# Independent Learning in Markov Potential Games: New Insights and Constraints

**Master Thesis**

**Author(s):**
Jordan, Philip

**Publication date:**
2023

**Permanent link:**
https://doi.org/10.3929/ethz-b-000646915

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# Independent Learning in Markov Potential Games: New Insights and Constraints

Master Thesis

Philip Jordan

December 7, 2023

Advisors: Dr. Anas Barakat,
Prof. Dr. Niao He

Department of Computer Science, ETH Zürich

**Abstract**

Markov games offer a formal mathematical framework for modeling multi-agent reinforcement learning problems. Markov Potential Games (MPGs) represent a subclass of mixed cooperative and competitive Markov games for which finding Nash equilibria is tractable. The recently introduced class of constrained Markov Potential Games (CMPGs) generalizes MPGs to model the case where the agents' reward maximization is subject to global constraints. Existing methods for learning Nash equilibria (NE) in MPGs can be categorized into centralized and independent learning algorithms. Notably, for learning $\varepsilon$-approximate NE, the best-known sample complexity achieved by centralized algorithms is significantly lower than for independent learning ($\mathcal{O}(\varepsilon^{-3})$ vs. $\mathcal{O}(\varepsilon^{-5})$). Moreover, no converging independent learning algorithm is known for CMPGs. Nevertheless, whether these gaps are inherent is unknown, i.e., no provable separation between centralized and independent learning has been shown. Continuing on this quest, our contributions are twofold: (a) We propose a new playerwise policy gradient (PG) algorithm that requires coordination among players, however, improves on iteration and sample complexity regarding the dependence on the number of players $m$. The proposed method also improves over the $m$-dependence in the complexity of previously known centralized algorithms. (b) In the constrained case, we make progress on closing the gap between centralized and independent learning by providing an independent policy gradient algorithm for learning approximate constrained Nash equilibria in CMPGs. Inspired by contemporary optimization literature, our algorithm performs proximal-point-like updates augmented with a regularized constraint set. Each proximal step is solved inexactly using a stochastic switching gradient algorithm. Under some technical constraint qualification conditions, we establish convergence guarantees towards constrained approximate Nash equilibria. We perform simulations to illustrate our results in two real-world applications of NE-learning in CMPGs.

# Contents

Chapter 1

# Introduction

In multi-agent reinforcement learning (MARL), several agents interact within a shared dynamic and uncertain environment evolving over time depending on the agents' individual strategic decisions. Each agent aims to maximize their own individual reward, which may, however, depend on all players'[1] decisions. The framework of stochastic games, a.k.a. Markov games, initiated by Shapley [Sha53] is certainly the most widely adopted mathematical framework for studying MARL.

This thesis studies Nash equilibrium learning in Markov games. More specifically, we will be concerned with three major aspects that can be seen as orthogonal additions to the Markov game setting: (a) requiring a potential structure leading to the notion of Markov Potential Games; (b) considering a centralized vs. an independent learning information structure; and (c) incorporating constraints into the game. The remainder of this section motivates and introduces each of these aspects informally.

Beyond the fully competitive Markov game setting that has been investigated comparatively more in the literature, the cooperative setting has been studied less. However, the ability to cooperate between learning agents is crucial to improve their joint welfare and achieve social welfare for artificial intelligence (see [DHB+20, DBH+21] for an extensive discussion of the need to promote cooperative AI). Markov Potential Games (MPGs) form a particular class of structured Markov games that has been actively investigated in recent years [MZZ18, LOPP22, FMOP22, ZRL22, SMB22, DWZJ22, ZMD+22, MWPS22, ZCLW23]. Interestingly, MPGs represent a class of mixed cooperative/competitive Markov games including, as a particular case, pure identical interest Markov games in which all the agents' reward and cost functions are identical. Moreover, while intractability results were very recently developed for general stochastic games [DGZ23], [JMS22], computing Nash equilibria for MPGs turns out to be tractable thanks to the potential structure of the game. Further, we shall mention that MPGs can be seen as a generalization of potential games, a particular class of strategic normal-form games studied in game theory. The latter class of games coincides with one-state MPGs. From this viewpoint, MPGs incorporate a dynamic stateful aspect into static potential games, which have been extensively studied since their introduction in [MS96], and which already have numerous applications, for instance, in wireless networks. Applications of MPGs include real-world problems such as routing games (transportation networks), wireless communications, congestion games, smart grids, traffic network systems with self-driving vehicles, and cloud computing.

Besides reward maximization, agents may also contend with satisfying constraints that are often dictated by multi-agent RL applications. Prominent such real-world applications

[1] We will use *player* and *agent* interchangeably.

include multi-robot control on cooperative tasks [GGC+23] as well as autonomous driving [SSSS16, LKT+23] where physical system constraints and safety considerations such as collision avoidance are of primary importance. In other applications, agents may be subject to soft constraints such as average users' total latency thresholds in wireless networks or average power constraints in signal transmission. Each agent seeks to maximize their reward while accounting for constraints that are coupled among agents. Constrained Markov games (CMGs) [AS00] offer a mathematical framework to model multi-agent RL problems incorporating coupled constraints. Analogous to the unconstrained case of MGs and MPGs, requiring a potential structure yields the recently introduced class of constrained Markov Potential Games (CMPGs) [ARHK23] for which computing (constrained) Nash equilibria is tractable.

Independent learning has recently attracted increasing attention thanks to its versatility as a learning protocol. We refer the reader to a recent nice survey on the topic [OSZ21]. In this protocol, agents can only observe the realized state and their own reward and action in each stage to individually optimize their return. In particular, each agent does not observe actions or policies from any other agent. This protocol offers several advantages including the following aspects: (a) Scaling: independent learning dynamics do not scale exponentially with the number of players in the game (also known as the curse of multi-agents); (b) Privacy protection: agents may avoid sharing their local data and information to protect their privacy and autonomy; (c) Communication cost: a central node that can bidirectionally communicate with all agents may not exist or may be too expensive to afford. Therefore, this protocol is particularly appealing in several applications where agents must make decisions independently in a decentralized manner. For example, dynamic load balancing, which consists of evenly assigning clients to servers in distributed computing, demands learning algorithms that minimize communication overhead to enable low-latency response times and scalability across large data centers. This task has been modeled as an MPG [YD22]. In other applications, such as the pollution tax model and the distributed energy marketplace detailed in Section 4.3, coordination is inherently ruled out due to the competitive nature of the players' interactions. Independent learning algorithms have been proposed for unconstrained multi-agent RL problems such as zero-sum Markov games [DFG20, SZL+21, CZM+23] as well as for unconstrained MPGs in a recent line of works [LOPP22, ZMD+22, ZRL22, DWZJ22, MWPS22].

In the case of MPGs, current results exhibit a gap between guarantees achievable by centralized (i.e., allowing for coordination/communication via a central entity, enabling, e.g., sharing of individual players' policies or turn-based updates) vs. independent learning algorithms. In particular, this gap can be observed in the following two ways:

1. The best known independent algorithm for learning $\varepsilon$-Nash equilibria in MPGs achieves an $\mathcal{O}(\varepsilon^{-4.5})$ sample complexity via a variance-reduced version [MYZB22] of independent policy gradient ascent [LOPP22, DWZJ22]. In contrast, the centralized, turn-based Nash-CA algorithm [SMB22] has sample complexity $\mathcal{O}(\varepsilon^{-3})$.

2. In the case of CMPGs, a centralized adaptation of Nash-CA has been proposed. However, the more challenging problem of independently learning constrained Nash equilibria has not been settled so far.

As central themes of this thesis, we pose the following two questions:

1. *Can **centralization** be leveraged to improve over the iteration and sample complexity of independent learning algorithms for MPGs?*

2. *Can we design an **independent** learning algorithm for **constrained** MPGs with non-asymptotic global convergence guarantees?*

## 1.1 Organization and Contributions

The remainder of this thesis is structured as follows. In Section 1.2, we discuss related work. Chapter 2 formally introduces (constrained) Markov Potential Games, defines the respective notions of Nash equilibria, and clarifies the game's information structure. In Chapters 3 and 4, we establish **our contributions** which can be summarized as follows:

- Chapter 3 addresses Nash equilibrium learning in unconstrained MPGs with a particular focus on the role of centralization vs. independence. We begin by describing and comparing existing methods. Then, we propose a centralized learning algorithm that combines ideas from coordinate ascent and policy gradient ascent. Our analysis shows that in an exact gradients setting, this new method achieves an iteration complexity that, unlike for existing algorithms, does not depend on the number of players involved in the game. Furthermore, we investigate a simultaneous best-response update and, as a partial negative result, show that naively performing such independent updates cannot guarantee global convergence.
- As our main contribution, in Chapter 4, we design an algorithm for independent learning of constrained $\varepsilon$-approximate Nash equilibria (NE) in CMPGs. Inspired by recent works in nonconvex optimization under nonconvex constraints, our algorithm implements an inexact proximal-point update augmented with a regularized constraint set. In particular, the inexact proximal step is computed using a stochastic gradient switching algorithm for solving the resulting subproblem where both the objective and the constraint functions are strongly convex. In particular, the different agents can run the algorithm independently without taking turns. Moreover, we analyze the proposed algorithm and establish its sample complexity to converge to an $\varepsilon$-approximate NE of the CMPG with polynomial dependence on problem parameters. Our analysis requires new technical developments that do not rely on results from the CMDP literature. In addition, we illustrate the performance of our algorithm on two simple CMPG applications: a pollution tax model and a marketplace for distributed energy resources.

Finally, we conclude the thesis in Chapter 5 by pointing out some limitations of our current results and by highlighting promising directions for future work. Proofs of all our results are deferred to Appendix A for Chapter 3 and to Appendix B for Chapter 4.

## 1.2 Related Work

We next discuss some closely related work. We first cover results regarding unconstrained and constrained MPGs. Then, we review related approaches for stateless, possibly constrained, potential games. Finally, we discuss the use of inexact proximal-point methods in recent optimization literature, which serves as inspiration for our independent NE-learning algorithm for CMPGs.

**Markov Potential Games (MPGs).** MPGs have been introduced as a natural extension of normal form potential games [MS96] to the dynamic setting starting with state-based potential games [Mar12] and later Markov games [MZZ18]. [LOPP22] introduced a variant of MPGs and proposed independent stochastic policy gradient methods with an $\mathcal{O}(\varepsilon^{-6})$ sample complexity to reach an $\varepsilon$-approximate NE. Similar results were shown in [ZRL22] with model-based algorithms. This result was later improved to an $O(\varepsilon^{-5})$ sample complexity for large state-action spaces with linear function approximation [DWZJ22] and further to $O(\varepsilon^{-4.5})$ by reducing the variance of the agent-wise stochastic policy gradients [MYZB22]. [ZMD+22] explored using the softmax policy parametrization

instead of the direct parametrization. In particular, they established an $\mathcal{O}(\varepsilon^{-2})$ iteration complexity in the deterministic setting and showed the benefits of using regularization to improve the convergence rate. [MWPS22] proposed a fully independent and decentralized two-timescale algorithm for MPGs with asymptotic guarantees where players may not even know the existence of other players. [NLKS22] provided verifiable structural assumptions under which a Markov game is an MPG and further provided several algorithms for solving MPGs in the deterministic setting. [SMB22] proposed an $\mathcal{O}(\varepsilon^{-3})$ sample complexity coordinate ascent algorithm (Nash-CA), which requires coordination among players. [GLM+23] recently introduced the class of $\alpha$-MPGs which relaxes the definition of MPGs by allowing $\alpha$-deviations with respect to (w.r.t.) the potential function. More recently, [ZCLW23] introduced a class of networked MPGs for which they proposed a localized actor-critic algorithm with linear function approximation. All the aforementioned works focused on the unconstrained setting.

**Constrained Markov Potential Games (CMPGs).** There has been a vast array of works in multi-agent RL with safety constraints in practice, see, e.g., [EABA+21, GGC+23] and the references therein. [AS00] defined constrained Markov games and provided sufficient conditions for the existence of stationary constrained NE. Non-asymptotic theoretical convergence guarantees to game-theoretic solution concepts for constrained multi-agent RL are relatively scarce in the literature. [CMZ22] introduced a notion of correlated equilibria for general constrained Markov games and provided a primal-dual algorithm for learning those equilibria. Unlike ours, their setting exhibits strong duality, enabling the use of primal-dual algorithms. [DWY+23] established regret guarantees for episodic two-player zero-sum constrained Markov games. [ARHK23] introduced the class of constrained MPGs. Inspired by Nash-CA [SMB22], they proposed a constrained variant of the algorithm which enjoys an $\mathcal{O}(\varepsilon^{-5})$ sample complexity. Crucially, this algorithm requires coordination between agents and cannot be implemented independently by the agents.

**(Constrained) Potential Games.** While unconstrained potential games have been studied extensively in the game theory literature [MS96, BR20, CCC22, CMS06, SMK18], only few results exist on their constrained counterpart. [Zhu08] studies structural properties of constrained potential games with coupled constraints, i.e., a stateless version of a setting that is otherwise similar to ours. Through a Lagrangian approach, it is observed that the solution to the respective constrained maximization problem with respect to the potential function constitutes a constrained Nash equilibrium — an observation that we, as well as [ARHK23], also build on when motivating our approach. However, our further insights significantly differ from [Zhu08] since we cannot hope to reach an optimal solution to the constrained potential maximization problem in our stateful nonconvex setting. Instead, we need to leverage the specific structure of CMPGs to argue that satisfying local approximate KKT conditions also suffices for attaining approximate constrained Nash equilibria. In particular, [Zhu08] also does not study independent learning and generally does not provide convergence guarantees on algorithms for reaching approximate equilibria.

**Inexact Proximal-Point Methods.** The idea of using inexact proximal-point methods to solve *nonconvex* problems has been fruitfully exploited in the literature for a couple of decades (see, e.g., [HS09, DG19]). A recent line of works ([BDL23, MLY20]; and also [JG23]) extended this idea in order to solve nonconvex optimization problems with nonconvex functional constraints. The initial nonconvex problem is transformed into a sequence of convex problems by adding quadratic regularization terms

to *both* the objective and constraints. These works also established convergence rates to Karush–Kuhn–Tucker (KKT) points under constraint qualification conditions. Our present work is inspired by this recent line of research. We point out, though, that we deal with a multi-agent RL problem and provide convergence guarantees to approximate constrained NE. In these regards, our independent algorithm design and our analysis require several new technical developments. Alternatives to inexact proximal methods for handling nonconvex constraints include second-order approaches [NW06, CGT15] and penalty methods [WMY17, FKLS21].

Chapter 2

# Preliminaries

In this preliminary chapter, we formally introduce the mathematical framework of Markov Potential Games and constrained Markov Potential Games and related concepts such as Nash equilibria used throughout the rest of the thesis. We also clarify the information structures considered, that is, the difference between centralized and independent learning.

**Notation.** Throughout the thesis, $\|\cdot\|$ denotes the standard Euclidean norm $\|\cdot\|_2$. Furthermore, given a set $X$, a totally ordered set $Y$, and a function $f : X \to Y$, we use the definition $\arg\min_{x\in X} f(x) := \{x \in X \mid \forall x' \in X : f(x) \leq f(x')\} \subset X$, and analogously define $\arg\max_{x\in X} f(x)$.

## 2.1 Markov Games

This first section defines our reinforcement learning setup in the context of (constrained) Markov games.

**Markov game.** An $m$-player Markov game is a tuple $\mathcal{G} = (\mathcal{S}, \mathcal{N}, \{\mathcal{A}_i, r_i\}_{i\in\mathcal{N}}, \mu, P, \kappa)$ with:

- A finite shared state space $\mathcal{S}$ of cardinality $S := |\mathcal{S}|$.
- A finite set of $m$ agents $\mathcal{N} := \{1, \ldots, m\}$.
- A finite set of actions $\mathcal{A}_i$ of cardinality $A_i := |\mathcal{A}_i|$ for all $i \in \mathcal{N}$ with $A_{\max} := \max_{i\in\mathcal{N}} A_i$. The joint action space is denoted by $\mathcal{A} := \prod_{i\in\mathcal{N}} \mathcal{A}_i$.
- A reward function $r_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$ for each agent $i \in \mathcal{N}$.
- A distribution $\mu$ over states from which the game's initial state is drawn.
- A probability transition kernel $P$: For any state $s \in \mathcal{S}$ and any joint action $a \in \mathcal{A}$, the game transitions from state $s$ to a state $s' \in \mathcal{S}$ with probability $P(s'|s, a)$ and the game terminates with probability $\kappa_{s,a} > 0$. We further define $\kappa := \min_{s\in\mathcal{S}, a\in\mathcal{A}} \kappa_{s,a}$ and $\gamma := 1 - \kappa$.

At each time step $t \geq 0$ of a given episode of the game, all the agents observe a shared state $s_t \in \mathcal{S}$ and choose a joint action $a_t \in \mathcal{A}$. Then, each agent $i \in \mathcal{N}$ receives a reward $r_i(s_t, a_t)$. The game either stops at time $t$ with probability $\kappa_{s_t,a_t}$ or proceeds by transitioning to a state $s_{t+1}$ drawn from the distribution $P(\cdot|s_t, a_t)$. We denote by $T_e$ the

random stopping time when the episode terminates.[1] For a similar setting, see [DFG20, GLMVG22].

Below, we extend Markov games to incorporate constraints by introducing an additional cost function.

**Constrained Markov game.** An $m$-player constrained Markov game is a tuple $\mathcal{G} = (\mathcal{S}, \mathcal{N}, \{\mathcal{A}_i, r_i, c_i\}_{i\in\mathcal{N}}, \alpha, \mu, P, \kappa)$ where $S, \mathcal{N}, A_i, r_i, \mu, P$, and $\kappa$ are defined as for Markov games above, and additionally, we have:

- A cost function $c_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$ for each agent $i \in \mathcal{N}$. Throughout this thesis, we will assume that all cost functions are identical across the agents and equal to a single cost function $c$. The case of multiple such common costs can be addressed with our approach with minor modifications. The case where cost functions may differ between players is more challenging and left for future work.
- A constraint threshold $\alpha \in \mathbb{R}$ that is used to define the set of feasible policies (see definition in paragraph on policies and value functions below).

The game interaction is analogous to the unconstrained case, except that after choosing a joint action $a_t \in \mathcal{A}$, each agent $i \in \mathcal{N}$, in addition to the reward $r_i(s_t, a_t)$, incurs a cost $c(s_t, a_t)$.

In the rest of this thesis, we aim to minimize both rewards and costs to align with conventions from the (constrained) optimization literature. The equivalence to the common RL reward maximization formulation follows from considering reward functions $1 - r_i$ instead of $r_i$ for each $i \in \mathcal{N}$.

Next, we introduce players' individual and joint policies. Moreover, we define value functions on joint policies in terms of both rewards and cost.

**Policies and Value Functions.** Each agent $i \in \mathcal{N}$ chooses their actions according to a randomized stationary policy denoted by $\pi_i \in \Pi^i := \Delta(\mathcal{A}_i)^{\mathcal{S}}$ where $\Delta(\mathcal{A}_i)$ is the probability simplex over the finite action space $\mathcal{A}_i$. The set of joint policies $\pi = (\pi_i)_{i\in\mathcal{N}}$ is denoted by $\Pi := \prod_{i\in\mathcal{N}} \Pi^i$ and we further use the notation $\pi_{-i} = (\pi_j)_{j\in\mathcal{N}\setminus\{i\}} \in \Pi^{-i} := \prod_{j\in\mathcal{N}\setminus\{i\}} \Pi^j$ for joint policies of all agents other than $i$. For any $u \in \{r_i \mid i \in \mathcal{N}\} \cup \{c\}$ and any joint policy $\pi \in \Pi$, we define the value function $V_u(\pi)$ for every state $s \in \mathcal{S}$ by $V_{u,s}(\pi) := \mathbb{E}[\sum_{t=0}^{T_e} u(s_t, a_t) | s_0 = s]$. The shorthand notation $V_u(\pi)$ will stand for $V_{u,\mu}(\pi) := \mathbb{E}_{s\sim\mu}[V_{u,s}(\pi)]$. For any policy $\pi \in \Pi$ and $s, s' \in \mathcal{S}$, the state visitation distribution is defined by $d_s^\pi(s') := \mathbb{E}[\sum_{t=0}^{T_e} \mathbb{1}_{\{s_t=s'\}} | s_0 = s]$ where $\mathbb{1}$ is the indicator function and we write $d_\mu^\pi(s') = \mathbb{E}_{s\sim\mu}[d_s^\pi(s')]$. For constrained MGs, we additionally define the set of feasible policies as $\Pi_c := \{\pi \in \Pi \mid V_c(\pi) \le \alpha\}$. Moreover, the set of feasible policies for agent $i \in \mathcal{N}$ when the policy of the other agents is fixed to $\pi_{-i} \in \Pi^{-i}$ is denoted by $\Pi_c^i(\pi_{-i}) := \{\pi_i \in \Pi^i \mid (\pi_i, \pi_{-i}) \in \Pi_c\}$.

As a solution concept, we will be studying convergence to the well-known game-theoretic concept of a Nash equilibrium — a state where no player can individually improve by deviating to a different strategy (i.e., policy). A similar equilibrium notion can also be defined for constrained games.

**Nash Equilibria.** For any $\varepsilon \ge 0$, a joint policy $\pi^* \in \Pi$ is called an $\varepsilon$-approximate constrained NE if for every $i \in \mathcal{N}$ and any policy $\pi_i' \in \Pi^i$, we have $V_{r_i}(\pi^*) - V_{r_i}(\pi_i', \pi_{-i}^*) \le \varepsilon$. When $\varepsilon = 0$, such a policy $\pi^*$ is called an NE policy, and no agent has an incentive to

[1] The discounted infinite horizon setting can also be addressed with minor adaptations.

unilaterally deviate from an NE policy $\pi^*$. A natural way of generalizing the concept of Nash equilibrium in an MG to the constrained setting is to restrict the $i$-th players allowed deviations to feasible policies, i.e., to $\Pi_c^i(\pi_{-i})$. Therefore, we define $\pi = (\pi_i, \pi_{-i}) \in \Pi_c$ to be a constrained $\varepsilon$-Nash equilibrium (constrained $\varepsilon$-NE) if for every $i \in \mathcal{N}$ and any policy $\pi_i' \in \Pi_c^i(\pi_{-i}^*)$, we have $V_{r_i}(\pi^*) - V_{r_i}(\pi_i', \pi_{-i}^*) \leq \varepsilon$. We refer the reader to [AS00] for the existence of stationary constrained NE.

Under common assumptions in complexity theory (Exponential Time Hypothesis for PPAD), finding Nash equilibria in general games, even approximately, requires an exponential number of game interactions [Das13, Rub16]. In the following, we introduce a potential structure leading to the definition of MPGs and CMPGs as tractable subclasses of MGs and constrained MGs, respectively.

**Potential Structure and (C)MPGs.** In an MPG ([MZZ18, LOPP22]), for each state $s \in \mathcal{S}$, there exists a so-called potential function $\Phi_s : \Pi \to \mathbb{R}$ such that for all $i \in \mathcal{N}$, it holds that

$$V_{r_i,s}(\pi_i, \pi_{-i}) - V_{r_i,s}(\pi_i', \pi_{-i}) = \Phi_s(\pi_i, \pi_{-i}) - \Phi_s(\pi_i', \pi_{-i}) \tag{2.1}$$

for any policies $(\pi_i, \pi_{-i}) \in \Pi$, and $\pi_i' \in \Pi^i$. We will also use the notation $\Phi(\pi) := \mathbb{E}_{s\sim\mu}[\Phi_s(\pi)]$. Notice that the fully cooperative setting, when all the reward functions of the players are identical, is a particular instance of an MPG. Note also that the potential function is typically unknown for the players interacting in the game. A constrained Markov Potential Game (CMPG), see also [ARHK23], is a constrained MG such that for all $i \in \mathcal{N}$, $\pi = (\pi_i, \pi_{-i}) \in \Pi, \pi_i' \in \Pi^i$, and $s \in \mathcal{S}$, there exists a potential $\Phi_s : \Pi \to \mathbb{R}$ such that equation (2.1) holds.

## 2.2 Information Structure

For multi-agent RL environments, it is important to clarify to what extent information exchange may happen among agents. In this thesis, we will consider the following two common information structures; see also [ZYB21] for a survey on various such settings.

**Independent Setting.** Suppose all players interact with the game by executing their policies for a fixed number of episodes. Notably, during the learning procedure, each player executes their policy at each episode of the game to sample a trajectory and exclusively observes their own trajectory $(s_t, a_{i,t}, r_i(s_t, a_t), c(s_t, a_t))_{0\leq t\leq T_e}$. In particular, in independent learning, a player does not have access to the policies of other players or their chosen actions, and there is no communication among players outside the game interaction. Such a protocol was previously considered, for instance, in two-player zero-sum Markov games [DFG20, CZM+23] as well as for unconstrained MPGs [LOPP22, DWZJ22, MWPS22].

**Centralized Setting.** In the centralized setting, game interaction happens as in the independent setting. However, we additionally assume the existence of a central controller that can aggregate information from the agents (such as their individual policies or obtained rewards) and, for instance, coordinate agents to update their policies in a turn-based manner. This setting is assumed, e.g., for learning Nash equilibria in MPGs via Nash-CA in [SMB22].

