)$$

If we add (3.25) to the above inequality and take $\eta = 1/\sqrt{K}$, then,

$$\sum_{k=1}^{K} (V_{r,1}^\pi(s_1) - V_{r,1}^k(s_1)) \leq C_3 H^{2.5} \sqrt{T \log |A|} + \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ l_{r,h}^k(s_h, a_h) \right] \quad (3.26)$$

where $C_3$ is an absolute constant. Finally, we combine (3.14) and (3.26) to complete the proof. □

By Lemma 19, the rest is to bound the last two terms in the right-hand side of (3.24). We next show two probability bounds for them in Lemma 20 and Lemma 21 separately.

**Lemma 20 (Model prediction error bound)** Let Assumption 8 hold. Fix $p \in (0, 1)$. If we set $\beta = C_1 \sqrt{dH^2 \log (d T / p)}$ in Algorithm 6, then with probability $1 - p/2$ it holds that

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \left( \mathbb{E}_{\pi^*} \left[ l_{r,h}^k(s_h, a_h) \right] - l_{r,h}^k(s_h^k, a_h^k) \right) \leq 4C_1 \sqrt{2 d^2 H^3 T \log (K + 1) \log \left( \frac{d T}{p} \right)} \quad (3.27)$$

where $C_1$ is an absolute constant and $T = HK$. 

100
Proof. By the UCB result (3.16), with probability \(1 - p/2\) for any \((k, h) \in [K] \times [H]\) and \((s, a) \in S \times A\), we have

\[-2(\Gamma^k_h + \Gamma^k_{r,h})(s, a) \leq \iota^k_{r,h}(x, a) \leq 0.\]

By the definition of \(\iota^k_{r,h}(s, a), |\iota^k_{r,h}(s, a)| \leq 2H\). Hence, it holds with probability \(1 - p/2\) that

\[-\iota^k_{r,h}(s, a) \leq 2 \min \left( H, (\Gamma^k_h + \Gamma^k_{r,h})(s, a) \right)\]

for any \((k, h) \in [K] \times [H]\) and \((s, a) \in S \times A\). Therefore, we have

\[
\sum_{k=1}^{K} \sum_{h=1}^{H} \left( \mathbb{E}_{\pi^*} \left[ \iota^k_{r,h}(s_h, a_h) \mid s_1 \right] - \iota^k_{r,h}(s_h^k, a_h^k) \right) \leq 2 \sum_{k=1}^{K} \sum_{h=1}^{H} \min \left( H, (\Gamma^k_h + \Gamma^k_{r,h})(s_h^k, a_h^k) \right)
\]

where \(\Gamma^k_h(\cdot, \cdot) = \beta(\varphi(\cdot, \cdot)\top \Lambda^k_h)^{-1}\varphi(\cdot, \cdot)^{1/2}\) and \(\Gamma^k_{r,h}(\cdot, \cdot) = \beta(\phi^k_{r,h}(\cdot, \cdot)\top \Lambda^k_{r,h})^{-1}\phi^k_{r,h}(\cdot, \cdot)^{1/2}\). Application of the Cauchy-Schwartz inequality shows that

\[
\sum_{k=1}^{K} \sum_{h=1}^{H} \min \left( H, (\Gamma^k_h + \Gamma^k_{r,h})(s_h^k, a_h^k) \right)
\leq \beta \sum_{k=1}^{K} \sum_{h=1}^{H} \min \left( \frac{H}{\beta}, \left( \varphi(s_h^k, a_h^k)\top \Lambda^k_h \right)^{-1}\varphi(s_h^k, a_h^k)^{1/2} + \left( \phi^k_{r,h}(s_h^k, a_h^k)\top \Lambda^k_{r,h} \right)^{-1}\phi^k_{r,h}(s_h^k, a_h^k)^{1/2} \right). \tag{3.28}
\]

Since we take \(\beta = C_1 \sqrt{dH^2 \log (dT/p)}\) with \(C_1 > 1\), we have \(H/\beta \leq 1\). The rest is to apply Lemma 52. First, for any \(h \in [H]\) it holds that

\[
\sum_{k=1}^{K} \phi^k_{r,h}(s_h^k, a_h^k)\top \Lambda^k_{r,h}^{-1}\phi^k_{r,h}(s_h^k, a_h^k) \leq 2 \log \left( \frac{\det(\Lambda^{K+1}_{r,h})}{\det(\Lambda^1_{r,h})} \right).
\]
Due to \( \| \phi^k_{r,h} \| \leq \sqrt{dH} \) in Assumption 8 and \( \Lambda^1_{r,h} = \lambda I \) in Algorithm 7, it is clear that for any \( h \in [H] \),

\[
\Lambda^{K+1}_{r,h} = \sum_{k=1}^{K} \phi^k_{r,h}(s^k_{h}, a^k_{h}) \phi^k_{r,h}(s^k_{h}, a^k_{h})^\top + \lambda I \preceq (dH^2K + \lambda)I.
\]

Thus,

\[
\log \left( \frac{\det (\Lambda^{K+1}_{r,h})}{\det (\Lambda^1_{r,h})} \right) \leq \log \left( \frac{\det ((dH^2K + \lambda)I)}{\det (\lambda I)} \right) \leq d \log \left( \frac{dH^2K + \lambda}{\lambda} \right).
\]

Therefore,

\[
\sum_{k=1}^{K} \phi^k_{r,h}(s^k_{h}, a^k_{h})^\top \left( \Lambda^k_{h} \right)^{-1} \phi^k_{r,h}(s^k_{h}, a^k_{h}) \leq 2d \log \left( \frac{dH^2K + \lambda}{\lambda} \right). \tag{3.29}
\]

Similarly, we can show that

\[
\sum_{k=1}^{K} \varphi(s^k_{h}, a^k_{h})^\top \left( \Lambda^k_{h} \right)^{-1} \varphi(s^k_{h}, a^k_{h}) \leq 2d \log \left( \frac{dK + \lambda}{\lambda} \right). \tag{3.30}
\]

Applying the above inequalities (3.29) and (3.30) to (3.28) leads to

\[
\sum_{k=1}^{K} \sum_{h=1}^{H} \min \left( H, (\Gamma^k_{h} + \Gamma^k_{r,h})(s^k_{h}, a^k_{h}) \right) \\
\leq \beta \sum_{h=1}^{H} \min \left( K, \sum_{k=1}^{K} \left( \varphi(s^k_{h}, a^k_{h})^\top \left( \Lambda^k_{h} \right)^{-1} \varphi(s^k_{h}, a^k_{h}) \right)^{1/2} + \left( \phi^k_{r,h}(s^k_{h}, a^k_{h})^\top \left( \Lambda^k_{r,h} \right)^{-1} \phi^k_{r,h}(s^k_{h}, a^k_{h}) \right)^{1/2} \right) \\
\leq \beta \sum_{h=1}^{H} \left( K \sum_{k=1}^{K} \varphi(s^k_{h}, a^k_{h})^\top \left( \Lambda^k_{h} \right)^{-1} \varphi(s^k_{h}, a^k_{h}) \right)^{1/2} + \left( K \sum_{k=1}^{K} \phi^k_{r,h}(s^k_{h}, a^k_{h})^\top \left( \Lambda^k_{r,h} \right)^{-1} \phi^k_{r,h}(s^k_{h}, a^k_{h}) \right)^{1/2} \\
\leq \beta \sum_{h=1}^{H} \sqrt{K} \left( 2d \log \left( \frac{dK + \lambda}{\lambda} \right) \right)^{1/2} + \left( 2d \log \left( \frac{dH^2K + \lambda}{\lambda} \right) \right)^{1/2}
\]

Finally, we set \( \beta = C_1 \sqrt{dH^2 \log (dT/p)} \) and \( \lambda = 1 \) to obtain (3.27). \( \square \)
Lemma 21 (Matingale bound) Fix $p \in (0, 1)$. In Algorithm 6, it holds with probability $1 - p/2$ that
\[
|M_{r,H,2}^K| \leq 4 \sqrt{H^2 T \log \left( \frac{4}{p} \right)} \tag{3.31}
\]
where $T = HK$.

PROOF. In the verification of (3.14) (see Appendix B.3), we introduce the following martingale,
\[
M_{r,H,2}^K = \sum_{k=1}^K \sum_{h=1}^H (D_{r,h,1}^k + D_{r,h,2}^k)
\]
where
\[
D_{r,h,1}^k = \left( I_{h}^k (Q_{r,h}^k - Q_{r,h}^{\pi_k,k}) \right) (s_h^k) - \left( Q_{r,h}^k - Q_{r,h}^{\pi_k,k} \right) (s_h^k, a_h^k)
\]
\[
D_{r,h,2}^k = \left( P_{h}^k V_{r,h+1}^k - P_{h}^k V_{r,h+1}^{\pi_k,k} \right) (s_h^k, a_h^k) - \left( V_{r,h+1}^k - V_{r,h+1}^{\pi_k,k} \right) (s_{h+1}^k)
\]
where \( (I_{h}^k f)(s) := \langle f(s, \cdot), \pi_h^k(\cdot|s) \rangle \).

Due to the truncation in line 11 of Algorithm 7, we know that $Q_{r,h}^k, Q_{r,h}^{\pi_k,k}, V_{r,h+1}^k, V_{r,h+1}^{\pi_k,k} \in [0, H]$. This shows that $|D_{r,h,1}^k|, |D_{r,h,2}^k| \leq 2H$ for all $(k, h) \in [K] \times [H]$. Application of the Azuma-Hoeffding inequality yields,
\[
P\left( |M_{r,H,2}^K| \geq s \right) \leq 2 \exp \left( \frac{-s^2}{16H^2 T} \right).
\]
For $p \in (0, 1)$, if we set $s = 4H \sqrt{T \log (4/p)}$, then the inequality (3.31) holds with probability at least $1 - p/2$. \(\square\)
We now are ready to show the desired regret bound. Applying (3.27) and (3.31) to the right-hand side of the inequality (3.24), we have

\[
\text{Regret}(K) \leq C_3 H^{2.5} \sqrt{T \log |A|} + 2C_1 \sqrt{2d^2 H^3 T \log (K + 1) \log \left( \frac{dT}{p} \right)} + 4 \sqrt{H^2 T \log \left( \frac{1}{p} \right)}
\]

with probability $1 - p$ where $C_1, C_3$ are absolute constants. Then, with probability $1 - p$ it holds that

\[
\text{Regret}(K) \leq C d H^{2.5} \sqrt{T \log \left( \frac{dT}{p} \right)}
\]

where $C$ is an absolute constant.

### 3.4.3 Proof of constraint violation

In Lemma 18 we have provided a useful upper bound on the total differences that are weighted by the dual update $Y^k$. To extract the constraint violation, we first refine Lemma 18 as follows.

**Lemma 22 (Policy improvement: refined primal-dual mirror descent)** Let Assumptions 7-8 hold. In Algorithm 6, if we set $\alpha = \sqrt{\log |A| / (H^2 \sqrt{K})}$, $\theta = 1/K$, and $\eta = 1/\sqrt{K}$, then

Then, for any $Y \in [0, \chi]$, with probability $1 - p/2$,

\[
\sum_{k=1}^{K} (V_{\pi^*_{r,1}}(s_1) - V_{\pi^*_{r,1}}(s_1)) + Y \sum_{k=1}^{K} \left(b - V^k_{g,1}(s_1)\right) \leq C_4 H^{2.5} \sqrt{T \log |A|}
\]  

(3.32)

where $C_4$ is an absolute constant, $T = HK$, and $\chi := H/\gamma$. 

104
Proof. By the dual update in line 10 in Algorithm 6 for any $Y \in [0, \chi]$ we have

$$|Y^{k+1} - Y|^2 = |\mathcal{P}_{[0, \chi]}(Y^k + \eta(b - V^k_{g,1}(s_1))) - \mathcal{P}_{[0, \chi]}(Y)|^2 \leq |Y^k + \eta(b - V^k_{g,1}(s_1)) - Y|^2 \leq (Y^k - Y)^2 + 2\eta(b - V^k_{g,1}(s_1))(Y^k - Y) + \eta^2 H^2$$

where we apply the non-expansiveness of projection in the first inequality and $|b - V^k_{g,1}(s_1)| \leq H$ for the last inequality. By summing the above inequality from $k = 1$ to $k = K$, we have

$$0 \leq |Y^{K+1} - Y|^2 = |Y^1 - Y|^2 + 2\eta \sum_{k=1}^{K} (b - V^k_{g,1}(s_1))(Y^k - Y) + \eta^2 H^2 K$$

which implies that

$$\sum_{k=1}^{K} (b - V^k_{g,1}(s_1))(Y^k - Y) \leq \frac{1}{2\eta} |Y^1 - Y|^2 + \frac{\eta}{2} H^2 K.$$

By adding the above inequality to (3.23) in Lemma 18 and noting that $V^\pi(r,1)(s_1) \geq b$ and the UCB result (3.16), we have

$$\sum_{k=1}^{K} (V^\pi_{r,1}(s_1) - V^k_{r,1}(s_1)) + \sum_{k=1}^{K} (b - V^k_{g,1}(s_1)) \leq \frac{\alpha(1 + \chi)^2 H^3 (K + 1)}{2} + \theta (1 + \chi) H^2 (K + 1) + \frac{\theta H (K + 1) \log |A|}{\alpha} + \frac{H \log |A|}{\alpha} + \frac{1}{2\eta} |Y^1 - Y|^2 + \frac{\eta}{2} H^2 K.$$

By taking $\chi = H/\gamma$, and $\alpha, \theta, \eta$ in the lemma, we complete the proof. \qed
According to Lemma 22, we can multiply (3.15) by $Y \geq 0$ and add it, together with (3.14), to (3.32),

$$
\sum_{k=1}^{K} (V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{\pi_k}(s_1)) + Y \sum_{k=1}^{K} (b - V_{g,1}^{\pi_k}(s_1)) \\
\leq C_4 H^{2.5} \sqrt{T \log |A|} - \sum_{k=1}^{K} \sum_{h=1}^{H} \eta_{r,h}(s_k^{(h)}, a_{h}) - Y \sum_{k=1}^{K} \sum_{h=1}^{H} \eta_{g,h}(s_k^{(h)}, a_{h}) + M_{r,H,2}^{K} + Y M_{g,H,2}^{K}.
$$

(3.33)

We now are ready to show the desired constraint violation bound. We note that there exists a policy $\pi'$ such that $V_{r,1}^{\pi'}(s_1) = \frac{1}{K} \sum_{k=1}^{K} V_{r,1}^{\pi_k}(s_1)$ and $V_{g,1}^{\pi'}(s_1) = \frac{1}{K} \sum_{k=1}^{K} V_{g,1}^{\pi_k}(s_1)$. By the occupancy measure method [11], $V_{r,1}^{\pi_k}(s_1)$ and $V_{g,1}^{\pi_k}(s_1)$ are linear in terms of an occupancy measure induced by policy $\pi^k$ and initial state $s_1$. Thus, an average of $K$ occupancy measures is still an occupancy measure that produces policy $\pi'$ with values $V_{r,1}^{\pi'}(s_1)$ and $V_{g,1}^{\pi'}(s_1)$. Particularly, we take $Y = 0$ when $\sum_{k=1}^{K} (b - V_{g,1}^{\pi_k}(s_1)) < 0$; otherwise $Y = \chi$. Therefore, we have

$$
V_{r,1}^{\pi^*}(s_1) - \frac{1}{K} \sum_{k=1}^{K} V_{r,1}^{\pi_k}(s_1) + \chi \left[ b - \frac{1}{K} \sum_{k=1}^{K} V_{g,1}^{\pi_k}(s_1) \right] + \\
= V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{\pi'}(s_1) + \chi \left[ b - V_{g,1}^{\pi'}(s_1) \right] + \\
\leq \frac{C_4 H^{2.5} \sqrt{T \log |A|}}{K} - \frac{1}{K} \sum_{k=1}^{K} \sum_{h=1}^{H} \eta_{r,h}(s_k^{(h)}, a_{h}) - \frac{\chi}{K} \sum_{k=1}^{K} \sum_{h=1}^{H} \eta_{g,h}(s_k^{(h)}, a_{h}) \\
+ \frac{1}{K} M_{r,H,2}^{K} + \frac{\chi}{K} M_{g,H,2}^{K} \\
\leq \frac{C_4 H^{2.5} \sqrt{T \log |A|}}{K} + \frac{1}{K} \sum_{k=1}^{K} \sum_{h=1}^{H} (\Gamma_{r,h}^k + \Gamma_{r,h}^k) (s_k^{(h)}, a_{h}) - \frac{\chi}{K} \sum_{k=1}^{K} \sum_{h=1}^{H} (\Gamma_{g,h}^k + \Gamma_{g,h}^k) (s_k^{(h)}, a_{h}) \\
+ \frac{1}{K} M_{r,H,2}^{K} + \frac{\chi}{K} M_{g,H,2}^{K}
$$

(3.34)

where we apply the UCB result (3.16) for the last inequality.

106
Finally, we recall two immediate results of Lemma 20 and Lemma 21. Fix $p \in (0, 1)$, the proof of Lemma 20 also shows that with probability $1 - p/2$,

$$\sum_{k=1}^{K} \sum_{h=1}^{H} (\Gamma^k_h + \Gamma^k_{c,h}) (s^k_h, a^k_h) \leq C_1 \sqrt{2d^2 H^3 \log (K + 1) \log \left( \frac{dT}{p} \right)}$$

(3.35)

and the proof of Lemma 21 shows that with probability $1 - p/2$,

$$\left| M^K_{g,H,2} \right| \leq 4 \sqrt{H^2 T \log \left( \frac{4}{p} \right)}.$$

If we take $\log |A| = O(d^2 \log^2 (dT/p))$, (3.34) implies that with probability $1 - p$ we have

$$V^{\pi^*}_{r,1}(s_1) - V^{\pi'}_{r,1}(s_1) + \chi \left[ b - V^{\pi'}_{g,1}(s_1) \right]_+ \leq C_5 d H^{2.5} \sqrt{T} \log \left( \frac{dT}{p} \right)$$

where $C_5$ is an absolute constant. Finally, by noting our choice of $\chi \geq 2Y^*$, we can apply Lemma 49 to conclude that

$$[\text{Violation}(K)]_+ \leq C' d H^{2.5} \sqrt{T} \log \left( \frac{dT}{p} \right)$$

with probability $1 - p$, where $C'$ is an absolute constant.
## 3.5 Further results on the tabular case

The proof of Theorem 16 is generic, since it is ready to achieve sublinear regret and constraint violation bounds as long as the policy evaluation is sample-efficient, e.g., the UCB design of optimism in the face of uncertainty. In what follows, we introduce another efficient policy evaluation for line 11 of Algorithm 6 in the tabular case. Let us first introduce some notation. For any \((h, k) \in [H] \times [K]\), any \((s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}\), and any \((s, a) \in \mathcal{S} \times \mathcal{A}\), we define two visitation counters \(n^k_h(s, a, s')\) and \(n^k_h(s, a)\) at step \(h\) in episode \(k\),

\[
\begin{align*}
    n^k_h(s, a, s') &= \sum_{\tau=1}^{k-1} \mathbb{1}\{(s, a, s') = (s^\tau_h, a^\tau_h, a_{h+1}^\tau)\} \\
    n^k_h(s, a) &= \sum_{\tau=1}^{k-1} \mathbb{1}\{(s, a) = (s^\tau_h, a^\tau_h)\}.
\end{align*}
\]  

(3.36)

This allows us to estimate reward function \(r\), utility function \(g\), and transition kernel \(P_h\) for episode \(k\) by

\[
\begin{align*}
    \hat{r}^k_h(s, a) &= \sum_{\tau=1}^{k-1} \mathbb{1}\{(s, a) = (s^\tau_h, a^\tau_h)\} \frac{r_h(s^\tau_h, a^\tau_h)}{n^k_h(s, a) + \lambda} \\
    \hat{g}^k_h(s, a) &= \sum_{\tau=1}^{k-1} \mathbb{1}\{(s, a) = (s^\tau_h, a^\tau_h)\} \frac{g_h(s^\tau_h, a^\tau_h)}{n^k_h(s, a) + \lambda} \\
    \hat{P}^k_h(s' | s, a) &= \frac{n^k_h(s, a, s')}{n^k_h(s, a) + \lambda}
\end{align*}
\]  

(3.37)

(3.38)

for all \((s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}\), \((s, a) \in \mathcal{S} \times \mathcal{A}\) where \(\lambda > 0\) is the regularization parameter.

Moreover, we introduce the bonus term \(\Gamma^k_h: \mathcal{S} \times \mathcal{A} \to \mathbb{R}, \Gamma^k_h(s, a) = \beta \left(n^k_h(s, a) + \lambda\right)^{-1/2}\) which adapts the counter-based bonus terms in \([19, 108]\), where \(\beta > 0\) is to be determined later.

Using the estimated transition kernels \(\{\hat{P}^k_h\}_{h=1}^H\), reward/utility functions \(\{\hat{r}^k_h, \hat{g}^k_h\}_{h=1}^H\), and the bonus terms \(\{\Gamma^k_h\}_{h=1}^H\), we now can estimate the state-action function via line 7 of Algorithm 8 for
any \((s, a) \in S \times A\), where \(\diamond = r\) or \(g\). Thus, \(V_{\diamond,h}^k(s) = \langle Q_{\diamond,h}^k(s, \cdot), \pi_h^k(\cdot \mid s) \rangle_A\). We summarize the above procedure in Algorithm 8. Using already estimated \(\{Q_{r,h}^k(\cdot, \cdot), Q_{g,h}^k(\cdot, \cdot)\}_{h=1}^H\), we execute the policy improvement and the dual update in Algorithm 6.

As Theorem 16, we provide theoretical guarantees in Theorem 23 which improves \((|S|, |A|)\) dependence in Theorem 16 for the tabular case and also matches \(|S|\) dependence in \([85, 187]\). It is worthy mentioning our Algorithm 6 is generic in handling an infinite state space.

**Theorem 23 (Tabular case: regret and constraint violation)** Let Assumption 7 hold and let Assumption 8 hold with features (3.6). Fix \(p \in (0, 1)\). In Algorithm 6 we set \(\alpha = \sqrt{\log |A|/(H^2 K)}\), \(\beta = C_1 H \sqrt{|S| \log(|S||A| T/p)}\), \(\eta = 1/\sqrt{K}\), \(\theta = 1/K\), and \(\lambda = 1\) where \(C_1\) is an absolute constant. Then, with probability \(1 - p\), the regret and the constraint violation in (3.3) satisfy

\[
\text{Regret}(K) \leq C |S| \sqrt{|A| H^5 T \log \left( \frac{|S||A|T}{p} \right)}
\]

\[
[\text{Violation}(K)]_+ \leq C' |S| \sqrt{|A| H^5 T \log \left( \frac{|S||A|T}{p} \right)}
\]

where \(C\) and \(C'\) are absolute constants.

**Proof.** See Appendix B.1 \(\square\)

### 3.6 Concluding remarks

We have developed a provably efficient safe reinforcement learning algorithm in the linear MDP setting. The algorithm extends the proximal policy optimization to constrained MDPs by incorporating the UCB exploration. We prove that the proposed algorithm achieves an \(\tilde{O}(\sqrt{T})\) regret and an \(\tilde{O}(\sqrt{T})\) constraint violation under mild conditions, where \(T\) is the total number of steps.
taken by the algorithm. Our algorithm works in the setting where reward/utility functions are
given by bandit feedback. To the best of our knowledge, our algorithm is the first provably effi-
cient online policy optimization algorithm for constrained MDPs in the function approximation
setting.
Part II

Reinforcement learning for multi-agent control systems
Chapter 4

Multi-agent temporal-difference learning for multi-agent MDPs

In this chapter, we study the policy evaluation problem in multi-agent reinforcement learning where a group of agents, with jointly observed states and private local actions and rewards, collaborate to learn the value function of a given policy via local computation and communication over a connected undirected network. This problem arises in various large-scale multi-agent systems, including power grids, intelligent transportation systems, wireless sensor networks, and multi-agent robotics. When the dimension of state-action space is large, the temporal-difference learning with linear function approximation is widely used. In this chapter, we develop a new distributed temporal-difference learning algorithm and quantify its finite-time performance. Our algorithm combines a distributed stochastic primal-dual method with a homotopy-based approach to adaptively adjust the learning rate in order to minimize the mean-square projected Bellman error by taking fresh online samples from a causal on-policy trajectory. We explicitly take into account the Markovian nature of sampling and improve the best-known finite-time error bound from $O(1/\sqrt{T})$ to $O(1/T)$, where $T$ is the total number of iterations.
4.1 Introduction

Temporal-difference (TD) learning is a central approach to policy evaluation in modern reinforcement learning (RL) \[214\]. It was introduced in \[213, 31, 26\] and significant advances have been made in a host of single-agent decision-making applications, including Atari or Go games \[165, 203\]. Recently, TD learning has been used to address multi-agent decision making problems for large-scale systems, including power grids \[163\], intelligent transportation systems \[120\], wireless sensor networks \[182\], and multi-agent robotics \[119\]. Motivated by these applications, we introduce an extension of TD learning to a distributed setting of policy evaluation. This setup involves a group of agents that communicate over a connected undirected network. All agents share a joint state and dynamics of state transition are governed by the local actions of agents which follow a local policy and own a private local reward. To maximize the total reward, i.e., the sum of all local rewards, it is essential to quantify performance that each agent achieves if it follows a particular policy while interacting with the environment and using only local data and information exchange with its neighbors. This task is usually referred to as a distributed policy evaluation problem and it has received significant recent attention \[148, 155, 127, 232, 48, 76, 77, 212, 196\].

In the context of distributed policy evaluation, several attempts have been made to extend TD algorithms to a multi-agent setup using linear function approximators. When the reward is global and actions are local, mean square convergence of a distributed primal-dual gradient temporal-difference (GTD) algorithm for minimizing mean-square projected Bellman error (MSPBE), with diffusion updates, was established in \[148\] and an extension to time-varying networks was made in \[208\]. In \[155\], the authors combined the gossip averaging scheme \[40\] with TD(0) and showed
asymptotic convergence. In the off-line setting, references [232] and [48] proposed different consensus-based primal-dual algorithms for minimizing a batch-sampled version of MSPBE with linear convergence; a fully asynchronous gossip-based extension was studied in [196] and its communication efficiency was analyzed in [188]. To understand/gain data efficiency, the recent focus of multi-agent TD learning research has shifted to finite-time or finite-sample performance analysis. For distributed TD(0) and TD(\(\lambda\)) with local rewards, \(O(1/T)\) error bound was established in [76] and [77], respectively. Linear convergence of distributed TD(0) to a neighborhood of the stationary point was proved in [212] and \(O(1/\sqrt{T})\) error bound for distributed GTD was shown in [127]. In [78], \(O(1/T^{2/3})\) error bound was provided for a distributed variant of two-time-scale stochastic approximation algorithm. Apart from [77, 212], other finite sample results rely on the i.i.d. state sampling in policy evaluation. In most RL applications, this assumption is overly restrictive because of the Markovian nature of state trajectory samples. In [96], an example was provided to demonstrate that i.i.d. sampling-based convergence guarantees can fail to hold when samples become correlated. It is thus relevant to examine how to design an online distributed learning algorithm for the policy evaluation problem (e.g., MSPBE minimization) in the Markovian setting. Such distributed learning algorithms are essential in multi-agent RL; e.g., see the distributed variant of policy gradient theorem [275] along with recent surveys [273, 272, 128].

In Section 4.2, we introduce a class of multi-agent stochastic saddle point problems that contain, as a special instance, minimization of a mean square projected Bellman error via distributed TD learning. In Section 4.3, we develop a homotopy-based online distributed primal-dual algorithm to solve this problem and establish a finite-time performance bound for the proposed algorithm. In Section 4.4, we prove the main result, in Section 4.5, we offer computational experiments
to demonstrate the merits and the effectiveness of our theoretical findings and, in Section 4.6, we close the chapter with concluding remarks.

4.2 Problem formulation and background

In this section, we formulate a multi-agent stochastic saddle point problem over a connected undirected network. The motivation for studying this class of problems comes from distributed reinforcement learning where a group of agents with jointly observed states and private local actions/rewards collaborate to learn the value function of a given policy via local computation and communication. We exploit the structure of the underlying optimization problem to demonstrate that it enables unbiased estimation of the saddle point objective from Markovian samples. Furthermore, we discuss an algorithm that is convenient for distributed implementation and finite-time performance analysis.

4.2.1 Multi-agent stochastic optimization problem

We consider a stochastic optimization problem over a connected undirected network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $N$ agents,

$$
\min_{x \in \mathcal{X}} \frac{1}{N} \sum_{j=1}^{N} f_j(x) \tag{4.1a}
$$

where $\mathcal{V} = \{1, \ldots, N\}$ is the set of nodes, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges, $x$ is the optimization variable, $\mathcal{X} \subset \mathbb{R}^d$ is a convex set, and $f_j: \mathcal{X} \to \mathbb{R}$ is a local objective function determined by,

$$
f_j(x) = \max_{y_j \in \mathcal{Y}} \mathbb{E}_{\xi \sim \Pi}[\Psi_j(x, y_j; \xi)]. \tag{4.1b}
$$
Here, $y_j$ is a local variable that belongs to a convex set $\mathcal{Y} \subset \mathbb{R}^d$ and $\Psi_j(x, y_j; \xi)$ is a stochastic function of a random variable $\xi$ which is distributed according to the stationary distribution $\Pi$ of a Markov chain. Equivalently, problem (4.1) can be cast as a multi-agent stochastic saddle point problem,

$$\min_{x \in \mathcal{X}} \max_{y_j \in \mathcal{Y}} \frac{1}{N} \sum_{j=1}^{N} \psi_j(x, y_j)$$

(4.2)

with the primal variable $x$ and the dual variable $y := (y_1, \ldots, y_N)$.

Each agent $j$ can only communicate with its neighbors over the network $\mathcal{G}$ and receive samples $\xi$ from a stochastic process that converges to the stationary distribution $\Pi$. Finite optimal solution $(x^*, y^*)$ to the saddle point problem (4.2) satisfies

$$x^* := \arg\min_{x \in \mathcal{X}} \frac{1}{N} \sum_{j=1}^{N} \psi_j(x, y_j^*)$$

$$y_j^* := \arg\max_{y_j \in \mathcal{Y}} \psi_j(x^*, y_j)$$

and the primary motivation for studying this class of problems comes from multi-agent reinforcement learning where a group of agents with jointly observed states and private local actions/rewards collaborate to learn the value function of a given policy via local computation and communication. Although our theory and algorithm can be readily extended to other settings, we restrict our attention to the Markovian structure in the context of policy evaluation. Formulation (4.1) arises in a host of large-scale multi-agent systems, e.g., in supervised learning [261] and in nonparametric regression [67], and we describe it next.
4.2.2 Multi-agent Markov decision process

Let us consider a control system described by a Markov decision process (MDP) over a connected undirected network $\mathcal{G}$ with $N$ agents, the state space $\mathcal{S}$, and the joint action space $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$. Without loss of generality we assume that, for each agent $j$, the local action space $\mathcal{A}_j$ is the same for all states. Let $P(s' \mid s, a)$ is the transition probability from state $s$ to state $s'$ under a joint action $a = (a_1, \ldots, a_N) \in \mathcal{A}$ and let $r_j(s, a)$ be the local reward received by agent $j$ that corresponds to the pair $(s, a)$. The multi-agent MDP can be represented by the tuple,

$$\left( \mathcal{S}, \{\mathcal{A}_j\}_{j=1}^N, P, \{r_j\}_{j=1}^N, \gamma \right)$$

where $\gamma \in (0, 1)$ is a fixed discount factor.

When the states, actions, and rewards are globally observable, a multi-agent MDP problem simplifies to the classical single-agent problem and the centralized controller can be utilized to solve it. In many applications (e.g., see [273 Section 12.4.1.2]), both the actions $a_j \in \mathcal{A}_j$ and the rewards $r_j(s, a)$ are private and every agent can only communicate with its neighbors over the network $\mathcal{G}$. It is thus critically important to extend single-agent TD learning algorithms to a setup in which only local information exchange is available.

We consider a cooperative learning task in which agents aim to maximize the total reward $(1/N) \sum_j r_j(s, a)$ and, in Fig. 4.1, we illustrate the interaction between the environment and the agents. Let $\pi: \mathcal{S} \times \mathcal{A} \to [0, 1]$ be a joint policy which specifies the probability to take an action
Each agent interacts with the environment by receiving a private reward and taking a local action. $a \in \mathcal{A}$ at state $s \in \mathcal{S}$. We define the global reward $R^\pi(s)$ at state $s \in \mathcal{S}$ under policy $\pi$ to be the expected value of the average of all local rewards,

$$R^\pi(s) = \frac{1}{N} \sum_{j=1}^{N} R^\pi_j(s)$$

(4.3)

where $R^\pi_j(s) := \mathbb{E}_{a \sim \pi(\cdot|s)} [r_j(s, a)]$.

For any fixed joint policy $\pi$, the multi-agent MDP becomes a Markov chain over $\mathcal{S}$ with the probability transition matrix $P^\pi \in \mathbb{R}^{||\mathcal{S}|| \times ||\mathcal{S}||}$, where $P^\pi_{s,s'} = \sum_{a \in \mathcal{A}} \pi(a | s) \mathbb{P}(s' | s, a)$ is the $(s,s')$-element of $P^\pi$. If the Markov chain associated with the policy $\pi$ is aperiodic and irreducible then, for any initial state, it converges to the unique stationary distribution $\Pi$ with a geometric rate [132]; see Assumption [14] for a formal statement.
4.2.3 Multi-agent policy evaluation and temporal-difference learning

Let the value function of a policy \( \pi, V^\pi: S \rightarrow \mathbb{R} \), be defined as the expected value of discounted cumulative rewards,

\[
V^\pi(s) = \mathbb{E} \left[ \sum_{p=0}^{\infty} \gamma^p R^\pi(s_p) \right| s_0 = s, \pi]
\]

where \( s_0 = s \) is the initial state. If we arrange \( V^\pi(s) \) and \( R^\pi(s) \) over all states \( s \in S \) into the vectors \( V^\pi \) and \( R^\pi \), the Bellman equation for \( V^\pi \) can be written as [185],

\[
V^\pi = \gamma \mathbb{P}^\pi V^\pi + \mathbb{R}^\pi. \tag{4.4}
\]

Since it is challenging to evaluate \( V^\pi \) directly for a large state space, we approximate \( V^\pi \) using a family of linear functions \( \{ V_x(s) = \phi(s)^T x, x \in \mathbb{R}^d \} \), where \( x \in \mathbb{R}^d \) is the vector of unknown parameters and \( \phi: S \rightarrow \mathbb{R}^d \) is a known dictionary consisting of \( d \) features. If we arrange \( \{ V_x(s) \}_{s \in S} \) into the vector \( V_x \in \mathbb{R}^{|S|} \), we have \( V_x = \Phi x \) where the \( i \)th row of the matrix \( \Phi \in \mathbb{R}^{|S| \times d} \) is given by \( \phi(s_i)^T \). Since the dictionary is a function determined by, e.g., polynomial basis, it is not restrictive to assume that the matrix \( \Phi \) has the full column rank [31].

The goal of policy evaluation now becomes to determine the vector of feature weights \( x \in \mathbb{R}^d \) so that \( V_x \) approximates the true value function \( V^\pi \). The objective of a typical TD learning method is to minimize the mean square Bellman error (MSBE) [215],

\[
\frac{1}{2} \| V_x - \gamma \mathbb{P}^\pi V_x - \mathbb{R}^\pi \|_D^2
\]

where \( D := \text{diag}\{ \Pi(s), s \in S \} \in \mathbb{R}^{|S| \times |S|} \) is a diagonal matrix determined by the stationary distribution \( \Pi \). As discussed in [216], the solution to the fixed point problem \( V_x = \gamma \mathbb{P}^\pi V_x + \mathbb{R}^\pi \)
may not exist because the right-hand-side may not stay in the column space of the matrix $\Phi$. To address this challenge, GTD algorithm \cite{216} was proposed to minimize the mean square projected Bellman error (MSPBE),

\[
f(x) := \frac{1}{2} \| P_\Phi (V_x - \gamma P^\pi V_x - R^\pi) \|_D^2
\]

via stochastic-gradient-type updates, where $P_\Phi := \Phi (\Phi^\top D \Phi)^{-1} \Phi^\top D$ is a projection operator onto the column space of $\Phi$. Equivalently, MSPBE can be compactly written as,

\[
f(x) = \frac{1}{2} \| Ax - b \|_{C^{-1}}^2 \tag{4.5a}
\]

where $A$, $b$, and $C$ are obtained by taking expectations over the stationary distribution $\Pi$,

\[
A := E_{s \sim \Pi} [\phi(s)(\phi(s) - \gamma \phi(s'))^\top] \\
b := E_{s \sim \Pi} [R^\pi(s)\phi(s)] \\
C := E_{s \sim \Pi} [\phi(s)\phi(s)^\top]. \tag{4.5b}
\]

**Assumption 9** There exists a feature matrix $\Phi$ such that the matrices $A$ and $C$ are full rank and positive definite, respectively.

In \cite[page 300]{31}, it was shown that the full column rank matrix $\Phi$ yields a full rank $A$ and a positive definite $C$ and that the objective function $f$ in (4.5) has a unique minimizer. Nevertheless, when $A$, $b$, and $C$ are replaced by their sampled versions it is challenging to solve (4.5) because their nonlinear dependence on the underlying samples introduces bias in the objective function.
In what follows, we address the sampling challenge by reformulating (4.5) in terms of a saddle-point objective.

Since the global reward $R^{\pi}(s)$ in (4.3) is determined by the average of all local rewards $R^{\pi}_j(s)$, we can express the vector $b$ as $b = (1/N) \sum_j b_j$, where $b_j := \mathbb{E}_{s \sim \Pi}[R^{\pi}_j(s)\phi(s)]$. Thus, the problem of minimizing MSPBE (4.5) can be cast as

$$\min_{x \in \mathcal{X}} \frac{1}{N} \sum_{j=1}^N \frac{1}{2} \|Ax - b_j\|_{C^{-1}}^2$$ \hspace{1cm} (4.6)

where $f_j(x) := \frac{1}{2} \|Ax - b_j\|_{C^{-1}}^2$ is the local MSPBE for the agent $j$ and $\mathcal{X} \subset \mathbb{R}^d$ is a convex set that contains the unique minimizer of $f(x) = (1/N) \sum_j f_j(x)$. Hence, it is sufficient to restrict $\mathcal{X}$ to be a compact set containing the minimizer.

A decentralized stochastic optimization problem (4.6) with $N$ private stochastic objectives involves products and inverses of the expectations; cf. (4.5b). This unique feature of MSPBE is not encountered in typical distributed optimization settings \cite{173,82} and it makes the problem of obtaining an unbiased estimator of the objective function from a few state samples challenging.

Using Fenchel duality, we can express each local MSPBE in (4.6) as

$$f_j(x) = \max_{y_j \in \mathcal{Y}} \left( y_j^\top (Ax - b_j) - \frac{1}{2} y_j^\top C y_j \right)$$ \hspace{1cm} (4.7)

where $y_j$ is a dual variable and $\mathcal{Y} \subset \mathbb{R}^d$ is a convex compact set such that $C^{-1}(Ax - b_j) \in \mathcal{Y}$ for all $x \in \mathcal{X}$. Since $C$ is a positive definite matrix and $\mathcal{X}$ is a compact set, such $\mathcal{Y}$ exists. In fact, one could take a ball centered at the origin with a radius greater than $(1/\lambda_{\min}(C)) \sup \|Ax - b_j\|$, where the supremum is taken over $x \in \mathcal{X}$ and $j \in \{1, \ldots, N\}$. Thus, we can reformulate (4.6) as
a decentralized stochastic saddle point problem (4.2) with objective
\[ \psi_j(x, y_j) = y_j^\top (Ax - b_j) - \frac{1}{2} y_j^\top C y_j \]
and compact convex domain sets \( \mathcal{X} \) and \( \mathcal{Y} \). By replacing expectations in the expressions for \( A, b_j, \) and \( C \) with their samples that arise from the stationary distribution \( \Pi \), we obtain an unbiased estimate of the saddle point objective \( \psi_j \). We also note that each agent \( j \) indeed takes a local MSPBE as its local objective function
\[ f_j(x) = \frac{1}{2} \| Ax - b_j \|_{C^{-1}}^2. \]

Since the stationary distribution is not known, it is not possible to directly estimate \( A, b, \) and \( C \). However, as we explain next, the policy evaluation problem allows correlated sampling according to a Markov process.

### 4.2.4 Standard stochastic primal-dual algorithm

When i.i.d. samples from the stationary distribution \( \Pi \) are available and a centralized controller exists, the stochastic approximation method can be used to compute the solution to (4.2) with a convergence rate \( O(1/\sqrt{T}) \) in terms of the primal-dual gap [175]. The stochastic primal-dual algorithm generates two pairs of vectors \((x'(t), y'(t))\) and \((x(t), y(t))\) that are contained in \( \mathcal{X} \times \mathcal{Y}^N \), where \( t \) is a positive integer, \( \mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d \) are convex projection sets. It is sufficient to take bounded sets \( \mathcal{X} \) and \( \mathcal{Y} \) containing the finite solution to (4.2). At iteration \( t \), the primal-dual updates are given by

\[
\begin{align*}
x'(t + 1) &= x'(t) - \eta(t) G_x(x(t), y(t); \xi_t) \\
x(t + 1) &= \mathcal{P}_X(x'(t + 1)) \\
y'(t + 1) &= y'(t) + \eta(t) G_y(x(t), y(t); \xi_t) \\
y(t + 1) &= \mathcal{P}_{\mathcal{Y}^N}(y'(t + 1))
\end{align*}
\]
where $\eta(t)$ is a non-increasing sequence of stepsizes, operators $P_X(x') := \arg\min_{x \in X} \|x - x'\|$ and $P_{Y^N}(y') := \arg\min_{y \in Y^N} \|y5 - y'\|$ are Euclidean projections onto $X$ and $Y^N$, and the sampled gradients are given by

\[
G_x(x(t), y(t); \xi_t) = \nabla_x \Psi(x(t), y(t); \xi_t) \\
G_y(x(t), y(t); \xi_t) = \nabla_y \Psi(x(t), y(t); \xi_t)
\]

In our multi-agent MDP setup, however, each agent receives samples $\xi_t$ from a Markov process whose state distribution at time $t$ is $P_t$, where $P_t$ converges to the unknown distribution $\Pi$ with a geometric rate. Thus, i.i.d. samples from the stationary distribution $\Pi$ are not available. Since i.i.d. sampling-based convergence guarantees may not hold for correlated samples [96], it is important to examine the ergodic stochastic optimization scenario in which samples are taken from a stochastic process [83]: a recent application for the centralized GTD can be found in [238]. In particular, we are interested in designing and analyzing distributed algorithms for stochastic saddle point problem (4.2) in the ergodic setting.

### 4.3 Algorithm and performance

We now present the main results of the chapter: a fast algorithm for the multi-agent learning. We propose a distributed stochastic primal-dual algorithm in Section 4.3.1, introduce underlying assumptions in Section 4.3.2 and establish a finite-time performance bound in Section 4.3.3.
Algorithm 9 Distributed Homotopy Primal-Dual Algorithm

Initialization: \( x_{j,1}(1) = x'_{j,1}(1) = 0, y_{j,1}(1) = y'_{j,1}(1) = 0 \) for all \( j \in \mathcal{V}; T_1, \eta_1, K \)

For \( k = 1 \) to \( K \) do

1. For \( t = 1 \) to \( T_k - 1 \) do
   - Primal update,
     \[
     x'_{j,k}(t+1) = \sum_{i=1}^{N} W_{ij} x'_{i,k}(t) - \eta_k G_{j,x}(x_{j,k}(t), y_{j,k}(t); \xi_{k,t}) \\
     x_{j,k}(t+1) = \mathcal{P}_X(x'_{j,k}(t+1)).
     \]
   - Dual update,
     \[
     y'_{j,k}(t+1) = y_{j,k}(t) + \eta_k G_{j,y}(x_{j,k}(t), y_{j,k}(t); \xi_{k,t}) \\
     y_{j,k}(t+1) = \mathcal{P}_Y(y'_{j,k}(t+1)).
     \]

2. Restart initialization,
   \[
   (x_{j,k+1}(1), y_{j,k+1}(1)) = (\hat{x}_{j,k}, \hat{y}_{j,k}) \quad \text{(see (4.9))} \\
   (x'_{j,k+1}(1), y'_{j,k+1}(1)) = (x_{j,k+1}(1), y_{j,k+1}(1)).
   \]

3. Update stepsize, horizon:
   \[
   \eta_{k+1} = \frac{1}{2} \eta_k, \quad T_{k+1} = 2T_k.
   \]

end for

Output: \((\hat{x}_{j,K}, \hat{y}_{j,K})\) for all \( j \in \mathcal{V} \)

4.3.1 Distributed homotopy primal-dual algorithm

In this section, we extend stochastic primal-dual algorithm (4.8) to the multi-agent learning setting. To solve the stochastic saddle point program (4.2) in a distributed manner, the algorithm operates \( 2N \) primal-dual pairs of vectors \( z_{j,k}(t) := (x_{j,k}(t), y_{j,k}(t)) \) and \( z'_{j,k}(t) := (x'_{j,k}(t), y'_{j,k}(t)) \),
which belong to projection set $X \times Y$. In the $k$th iteration round at time $t$, the $j$th agent computes local gradient using the private objective $\Psi_j(z_{j,k}(t); \xi_{k,t}),$

$$G_j(z_{j,k}(t); \xi_{k,t}) := \begin{bmatrix} G_{j,x}(z_{j,k}(t); \xi_{k,t}) \\ G_{j,y}(z_{j,k}(t); \xi_{k,t}) \end{bmatrix}$$

and receives the vectors $\{x'_i(t), i \in N_j\}$ from its neighbors $N_j$. Here, $G_{j,x}(z_{j,k}(t); \xi_{k,t})$ and $G_{j,y}(z_{j,k}(t); \xi_{k,t})$ are gradients of $\Psi_j(z_{j,k}(t); \xi_{k,t})$ with respect to $x_{j,k}(t)$ and $y_{j,k}(t)$, respectively.

The primal iterate $x_{j,k}(t)$ is updated using a convex combination of the vectors $\{x'_i(t), i \in N_j\}$ and the dual iterate $y_{j,k}(t)$ is modified using the mirror descent update. We model the convex combination as a mixing process over the graph $G$ and assume that the mixing matrix $W$ is a doubly stochastic,

$$\sum_{i=1}^{N} W_{ij} = \sum_{i \in N_j} W_{ij} = 1, \text{ for all } j \in \mathcal{V}$$
$$\sum_{j=1}^{N} W_{ij} = \sum_{j \in N_i} W_{ij} = 1, \text{ for all } i \in \mathcal{V}$$

where $W_{ij} > 0$ for $(i,j) \in \mathcal{E}$. For a given learning rate $\eta_k$, each agent updates primal and dual variables according to Algorithm 9 where $\mathcal{P}_X(\cdot)$ and $\mathcal{P}_Y(\cdot)$ are Euclidean projections onto bounded sets $X$ and $Y$. In practice, these projections are easily computable, $\mathcal{P}_{\|\theta\| \leq r}(\theta') = r\theta'/\|\theta'\|$ when $\|\theta'\| > r$ and is simply $\theta'$ otherwise.

The iteration counters $k$ and $t$ are used in our Distributed Homotopy Primal-Dual (DHPD) Algorithm, i.e., Algorithm 9. The homotopy approach varies certain parameter for multiple rounds, where each round takes an estimated solution from the previous round as a starting point. We use the learning rate as a homotopy parameter in our algorithm. At the initial round $k = 1$,
problem (4.2) is solved with a large learning rate $\eta_1$ and, in subsequent iterations, the learning rate is gradually decreased until a desired error tolerance is reached.

For a fixed learning rate $\eta_k$, we employ the distributed stochastic primal-dual method to solve (4.2) and obtain an approximate solution that is given by a time-running average of primal-dual pairs,

$$
\hat{x}_{j,k} := \frac{1}{T_k} \sum_{t=1}^{T_k} x_{j,k}(t), \quad \hat{y}_{j,k} := \frac{1}{T_k} \sum_{t=1}^{T_k} y_{j,k}(t).
$$

(4.9)

These are used as initial points for the next learning rate $\eta_{k+1}$. At round $k$, each agent $j$ performs primal-dual updates with $T_k$ iterations, indexed by time $t$. At next round $k+1$, we initialize primal and dual iterations using the previous approximate solutions $\hat{x}_{j,k}$ and $\hat{y}_{j,k}$, reduce the learning rate by half, $\eta_{k+1} = \eta_k / 2$, and double the number of the inner-loop iterations, $T_{k+1} = 2T_k$. The number of inner iterations in the $k$th round is $T_k$ and the number of total rounds is $K$.

The homotopy approach not only provides outstanding practical performance but it also facilitates an effective iteration complexity analysis [248]. In particular, for stochastic strongly convex programs, the rate faster than $O(1/\sqrt{T})$ was established in [224, 258]; other fast rate results can be found in [252, 243]. To the best of our knowledge, we are the first to show that the homotopy approach can be used to solve distributed stochastic saddle-point problems with convergence rate better than $O(1/\sqrt{T})$.

In Section 4.3.3 we use the primal optimality gap,

$$
\text{err}(\hat{x}_{i,k}) := \frac{1}{N} \sum_{j=1}^{N} \left( f_j(\hat{x}_{i,k}) - f_j(x^*) \right)
$$

(4.10)
to quantify the distance of the running local average, \( \hat{x}_{i,k} := \frac{1}{T_k} \sum_{t=1}^{T_k} x_{i,k}(t) \), for the \( i \)th agent from the optimal solution \( x^* \). The primal optimality gap measures performance of each agent in terms of MSPBE which is described by the global objective function (4.5), or equivalently by (4.6).

**Remark 5 (Multi-agent policy evaluation)** For the multi-agent policy evaluation problem, we take \( \psi_j(x, y_j) = y_j^\top (Ax - b_j) - \frac{1}{2} y_j^\top Cy_j \). Since \( \psi_j \) depends linearly on \( A, b_j, \) and \( C \), by replacing expectations in (4.5b) with the corresponding samples we obtain unbiased gradients,

\[
G_{j,x}(z_{j,k}(t); \xi_{k,t}) = (\phi(s) - \gamma \phi(s')) \phi(s)^\top y_{j,k}(t)
\]
\[
G_{j,y}(z_{j,k}(t); \xi_{k,t}) = \phi(s)(\phi(s) - \gamma \phi(s')) \phi(s)^\top x_{j,k}(t) - R^\pi_j(s) \phi(s) - \phi(s) \phi(s)^\top y_{j,k}(t).
\]

where \( s, s' \) are two consecutive states that evolve according to the underlying Markov process \( \xi_{k,t} \) indexed by time \( t \) and the iteration round \( k \). In each iteration, Algorithm 9 requires \( O(dN^2) \) operations where \( d \) is the problem dimension (or the feature dimension in linear approximation) and \( N \) is the total number of agents. For a single-agent problem, \( O(d) \) operations are required which is consistent with GTD algorithm [216].

### 4.3.2 Assumptions

We now formally state assumptions required to establish our main result in Theorem 24 that quantifies finite-time performance of Algorithm 9 for stochastic primal-dual optimization problem (4.2).

**Assumption 10 (Convex compact projection sets)** The projection sets \( X \) and \( Y \) contain the origin in \( \mathbb{R}^d \) and the finite solution to (4.2), and they are convex and compact with radius \( r > 0 \), i.e.,

\[
\sup_{x \in X, y \in Y} \|(x, y)\|^2 \leq r^2.
\]
Assumption 11 (Convexity and concavity) The function $\psi_j(x, y_j)$ in (4.2) is convex in $x$ for any fixed $y_j \in \mathcal{Y}$, and is strongly concave in $y_j$ for any fixed $x \in \mathcal{X}$, i.e., for any $x, x' \in \mathcal{X}$ and $y_j, y'_j \in \mathcal{Y}$, there exists $L_y > 0$ such that

$$
\psi_j(x, y_j) \geq \psi_j(x', y_j) + \langle \nabla_x \psi_j(x', y_j), x - x' \rangle
$$

$$
\psi_j(x, y_j) \leq \psi_j(x, y'_j) - \langle \nabla_y \psi_j(x, y'_j), y_j - y'_j \rangle - \frac{L_y}{2} \|y_j - y'_j\|^2.
$$

Moreover, $f_j(x) := \max_{y_j \in \mathcal{Y}} \psi_j(x, y_j)$ is strongly convex, i.e., for any $x, x' \in \mathcal{X}$, there exists $L_x > 0$ such that

$$
f_j(x) \geq f_j(x') + \langle \nabla f_j(x'), x - x' \rangle + \frac{L_x}{2} \|x - x'\|^2.
$$

Assumption 12 (Bounded gradient) For any $t$ and $k$, there is a positive constant $c$ such that the sample gradient $G_j(x, y_j; \xi_{k,t})$ satisfies,

$$
\|G_j(x, y_j; \xi_{k,t})\| \leq c
$$

for all $x \in X, y_j \in Y$ with probability one.

Remark 6 Jensen’s inequality can be combined with Assumption 12 to show that the population gradient is also bounded, i.e., $\|g_j(x, y_j)\| \leq c$, for all $x \in X$ and $y_j \in Y$, where

$$
g_j(x, y_j) := \begin{bmatrix} g_{j,x}(x, y_j) \\ g_{j,y}(x, y_j) \end{bmatrix} = \begin{bmatrix} \nabla_x \psi_j(x, y_j) \\ \nabla_y \psi_j(x, y_j) \end{bmatrix}.
$$
**Assumption 13 (Lipschitz gradient)** For any \( t \) and \( k \), there exists a positive constant \( L \) such that for any \( x, x' \in X \) and \( y_j, y'_j \in Y \), we have

\[
\| G_j(x, y_j; \xi_{k,t}) - G_j(x', y_j; \xi_{k,t}) \| \leq L \| x - x' \| \\
\| G_j(x, y_j; \xi_{k,t}) - G_j(x, y'_j; \xi_{k,t}) \| \leq L \| y_j - y'_j \|
\]

with probability one.

We also recall some important concepts from probability theory. The total variation distance between distributions \( P \) and \( Q \) on a set \( \Xi \subset \mathbb{R}^{|S|} \) is given by

\[
d_{tv}(P, Q) := \int_{\Xi} |p(\xi) - q(\xi)| \, d\mu(\xi) = 2 \sup_{E \subset \Xi} |P(E) - Q(E)|
\]

where distributions \( P \) and \( Q \) (with densities \( p \) and \( q \)) are continuous with respect to the Lebesgue measure \( \mu \), and the supremum is taken over all measurable subsets of \( \Xi \).

We use the notion of mixing time to evaluate the convergence speed of a sequence of probability measures generated by a Markovian process to its (unique) stationary distribution \( \Pi \), whose density \( \pi \) is assumed to exist. Let \( \mathcal{F}_{k,t} \) be the \( \sigma \)-field generated by the first \( t \) samples at round \( k \), \( \{\xi_{k,1}, \ldots, \xi_{k,t}\} \), drawn from \( \{P_{k,1}, \ldots, P_{k,t}\} \), where \( P_{k,t} \) is the probability measure of the Markovian process at time \( t \) and round \( k \). Let \( P_{k,t}^{[s]} \) be the distribution of \( \xi_{k,t} \) conditioned on \( \mathcal{F}_{k,s} \) (i.e., given samples up to time slot \( s \), \( \{\xi_{k,1}, \ldots, \xi_{k,s}\} \)) at round \( k \), whose density \( p_{k,t}^{[s]} \) also exists. The mixing time for a Markovian process is defined as follows [83].
Definition 1  The total variation mixing time \( \tau_{tv}(P_k^{[s]}, \varepsilon) \) of the Markovian process conditioned on the \( \sigma \)-field of the initial \( s \) samples \( \mathcal{F}_{k,s} = \sigma(\xi_{k,1}, \ldots, \xi_{k,s}) \) is the smallest positive integer \( t \) such that

\[
d_{tv}(P_{k,s+t}^{[s]}, \Pi) \leq \varepsilon,
\]

where

\[
\tau_{tv}(P_k^{[s]}, \varepsilon) = \inf \left\{ t - s \middle| t \in \mathbb{N}, \int_{\Xi} |p_{k,t}^{[s]}(\xi) - \pi(\xi)| \, d\mu(\xi) \leq \varepsilon \right\}.
\]

The mixing time \( \tau_{tv}(P_k^{[s]}, \varepsilon) \) measures the number of additional steps required until the distribution of \( \xi_{k,t} \) is within \( \varepsilon \) neighborhood of the stationary distribution \( \Pi \) given the initial \( s \) samples, \( \{\xi_{k,1}, \ldots, \xi_{k,s}\} \).

Assumption 14  The underlying Markov chain is irreducible and aperiodic, i.e., there exists \( \Gamma > 0 \) and \( \rho \in (0, 1) \) such that \( \mathbb{E}[d_{tv}(P_{k,t}^{[t]}, \Pi)] \leq \Gamma \rho^\tau \) for all \( \tau \in \mathbb{N} \) and all \( k \).

Furthermore, we have

\[
\tau_{tv}(P_k^{[s]}, \varepsilon) \geq \left\lceil \frac{\log (\Gamma/\varepsilon)}{\log \rho} \right\rceil + 1, \quad \text{for all } k, s \in \mathbb{N}
\]  \hspace{1cm} (4.11)

where \( \lceil \cdot \rceil \) is the ceiling function and \( \varepsilon \leq \Gamma \) specifies the distance to the stationarity; also see [132, Theorem 4.9].

4.3.3 Finite-time performance bound

For stochastic saddle point problem (4.2), we establish a finite-time error bound in terms of the average primal optimality gap in Theorem [24] where the total number of iterations in Algorithm [9] is given by

\[
T := \sum_{k=1}^{K} T_k = (2^K - 1)T_1.
\]
Theorem 24 Let Assumptions 10–14 hold. Then, for any $\eta_1 \geq 1/(4/L_y + 2/L_x)$ and any $T_1$ and $K$ that satisfy

$$T_1 \geq \tau := \left\lceil \frac{\log (\Gamma T)}{\log \rho} \right\rceil + 1$$

the output $\hat{x}_{j,K}$ of Algorithm 9 provides the solution to problem (4.2) with the following upper bound

$$c(rL + c) \left( \frac{C_1 \log^2(T\sqrt{N})}{1 - \sigma_2(W)} + C_2(1 + T_1) \right)$$

on $(1/N) \sum_{j=1}^{N} \mathbb{E}[err(\hat{x}_{j,K})]$, where the primal optimality gap $err(\hat{x}_{j,K})$ is defined in (4.10), $r$ is the bound on feasible sets $X$ and $Y$ in Assumption 10, $c$ is the bound on sample gradients in Assumption 12, $C_1$ and $C_2$ are constants independent of $T$, $\sigma_2(W)$ is the second largest eigenvalue of $W$, and $N$ is the total number of agents.

Remark 7 (Multi-agent policy evaluation) Assumption 9 guarantees that $A$ is full rank and that $C$ is positive definite. For $\psi_j(x, y_j) = y_j^\top (Ax - b_j) - \frac{1}{2}y_j^\top Cy_j$, since all features and rewards are bounded, Assumptions 11–13 hold with

$$L_x = \sigma_{\text{max}}(A^\top A)/\sigma_{\text{min}}(C), \quad L_y = \sigma_{\text{min}}(C)$$

$$c \geq (2\beta_1 + \beta_2)r + \beta_0$$

$$L \geq \max(\sqrt{\beta_1^2 + \beta_2^2}, \beta_1)$$
where $\beta_0, \beta_1$ and $\beta_2$ provide upper bounds to $\| R^\pi_j(s) \phi(s) \| \leq \beta_0$, $\| \phi(s)(\phi(s) - \gamma \phi(s'))^\top \| \leq \beta_1$, and $\| \phi(s) \phi(s)^\top \| \leq \beta_2$. The unique minimizer of (4.5) and the expression (4.7) that results from Fenchel duality validate Assumption 10 with

$$r \geq \frac{2\beta_0}{\sigma_{\text{min}}(C)} \left( \frac{\beta_0^2 \sigma_{\text{max}}(C)}{\sigma_{\text{min}}(A^\top A) \sigma_{\text{min}}(C)} + 1 \right).$$

In practice, when some prior knowledge about the model is available, e.g., when generative models or simulators can be utilized, samples from a near stationary state distribution under a given policy can be used to estimate these parameters.