Chapter 3

# Centralized vs. Independent Learning in Markov Potential Games

This chapter addresses Nash equilibrium learning in unconstrained Markov Potential Games, with a focus on the role of centralization in the form of coordination among agents. We proceed by first introducing the two main approaches present in the literature leading to independent (Section 3.1, PGA), and turn-based (Section 3.2, Nash-CA) algorithms, respectively. Building on these existing results, we then explore two new directions: (a) In Section 3.3, by allowing for coordination during playerwise policy gradient updates, we obtain a centralized algorithm that does not have an $m$-dependence in its iteration complexity and may hence be more suitable than existing methods for learning in MPGs with a large number of players $m$. (b) In Section 3.4, we discuss ideas towards a decentralized version of Nash-CA and give a negative result for an approach via simultaneous best-response updates. We refer to Table 3.1 for a schematic illustration of the methods considered in this chapter.

| | centralized | independent |
|---|---|---|
| policy gradient | centralized PGA<br>Section 3.3<br>$\mathcal{O}(\varepsilon^{-2})$ | independent PGA<br>[LOPP22],[ZRL22],[DWZJ22]<br>$\mathcal{O}(m\varepsilon^{-2})$ |
| best-response | Nash-CA<br>[SMB22]<br>$\mathcal{O}(m\varepsilon^{-1})$ | simultaneous best-response<br>Section 3.4<br>no global convergence |

**Table 3.1:** Overview of existing and proposed methods with iteration complexities in a deterministic setting, i.e., with access to exact gradients and value functions. $\mathcal{O}(\cdot)$ hides polynomial dependencies in $S, A_{\max}, D, \gamma$, and $\Phi_{\max}$ (which may depend on $m$ in the worst case) where the distribution mismatch coefficient $D := \max_{\pi\in\Pi} \|d_\mu^\pi/\mu\|_\infty$ is assumed to be finite. 'Centralized' and 'independent' refer to the information structures introduced in Section 2.2.

## 3.1 Independent Policy Gradient Ascent

In this section, we describe ideas for proving convergence of independent policy gradient ascent[1] (PGA) as initially developed in [LOPP22]. Our centralized PGA method, introduced in the next section, will build on these insights. Starting with some initial

[1]To be consistent with common RL algorithm naming, we stick to calling the algorithm PG ascent instead of descent, even though due to our reward minimization convention introduced in Chapter 2 we perform a descent.

policy $\pi^{(0)} \in \Pi$, and choosing step size $\eta > 0$, suppose each player $i \in \mathcal{N}$ independently performs the update

$$\pi_i^{(t+1)} = \mathcal{P}_{\Pi^i}\left(\pi_i^{(t)} - \eta \nabla_{\pi_i} V_{r_i}(\pi^{(t)})\right) \tag{3.1}$$

where $\mathcal{P}_{\Pi^i}(\cdot)$ is the projection onto player $i$'s policy space. Then, due to the separability of the projection operator and the crucial fact that for any $\pi \in \Pi$, $\nabla_{\pi_i} V_{r_i}(\pi) = \nabla_{\pi_i}\Phi(\pi)$, above update is equivalent to running a full PGA step on the potential function, i.e.,

$$\pi^{(t+1)} = \mathcal{P}_{\Pi}\left(\pi^{(t)} - \eta \nabla_{\pi}\Phi(\pi^{(t)})\right).$$

Even though $\Phi$ is not known to the players, together with $\frac{2m\gamma A_{\max}}{(1-\gamma)^3}$-smoothness of $\Phi$ (see Lemma 4.4 of [LOPP22]), we can leverage this equivalence to analyze independent PGA by invoking known results for convergence to stationary points from nonconvex optimization. Using a playerwise version of gradient dominance, [LOPP22] show that if $\pi \in \Pi$ is $\varepsilon$-stationary, i.e., if

$$\max_{(\pi_1+\delta_1,\ldots,\pi_m+\delta_m)\in\Pi, \sum_{i\in\mathcal{N}}\|\delta_i\|^2\leq 1} \sum_{i\in\mathcal{N}} \delta_i^\top \nabla_{\pi_i}\Phi(\pi) \leq \varepsilon,$$

then $\pi$ is a $\frac{\sqrt{S}D\varepsilon}{1-\gamma}$-NE. For completeness, we restate their iteration complexity result for the deterministic case with access to exact value function gradients.

**Theorem 3.1 (Theorem 4.5, [LOPP22])** *Let $\varepsilon > 0$, and suppose the distribution mismatch coefficient $D := \max_{\pi\in\Pi} \|d_\mu^\pi/\mu\|_\infty$ is finite. Then, starting with an arbitrary initial policy $\pi^{(0)} \in \Pi$ and after running $T = \frac{16m\gamma D^2 S A_{\max}\Phi_{\max}}{(1-\gamma)^5\varepsilon^2}$ iterations of independent PGA as in (3.1) with step size $\eta = \frac{(1-\gamma)^3}{2m\gamma A_{\max}}$, there exists $t \in [T]$ such that $\pi^{(t)}$ is an $\varepsilon$-NE.*

In particular, we point out that smaller step sizes are required for MPGs with a larger number of players, introducing an $m$-dependence into the iteration complexity.

## 3.2 A Turn-Based Best-Response Algorithm

Next, we describe a different approach for learning Nash equilibria in MPGs introduced by [SMB22] that uses coordination to let players improve their policies in a turn-based manner. The key observation is that while $\pi_{-i}$ remains fixed for some $i \in \mathcal{N}, \pi_{-i} \in \Pi^{-i}$, the problem for player $i$ to find a best-response policy, i.e. a policy in

$$\mathrm{BR}_i(\pi_{-i}) := \arg\min_{\pi_i'\in\Pi^i} V_{r_i}(\pi_i', \pi_{-i}) \subset \Pi^i,$$

reduces to solving a single-agent MDP. Observe that due to the potential structure, for MPGs, it also holds that $\mathrm{BR}_i(\pi_{-i}) = \arg\min_{\pi_i'\in\Pi^i} \Phi(\pi_i', \pi_{-i})$. Therefore, as long as $\pi^{(t)}$ is not an $\varepsilon$-NE, there exists $i \in \mathcal{N}$ and $\pi_i' \in \mathrm{BR}_i(\pi_{-i}^{(t)})$ such that

$$V_{r_i}(\pi^{(t)}) - V_{r_i}(\pi_i', \pi_{-i}^{(t)}) = \Phi(\pi^{(t)}) - \Phi(\pi_i', \pi_{-i}^{(t)}) > \varepsilon. \tag{3.2}$$

Due to boundedness of $\Phi$, this coordinate descent terminates after at most $T = \mathcal{O}(\Phi_{\max}/\varepsilon)$ update steps, if we can ensure (3.2) to hold for all $0 \leq t \leq T-1$ (see also proof of Theorem 7, [SMB22]). This can be done by finding playerwise best-response policies in a turn-based manner and comparing the respective value function improvements, as detailed by Algorithm 1.

**Algorithm 1** NASH-CA ([SMB22])

1: **initialization:** $\pi^{(0)} \in \Pi$ arbitrary
2: **for** $t = 0, \dots, T-1$ **do**
3: **for** $i \in \mathcal{N}$ **do**
4: find $\tilde{\pi}_i^{(t)} \in \mathrm{BR}_i(\pi_{-i}^{(t)})$ and evaluate $V_{r_i}(\tilde{\pi}_i^{(t)})$
5: $\varepsilon_i^{(t)} = V_{r_i}(\pi^{(t)}) - V_{r_i}(\tilde{\pi}_i^{(t)}, \pi_{-i}^{(t)})$
6: **if** $\max_{i \in \mathcal{N}} \varepsilon_i^{(t)} > \varepsilon$ **then**
7: $\pi^{(t+1)} = (\tilde{\pi}_{i_t}^{(t)}, \pi_{-i_t}^{(t)})$ for some $i_t \in \arg\max_{i \in \mathcal{N}} \varepsilon_i^{(t)}$
8: **else**
9: **return** $\pi^{(t)}$

Line 4 requires solving an MDP. In a stochastic setting, this can be implemented, e.g., via a confidence bound version of value iteration, resulting in an overall sample complexity of $\mathcal{O}(m/\varepsilon^3)$ where $\mathcal{O}(\cdot)$ hides polynomial dependencies in $S, A_{\max}, \Phi_{\max}$, and $\gamma$. If we instead assume access to exact value function evaluation, standard value or policy iteration algorithms as subroutines yield an overall $\mathcal{O}(m/\varepsilon)$ iteration complexity.

## 3.3 Centralized Policy Gradient Ascent

An undesirable property of both independent PGA and NASH-CA is that the iteration complexity scales with the number of players $m$. In this section, we propose a centralized version of a playerwise PGA that does not have any $m$-dependence in the iteration complexity, i.e., that may be more suitable than existing methods for MPGs with a large number of players.

For independent PGA, see Theorem 3.1, the $m$-dependence originates from the smoothness parameter of $\Phi$. We further observe that for any $i \in \mathcal{N}$ and $\pi_{-i} \in \Pi^{-i}$, the function $\Phi(\cdot, \pi_{-i})$ is $\frac{2\gamma A_i}{(1-\gamma)^3}$-smooth. This motivates the use of coordination for selecting only one player per iteration to update its policy. In the context of large-scale optimization, see e.g. [Nes12], similar methods have been proposed under the name coordinate descent to avoid full gradient computations in high-dimensional spaces. The update direction is then usually selected randomly. In our setting, however, partial gradients are computed simultaneously at each player; hence, we can afford to determine an optimal descent direction. Algorithm 2 outlines our approach in detail. Note that to select $i_t$ in Line 6, the $\Delta_i^{(t)}$'s need to be communicated among agents.

**Algorithm 2** Centralized PGA (exact gradients setting)

1: **initialization:** $\pi^{(0)} \in \Pi$ arbitrary and $\eta = \frac{(1-\gamma)^3}{2\gamma A_{\max}}$ for $i \in \mathcal{N}$
2: **for** $t = 0, \dots, T-1$ **do**
3: **for** $i \in \mathcal{N}$ simultaneously **do**
4: $\widetilde{\pi}_i^{(t+1)} = \mathcal{P}_{\Pi^i}\left(\pi_i^{(t)} - \eta \nabla_{\pi_i} V_i(\pi^{(t)})\right)$
5: $\Delta_i^{(t)} = \|\widetilde{\pi}_i^{(t+1)} - \pi_i^{(t)}\|$
6: $\pi^{(t+1)} = (\widetilde{\pi}_{i_t}^{(t+1)}, \pi_{-i_t}^{(t)})$ where $i_t \in \arg\max_{i \in \mathcal{N}} \Delta_i^{(t)}$

Note that unlike in NASH-CA, Algorithm 1, a player's turn consists of doing only a single gradient step instead of finding a best-response policy. Since gradients, unlike best responses, can be determined simultaneously, we save an $m$-factor in the inner loop.

Next, we state and prove our iteration complexity result for centralized PGA as in Algorithm 2.

**Theorem 3.2** *Let $\varepsilon > 0$, suppose the distribution mismatch coefficient $D := \max_{\pi \in \Pi} \|d_\mu^\pi / \mu\|_\infty$ is finite, and choose step size $\eta = \frac{(1-\gamma)^3}{2\gamma A_{\max}}$. Then, starting with an arbitrary initial policy $\pi^{(0)} \in \Pi$, after running centralized PGA as in Algorithm 2 for $T = \frac{16\gamma D^2 S A_{\max} \Phi_{\max}}{(1-\gamma)^5 \varepsilon^2}$ iterations, there exists $t \in [T]$ such that $\pi^{(t)}$ is an $\varepsilon$-NE.*

The proof of Theorem 3.2 is provided in Appendix A. Note that using techniques analogous to, e.g., [LOPP22, DWZJ22, ZRL22] or what we present in Chapter 4, the above insights carry over to the stochastic finite sample setting. We do not elaborate on such results here, as all required ideas are already present in existing literature.

## 3.4 Towards Simultaneous Best-Response Algorithms

### 3.4.1 Challenges and Counterexample

A natural idea towards improving the sample complexity of independent Nash equilibrium learning is to remove coordination from Nash-CA, e.g., by performing simultaneous instead of turn-based best-response updates. Formally, for $\pi \in \Pi$, let $\mathrm{BR}(\pi) := (\mathrm{BR}_1(\pi_{-1}), \ldots, \mathrm{BR}_m(\pi_{-m})) \in \Pi$. Then the simultaneous best-response update is given by choosing $\pi^{(t+1)} \in \mathrm{BR}(\pi^{(t)})$.

However, Proposition 3.3 points out a fundamental problem when naively performing such simultaneous updates: We construct a two-player single-state Markov cooperative game (MCG), i.e., an MPG where $r_1 = \cdots = r_m$, and a policy $\pi \in \Pi$, such that all players $i \in \mathcal{N}$, if $\pi_{-i}$ remains fixed, may improve everyone's value by updating $\pi_i$ to $\mathrm{BR}_i(\pi_{-i})$, but if a simultaneous best-response update is performed, the value worsens.

**Proposition 3.3** *There exists an MCG and a policy $\pi \in \Pi$ such that for two players $i \neq j \in \mathcal{N}$, it holds that $V(BR_i(\pi_i), \pi_{-i}) < V(\pi)$ and $V(BR_j(\pi_j), \pi_{-j}) < V(\pi)$, but $V(BR(\pi)) \geq V(\pi)$, where we may use the notation $V = V_{r_1} = \ldots = V_{r_m}$ due to the MCG property.*

**Proof** Consider a two-player single-state game that, with probability 1, terminates after one step, i.e. $\gamma = 0$. Let $\mathcal{S} = \{s\}$ and $\mathcal{A}_1 = \mathcal{A}_2 = \{x, y\}$. Denote $r := r_1 = r_2$ and choose for $a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2$,

$$r(s, (a_1, a_2)) = \begin{cases} 0 & \text{if } a_1 = a_2 \\ 1 & \text{else.} \end{cases}$$

Let $\pi \in \Pi$ be the policy such that $\pi_1(x \mid s) = 1$, $\pi_2(y \mid s) = 1$, and let $\pi' \in \Pi$ be the policy such that $\pi_1'(y \mid s) = 1$, $\pi_2'(x \mid s) = 1$. Note that $\mathrm{BR}_1(\pi_2) = \pi_1'$ and $\mathrm{BR}_2(\pi_1) = \pi_2'$. Then we have

$$\begin{aligned} &V(\mathrm{BR}_1(\pi_2), \pi_2) = 0 < 1 = V(\pi) \quad \text{and} \\ &V(\pi_1, \mathrm{BR}_2(\pi_1)) = 0 < 1 = V(\pi)\,. \end{aligned}$$

Moreover, it holds that

$$V(\mathrm{BR}(\pi)) = V(\mathrm{BR}_1(\pi_2), \mathrm{BR}_2(\pi_1)) = 1 \geq V(\pi). \qquad \square$$

In the example of Proposition 3.3, one can further observe that $\mathrm{BR}(\mathrm{BR}(\pi)) = \pi$. We conclude that simultaneous best-response updates do not converge for all MCGs (and

hence MPGs) and all initial policies. However, this oscillating behavior observed here for such full simultaneous best-response updates may be due to "overshooting" equilibrium policies. Simultaneously taking small steps towards $\mathrm{BR}(\pi)$ may still yield a converging independent algorithm.

### 3.4.2 Related Ideas from Game Theory and Questions for Future Work

In game theory, it is common to study strategic normal-form games from a dynamical systems perspective by considering the continuous-time dynamics of a game under a certain strategy as a solution to a differential equation, see e.g. [CM14, Hop99, Mat92]. For instance, the best-response dynamics in potential games have been investigated in [SMK18] and is given by the dynamical system

$$\frac{d}{dt}\pi(t) \in \mathrm{BR}(\pi(t)) - \pi(t). \tag{3.3}$$

By definition, Nash equilibria of the game coincide with equilibrium points of these dynamics, i.e., with policies $\pi$ such that $\pi \in \mathrm{BR}(\pi)$.

It is shown in [SMK18] that in potential games, the best-response (BR) dynamics (3.3) converge to a Nash equilibrium for all initial policies (and, in fact, to a pure-strategy NE for almost all initial policies). Moreover, it is proven that the BR dynamics converge at an exponential rate, however, only locally for initial policies in a region around an equilibrium point. Whether similar techniques can be used to show global convergence is an interesting direction for future work. Moreover, this brings up the following questions:

- Can results for continuous time best-response dynamics be discretized, e.g., by using a step size or some form of regularization (see also smoothed fictitious play, [SP19]) to prevent the issue with the naive discrete simultaneous best-response update presented in Proposition 3.3?
- How can agents compute a simultaneous best-response update in an MPG independently, i.e., without taking turns as in Nash-CA? Recent work (see [MWPS22]) makes progress in this direction. However, such current results only show asymptotic convergence.

Chapter 4

# Learning in Constrained Markov Potential Games

In this chapter, we focus on learning in *constrained* Markov Potential Games, where agents, besides reward maximization, have to contend with satisfying global constraints that may depend on the joint behavior of all agents. While we have seen independent learning algorithms for unconstrained MPGs in the previous chapter, for CMPGs, existing algorithms with convergence guarantees require coordination among players. Indeed, inspired by [SMB22], [ARHK23] recently proposed a coordinate ascent algorithm for CMPGs in which each agent updates their policy in turn. At each time step, other agents' policies are fixed while the updating agent faces a constrained Markov Decision Process (CMDP) to solve. When this coordination is not possible, as in the independent learning protocol, the problem becomes more challenging as the environment is no longer stationary from the viewpoint of each agent and the problem does not reduce to solving a CMDP at each time step. This motivates the following question, which we answer affirmatively in this chapter:

> *Can we design an **independent** learning algorithm for **constrained** MPGs with non-asymptotic global convergence guarantees?*

Moreover, we refer the reader to Table 4.1 for a schematic positioning of our work in the recent literature.

| | centralized | independent |
|---|---|---|
| MPG | Nash-CA [SMB22] | independent PGA [LOPP22] [ZRL22] [DWZJ22] |
| CMPG | CA-CMPG [ARHK23] | Algorithm 3 *this work* |

**Table 4.1:** Positioning of our work in the literature; 'centralized' means that the algorithm requires coordination between players who take turns in updating their policy; for 'independent' learning, see Chapter 2.

## 4.1 An Independent Algorithm for Constrained MPGs

This first section presents our independent iProxCMPG algorithm for learning constrained NE in CMPGs.

### 4.1.1 Motivation and Challenges

Before describing our approach, we discuss an alternative, natural but unsuccessful, approach to motivate our algorithm design. This will allow us to highlight the challenges arising from the combination of (a) the presence of coupled constraints, (b) the multi-player setting, and (c) the independent learning protocol.

Our starting point is the known result that any maximizer of the potential function is an NE of the game. This result was initially proved by [MS96] for normal form potential games and later generalized to MPGs by [LOPP22] and to constrained MPGs more recently ([ARHK23]). Therefore, in order to find an (approximate) constrained NE for our CMPG[1], we will consider solving the following constrained optimization problem:

$$\min_{\pi \in \Pi_c} \Phi(\pi)\,, \tag{4.1}$$

where $\Phi$ is the potential function for our CMPG using the notations introduced in Chapter 2. This problem involves a nonconvex objective with a nonconvex constraint since the value function is a nonconvex function of the policy in general (see, e.g., Lemma 1 in [AKLM21]). However, although nonconvex optimization problems with nonconvex constraints are notoriously hard, it turns out that problem (4.1) is still tractable in the single agent setting. In this case, the problem boils down to a CMDP problem. Despite its nonconvexity, the problem can be recast as a linear program in the space of occupancy measures, which is a convex set (see Chapter 3 in [Alt99]). Then, strong duality permits the design of primal-dual policy gradient algorithms to solve the problem with convergence guarantees (see, e.g., [PCCFR19]).

Given those positive results for single-agent CMDPs, a natural approach is to derive a primal-dual algorithm for our multi-agent problem (4.1) as it was proposed by [DDJB20]. In the latter work, a primal-dual policy gradient algorithm was proposed using the Lagrangian function $\mathcal{L}(\pi, \lambda) := \Phi(\pi) + \lambda(V_c(\pi) - \alpha)$ where $\lambda \geq 0$ is a Lagrange multiplier. This algorithm can then be run independently by the different agents using existing independent learning algorithms for the unconstrained setting ([LOPP22, ZRL22, DWZJ22]). Unfortunately, it has been recently shown by [ARHK23] that strong duality does not hold in general for the CMPG problem. Consequently, it is unclear how to obtain guarantees for convergence to constrained NE using this duality approach. This is due to the multi-agent nature of our problem. In particular, since the constraint couples the agents' individual policies, the set of state-action occupancy measures induced by the joint policies of the players cannot be obviously split into several convex problems involving the occupancy measures induced by each of the players' policies. The well-known challenge onon-stationarityity of the environment in multi-agent RL makes the design of independent learning algorithms difficult. As a remedy, [ARHK23] resort to coordination among players and propose a coordinate ascent algorithm for CMDPs. At each time step and for every player $i$, by fixing the policy of other players but player $i$ to $\pi_{-i}$, player $i$ can learn a "best-response" policy by solving a CMDP since the environment now becomes stationary from agent $i$'s viewpoint.

### 4.1.2 Proximal-policy Update with Regularized Constraint

We now describe our approach, which takes a different route. Our algorithm is inspired by recent work in nonconvex optimization under nonconvex constraints ([BDL23, MLY20, JG23]). Following their ideas, we consider the following proximal update with penalized

[1] Approximate KKT points of this problem will be related to approximate constrained NE of our CMPG.

constraints:

$$\pi^{(t+1)} = \arg\min_{\pi \in \Pi} \Big\{ \Phi(\pi) + \frac{1}{2\eta} \left\| \pi - \pi^{(t)} \right\|^2 \Big| \; V_c(\pi) + \frac{1}{2\eta} \left\| \pi - \pi^{(t)} \right\|^2 + \beta \leq \alpha \Big\} \tag{4.2}$$

where $\pi^{(0)}$ is a given initial joint policy, $\eta > 0$ is a step size and $\beta > 0$ an additional slack. Observe that $V_c(\pi^{(t+1)}) + \beta - \alpha \leq -\|\pi^{(t+1)} - \pi^{(t)}\|^2/2\eta$ . Hence, the policy $\pi^{(t)}$ is feasible with slack $\beta$, i.e., $V_c(\pi^{(t)}) + \beta \leq \alpha$, for every $t \geq 0$. We introduce two additional notations for convenience. Define for any joint policies $\pi, \pi' \in \Pi$, and $\eta > 0$,

$$\begin{aligned} \Phi_{\eta,\pi'}(\pi) &:= \Phi(\pi) + \frac{1}{2\eta} \left\| \pi - \pi' \right\|^2, \\ V^c_{\eta,\pi'}(\pi) &:= V_c(\pi) + \frac{1}{2\eta} \left\| \pi - \pi' \right\|^2, \\ \Pi^c_{\eta,\pi'} &:= \left\{ \pi \in \Pi \mid V^c_{\eta,\pi'}(\pi) + \beta \leq \alpha \right\} . \end{aligned}$$

Our update rule in (4.2) can then be rewritten as:

$$\pi^{(t+1)} = \arg\min_{\pi \in \Pi^c_{\eta,\pi^{(t)}}} \Phi_{\eta,\pi^{(t)}}(\pi) . \tag{4.3}$$

We immediately observe that the above update rule is well-defined since $\Phi_{\eta,\pi^{(t)}}$ and $V^c_{\eta,\pi^{(t)}}$ are strongly convex for every $t \geq 0$ for a suitable step size $\eta$. This is in contrast with the original problem where both the potential function $\Phi$ and the constraint function $V_c$ are smooth but nonconvex. We also remark that if $\pi^{(t)}$ converges, then the regularization term $\left\| \pi^{(t+1)} - \pi^{(t)} \right\|$ becomes small and the surrogate feasible region $\Pi^c_{\eta,\pi^{(t)}}$ approaches the original constraint set $\Pi_c$ up to the additional slack $\beta$.

Now, we discuss how to solve the proximal problem in (4.3) defining our main update rule. To solve this strongly convex problem with strongly convex constraint, we adopt a gradient switching algorithm proposed in [LZ20]. At each iteration $k$, our algorithm performs a projected gradient descent step along either the gradient of the (regularized) objective or the gradient of the constraint function depending on whether an estimate of the constraint function satisfies the relaxed constraint $V_c(\pi^{(t,k)}) + \beta - \alpha \leq \delta_k$ where $(\delta_k)$ is a decreasing sequence converging to zero and hence progressively enforcing the constraint. However, it is not immediate from the above procedure how to obtain an independent learning algorithm specifying an update rule for each player without coordination between the players. Recall, for instance that the potential function $\Phi$ is unknown to the players in general, and full gradients of both the potential and constraint functions w.r.t. the joint policy cannot be available to each agent since we exclude coordination and centralization. To obtain our independent iProxCMPG, we propose to use agent-wise updates where each agent runs the gradient switching algorithm independently using only partial gradients of the potential and constraint functions w.r.t. their individual policy. Notice that our subroutine algorithm deviates from the one proposed in [LZ20] in that we use the estimate of the constraint function $V_c$ *instead* of the *regularized* constraint function $V_{\eta,\pi^{(t)}}$. This is because the regularized constraint function involves the *joint policy* in the regularization while the constraint value function can be estimated independently.

### 4.1.3 Full Algorithm and Stochastic Setting

Below, we state the full iProxCMPG algorithm in the exact gradients case. Subsequently, we describe how gradients and constraint function values can be estimated from trajectory sampled in the stochastic setting.

**Algorithm 3** iProxCMPG: **i**ndependent **Prox**imal-policy algorithm for **CMPG**s

1: **initialization:** $\pi^{(0)} \in \Pi^{\xi}$ s.t. $V_c(\pi^{(0)}) < \alpha$ and suitably chosen $\eta, \beta, T, K, \{(\nu_k, \delta_k, \rho_k)\}_{0 \le k \le K}$
2: **for** $t = 0, \dots, T-1$ **do**
3: $\quad \pi_i^{(t,0)} = \pi_i^{(t)}$ for $i \in \mathcal{N}$
4: $\quad$ **for** $k = 0, \dots, K-1$ and $i \in \mathcal{N}$ simultaneously **do**
5: $$\pi_i^{(t,k+1)} = \begin{cases} \mathcal{P}_{\Pi^{i,\xi}}\left[\pi_i^{(t,k)} - \nu_k \hat{\nabla}_{\pi_i} V^{r_i}_{\eta,\pi^{(t)}}(\pi^{(t,k)})\right] & \text{if } \hat{V}_c(\pi^{(t,k)}) + \beta - \alpha \le \delta_k \\ \mathcal{P}_{\Pi^{i,\xi}}\left[\pi_i^{(t,k)} - \nu_k \hat{\nabla}_{\pi_i} V^{c}_{\eta,\pi^{(t)}}(\pi^{(t,k)})\right] & \text{otherwise} \end{cases}$$
6: $\quad \mathcal{B}^{(t)} = \{\lfloor K/2 \rfloor \le k \le K \mid \hat{V}_c(\pi^{(t,k)}) \le \delta_k\}$
7: $\quad \pi_i^{(t+1)} = \pi_i^{(t,\hat{k})}$ where $\hat{k} = \begin{cases} 1 & \text{if } \mathcal{B}^{(t)} = \emptyset \\ \mathbb{P}(\hat{k} = k) = \left(\sum_{k \in \mathcal{B}^{(t)}} \rho_k\right)^{-1} \rho_k & \text{else} \end{cases}$
8: **output:** $\pi_i^{(T)}$ for $i \in \mathcal{N}$

**Remark 4.1** *For our analysis, the index $\hat{k}$ sampled in line 7 of Algorithm 3 is supposed to be picked the same by all the players.*

**Stochastic setting.** When exact gradients and value functions are not available, we estimate them using sampled trajectories. For each joint policy $\pi^{(t,k)}$, every player $i$ samples a trajectory $\tau_i := (s_j^{(t,k)}, a_{i,j}^{(t,k)}, r_{i,j}^{(t,k)}, c_j^{(t,k)})_{0 \le j \le T_e}$ of length $T_e + 1$ by executing their own policy $\pi_i^{(t,k)}$. Here, $s_0^{(t,k)} \sim \mu$ and $r_{i,j}^{(t,k)}, c_j^{(t,k)}$ respectively refer to the reward and cost incurred by the $i$-th player at the $j$-th step. The gradients $\nabla_{\pi_i} V_{r_i}(\pi^{(t,k)})$ and $\nabla_{\pi_i} V_c(\pi^{(t,k)})$ are replaced by their sample estimates

$$\begin{aligned} \hat{\nabla} V^{r_i}_{\pi_i}(\pi^{(t,k)}) &:= R_i^{(T_e,t,k)}\, \psi^{T_e}_{\pi_i^{(t,k)}}, \\ \hat{\nabla} V^{c}_{\pi_i}(\pi^{(t,k)}) &:= C^{(T_e,t,k)}\, \psi^{T_e}_{\pi_i^{(t,k)}}, \end{aligned} \tag{4.4}$$

where $R_i^{(T_e,t,k)} := \sum_{j=0}^{T_e} r_{i,j}^{(t,k)}$, $C^{(T_e,t,k)} := \sum_{j=0}^{T_e} c_j^{(t,k)}$ and

$$\psi^{T_e}_{\pi_i^{(t,k)}} := \sum_{j=0}^{T_e} \nabla_{\pi_i} \log \pi_i^{(t,k)}\left(a_{i,j}^{(t,k)} \mid s_j^{(t,k)}\right).$$

Each agent estimates $V_c(\pi^{(t,k)})$ by $\hat{V}_c(\pi^{(t,k)}) := C^{(T_e,t,k)}$ independently, using the cost feedback information they receive.

## 4.2 Convergence Analysis and Sample Complexity

In this section, we establish the iteration complexity of Algorithm 3 in the deterministic setting before stating its sample complexity in the stochastic setting. We first introduce our assumptions. The first one guarantees the existence of a strictly feasible policy that is available to the agents for initialization.

**Assumption 4.2** *The initial policy $\pi^{(0)}$ satisfies $V_c(\pi^{(0)}) < \alpha$.*

A few remarks are in order regarding this assumption:

- Similar assumptions have been made in the related constrained optimization literature when dealing with nonconvex constraints ([BDL23, MLY20, JG23]). Otherwise, satisfying a constraint may require finding a global minimizer which is computationally intractable in a general nonconvex setting. In our case, this corresponds to finding the global minimizer of a potential function in a fully cooperative unconstrained MPG. While this can be achieved in a single agent setting thanks to the gradient dominance property ([AKLM21, Xia22]), such a global optimality result is not available in the literature for our multi-agent setting to the best of our knowledge.
- While finding a strictly feasible policy is involved in general, it may be possible to find such a policy in some special cases, such as when the state space can be factored, the probability transitions are independent across agents, and the constraint cost functions are separable (see examples 1 and 2 in [ARHK23] for more details).

In addition to initial feasibility, we require that Slater's condition holds for each subproblem given by a proximal-point update. This is ensured by the following uniform Slater's condition.

**Assumption 4.3** *Let $\eta = \frac{1}{2L_\Phi}$ where $L_\Phi$ is the smoothness parameter of $\Phi$. Then, there exists $\zeta > 0$ such that for any strictly feasible $\pi' \in \Pi$, i.e., $V_c(\pi') < \alpha$, there exists $\pi \in \Pi$ with $V^c_{\eta,\pi'}(\pi) \leq \alpha - \zeta$.*

We make the following comments:

- First, we point out that a strictly feasible $\pi'$ satisfies $V^c_{\eta,\pi'}(\pi') = V_c(\pi') < \alpha$, i.e., the existence of a strictly feasible policy for the regularized constraint function $V^c_{\eta,\pi'}$ is trivially given. Assumption 4.3 additionally ensures that strict feasibility holds with slack $\zeta$ where $\zeta$ is independent of $\pi'$.
- Similar constraint qualification conditions have been widely used in the nonconvex constrained optimization literature, see [BDL22], Table 1 for an overview. In particular, Assumption 4.3 is similar to the uniform Slater's condition of [MLY20]. Assumption 3 in [BDL23] is a strong feasibility assumption which implies Assumption 4.3, and hence could also replace it here. Strong feasibility assumes existence of a policy $\pi$ such that $V_c(\pi) \leq \alpha - \frac{\mathrm{diam}(\Pi)^2}{\eta}$ where $\mathrm{diam}(\Pi) := \max_{\pi,\pi' \in \Pi} \|\pi - \pi'\|$.
- A uniform strict feasibility assumption similar to Assumption 4.3 was used for centralized NE-learning, see [ARHK23], Assumption 2.

### 4.2.1 Exact Gradients Case

In the noiseless setting with access to exact gradients, we achieve the following iteration complexity result.

**Theorem 4.4** *Let Assumptions 4.2 and 4.3 hold and let the distribution mismatch coefficient $D := \max_{\pi \in \Pi} \left\| d^\pi_\mu / \mu \right\|_\infty$ be finite. For any $\varepsilon > 0$, after running iProxCMPG, Algorithm 3, for $\xi = 0$, suitably chosen $\eta, \beta, T, K$, and $\{(\nu_k, \delta_k, \rho_k)\}_{0 \leq k \leq K}$, there exists $t \in [T]$, such that $\pi^{(t)}$ is a constrained $\varepsilon$-NE. The total iteration complexity is given by $\mathcal{O}(\varepsilon^{-4})$ where $\mathcal{O}(\cdot)$ hides polynomial dependencies in $m, S, A_{\max}, D, 1-\gamma$, and $\zeta$.*

The full proof of Theorem 4.4 is deferred to Appendix B.2.1. We briefly outline the key steps below.

**Proof (Idea)** First, we show that $K = \mathcal{O}(\varepsilon^{-2})$ iterations of the inner loop yield a policy that is feasible and achieves potential value sufficiently close to the exact proximal update (4.3). For $T = \mathcal{O}(\varepsilon^{-2})$, standard arguments then imply existence of $t \in [T]$ such that $\|\pi^{(t+1)} - \pi^{(t)}\| = \mathcal{O}(\varepsilon)$. It can further be shown that such $\pi^{(t+1)}$ satisfies a particular form of approximate CMPG-specific KKT conditions for the original constrained optimization problem (4.1). We then leverage the multi-agent structure to argue that for all $i \in \mathcal{N}$, similar KKT conditions also hold w.r.t. the playerwise problem $\min_{\pi_i \in \Pi_c^i(\pi_{-i}^{(t+1)})} V_{r_i}(\pi_i, \pi_{-i}^{(t+1)})$ where $\pi_{-i}^{(t+1)}$ is fixed. Finally, using playerwise gradient dominance (see e.g., Lemma D.3 in [LOPP22] or Lemma 2 in [GLMVG22]), one can bound the duality gap of player $i$'s constrained problem for all $i \in \mathcal{N}$ which implies that $\pi^{(t+1)}$ is a constrained $\varepsilon$-NE. The total iteration complexity is given by $T \cdot K = \mathcal{O}(\varepsilon^{-4})$. □

### 4.2.2 Finite Sample Case

In the stochastic setting, when exact gradients are not available, the variance of the stochastic policy gradients in (4.4) can be unbounded if the policies get closer to the boundaries of the simplex (see, e.g., Eq. (13) in [GLMVG22]). Therefore, we consider exploratory $\xi$-greedy policies to address this issue as in prior work ([DFG20, LOPP22, DWZJ22, GLMVG22]). Define for any $\xi \geq 0, i \in \mathcal{N}$ the subset of $\xi$-greedy policies

$$\Pi^{i,\xi} := \{\pi \in \Pi \mid \forall s \in \mathcal{S} : \pi_i(\cdot \mid s) \geq \xi / A_i\} ,$$

which is used in Algorithm 3. We are now ready to state our sample complexity result.

**Theorem 4.5** *Let Assumptions 4.2 and 4.3 hold, and let $D$ (as in Theorem 4.4) be finite. Then, for any $\varepsilon > 0$, after running iProxCMPG based on finite sample estimates (see Algorithm 4) for suitably chosen $\eta, \beta, \xi, T, K, B$, and $\{(\nu_k, \delta_k, \rho_k)\}_{0 \leq k \leq K}$, there exists $t \in [T]$, such that in expectation, $\pi^{(t)}$ is a constrained $\varepsilon$-NE. The total sample complexity is given by $\tilde{\mathcal{O}}(\varepsilon^{-7})$ where $\tilde{\mathcal{O}}(\cdot)$ hides polynomial dependencies in $m, S, A_{\max}, D, 1-\gamma$, and $\zeta$, as well as logarithmic dependencies in $1/\varepsilon$.*

We refer the reader to Appendix B.2.2 for the proof of Theorem 4.5. Below, we briefly explain how we obtain our sample complexity result.

**Proof (Idea)** As in the exact gradients case, we require $T = \mathcal{O}(\varepsilon^{-2})$ iterations of the outer loop. In the stochastic setting, our independent implementation of the CSA algorithm ([LZ20]) still converges at a $\mathcal{O}(1/K)$-rate due to strong convexity, but requires sampling a batch of size $B = \mathcal{O}(\varepsilon^{-2})$ for estimating constraint function values at each iteration. To counteract the variance of $\xi$-greedy gradient estimates (which in our case grows as $\mathcal{O}(\varepsilon^{-1})$), we need to set $K = \mathcal{O}(\varepsilon^{-3})$. All in all, we end up with sample complexity $T \cdot K \cdot B = \mathcal{O}(\varepsilon^{-7})$ for proving existence of $t \in [T]$ such that $\mathbb{E}\left[\left\|\pi^{(t)} - \pi^{(t+1)}\right\|\right] = \mathcal{O}(\varepsilon)$. Using similar arguments as for Theorem 4.4, this implies that $\pi^{(t+1)}$ is a constrained $\varepsilon$-NE in expectation. □

**Remark 4.6** *Comparing our result to the state-of-the-art in the unconstrained case ($\mathcal{O}(\varepsilon^{-5})$, [DWZJ22]), accounting for constraints comes at a cost, increasing the sample complexity by a $\mathcal{O}(\varepsilon^{-2})$-factor. In the centralized setting, a similar gap can be observed between best-known results for unconstrained ($\mathcal{O}(\varepsilon^{-3})$, [SMB22]) vs. constrained ($\mathcal{O}(\varepsilon^{-5})$, [ARHK23]) NE-learning. Whether this $\mathcal{O}(\varepsilon^{-2})$-gap can be narrowed is an interesting open question for both centralized and independent learning.*

## 4.3 Real-World Applications and Simulations

We test our iProxCMPG algorithm in two simple applications that can be modeled as CMPGs and for which coordination among players is unrealistic. Both examples are inspired by unconstrained variants presented in [NLKS22] who study MPGs.

### 4.3.1 Pollution Tax Model

Consider a simple environment with $m$ agents representing, e.g., factories, two states, *pollution-free* and *polluted*, and two actions, *clean* and *dirty*, corresponding to low and high production volume. Starting in the *pollution-free* state, in each round, the environment transitions to the *polluted* state if and only if at least one agent chooses *dirty*. Each agent's reward is the sum of its profit minus a pollution tax. In either state, the profit is $P_c$ when choosing *clean* and $P_d$ when choosing *dirty*. The pollution tax is zero in the *pollution-free*, and $T_p$ in the *polluted* state. As pointed out by [NLKS22], due to rewards being separable in the sense that $r_i(s, a_i, a_{-i}) = r_i'(s) + r_i''(a_i, a_{-i})$ and state transition probabilities being state independent, the pollution tax model satisfies a sufficient condition under which a Markov game is an MPG. For our simulations, we set $P_c = 2, P_d = 4$, and $T_p = 4$. Due to the lack of incentives for agents to cooperate when promoting environmental sustainability, requiring coordination is unrealistic in this example. Moreover, note that the purpose of the pollution tax is to counteract pollution by penalizing *dirty* actions. However, in practice, there may be additional global requirements on the minimum total production volume. To model this as a CMPG, we charge a cost $C$ per agent that chooses *clean* and impose the constraint $V_c(\pi) \leq \alpha_C$ for appropriately chosen $\alpha_C$.

We run iProxCMPG on the resulting $m$-agent CMPG for $m \in \{2, 4, 8\}$ and with $C = 1, \alpha_C = 12$. Fig. 4.1 shows mean and standard deviation (shaded region) across independent runs of per-iteration potential and constraint values. We observe convergence to a constrained NE under which the minimum production requirements are approximately satisfied.

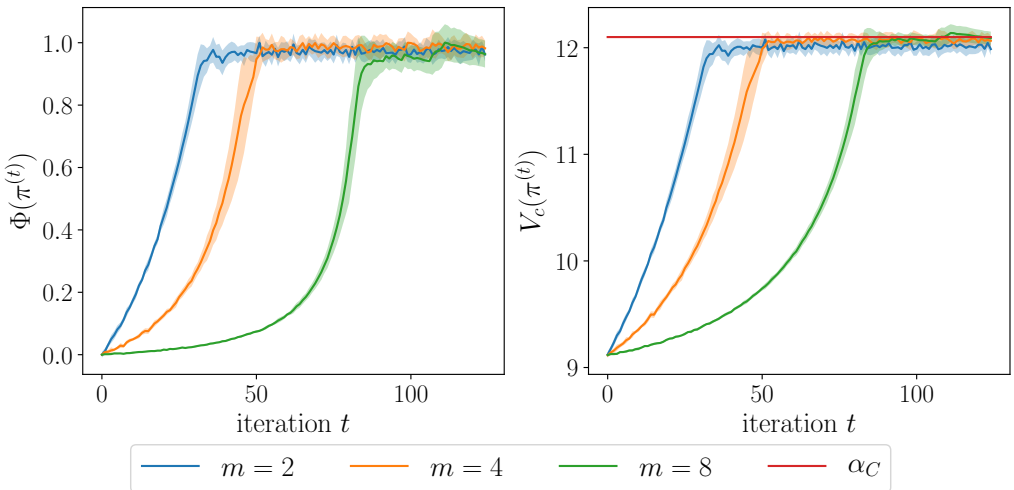


**Figure 4.1:** Potential (left, scaled to $[0, 1]$) and constraint (right) values of iProxCMPG for the $m$-agent pollution tax model.

### 4.3.2 Marketplace for Distributed Energy Resources

As more and more small-scale electricity producers enter the electrical grids, a marketplace emerges. Each participant needs to decide how much energy to sell, given the current supply and demand. The competitive nature of such marketplaces motivates study-

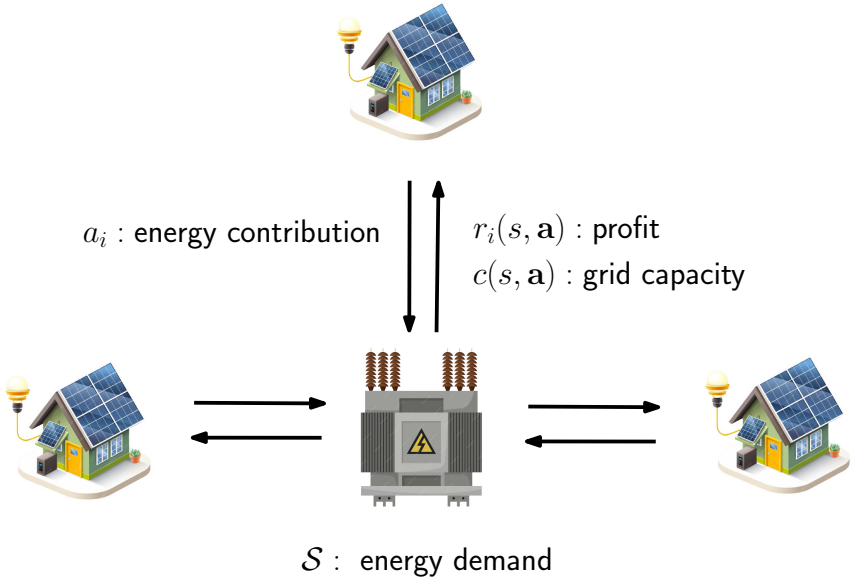


**Figure 4.2:** Schematic illustration of a distributed energy marketplace modeled as a CMPG.

ing the convergence of independent algorithms to NEs under the constraints imposed by market rules. The CMPG we consider has states $\mathcal{S} = \{0, \ldots, S-1\}$ indicating the grid's current energy demand from high at 0 to low at $S-1$. Action $a_i \in \mathcal{A}_i = \{0, \ldots, A_i - 1\}$ represents the units of energy agent $i$ contributes, for which it is rewarded with profit $r_i(s, a_i, a_{-i}) = c_0 a_i^2 - c_1 a_i^2 \sum_{i \in \mathcal{N}} a_i - a_i c_2^s$ where $c_0, c_1, c_2$ are model parameters. State transitions are modeled by first sampling $w \sim \mathcal{U}(\{0, 1, \ldots, W\})$ which models uncertainty due to e.g. weather, and then setting $s' = \max\{0, \min\{S-1, \sum_{i \in \mathcal{N}} a_i - w\}\}$ with probability 0.9 and $s' = w$ otherwise. For our simulations, we set $S = A = W = 5, c_0 = 2, c_1 = 0.25$, and $c_2 = 1.25$. [NLKS22] show that the described game is indeed an MPG with $\Phi(\pi) = \mathbb{E}_{\pi, s_0 \sim \mu}[\sum_{t=0}^{T_e} \varphi_{s_t}(a_t)]$ and $\varphi_s(a_i, a_{-i}) = c_0 \sum_{i \in \mathcal{N}} a_i - c_1 \sum_{i \in \mathcal{N}} a_i^2 - c_1 \sum_{1 \le i < j \le m} a_i a_j - m c_2^s$.

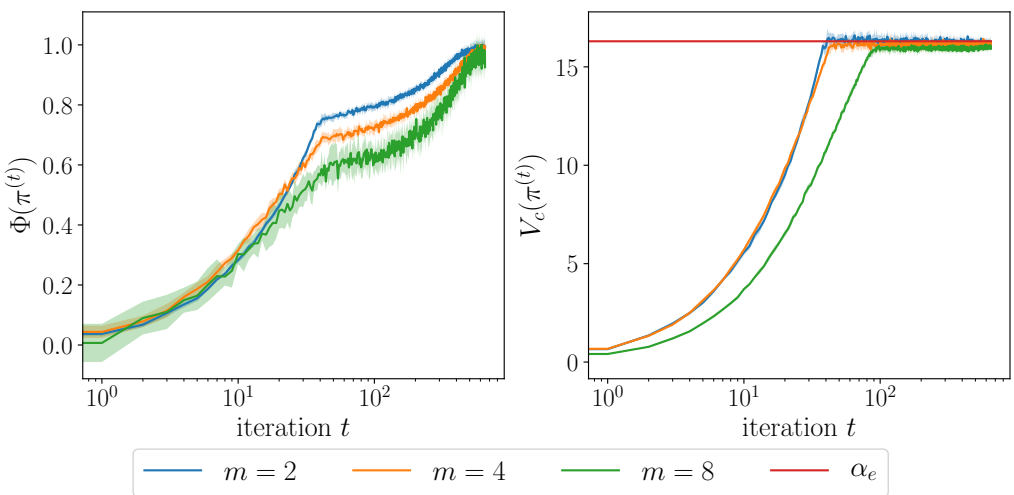


**Figure 4.3:** Potential (left) and constraint (right) values of iProxCMPG for the $m$-agent energy marketplace.

We extend this game into a CMPG by having the system incur a cost per unit of energy provided to the grid, i.e., by defining $c(s, a) = \sum_{i \in \mathcal{N}} a_i$ for all $s \in \mathcal{S}$, and requiring $V_c(\pi) \le \alpha_e$ where we set $\alpha_e = 16$. Fig. 4.3 shows convergence to a constrained NE where players satisfy the energy provision bound on average.

Chapter 5

# Conclusion, Limitations, and Future Work

In this thesis, we studied Nash equilibrium learning in possibly constrained Markov Potential Games with a focus on centralized vs. independent algorithms. In the unconstrained case, we proposed a new centralized policy gradient-based algorithm that leverages ideas from coordinate descent as previously used in large-scale optimization. Our algorithm eliminates the $m$-dependence (number of players) present in the iteration and sample complexity of existing centralized and independent methods and may therefore improve convergence for MPGs with a large number of players. We further give partial negative results for a seemingly promising simultaneous (and thus independent) best-response algorithm. As our main contribution, in the constrained case, we proposed and analyzed the first independent learning algorithm with provable convergence to Nash equilibria and illustrated its practical applicability in two simulations of real-world environments.

Even though learning in MPGs has by now been studied in numerous works, general answers to fundamental questions on the role of centralization vs. independence remain elusive. So far, to the best of our knowledge, no meaningful separation between the two information settings could be shown. While learning in constrained MPGs seemed challenging without centralization, our independent iProxCMPG algorithm eliminates this candidate for showing such separation. On the unconstrained side, however, our improved centralized algorithm emerges as a new candidate for a complexity bound that could potentially only be achieved via coordination.

More concretely, in the unconstrained setting, we list the following directions for future work:

- By introducing coordination into an initially independent playerwise policy gradient method, we decrease its complexity by an $m$-factor. Does this improvement critically depend on centralization? Or is there also an independent NE-learning algorithm for MPGs whose iteration/sample complexity does not depend on the number of players $m$?
- Can the iteration/sample complexity gap between centralized and independent learning ($\mathcal{O}(\varepsilon^{-1})$ vs. $\mathcal{O}(\varepsilon^{-2})$ for centralized, $\mathcal{O}(\varepsilon^{-3})$ vs. $\mathcal{O}(\varepsilon^{-5})$ for independent) be narrowed? In particular, can the $\mathcal{O}(\varepsilon^{-3})$ sample complexity of Nash-CA be matched by independent learning, and are there lower bounds improving over the best known $\mathcal{O}(\varepsilon^{-2})$ (even for the centralized case)?
- Regarding the above two points: Can one prove lower bounds that specifically capture the difficulty of *independent* learning?

Our result on independent learning in constrained MPGs additionally raises the following questions:

- Can the $\mathcal{O}(\varepsilon^{-6})$ sample complexity of iProxCMPG be improved to match (a) the best known unconstrained rate of $\mathcal{O}(\varepsilon^{-4.5})$, or (b) the best known centralized constrained rate of $\mathcal{O}(\varepsilon^{-5})$ (which itself may leave room for improvement)?
- Our algorithm and theoretical guarantees require the agents to run the same algorithm: This may be seen as implicit coordination between agents. Can one design fully independent learning dynamics for our constrained setting, where the players may not even be aware of the existence of other players?[1]
- Can one go beyond the class of CMPGs for learning constrained Nash equilibria?
- Can function approximation be used to scale to large state-action spaces beyond the tabular setting?

[1]Approaches (with asymptotic convergence results) in this direction have been made for unconstrained MPGs in [MWPS22].