**Remark 8 (Optimal performance bound and selection of parameters)** We can find $T_1$ and $K$ such that condition (4.12) holds, as long as $T > \tau = \lceil \log (\Gamma T)/|\log \rho| \rceil$. In particular, choosing $T_1 = \tau$ and $K = \log(1 + T/\tau)$ gives the desired $O(\log^2(T \sqrt{N})/T)$ scaling of finite-time performance bound (4.13). In general, to satisfy condition (4.12), we can choose $T_1$ and $K$ such that $T_1 \geq \lceil (K + \log(\Gamma T_1))/|\log \rho| \rceil + 1$. Hence, performance bound (4.13) scales as $O((K^2 + \log^2 T_1)/T)$.

Time-running average (4.9) is used as the output of our algorithm and when the algorithm is terminated in an inner loop $K$, the time-running average from previous inner loop can be used as an output and our performance bound holds for $K - 1$.

**Remark 9 (Mixing time)** When $\epsilon = 1/T$, $\tau = \lceil \log (\Gamma T)/|\log \rho| \rceil$ provides a lower bound on the mixing time (cf. (4.11)), where $\Gamma > 0$ and $\rho \in (0, 1)$ are given in Assumption 14. Thus, performance bound (4.13) in Theorem 24 depends on how fast the process $P^{|s|}_k$ reaches $1/T$ mixing via $\tau$. In particular, when samples are independent and identically distributed, 0-mixing is reached in one step, i.e., we have $\tau = 1$ and $\epsilon = 0$. Hence, by setting $T_1 = 1$, performance bound (4.13) simplifies to $O(\log^2(T)/T)$.
Remark 10 (Network size and topology) In finite-time performance bound (4.13), the dependence on the network size $N$ and the spectral gap $1 - \sigma_2(W)$ of the mixing matrix $W$ is quantified by $\log^2(T\sqrt{N})/(1 - \sigma_2(W))$. We note that $W$ can be expressed using the Laplacian $L$ of the underlying graph, $W = I - D^{1/2}LD^{1/2}/(\delta_{\text{max}} + 1)$, where $D := \text{diag}(\delta_1, \ldots, \delta_N)$, $\delta_i$ is the degree of node $i$, and $\delta_{\text{max}} := \max_i \delta_i$. The algebraic connectivity of the network $\lambda_{N-1}(L)$, i.e., the second smallest eigenvalue of $L$, can be used to bound $\sigma_2(W)$. In particular, for a ring with $N$ nodes, we have $\sigma_2(W) = \Theta(1/N^2)$; for other topologies, see [82, Section 6].

4.4 Finite-time performance analysis

In this section, we study finite-time performance of the distributed homotopy primal-dual algorithm described in Algorithm 9. We define auxiliary quantities in Section 4.4.1, present useful lemmas in Section 4.4.2, and provide the proof of Theorem 24 in Section 4.4.3.

4.4.1 Setting up the analysis

We first introduce three types of averages that are used to describe sequences generated by the primal update of Algorithm 9. The average value of $x_{j,k}(t)$ over all agents is denoted by $\bar{x}_k(t) := \frac{1}{N} \sum_{j=1}^{N} x_{j,k}(t)$, the time-running average of $x_{j,k}(t)$ is $\hat{x}_{j,k} := \frac{1}{T_k} \sum_{t=1}^{T_k} x_{j,k}(t)$, the averaged time-running average of $x_{j,k}(t)$ is given by $\bar{x}_k := \frac{1}{N} \sum_{j=1}^{N} \hat{x}_{j,k} = \frac{1}{T_k} \sum_{t=1}^{T_k} \bar{x}_k(t)$, and two auxiliary averaged sequences are, respectively, given by $\bar{x}_k'(t) := \frac{1}{N} \sum_{j=1}^{N} x_{j,k}'(t)$ and $x_k(t) := \mathcal{P}_X(\bar{x}_k'(t))$.

Since the mixing matrix $W$ is doubly stochastic, the primal update $\bar{x}_k'(t)$ has a simple ‘centralized’ form,

\[ x_k'(t+1) = \bar{x}_k'(t) - \frac{\eta_k}{N} \sum_{j=1}^{N} G_{j,x}(z_{j,k}(t); \xi_{k,t}). \] (4.14)
We now utilize the approach similar to the network averaging analysis in [82] to quantify how well the agent $j$ estimates the network average at round $k$.

**Lemma 25** Let Assumption 12 hold, let $W$ be a doubly stochastic mixing matrix over graph $G$, let $\sigma_2(W)$ denote its second largest singular value, and let a sequence $x_{j,k}(t)$ be generated by Algorithm 9 for agent $j$ at round $k$. Then,

$$\frac{1}{T_k} \sum_{t=1}^{T_k} \mathbb{E} \| x_{j,k}(t) - \bar{x}_k(t) \| \leq 2\Delta_k$$

where $\Delta_k$ is given by

$$\Delta_k := \frac{2\eta k c \log(\sqrt{NT_k})}{1 - \sigma_2(W)} + \frac{4c}{T_k} \left( \frac{\log(\sqrt{NT_k})}{1 - \sigma_2(W)} + 1 \right) \sum_{t=1}^{k} \eta_t T_t + 2\eta_k c$$

**Proof.** See Appendix C.1.

The dual update $y'_{j,k}(t+1)$ behaves in a similar way as (4.14). We utilize the following classical online gradient descent to analyze the behavior of $\bar{x}'_k(t)$ and $y_{j,k}'(t)$ under projections $P_X(\cdot)$ and $P_Y(\cdot)$.

**Lemma 26 ([285])** Let $U$ be a convex closed subset of $\mathbb{R}^d$, let $\{g(t)\}_{t=1}^T$ be an arbitrary sequence in $\mathbb{R}^d$, and let sequences $w(t)$ and $u(t)$ be generated by the projection, $w(t+1) = w(t) - \eta g(t)$ and $u(t+1) = P_U(w(t+1))$, where $u(1) \in U$ is the initial point and $\eta > 0$ is the learning rate. Then, for any fixed $u^* \in U$,

$$\sum_{t=1}^{T} \langle g(t), u(t) - u^* \rangle \leq \frac{\| u(1) - u^* \|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \| g(t) \|^2.$$
If \( \hat{y}_{j,k}^* := \arg\max_{y_j \in Y} \psi_j(\hat{x}_{i,k}, y_j) \), using Fenchel dual (4.7), we have \( f_j(\hat{x}_{i,k}) = \psi_j(\hat{x}_{i,k}, \hat{y}_{j,k}^*) \) and \( f_j(x^*) = \psi_j(x^*, y_j^*) \). This allows us to express primal optimality gap (4.10) in terms of a primal-dual objective,

\[
\text{err}(\hat{x}_{i,k}) = \frac{1}{N} \sum_{j=1}^{N} \left( \psi_j(\hat{x}_{i,k}, y_{j,k}^*) - \psi_j(x^*, y_j^*) \right). 
\tag{4.15}
\]

Let \( \hat{x}_k^* := \arg\min_{x \in X} \frac{1}{N} \sum_{j=1}^{N} \psi_j(x, \hat{y}_{j,k}) \). To analyze optimality gap (4.15), we introduce a surrogate gap,

\[
\text{err}'(\hat{x}_{i,k}, \hat{y}_k) := \frac{1}{N} \sum_{j=1}^{N} \left( \psi_j(\hat{x}_{i,k}, \hat{y}_{j,k}^*) - \psi_j(x^*, \hat{y}_{j,k}) \right) 
\tag{4.16}
\]

as well as average primal optimality (4.10) and surrogate (4.16) gaps,

\[
\overline{\text{err}}_k := \frac{1}{N} \sum_{j=1}^{N} \mathbb{E} [ \text{err}(\hat{x}_{j,k}) ], \\
\overline{\text{err}}'_k := \frac{1}{N} \sum_{j=1}^{N} \mathbb{E} [ \text{err}'(\hat{x}_{j,k}, \hat{y}_k) ].
\]

In Lemma 27, we establish relation between the surrogate gap \( \text{err}'(\hat{x}_{i,k}, \hat{y}_k) \) and the primal optimality gap \( \text{err}(\hat{x}_{i,k}) \).

**Lemma 27** Let \( \hat{x}_{i,k} \) and \( \hat{y}_k \) be generated by Algorithm \( 6 \) for agent \( i \) at round \( k \). Then,

\[
0 \leq \text{err}(\hat{x}_{i,k}) \leq \text{err}'(\hat{x}_{i,k}, \hat{y}_k) \text{ and } 0 \leq \overline{\text{err}}_k \leq \overline{\text{err}}'_k.
\]

**Proof.** See Appendix C.3. \( \square \)
Lemma 28 Let Assumption 11 hold and let \( \hat{x}_{i,k} \) and \( \hat{y}_k \) be generated by Algorithm 9 for agent \( i \) at round \( k \). Then,

\[
\text{err}(\hat{x}_{i,k}) \geq \frac{L_x}{2} \|x^* - \hat{x}_{i,k}\|^2
\]

(4.17a)

\[
\text{err}'(\hat{x}_{i,k}, \hat{y}_k) \geq \frac{L_y}{2N} \sum_{j=1}^{N} \|y_j^* - \hat{y}_{j,k}\|^2
\]

(4.17b)

\[
\text{err}'(\hat{x}_{i,k}, \hat{y}_k) \geq \frac{L_y}{2N} \sum_{j=1}^{N} \|y_j^* - \hat{y}_{j,k}^*\|^2.
\]

(4.17c)

Proof. See Appendix C.4.

We are now ready to provide an overview of our remaining analysis. In Lemma 29, we use a sum of network errors (NET) and local primal-dual gaps (PDG) to bound surrogate gap (4.16). In Lemma 30, we provide a bound on local primal-dual gaps by a sum of local dual gaps and a term that depends on mixing time. We combine Lemma 29 and Lemma 30 and apply the restarting strategy to get a recursion on the surrogate gap. Finally, in Section 4.4.3, we complete the proof by utilizing induction on the round \( k \).

4.4.2 Useful lemmas

We utilize convexity and concavity of \( \psi_j \) with respect to primal and dual variables to decompose surrogate gap (4.16) into parts that quantify the influence of network errors (NET) and local primal-dual gaps (PDG), respectively.

Lemma 29 Let Assumptions 11 and 12 hold and let \( \hat{x}_{i,k} \) and \( \hat{y}_k \) be generated by Algorithm 9 for agent \( i \) at round \( k \). Then,

\[
\text{err}'(\hat{x}_{i,k}, \hat{y}_k) \leq \text{NET} + \text{PDG}
\]
where

\[
\begin{align*}
\text{NET} &= \frac{c}{T_k} \sum_{t=1}^{T_k} \left( \|x_{i,k}(t) - \bar{x}_k(t)\| + \frac{1}{N} \sum_{j=1}^{N} \|x_{j,k}(t) - \bar{x}_k(t)\| \right) \\
\text{PDG} &= \frac{1}{NT_k} \sum_{t=1}^{T_k} \sum_{j=1}^{N} \left( \psi_j(x_{j,k}(t), \hat{y}_{j,k}^*) - \psi_j(x^*, y_{j,k}(t)) \right).
\end{align*}
\]

**Proof.** Applying the mean value theorem and boundedness of the gradient of \( \psi_j(x, \hat{y}_{j,k}^*) \) with respect to \( x \), we have

\[
\begin{align*}
\psi_j(\hat{x}_{i,k}, \hat{y}_{j,k}^*) - \psi_j(\bar{x}_k, \hat{y}_{j,k}^*) &\leq c \|\hat{x}_{i,k} - \bar{x}_k\| \\
&\leq \frac{c}{T_k} \sum_{t=1}^{T_k} \|x_{i,k}(t) - \bar{x}_k(t)\|
\end{align*}
\]

where the second inequality follows from the Jensen’s inequality. Then, breaking \( \text{err}'(\hat{x}_{i,k}, \hat{y}_k) \) in (4.16) by adding and subtracting \( \frac{1}{N} \sum_{j=1}^{N} \psi_j(\bar{x}_k, \hat{y}_{j,k}^*) \), we bound \( \text{err}'(\hat{x}_{i,k}, \hat{y}_k) \) by

\[
\text{err}'(\hat{x}_{i,k}, \hat{y}_k) \leq \frac{c}{T_k} \sum_{t=1}^{T_k} \|x_{i,k}(t) - \bar{x}_k(t)\| + \frac{1}{N} \sum_{j=1}^{N} \left( \psi_j(\bar{x}_k, \hat{y}_{j,k}^*) - \psi_j(x^*, y_{j,k}(t)) \right).
\]

Next, we find a simple bound for the second sum. We recall that \( \bar{x}_k := \frac{1}{T_k} \sum_{t=1}^{T_k} \bar{x}_k(t) \) and \( \hat{y}_{j,k} := \frac{1}{T_k} \sum_{t=1}^{T_k} y_{j,k}(t) \) and apply the Jensen’s inequality twice to obtain,

\[
\frac{1}{NT_k} \sum_{t=1}^{T_k} \sum_{j=1}^{N} \left( \psi_j(\bar{x}_k(t), \hat{y}_{j,k}^*) - \psi_j(x^*, y_{j,k}(t)) \right). \tag{4.18}
\]

Similarly, we have \( \psi_j(\bar{x}_k(t), \hat{y}_{j,k}^*) - \psi_j(x_{j,k}(t), \hat{y}_{j,k}^*) \leq c \|\bar{x}_k(t) - x_{j,k}(t)\|. \) Breaking (4.18) by adding and subtracting \( \frac{1}{NT_k} \sum_{t=1}^{T_k} \sum_{j=1}^{N} \psi_j(x_{j,k}(t), \hat{y}_{j,k}^*) \), we further bound (4.18) by

\[
\frac{c}{NT_k} \sum_{t=1}^{T_k} \sum_{j=1}^{N} \|x_{j,k}(t) - \bar{x}_k(t)\| + \text{PDG}.
\]
The proof is completed by combining all above bounds. □

Lemma 29 establishes a bound for the surrogate gap \( \text{err}'(\hat{x}_{i,k}, \hat{y}_k) \) for agent \( i \) at round \( k \). The terms in NET describe the accumulated network error that measures the deviation of each agent’s estimate from the average. On the other hand, PDG determines the average of primal-dual gaps incurred by local agents that are commonly used in the analysis of primal-dual algorithms [175, 174].

Next, we utilize the Markov mixing property to control the average of local primal-dual gaps PDG. First, we break the difference \( \psi_j(x_{j,k}(t), \hat{y}_{j,k}(t)) - \psi_j(x_{j,k}(t), y_{j,k}(t)) \) into a sum of \( \psi_j(x_{j,k}(t), \hat{y}_{j,k}(t)) - \psi_j(x_{j,k}(t), y_{j,k}(t)) \) and \( \psi_j(x_{j,k}(t), y_{j,k}(t)) - \psi_j(x^*, y_{j,k}(t)) \). We now can utilize convexity and concavity of \( \psi_j(x, y_j) \) to deal with these terms separately. Dividing the sum indexed by \( t \) into two intervals, \( 1 \leq t \leq T_k - \tau \) and \( T_k - \tau + 1 \leq t \leq T_k \), the PDG term can be bounded by

\[
\text{PDG} \leq \text{PDG}^+ + \text{PDG}^-,
\]

where

\[
\text{PDG}^+ = \frac{1}{NT_k} \sum_{t=1}^{T_k-\tau} \sum_{j=1}^{N} \left( \langle g_{j,y}(z_{j,k}(t)), \hat{y}_{j,k} - y_{j,k}(t) \rangle + \langle g_{j,x}(z_{j,k}(t)), x_{j,k}(t) - x^* \rangle \right)
\]

\[
\text{PDG}^- = \frac{1}{NT_k} \sum_{t=T_k-\tau+1}^{T_k} \sum_{j=1}^{N} \left( \langle g_{j,y}(z_{j,k}(t)), \hat{y}_{j,k} - y_{j,k}(t) \rangle + \langle g_{j,x}(z_{j,k}(t)), x_{j,k}(t) - x^* \rangle \right).
\]

Here, \( \tau \) is the mixing time of the ergodic sequence \( \xi_{k,1}, \ldots, \xi_{k,t} \) at round \( k \). The intuition behind this is that, given the initial \( t - \tau \) samples \( \xi_{k,1}, \ldots, \xi_{k,t-\tau} \), the sample \( \xi_{k,t} \) is almost a sample that arises from the stationary distribution \( \Pi \). With this in mind, we next show that an appropriate breakdown of the term PDG\(^+\) enables applications of the martingale concentration from Lemma 55 and mixing time property (4.11), thereby producing a gradient-free bound on primal-dual gaps.
Lemma 30  Let Assumptions 10–13 hold. For $T_k$ that satisfies $T_k \geq 1 + \lceil \log(\Gamma T_k)/|\log \rho| \rceil = \tau$, we have

$$
\mathbb{E}[\text{PDG}] \leq \frac{c}{N} \left( \frac{8}{\sqrt{T_k - \tau}} + \frac{1}{T_k} \right) \sum_{j=1}^{N} \mathbb{E}\left[ \| \hat{y}_{j,k}^* - y_j^* \| \right] + \mathbb{E}[\text{MIX}]
$$

where

$$
\text{MIX} = \frac{2rL + \sqrt{2c}}{NT_k} \sum_{t=\tau+1}^{T_k} \sum_{j=1}^{N} \| z_{j,k}(t - \tau) - z_{j,k}(t) \| + \frac{1}{2\eta_k T_k} \left( \| x^* - x_k(\tau + 1) \|^2 + \frac{1}{N} \sum_{j=1}^{N} \| \hat{y}_{j,k}^* - y_{j,k}(\tau + 1) \|^2 \right) + \frac{c}{NT_k} \sum_{t=\tau+1}^{T_k} \sum_{j=1}^{N} \| x_{j,k}(t) - x_k(t) \| + \frac{2rc(\tau + 1)}{T_k} + \eta_k c^2.
$$

Proof. Using Assumption 12, $\text{PDG}^-$ can be upper bounded by

$$
\text{PDG}^- \leq \frac{c}{NT_k} \sum_{t=T_k-\tau+1}^{T_k} \sum_{j=1}^{N} \left( \| \hat{y}_{j,k}^* - y_{j,k}(t) \| + \| x_{j,k}(t) - x^* \| \right).
$$

Since the domain is bounded, this term is upper bounded by $2rc\tau/T_k$.

Next, we deal with $\text{PDG}^+$. We divide each $\langle g_{j,y}(z_{j,k}(t)), \hat{y}_{j,k}^* - y_{j,k}(t) \rangle + \langle g_{j,x}(z_{j,k}(t)), x_{j,k}(t) - x^* \rangle$ into a sum of five terms (4.19a)–(4.19e) by adding and subtracting $G_{j,x}(z_{j,k}(t); \xi_{k,t+\tau})$ and $G_{j,y}(z_{j,k}(t); \xi_{k,t+\tau})$ into the first arguments of two inner products, respectively, and then inserting $y_j^*$ into the second argument for the first resulting inner product,

$$
\langle g_{j,y}(z_{j,k}(t)) - G_{j,y}(z_{j,k}(t); \xi_{k,t+\tau}), \hat{y}_{j,k}^* - y_j^* \rangle \quad (4.19a)
$$

$$
\langle g_{j,y}(z_{j,k}(t)) - G_{j,y}(z_{j,k}(t); \xi_{k,t+\tau}), y_j^* - y_{j,k}(t) \rangle \quad (4.19b)
$$
\begin{align}
\langle G_{j,y}(z_{j,k}(t); \xi_{k,t+\tau}), \hat{y}_{j,k}^* - y_{j,k}(t) \rangle & \quad \quad \quad (4.19c) \\
\langle g_{j,x}(z_{j,k}(t)) - G_{j,x}(z_{j,k}(t); \xi_{k,t+\tau}), x_{j,k}(t) - x^* \rangle & \quad (4.19d) \\
\langle G_{j,x}(z_{j,k}(t); \xi_{k,t+\tau}), x_{j,k}(t) - x^* \rangle. & \quad (4.19e)
\end{align}

We sum each of (4.19a)–(4.19e) over \( t = 1, \ldots, T_k - \tau \) and \( j = 1, \ldots, N \), divide it by \( NT_k \), and represent each of them using \( S_1 \) to \( S_5 \). Thus, \( \mathbb{E}[\mathbf{PDG}^+] = \mathbb{E}[S_1 + S_2 + S_3 + S_4 + S_5] \). We next bound each term separately.

Bounding the term \( \mathbb{E}[S_1] \): For agent \( j \), we have a martingale difference sequence \( \{X_j(t)\}_{i=1}^{T_k - \tau} \),

\[
X_j(t) = g_{j,y}(z_{j,k}(t)) - G_{j,y}(z_{j,k}(t); \xi_{k,t+\tau}) - E_{j,k}(t)
\]

where \( E_{j,k}(t) := \mathbb{E}[g_{j,y}(z_{j,k}(t)) - G_{j,y}(z_{j,k}(t); \xi_{k,t+\tau}) | \mathcal{F}_{k,t}] \) and \( M = 4c \) in Lemma 55. This allows us to rewrite \( S_1 \) as

\[
S_1 = \frac{1}{N} \sum_{j=1}^{N} \left( \frac{1}{T_k} \sum_{t=1}^{T_k-\tau} X_j(t), \hat{y}_{j,k}^* - y_{j,k} \right) + \frac{1}{NT_k} \sum_{j=1}^{N} \sum_{t=1}^{T_k-\tau} \langle E_{j,k}(t), \hat{y}_{j,k}^* - y_{j,k} \rangle.
\]

Since \( (T_k - \tau)/T_k \leq 1 \), Lemma 55 implies

\[
\mathbb{E} \left[ \left\| \frac{1}{T_k} \sum_{t=1}^{T_k-\tau} X_j(t) \right\|^2 \right] \leq \frac{64c^2}{T_k - \tau}.
\]
Using Assumption 12 and the mixing time property (4.11), we can bound $\|E_{j,k}(t)\|$ by

$$
\|E_{j,k}(t)\| = \left\| \int G_{i,y}(z_{i,k}(t); \xi)(\pi(\xi) - p_{k,t,\tau}^{[t]}(\xi)) \, d\mu(\xi) \right\|
\leq c \int \left| \pi(\xi) - p_{k,t,\tau}^{[t]}(\xi) \right| \, d\mu(\xi)
= c \, d_{tv}(P_{k,t,\tau}^{[t]}, \Pi).
$$

(4.20)

Applying the triangle and Cauchy-Schwartz inequalities to $E[S_1]$ and using (4.20) lead to,

$$
E[S_1] \leq \frac{1}{N} \frac{8c}{\sqrt{\tau - T_k}} \sum_{j=1}^{N} \left\| \hat{y}_{j,k} - y_j^* \right\| + \frac{c}{NT_k} \sum_{j=1}^{N} \sum_{t=1}^{T_k-\tau} \mathbb{E}\left[ \, d_{tv}(P_{k,t,\tau}^{[t]}, \Pi) \right] \left\| \hat{y}_{j,k} - y_j^* \right\|
\leq \frac{c}{N} \left( \frac{8}{\sqrt{\tau - T_k}} + \frac{1}{T_k} \right) \sum_{j=1}^{N} \left\| \hat{y}_{j,k} - y_j^* \right\|.
$$

where the last inequality follows from the mixing time property: if we choose $T_k$ such that $\tau = 1 + \lceil \log(\Gamma T_k)/\log |\rho| \rceil \geq \tau_{tv}(P_{k}^{[t]}), 1/T_k),$ then we have $E[d_{tv}(P_{k,t,\tau}^{[t]}, \Pi)] \leq 1/T_k.$

**Bounding the terms** $E[S_2]$ and $E[S_4]$: Using the Cauchy-Schwartz inequality and (4.20), we can bound $E[S_2]$ by

$$
E[S_2] \leq \frac{c}{NT_k} \sum_{t=1}^{T_k-\tau} \sum_{j=1}^{N} \mathbb{E}\left[ \, d_{tv}(\Pi, P_{k,t,\tau}^{[t]}), \Pi \right] \left\| y_j^* - y_{j,k}(t) \right\|
\leq \frac{c}{NT_k^2} \sum_{t=1}^{T_k-\tau} \sum_{j=1}^{N} \mathbb{E}\left[ \left\| y_j^* - y_{j,k}(t) \right\| \right]
\leq \frac{r c}{T_k}.
$$

Similarly, we have

$$
E[S_4] \leq \frac{c}{NT_k^2} \sum_{t=1}^{T_k-\tau} \sum_{j=1}^{N} \mathbb{E}\left[ \left\| x_{j,k}(t) - x^* \right\| \right] \leq \frac{r c}{T_k}.
$$
Bounding the terms $\mathbb{E}[S_3]$ and $\mathbb{E}[S_5]$: We re-index the sum in $S_3$ over $t$ and write it as

$$
\frac{1}{NT_k} \sum_{t=\tau+1}^{T_k} \sum_{j=1}^{N} \langle G_{j,y}(z_{j,k}(t-\tau); \xi_{k,t}), \hat{y}_{j,k}^* - y_{j,k}(t-\tau) \rangle
$$

where each $\langle G_{j,y}(z_{j,k}(t-\tau); \xi_{k,t}), \hat{y}_{j,k}^* - y_{j,k}(t-\tau) \rangle$ can be split into a sum of the following three inner products,

$$
\langle G_{j,y}(z_{j,k}(t-\tau); \xi_{k,t}), \hat{y}_{j,k}^* - y_{j,k}(t) \rangle
$$

$$
\langle G_{j,y}(z_{j,k}(t); \xi_{k,t}), \hat{y}_{j,k}^* - y_{j,k}(t) \rangle
$$

$$
\langle G_{j,y}(z_{j,k}(t); \xi_{k,t}), y_{j,k}(t) - y_{j,k}(t-\tau) \rangle.
$$

Combining the Lipschitz continuity of the gradient $\|G_{j,y}(z_{j,k}(t-\tau); \xi_{k,t}) - G_{j,y}(z_{j,k}(t); \xi_{k,t})\| \leq L\|z_{j,k}(t-\tau) - z_{j,k}(t)\|$ with the domain/gradient boundedness yields,

$$
S_3 \leq \frac{rL}{NT_k} \sum_{t=\tau+1}^{T_k} \sum_{j=1}^{N} \|z_{j,k}(t-\tau) - z_{j,k}(t)\|
$$

$$
+ \frac{1}{NT_k} \sum_{t=\tau+1}^{T_k} \sum_{j=1}^{N} \langle G_{j,y}(z_{j,k}(t); \xi_{k,t}), \hat{y}_{j,k}^* - y_{j,k}(t) \rangle
$$

$$
+ \frac{c}{NT_k} \sum_{t=\tau+1}^{T_k} \sum_{j=1}^{N} \|y_{j,k}(t) - y_{j,k}(t-\tau)\|.
$$

Similarly, we have

$$
S_5 \leq \frac{rL}{NT_k} \sum_{t=\tau+1}^{T_k} \sum_{j=1}^{N} \|z_{j,k}(t-\tau) - z_{j,k}(t)\|
$$

$$
+ \frac{c}{NT_k} \sum_{t=\tau+1}^{T_k} \sum_{j=1}^{N} \|x_{j,k}(t-\tau) - x_{j,k}(t)\|
$$

$$
+ \frac{1}{NT_k} \sum_{t=\tau+1}^{T_k} \sum_{j=1}^{N} \langle G_{j,x}(z_{j,k}(t); \xi_{k,t}), x_{j,k}(t) - x^* \rangle.
$$
Inserting $z_k(t)$ into the second argument of the inner product in the above bound on $S_5$ and using inequality $(\|a\| + \|b\|)^2 \leq 2\|a\|^2 + 2\|b\|^2$, we bound $S_3 + S_5$ by

$$S_3 + S_5 \leq \frac{2rL + \sqrt{2}C}{NT_k} \sum_{t=\tau+1}^{T_k} \sum_{j=1}^{N} \|z_{j,k}(t - \tau) - z_{j,k}(t)\|$$

$$+ \frac{1}{NT_k} \sum_{t=\tau+1}^{T_k} \sum_{j=1}^{N} \langle G_{j,y}(z_{j,k}(t); \xi_{k,t}), \hat{y}_{j,k}^* - y_{j,k}(t) \rangle$$

$$+ \frac{1}{NT_k} \sum_{t=\tau+1}^{T_k} \sum_{j=1}^{N} \langle G_{j,x}(z_{j,k}(t); \xi_{k,t}), x_{j,k}(t) - \hat{x}_k(t) \rangle$$

$$+ \frac{1}{T_k} \sum_{t=\tau+1}^{T_k} \left\langle \frac{1}{N} \sum_{j=1}^{N} G_{j,x}(z_{j,k}(t); \xi_{k,t}), \hat{x}_k(t) - x^* \right\rangle.$$

On the right-hand side of the above inequality, the third term can be bounded by applying the Cauchy-Schwartz inequality and the gradient boundedness; for the second term and the fourth term, application of Lemma 26 yields,

$$\sum_{t=\tau+1}^{T_k} \langle G_{j,y}(z_{j,k}(t); \xi_{k,t}), \hat{y}_{j,k}^* - y_{j,k}(t) \rangle \leq \frac{\|\hat{y}_{j,k}^* - y_{j,k}(\tau + 1)\|^2}{\frac{\eta_k}{2}} + \frac{\eta_k}{2} \sum_{t=\tau+1}^{T_k} \|G_{j,y}(z_{j,k}(t); \xi_{k,t})\|^2$$

$$\sum_{t=\tau+1}^{T_k} \left\langle \frac{1}{N} \sum_{j=1}^{N} G_{j,x}(z_{j,k}(t); \xi_{k,t}), \hat{x}_k(t) - x^* \right\rangle \leq \frac{\|\hat{x}_k(\tau + 1) - x^*\|^2}{2\eta_k} + \frac{\eta_k}{2} \sum_{t=\tau+1}^{T_k} \left\| \frac{1}{N} \sum_{j=1}^{N} G_{j,x}(z_{j,k}(t); \xi_{k,t}) \right\|^2.$$
which allows us to bound $S_3 + S_5$ by

\[
S_3 + S_5 \leq \frac{2rL + \sqrt{2c}}{NT_k} T_k \sum_{t=\tau+1}^N \sum_{j=1}^N \|z_{j,k}(t - \tau) - z_{j,k}(t)\|
\]

\[
+ \frac{1}{2\eta_k T_k} \left( \|x_k(\tau + 1) - x^*\|^2 + \frac{1}{N} \sum_{j=1}^N \|y^*_j - y_j(\tau + 1)\|^2 \right)
\]

\[
+ \frac{c}{NT_k} T_k \sum_{t=\tau+1}^N \sum_{j=1}^N \|x_{j,k}(t) - x_k(t)\| + \eta_k c^2.
\]

Taking expectation of $S_3 + S_5$ and adding previous bounds on $\mathbb{E}[S_1]$, $\mathbb{E}[S_2]$, and $\mathbb{E}[S_4]$ to it lead to the final bound on $\mathbb{E}[\text{PDG}^+]$. The proof is completed by adding the established upper bounds on $\mathbb{E}[\text{PDG}^+]$ and $\mathbb{E}[\text{PDG}^-]$.

Lemma 30 is based on the ergodic analysis of the mixing process. We may set $\tau = 0$ to take the traditional stochastic gradient method with i.i.d. samples. In fact, by combining results of Lemmas 29 and 30 we can obtain a loose bound $O(1/\sqrt{T_k})$ bound for the surrogate gap $\text{err}'(\hat{x}_{i,k}, \hat{y}_k)$. Instead, we combine the strong concavity of $\psi_j(x, y_j)$ in terms of $y_j$ with Lemma 30 to establish $O(1/T_k)$ bound on the surrogate gap.

Lemma 31 Let Assumptions 10–13 hold. For $\eta_k$ and $T_k$ that satisfy $L_x \eta_k T_k \geq 16$ and $T_k \geq 1 + \lceil \log(\Gamma T_k)/\log \rho \rceil = \tau$, we have

\[
\mathbb{E}[\text{err}'(\hat{x}_{i,k}, \hat{y}_k)] \leq 4 \mathbb{E}[\text{NET}'] + \frac{4c^2}{L_y} \left( \frac{4}{\sqrt{T_k - \tau}} + \frac{1}{T_k} \right)^2
\]

\[
+ 16\eta_k (r L + c) \tau + 4\eta_k c^2 + \frac{8\eta_k c^2 \tau}{T_k} + \frac{8rc(\tau + 1)}{T_k}
\]

\[
+ \frac{16 \mathbb{E}[\text{err}'(\hat{x}_{i,k-1}, \hat{y}_{k-1})]}{L_y \eta_k T_k} + \sum_{j=1}^N \frac{8 \mathbb{E}[\text{err}'(\hat{x}_{j,k-1}, \hat{y}_{k-1})]}{L_x \eta_k N T_k}
\]
where

\[ \text{NET}' := \text{NET} + \frac{4 (rL + c)}{NT_k} \sum_{t=1}^{T_k} \sum_{j=1}^{N} \|x_{j,k}(t) - x_k(t)\|. \]

**Proof.** Taking expectation of (4.17c) yields,

\[
\mathbb{E}[\text{err}'(\hat{x}_{i,k}, \hat{y}_k)] \geq \frac{L_y}{2N} \sum_{j=1}^{N} \mathbb{E}[\|y^*_j - \hat{y}^*_{j,k}\|^2] \\
\geq \frac{L_y}{2} \left( \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}[\|y^*_j - \hat{y}^*_{j,k}\|] \right)^2
\]

where we use \( \mathbb{E}[X^2] \geq \mathbb{E}[X]^2 \) along with \((1/N) \sum_{i=1}^{N} a_i^2 \geq ((1/N) \sum_{i=1}^{N} a_i)^2 \) to establish the second inequality. Substituting the above inequality into the result in Lemma 30 and combining it with Lemma 29 yield,

\[
\mathbb{E}[\text{err}'(\hat{x}_{i,k}, \hat{y}_k)] \leq \mathbb{E}[\text{NET}] + \mathbb{E}[\text{MIX}] + \frac{\sqrt{2c}}{L_y} \left( \frac{4}{\sqrt{T_k} - \tau} + \frac{1}{T_k} \right) \mathbb{E}[\text{err}'(\hat{x}_{i,k}, \hat{y}_k)]^{1/2}.
\]

Thus, \( \zeta := \mathbb{E}[\text{err}'(\hat{x}_{i,k}, \hat{y}_k)]^{1/2} \) satisfies the quadratic inequality \( \zeta^2 \leq a\zeta + b \). Combining the values of \( \zeta \) that satisfy this inequality with \((\|a\| + \|b\|)^2 \leq 2\|a\|^2 + 2\|b\|^2 \) leads to,

\[
\mathbb{E}[\text{err}'(\hat{x}_{i,k}, \hat{y}_k)] \leq 2\mathbb{E}[\text{NET}] + 2\mathbb{E}[\text{MIX}] + \frac{2c^2}{L_y} \left( \frac{4}{\sqrt{T_k} - \tau} + \frac{1}{T_k} \right)^2.
\] (4.21)

The remaining task is to express expectations on the right-hand side of (4.21) in terms of previously introduced terms. Lemma 25 provides a bound on \( \mathbb{E}[\text{NET}] \) and, next, we evaluate the
terms in $E[\text{MIX}]$. Combination of the triangle inequality with non-expansiveness of projection allows us to bound $E[\|z_{j,k}(t - \tau) - z_{j,k}(t)\|]$ by

$$E[\|z_{j,k}(t - \tau) - z_{j,k}(t)\|] \leq E[\|y'_{j,k}(t - \tau) - y'_{j,k}(t)\|]\ + E[\|x_{j,k}(t - \tau) - x_{j,k}(t)\|]$$

where $E[\|x_{j,k}(t - \tau) - x_{j,k}(t)\|]$ can be further bounded by a sum of three terms:

$$E[\|x_{j,k}(t - \tau) - x_{k}(t - \tau)\|], E[\|x_{k}(t - \tau) - x_{k}(t)\|], \text{ and } E[\|x_{k}(t) - x_{j,k}(t)\|].$$

Similarly, $\|x_{k}(t - \tau) - x_{k}(t)\| \leq \|x'_{k}(t - \tau) - x'_{k}(t)\|$. By the gradient boundedness, it is clear from the primal-dual updates that $E[\|y'_{j,k}(t - \tau) - y'_{j,k}(t)\|], E[\|x_{k}(t - \tau) - x_{k}(t)\|] \leq \eta k c \tau$. Thus, we have

$$\frac{1}{NT_k} \sum_{t = \tau + 1}^{T_k} \sum_{j = 1}^{N} E[\|z_{j,k}(t - \tau) - z_{j,k}(t)\|]$$

$$\leq 2\eta k c \tau + \frac{1}{NT_k} \sum_{t = \tau + 1}^{T_k} \sum_{j = 1}^{N} E[\|x_{j,k}(t - \tau) - x_{k}(t - \tau)\|]$$

$$+ \frac{1}{NT_k} \sum_{t = \tau + 1}^{T_k} \sum_{j = 1}^{N} E[\|x_{k}(t) - x_{j,k}(t)\|]$$

$$\leq 2\eta k c \tau + \frac{2}{NT_k} \sum_{t = 1}^{T_k} \sum_{j = 1}^{N} E[\|x_{j,k}(t) - x_{k}(t)\|]$$
where we sum over $t$ from $1$ to $T_k$ instead of from $\tau + 1$ to $T_k$ in the last inequality. Now, we turn to next two expectations as follows,

$$
\mathbb{E}[\|x^* - x_k(\tau + 1)\|^2] \\
= \mathbb{E}[\|x_k(\tau + 1) - x_k(1) + x_k(1) - x^*\|^2] \\
\leq 2 \mathbb{E}[\|x_k(\tau + 1) - x_k(1)\|^2] + 2 \mathbb{E}[\|x_k(1) - x^*\|^2] \\
\leq 2 \mathbb{E}[\|\bar{x}_k(\tau + 1) - \bar{x}_k(1)\|^2] + 2 \mathbb{E}[\|x_k(1) - x^*\|^2] \\
\leq 2 \eta_k^2 c^2 \tau^2 + \frac{2}{N} \sum_{j=1}^{N} \mathbb{E}[\|x'_{j,k}(1) - x^*\|^2] \\
= 2 \eta_k^2 c^2 \tau^2 + \frac{2}{N} \sum_{j=1}^{N} \mathbb{E}[\|x_{j,k}(1) - x^*\|^2]
$$

where we apply the inequality $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and the non-expansiveness of projection for the first and the second inequalities; the third inequality is because of (4.14), the gradient
boundedness, and application of the Jensen’s inequality to $\| \cdot \|^2$ with $x_k(1) := \frac{1}{N} \sum_{j=1}^{N} x_{j,k}^*(1)$; and the last equality follows the initialization $x_{j,k}^*(1) = x_{j,k}(1)$. Similarly, we derive

$$\mathbb{E}[\| \hat{y}_{j,k}^* - y_{j,k}(\tau + 1)\|^2]$$

$$= \mathbb{E}[\| \hat{y}_{j,k}^* - y_{j,k}(1) + y_{j,k}(1) - y_{j,k}(\tau + 1)\|^2]$$

$$\leq 2 \mathbb{E}[\| \hat{y}_{j,k}^* - y_{j,k}(1)\|^2] + 2 \mathbb{E}[\| y_{j,k}(1) - y_{j,k}(\tau)\|^2]$$

$$\leq 2 \mathbb{E}[\| \hat{y}_{j,k}^* - y_{j,k}(1)\|^2] + 2 \mathbb{E}[\| y_{j,k}^*(1) - y_{j,k}^*(\tau)\|^2]$$

$$\leq 2 \mathbb{E}[\| \hat{y}_{j,k}^* - y_{j,k}(1)\|^2] + 2 \eta_k^2 c^2 \tau^2$$

$$= 2 \mathbb{E}[\| \hat{y}_{j,k}^* - y_{j,k}^* + y_{j,k}^* - y_{j,k}(1)\|^2] + 2 \eta_k^2 c^2 \tau^2$$

$$\leq 4 \mathbb{E}[\| \hat{y}_{j,k}^* - y_{j,k}^*\|^2] + 4 \mathbb{E}[\| y_{j,k}^* - y_{j,k}(1)\|^2] + 2 \eta_k^2 c^2 \tau^2.$$

From (4.17c) we know that

$$\frac{1}{N} \sum_{j=1}^{N} \mathbb{E}[\| \hat{y}_{j,k}^* - y_{j,k}^*\|^2] \leq \frac{2}{L_y} \mathbb{E}[\text{err}'(\hat{x}_{i,k}, \hat{y}_k)].$$

Now, we collect above inequalities for (4.21). For notational convenience, we sum over $t$ from 1 until $T_k$ instead of from $\tau + 1$ to $T_k$ and combine similar terms to obtain the following bound on $\mathbb{E}[\text{MIX}].$

$$\mathbb{E}[\text{MIX}] \leq 4\eta_k (rL + c)\tau + \frac{4(rL + c)}{NT_k} \sum_{t=1}^{T_k} \sum_{j=1}^{N} \mathbb{E}[\| x_{j,k}(t) - x_k(t)\|]$$

$$+ \frac{1}{\eta_k NT_k} \sum_{j=1}^{N} (2\mathbb{E}[\| y_{j,k}^* - y_{j,k}(1)\|^2] + \mathbb{E}[\| x_{j,k}(1) - x^*\|^2])$$

$$+ \frac{4 \mathbb{E}[\text{err}'(\hat{x}_{i,k}, \hat{y}_k)]}{L_y \eta_k T_k} + \frac{2 \eta_k^2 c^2 \tau^2}{T_k} + \frac{2rc(\tau + 1)}{T_k} + \eta_k c^2.$$
We note the restarting scheme of Algorithm 9, i.e., $x_{j,k}(1) = \hat{x}_{j,k-1}$ and $y_{j,k}(1) = \hat{y}_{j,k-1}$. By (4.17a), (4.17b), and Lemma 27, we have

$$\mathbb{E}[\|x_{j,k}(1) - x^*\|^2] = \mathbb{E}[\|\hat{x}_{j,k-1} - x^*\|^2]$$

$$\leq \frac{2}{L_x} \mathbb{E}[\text{err}(\hat{x}_{j,k-1})]$$

$$\leq \frac{2}{L_x} \mathbb{E}[\text{err}'(\hat{x}_{j,k-1}, \hat{y}_{k-1})].$$

$$\frac{1}{N} \sum_{j=1}^{N} \mathbb{E}[\|y^*_j - y_{j,k}(1)\|^2] = \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}[\|y^*_j - \hat{y}_{j,k-1}\|^2]$$

$$\leq \frac{2}{L_y} \mathbb{E}[\text{err}'(\hat{x}_{i,k-1}, \hat{y}_{k-1})].$$

Combining above inequalities with the bound on $\mathbb{E}[	ext{MIX}]$ yields a bound on $\mathbb{E}[\text{err}'(\hat{x}_{i,k}, \hat{y}_k)].$

Finally, we utilize $L_x \eta_k T_k \geq 16$ to finish the proof.

4.4.3 Proof of main theorem

The proof of Theorem 24 is based on the result in Lemma 31. We leave the mixing time $\tau$ to be determined so that it works for every round $k$ and focus on the averaged surrogate gap $\text{err}'_k := \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}[\text{err}'(\hat{x}_{j,k}, \hat{y}_k)].$

Let

$$H_k := \frac{4c^2}{L_y} \left( \frac{8}{\sqrt{T_k - \tau}} + \frac{1}{T_k} \right)^2$$

$$E_k := 8 (4rL + 6c) \Delta_k + 16 \eta_k c (rL + c) \tau + \frac{8rc (\tau + 1)}{T_k} + 4 \eta_k c^2.$$
We first apply Lemma 25 to obtain \( \mathbb{E}[\text{NET}'] \leq 2(4rL + 6c)\Delta_k \) for each round \( k \). With previous simplified notation, we apply this inequality for the bound in Lemma 31 and then take average over \( i = 1, \ldots, N \) on both sides to obtain,

\[
\text{err}'_k \leq H_k + E_k + \frac{8\eta_k c^2 \tau^2}{T_k} + \frac{\text{err}'_{k-1}}{L' \eta_k T_k} \tag{4.22}
\]

where \( 1/L' := 16/L_y + 8/L_x \). In Algorithm 9, we have updates \( \eta_k = \eta_{k-1}/2 \) and \( T_k = 2T_{k-1} \).

Since \( \eta_1 \geq 4/L' \) and \( T_1 \geq 1 \), we have \( L' \eta_k T_k \geq 4 \) for all \( k \geq 1 \). Clearly, the assumption \( L_x \eta_k T_k \geq 16 \) holds in Lemma 31. Therefore, (4.22) can be simplified into

\[
\text{err}'_k \leq H_k + E_k + \frac{8\eta_k c^2 \tau^2}{T_k} + \frac{\text{err}'_{k-1}}{4}. \tag{4.23}
\]

Simple comparisons show that \( H_k \leq H_{k-1}/2 \) and \( E_k \geq E_{k-1}/2 \). Thus, \( \eta_k/T_k = (1/4)^{k-1} \eta_1/T_1 \) and \( H_k \leq H_1/2^{k-1} \) are true for all \( k \geq 1 \). If we set \( \tau = 1 + \lceil \log(\Gamma) / \log |\rho| \rceil \leq T_1 \) with suitable \( T_1 \) and \( K \), Lemma 31 applies at any round \( k \). Starting from the final round \( k = K \), we repeat (4.23) to obtain,

\[
\text{err}'_K \leq H_K + E_K + \frac{2\eta_1 c^2 \tau^2}{4^{K-2} T_1} + \frac{\text{err}'_{K-1}}{4} \leq \frac{1}{2^{K-2}} \sum_{k=1}^{K} \left( \frac{1}{2} \right)^k H_1 + \sum_{k=1}^{K} \left( \frac{1}{2} \right)^k E_K + \frac{2K \eta_1 c^2 \tau^2}{4^{K-2} T_1} + \left( \frac{1}{4} \right)^K \text{err}'_0 \leq \frac{8H_1 T_1}{T} + 2E_K + \frac{32K \eta_1 c^2 \tau^2 T_1}{T^2} + \frac{r^2 T_1^2}{T^2}.
\]
where $T = \sum_{k=1}^{K} T_k = (2^K - 1)T_1 \leq 2^K T_1$ and $\overline{err}_0' \leq \overline{err}_0 = \sup_{x, y} \| (x, y) \|^2 \leq r^2$. We now substitute $H_1$ and $E_K$ into this bound and bound $\overline{err}_K'$ by

$$\frac{32 c^2 T_1}{L_y T} \left( \frac{8}{\sqrt{T_1} - \tau} + \frac{1}{T_1} \right)^2 + 16 (4rL + 6c) \Delta_K + \frac{32 K \eta_1 c^2 \tau^2 T_1}{T^2} + \frac{r^2 T_1}{T^2} + \frac{16 r c (\tau + 1)}{T_K} + 32 \eta_K c (2rL + \sqrt{2c}) \tau + 8 \eta_K c^2.$$

Since $T \leq 2^K T_1$, we have $\eta_K = \eta_1 / 2^K \leq \eta_1 T_1 / T$ and $T_K = 2^{K-1} T_1 \geq T/2$. Since $T = (2^K - 1)T_1 \geq 2^{K-1} T_1$, we have $K \leq 1 + \log(T/T_1)$. Thus, $\sum_{k=1}^{K} \eta_k T_k = K \eta_1 T_1 \leq \eta_1 T_1 (1 + \log(T/T_1))$. Therefore, the above bound has the following order,

$$C_1 \frac{c (c + rL) \log^2(\sqrt{N} T)}{T (1 - \sigma^2(W))} + C_2 \frac{c (c + rL)(1 + T_1)}{T},$$

where $C_1$ are $C_2$ are absolute constants. Since $\overline{err}_K \leq \overline{err}_K'$, it also bounds $\overline{err}_K$. The proof is completed by combining $\eta_1 \geq 4/L'$ with $T_1 \geq \tau$.

### 4.5 Computational experiments

We first utilize a modified Mountain Car Task [214, Example 10.1] for multi-agent policy evaluation problem. We generate the dataset using the approach presented in [232], obtain a policy by running SARSA with $d = 300$ features, and sample the trajectories of states and actions according to the policy. The discount factor is set to $\gamma = 0.95$. We simulate the communication network with $N$ agents using the Erdős-Rényi graph with connectivity 0.1. At every time instant, each agent observes a local reward that is generated as a random proportion of the total reward. Since the stationary distribution $\Pi$ is unknown, we use sampled averages from the entire dataset to
compute sampled versions $\hat{A}$, $\hat{b}$, and $\hat{C}$ of $A$, $b$, and $C$. We then formulate an empirical MSPBE as $(1/2)\|\hat{A}x - \hat{b}\|_2^2$ and compute the optimal empirical MSPBE. We use this empirical MSPBE as an approximation of the population MSPBE to calculate the optimality gap. The dataset contains 85453 samples and we run our online algorithm over one trajectory of 30000 samples using multiple passes. We set an initial restart time to $T_1 = 10^5$ and a restart round to $K = 4$ to insure $T_1 \simeq O(K + \log T_1)$, take large bounds for Euclidean projections, and choose different learning rates $\eta$.

Figure 4.2: Performance comparison for the centralized problem with $N = 1$. Our algorithm with stepsize $\eta = 0.1$ and $K = 4$ (—) achieves a smaller optimality gap than stochastic primal-dual algorithm with stepsize: $\eta = 0.1$ (---), $\eta = 0.05$ (----), $\eta = 0.025$ (-----), and $\eta = 1/\sqrt{t}$ (–). It also provides a smaller optimality gap than the approach that utilizes pre-collected i.i.d. samples in a buffer (—).

We compare the performance of Algorithm 9 (DHPD) with stochastic primal-dual (SPD) algorithm under different settings. For $N = 1$, SPD corresponds to GTD algorithm in [138, 238, 222], and for $N > 1$, SPD becomes the multi-agent GTD algorithm [127]. We show computational results for $N = 1$ and $N = 10$ in Fig. 4.2 and Fig. 4.3 respectively. The optimality gap is the difference between empirical MSPBE and the optimal one. Our algorithm achieves a smaller
optimality gap than SPD in all cases, thereby demonstrating its sample efficiency. In computational experiments with simple diminishing stepsizes, the algorithm converges relatively slow as is typical in stochastic optimization. Also, our online algorithm competes well against the approach that utilizes pre-collected i.i.d. samples (instead of true i.i.d. samples from the stationary distribution) in a fixed buffer.

In our second computational experiment, we test randomly generated multi-agent MDPs for a ring network with $N = 5$ agents that utilize a fixed policy [127, Example 1]. This leads to a Markov chain with $S = \{1, 2, 3, 4\}$, $\gamma = 0.95$, $\phi(s) = [\phi_1(s) \ \phi_2(s) \ \phi_3(s) \ \phi_4(s)]^T \in \mathbb{R}^4$ where $\phi_i(s) = e^{-(s-i)^2}$, and $r_j(s) = 1 (s = 4)$. In Fig. 4.4 we demonstrate that our online algorithm with an initial restart time $T_1 = 20000$ and a restart round $K = 3$ outperforms SPD algorithm that utilizes diminishing stepsizes or a replay buffer.

![Graph showing performance comparison](image)

Figure 4.3: Performance comparison for the distributed case with $N = 10$. Our algorithm with stepsize $\eta = 0.05$ and $K = 4$ (---) achieves a smaller optimality gap than stochastic primal-dual algorithm with stepsize: $\eta = 0.05$ (---), $\eta = 0.025$ (---), $\eta = 0.0125$ (---), and $\eta = 1/\sqrt{t}$ (—). It also provides a smaller optimality gap than the approach that utilizes pre-collected iid samples in a buffer (—).
optimality gap

iteration count

Figure 4.4: Performance comparison for the distributed case with $N = 5$. Our algorithm with stepsize $\eta = 0.1$ and $K = 3$ (– -) achieves a smaller optimality gap than stochastic primal-dual algorithm with stepsize: $\eta = 0.5$ (⋯), $\eta = 0.25$ (---), $\eta = 0.125$ (····), and $\eta = 1/\sqrt{t}$ (—). It also provides a smaller optimality gap than the approach that utilizes pre-collected iid samples in a buffer (—).

4.6 Concluding remarks

In this chapter, we begin the multi-agent temporal-difference learning with a distributed primal-dual stochastic saddle point problem. We have proposed a new online distributed homotopy-based primal-dual algorithm for minimizing the mean square projected Bellman error under the Markovian setting and establish an $O(1/T)$ error bound. Our result improves the best known $O(1/\sqrt{T})$ error bound for general stochastic primal-dual algorithms and it demonstrates that distributed saddle point programs can be solved efficiently even in applications with limited time budgets.
Chapter 5

Independent policy gradient for Markov potential games

In this chapter, we study global non-asymptotic convergence properties of policy gradient methods for multi-agent reinforcement learning problems in Markov potential games (MPGs). To learn a Nash equilibrium of an MPG in which the size of state space and/or the number of players can be very large, we propose new independent policy gradient algorithms that are run by all players in tandem. When there is no uncertainty in the gradient evaluation, we show that our algorithm finds an $\epsilon$-Nash equilibrium with $O(1/\epsilon^2)$ iteration complexity which does not explicitly depend on the state space size. When the exact gradient is not available, we establish $O(1/\epsilon^5)$ sample complexity bound in a potentially infinitely large state space for a sample-based algorithm that utilizes function approximation. Moreover, we identify a class of independent policy gradient algorithms that enjoy convergence for both zero-sum Markov games and Markov cooperative games with the players that are oblivious to the types of games being played. Finally, we provide computational experiments to corroborate the merits and the effectiveness of our theoretical developments.
5.1 Introduction

Multi-agent reinforcement learning (multi-agent RL) studies how multiple players learn to maximize their long-term returns in a setup where players’ actions influence the environment and other agents’ returns [46, 273]. Recently, multi-agent RL has achieved significant success in various multi-agent learning scenarios, e.g., competitive game-playing [203, 202, 229], autonomous robotics [198, 133], and economic policy-making [283, 223]. In the framework of stochastic games [200, 90], most results are established for fully-competitive (i.e., two-player zero-sum) games; e.g., see [64, 239, 49]. However, to achieve social welfare for AI [60, 59, 209], it is imperative to establish theoretical guarantees for multi-agent RL in Markov games with cooperation.

Policy gradient methods [245, 217] have received a lot of attention for both single-agent [32, 8] and multi-agent RL problems [274, 64, 239]. Independent policy gradient [273, 177] is probably the most practical protocol in multi-agent RL, where each player behaves myopically by only observing her own rewards and actions (as well as the system states), while individually optimizing its own policy. More importantly, independent learning dynamics do not scale exponentially with the number of players in the game. Recently, [64, 131, 277] have in fact shown that multi-agent RL players could perform policy gradient updates independently, while enjoying global non-asymptotic convergence. However, these results are only focused on the basic tabular setting in which the value functions are represented by tables; they do not carry over to large-scale multi-agent RL problems in which the state space size is potentially infinite and the number of players is large. This motivates the following question:

Can we design independent policy gradient methods for large-scale Markov games, with non-asymptotic global convergence guarantees?
In this chapter, we provide the first affirmative answer to this question for a class of mixed cooperative/competitive Markov games: Markov potential games [147, 131, 277].

In Section 5.2 we introduce Markov potential games, Nash equilibrium, and provide necessary background material. In Section 5.3 we present an independent learning protocol. We propose an independent policy gradient method for Markov potential games in Section 5.4 and establish the Nash regret analysis in Section 5.5. In Section 5.6 we establish an extension of our method and analysis to the linear function approximation setting. In Section 5.7 we establish game-agnostic convergence of an optimistic independent policy gradient method for both Markov cooperative games and zero-sum Markov games. We provide computational experiments to demonstrate the merits and the effectiveness of our theoretical findings in Section 5.8 and we close the chapter with concluding remarks in Section 5.9.

### 5.2 Markov potential games

We consider an $N$-player, infinite-horizon, discounted Markov potential game (MPG),

$$\text{MPG} \left( S, \{A_i\}_{i=1}^N, P, \{r_i\}_{i=1}^N, \gamma, \rho \right)$$

where $S$ is a state space, $A_i$ is an action space for the $i$th player, with the joint action space of $N \geq 2$ players denoted as $\mathcal{A} := A_1 \times \ldots \times A_N$, $P$ is a transition probability measure specified by a distribution $P(\cdot | s, a)$ over $S$ if $N$ players jointly take an action $a$ from $\mathcal{A}$ in state $s$, $r_i: S \times \mathcal{A} \to [0, 1]$ is an immediate reward function for the $i$th player, $\gamma \in [0, 1)$ is a discount factor, and $\rho$ is an initial state distribution over $S$. We assume that all action spaces are finite with the
same size $A = A_i = |A_i|$ for all $i = 1, \ldots, N$. It is straightforward to apply our analysis to the general case in which players’ finite action spaces have different sizes.

For the $i$th player, $\Delta(A_i)$ represents the probability simplex over the action set $A_i$. A stochastic policy for player $i$ is given by $\pi_i: S \rightarrow \Delta(A_i)$ that specifies the action distribution $\pi_i(\cdot | s) \in \Delta(A_i)$ for each state $s \in S$. The set of stochastic policies for player $i$ is denoted by $\Pi_i := (\Delta(A_i))^{|S|}$, the joint probability simplex is given by $\Delta(A) := \Delta(A_1) \times \ldots \times \Delta(A_N)$, and the joint policy space is $\Pi := (\Delta(A))^{|S|}$. A Markov product policy $\pi := \{\pi_i\}_{i=1}^N \in \Pi$ for $N$ players consists of the policy $\pi_i \in \Pi$ for all players $i = 1, \ldots, N$. We use the shorthand $\pi_{-i} = \{\pi_k\}_{k=1, k \neq i}^N$ to represent the policy of all but the $i$th player. We denote by $V^\pi_i: S \rightarrow \mathbb{R}$ the $i$th player value function under the joint policy $\pi$, starting from an initial state $s^{(0)} = s$:

$$V^\pi_i(s) := \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t r_i(s^{(t)}, a^{(t)}) \mid s^{(0)} = s \right]$$

where the expectation $\mathbb{E}^\pi$ is over $a^{(t)} \sim \pi(\cdot | s^{(t)})$ and $s^{(t+1)} \sim \mathbb{P}(\cdot | s^{(t)}, a^{(t)})$. Finally, $V^\pi_i(\mu)$ denotes the expected value function of $V^\pi_i(s)$ over a state distribution $\mu$, $V^\pi_i(\mu) := \mathbb{E}_{s \sim \mu}[V^\pi_i(s)]$.