# Bibliography


[AKLM21] Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021. 15, 18, 40, 42, 44

[Alt99] Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC press, 1999. 15

[ARHK23] Pragnya Alatur, Giorgia Ramponi, Niao He, and Andreas Krause. Provably learning nash policies in constrained markov potential games. In *Sixteenth European Workshop on Reinforcement Learning*, 2023. 2, 4, 8, 14, 15, 18, 19

[AS00] Eitan Altman and Adam Shwartz. Constrained Markov Games: Nash Equilibria. In Jerzy A. Filar, Vladimir Gaitsgory, and Koichi Mizukami, editors, *Advances in Dynamic Games and Applications*, Annals of the International Society of Dynamic Games, pages 213–221, Boston, MA, 2000. Birkhäuser. 2, 4, 8

[BDL22] Digvijay Boob, Qi Deng, and Guanghui Lan. Level constrained first order methods for function constrained optimization. *arXiv preprint arXiv:2205.08011*, 2022. 18

[BDL23] Digvijay Boob, Qi Deng, and Guanghui Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, 197(1):215–279, 2023. 4, 15, 18, 36, 46

[BR20] Yakov Babichenko and Aviad Rubinstein. Communication complexity of Nash equilibrium in potential games, November 2020. arXiv:2011.06660 [cs]. 4

[Bub15] Sébastien Bubeck. Convex Optimization: Algorithms and Complexity, November 2015. arXiv:1405.4980 [cs, math, stat]. 30

[CCC22] Shicong Cen, Fan Chen, and Yuejie Chi. Independent Natural Policy Gradient Methods for Potential Games: Finite-time Global Convergence with Entropy Regularization, August 2022. arXiv:2204.05466 [cs, math]. 4

[CGT15] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. On the evaluation complexity of constrained nonlinear least-squares and general constrained nonlinear optimization using second-order methods. *SIAM Journal on Numerical Analysis*, 53(2):836–851, 2015. 5

[CM14] Andres Cortés and Sonia Martínez. Self-triggered best-response dynamics for continuous games. *IEEE Transactions on Automatic Control*, 60(4):1115–1120, 2014. 13

[CMS06] George Christodoulou, Vahab S Mirrokni, and Anastasios Sidiropoulos. Convergence and approximation in potential games. In *STACS 2006: 23rd Annual Symposium on Theoretical Aspects of Computer Science, Marseille, France, February 23-25, 2006. Proceedings 23*, pages 349–360. Springer, 2006. 4

[CMZ22] Ziyi Chen, Shaocong Ma, and Yi Zhou. Finding correlated equilibrium of constrained markov game: A primal-dual approach. In *Advances in Neural Information Processing Systems*, 2022. 4

[CZM+23] Zaiwei Chen, Kaiqing Zhang, Eric Mazumdar, Asuman Ozdaglar, and Adam Wierman. A finite-sample analysis of payoff-based independent learning in zero-sum stochastic games. *arXiv preprint arXiv:2303.03100*, 2023. 2, 8

[Das13] Constantinos Daskalakis. On the Complexity of Approximating a Nash Equilibrium. *ACM Transactions on Algorithms*, 9(3):23:1–23:35, June 2013. 8

[DBH+21] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative ai: machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021. 1

[DDJB20] Raghuram Bharadwaj Diddigi, Sai Koti Reddy Danda, Prabuchandran K. J., and Shalabh Bhatnagar. Actor-Critic Algorithms for Constrained Multi-agent Reinforcement Learning, July 2020. arXiv:1905.02907 [cs]. 15

[DFG20] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540, 2020. 2, 7, 8, 19, 45

[DG19] Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019. 4

[DGZ23] Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of markov equilibrium in stochastic games. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4180–4234. PMLR, 2023. 1

[DHB+20] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*, 2020. 1

[DWY+23] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient generalized lagrangian policy optimization for safe multi-agent reinforcement learning. In *Learning for Dynamics and Control Conference*, pages 315–332. PMLR, 2023. 4

[DWZJ22] Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. Independent Policy Gradient for Large-Scale Markov Potential Games: Sharper Rates, Function Approximation, and Game-Agnostic Convergence. In *Proceedings of the 39th International Conference on Machine Learning*, pages 5166–5220. PMLR, June 2022. ISSN: 2640-3498. 1, 2, 3, 8, 9, 12, 14, 15, 19

[EABA+21] Ingy ElSayed-Aly, Suda Bharadwaj, Christopher Amato, Rüdiger Ehlers, Ufuk Topcu, and Lu Feng. Safe multi-agent reinforcement learning via shielding. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 483–491, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems. 4

[FKLS21] Francisco Facchinei, Vyacheslav Kungurtsev, Lorenzo Lampariello, and Gesualdo Scutari. Ghost penalties in nonconvex constrained optimization: Diminishing stepsizes and iteration complexity. *Mathematics of Operations Research*, 46(2):595–627, 2021. 5

[FMOP22] Roy Fox, Stephen M Mcaleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in markov potential games. In *International Conference on Artificial Intelligence and Statistics*, pages 4414–4425. PMLR, 2022. 1

[GGC+23] Shangding Gu, Jakub Grudzien Kuba, Yuanpei Chen, Yali Du, Long Yang, Alois Knoll, and Yaodong Yang. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319:103905, 2023. 2, 4

[GLM+23] Xin Guo, Xinyu Li, Chinmay Maheshwari, Shankar Sastry, and Manxi Wu. Markov *alpha*-potential games: Equilibrium approximation and regret analysis. *arXiv preprint arXiv:2305.12553*, 2023. 4

[GLMVG22] Angeliki Giannou, Kyriakos Lotidis, Panayotis Mertikopoulos, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. On the convergence of policy gradient methods to nash equilibria in general stochastic games. *Advances in Neural Information Processing Systems*, 35:7128–7141, 2022. 7, 19

[Hop99] Ed Hopkins. A note on best response dynamics. *Games and Economic Behavior*, 29(1-2):138–150, 1999. 13

[HS09] Warren Hare and Claudia Sagastizábal. Computing proximal points of nonconvex functions. *Mathematical Programming*, 116(1-2):221–258, 2009. 4

[JG23] Zhichao Jia and Benjamin Grimmer. First-Order Methods for Nonsmooth Nonconvex Functional Constrained Optimization with or without Slater Points, April 2023. arXiv:2212.00927 [math]. 4, 15, 18, 36

[JMS22] Yujia Jin, Vidya Muthukumar, and Aaron Sidford. The complexity of infinite-horizon general-sum stochastic games. *arXiv preprint arXiv:2204.04186*, 2022. 1

[LKT+23] Mushuang Liu, Ilya Kolmanovsky, H Eric Tseng, Suzhou Huang, Dimitar Filev, and Anouck Girard. Potential game-based decision-making for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 2

[LOPP22] Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 8, 9, 10, 12, 14, 15, 19, 31, 34, 44, 45

[LZ20] Guanghui Lan and Zhiqiang Zhou. Algorithms for stochastic optimization with function or expectation constraints. *Computational Optimization and Applications*, 76(2):461–498, 2020. 16, 19, 46, 47, 48, 49

[Mar12] Jason R Marden. State based potential games. *Automatica*, 48(12):3075–3088, 2012. 3

[Mat92] Akihiko Matsui. Best response dynamics and socially stable strategies. *Journal of Economic Theory*, 57(2):343–362, 1992. 13

[MLY20] Runchao Ma, Qihang Lin, and Tianbao Yang. Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints. pages 6554–6564. PMLR, 2020. 4, 15, 18, 38

[MS96] Dov Monderer and Lloyd S Shapley. Potential games. *Games and economic behavior*, 14(1):124–143, 1996. 1, 3, 4, 15

[MWPS22] Chinmay Maheshwari, Manxi Wu, Druv Pai, and Shankar Sastry. Independent and decentralized learning in markov potential games. *arXiv preprint arXiv:2205.14590*, 2022. 1, 2, 4, 8, 13, 23

[MYZB22] Weichao Mao, Lin Yang, Kaiqing Zhang, and Tamer Basar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 15007–15049. PMLR, 2022. 2, 3

[MZZ18] Sergio Valcarcel Macua, Javier Zazo, and Santiago Zazo. Learning parametric closed-loop policies for markov potential games. In *International Conference on Learning Representations*, 2018. 1, 3, 8

[Nes12] Yu. Nesterov. Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems. *SIAM Journal on Optimization*, 22(2):341–362, January 2012. 11

[NLKS22] Dheeraj Narasimha, Kiyeob Lee, Dileep Kalathil, and Srinivas Shakkottai. Multi-agent learning via markov potential games in marketplaces for distributed energy resources. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6350–6357. IEEE, 2022. 4, 20, 21

[NW06] Jorge Nocedal and Stephen J Wright. Quadratic programming. *Numerical optimization*, pages 448–492, 2006. 5

[OSZ21] Asuman Ozdaglar, Muhammed O Sayin, and Kaiqing Zhang. Independent learning in stochastic games. *Invited chapter for the International Congress of Mathematicians 2022 (ICM 2022), arXiv preprint arXiv:2111.11743*, 2021. 2

[PCCFR19] Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32, 2019. 15

[Pol67] Boris Teodorovich Polyak. A general method for solving extremal problems. In *Doklady Akademii Nauk*, volume 174, pages 33–36. Russian Academy of Sciences, 1967. 46

[Rub16] Aviad Rubinstein. Settling the complexity of computing approximate two-player Nash equilibria, August 2016. arXiv:1606.04550 [cs]. 8

[Sha53] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953. 1

[SMB22] Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? In *International Conference on Learning Representations*, 2022. 1, 2, 4, 8, 9, 10, 11, 14, 19

[SMK18] Brian Swenson, Ryan Murray, and Soummya Kar. On Best-Response Dynamics in Potential Games, February 2018. arXiv:1707.06465 [cs, math]. 4, 13

[SP19] Brian Swenson and H. Vincent Poor. Smooth Fictitious Play in $n \times 2$ Potential Games, November 2019. arXiv:1912.00251 [cs]. 13

[SSSS16] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016. 2

[SZL+21] Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Basar, and Asuman Ozdaglar. Decentralized q-learning in zero-sum markov games. *Advances in Neural Information Processing Systems*, 34:18320–18334, 2021. 2

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018. 46

[WMY17] Xiao Wang, Shiqian Ma, and Ya-xiang Yuan. Penalty methods with stochastic approximation for stochastic nonlinear programming. *Mathematics of computation*, 86(306):1793–1820, 2017. 5

[Xia22] Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022. 18

[YD22] Zhiyuan Yao and Zihan Ding. Learning distributed and fair policies for network load balancing as markov potential game. *Advances in Neural Information Processing Systems*, 35:28815–28828, 2022. 2

[ZCLW23] Zhaoyi Zhou, Zaiwei Chen, Yiheng Lin, and Adam Wierman. Convergence rates for localized actor-critic in networked Markov potential games. In Robin J. Evans and Ilya Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2563–2573. PMLR, 31 Jul–04 Aug 2023. 1, 4

[Zhu08] Quanyan Zhu. A Lagrangian approach to constrained potential games: Theory and examples. In *47th IEEE Conference on Decision and Control*, pages 2420–2425, December 2008. ISSN: 0191-2216. 4

[ZMD+22] Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. On the global convergence rates of decentralized softmax gradient play in markov potential games. *Advances in Neural Information Processing Systems*, 35:1923–1935, 2022. 1, 2, 3, 44

[ZRL22] Runyu (Cathy) Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: Stationary points and local geometry. *IFAC-PapersOnLine*, 55(30):73–78, 2022. 25th International Symposium on Mathematical Theory of Networks and Systems MTNS 2022. 1, 2, 3, 9, 12, 14, 15

[ZYB21] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021. 8

Appendix A

# Proofs for Chapter 3

In this part of the appendix, we provide the proof of our centralized policy gradient algorithm presented in Chapter 3. First, we restate Theorem 3.2.

**Theorem 3.2** Let $\varepsilon > 0$, suppose the distribution mismatch coefficient $D := \max_{\pi \in \Pi} \|d_\mu^\pi / \mu\|_\infty$ is finite, and choose step size $\eta = \frac{(1-\gamma)^3}{2\gamma A_{\max}}$. Then, starting with an arbitrary initial policy $\pi^{(0)} \in \Pi$, after running coordinated PGA as in Algorithm 2 for $T = \frac{16\gamma D^2 S A_{\max} \Phi_{\max}}{(1-\gamma)^5 \varepsilon^2}$ iterations, there exists $t \in [T]$ such that $\pi^{(t)}$ is an $\varepsilon$-NE. □

**Proof** First, it follows from Lemma B.9 that the function $\Phi(\cdot, \pi_{-i})$ is $\frac{2\gamma A_{\max}}{(1-\gamma)^3}$-smooth for any $i \in \mathcal{N}$ and $\pi_{-i} \in \Pi_{-i}$. Using the definition of $i_t$ in Algorithm 2, and a standard descent lemma (Lemma 3.6, [Bub15]) together with our choice of $\eta$, for any $t \geq 0$, it holds that

$$\begin{aligned}
\Phi\left(\pi^{(t)}\right) - \Phi\left(\pi^{(t+1)}\right) &= \Phi\left(\pi_{i_t}^{(t)}, \pi_{-i_t}^{(t)}\right) - \Phi\left(\pi_{i_t}^{(t+1)}, \pi_{-i_t}^{(t+1)}\right) \\
&\geq \frac{(1-\gamma)^3}{4\gamma A_{\max}} \left\|\pi_{i_t}^{(t+1)} - \pi_{i_t}^{(t)}\right\|^2 \\
&= \frac{(1-\gamma)^3}{4\gamma A_{\max}} \left\|\pi^{(t+1)} - \pi^{(t)}\right\|^2
\end{aligned}$$

where we point out that $\pi_{-i_t}^{(t)} = \pi_{-i_t}^{(t+1)}$. The last equality is due to the fact that only the $i_t$-th player's policy is updated in iteration $t$. Summing this inequality over all iterations yields

$$\frac{(1-\gamma)^3}{4\gamma A_{\max}} \sum_{t=0}^{T-1} \left\|\pi^{(t+1)} - \pi^{(t)}\right\|^2 \leq \Phi\left(\pi^{(0)}\right) - \Phi\left(\pi^{(T)}\right) \leq \Phi_{\max}.$$

which implies that

$$\frac{1}{T} \sum_{t=0}^{T-1} \left\|\pi^{(t+1)} - \pi^{(t)}\right\|^2 \leq \frac{4\Phi_{\max}\gamma A_{\max}}{(1-\gamma)^3 T}.$$

Therefore, with our choice of $T = \frac{16\gamma D^2 S A_{\max} \Phi_{\max}}{(1-\gamma)^5 \varepsilon^2}$, there exists $0 \leq t \leq T-1$ such that $\left\|\pi^{(t+1)} - \pi^{(t)}\right\| \leq \frac{\varepsilon(1-\gamma)}{2D\sqrt{S}}$. Moreover, due to our choice of $i_t$, we know that for all $j \in \mathcal{N} \setminus \{i_t\}$, it holds that $\left\|\tilde{\pi}_j^{(t+1)} - \pi_j^{(t)}\right\| \leq \left\|\pi_{i_t}^{(t+1)} - \pi_{i_t}^{(t)}\right\| \leq \frac{\varepsilon(1-\gamma)}{2D\sqrt{S}}$. Using Lemma D.2

of [LOPP22], it follows that for all $i \in \mathcal{N}$, $\pi^{(t+1)}$ is a $\frac{\varepsilon(1-\gamma)}{2D\sqrt{S}}$-stationary point for $V_{r_i}$, i.e., that for all $i \in \mathcal{N}$,

$$\max_{\pi_i' \in \Pi^i} \left\langle \pi_i^{(t+1)} - \pi_i', \nabla_{\pi_i} V_{r_i}(\pi^{(t+1)}) \right\rangle \leq \frac{\varepsilon(1-\gamma)}{2D}.$$

Finally, suppose player $i \in \mathcal{N}$ deviates to some $\pi_i^*$ and let $\pi^* = (\pi_i^*, \pi_{-i}^{(t+1)})$. Then, applying the definition of the potential function and playerwise gradient dominance (see [LOPP22], Lemma 4.3), we get

$$\begin{aligned} V_{r_i}(\pi^{(t+1)}) - V_{r_i}(\pi^*) &= \Phi(\pi^{(t+1)}) - \Phi(\pi^*) \\ &\leq \frac{1}{1-\gamma} \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty \max_{\pi' = (\pi_i', \pi_{-i}^{(t+1)})} \left\langle \pi^{(t+1)} - \pi', \nabla_{\pi_i} \Phi(\pi^{(t+1)}) \right\rangle \\ &\leq \varepsilon \end{aligned}$$

which completes the proof. □

Appendix B

# Proofs and Details for Chapter 4

## B.1 iProxCMPG: Full Stochastic Algorithm

In this section, for the convenience of the reader, we report the full pseudo-code of Algorithm 3 in the stochastic setting where exact gradients are not available. See Algorithm 4.

**Algorithm 4** iProxCMPG: independent proximal-policy algorithm for CMPGs

1: **initialization:** $\pi^{(0)} \in \Pi^{\xi}$ s.t. $V_c(\pi^{(0)}) < \alpha$ and suitably chosen $\eta, \beta, \xi, T, K, \{(\nu_k, \delta_k, \rho_k)\}_{0 \le k \le K}$
2: **for** $t = 0, \dots, T-1$ **do**
3: $\quad \pi_i^{(t,0)} = \pi_i^{(t)}$
4: $\quad$ **for** $k = 0, \dots, K-1$ and $i \in \mathcal{N}$ simultaneously **do**
5: $\quad\quad$ sample $B$ trajectories $\{\{(a_{i,j}^{(b)}, s_j^{(b)}, r_{i,j}^{(b)}, c_j^{(b)})\}_{j=0}^{\hat{T}_e^{(b)}}\}_{b=1}^{B}$ by following $\pi_i^{(t,k)}$
6: $\quad\quad$ set $\hat{V}_{r_i}(\pi^{(t,k)}) = \frac{1}{B}\sum_{b=1}^{B}\sum_{j=0}^{\hat{T}_e^{(b)}} r_{i,j}^{(b)}$ and $\hat{V}_c(\pi^{(t,k)}) = \frac{1}{B}\sum_{b=1}^{B}\sum_{j=0}^{\hat{T}_e^{(b)}} c_j^{(b)}$
7: $\quad\quad \hat{\nabla} V_{\pi_i}^{r_i}(\pi^{(t,k)}) = \hat{V}_{r_i}(\pi^{(t,k)}) \cdot \frac{1}{B}\sum_{b=1}^{B}\sum_{j=1}^{\hat{T}_e^{(b)}} \nabla \log \pi_i(a_{i,j}^{(b)} \mid s_j^{(b)})$
8: $\quad\quad \hat{\nabla} V_{\pi_i}^{c}(\pi^{(t,k)}) = \hat{V}_{c}(\pi^{(t,k)}) \cdot \frac{1}{B}\sum_{b=1}^{B}\sum_{j=1}^{\hat{T}_e^{(b)}} \nabla \log \pi_i(a_{i,j}^{(b)} \mid s_j^{(b)})$
9: $\quad\quad \pi_i^{(t,k+1)} = \begin{cases} \mathcal{P}_{\Pi^{i,\xi}}\left[\pi_i^{(t,k)} - \nu_k \hat{\nabla}_{\pi_i} V_{r_i}(\pi^{(t,k)}) - \frac{\nu_k}{\eta}(\pi_i^{(t,k)} - \pi_i^{(t)})\right] & \text{if } \hat{V}_c(\pi^{(t,k)}) + \beta - \alpha \le \delta_k \\ \mathcal{P}_{\Pi^{i,\xi}}\left[\pi_i^{(t,k)} - \nu_k \hat{\nabla}_{\pi_i} V_{c}(\pi^{(t,k)}) - \frac{\nu_k}{\eta}(\pi_i^{(t,k)} - \pi_i^{(t)})\right] & \text{otherwise} \end{cases}$
10: $\quad \mathcal{B}^{(t)} = \{\lfloor K/2 \rfloor \le k \le K \mid \hat{V}_c(\pi^{(t,k)}) \le \delta_k\}$
11: $\quad \pi_i^{(t+1)} = \pi_i^{(t,\hat{k})}$ where $\hat{k} = 1$ if $\mathcal{B}^{(t)} = \emptyset$ and else sampled s.t. for $k \in \mathcal{B}^{(t)}$, $\mathbb{P}(\hat{k} = k) = \left(\sum_{k \in \mathcal{B}^{(t)}} \rho_k\right)^{-1} \rho_k$
12: **output:** $\pi_i^{(T)}$ for $i \in \mathcal{N}$

**Remark B.1** *For our analysis, the index $\hat{k}$ sampled in line 11 of Algorithm 4 is supposed to be picked the same by all the players.*

## B.2 Proofs for Section 4.2

**Notation.** For any integer $n \ge 1$, we use the notation $[n] := \{1, \dots, n\}$ throughout the proofs.

In this section, we provide complete proofs of our main results. We begin with the exact gradients case before addressing the more involved finite sample case.

### B.2.1 Proof of Theorem 4.4 – Exact Gradients Case

First, we restate Theorem 4.4.

**Theorem 4.4** Let Assumptions 4.2 and 4.3 hold and let the distribution mismatch coefficient $D := \max_{\pi\in\Pi} \left\| d_\mu^\pi / \mu \right\|_\infty$ be finite. For any $\varepsilon > 0$, after running iProxCMPG, Algorithm 3, with $\xi = 0$, suitably chosen $\eta, \beta, T, K$, and $\{(\nu_k, \delta_k, \rho_k)\}_{0\le k\le K}$, there exists $t \in [T]$, such that $\pi^{(t)}$ is a constrained $\varepsilon$-NE in expectation[1]. The total iteration complexity is given by $\mathcal{O}\left(\varepsilon^{-4}\right)$ where $\mathcal{O}(\cdot)$ hides polynomial dependencies in $m, S, A_{\max}, D, 1-\gamma$, and $\zeta$. □

Before analyzing the outer loop of Algorithm 3, we begin by focusing on the proximal-point update step. We first introduce some useful notation. Then, we explain how we can use the switching gradient algorithm in Appendix B.3 for approximately solving the proximal-point update step independently. We proceed by establishing guarantees that will be important in the analysis of the outer loop of Algorithm 3.

**Notation.** Recall that for any policies $\pi, \pi' \in \Pi$ and any $\eta > 0$,

$$\Phi_{\eta,\pi'}(\pi) = \Phi(\pi) + \frac{1}{2\eta}\left\|\pi - \pi'\right\|^2,$$
$$V^c_{\eta,\pi'}(\pi) = V_c(\pi) + \frac{1}{2\eta}\left\|\pi - \pi'\right\|^2$$
$$\Pi^c_{\eta,\pi'} = \left\{\pi \in \Pi \mid V^c_{\eta,\pi'}(\pi) \le \alpha - \beta\right\}.$$

Moreover, recall the following constrained optimization problem:

$$\min_{\pi\in\Pi^c_{\eta,\pi'}} \Phi_{\eta,\pi'}(\pi). \tag{ProxPb($\eta,\pi'$)}$$

In the following "$\lesssim$" denotes inequality up to numerical constants. Moreover, let $L_\Phi$ be the smoothness constant of the functions $\Phi$ and $V_c$ (see Lemma B.9) and let $\Phi_{\max}$ be an upper bound[2] on $\Phi$. Recall that under Assumption 4.2, the initial policy $\pi^{(0)}$ is strictly feasible. We denote the respective slack by $\bar{\zeta}_0 > 0$, i.e., $\bar{\zeta}_0 := \alpha - V_c(\pi^{(0)})$.

Next, we state and prove the guarantees provided by our proximal-point update subroutine.

**Lemma B.2** *Let Assumption 4.2 hold and let $0 < \bar{\varepsilon} \le \bar{\zeta}_0$. Set $\beta = \bar{\varepsilon}$, $\eta = \frac{1}{2L_\Phi}$, and $\xi = 0$. Denote by $\tilde{\pi}^{(t+1)}$ the unique optimal solution to (ProxPb($\eta, \pi^{(t)}$)). There exist $K = \mathcal{O}\left(\bar{\varepsilon}^{-2}\right)$ and suitable choices of $\{(\nu_k, \delta_k, \rho_k)\}_{0\le k\le K}$, such that lines 4-6 of Algorithm 3 guarantee that for any $t \in [T-1]$,*

$$\begin{aligned}\mathbb{E}\left[\Phi_{\eta,\pi^{(t)}}(\pi^{(t+1)}) - \Phi_{\eta,\pi^{(t)}}(\tilde{\pi}^{(t+1)})\right] &\le \bar{\varepsilon}^2,\\ \mathbb{E}\left[V_c(\pi^{(t+1)})\right] &\le \alpha,\end{aligned} \tag{B.1}$$

*where the expectation is with respect to the randomness induced by the sampling of $\hat{k}$ in line 11 of Algorithm 4.*

[1]Notice that here we take the expectation w.r.t. the randomness which is induced by the sampling of $\hat{k}$ in line 11 of Algorithm 4.

[2]Such a bound is always trivially available.

**Proof** We divide the proof into two steps.

- **Step 1: Equivalent centralized update rule for our algorithm.** First, we argue that independently running the subroutine given by the inner loop of Algorithm 3, i.e., lines 4-6, is equivalent to a centralized execution of the stochastic switching subgradient algorithm (see Algorithm 5) applied to our proximal-point update problem. Crucially, as observed by [LOPP22], Proposition B.1, for any $i \in \mathcal{N}$ and $\pi \in \Pi$, it holds that $\nabla_{\pi_i}\Phi(\pi) = \nabla_{\pi_i}V_{r_i}(\pi)$. We can extend this observation to our regularized potential and value functions, namely for any $\pi' \in \Pi$,

$$\begin{aligned}\nabla_{\pi_i}\Phi_{\eta,\pi'}(\pi) &= \nabla_{\pi_i}\Phi(\pi) + \frac{1}{\eta}\left(\pi_i - \pi_i'\right)\\ &= \nabla_{\pi_i}V_{r_i}(\pi) + \frac{1}{\eta}\left(\pi_i - \pi_i'\right),\end{aligned}$$

  which is an expression that can be evaluated independently by player $i$, since access to the joint policy $\pi$ is not required. Together with separability of the projection operator $\mathcal{P}_{\Pi^\xi}$, see e.g. [LOPP22], Lemma D.1, we have

$$\left(\mathcal{P}_{\Pi^{i,\xi}}\left[\pi_i^{(t,k)} - \nu_k\nabla_{\pi_i}V^{r_i}_{\eta,\pi^{(t,k)}}(\pi^{(t,k)})\right]\right)_{i\in\mathcal{N}} = \mathcal{P}_{\Pi^\xi}\left[\pi^{(t,k)} - \nu_k\nabla_\pi\Phi_{\eta,\pi^{(t,k)}}(\pi^{(t,k)})\right],$$

  and similarly, for the constraint value function,

$$\left(\mathcal{P}_{\Pi^{i,\xi}}\left[\pi_i^{(t,k)} - \nu_k\nabla_{\pi_i}V^{c}_{\eta,\pi^{(t,k)}}(\pi^{(t,k)})\right]\right)_{i\in\mathcal{N}} = \mathcal{P}_{\Pi^\xi}\left[\pi^{(t,k)} - \nu_k\nabla_\pi V^{c}_{\eta,\pi^{(t,k)}}(\pi^{(t,k)})\right].$$

  Moreover, since $V_c(\pi^{(t,k)})$ can be estimated equally by each player due to the cooperative nature of our constraint, we can conclude that Algorithm 3 is equivalent to a centralized version where the independent, simultaneous update in line 5 is replaced by the following centralized version:

$$\pi^{(t,k+1)} = \begin{cases}\mathcal{P}_{\Pi^\xi}\left[\pi^{(t,k)} - \nu_k\hat{\nabla}_\pi\Phi_{\eta,\pi^{(t,k)}}(\pi^{(t,k)})\right] & \text{if } \hat{V}_c(\pi^{(t,k)}) + \beta - \alpha \le \delta_k,\\ \mathcal{P}_{\Pi^\xi}\left[\pi^{(t,k)} - \nu_k\hat{\nabla}_\pi V^{c}_{\eta,\pi^{(t,k)}}(\pi^{(t,k)})\right] & \text{otherwise.}\end{cases}$$

- **Step 2: Induction on $t$.** Next, to prove the claimed guarantee for all $t \in [T-1]$, we proceed by induction on $t$. We will invoke results on the stochastic switching gradient algorithm (see CSA, Algorithm 5) that are separately presented in Appendix B.3 in the context of constrained optimization. By Assumption 4.2, since $\bar{\varepsilon} \le \bar{\zeta}_0$ and $\beta = \bar{\varepsilon}$, we have $V_c(\pi^{(0)}) \le \alpha - \beta$. That is, for $t = 0$, the initial feasibility condition of our CSA result, Theorem B.18 in Appendix B.3, holds for $\pi^{(t)}$. Note further that in our deterministic case, Assumption B.14 (which is required for Theorem B.18) holds, since by Lemma B.9 we have a bound on objective and constraint gradient norms.