In an MPG, at any state $s \in S$, there exists a global function – the potential function $\Phi^\pi(s)$: $\Pi \times S \rightarrow \mathbb{R}$ – that captures the incentive of all players to vary their policies at state $s$,

$$V^\pi_{i, \pi_{-i}}(s) - V^{\pi_i', \pi_{-i}}(s) = \Phi^\pi_{i, \pi_{-i}}(s) - \Phi^{\pi_i', \pi_{-i}}(s)$$

for any policies $\pi_i, \pi_i' \in \Pi_i$ and $\pi_{-i} \in \Pi_{-i}$. Let $\Phi^\pi(\mu) := \mathbb{E}_{s \sim \mu}[\Phi^\pi(s)]$ be the expected potential function over a state distribution $\mu$. Thus, $V^\pi_{i, \pi_{-i}}(\mu) - V^{\pi_i', \pi_{-i}}(\mu) = \Phi^\pi_{i, \pi_{-i}}(\mu) - \Phi^{\pi_i', \pi_{-i}}(\mu)$. There always exists a constant $C_\Phi > 0$ such that $|\Phi^\pi(\mu) - \Phi^{\pi'}(\mu)| \leq C_\Phi$ for any $\pi, \pi', \mu$; see a
trivial upper bound in Lemma 68 in Appendix D. An important subclass of MPGs is given by Markov cooperative games in which all players share the same reward function $r = r_i$ for all $i = 1, \ldots, N$.

We also denote by $Q^\pi_i : S \times A \to \mathbb{R}$ the state-action value function under policy $\pi$, starting from an initial state-action pair $(s^{(0)}, a^{(0)}) = (s, a)$:

$$Q^\pi_i(s, a) := \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t r_i(s^{(t)}, a^{(t)}) \mid s^{(0)} = s, a^{(0)} = a \right].$$

The value function can be equivalently expressed as $V^\pi_i(s) = \sum_{a' \in A} \pi(a' \mid s) Q^\pi_i(s, a')$. For each player $i$, by averaging out $\pi_{-i}$, we can define the averaged state-action value function $\bar{Q}^\pi_i, \pi_{-i}$:

$$\bar{Q}^\pi_i, \pi_{-i}(s, a_i) := \sum_{a_{-i} \in A_{-i}} \pi_{-i}(a_{-i} \mid s) Q^\pi_i, \pi_{-i}(s, a_i, a_{-i})$$

where $A_{-i}$ is the set of actions of all but the $i$th player. We use the shorthand $\bar{Q}^\pi_i$ for $\bar{Q}^\pi_i, \pi_{-i}$ when $\pi_i$ and $\pi_{-i}$ are from the same joint policy $\pi$. It is straightforward to see that $V^\pi_i$, $Q^\pi_i$, and $\bar{Q}^\pi_i$ are bounded between 0 and $1/(1 - \gamma)$.

We recall the notion of (Markov perfect stationary) Nash equilibrium [90]. A joint policy $\pi^*$ is called a Nash equilibrium if for each player $i = 1, \ldots, N$,

$$V^{\pi^*, \pi_{-i}}_i(s) \geq V^{\pi_i, \pi_{-i}^*}_i(s), \text{ for all } \pi_i \in \Pi_i, s \in S$$

and called an $\epsilon$-Nash equilibrium if for $i = 1, \ldots, N$,

$$V^{\pi^*, \pi_{-i}}_i(s) \geq V^{\pi_i, \pi^*_{-i}}_i(s) - \epsilon, \text{ for all } \pi_i \in \Pi_i, s \in S.$$
Nash equilibria for MPGs with finite states and actions always exist \[90\]. When the state space is infinite, we assume the existence of a Nash equilibrium; see \[219, 150, 151, 12\] for cases with countable or compact state spaces.

Given policy \( \pi \) and initial state \( s^{(0)} \), we define the discounted state visitation distribution,

\[
d^\pi_{s^{(0)}}(s) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr^\pi(s^{(t)} = s | s^{(0)}).
\]

For a state distribution \( \mu \), define \( d^\pi_\mu(s) = \mathbb{E}_{s^{(0)} \sim \mu}[d^\pi_{s^{(0)}}(s)] \). By definition, \( d^\pi_\mu(s) \geq (1 - \gamma)\mu(s) \) for any \( \mu \) and \( s \).

**Remark 11** It is useful to introduce a variant of the performance difference lemma \[8\] for multiple players; for other versions, see \[274, 64, 277, 131\]. For the \( i \)th player, if we fix the policy \( \pi_{-i} \) and any state distribution \( \mu \), then for any two policies \( \hat{\pi}_i \) and \( \bar{\pi}_i \),

\[
V^{\hat{\pi}_i, \pi_{-i}}_i(\mu) - V^{\bar{\pi}_i, \pi_{-i}}_i(\mu) = \frac{1}{1 - \gamma} \sum_{s,a_i} d^{\hat{\pi}_i, \pi_{-i}}(s) \cdot (\hat{\pi}_i - \bar{\pi}_i)(a_i | s) Q^{\bar{\pi}_i, \pi_{-i}}_i(s,a_i)
\]

where \( Q^{\bar{\pi}_i, \pi_{-i}}_i(s,a_i) = \sum_{a_{-i}} \pi_{-i}(a_{-i} | s) Q^{\bar{\pi}_i, \pi_{-i}}_i(s,a_i,a_{-i}) \).

It is common to use the distribution mismatch coefficient to measure the exploration difficulty in policy optimization \[8\]. We next define a distribution mismatch coefficient for MPGs \[131\] in Definition 2 and its minimax variant in Definition 3.

**Definition 2 (Distribution mismatch coefficient)** For any distribution \( \mu \in \Delta(S) \) and policy \( \pi \in \Pi \), the distribution mismatch coefficient \( \kappa_\mu \) is the maximum distribution mismatch of \( \pi \) relative to \( \mu \),

\[
\kappa_\mu := \sup_{\pi \in \Pi} \left\| d^\pi_\mu / \mu \right\|_\infty,
\]

where the division \( d^\pi_\mu / \mu \) is evaluated in a componentwise manner.
Definition 3 (Minimax distribution mismatch coefficient) For any distribution $\mu \in \Delta(S)$, the minimax distribution mismatch coefficient $\tilde{\kappa}_\mu$ is the minimax value of the distribution mismatch of $\pi$ relative to $\nu$, $\tilde{\kappa}_\mu := \inf_{\nu \in \Delta(S)} \sup_{\pi \in \Pi} \|d^{\pi}_\mu/\nu\|_\infty$, where the division $d^{\pi}_\mu/\nu$ is evaluated in a componentwise manner.

Other notation. We denote by $\|\cdot\|$ the $\ell_2$-norm of a vector or the spectral norm of a matrix. The inner product of a function $f: S \times A \to \mathbb{R}$ with $p \in \Delta(A)$ at fixed $s \in S$ is given by

$$\langle f(s, \cdot), p(\cdot) \rangle_A := \sum_{a \in A} f(s, a) p(a).$$

The $\ell_2$-norm projection operator onto a convex set $\Omega$ is defined as $P_\Omega(x) := \text{argmin}_{x' \in \Omega} \|x' - x\|$. For functions $f$ and $g$, we write $f(n) = O(g(n))$ if there exists $N < \infty$ and $C < \infty$ such that $f(n) \leq C g(n)$ for $n \geq N$, and write $f(n) = \tilde{O}(g(n))$ if $\log g(n)$ appears in $O(\cdot)$. We use \(\prec\) and \(\succ\) to denote \(\leq\) and \(\geq\) up to a constant.

5.3 Independent learning protocol

Algorithm 10 Independent policy gradient ascent

1: **Parameters:** $\eta > 0.$
   **Initialization:** Let $\pi_i^{(1)}(a_i | s) = 1/A$ for $s \in S$, $a_i \in A_i$ and $i = 1, \ldots, N$.
2: **for** step $t = 1, \ldots, T$ **do**
3:   **for** player $i = 1, \ldots, N$ (in parallel) **do**
4:     Define player $i$’s policy on $s \in S$,
5:     $$\pi_i^{(t+1)}(\cdot | s) := \argmax_{\pi_i(\cdot | s) \in \Delta(A_i)} \left\{ \langle \pi_i(\cdot | s), \bar{Q}_i^{(t)}(s, \cdot) \rangle_{A_i} - \frac{1}{2\eta} \|\pi_i(\cdot | s) - \pi_i^{(t)}(\cdot | s)\|^2 \right\}$$
6:     where $\bar{Q}_i^{(t)}(s, a_i)$ is a shorthand for $Q_i^{(t)}(s, a_i)$ (defined in Definition 5.2).
5:   **end for**
6: **end for**

We examine an independent learning setting \[273, 64, 177\] in which all players repeatedly execute their own policy and update rules individually. At each time $t$, all players propose their
own policies $\pi_i^{(t)}: S \to \Delta(A_i)$ with the player index $i = 1, \ldots, N$, while a game oracle can either evaluate each player’s policy or generate a set of sample trajectories for each player. In repeating such protocol for $T$ times, each player behaves myopically in optimizing its own policy.

To evaluate the learning performance, we introduce a notion of regret,

$$\text{Nash-Regret}(T) := \frac{1}{T} \sum_{t=1}^{T} \max_i \left( \max_{\pi_i'} V_i^{\pi_i', \pi_{-i}^{(t)}}(\rho) - V_i^{\pi_i^{(t)}}(\rho) \right)$$

which averages the worst player’s local gaps in $T$ iterations: $\max_{\pi_i'} V_i^{\pi_i', \pi_{-i}^{(t)}}(\rho) - V_i^{\pi_i^{(t)}}(\rho)$ for $t = 1, \ldots, T$, where $\max_{\pi_i'} V_i^{\pi_i', \pi_{-i}^{(t)}}(\rho)$ is the $i$th player best response given $\pi_{-i}^{(t)}$. In Nash-Regret($T$), we compare the learnt joint policy $\pi^{(t)}$ with the best policy that the $i$th player can take by fixing $\pi_{-i}^{(t)}$. We notice that Nash-Regret is closely related to the notion of dynamic regret [285] in which the regret comparator changes over time. This is a suitable notion because the environment is non-stationary from the perspective of an independent learner [156, 273].

To obtain an $\epsilon$-Nash equilibrium $\pi^{(t*)}$ with a tolerance $\epsilon > 0$, our goal is to show the following average performance,

$$\text{Nash-Regret}(T) = \epsilon.$$

The existence of such $t*$ is straightforward,

$$t* := \arg\min_{1 \leq t \leq T} \max_i \left( \max_{\pi_i'} V_i^{\pi_i', \pi_{-i}^{(t)}}(\rho) - V_i^{\pi_i^{(t)}}(\rho) \right).$$

Since each summand above is non-negative, $V_i^{\pi^{(t*)}}(\rho) \geq V_i^{\pi_i', \pi_{-i}^{(t*)}}(\rho) - \epsilon$ for any $\pi_i'$ and $i = 1, \ldots, N$, which implies that $\pi^{(t*)}$ is an $\epsilon$-Nash equilibrium.
5.4 Independent policy gradient methods

In this section, we assume that we have access to exact gradient and examine a gradient-based method for learning a Nash equilibrium in Markov potential/cooperative games.

5.4.1 Policy gradient for Markov potential games

A natural independent learning scheme for MPGs is to let every player independently perform policy gradient ascent \[131, 277\]. In this approach, the \(i\)th player updates its policy according to the gradient of the value function with respect to the policy parameters,

\[
\pi^{(t+1)}_i(\cdot | s) \leftarrow P_{\Delta(A_i)} \left( \pi^{(t)}_i(\cdot | s) + \eta \left. \frac{\partial V^\pi_i(\rho)}{\partial \pi_i(a_i | s)} \right|_{\pi=\pi^{(t)}} \right)
\]

where the calculation for the gradient in (5.3) can be found in \[8, 131, 277\].

Update rule (5.3) may suffer from a slow learning rate for some states. Since the gradient with respect to \(\pi_i(a_i | s)\) scales with \(d^\pi_\rho(s)\) – which may be small if the current policy \(\pi\) has small visitation frequency to \(s\) – the corresponding states may experience slow learning progress. To address this issue, we propose the following update rule (equivalent to (5.2) in Algorithm 10):

\[
\pi^{(t+1)}_i(\cdot | s) \leftarrow P_{\Delta(A_i)} \left( \pi^{(t)}_i(\cdot | s) + \eta \bar{Q}^\pi_i(s, a_i) \right)
\]

which essentially removes the \(d^\pi_\rho(s)/(1 - \gamma)\) factor in standard policy gradient (5.3) and alleviates the slow-learning issue. Interestingly, update rule (5.4) for the single-player MDP has also...
been studied in [247], concurrently. However, since the optimal value is not unique, the analysis of [247] does not apply to our multi-player case for which many Nash policies exist and the set that contains them is non-convex [131]. We also note that regularized variants of (5.4) for the single-player MDP appeared in [123, 270].

Furthermore, in contrast to (5.3), our update rule (5.4) is invariant to the initial state distribution \( \rho \). This allows us to establish performance guarantees simultaneously for all \( \rho \) in a similar way as typically done for natural policy gradient (NPG) and other policy mirror descent algorithms for single-player MDPs [8, 123, 270].

Theorem 32 establishes performance guarantees for Algorithm 10; see Section 5.5.2 for proof.

**Theorem 32 (Nash-Regret bound for Markov potential games) For MPG (5.1) with an initial state distribution \( \rho \), if all players independently perform the policy update in Algorithm 10 then, for two different choices of stepsize \( \eta \), we have**

\[
\text{Nash-Regret}(T) \lesssim \begin{cases} 
\sqrt{\kappa_{\rho}A}N^{2}C_{\Phi}^{2}T^{1/4}, & \eta = \frac{(1-\gamma)^{2}}{NA\sqrt{T}}, \\
\min(\kappa_{\rho},S)^{2}\sqrt{ANC_{\Phi}} \sqrt{T}, & \eta = \frac{(1-\gamma)^{4}}{8\min(\kappa_{\rho},S)^{3NA}}.
\end{cases}
\]

Depending on the stepsize \( \eta \), Theorem 32 provides two rates for the average Nash regret: \( T^{-1/4} \) and \( T^{-1/2} \). The technicalities behind these choices will be explained later and, to obtain an \( \epsilon \)-Nash equilibrium, our two bounds suggest respective iteration complexities,

\[
\frac{\kappa_{\rho}A}{(1-\gamma)^{9}}\epsilon^{4} \quad \text{and} \quad \frac{\min(\kappa_{\rho},S)^{4}AN}{(1-\gamma)^{6}}\epsilon^{2}.
\]
Compared with the iteration complexity guarantees in [131, 277], our bounds in Theorem 32 improve the dependence on the distribution mismatch coefficient $\kappa_\rho$ and the state space size $S = |S|$. Since our minimax distribution mismatch coefficient $\tilde{\kappa}_\rho$ satisfies

$$\tilde{\kappa}_\rho \leq \min(\kappa_\rho, S) \leq \kappa_\rho$$

our $\tilde{\kappa}_\rho$-dependence or $\min(\kappa_\rho, S)$-dependence are less restrictive than the explicit $S$-dependence in [131, 277]. Importantly, this permits our bounds to work for systems with large number of states, and makes Algorithm 10 suitable for sample-based scenario with function approximation (see Section 5.6). With polynomial dependence on the number of players $N$ instead of exponential, Algorithm 10 overcomes the curse of multiagents [111, 205]. In terms of problem parameters $(\gamma, A, N, C_\Phi)$, our iteration complexity either improves or becomes slightly worse.

**Remark 12 (Infinite state space)** When the state space is infinite, explicit $S$-dependence disappears in our iteration complexities. Implicit $S$-dependence only exists in the distribution mismatch coefficient $\kappa_\rho$ or $\tilde{\kappa}_\rho$. However, it is easy to bound $\kappa_\rho$ by devising an initial state distribution without introducing constraints on the MDP dynamics. For instance, in MPGs with agent-independent transitions (in which every state is a potential game and transitions do not depend on actions [131]), if we select $\rho$ to be the stationary state distribution $\rho_\pi$, then $\kappa_\rho = 1$ regardless of the state-space size $S$.

**Remark 13 (Our key techniques)** A key step of the analysis is to quantify the policy improvement regarding the potential function $\Phi$ in each iteration. Similar to the standard descent lemma in
optimization [15], applying the projected policy gradient algorithm to a smooth \( \Phi \) yields the following ascent property (cf. Eq. (9) in [131] and Lemmas 11 and 12 in [277]),

\[
\Phi_{\pi^{(t+1)}} (\mu) - \Phi_{\pi^{(t)}} (\mu) \geq \frac{1}{\beta} \sum_{i=1}^{N} \sum_{s} \| \pi_{i}^{(t+1)} (\cdot | s) - \pi_{i}^{(t)} (\cdot | s) \|^{2}
\]

where \( \beta > 0 \) is related to the smoothness constant (or the second-order derivative) of the potential function. However, since the search direction in our policy update is not the standard search direction utilized in policy gradient, this ascent analysis does not apply to our algorithm.

To obtain such improvement bound, it is crucial to analyze the joint policy improvement. Let us consider two players \( i \) and \( j \): player \( i \) changes its policy from \( \pi_{i} \) to \( \pi_{i}' \) to maximize its own reward based on the current policy profile \((\pi_{i}, \pi_{j})\) and player \( j \) changes its policy from \( \pi_{j} \) to \( \pi_{j}' \) in its own interest. What is the overall progress after they independently change their policies from \((\pi_{i}, \pi_{j})\) to \((\pi_{i}', \pi_{j}')\)? One method of capturing the joint policy improvement exploits the smoothness of the potential function, which is useful in the standard policy gradient ascent method [131, 277]. In our analysis, we connect the joint policy improvement with the individual policy improvement via the performance difference lemma. In particular, as shown in Lemma 38, Lemma 34 and Lemma 37 provide an effective means for analyzing the joint policy improvement. The proposed approach could be of independent interests for analyzing other Markov games.

In Lemma 38, we obtain two different joint policy improvement bounds by dealing with the cross terms in two different ways (see the proofs for details). Hence, we establish two different Nash-Regret bounds in Theorem 32: one has better dependence on \( T \) while the other has better dependence on \( \kappa_{\mu} \). Even though, it is an open issue how to achieve the best of the two, we next show that this is indeed possible for a special case: Markov cooperative games.
5.4.2 Faster rates for Markov cooperative games

When all players use the same reward function, i.e., \( r = r_i \) for all \( i = 1, \ldots, N \), MPG (5.1) reduces to a Markov cooperative game. In this case, \( V_i^\pi = V^\pi \) and \( Q_i^\pi = Q^\pi \) for all \( i = 1, \ldots, N \) and Algorithm 10 works immediately. Thus, we continue to use Nash-Regret \( i(T) \) that is defined through \( V_i^{\pi_i', \pi_t} = V_{\pi_t}^{\pi_t}, \pi_t \) and \( V_i^{\pi_t} = V^{\pi_t} \).

Algorithm 33 provides a Nash-Regret bound for Markov cooperative games; see Section 5.5.3 for proof.

**Theorem 33 (Nash-Regret bound for Markov cooperative games)** For MPG (5.1) with identical rewards and an initial state distribution \( \rho \), if all players independently perform the policy update in Algorithm 10 with stepsize \( \eta = (1 - \gamma)/(2NA) \) then,

\[
\text{Nash-Regret}(T) \lesssim \frac{\sqrt{\tilde{\kappa}_\rho AN}}{(1 - \gamma)^2 \sqrt{T}}.
\]

For Markov cooperative games, Theorem 33 achieves the best of the two bounds in Theorem 32 and an \( \epsilon \)-Nash equilibrium is achieved with the following iteration complexity,

\[
\tilde{\kappa}_\rho \frac{AN}{(1 - \gamma)^4 \epsilon^2}.
\]

This iteration complexity improves the ones provided in [131, 277] in several aspects. In particular, we have introduced the minimax distribution mismatch coefficient \( \tilde{\kappa}_\rho \), which is upper bounded by \( \kappa_\rho \). When we take this lower bound, our bound improves the \( \kappa_\rho \)-dependence in [131, 277] from \( \kappa_\rho^2 \) to \( \kappa_\rho \). We note that if we view the Markov cooperative game as an MPG, then the
value function $V^\pi$ serves as a potential function $\Phi$ which is bounded between 0 and $1/(1 - \gamma)$. Thus, our $(1 - \gamma)$-dependence matches the one in [277] and improves the one in [131] by $(1 - \gamma)^2$.

5.5 Nash regret analysis

In this section, we study Nash regret of Algorithm 10 for Markov potential games and Markov cooperative games. We present useful lemmas in Section 5.5.1 and provide the proof of Theorem 32 in Section 5.5.2 and the proof of Theorem 33 in Section 5.5.3.

5.5.1 Setting up the analysis

We first introduce a decomposition of the difference of multivariate functions, which is useful to decompose the difference of potential functions $\Phi^\pi(\mu)$ at two different policies for any state distribution $\mu$.

Let $\Psi^\pi : \Pi \to \mathbb{R}$ be any multivariate function mapping a policy $\pi \in \Pi$ to a real number. In Lemma 34, we show that the difference $\Psi^{\pi'} - \Psi^\pi$ at any two policies $\pi, \pi'$ equals to a sum of several partial differences. For $i, j \in \{1, \ldots, N\}$ with $i < j$, we denote by “$i \sim j$” the set of indices $\{k | i < k < j\}$, “$< i$” the set of indices $\{k | k = 1, \ldots, i - 1\}$, and “$> j$” the set of indices $\{k | k = j + 1, \ldots, N\}$. We use the shorthand $\pi_I := \{\pi_k\}_{k \in I}$ to represent the joint policy for all players $k \in I$. For example, when $I = i \sim j$, $\pi_I = \{\pi_k\}_{k=i+1}^{j-1}$ is a joint policy for players from $i + 1$ to $j - 1$; $\pi_{<i, i \sim j}$, $\pi_{<i}$, and $\pi_{> j}$ can be introduced similarly.
Lemma 34 (Multivariate function difference) For any function $\Psi: \Pi \to \mathbb{R}$, and any two policies $\pi, \pi' \in \Pi$,

$$\Psi^\pi - \Psi^\pi' = \sum_{i=1}^{N} \left( \Psi_{\pi', \pi_{-i}}^i - \Psi^\pi \right) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left( \Psi_{\pi_{<i, i\sim j}}, \pi_{>j}^i, \pi_j^j, \pi_i^j \right) - \Psi_{\pi_{<i, i\sim j}}, \pi_{>j}^i, \pi_i, \pi_j^j \right) + \Psi_{\pi_{<i, i\sim j}}, \pi_{>j}^i, \pi_i, \pi_j^j \right).$$

(5.5)

Proof. See Appendix D.1.

It is useful to introduce the following two difference bounds.

Lemma 35 (State-action value function difference) Suppose $i < j$ for $i, j = 1, \ldots, N$. Let $\tilde{\pi}_{-ij}$ be the policy for all players but $i, j$ and $\pi_i$ be the policy for player $i$. For any two policies for player $j$: $\pi_j$ and $\pi_j'$, we have

$$\max_s \| \bar{Q}_{\tilde{\pi}_{-ij}, \pi_i, \pi_j^j}^i (s, \cdot) - \bar{Q}_{\tilde{\pi}_{-ij}, \pi_i, \pi_j}^j (s, \cdot) \|_\infty \leq \frac{1}{(1 - \gamma)^2} \max_s \| \pi_j' (\cdot | s) - \pi_j (\cdot | s) \|_1.$$

Proof. See Appendix D.2.

Lemma 36 (Visitation measure difference) Let $\pi$ and $\pi'$ be two policies for an MDP, and $\mu$ be an initial state distribution. Then,

$$\sum_s \left| d_{\mu}^\pi (s) - d_{\mu}^\pi' (s) \right| \leq \max_s \| \pi (\cdot | s) - \pi' (\cdot | s) \|_1.$$

Proof. See Appendix D.3.
5.5.2 Nash regret analysis for Markov potential games

We first extend the 1st-order performance difference in Remark 11 to the 2nd-order performance difference, which is useful to measure the joint policy improvement from multiple players.

Lemma 37 (The 2nd-order performance difference) In a two-player common-reward Markov game with state space $S$ and action sets $A_1, A_2$, let $\Pi_1 = (\Delta(A_1))^{\|S\|}$ and $\Pi_2 = (\Delta(A_2))^{\|S\|}$ be player 1 and player 2’s policy sets, respectively. Then, for any $x, x' \in \Pi_1$ and $y, y' \in \Pi_2$,

\[
V^{x,y}(\mu) - V^{x',y}(\mu) - V^{x,y'}(\mu) + V^{x',y'}(\mu) \leq \frac{2\kappa^2 \mu^2 A^2}{(1 - \gamma)^4} \sum_s \|d^{x',y'}(s) (\|x(\cdot|s) - x'(\cdot|s)\|^2 + \|y(\cdot|s) - y'(\cdot|s)\|^2)\| \]

where $\kappa_{\mu}$ is the distribution mismatch coefficient relative to $\mu$.

Proof. See Appendix D.4.

We now apply Lemma 34 to the potential function $\Phi^\pi(\mu)$ at two consecutive policies $\pi^{(t+1)}$ and $\pi^{(t)}$ in Algorithm 10, where $\mu$ is an initial state distribution. We use the shorthand $\Phi^{(t)}(\mu)$ for $\Phi^{\pi^{(t)}}(\mu)$, the value of potential function at policy $\pi^{(t)}$.

Lemma 38 (Policy improvement: Markov potential games) For MPG (5.1) with any initial state distribution $\mu$, the difference of potential functions $\Phi^\pi(\mu)$ at two consecutive policies $\pi^{(t+1)}$ and $\pi^{(t)}$ in Algorithm 10, $\Phi^{(t+1)}(\mu) - \Phi^{(t)}(\mu)$ can be lower bounded by either (i) or (ii),

\[
\text{(i) } \frac{1}{2\eta(1 - \gamma)} \sum_{i=1}^N \sum_s d^{(t+1),\pi^{(t)}}_i (s) \left\| \pi^{(t+1)}_i (\cdot|s) - \pi^{(t)}_i (\cdot|s) \right\|^2 - \frac{4\eta^2 A^2 N^2}{(1 - \gamma)^5} \\
\text{(ii) } \frac{1}{2\eta(1 - \gamma)} \sum_{i=1}^N \sum_s d^{(t+1),\pi^{(t)}}_i (s) \left( 1 - \frac{4\eta\kappa_{\mu}^3 AN}{(1 - \gamma)^4} \right) \left\| \pi^{(t+1)}_i (\cdot|s) - \pi^{(t)}_i (\cdot|s) \right\|^2
\]
where \( \eta \) is the stepsize, \( N \) is the number of players, \( A \) is the size of one player's action space, and \( \kappa_\mu \) is the distribution mismatch coefficient relative to \( \mu \).

**Proof.** We let \( \pi' = \pi^{(t+1)} \) and \( \pi = \pi^{(t)} \) for brevity. By Lemma \ref{lem:diff} with \( \Psi_\pi = \Phi_\pi(\mu) \), it is equivalent to analyze

\[
\Phi^{(t+1)}(\mu) - \Phi^{(t)}(\mu) = \text{Diff}_\alpha + \text{Diff}_\beta
\]

(5.6)

\[
\text{Diff}_\alpha = \sum_{i=1}^{N} \left( \Phi_{\pi_{i}',\pi_{i}^{-i}}(\mu) - \Phi_{\pi}(\mu) \right)
\]

\[
\text{Diff}_\beta = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left( \Phi_{\pi_{<i,i\sim j},\pi_{>j}',\pi_{j}',\pi_{i}',\pi_{j}'}(\mu) - \Phi_{\pi_{<i,i\sim j},\pi_{>j}',\pi_{i}',\pi_{j}'}(\mu) \right. \\
\left. - \Phi_{\pi_{<i,i\sim j},\pi_{>j}',\pi_{i}',\pi_{j}'}(\mu) + \Phi_{\pi_{<i,i\sim j},\pi_{>j}',\pi_{i}',\pi_{j}'}(\mu) \right)
\]

**Bounding \text{Diff}_\alpha.** By the property of the potential function \( \Phi_{\pi}(\mu) \),

\[
\Phi_{\pi_{i}',\pi_{i}^{-i}}(\mu) - \Phi_{\pi}(\mu) = V_{\pi_{i}',\pi_{i}^{-i}}(\mu) - V_{\pi}(\mu) \\
= \frac{1}{1-\gamma} \sum_{s,a_{i}} d_{\mu_{i}}^{\pi_{i}',\pi_{i}^{-i}}(s) \left( \pi_{i}'(a_{i} \mid s) - \pi_{i}(a_{i} \mid s) \right) Q_{i}^{\pi_{i}',\pi_{i}^{-i}}(s,a_{i})
\]

(5.7)

where the second equality is due to the performance difference in Remark \ref{rem:performance} using \( \hat{\pi}_{i} = \pi_{i}' \) and \( \bar{\pi}_{i} = \pi_{i} \). The optimality of \( \pi_{i}' = \pi_{i}^{(t+1)} \) in line 4 of Algorithm \ref{alg:algorithm} leads to

\[
\langle \pi_{i}'(\cdot \mid s), \bar{Q}_{i}^{\pi_{i}',\pi_{i}^{-i}}(s, \cdot) \rangle_{A_{i}} - \frac{1}{2\eta} \| \pi_{i}'(\cdot \mid s) - \pi_{i}(\cdot \mid s) \|^2 \geq \langle \pi_{i}(\cdot \mid s), \bar{Q}_{i}^{\pi_{i},\pi_{i}^{-i}}(s, \cdot) \rangle_{A_{i}}.
\]

(5.8)

Combining (5.7) and (5.8) yields

\[
\Phi_{\pi_{i}',\pi_{i}^{-i}}(\mu) - \Phi_{\pi}(\mu) \geq \frac{1}{2\eta(1-\gamma)} \sum_{s} d_{\mu_{i}}^{\pi_{i}',\pi_{i}^{-i}}(s) \| \pi_{i}'(\cdot \mid s) - \pi_{i}(\cdot \mid s) \|^2.
\]
Therefore,
\[ \text{Diff}_\alpha \geq \frac{1}{2\eta(1 - \gamma)} \sum_{i=1}^{N} \sum_s d_{\mu}^{(t+1)} \| \pi_{i}^{(t)}(\cdot | s) - \pi_{i}^{(t+1)}(\cdot | s) \|^2. \] (5.9)

**Bounding \text{Diff}_\beta.** For simplicity, we denote \( \bar{\pi}_{-ij} \) as the joint policy of players \( N \setminus \{i, j\} \) where players \( < i \) and \( i \sim j \) use \( \pi \) and players \( > j \) use \( \pi' \). For each summand in \( \text{Diff}_\beta \),

\[
\Phi \bar{\pi}_{-ij}, \pi'_i, \pi'_j(\mu) - \Phi \bar{\pi}_{-ij}, \pi'_i, \pi_j(\mu) + \Phi \bar{\pi}_{-ij}, \pi_i, \pi'_j(\mu)
\]

\[ \overset{(a)}{=} V_i \bar{\pi}_{-ij}, \pi'_i, \pi'_j(\mu) - V_i \bar{\pi}_{-ij}, \pi_i, \pi_j(\mu) + V_i \bar{\pi}_{-ij}, \pi_i, \pi'_j(\mu)
\]

\[ \overset{(b)}{=} \frac{1}{1 - \gamma} \sum_{s, a_i} d_{\mu} \pi_{-ij}, \pi'_i, \pi'_j(s)(\pi'_i(a_i | s) - \pi_i(a_i | s)) Q_i \pi_{-ij}, \pi_i, \pi'_j(s, a_i)
\]

\[ - \frac{1}{1 - \gamma} \sum_{s, a_i} d_{\mu} \pi_{-ij}, \pi'_i, \pi_j(s)(\pi'_i(a_i | s) - \pi_i(a_i | s)) Q_i \pi_{-ij}, \pi_i, \pi_j(s, a_i)
\]

\[ \overset{(c)}{=} \frac{1}{1 - \gamma} \sum_{s} d_{\mu} \pi_{-ij}, \pi'_i, \pi'_j(s) || \pi'_i(\cdot | s) - \pi_i(\cdot | s) ||_1 Q_i \pi_{-ij}, \pi_i, \pi'_j(\cdot, \cdot) - Q_i \pi_{-ij}, \pi_i, \pi_j(\cdot, \cdot) ||_\infty
\]

\[ \overset{(d)}{\geq} \frac{1}{(1 - \gamma)^3} \left( \max_s || \pi'_i(\cdot | s) - \pi_i(\cdot | s) ||_1 \right)\left( \max_s || \pi'_j(\cdot | s) - \pi_j(\cdot | s) ||_1 \right)
\]

\[ \overset{(e)}{=} \frac{1}{(1 - \gamma)^2} \left( \max_s || \pi'_j(\cdot | s) - \pi_j(\cdot | s) ||_1 \right)\left( \max_s || \pi'_i(\cdot | s) - \pi_i(\cdot | s) ||_1 \right)
\]

\[ \overset{(f)}{\geq} \frac{8\eta^2 A^2}{(1 - \gamma)^5} \]

172
where (a) is due to the property of the potential function, (b) is due to the performance difference in Remark 11; for (c), we use Lemma 35, Lemma 36 and the fact that \( \sum_s d_{\mu, \pi_i, \pi_j}^\pi (s) = 1 \) and \( \| Q_i^\pi - \pi_i, \pi_j (s, \cdot) \|_\infty \leq \frac{1}{1 - \gamma} \); The last inequality (d) follows a direct result from the optimality of \( \pi_i' = \pi_i^{(t+1)} \) given by (5.8) and \( \| \cdot \| \leq \sqrt{A} \| \cdot \|_\infty \) and \( \| \cdot \|_1 \leq \sqrt{A} \| \cdot \| ).

\[
\| \pi_i' (\cdot, \cdot) - \pi_i (\cdot, \cdot) \|_2^2 \leq 2 \eta (\pi_i^{(t+1)} (\cdot, \cdot) - \pi_i (\cdot, \cdot), Q_i^\pi - \pi_i, \pi_j (s, \cdot) )_{\mathcal{A}_i} \\
\leq 2 \eta \| \pi_i^{(t+1)} (\cdot, \cdot) - \pi_i (\cdot, \cdot) \| \| Q_i^\pi - \pi_i, \pi_j (s, \cdot) \|
\]

and thus

\[
\| \pi_i^{(t+1)} (\cdot, \cdot) - \pi_i (\cdot, \cdot) \|_1 \leq 2 \eta \frac{\sqrt{A}}{1 - \gamma} \\
\| \pi_j^{(t+1)} (\cdot, \cdot) - \pi_j (\cdot, \cdot) \|_1 \leq 2 \eta A \frac{\sqrt{A}}{1 - \gamma}.
\]

Therefore,

\[
\text{Diff} \geq - \frac{N(N - 1)}{2} \times \frac{8 \eta^2 A^2}{(1 - \gamma)^5} \geq - \frac{4 \eta^2 A^2 N^2}{(1 - \gamma)^5}.
\]

We now complete the proof of (i) by combining (5.6), (5.9), and (5.10).

Alternatively, by Lemma 37, we can bound each summand of \text{Diff} by

\[
\Phi_i^{\pi, \pi_i', \pi_j'} (\mu) - \Phi_i^{\pi, \pi_i, \pi_j'} (\mu) - \Phi_i^{\pi, \pi_i', \pi_j} (\mu) + \Phi_i^{\pi, \pi_i, \pi_j} (\mu) = V_i^{\pi, \pi_i, \pi_j} (\mu) - V_i^{\pi, \pi_i', \pi_j'} (\mu) - V_i^{\pi, \pi_i', \pi_j} (\mu) + V_i^{\pi, \pi_i, \pi_j'} (\mu) \\
\geq - \frac{2 \eta^2 A}{(1 - \gamma)^4} \sum_s d_{\mu, \pi_i, \pi_j} (s) \left( \| \pi_i (\cdot, \cdot) - \pi_i' (\cdot, \cdot) \|^2 + \| \pi_j (\cdot, \cdot) - \pi_j' (\cdot, \cdot) \|^2 \right).
\]
Thus,

\[
\text{Diff}_\beta \geq - \frac{2\kappa^2 \mu A}{(1 - \gamma)^4} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \mu d_{\mu} \bar{\pi}_{i+j, \pi_i, \pi_j}(s) \left( \| \pi_i(\cdot | s) - \pi_i'(\cdot | s) \|^2 + \| \pi_j(\cdot | s) - \pi_j'(\cdot | s) \|^2 \right)
\]

\[
\geq - \frac{2\kappa^3 \mu N A}{(1 - \gamma)^5} \sum_{i=1}^{N} \sum_{s} \mu^{(t+1), \pi_i(t)}(s) \| \pi_i(t)(\cdot | s) - \pi_i'(t)(\cdot | s) \|^2
\]

where the last inequality is due to \( \frac{d_{\mu}^2(s)}{d_{\mu}^2(s)} \leq \frac{\kappa}{1 - \gamma} \) for any \( \pi, \pi', s \).

Combining the inequality above with (5.6) and (5.9) finishes the proof of (ii).

\[\square\]

**Proof.** [Theorem 32]

By the optimality of \( \pi_i(t+1) \) in line 4 of Algorithm 10,

\[
\langle \pi_i'(\cdot | s) - \pi_i(t+1)(\cdot | s), \eta Q_i(t)(s, \cdot) - \pi_i(t+1)(\cdot | s) + \pi_i(t)(\cdot | s) \rangle_{A_i} \leq 0, \text{ for any } \pi_i' \in \Pi_i.
\]

Hence, if \( \eta \leq \frac{1 - \gamma}{\sqrt{\lambda}} \), then for any \( \pi_i' \in \Pi_i \),

\[
\langle \pi_i'(\cdot | s) - \pi_i(t)(\cdot | s), \bar{Q}_i(t)(s, \cdot) \rangle_{A_i}
\]

\[
= \langle \pi_i'(\cdot | s) - \pi_i(t+1)(\cdot | s), \bar{Q}_i(t)(s, \cdot) \rangle_{A_i} + \langle \pi_i(t+1)(\cdot | s) - \pi_i(t)(\cdot | s), \bar{Q}_i(t)(s, \cdot) \rangle_{A_i}
\]

\[
\leq \frac{1}{\eta} \langle \pi_i'(\cdot | s) - \pi_i(t+1)(\cdot | s), \pi_i(t+1)(\cdot | s) - \pi_i(t)(\cdot | s) \rangle_{A_i}
\]

\[\leq \frac{2}{\eta} \\| \pi_i(t+1)(\cdot | s) - \pi_i(t)(\cdot | s) \| + \\| \pi_i(t+1)(\cdot | s) - \pi_i(t)(\cdot | s) \| \\| \bar{Q}_i(t)(s, \cdot) \|
\]

\[\leq \frac{3}{\eta} \\| \pi_i(t+1)(\cdot | s) - \pi_i(t)(\cdot | s) \|
\]
where in (a) we apply the Cauchy–Schwarz inequality and that \( \|p-p'\| \leq \|p-p'\|_1 \leq 2 \) for any two distributions \( p \) and \( p' \); (b) is because of \( \|Q_i(t)|s,\cdot\| \leq \frac{\sqrt{A}}{1-\gamma} \) and \( \eta \leq \frac{1-\gamma}{\sqrt{A}} \).

Therefore, for any initial distribution \( \rho \),

\[
\sum_{t=1}^{T} \max_{i} \left( \max_{\pi_i} V_{i}^{\pi_i,\pi_i(t)} (\rho) - V_{i}^{\pi_i(t)} (\rho) \right)
\]

\[\overset{(a)}{=} \frac{1}{1-\gamma} \sum_{t=1}^{T} \max_{\pi_i} \sum_{s,a_i} d_{\rho}^{\pi_i,\pi_i(t)} (s) \left( \pi_i'(a_i \mid s) - \pi_i^{(t)}(a_i \mid s) \right) \tilde{Q}_i^{(t)} (s, a_i) \]

\[\overset{(b)}{\leq} \frac{3}{\eta(1-\gamma)} \sum_{t=1}^{T} \sum_{s} d_{\rho}^{\pi_i,\pi_i(t)} (s) \left\| \pi_i^{(t+1)}(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s) \right\| \]

\[\overset{(c)}{\leq} \sqrt{\sup_{\pi \in \Pi} \frac{d_\rho^\pi}{\nu}} \sqrt{\frac{T}{\eta(1-\gamma)^{\frac{3}{2}}} \sum_{t=1}^{T} \sum_{s} d_{\rho}^{\pi_i,\pi_i(t)} (s) \sqrt{\sum_{t=1}^{T} \sum_{s} d_{\rho}^{\pi_i,\pi_i(t)} (s) \left( \pi_i^{(t+1)}(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s) \right)^2}} \]

\[\overset{(d)}{\leq} \sqrt{\sup_{\pi \in \Pi} \frac{d_\rho^\pi}{\nu}} \sqrt{T} \times \sqrt{\frac{T}{\eta(1-\gamma)^{\frac{3}{2}}} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{s} d_{\rho}^{\pi_i,\pi_i(t)} (s) \left( \pi_i^{(t+1)}(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s) \right)^2} \]

(5.11)

where (a) is due to the performance difference in Remark 11 and we slightly abuse the notation \( i \) to represent \( \arg\max_{\pi_i} \), in (b) we slightly abuse the notation \( \pi_i' \) to represent \( \arg\max_{\pi_i} \), in (c) we choose an arbitrary \( \nu \in \Delta(S) \) and use the following inequality:

\[
\frac{d_{\rho}^{\pi_i,\pi_i(t)} (s)}{d_{\nu}^{\pi_i,\pi_i(t)} (s)} \leq \frac{d_{\rho}^{\pi_i,\pi_i(t)} (s)}{(1-\gamma)\nu(s)} \leq \frac{\sup_{\pi \in \Pi} \|d_\rho^\pi\|_\infty}{1-\gamma}. \]

We apply the Cauchy–Schwarz inequality in (d), and finally we replace \( i \) (\( \arg\max_{\pi_i} \) in (a)) in the last square root term in (e) by the sum over all players.
If we proceed (5.11) with $\nu = \arg\min_{\nu \in \Delta(S)} \max_{\pi \in \Pi} \|d_\rho^\pi/\nu\|_\infty$, then,

$$
\sum_{t=1}^{T} \max_{i} \left( \max_{\pi'_t} V_{i}^{\pi'_t, \pi_{t-1}^i}(\rho) - V_i^{\pi(t)}(\rho) \right)
\leq \sqrt{\frac{\kappa_\rho}{\eta(1-\gamma)^{\frac{3}{2}}} \sqrt{T} \times \sqrt{2\eta(1-\gamma) (\Phi(T+1)(\nu) - \Phi(1)(\nu)) + \frac{4\eta^3 A^2 N^2}{(1-\gamma)^4} T}}
\leq \sqrt{\frac{\kappa_\rho T \sqrt{C_\Phi}}{\eta(1-\gamma)^2}} + \sqrt{\frac{\kappa_\rho T^2 A^2 N^2}{(1-\gamma)^7}}
$$

where in (a) we apply the first bound (i) in Lemma 38 (with $\mu = \nu$) and use Definition 3 $\tilde{\kappa}_\rho = \min_{\nu \in \Delta(S)} \max_{\pi \in \Pi} \|d_\rho^\pi/\nu\|_\infty$, and in (b) we use $|\Phi^\pi(\nu) - \Phi^{\pi'}(\nu)| \leq C_\Phi$ for any $\pi, \pi'$, and further simplify the bound in (b). We complete the proof for the first bound by taking stepsize $\eta = \frac{(1-\gamma)^2\sqrt{C_\Phi}}{N\sqrt{A}}$ (by the upper bound of $C_\Phi$ given in Lemma 68, the condition $\eta \leq \frac{1-\gamma}{\sqrt{A}}$ is satisfied).

If we proceed (5.11) with the second bound (ii) in Lemma 38 with the choice of $\eta \leq \frac{(1-\gamma)^4}{8\kappa_\rho^4 N A}$, then,

$$
\sum_{t=1}^{T} \max_{i} \left( \max_{\pi'_t} V_{i}^{\pi'_t, \pi_{t-1}^i}(\rho) - V_i^{\pi(t)}(\rho) \right)
\leq \sqrt{\sup_{\pi \in \Pi} \|d_\rho^\pi/\nu\|_\infty} \frac{\sqrt{\kappa_\rho T C_\Phi}}{\eta(1-\gamma)^2}
\leq \sqrt{\sup_{\pi \in \Pi} \|d_\rho^\pi/\nu\|_\infty} \frac{T \sqrt{C_\Phi}}{\eta(1-\gamma)^2}.
$$

We next discuss two special choices of $\nu$ for proving our bound. First, if $\nu = \rho$, then $\eta \leq \frac{(1-\gamma)^4}{8\kappa_\rho^4 N A}$.

By letting $\eta = \frac{(1-\gamma)^4}{8\kappa_\rho^4 N A}$, the last square root term can be bounded by $O\left(\sqrt{\frac{\kappa_\rho^4 N A T C_\Phi}{(1-\gamma)^6}}\right)$. Second, if $\nu = \text{Unif}_S$, the uniform distribution over $S$, then $\kappa_\nu \leq \frac{1}{S}$, which allows a valid choice $\eta = \frac{(1-\gamma)^4}{8S^4 N A} \leq \frac{(1-\gamma)^4}{8\kappa_\rho^4 N A}$. Hence, we can bound the last square root term by $O\left(\sqrt{\frac{S^4 N A T C_\Phi}{(1-\gamma)^6}}\right)$. Since $\nu$ is arbitrary, combining these two special choices completes the proof.
5.5.3 Nash regret analysis for Markov cooperative games

We first establish policy improvement regarding the state-action value function at two consecutive policies \( \pi^{(t+1)} \) and \( \pi^{(t)} \) in Algorithm 10.

**Lemma 39 (Policy improvement: Markov cooperative games)** For MPG (5.1) with identical rewards and an initial state distribution \( \rho > 0 \), if all players independently perform the policy update in Algorithm 10 with stepsize \( \eta \leq \frac{1 - \gamma}{2N} \), then for any \( t \) and any \( s \),

\[
\mathbb{E}_{a \sim \pi^{(t+1)}(\cdot | s)} \left[ Q^{(t)}(s, a) \right] - \mathbb{E}_{a \sim \pi^{(t)}(\cdot | s)} \left[ Q^{(t)}(s, a) \right] \geq \frac{1}{4\eta} \sum_{i=1}^{N} \left\| \pi^{(t+1)}_{i}(\cdot | s) - \pi^{(t)}_{i}(\cdot | s) \right\|^{2}
\]

where \( \eta \) is the stepsize and \( N \) is the number of players.

**Proof.** Fixing the time \( t \) and the state \( s \), we apply 34 to \( \Psi^{\pi} = \mathbb{E}_{a \sim \pi(\cdot | s)} \left[ Q^{(t)}(s, a) \right] \), where \( Q^{(t)} := Q^{\pi^{(t)}} \) (recall that \( \pi \) is a joint policy of all players). By Lemma 34 for any two policies \( \pi' \) and \( \pi \),

\[
\mathbb{E}_{a \sim \pi'(\cdot | s)} \left[ Q^{(t)}(s, a) \right] - \mathbb{E}_{a \sim \pi(\cdot | s)} \left[ Q^{(t)}(s, a) \right] = \sum_{i=1}^{N} \left( \mathbb{E}_{a_i \sim \pi_i'(\cdot | s), a_{-i} \sim \pi_{-i}(\cdot | s)} \left[ Q^{(t)}(s, a) \right] - \mathbb{E}_{a_i \sim \pi_i(\cdot | s)} \left[ Q^{(t)}(s, a) \right] \right)
\]
\[
+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left( \mathbb{E}_{a_i \sim \pi_i'(\cdot | s), a_j \sim \pi_j'\sim \pi_i'(\cdot | s), a_{-ij} \sim \pi_{-ij}(\cdot | s)} \left[ Q^{(t)}(s, a) \right] - \mathbb{E}_{a_i \sim \pi_i(\cdot | s), a_j \sim \pi_j(\cdot | s), a_{-ij} \sim \pi_{-ij}(\cdot | s)} \left[ Q^{(t)}(s, a) \right] \right)
\]

(5.12)
where $\pi_{-ij}$ is a joint policy of players $N\setminus\{i, j\}$ in which players $< i$ and $i \sim j$ use $\pi$, and players $> j$ use $\pi'$. Particularly, we choose $\pi' = \pi^{(t+1)}$ and $\pi = \pi^{(t)}$. Thus, we can reduce (5.12) into

$$
\mathbb{E}_{a \sim \pi'(|s)} \left[ Q^{(t)}(s, a) \right] - \mathbb{E}_{a \sim \pi(|s)} \left[ Q^{(t)}(s, a) \right] 
\begin{align*}
&= \sum_{i=1}^{N} \sum_{a_i} (\pi'_i(a_i | s) - \pi_i(a_i | s)) Q^{(t)}_i(s, a_i) \\
&\quad + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{a_i, a_j} (\pi'_i(a_i | s) - \pi_i(a_i | s)) (\pi'_j(a_j | s) - \pi_j(a_j | s)) \mathbb{E}_{a_{-ij} \sim \pi_{-ij}(|s)} \left[ Q^{(t)}(s, a) \right] \\
&\geq (a) \sum_{i=1}^{N} \frac{1}{2\eta} \left\| \pi'_i(\cdot | s) - \pi_i(\cdot | s) \right\|^2 \\
&\quad - \frac{1}{1-\gamma} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{a_i, a_j} \left| \pi'_i(a_i | s) - \pi_i(a_i | s) \right| \left| \pi'_j(a_j | s) - \pi_j(a_j | s) \right| \\
&\geq (b) \sum_{i=1}^{N} \frac{1}{2\eta} \left\| \pi'_i(\cdot | s) - \pi_i(\cdot | s) \right\|^2 \\
&\quad - \frac{A}{2(1-\gamma)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left( \left\| \pi'_i(\cdot | s) - \pi_i(\cdot | s) \right\|^2 + \left\| \pi'_j(\cdot | s) - \pi_j(\cdot | s) \right\|^2 \right) \\
&= \sum_{i=1}^{N} \frac{1}{2\eta} \left\| \pi'_i(\cdot | s) - \pi_i(\cdot | s) \right\|^2 - \frac{(N-1)A}{2(1-\gamma)} \sum_{i=1}^{N} \left\| \pi'_i(\cdot | s) - \pi_i(\cdot | s) \right\|^2 \\
&\geq (c) \sum_{i=1}^{N} \frac{1}{4\eta} \left\| \pi'_i(\cdot | s) - \pi_i(\cdot | s) \right\|^2
\end{align*}
$$

where $(a)$ is due to the optimality condition (5.8) and $Q^{(t)}(s, a) \leq \frac{1}{1-\gamma}$, $(b)$ is due to $\langle x, y \rangle \leq \frac{\|x\|^2 + \|y\|^2}{2}$, and $(c)$ follows the choice of $\eta \leq \frac{1-\gamma}{2NA}$.

**Proof.** [Proof of Theorem 33]

By the performance difference in Remark 11 and Lemma 39, we have for any $\nu \in \Delta(S)$,

$$
V^{(t+1)}(\nu) - V^{(t)}(\nu) = \frac{1}{1-\gamma} \sum_{s, a} d^{(t+1)}(s) \left( \pi^{(t+1)}(a | s) - \pi^{(t)}(a | s) \right) Q^{(t)}(s, a) \\
\quad \geq \frac{1}{4\eta(1-\gamma)} \sum_{i=1}^{N} \sum_{s} d^{(t+1)}(s) \left\| \pi^{(t+1)}_i(\cdot | s) - \pi^{(t)}_i(\cdot | s) \right\|^2.
$$

(5.13)
By the same argument as the proof of Theorem 32,

\[
\sum_{t=1}^{T} \max_i \left( \max_{\pi_i} V_{\pi_i, \pi_{i-1}}^{(t)}(\rho) - V_{\pi_i}^{(t)}(\rho) \right)
\]

\[(a)\] \[\frac{3}{\eta(1 - \gamma)} \sum_{t=1}^{T} \sum_s d_{\rho, \pi_i}^{(t)}(s) \left\| \pi_{i+1}(s) - \pi_i(t) \right\|\]

\[(b)\] \[\frac{\sqrt{\kappa_{\rho}}}{\eta(1 - \gamma)^{3/2}} \sum_{t=1}^{T} \sum_s \sqrt{d_{\rho, \pi_i}^{(t)}(s) \times d_{\nu}^{(t+1)}(s)} \left\| \pi_{t+1}(\cdot | s) - \pi_i(t) \right\|\]

\[(c)\] \[\frac{\sqrt{\kappa_{\rho}}}{\eta(1 - \gamma)^{3/2}} \sqrt{T} \times \left[ \sum_{t=1}^{T} \sum_{i=1}^{N} d_{\rho, \pi_i}^{(t+1)}(s) \left\| \pi_{t+1}(\cdot | s) - \pi_i(t) \right\|^2 \right]\]

\[(d)\] \[\frac{\sqrt{\kappa_{\rho}}}{\eta(1 - \gamma)^{3/2}} \sqrt{T} \times \sqrt{4\eta(1 - \gamma) \left( V^{(T+1)}(\nu) - V^{(1)}(\nu) \right)}\]

where in \(a\) we slightly abuse the notation \(i\) to represent \(\arg\max_i\) as in (5.11), in \(b\) we take \(\nu = \arg\min_{\nu \in \Delta(S)} \max_{\pi \in \Pi} \|d_{\rho, \pi}^\nu\|_\infty\) and use the definition of \(\kappa_{\rho}\) from Definition 3, and we replace \(i\) (\(\arg\max_i\) in \(a\)) in the last square root term in \(c\) by the sum over all players, and we apply (5.13) in \(d\).

Finally, we complete the proof by taking stepsize \(\eta = \frac{1 - \gamma}{2N\lambda} \) and using \(V^{(T+1)}(\nu) - V^{(1)}(\nu) \leq \frac{1}{1 - \gamma}\). \(\square\)
5.6 Independent policy gradient with function approximation

We next remove the exact gradient requirement and apply Algorithm 10 to the linear function approximation setting. In what follows, we assume that the averaged action value function is linear in a given feature map.

Algorithm 11 Independent policy gradient with linear function approximation

1: Parameters: $K$, $W$, and $\eta > 0$.
2: Initialization: Let $\pi^{(1)}_i(a_i | s) = 1/A$ for $s \in S$, $a_i \in A_i$ and $i = 1, \ldots, N$.
3: for step $t = 1, \ldots, T$ do
4: // Phase 1 (data collection)
5: for round $k = 1, \ldots, K$ do
6: For each $i \in [N]$, sample $h_i \sim \text{GEOMETRIC}(1 - \gamma)$ and $h'_i \sim \text{GEOMETRIC}(1 - \gamma)$.
7: Draw an initial state $\bar{s}^{(0)} \sim \rho$.
8: Continuing from $\bar{s}^{(0)}$, let all players interact with each other using $\{\pi^{(t)}_i\}_{i=1}^N$ for $H = \max_i(h_i + h'_i)$ steps, which generates a state-joint-action-reward trajectory $\bar{s}^{(0)}, \bar{a}^{(0)}, \bar{r}^{(0)}, \bar{s}^{(1)}, \bar{a}^{(1)}, \bar{r}^{(1)}, \ldots, \bar{s}^{(H)}, \bar{a}^{(H)}, \bar{r}^{(H)}$.
9: Define for every player $i \in [N],$

\[
\begin{align*}
  s^{(k)}_i &= \bar{s}^{(h_i)}, \\
  a^{(k)}_i &= \bar{a}^{(h_i)}_i, \\
  R^{(k)}_i &= \sum_{h=h_i}^{h_i+h'_i-1} \bar{r}^{(h)}_i.
\end{align*}
\]

(5.14)

10: end for
11: // Phase 2 (policy update)
12: for player $i = 1, \ldots, N$ (in parallel) do
13: Compute $\hat{w}^{(t)}_i$ as

\[
\hat{w}^{(t)}_i \approx \arg \min_{||w_i|| \leq W} \sum_{k=1}^{K} \left( R^{(k)}_i - \langle \phi_i(s^{(k)}_i, a^{(k)}_i), w_i \rangle \right)^2.
\]

(5.15)

14: Define $\hat{Q}^{(t)}_i(s, \cdot) := \langle \phi_i(s, \cdot), \hat{w}^{(t)}_i \rangle$ and player $i$’s policy for $s \in S,$

\[
\pi^{(t+1)}_i(\cdot | s) = \arg \max_{\pi_i(\cdot | s) \in \Delta(A_i)} \left\{ \langle \pi_i(\cdot | s), \hat{Q}^{(t)}_i(s, \cdot) \rangle_{A_i} - \frac{1}{2\eta} \| \pi_i(\cdot | s) - \pi^{(t)}_i(\cdot | s) \|^2 \right\}.
\]

(5.16)

15: end for
16: end for
Assumption 15 (Linear averaged \(Q\)) In MPG (5.1), for each player \(i\), there is a feature map \(\phi_i : S \times A_i \to \mathbb{R}^d\), such that for any \((s, a_i) \in S \times A_i\) and any policy \(\pi \in \Pi\),

\[
\bar{Q}_i^\pi(s, a_i) = \langle \phi_i(s, a_i), w_i^\pi \rangle, \text{ for some } w_i^\pi \in \mathbb{R}^d.
\]

Moreover, \(\|\phi_i\| \leq 1\) for all \(s, a_i\), and \(\|w_i^\pi\| \leq W\) for all \(\pi\).

Without loss of generality, we can assume \(W \leq \sqrt{d}/(1 - \gamma)\); see Lemma 8 in [240]. Assumption 15 is a multi-agent generalization of the standard linear \(Q\) assumption [4] for single-player MDPs. It is different from the multi-agent linear MDP assumption [250, 81] in which both transition and reward functions are linear in given feature maps. In contrast, Assumption 15 qualifies each player to estimate its averaged action value function without observing other players’ actions. A special case of Assumption 15 is the tabular case in which the sizes of state/action spaces are finite, and where we can select \(\phi_i\) to be an indicator function. Since the feature map \(\phi_i\) is locally-defined coordination between players is avoided [282].

Remark 14 (Function approximation) Since RL with function approximation is often statistically hard, e.g., see [244, 236] for hardness results, assuming regularity of underlying MDPs is necessary for the application of function approximation to multi-agent RL in which either the value function [250, 81, 107, 104] or the policy [282] is approximated. Because of restrictive function approximation power, the main challenge is the entanglement of policy improvement (or optimization) and policy evaluation (or approximation) errors. In Theorem 40 and Theorem 41 we show that optimization and approximation errors are decoupled under Assumption 15 so that we can control them, separately. Our analysis can be generalized to some neural networks, e.g., overparametrized neural
networks [139], a rich function class that allows splitting optimization and approximation errors, which we leave for future work.

We formally present our algorithm in Algorithm 11. At each step $t$, there are two phases. In Phase 1, the players begin with the initial state $\bar{s}^{(0)} \sim \rho$ and simultaneously execute their current policies $\{\pi_i^{(t)}\}_{i=1}^N$ to interact with the environment for $K$ rounds. In each round $k$, we terminate the interaction at step $H = \max_i (h_i + h'_i)$, where $h_i$ and $h'_i$ are sampled from a geometric distribution $\text{GEOMETRIC}(1 - \gamma)$, independently; the state at $h_i$ naturally follows $\bar{s}^{(h_i)} \sim d_{\pi_i^{(t)}}$. By collecting rewards from step $h_i$ to $h_i + h'_i - 1$, as shown in (5.14), we can justify $\mathbb{E}[R_i^{(k)}] = \bar{Q}_i^{(t)}(\bar{s}^{(h_i)}, \bar{a}_i^{(h_i)})$ where $\bar{Q}_i^{(t)}(\cdot, \cdot) := \bar{Q}^{\pi_i^{(t)}}(\cdot, \cdot)$ and $\bar{a}_i^{(h_i)} \sim \pi_i^{(t)}(\cdot | \bar{s}^{(h_i)})$, in Appendix D.5.1 In the end of round $k$, we collect a sample tuple: $(s_i^{(k)}, a_i^{(k)}, R_i^{(k)})$ in (5.14) for each player $i$.

After each player collects $K$ samples, in Phase 2, they use these samples to estimate $\bar{Q}_i^{(t)}(\cdot, \cdot)$, which is required for policy updates. By Assumption 15

$$ Q_i^{(t)}(s, a_i) = \langle \phi_i(s, a_i), w_i^{(t)} \rangle, \forall (s, a_i) \in \mathcal{S} \times \mathcal{A}_i $$

where $w_i^{(t)}$ represents $w_i^{\pi_i^{(t)}}$. Our goal is to obtain a solution $\hat{w}_i^{(t)} \approx w_i^{(t)}$ using samples, and estimate $\bar{Q}_i^{(t)}(s, a_i)$ via

$$ \hat{Q}_i^{(t)}(s, a_i) := \langle \phi_i(s, a_i), \hat{w}_i^{(t)} \rangle, \forall (s, a_i) \in \mathcal{S} \times \mathcal{A}_i. \quad (5.17) $$

We note that $\mathbb{E}[R_i^{(k)}] = Q_i^{(t)}(s_i^{(k)}, a_i^{(k)}) = \langle \phi_i(s_i^{(k)}, a_i^{(k)}), w_i^{(t)} \rangle$. To obtain $\hat{w}_i^{(t)}$, the standard approach is to solve linear regression (5.15).
We measure the estimation quality of $\hat{w}_t^i$ via the expected regression loss,

$$L_t^i(w_i) = \mathbb{E}_{(s,a_i) \sim \nu_t^i} \left[ \left( \bar{Q}_t^i(s,a_i) - \langle \phi_i(s,a_i), w_i \rangle \right)^2 \right]$$

where $\nu_t^i(s,a_i) := \frac{d_t^i(s) \circ \pi_t^i(a_i | s)}{\rho(s) \circ \pi_t^i(a_i | s)}$ and $L_t^i(w_t^i) = 0$ by Assumption 15. We make the following assumption for the expected regression loss of $\hat{w}_t^i$.

**Assumption 16 (Bounded statistical error)** Fix a state distribution $\rho$. For any sequence of iterates $\hat{w}_t^1, \ldots, \hat{w}_t^N$ for $i = 1, \ldots, N$ that are generated by Algorithm 11, there exists an $\epsilon_{\text{stat}} < \infty$ such that

$$\mathbb{E}[L_t^i(\hat{w}_t^i)] \leq \epsilon_{\text{stat}}$$

for all $i$ and $t$, where the expectation is on randomness in generating $\hat{w}_t^i$.

The bound for $\epsilon_{\text{stat}}$ can be established using standard linear regression analysis and it is given by $\epsilon_{\text{stat}} = O\left( \frac{dW^2}{K(1-\gamma)^2} \right)$. This bound can be achieved by applying the stochastic projected gradient descent method [101, 58] to the regression problem.

We next show how the approximation error affect the convergence. We take an expectation over the randomness of approximately computing $\hat{w}_t^i$ as Assumption 16.

After obtaining $\hat{Q}_t^i(\cdot, \cdot)$, we update the polices in (5.16) which is different from the update in Algorithm 10 in two aspects: (i) the gradient direction $\hat{Q}_t^i(\cdot, \cdot)$ is the estimated version of $\bar{Q}_t^i(\cdot, \cdot)$; and (ii) the Euclidean projection set becomes $\Delta_\xi(A_i) := \left\{ (1 - \xi) \pi_i(\cdot | s) + \xi \text{Unif}_{A_i}, \forall \pi_i(\cdot | s) \right\}$ that introduces $\xi$-greedy policies for exploration [131, 277], where $\xi \in (0, 1)$.

Theorem 40 establishes performance guarantees for Algorithm 11 see Appendix D.5.2 for proof.
Theorem 40 (Nash-Regret bound for Markov potential games) Let Assumption 15 hold for MPG (5.1) with an initial state distribution $\rho$. If all players independently run Algorithm 11 with $
abla = \min \left( \left( \frac{\kappa_\rho^2 N A \epsilon_{\text{stat}}}{(1 - \gamma)^2 W^2} \right)^{\frac{1}{3}}, \frac{1}{2} \right)$ and Assumption 16 holds, then

$$E \left[ \text{Nash-Regret}(T) \right] \lesssim \mathcal{R}(\eta) + \left( \frac{\kappa_\rho^2 W A N \epsilon_{\text{stat}}}{(1 - \gamma)^5} \right)^{\frac{1}{3}}$$

$$\mathcal{R}(\eta) = \begin{cases} \sqrt{\kappa_\rho W N (AC_\Phi)^{\frac{1}{2}}} & \eta = \frac{(1 - \gamma)^{\frac{1}{2}} \sqrt{C_\Phi}}{WN \sqrt{AT}} \\ \frac{\kappa_\rho \sqrt{A N C_\Phi}}{(1 - \gamma)^3 \sqrt{T}} & \eta = \frac{(1 - \gamma)^{\frac{4}{16}} \kappa_\rho^3 N A}{16} \end{cases}$$

Theorem 40 shows the additive effect of the function approximation error $\epsilon_{\text{stat}}$ on the Nash regret of Algorithm 11. When $\epsilon_{\text{stat}} = 0$, Theorem 40 matches the rates in Theorem 32 in the exact gradient case. As in Algorithm 10, even though update rule (5.16) iterates over all $s \in S$, we do not need to assume a finite state space $S$. In fact, (5.16) only “defines” a function $\pi_i(t) (\cdot | s)$ instead of “calculating” it. This is commonly used in policy optimization with function approximation, e.g., [47, 146]. To execute this algorithm, $\pi_i(t) (\cdot | s)$ only needs to be evaluated if necessary, e.g., when the state $s$ is visited in Phase 1 of Algorithm 11.

When we apply stochastic projected gradient updates to (5.15), Algorithm 11 becomes a sample-based algorithm and existing stochastic projected gradient results directly apply. Depending on the stepsize choice, an $\epsilon$-Nash equilibrium is achieved with sample complexities (see Corollary 1 in Appendix D.5.4),

$$TK = O \left( \frac{1}{\epsilon^7} \right) \text{ and } O \left( \frac{1}{\epsilon^5} \right)$$

Compared with the sample complexity guarantees for the tabular MPG case [131, 277], our sample complexity guarantees hold for MPGs with potentially infinitely large state spaces. When we specialize Assumption 15 to the tabular case, our second sample complexity improves the sample complexity in [131, 277] from $O(1/\epsilon^6)$ to $O(1/\epsilon^5)$.

As before, we get improved performance guarantees when we apply Algorithm 11 to Markov cooperative games.

**Theorem 41 (Nash-Regret bound for Markov cooperative games)** Let Assumption 15 hold for MPG (5.1) with identical rewards and an initial state distribution $\rho > 0$. If all players independently perform the policy update in Algorithm 11 with stepsize $\eta = (1 - \gamma)/(2NA)$ and exploration rate $\xi = \min\left(\left(\frac{\kappa^2 NA\epsilon_{stat}}{1 - \gamma}\right)^\frac{1}{2}, \frac{1}{2}\right)$, with Assumption 16,

$\mathbb{E}[\text{Nash-Regret}(T)] \lesssim \frac{\sqrt{\kappa^2 AN}}{(1 - \gamma)^2 \sqrt{T}} + \left(\frac{\kappa^2 W AN\epsilon_{stat}}{(1 - \gamma)^5}\right)^\frac{1}{2}$.

We prove Theorem 41 in Appendix D.5.3 and show sample complexity $TK = O(1/\epsilon^5)$ in Corollary 2 of Appendix D.5.4.

### 5.7 Game-agnostic convergence

In Section 5.4 and Section 5.6, we have shown that our independent policy gradient method converges (in best-iterate sense) to a Nash equilibrium of MPGs. For the same algorithm in two-player case, however, [24] showed that players’ policies can diverge for zero-sum matrix games (a single-state case of zero-sum Markov games). A natural question arises:
Does there exist a simple gradient-based algorithm that provably converges to a Nash equilibrium in both potential/cooperative and zero-sum games?

Unfortunately, classical MWU and optimistic MWU updates do not converge to a Nash equilibrium in zero-sum and coordination games simultaneously [53]. Recently, this question was partially answered by [130, 129] in which the authors established last-iterate convergence of Q-learning dynamics to a quantal response equilibrium for both zero-sum and potential/cooperative matrix games. In this work, we provide an affirmative answer to this question for general Markov games that cover matrix games. Specifically, we next show that optimistic gradient descent/ascent with a smoothed critic (see Algorithm 12) – an algorithm that converges to a Nash equilibrium in two-player zero-sum Markov games [239] – also converges to a Nash equilibrium in Markov cooperative games.

We now setup notation for tabular two-player Markov cooperative games with $N = 2$, $r = r_1 = r_2$, $A = |A_1| = |A_2|$, and $S = |S|$. For convenience, we use $x_s \in \mathbb{R}^A$ and $y_s \in \mathbb{R}^A$ to denote policies $\pi_1(\cdot \mid s)$ and $\pi_2(\cdot \mid s)$ taken at state $s \in S$, and $Q^\pi_s \in \mathbb{R}^{A \times A}$ to denote $Q^\pi(s, a_1, a_2)$ with $a_1 \in A_1$ and $a_2 \in A_2$. We describe our policy update (5.18) in Algorithm 12: the next iterate $(\bar{x}_s^{(t+1)}, \bar{y}_s^{(t+1)})$ is obtained from two steps of policy gradient ascent with an intermediate iterate $(\bar{x}_s^{(t+1)}, \bar{y}_s^{(t+1)})$. Motivated by [239], we introduce a critic $Q^{(t)}_s$ to learn the value function at each state $s$ using the learning rate $\alpha^{(t)}$. When the critic is ideal, i.e., $Q^{(t)}_s = Q^\pi_s$, where $Q^\pi_s$ is a matrix form of $Q^\pi(s, a_1, a_2)$ for $a_1 \in A_1$ and $a_2 \in A_2$, we can view Algorithm 12 as a two-player case of Algorithm 10.

In Theorem 42, we establish asymptotic last-iterate convergence of Algorithm 12 in Markov cooperative games; see Appendix D.6.1 for proof.
Algorithm 12 Independent optimistic policy gradient ascent

1: **Parameters:** $0 < \eta \leq \frac{1-\gamma}{32\sqrt{A}}$ and a non-increasing sequence $\{\alpha(t)\}_{t=1}^{\infty}$ that satisfies

$$0 < \alpha(t) \leq \frac{1}{6} \quad \text{for all } t \quad \text{and} \quad \sum_{t=t'}^{\infty} \alpha(t) = \infty \quad \text{for any } t'.$$

2: **Initialization:** Let $x_s^{(1)} = \bar{x}_s^{(1)} = y_s^{(1)} = \bar{y}_s^{(1)} = 1/A$ and $\mathcal{V}_s^{(0)} = 0$ for all $s \in S$.

3: **for** step $t = 1, 2, \ldots$ **do**

4: Define $Q^{(t)}_s \in \mathbb{R}^{A \times A}$ for all $s \in S$,

$$Q^{(t)}_s(a_1, a_2) = r(s, a_1, a_2) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot \mid s, a_1, a_2)} \left[ \mathcal{V}^{(t-1)}_{s'} \right].$$

5: Define two players’ policies for $s \in S$,

$$\bar{x}_s^{(t+1)} = \arg\max_{x_s \in \Delta(A_1)} \left\{ x_s^\top Q^{(t)}_s y_s - \frac{1}{2\eta} \| x_s - \bar{x}_s^{(t)} \|^2 \right\},$$

$$x_s^{(t+1)} = \arg\max_{x_s \in \Delta(A_1)} \left\{ x_s^\top Q^{(t)}_s y_s - \frac{1}{2\eta} \| x_s - \bar{x}_s^{(t+1)} \|^2 \right\},$$

$$\bar{y}_s^{(t+1)} = \arg\max_{y_s \in \Delta(A_2)} \left\{ (x_s^{(t)})^\top Q^{(t)}_s y_s - \frac{1}{2\eta} \| y_s - \bar{y}_s^{(t)} \|^2 \right\},$$

$$y_s^{(t+1)} = \arg\max_{y_s \in \Delta(A_2)} \left\{ (x_s^{(t)})^\top Q^{(t)}_s y_s - \frac{1}{2\eta} \| y_s - \bar{y}_s^{(t+1)} \|^2 \right\} \quad (5.18)$$

$$\mathcal{V}_s^{(t)} = (1 - \alpha(t))\mathcal{V}_s^{(t-1)} + \alpha(t)(x_s^{(t)})^\top Q^{(t)}_s y_s^{(t)}.$$  

6: **end for**

Theorem 42 (Last-iterate convergence for two-player Markov cooperative games) For

MPG [5.1] with two players and identical rewards, if both players run Algorithm 12 with $0 < \eta < (1 - \gamma)/(32\sqrt{A})$ and a non-increasing $\{\alpha(t)\}_{t=1}^{\infty}$ that satisfies $0 < \alpha(t) < 1/6$ and $\sum_{t=t'}^{\infty} \alpha(t) = \infty$ for any $t' \geq 0$, then the policy pair $(x^{(t)}, y^{(t)})$ converges to a Nash equilibrium when $t \to \infty$.

Last-iterate convergence in Theorem 42 is measured by the local gaps $\max_{x'} (V^{x', y^{(t)}}(\rho) - V^{x^{(t)}, y^{(t)}}(\rho))$ and $\max_{y'} (V^{x^{(t)}, y'}(\rho) - V^{x^{(t)}, y^{(t)}}(\rho))$, i.e., a policy pair $(x^{(t)}, y^{(t)})$ constitutes an approximate Nash policy for large $t$. The condition on algorithm parameters $\eta$ and $\alpha(t)$ in Theorem 42 is mild in sense that it is straightforward to take a pair of such parameters that ensures
last-iterate convergence in zero-sum Markov games \cite{239}. Hence, Algorithm \texttt{12} enjoys last-
iterate convergence in both two-player Markov cooperative and zero-sum competitive games.

Compared with the result \cite{92}, our proof of Theorem \texttt{42} utilizes gap convergence instead of
point-wise policy convergence that is restricted to isolated /fixed points of the algorithm

dynamics. Moreover, our algorithm works for both cooperative and competitive Markov games.

In the following Theorem \texttt{43} we further strengthen our result of Theorem \texttt{42} and show the
sublinear Nash-Regret bounds for Algorithm \texttt{12} in both two-player Markov cooperative and zero-
sum competitive games; see Appendix \texttt{D.6.2} for proof.

\textbf{Theorem 43 (Nash-Regret bound for Markov cooperative/zero-sum games)} When both
players in two-player Markov games running Algorithm \texttt{12} with $\alpha(t) = \frac{1}{6 \sqrt{t}}$ and $\eta = \frac{(1-\gamma)^2}{32 \sqrt{SA}}$, inde-
pendently, we have

(i) if two players have identical rewards ($r_1 = r_2 = r$), then,

$$
\frac{1}{T} \sum_{t=1}^{T} \max_{x',y'} \left( V^{x',y'}(\rho) + V^{x(t),y}(\rho) - 2V^{x(t),y(t)}(\rho) \right) \lesssim \frac{(SA)^{1/4}}{(1-\gamma)^{3/2} T^{1/6}}.
$$

(ii) if two players have zero-sum rewards ($r_1 = -r_2 = r$), then,

$$
\frac{1}{T} \sum_{t=1}^{T} \max_{x',y'} \left( V^{x',y'(t)}(\rho) - V^{x(t),y}(\rho) \right) \lesssim \frac{(SA)^{1/2}}{(1-\gamma)^{15/4} T^{1/6}}.
$$

For two-player Markov cooperative/competitive games, Theorem \texttt{43} establishes the same rate
$T^{-1/6}$ for the Nash regret and the average duality gap, respectively. Alternatively, independent
players in Algorithm \texttt{12} can find an $\epsilon$-Nash equilibrium after $O(1/\epsilon^6)$ iterations, no matter which
types of games are being played. To the best of our knowledge, Theorem \texttt{43} appears to be the first
game-agnostic convergence for Markov cooperative/competitive games with finite-time performance guarantees. We leave the extension to more general Markov games for future work.

5.8 Computational experiments

To demonstrate the merits and the effectiveness of our approach, we examine an MDP in which every state defines a congestion game. This example is borrowed from [36] and it includes MPG as a special case. For illustration, we consider the state space \( S = \{ \text{safe, distancing} \} \) and action space \( \mathcal{A}_i = \{ A, B, C, D \} \), and the number of players \( N = 8 \). In each state \( s \in S \), the reward for player \( i \) taking an action \( a \in \mathcal{A}_i \) is the \( w_s^a \)-weighted number of players using the action \( a \), where \( w_s^a \) specifies the action preference \( w_s^A < w_s^B < w_s^C < w_s^D \). The reward in state \( \text{distancing} \) is less than that in state \( \text{safe} \) by a large amount \( c > 0 \). For state transition, if more than half of players find themselves using the same action, then the state transits to the state \( \text{distancing} \); transition back to the state \( \text{safe} \) whenever no more than half of players take the same action.

In our experiments, we implement our independent policy gradient method based on the code for the projected stochastic gradient ascent \([131]\). At each iteration, we collect a batch of 20 trajectories to estimate the state-action value function and (or) the stationary state distribution under current policy. We choose the discount factor \( \gamma = 0.99 \), and different the stepsize \( \eta \), and initial state distributions. We note that stepsize \( \eta \geq 0.001 \) does not provide convergence of the projected stochastic gradient ascent \([131]\).

We report our computational results showing that our independent policy gradient with a large stepsize (green curve) quickly converges to a Nash equilibrium for a broad range of initial distributions. We first verify that our independent policy gradient with \( \eta = 0.001 \) still converges
In Figure 5.1, we see an improved convergence of our independent policy gradient using a larger stepsize, e.g., \( \eta = 0.002 \). We also remark that the learnt policies for all these experiments can generate the same Nash policy that matches the result in \([131]\).

![Graphs](image)

**Figure 5.1:** Convergence performance. (a) Learning curves for our independent policy gradient (—) with stepsize \( \eta = 0.001 \) and the projected stochastic gradient ascent (—) with \( \eta = 0.0001 \) \([131]\). Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval. (b) Learning curves for six individual runs of our independent policy gradient (solid line) and the projected stochastic gradient ascent (dash line) three each. (c) Distribution of players in one of two states taking four actions. In (a) and (b), we measure the accuracy by the absolute distance of each iterate to the converged Nash policy, i.e., 
\[
\frac{1}{N} \sum_{i=1}^{N} \| \pi_i^{(t)} - \pi_i^{\text{Nash}} \|_1.
\]
In our computational experiments, the initial distribution \( \rho \) is uniform.

We next examine how sensitive the performance of algorithms depends on initial state distributions. As discussed in Section 5.4, our independent policy gradient method \([5.4]\) is different
Figure 5.2: Convergence performance. (a) Learning curves for our independent policy gradient (—) with stepsize $\eta = 0.002$ and the projected stochastic gradient ascent (—) with $\eta = 0.0001$ [131]. Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval. (b) Learning curves for six individual runs of our independent policy gradient (solid line) and the projected stochastic gradient ascent (dash line) three each. (c) Distribution of players in one of two states taking four actions. In (a) and (b), we measure the accuracy by the absolute distance of each iterate to the converged Nash policy, i.e.,

$$\frac{1}{N} \sum_{i=1}^{N} \| \pi_i^{(t)} - \pi_i^{\text{Nash}} \|_1.$$  

In our computational experiments, the initial distribution $\rho$ is uniform. From the projected policy gradient (5.3) by removing the dependence on the initial state distribution. In the policy gradient theory [8], convergence of projected policy gradient methods is often restricted by how explorative the initial state distribution is. To be fair, we choose stepsize $\eta = 0.001$ for our algorithm since it achieves a similar performance as the projected stochastic gradient ascent [131] in Figure 5.1. We choose two different initial state distributions.
Figure 5.3: Convergence performance. (a) Learning curves for our independent policy gradient (—) with stepsize $\eta = 0.001$ and the projected stochastic gradient ascent (—) with $\eta = 0.0001$ [131]. Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval. (b) Learning curves for six individual runs of our independent policy gradient (solid line) and the projected stochastic gradient ascent (dash line) three each. (c) Distribution of players in one of two states taking four actions. In (a) and (b), we measure the accuracy by the absolute distance of each iterate to the converged Nash policy, i.e.,

$$\frac{1}{N} \sum_{i=1}^{N} \| \pi_i^{(t)} - \pi_i^{\text{Nash}} \|_1.$$  

In our computational experiments, the initial distribution is nearly degenerate $\rho = (0.9999, 0.0001)$. 

$\rho = (0.9999, 0.0001)$ and $\rho = (0.0001, 0.9999)$ and report our computational results in Figure 5.3 and Figure 5.4 respectively. Compared Figure 5.3 with Figure 5.1 both algorithms become a bit slower, but our algorithm is relatively insusceptible to the change of $\rho$. This becomes more clear in Figure 5.4 for another $\rho = (0.0001, 0.9999)$. This demonstrates that practical performance of our independent policy gradient method indeed is invariant to the initial distribution $\rho$. 


Figure 5.4: Convergence performance. (a) Learning curves for our independent policy gradient (—) with stepsize $\eta = 0.001$ and the projected stochastic gradient ascent (—) with $\eta = 0.0001$. Each solid line is the mean of trajectories over three random seeds and each shaded region displays the confidence interval. (b) Learning curves for six individual runs of our independent policy gradient (solid line) and the projected stochastic gradient ascent (dash line) three each. (c) Distribution of players in one of two states taking four actions. In (a) and (b), we measure the accuracy by the absolute distance of each iterate to the converged Nash policy, i.e., $\frac{1}{N} \sum_{i=1}^{N} \| \pi_i(t) - \pi_i^{Nash} \|_1$. In our computational experiments, the initial distribution is nearly degenerate $\rho = (0.0001, 0.9999)$.

### 5.9 Concluding remarks

We have proposed new independent policy gradient algorithms for learning a Nash equilibrium of Markov potential games (MPGs) when the size of state space and/or the number of players are large. In the exact gradient case, we show that our algorithm finds an $\epsilon$-Nash equilibrium
with $O(1/\epsilon^2)$ iteration complexity. Such iteration complexity does not explicitly depend on the state space size. In the sample-based case, our algorithm works in the function approximation setting, and we prove $O(1/\epsilon^5)$ sample complexity in a potentially infinitely large state space. This appears to be the first result for learning MPGs with function approximation. Moreover, we identify a class of independent policy gradient algorithms that enjoys last-iterate convergence and sublinear Nash regret for both zero-sum Markov games and Markov cooperative games (a special case of MPGs). This finding sheds light on an open question in the literature on the existence of such an algorithm.
Chapter 6

Discussion and future directions

We have established several RL algorithms for constrained and multi-agent control systems with theoretical convergence guarantees. We hope our results will serve as general frameworks/tools for considering real-world systems in standard RL setups. While there are a number of future directions motivated by our work, we list the most compelling ones in this chapter.

6.1 Policy gradient primal-dual algorithms

Since our work [73] published at NeurIPS 2020, there is a line of studies on policy gradient primal-dual algorithms for constrained MDPs. The first focus is the two-time scale scheme for updating primal and dual variables: [141, 134, 260] have shown the fast convergence rate by incorporating modifications to the objective function or the update of the dual variable into the algorithm design. It is relevant to examine when can we improve the convergence of single-time scale primal-dual algorithms. The second focus is the zero constraint violation during the training: [22] has shown that the constraint violation can be reduced to zero using the pessimism for constraint satisfaction, which is an important design to implement for practical algorithms. Moreover, all
these methods are limited for small tabular constrained MDPs, which leads to an open question for constrained MDPs with large state/action spaces.

A critical underlying assumption in policy gradient primal-dual algorithms is that the state space is already well-explored. Without this assumption, vanishing policy gradients often yield poor sample efficiency [8]. It is important to study strategic exploration for policy gradient primal-dual algorithms, e.g., policy cover directed exploration [9, 88, 266] and $\epsilon$-greedy [269].

In practice, the oscillatory behavior commonly arises in primal-dual methods [210]. It would be useful to establish a tighter understanding of the primal-dual optimization geometry for constrained MDPs.

6.2 Provably efficient RL for constrained MDPs

Robustness has received a lot of attention in policy optimization algorithms, especially in handling adversarial rewards/utilities [190, 109, 47, 146, 99] and model misspecification [110]. We believe that techniques from the adversarial MDP literature [47, 99, 146] allow us to derive similar regret and constraint violation bounds for constrained MDPs, which is an important future direction.

Beyond linear kernel MDPs, the confidence-interval exploration has been widely used for other types of dynamics, e.g., linear MDPs [110] and linear/nonlinear regulators [6, 114]. It remains to be seen if/how provably efficient RL algorithms can be designed for similar constrained dynamics.
In practice, we often encounter general function approximation beyond linear functions \[103, 106, 79, 279\]. It would be useful to design provably efficient RL algorithms for constrained MDPs with nonlinear function approximation.

### 6.3 Multi-agent policy evaluation in other settings

Our developed framework of distributed policy evaluation is based on the linear function approximation. It is of interest to study nonlinear function approximators \([281, 237]\). Convex-concave property of the primal-dual formulation no longer holds in this setup and different analysis is required.

Our network setting is restricted to fully connected consensus networks. It is of interest to examine transient performance of a distributed algorithms that, instead of our consensus-based approach, utilize a network diffusion strategy \([192]\). Our algorithm requires synchronous communication over a network with doubly stochastic matrices. This can be restrictive in applications that involve directed networks, communication delays, and time-varying channels. It is thus relevant to examine the effect of alternative communication schemes.

Our multi-agent MDP assumes that agents function without any failures, and it is of interest to examine a setup in which agents may experience communication/computation failures with some of them acting maliciously \([246]\).
6.4 Multi-agent policy gradient methods

A natural future direction is to extend techniques that offer faster rates for the single-agent policy gradient methods [123, 270, 247] to independent multi-agent learning for Markov cooperative/potential games.

In terms of improving sample efficiency, it is of interest to study the exploration techniques in single-agent RL [111, 186] for Markov zero-sum/cooperative/potential games with or without function approximation.

Another important direction is to investigate independent policy gradient methods for other classes of large-scale Markov games with continuous state/action spaces, e.g., linear quadratic Markov zero-sum/cooperative/potential games [268, 157, 44, 274, 97].
Bibliography


policy evaluation in distributed reinforcement learning over networks”. In: arXiv

[197] Shai Shalev-Shwartz and Shai Ben-David. Understanding Machine Learning: From Theory

[198] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. “Safe, multi-agent,
reinforcement learning for autonomous driving”. In: arXiv preprint arXiv:1610.03295
(2016).

optimization: Global convergence and faster rates for regularized MDPs”. In: Proceedings

[200] Lloyd S. Shapley. “Stochastic games”. In: Proceedings of the National Academy of Sciences

[201] Li Shen, Long Yang, Shixiang Chen, Bo Yuan, Xueqian Wang, and Dacheng Tao.
"Penalized proximal policy optimization for safe reinforcement learning”. In: arXiv

[202] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai,
Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. “A
general reinforcement learning algorithm that masters chess, shogi, and Go through

[203] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van
Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam,
Marc Lanctot, et al. “Mastering the game of Go with deep neural networks and tree


[205] Ziang Song, Song Mei, and Yu Bai. “When can we learn general-sum Markov games
with a large number of players sample-efficiently?” In: Proceedings of the International


Appendices
Appendix A

Supporting proofs in Chapter 2

A.1 Proof of Lemma 1

The proof of (i) is standard; e.g., see [11, Theorem 3.6] or [179, Theorem 1] or [180, Theorem 3]. The proof of (ii) builds on the constrained optimization [27, Section 8.5]. Let \( \Lambda_a := \{ \lambda \geq 0 \mid V_D^\lambda(\rho) \leq a \} \) be a sublevel set of the dual function for \( a \in \mathbb{R} \). For any \( \lambda \in \Lambda_a \), we have
\[
a \geq V_D^\lambda(\rho) \geq V_r^\pi(\rho) + \lambda (V_g^\pi(\rho) - b) \geq V_r^\pi(\rho) + \lambda \xi
\]
where \( \bar{\pi} \) is a Slater point. Thus, \( \lambda \leq (a - V_r^\pi(\rho)) / \xi \). If we take \( a = V^*_r(\rho) = V^*_D \), then \( \Lambda_a = \Lambda^* \) which proves (ii).

A.2 Proof of Lemma 2

By the definition of \( v(\tau) \), we have \( v(0) = V^*_r(\rho) \). We also note that \( v(\tau) \) is concave (see the proof of [179, Proposition 1]). First, we show that \( -\lambda^* \in \partial v(0) \). By the definition of \( V_{L,\lambda}^\pi(\rho) \) and the strong duality in Lemma 1 for all \( \pi \in \Pi \),
\[
V_{L,\lambda}^\pi(\rho) \leq \max_{\pi \in \Pi} V_{L,\lambda}^\pi(\rho) = V_D^*(\rho) = V_r^*(\rho) = v(0).
\]
Hence, for any \( \pi \in \{ \pi \in \Pi \mid V_g^\pi(\rho) \geq b + \tau \} \),
\[
v(0) - \tau \lambda^* \geq V_{L,\lambda}^\pi(\rho) - \tau \lambda^*
= V_r^\pi(\rho) + \lambda^* (V_g^\pi(\rho) - b) - \tau \lambda^*
= V_r^\pi(\rho) + \lambda^* (V_g^\pi(\rho) - b - \tau)
\geq V_r^\pi(\rho).
\]
Maximizing the right-hand side of this inequality over \( \{ \pi \in \Pi \mid V_g^\pi(\rho) \geq b + \tau \} \) yields
\[
v(0) - \tau \lambda^* \geq v(\tau) \quad (A.1)
\]
and, thus, \( -\lambda^* \in \partial v(0) \).
On the other hand, if we take \( \tau = - (b - V_g^\pi(\rho))_+ \), then
\[
V_r^\pi(\rho) \leq v(\tau) \quad \text{and} \quad V_\rho^* (\rho) = v(0) \leq v(\tau).
\] (A.2)

Combining (A.1) and (A.2) yields \( V_r^\pi(\rho) - V_\rho^* (\rho) \leq -\tau \lambda^* \). Thus,
\[
(C - \lambda^*) |\tau| = - \lambda^* |\tau| + C |\tau| = \tau \lambda^* + C |\tau| \leq V_\rho^* (\rho) - V_r^\pi(\rho) + C |\tau|
\]
which completes the proof by applying the assumed condition on \( \pi \).

### A.3 Proof of Lemma 3

We prove Lemma 3 by providing a concrete constrained MDP example as shown in Figure 2.1. States \( s_3, s_4 \), and \( s_5 \) are terminal states with zero reward and utility. We consider non-trivial state \( s_1 \) with two actions: \( a_1 \) moving ‘up’ and \( a_2 \) going ‘right’, and the associated value functions are given by
\[
V_r^\pi(s_1) = \pi(a_2 \mid s_1)\pi(a_1 \mid s_2) \\
V_g^\pi(s_1) = \pi(a_1 \mid s_1) + \pi(a_2 \mid s_1)\pi(a_1 \mid s_2).
\]

We consider the following two policies \( \pi^{(1)} \) and \( \pi^{(2)} \) using the softmax parametrization (2.5),
\[
\theta^{(1)} = (\log 1, \log x, \log x, \log 1) \\
\theta^{(2)} = (-\log 1, -\log x, -\log x, -\log 1)
\]
where the parameter takes form of \( (\theta_{s_1,a_1}, \theta_{s_1,a_2}, \theta_{s_2,a_1}, \theta_{s_2,a_2}) \) with \( x > 0 \).

First, we show that \( V_r^\pi \) is not concave. We compute that
\[
\pi^{(1)}(a_1 \mid s_1) = \frac{1}{1 + x}, \quad \pi^{(1)}(a_2 \mid s_1) = \frac{x}{1 + x}, \quad \pi^{(1)}(a_1 \mid s_2) = \frac{x}{1 + x} \\
V_r^{(1)}(s_1) = \left( \frac{x}{1 + x} \right)^2, \quad V_g^{(1)}(s_1) = \frac{1 + x + x^2}{(1 + x)^2} \\
\pi^{(2)}(a_1 \mid s_1) = \frac{x}{1 + x}, \quad \pi^{(2)}(a_2 \mid s_1) = \frac{1}{1 + x}, \quad \pi^{(2)}(a_1 \mid s_2) = \frac{1}{1 + x} \\
V_r^{(2)}(s_1) = \left( \frac{1}{1 + x} \right)^2, \quad V_g^{(2)}(s_1) = \frac{1 + x + x^2}{(1 + x)^2}.
\]

Now, we consider policy \( \pi^{(\zeta)} \),
\[
\zeta \theta^{(1)} + (1 - \zeta) \theta^{(2)} = (\log 1, \log (x^{2\zeta-1}), \log (x^{2\zeta-1}), \log 1)
\]
for some \( \zeta \in [0, 1] \), which is defined on the segment between \( \theta^{(1)} \) and \( \theta^{(2)} \). Therefore,
\[
\pi^{(1)}(a_1 \mid s_1) = \frac{1}{1 + x^{2\zeta-1}}, \quad \pi^{(1)}(a_2 \mid s_1) = \frac{x^{2\zeta-1}}{1 + x^{2\zeta-1}}, \quad \pi^{(1)}(a_1 \mid s_2) = \frac{x^{2\zeta-1}}{1 + x^{2\zeta-1}}
\]
\[ V_r^{(\zeta)}(s_1) = \left( \frac{x^{2\zeta-1}}{1 + x^{2\zeta-1}} \right)^2, \quad V_g^{(\zeta)}(s_1) = \frac{1 + x^{2\zeta-1} + (x^{2\zeta-1})^2}{(1 + x^{2\zeta-1})^2}. \]

When \( x = 3 \) and \( \zeta = \frac{1}{2} \),
\[ \frac{1}{2} V_r^{(1)}(s_1) + \frac{1}{2} V_r^{(2)}(s_1) = \frac{5}{16} > V_r^{\frac{1}{2}}(s_1) = \frac{4}{16} \]
which implies that \( V_r^\pi \) is not concave.

When \( x = 10 \) and \( \zeta = \frac{1}{2} \),
\[ V_g^{(1)}(s_1) = V_g^{(2)}(s_1) \geq 0.9 \quad \text{and} \quad V_g^{\frac{1}{2}}(s_1) = 0.75 \]
which shows that if we take constraint offset \( b = 0.9 \), then \( V_g^{(1)}(s_1) = V_g^{(2)}(s_1) \geq b \), and \( V_g^{\frac{1}{2}}(s_1) < b \) in which the policy \( \pi^{\frac{1}{2}} \) is infeasible. Therefore, the set \( \{ \theta \mid V_g^\pi(s) \geq b \} \) is not convex.

### A.4 Proof of Theorem 4

Let us first recall the notion of occupancy measure \([11]\). An occupancy measure \( q^\pi \) of a policy \( \pi \) is defined as a set of distributions generated by executing \( \pi \),
\[ q^\pi_{s,a} = \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s, a_t = a \mid \pi, s_0 \sim \rho) \tag{A.3} \]
for all \( s \in S \), \( a \in A \). We put all \( q^\pi_{s,a} \) together as \( q^\pi \in \mathbb{R}^{|S| \times |A|} \) and \( q^\pi_a = [q^\pi_{1,a}, \ldots, q^\pi_{|S|,a}]^\top \), for brevity. For an action \( a \), we collect transition probabilities \( \mathbb{P}(s' \mid s, a) \) for all \( s', s \in S \) to have the shorthand notation \( P_a \in \mathbb{R}^{|S| \times |S|} \). The occupancy measure \( q^\pi \) has to satisfy a set of linear constraints given by \( Q := \{ q^\pi \in \mathbb{R}^{|S| \times |A|} \mid \sum_{a \in A} (I - \gamma P_a^\top) q^\pi_a = \rho \text{ and } q^\pi \succeq 0 \} \). With a slight abuse of notation, we write \( r \in [0,1]^{|S| \times |A|} \) and \( g \in [0,1]^{|S| \times |A|} \). Thus, the value functions \( V_r^\pi, V_g^\pi: S \to \mathbb{R} \) under the initial state distribution \( \rho \) are linear in \( q^\pi \):
\[ V_r^\pi(\rho) = \langle q^\pi, r \rangle := F_r(q^\pi) \quad \text{and} \quad V_g^\pi(\rho) = \langle q^\pi, g \rangle := F_g(q^\pi). \]

We are now in a position to consider the primal problem (2.3) as a linear program,
\[ \max_{q^\pi \in Q} F_r(q^\pi) \quad \text{subject to} \quad F_g(q^\pi) \geq b \tag{A.4} \]
where the maximization is over all occupancy measures \( q^\pi \in Q \). Once we compute a solution \( q^\pi \), the associated policy solution \( \pi \) can be recovered via
\[ \pi(a \mid s) = \frac{q^\pi_{s,a}}{\sum_{a \in A} q^\pi_{s,a}} \quad \text{for all } s \in S, a \in A. \tag{A.5} \]
Abstractly, we let \( \pi^\theta : \mathcal{Q} \to \Delta(\mathcal{A})^{\mathcal{S}} \) be a mapping from an occupancy measure \( q^\pi \) to a policy \( \pi \). Similarly, as defined by (A.3) we let \( q^\pi : \Delta(\mathcal{A})^{\mathcal{S}} \to \mathcal{Q} \) be a mapping from a policy \( \pi \) to an occupancy measure \( q^\pi \). Clearly, \( q^\pi = (\pi^\theta)^{-1} \).

Despite the non-convexity essence of (2.3) in policy space, the reformulation (A.4) reveals underlying convexity in occupancy measure \( q^\pi \). In Lemma 44 we exploit this convexity to show the average policy improvement over \( T \) steps.

**Lemma 44 (Bounded average performance)** Let assumptions in Theorem 4 hold. Then, the iterates \((\theta(t), \lambda(t))\) generated by PG-PD method (2.10) satisfy

\[
\frac{1}{T} \sum_{t=0}^{T-1} Z(t) \left( F_r(q^\theta) - F_r(q^{\theta(t)}) \right) + \frac{1}{T} \sum_{t=0}^{T-1} \lambda(t) \left( F_g(q^\theta) - F_g(q^{\theta(t)}) \right) \leq \frac{D_\theta L_\theta}{T^{1/4}}
\]

where \( D_\theta := \frac{8|\mathcal{S}|}{(1-\gamma)^2} \|q^\pi^* / \rho\|^2_\infty \) and \( L_\theta := \frac{2|\mathcal{A}|(1+3/\xi)}{(1-\gamma)^2} \).

**Proof.** From the dual update in (2.10) we have \( 0 \leq \lambda(t) \leq 2/(1 - \gamma) \xi \). From the smooth property of the value functions under the direct policy parametrization (8 Lemma D.3) we have

\[
\left| F_r(q^\theta) - F_r(q^{\theta(t)}) - \langle \nabla_\theta F_r(q^{\theta(t)}), \theta - \theta^{(t)} \rangle \right| \leq \frac{\gamma |\mathcal{A}|}{(1-\gamma)^3} \|\theta - \theta^{(t)}\|^2.
\]

If we fix \( \lambda^{(t)} \geq 0 \), then

\[
\left| (F_r + \lambda^{(t)} F_g)(q^\theta) - (F_r + \lambda^{(t)} F_g)(q^{\theta(t)}) - \langle \nabla_\theta F_r(q^{\theta(t)}), \lambda^{(t)} \nabla_\theta F_g(q^{\theta(t)}), \theta - \theta^{(t)} \rangle \right|
\leq \frac{L_\theta}{2} \|\theta - \theta^{(t)}\|^2.
\]

Thus,

\[
(F_r + \lambda^{(t)} F_g)(q^\theta) \geq (F_r + \lambda^{(t)} F_g)(q^{\theta(t)}) + \langle \nabla_\theta F_r(q^{\theta(t)}), \lambda^{(t)} \nabla_\theta F_g(q^{\theta(t)}), \theta - \theta^{(t)} \rangle - \frac{L_\theta}{2} \|\theta - \theta^{(t)}\|^2
\]

\[
\geq (F_r + \lambda^{(t)} F_g)(q^\theta) - L_\theta \|\theta - \theta^{(t)}\|^2.
\]

We note that the primal update in (2.10) is equivalent to

\[
\theta^{(t+1)} = \operatorname{argmax}_{\theta \in \Theta} \left\{ \frac{V_r^{\theta(t)}(\rho)}{\theta} + \lambda^{(t)} V_g^{\theta(t)}(\rho)
+ \langle \nabla_\theta V_r^{\theta(t)}(\rho), \lambda^{(t)} \nabla_\theta V_g^{\theta(t)}(\rho), \theta - \theta^{(t)} \rangle - \frac{1}{2\eta_1} \|\theta - \theta^{(t)}\|^2 \right\}.
\]
By taking \( \eta_t = 1/L_\theta \) and \( \theta = \theta^{(t+1)} \) in (A.7),

\[
(F_r + \lambda^{(t)} F_g)(q^{\theta^{(t+1)}})
\]

\[
\geq \maximize_{\theta \in \Theta} \left\{ (F_r + \lambda^{(t)} F_g)(q^{\theta(t)}) + \langle \nabla_\theta F_r(q^{\theta(t)}), \lambda^{(t)} \nabla_\theta F_g(q^{\theta(t)}), \theta - \theta^{(t)} \rangle - \frac{L_\theta}{2} \| \theta - \theta^{(t)} \|^2 \right\} \tag{A.8}
\]

\[
\geq \maximize_{\theta \in \Theta} \left\{ (F_r + \lambda^{(t)} F_g)(q^{\theta(t)}) - L_\theta \| \theta - \theta^{(t)} \|^2 \right\}
\]

\[
\geq \maximize_{\alpha \in [0,1]} \left\{ (F_r + \lambda^{(t)} F_g)(q^{\theta_{\alpha}}) - L_\theta \| \theta_{\alpha} - \theta^{(t)} \|^2 \right\}
\]

where \( \theta_{\alpha} := \pi^q(\alpha q^{\theta^*} + (1 - \alpha)q^{\theta(t)}) \), we apply (A.7) for the second inequality, and the last inequality is due to \( \pi^q \circ q^\theta = \text{id}_{S,A} \) and linearity of \( q^\theta \) in \( \theta \). Since \( F_r \) and \( F_g \) are linear in \( q^\theta \), we have

\[
(F_r + \lambda^{(t)} F_g)(q^{\theta^*}) = \alpha(F_r + \lambda^{(t)} F_g)(q^{\theta_{\alpha}}) + (1 - \alpha)(F_r + \lambda^{(t)} F_g)(q^{\theta(t)}). \tag{A.9}
\]

By the definition of \( \pi^q \),

\[
(\pi^q(q) - \pi^q(q'))_{sa} = \frac{1}{\sum_{a \in A} q_{sa}} (q_{sa} - q'_{sa}) + \frac{\sum_{a \in A} q_{sa} - \sum_{a \in A} q_{sa}}{\sum_{a \in A} q_{sa} \sum_{a \in A} q_{sa}} q'_{sa}
\]

which, together with \( \| x + y \|^2 \leq 2 \| x \|^2 + 2 \| y \|^2 \), gives

\[
\| \pi^q(q) - \pi^q(q') \|^2
\]

\[
\leq 2 \sum_{s \in S} \sum_{a \in A} \frac{(q_{sa} - q'_{sa})^2}{(\sum_{a \in A} q_{sa})^2} + 2 \sum_{s \in S} \sum_{a \in A} \left( \frac{\sum_{a \in A} q_{sa} - \sum_{a \in A} q_{sa}}{\sum_{a \in A} q_{sa} \sum_{a \in A} q_{sa}} \right)^2 (q'_{sa})^2
\]

\[
\leq 2 \sum_{s \in S} \frac{1}{(\sum_{a \in A} q_{sa})^2} \left( \sum_{a \in A} (q_{sa} - q'_{sa})^2 + \left( \sum_{a \in A} q'_{sa} - \sum_{a \in A} q_{sa} \right)^2 \right).
\]

Therefore,

\[
\| \theta_{\alpha} - \theta^{(t)} \|^2
\]

\[
= \| \pi^q \left( \alpha q^{\theta^*} + (1 - \alpha)q^{\theta(t)} \right) - \pi^q (q^{\theta(t)}) \|^2
\]

\[
\leq \sum_{s \in S} \frac{2 \alpha^2}{(\sum_{a \in A} q^{\theta(t)}_{sa})^2} \left( \sum_{a \in A} (q^{\theta^*}_{sa} - q^{\theta(t)}_{sa})^2 + \left( \sum_{a \in A} q^{\theta(t)}_{sa} - \sum_{a \in A} q^{\theta^*}_{sa} \right)^2 \right)
\]
in which the upper bound further can be relaxed into

\[
\sum_{s \in S} \frac{4\alpha^2}{(\sum_{a \in A} q_{sa}^{\theta(s)})^2} \left( \left( \sum_{a \in A} q_{sa}^{\theta^*} \right)^2 + \left( \sum_{a \in A} q_{sa}^{\theta(t)} \right)^2 \right) = \sum_{s \in S} \frac{4\alpha^2}{(\sum_{a \in A} q_{sa}^{\theta(s)})^2} \left( \left( \sum_{a \in A} q_{sa}^{\theta^*} \right)^2 + \left( \sum_{a \in A} q_{sa}^{\theta(t)} \right)^2 \right)
\]

\[
\leq 4\alpha^2 |S| \sum_{s \in S} \left( \left( \sum_{a \in A} q_{sa}^{\theta^*} \right)^2 + \left( \sum_{a \in A} q_{sa}^{\theta(t)} \right)^2 \right) \leq 4\alpha^2 |S| \sum_{s \in S} \left( \frac{d_{\theta}^{\pi(s)}}{\|d_{\theta}^{\pi(t)}\|_\infty} \right)^2 \leq \alpha^2 D_{\theta}
\]

where we apply \(d_{\theta}^{\pi(t)} \geq (1 - \gamma)\rho\) componentwise in the second inequality.

We now apply (A.9) and (A.10) to (A.8),

\[
(F_r + \lambda(t) F_g)(q^{\theta^*}) - (F_r + \lambda(t) F_g)(q^{\theta(t)}) \leq \min_{\alpha \in [0, 1]} \left\{ L_\theta \|\alpha - \theta\|^2 + (F_r + \lambda(t) F_g)(q^{\theta^*}) - (F_r + \lambda(t) F_g)(q^{\theta^*}) \right\}
\]

\[
\leq \min_{\alpha \in [0, 1]} \left\{ \alpha^2 D_{\theta} L_\theta + (1 - \alpha)((F_r + \lambda(t) F_g)(q^{\theta^*}) - (F_r + \lambda(t) F_g)(q^{\theta(t)})) \right\}
\]

which further implies

\[
(F_r + \lambda(t+1) F_g)(q^{\theta^*}) - (F_r + \lambda(t+1) F_g)(q^{\theta(t+1)}) \leq \min_{\alpha \in [0, 1]} \left\{ \alpha^2 D_{\theta} L_\theta + (1 - \alpha)((F_r + \lambda(t) F_g)(q^{\theta^*}) - (F_r + \lambda(t) F_g)(q^{\theta(t)})) \right\}
\]

(A.11)

We check the right-hand side of the inequality (A.11). By the dual update in (2.10), it is easy to see that

\[
-(\lambda(t) - \lambda(t+1))(F_g(q^{\theta^*}) - F_g(q^{\theta(t+1)})) \leq |\lambda(t) - \lambda(t+1)|/(1 - \gamma) \leq \eta_2/(1 - \gamma)^2.
\]

we discuss three cases: (i) when \(\alpha(t) < 0\), we set \(\alpha = 0\) for (A.11),

\[
(F_r + \lambda(t+1) F_g)(q^{\theta^*}) - (F_r + \lambda(t+1) F_g)(q^{\theta(t+1)}) \leq \frac{D_{\theta} L_\theta}{2\sqrt{T}};
\]

(A.12)

(ii) when \(\alpha(t) > 1\), we set \(\alpha = 1\) that leads to

\[
(F_r + \lambda(t+1) F_g)(q^{\theta^*}) - (F_r + \lambda(t+1) F_g)(q^{\theta(t+1)}) \leq \frac{D_{\theta} L_\theta}{2\sqrt{T}}, \text{i.e., } \alpha^{(t+1)} \leq 3/4.
\]

Thus, this case reduces to the next case (iii): \(0 \leq \alpha(t) \leq 1\) in which we can express (A.11) as

\[
(F_r + \lambda(t+1) F_g)(q^{\theta^*}) - (F_r + \lambda(t+1) F_g)(q^{\theta(t+1)}) \leq \left( 1 - \frac{(F_r + \lambda(t) F_g)(q^{\theta^*}) - (F_r + \lambda(t) F_g)(q^{\theta(t)})}{4D_{\theta} L_\theta} \right) \times \left( (F_r + \lambda(t) F_g)(q^{\theta^*}) - (F_r + \lambda(t) F_g)(q^{\theta(t)}) \right)
\]

\[
+ \frac{D_{\theta} L_\theta}{2\sqrt{T}}
\]

(A.13)
or equivalently,

\[ \alpha^{(t+1)} \leq \left(1 - \frac{\alpha^{(t)}}{2}\right) \alpha^{(t)} + \frac{1}{4\sqrt{T}}. \]  

(A.13)

By choosing \( \lambda^{(0)} = 0 \) and \( \theta^{(0)} \) such that \( V_{\pi}^{\theta^{(0)}}(\rho) \geq V_{\pi}^{\theta^*}(\rho) \), we know that \( \alpha^{(0)} \leq 0 \). Thus, \( \alpha^{(1)} \leq 1/(4\sqrt{T}) \). By (A.12), the case \( \alpha^{(1)} \leq 0 \) is trivial. Without loss of generality, we assume that \( 0 \leq \alpha^{(t)} \leq 1/T^{1/4} \leq 1 \). By induction over \( t \) for (A.13),

\[ \alpha^{(t+1)} \leq \left(1 - \frac{\alpha^{(t)}}{2}\right) \alpha^{(t)} + \frac{1}{4\sqrt{T}} \leq \frac{1}{T^{1/4}}. \]  

(A.14)

By combining (A.12) and (A.14), and averaging over \( t = 0, 1, \cdots, T - 1 \), we get the desired bound. \( \square \)

**Proof.** [Proof of Theorem 4]

**Bounding the optimality gap.** By the dual update (2.10) and \( \lambda^{(0)} = 0 \), it is convenient to bound \( (\lambda^{(T)})^2 \) by

\[
(\lambda^{(T)})^2 = \sum_{t=0}^{T-1} ((\lambda^{(t+1)})^2 - (\lambda^{(t)})^2)
\]

\[= 2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)}(b - F_g(q^{\theta^{(t)}})) + \eta_2^2 \sum_{t=0}^{T-1} (F_g(q^{\theta^{(t)}}) - b)^2
\]

\[\leq 2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)}(F_g(q^*) - F_g(q^{\theta^{(t)}})) + \eta_2^2 \frac{T}{(1 - \gamma)^2}
\]

where the inequality is due to the feasibility of the optimal policy \( \pi^* \) or the associated occupancy measure \( q^* = q^{\theta^*} : F_g(q^*) \geq b \), and \( |F_g(q^{\theta^{(t)}}) - b| \leq 1/(1 - \gamma) \). The above inequality further implies

\[-\frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)}(F_g(q^*) - F_g(q^{\theta^{(t)}})) \leq \frac{\eta_2^2}{2(1 - \gamma)^2}.
\]

By substituting the above inequality into (A.6) in Lemma 44, we show the desired optimality gap bound, where we take \( \eta_2 = (1 - \gamma)^2 D_q L_g/(2\sqrt{T}) \).

**Bounding the constraint violation.** From the dual update in (2.10) we have for any \( \lambda \in [0, 2/((1 - \gamma)\xi)] \),

\[|\lambda^{(t+1)} - \lambda|^2 \overset{(a)}{\leq} |\lambda^{(t)} - \eta_2(F_g(q^{\theta^{(t)}}) - b) - \lambda|^2
\]

\[\overset{(b)}{\leq} |\lambda^{(t)} - \lambda|^2 - 2\eta_2(F_g(q^{\theta^{(t)}}) - b)(\lambda^{(t)} - \lambda) + \frac{\eta_2^2}{(1 - \gamma)^2}
\]
where \((a)\) is due to the non-expansiveness of projection \(P_\lambda\) and \((b)\) is due to \((F_g(q^{\theta(t)}) - b)^2 \leq 1/(1 - \gamma)^2\). Summing it up from \(t = 0\) to \(t = T - 1\), and dividing it by \(T\), yield

\[
\frac{1}{T} |\lambda(T) - \lambda|^2 - \frac{1}{T} |\lambda(0) - \lambda|^2 \leq -\frac{2\eta_2}{T} \sum_{t=0}^{T-1} (F_g(q^{\theta(t)}) - b)(\lambda(t) - \lambda) + \frac{\eta_2^2}{(1 - \gamma)^2}
\]

which further implies,

\[
\frac{1}{T} \sum_{t=0}^{T-1} (F_g(q^{\theta(t)}) - b)(\lambda(t) - \lambda) \leq \frac{|\lambda(0) - \lambda|^2}{2\eta_2 T} + \frac{\eta_2}{2(1 - \gamma)^2}.
\]

We note that \(F_g(q^{\theta^*}) \geq b\). By adding the inequality above to (A.6) in Lemma 44 from both sides,

\[
\frac{1}{T} \sum_{t=0}^{T-1} (F_r(q^{\theta^*}) - F_r(q^{\theta(t)})) + \frac{\lambda}{T} \sum_{t=0}^{T-1} (b - F_g(q^{\theta(t)})) \leq \frac{D_\theta L_\theta}{T^{1/4}} + \frac{1}{2\eta_2 T} |\lambda(0) - \lambda|^2 + \frac{\eta_2}{2(1 - \gamma)^2}.
\]

We choose \(\lambda = \frac{-2}{(1 - \gamma)\xi}\) if \(\sum_{t=0}^{T-1} (b - F_g(q^{\theta(t)})) \geq 0\); otherwise \(\lambda = 0\). Thus,

\[
F_r(q^{\theta^*}) - F_r(q') + \frac{2}{(1 - \gamma)\xi} [b - F_g(q')]_+ \leq \frac{D_\theta L_\theta}{T^{1/4}} + \frac{1}{2\eta_2 (1 - \gamma)^{2}\xi T} + \frac{\eta_2}{2(1 - \gamma)^2}
\]

where there exists \(q'\) such that \(F_r(q') := \frac{1}{T} \sum_{t=0}^{T-1} F_r(q^{\theta(t)})\) and \(F_g(q') := \frac{1}{T} \sum_{t=0}^{T-1} F_g(q^{\theta(t)})\) by the definition of occupancy measure.

Application of Lemma 2 with \(2/((1 - \gamma)\xi) \geq 2\lambda\) yields

\[
[b - F_g(q')]_+ \leq \frac{(1 - \gamma)\xi D_\theta L_\theta}{T^{1/4}} + \frac{1}{2\eta_2 (1 - \gamma)^2 T} + \frac{\eta_2}{2(1 - \gamma)}
\]

which readily leads to the desired constraint violation bound by noting that

\[
\frac{1}{T} \sum_{t=0}^{T-1} (b - F_g(q^{\theta(t)})) = b - F_g(q')
\]

and taking \(\eta_2 = \frac{8|\lambda| |\xi| (1 + 2/\xi)}{(1 - \gamma)^3 \sqrt{T} \|d_{\rho}^{\pi^*}\|_2^2} \|d_{\rho}^{\pi^*}/\rho\|_\infty^2 \geq (1 - \gamma)^2\).

### A.5 Proof of Lemma 5

The dual update follows Lemma 1. Since \(\lambda \leq (V^*_r(\rho) - V^*_r(\rho))/\xi\) with \(0 \leq V^*_r, V^*_r \leq 1/(1 - \gamma)\), we take projection interval \(\Lambda = [0, 2/((1 - \gamma)\xi)]\) such that upper bound \(2/((1 - \gamma)\xi)\) is such that \(2/((1 - \gamma)\xi) \geq 2\lambda\).
We now verify the primal update. We expand the primal update in (2.11) into the following form,
\[
\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta V_{r}^{\theta(t)}(\rho) + \eta \lambda^{(t)} F_\rho^1(\theta^{(t)}) \nabla_\theta V_{\theta}^{\theta(t)}(\rho).
\] (A.15)

We now deal with: \( F_\rho^1(\theta^{(t)}) \nabla_\theta V_{r}^{\theta(t)}(\rho) \) and \( F_\rho^1(\theta^{(t)}) \nabla_\theta V_{\theta}^{\theta(t)}(\rho) \). For the first one, the proof begins with solutions to the following approximation error minimization problem:

\[
\text{minimize}_{w \in \mathbb{R}^{[S]|A|}} E_r(w) := \mathbb{E}_{s \sim d_{\theta}^w, a \sim \pi_\theta(s | s)} \left[ (A_{r,\theta}^w(s, a) - w \nabla_\theta \log \pi_\theta(a | s))^2 \right].
\]

Using the Moore-Penrose inverse, the optimal solution reads,
\[
w_r^* = F_\rho^1(\theta) \mathbb{E}_{s \sim d_{\theta}^w, a \sim \pi_\theta(s | s)} \left[ \nabla_\theta \log \pi_\theta(a | s) A_{r,\theta}^w(s, a) \right] = (1 - \gamma) F_\rho^1(\theta) \nabla_\theta V_{r}^{\pi_\theta \lambda}(\rho)
\]
where \( F_\rho(\theta) \) is the Fisher information matrix induced by \( \pi_\theta \). One key observation from this solution is that \( w_r^* \) is parallel to the NPG direction \( F_\rho^1(\theta) \nabla_\theta V_{r}^{\pi_\theta \lambda}(\rho) \).

On the other hand, it is easy to verify that \( A_{r,\theta}^w \) is a minimizer of \( E_r(w) \). The softmax policy (2.5) implies that
\[
\frac{\partial \log \pi_\theta(a | s)}{\partial \theta_{s',a'}} = \mathbb{I}[s = s'] (\mathbb{I}[a = a'] - \pi_\theta(a' | s))
\] (A.16)
where \( \mathbb{I}[E] \) is the indicator function of event \( E \) being true. Thus, we have
\[
w_r^* \nabla_\theta \log \pi_\theta(a | s) = w_s,a - \sum_{a' \in A} w_{s,a'} \pi_\theta(a' | s).
\]

The above equality together with the fact: \( \sum_{a \in A} \pi_\theta(a | s) A_{r,\theta}^w(s, a) = 0 \), yields \( E_r(A_{r,\theta}^w) = 0 \). However, \( A_{r,\theta}^w \) may not be the unique minimizer. We consider the following general form of possible solutions,
\[A_{r,\theta}^w + u, \text{ where } u \in \mathbb{R}^{[S]|A|}.\]

For any state \( s \) and action \( a \) such that \( s \) is reachable under \( \rho \), using (A.16) yields
\[u \nabla_\theta \log \pi_\theta(a | s) = u_s,a - \sum_{a' \in A} u_{s,a'} \pi_\theta(a' | s).
\]

Here, we make use of the following fact: \( \pi_\theta \) is a stochastic policy with \( \pi_\theta(a | s) > 0 \) for all actions \( a \) in each state \( s \), so that if a state is reachable under \( \rho \), then it will also be reachable using \( \pi_\theta \). Therefore, we require zero derivative at each reachable state:
\[
u \nabla_\theta \log \pi_\theta(a | s) = 0
\]
for all \( s, a \) so that \( u_{s,a} \) is independent of the action and becomes a constant \( c_s \) for each \( s \). Therefore, the minimizer of \( E_r(w) \) is given up to some state-dependent offset,
\[
F_\rho^1(\theta) \nabla_\theta V_{r}^{\pi_\theta}(\rho) = \frac{A_{r,\theta}^w}{1 - \gamma} + u
\] (A.17)
where \( u_{s,a} = c_s \) for some \( c_s \in \mathbb{R} \) for each state \( s \) and action \( a \).

We can repeat the above procedure for \( F^\dagger_\rho (\theta^{(t)}) \nabla_\theta V^{\theta^{(t)}} (\rho) \) and show,

\[
F^\dagger_\rho (\theta) \nabla_\theta V^{\pi_\theta} (\rho) = \frac{A^{\pi_\theta}_g}{1 - \gamma} + v
\]

(A.18)

where \( v_{s,a} = d_s \) for some \( d_s \in \mathbb{R} \) for each state \( s \) and action \( a \).