  Hence, we can apply Theorem B.18 in the deterministic setting, i.e., with batch size $J = 1$ and access to exact gradients and constraint function values, to $\Phi_{\eta,\pi^{(t)}}$ and $V^c_{\eta,\pi^{(t)}}$ with $\mu = L_\Phi$ and

$$M^2 \lesssim \max\left\{M_G^2 + \mu_G^2\Delta^4, M_F^2 + \mu_F^2\Delta^4\right\} \lesssim M_c^2 + L_\Phi^2\mathrm{diam}(\Pi)^4,$$

  in the notation of Theorem B.18. After plugging in the bounds on $M_c, L_\Phi$, and $\mathrm{diam}(\Pi)$ from Lemma B.9, and choosing $K$ as in the statement of this lemma,

Theorem B.18 implies the desired bounds on constraint violation and optimality gap w.r.t. $\tilde{\pi}^{(t+1)}$ in (B.1). This concludes the base case of the induction.

As induction hypothesis, suppose now that (B.1) holds for some $t \in [T-1]$. Then, due to $\beta \geq \bar{\varepsilon}$, $V_c(\pi^{(t+1)}) + \beta \leq \alpha + \bar{\varepsilon}$ implies that the initial feasibility condition of Theorem B.18 is satisfied and hence with the same argument as above regarding Assumption B.14, we can apply Theorem B.18 to conclude that at the end of iteration $t+1$ of Algorithm 3, the inner loop guarantees that

$$\mathbb{E}\left[\Phi_{\eta,\pi^{(t+2)}}(\pi^{(t+2)}) - \Phi_{\eta,\pi^{(t+2)}}(\tilde{\pi}^{(t+2)})\right] \leq \bar{\varepsilon}^2,$$
$$\mathbb{E}\left[V_c(\pi^{(t+2)})\right] \leq \alpha,$$

i.e., the inductive hypothesis also holds for $t+1$. □

We next determine the number of iterations of the outer loop of Algorithm 3 required for convergence in the following sense.

**Lemma B.3** *Let $\varepsilon > 0$ and set $\eta = \frac{1}{2L_\Phi}$. Suppose $K$ is chosen such that the guarantee from Lemma B.2 holds for $\bar{\varepsilon}^2 = \frac{\varepsilon^2}{4\eta}$. Then, after $T = \frac{4\eta\Phi_{\max}}{\varepsilon^2}$ iterations of the outer loop of Algorithm 3 where $\Phi_{\max}$ is an upper bound of the potential function (i.e., $\forall \pi \in \Pi, \Phi(\pi) \leq \Phi_{\max}$), there exists $0 \leq t \leq T-1$ such that $\|\pi^{(t+1)} - \pi^{(t)}\| \leq \varepsilon$.*

**Proof** Let $\mathcal{F}_t$ denote the $\sigma$-field generated by the random variables given by the iterates $\pi^{(t)}$ up to iteration $t$. Notice that this randomness is induced by the sampling of $\hat{k}$ in line 11 of Algorithm 4. By Lemma B.2, the inner loop of Algorithm 3 guarantees that for any $0 \leq t \leq T-1$,

$$\begin{aligned}
\mathbb{E}\left[\Phi(\pi^{(t+1)}) + \frac{1}{2\eta}\|\pi^{(t+1)} - \pi^{(t)}\|^2 \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\Phi_{\eta,\pi^{(t)}}(\pi^{(t+1)}) \mid \mathcal{F}_t\right] \\
&\leq \mathbb{E}\left[\Phi_{\eta,\pi^{(t)}}(\tilde{\pi}^{(t+1)}) \mid \mathcal{F}_t\right] + \bar{\varepsilon}^2 \\
&\leq \Phi_{\eta,\pi^{(t)}}(\pi^{(t)}) + \bar{\varepsilon}^2 \\
&= \Phi(\pi^{(t)}) + \bar{\varepsilon}^2
\end{aligned}$$

where the second inequality is due to $\mathbb{E}\left[V^c_{\eta,\pi^{(t)}}(\pi^{(t)}) \mid \mathcal{F}_t\right] = \mathbb{E}\left[V_c(\pi^{(t)})\right] \leq \alpha$. Taking total expectation in the above inequality, we obtain

$$\mathbb{E}\left[\|\pi^{(t+1)} - \pi^{(t)}\|^2\right] \leq 2\eta\left(\mathbb{E}\left[\Phi(\pi^{(t)})\right] - \mathbb{E}\left[\Phi(\pi^{(t+1)})\right]\right).$$

Summing the above inequality over $0 \leq t \leq T-1$, using the upper bound $\Phi_{\max}$ on the potential function and plugging in our choices of $\eta$, $T$, and $\bar{\varepsilon}$, we obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\pi^{(t+1)} - \pi^{(t)}\|^2\right] &\leq 2\eta\left(\bar{\varepsilon}^2 + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\Phi(\pi^{(t)})\right] - \mathbb{E}\left[\Phi(\pi^{(t+1)})\right]\right) \\
&\leq \frac{2\eta\Phi_{\max}}{T} + 2\eta\bar{\varepsilon}^2 \\
&\leq \varepsilon^2.
\end{aligned}$$

Using Jensen's inequality, we conclude that there exists $t \in [T-1]$ such that

$$\mathbb{E}\left[\|\pi^{(t+1)} - \pi^{(t)}\|\right] \leq \varepsilon.$$ □

Next, we aim to prove that the event $\|\pi^{(t+1)} - \pi^{(t)}\| \leq \varepsilon$ implies Nash-gap$(\pi^{(t+1)}) = \mathcal{O}(\varepsilon)$ where the constrained Nash-gap is defined as

$$\text{Nash-gap}(\pi^*) := \max_{i \in \mathcal{N}} \max_{\pi_i' \in \Pi_c^i(\pi_{-i}^*)} V_{r_i}(\pi^*) - V_{r_i}(\pi_i', \pi_{-i}^*) \,. \tag{B.2}$$

We can then argue that $\mathbb{E}[\|\pi^{(t+1)} - \pi^{(t)}\|] \leq \varepsilon$ implies $\mathbb{E}[\text{Nash-gap}(\pi^{(t+1)})] = \mathcal{O}(\varepsilon)$, i.e., that the policy $\pi^{(t+1)}$ is a constrained $\mathcal{O}(\varepsilon)$-NE in expectation.

Towards this goal, we first show that a policy $\pi^{(t+1)}$ satisfying $\|\pi^{(t+1)} - \pi^{(t)}\| = \mathcal{O}(\varepsilon)$ (as in the previous lemma) is a $\mathcal{O}(\varepsilon)$-$\widetilde{\text{KKT}}$ policy for our initial constrained minimization problem. The $\varepsilon$-$\widetilde{\text{KKT}}$ conditions are a slight modification of the standard $\varepsilon$-KKT conditions adapted to our specific requirements (see Definition B.19 and Definition B.20 in Appendix B.4). In the following lemma, we will be referring to (in-)exact solutions as well as KKT and $\widetilde{\text{KKT}}$ conditions for different problems. Therefore, we first introduce additional useful notation for clarity.

**Notation.** We refer to the following constrained optimization problem as (InitPb):

$$\min_{\pi \in \Pi} \Phi(\pi) \,. \tag{InitPb}$$

For the previously introduced (ProxPb$(\eta, \pi^{(t)})$), we distinguish between the inexact solution resulting from the update which we denote by $\pi^{(t+1)}$, and the exact solution which will be denoted by $\tilde{\pi}^{(t+1)}$ in the proof below. Furthermore, we define the Lagrangians for the two problems as

$$\mathcal{L}(\pi, \lambda) = \Phi(\pi) + \lambda \left(V_c(\pi) - \alpha\right) \,, \tag{InitPb–$\mathcal{L}$}$$

$$\mathcal{L}_{\eta,\pi'}(\pi, \lambda) = \Phi_{\eta,\pi'}(\pi) + \lambda \left(V_{\eta,\pi'}^c(\pi) - \alpha + \beta\right) \,. \tag{ProxPb$(\eta, \pi')$–$\mathcal{L}$}$$

Using Lemma B.2 and Lemma B.3, the following lemma shows that Algorithm 3 is guaranteed to generate an $\mathcal{O}(\varepsilon)$-$\widetilde{\text{KKT}}$ policy. Parts of the proof have appeared in a similar form in the optimization literature (see Lemma 3.5 and Theorem 3.2 in [JG23], and Theorem 5 in [BDL23]). The lemma below differs from these results since we are in a smooth setting and prove convergence w.r.t. our notion of $\widetilde{\text{KKT}}$ conditions rather than towards a point that is near an $\varepsilon$-KKT point. Moreover, our guarantee for the proximal update subroutine is somewhat weaker due to the relaxed constraint satisfaction condition that we use to switch between update types in the inner loop, see Lemma B.2. Additionally, in order to achieve exact primal feasibility (instead of $\varepsilon$-approximate), we employ a feasibility margin $\beta$.

**Lemma B.4** *Let Assumptions 4.2 and 4.3 hold. Let $\varepsilon > 0$ and choose $\bar{\varepsilon}, K, \beta$ as in Lemma B.2 and B.3. If $\pi^{(t+1)}$ is a policy such that $\|\pi^{(t+1)} - \pi^{(t)}\| \leq \varepsilon$ for some $t \in [T-1]$, then $\pi^{(t+1)}$ is a $(C_{KKT}\,\varepsilon)$-$\widetilde{\text{KKT}}$ policy of (InitPb) where $C_{KKT}$ is a positive constant such that $C_{KKT} \lesssim \frac{m^{2.5} A_{\max}^{1.5} S}{(1-\gamma)^{4.5}\sqrt{\zeta}}$.*

**Proof** First, note that (ProxPb$(\eta, \pi^{(t)})$) is a strongly convex optimization problem with strongly convex constraints, which is sufficient for the existence of a unique optimum $\tilde{\pi}^{(t+1)}$. Since by Assumption 4.3, Slater's condition holds for (ProxPb$(\eta, \pi')$) for any $\pi' \in \Pi$, strong duality is given for (ProxPb$(\eta, \pi^{(t)})$) and hence there exists a finite dual variable $\tilde{\lambda}^{(t+1)} \geq 0$ forming a KKT pair with $\tilde{\pi}^{(t+1)}$. We first claim that $\|\tilde{\pi}^{(t+1)} - \pi^{(t+1)}\| \leq \varepsilon$. This can be seen as follows: By optimality of $(\tilde{\pi}^{(t+1)}, \tilde{\lambda}^{(t+1)})$ for (ProxPb$(\eta, \pi^{(t)})$), we have

$$\tilde{\lambda}^{(t+1)} \left(V_{\eta,\pi^{(t)}}^c(\tilde{\pi}^{(t+1)}) - \alpha + \beta\right) = 0 \,, \tag{B.3}$$

$$\left\langle \nabla_\pi \mathcal{L}_{\eta,\pi^{(t)}}(\tilde{\pi}^{(t+1)}, \tilde{\lambda}^{(t+1)}), \pi^{(t+1)} - \tilde{\pi}^{(t+1)} \right\rangle \geq 0. \tag{B.4}$$

From Lemma B.9, we know that $\mathcal{L}_{\eta,\pi^{(t)}}(\cdot,\tilde{\lambda}^{(t+1)})$ is $L_\Phi(1+\tilde{\lambda}^{(t)})$-strongly convex. Therefore, after rearranging the standard strong convexity lower bound, we get

$$\begin{aligned}
&\frac{L_\Phi}{2}\left(1+\tilde{\lambda}^{(t+1)}\right)\|\tilde{\pi}^{(t+1)}-\pi^{(t+1)}\|^2 \\
&\quad\le \mathcal{L}_{\eta,\pi^{(t)}}(\pi^{(t+1)},\tilde{\lambda}^{(t+1)})-\mathcal{L}_{\eta,\pi^{(t)}}(\tilde{\pi}^{(t+1)},\tilde{\lambda}^{(t+1)}) \\
&\qquad -\left\langle \nabla_\pi \mathcal{L}_{\eta,\pi^{(t)}}(\tilde{\pi}^{(t+1)},\tilde{\lambda}^{(t+1)}),\pi^{(t+1)}-\tilde{\pi}^{(t+1)}\right\rangle \\
&\quad\overset{(a)}{\le} \underbrace{\Phi_{\eta,\pi^{(t)}}(\pi^{(t+1)})-\Phi_{\eta,\pi^{(t)}}(\tilde{\pi}^{(t+1)})}_{\le\bar{\varepsilon}^2}+\tilde{\lambda}^{(t+1)}\underbrace{\left(V_c(\pi^{(t+1)})-\alpha-\beta\right)}_{\le 0} \\
&\quad\overset{(b)}{\le} \bar{\varepsilon}^2,
\end{aligned}$$

where step (a) follows by applying (B.3) and (B.4), and step (b) by Lemma B.2, i.e. the guarantee for $\pi^{(t+1)}$ provided by the algorithm's inner loop. Then, it follows from the previous inequality that

$$\|\tilde{\pi}^{(t+1)}-\pi^{(t+1)}\| \le \sqrt{\frac{2\bar{\varepsilon}^2}{L_\Phi}} \le \frac{\varepsilon}{\sqrt{2}} \le \varepsilon. \tag{B.5}$$

Using the fact that $(\tilde{\pi}^{(t+1)},\tilde{\lambda}^{(t+1)})$ is a KKT pair for (ProxPb($\eta,\pi^{(t)}$)), we now argue that $(\pi^{(t+1)},\tilde{\lambda}^{(t+1)})$ is a $(C_{\text{KKT}}\,\varepsilon)$-$\widetilde{\text{KKT}}$ pair for (InitPb) (see Definition B.20 in Appendix B.4). We check each one of the requirements of the definition in what follows.

- **Exact primal feasibility:** By Lemma B.2, we know that $V_c(\pi^{(t+1)}) \le \alpha$ for any $0 \le t \le T-1$.
- **Dual feasibility:** This immediately holds by dual feasibility of $(\tilde{\pi}^{(t+1)},\tilde{\lambda}^{(t+1)})$ for (ProxPb($\eta,\pi^{(t)}$)).
- **Complementary slackness:** In the case of $\tilde{\lambda}^{(t+1)}=0$, we clearly have

  $$|\tilde{\lambda}^{(t+1)}\left(V_c(\pi^{(t+1)})-\alpha\right)| = 0 \le \varepsilon.$$

  Otherwise, we have

  $$\begin{aligned}
  V_c(\pi^{(t+1)}) &\overset{(a)}{\ge} V_c(\tilde{\pi}^{(t+1)})-M_c\varepsilon \\
  &\overset{(b)}{=} \alpha-\beta-\frac{1}{2\eta}\|\tilde{\pi}^{(t+1)}-\pi^{(t)}\|^2-M_c\varepsilon \\
  &\overset{(c)}{\ge} \alpha-\frac{\varepsilon^2}{\eta}-\left(M_c+\frac{1}{2\sqrt{\eta}}\right)\varepsilon,
  \end{aligned} \tag{B.6}$$

  where (a) follows from $M_c$ Lipschitz continuity of $V_c$ (see Lemma B.9-item (1)) and Eq. (B.5), (b) stems from complementary slackness of $(\tilde{\pi}^{(t+1)},\tilde{\lambda}^{(t+1)})$ for (ProxPb($\eta,\pi^{(t)}$)) which states that $V^c_{\eta,\pi^{(t)}}(\tilde{\pi}^{(t+1)})-\alpha+\beta=0$. To obtain inequality (c), observe that using the bound from (B.5), our assumption on $\|\pi^{(t+1)}-\pi^{(t)}\|$, and the triangle inequality, we have

  $$\|\tilde{\pi}^{(t+1)}-\pi^{(t)}\| \le \|\tilde{\pi}^{(t+1)}-\pi^{(t+1)}\|+\|\pi^{(t+1)}-\pi^{(t)}\| \le 2\varepsilon.$$

Combining (B.6) with the upper bound $V_c(\pi^{(t+1)}) \leq \alpha$ from primal feasibility, we get

$$\left|\tilde{\lambda}^{(t+1)}\left(V_c(\pi^{(t+1)}) - \alpha\right)\right| \leq \tilde{\lambda}^{(t+1)}\left(\frac{\varepsilon^2}{\eta} + M_c\varepsilon + \frac{\varepsilon}{2\sqrt{\eta}}\right). \tag{B.7}$$

We now show that the dual variable $\tilde{\lambda}^{(t+1)}$ is bounded by a constant depending on $\zeta$ using Assumption 4.3 and strong duality. Indeed, we have

$$\tilde{\lambda}^{(t+1)} \leq \frac{\left\|\nabla_\pi \Phi(\tilde{\pi}^{(t+1)})\right\| + \eta^{-1}\left\|\tilde{\pi}^{(t+1)} - \pi^{(t)}\right\|}{\sqrt{\zeta\eta^{-1}}} \leq \frac{M_c + 4\varepsilon\eta^{-1}}{\sqrt{\zeta\eta^{-1}}}, \tag{B.8}$$

where the first inequality follows from the proof of Lemma 1 in [MLY20], whereas the second inequality uses Lipschitzness of the potential function (see Lemma B.9) and the fact that $\left\|\tilde{\pi}^{(t+1)} - \pi^{(t)}\right\| \leq 2\varepsilon$. Combining (B.7) and (B.8), and using the bounds on $M_c$ and $L_\Phi$ from Lemma B.9, we obtain the desired $C_{\text{KKT}}\varepsilon$-complementary slackness.

- **Variational Lagrangian stationarity:** Suppose by contradiction that the Lagrangian stationarity condition that comes with the $\frac{2\varepsilon(1+\tilde{\lambda}^{(t+1)})}{\eta}$-KKT conditions does not hold for $\tilde{\pi}^{(t+1)}$ and (InitPb). Then there exists $\nu \in N_\Pi(\tilde{\pi}^{(t+1)}, \tilde{\lambda}^{(t+1)})$ (normal cone to the convex set of policies $\Pi$) such that

$$\nabla_\pi \mathcal{L}_{\eta,\pi^{(t)}}(\tilde{\pi}^{(t+1)}, \tilde{\lambda}^{(t+1)}) + \nu = 0 \quad \text{and}$$
$$\left\|\nabla_\pi \mathcal{L}(\tilde{\pi}^{(t+1)}, \tilde{\lambda}^{(t+1)}) + \nu\right\| > \frac{2\varepsilon(1+\tilde{\lambda}^{(t+1)})}{\eta}$$

where the equality is by Lagrangian stationarity of $\tilde{\pi}^{(t+1)}$ for (ProxPb($\eta, \pi^{(t)}$)) and the inequality is due to the above-assumed lack of Lagrangian stationarity of $\tilde{\pi}^{(t+1)}$ for (InitPb). Plugging in the definition of $\mathcal{L}_{\eta,\pi^{(t)}}(\tilde{\pi}^{(t+1)}, \tilde{\lambda}^{(t+1)})$ and combining the equality and inequality above, one can conclude that

$$\frac{2\varepsilon(1+\tilde{\lambda}^{(t+1)})}{\eta} < \left\|\nabla_\pi \mathcal{L}(\tilde{\pi}^{(t+1)}, \tilde{\lambda}^{(t+1)}) + \nu\right\| = \frac{1+\tilde{\lambda}^{(t+1)}}{\eta}\left\|\tilde{\pi}^{(t+1)} - \pi^{(t)}\right\|,$$

which contradicts the inequality $\left\|\tilde{\pi}^{(t+1)} - \pi^{(t)}\right\| \leq 2\varepsilon$. Hence using the bound on $\tilde{\lambda}^{(t+1)}$ from (B.8), the policy $\tilde{\pi}^{(t+1)}$ is a $\tilde{C}\varepsilon$-KKT policy for (InitPb) with $\tilde{C} = \frac{2}{\eta}\left(1 + \frac{M_c + 2\varepsilon\eta^{-1}}{\sqrt{\zeta\eta^{-1}}}\right)$.

By Lemma B.21, this implies that

$$\max_{\pi' \in \Pi}\left\langle \tilde{\pi}^{(t+1)} - \pi', \nabla_\pi \mathcal{L}(\tilde{\pi}^{(t+1)}, \tilde{\lambda}^{(t+1)})\right\rangle \leq \text{diam}(\Pi)\tilde{C}\varepsilon. \tag{B.9}$$

Then, in view of showing the variational Lagrangian stationarity for the pair $(\pi^{(t+1)}, \tilde{\lambda}^{(t+1)})$

for (InitPb), we can write

$$\begin{aligned}
&\max_{\pi' \in \Pi} \left\langle \pi^{(t+1)} - \pi', \nabla_\pi \mathcal{L}(\pi^{(t+1)}, \tilde{\lambda}^{(t+1)}) \right\rangle \\
&\quad = \max_{\pi' \in \Pi} \left\langle \tilde{\pi}^{(t+1)} - \pi', \nabla_\pi \mathcal{L}(\pi^{(t+1)}, \tilde{\lambda}^{(t+1)}) \right\rangle \qquad \text{(B.10)} \\
&\qquad + \left\langle \pi^{(t+1)} - \tilde{\pi}^{(t+1)}, \nabla_\pi \mathcal{L}(\pi^{(t+1)}, \tilde{\lambda}^{(t+1)}) \right\rangle \\
&\quad \leq \max_{\pi' \in \Pi} \left\langle \tilde{\pi}^{(t+1)} - \pi', \nabla_\pi \mathcal{L}(\pi^{(t+1)}, \tilde{\lambda}^{(t+1)}) \right\rangle \qquad \text{(B.11)} \\
&\qquad + \left\| \pi^{(t+1)} - \tilde{\pi}^{(t+1)} \right\| \cdot \left\| \nabla_\pi \mathcal{L}(\pi^{(t+1)}, \tilde{\lambda}^{(t+1)}) \right\| \\
&\quad \leq \max_{\pi' \in \Pi} \left\langle \tilde{\pi}^{(t+1)} - \pi', \nabla_\pi \mathcal{L}(\pi^{(t+1)}, \tilde{\lambda}^{(t+1)}) \right\rangle + \varepsilon(1 + \tilde{\lambda}^{(t+1)}) M_c \,. \qquad \text{(B.12)}
\end{aligned}$$

We now bound the first term in the above inequality by using $2L_\Phi(1 + \tilde{\lambda}^{(t+1)})$-smoothness of $\nabla_\pi \mathcal{L}(\cdot, \tilde{\lambda}^{(t+1)})$ and (B.9). Using the fact that

$$\max_{\pi \in \Pi} (A(\pi) + B(\pi)) \leq \max_{\pi \in \Pi} A(\pi) + \max_{\pi \in \Pi} B(\pi)$$

for any functions $A(\pi), B(\pi)$, we have

$$\begin{aligned}
&\max_{\pi' \in \Pi} \left\langle \tilde{\pi}^{(t+1)} - \pi', \nabla_\pi \mathcal{L}(\pi^{(t+1)}, \tilde{\lambda}^{(t+1)}) \right\rangle \\
&\leq \max_{\pi' \in \Pi} \left\langle \tilde{\pi}^{(t+1)} - \pi', \nabla_\pi \mathcal{L}(\pi^{(t+1)}, \tilde{\lambda}^{(t+1)}) - \nabla_\pi \mathcal{L}(\tilde{\pi}^{(t+1)}, \tilde{\lambda}^{(t+1)}) \right\rangle \\
&\qquad + \max_{\pi' \in \Pi} \left\langle \tilde{\pi}^{(t+1)} - \pi', \nabla_\pi \mathcal{L}(\tilde{\pi}^{(t+1)}, \tilde{\lambda}^{(t+1)}) \right\rangle \\
&\leq \max_{\pi' \in \Pi} \left\| \tilde{\pi}^{(t+1)} - \pi' \right\| \cdot 2L_\Phi(1 + \tilde{\lambda}^{(t+1)}) \left\| \tilde{\pi}^{(t+1)} - \pi^{(t+1)} \right\| + \mathrm{diam}(\Pi)\tilde{C}\varepsilon \\
&\leq \left( 2\mathrm{diam}(\Pi) L_\Phi (1 + \tilde{\lambda}^{(t+1)}) + \mathrm{diam}(\Pi)\tilde{C} \right) \varepsilon \,.
\end{aligned}$$

Combining the above inequality with (B.10), we obtain

$$\begin{aligned}
&\max_{\pi' \in \Pi} \left\langle \pi^{(t+1)} - \pi', \nabla_\pi \mathcal{L}(\pi^{(t+1)}, \tilde{\lambda}^{(t+1)}) \right\rangle \\
&\qquad \leq \left( (1 + \tilde{\lambda}^{(t+1)}) M_c + 2\mathrm{diam}(\Pi) L_\Phi (1 + \tilde{\lambda}^{(t+1)}) + \mathrm{diam}(\Pi)\tilde{C} \right) \varepsilon.
\end{aligned}$$

Finally, we use the bound on $\tilde{\lambda}^{(t+1)}$ from (B.8), as well as bounds on $\mathrm{diam}(\Pi)$, $L_\Phi$, and $M_c$ from Lemma B.9, to conclude that $\pi^{(t+1)}$ is a $C_{\mathrm{KKT}}\varepsilon$-$\widetilde{\mathrm{KKT}}$ policy for (InitPb). □

To complete the analysis, it now remains to show that an $\mathcal{O}(\varepsilon)$-$\widetilde{\mathrm{KKT}}$ policy of (InitPb) is a constrained $\mathcal{O}(\varepsilon)$-NE. For this, we leverage the playerwise gradient domination property satisfied by the potential function and the constraint value function. We first introduce some notations.

**Notation.** For each player $i \in \mathcal{N}$ and each policy $\pi_{-i} \in \Pi^{-i}$, consider the playerwise constrained optimization problem given by

$$\min_{\pi_i \in \Pi_c^i(\pi_{-i})} V_{r_i}(\pi_i, \pi_{-i}) \,. \qquad \text{(PlayerPb}(\pi_{-i}))$$

The respective Lagrangian $\mathcal{L}_{\pi_{-i}} : \Pi^i \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ is defined for every $\pi_i \in \Pi^i$ and every $\lambda \geq 0$ by

$$\mathcal{L}_{\pi_{-i}}(\pi_i, \lambda) = \Phi(\pi_i, \pi_{-i}) + \lambda \left( V_c(\pi_i, \pi_{-i}) - \alpha \right) \,. \qquad \text{(PlayerPb}(\pi_{-i})\text{–}\mathcal{L})$$

**Lemma B.5** *Let $\pi \in \Pi$ be an $\varepsilon$-$\widetilde{\text{KKT}}$ policy of* (InitPb)*. Then $\pi$ is a constrained $C_{NE}$ $\varepsilon$-NE where $C_{NE} \lesssim \frac{D}{1-\gamma} + 1$.*

**Proof** The proof of the lemma proceeds in two steps:

- **Step 1.** We show that if $\pi$ is an $\mathcal{O}(\varepsilon)$-$\widetilde{\text{KKT}}$ policy of (InitPb), then for all $i \in \mathcal{N}$, $\pi_i$ is an $\mathcal{O}(\varepsilon)$-$\widetilde{\text{KKT}}$ policy of (PlayerPb($\pi_{-i}$)).
- **Step 2.** We conclude that each player cannot significantly improve its policy $\pi_i$ while staying within $\Pi_c^i(\pi_{-i})$ which means $\pi$ is a constrained $\mathcal{O}(\varepsilon)$-NE.

We provide a proof of each one of the steps successively.

- **Step 1:** Let $\lambda \geq 0$ be a dual variable such that $(\pi, \lambda)$ is an $\varepsilon$-$\widetilde{\text{KKT}}$ pair of (InitPb), and let $i \in \mathcal{N}$ be arbitrary. We show that $(\pi_i, \lambda)$ is an $\varepsilon$-$\widetilde{\text{KKT}}$ pair of (PlayerPb($\pi_{-i}$)) by checking that the respective $\widetilde{\text{KKT}}$ conditions hold. For dual and exact primal feasibility, as well as complementary slackness, this is immediate since the conditions are equivalent for (InitPb) and (PlayerPb($\pi_{-i}$)). For variational Lagrangian stationarity, observe that

$$\begin{aligned}\max_{\pi_i' \in \Pi^i} \left\langle \pi_i - \pi_i', \nabla_{\pi_i} \mathcal{L}_{\pi_{-i}}(\pi_i, \lambda) \right\rangle &= \max_{\pi_i' \in \Pi^i} \left\langle \pi - (\pi_i', \pi_{-i}), \nabla_\pi \mathcal{L}(\pi, \lambda) \right\rangle \\ &\leq \max_{\pi' \in \Pi} \left\langle \pi - \pi', \nabla_\pi \mathcal{L}(\pi, \lambda) \right\rangle \\ &\leq \varepsilon\,,\end{aligned}$$

  where the first equality is due to $\nabla_{\pi_i} \mathcal{L}(\pi, \lambda) = \nabla_{\pi_i} \mathcal{L}_{\pi_{-i}}(\pi_i, \lambda)$ and the fact that all terms except for $\pi_i - \pi_i'$ vanish in the first argument of the scalar product. The second inequality is because $(\pi_i', \pi_{-i}) \in \Pi$, and the final step is by Lagrangian stationarity of $\pi$ for (InitPb).

- **Step 2:** Let $i \in \mathcal{N}$ and consider the MDP $M_\lambda$, $\lambda \geq 0$, with state space $\mathcal{S}$, action space $\mathcal{A}_i$, probability transition kernel $P_\lambda$, reward $r_\lambda$, and initial distribution $\mu$ where

$$\begin{aligned}P_\lambda(s' \mid s, a_i) &:= \mathbb{E}_{a_{-i} \sim \pi_{-i}((a_i, \cdot)|s)} \left[ P(s' \mid s, (a_i, a_{-i})) \right] \\ r_\lambda(s, a_i) &:= \mathbb{E}_{a_{-i} \sim \pi_{-i}((a_i, \cdot)|s)} \left[ r_i(s, (a_i, a_{-i})) + \lambda\, c(s, (a_i, a_{-i})) \right].\end{aligned}$$

  Observe that $\mathcal{L}_{\pi_{-i}}(\pi_i, \lambda)$ is the value function associated to the policy $\pi_i$ in the MDP $M_\lambda$, and hence gradient domination holds [AKLM21], i.e., we have

$$\mathcal{L}_{\pi_{-i}}(\pi_i, \lambda) - \min_{\tilde{\pi}_i \in \Pi^i} \mathcal{L}_{\pi_{-i}}(\tilde{\pi}_i, \lambda) \leq \frac{D}{1-\gamma} \max_{\pi_i' \in \Pi^i} \left\langle \pi_i - \pi_i', \nabla_{\pi_i} \mathcal{L}_{\pi_{-i}}(\pi_i, \lambda) \right\rangle\,,$$

  where $D$ is the distribution mismatch coefficient, supposed to be finite. Using Proposition B.22 in Appendix B.4 for $C_1 = 0$, and using the definition of the playerwise primal optimum, see (PlayerPb($\pi_{-i}$)), we then get

$$V_{r_i}(\pi) - \min_{\pi_i^* \in \Pi_{c_i}(\pi_{-i})} V_i(\pi_i^*, \pi_{-i}) \leq \left( \frac{D}{1-\gamma} + 1 \right) \varepsilon.$$

  Since additionally we have exact primal feasibility of $\pi$ for (InitPb), the result follows by definition of the constrained $\varepsilon$-NE. □

Finally, we put together above lemmas to prove the main theorem.

**Proof (Proof of Theorem 4.4)** Let $\bar{\varepsilon}^2 = \frac{L_\Phi \varepsilon^2}{C_{\text{KKT}}^2 C_{\text{NE}}^2}$. Then with $K = \mathcal{O}\left(\varepsilon^{-2}\right)$, and $T = \mathcal{O}\left(\varepsilon^{-2}\right)$, Lemma B.2 and Lemma B.3 imply that there exists $0 \leq t \leq T-1$ such that $\mathbb{E}[\|\pi^{(t+1)} - \pi^{(t)}\|] \leq \frac{\varepsilon}{C_{\text{KKT}} C_{\text{NE}}}$. We use Lemma B.4 to conclude that $\pi^{(t+1)}$ is a $\varepsilon/C_{\text{NE}}$-$\widetilde{\text{KKT}}$ policy of (InitPb). Then, by Lemma B.5, $\pi^{(t+1)}$ is a constrained $\varepsilon$-NE. The total iteration complexity is bounded by $T \cdot K = \mathcal{O}\left(\varepsilon^{-4}\right)$. □

### B.2.2 Proof of Theorem 4.5 – Finite Sample Case

Moving on to the stochastic setting, we first restate Theorem 4.5.

**Theorem 4.5** Let Assumptions 4.2 and 4.3 hold, and let $D$ (as in Theorem 4.4) be finite. Then, for any $\varepsilon > 0$, after running iProxCMPG based on finite sample estimates (see Algorithm 4) for suitably chosen $\eta, \beta, \xi, T, K, B$, and $\{(\nu_k, \delta_k, \rho_k)\}_{0 \leq k \leq K}$, there exists $t \in [T]$, such that in expectation, $\pi^{(t)}$ is a constrained $\varepsilon$-NE. The total sample complexity is given by $\tilde{\mathcal{O}}\left(\varepsilon^{-7}\right)$ where $\tilde{\mathcal{O}}(\cdot)$ hides polynomial dependencies in $m, S, A_{\max}, D, 1-\gamma$, and $\zeta$, as well as logarithmic dependencies in $1/\varepsilon$. □

Similar to the deterministic case, we begin by proving the guarantees provided by the inner loop of Algorithm 4.

**Lemma B.6** *Let Assumption 4.2 hold, let $\bar{\varepsilon} > 0$ and set $\beta = \bar{\varepsilon}, \eta = \frac{1}{2L_\Phi}, \xi = \bar{\varepsilon}\sqrt{2\eta}$. Then, there exist $K = \tilde{\mathcal{O}}\left(\bar{\varepsilon}^{-3}\right)$, $B = \tilde{\mathcal{O}}\left(\bar{\varepsilon}^{-2}\right)$, and suitable choices of $\{(\nu_k, \delta_k, \rho_k)\}_{0 \leq k \leq K}$, such that lines 4-11 of Algorithm 4 guarantee that for any $t \in [T-1]$,*

$$\begin{aligned} \mathbb{E}\left[\Phi_{\eta,\pi^{(t)}}(\pi^{(t+1)}) - \Phi_{\eta,\pi^{(t)}}(\tilde{\pi}^{(t+1)})\right] &\leq \bar{\varepsilon}^2, \\ \mathbb{E}\left[V_c(\pi^{(t+1)})\right] &\leq \alpha\,, \end{aligned} \tag{B.13}$$

*where $\tilde{\pi}^{(t+1)}$ denotes the unique optimal solution to (ProxPb($\eta, \pi^{(t)}$)).*

**Proof** The result follows similarly as for Lemma B.2 in the deterministic case. We hence only point out differences. In order to ensure bounded norms of gradient estimates, we use $\xi$-greedy policies. Then, according to Lemma B.11, the second moment of value and constraint gradient estimates is bounded by $\mathcal{O}(1/\xi)$. The concentration result shown in Lemma B.12 ensures that constraint value estimates follow a sub-exponential distribution. Therefore, Assumption B.14, see Appendix B.3 on guarantees for our subroutine, is satisfied. We can thus apply the respective Theorem B.18 for optimizing over $\Pi^\xi$, and with $\mu = L_\Phi$ and $M^2 \lesssim \max\left\{M_G^2 + \mu_G^2\Delta^4, M_F^2 + \mu_F^2\Delta^4\right\} \lesssim \frac{24A_{\max}^2}{\xi(1-\gamma)^4} + L_\Phi^2 \text{diam}(\Pi)^4$. After plugging bounds on $L_\Phi$, and $\text{diam}(\Pi)$ from Lemma B.9, and choosing $K$ and $B$ as stated, Theorem B.18 implies the desired bounds via the same arguments as in the proof of Lemma B.2. □

Next, we analyze the convergence of our main proximal-point method, Algorithm 4. More concretely, we bound the required sample complexity for ensuring that for some $\varepsilon > 0$, there exist iterates $\pi^{(t)}, \pi^{(t+1)}$ such that $\|\pi^{(t)} - \pi^{(t+1)}\| = \mathcal{O}(\varepsilon)$. In the following, we will then prove that this implies convergence to a constrained $\mathcal{O}(\varepsilon)$-NE.

Similarly to the deterministic case, we next determine the number of updates needed until convergence in the following sense. The next lemma is analogous to its deterministic counterpart Lemma B.3.

**Lemma B.7** *Let $\varepsilon > 0$ and set $\eta = \frac{1}{2L_\Phi}$. Suppose $K$ is chosen such that the guarantee from Lemma B.6 holds for $\bar{\varepsilon}^2 = \frac{\varepsilon^2}{4\eta}$. Then after $T = \frac{4\eta\Phi_{\max}}{\varepsilon^2}$ iterations of the outer loop of Algorithm 4 where $\Phi_{\max}$ is an upper bound of the potential function (i.e., $\forall \pi \in \Pi, \Phi(\pi) \leq \Phi_{\max}$), there exists $0 \leq t \leq T-1$ such that $\mathbb{E}[\|\pi^{(t+1)} - \pi^{(t)}\|] \leq \varepsilon$.*

**Proof** The proof follows the same lines as the proof of Lemma B.3 upon replacing Lemma B.2 by Lemma B.6. We do not reproduce it here for conciseness.

Next, we prove that the event $\|\pi^{(t+1)} - \pi^{(t)}\| \leq \varepsilon$ implies Nash-gap$(\pi^{(t+1)}) = \mathcal{O}(\varepsilon)$, in order to argue that $\mathbb{E}[\|\pi^{(t+1)} - \pi^{(t)}\|] \leq \varepsilon$ implies $\mathbb{E}[\text{Nash-gap}(\pi^{(t+1)})] = \mathcal{O}(\varepsilon)$.

Recall that in Lemma B.4 we have already shown $\|\pi^{(t+1)} - \pi^{(t)}\| \leq \varepsilon$ to imply that $\pi^{(t+1)}$ is a $C_{\text{KKT}}\varepsilon$-$\widetilde{\text{KKT}}$ policy of (InitPb) which equivalently holds in the stochastic $\xi$-greedy setting. Arguing that a $\varepsilon$-$\widetilde{\text{KKT}}$ policy is a constrained $\mathcal{O}(\varepsilon)$-NE, however, requires an adapted proof, since here in each iteration we solve the subproblem over $\Pi^{\xi}$ instead of $\Pi$, i.e., the $\widetilde{\text{KKT}}$ conditions hold w.r.t. $\Pi^{\xi}$. The following lemma is an adjustment of Lemma B.5 for this fact.

**Lemma B.8** *Let $\pi \in \Pi^{\xi}$ be an $\varepsilon$-$\widetilde{KKT}$ policy of* (InitPb) *(where $\widetilde{KKT}$ are w.r.t. $\Pi^{\xi}$) and $\xi = \varepsilon$. Then $\pi$ is a constrained $\hat{C}_{NE}$ $\varepsilon$-NE where $\hat{C}_{NE} \lesssim \frac{D}{1-\gamma} + \frac{m\sqrt{S}A_{\max}D}{(1-\gamma)^{4.5}} + 1$.*

**Proof** We divide the proof into two steps:

- **Step 1:** Analogously to step 1 of Lemma B.5, one can show that $(\pi_i, \lambda)$ is an $\varepsilon$-$\widetilde{\text{KKT}}$ pair of (PlayerPb($\pi_{-i}$)).
- **Step 2:** Let $i \in \mathcal{N}$ and consider the MDP $\tilde{M}_\lambda$ (for $\lambda \geq 0$) with state space $\mathcal{S}$, action space $\mathcal{A}_i$, transition probability kernel $\tilde{P}$, reward $\tilde{r}_\lambda$, discount factor $\gamma$, and initial distribution $\mu$ where

$$\begin{aligned}\tilde{P}(s' \mid s, a_i) &:= \mathbb{E}_{a_{-i}\sim\pi_{-i}(\cdot|s)}\left[P(s' \mid s, (a_i, a_{-i}))\right] ,\\ \tilde{r}_\lambda(s, a_i) &:= \mathbb{E}_{a_{-i}\sim\pi_{-i}(\cdot|s)}\left[r_i(s, (a_i, a_{-i})) + \lambda\, c(s, (a_i, a_{-i}))\right] .\end{aligned}$$

  Observe that $\mathcal{L}_{\pi_{-i}}(\pi_i, \lambda)$ is the value function associated to the policy $\pi_i$ for $\tilde{M}_\lambda$, and hence gradient domination holds [AKLM21]. In our particular case of $\pi$ being a $\xi$-greedy policy, we can also show a similar inequality, even w.r.t. the non-$\xi$-greedy optimum. Let

$$\begin{aligned}\hat{\pi}_i &\in \operatorname*{arg\,max}_{\pi_i' \in \Pi^i} \left\langle \pi_i - \pi_i', \nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i, \lambda)\right\rangle ,\\ \hat{\pi}_i^{\xi} &\in \operatorname*{arg\,max}_{\pi_i' \in \Pi^{i,\xi}} \left\langle \pi_i - \pi_i', \nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i, \lambda)\right\rangle .\end{aligned}$$

  Then, we have

$$\begin{aligned}&\mathcal{L}_{\pi_{-i}}(\pi_i, \lambda) - \min_{\pi_i^* \in \Pi^i} \mathcal{L}_{\pi_{-i}}(\pi_i^*, \lambda)\\ &\quad\leq \frac{1}{1-\gamma}\left\|\frac{d_\mu^{\pi_i^*, \pi_{-i}}}{\mu}\right\|_\infty \left\langle \pi_i - \hat{\pi}_i \nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i, \lambda)\right\rangle\\ &\quad= \frac{1}{1-\gamma}\left\|\frac{d_\mu^{\pi_i^*, \pi_{-i}}}{\mu}\right\|_\infty \left\langle \pi_i - \hat{\pi}_i^{\xi}, \nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i, \lambda)\right\rangle\\ &\qquad+ \left\langle \hat{\pi}_i^{\xi} - \hat{\pi}_i, \nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i, \lambda)\right\rangle .\end{aligned}$$

We further bound the last term above as follows

$$\begin{aligned}
&\left\langle \hat{\pi}_i^{\xi} - \hat{\pi}_i, \nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i,\lambda)\right\rangle \\
&\quad = \max_{\pi_i^{\xi}\in\Pi^{i,\xi}} \left\langle \pi_i^{\xi}, \nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i,\lambda)\right\rangle - \max_{\pi_i\in\Pi^i}\left\langle \pi_i, \nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i,\lambda)\right\rangle \\
&\quad \overset{(a)}{\leq} \xi\sqrt{S}\left\|\nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i,\lambda)\right\| \\
&\quad \overset{(b)}{\leq} \xi\sqrt{S}(1+\lambda)M_c\,.
\end{aligned}$$

In the above inequalities, (a) follows from using Lemma B.10 to obtain

$$\begin{aligned}
&\max_{\pi_i^{\xi}\in\Pi^{i,\xi}} \left\langle \pi_i^{\xi}, \nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i,\lambda)\right\rangle \\
&\quad \leq \max_{\pi_i\in\Pi^i}\left\langle \pi_i, \nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i,\lambda)\right\rangle + \sum_{a_i,s}\frac{\xi}{A_i}[\nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i,\lambda)](a_i\mid s)\,.
\end{aligned}$$

Using the fact that for any $x\in\mathbb{R}^d$, $\|x\|_1\leq\sqrt{d}\|x\|_2$, we further get

$$\begin{aligned}
\sum_{a_i,s}\frac{\xi}{A_i}[\nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i,\lambda)](a_i\mid s) &= \frac{\xi}{A_i}\|\nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i,\lambda)\|_1 \\
&\leq \frac{\xi}{A_i}\sqrt{A_iS}\,\|\nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i,\lambda)\| \\
&\leq \zeta\sqrt{S}\,\|\nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i,\lambda)\|\,.
\end{aligned}$$

The bound used in (b) follows from Lipschitz continuity, see Lemma B.9. We conclude that

$$\mathcal{L}_{\pi_{-i}}(\pi_i,\lambda) - \min_{\pi_i^*\in\Pi^i}\mathcal{L}_{\pi_{-i}}(\pi_i^*,\lambda) \tag{B.14}$$

$$\leq \frac{1}{1-\gamma}\left\|\frac{d_\mu^{\pi_i^*,\pi_{-i}}}{\mu}\right\|_\infty\left[\left\langle \pi_i - \hat{\pi}_i\nabla_{\pi_i}\mathcal{L}_{\pi_{-i}}(\pi_i,\lambda)\right\rangle + \xi\sqrt{S}(1+\lambda)M_c\right]. \tag{B.15}$$

Applying Proposition B.22, see Appendix B.4.2, with $\xi=\varepsilon$, bounding the distribution mismatch coefficient by $D$, and using the definition of the playerwise primal optimum, see (PlayerPb($\pi_{-i}$)), we then get

$$\begin{aligned}
V_i(\pi) - \min_{\pi_i^*\in\Pi_{c_i}(\pi_{-i})} V_i(\pi_i^*,\pi_{-i}) &\leq \left(\frac{D}{1-\gamma} + \frac{(1+\lambda)\sqrt{S}M_cD}{1-\gamma} + 1\right)\varepsilon \\
&\leq \left(\frac{D}{1-\gamma} + \frac{m\sqrt{S}A_{\max}D}{(1-\gamma)^{4.5}} + 1\right)\varepsilon.
\end{aligned}$$

where for the second inequality we use (B.8) and our bounds on $M_c$ and $L_\Phi$ from Lemma B.9. Since additionally, we have exact primal feasibility of $\pi$ for (InitPb) by the $\widetilde{\text{KKT}}$ conditions, the result follows by definition of a constrained $\varepsilon$-NE. □

Finally, we put together the above lemmas to prove our main theorem in the stochastic setting.

**Proof (Proof of Theorem 4.5)** Let $\bar{\varepsilon}^2 = \frac{L_\Phi\varepsilon^2}{C_{\text{KKT}}^2\hat{C}_{\text{NE}}^2}$. Then with $K=\tilde{\mathcal{O}}\left(\varepsilon^{-3}\right)$, $B=\tilde{\mathcal{O}}\left(\varepsilon^{-2}\right)$, and $T=\mathcal{O}\left(\varepsilon^{-2}\right)$, Lemma B.2 and Lemma B.3 imply that there exists $t\in[T-1]$ such that $\mathbb{E}[\|\pi^{(t+1)}-\pi^{(t)}\|]\leq\frac{\varepsilon}{C_{\text{KKT}}\hat{C}_{\text{NE}}}$. We use Lemma B.4 to conclude that $\pi^{(t+1)}$ is a $\varepsilon/\hat{C}_{\text{NE}}$-$\widetilde{\text{KKT}}$ policy of (InitPb) in expectation. Then, by Lemma B.5, $\pi^{(t+1)}$ is a constrained $\varepsilon$-NE in expectation. The total sample complexity is bounded by $T\cdot K\cdot B=\tilde{\mathcal{O}}\left(\varepsilon^{-7}\right)$. □

### B.2.3 Other Technical Lemmas

The next lemma collects standard regularity properties of the value and potential functions.