Substituting (A.17) and (A.18) into the primal update (A.15) yields,

\[
\theta^{(t+1)} = \theta^{(t)} + \frac{\eta_1}{1 - \gamma} (A^{(t)} + \lambda^{(t)} A^{(t)}_g) + \eta_1 (u + \lambda^{(t)} v)
\]

\[
\pi^{(t+1)}(a \mid s) = \frac{\pi^{(t)}(a \mid s) \exp \left( \frac{m}{1 - \gamma} \left( A^{(t)}(s,a) + \lambda^{(t)} A^{(t)}_g (s,a) \right) + \eta_1 (c_s + \lambda^{(t)} d_s) \right)}{Z^{(t)}(s)}
\]

where the second equality also utilizes the normalization term \( Z^{(t)}(s) \). Finally, we complete the proof by setting \( c_s = d_s = 0 \).

### A.6 Sample-based algorithm with function approximation

We describe a sample-based NPG-PD algorithm with function approximation in Algorithm 1. We note the computational complexity of Algorithm 1: each round has expected length \( 2/(1 - \gamma) \) so the expected number of total samples is \( 4KT/(1 - \gamma) \); the total number of gradient computations \( \nabla_\theta \log \pi^{(t)}(a \mid s) \) is \( 2KT \); the total number of scalar multiplies, divides, and additions is \( O(dKT + KT/(1 - \gamma)) \).

The following unbiased estimates that are useful in our analysis.

\[
\mathbb{E} \left[ \hat{V}^{(t)}_g (s) \right] = \mathbb{E} \left[ \sum_{k=0}^{K'-1} g(s_k,a_k) \mid \theta^{(t)}, s_0 = s \right]
\]

\[
= \mathbb{E} \left[ \sum_{k=0}^{\infty} \mathbb{1} [K' - 1 \geq k \geq 0] g(s_k,a_k) \mid \theta^{(t)}, s_0 = s \right]
\]

\[
= \sum_{k=0}^{\infty} \mathbb{E} \left[ \mathbb{E}_{K'} [\mathbb{1} [K' - 1 \geq k \geq 0]] g(s_k,a_k) \mid \theta^{(t)}, s_0 = s \right]
\]

\[
= \sum_{k=0}^{\infty} \mathbb{E} \left[ \gamma^k g(s_k,a_k) \mid \theta^{(t)}, s_0 = s \right]
\]

\[
= \mathbb{E} \left[ \sum_{k=0}^{\infty} \gamma^k g(s_k,a_k) \mid \theta^{(t)}, s_0 = s \right]
\]

\[
= \hat{V}^{(t)}_g (s)
\]

where we apply the Monotone Convergence Theorem and the Dominated Convergence Theorem for (a) and swap the expectation and the infinite sum in (c), and in (b) we apply

\[
\mathbb{E}_{K'} [\mathbb{1} [K' - 1 \geq k \geq 0]] = 1 - P(K' < k) = \gamma^k
\]
since $K' \sim \text{Geometric}(1 - \gamma)$, a geometric distribution.

By a similar argument as above,

\[
\mathbb{E} \left[ \hat{Q}_r^{(t)}(s, a) \right] = \mathbb{E} \left[ \sum_{k=0}^{K' - 1} r(s_k, a_k) | \theta^{(t)}, s_0 = s, a_0 = a \right]
\]
\[
= \mathbb{E} \left[ \sum_{k=0}^{\infty} [K' - 1 \geq k \geq 0] r(s_k, a_k) | \theta^{(t)}, s_0 = s, a_0 = a \right]
\]
\[
= \sum_{k=0}^{\infty} \mathbb{E} \left[ \mathbb{E}_{K'} [K' - 1 \geq k \geq 0] r(s_k, a_k) | \theta^{(t)}, s_0 = s, a_0 = a \right]
\]
\[
= \sum_{k=0}^{\infty} \mathbb{E} \left[ \gamma^k r(s_k, a_k) | \theta^{(t)}, s_0 = s, a_0 = a \right]
\]
\[
= \mathbb{E} \left[ \sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) | \theta^{(t)}, s_0 = s, a_0 = a \right]
\]
\[
= Q_r^{(t)}(s, a).
\]

Therefore,

\[
\mathbb{E} \left[ \hat{A}_r^{(t)}(s, a) \right] = \mathbb{E} \left[ \hat{Q}_r^{(t)}(s, a) \right] - \mathbb{E} \left[ \hat{V}_r^{(t)}(s) \right] = Q_r^{(t)}(s, a) - V_r^{(t)}(s) = A_r^{(t)}(s, a).
\]

We also provide a bound on the variance of $\hat{V}_g^{(t)}(s)$,

\[
\text{Var} \left[ \hat{V}_g^{(t)}(s) \right] = \mathbb{E} \left[ \left( \hat{V}_g^{(t)}(s) - V_g^{(t)}(s) \right)^2 | \theta^{(t)}, s_0 = s \right]
\]
\[
= \mathbb{E} \left[ \left( \sum_{k=0}^{K' - 1} g(s_k, a_k) - V_g^{(t)}(s) \right)^2 | \theta^{(t)}, s_0 = s \right]
\]
\[
= \mathbb{E}_{K'} \left[ \mathbb{E} \left[ \left( \sum_{k=0}^{K' - 1} g(s_k, a_k) - V_g^{(t)}(s) \right)^2 \right] | K' \right]
\]
\[
\stackrel{(a)}{\leq} \mathbb{E}_{K'} \left[ (K')^2 | K' \right]
\]
\[
\stackrel{(b)}{=} \frac{1}{(1 - \gamma)^2}
\]

where (a) is due to $0 \leq g(x_k, a_k) \leq 1$ and $V_g^{(t)}(s) \geq 0$ and (b) is clear from $K' \sim \text{Geometric}(1 - \gamma)$.

By the sampling scheme of Algorithm 2 we can show that $G_{r,k}$ is an unbiased estimate of the population gradient $\nabla_{\theta} E_r^{(t)}(w_r; \theta^{(t)})$,

\[
\mathbb{E}_{(s,a) \sim d^{(t)}} \left[ G_{o,k} \right] = 2 \mathbb{E} \left[ \left( \begin{array}{c} \nabla_{\theta} \log \pi_{\theta}^{(t)}(a | s) - A_r^{(t)}(s, a) \\ \nabla_{\theta} \log \pi_{\theta}^{(t)}(a | s) - A_r^{(t)}(s, a) \end{array} \right) \nabla_{\theta} \log \pi_{\theta}^{(t)}(a | s) \right]
\]
\[
= 2 \mathbb{E} \left[ \left( w_{r,k}^T \nabla_{\theta} \log \pi_{\theta}^{(t)}(a | s) - \mathbb{E} \left[ A_r^{(t)}(s, a) \right] \right) \nabla_{\theta} \log \pi_{\theta}^{(t)}(a | s) \right]
\]
\[
= 2 \mathbb{E} \left[ \left( w_{r,k}^T \nabla_{\theta} \log \pi_{\theta}^{(t)}(a | s) - A_r^{(t)}(s, a) \right) \nabla_{\theta} \log \pi_{\theta}^{(t)}(a | s) \right]
\]
\[
= \nabla_{w_r} E_r^{(t)}(w_r; \theta^{(t)}).
\]
A.7 Proof of Theorem 13

We first adapt Lemma 11 to the sample-based case as follows.

**Lemma 45 (Sample-based regret/violation lemma)** Let Assumption 11 hold and let us fix a state distribution $\rho$ and $T > 0$. Assume that $\log \pi_{\theta}(a | s)$ is $\beta$-smooth in $\theta$ for any $(s, a)$. If the iterates $(\pi^{(t)}, \lambda^{(t)})$ generated by the Algorithm 1 with $\theta^{(0)} = 0$, $\lambda^{(0)} = 0$, $\eta_1 = \eta_2 = 1/\sqrt{T}$, and $\|\hat{w}_r^{(t)}\|, \|\hat{w}_g^{(t)}\| \leq W$, then,

$$
\mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} (V_r^{(t)}(\rho) - V_r^{(t)}(\rho)) \right] \leq \frac{C_5}{(1 - \gamma)^5} \frac{1}{\sqrt{T}} + \sum_{t=0}^{T-1} \mathbb{E} \left[ \text{err}_r^{(t)}(\pi^*) \right] + \sum_{t=0}^{T-1} \frac{2\mathbb{E} \left[ \text{err}_g^{(t)}(\pi^*) \right]}{(1 - \gamma)^2} \\
\mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} (b - V_g^{(t)}(\rho)) \right] + \sum_{t=0}^{T-1} \frac{1}{T} \xi \mathbb{E} \left[ \text{err}_g^{(t)}(\pi^*) \right] + \sum_{t=0}^{T-1} \frac{2\mathbb{E} \left[ \text{err}_g^{(t)}(\pi^*) \right]}{(1 - \gamma)^2} 
$$

where $C_5 = 2 + \log |A| + 5B^2/\xi$, $C_6 = (2 + \log |A| + B^2)\xi + (2 + 4B^2)/\xi$, and

$$
\hat{\text{err}}^{(t)}(\pi) := \mathbb{E}_{s \sim d_0^s} \mathbb{E}_{a \sim \pi^{(t)}(s)} \left[ A^{(t)}(s, a) - (\hat{\theta}^{(t)}_r)^\top \nabla_\theta \log \pi^{(t)}(a | s) \right], \quad \text{where } \diamond = r \text{ or } g.
$$

**Proof.** The smoothness of log-linear policy in conjunction with an application of Taylor’s theorem to $\log \pi^{(t)}(a | s)$ yield

$$
\log \frac{\pi^{(t)}(a | s)}{\pi^{(t+1)}(a | s)} + (\theta^{(t+1)} - \theta^{(t)})^\top \nabla_\theta \log \pi^{(t)}(a | s) \leq \frac{\beta}{2} \|\theta^{(t+1)} - \theta^{(t)}\|^2
$$

where $\theta^{(t+1)} - \theta^{(t)} = \frac{\eta_1}{1 - \gamma} \hat{w}^{(t)}$. We unload $d_0^s$ as $d^*$ since $\pi^*$ and $\rho$ are fixed. Therefore,

$$
\mathbb{E}_{s \sim d^*} \left( D_{\text{KL}}(\pi^{*} (\cdot | s) \parallel \pi^{(t)}_\theta (\cdot | s)) - D_{\text{KL}}(\pi^{*} (\cdot | s) \parallel \pi^{(t+1)}_\theta (\cdot | s)) \right) = -\mathbb{E}_{s \sim d^*} \mathbb{E}_{a \sim \pi^{*} (\cdot | s)} \log \frac{\pi^{(t)}_\theta (a | s)}{\pi^{(t+1)}_\theta (a | s)} \\
\geq \eta_1 \mathbb{E}_{s \sim d^*} \mathbb{E}_{a \sim \pi^{*} (\cdot | s)} \left[ (\hat{\theta}^{(t)}_r)^\top \nabla_\theta \log \pi^{(t)}_\theta (a | s) \right] - \frac{\beta}{2} \frac{\eta_1^2}{(1 - \gamma)^2} \|\hat{w}^{(t)}\|^2
$$

$$
= \eta_1 \mathbb{E}_{s \sim d^*} \mathbb{E}_{a \sim \pi^{*} (\cdot | s)} \left[ (\hat{\theta}^{(t)}_r)^\top \nabla_\theta \log \pi^{(t)}_\theta (a | s) \right] - \frac{\beta}{2} \frac{\eta_1^2}{(1 - \gamma)^2} \|\hat{w}^{(t)}\|^2
$$

$$
\geq \eta_1 (1 - \gamma) (V_r^*(\rho) - V_r^{(t)}(\rho)) + \eta_1 (1 - \gamma) \lambda^{(t)} (V_r^*(\rho) - V_g^{(t)}(\rho)) - \eta_1 \text{err}_r^{(t)}(\pi^*) - \eta_1 \lambda^{(t)} \text{err}_g^{(t)}(\pi^*) - \beta \frac{\eta_1^2 W^2}{(1 - \gamma)^2} - \beta \frac{\eta_1^2 W^2}{(1 - \gamma)^2} \lambda^{(t)}
$$

236
where \( \tilde{w}^{(t)} = \hat{w}^{(t)} + \lambda^{(t)} \hat{w}^{(t)} \) for a given \( \lambda^{(t)} \), in the last inequality we apply the performance difference lemma, notation of \( \hat{\text{err}}^{(t)}(\pi^*) \) and \( \text{err}^{(t)}(\pi^*) \), and \( \| \hat{w}^{(t)} \|, \| \hat{w}^{(t)} \| \leq W \).

Rearranging the inequality above leads to,

\[
\begin{align*}
V^*_r(\rho) - V^{(t)}_r(\rho) &\leq \frac{1}{1 - \gamma} \left( \frac{1}{\eta_1} \mathbb{E}_{s \sim d^*} \left( D_{\text{KL}}(\pi^*(\cdot | s) \| \pi^{(t)}(\cdot | s)) - D_{\text{KL}}(\pi^*(\cdot | s) \| \pi^{(t+1)}(\cdot | s)) \right) \right) \\
&\quad + \frac{1}{1 - \gamma} \tilde{\text{err}}^{(t)}(\pi^*) + \frac{2}{(1 - \gamma)^2} \text{err}^{(t)}(\pi^*) + \beta \frac{\eta_1 W^2}{(1 - \gamma)^3} + \beta \frac{4\eta_1 W^2}{(1 - \gamma)^5} - \lambda^{(t)} (V^*_g(\rho) - V^{(t)}_g(\rho))
\end{align*}
\]

where we utilize \( 0 \leq \lambda^{(t)} \leq 2/((1 - \gamma)\xi) \) from the dual update of Algorithm 1.

Therefore,

\[
\begin{align*}
\frac{1}{T} \sum_{t=0}^{T-1} (V^*_r(\rho) - V^{(t)}_r(\rho)) &\leq \frac{1}{(1 - \gamma)\eta_1 T} \sum_{t=0}^{T-1} \left( \mathbb{E}_{s \sim d^*} \left( D_{\text{KL}}(\pi^*(\cdot | s) \| \pi^{(t)}(\cdot | s)) - D_{\text{KL}}(\pi^*(\cdot | s) \| \pi^{(t+1)}(\cdot | s)) \right) \right) \\
&\quad + \frac{1}{(1 - \gamma)T} \sum_{t=0}^{T-1} \tilde{\text{err}}^{(t)}(\pi^*) + \frac{2}{(1 - \gamma)^2} \sum_{t=0}^{T-1} \text{err}^{(t)}(\pi^*) + \beta \frac{\eta_1 W^2}{(1 - \gamma)^3} + \beta \frac{4\eta_1 W^2}{(1 - \gamma)^5} \\
&\quad - \frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)} (V^*_g(\rho) - V^{(t)}_g(\rho)) \\
&\leq \frac{\log |A|}{(1 - \gamma)\eta_1 T} + \frac{1}{(1 - \gamma)T} \sum_{t=0}^{T-1} \tilde{\text{err}}^{(t)}(\pi^*) + \frac{2}{(1 - \gamma)^2} \sum_{t=0}^{T-1} \text{err}^{(t)}(\pi^*) \\
&\quad + \beta \frac{\eta_1 W^2}{(1 - \gamma)^3} + \beta \frac{4\eta_1 W^2}{(1 - \gamma)^5} + \frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)} (V^*_g(\rho) - V^{(t)}_g(\rho))
\end{align*}
\]

where in the last inequality we take a telescoping sum of the first sum and drop a non-positive term. Taking the expectation over the randomness in sampling on both sides of the inequality above yields

\[
\begin{align*}
\mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} (V^*_r(\rho) - V^{(t)}_r(\rho)) \right] &\leq \mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)} (V^*_g(\rho) - V^{(t)}_g(\rho)) \right] \\
&\leq \frac{\log |A|}{(1 - \gamma)\eta_1 T} + \frac{1}{(1 - \gamma)T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \tilde{\text{err}}^{(t)}(\pi^*) \right] + \frac{2}{(1 - \gamma)^2} \sum_{t=0}^{T-1} \mathbb{E} \left[ \text{err}^{(t)}(\pi^*) \right] + \beta \frac{\eta_1 W^2}{(1 - \gamma)^3} + \beta \frac{4\eta_1 W^2}{(1 - \gamma)^5} \tag{A.19}
\end{align*}
\]
Proving the first inequality. From the dual update in Algorithm 1 we have

\[
0 \leq (\lambda(T))^2 = \sum_{t=0}^{T-1} ((\lambda^{(t+1)})^2 - (\lambda^{(t)})^2) \\
\leq \sum_{t=0}^{T-1} ((\lambda^{(t)} - \eta_2 (\hat{V}_g^{(t)}(\rho) - b))^2 - (\lambda^{(t)})^2) \\
= 2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)} (b - \hat{V}_g^{(t)}(\rho)) + \eta_2^2 \sum_{t=0}^{T-1} (\hat{V}_g^{(t)}(\rho) - b)^2 \\
\leq 2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)} (V_g^{*}(\rho) - \hat{V}_g^{(t)}(\rho)) + 2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)} (\hat{V}_g^{(t)}(\rho) - \hat{V}_g^{(t)}(\rho)) \\
+ \eta_2^2 \sum_{t=0}^{T-1} (\hat{V}_g^{(t)}(\rho) - b)^2
\]

where the second inequality is due to the feasibility of the policy $\pi^*$: $V_g^{*}(\rho) \geq b$. Since $\hat{V}_g^{(t)}(\rho)$ is a population quantity and $\hat{V}_g^{(t)}(\rho)$ is an estimate that is independent of $\lambda^{(t)}$ given the past history, $\lambda^{(t)}$ is independent of $V_g^{(t)}(\rho) - \hat{V}_g^{(t)}(\rho)$ at time $t$ and thus $\mathbb{E} [\lambda^{(t)} (V_g^{(t)}(\rho) - \hat{V}_g^{(t)}(\rho))] = 0$ due to the fact $\mathbb{E} [\hat{V}_g^{(t)}(\rho)] = V_g^{(t)}(\rho)$; see it in Appendix A.6. Therefore,

\[
- \mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)} (V_g^{*}(\rho) - \hat{V}_g^{(t)}(\rho)) \right] \leq \mathbb{E} \left[ \frac{\eta_2}{2T} \sum_{t=0}^{T-1} (\hat{V}_g^{(t)}(\rho) - b)^2 \right] \leq \frac{2\eta_2}{(1 - \gamma)^2} \tag{A.20}
\]

where in the second inequality we drop a non-positive term and use the fact

\[
\mathbb{E} \left[ (\hat{V}_g^{(t)}(\rho))^2 \right] = \text{Var} \left[ \hat{V}_g^{(t)}(\rho) \right] + \mathbb{E} \left[ (\hat{V}_g^{(t)}(\rho))^2 \right] \leq \frac{2}{(1 - \gamma)^2}
\]

where the inequality is due to that $\text{Var} [\hat{V}_g^{(t)}(\rho)] \leq 1/(1 - \gamma)^2$; see it in Appendix A.6 and $\mathbb{E} [\hat{V}_g^{(t)}(\rho)] = V_g^{(t)}(\rho)$, where $0 \leq V_g^{(t)}(\rho) \leq 1/(1 - \gamma)$.

Adding the inequality (A.20) to (A.19) on both sides and taking $\eta_1 = \eta_2 = 1/\sqrt{T}$ yield the first inequality.
**Proving the second inequality.** From the dual update in Algorithm 1 we have for any $\lambda \in \Lambda := [0, 1/(1-\gamma)\xi]$, 

\[
\mathbb{E} \left[ |\lambda^{(t+1)} - \lambda|^2 \right] 
= \mathbb{E} \left[ P_\Lambda \left( \lambda^{(t)} - \eta_2 (\hat{V}_g^{(t)}(\rho) - b) \right) - P_\Lambda(\lambda) \right]^2 
\leq (a) \mathbb{E} \left[ |\lambda^{(t)} - \eta_2 (\hat{V}_g^{(t)}(\rho) - b) - \lambda|^2 \right] 
= \mathbb{E} \left[ |\lambda^{(t)} - \lambda|^2 \right] - 2\eta_2 \mathbb{E} \left[ (\hat{V}_g^{(t)}(\rho) - b)(\lambda^{(t)} - \lambda) \right] + \frac{3\eta_2^2}{(1-\gamma)^2}
\leq (b) \mathbb{E} \left[ |\lambda^{(t)} - \lambda|^2 \right] - 2\eta_2 \mathbb{E} \left[ (\hat{V}_g^{(t)}(\rho) - b)(\lambda^{(t)} - \lambda) \right] + \frac{3\eta_2^2}{(1-\gamma)^2}
\]

where (a) is due to the non-expansiveness of projection $P_\Lambda$ and (b) is due to $\mathbb{E} \left[ (\hat{V}_g^{(t)}(\rho) - b)^2 \right] \leq 2/(1-\gamma)^2 + 1/(1-\gamma)^2$. Summing it up from $t = 0$ to $t = T-1$ and dividing it by $T$ yield 

\[
0 \leq \frac{1}{T} \mathbb{E} \left[ |\lambda^{(T)} - \lambda|^2 \right] 
\leq \frac{1}{T} \mathbb{E} \left[ |\lambda^{(0)} - \lambda|^2 \right] - \frac{2\eta_2}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ (\hat{V}_g^{(t)}(\rho) - b)(\lambda^{(t)} - \lambda) \right] + \frac{3\eta_2^2}{(1-\gamma)^2}
\]

which further implies that 

\[
\mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} (V_g^{(t)}(\rho) - b)(\lambda^{(t)} - \lambda) \right] \leq \frac{1}{2\eta_2 T} \mathbb{E} \left[ |\lambda^{(0)} - \lambda|^2 \right] + \frac{2\eta_2}{(1-\gamma)^2}
\]

where we use $\mathbb{E} \left[ \hat{V}_g^{(t)}(\rho) \right] = V_g^{(t)}(\rho)$ and $\lambda^{(t)}$ is independent of $\hat{V}_g^{(t)}(\rho)$ given the past history. We now add the above inequality into (A.19) on both sides and utilize $V_g^*(\rho) \geq b$, 

\[
\mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} (V_r^*(\rho) - V_r^{(t)}(\rho)) \right] + \lambda \mathbb{E} \left[ \frac{1}{T} \sum_{t=0}^{T-1} (b - V_g^{(t)}(\rho)) \right] 
\leq \frac{\log |\mathcal{A}|}{(1-\gamma)\eta_1 T} + \frac{1}{(1-\gamma)T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \text{err}_r^{(t)}(\pi^*) \right] + \frac{2}{(1-\gamma)^2\xi T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \text{err}_g^{(t)}(\pi^*) \right] 
+ \beta \frac{\eta_1 W^2}{(1-\gamma)^3} + \beta \frac{4\eta_1 W^2}{(1-\gamma)^5\xi^2} + \frac{1}{2\eta_2 T} \mathbb{E} \left[ |\lambda^{(0)} - \lambda|^2 \right] + \frac{2\eta_2}{(1-\gamma)^2}.
\]
By taking \( \lambda = \frac{2}{(1-\gamma)\xi} \) when \( \sum_{t=0}^{T-1} (b - V_g^{(t)}(\rho)) \geq 0 \); otherwise \( \lambda = 0 \), we reach

\[
\mathbb{E} \left[ V_r^*(\rho) - \frac{1}{T} \sum_{t=0}^{T-1} V_r^{(t)}(\rho) \right] + \frac{2}{(1-\gamma)\xi} \mathbb{E} \left[ b - \frac{1}{T} \sum_{t=0}^{T-1} V_g^{(t)}(\rho) \right] \\
\leq \frac{\log |\mathcal{A}|}{(1-\gamma)\eta_1 T} + \frac{1}{(1-\gamma)T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \text{err}_r^{(t)}(\pi^*) \right] + \frac{2}{(1-\gamma)^2 T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \text{err}_g^{(t)}(\pi^*) \right] \\
+ \beta \frac{\eta_1 W^2}{(1-\gamma)^3} + \beta \frac{4\eta_1 W^2}{(1-\gamma)^5} \xi^2 + \frac{2}{\eta_2 (1-\gamma)^2 T} + \frac{2\eta_2}{(1-\gamma)^2}.
\]

Since \( V_r^{(t)}(\rho) \) and \( V_g^{(t)}(\rho) \) are linear functions in the occupancy measure \([11, \text{Chapter 10}]\), there exists a policy \( \pi' \) such that \( V_r^{\pi'}(\rho) = \frac{1}{T} \sum_{t=0}^{T-1} V_r^{(t)}(\rho) \) and \( V_g^{\pi'}(\rho) = \frac{1}{T} \sum_{t=0}^{T-1} V_g^{(t)}(\rho) \). Hence,

\[
\mathbb{E} \left[ V_r^*(\rho) - V_r^{\pi'}(\rho) \right] + \frac{2}{(1-\gamma)\xi} \mathbb{E} \left[ b - V_g^{\pi'}(\rho) \right] + \\
\leq \frac{\log |\mathcal{A}|}{(1-\gamma)\eta_1 T} + \frac{1}{(1-\gamma)T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \text{err}_r^{(t)}(\pi^*) \right] + \frac{2}{(1-\gamma)^2 T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \text{err}_g^{(t)}(\pi^*) \right] \\
+ \beta \frac{\eta_1 W^2}{(1-\gamma)^3} + \beta \frac{4\eta_1 W^2}{(1-\gamma)^5} \xi^2 + \frac{2}{\eta_2 (1-\gamma)^2 T} + \frac{2\eta_2}{(1-\gamma)^2}.
\]

Application of Lemma 2 with \( 2/((1-\gamma)\xi) \geq 2\lambda^* \) yields

\[
\mathbb{E} \left[ b - V_g^{\pi'}(\rho) \right] \leq \frac{\xi \log |\mathcal{A}|}{\eta_1 T} + \xi \sum_{t=0}^{T-1} \mathbb{E} \left[ \text{err}_r^{(t)}(\pi^*) \right] + \frac{2}{(1-\gamma)T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \text{err}_g^{(t)}(\pi^*) \right] \\
+ \beta \frac{\eta_1 \xi W^2}{(1-\gamma)^2} + \beta \frac{4\eta_1 W^2}{(1-\gamma)^4} \xi^2 + \frac{2}{\eta_2 (1-\gamma)^2 T} + \frac{2\eta_2}{(1-\gamma)^2}.
\]

which leads to our constraint violation bound by taking \( \eta_1 = \eta_2 = 1/\sqrt{T} \). \( \Box \)

**Proof.** [Proof of Theorem 13]

By Lemma 45 we only need to consider the randomness in sequences of \( \hat{w}^{(t)} \) and bound \( \mathbb{E} \left[ \text{err}_r^{(t)}(\pi^*) \right] \) for \( \odot = r \) or \( g \). Application of the triangle inequality yields

\[
\hat{\text{err}}_r^{(t)}(\pi^*) \leq \left| \mathbb{E}_{s \sim d^\rho} \mathbb{E}_{a \sim \pi^*} \left[ A_r^{(t)}(s,a) - (\hat{w}_r^{(t)})^T \nabla \log \pi_r^{(t)}(a | s) \right] \right| \\
+ \left| \mathbb{E}_{s \sim d^\rho} \mathbb{E}_{a \sim \pi^*} \left[ (\hat{w}_r^{(t)} - w_r^{(t)})^T \nabla \log \pi_r^{(t)}(a | s) \right] \right|
\] (A.21)
where \( w_{r,*}^{(t)} \in \text{argmin}_{\|w_r\|_2 \leq W} E_r^{(t)}(w_r; \theta^{(t)}) \). We next bound each term in the right-hand side of (A.21), separately. For the first term,

\[
\mathbb{E}_{s \sim d_\nu^t} \mathbb{E}_{a \sim \pi^*(\cdot | s)} \left[ A_r^{(t)}(s, a) - (w_{r,*}^{(t)})^\top \nabla_\theta \log \pi_\theta^{(t)}(a | s) \right] \\
\leq \sqrt{\mathbb{E}_{s \sim d_\nu^t} \mathbb{E}_{a \sim \pi^*(\cdot | s)} \left[ \left( A_r^{(t)}(s, a) - (w_{r,*}^{(t)})^\top \nabla_\theta \log \pi_\theta^{(t)}(a | s) \right)^2 \right]} \\
= \sqrt{E_r^{\nu^*}\left( w_{r,*}^{(t)}; \theta^{(t)} \right)}.
\]

Similarly,

\[
\mathbb{E}_{s \sim d_\nu^t} \mathbb{E}_{a \sim \pi^*(\cdot | s)} \left[ (w_{r,*}^{(t)} - \hat{w}_r^{(t)})^\top \nabla_\theta \log \pi_\theta^{(t)}(a | s) \right] \\
\leq \mathbb{E}_{s \sim d_\nu^t} \mathbb{E}_{a \sim \pi^*(\cdot | s)} \left[ \left( (w_{r,*}^{(t)} - \hat{w}_r^{(t)})^\top \nabla_\theta \log \pi_\theta^{(t)}(a | s) \right)^2 \right] \\
= \sqrt{\|w_{r,*}^{(t)} - \hat{w}_r^{(t)}\|^2_{\Sigma_{\nu^*}^{(t)}}}.
\]

We let \( \kappa^{(t)} := \left\| (\Sigma_{\nu_0}^{(t)})^{-1/2} \Sigma_{\nu}^{(t)} (\Sigma_{\nu_0}^{(t)})^{-1/2} \right\|_2 \) be the relative condition number at time \( t \). Thus,

\[
\|w_{r,*}^{(t)} - \hat{w}_r^{(t)}\|^2_{\Sigma_{\nu^*}^{(t)}} \leq \left\| (\Sigma_{\nu_0}^{(t)})^{-1/2} \Sigma_{\nu}^{(t)} (\Sigma_{\nu_0}^{(t)})^{-1/2} \right\| \|w_{r,*}^{(t)} - \hat{w}_r^{(t)}\|_{\Sigma_{\nu}^{(t)}}^2 \\
\leq (a) \frac{\kappa^{(t)}}{1 - \gamma} \|w_{r,*}^{(t)} - \hat{w}_r^{(t)}\|_{\Sigma_{\nu}^{(t)}}^2 \\
\leq (b) \frac{\kappa^{(t)}}{1 - \gamma} \left( E_r^{\nu^*}\left( \hat{w}_r^{(t)}; \theta^{(t)} \right) - E_r^{\nu^*}\left( w_{r,*}^{(t)}; \theta^{(t)} \right) \right)
\]

where we use \( (1 - \gamma) \nu_0 \leq \nu_{\nu_0}^{(t)} := \nu^{(t)} \) in \( (a) \), and we get \( (b) \) due to that the first-order optimality condition for \( w_{r,*}^{(t)} \),

\[
(w_r - w_{r,*}^{(t)})^\top \nabla_\theta E_r^{\nu^{(t)}}(w_{r,*}^{(t)}; \theta^{(t)}) \geq 0, \text{ for any } w_r \text{ satisfying } \|w_r\| \leq W.
\]

Further implies that

\[
E_r^{\nu^{(t)}}\left( w_r; \theta^{(t)} \right) - E_r^{\nu^{(t)}}\left( w_{r,*}^{(t)}; \theta^{(t)} \right) \\
= \mathbb{E}_{s,a \sim \nu^{(t)}} \left[ A_r^{(t)}(s, a) - \phi_{s,a}^T w_{r,*}^{(t)} + \phi_{s,a}^T w_{r,*}^{(t)} - \phi_{s,a}^T w_r \right]^2 \\
= 2 \left( w_{r,*}^{(t)} - w_r \right)^\top \mathbb{E}_{s,a \sim \nu^{(t)}} \left[ \left( A_r^{(t)}(s, a) - \phi_{s,a}^T w_{r,*}^{(t)} \right) \phi_{s,a} \right] \\
+ \mathbb{E}_{s,a \sim \nu^{(t)}} \left[ \left( \phi_{s,a}^T w_{r,*}^{(t)} - \phi_{s,a}^T w_r \right)^2 \right] \\
= (w_r - w_{r,*}^{(t)})^\top \nabla_\theta E_r^{\nu^{(t)}}(w_{r,*}^{(t)}; \theta^{(t)}) + \|w_r - w_{r,*}^{(t)}\|^2_{\Sigma_{\nu}^{(t)}} \\
\geq \|w_r - w_{r,*}^{(t)}\|^2_{\Sigma_{\nu}^{(t)}}.
\]

241
Taking an expectation over (A.24) from both sides yields

\[
E \left[ \left\| w_{r,s}^t - w_{r,s}^t \right\|^2 \right] \leq E \left[ \frac{1}{1 - \gamma} E \left[ E_{\nu}^{(t)} \left( \nu^t; \theta^t \right) - E_{\nu}^{(t)} \left( w^t_{r,s}; \theta^t \right) \right] \right] \\
\leq E \left[ \frac{\kappa(t)}{1 - \gamma} \sqrt{W} \right] \\
\leq \left( \frac{\kappa(t)}{1 - \gamma} \right) \sqrt{\frac{K}{\kappa(t)}} \tag{A.25}
\]

where (a) is due to the standard SGD result [197 Theorem 14.8]: for \( \alpha = W/(G\sqrt{K}) \),

\[
E_{r,est} = E \left[ E_{\nu}^{(t)} \left( \nu^t; \theta^t \right) - E_{\nu}^{(t)} \left( w^t_{r,s}; \theta^t \right) \right] \leq \frac{GW}{\sqrt{K}}
\]

and (b) follows Assumption 5.

Substitution of (A.23), (A.25) into the right-hand side of (A.21) yields an upper bound on \( E \left[ err^t(\pi^*) \right] \). By the same reasoning, we can establish a similar bound on \( E \left[ err_g^t(\pi^*) \right] \). Finally, application of these upper bounds to Lemma 45 leads to our desired results.

\[ \square \]

### A.8 Proof of Theorem 14

By \( \left\| \phi_{s,a} \right\| \leq B \), for the log-linear policy class, \( \log \pi_\theta(a \mid s) \) is \( \beta \)-smooth with \( \beta = B^2 \). By Lemma 45, we only need to consider the randomness in sequences of \( \hat{w}(t) \) and the error bounds for \( E \left[ err^t(\pi^*) \right] \) and \( E \left[ err_g^t(\pi^*) \right] \). We first use (A.21) and consider the following cases. By (2.26) and \( A^t(s,a) = Q^t(s,a) - E_{a' \sim \pi^t(\cdot \mid s)} Q^t(s,a') \),

\[
E_{s \sim d^t_\rho, a \sim \pi^*(\cdot \mid s)} \left[ A^t(s,a) - (w_{r,s}^t)\top \nabla_\theta \log \pi^t(a \mid s) \right] \\
= E_{s \sim d^t_\rho, a \sim \pi^*(\cdot \mid s)} \left[ Q^t(s,a) - \phi_{s,a}^\top w_{r,s}^t \right] \\
- E_{s \sim d^t_\rho, a' \sim \pi^t(\cdot \mid s)} \left[ Q^t(s,a') - \phi_{s,a'}^\top w_{r,s}^t \right] \\
\leq \sqrt{E_{s \sim d^t_\rho, a \sim \pi^*(\cdot \mid s)} \left[ Q^t(s,a) - \phi_{s,a}^\top w_{r,s}^t \right]^2} \\
+ \sqrt{E_{s \sim d^t_\rho, a' \sim \pi^t(\cdot \mid s)} \left[ Q^t(s,a') - \phi_{s,a'}^\top w_{r,s}^t \right]^2} \\
\leq 2 \sqrt{|A| E_{s \sim d^t_\rho, a \sim \text{Unif}_A} \left[ \left( Q^t(s,a) - \phi_{s,a}^\top w_{r,s}^t \right)^2 \right]} \\
= 2 \sqrt{|A| E_{r,s}^{(t)}(w_{r,s}^t; \theta^t)}.
\]
Similarly,
\[
\begin{align*}
\mathbb{E}_{s \sim d_{\rho}^{\pi} \mathbb{E}_{a \sim \pi^* \cdot | s}} & \left[ (w_{r, \star}^{(t)} - \dot{w}_{r}^{(t)})^\top \nabla_{\theta} \log \pi^{(t)}(a | s) \right] \\
& = \mathbb{E}_{s \sim d_{\rho}^{\pi} \mathbb{E}_{a \sim \pi^* \cdot | s}} \left[ (w_{r, \star}^{(t)} - \dot{w}_{r}^{(t)})^\top \phi_{s,a} \right] \\
& - \mathbb{E}_{s \sim d_{\rho}^{\pi} \mathbb{E}_{a \sim \pi^{(t)} \cdot | s}} \left[ (w_{r, \star}^{(t)} - \dot{w}_{r}^{(t)})^\top \phi_{s,a'} \right] \\
& \leq 2 \sqrt{|A| \mathbb{E}_{s \sim d_{\rho}^{\pi} \mathbb{E}_{a \sim \text{Unif}_A}} \left[ \left( (w_{r, \star}^{(t)} - \dot{w}_{r}^{(t)})^\top \phi_{s,a} \right)^2 \right]}
\end{align*}
\]
\[(A.27)\]

where \( \Sigma_{\nu^*} := \mathbb{E}_{(s,a) \sim \nu^*} \left[ \phi_{s,a} \phi_{s,a}^\top \right] \). By the definition of \( \kappa \),
\[
\|w_{r, \star}^{(t)} - \dot{w}_{r}^{(t)}\|_{\Sigma_{\nu^*}}^2 \leq \kappa \|w_{r, \star}^{(t)} - \dot{w}_{r}^{(t)}\|_{\Sigma_{\nu_0}}^2 \leq \frac{\kappa}{1 - \gamma} \|w_{r, \star}^{(t)} - \dot{w}_{r}^{(t)}\|_{\Sigma_{\nu^*(t)}}^2 \tag{A.28}
\]

where we use \( (1 - \gamma)\nu_0 \leq \nu_{\pi^*(t)} := \nu^{(t)} \) in the second inequality. We note that
\[
w_{r, \star}^{(t)} \in \arg\min_{\|w_r\| \leq W} \mathcal{E}_{\nu^{(t)}}^{(t)}(w_r; \theta^{(t)}).
\]

Application of the first-order optimality condition for \( w_{r, \star}^{(t)} \) yields
\[
(w_r - w_{r, \star}^{(t)})^\top \nabla_{\theta} \mathcal{E}_{\nu^{(t)}}^{(t)}(w_{r, \star}^{(t)}; \theta^{(t)}) \geq 0, \text{ for any } w_r \text{ satisfying } \|w_r\| \leq W.
\]

Thus,
\[
\mathcal{E}_{\nu^{(t)}}^{(t)}(w_r; \theta^{(t)}) - \mathcal{E}_{\nu^{(t)}}^{(t)}(w_{r, \star}^{(t)}; \theta^{(t)})
\]
\[
= \mathbb{E}_{s,a \sim \nu^{(t)}} \left[ (Q_{\nu^{(t)}}^{(t)}(s, a) - \phi_{s,a}^\top w_{r, \star}^{(t)} + \phi_{s,a}^\top w_{r, \star}^{(t)} - \phi_{s,a}^\top w_r) \right]^2 - \mathcal{E}_{\nu^{(t)}}^{(t)}(w_{r, \star}^{(t)}; \theta^{(t)})
\]
\[
= 2 (w_{r, \star}^{(t)} - w_r)^\top \mathbb{E}_{s,a \sim \nu^{(t)}} \left[ Q_{\nu^{(t)}}^{(t)}(s, a) - \phi_{s,a}^\top w_{r, \star}^{(t)} \right] \phi_{s,a}
\]
\[
+ \mathbb{E}_{s,a \sim \nu^{(t)}} \left[ \phi_{s,a}^\top w_{r, \star}^{(t)} - \phi_{s,a}^\top w_r \right]^2
\]
\[
= (w_r - w_{r, \star}^{(t)})^\top \nabla_{\theta} \mathcal{E}_{\nu^{(t)}}^{(t)}(w_{r, \star}^{(t)}; \theta^{(t)}) + \|w_r - w_{r, \star}^{(t)}\|_{\Sigma_{\nu^{(t)}}}^2.
\]

Taking \( w_r = \dot{w}_{r}^{(t)} \) in the inequality above and combining it with (A.28) and (A.27) yield
\[
\begin{align*}
\mathbb{E}_{s \sim d_{\rho}^{\pi} \mathbb{E}_{a \sim \pi^* \cdot | s}} & \left[ (w_{r, \star}^{(t)} - \dot{w}_{r}^{(t)})^\top \nabla_{\theta} \log \pi^{(t)}(a | s) \right] \\
& \leq 2 \sqrt{|A| \kappa \left( \mathcal{E}_{\nu^{(t)}}^{(t)}(\dot{w}_{r}^{(t)}; \theta^{(t)}) - \mathcal{E}_{\nu^{(t)}}^{(t)}(w_{r, \star}^{(t)}; \theta^{(t)}) \right)}.
\end{align*}
\]
\[(A.29)\]
We now substitute (A.26) and (A.29) into the right-hand side of (A.21),

\[
\mathbb{E} \left[ \text{err}^{(t)}(\pi^*) \right] \leq 2 \sqrt{|A|} \mathbb{E} \left[ \mathcal{E}_{r}^{\nu} (w_r; \theta(t)) \right] + 2 \sqrt{\frac{|A|}{1-\gamma}} \mathbb{E} \left[ \mathcal{E}_{r}^{\nu} (\hat{w}_r; \theta(t)) - \mathcal{E}_{r}^{\nu} (w_r^*; \theta(t)) \right]
\]

\[
\leq 2 \sqrt{|A|} \mathbb{E} \left[ \mathcal{E}_{r}^{\nu} (w_r^*; \theta(t)) \right] + 2 \sqrt{\frac{|A|}{1-\gamma}} \frac{GW}{\sqrt{K}}
\]

where the second inequality is due to the standard SGD result [197, Theorem 14.8]: for \( \alpha = W/(G\sqrt{K}) \),

\[
\mathcal{E}_{r,\text{est}}^{(t)} = \mathbb{E} \left[ \mathcal{E}_{r}^{\nu} (\hat{w}_r; \theta(t)) - \mathcal{E}_{r}^{\nu} (w_r^*; \theta(t)) \right] \leq \frac{GW}{\sqrt{K}}.
\]

By the same reasoning, we can find a similar bound on \( \mathbb{E} \left[ \text{err}^{(t)}(\pi^*) \right] \). Finally, our desired results follow by applying Assumption 2 and Lemma 45.
Appendix B

Supporting proofs in Chapter 3

B.1 Proof of Theorem 23

As we see in the proof of Theorem 16, our final regret or constraint violation bounds are dominated by the accumulated bonus terms, which come from the design of ‘optimism in the face of uncertainty.’ This framework provides a powerful flexibility for Algorithm 6 to incorporate other optimistic policy evaluation methods. In what follows, we introduce Algorithm 6 with a variant of optimistic policy evaluation.

We repeat notation for readers’ convenience. For any \((h, k) \in [H] \times [K]\), any \((s, a, s') \in S \times A \times S\), and any \((s, a) \in S \times A\), we define two visitation counters \(n^k_h(s, a, s')\) and \(n^k_h(s, a)\) at step \(h\) in episode \(k\),

\[
 n^k_h(s, a, s') = \sum_{\tau=1}^{k-1} 1\{ (s, a, s') = (s_{\tau h}, a_{\tau h}, a_{\tau h+1}) \} \quad \text{and} \quad n^k_h(s, a) = \sum_{\tau=1}^{k-1} 1\{ (s, a) = (s_{\tau h}, a_{\tau h}) \}.
\]

This allows us to estimate transition kernel \(\hat{P}^k_h\), reward function \(\hat{r}^k_h\), and utility function \(\hat{g}^k_h\) for episode \(k\) by

\[
 \hat{P}^k_h(s', s | a) = \frac{n^k_h(s, a, s')}{n^k_h(s, a) + \lambda}, \quad \text{for all } (s, a, s') \in S \times A \times S
\]

\[
 \hat{r}^k_h(s, a) = \frac{1}{n^k_h(s, a) + \lambda} \sum_{\tau=1}^{k-1} 1\{ (s, a) = (s_{\tau h}, a_{\tau h}) \} r_h(s_{\tau h}, a_{\tau h}), \quad \text{for all } (s, a) \in S \times A
\]

\[
 \hat{g}^k_h(s, a) = \frac{1}{n^k_h(s, a) + \lambda} \sum_{\tau=1}^{k-1} 1\{ (s, a) = (s_{\tau h}, a_{\tau h}) \} g_h(s_{\tau h}, a_{\tau h}), \quad \text{for all } (s, a) \in S \times A
\]

where \(\lambda > 0\) is the regularization parameter. Moreover, we introduce the bonus term \(\Gamma^k_h: S \times A \to \mathbb{R}\),

\[
 \Gamma^k_h(s, a) = \beta \left( n^k_h(s, a) + \lambda \right)^{-1/2}
\]
which adapts the counter-based bonus terms in \[19, 108\], where \(\beta > 0\) is to be determined later. Using the estimated transition kernels \(\{\hat{p}_h^k\}_{h=1}^H\), the estimated reward/utility functions \(\{\hat{g}_h^k\}_{h=1}^H\), and the bonus terms \(\{\Gamma_h^k\}_{h=1}^H\), we now can estimate the action-value function via

\[
Q_{o,h}^k(s, a) = \min \left( \hat{\phi}_h^k(s, a) + \sum_{s' \in \mathcal{S}} \hat{P}_h(s' | s, a) V_{o,h+1}^k(s') + 2\Gamma_h^k(s, a), H - h + 1 \right)
\]

for any \((s, a) \in \mathcal{S} \times \mathcal{A}\), where \(\phi = r\) or \(g\). Thus, \(V_{o,h}^k(s) = \langle Q_{o,h}^k(s, \cdot), \pi_h^k(\cdot | s) \rangle_{\mathcal{A}}\). We summarize the above procedure in Algorithm 8. Using already estimated \(\{Q_{r,h}^k(\cdot, \cdot), Q_{g,h}^k(\cdot, \cdot)\}_{h=1}^H\), we can execute the policy improvement and the dual update in Algorithm 6.

Similar to Theorem 16, we prove the following regret and constraint violation bounds.

Proof. [Proof of Theorem 23] The proof is similar to Theorem 16. Since we only change the policy evaluation, all previous policy improvement results still hold. By Lemma 19, we have

\[
\text{Regret}(H) = C_3 H^{2.5} \sqrt{T \log |\mathcal{A}|} + \sum_{k=1}^K \sum_{h=1}^H \left( \mathbb{E}_{\pi^*} [\ell_{r,h}(s_h, a_h)] - \ell_{r,h}(s_h, a_h) \right) + M_{r,H,2}^K
\]

where \(\ell_{r,h}^k\) is the model prediction error given by (3.8) and \(\{M_{r,m}^k\}_{(k,m) \in [K] \times [H] \times [2]}\) is a martingale adapted to the filtration \(\mathcal{F}_{h,m}^k\) in terms of time index \(t\) defined in (3.13). By Lemma 21, it holds with probability \(1 - p/3\) that \(|M_{r,H,2}^K| \leq 4\sqrt{H^2 T \log(4/p)}\). The rest is to bound the double sum term. As shown in Appendix B.3.2, with probability \(1 - p/2\) it holds that for any \((k, h) \in [K] \times [H]\) and \((s, a) \in \mathcal{S} \times \mathcal{A}\),

\[
-4\Gamma_h^k(s, a) \leq \ell_{r,h}^k(s, a) \leq 0.
\]

Together with the choice of \(\Gamma_h^k\), we have

\[
\sum_{k=1}^K \sum_{h=1}^H \left( \mathbb{E}_{\pi^*} [\ell_{r,h}(s_h, a_h) | s_1] - \ell_{r,h}(s_h, a_h) \right) \leq 4 \sum_{k=1}^K \sum_{h=1}^H \Gamma_h^k(s_h, a_h) = 4\beta \sum_{k=1}^K \sum_{h=1}^H (n_h^k(s_h, a_h) + \lambda)^{-1/2}.
\]

Define mapping \(\phi: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^{(|\mathcal{S}| \times |\mathcal{A}|)}\) as \(\phi(s, a) = e_{(s,a)}\), we can utilize Lemma 52. For any \((k, h) \in [K] \times [H]\), we have

\[
\bar{\Lambda}_h^k = \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) \phi(s_h^\tau, a_h^\tau)^\top + \lambda I \in \mathbb{R}^{(|\mathcal{S}| \times |\mathcal{A}| \times |\mathcal{S}| \times |\mathcal{A}|)}
\]

\[
\Gamma_h^k(s, a) = \beta (n_h^k(s, a) + \lambda)^{-1/2} = \beta \sqrt{\phi(s, a)(\bar{\Lambda}_h^k)^{-1} \phi(s, a)^\top}
\]
where $\bar{A}^k_h$ is a diagonal matrix whose the $(s, a)$th diagonal entry is $n^k_h(s, a) + \lambda$. Therefore, we have

$$\sum_{k=1}^K \sum_{h=1}^H (\mathbb{E}_{\pi^*} [\ell^k_{r,h}(s_h, a_h)] - \ell^k_{r,h}(s_h^k, a_h^k)) \leq 4\beta \sum_{k=1}^K \sum_{h=1}^H \left( \bar{\phi}(s_h^k, a_h^k)(\bar{A}^k_h)^{-1}\bar{\phi}(s_h^k, a_h^k)^\top \right)^{1/2} \leq 4\beta \sum_{h=1}^H \left( K \sum_{k=1}^K \bar{\phi}(s_h^k, a_h^k)(\bar{A}^k_h)^{-1}\bar{\phi}(s_h^k, a_h^k)^\top \right)^{1/2} \leq 4\beta \sqrt{2K} \sum_{h=1}^H \log^{1/2} \left( \frac{\det (\bar{A}_h^K + 1)}{\det \bar{A}_h^1} \right)$$

where we apply the Cauchy-Schwartz inequality for the second inequality and Lemma 52 for the third inequality. Notice that $(K + \lambda) I \geq \bar{A}_h^K$ and $\bar{A}_h^1 = \lambda I$. Hence,

$$\text{Regret}(K) = C_3 H^{2.5} \sqrt{T \log |A| + 4\beta \sqrt{2|S||A|HT} \log \left( \frac{K + \lambda}{\lambda} \right) + 4 \sqrt{H^2T \log \left( \frac{6}{\lambda} \right)}}.$$

Notice that $\log |A| \leq O(|S|^2|A| \log^2(|S||A|T/p))$. We conclude the desired regret bound by setting $\lambda = 1$ and $\beta = C_1 H \sqrt{|S| \log(|S||A|T/p)}$.

For the constraint violation analysis, Lemmas 22 still holds. Similar to (3.34), we have

$$V_{r,1}^*(s_1) - V_{r,1}^g(s_1) + \chi \left[ b - V_{g,1}^g(s_1) \right]_+ \leq \frac{C_4 H^{2.5} \sqrt{T \log |A|}}{K} + \frac{4K}{K} \sum_{k=1}^K \sum_{h=1}^H \Gamma^k_h(s_h^k, a_h^k) + \frac{4 \chi}{K} \sum_{k=1}^K \sum_{h=1}^H \Gamma^k_h(s_h^k, a_h^k) + \frac{1}{K} M_{r,H,2}^K + \frac{\chi}{K} |M_{g,H,2}^K|$$

where $V_{r,1}^*(s_1) = \frac{1}{K} \sum_{k=1}^K V_{r,1}^k(s_1)$ and $V_{g,1}^g(s_1) = \frac{1}{K} \sum_{k=1}^K V_{g,1}^k(s_1)$. Similar to Lemma 21, it holds with probability $1 - p/3$ that $|M_{r,H,2}^K| \leq 4 \sqrt{H^2T \log(6/p)}$ for $\phi = r$ or $g$. As shown in Appendix B.3.2, with probability $1 - p/3$ it holds that $-4\Gamma^k_h(s, a) \leq \ell^k_{s,h}(s, a) \leq 0$ for any $(k, h) \in [K] \times [H]$ and $(s, a) \in S \times A$. Therefore, we have

$$V_{r,1}^*(s_1) - V_{r,1}^g(s_1) + \chi \left[ b - V_{g,1}^g(s_1) \right]_+ \leq \frac{C_4 H^{2.5} \sqrt{T \log |A|}}{K} + \frac{4(1 + \chi)\beta \sqrt{2|S||A|HT}}{K} \log \left( \frac{K + \lambda}{\lambda} \right) + \frac{4(1 + \chi)}{K} \sqrt{H^2T \log \left( \frac{6}{p} \right)}$$

which leads to the desired constraint violation bound due to Lemma 49 and we set $\lambda$ and $\beta$ as previously. □
B.2 Proof of Formulas (3.9) and (3.11)

For any \((k, h) \in [K] \times [H]\), we recall the definitions of \(V_{r,h}^{\pi_s^*}\) in the Bellman equations (3.1) and \(V_{r,h}^{\pi_s}\) from line 12 in Algorithm 7:

\[
V_{r,h}^{\pi_s^*}(s) = \langle Q_{h}^{\pi_s^*}(s, \cdot), \pi_h^*(\cdot | s) \rangle \quad \text{and} \quad V_{r,h}^{\pi_s}(s) = \langle Q_{h}^{\pi_s}(s, \cdot), \pi_h^*(\cdot | s) \rangle.
\]

We can expand the difference \(V_{r,h}^{\pi_s^*}(s) - V_{r,h}^{\pi_s}(s)\) into

\[
V_{r,h}^{\pi_s^*}(s) - V_{r,h}^{\pi_s}(s) = \langle Q_{h}^{\pi_s^*}(s, \cdot), \pi_h^*(\cdot | s) \rangle - \langle Q_{h}^{\pi_s}(s, \cdot), \pi_h^*(\cdot | s) \rangle
= \langle Q_{h}^{\pi_s^*}(s, \cdot) - Q_{h}^{\pi_s}(s, \cdot), \pi_h^*(\cdot | s) \rangle + \langle Q_{h}^{\pi_s}(s, \cdot) - \pi_h^*(\cdot | s) \rangle
= \langle Q_{h}^{\pi_s^*}(s, \cdot) - Q_{h}^{\pi_s}(s, \cdot), \pi_h^*(\cdot | s) \rangle + \xi_h^k(s)
\]

where \(\xi_h^k(s) := \langle Q_{h}^{\pi_s^*}(s, \cdot), \pi_h^*(\cdot | s) - \pi_h^*(\cdot | s) \rangle\).

Recall the equality in the Bellman equations (3.1) and the model prediction error,

\[
Q_{r,h}^{\pi_s^*} = r_h + \mathbb{P}_h V_{r,h+1}^{\pi_s^*} \quad \text{and} \quad \iota_{r,h}^k = r_h + \mathbb{P}_h V_{r,h+1}^{\pi_s} - Q_{r,h}.
\]

As a result of the above two, it is easy to see that

\[
Q_{r,h}^{\pi_s^*} - Q_{r,h}^{\pi_s} = \mathbb{P}_h (V_{r,h+1}^{\pi_s^*} - V_{r,h+1}^{\pi_s}) + \iota_{r,h}^k.
\]

Substituting the above difference into the right-hand side of (B.2) yields,

\[
V_{r,h}^{\pi_s^*}(s) - V_{r,h}^{\pi_s}(s) = \langle \mathbb{P}_h (V_{r,h+1}^{\pi_s^*} - V_{r,h+1}^{\pi_s}), \pi_h^*(\cdot | s) \rangle + \langle \iota_{r,h}^k(s), \pi_h^*(\cdot | s) \rangle + \xi_h^k(s),
\]

which displays a recursive formula over \(h\). Thus, we expand \(V_{r,1}^{\pi_s^*}(s_1) - V_{r,1}^{\pi_s}(s_1)\) recursively with \(x = s_1\) as

\[
V_{r,1}^{\pi_s^*}(s_1) - V_{r,1}^{\pi_s}(s_1) = \langle \mathbb{P}_1 (V_{r,2}^{\pi_s^*} - V_{r,2}^{\pi_s})(s_1, \cdot), \pi_2^*(\cdot | s_1) \rangle + \langle \iota_{r,1}^k(s_1, \cdot), \pi_1^*(\cdot | s_1) \rangle + \xi_{r,1}^k(s_1)
= \langle \mathbb{P}_1 \langle \mathbb{P}_2 (V_{r,3}^{\pi_s^*} - V_{r,3}^{\pi_s})(x_2, \cdot), \pi_3^*(\cdot | x_2) \rangle \rangle (s_1, \cdot), \pi_2^*(\cdot | s_1) \rangle
+ \langle \mathbb{P}_1 \langle \iota_{r,2}^k(x_2, \cdot), \pi_2^*(\cdot | x_2) \rangle \rangle (s_1, \cdot), \pi_1^*(\cdot | s_1) \rangle + \langle \iota_{r,1}^k(s_1, \cdot), \pi_1^*(\cdot | s_1) \rangle
+ \langle \mathbb{P}_1 \xi_{r,2}^k(s_1, \cdot), \pi_1^*(\cdot | s_1) \rangle + \xi_{r,1}^k(s_1).
\]

For notational simplicity, for any \((k, h) \in [K] \times [H]\), we define an operator \(\mathcal{I}_h\) for function \(f : S \times A \rightarrow \mathbb{R}\),

\[
(\mathcal{I}_hf)(s) = \langle f(s, \cdot), \pi_h^*(\cdot | s) \rangle.
\]

248
With this notation, repeating the above recursion \([B.3]\) over \(h \in [H]\) yields

\[
V_{r,h}^\pi(s_1) - V_{r,1}^k(s_1) = I_1^k P_1 I_2^k P_2^k (V_{r,3}^\pi - V_{r,1}^k) + I_1^k P_1 I_2^k t_{r,2}^k + I_1^k t_{r,1}^k + I_1^k P_1 \xi_2^k + \xi_1^k
\]

\[
= I_1^k P_1 I_2^k P_2^k I_3^k P_3^k (V_{r,4}^\pi - V_{r,1}^k) + I_1^k P_1 I_2^k P_2^k I_3^k t_{r,3}^k + I_1^k P_1 I_2^k t_{r,2}^k + I_1^k t_{r,1}^k + I_1^k I_2^k P_2^k \xi_3^k + I_1^k P_1 \xi_2^k + \xi_1^k
\]

\[
\vdots
\]

\[
= \left( \prod_{h=1}^{H} I_h P_h \right) (V_{r,H+1}^\pi - V_{r,H+1}^k) + \sum_{h=1}^{H-1} \left( \prod_{i=1}^{h} I_i P_i \right) I_{h+1}^k + \sum_{h=1}^{H-1} \left( \prod_{i=1}^{h} I_i P_i \right) \xi_h^k.
\]

Finally, notice that \(V_{r,H+1}^\pi = V_{r,H+1}^k = 0\), we use the definitions of \(P_h\) and \(I_h\) to conclude \([3.9]\). Similarly, we can also use the above argument to verify \([3.11]\).