**Lemma B.9** *The following statements hold true.*

1. *The functions $\Phi$ and $V_c$ are $M_c$-Lipschitz continuous over $\Pi$ with $M_c = \frac{\sqrt{mA_{\max}}}{(1-\gamma)^2}$. This immediately implies that $\|\nabla\Phi(\pi)\| \leq M_c$ and $\|\nabla V_c(\pi)\| \leq M_c$, for all $\pi \in \Pi$.*
2. *For any $i \in \mathcal{N}$ and any $\pi_{-i} \in \Pi_{-i}$, the function $V_{r_i}(\cdot, \pi_{-i})$ is $L_i$-smooth with $L_i = \frac{2\gamma A_i}{(1-\gamma)^3}$ and hence $L_i$-weakly convex.*
3. *The functions $\Phi$ and $V_c$ are $L_\Phi$-smooth with $L_\Phi = m \cdot \max_i L_i = \frac{2m\gamma A_{\max}}{(1-\gamma)^3}$ and hence $L_\Phi$-weakly convex.*
4. *For $\eta = \frac{1}{2L_\Phi}$ and any $\pi' \in \Pi$, the regularized function $\Phi_{\eta,\pi'}(\pi) = \Phi(\pi) + L_\Phi \|\pi - \pi'\|^2$ is $L_\Phi$-strongly convex and the functions $\Phi_{\eta,\pi'}, V^c_{\eta,\pi'}$ are both $2L_\Phi$-smooth.*
5. *For any $\lambda \in \mathbb{R}, \pi' \in \Pi$ and $\eta = \frac{1}{2L_\Phi}$, $\mathcal{L}(\cdot, \lambda) = \Phi(\cdot) + \lambda V_c(\cdot)$ is $L_\Phi(1+\lambda)$-smooth, and $\mathcal{L}_{\eta,\pi'}(\cdot, \lambda) = \Phi_{\eta,\pi'}(\cdot) + \lambda V^c_{\eta,\pi'}(\cdot)$ is $2L_\Phi(1+\lambda)$-smooth. Hence $\mathcal{L}_{\eta,\pi'}(\cdot, \lambda)$ is also $L_\Phi(1+\lambda)$-strongly convex.*

**Proof** Item 2 has been proved in [AKLM21], Lemma D.3. Item 3 has been reported in [LOPP22], Lemma D.4. Item 4 immediately follows from item 3. We now prove item 5 for $\mathcal{L}$, the result for $\mathcal{L}_{\eta,\pi'}$ follows similarly. Using the definition of the Lagrangian and the triangle inequality, for any $\lambda \in \mathbb{R}$ and $\pi, \pi' \in \Pi$,

$$\begin{aligned}
\left\|\nabla_\pi \mathcal{L}(\pi, \lambda) - \nabla_\pi \mathcal{L}(\pi', \lambda)\right\| &\leq \left\|\nabla\Phi(\pi) - \nabla\Phi(\pi')\right\| + \lambda \left\|\nabla V_c(\pi) - \nabla V_c(\pi')\right\| \\
&\leq L_\Phi \left\|\pi - \pi'\right\| + \lambda L_\Phi \left\|\pi - \pi'\right\| \\
&\leq L_\Phi(1+\lambda) \left\|\pi - \pi'\right\|.
\end{aligned}$$

To show item 1, Lipschitz continuity of $\Phi$ and $V_c$, observe that for any $i \in \mathcal{N}$, $\pi \in \Pi$ and $\pi'_i \in \Pi^i$, by using Lemma 32 of [ZMD+22] in the second step, we have

$$\begin{aligned}
\left|\Phi(\pi_i, \pi_{-i}) - \Phi(\pi'_i, \pi_{-i})\right| &= \left|V_{r_i}(\pi_i, \pi_{-i}) - V_{r_i}(\pi'_i, \pi_{-i})\right| \\
&\leq \frac{1}{(1-\gamma)^2} \max_{s\in\mathcal{S}} \left\|\pi_i(\cdot \mid s) - \pi'_i(\cdot \mid s)\right\|_1 \\
&\leq \frac{\sqrt{A_i}}{(1-\gamma)^2} \max_{s\in\mathcal{S}} \left\|\pi_i(\cdot \mid s) - \pi'_i(\cdot \mid s)\right\|_2 \\
&\leq \frac{\sqrt{A_i}}{(1-\gamma)^2} \left\|\pi_i - \pi'_i\right\|_2
\end{aligned}$$

where in the third step we use the fact that for any $x \in \mathbb{R}^d$, $\|x\|_1 \leq \sqrt{d}\,\|x\|_2$. Then, the following decomposition yields the result: For any $\pi, \pi' \in \Pi$,

$$\begin{aligned}
\left|\Phi(\pi) - \Phi(\pi')\right| &= \left|\sum_{i\in\mathcal{N}} \Phi(\pi'_1, \ldots, \pi'_{i-1}, \pi_i, \pi_{i+1}, \ldots, \pi_m) - \Phi(\pi'_1, \ldots, \pi'_{i-1}, \pi'_i, \pi_{i+1}, \ldots, \pi_m)\right| \\
&\leq \sum_{i\in\mathcal{N}} \left|\Phi(\pi'_1, \ldots, \pi'_{i-1}, \pi_i, \pi_{i+1}, \ldots, \pi_m) - \Phi(\pi'_1, \ldots, \pi'_{i-1}, \pi'_i, \pi_{i+1}, \ldots, \pi_m)\right| \\
&\leq \frac{1}{(1-\gamma)^2} \sum_{i\in\mathcal{N}} \sqrt{A_i} \left\|\pi_i - \pi'_i\right\| \\
&\leq \frac{\sqrt{mA_{\max}}}{(1-\gamma)^2} \left\|\pi - \pi'\right\|,
\end{aligned}$$

where in the second inequality we apply the above result for playerwise deviations, and the last step is again due to the fact that for any $x \in \mathbb{R}^d$, $\|x\|_1 \leq \sqrt{d}\,\|x\|_2$. The result follows similarly for $V_c$. □

The next lemma is an immediate result showing that any $\xi$-greedy playerwise policy $\pi_i \in \Pi^{i,\xi}$ (see definition in the main part p. 7 which defines this set as a set of lower bounded policies away from zero) can be represented as a convex combination of a uniform distribution over the action space $\mathcal{A}_i$ and a policy $\pi_i \in \Pi^i$.

**Lemma B.10** *For any $\xi > 0$, $i \in \mathcal{N}$,*

$$\Pi^{i,\xi} \subseteq \left\{\pi_i \in \Pi^i \mid \exists \pi_i' \in \Pi^i, \forall a_i \in \mathcal{A}_i, \forall s \in \mathcal{S}:\ \pi_i(a_i \mid s) = \xi/A_i + (1-\xi)\pi_i'(a_i \mid s)\right\}.$$

**Proof** Let $\xi > 0$, $i \in \mathcal{N}$ and let $\pi_i^\xi \in \Pi^{i,\xi}$. Then for all $a_i \in \mathcal{A}_i$ and $s \in \mathcal{S}$, set

$$\pi_i(a_i \mid s) := \frac{\pi_i^\xi(a_i \mid s) - \xi/A_i}{1-\xi}.$$

Indeed $\pi_i \in \Pi^i$, since for all $a_i \in \mathcal{A}_i$ and $s \in \mathcal{S}$, we have $\pi_i(a_i \mid s) \geq 0$ due to $\pi_i^\xi(a_i \mid s) \geq \xi$ and

$$\sum_{a_i \in \mathcal{A}_i} \pi_i(a_i \mid s) = \frac{1}{1-\xi}\left(\underbrace{\sum_{a_i \in \mathcal{A}_i} \pi_i^\xi(a_i \mid s)}_{=1} - \underbrace{\sum_{a_i \in \mathcal{A}_i} \xi/A_i}_{=\xi}\right) = 1. \qquad \square$$

The following lemma shows that the estimators we use for the playerwise policy gradients are unbiased and enjoy a bounded variance.

**Lemma B.11 ([DFG20, LOPP22])** *For any $\xi > 0$ and $\pi \in \Pi^\xi$, we have*

$$\begin{aligned}
&\mathbb{E}_\pi\left[\hat{\nabla} V_{\pi_i}^{r_i}(\pi)\right] = \nabla_{\pi_i} V_{r_i}(\pi) = \nabla_{\pi_i}\Phi(\pi)\,,\\
&\mathbb{E}_\pi\left[\left\|\hat{\nabla} V_{\pi_i}^{r_i}(\pi)\right\|^2\right] \leq \frac{24 A_{\max}^2}{\xi(1-\gamma)^4}\,,\\
&\mathbb{E}_\pi\left[\hat{V}_c(\pi)^2\right] \leq \frac{1}{(1-\gamma)^2}\,.
\end{aligned}$$

*The same holds for $\hat{\nabla} V_{\pi_i}^c(\pi)$ w.r.t. $\nabla_{\pi_i} V_c(\pi)$.*

Finally, the following lemma shows that our constraint function estimates concentrate around their mean.

**Lemma B.12** *For $\pi \in \Pi^\xi$, let $\hat{V}_c^{(1)}, \ldots, \hat{V}_c^{(B)}$ be independent copies of $\hat{V}_c(\pi)$, and let $\hat{V}_c := \frac{1}{B}\sum_{i=1}^B \hat{V}_c^{(i)}$. Then there exists $C > 0$ such that for any $\lambda \geq 0$,*

$$\mathbb{P}\left(\left|\hat{V}_c - V_c(\pi)\right| > \frac{\lambda}{\sqrt{B}}\right) \leq 4\exp\left(-C(1-\gamma)\lambda\right) + 2\exp\left(-C^2(1-\gamma)^2\lambda^2\right).$$

**Proof** For $i \in [B]$, we decompose $\hat{V}_c^{(i)} = \hat{c}_0^{(i)} + \hat{V}_{c,\geq 1}^{(i)}$ where $\hat{c}_0^{(i)}$ is the cost incurred at step 0, and let $\hat{c}_0 := \frac{1}{B}\sum_{i=1}^B c_0^{(i)}$, $\hat{V}_{c,\geq 1} := \frac{1}{B}\sum_{i=1}^B \hat{V}_{c,\geq 1}^{(i)}$. Since $0 \leq \hat{c}_0 \leq 1$, by Hoeffding's inequality, there exists $C_0 > 0$ such that for any $\lambda \geq 0$,

$$\mathbb{P}\left(|\hat{c}_0 - \mathbb{E}[\hat{c}_0]| > \frac{\lambda}{2\sqrt{B}}\right) \leq 2\exp\left(-C_0\lambda^2\right). \tag{B.16}$$

Moreover, note that since for all $s \in \mathcal{S}, a \in \mathcal{A}$, $0 \leq c(s,a) \leq 1$, we have that for all $i \in [B]$, $V_{c,\geq 1}^{(i)} \leq T_e^{(i)}$ where $T_e^{(i)}$ is the stopping time of the respective episode and an independent copy of $T_e$. Assuming $\kappa_{s,a} = \min_{s\in\mathcal{S},a\in\mathcal{A}} \kappa_{s,a} = 1-\gamma$ for all $s \in \mathcal{S}, a \in \mathcal{A}$, $T_e$ follows a geometric distribution with parameter $1-\gamma$. By definition of the geometric distribution and elementary computations, we get for any $\lambda \geq 0$,

$$\begin{aligned}\mathbb{P}(T_e \geq \lambda) \leq \gamma^{\lceil\lambda\rceil} \leq \exp(\lceil\lambda\rceil \log\gamma) &\leq \exp\left(-\lceil\lambda\rceil \frac{1-\gamma}{3}\right) \\ &\leq \exp\left(-(1-\gamma)\lambda/3\right)\end{aligned}$$

which by a standard characterization of sub-exponential random variables, see Proposition 2.7.1 in [Ver18], implies that $T_e$, and therefore also $V_{c,\geq 1}^{(i)}$ for all $i \in [B]$ are sub-exponential. Moreover, by the so-called centering lemma for sub-exponential distributions, see Section 2.7 in [Ver18], for any random variable $X$ that is sub-exponential with parameter $\sigma$, there exists an absolute constant $c$ such that $X - \mathbb{E}[X]$ is sub-exponential with parameter $c\sigma$. Thus for all $i \in [B]$, $V_{c,\geq 1}^{(i)} - \mathbb{E}[\hat{V}_{c,\geq 1}^{(i)}]$ is sub-exponential with parameter in $\mathcal{O}(1/(1-\gamma))$. Then, we apply Bernstein's inequality, see Theorem 2.8.1 in [Ver18], to show that there exist $C_1, C_2 > 0$ such that for any $\lambda \geq 0$,

$$\mathbb{P}\left(\left|\hat{V}_{c,\geq 1} - \mathbb{E}[\hat{V}_{c,\geq 1}]\right| > \frac{\lambda}{2\sqrt{B}}\right) \leq 2\exp\left(-C_1(1-\gamma)\lambda\right) + 2\exp\left(-C_2^2(1-\gamma)^2\lambda^2\right). \tag{B.17}$$

Finally, using a union bound, we combine (B.16) and (B.17) to get the desired bound. □

### B.2.4 Strong Feasibility Implies Uniform Slater's Condition

As a remark on Assumption 4.3, we claimed that this uniform Slater's condition is weaker than the strong feasibility assumption introduced in [BDL23]. Here, we provide the simple proof.

**Proposition B.13** *Let strong feasibility hold, i.e., suppose there exists $\pi \in \Pi$ such that $V_c(\pi) \leq \alpha - \frac{diam(\Pi)^2}{\eta}$. Then, the uniform Slater's condition, see Assumption 4.3, holds.*

**Proof** Clearly, for any $\pi' \in \Pi$, we have $\|\pi - \pi'\|^2 \leq \mathrm{diam}(\Pi)^2$. Therefore,

$$V_{\eta,\pi'}^c(\pi) = V^c(\pi) + \frac{1}{2\eta}\left\|\pi - \pi'\right\|^2 \leq \alpha - \frac{\mathrm{diam}(\Pi)^2}{\eta}$$

and hence the uniform Slater's condition holds with $\zeta = \frac{\mathrm{diam}(\Pi)^2}{\eta} > 0$. □

## B.3 Strongly Convex Stochastic Optimization with Strongly Convex Expectation Constraint

In this section, we describe a stochastic gradient switching algorithm for stochastic constrained optimization under expectation constraints. Up to the modification of using a relaxed constraint (which is crucial for enabling its independent implementation), our algorithm and analysis follow the Cooperative Stochastic Approximation (CSA) algorithm presented in [LZ20] which is inspired by Polyak's subgradient method [Pol67]. [LZ20] hint at the fact that a $1/K$ convergence rate of the CSA algorithm can be shown in the case of strongly convex objective and under expectation constraints. Here we explicitly carry out this analysis by deriving a result in expectation and under a somewhat weaker assumption on the distribution of the constraint function estimates.

Let $X \subseteq \mathbb{R}^d$ be a convex and compact set with diameter $\Delta := \max_{x,x' \in X} \|x - x'\|$. Suppose $\theta$ are random vectors supported on $\Theta \subset \mathbb{R}^p$, and let $F : X \times \Theta \to \mathbb{R}, G : X \times \Theta \to \mathbb{R}$ be functions such that $F(\cdot, \theta)$ and $G(\cdot, \theta)$ are $\mu_F$ and $\mu_G$-weakly convex, respectively. For any $x' \in X$, we define $F_{\mu,x'}(x,\theta) := F(x,\theta) + \mu_F \|x - x'\|^2$ and $G_{\mu,x'}(x,\theta) := G(x,\theta) + \mu_G \|x - x'\|^2$. Let $f(x) := \mathbb{E}_\theta[F(x,\theta)], g(x) := \mathbb{E}_\theta[G(x,\theta)]$ (where expectations are supposed to be well-defined and finite) and $f_{\mu,x'}(x) := f(x) + \mu_F \|x - x'\|^2, g_{\mu,x'}(x) := g(x) + \mu_G \|x - x'\|^2$ for every $x \in X$.

The problem we aim to solve[3] is given by

$$\begin{aligned} \min_{x \in X} f_{\mu,x'}(x) &:= \mathbb{E}_\theta[F_{\mu,x'}(x,\theta)] \\ \text{s.t. } g_{\mu,x'}(x) &:= \mathbb{E}_\theta[G_{\mu,x'}(x,\theta)] \leq 0. \end{aligned} \tag{B.18}$$

Recall that such a problem needs to be solved at each time step in our iProxCMPG algorithm. The point $x'$ is arbitrarily fixed throughout the rest of this section.

Suppose we are only given access to first-order information of $f_{\mu,x'}, g_{\mu,x'}$ and zeroth-order information of $g$ via a stochastic oracle that outputs unbiased and bounded-variance estimates.

**Assumption B.14** *For every $x \in X$, the estimators $F'_{\mu,x'}(x,\theta), G'_{\mu,x'}(x,\theta)$ and $G(x,\theta)$ are unbiased, i.e., $\mathbb{E}_\theta\left[F'_{\mu,x'}(x,\theta)\right] = \nabla f_{\mu,x'}(x), \mathbb{E}_\theta\left[G'_{\mu,x'}(x,\theta)\right] = \nabla g_{\mu,x'}(x)$ and $\mathbb{E}_\theta\left[G(x,\theta)\right] = g(x)$. Moreover, there exist $M_F, M_G > 0$ such that*

$$\mathbb{E}_\theta\left[\|F'(x,\theta)\|^2\right] \leq M_F^2; \quad \mathbb{E}_\theta\left[\|G'(x,\theta)\|^2\right] \leq M_G^2.$$

*Furthermore, we suppose to have access to independent unbiased estimators $\hat{G}^{(1)}, \ldots, \hat{G}^{(J)}$ of $G(x,\cdot)$ for which there exists $\sigma > 0$ such that for any $\lambda \geq 0$, it holds that*

$$\mathbb{P}_\theta\left(|\hat{G} - g(x)| > \lambda/\sqrt{J}\right) \leq 4\exp\left(-\lambda/\sigma\right) + 2\exp\left(-\lambda^2/\sigma^2\right), \tag{B.19}$$

*where $\hat{G} := \frac{1}{J}\sum_{j=1}^J \hat{G}^{(j)}$.*

It can be easily seen that Assumption B.14 also implies existence of $\tilde{M}_F, \tilde{M}_G$ such that

$$\begin{aligned} \mathbb{E}_\theta\left[\left\|F'_{\mu,x'}(x,\theta)\right\|^2\right] &\leq 2\mathbb{E}_\theta\left[\|F'(x,\theta)\|^2\right] + 2\mu_F^2\|x - x'\|^4 \leq 2M_F^2 + 2\mu_F^2\Delta^4 =: \tilde{M}_F^2, \\ \mathbb{E}_\theta\left[\left\|G'_{\mu,x'}(x,\theta)\right\|^2\right] &\leq 2\mathbb{E}_\theta\left[\|G'(x,\theta)\|^2\right] + 2\mu_G^2\|x - x'\|^4 \leq 2M_G^2 + 2\mu_G^2\Delta^4 =: \tilde{M}_G^2. \end{aligned}$$

**Remark B.15** *Notice that the concentration requirement of (B.19) is relaxed compared to the sub-Gaussian assumption made in [LZ20] which is too strong to hold in our case. We refer the reader to Lemma B.12 where we prove that this weaker tail bound assumption holds for our constraint function estimates.*

### B.3.1 A Primal Switching Gradient Algorithm

Algorithm 5 is designed as a primal algorithm that switches between taking a step along the objective or constraint gradient, depending on whether the constraint is currently (estimated to be) satisfied or not.

[3]Note that the final guarantees we will obtain are actually in terms of a relaxed constraint satisfaction bound. This is due to our modification of the original CSA algorithm.

**Algorithm 5** CSA (adapted from [LZ20])

1: **initialization:** $x_1 \in X$ s.t. $g(x_1) \leq \varepsilon$ and $\{\delta_k\}_{k\in[N]}, \{\nu_k\}_{k\in[N]}, \{\rho_k\}_{k\in[N]}, s \in [N]$
2: **for** $k = 1, \ldots, N-1$ **do**
3: sample $\hat{G}_k^{(1)}, \ldots, \hat{G}_k^{(J)}$ from $G(x_k, \cdot)$ and set $\hat{G}_k = \frac{1}{J}\sum_{j=1}^{J} \hat{G}_k^{(j)}$
4: $x_{k+1} = \begin{cases} \mathcal{P}_X\left[x_k - \nu_k F'_{\mu,x'}(x_k, \theta_k)\right] & \text{if } \hat{G}_k \leq \delta_k \\ \mathcal{P}_X\left[x_k - \nu_k G'_{\mu,x'}(x_k, \theta_k)\right] & \text{else} \end{cases}$
5: let $\mathcal{B}_s := \{s \leq k \leq N \mid \hat{G}_k \leq \delta_k\}$
6: **output:** $x_{\hat{k}}$ where $\hat{k} = 1$ if $\mathcal{B}_s = \emptyset$ and otherwise sampled s.t. for $k \in \mathcal{B}_s$,
7: $\mathbb{P}(\hat{k} = k) = \left(\sum_{k\in\mathcal{B}_s} \rho_k\right)^{-1} \rho_k$

In the analysis, we will denote $\mathcal{M}_s := \{s \leq k \leq N \mid k \notin \mathcal{B}_s\}$ and $\mathcal{B} := \mathcal{B}_1$, $\mathcal{M} := \mathcal{M}_1$.

We point out the following differences between Algorithm 5 and the original CSA algorithm, see [LZ20], Algorithm 1.

(a) We relax the switching condition in line 4 by using an estimate of $g(x_k)$ instead of $g_{\mu,x'}(x_k)$ if we were to exactly use the algorithm proposed in [LZ20]. This modification is crucial for our application as a subroutine of an *independent* learning algorithm, as described in the proof of Lemma B.2, see Appendix B.2.1. As a result, compared to [LZ20], we get a weaker guarantee in terms of constraint violation which however is still sufficient for our purposes.

(b) Instead of constructing the output as a $\rho_k$-weighted average over iterates $x_k$, we sample an iterate from a $\rho_k$-weighted distribution, see line 6. This is because our relaxed constraint function $g$ is not necessarily convex (unlike $g_{\mu,x'}$) and hence we cannot easily bound the constraint value at an average over iterates.

### B.3.2 Convergence and Sample Complexity Guarantee

The following analysis uses the techniques presented in [LZ20] applied to the strongly convex case with expectation constraint, under our modified Assumption B.14 and Algorithm 5. The proofs follow along the same lines, we highlight differences when appropriate.

First, we establish a basic recursion about CSA iterates that will be used repeatedly throughout the rest of the analysis.

**Proposition B.16** *For any $s \in [N]$, $x \in X$, and $a_s$ as defined by (B.20), it holds that*

$$\sum_{k\in\mathcal{M}_s} \rho_k \left(G_{\mu,x'}(x_k, \theta_k) - G_{\mu,x'}(x, \theta_k)\right) + \sum_{k\in\mathcal{B}_s} \rho_k \left(F_{\mu,x'}(x_k, \theta_k) - F_{\mu,x'}(x, \theta_k)\right)$$
$$\leq (1 - a_s)\Delta^2 + \frac{1}{2}\sum_{k\in\mathcal{B}_s} \rho_k \nu_k \left\|F'_{\mu,x'}(x_k, \theta_k)\right\|^2 + \frac{1}{2}\sum_{k\in\mathcal{B}_s} \rho_k \nu_k \left\|G'_{\mu,x'}(x_k, \theta_k)\right\|^2.$$

**Proof** Let $s \in [N]$ and $k \in \mathcal{B}_s$. Then, by non-expansiveness of the projection $\mathcal{P}_X$ and strong convexity,

$$\begin{aligned}
\|x_{k+1} - x\|^2 &\leq \|x_k - x\|^2 - \nu_k \left\langle F'_{\mu,x'}(x_k, \theta_k), x_k - x \right\rangle + \frac{1}{2}\nu_k^2 \left\|F'_{\mu,x'}(x_k, \theta_k)\right\|^2 \\
&\leq \|x_k - x\|^2 - \nu_k \left[F_{\mu,x'}(x_k, \theta_k) - F_{\mu,x'}(x, \theta_k) + \frac{\mu_F}{2}\|x_k - x\|^2\right] + \frac{1}{2}\nu_k^2 \left\|F'_{\mu,x'}(x_k, \theta_k)\right\|^2 \\
&\leq \left(1 - \frac{\nu_k \mu_F}{2}\right)\|x_k - x\|^2 - \nu_k \left[F_{\mu,x'}(x_k, \theta_k) - F_{\mu,x'}(x, \theta_k)\right] + \frac{1}{2}\nu_k^2 \left\|F'_{\mu,x'}(x_k, \theta_k)\right\|^2.
\end{aligned}$$

Similarly, if $k \in \mathcal{M}_s$,

$$\|x_{k+1} - x\|^2 \leq \left(1 - \frac{\nu_k \mu_G}{2}\right) \|x_k - x\|^2 - \nu_k \left[G_{\mu,x'}(x_k, \theta_k) - G_{\mu,x'}(x, \theta_k)\right] + \frac{1}{2}\nu_k^2 \left\|G'_{\mu,x'}(x_k, \theta_k)\right\|^2.$$

After defining

$$a_k = \begin{cases} \mu_F \nu_k & \text{if } k \in \mathcal{B} \\ \mu_G \nu_k & \text{if } k \in \mathcal{M} \end{cases}; \qquad A_k = \begin{cases} 1 & \text{if } k = 1 \\ (1 - a_k) A_{k-1} & \text{if } k \geq 2 \end{cases}; \qquad \rho_k = \frac{\nu_k}{A_k}; \tag{B.20}$$

the result follows by application of Lemma 21, [LZ20]. □

The next lemma provides a condition on $\{\nu_k, \delta_k, \rho_k\}_{s \leq k \leq N}$ that guarantees either low regret in terms of objective value or that a large number of iterates satisfy the constraint with high probability.

**Lemma B.17** *Let $x^*$ be an optimal solution of (B.18). If for some $s \in [N]$ and $\lambda \geq 0$,*

$$\frac{N - s + 1}{2} \min_{k \in \mathcal{M}_s} \rho_k \delta_k > (1 - a_s)\Delta^2 + \frac{1}{2} \sum_{k \in \mathcal{M}_s} \rho_k \nu_k \tilde{M}_G^2 + \frac{1}{2} \sum_{k \in \mathcal{B}_s} \rho_k \nu_k \tilde{M}_F^2 + \frac{\lambda}{\sqrt{J}} \sum_{k \in \mathcal{M}_s} \rho_k, \tag{B.21}$$

*then one of the following statements holds,*

*(a)* $\mathbb{P}_\theta(|\mathcal{B}_s| \geq (N - s + 1)/2) \geq 1 - |\mathcal{M}_s| \left(4 \exp\left(-\lambda/\sigma\right) + 2 \exp\left(-\lambda^2/\sigma^2\right)\right)$*, or,*

*(b)* $\sum_{k \in \mathcal{B}_s} \rho_k \left(f_{\mu,x'}(x_k) - f_{\mu,x'}(x^*)\right) \leq 0$.

Note that unlike in [LZ20], due to our modified choice of Algorithm 5's output, well-definedness of $x_{\hat{k}}$ does not require $\mathcal{B}_s \neq \emptyset$.

**Proof** In Proposition B.16, set $x = x^*$, take expectation w.r.t. $\theta$ on both sides, and apply $\mathbb{E}_\theta \left\|F'_{\mu,x'}(x, \theta)\right\|^2 \leq \tilde{M}_F^2$, $\mathbb{E}_\theta \left\|G'_{\mu,x'}(x, \theta)\right\|^2 \leq \tilde{M}_G^2$. Then,

$$\begin{aligned} &\sum_{k \in \mathcal{M}_s} \rho_k \left(g_{\mu,x'}(x_k) - g_{\mu,x'}(x^*)\right) + \sum_{k \in \mathcal{B}_s} \rho_k \left(f_{\mu,x'}(x_k) - f_{\mu,x'}(x^*)\right) \\ &\qquad \leq (1 - a_s)\Delta^2 + \frac{1}{2} \sum_{k \in \mathcal{M}_s} \rho_k \nu_k \tilde{M}_G^2 + \frac{1}{2} \sum_{k \in \mathcal{B}_s} \rho_k \nu_k \tilde{M}_F^2. \end{aligned} \tag{B.22}$$

If $\sum_{k \in \mathcal{B}_s} \rho_k \left(f_{\mu,x'}(x_k) - f_{\mu,x'}(x^*)\right) \leq 0$, then (b) holds. Otherwise, we make three observations. First, we have that $g_{\mu,x'}(x^*) \leq 0$. Second, it holds that $g(x_k) \leq g_{\mu,x'}(x_k)$. Third, for $k \in \mathcal{M}_s$, by Assumption B.14 and due to $\hat{G}_k > \delta_k$, we get

$$\mathbb{P}_\theta \left(g(x_k) < \delta_k - \frac{\lambda}{\sqrt{J}}\right) \leq 4 \exp\left(-\lambda/\sigma\right) + 2 \exp\left(-\lambda^2/\sigma^2\right). \tag{B.23}$$

By a union bound this inequality holds for all $k \in \mathcal{M}_s$ with probability at most $|\mathcal{M}_s| \left(4 \exp\left(-\lambda/\sigma\right) + 2 \exp\left(-\lambda^2/\sigma^2\right)\right)$. Combining these three observations with (B.22) yields that with probability at least $1 - |\mathcal{M}_s| \left(4 \exp\left(-\lambda/\sigma\right) + 2 \exp\left(-\lambda^2/\sigma^2\right)\right)$, it holds that

$$\sum_{k \in \mathcal{M}_s} \rho_k \delta_k \leq (1 - a_s)\Delta^2 + \frac{1}{2} \sum_{k \in \mathcal{M}_s} \rho_k \nu_k \tilde{M}_G^2 + \frac{1}{2} \sum_{k \in \mathcal{B}_s} \rho_k \nu_k \tilde{M}_F^2 + \frac{\lambda}{\sqrt{J}}.$$

Above inequality then implies (a) because if $|\mathcal{B}_s| < (N - s + 1)/2$, i.e., $|\mathcal{M}_s| \geq (N - s + 1)/2$, then condition (B.21) implies that

$$\sum_{k \in \mathcal{M}_s} \rho_k \delta_k \geq \frac{N - s + 1}{2} \min_{k \in \mathcal{M}_s} \rho_k \delta_k > (1 - a_s)\Delta^2 + \frac{1}{2} \sum_{k \in \mathcal{M}_s} \rho_k \nu_k \tilde{M}_G^2 + \frac{1}{2} \sum_{k \in \mathcal{B}_s} \rho_k \nu_k \tilde{M}_F^2 + \frac{\lambda}{\sqrt{J}} \sum_{k \in \mathcal{M}_s} \rho_k,$$

which is a contradiction. □

Next, we state and prove the main guarantees provided by Algorithm 5.

**Theorem B.18** *Under Assumption B.14, let $\varepsilon > 0$, suppose $x_1$ is such that $g(x_1) \le \varepsilon$, and let $f_{\max} > 0$ such that for all $x \in X$, $0 \le f(x) \le f_{\max}$. Choose $s = N/2, \lambda = \sigma^2 \log(N^2/(4f_{\max}))$, set $M = \max\{\tilde{M}_G, \tilde{M}_F\}$, $\mu = \min\{\mu_G, \mu_F\}$, and*

$$\nu_k = \begin{cases} \frac{2}{\mu_F(k+1)} & \text{if } k \in \mathcal{B} \\ \frac{2}{\mu_G(k+1)} & \text{if } k \in \mathcal{M} \end{cases}; \qquad \delta_k = \frac{\lambda}{\sqrt{J}} + \frac{1}{2k}\left(\frac{4\Delta^2}{k} + \frac{16M^2}{\mu^2}\right) \cdot \begin{cases} \mu_F & \text{if } k \in \mathcal{B} \\ \mu_G & \text{if } k \in \mathcal{M} \end{cases};$$

$$a_k = \begin{cases} \mu_F \nu_k & \text{if } k \in \mathcal{B} \\ \mu_G \nu_k & \text{if } k \in \mathcal{M} \end{cases}; \qquad A_k = \begin{cases} 1 & \text{if } k = 1 \\ (1-a_k)A_{k-1} & \text{if } k \ge 2 \end{cases}; \qquad \rho_k = \frac{\nu_k}{A_k}$$

$$N = \max\left\{\frac{64\mu_F M^2}{\mu^2\varepsilon^2}, \frac{\sqrt{32\Delta^2\mu_F}}{\varepsilon}, \frac{32\sigma\mu_F}{\mu\varepsilon^2}\right\}; \qquad J = \max\left\{\frac{9\lambda^2}{\varepsilon^2}, \frac{32\sigma\mu_F}{\mu\varepsilon^2}\right\}.$$

*Then Algorithm 5 guarantees that*

$$\mathbb{E}\left[f_{\mu,x'}(x_{\hat{k}}) - f_{\mu,x'}(x^*)\right] \le \varepsilon^2, \tag{B.24}$$

$$\mathbb{E}\left[g(x_{\hat{k}})\right] \le \varepsilon. \tag{B.25}$$

**Proof** First, we observe that for any $k \in \mathcal{M}_s$,

$$\begin{aligned} \mathbb{E}\left[\sqrt{J}\left(g(x_k) - \delta_k\right)\right] &= \int_0^\infty \left(1 - \mathbb{P}\left(\sqrt{J}\left(g(x_k) - \delta_k\right) \le z\right)\right) dz \\ &\quad - \int_{-\infty}^0 \mathbb{P}\left(\sqrt{J}\left(g(x_k) - \delta_k\right) \le z\right) dz \\ &\ge -\int_{-\infty}^0 4\exp\left(z/\sigma\right) + 2\exp\left(z^2/\sigma^2\right) dz \\ &\ge -6\sigma \end{aligned} \tag{B.26}$$

where the first inequality is by (B.23). Therefore, we have $\mathbb{E}[g(x_k)] \ge \delta_k - \frac{6\sigma}{\sqrt{J}}$. Moreover, by an argument similar to our derivation in Lemma B.12 but with Bernstein's inequality applied to the sum $\left(\sum_{k\in\mathcal{M}_s}\rho_k\right)^{-1}\sum_{k\in\mathcal{M}_s}\rho_k\hat{G}_k$,

$$\mathbb{P}\left(\sum_{k\in\mathcal{M}_s}\rho_k\left(\hat{G}_k - g(x_k)\right) > \frac{\lambda}{\sqrt{J|\mathcal{M}_s|}}\sum_{k\in\mathcal{M}_s}\rho_k\right) \le 4\exp\left(-\lambda/\sigma\right) + 2\exp\left(-\lambda^2/\sigma^2\right).$$

Therefore, following (B.26), we get

$$\mathbb{E}\left[\sum_{k\in\mathcal{M}_s}\rho_k g(x_k)\right] \ge \sum_{k\in\mathcal{M}_s}\rho_k\delta_k - \frac{6\sigma}{\sqrt{J|\mathcal{M}_s|}}\sum_{k\in\mathcal{M}_s}\rho_k. \tag{B.27}$$

Next, we derive (B.24). Note that (B.21) holds for our choices of $s, \nu_k, \delta_k, \rho_k$. Then, if part (b) of Lemma B.17 holds, we have

$$\begin{aligned} \mathbb{E}\left[f(x_{\hat{k}}) - f(x^*)\right] &= \mathbb{E}_{\hat{k}}\left[\mathbb{E}\left[f(x_k) - f(x^*) \mid \hat{k} = k\right]\right] \\ &\le \left(\sum_{k\in\mathcal{B}_s}\rho_k\right)^{-1}\sum_{k\in\mathcal{B}_s}\rho_k\mathbb{E}\left[f(x_k) - f(x^*)\right] \\ &\le 0. \end{aligned}$$

Otherwise, if part (a) holds, then using the above bound on $\mathbb{E}[g(x_k)]$ together with the convexity of $f_{\mu,x'}$, (B.22) and (B.27), it follows that

$$\begin{aligned}
&\sum_{k\in\mathcal{M}_s}\rho_k\delta_k - \frac{6\sigma}{\sqrt{J|\mathcal{M}_s|}}\sum_{k\in\mathcal{M}_s}\rho_k + \sum_{k\in\mathcal{B}_s}\rho_k\mathbb{E}\left[f_{\mu,x'}(x_{\hat{k}}) - f_{\mu,x'}(x^*)\right] \\
&\quad\le \sum_{k\in\mathcal{M}_s}\rho_k\mathbb{E}\left[g(x_k)\right] + \sum_{k\in\mathcal{B}_s}\rho_k\mathbb{E}\left[f_{\mu,x'}(x_{\hat{k}}) - f_{\mu,x'}(x^*)\right] \\
&\quad\le \sum_{k\in\mathcal{M}_s}\rho_k\mathbb{E}\left[g(x_k)\right] + \sum_{k\in\mathcal{B}_s}\rho_k\mathbb{E}\left[f_{\mu,x'}(x_k) - f_{\mu,x'}(x^*)\right] \\
&\quad\le (1-a_s)\Delta^2 + \frac{1}{2}\sum_{k\in\mathcal{M}_s}\rho_k\nu_k\tilde{M}_G^2 + \frac{1}{2}\sum_{k\in\mathcal{B}_s}\rho_k\nu_k\tilde{M}_F^2.
\end{aligned}$$

Denote by $E_{\mathcal{B}_s}$ the event that $|\mathcal{B}_s| \ge (N-s+1)/2$. Then, using the law of total expectation, our choice of $\lambda = \sigma^2\log(N^2/(4f_{\max}))$, $\rho_k\delta_k \ge 0$, and above inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[f(x_{\hat{k}}) - f(x^*)\right] \\
&\quad\le \mathbb{E}\left[f(x_{\hat{k}}) - f(x^*) \mid E_{\mathcal{B}_s}\right]\cdot\underbrace{\mathbb{P}\left(E_{\mathcal{B}_s}\right)}_{\le 1} + \mathbb{E}\left[f(x_{\hat{k}}) - f(x^*) \mid \overline{E}_{\mathcal{B}_s}\right]\cdot\underbrace{\mathbb{P}\left(\overline{E}_{\mathcal{B}_s}\right)}_{\le|\mathcal{M}_s|(4\exp(-\lambda/\sigma)+2\exp(-\lambda^2/\sigma^2))} \\
&\quad\le \left(\sum_{k\in\mathcal{B}_s}\rho_k\right)^{-1}\left((1-a_s)\Delta^2 + \frac{1}{2}\sum_{k\in\mathcal{M}_s}\rho_k\nu_k\tilde{M}_G^2 + \frac{1}{2}\sum_{k\in\mathcal{B}_s}\rho_k\nu_k\tilde{M}_F^2 + \frac{6\sigma}{\sqrt{J|\mathcal{M}_s|}}\sum_{k\in\mathcal{M}_s}\rho_k\right) + \frac{1}{N} \\
&\quad\le \left(\frac{N-s+1}{2}\min_{k\in\mathcal{B}_s}\rho_k\right)^{-1}\left((1-a_s)\Delta^2 + \frac{1}{2}\sum_{k\in\mathcal{M}_s}\rho_k\nu_k\tilde{M}_G^2 + \frac{1}{2}\sum_{k\in\mathcal{B}_s}\rho_k\nu_k\tilde{M}_F^2 + \frac{6\sigma}{\sqrt{J|\mathcal{M}_s|}}\sum_{k\in\mathcal{M}_s}\rho_k\right) + \frac{1}{N}.
\end{aligned}$$

In order to show the constraint violation bound, note that by a similar argument as (B.26), for any $k \in \mathcal{B}_s$, $\mathbb{E}[g(x_k)] \le \delta_k + \frac{6\sigma}{\sqrt{J}}$, and therefore

$$\mathbb{E}\left[g(x_{\hat{k}})\right] = \mathbb{E}_{\hat{k}}\left[\mathbb{E}\left[g(x_k) \mid k = \hat{k}\right]\right] \le \mathbb{E}_{\hat{k}}\left[\delta_{\hat{k}}\right] + \frac{6\sigma}{\sqrt{J}} = \frac{\sum_{k\in\mathcal{B}_s}\rho_k\delta_k}{\sum_{k\in\mathcal{B}_s}\rho_k} + \frac{6\sigma}{\sqrt{J}}.$$

In order to derive the guarantees (B.24) and (B.25), we plug in the choices of $\nu_k, \delta_k, a_k, \rho_k, N$, and $J$ stated in Theorem B.18. Observing that for any $s \le k \le N$, we have $A_k = \frac{2}{k(k+1)}$, that for any $k \in \mathcal{B}$, we have $\rho_k = \frac{2k}{\mu_F}$ as well as $\rho_k\nu_k = \frac{4}{\mu_F^2}$, and for any $k \in \mathcal{M}$, $\rho_k = \frac{2k}{\mu_G}, \rho_k\nu_k = \frac{4}{\mu_G^2}$.

$$\begin{aligned}
\mathbb{E}\left[f(x_{\hat{k}}) - f(x^*)\right] &\le \frac{\Delta^2 + 2N\mu_F^{-2}\tilde{M}_F^2 + 2N\mu_G^{-2}\tilde{M}_G^2 + \frac{2\sigma}{\sqrt{J}}N^{3/2}\mu_G^{-1}}{N^2/4\cdot\mu_F^{-1}} \\
&\le \frac{\Delta^2 + 4N\mu^{-2}M^2 + \frac{2\sigma}{\sqrt{J}}N^{3/2}\mu^{-1}}{N^2/4\cdot\mu_F^{-1}} + \frac{1}{N} \\
&\le \frac{4\mu_F\Delta^2}{N^2} + \frac{16\mu_F\mu^{-2}M^2}{N} + \frac{8\sigma\mu^{-1}\mu_F}{\sqrt{JN}} + \frac{1}{N} \\
&\le \varepsilon^2/4 + \varepsilon^2/4 + \varepsilon^2/4 + \varepsilon^2/4.
\end{aligned}$$

Moreover, for the constraint bound, it holds that

$$\begin{aligned}\mathbb{E}\left[g(x_{\hat{k}})\right] &\leq \frac{\sum_{k\in\mathcal{B}_s}\left(4\Delta^2/k + 16M^2/\mu^2\right)}{\sum_{k\in\mathcal{B}_s} 2k/\mu_F} + \frac{6\sigma}{\sqrt{J}} \\ &\leq \frac{8\Delta^2\mu_F}{N^2} + \frac{16M^2\mu_F}{\mu^2 N} + \frac{6\sigma}{\sqrt{J}} \\ &\leq \varepsilon.\end{aligned}$$

□

## B.4 Background in Constrained Optimization and a Novel Technical Lemma

**Notation.** For any non-empty subset $Y \subset \mathbb{R}^d$ and any vector $x \in \mathbb{R}^d$, the distance from $x$ to the set $Y$ is defined as $\mathrm{dist}(x, Y) := \inf_{y\in Y} \|x - y\|$ where $\|\cdot\|$ is the standard 2-norm of the Euclidean space $\mathbb{R}^d$.

In this section, we recall some useful definitions for constrained optimization. In particular, we recall the definition of an approximate Karush-Kuhn-Tucker (KKT) point and a variation thereof. Then we prove a new technical result that will be useful in our analysis.

### B.4.1 Approximate KKT Points in Constrained Optimization

Let $X \subset \mathbb{R}^d$ be a closed convex set. Consider the following constrained optimization problem:

$$\begin{aligned} P^* = \min_{x\in X} & f(x) \\ \text{s.t. } & f_c(x) \leq 0\,, \end{aligned} \tag{ConstrOpt}$$

where $f, f_c : X \to \mathbb{R}$ are differentiable (possibly nonconvex) functions.

The associated Lagrangian function $\mathcal{L} : X \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ is defined for any $x \in X, \lambda \geq 0$ by $\mathcal{L}(x, \lambda) = f(x) + \lambda f_c(x)$. The primal and dual problems can be written as

$$\begin{aligned} P^* &= \inf_{x\in X} \sup_{\lambda\geq 0} \mathcal{L}(x, \lambda)\,, \\ D^* &= \sup_{\lambda\geq 0} \underbrace{\inf_{x\in X} \mathcal{L}(x, \lambda)}_{=:d(\lambda)}\,. \end{aligned}$$

By weak duality, we know that $P^* \geq D^*$.

For any $x \in \mathbb{R}^d$, the normal cone to the set $X$ at $x$ is defined by:

$$N_X(x) := \left\{ g \in \mathbb{R}^d \mid \forall y \in X, \langle g, y - x\rangle \leq 0 \right\}.$$

**Definition B.19** *Let $\varepsilon \geq 0$. A point $x \in X$ is an $\varepsilon$-KKT point of* (ConstrOpt) *if there exists a real $\lambda$ such that the following conditions hold:*

$$\begin{aligned} f_c(x) &\leq \varepsilon\,, && \text{(primal feasibility)} \\ \lambda &\geq 0\,, && \text{(dual feasibility)} \\ |\lambda f_c(x)| &\leq \varepsilon\,, && \text{(complementary slackness)} \\ \mathit{dist}\left(\nabla_x \mathcal{L}(x, \lambda), -N_X(x)\right) &\leq \varepsilon\,. && \text{(Lagrangian stationarity)} \end{aligned}$$

*We also call $(x, \lambda)$ an $\varepsilon$-KKT pair. The point $x$ is simply a KKT point of* (ConstrOpt) *if moreover $\varepsilon = 0$.*

We additionally define a slight modification of the above standard KKT conditions which turns out to be useful in our analysis. More precisely, the definition replaces approximate Lagrangian stationarity by a variational form thereof. Moreover, primal feasibility is now supposed to be exact. Other conditions remain unchanged.

**Definition B.20** *Let $\varepsilon \geq 0$. A point $\tilde{x} \in X$ is an $\varepsilon$-$\widetilde{KKT}$ point of* (ConstrOpt) *if there exists a real $\tilde{\lambda}$ such that the following conditions hold:*

$$
\begin{aligned}
f_c(\tilde{x}) &\leq 0, && \text{(exact primal feasibility)}\\
\tilde{\lambda} &\geq 0, && \text{(dual feasibility)}\\
\left|\tilde{\lambda} f_c(\tilde{x})\right| &\leq \varepsilon, && \text{(complementary slackness)}\\
\max_{x' \in X} \left\langle \tilde{x} - x', \nabla_x \mathcal{L}(\tilde{x}, \tilde{\lambda}) \right\rangle &\leq \varepsilon. && \text{(variational Lagrangian stationarity)}
\end{aligned}
$$

*In particular, the point $\tilde{x}$ is said to be a $\widetilde{KKT}$ point of* (ConstrOpt) *when $\varepsilon = 0$.*

The next lemma connects the first stationarity condition with a variational form thereof. In particular, this result allows us to connect the two definitions of approximate KKT points above.

**Lemma B.21** *Let $X \subseteq \mathbb{R}^d$ be a convex and compact set. Let $\varepsilon > 0$ and let $x, g \in \mathbb{R}^d$. If dist $(g, -N_X(x)) \leq \varepsilon$, then $\max_{x' \in X} \langle x - x', g \rangle \leq \Delta\varepsilon$, where $\Delta := \max_{x, x' \in X} \|x - x'\|$ is the diameter of the set $X$.*

**Proof** Let $y_0 \in -N_X(x)$. For any $x' \in X$, we have

$$
\begin{aligned}
\langle x - x', g \rangle &= \langle x - x', g - y_0 \rangle + \langle -y_0, x' - x \rangle, \\
&\leq \langle x' - x, g - y_0 \rangle, \\
&\leq \|x' - x\| \cdot \|g - y_0\|,
\end{aligned}
$$

where the first inequality follows from the fact that $y_0 \in -N_X(x)$, the second inequality stems from the Cauchy-Schwarz inequality. Taking the infimum with respect to $y_0$ in the last inequality gives the desired inequality since dist $(g, -N_X(x)) = \inf_{y \in -N_X(x)} \|g - y\| \leq \varepsilon$. □

### B.4.2 A Novel Technical Lemma for Approximate Optimality Under Gradient Dominance

We now state our technical lemma. This result shows that an approximate KKT point of (ConstrOpt) at which a gradient domination inequality holds for the Lagrangian function is approximately optimal for the objective function to be minimized.

**Proposition B.22** *Let $\varepsilon > 0$ and let $\tilde{x} \in X$ be an $\varepsilon$-$\widetilde{KKT}$ point of* (ConstrOpt)*. Suppose there exist constants $C_0, C_1 \geq 0$ such that the Lagrangian function associated to* (ConstrOpt) *satisfies for all $\lambda \geq 0$ and for all $x \in X$,*

$$
\mathcal{L}(x, \lambda) - \mathcal{L}(x_\lambda^*, \lambda) \leq C_0 \max_{x' \in X} \left\langle x - x', \nabla_x \mathcal{L}(\tilde{x}, \tilde{\lambda}) \right\rangle + C_1 \varepsilon, \tag{B.28}
$$

*where $x_\lambda^*$ is a minimizer of $\mathcal{L}(\cdot, \lambda)$. Then, we have*

$$
f(\tilde{x}) - P^* \leq (C_0 + C_1 + 1)\varepsilon.
$$

**Proof** Let $(\tilde{x}, \tilde{\lambda})$ be an $\varepsilon$-$\widetilde{\text{KKT}}$ pair. Then, we have

$$\begin{aligned} D^* &\overset{(a)}{=} \max_{\lambda \geq 0} d(\lambda) \geq d(\tilde{\lambda}) \\ &\overset{(b)}{=} \min_{x \in X} \mathcal{L}(x, \tilde{\lambda}) \\ &\overset{(c)}{\geq} \mathcal{L}(\tilde{x}, \tilde{\lambda}) - (C_0 + C_1)\varepsilon \\ &= f(\tilde{x}) + \tilde{\lambda} f_c(\tilde{x}) - (C_0 + C_1)\varepsilon \\ &\overset{(d)}{\geq} f(\tilde{x}) - (C_0 + C_1 + 1)\varepsilon \end{aligned}$$

where (a) and (b) are by definition, and (d) is due to complementary slackness. To see (c), observe that by Lagrangian stationarity and (B.28),

$$\varepsilon \geq \max_{x' \in X} \left\langle \tilde{x} - x', \nabla_x \mathcal{L}(\tilde{x}, \tilde{\lambda}) \right\rangle \geq \frac{1}{C_0} \left( \mathcal{L}(\tilde{x}, \tilde{\lambda}) - \mathcal{L}(x^*_{\tilde{\lambda}}, \tilde{\lambda}) - C_1 \varepsilon \right) ,$$

which implies that[4]

$$\mathcal{L}(\tilde{x}, \tilde{\lambda}) - \mathcal{L}(x^*_{\tilde{\lambda}}, \tilde{\lambda}) \leq (C_0 + C_1)\varepsilon.$$

Finally, we use weak duality, i.e. $P^* \geq D^*$, to conclude that

$$f(\tilde{x}) - P^* = \underbrace{f(\tilde{x}) - D^*}_{\leq (C_0 + C_1 + 1)\varepsilon} + \underbrace{D^* - P^*}_{\leq 0} \leq (C_0 + C_1 + 1)\varepsilon. \qquad \square$$

## B.5 Additional Details About Simulations

We provide additional details regarding the implementation of our iProxCMPG (Algorithm 4) in practice:

(a) In our experiments, each episode terminates after a fixed number of steps $T_e = 10$ corresponding to a discount factor $\gamma = 0.9$.

(b) In order to reduce the variance and enable the usage of larger step sizes, all constraint and value (gradient) estimates are obtained by sampling a batch of $B$ trajectories.

(c) For the subroutine, i.e. as a solution to the proximal-point update, we do not consider a $\rho_k$-weighted average over iterates but simply use the last iterate $\pi^{(t,K)}$.

(d) We choose $\delta_k = 0$ for all $k \in \mathbb{N}$.

**Notation.** As used in the main part, $\mathcal{U}(\{1, \cdots, W\})$ refers to the uniform distribution over the finite set $\{1, \cdots, W\}$ where $W \geq 2$ is an integer.

**Hyperparameters.** We report hyperparameter choices for our simulations in Table B.1. Note that to ensure convergence, as indicated by our theoretical results, a larger number of players $m$ requires smaller step sizes and larger sample batches.

[4] If $C_0 = 0$, the same inequality immediately holds from (B.28).

| Hyperparameters | Number of players $m$ | Pollution tax | Energy marketplace |
|---|---|---|---|
| Step size $\eta$ (outer loop) | - | 0.1 | 0.1 |
| Step size $\nu_k$ (inner loop) | 2 | 0.005 | 0.002 |
| | 4 | 0.002 | 0.001 |
| | 8 | 0.0007 | 0.0003 |
| Sample batch size $B$ | 2 | 1000 | 100 |
| | 4 | 1000 | 150 |
| | 8 | 2500 | 200 |
| $K$ (#iterations inner loop) | - | 20 | 20 |
| $T$ (#iterations outer loop) | - | 20 | 60 |
| Discount factor $\gamma$ | - | 0.9 | 0.9 |
| Episode length $T_e$ | - | 10 | 10 |

**Table B.1:** Overview of hyperparameters used in our simulations.

**Error Bars and Reproducibility.** The plots in Figs. 4.1 and 4.3 show the means of estimated potential values across 10 independent runs, and the corresponding shaded region displays the respective standard deviation. Obtaining results for all presented experiments thus requires simulating 60 runs in total. All experiments are fully reproducible using the provided code and specified seeds.

**Computing Infrastructure.** In order to reduce computation time by executing all runs in parallel, we conducted the simulations within less than 4 hours on a cluster of 15 4-core *Intel(R) Xeon(R) CPU E3-1284L v4* clocked at 2.90GHz and equipped with 8Gbs of memory.

# Acknowledgments

First, I would like to thank Prof. Niao He for giving me the great opportunity to conduct my thesis within the Optimization and Decision Intelligence Group in the Department of Computer Science at ETH Zürich.

Moreover, I sincerely thank Dr. Anas Barakat for his invaluable insights and guidance. His profound interest, infectious enthusiasm, and passion for gaining a deep understanding of the subject matter have greatly inspired me throughout my time working on this thesis. Investing the time and effort to let our frequent discussions wander from high-level ideas down to technical nuances was not just fun but also absolutely vital for turning vague intuitions into concrete algorithms and proofs that otherwise may have never seen the light of day. I admire his relentless optimism and patience in the face of the series of obstacles and challenges we encountered, which motivated me to keep seeking and ultimately finding solutions. Therefore, I see this thesis and the findings presented therein as the result of a truly joint effort. This holds true in particular for the contents of Chapter 4, which led to the paper "Independent Learning in Constrained Markov Potential Games". The lessons learned during the process of collaboratively crafting and submitting this paper are invaluable to me, and I am very grateful for getting this opportunity while working on my thesis.