**B.3 Proof of Formulas \([3.14]\) and \([3.15]\)**

We recall the definition of \(V_{r,h}^\pi\) and define an operator \(I_h^k\) for function \(f: S \times A \to \mathbb{R}\),

\[
V_{r,h}^k(s) = \langle Q_h^\pi(s, \cdot), \pi_h^k(\cdot | s) \rangle \text{ and } (I_h^k f)(s) = \langle f(s, \cdot), \pi_h^k(\cdot | s) \rangle.
\]

We expand the model prediction error \(t_{r,h}^k\) into,

\[
t_{r,h}^k(s_h, a_h) = r_h(s_h, a_h) + (P_h V_{r,h+1}^\pi(s_h, a_h)) - Q_{r,h}^k(s_h, a_h)
\]

\[
= (r_h(s_h, a_h) + (P_h V_{r,h+1}^\pi(s_h, a_h)) - Q_{r,h}^k(s_h, a_h)) + (Q_{r,h}^k(s_h, a_h) - Q_{r,h}^k(s_h, a_h))
\]

\[
= (P_h V_{r,h+1}^\pi - P_h V_{r,h+1}^\pi)(s_h, a_h) + (Q_{r,h}^\pi(s_h, a_h) - Q_{r,h}^k(s_h, a_h))
\]

where we use the Bellman equation \(Q_{r,h}^\pi(s_h, a_h) = r_h(s_h, a_h) + (P_h V_{r,h+1}^\pi(s_h, a_h))\) in the last equality. With the above formula, we expand the difference \(V_{r,1}^k(s_1) - V_{r,1}^\pi(s_1)\) into

\[
V_{r,h}^k(s_h) - V_{r,h}^\pi(s_h) = (I_h^k(Q_{r,h}^k - Q_{r,h}^\pi))(s_h) - t_{r,h}^k(s_h, a_h)
\]

\[
+ (P_h V_{r,h+1}^k - P_h V_{r,h+1}^\pi)(s_h, a_h) + (Q_{r,h}^\pi - Q_{r,h}^k)(s_h, a_h)
\]

Let

\[
D_{r,h,1}^k := (I_h^k(Q_{r,h}^k - Q_{r,h}^\pi))(s_h) - (Q_{r,h}^k - Q_{r,h}^\pi)(s_h, a_h),
\]

\[
D_{r,h,2}^k := (P_h V_{r,h+1}^k - P_h V_{r,h+1}^\pi)(s_h, a_h) - (V_{r,h+1}^k - V_{r,h+1}^\pi)(s_{h+1}).
\]
Therefore, we have the following recursive formula over $h$,

\[
V^k_r(s^k_h) - V^\pi^k_r(s^k_h) = D^k_{r,h,1} + D^k_{r,h,2} + \left( V^k_r(s^k_{h+1}) - V^\pi^k_r(s^k_{h+1}) \right) (s^k_{h+1}) - l^k_{r,h}(s^k_h, a^k_h).
\]

Notice that $V^\pi^k_{r,H+1} = V^k_{r,H+1} = 0$. Summing the above equality over $h \in [H]$ yields

\[
V^k_r(s_1) - V^\pi^k_r(s_1) = \sum_{h=1}^{H} \left( D^k_{r,h,1} + D^k_{r,h,2} \right) - \sum_{h=1}^{H} l^k_{r,h}(s^k_h, a^k_h).
\]

(B.4)

Following the definitions of $\mathcal{F}^k_{h,1}$ and $\mathcal{F}^k_{h,2}$, we know $D^k_{r,h,1} \in \mathcal{F}^k_{h,1}$ and $D^k_{r,h,2} \in \mathcal{F}^k_{h,2}$. Thus, for any $(k, h) \in [K] \times [H]$,

\[
\mathbb{E} [D^k_{r,h,1} | \mathcal{F}^k_{h-1,2}] = 0 \quad \text{and} \quad \mathbb{E} [D^k_{r,h,2} | \mathcal{F}^k_{h,1}] = 0.
\]

Notice that $t(k, 0, 2) = t(k - 1, H, 2) = 2H(k - 1)$. Clearly, $\mathcal{F}^k_{0,2} = \mathcal{F}^{k-1}_{H,2}$ for any $k \geq 2$. Let $\mathcal{F}^0_{0,2}$ be empty. We define a martingale sequence,

\[
M^k_{r,h,m} = \sum_{\tau = 1}^{k-1} \sum_{i = 1}^{H} (D^\tau_{r,i,1} + D^\tau_{r,i,2}) + \sum_{i = 1}^{h-1} (D^k_{r,i,1} + D^k_{r,i,2}) + \sum_{\ell = 1}^{m} D^k_{r,h,\ell}
\]

\[
= \sum_{(\tau,i,\ell) \in [K] \times [H] \times [2], t(\tau,i,\ell) \leq t(k,h,m)} D^\tau_{r,i,\ell}
\]

where $t(k, h, m) := 2(k - 1)H + 2(h - 1) + m$ is the time index. Clearly, this martingale is adapted to the filtration $\{\mathcal{F}^k_{h,m}\}_{(k,h,m) \in [K] \times [H] \times [2]}$, and particularly,

\[
\sum_{k=1}^{K} \sum_{h=1}^{H} (D^k_{r,h,1} + D^k_{r,h,2}) = M^K_{r,H,2}.
\]

Finally, we combine the above martingale with (B.4) to obtain (3.14). It is similar for (3.15).

**B.3.1 Proof of Formula (3.16)**

We recall the definition of the feature map $\phi^k_{r,h}$,

\[
\phi^k_{r,h}(s, a) = \int_{s'} \psi(s, a, s') V^k_{r,h+1}(s') ds'.
\]
for any \((k, h) \in [K] \times [H]\) and \((s, a) \in \mathcal{S} \times \mathcal{A}\). By Assumption 8 we have

\[
(P_h V_{r,h+1}^k) (s, a) = \int_{\mathcal{S}} \psi(s, a, s') \top \theta_h \cdot V_{r,h+1}^k(s') ds'
\]

\[
= \phi_{r,h}^k(s, a) \top \theta_h
\]

\[
= \phi_{r,h}^k(s, a) \top (\Lambda_{r,h}^k)^{-1} \Lambda_{r,h}^k \theta_h
\]

\[
= \phi_{r,h}^k(s, a) \top (\Lambda_{r,h}^k)^{-1} \left( \sum_{\tau=1}^{k-1} \phi_{r,h}^\tau(s_h^\tau, a_h^\tau) \phi_{r,h}^\tau(s_h^\tau, a_h^\tau) \top \theta_h + \lambda \theta_h \right)
\]

\[
= \phi_{r,h}^k(s, a) \top (\Lambda_{r,h}^k)^{-1} \left( \sum_{\tau=1}^{k-1} \phi_{r,h}^\tau(s_h^\tau, a_h^\tau) \cdot (P_h V_{r,h+1}^\tau) (s_h^\tau, a_h^\tau) + \lambda \theta_h \right)
\]

where the second equality is due to the definition of \(\phi_{r,h}^k\), we exploit \(\Lambda_{r,h}^k\) from line 4 of Algorithm 7 in the fourth equality, and we recursively replace \(\phi_{r,h}^\tau(s_h^\tau, a_h^\tau) \top \theta_h\) by \((P_h V_{r,h+1}^\tau) (s_h^\tau, a_h^\tau)\) for all \(\tau \in [k-1]\) in the last equality.

We recall the update \(w_{r,h}^k = (\Lambda_{r,h}^k)^{-1} \sum_{\tau=1}^{k-1} \phi_{r,h}^\tau(s_h^\tau, a_h^\tau) V_{r,h+1}^\tau (s_h^\tau, a_h^\tau)\) from line 5 of Algorithm 7. Therefore,

\[
\left| \phi_{r,h}^k(s, a) \top w_{r,h}^k - (P_h V_{r,h+1}^k) (s, a) \right|
\]

\[
= \left| \phi_{r,h}^k(s, a) \top (\Lambda_{r,h}^k)^{-1} \sum_{\tau=1}^{k-1} \phi_{r,h}^\tau(s_h^\tau, a_h^\tau) \cdot (V_{r,h+1}^\tau(s_h^\tau, a_h^\tau) - (P_h V_{r,h+1}^\tau) (s_h^\tau, a_h^\tau)) \right|
\]

\[
+ \left| \lambda \cdot \phi_{r,h}^k(s, a) \top (\Lambda_{r,h}^k)^{-1} \theta_h \right|
\]

\[
\leq \left( \phi_{r,h}^k(s, a) \top (\Lambda_{r,h}^k)^{-1} \phi_{r,h}^k(s, a) \right)^{1/2} \left| \sum_{\tau=1}^{k-1} \phi_{r,h}^\tau(s_h^\tau, a_h^\tau) \left( V_{r,h+1}^\tau(s_h^\tau, a_h^\tau) - (P_h V_{r,h+1}^\tau)(s_h^\tau, a_h^\tau) \right) \right| (\Lambda_{r,h}^k)^{-1}
\]

\[
+ \lambda \left( \phi_{r,h}^k(s, a) \top (\Lambda_{r,h}^k)^{-1} \phi_{r,h}^k(s, a) \right)^{1/2} \left\| \theta_h \right\| (\Lambda_{r,h}^k)^{-1}
\]

for any \((k, h) \in [K] \times [H]\) and \((s, a) \in \mathcal{S} \times \mathcal{A}\), where we apply the Cauchy-Schwarz inequality twice in the inequality. By Lemma 51 set \(\lambda = 1\), with probability \(1 - p/2\) it holds that

\[
\left| \sum_{\tau=1}^{k-1} \phi_{r,h}^\tau(s_h^\tau, a_h^\tau) \cdot (V_{r,h+1}^\tau(s_h^\tau, a_h^\tau) - (P_h V_{r,h+1}^\tau)(s_h^\tau, a_h^\tau)) \right| (\Lambda_{r,h}^k)^{-1}
\]

\[
\leq C \sqrt{dH^2 \log \left( \frac{dT}{p} \right)}.
\]

Also notice that \(\Lambda_{r,h}^k \succeq \lambda I\) and \(\left\| \theta_h \right\| \leq \sqrt{d}\), thus \(\left\| \theta_h \right\| (\Lambda_{r,h}^k)^{-1} \leq \sqrt{\lambda d}\). Thus, by taking an appropriate absolute constant \(C\), we obtain that

\[
\left| \phi_{r,h}^k(s, a) \top w_{r,h}^k - (P_h V_{r,h+1}^k) (s, a) \right| \leq C \left( \phi_{r,h}^k(s, a) \top (\Lambda_{r,h}^k)^{-1} \phi_{r,h}^k(s, a) \right)^{1/2} \sqrt{dH^2 \log \left( \frac{dT}{p} \right)}
\]

for any \((k, h) \in [K] \times [H]\) and \((s, a) \in \mathcal{S} \times \mathcal{A}\) under the event of Lemma 51.
We now set $C > 1$ and $\beta = C \sqrt{dH^2 \log (dT/p)}$. By the exploration bonus $\Gamma^k_{r,h}$ in line 7 of Algorithm\textsuperscript{7} with probability $1 - p/2$ it holds that
\[
|\phi^k_{r,h}(s, a)^\top w^k_{r,h} - (\mathbb{P}_h V^k_{r,h+1})(x, a)| \leq \Gamma^k_{r,h}(s, a) \tag{B.5}
\]
for any $(k, h) \in [K] \times [H]$ and $(s, a) \in S \times A$.

We note that reward/utility functions are fixed over episodes, $r_h(s^*_h, a^*_h) := \varphi(s^*_h, a^*_h)^\top \theta_{r,h}$ For the difference $\varphi(s, a)^\top u^k_{r,h} - r_h(s, a)$, we have
\[
\begin{align*}
|\varphi(s, a)^\top u^k_{r,h} - r_h(s, a)| &= |\varphi(s, a)^\top u^k_{r,h} - \varphi(s, a)^\top \theta_{r,h}| \\
&= |\varphi(s, a)^\top (\Lambda^k_h)^{-1} \left( \sum_{\tau=1}^{k-1} \varphi(s^\tau_h, a^\tau_h) r_h(s^\tau_h, a^\tau_h) - \Lambda^k_h \theta_{r,h} \right)| \\
&= |\varphi(s, a)^\top (\Lambda^k_h)^{-1} \left( \sum_{\tau=1}^{k-1} \varphi(s^\tau_h, a^\tau_h) (r_h(s^\tau_h, a^\tau_h) - \varphi(s^\tau_h, a^\tau_h)^\top \theta_{r,h}) + \lambda \theta_{r,h} \right)| \\
&= \lambda |\varphi(s, a)^\top (\Lambda^k_h)^{-1} \theta_{r,h}| \\
&\leq \lambda (|\varphi(s, a)^\top (\Lambda^k_h)^{-1} \varphi(s, a)|)^{1/2} \|\theta_{r,h}\|_{(\Lambda^k_h)^{-1}}
\end{align*}
\]
where we apply the Cauchy-Schwartz inequality in the inequality. Notice that $\Lambda^k_h \geq \lambda I$ and $\|\theta_{r,h}\| \leq \sqrt{d}$, thus $\|\theta_{r,h}\|_{(\Lambda^k_h)^{-1}} \leq \sqrt{\lambda d}$. Hence, if we set $\lambda = 1$ and $\beta = C \sqrt{dH^2 \log (dT/p)}$, then any $(k, h) \in [K] \times [H]$ and $(s, a) \in S \times A$,
\[
|\varphi(s, a)^\top u^k_{r,h} - r_h(s, a)| \leq \Gamma^k_{r,h}(s, a). \tag{B.6}
\]

We recall the model prediction error $r^k_{r,h} := r_h + \mathbb{P}_h V^k_{r,h+1} - Q^k_{r,h}$ and the estimated state-action value function $Q^k_{r,h}$ in line 11 of Algorithm\textsuperscript{7}
\[
Q^k_{r,h}(s, a) = \min \left( \varphi(s, a)^\top u^k_{r,h} + \phi^k_{r,h}(s, a)^\top w^k_{r,h} + (\Lambda^k_h + \Gamma^k_{r,h})(s, a), H - h + 1 \right)^+
\]
for any $(k, h) \in [K] \times [H]$ and $(s, a) \in S \times A$. By (B.5) and (B.6), we first have
\[
\phi^k_{r,h}(s, a)^\top w^k_{r,h} + \Gamma^k_{r,h}(s, a) \geq 0 \quad \text{and} \quad \varphi(s, a)^\top u^k_{r,h} + \Gamma^k_{r,h}(s, a) \geq 0.
\]

Then, we can show that
\[
\begin{align*}
-w^k_{r,h}(s, a) &\leq Q^k_{r,h}(s, a) - (r_h + \mathbb{P}_h V^k_{r,h+1})(s, a) \\
&\leq \varphi(s, a)^\top u^k_{r,h} + \phi^k_{r,h}(s, a)^\top w^k_{r,h} + (\Lambda^k_h + \Gamma^k_{r,h})(s, a) - (r_h + \mathbb{P}_h V^k_{r,h+1})(s, a) \\
&\leq (\varphi(s, a)^\top u^k_{r,h} - r_h(s, a)) + \Gamma^k_{r,h}(s, a) + 2\Gamma^k_{r,h}(s, a)
\end{align*}
\]
for any $(k, h) \in [K] \times [H]$ and $(s, a) \in S \times A$. 252
Therefore, (B.7) reduces to
\[-t^k_{r,h}(s, a) \leq 2\Gamma^k_h(s, a) + 2\Gamma^k_{r,h}(s, a) = 2(\Gamma^k_h + \Gamma^k_{r,h})(s, a) .\]

On the other hand, notice that \((r^k_h + P^k_h V^k_{r,h+1})(s, a) \leq H - h + 1\), thus
\[t^k_{r,h}(s, a) = (r^k_h + P^k_h V^k_{r,h+1})(s, a) - Q^k_{r,h}(s, a) \leq (r^k_h + P^k_h V^k_{r,h+1})(s, a) - \min \{ \varphi(s, a)\top w^k_{r,h} + \phi^k_{r,h}(s, a)\top w^k_{r,h} + (\Gamma^k_h + \Gamma^k_{r,h})(s, a), H - h + 1 \}^+ \]
\[\leq \max (r^k_h(s, a) - \varphi(s, a)\top w^k_{r,h} - \Gamma^k_h(s, a) + (P^k_h V^k_{r,h+1})(s, a) - \phi^k_{r,h}(s, a)\top w^k_{r,h} - \Gamma^k_{r,h}(s, a), 0)^+ \]
\[\leq 0\]
for any \((k, h) \in [K] \times [H]\) and \((s, a) \in S \times A\).

Therefore, we have proved that with probability \(1 - p/2\) it holds that
\[-2(\Gamma^k_h + \Gamma^k_{r,h})(s, a) \leq t^k_{r,h}(s, a) \leq 0\]
for any \((k, h) \in [K] \times [H]\) and \((s, a) \in S \times A\).

Similarly, we can show another inequality \(-2(\Gamma^k_h + \Gamma^k_{g,h})(s, a) \leq t^k_{g,h}(s, a) \leq 0\).

**B.3.2 Proof of Formula (B.1)**

Let \(V = \{V : S \rightarrow [0, H]\}\) be a set of bounded function on \(S\). Fo any \(V \in V\), we consider the difference between \(\sum_{s' \in S} \bar{P}^k_h(s' | s, a)V(s')\) and \(\sum_{s' \in S} P^k_h(s' | s, a)V(s')\) as follows,

\[
\left( n^k_h(s, a) + \lambda \right)^{1/2} \left| \sum_{s' \in S} \left( \bar{P}^k_h(s' | s, a)V(s') - P^k_h(s' | s, a)V(s') \right) \right| \\
= \left( n^k_h(s, a) + \lambda \right)^{-1/2} \left| \sum_{s' \in S} n^k_h(s, a, s')V(s') - \left( n^k_h(s, a) + \lambda \right)(P^k_h V)(s, a) \right| \\
\leq \left( n^k_h(s, a) + \lambda \right)^{-1/2} \left| \sum_{s' \in S} n^k_h(s, a, s')V(s') - n^k_h(s, a)(P^k_h V)(s, a) \right| \\
+ \left( n^k_h(s, a) + \lambda \right)^{-1/2} \left| \lambda(P^k_h V)(s, a) \right| \\
= \left( n^k_h(s, a) + \lambda \right)^{-1/2} \left| \sum_{\tau = 1}^{k-1} 1\{ (s, a) = (s^\tau_h, a^\tau_h) \} \left( (V(x^\tau_h) - (P^k_h V)(s, a) \right) \\
+ \left( n^k_h(s, a) + \lambda \right)^{-1/2} \left| \lambda(P^k_h V)(s, a) \right| \\
for any \((k, h) \in [K] \times [H]\) and \((s, a) \in S \times A\), where we apply the triangle inequality for the inequality.
Let $\eta_k^\tau := V(x_{h+1}^\tau) - (\P_h V)(s_h^\tau, a_h^\tau)$. Conditioning on the filtration $\mathcal{F}_{h+1}$, $\eta_k^\tau$ is a zero-mean and $H/2$-subGaussian random variable. By Lemma 50, we use $Y = \lambda I$ and $X_\tau = 1\{(s, a) = (s_h^\tau, a_h^\tau)\}$ and thus with probability at least $1 - \delta$ it holds that

$$
(n_h^k(s, a) + \lambda)^{-1/2} \sum_{\tau = 1}^{k-1} 1\{(s, a) = (s_h^\tau, a_h^\tau)\} (V(x_{h+1}^\tau) - (\P_h V)(s, a)) \leq \sqrt{2} \log \left( \frac{(n_h^k(s, a) + \lambda)^{1/2} \lambda^{-1/2}}{\delta / H} \right) \leq \sqrt{2} \log \left( \frac{\lambda}{\delta} \right)
$$

for any $(k, h) \in [K] \times [H]$. Also, since $0 \leq V \leq H$, we have

$$(n_h^k(s, a) + \lambda)^{-1/2} |\lambda(\P_h V)(s, a)| \leq \sqrt{\lambda H}.$$ 

By returning to (B.8) and setting $\lambda = 1$, with probability at least $1 - \delta$ it holds that

$$(n_h^k(s, a) + \lambda) \left| \sum_{s' \in S} \left( \P_h^k(s' | s, a)V(s') - \P_h(s' | s, a)V(s) \right) \right|^2 \leq H^2 \left( \log \left( \frac{T}{\delta} \right) + 2 \right)$$

for any $k \geq 1$.

Let $d(V, V') = \max_{s \in S} |V(s) - V'(s)|$ be a distance on $\mathcal{V}$. For any $\epsilon$, an $\epsilon$-covering $\mathcal{V}_\epsilon$ of $\mathcal{V}$ with respect to distance $d(\cdot, \cdot)$ satisfies

$$|\mathcal{V}_\epsilon| \leq \left( 1 + \frac{2\sqrt{|S|}H}{\epsilon} \right)^{|S|}.$$ 

Thus, for any $V \in \mathcal{V}$, there exists $V' \in \mathcal{V}_\epsilon$ such that $\max_{s \in S} |V(s) - V'(s)| \leq \epsilon$. By the triangle inequality, we have

$$
(n_h^k(s, a) + \lambda)^{-1/2} \sum_{s' \in S} \left( \P_h^k(s' | s, a)V(s') - \P_h(s' | s, a)V(s) \right) \leq (n_h^k(s, a) + \lambda)^{-1/2} \sum_{s' \in S} \left( \P_h^k(s' | s, a)V'(s') - \P_h(s' | s, a)V'(s) \right) + (n_h^k(s, a) + \lambda)^{1/2} \sum_{s' \in S} \left( \P_h^k(s' | s, a)(V(s') - V'(s')) - \P_h(s' | s, a)(V(s') - V'(s')) \right) \leq (n_h^k(s, a) + \lambda)^{1/2} \sum_{s' \in S} \left( \P_h^k(s' | s, a)V'(s') - \P_h(s' | s, a)V'(s) \right) + 2(n_h^k(s, a) + \lambda)^{-1/2} \epsilon.
$$
Furthermore, we choose $\delta = (p/3) / (|\mathcal{V}| |\mathcal{S}| |\mathcal{A}|)$ and take a union bound over $V \in \mathcal{V}$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$. By (B.9), with probability at least $1 - p/2$ it holds that

$$
\sup_{V \in \mathcal{V}} \left\{ \left( n^k_h(s, a) + \lambda \right)^{1/2} \sum_{s' \in \mathcal{S}} \left( \mathbb{P}_{h}^k(s' | s, a)V(s') - \mathbb{P}_{h}(s' | s, a)V(s') \right) \right\} \\
\leq \sqrt{H^2 \left( \log \left( \frac{T}{\delta} \right) + 2 \right)} + 2 \left( n^k_h(s, a) + \lambda \right)^{-1/2} \frac{H}{K} \\
\leq \sqrt{2H^2 \left( \log |\mathcal{V}| + \log \left( \frac{2|\mathcal{S}| |\mathcal{A}| T}{p} \right) + 2 \right)} + 2 \left( n^k_h(s, a) + \lambda \right)^{-1/2} \frac{H}{K} \\
\leq C_1 H \sqrt{|\mathcal{S}| \log \left( \frac{|\mathcal{S}| |\mathcal{A}| T}{p} \right)} := \beta
$$

for all $(k, h)$ and $(s, a)$, where $C_1$ is an absolute constant. We recall our choice of $\Gamma^k_h$ and $\beta$. Hence, with probability at least $1 - p/2$ it holds that

$$
\left| \sum_{s' \in \mathcal{S}} \left( \mathbb{P}_{h}^k(s' | s, a)V(s') - \mathbb{P}_{h}(s' | s, a)V(s') \right) \right| \leq \beta \left( n^k_h(s, a) + \lambda \right)^{-1/2} := \Gamma^k_h(s, a)
$$

for any $(k, h) \in [K] \times [H]$ and $(s, a) \in |\mathcal{S}| \times |\mathcal{A}|$, where $\beta := C_1 H \sqrt{|\mathcal{S}| \log(|\mathcal{S}| |\mathcal{A}| T/p)}$.

We recall the definition $r_{h}(s, a) = e^r_{s,a} \theta_{r,h}$. By our estimation $\hat{r}^k_{h}(s, a)$ in Algorithm 8, we have

$$
\hat{r}^k_{h}(s, a) = \frac{1}{n^k_h(s, a) + \lambda} \sum_{\tau=1}^{k-1} \mathbb{1} \{ (s, a) = (s^\tau, a^\tau_h) \} [\theta_{r,h}(s^\tau_h, a^\tau_h)]
$$

and thus

$$
|\hat{r}^k_{h}(s, a) - r_{h}(s, a)| \\
= |\hat{r}^k_{h}(s, a) - [\theta_{r,h}]_{(s,a)}| \\
= \left( n^k_h(s, a) + \lambda \right)^{-1} \sum_{\tau=1}^{k-1} \mathbb{1} \{ (s, a) = (s^\tau, a^\tau_h) \} \left[ [\theta_{r,h}(s^\tau_h, a^\tau_h)] - [\theta_{r,h}]_{(s, a)} \right] \\
= \left( n^k_h(s, a) + \lambda \right)^{-1} \left[ \lambda [\theta_{r,h}]_{(s,a)} \right] \\
\leq \left( n^k_h(s, a) + \lambda \right)^{-1} \lambda \\
\leq \left( n^k_h(s, a) + \lambda \right)^{-1/2} \lambda \\
\leq \Gamma^k_h(s, a)
$$

where we utilize $\lambda = 1$ and $\beta \geq 1$ in the inequalities.
We now are ready to check the model prediction error $i_r^k(s,a)$ defined by (3.8),

$$i_r^k(s,a) = Q^k_{r,h}(s,a) - (r_h + \mathbb{P}_h V_{r,h+1}^k)(s,a)$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. On the other hand, notice that $(r_h + \mathbb{P}_h V_{r,h+1}^k)(s,a) \leq H - h + 1$, thus

$$i_r^k(s,a) = (r_h + \mathbb{P}_h V_{r,h+1}^k)(s,a) - Q^k_{r,h}(s,a)$$

for any $(k,h) \in [K] \times [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Hence, we complete the proof of (B.1).

### B.4 Supporting lemmas from optimization

We rephrase some optimization results from the literature for our constrained problem (3.2),

$$\max_{\pi \in \Delta(\mathcal{A}|S,H)} \{ V^\pi_{r,1}(s_1) \mid V^\pi_{g,1}(s_1) \geq b \}$$

in which we maximize over all policies and $b \in (0,H]$. Let the optimal solution be $\pi^*$ such that

$$V^\pi_{r,1}(s_1) = \max_{\pi \in \Delta(\mathcal{A}|S,H)} \{ V^\pi_{r,1}(s_1) \mid V^\pi_{g,1}(s_1) \geq b \}.$$ 

Let the Lagrangian be $V^\pi_Y(s_1) := V^\pi_{r,1}(s_1) + Y(V^\pi_{g,1}(s_1) - b)$, where $Y \geq 0$ is the Lagrange multiplier or dual variable. The associated dual function is defined as

$$V^Y_D(s_1) := \max_{\pi \in \Delta(\mathcal{A}|S,H)} V^\pi_Y(s_1) := V^\pi_{r,1}(s_1) + Y(V^\pi_{g,1}(s_1) - b)$$

and the optimal dual is $Y^* := \arg\min_{Y \geq 0} V^Y_D(s_1)$, $V^Y_D(s_1) := \min_{Y \geq 0} V^Y_D(s_1)$.

We recall that the problem (3.2) enjoys strong duality under the strict feasibility condition (also called Slater condition). The proof follows [180, Proposition 1] in the finite-horizon setting.
Assumption 17 (Slater condition) There exists $\gamma > 0$ and $\bar{\pi}$ such that $V_{g,1}(s_1) - b \geq \gamma$.

Lemma 46 (Strong duality) [180, Proposition 1] If the Slater condition holds, then the strong duality holds,

$$V_{r,1}^{\pi^*}(s_1) = V_Y^{Y^*}(s_1).$$

Strong duality implies that the optimal solution to the dual problem: minimize $Y \geq 0$ $V_Y D(s_1)$ is obtained at $Y^*$. Denote the set of all optimal dual variables as $\Lambda^*$. Under the Slater condition, a useful property of the dual variable is that the sublevel sets are bounded [27, Section 8.5].

Lemma 47 (Boundedness of sublevel sets of the dual function) Let Slater condition hold. Fix $C \in \mathbb{R}$. For any $Y \in \{Y \geq 0 \mid V_Y D(s_1) \leq C\}$, it holds that

$$Y \leq \frac{1}{\gamma} \left( C - V_{r,1}^{\pi^*}(s_1) \right).$$

Proof. By $Y \in \{Y \geq 0 \mid V_Y D(s_1) \leq C\}$,

$$C \geq V_Y^{Y^*}(s_1) \geq V_{r,1}^{\pi^*}(s_1) + Y (V_{g,1}^{\pi^*}(s_1) - b) \geq V_{r,1}^{\pi^*}(\rho) + Y \gamma$$

where we use the Slater point $\bar{\pi}$ in the last inequality. We complete the proof by noting $\gamma > 0$. □

Corollary 48 (Boundedness of $Y^*$) If we take $C = V_{r,1}^{\pi^*}(s_1) = V_Y^{Y^*}(s_1)$, then $\Lambda^* = \{Y \geq 0 \mid V_Y D(s_1) \leq C\}$. Thus, for any $Y \in \Lambda^*$,

$$Y \leq \frac{1}{\gamma} \left( V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{\pi^*}(s_1) \right).$$

Another useful theorem from the optimization [27, Section 3.5] is given as follows. It describes that the constraint violation $b - V_{g,1}^{\pi^*}(s_1)$ can be bounded similarly even if we have some weak bound. We next state and prove it for our problem, which is used in our constraint violation analysis.

Lemma 49 (Constraint violation) Let the Slater condition hold and $Y^* \in \Lambda^*$. Let $C^* \geq 2Y^*$. Assume that $\pi \in \Delta(A \mid S, H)$ satisfies

$$V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{\pi}(s_1) + C^* \left[ b - V_{g,1}^{\pi^*}(s_1) \right]_+ \leq \delta.$$  

Then,

$$\left[ b - V_{g,1}^{\pi^*}(s_1) \right]_+ \leq \frac{2\delta}{C^*}$$

where $[x]_+ = \max(x, 0)$.

Proof. Let

$$v(\tau) = \maximize_{\pi \in \Delta(A \mid S, H)} \{ V_{r,1}^{\pi^*}(s_1) \mid V_{g,1}^{\pi^*}(s_1) \geq b + \tau \}.$$
By definition, \( v(0) = V^\pi_{r,1}(s_1) \). It has been shown as a special case of [180, Proposition 1] that \( v(\tau) \) is concave. First, we show that \(-Y^* \in \partial v(0)\). By the Lagrangian and the strong duality,

\[
V_L^{\pi,Y^*}(s_1) \leq \max_{\pi \in \Delta(A|S,H)} V_L^{\pi,Y^*}(s_1) = V_D^{Y^*}(s_1) = V_{r,1}^{\pi^*}(s_1) = v(0)
\]

for all \( \pi \in \Delta(A|S,H) \). For any \( \pi \in \{ \pi \in \Delta(A|S,H) | V_{g,1}^{\pi}(s_1) \geq b + \tau \} \),

\[
v(0) - \tau Y^* \geq \mathcal{L}(\pi,Y^*) - \tau Y^* \\
= V_{r,1}^\pi(s_1) + Y^*(V_{g,1}^\pi(s_1) - b) - \tau Y^* \\
= V_{r,1}^\pi(s_1) + Y^*(V_{g,1}^\pi(s_1) - b - \tau) \\
\geq V_{r,1}^\pi(s_1).
\]

If we maximize the right-hand side of above inequality over \( \pi \in \{ \pi \in \Delta(A|S,H) | V_{g,1}^{\pi}(s_1) \geq b + \tau \} \), then

\[
v(0) - \tau Y^* \geq v(\tau)
\]

which show that \(-Y^* \in \partial v(0)\). On the other hand, if we take \( \tau = \tilde{\tau} := -(b - V_{g,1}^\pi(s_1))^+ \), then

\[
V_{r,1}^\pi(s_1) \leq v(\tilde{\tau}) \text{ and } V_{r,1}^{\pi^*}(s_1) = v(0) \leq v(\tilde{\tau}).
\]

Combing the above two yields

\[
V_{r,1}^\pi(s_1) - V_{r,1}^{\pi^*}(s_1) \leq -\tilde{\tau} Y^*.
\]

Thus,

\[
(C^* - Y^*) |\tilde{\tau}| = -Y^* |\tilde{\tau}| + C^* |\tilde{\tau}| \\
= \tilde{\tau} Y^* + C^* |\tilde{\tau}| \\
\leq V_{r,1}^{\pi^*}(s_1) - V_{r,1}^\pi(s_1) + C^* |\tilde{\tau}|.
\]

By our assumption and \( |\tilde{\tau}| = [b - V_{g,1}^\pi(\rho)]^+ \),

\[
[b - V_{g,1}^\pi(s_1)]^+ \leq \frac{\delta}{C^* - Y^*} \leq \frac{2\delta}{C^*}.
\]

\[ \square \]

**B.5 Other supporting lemmas**

First, we state a useful concentration inequality for the standard self-normalized processes.

**Lemma 50 (Concentration of self-normalized processes)** Let \( \{F_t\}_{t=0}^\infty \) be a filtration and \( \{\eta_t\}_{t=0}^\infty \) be a \( \mathbb{R} \)-valued stochastic process such that \( \eta_t \) is \( F_t \)-measurable for any \( t \geq 0 \). Assume that for any \( t \geq 0 \), conditioning on \( F_t \), \( \eta_t \) is a zero-mean and \( \sigma \)-subGaussian random variable with the variance proxy \( \sigma^2 > 0 \), i.e., \( \mathbb{E} \left[ e^{\lambda \eta_t} | F_t \right] \leq e^{\lambda^2 \sigma^2/2} \) for any \( \lambda \in \mathbb{R} \). Let \( \{X_t\}_{t=1}^\infty \) be an \( \mathbb{R}^d \)-valued
stochastic process such that \( X_t \) is \( \mathcal{F}_t \)-measurable for any \( t \geq 0 \). Let \( Y \in \mathbb{R}^{d \times d} \) be a deterministic and positive-definite matrix. For any \( t \geq 0 \), we define

\[
\bar{Y}_t := Y + \sum_{\tau = 1}^{t} X_{\tau} X_{\tau}^T \quad \text{and} \quad S_t = \sum_{\tau = 1}^{t} \eta_{\tau} X_{\tau}.
\]

Then, for any fixed \( \delta \in (0, 1) \), it holds with probability at least \( 1 - \delta \) that

\[
\|S_t\|_2^2 (\bar{Y}_t)^{-1} - \frac{1}{2} \leq 2 \sigma^2 \log \left( \frac{\det(\bar{Y}_t)^{1/2} \det(Y)^{-1/2}}{\delta} \right)
\]

for any \( t \geq 0 \).

\textbf{Proof.} See the proof of Theorem 1 in [3]. \( \square \)

The above concentration inequality can be customized to our setting in the following form without using covering number arguments as in [110].

\textbf{Lemma 51} Let \( \lambda = 1 \) in Algorithm 7. Fix \( \delta \in (0, 1) \). Then, for any \( (k, h) \in [K] \times [H] \) it holds for \( \diamond = r \) or \( g \) that

\[
\left\| \sum_{\tau = 1}^{k-1} \phi_{\diamond, h}^\tau (s_{h,1}^\tau, a_{h,1}^\tau) (V_{\diamond, h+1}^k (s_{h+1}^\tau) - (P_{h} V_{\diamond, h+1}^k (s_{h+1}^\tau, a_{h}^\tau))) \right\|_{(\Lambda_{\diamond, h}^k)^{-1}} \leq C \sqrt{dH^2 \log \left( \frac{dT}{\delta} \right)}
\]

with probability at least \( 1 - \delta/2 \) where \( C > 0 \) is an absolute constant.

\textbf{Proof.} See the proof of Lemma D.1 in [47]. \( \square \)

\textbf{Lemma 52 (Elliptical potential lemma)} Let \( \{\phi_t\}_{t=1}^{\infty} \) be a sequence of functions in \( \mathbb{R}^d \) and \( \Lambda_0 \in \mathbb{R}^{d \times d} \) be a positive definite matrix. Let \( \Lambda_t = \Lambda_0 + \sum_{i=1}^{t-1} \phi_i \phi_i^\top \). Assume \( \|\phi_i\|_2 \leq 1 \) and \( \lambda_{\min}(\Lambda_0) \geq 1 \). Then for any \( t \geq 1 \) it holds that

\[
\log \left( \frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)} \right) \leq \sum_{i=1}^{t} \phi_i^\top \Lambda_i^{-1} \phi_i \leq 2 \log \left( \frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)} \right).
\]

\textbf{Proof.} See the proof of Lemma D.2 in [110] or [47]. \( \square \)

\textbf{Lemma 53 (Pushback property of KL-divergence)} Let \( f : \Delta \rightarrow \mathbb{R} \) be a concave function where \( \Delta \) is a probability simplex in \( \mathbb{R}^d \). Let \( \Delta^o \) be the interior of \( \Delta \). Let \( x^* = \arg\max_{x \in \Delta} f(x) - \alpha^{-1} D_{KL}(x, y) \) for a fixed \( y \in \Delta^o \) and \( \alpha > 0 \). Then, for any \( z \in \Delta \),

\[
f(x^*) - \frac{1}{\alpha} D_{KL}(x^*, y) \geq f(z) - \frac{1}{\alpha} D_{KL}(z, y) + \frac{1}{\alpha} D_{KL}(z, x^*).
\]

\textbf{Proof.} See the proof of Lemma 14 in [242]. \( \square \)
Lemma 54 (Bounded KL-divergence difference) Let \( \pi_1, \pi_2 \) be two probability distributions in \( \Delta(\mathcal{A}) \). Let \( \tilde{\pi}_2 = (1 - \theta) \pi_2 + \theta |\mathcal{A}| \) where \( \theta \in (0, 1] \). Then,
\[
D_{KL}(\pi_1 | \tilde{\pi}_2) - D_{KL}(\pi_1 | \pi_2) \leq \theta \log |\mathcal{A}|.
\]
Moreover, we have a uniform bound, \( D_{KL}(\pi_1 | \tilde{\pi}_2) \leq \log(|\mathcal{A}|/\theta) \).

Proof. See the proof of Lemma 31 in [242].
Appendix C

Supporting proofs in Chapter 4

C.1 Proof of Lemma 25

We begin with the triangle inequality, \( \|x_{j,k}(t) - \bar{x}_k(t)\| \leq \|x_{j,k}(t) - x_k(t)\| + \|x_k(t) - \bar{x}_k(t)\| \) where \( x_k(t) = P_X(\bar{x}_k(t)) \) and \( \bar{x}_k(t) = \frac{1}{N} \sum_{j=1}^{N} x_{j,k}(t) \). First, the non-expansiveness of projection \( P_X \) implies that

\[
\|x_{j,k}(t) - x_k(t)\| \leq \|x'_{j,k}(t) - \bar{x}'_k(t)\| \tag{C.1}
\]

where \( x'_{j,k}(t) = P_X(x'_{j,k}(t)) \). Second, by the convexity of norm \( \|\cdot\| \) and non-expansiveness of projection \( P_X \),

\[
\|x_k(t) - \bar{x}_k(t)\| \leq \frac{1}{N} \sum_{j=1}^{N} \|P_X(x'_k(t)) - P(x'_{j,k}(t))\| \tag{C.2}
\]

Next, we focus on \( x'_{j,k}(t) \) and \( \bar{x}'_k(t) \). Let \([W^*]_j\) be the \( j \)th row of \( W^* \) and \([W^*]_{ji}\) be the \((j,i)\)th element of \( W^* \). For any \( t \geq 2 \), the primal update \( x'_{j,k}(t) \) of Algorithm 9 can be expressed as

\[
x'_{j,k}(t) = \sum_{i=1}^{N} [W^{t-1}]_{ij} x'_{i,k}(1) - \eta_k \sum_{s=2}^{t-1} \sum_{i=1}^{N} [W^{t-s+1}]_{ij} G_i(x(z_{i,k}(s-1); \xi_{k,s-1}) - \eta_k G_j(x(z_{j,k}(t-1); \xi_{k,t-1}) \tag{C.3}
\]

and \( x'_{j,k}(2) = \sum_{i=1}^{N} [W]_{ij} x'_{i,k}(1) - \eta_k G_j(x(z_{j,k}(1); \xi_{k,1}) \). Similar to the argument of [82, Eqs. (26) and (27)], we utilize the gradient boundedness to bound the difference of (4.14) and (C.3) by

\[
\|x'_{j,k}(t) - \bar{x}'_k(t)\| \leq \sum_{i=1}^{N} \left\| \frac{1}{N} - [W^{t-1}]_{ij} \right\| \|x'_{i,k}(1)\| + \eta_k c \sum_{s=2}^{t-1} \left\| \frac{1}{N} - [W^{t-s+1}]_{ij} \right\|_1 + 2 \eta_k c. \tag{C.4}
\]
Application of the Markov chain property of mixing matrix [82] on the second sum in (C.4) yields,

\[
\sum_{t=2}^{T_k-1} \left\| \frac{1}{N} - [W^{t-s+1}]_{ji} \right\|_1 \leq \frac{2 \log(\sqrt{N} T_k)}{1 - \sigma_2(W)}.
\]  
(C.5)

For \( x_{i,k}^j(1) = \frac{1}{T_k} \sum_{t=1}^{T_k-1} x_{i,k-1}(t) \) where \( x_{i,k-1}(t) = P(x_{i,k-1}(t)) \) and \( 0 \in X \), we utilize the non-expansiveness of projection to bound it as

\[
\|x_{i,k}^j(1)\| \leq \frac{1}{T_k} \sum_{t=1}^{T_k-1} \|x_{i,k-1}^j(t)\|.
\]

Using (C.3) at round \( k-1 \), we utilize the property of doubly stochastic \( W \) to have

\[
\|x_{i,k-1}^j(t)\| \leq \sum_{j=1}^{N} [W^{t-1}]_{ji} \|x_{j,k-1}^j(1)\| + 2\eta_{k-1} T_{k-1} c.
\]

Repeating this inequality for \( k-2, k-3, \ldots, 1 \) yields,

\[
\|x_{i,k}^j(1)\| \leq 2 \sum_{l=1}^{k-1} \eta_l T_l c.
\]  
(C.6)

where we use \( x_{j,1}^j(1) = 0 \) for all \( j \in V \).

Now, we are ready to show the desired result. Notice that \( \|x_{j,k}^j(1) - \bar{x}_{k}^j(1)\| \leq \|x_{j,k}^j(1)\| + \frac{1}{N} \sum_{i=1}^{N} \|x_{j,k}^j(1)\| \). We collect (C.5) and (C.6) for (C.4), and average it over \( t = 1, \ldots, T_k \) to obtain,

\[
\frac{1}{T_k} \sum_{i=1}^{T_k} \|x_{j,k}(t) - \bar{x}_{k}(t)\| = \frac{1}{T_k} \sum_{t=2}^{T_k} \|x_{j,k}^j(t) - \bar{x}_{k}^j(t)\| + \frac{1}{T_k} \|x_{j,k}^j(1) - \bar{x}_{k}^j(1)\|
\leq \frac{2\eta_{k-1} \log(\sqrt{N} T_k)}{1 - \sigma_2(W)} + \frac{4}{T_k} \sum_{l=1}^{k-1} \eta_l T_l c + 2\eta_{k-1} c
\]

Bounding the sum \( \sum_{t=2}^{T_k} \sum_{i=1}^{N} \left\| \frac{1}{N} - [W^{t-1}]_{ji} \right\| \) by (C.5) and application of (C.1) and (C.2) complete the proof.

### C.2 Martingale concentration bound

We state a useful result about martingale sequence.
Lemma 55 Let \( \{X(t)\}_{t=1}^T \) be a martingale difference sequence in \( \mathbb{R}^d \), and let \( \|X(t)\| \leq M \). Then,

\[
\mathbb{E} \left[ \left\| \frac{1}{T} \sum_{t=1}^T X(t) \right\|^2 \right] \leq \frac{4M^2}{T}.
\] (C.7)

Proof. We recall the classic concentration result in [183]: for any \( \delta \geq 0 \), we have

\[
P \left( \left\| \frac{1}{T} \sum_{t=1}^T X(t) \right\|^2 \geq \frac{4M^2\delta}{T} \right) \leq e^{-\delta}.
\]

The left-hand side of (C.7) can be expanded into

\[
\int_0^\infty P \left( \left\| \frac{1}{T} \sum_{t=1}^T X(t) \right\|^2 \geq s \right) ds = \frac{4M^2}{T} \int_0^\infty P \left( \left\| \frac{1}{T} \sum_{t=1}^T X(t) \right\|^2 \geq \frac{4M^2\delta}{T} \right) d\delta
\]

\[
\leq \frac{4M^2}{T} \int_0^\infty e^{-\delta} d\delta
\]

\[
\leq \frac{4M^2}{T}.
\]

C.3 Proof of Lemma 27

It is clear that \( \text{err} (\hat{x}_{i,k}) \geq 0 \) from the optimality of \( x^* \) in (4.10). The optimality of \( y^*_j \) yields,

\[
\frac{1}{N} \sum_{j=1}^N \psi_j (x^*, \hat{y}_j) \geq \frac{1}{N} \sum_{j=1}^N \psi_j (x^*, \hat{y}_{j,k}).
\]

Thus, using (4.15) and (4.16), we have \( \text{err} (\hat{x}_{i,k}) \leq \text{err}' (\hat{x}_{i,k}, \hat{y}_k) \).

C.4 Proof of Lemma 28

To show (4.17a), we apply the strong convexity of \( f_j(x) \),

\[
\text{err} (\hat{x}_{i,k}) \geq \frac{1}{N} \sum_{j=1}^N \langle \nabla f_j(x^*), \hat{x}_{i,k} - x^* \rangle + \frac{L_x}{2} \| \hat{x}_{i,k} - x^* \|^2
\]

\[
\geq \frac{L_x}{2} \| \hat{x}_{i,k} - x^* \|^2
\]

where the second inequality is due to the optimality of \( x^* \).
To show (4.17b), we subtract and add $\frac{1}{N} \sum_{j=1}^{N} \psi_j(x^*, y_j^*)$ into (4.16) and apply the strong concavity of $\psi_j(x, y_j)$ in $y_j$,

$$
\text{err}'(\hat{x}_{i,k}, \hat{y}_k) = \text{err}(\hat{x}_{i,k}) + \frac{1}{N} \sum_{j=1}^{N} (\psi_j(x^*, y_j^*) - \psi_j(x^*, \hat{y}_{j,k})) \\
\geq \frac{1}{N} \sum_{j=1}^{N} (\psi_j(x^*, y_j^*) - \psi_j(x^*, \hat{y}_{j,k})) \\
\geq \frac{L_y}{2N} \sum_{j=1}^{N} \|y_j^* - \hat{y}_{j,k}\|^2
$$

where $\text{err}(\hat{x}_{i,k}) \geq 0$ is omitted in the first inequality, and the second inequality follows the strong concavity assumption of $\psi_j(\hat{x}_{i,k}, y_j)$ in $y_j$ and the optimality of $y_j^* = \arg\max_{y_j \in Y_j} \psi_j(x^*, y_j)$.

To show (4.17c), because of the optimality of $x^*$ and $y_j^*$, we first have $\psi_j(\hat{x}_{i,k}, y_j^*) - \psi_j(x^*, y_j^*) \geq 0$, and $\psi_j(x^*, y_j^*) - \psi_j(x^*, \hat{y}_{j,k}) \geq 0$. This further shows that

$$
\psi_j(\hat{x}_{i,k}, \hat{y}_{j,k}^*) - \psi_j(x^*, \hat{y}_{j,k}) = (\psi_j(\hat{x}_{i,k}, \hat{y}_{j,k}^*) - \psi_j(\hat{x}_{i,k}, y_j^*)) + (\psi_j(\hat{x}_{i,k}, y_j^*) - \psi_j(x^*, y_j^*)) + (\psi_j(x^*, y_j^*) - \psi_j(x^*, \hat{y}_{j,k})) \\
\geq \psi_j(\hat{x}_{i,k}, \hat{y}_{j,k}^*) - \psi_j(\hat{x}_{i,k}, y_j^*)
$$

Hence, we have

$$
\text{err}'(\hat{x}_{i,k}, \hat{y}_k) \geq \frac{1}{N} \sum_{j=1}^{N} (\psi_j(\hat{x}_{i,k}, \hat{y}_{j,k}^*) - \psi_j(\hat{x}_{i,k}, y_j^*)) \\
\geq \frac{L_y}{2N} \sum_{j=1}^{N} \|y_j^* - \hat{y}_{j,k}\|^2
$$

where the second inequality follows from the strong concavity assumption of $\psi_j(\hat{x}_{i,k}, y_j)$ in $y_j$ and the optimality of $\hat{y}_{j,k}^* = \arg\max_{y_j \in Y_j} \psi_j(\hat{x}_{i,k}, y_j)$. 

264
Appendix D

Supporting proofs in Chapter 5

D.1 Proof of Lemma 34

We prove (5.5) by induction on the number of players \( N \). In the basic step: \( N = 2 \), the right-hand side of (5.5) becomes
\[
\left( \Psi_{\pi_1, \pi_2} - \Psi_{\pi_1, \pi_2} \right) + \left( \Psi_{\pi_1, \pi_2} - \Psi_{\pi_1, \pi_2} \right) + \left( \Psi_{\pi_1, \pi_2} - \Psi_{\pi_1, \pi_2} + \Psi_{\pi_1, \pi_2} \right)
\]
which equals to the left-hand side: \( \Psi_{\pi_1', \pi_2'} - \Psi_{\pi_1, \pi_2} \).

Assume the equality (5.5) holds for \( N \) players. We next consider the induction step for \( N + 1 \) players. By subtracting and adding \( \Psi_{\pi_{\leq N}, \pi_{N+1}} \),
\[
\Psi_{\pi'} - \Psi_{\pi} = \left( \Psi_{\pi_{\leq N}, \pi_{N+1}} - \Psi_{\pi_{\leq N}, \pi_{N+1}} \right) + \left( \Psi_{\pi_{\leq N}, \pi_{N+1}} - \Psi_{\pi_{\leq N}, \pi_{N+1}} \right).
\]
(D.1)

In (D.1), we use the shorthand \( \pi_{\leq N} \) and \( \pi_{\leq N} \) for \( \{\pi_k\}_{k=1}^N \) and \( \{\pi_k\}_{k=1}^N \), respectively. We note that \( \text{Diff}_{\leq N} \) or \( \text{Diff}_{N+1} \) can be viewed as a function for \( N \) players if we fix the \((N+1)\)th policy. For the first term \( \text{Diff}_{\leq N} \),
\[
\text{Diff}_{\leq N} = \sum_{i=1}^{N} \left( \Psi_{\pi_i', \pi_{i\sim N+1}, \pi_{N+1}'} - \Psi_{\pi_{i\leq N}, \pi_{N+1}'} \right)
\]
\[
+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left( \Psi_{\pi_{i\sim j}, \pi_{j}, \pi_{j}, \pi_{N+1}} - \Psi_{\pi_{i\sim j}, \pi_{j}, \pi_{j}, \pi_{N+1}} \right)
\]
\[
- \Psi_{\pi_{i\sim j}, \pi_{j}, \pi_{j}, \pi_{N+1}} + \Psi_{\pi_{i\sim j}, \pi_{j}, \pi_{j}, \pi_{N+1}} \right)
\]
We note that

\[ \text{Diff} \leq N = \sum_{i=1}^{N} \left( \Psi \pi'_{i}, \pi <_{i} \sim N_{i+1}, \pi N_{i+1} - \Psi \pi \right) \]

\[ + \sum_{i=1}^{N} \left( \Psi \pi'_{i}, \pi <_{i} \sim N_{i+1}, \pi N_{i+1} - \Psi \pi'_{i}, \pi <_{i} \sim N_{i+1}, \pi N_{i+1} + \Psi \pi^{ \leq N, \pi N+1} \right) \]

\[ + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left( \Psi \pi <_{i} \sim j, \pi >_{j}, \pi'_{i}, \pi'_{j}, \pi'_{N+1} - \Psi \pi <_{i} \sim j, \pi >_{j}, \pi_{i}, \pi'_{j}, \pi'_{N+1} \right) \]

\[ - \Psi \pi <_{i} \sim j, \pi >_{j}, \pi'_{i}, \pi'_{j}, \pi'_{N+1} + \Psi \pi <_{i} \sim j, \pi >_{j}, \pi_{i}, \pi_{j}, \pi'_{N+1} \]}

where we use \( \pi'_{j} \) to represent \( \{ \pi'_{k}\}_{k=j+1}^{N} \).

Adding \( \text{Diff}_{N+1} \) to the last equivalent expression of \( \text{Diff} \leq N \) above yields

\[ \text{Diff} \leq N + \text{Diff}_{N+1} = \sum_{i=1}^{N+1} \left( \Psi \pi'_{i}, \pi_{i} - \Psi \pi \right) \]

\[ + \sum_{i=1}^{N} \sum_{j=N+1}^{N+1} \left( \Psi \pi <_{i} \sim j, \pi >_{j}, \pi'_{i}, \pi'_{j} - \Psi \pi <_{i} \sim j, \pi >_{j}, \pi_{i}, \pi'_{j} \right) \]

\[ + \sum_{i=1}^{N} \sum_{j=i+1}^{N+1} \left( \Psi \pi <_{i} \sim j, \pi >_{j}, \pi_{i}, \pi'_{j}, \pi'_{N+1} - \Psi \pi <_{i} \sim j, \pi >_{j}, \pi_{i}, \pi'_{j}, \pi'_{N+1} \right) \]

\[ - \Psi \pi <_{i} \sim j, \pi >_{j}, \pi_{i}, \pi'_{j}, \pi'_{N+1} + \Psi \pi <_{i} \sim j, \pi >_{j}, \pi_{i}, \pi_{j}, \pi'_{N+1} \]}

where the first equality has a slight abuse of the notation: \( \pi'_{j} \) represents \( \{ \pi'_{k}\}_{k=j+1}^{N+1} \) in the first double sum and \( \pi'_{j}^{N} \) represents \( \{ \pi'_{k}\}_{k=j+1}^{N} \) in the second double sum. Therefore, (5.5) holds for \( N + 1 \) players. The proof is completed by induction.

### D.2 Proof of Lemma 35

We note that \( Q_{i}^{\pi_{i}, \pi_{j}}(s, \cdot) \) and \( Q_{i}^{\pi_{i}, \pi_{j}^{N}}(s, \cdot) \) are averaged action value functions for player \( i \) using policy \( \pi_{i} \), but they have different underlying averaged MDPs because of different policies executed by player \( j \). Hence, we can directly apply Lemma 69. Specifically, let \((r, p)\) be the
averaged reward and transition functions for player $i$ induced by $(\tilde{\pi}_{-ij}, \pi_j)$, and $(\tilde{r}, \tilde{p})$ be those induced by $(\tilde{\pi}_{-ij}, \pi_j')$. Then,

$$|r(s, a_i) - \tilde{r}(s, a_i)|$$

$$= \left| \sum_{a_j, a_{-ij}} r(s, a_i, a_j, a_{-ij}) \left( \pi_j(a_j | s) - \pi_j'(a_j | s) \right) \tilde{\pi}_{-ij}(a_{-ij} | s) \right|$$

$$\leq \|\pi_j(\cdot | s) - \pi_j'(\cdot | s)\|_1$$

and

$$\|p(\cdot | s, a_i) - \tilde{p}(\cdot | s, a_i)\|_1$$

$$= \sum_{s'} \left| \sum_{a_j, a_{-ij}} p(s' | s, a_i, a_j, a_{-ij}) \left( \pi_j(a_j | s) - \pi_j'(a_j | s) \right) \tilde{\pi}_{-ij}(a_{-ij} | s) \right|$$

$$\leq \sum_{s'} \left| \sum_{a_j, a_{-ij}} p(s' | s, a_i, a_j, a_{-ij}) \pi_{-ij}(a_{-ij} | s) \left( \pi_j(a_j | s) - \pi_j'(a_j | s) \right) \right|$$

$$\leq \|\pi_j(\cdot | s) - \pi_j'(\cdot | s)\|_1.$$
We also note that $Q^\pi(\cdot, \cdot; s^\sharp)$ is the action value function associated with the reward function $r(s, a) = (1 - \gamma) 1_{\{s = s^\sharp\}}$. Thus,

$$\sum_{s^\sharp} Q^\pi(\cdot, \cdot; s^\sharp) = \sum_{s^\sharp} E \left[ (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t 1_{\{s_t = s^\sharp\}} \mid (s_0, a_0) = (s, a), \pi' \right] = E \left[ (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t 1_{\{s_t = s^\sharp\}} \mid (s_0, a_0) = (s, a), \pi' \right] = 1.$$  

Therefore, we can arrange (D.2) as follows,

$$\sum_{s^\sharp} \left| d^\pi_{\mu}(s^\sharp) - d^\pi'(s^\sharp) \right| \leq \sum_{s, a} d^\pi_{\mu}(s) \mid \pi(a \mid s) - \pi'(a \mid s) \mid \leq \sum_{s} d^\pi_{\mu}(s) \| \pi(\cdot \mid s) - \pi'(\cdot \mid s) \|_1 \leq \max_s \| \pi(\cdot \mid s) - \pi'(\cdot \mid s) \|_1.$$  

### D.4 Proof of Lemma 37

We consider a two-player common-reward Markov game with state space $S$ and action sets $A_1, A_2$. Let $r : S \times A_1 \times A_2 \to [0, 1]$ be the reward function, and $p : S \times A_1 \times A_2 \to \Delta(S)$ be the transition function. Let $\Pi_1 = (\Delta(A_1))^{\mid S \mid}$ and $\Pi_2 = (\Delta(A_2))^{\mid S \mid}$ be player 1 and player 2’s policy sets, respectively.

For any $x, x' \in \Pi_1$, we define the following non-stationary policies:

- $\bar{x}_i$: a Player 1’s policy where in steps from 0 to $i - 1$, $x'$ is executed; in steps from $i$ to $\infty$, $x$ is executed.

With this definition, $\bar{x}_0 = x$ and $\bar{x}_\infty = x'$. We define $\bar{y}_i$ similarly. Since $\bar{x}_i$ is non-stationary, we specify its action distribution as $\bar{x}_i(\cdot \mid s, h)$ where $h$ is the step index. The joint value function for these non-stationary policies can be defined as usual:

$$V^{\bar{x}_i, \bar{y}_j}(\mu) := E \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) \mid s_0 \sim \mu, a_t \sim \bar{x}_i(\cdot \mid s_t, t), b_t \sim \bar{y}_j(\cdot \mid s_t, t) \right].$$

For simplicity, we omit the initial distribution $\mu$ in writing the value function. We first show that for any $H \in \mathbb{N}$,

$$V^{\bar{x}_{0}, \bar{y}_0} - V^{\bar{x}_{H}, \bar{y}_0} - V^{\bar{x}_{0}, \bar{y}_H} + V^{\bar{x}_H, \bar{y}_H} = \sum_{i=0}^{H-1} \sum_{j=0}^{H-1} (V^{\bar{x}_i, \bar{y}_j} - V^{\bar{x}_{i+1}, \bar{y}_j} - V^{\bar{x}_i, \bar{y}_{j+1}} + V^{\bar{x}_{i+1}, \bar{y}_{j+1}}).$$
In fact, the right-hand side above is equal to

\[
\sum_{j=0}^{H-1} \sum_{i=0}^{H-1} (V^{x_i,y_j} - V^{x_{i+1},y_j}) + \sum_{j=0}^{H-1} \sum_{i=0}^{H-1} (-V^{x_i,y_{j+1}} + V^{x_{i+1},y_{j+1}})
\]

\[
= \sum_{j=0}^{H-1} (V^{x_0,y_j} - V^{x_H,y_j}) + \sum_{j=0}^{H-1} (-V^{x_0,y_{j+1}} + V^{x_H,y_{j+1}})
\]

\[
= \sum_{j=0}^{H-1} (V^{x_0,y_j} - V^{x_0,y_{j+1}}) + \sum_{j=0}^{H-1} (-V^{x_H,y_j} + V^{x_H,y_{j+1}})
\]

\[
= V^{x_0,y_0} - V^{x_0,y_H} - V^{x_H,y_0} + V^{x_H,y_H}.
\]

Sending \( H \) to infinity and recalling that \( x_0 = x, x_\infty = x', y_0 = y, y_\infty = y' \) lead to

\[
V^{x,y} - V^{x',y} - V^{x,y'} + V^{x',y'} = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} (V^{x_i,y_j} - V^{x_{i+1},y_j} - V^{x_i,y_{j+1}} + V^{x_{i+1},y_{j+1}}).
\]

We next focus on the particular summand above with index \((i, j)\) and discuss three cases.

**Case 1: \( i < j \).** We first re-write \( V^{x_i,y_j} - V^{x_{i+1},y_j} \). Notice that the value difference between the policy pairs \((x_i, y_j)\) and \((x_{i+1}, y_j)\) starts at step \( i \), since both policy pairs are equal to \((x', y')\) from step 0 to step \( i - 1 \). At the \( i \)th step, \( x \) changes to \( x' \) while \( x_{i+1} \) remains as \( x' \). Therefore,

\[
V^{x_i,y_j} - V^{x_{i+1},y_j} = \frac{1}{1 - \gamma} \sum_{s,a,b} d_{\mu}^{x',y'}(s;i)x(a \mid s)y'(b \mid s) (r(s, a, b) + \Theta(s, a, b; i, j))
\]

\[
- \frac{1}{1 - \gamma} \sum_{s,a,b} d_{\mu}^{x,y}(s;i)x'(a \mid s)y'(b \mid s) (r(s, a, b) + \Theta(s, a, b; i, j))
\]

\[
= \frac{1}{1 - \gamma} \sum_{s,a,b} d_{\mu}^{x,y}(s;i) \left( (x(a \mid s) - x'(a \mid s))y'(b \mid s) (r(s, a, b) + \Theta(s, a, b; i, j)) \right)
\]

where we define (also note that \( d_{\mu}^{x,y}(s) = \sum_{i=0}^{\infty} d_{\mu}^{x,y}(s;i) \))

\[
d_{\mu}^{x,y}(s;i) := (1 - \gamma)E \left[ \gamma^i I[s_i = s] \mid s_0 \sim \mu \right]
\]

\[
\Theta(s, a, b; i, j) := E \left[ \sum_{t=i+1}^{\infty} \gamma^{t-i} r(s_t, a_t, b_t) \mid s_{i+1} \sim p(\cdot \mid s, a, b, x_i, y_j) \right].
\]
Similarly,
\[
V^{x_i, y_{j+1}} - V^{x_{i+1}, y_{j+1}} \\
= \frac{1}{1 - \gamma} \sum_{s,a,b} d^{x',y'}(s; i) \left( x(a | s) - x'(a | s) \right) y'(b | s) \left( r(s, a, b) + \Theta(s, a, b; i, j + 1) \right).
\]

We notice that the difference \(\Theta(s, a, b; i, j) - \Theta(s, a, b; i, j + 1)\) is equivalent to
\[
\frac{\gamma}{1 - \gamma} \sum_{\bar{\alpha}, \bar{\beta}} d^{x,y'}(\bar{\alpha}; j - i - 1) x(\bar{\alpha} | \bar{\beta}) y'(\bar{\beta} | \bar{\beta}) Q^{x,y}(\bar{\beta}, \bar{\alpha}, \bar{\beta}).
\]

Hence,
\[
V^{x_i, y_j} - V^{x_{i+1}, y_j} - V^{x_i, y_{j+1}} + V^{x_{i+1}, y_{j+1}} \\
= \frac{\gamma}{(1 - \gamma)^2} \sum_{s,a,b} \sum_{\bar{\alpha}, \bar{\beta}} d^{x',y'}(s; i) d^{x,y'}(\bar{\alpha}; j - i - 1) \left( x(a | s) - x'(a | s) \right) y'(b | s) x(\bar{\alpha} | \bar{\beta}) y'(\bar{\beta} | \bar{\beta}) \left( y(\bar{\beta} | \bar{\beta}) - y'(\bar{\beta} | \bar{\beta}) \right)
\]

\[
\leq \frac{\gamma A}{(1 - \gamma)^3} \sum_{s,a} \sum_{\bar{\alpha}, \bar{\beta}} d^{x,y'}(s; i) d^{x,y'}(\bar{\alpha}; j - i - 1) y'(b | s) x(\bar{\alpha} | \bar{\beta}) \left( y(\bar{\beta} | \bar{\beta}) - y'(\bar{\beta} | \bar{\beta}) \right)^2
\]

(bound \(|Q^{x,y}(\cdot, \cdot, \cdot)|\) by \(\frac{1}{1 - \gamma}\) and use AM-GM)

\[
= \frac{\gamma A}{2(1 - \gamma)^3} \sum_{s,a} \sum_{\bar{\alpha}, \bar{\beta}} d^{x,y'}(s; i) d^{x,y'}(\bar{\alpha}; j - i - 1) \left( x(a | s) - x'(a | s) \right)^2
\]

(define \(p(\cdot | s, a, y) = \sum_b p(\cdot | s, a, b) y(b | s)\))

\[
+ \frac{\gamma A}{2(1 - \gamma)^3} \sum_{s,a} \sum_{\bar{\alpha}, \bar{\beta}} d^{x,y'}(s; i) d^{x,y'}(\bar{\alpha}; j - i - 1) \left( y(\bar{\beta} | \bar{\beta}) - y'(\bar{\beta} | \bar{\beta}) \right)^2.
\]

(uniform distribution \(\text{Unif}_A = \frac{1}{A} I\))
Summing the inequality above over $i < j$ yields

\[
\sum_{i=0}^{\infty} \sum_{j=i+1}^{\infty} \left( V^{\bar{x},\bar{y}}_{ij} - V^{\bar{x},\bar{y}}_{i+1,j} - V^{\bar{x},\bar{y}}_{i,j+1} + V^{\bar{x},\bar{y}}_{i+1,j+1} \right)
\]

\[
\leq \frac{\gamma A}{2(1-\gamma)^3} \sum_{i=0}^{\infty} \sum_{j=i+1}^{\infty} d^{x,y}_{\mu}(s; \bar{\pi})(x(a \mid s) - x'(a \mid s))^2 \left( \sum_{\bar{s}, a, s', y} d^{x,y}_{p(\bar{s} \mid s, a, y)}(\bar{s}; j - i - 1) \right)
\]

\[
+ \frac{\gamma A}{2(1-\gamma)^3} \sum_{i=0}^{\infty} \sum_{j=i+1}^{\infty} \sum_{s, \bar{y}, \bar{\pi}} d^{x,y}_{\mu}(s; \bar{\pi}) \left( y(\bar{b} \mid \bar{s}) - y'(\bar{b} \mid \bar{s}) \right)^2 \left( \sum_{\bar{s}, a, s', y} d^{x,y}_{p(\bar{s} \mid s, \bar{y}, a, y)}(\bar{s}; j - i - 1) \right)
\]

\[
= \frac{\gamma A}{2(1-\gamma)^3} \sum_{s} d^{x,y}_{\mu}(s) \left( x(a \mid s) - x'(a \mid s) \right)^2 + \frac{\gamma A}{2(1-\gamma)^3} \sum_{s} d^{x,y}_{\mu}(s) \left( y(\cdot \mid s) - y'(\cdot \mid s) \right)^2
\]

\[
(\text{use the property } \sum_{i=0}^{\infty} d^{x,y}_{\mu}(s; \bar{\pi}) = d^{x,y}_{\mu}(s))
\]

where $\mu'$ is a state distribution that generates the state by the following procedure: first sample a state $s_0$ according to $d^{x,y}_{\mu'}(\cdot)$, then execute $(\text{Unif}_A, y') = (\frac{1}{A} \mathbb{1}, y')$ for one step, and then output the next state.

By Lemma \(70\) (with $\pi = (x', y')$, $\bar{\pi}' = (x, y')$, and $\bar{\pi} = (\text{Unif}_A, y')$), we have $d^{x,y}_{\mu}(s) \leq \frac{d^{x,y}_{\mu'}(s)}{\mu(1-\gamma)} \leq \frac{r^2}{\gamma(1-\gamma)}$. Therefore,

\[
\sum_{i=0}^{\infty} \sum_{j=i+1}^{\infty} \left( V^{\bar{x},\bar{y}}_{ij} - V^{\bar{x},\bar{y}}_{i+1,j} - V^{\bar{x},\bar{y}}_{i,j+1} + V^{\bar{x},\bar{y}}_{i+1,j+1} \right)
\]

\[
\leq \frac{r^2 A}{2(1-\gamma)^4} \sum_{s} d^{x,y}_{\mu}(s) \left( \| x(\cdot \mid s) - x'(\cdot \mid s) \|^2 + \| y(\cdot \mid s) - y'(\cdot \mid s) \|^2 \right).
\]

**Case 2: $i > j$.** This case is symmetric to the case of $i < j$, and can be handled similarly.
Case 3: \( i = j \). In this case,

\[
\sum_{i=0}^{\infty} (V^{x_i, y_i} - V^{x_{i+1}, y_{i+1}})
= \frac{1}{1 - \gamma} \sum_{s,a,b} d_{\mu}^{x',y'}(s) \left( x'(a \mid s) - x(a \mid s) \right) \left( y'(a \mid s) - y(a \mid s) \right) Q^{x,y}(s, a, b)
= \frac{1}{1 - \gamma} \sum_{s,a,b} d_{\mu}^{x'}(s) \left( x'(a \mid s) - x(a \mid s) \right) \left( y'(a \mid s) - y(a \mid s) \right) Q^{x,y}(s, a, b)
\leq \frac{1}{2(1 - \gamma)^2} \sum_{s,a,b} d_{\mu}^{x',y'}(s) \left( x'(a \mid s) - x(a \mid s) \right)^2 + \frac{1}{2(1 - \gamma)^2} \sum_{s,a,b} d_{\mu}^{x',y'}(s) \left( y'(a \mid s) - y(a \mid s) \right)^2
= \frac{A}{2(1 - \gamma)^2} \sum_s d_{\mu}^{x',y'}(s) \left\| x'(\cdot \mid s) - x(\cdot \mid s) \right\|^2 + \frac{A}{2(1 - \gamma)^2} \sum_s d_{\mu}^{x',y'}(s) \left\| y'(\cdot \mid s) - y(\cdot \mid s) \right\|^2.
\]

Summing the bounds in all three cases above completes the proof.

D.5 Proofs for Section 5.6

In this section, we provide proofs of Theorem 40 and Theorem 41 in Appendix D.5.2 and Appendix D.5.3 respectively.

D.5.1 Unbiased estimate

We consider the \( k \)th sampling in the data collection phase of Algorithm 11. By the sampling model in lines 6-8 of Algorithm 11, it is straightforward to see that \( \bar{s}(h_i) \sim d_{\pi}(i) \) for player \( i \). Then, we take \( \bar{a}(h_i) \sim \pi_i^{(k)}(\cdot \mid s(h_i)) \) at step \( h_i \) for player \( i \). Each player \( i \) begins with such \((\bar{s}(h_i), \bar{a}(h_i))\) while
all players execute the policy \( \{ \pi_i^{(t)} \}_{i=1}^N \) with the termination probability \( 1 - \gamma \). Once terminated, we add all rewards collected in \( R_i^{(k)} \). We next show that \( \mathbb{E}[R_i^{(k)}] = \bar{Q}_i^\pi(\bar{s}(h_i), \bar{a}_i(h_i)) \),

\[
\mathbb{E}[R_i^{(k)}]
\]

\[
= \mathbb{E} \left[ \sum_{h = h_i}^{h_{i+1} - 1} \bar{r}_i^{(h)} \left| \bar{s}(h_i), \bar{a}_i(h_i), \bar{a}_{-i} \sim \pi_i^{(t)}(\cdot | \bar{s}(h_i)), h_i' \sim \text{GEOMETRIC}(1 - \gamma) \right. \right]
\]

\[
= \sum_{h = 0}^{h_{i+1} - 1} \bar{r}_i^{(h)} \left[ \mathbb{E} \left[ \mathbb{E}_{h_i'} \left[ \mathbb{E}_{h} \left( \bar{s}(h_i), \bar{a}_i(h_i), \bar{a}_{-i} \sim \pi_i^{(t)}(\cdot | \bar{s}(h_i)), h_i' \sim \text{GEOMETRIC}(1 - \gamma) \right) \right] \right] \right]
\]

\[
= \sum_{h = 0}^{h_{i+1} - 1} \bar{r}_i^{(h)} \left[ \mathbb{E} \left[ \mathbb{E}_{h_i'} \left[ \mathbb{E}_{h} \left( \bar{s}(h_i), \bar{a}_i(h_i), \bar{a}_{-i} \sim \pi_i^{(t)}(\cdot | \bar{s}(h_i)), h_i' \sim \text{GEOMETRIC}(1 - \gamma) \right) \right] \right] \right]
\]

\[
= \sum_{h = 0}^{h_{i+1} - 1} \bar{r}_i^{(h)} \left[ \mathbb{E} \left[ \mathbb{E}_{h_i'} \left[ \mathbb{E}_{h} \left( \bar{s}(h_i), \bar{a}_i(h_i), \bar{a}_{-i} \sim \pi_i^{(t)}(\cdot | \bar{s}(h_i)), h_i' \sim \text{GEOMETRIC}(1 - \gamma) \right) \right] \right] \right]
\]

\[
= \sum_{h = 0}^{h_{i+1} - 1} \bar{r}_i^{(h)} \left[ \mathbb{E} \left[ \mathbb{E}_{h_i'} \left[ \mathbb{E}_{h} \left( \bar{s}(h_i), \bar{a}_i(h_i), \bar{a}_{-i} \sim \pi_i^{(t)}(\cdot | \bar{s}(h_i)), h_i' \sim \text{GEOMETRIC}(1 - \gamma) \right) \right] \right] \right]
\]

where in (a) we change the range of index \( h \) while using the same initial state and action, (b) is due to the tower property, (c) follows that \( \mathbb{E}_{h_i'} \left[ \mathbb{E}_{h} \right] = 1 - (1 - (1 - p)^h) = \gamma^h \), where \( p = 1 - \gamma \), and we also apply the monotone convergence and dominated convergence theorems for swapping the sum and the expectation.

### D.5.2 Proof of Theorem 40

We apply Lemma 34 to the potential function \( \Phi^\pi(\rho) \) at two consecutive policies \( \pi^{(t+1)} \) and \( \pi^{(t)} \) in Algorithm 11, where \( \rho \) is the initial state distribution. We use the shorthand \( \Phi^{(t)}(\rho) \) for \( \Phi^{\pi^{(t)}}(\rho) \), the value of potential function at policy \( \pi^{(t)} \). The proof extends Lemma 38 by accounting for the statistical error in Assumption 16.
Lemma 56 (Policy improvement: Markov potential games) Let Assumption 15 hold. In Algorithm 11, the difference of potential functions $\Phi^\pi(\rho)$ at two consecutive policies $\pi_i^{(t+1)}$ and $\pi_i^{(t)}$, $\Phi_i^{(t+1)}(\rho) - \Phi_i^{(t)}(\rho)$ is lower bounded by either (i) or (ii).

\[
(i) \quad \frac{1}{4\eta(1-\gamma)} \sum_{i} \sum_s d_{\rho}^{\pi_i^{(t+1)}-\pi_i^{(t)}}(s) \left( \frac{1}{1-\gamma} \right) \| \pi_i^{(t+1)}(\cdot|s) - \pi_i^{(t)}(\cdot|s) \|^2 - \frac{4\eta^2 A W^2 N^2}{(1-\gamma)^3}
\]

\[
(ii) \quad \frac{1}{4\eta(1-\gamma)} \sum_{i} \sum_s d_{\rho}^{\pi_i^{(t+1)}-\pi_i^{(t)}}(s) \left( 1 - \frac{4\eta\kappa^CA}{(1-\gamma)^4} \right) \| \pi_i^{(t+1)}(\cdot|s) - \pi_i^{(t)}(\cdot|s) \|^2
\]

where $\eta$ is the stepsize, $N$ is the number of players, $A$ is the size of one player’s action space, $W$ is the 2-norm bound of $\hat{w}_i^{(t)}$, and $\kappa_\rho$ is the distribution mismatch coefficient relative to $\rho$.

**Proof.** We let $\pi' = \pi_i^{(t+1)}$ and $\pi = \pi_i^{(t)}$ for brevity. We first express $\Phi_i^{(t+1)}(\rho) - \Phi_i^{(t)}(\rho) = \text{Diff}_0 + \text{Diff}_\beta$, where $\text{Diff}_0$ and $\text{Diff}_\beta$ are given as those in (5.6).

**Bounding Diff$_0$.** By the property of the potential function $\Phi^\pi(\rho)$ and Remark 11

\[
\Phi_i^{\pi_i^{(t+1)}-\pi_i^{(t)}}(\rho) - \Phi_i^{\pi}(\rho) = V_i^{\pi_i^{(t+1)}-\pi_i^{(t)}}(\rho) - V_i^{\pi}(\rho)
\]

\[
= \frac{1}{1-\gamma} \sum_s d_{\rho}^{\pi_i^{(t+1)}-\pi_i^{(t)}}(s) \langle \pi_i^{(t+1)}(\cdot|s) - \pi_i^{(t)}(\cdot|s) \rangle \hat{Q}_i^{\pi_i^{(t+1)}-\pi_i^{(t)}}(s, a_i).
\]

The optimality of $\pi_i^{(t+1)} = \pi_i^{(t)}$ in line 14 of Algorithm 11 leads to

\[
\langle \pi_i^{(t)}(\cdot|s), \hat{Q}_i^{(t)}(s, \cdot) \rangle_{A_i} - \frac{1}{2\eta} \| \pi_i^{(t)}(\cdot|s) - \pi_i^{(t)}(\cdot|s) \|^2 \geq \langle \pi_i^{(t)}(\cdot|s), \hat{Q}_i^{(t)}(s, \cdot) \rangle_{A_i}.
\]

(D.3)

Hence,

\[
\Phi_i^{\pi_i^{(t+1)}-\pi_i^{(t)}}(\rho) - \Phi_i^{\pi}(\rho) \geq \frac{1}{2\eta(1-\gamma)} \sum_s d_{\rho}^{\pi_i^{(t+1)}-\pi_i^{(t)}}(s) \| \pi_i^{(t+1)}(\cdot|s) - \pi_i^{(t)}(\cdot|s) \|^2
\]

\[
+ \frac{1}{1-\gamma} \sum_s d_{\rho}^{\pi_i^{(t+1)}-\pi_i^{(t)}}(s) \langle \pi_i^{(t+1)}(\cdot|s) - \pi_i^{(t)}(\cdot|s), \hat{Q}_i^{\pi_i^{(t+1)}-\pi_i^{(t)}}(s, \cdot) - \hat{Q}_i^{(t)}(s, \cdot) \rangle_{A_i}
\]

Therefore,

\[
\text{Diff}_0 \geq \frac{1}{2\eta(1-\gamma)} \sum_{i} \sum_s d_{\rho}^{\pi_i^{(t+1)}-\pi_i^{(t)}}(s) \| \pi_i^{(t+1)}(\cdot|s) - \pi_i^{(t)}(\cdot|s) \|^2
\]

\[
+ \frac{1}{1-\gamma} \sum_{i} \sum_s d_{\rho}^{\pi_i^{(t+1)}-\pi_i^{(t)}}(s) \langle (\pi_i^{(t+1)} - \pi_i^{(t)})(\cdot|s), \hat{Q}_i^{(t)}(s, \cdot) - \hat{Q}_i^{(t)}(s, \cdot) \rangle_{A_i}.
\]
However,
\[
\sum_{s} d_{\rho}^{(t+1), \pi^{(t)}_{i}}(s) \langle \pi^{(t+1)}_{i} - \pi^{(t)}_{i} \rangle(s) , Q^{(t)}_{i}(s, \cdot) - \hat{Q}^{(t)}_{i}(s, \cdot) \rangle_{A_{i}} \\
\geq \frac{1}{4\eta(1-\gamma)} \sum_{i=1}^{N} \sum_{s} d_{\rho}^{(t+1), \pi^{(t)}_{i}}(s) \left\| \pi^{(t+1)}_{i}(\cdot | s) - \pi^{(t)}_{i}(\cdot | s) \right\|^{2} - \frac{\eta}{1-\gamma} \sum_{i=1}^{N} \sum_{s} d_{\rho}^{(t+1), \pi^{(t)}_{i}}(s) \left\| \hat{Q}^{(t)}_{i}(s, \cdot) - \hat{Q}^{(t)}_{i}(s, \cdot) \right\|^{2}
\]
\[
(\text{D.4})
\]

where (a) follows the inequality \( \langle x, y \rangle \leq \left\| x \right\| \left\| y \right\| \) for \( \eta > 0 \), and we choose \( \eta' = 2\eta \) in (b).

Therefore,
\[
\text{Diff}_{\alpha} \geq \frac{1}{4\eta(1-\gamma)} \sum_{i=1}^{N} \sum_{s} d_{\rho}^{(t+1), \pi^{(t)}_{i}}(s) \left\| \pi^{(t+1)}_{i}(\cdot | s) - \pi^{(t)}_{i}(\cdot | s) \right\|^{2} - \frac{\eta}{1-\gamma} \sum_{i=1}^{N} \sum_{s} d_{\rho}^{(t+1), \pi^{(t)}_{i}}(s) \left\| \hat{Q}^{(t)}_{i}(s, \cdot) - \hat{Q}^{(t)}_{i}(s, \cdot) \right\|^{2}.
\]

Bounding Diff\(\beta\). For simplicity, we denote \( \tilde{\pi}_{-ij} \) as the joint policy of players \( N \setminus \{i, j\} \) where players \( < i \) and \( i \sim j \) use \( \pi \) and players \( > j \) use \( \pi' \). As done in the proof of Lemma 38, we can bound each summand in \( \text{Diff}_{\beta} \) except for the last step from (c) to (d),
\[
\Phi_{\tilde{\pi}_{-ij}, \pi_{i}', \pi_{j}'}(\rho) - \Phi_{\tilde{\pi}_{-ij}, \pi_{i}, \pi_{j}}(\rho) - \Phi_{\tilde{\pi}_{-ij}, \pi_{i}', \pi_{j}'}(\rho) + \Phi_{\tilde{\pi}_{-ij}, \pi_{i}, \pi_{j}}(\rho)
\geq \frac{1}{(1-\gamma)^{3}} \left( \max_{s} \left\| \pi_{i}'(\cdot | s) - \pi_{i}(\cdot | s) \right\|_{1} \right) \left( \max_{s} \left\| \pi_{j}'(\cdot | s) - \pi_{j}(\cdot | s) \right\|_{1} \right)
- \frac{1}{(1-\gamma)^{2}} \left( \max_{s} \left\| \pi_{j}'(\cdot | s) - \pi_{j}(\cdot | s) \right\|_{1} \right) \left( \max_{s} \left\| \pi_{i}'(\cdot | s) - \pi_{i}(\cdot | s) \right\|_{1} \right)
\geq \frac{8\eta^{2}AW^{2}}{(1-\gamma)^{3}}
\]
\[
(\text{D.4})
\]

where (d) follows a direct result from the optimality of \( \pi^{(t+1)}_{j} \) given by (D.3),
\[
\left\| \pi^{(t+1)}_{j}(\cdot | s) - \pi^{(t)}_{j}(\cdot | s) \right\| \leq 2\eta \left\| \hat{Q}^{(t)}_{i}(s, \cdot) \right\| \leq 2\eta W
\]
and that \( \left\| \cdot \right\|_{1} \leq \sqrt{A} \left\| \cdot \right\| \). Therefore,
\[
\text{Diff}_{\beta} \geq - \frac{4\eta^{2}AW^{2}N^{2}}{(1-\gamma)^{3}}.
\]
\[
(\text{D.5})
\]
We now complete the proof of (i) by combining (D.4) and (D.5) and we also employ that

\[
\sum_s d_{\rho}^{(t+1), \pi_i}^i(s) \|Q_i^{(t)}(s, \cdot) - \hat{Q}_i^{(t)}(s, \cdot)\|^2 \\
\leq (a) \frac{\kappa_{\rho}}{1 - \gamma} \sum_s d_{\rho}^{(t), \pi_i}^i(s) \|Q_i^{(t)}(s, \cdot) - \hat{Q}_i^{(t)}(s, \cdot)\|^2 \\
\leq (b) \frac{\kappa_{\rho} A}{(1 - \gamma) \xi} L_i^{(t)}(\bar{w}_i^{(t)})
\]

where (a) follows the definition of \(\kappa_{\rho}\) and (b) is the definition of \(L_i^{(t)}(\bar{w}_i^{(t)})\):

\[
L_i^{(t)}(\bar{w}_i^{(t)}) := \mathbb{E}_{s \sim d_{\rho}(\cdot, a_i, \pi_i^{(t)})(\cdot | s)} \left[ (Q_i^{(t)}(s, a_i) - \hat{Q}_i^{(t)}(s, a_i))^2 \right] \\
\geq \frac{\xi}{A} \mathbb{E}_{s \sim d_{\rho}(\cdot, a_i, \pi_i^{(t)})(\cdot | s)} (Q_i^{(t)}(s, a_i) - \hat{Q}_i^{(t)}(s, a_i))^2.
\]

Alternatively, as done in Lemma 38 we can apply Lemma 37 to each summand of \(\text{Diff}_\beta\) and show that

\[
\text{Diff}_\beta \geq -\frac{2\kappa_{\rho}^3 N A}{(1 - \gamma)^5} \sum_{i=1}^{N} d_{\rho}^{(t+1), \pi_i^{(t)}(\cdot | s)} \|\pi_i^{(t)}(\cdot | s) - \pi_i^{(t+1)}(\cdot | s)\|^2.
\]

Combining the inequality above with (D.4) finishes the proof of (ii).

**Proof.** [Theorem 40] By the optimality of \(\pi_i^{(t+1)}\) in line 14 of Algorithm 11 for any \(\pi_i' \in \Pi_i\),

\[
\left\langle (1 - \xi)\pi_i'(\cdot | s) + \xi \frac{1}{A} - \pi_i^{(t+1)}(\cdot | s), \eta \hat{Q}_i^{(t)}(s, \cdot) - \pi_i^{(t+1)}(\cdot | s) + \pi_i^{(t)}(\cdot | s) \right\rangle_{A_i} \leq 0
\]

which leads to

\[
\left\langle \pi_i'(\cdot | s) - \pi_i^{(t+1)}(\cdot | s), \eta \hat{Q}_i^{(t)}(s, \cdot) \right\rangle_{A_i} \\
\leq \left\langle \pi_i'(\cdot | s) - \pi_i^{(t+1)}(\cdot | s), \pi_i^{(t+1)}(\cdot | s) - \pi_i^{(t)}(\cdot | s) \right\rangle_{A_i} \\
+ \frac{\xi}{1 - \xi} \left\langle \pi_i^{(t+1)}(\cdot | s) - \frac{1}{A} 1, \eta \hat{Q}_i^{(t)}(s, \cdot) - \pi_i^{(t+1)}(\cdot | s) + \pi_i^{(t)}(\cdot | s) \right\rangle \\
\leq \|\pi_i^{(t+1)}(\cdot | s) - \pi_i^{(t)}(\cdot | s)\| + \eta \xi W
\]
where the last inequality is because of $\|\hat{Q}_i^{(t)}(s, \cdot)\| \leq W$ and $\xi \leq \frac{1}{2}$. Hence, if $\eta \leq \frac{1}{W}$, then for any $\pi_i' \in \Pi_i$,

$$\langle \pi_i'(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s), \bar{Q}_i^{(t)}(s, \cdot) \rangle_{\mathcal{A}_i}$$

$$= \langle \pi_i'(\cdot \mid s) - \pi_i^{(t+1)}(\cdot \mid s), \bar{Q}_i^{(t)}(s, \cdot) \rangle_{\mathcal{A}_i} + \langle \pi_i^{(t+1)}(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s), \bar{Q}_i^{(t)}(s, \cdot) \rangle_{\mathcal{A}_i}$$

$$+ \langle \pi_i^{(t)}(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s), \bar{Q}_i^{(t)}(s, \cdot) - \hat{Q}_i^{(t)}(s, \cdot) \rangle_{\mathcal{A}_i}$$

$$(a) \quad \frac{1}{\eta} \|\pi_i^{(t+1)}(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s)\| \leq \frac{4}{1-\gamma} + \xi W + \|\pi_i^{(t+1)}(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s)\|\|\hat{Q}_i^{(t)}(s, \cdot)\|$$

$$(b) \quad \frac{1}{\eta} \|\pi_i^{(t+1)}(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s)\| + \xi W$$

where we apply (D.7) and the Cauchy-Schwarz inequality in (a), and (b) is because $\|\hat{Q}_i^{(t)}(s, \cdot)\| \leq W$ and $\eta \leq \frac{1}{W}$. As done in the proof of Theorem 32, the different steps begin from (b) in (5.11).

$$\sum_{t=1}^{T} \max_{\pi_i'} \left( \max_{\pi_i} V_i^{\pi_i', \pi_i^{(t)}}(\rho) - V_i^{\pi_i^{(t)}}(\rho) \right)$$

$$\leq \frac{\xi TW}{1-\gamma}$$

$$+ \frac{1}{\eta(1-\gamma)} \sum_{t=1}^{T} \sum_{s} d_{\rho}^{\pi_i', \pi_i^{(t)}}(s) \|\pi_i^{(t+1)}(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s)\|$$

$$+ \frac{\kappa_{\rho}}{1-\gamma} \sum_{t=1}^{T} \sum_{s} d_{\rho}^{\pi_t^{(t)}(\cdot \mid s), \pi_i^{(t)}(s, \cdot)}(s) \pi_i^{(t)}(\cdot \mid s) - \hat{Q}_i^{(t)}(s, \cdot) \rangle_{\mathcal{A}_i}$$

$$\leq \frac{\sqrt{\kappa_{\rho}}}{\eta(1-\gamma)^{3/2}} \sum_{t=1}^{T} \sum_{s} d_{\rho}^{\pi_t^{(t)}(\cdot \mid s), \pi_i^{(t)}(s, \cdot)}(s) \|\pi_i^{(t+1)}(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s)\|$$

$$+ \frac{\xi TW}{1-\gamma} + \frac{\kappa_{\rho}}{1-\gamma} \sum_{t=1}^{T} \sqrt{\frac{AL_i^{(t)}(\tilde{w}_i^{(t)})}{\xi}}$$

$$(c) \quad \frac{\sqrt{\kappa_{\rho}}}{\eta(1-\gamma)^{3/2}} \sqrt{T} \times \left[ \sum_{t=1}^{T} \sum_{s} d_{\rho}^{\pi_t^{(t)}(\cdot \mid s), \pi_i^{(t)}(s, \cdot)}(s) \pi_i^{(t+1)}(\cdot \mid s) - \pi_i^{(t)}(\cdot \mid s) \|^2 \right. + \frac{\xi TW}{1-\gamma}$$

$$+ \frac{\kappa_{\rho}}{1-\gamma} \sum_{t=1}^{T} \sqrt{\frac{AL_i^{(t)}(\tilde{w}_i^{(t)})}{\xi}}$$

(D.8)
where we slightly abuse the notation $\pi'_i$ in (b) to represent $\arg\max_{\pi'_i}$ and $i$ represents $\arg\max_i$ as in (5.11), (c) is due to the definition of the distribution mismatch coefficient (see it in Definition 2):

$$\frac{d_{\rho}^{\pi'_i, \pi_{-i}^{(t)}}(s)}{d_{\rho}^{\pi'^{(t+1)}_{-i}, \pi_{-i}^{(t)}}(s)} \leq \frac{d_{\rho}^{\pi'_i, \pi_{-i}^{(t)}}(s)}{(1 - \gamma)\rho(s)} \leq \frac{\kappa_{\rho}}{1 - \gamma}$$

(d) follows the Cauchy–Schwarz inequality, the inequality $\sqrt{\sum_i x_i} \leq \sum_i x_i$ for any $x_i \geq 0$, the Jensen’s inequality, and the definition of $L_i^{(t)}(\hat{w}_i^{(t)})$,

$$|\sum_s d_{\rho}^{\pi^{(t)}}(s)\langle \pi'_i(\cdot | s) - \pi_i^{(t)}(\cdot | s), \bar{Q}_i^{(t)}(s, \cdot) - \hat{Q}_i^{(t)}(s, \cdot) \rangle_{\mathcal{A}_i}| \leq \sqrt{\sum_s d_{\rho}^{\pi^{(t)}}(s)} \sum_s d_{\rho}^{\pi^{(t)}}(s) \sum_{a_i} \left( Q_i^{(t)}(s, a_i) - \hat{Q}_i^{(t)}(s, a_i) \right)^2 \leq \sqrt{AL_i^{(t)}(\hat{w}_i^{(t)})}$$

where $\hat{Q}_i^{(t)}(s, a_i) = \langle \phi_i(s, a_i), \hat{w}_i^{(t)} \rangle$, and we replace $i (\arg\max_i$ in (b)) in the square root term in (c) by the sum over all players.

If we proceed (D.8) with the first bound (i) in Lemma 56 then,

$$\mathbb{E} \left[ \sum_{t=1}^{T} \max_i \left( \max_{\pi'_i} V_i^{\pi'_i, \pi_{-i}^{(t)}}(\rho) - V_i^{\pi^{(t)}}(\rho) \right) \right] \approx (a) \sqrt{\frac{\kappa_{\rho} T}{\eta(1 - \gamma)^{\frac{3}{2}}}} \eta(1 - \gamma)(\Phi^{(N+1)} - \Phi^{(1)}) + \frac{\eta^3 A W^2 N^2}{(1 - \gamma)^2} T + \frac{\eta^2 \kappa_{\rho} A}{(1 - \gamma)} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E} \left[ L_i^{(t)}(\hat{w}_i^{(t)}) \right]$$

$$+ \frac{\xi T W}{1 - \gamma} + \frac{\kappa_{\rho}}{1 - \gamma} \sum_{t=1}^{T} \sqrt{\frac{\mathbb{E} \left[ L_i^{(t)}(\hat{w}_i^{(t)}) \right]}{\xi}}$$

$$\approx (b) \sqrt{\frac{\kappa_{\rho} T C_{\Phi}}{\eta(1 - \gamma)^2}} + T \left[ \frac{\eta \kappa_{\rho} A W^2 N^2}{(1 - \gamma)^5} \right] + \frac{\kappa_{\rho} T}{(1 - \gamma)^2} \sqrt{\frac{A N \epsilon_{\text{stat}}}{\xi}} + \frac{\xi T W}{1 - \gamma}$$

where we apply the first bound (i) in Lemma 56 and the telescoping sum for (a), and we use the boundedness of the potential function: $|\Phi^\pi - \Phi^\pi'| \leq C_{\Phi}$ for any $\pi$ and $\pi'$, and further simplify the bound in (f) by Assumption 16 We complete the proof of (i) by taking stepsize $\eta = \frac{(1 - \gamma)^{3/2} \sqrt{C_{\Phi}}}{WN^3 A T}$ and exploration rate $\xi \leq \left( \frac{\kappa_{\rho}^2 N A \epsilon_{\text{stat}}}{(1 - \gamma)^2 WN^3} \right)^{\frac{1}{3}}$.  

278
If we proceed with the first bound (ii) in Lemma 56 with the choice of \( \eta \leq \frac{(1-\gamma)^4}{16\kappa^3N\rho} \), then,

\[
E \left[ \sum_{t=1}^{T} \max_{\pi_i^t} \left( \max_{\pi_i} V_{i,\pi_i}^{\pi_i^t}(\rho) - V_{i}^{\pi_i^t}(\rho) \right) \right] \\
\leq \frac{\sqrt{\kappa_0T}}{\eta(1-\gamma)^{3/2}} \sqrt{\eta(1-\gamma)(\Phi(N+1) - \Phi(1)) + \frac{\eta^2\kappa_0A}{(1-\gamma)\xi} \sum_{t=1}^{T} \sum_{i=1}^{N} E \left[ L_i^{\pi_i^t}(\hat{\omega}_i^{\pi_i^t}) \right]} \\
+ \frac{\xi TW}{1-\gamma} + \frac{\kappa_0}{1-\gamma} \sum_{t=1}^{T} \sqrt{\frac{A E \left[ L_i^{\pi_i^t}(\hat{\omega}_i^{\pi_i^t}) \right]}{\xi}} \\
\geq \frac{1}{8\eta} \sum_{i=1}^{N} \left\| \pi_i^{(t+1)}(\cdot | s) - \pi_i^{(t)}(\cdot | s) \right\|^2 \\
- \eta \sum_{i=1}^{N} \left\| Q_i^{(t)}(s, \cdot) - \hat{Q}_i^{(t)}(s, \cdot) \right\|^2 \\
\]

which completes the proof if we choose \( \eta = \frac{(1-\gamma)^4}{16\kappa^3N\rho} \) and exploration rate \( \xi \leq \left( \frac{\kappa_0^2NA\epsilon_{\text{stat}}}{(1-\gamma)^{3/2}} \right)^{1/3} \). □

### D.5.3 Proof of Theorem 41

We first establish policy improvement regarding the \( Q \)-function at two consecutive policies \( \pi^{(t+1)} \) and \( \pi^{(t)} \) in Algorithm 11.

**Lemma 57 (Policy improvement: Markov cooperative games)** For MPG (5.1) with identical rewards and an initial state distribution \( \rho > 0 \), if all players independently perform the policy update in Algorithm 11 with stepsize \( \eta \leq \frac{1-\gamma}{2N} \), then for any \( t \) and any \( s \),

\[
E_{a \sim \pi^{(t+1)}(\cdot | s)} \left[ Q^{(t)}(s, a) \right] - E_{a \sim \pi^{(t)}(\cdot | s)} \left[ Q^{(t)}(s, a) \right] \geq \frac{1}{8\eta} \sum_{i=1}^{N} \left\| \pi_i^{(t+1)}(\cdot | s) - \pi_i^{(t)}(\cdot | s) \right\|^2 \\
- \eta \sum_{i=1}^{N} \left\| Q_i^{(t)}(s, \cdot) - \hat{Q}_i^{(t)}(s, \cdot) \right\|^2 \\
\]

where \( \eta \) is the stepsize and \( N \) is the number of players,
\textbf{Proof.} As done in the proof of Lemma \ref{lem:39}, we let $\Psi^\pi := \mathbb{E}_{a \sim \pi(\cdot \mid x)} \left[ Q(t)(s, a) \right]$ and \eqref{eq:5.12} holds, where $Q(t) := Q^\pi(t)$. By taking $\pi' = \pi(t + 1)$ and $\pi = \pi(t)$ for \eqref{eq:5.12},

\[
\mathbb{E}_{a \sim \pi^\pi(\cdot \mid x)} \left[ Q(t)(s, a) \right] - \mathbb{E}_{a \sim \pi(\cdot \mid x)} \left[ Q(t)(s, a) \right]
= \sum_{i=1}^{N} \sum_{a_i} \left( \pi_i'(a_i \mid s) - \pi_i(a_i \mid s) \right) \hat{Q}_i(t)(s, a_i) \\
+ \sum_{i=1}^{N} \sum_{a_i} \left( \pi_i'(a_i \mid s) - \pi_i(a_i \mid s) \right) \left( \hat{Q}_i(t)(s, a_i) - \hat{Q}_i(t')(s, a_i) \right) \\
+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{a_i, a_j} \left( \pi_i'(a_i \mid s) - \pi_i(a_i \mid s) \right) \left( \pi_j'(a_j \mid s) - \pi_j(a_j \mid s) \right) \mathbb{E}_{a \sim \pi_{i,j}^t \sim \pi_{-i,j}(\cdot \mid x)} \left[ Q(t)(s, a) \right]
\]
\[
\overset{(a)}{=} \sum_{i=1}^{N} \frac{1}{2\eta} \left\| \pi_i'(\cdot \mid s) - \pi_i(\cdot \mid s) \right\|^2 \\
- \sum_{i=1}^{N} \left( \frac{1}{2\eta} \left\| \pi_i'(\cdot \mid s) - \pi_i(\cdot \mid s) \right\|^2 + \frac{\eta'}{2} \left\| \tilde{Q}_i(t)(s, \cdot) - \hat{Q}_i(t)(s, \cdot) \right\|^2 \right) \\
- \frac{1}{1-\gamma} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{a_i, a_j} \left| \pi_i'(a_i \mid s) - \pi_i(a_i \mid s) \right| \left| \pi_j'(a_j \mid s) - \pi_j(a_j \mid s) \right|
\]
\[
\overset{(b)}{=} \sum_{i=1}^{N} \frac{1}{4\eta} \left\| \pi_i'(\cdot \mid s) - \pi_i(\cdot \mid s) \right\|^2 - \sum_{i=1}^{N} \eta \left\| \tilde{Q}_i(t)(s, \cdot) - \hat{Q}_i(t)(s, \cdot) \right\|^2 \\
- \frac{1}{2(1-\gamma)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left( \left\| \pi_i'(\cdot \mid s) - \pi_i(\cdot \mid s) \right\|^2 + \left\| \pi_j'(\cdot \mid s) - \pi_j(\cdot \mid s) \right\|^2 \right)
\]
\[
= \sum_{i=1}^{N} \frac{1}{4\eta} \left\| \pi_i'(\cdot \mid s) - \pi_i(\cdot \mid s) \right\|^2 - \sum_{i=1}^{N} \eta \left\| \tilde{Q}_i(t)(s, \cdot) - \hat{Q}_i(t)(s, \cdot) \right\|^2 \\
- \frac{N-1}{2(1-\gamma)} \sum_{i=1}^{N} \left\| \pi_i'(\cdot \mid s) - \pi_i(\cdot \mid s) \right\|^2
\]
\[
\overset{(c)}{=} \sum_{i=1}^{N} \frac{1}{8\eta} \left\| \pi_i'(\cdot \mid s) - \pi_i(\cdot \mid s) \right\|^2 - \sum_{i=1}^{N} \eta \left\| \tilde{Q}_i(t)(s, \cdot) - \hat{Q}_i(t)(s, \cdot) \right\|^2
\]

where (a) is due to the optimality condition \eqref{eq:D3}, the inequality $\langle x, y \rangle \leq \frac{\|x\|^2}{2\eta'} + \frac{\eta'\|y\|^2}{2}$ for $\eta' > 0$, and $Q(t)(s, a) \leq \frac{1}{1-\gamma}$, (b) is due to $\langle x, y \rangle \leq \frac{\|x\|^2 + \|y\|^2}{2}$ and $\eta' = 2\eta$, and (c) follows the choice of $\eta \leq \frac{1}{4N}$. \hfill \square
\[ V^{(t+1)}(\rho) - V^t(\rho) = \frac{1}{1-\gamma} \sum_{s,a} d^{(t+1)}_\rho(s) \left( \pi^{(t+1)}(a \mid s) - \pi^t(a \mid s) \right) Q^t(s,a) \]
\[ \geq \frac{1}{8\eta(1-\gamma)} \sum_{i=1}^N \sum_{s} d^{(t+1)}_\rho(s) \| \pi^{(t+1)}_i(s \mid s) - \pi^t_i(s \mid s) \|^2 \]
\[ - \frac{\eta}{1-\gamma} \sum_{i=1}^N \sum_{s} d^{(t+1)}_\rho(s) \| \hat{Q}^t_i(s, \cdot) - \hat{Q}^t_i(s, \cdot) \|^2 \]
\[ \geq \frac{1}{8\eta(1-\gamma)} \sum_{i=1}^N \sum_{s} d^{(t+1)}_\rho(s) \| \pi^{(t+1)}_i(s \mid s) - \pi^t_i(s \mid s) \|^2 \]
\[ - \frac{\eta \kappa \rho A}{\xi(1-\gamma)^2} \sum_{i=1}^N L_i^{(t)}(\hat{w}_i^{(t)}) \]

where the last inequality is due to that
\[ \sum_{s} d^{(t+1)}_\rho(s) \| \hat{Q}^{\pi^t}_i(s, \cdot) - \hat{Q}^{\pi^t}_i(s, \cdot) \|^2 \]
\[ \leq \frac{\kappa \rho}{1-\gamma} \sum_{s} d^{(t)}_\rho(s) \| \tilde{Q}^{\pi^t}_i(s, \cdot) - \tilde{Q}^{\pi^t}_i(s, \cdot) \|^2 \]
\[ = \frac{\kappa \rho A}{(1-\gamma)\xi} \sum_{s} d^{(t)}_\rho(s) \sum_{a_i} \frac{\xi}{A} \left( Q^{(t)}_i(s, a_i) - \langle \phi_i(s, a_i), \hat{w}_i^{(t)} \rangle \right)^2 \]
\[ \leq \frac{\kappa \rho A}{(1-\gamma)\xi} \mathbb{E} \sum_{s} d^{(t)}_\rho(s) \sum_{a_i} \left( Q^{(t)}_i(s, a_i) - \langle \phi_i(s, a_i), \hat{w}_i^{(t)} \rangle \right)^2 \]
\[ = \frac{\kappa \rho A}{(1-\gamma)\xi} L_i^{(t)}(\hat{w}_i^{(t)}) \]

where (a) follows the definition of \( \kappa_\rho \) and (b) is the definition of \( L_i^{(t)}(\hat{w}_i^{(t)}) \).

By the same argument as the proof of Theorem 40,
\[ \sum_{i=1}^T \max_{\pi_i} \left( \max_{\pi_i} V^{\pi_i, \pi_{-i}^{(t)}}(\rho) - V^{(t)}(\rho) \right) \]
\[ \leq \frac{\sqrt{\kappa \rho}}{\eta(1-\gamma)^{\frac{3}{2}}} \sqrt{\sum_{i=1}^T \sum_{s} d^{\pi_i, \pi_{-i}^{(t)}}_\rho(s) \sum_{i=1}^T \sum_{i=1}^N \sum_{s} d^{(t+1)}_\rho(s) \| \pi^{(t+1)}_i(s \mid s) - \pi^t_i(s \mid s) \|^2} \]
\[ + \frac{\xi TW}{1-\gamma} + \frac{\kappa \rho}{1-\gamma} \sum_{i=1}^T \sqrt{\frac{A L_i^{(t)}(\hat{w}_i^{(t)})}{\xi}} \]

281
By taking expectation and the Jensen’s inequality,
\[
\mathbb{E}
\left[
\sum_{t=1}^{T}
\max_{i}
\left(
\max_{\pi'_t}
V_{\pi'_t, \pi_{t-1}}(\rho) - V^{\pi_t}(\rho)
\right)
\right]
\lesssim
\sqrt{\frac{\kappa_T}{\eta(1-\gamma)^{3/2}}}
8\eta(1-\gamma)(V^{(N+1)} - V^{(1)})
+ \frac{8\eta^2\kappa_A}{(1-\gamma)^2}\xi
\sum_{t=1}^{T} T
\mathbb{E}
\left[
L_i^{(t)}(\hat{w}_i^{(t)})
\right]
+ \xi TW + \frac{\kappa_T}{1-\gamma}
\sum_{t=1}^{T}
\sqrt{\frac{\lambda}{\xi}}
\mathbb{E}
\left[
L_{i}^{(t)}(\hat{w}_{i}^{(t)})
\right]
\lesssim
\sqrt{\frac{8\kappa_T}{\eta(1-\gamma)^{3/2}}}
+ \frac{\kappa_T}{1-\gamma}
\sqrt{\frac{8AN}{(1-\gamma)^2}\xi_{stat} + \xi TW}.
\]

We complete the proof by taking stepsize \( \eta = \frac{1-\gamma}{2N_{A}} \), exploration rate \( \xi \leq \left( \frac{\kappa_{N_{A}}}{(1-\gamma)^2W^{2}} \right)^{\frac{1}{3}} \), and using \( V^{(N+1)} - V^{(1)} \leq \frac{1}{1-\gamma} \).

**D.5.4 Sample complexity**

We present our sample complexity guarantees for Algorithm 11 in which the regression problem (5.15) in each iteration is approximately solved by the stochastic projected gradient descent (D.20). We measure the sample complexity by the total number of trajectory samples \( TK \), where \( T \) is the number of iterations and \( K \) is the batch size of trajectories.

**Corollary 1 (Sample complexity for Markov potential games)** Assume the setting in Theorem 40 by excluding Assumption 16. Suppose we compute \( \hat{w}_{i}^{(k)} := \frac{1}{K} \sum_{k=1}^{K} \beta_{i}^{(k)} \hat{w}_{i}^{(k)} \) via stochastic projected gradient descent (D.20) with stepsize \( \lambda^{(k)} = \frac{2}{2+k} \) and \( \beta_{i}^{(k)} = \frac{1/\lambda^{(k)}}{\sum_{n=1}^{N} 1/\lambda^{(n)}} \). Then, if we choose stepsize \( \eta = \frac{(1-\gamma)^{3/2}/C_{\Phi}}{WN_{A}T} \) and exploration rate \( \xi = \min \left( \left( \frac{\kappa_{N_{A}}}{(1-\gamma)^{2}K} \right)^{\frac{1}{3}}, \frac{1}{2} \right) \), then,
\[
\mathbb{E} \left[ \text{Nash-Regret}(T) \right] \lesssim \frac{\sqrt{\kappa_{T}WN(AC_{\Phi})^{\frac{3}{2}}}}{(1-\gamma)^{\frac{3}{2}} T^{\frac{1}{2}}} + \frac{W(\kappa_{T}^{2}AnD)^{\frac{3}{2}}}{(1-\gamma)^{\frac{5}{2}} K^{\frac{2}{3}}}.
\]

Furthermore, if we choose stepsize \( \eta = \frac{(1-\gamma)^{4}}{16\kappa_{T}^{2}N_{A}} \) and exploration rate \( \xi = \min \left( \left( \frac{\kappa_{T}^{2}AnD}{(1-\gamma)^{2}K} \right)^{\frac{1}{3}}, \frac{1}{2} \right) \), then,
\[
\mathbb{E} \left[ \text{Nash-Regret}(T) \right] \lesssim \frac{\kappa_{T}^{2}\sqrt{ANC_{\Phi}}}{(1-\gamma)^{3/2} T^{1/2}} + \frac{W(\kappa_{T}^{2}AnD)^{3/2}}{(1-\gamma)^{5/2} K^{1/3}}.
\]

Moreover, their sample complexity guarantees are \( TK = O(\frac{1}{\epsilon}) \) or \( TK = O(\frac{1}{\epsilon^7}) \), respectively, for obtaining an \( \epsilon \)-Nash equilibrium.
By the unbiased estimate in Appendix D.5.1, the stochastic gradient $\hat{\nabla}_i(t)$ in (D.20) is also unbiased. We note the variance of the stochastic gradient is bounded by $\frac{1}{(1-\gamma)^2}$. By Lemma 71, if we choose $\lambda(k) = \frac{2}{2+k}$ and $\beta_k^{(K)} = \frac{1/\lambda(k)}{\sum_{r=1}^{K} 1/\lambda(r)}$, then

$$\mathbb{E} \left[ L_i(t)(\hat{w}_i(t)) \right] - L_i(t)(w_i(t)) \leq \frac{dW^2}{(1-\gamma)^2 K},$$

where $L_i(t)(w_i(t)) = 0$ by Assumption 15. Therefore, substitution of $\epsilon_{\text{stat}} \leq \frac{dW^2}{(1-\gamma)^2 K}$ into Theorem 40 yields desired results.

Finally, we let the upper bound on Nash-Regret $(T)$ be $\epsilon > 0$ and calculate the sample complexity $TK = O\left(\frac{1}{\epsilon^2}\right)$ or $TK = O\left(\frac{1}{\epsilon^5}\right)$, respectively. □

**Corollary 2 (Sample complexity for Markov cooperative games)** Excepting Assumption 16, assume the setting in Theorem 41. Suppose we compute $\hat{w}_i(t) := \frac{1}{K} \sum_{k=1}^{K} \beta_k^{(K)} w_i^{(k)}$ via stochastic projected gradient descent (D.20) with stepsize $\lambda(k) = \frac{2}{2+k}$ and $\beta_k^{(K)} = \frac{1/\lambda(k)}{\sum_{r=1}^{K} 1/\lambda(r)}$. Then, if we choose stepsize $\eta = \frac{1-\gamma}{WNA\sqrt{T}}$ and exploration rate $\xi = \min \left( \left( \frac{\kappa^2 AN}{(1-\gamma)^3 K} \right)^{\frac{1}{3}}, \frac{1}{2} \right)$, then,

$$\mathbb{E} \left[ \text{Nash-Regret}(T) \right] \lesssim \frac{\sqrt{\kappa \rho AN}}{(1-\gamma)^2 \sqrt{T}} + \frac{W(\kappa^2 ANd)^{\frac{1}{3}}}{(1-\gamma)^{\frac{5}{2}} K^{\frac{1}{2}}},$$

Moreover, the sample complexity guarantee is $TK = O\left(\frac{1}{\epsilon^2}\right)$ for obtaining an $\epsilon$-Nash equilibrium.

Proof. The proof follows the proof steps of Corollary 1 above. □

### D.6 Proofs for Section 5.7

In this section, we prove Theorem 42 and Theorem 43 in D.6.1 and D.6.2, respectively.

**D.6.1 Proof of Theorem 42**

It is convenient to introduce an auxiliary sequence $\{\alpha(t,\tau)\}_{\tau=0}^{\infty}$ associated with the learning rate $\{\alpha(t)\}_{t=1}^{\infty}$,

$$\alpha(t,\tau) := \begin{cases} \prod_{j=1}^{t} (1 - \alpha^{(j)}), & \text{for } \tau = 0 \\ \alpha^{(\tau)} \prod_{j=\tau+1}^{t} (1 - \alpha^{(j)}), & \text{for } 1 \leq \tau \leq t \\ 0, & \text{for } \tau > t. \end{cases} \quad \text{(D.9)}$$

It is straightforward to verify that $\sum_{\tau=0}^{t-1} \alpha(t-1,\tau) = 1$ for $t \geq 1$.

**Lemma 58** In Algorithm 12, $\psi^{(t)}(s) = \sum_{\tau=1}^{t} \alpha(t,\tau) (x^{(\tau)}_s)^\top Q^{(\tau)}_s y^{(\tau)}_s$ for all $s, t$.  

283
We next deal with \( \text{Diff} \). The optimality of \( \bar{x}_{s}^{(t+1)} \) implies that for any \( x_{s}' \in \Delta(A_1) \),
\[
(x_{s}'^{(t+1)})^\top Q_{s}^{(t)} y_{s}^{(t)} - \frac{1}{2\eta} \|x_{s}'^{(t+1)} - \bar{x}_{s}^{(t+1)}\|^2 \geq (x_{s}'^{(t)})^\top Q_{s}^{(t)} y_{s}^{(t)} - \frac{1}{2\eta} \|x_{s}' - \bar{x}_{s}^{(t+1)}\|^2 + \frac{1}{2\eta} \|x_{s}' - x_{s}^{(t+1)}\|^2
\]
which implies that, by taking \( x_{s}' = \bar{x}_{s}^{(t+1)} \),
\[
(x_{s}^{(t+1)})^\top Q_{s}^{(t)} y_{s}(t) \geq \frac{1}{\eta} \|x_{s}^{(t+1)} - \bar{x}_{s}^{(t+1)}\|^2. \tag{D.10}
\]
The optimality of \( \bar{x}_{s}^{(t+1)} \) implies that for any \( x_{s}' \in \Delta(A_1) \),
\[
(x_{s}'^{(t+1)})^\top Q_{s}^{(t)} y_{s}^{(t)} - \frac{1}{2\eta} \|x_{s}'^{(t+1)} - \bar{x}_{s}^{(t+1)}\|^2 \geq (x_{s}'^{(t)})^\top Q_{s}^{(t)} y_{s}(t) - \frac{1}{2\eta} \|x_{s}' - \bar{x}_{s}^{(t+1)}\|^2 + \frac{1}{2\eta} \|x_{s}' - x_{s}^{(t+1)}\|^2
\]
which implies that, by taking $x'_s = x_s^{(t)}$,

$$(x_s^{(t+1)} - x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)} \geq \frac{1}{2\eta} \|x_s^{(t+1)} - \bar{x}_s^{(t)}\|^2 - \frac{1}{2\eta} \|x_s^{(t)} - \bar{x}_s^{(t)}\|^2. \quad \text{(D.11)}$$

Combining the two inequalities above yields

$$\text{Diff}_x = (x_s^{(t+1)} - x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)} \geq \frac{1}{\eta} \|x_s^{(t+1)} - \bar{x}_s^{(t+1)}\|^2 + \frac{1}{2\eta} \|x_s^{(t+1)} - \bar{x}_s^{(t)}\|^2 - \frac{1}{2\eta} \|x_s^{(t)} - \bar{x}_s^{(t)}\|^2. \quad \text{(D.12)}$$

Bounding Diff$_y$. Similarly,

$$\text{Diff}_y = (x_s^{(t)})^\top Q_s^{(t)} (y_s^{(t+1)} - y_s^{(t)}) \geq \frac{1}{\eta} \|y_s^{(t+1)} - \bar{y}_s^{(t+1)}\|^2 + \frac{1}{2\eta} \|y_s^{(t+1)} - \bar{y}_s^{(t)}\|^2 - \frac{1}{2\eta} \|y_s^{(t)} - \bar{y}_s^{(t)}\|^2. \quad \text{(D.13)}$$

Bounding Diff$_{xy}$. By the AM-GM and Cauchy-Schwarz inequalities,

$$\text{Diff}_{xy} \geq -\frac{\sqrt{A}}{2(1 - \gamma)} \|x_s^{(t+1)} - x_s^{(t)}\|^2 - \frac{\sqrt{A}}{2(1 - \gamma)} \|y_s^{(t+1)} - y_s^{(t)}\|^2$$

\(\geq\)

$$- \frac{3\sqrt{A}}{2(1 - \gamma)} \left( \|x_s^{(t+1)} - \bar{x}_s^{(t+1)}\|^2 + \|x_s^{(t+1)} - \bar{x}_s^{(t)}\|^2 + \|x_s^{(t)} - x_s^{(t)}\|^2 + \|y_s^{(t+1)} - \bar{y}_s^{(t+1)}\|^2 + \|y_s^{(t+1)} - \bar{y}_s^{(t)}\|^2 + \|y_s^{(t)} - y_s^{(t)}\|^2 \right)$$

\(\geq\)

$$- \frac{1}{16\eta} \left( \|x_s^{(t+1)} - \bar{x}_s^{(t+1)}\|^2 + \|x_s^{(t+1)} - \bar{x}_s^{(t)}\|^2 + \|x_s^{(t)} - x_s^{(t)}\|^2 + \|y_s^{(t+1)} - \bar{y}_s^{(t+1)}\|^2 + \|y_s^{(t+1)} - \bar{y}_s^{(t)}\|^2 + \|y_s^{(t)} - y_s^{(t)}\|^2 \right)$$

where (a) follows $\|x + y + z\|^2 \leq 3\|x\|^2 + 3\|y\|^2 + 3\|z\|^2$ and (b) is by $\eta \leq \frac{1 - \gamma}{32\sqrt{A}}$.

Finally, we complete the proof by summing up the bounds above for Diff$_x$, Diff$_y$, and Diff$_{xy}$.

\(\square\)

**Lemma 60** In Algorithm $\mathbb{12}$, for all $t$ and $s$, the following two inequalities hold:

(i) $\mathcal{V}_s^{(t)} \geq \mathcal{V}_s^{(t-1)}$;

(ii) $(x_s^{(t+1)})^\top Q_s^{(t+1)} y_s^{(t+1)} - (x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)} \geq \frac{15}{16\eta} \|z_s^{(t+1)} - \bar{z}_s^{(t+1)}\|^2 + \frac{7}{16\eta} \|\bar{z}_s^{(t+1)} - \bar{z}_s^{(t)}\|^2 - \frac{9}{16\eta} \|\bar{z}_s^{(t)} - \bar{z}_s^{(t)}\|^2$.
PROOF. We first note that (ii) is a consequence of \(59\) and (i),

\[
(x^{(t+1)}_s)\top Q^{(t+1)}_s y^{(t+1)}_s - (x^{(t)}_s)\top Q^{(t)}_s y^{(t)}_s \\
= (x^{(t+1)}_s)\top Q^{(t+1)}_s y^{(t+1)}_s - (x^{(t+1)}_s)\top Q^{(t)}_s y^{(t+1)}_s + (x^{(t+1)}_s)\top Q^{(t)}_s y^{(t+1)}_s - (x^{(t)}_s)\top Q^{(t)}_s y^{(t)}_s \\
\geq \min_{s'} \gamma (\psi^{(t)}_{s'} - \psi^{(t-1)}_{s'}) + \frac{15}{16\eta} \|z^{(t+1)}_s - \bar{z}^{(t+1)}_s\|^2 + \frac{7}{16\eta} \|\bar{z}^{(t+1)}_s - \bar{z}^{(t)}_s\|^2 - \frac{9}{16\eta} \|\bar{z}^{(t)}_s - \bar{z}^{(t-1)}_s\|^2 \\
\geq \frac{15}{16\eta} \|\bar{z}^{(t+1)}_s - \bar{z}^{(t+1)}_s\|^2 + \frac{7}{16\eta} \|\bar{z}^{(t+1)}_s - \bar{z}^{(t)}_s\|^2 - \frac{9}{16\eta} \|\bar{z}^{(t)}_s - \bar{z}^{(t-1)}_s\|^2
\]

where (a) is due to Lemma \(59\) and the update of \(Q^{(t)}_s\) in Algorithm \(12\)

\[
Q^{(t+1)}_s(a_1, a_2) - Q^{(t)}_s(a_1, a_2) = \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot | s, a_1, a_2)} [\psi^{(t)}_{s'} - \psi^{(t-1)}_{s'}]
\]

and (b) follows (i).

Therefore, it suffices to prove (i). We prove it by induction. Define \(\zeta_s^{(t)} := \|z^{(t)}_s - \bar{z}^{(t)}_s\|^2\) and \(\lambda_s^{(t)} := \|\bar{z}^{(t+1)}_s - \bar{z}^{(t)}_s\|^2\). For notational simplicity, define \(Q^{(0)}_s = 0_{A \times A}, z^{(0)}_s = \bar{z}^{(0)}_s = \frac{1}{4} I = z^{(1)}_s = \bar{z}^{(1)}_s\). Thus, (ii) holds for \(t = 0\) and (i) holds for \(t = 1\). We note that for \(t \geq 2\),

\[
\psi^{(t)}_s - \psi^{(t-1)}_s \\
\geq \alpha^{(t)} \left( x^{(t)}_s Q^{(t)}_s y^{(t)}_s - \psi^{(t-1)}_s \right) \\
\geq \alpha^{(t)} \left( \sum_{\tau = 0}^{t-1} \alpha^{(t-1, \tau)} \left( x^{(t)}_s Q^{(t)}_s y^{(t)}_s - x^{(\tau)}_s Q^{(\tau)}_s y^{(\tau)}_s \right) \right) \\
\geq \alpha^{(t)} \left( \sum_{\tau = 0}^{t-1} \alpha^{(t-1, \tau)} \sum_{i = \tau}^{t-1} \left( x^{(i+1)}_s Q^{(i)}_s y^{(i+1)}_s - x^{(i)}_s Q^{(i)}_s y^{(i)}_s \right) \right) \\
= \alpha^{(t)} \left( \sum_{\tau = 0}^{t-1} \alpha^{(t-1, \tau)} \sum_{i = \tau}^{t-1} \left( \alpha^{(i+1)}_s Q^{(i+1)}_s y^{(i+1)}_s - x^{(i)}_s Q^{(i)}_s y^{(i)}_s - \frac{15}{16\eta} \lambda^{(i+1)}_s - \frac{7}{16\eta} \lambda^{(i)}_s + \frac{9}{16\eta} \zeta^{(i)}_s \right) \right) \\
\geq \alpha^{(t)} \left( \sum_{\tau = 0}^{t-1} \sum_{i = \tau}^{t-1} \left( \frac{15}{16\eta} \zeta^{(i+1)}_s + \frac{7}{16\eta} \lambda^{(i)}_s - \frac{9}{16\eta} \zeta^{(i)}_s \right) \right) \\
\geq \alpha^{(t)} \sum_{i = 1}^{t} \zeta^{(i)}_s \left( \frac{15}{16\eta} \sum_{\tau = 0}^{i-1} \alpha^{(t-1, \tau)} - \frac{9}{16\eta} \sum_{\tau = 0}^{i-1} \alpha^{(t-1, \tau)} \right) \\
\geq 0
\]
where (a) follows the update of \( y_s^{(t)} \) in Algorithm 12, we apply Lemma 58 and \( \sum_{\tau=0}^{t-1} \alpha^{(t-1,\tau)} = 1 \) in (b), (c) follows the induction hypothesis (ii), (d) is due to that \( \zeta_s^{(0)} = \zeta_s^{(1)} = 0 \), and we apply Lemma 64 for (e).

**Lemma 61** For every \( s \in S \), the following quantities in Algorithm 12 all converge to some fixed values when \( t \to \infty \):

1. \( V_s^{(t)} \);
2. \( \| z_s^{(t)} - z_s^{(t)} \|^2 + \| z_s^{(t)} - z_s^{(t-1)} \|^2 \) (converges to zero);
3. \( (x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)} \).

**Proof.** Establishing (i). By (i) in Lemma 60, \( \{ V_s^{(t)} \}_{t=0}^\infty \) is a bounded increasing sequence. By the monotone convergence theorem, it is convergent. Therefore, (i) holds.

Establishing (ii). By summing up the inequality (ii) in Lemma 60 over \( t \) and using the fact that \( z_s^{(1)} = z_s^{(1)} \),

\[
\sum_{\tau=1}^{t} \left( \frac{6}{16\eta} \| z_s^{(\tau+1)} - z_s^{(\tau+1)} \|^2 + \frac{7}{16\eta} \| z_s^{(\tau+1)} - z_s^{(\tau)} \|^2 \right) \leq (x_s^{(t+1)})^\top Q_s^{(t+1)} y_s^{(t+1)} - (x_s^{(1)})^\top Q_s^{(1)} y_s^{(1)} \\
\leq \frac{1}{1-\gamma}
\]

which implies that \( \frac{6}{16\eta} \| z_s^{(\tau+1)} - z_s^{(\tau+1)} \|^2 + \frac{7}{16\eta} \| z_s^{(\tau+1)} - z_s^{(\tau)} \|^2 \) must converge to zero when \( \tau \to \infty \), which further implies (ii).

Establishing (iii). By (ii) in Lemma 60

\[
(x_s^{(t+1)})^\top Q_s^{(t+1)} y_s^{(t+1)} - \frac{15}{16\eta} \| z_s^{(t+1)} - z_s^{(t+1)} \|^2 \\
\geq \left( (x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)} - \frac{15}{16\eta} \| z_s^{(t)} - z_s^{(t)} \|^2 \right) + \frac{7}{16\eta} \| z_s^{(t+1)} - z_s^{(t)} \|^2 + \frac{6}{16\eta} \| z_s^{(t)} - z_s^{(t)} \|^2.
\]

Therefore,

\[
(x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)} - \frac{15}{16\eta} \| z_s^{(t)} - z_s^{(t)} \|^2
\]

converges to a fixed value (increasing and upper bounded). In (ii), we have shown that \( \| z_s^{(t)} - z_s^{(t)} \|^2 \) converges to zero. Therefore, \( (x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)} \) must also converge. Therefore, (iii) holds. \[\square\]

**Lemma 62** In Algorithm 12, for every \( s \in S \), \( \lim_{t \to \infty} V_s^{x(t),y(t)} \) exists, and

\[
\lim_{t \to \infty} V_s^{(t)} = \lim_{t \to \infty} V_s^{x(t),y(t)}.
\]

**Proof.** By Lemma 61, \( V_s^{(t)} \) and \( (x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)} \) both are convergent. Let \( V_s^* := \lim_{t \to \infty} V_s^{(t)} \) and \( \sigma_s^* := \lim_{t \to \infty} (x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)} \). We next show \( V_s^* = \sigma_s^* \) by contradiction. Assume that there exists
\(\epsilon > 0\) such that \(|\mathcal{V}_s^* - \sigma_s^*| = \epsilon\). Since \((x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)}\) converges to \(\sigma_s^*\), there exists some \(t_0 > 0\) such that for all \(t \geq t_0\),

\[
|x_s^{(t)}|^\top Q_s^{(t)} y_s^{(t)} - \sigma_s^*| \leq \frac{\epsilon}{3}. \tag{D.14}
\]

By our choice of \(\alpha^{(t)}\), \(\sum_{t'=t}^{\infty} \alpha^{(t')} = \infty\) for any \(t'\). Thus, there exists \(t_1 > 0\) such that for all \(t \geq t_1\) and all \(\tau \leq t_0\),

\[
\alpha^{(t,\tau)} \leq \prod_{i=\tau+1}^{t} (1 - \alpha^{(i)}) \leq \exp\left(-\sum_{i=t_0+1}^{t} \alpha^{(i)}\right) \leq \frac{\epsilon(1 - \gamma)}{3t_0} \tag{D.15}
\]

where \(\log(1 - x) \leq -x\) for \(x \in (0, 1)\) is used in \((a)\). By the update of \(\mathcal{V}_s^{(t)}\) in Algorithm 12 for all \(t \geq \max(t_0, t_1)\),

\[
|\mathcal{V}_s^{(t)} - \sigma_s^*| = \left|\sum_{\tau=0}^{t} \alpha^{(t,\tau)} \left((x_s^{(\tau)})^\top Q_s^{(\tau)} y_s^{(\tau)} - \sigma_s^*\right)\right|
\]

\[
\leq \sum_{\tau=0}^{t_0-1} \alpha^{(t,\tau)} \left((x_s^{(\tau)})^\top Q_s^{(\tau)} y_s^{(\tau)} - \sigma_s^*\right) + \left|\sum_{\tau=t_0}^{t} \alpha^{(t,\tau)} \left((x_s^{(\tau)})^\top Q_s^{(\tau)} y_s^{(\tau)} - \sigma_s^*\right)\right|
\]

\[
\leq \left(\sum_{\tau=0}^{t_0-1} \alpha^{(t,\tau)}\right) \times \frac{1}{1 - \gamma} + \left(1 - \sum_{\tau=1}^{t_0-1} \alpha^{(t,\tau)}\right) \times \frac{\epsilon}{3}
\]

\[
\leq t_0 \max_{\tau \leq t_0} \alpha^{(t,\tau)} \times \frac{1}{1 - \gamma} + \frac{\epsilon}{3}
\]

\[
\leq \frac{2\epsilon}{3}
\]

where we apply the triangle inequality for \((a)\), \((b)\) is due to \(\text{(D.14)}\) and \(\sum_{\tau=1}^{t} \alpha^{(t,\tau)} = 1\), and \((c)\) follows \(\text{(D.15)}\). Since \(|\mathcal{V}_s^* - \sigma_s^*| = \epsilon\), it is impossible that \(\mathcal{V}_s^{(t)}\) converges to \(\mathcal{V}_s^*\), and it must be that \(\mathcal{V}_s^* = \sigma_s^*\). Therefore, \(\mathcal{V}_s^{(t)} - (x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)}\) converges to zero as \(t \to \infty\).

Equivalently, \(\mathcal{V}_s^{(t)} - (x_s^{(t)})^\top Q_s^{(t)} y_s^{(t)}\) can be expressed as

\[
(\mathcal{V}_s^{(t)} - \mathcal{V}_s^{(t-1)}) + \mathcal{V}_s^{(t-1)} - \sum_{a_1,a_2} x_s^{(t)}(a_1)y_s^{(t)}(a_2) \left(r(s,a_1,a_2) + \gamma E_{s' \sim \mathbb{P}(.|s,a_1,a_2)} \left[\mathcal{V}_s^{(t-1)}\right]\right).
\]

By letting \(t \to 0\), since \(\mathcal{V}_s^{(t)} - \mathcal{V}_s^{(t-1)} \to 0\), thus,

\[
\mathcal{V}_s^{(t-1)} - \sum_{a_1,a_2} x_s^{(t)}(a_1)y_s^{(t)}(a_2) \left(r(s,a_1,a_2) + \gamma E_{s' \sim \mathbb{P}(.|s,a_1,a_2)} \left[\mathcal{V}_s^{(t-1)}\right]\right).
\]

also converges to zero. Hence, \(\mathcal{V}_s^{(t)}\) converges to the unique fixed point of the Bellman equation. By the uniqueness, \(\mathcal{V}_s^{(t-1)} - V_{x^{(t)},y^{(t)}}\) converges to zero. Therefore, \(\lim_{t \to \infty} V_{x^{(t)},y^{(t)}} = \lim_{t \to \infty} \mathcal{V}_s^{(t-1)} = \mathcal{V}_s^*\). \(\square\)
Lemma 63. In Algorithm 12 for every $s$,
\[
\lim_{t \to \infty} \max_{x'} \left( x'_{s} - x_{s}^{(t)} \right)^{\top} Q_{s}^{(t)} y_{s}^{(t)} = 0.
\]

**Proof.** By the optimality of $x_{s}^{(t+1)}$,
\[
\left\langle x'_{s} - x_{s}^{(t+1)} , \eta Q_{s}^{(t)} y_{s}^{(t)} - x_{s}^{(t+1)} + \bar{x}_{s}^{(t+1)} \right\rangle \leq 0, \text{ for any } x'.
\]
Rearranging the inequality yields, for any $x_{s}'$,
\[
\left\langle x'_{s} - x_{s}^{(t+1)} , Q_{s}^{(t)} y_{s}^{(t)} \right\rangle \leq \frac{1}{\eta} \left\langle x'_{s} - x_{s}^{(t+1)} , x_{s}^{(t+1)} - x_{s}^{(t)} \right\rangle
\]
\[
+ \frac{1}{\eta} \left\langle x_{s}^{(t+1)} - x_{s}^{(t)} , \eta Q_{s}^{(t)} y_{s}^{(t)} - x_{s}^{(t+1)} + \bar{x}_{s}^{(t+1)} \right\rangle
\]
\[
\leq \frac{1}{\eta} \left\| x_{s}^{(t+1)} - x_{s}^{(t)} \right\|
\]
\[
\leq \frac{1}{\eta} \left( \left\| x_{s}^{(t+1)} - \bar{x}_{s}^{(t+1)} \right\| + \left\| \bar{x}_{s}^{(t+1)} - x_{s}^{(t)} \right\| + \left\| x_{s}^{(t)} - x_{s}^{(t)} \right\| \right).
\]

By (ii) of Lemma 61 the right-hand side above converges to zero, which completes the proof. \(\square\)

Lemma 64. Let $\{\alpha^{(t)}\}_{t=1}^{\infty}$ be a non-increasing sequence that satisfies $0 < \alpha^{(t)} \leq \frac{1}{6}$ for all $t$. Then for any $t \geq i \geq 2$,
\[
\sum_{\tau=0}^{i} \alpha^{(t,\tau)} \leq \frac{5}{3} \sum_{\tau=0}^{i-1} \alpha^{(t,\tau)}.
\]

**Proof.** Equivalently, we prove
\[
\alpha^{(t,i)} \leq \frac{2}{3} \sum_{\tau=0}^{i-1} \alpha^{(t,\tau)}.
\]
If suffices to show that $\alpha^{(t,i)} \leq \frac{2}{3} \alpha^{(t,i-1)} + \frac{2}{3} \alpha^{(t,i-2)}$. We have the following two cases. 

**Case 1:** $i > 2$. By the definition of $\alpha^{(t,\tau)}$ and the monotonicity of $0 < \alpha^{(t)} \leq \frac{1}{6}$,
\[
\frac{\alpha^{(t,i)}}{\alpha^{(t,i-1)}} = \frac{\alpha^{(i)}}{\prod_{j=i+1}^{t} (1 - \alpha^{(j)})} = \frac{\alpha^{(i)}}{\prod_{j=i}^{t-1} (1 - \alpha^{(j)})} \leq \frac{1}{1 - \alpha^{(i)}} \leq \frac{1}{1 - \frac{1}{6}} = \frac{6}{5}
\]
\[
\frac{\alpha^{(t,i)}}{\alpha^{(t,i-2)}} = \frac{\alpha^{(i)}}{(1 - \alpha^{(i)}) (1 - \alpha^{(i-1)})} \leq \frac{36}{25}.
\]
Therefore,
\[
\frac{2}{3} \alpha^{(t,i-1)} + \frac{2}{3} \alpha^{(t,i-2)} \geq \frac{2}{3} \left( \frac{5}{6} + \frac{25}{36} \right) \alpha^{(t,i)} \geq \alpha^{(t,i)}.
\]

289
Case 2: $i = 2$. By the definition of $\alpha^{(t,\tau)}$ and the monotonicity of $0 < \alpha^{(t)} \leq \frac{1}{6}$,

\[
\frac{\alpha^{(t,2)}}{\alpha^{(t,0)}} = \frac{\alpha^{(2)} \prod_{j=3}^{t} (1 - \alpha^{(j)})}{\prod_{j=1}^{t} (1 - \alpha^{(j)})} = \frac{\alpha^{(2)}}{(1 - \alpha^{(1)})(1 - \alpha^{(2)})} \leq \frac{1}{6} \times \frac{5}{6} = \frac{5}{25}.
\]

Therefore,

\[
\frac{2}{3} \alpha^{(t,1)} + \frac{2}{3} \alpha^{(t,0)} \geq \frac{2}{3} \times \frac{25}{6} \alpha^{(t,2)} \geq \alpha^{(t,2)}.
\]

**Proof.** [Proof of Theorem 42]

\[
\begin{align*}
\max_{x'} & \left( V^{x',y'}(\rho) - V^{x^{(t)},y^{(t)}}(\rho) \right) \\
= & \max_{x'} \frac{1}{1 - \gamma} \sum_s d_{x'}^{x^{(t)},y^{(t)}}(s) \left( x'_s - x_s \right)^\top Q^{x^{(t)},y^{(t)}} y_s^{(t)} \\
\leq & \max_{x'} \frac{1}{1 - \gamma} \sum_s d_{x'}^{x^{(t)},y^{(t)}}(s) \left( x'_s - x_s \right)^\top Q y_s^{(t)} \\
+ & \max_{x'} \frac{1}{1 - \gamma} \sum_s d_{x'}^{x^{(t)},y^{(t)}}(s) \left( x'_s - x_s \right)^\top \left( Q^{x^{(t)},y^{(t)}} - Q^{x^{(t)},y^{(t)}} \right) y_s^{(t)}.
\end{align*}
\]

By Lemma 63 $\text{Diff}_P \to 0$ when $t \to \infty$. For $\text{Diff}_Q$, we notice that

\[
\left| Q^{x^{(t)},y^{(t)}} - Q^{x^{(t)},y^{(t)}} \right| \leq \gamma \max_{s'} \left| V^{x^{(t)},y^{(t)}} - V^{x^{(t)},y^{(t)}} \right|
\]

which converges to zero by Lemma 62. Therefore, $\text{Diff}_Q \to 0$ when $t \to \infty$. Therefore, $(x^{(t)}, y^{(t)})$ converges to a Nash equilibrium when $t \to \infty$. □

**D.6.2 Proof of Theorem 43**

We first introduce a corollary of Lemma 60.

**Corollary 3** In Algorithm 12, for every state $s$, and any $T > 0$,

\[
\sum_{t=1}^{T} \left( \| \bar{z}_s^{(t+1)} - \bar{z}_s^{(t)} \|^2 + \| \bar{z}_s^{(t+1)} - z_s^{(t)} \|^2 \right) \leq \frac{8\eta}{1 - \gamma}
\]

where $z_s^{(t)} = (x_s^{(t)}, y_s^{(t)})$ and $\bar{z}_s^{(t)} = (\bar{x}_s^{(t)}, \bar{y}_s^{(t)})$. 

290
Theorem 60

By (ii) of Lemma 60,

\[(x^{(t+1)}_s)^\top Q^{(t+1)}_s y^{(t+1)}_s - (x^{(t)}_s)^\top Q^{(t)}_s y^{(t)}_s + \frac{15}{16\eta} \left( \|z^{(t)}_s - \bar{z}^{(t)}_s\|^2 - \|z^{(t+1)}_s - \bar{z}^{(t+1)}_s\|^2 \right) \geq \frac{7}{16\eta} \|\bar{z}^{(t+1)}_s - \bar{z}^{(t)}_s\|^2 + \frac{6}{16\eta} \|z^{(t)}_s - \bar{z}^{(t)}_s\|^2 \]
\[\geq \frac{6}{16\eta} \|\bar{z}^{(t+1)}_s - \bar{z}^{(t)}_s\|^2 + \frac{6}{16\eta} \|z^{(t)}_s - \bar{z}^{(t)}_s\|^2.\]

Thus, by the inequality \(\|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2\),

\[\|\bar{z}^{(t+1)}_s - \bar{z}^{(t)}_s\|^2 + \|\bar{z}^{(t+1)}_s - \bar{z}^{(t)}_s\|^2 \leq 3\|\bar{z}^{(t+1)}_s - \bar{z}^{(t)}_s\|^2 + 2\|z^{(t)}_s - \bar{z}^{(t)}_s\|^2 \leq 8\eta \left( (x^{(t+1)}_s)^\top Q^{(t+1)}_s y^{(t+1)}_s - (x^{(t)}_s)^\top Q^{(t)}_s y^{(t)}_s \right) \]
\[+ \frac{15}{2} \left( \|z^{(t)}_s - \bar{z}^{(t)}_s\|^2 - \|z^{(t+1)}_s - \bar{z}^{(t+1)}_s\|^2 \right)\]

which yields our desired result if we sum it over \(t\), use \((x^{(T+1)}_s)^\top Q^{(T+1)}_s y^{(T+1)}_s \leq \frac{1}{1-\gamma}\) and \(z^{(1)}_s = \bar{z}^{(1)}_s\), and ignore a negative term. \(\square\)

Lemma 65

In Algorithm 12, the gap between the critic \(Q^{(t)}_s\) and the true \(Q^{(t)}_s\) satisfies

\[\sum_{t=1}^{T} \max_s \|Q^{(t)}_s - Q^{(t)}_s\|_\infty \leq \frac{A}{(\alpha(T))^{2}(1-\gamma)^6} \sum_{t=1}^{T} \max_s \left( \|x^{(t)}_s - x^{(t-1)}_s\|^2 + \|y^{(t)}_s - y^{(t-1)}_s\|^2 \right).\]

Proof. For notational simplicity, define \(Q^{(0)}_s = Q^{(0)}_s = 0_{A \times A}\).
\[
\max_s \left\| Q_s^{(t)} - Q_s^{(t)} \right\|_\infty^2 \\
\leq \max_{s,a_1,a_2} \left| Q_s^{(t)}(a_1, a_2) - Q_s^{(t)}(a_1, a_2) \right|^2 \\
\leq \max_{s,a_1,a_2} \left[ r(s, a_1, a_2) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot | s, a_1, a_2)} \left[ (x_{s'})^\top Q_s^{(t)} y_{s'} \right] \\
- \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} \left( r(s, a_1, a_2) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot | s, a_1, a_2)} \left[ (x_{s'}^{(\tau)})^\top Q_s^{(\tau)} y_{s'}^{(\tau)} \right] \right) \right]^2 \\
\leq \gamma^2 \max_{s} \left\| \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} \left( (x_{s'}^{(\tau)})^\top Q_s^{(\tau)} y_{s'}^{(\tau)} - (x_{s'}^{(\tau)})^\top Q_s^{(\tau)} y_{s'}^{(\tau)} \right) \right\|^2 \\
\leq \frac{6\gamma^2}{1 - \gamma} \max_{s} \left( \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} (x_s^{(t)})^\top (Q_s^{(t)} - Q_s^{(\tau)}) y_s^{(t)} \right)^2 \\
+ \frac{2\gamma^2}{1 + \gamma} \max_{s} \left( \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} (x_s^{(t)})^\top (Q_s^{(\tau)} - Q_s^{(\tau)}) y_s^{(t)} \right)^2 \\
+ \frac{6\gamma^2}{1 - \gamma} \max_{s} \left( \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} (x_s^{(t)})^\top Q_s^{(\tau)} (y_s^{(t)} - y_s^{(\tau)}) \right)^2 \\
+ \frac{6\gamma^2}{1 - \gamma} \max_{s} \left( \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} (x_s^{(t)} - x_s^{(\tau)})^\top Q_s^{(\tau)} y_s^{(\tau)} \right)^2 \\
\leq \frac{2\gamma^2}{1 + \gamma} \max_{s} \left( \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} \left( \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} (x_s^{(\tau)})^\top (Q_s^{(\tau)} - Q_s^{(\tau)}) y_s^{(t)} \right) \right)^2 \\
+ c' \max_{s} \left( \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} \left( \| x_s^{(t)} - x_s^{(\tau)} \|_1 + \| y_s^{(t)} - y_s^{(\tau)} \|_1 \right) \right)^2 \\
\leq \frac{2\gamma^2}{1 + \gamma} \max_{s} \left[ \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} \| Q_s^{(\tau)} - Q_s^{(\tau)} \|_2^2 \right] + c' \max_{s} \left( \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} \sum_{h = \tau + 1}^{t} \text{diff}^{(h)}_s \right)^2 \\
\leq \gamma \max_{s} \left[ \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} \| Q_s^{(\tau)} - Q_s^{(\tau)} \|_2^2 \right] + c' \max_{s} \left( \sum_{h = 1}^{t} \sum_{\tau = 0}^{h-1} \alpha^{(t-1,\tau)} \sum_{h = \tau + 1}^{t} \text{diff}^{(h)}_s \right)^2 \\
\leq \gamma \max_{s} \left[ \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} \| Q_s^{(\tau)} - Q_s^{(\tau)} \|_2^2 \right] + c' \max_{s} \left( \sum_{h = 1}^{t} \delta^{(t-1,h-1)} \text{diff}^{(h)}_s \right)^2 \\
\leq \gamma \max_{s} \left[ \sum_{\tau = 0}^{t-1} \alpha^{(t-1,\tau)} \| Q_s^{(\tau)} - Q_s^{(\tau)} \|_2^2 \right] + c' \max_{s} \left( \sum_{h = 1}^{t} (1 - \alpha^{(t)})^{t-h} \text{diff}^{(h)}_s \right)^2 \\
\]

where in (a) we apply \((x + y + z + w)^2 \leq \frac{6x^2}{1 - \gamma} + \frac{2y^2}{1 + \gamma} + \frac{6z^2}{1 - \gamma} + \frac{6w^2}{1 - \gamma}\) from the Cauchy-Schwarz inequality, in (b) we use Lemma 67 and obtain \(c' = O \left( \frac{1}{(1 - \gamma)^t} \right)\), in (c) we introduce notation,

\[
\text{diff}^{(h)}_s := \| x_s^{(h)} - x_s^{(h-1)} \|_1 + \| y_s^{(h)} - y_s^{(h-1)} \|_1
\]

292
and in (d) we introduce notation,

$$\delta^{(t,\tau)} = \prod_{i=\tau+1}^{t} (1 - \alpha^{(i)}).$$

and apply Lemma 35 of [239], (e) is due to that \{\alpha^{(t)}\}_{t=0}^{\infty} is a non-increasing sequence. Application of Lemma 33 of [239] to the recursion relation above yields

$$\max_{s} \|Q_{s}^{(t)} - Q_{s}^{(t)}\|_{2}^{2} \leq c^{\prime} \sum_{\tau=1}^{t} \beta^{(t,\tau)} \max_{s} \left( \sum_{q=1}^{\tau} (1 - \alpha^{(\tau)} \tau - q \text{diff}_{s}) \right)^{2}$$

where $$\beta^{(t,\tau)} := \alpha^{(\tau)} \prod_{i=\tau}^{t-1} (1 - \alpha^{(i)} + \alpha^{(i)} \gamma)$$ for \(1 \leq \tau < t\), and \(\beta^{(t,t)} := 1\), and \(\text{diff}^{(t)} := \max_{s} \text{diff}_{s}^{(t)}\). The right-hand side of (D.16) can be further upper bounded by

$$c^{\prime} \sum_{\tau=1}^{t} \beta^{(t,\tau)} \left( \sum_{q=1}^{\tau} (1 - \alpha^{(\tau)} \tau - q \text{diff}^{(q)}) \right)^{2}$$

(a) is due to that \(\sum_{q=1}^{\tau} (1 - \alpha^{(\tau)} \tau - q \text{diff}^{(q)})^2 \leq \frac{1}{\alpha^{(\tau)}}\).

$$c^{\prime} \sum_{\tau=1}^{t} \beta^{(t,\tau)} \left( \sum_{q=1}^{\tau} (1 - \alpha^{(\tau)} \tau - q \text{diff}^{(q)}) \right)^{2}$$

$$\leq c^{\prime} \sum_{\tau=1}^{t} \frac{\beta^{(t,\tau)}}{\alpha^{(\tau)}} \sum_{q=1}^{\tau} (1 - \alpha^{(\tau)} \tau - q \text{diff}^{(q)})^2$$

$$= c^{\prime} \sum_{q=1}^{t} \sum_{\tau=q}^{t} \beta^{(t,\tau)} \left( 1 - \alpha^{(\tau)} \tau - q \text{diff}^{(q)})^2 \right)$$

$$= c^{\prime} \sum_{q=1}^{t} \left[ \sum_{\tau=q}^{t-1} (1 - \alpha^{(\tau)} + \alpha^{(\tau)} \gamma) \text{diff}^{(q)}(1 - \alpha^{(\tau)} \tau - q \text{diff}^{(q)})^2 + \frac{(1 - \alpha^{(\tau)} \tau - q \text{diff}^{(q)})^2}{\alpha^{(\tau)}} \right]$$

$$= c^{\prime} \sum_{q=1}^{t} \left[ (1 - \alpha^{(t)} \gamma) \text{diff}^{(q)}(1 - \alpha^{(t)} \tau - q \text{diff}^{(q)})^2 + \frac{(1 - \alpha^{(\tau)} \tau - q \text{diff}^{(q)})^2}{\alpha^{(\tau)}} \right]$$

$$\leq \frac{2c^{\prime}}{\alpha^{(t)} \gamma} \sum_{q=1}^{t} (1 - \alpha^{(t)} + \alpha^{(t)} \gamma) \text{diff}^{(q)}(1 - \alpha^{(t)} \tau - q \text{diff}^{(q)})^2$$

where (a) is due to that \(\sum_{q=1}^{\tau} (1 - \alpha^{(\tau)} \tau - q \text{diff}^{(q)})^2 \leq \frac{1}{\alpha^{(\tau)}}\).
Substitution of the upper bound above into (D.16) yields,

\[
\sum_{t=1}^{T} \max_s \|Q^{(t)}_s - Q^{(t)}_s\|_2 \lesssim \sum_{t=1}^{T} \frac{c'}{\alpha(t)} \sum_{q=1}^{t} (1 - \alpha^{(t)} + \alpha^{(t)} \gamma)^{t-q} (\text{diff}(q))^2
\]

\[
\leq \sum_{q=1}^{T} \frac{c'}{\alpha(T)} (1 - \alpha^{(T)} + \alpha^{(T)} \gamma)^{T-q} (\text{diff}(q))^2
\]

\[
\leq \left( \frac{c'}{(\alpha(T))^2(1 - \gamma)} \right) \sum_{q=1}^{T} (\text{diff}(q))^2
\]

where \(a\) is due to that \(\alpha^{(t)}\) is non-increasing, and \(b\) is due to that \(\sum_{t=q}^{T} (1 - \alpha^{(T)} + \alpha^{(T)} \gamma)^{T-q} \leq \frac{1}{\alpha^{(T)}(1 - \gamma)}\).

Finally, using the definition of \(c'\) and applying \(\| \cdot \|_1 \leq \sqrt{A} \| \cdot \|\) to \(\text{diff}(q)\) lead to the desired result. \(\square\)

**Proof.** [Proof of Theorem 43]

The proof consists of two parts: Markov cooperative games and Markov competitive games, separately.

**Markov cooperative games.** Fix \(s\), the optimality of \(\bar{x}_s^{(t+1)}\) in Algorithm 12 yields

\[
\left\langle \eta Q^{(t)}_s y^{(t)}_s - (\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}), x'_s - \bar{x}_s^{(t+1)} \right\rangle \leq 0, \quad \text{for any } x'_s \in \Delta(A_1). \tag{D.17}
\]

Thus, for any \(x'_s \in \Delta(A_1),\)

\[
(x'_s - x_s^{(t)})^\top Q^{(t)}_s y^{(t)}_s = (x'_s - \bar{x}_s^{(t+1)})^\top Q^{(t)}_s y^{(t)}_s + (x'_s - \bar{x}_s^{(t+1)})^\top (Q^{(t)}_s - Q^{(t)}_s) y^{(t)}_s + (\bar{x}_s^{(t+1)} - x_s^{(t)})^\top Q^{(t)}_s y^{(t)}_s
\]

\[
\leq \left( \frac{1}{\eta} \right) (x'_s - \bar{x}_s^{(t+1)})^\top (\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}) + \|Q^{(t)}_s - Q^{(t)}_s\| + \frac{\sqrt{A}}{1 - \gamma} \|\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}\|
\]

\[
\leq \left( \frac{1}{\eta} \right) (\|\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}\| + \|\bar{x}_s^{(t+1)} - \bar{x}_s^{(t)}\|) + \|Q^{(t)}_s - Q^{(t)}_s\|
\]

\[\leq \left( \frac{c'}{(\alpha(T))^2(1 - \gamma)} \right) \sum_{q=1}^{T} (\text{diff}(q))^2
\]
where we use (D.17) and $\|Q_s^{(t)}\| \leq \frac{\sqrt{A}}{1 - \gamma}$ in (a), and (b) is due to the Cauchy-Schwarz inequality and the choice of $\eta \leq \frac{1 - \gamma}{32\sqrt{A}}$. Hence,

$$
\sum_{t=1}^{T} \left( \max_{x'} V^{x', y^{(t)}}(\rho) - V^{x^{(t)}, y^{(t)}}(\rho) \right)
$$

$$
= \frac{1}{1 - \gamma} \sum_{t=1}^{T} \max_{x'} \sum_{s} d^{x', y^{(t)}}(s) (x'_s - x^{(t)}_s) \top Q_s^{(t)} y_s
$$

$$
\leq \frac{1}{\eta(1 - \gamma)} \sum_{t=1}^{T} \sum_{s} d^{x', y^{(t)}}(s) (\|\bar{x}^{(t+1)}_s - \bar{x}^{(t)}_s\| + \|\bar{x}^{(t+1)}_s - x^{(t)}_s\|)
$$

$$
+ \frac{1}{(1 - \gamma)} \sum_{t=1}^{T} \sum_{s} d^{x', y^{(t)}}(s) \|Q_s^{(t)} - Q_s^{(t)}\|
$$

$$
\leq \frac{1}{\eta(1 - \gamma)} \sqrt{\sum_{t=1}^{T} \sum_{s} d^{x', y^{(t)}}(s) \left( \|\bar{x}^{(t+1)}_s - \bar{x}^{(t)}_s\|^2 + \|\bar{x}^{(t+1)}_s - x^{(t)}_s\|^2 \right)}
$$

$$
+ \frac{1}{(1 - \gamma)} \sum_{t=1}^{T} \sum_{s} d^{x', y^{(t)}}(s) \|Q_s^{(t)} - Q_s^{(t)}\|
$$

$$
\leq \frac{\sqrt{T}}{\eta(1 - \gamma)} \sqrt{\frac{\eta S}{1 - \gamma} + \frac{1}{1 - \gamma}} \sqrt{\sum_{t=1}^{T} \sum_{s} d^{x', y^{(t)}}(s) \left( \sum_{t=1}^{T} \sum_{s} \|Q_s^{(t)} - Q_s^{(t)}\|^2 \right)}
$$

$$
\leq \frac{\sqrt{T}}{\eta(1 - \gamma)^3} + \frac{\sqrt{T}}{(1 - \gamma)^4} \sqrt{\frac{SA}{A(T)^2} \sum_{t=1}^{T} \max_s \left( \|x^{(t)}_s - x^{(t-1)}_s\|^2 + \|y^{(t)}_s - y^{(t-1)}_s\|^2 \right)}
$$

$$
\leq \frac{\sqrt{T}}{\eta(1 - \gamma)^3} + \frac{\sqrt{T}}{1 - \gamma} \sqrt{\frac{\eta S^2 A}{(A(T))^2 (1 - \gamma)^7}}
$$
where (a) is due to the Cauchy-Schwarz inequality for (a), (b) follows the state distribution $d_{\rho}^{x',y'(t)}(s)$, (c) is due to Corollary 3 (d) is because of Lemma 65 and (e) again is due to Corollary 3.

By taking $\eta = \frac{(1-\gamma)^2}{32\sqrt{2A}}$ and $\alpha(t) = \frac{1}{6\sqrt{t}}$, the last upper bound above is of order,

$$O \left( \frac{(S^3A)^{\frac{1}{2}}T}{(1-\gamma)^2 \alpha(T)} \right) = O \left( \frac{(S^3A)^{\frac{1}{2}}T^\frac{3}{2}}{(1-\gamma)^2} \right).$$

**Markov competitive game.** We start from an intermediate step in the proof of Theorem 1 of [239]. Specifically, they have shown that if both players use Algorithm 12 in a two-player zero-sum Markov game, then,

$$\sum_{t=1}^{T} \left( \max_{x',y'} V^{x',y'(t)}(\rho) - V^{x(t),y'(t)}(\rho) \right) = O \left( \frac{S\sqrt{C_{a}C_{\beta}T}}{\eta(1-\gamma)} \right)$$

where $C_{a} := 1 + \sum_{t=1}^{T} \alpha(t)$ and $C_{\beta}$ is an upper bound for $\sum_{t=\tau}^{T} \beta(t,\tau)$ with $\beta(t,\tau) := \alpha(t) \prod_{i=\tau}^{t-1} (1 - \alpha(t) + \alpha(t)\gamma)$ if $\tau < t$ and $\beta(t,t) := 1$. We next calculate the upper bounds for $C_{a}$ and $C_{\beta}$.

**Bounding $C_{a}$.** Recall that $\alpha(t) = \frac{1}{6} t^{-\frac{1}{3}}$. By the definition of $C_{a}$,

$$C_{a} = 1 + \frac{1}{6} \sum_{t=1}^{T} t^{-\frac{1}{3}} = O \left( T^{\frac{2}{3}} \right).$$

**Bounding $C_{\beta}$.** Using $\alpha(t) = \frac{1}{6} t^{-\frac{1}{3}}$, for any $\tau \geq 1$, we have

$$\sum_{t=\tau}^{T} \beta(t,\tau) \leq 1 + \sum_{t=\tau+1}^{T} \alpha(t) \prod_{i=\tau}^{t-1} (1 - \alpha(t) + \alpha(t)\gamma)$$

$$= 1 + \frac{1}{6} \sum_{t=\tau+1}^{T} \tau^{-\frac{1}{3}} \left(1 - \frac{1}{6} t^{-\frac{1}{3}} (1 - \gamma)\right)^{t-\tau}$$

$$\leq 1 + \frac{1}{6} \sum_{t=\tau+1}^{t_{0}} \tau^{-\frac{1}{3}} + \frac{1}{6} \sum_{t=t_{0}+1}^{T} \tau^{-\frac{1}{3}} \left(1 - \frac{1}{6} t^{-\frac{1}{3}} (1 - \gamma)\right)^{t-\tau}$$

(for some $t_{0}$ defined below)

$$\leq 1 + \frac{1}{6} \tau^{-\frac{1}{3}} (t_{0} - \tau) + \frac{1}{6} \tau^{-\frac{1}{3}} \sum_{t=t_{0}+1}^{T} \exp \left( -\frac{1}{6} t^{-\frac{1}{3}} (1 - \gamma) (t - \tau) \right). \quad (\text{D.18})$$

Define $t_{0} := \tau + H(\tau + c)^{\frac{1}{3}} \ln(\tau + c) + c$, where $H := \frac{48}{1 - \gamma}$ and $c := 2 \left( \frac{2H}{1 - \gamma} \ln \frac{H}{1 - \gamma} \right)^{\frac{1}{3}}$ (if $t_{0} > T$, we simply ignore the second term in (D.18)). By Lemma 66 with $q = \frac{1}{3}$, for all $t \geq t_{0}$,

$$t - \tau \geq \frac{H}{2} \left( \frac{t}{2} \right)^{\frac{1}{3}} \ln \left( \frac{t}{2} \right).$$

296
Hence, we can continue to bound the right-hand side of (D.18) by

\[ O\left(H \left(\frac{\tau + c}{\tau}\right)^{\frac{1}{3}} \ln(\tau + c) + \frac{c}{\tau^{\frac{1}{3}}} + \frac{1}{6} \tau^{-\frac{1}{3}} \sum_{t = t_0 + 1}^{T} \exp \left(-\frac{1}{12} t^{-\frac{1}{3}} (1 - \gamma) \frac{H}{2} \left(\frac{t}{2}\right)^{\frac{1}{3}} \ln \left(\frac{t}{2}\right)\right) \right) \]

\[ \leq \tilde{O}(H(1 + c)^{\frac{1}{3}} + c) + \frac{1}{6} \tau^{-\frac{1}{3}} \sum_{t = t_0 + 1}^{T} \frac{2}{t} \]

\[ = \tilde{O} \left(\frac{1}{(1 - \gamma)^{\frac{3}{2}}}\right) \]

which proves that \( C_\beta = \tilde{O} \left(\frac{1}{(1 - \gamma)^{\frac{3}{2}}}\right) \).

Therefore,

\[ \sum_{t = 1}^{T} \left(\max_{x', y'} V^{x', y'}(\rho) - V^{x(t), y'}(\rho)\right) = O \left(\frac{S \sqrt{C_\alpha C_\beta T}}{\eta(1 - \gamma)}\right) = \tilde{O} \left(\frac{ST^{\frac{5}{6}}}{\eta(1 - \gamma)^{\frac{7}{4}}}\right) \]

which completes the proof by taking \( \eta = \frac{(1 - \gamma)^2}{32\sqrt{SA}} \). \( \square \)

**Lemma 66** Fix \( \tau \in \mathbb{N}, 0 < q < 1, H \geq 1 \). Let

\[ t_0 := \tau + H(\tau + c)^{q} \ln(\tau + c) + c \]

where \( c := 2 \left(\frac{2H}{1-q} \ln \left(\frac{H}{1-q}\right)\right)^{\frac{1}{1-q}} \). Then for all \( t \geq t_0 \),

\[ t - H \left(\frac{t}{2}\right)^{q} \ln \left(\frac{t}{2}\right) \geq \tau. \]

**Proof.** We first show that for all \( t \geq c \),

- \( H t^q \ln t \leq t; \)
- \( t - H \left(\frac{t}{2}\right)^{q} \ln \left(\frac{t}{2}\right) \) is non-decreasing.

To show the two items above, we apply Lemma A.1 of [197] which states that \( x \geq 2a \ln(a) \Rightarrow x \geq a \ln(x) \) for any \( a > 0 \). By the definition of \( c \), for all \( t \geq c \), \( t^{1-q} \geq \left(\frac{t}{2}\right)^{1-q} \geq \frac{2H}{1-q} \ln \frac{H}{1-q} \) and thus

\[ t^{1-q} \geq \frac{H}{1-q} \ln(t^{1-q}) = H \ln t \]

which proves the first item, and that

\[ \left(\frac{t}{2}\right)^{1-q} \geq \frac{H}{1-q} \ln \left(\frac{t}{2}\right)^{1-q} = H \ln \frac{t}{2} \]
which gives
\[
\frac{d}{dt} \left( t - \frac{H}{2} \left( \frac{t}{2} \right)^q \ln \left( \frac{t}{2} \right) \right) = 1 - \frac{H}{2} \cdot \frac{q}{2} \left( \frac{t}{2} \right)^{q-1} \ln \left( \frac{t}{2} \right) - \frac{H}{2} \cdot \left( \frac{t}{2} \right)^q \frac{1}{t} \\
\geq 1 - \frac{1}{2} \cdot \frac{q}{2} - \frac{1}{2} \geq 0
\]
which proves the second item.
By the first item and the definition of \( t_0 \), \( t_0 \leq (\tau + (\tau + c) + c = 2\tau + 2c \). Then by the second item, for all \( t \geq t_0 \) we have
\[
t - \frac{H}{2} \left( \frac{t}{2} \right)^q \ln \left( \frac{t}{2} \right) \geq t_0 - \frac{H}{2} \left( \frac{t_0}{2} \right)^q \ln \left( \frac{t_0}{2} \right) \geq t_0 - \frac{H}{2} (\tau + c)^q \ln (\tau + c) \geq \tau
\]
which completes the proof.

Lemma 67  For any two policies \((x', y')\) and \((x, y)\),
\[
\max_s \|Q_{s}^{x', y'} - Q_{s}^{x, y}\|_{\text{max}} \leq \frac{\gamma}{(1 - \gamma)^2} \max_{s'} \left( \|x_{s'}' - x_{s'}\|_1 + \|y_{s'}' - y_{s'}\|_1 \right).
\]

PROOF. By the definition,
\[
\|Q_{s}^{x', y'} - Q_{s}^{x, y}\|_{\text{max}} \overset{(a)}{=} \max_{a_1, a_2} |Q_{s}^{x', y'}(a_1, a_2) - Q_{s}^{x, y}(a_1, a_2)| \\
\leq \gamma \sum_{s'} \mathbb{P}(s' \mid s, a_1, a_2) \left| (x_{s'}')^T Q_{s'}^{x', y'} y_{s'}' - (x_{s'})^T Q_{s'}^{x, y} y_{s'} \right| \\
\leq \gamma \max_{s'} \left( (x_{s'})^T Q_{s'}^{x', y'} y_{s'}' - (x_{s'})^T Q_{s'}^{x, y} y_{s'} \right) \overset{\text{Qiff}}{\leq} \gamma \max_{s'} \left( (x_{s'})^T Q_{s'}^{x', y'} y_{s'}' - (x_{s'})^T Q_{s'}^{x, y} y_{s'} \right)
\]
where \(a_1\) and \(a_2\) achieve the maximum in \((a)\), and \((b)\) is due to the Bellman equation,
\[
Q_{s}^{x, y}(a_1, a_2) = r(s, a_1, a_2) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot \mid s, a_1, a_2)} [V_{s'}^{x, y}]
\]
\[
= r(s, a_1, a_2) + \gamma \sum_{a_1', a_2'} \mathbb{P}(s' \mid s, a_1, a_2) \sum_{a_1', a_2'} x_{s'}(a_1') y_{s'}(a_2') Q_{s'}^{x', y'}(a_1', a_2').
\]
Fix \(s'\), we next subtract and add \((x_{s'})^T Q_{s'}^{x', y'} y_{s'}\) in Qiff and apply \(|a + b| \leq |a| + |b|\) to reach,
\[
\text{Qiff} \leq \left| (x_{s'})^T Q_{s'}^{x', y'} y_{s'}' - (x_{s'})^T Q_{s'}^{x, y} y_{s'} \right| + \left| (x_{s'})^T Q_{s'}^{x', y'} y_{s'}' - (x_{s'})^T Q_{s'}^{x, y} y_{s'} \right| \\
\leq \frac{1}{1 - \gamma} \left| \sum_{a_1', a_2'} (x_{s'}(a_1') y_{s'}(a_2') - x_{s'}(a_1') y_{s'}(a_2')) Q_{s'}^{x', y'}(a_1', a_2') \right|
\]
\[
+ \left| (x_{s'})^T \left( Q_{s'}^{x', y'} - Q_{s'}^{x, y} \right) y_{s'} \right|
\leq \frac{1}{1 - \gamma} \sum_{a_1', a_2'} |x_{s'}(a_1') y_{s'}(a_2') - x_{s'}(a_1') y_{s'}(a_2')| + \left| (x_{s'})^T \left( Q_{s'}^{x', y'} - Q_{s'}^{x, y} \right) y_{s'} \right|.
\]
We also notice that

\[ \| x'_s \circ y'_s - x'_s \circ y'_s \|_1 := \sum_{a'_1, a'_2} | x'_s(a'_1) y'_s(a'_2) - x'_s(a'_1) y'_s(a'_2) | \]

\[ \leq \sum_{a'_1, a'_2} | x'_s(a'_1) y'_s(a'_2) - x'_s(a'_1) y'_s(a'_2) | + \sum_{a'_1, a'_2} | x'_s(a'_1) y'_s(a'_2) - x'_s(a'_1) y'_s(a'_2) | \]

\[ \leq \sum_{a'_1} | x'_s(a'_1) - x'_s(a'_1) | + \sum_{a'_2} | y'_s(a'_2) - y'_s(a'_2) | \]

\[ = \| x'_s - x'_s \|_1 + \| y'_s - y'_s \|_1 \]

and

\[ \left| (x'_s) \top \left( Q^{x',y'}_{s'} - Q^{x,y}_{s'} \right) y'_s \right| \leq \max_{a_1, a_2} \left| Q^{x',y'}_{s'}(a_1, a_2) - Q^{x,y}_{s'}(a_1, a_2) \right| := \| Q^{x',y'}_{s'} - Q^{x,y}_{s'} \|_{\text{max}}. \]

By substituting the upper bound on \( Q_{\text{iff}} \) above into (D.19),

\[ \| Q^{x',y'}_{s'} - Q^{x,y}_{s'} \|_{\text{max}} \]

\[ \leq \gamma \max_{s'} \frac{1}{1 - \gamma} (\| x'_s - x'_s \|_1 + \| y'_s - y'_s \|_1) + \gamma \max_{s'} \| Q^{x',y'}_{s'} - Q^{x,y}_{s'} \|_{\text{max}}. \]

Therefore,

\[ \max_s \| Q^{x',y'}_{s'} - Q^{x,y}_{s'} \|_{\text{max}} \]

\[ \leq \gamma \max_{s'} \frac{1}{1 - \gamma} (\| x'_s - x'_s \|_1 + \| y'_s - y'_s \|_1) + \gamma \max_{s'} \| Q^{x',y'}_{s'} - Q^{x,y}_{s'} \|_{\text{max}}. \]

which yields the desired result. \( \square \)

D.7 Other auxiliary lemmas

In this section, we provide other auxiliary lemmas that are helpful in our analysis.

D.7.1 Auxiliary lemma for potential functions

Lemma 68 For any \( N \)-player Markov potential game with instantaneous reward bounded in \([0, 1]\), it holds that

\[ \left| \Phi^\pi(\mu) - \Phi^{\pi'}(\mu) \right| \leq \frac{N}{1 - \gamma} \]

for any \( \pi, \pi' \in \Pi \) and \( \mu \in \Delta(S) \).
**Proof.** By the potential property,

\[
\Phi^\pi(\mu) - \Phi^{\pi'}(\mu) = (\Phi^\pi - \Phi^{\pi'}; \pi - 1) + (\Phi^{\pi'}; (1, 2) - \Phi^{\pi'}; (1, 2)) + \ldots + (\Phi^{\pi}; (1, 2) - \Phi^{\pi'}; N - \Phi^{\pi'}; N)
\]

\[
= (V^\pi_N - V^\pi'; N) + (V^{\pi'}; 1, 2) - V^{\pi'}; 1, 2) + \ldots + (V^{\pi}; N - V^{\pi'}; N)
\]

\[
\leq \frac{V^\pi - V^{\pi'} N}{1 - \gamma}
\]

where the last inequality is due to \( V^\pi_i - V^\pi' i \leq \frac{1}{1 - \gamma} \) for any \( \pi \) and \( \pi' \). By symmetry, \( \Phi^{\pi'}(\mu) - \Phi^\pi(\mu) \leq \frac{N}{1 - \gamma} \).

---

**D.7.2 Auxiliary lemma for single-player MDPs**

**Lemma 69 (State-action value function difference)** Suppose that two MDPs have the same state/action spaces, but different reward and transition functions: \((r, p)\) and \((\tilde{r}, \tilde{p})\). Then, for a given policy \(\pi\), two action value functions associated with two MDPs satisfy

\[
\max_{s, a} |Q^\pi(s, a) - \tilde{Q}^\pi(s, a)| \leq \frac{1}{1 - \gamma} \max_{s, a} |r(s, a) - \tilde{r}(s, a)|
\]

\[
+ \frac{\gamma}{(1 - \gamma)^2} \max_{s, a} \| p(\cdot | s, a) - \tilde{p}(\cdot | s, a) \|_1.
\]

**Proof.** By the Bellman equations,

\[
Q^\pi(s, a) = r(s, a) + \gamma \sum_{s', a'} p(s' | s, a) \pi(a' | s') Q^\pi(s', a')
\]

\[
\tilde{Q}^\pi(s, a) = \tilde{r}(s, a) + \gamma \sum_{s', a'} \tilde{p}(s' | s, a) \pi(a' | s') \tilde{Q}^\pi(s', a').
\]

Subtracting equalities above on both sides yields

\[
|Q^\pi(s, a) - \tilde{Q}^\pi(s, a)|
\]

\[
\leq |r(s, a) - \tilde{r}(s, a)| + \gamma \left| \sum_{s', a'} p(s' | s, a) \pi(a' | s') Q^\pi(s', a') \right|
\]

\[
+ \gamma \left| \sum_{s', a'} \tilde{p}(s' | s, a) \pi(a' | s') \left( Q^\pi(s', a') - \tilde{Q}^\pi(s', a') \right) \right|
\]

\[
\leq |r(s, a) - \tilde{r}(s, a)| + \frac{\gamma}{1 - \gamma} \left( p(\cdot | s, a) - \tilde{p}(\cdot | s, a) \right) + \gamma \max_{s', a'} \left| Q^\pi(s', a') - \tilde{Q}^\pi(s', a') \right|.
\]

Taking the maximum over \((s, a)\) leads to

\[
\max_{s, a} |Q^\pi(s, a) - \tilde{Q}^\pi(s, a)|
\]

\[
\leq \max_{s, a} |r(s, a) - \tilde{r}(s, a)| + \frac{\gamma}{1 - \gamma} \max_{s, a} \| p(\cdot | s, a) - \tilde{p}(\cdot | s, a) \|_1
\]

\[
+ \gamma \max_{s, a} \left| Q^\pi(s, a) - \tilde{Q}^\pi(s, a) \right|.
\]
which leads to the desired inequality after rearrangement.

\[ \square \]

### D.7.3 Auxiliary lemma for multi-player MDPs

**Lemma 70** Let \( \pi, \pi' \) and \( \bar{\pi} \) be three policies, and \( \mu \) be some initial distribution. Let \( \mu' \) be a state distribution that generates a state according to the following: first sample an \( s_0 \) from \( d^\mu(\cdot) \), then execute \( \bar{\pi} \) for one step, and then output the next state. Then,

\[
\left\| \frac{d^{\pi'}}{\mu} \right\|_\infty \leq \frac{\kappa^2}{\gamma} \left( \sup_{\bar{\pi}} \left\| \frac{d^{\bar{\pi}}}{\mu} \right\|_\infty \right)^2.
\]

**Proof.** For a particular state \( s^\sharp \), we view the supremum \( \sup_{\bar{\pi}} \frac{d^{\pi'}(s^\sharp)}{\mu(s^\sharp)} \) as the optimal value of an MDP whose reward function is \( r(s, a) = \frac{1 - \gamma}{\mu(s^\sharp)} 1[s = s^\sharp] \) and initial state is generated by \( \mu \). The optimal value of this MDP is upper bounded by \( \kappa_\mu \) by Definition\( \text{2} \). We next consider the following non-stationary policy for this MDP: first execute \( \bar{\pi} \) for one step, and then execute \( \pi' \) in the rest of the steps. The discounted value of this non-stationary policy is lower bounded by

\[
\gamma \sum_{s} \Pr (s_1 = s \mid s_0 \sim \mu, a_0 \sim \bar{\pi}(\cdot \mid s_0)) \frac{d^{\pi'}(s^\sharp)}{\mu(s^\sharp)} = \gamma \sum_{s_0, a_0, s} \mu(s_0) \bar{\pi}(a_0 \mid s_0) p(s \mid s_0, a_0) d^{\pi'}(s^\sharp) \mu(s^\sharp).
\]

We can upper and lower bound the discounted sum above as the following:

\[
\frac{\gamma}{\kappa_\mu} \sum_{s_0, a_0, s} d^{\pi'}(s^\sharp) \mu(s^\sharp) \leq \gamma \sum_{s_0, a_0, s} \mu(s_0) \bar{\pi}(a_0 \mid s_0) p(s \mid s_0, a_0) d^{\pi'}(s^\sharp) \mu(s^\sharp) \leq \kappa_\mu
\]

where the right inequality is due to that this discounted value must be upper bounded by the optimal value of this MDP, which has an upper bound \( \kappa_\mu \), and the left inequality is by the definition of \( \kappa_\mu \). Now notice that

\[
\mu'(s) = \sum_{s_0, a_0} d^{\pi'}(s_0) \bar{\pi}(a_0 \mid s_0) p(s \mid s_0, a_0)
\]

by the definition of \( \mu' \). Plugging this into the previous inequality, we get

\[
\frac{\gamma}{\kappa_\mu} \times \frac{d^{\pi'}(s^\sharp)}{\mu(s^\sharp)} \leq \kappa_\mu.
\]

Since this holds for any \( s^\sharp \), this gives

\[
\left\| \frac{d^{\pi'}}{\mu} \right\|_\infty \leq \frac{\kappa^2}{\gamma}.
\]

\[ \square \]
**Algorithm 13** Stochastic projected gradient descent with weighted averaging

1. **Parameters**: \( W, \lambda^{(k)}, \) and \( \beta_k^{(K)} \).
2. **Input**: Stepsize \( \alpha \), total number of iterations \( K > 0 \).
3. **Initialization**: \( w^{(0)} = 0 \).
4. **for** step \( k = 1, \ldots, K \) **do**
5. Draw \( \nabla^{(k)} \) form a distribution such that \( \mathbb{E}[\nabla^{(k)} | w^{(k)}] \in \partial f(w^{(k)}) \).
6. Update \( w^{(k+1)} = P_{\|w\| \leq W} \left( w^{(k)} - \lambda^{(k)} \nabla^{(k)} \right) \).
7. **end for**
8. **Output**: \( \sum_{k=0}^{K} \beta_k^{(K)} w^{(k)} \).

**D.7.4 Auxiliary lemma for stochastic projected gradient descent**

Algorithm [11] serves a sample-based algorithm if we solve the empirical risk minimization problem (5.15) via a stochastic projected gradient descent,

\[
  w^{(k+1)}_i = P_{\|w\| \leq W} \left( w^{(k)}_i - \lambda^{(k)} \hat{\nabla}^{(t)}_i (s^{(k)}_i, a^{(k)}_i) \right)
\]  

(D.20)

where \( \hat{\nabla}^{(t)}_i := 2(\langle \phi_i, w^{(k)}_i \rangle - R^{(k)}_i)\phi_i \) is the \( k \)-th gradient of (5.15) and \( \lambda^{(k)} > 0 \) is the stepsize. We assume that the smallest eigenvalue of correlation matrix \( \mathbb{E}_{s,a_i} [\phi_i(s,a_i)\phi_i(s,a_i)^\top] \) is positive.

For a constrained convex optimization, minimize \( w \in \{ w \mid \|w\| \leq W \} f(w) \), where \( f(w) \) is a convex function and \( W > 0 \), we consider a basic method for solving this problem: the stochastic projected gradient descent in Algorithm [13] where \( P_{\|w\| \leq W} \) is a Euclidean projection in \( \mathbb{R}^d \) to the constraint set \( \|w\| \leq W \).

**Lemma 71** Let \( w^* := \argmin_{w \in \{ w \mid \|w\| \leq W \}} f(w) \). Suppose \( \text{Var}(\nabla^{(k)}) \leq \sigma^2 \). If we run Algorithm [13] with stepsize \( \lambda^{(k)} = O\left( \frac{1}{1+k} \right) \) and \( \beta_k^{(K)} = \frac{1}{\lambda^{(k)}} \sum_{r=0}^{K} 1/\lambda^{(r)} \), then,

\[
  \mathbb{E} \left[ f \left( \sum_{k=0}^{K} \beta_k^{(K)} w^{(k)} \right) \right] - f(w^*) \lesssim \frac{\sigma^2 W^2 d}{K}.
\]

**PROOF.** See the proof of Theorem 1 in [58].