# Provably Efficient Algorithms for Multi-Objective Competitive RL

**Tiancheng Yu** [1] **Yi Tian** [1] **Jingzhao Zhang** [1] **Suvrit Sra** [1]

## Abstract

We study multi-objective reinforcement learning (RL) where an agent's reward is represented as a vector. In settings where an agent competes against opponents, its performance is measured by the distance of its average return vector to a target set. We develop statistically and computationally efficient algorithms to approach the associated target set. Our results extend Blackwell's approachability theorem (Blackwell, 1956) to tabular RL, where strategic exploration becomes essential. The algorithms presented are adaptive; their guarantees hold even without Blackwell's approachability condition. If the opponents use fixed policies, we give an improved rate of approaching the target set while also tackling the more ambitious goal of simultaneously minimizing a scalar cost function. We discuss our analysis for this special case by relating our results to previous works on constrained RL. To our knowledge, this work provides the first provably efficient algorithms for vector-valued Markov games and our theoretical guarantees are near-optimal.

## 1. Introduction

What can a player expect to achieve in competitive games when pursuing multiple objectives? If the player has a single objective, the answer is clear from von Neumann's minimax theorem (Neumann, 1928): the player can follow a fixed strategy to ensure that its cost is no worse than a certain threshold, the *minimax value* of the game, no matter how the opponents play. But if the player has multiple objectives, the answer is less clear and it must define some tradeoffs. One important way to capture tradeoffs is to define a certain *target set* of vectors, and then to ensure that player's vector of returns lies in this set. The player's performance can then be measured via the distance of its reward vector from the target set. In 1956, Blackwell showed that in a repeated game, the player of interest can make the distance of its average return to a target set small as long as this set satisfies a condition called *approachability* (Blackwell, 1956).

The approachability theorem applies to multi-objective games with a decision horizon of a *single* time step. However, in many practical domains such as robotics, self-driving, video games, and recommendation systems, the decision horizons span *multiple* time steps. For example, in a robot control task, we may hope the robot arm reaches a certain region in a 3D space; while, in self-driving, we may hope the car takes care of speed, safety and comfort simultaneously. In these problems, the state of the decision process transitions based on both the actions taken by the players and the unknown dynamics. Though a generalization (Assumption 3) of Blackwell's approachability condition Blackwell (1956) is relatively direct, efficient exploration and the need to learn the unknown transitions is what poses a challenge in the multiple time step setting.

This challenge motivates us to ask: *How can a player approach a target set that satisfies a generalized notion of approachability?* We answer this question by modeling multi-objective competitive reinforcement learning (RL) as an online learning problem in a vector-valued Markov game (MG), for which we provide efficient algorithms as instances of a generic meta-algorithm that we propose.

Going one step further, we can ask a more ambitious question: *Can we minimize a scalar cost function while also satisfying approachability?* Our answer is affirmative if the opponents play fixed policies; equivalently, if the agent interacts with a fixed environment (without opponents), in which case the model reduces to a vector-valued Markov decision process (MDP). In this setting, the target set can be viewed as a set of constraints, and our results improve on the rich literature on constrained MDP in multiple aspects.

In Table 1 we give a comparison of different multi-objective RL settings. Our work can be seen as a generalization of both (Blackwell, 1956) and (Agrawal & Devanur, 2014) to cases with an $H$ step horizon.

**Summary of our contributions.**

- For online learning in vector-valued Markov games, we propose two provably efficient algorithms to approach a target set under a generic framework. Strategic exploration is essential to obtain statistical efficiency (Theo-

[1]Department of EECS, MIT, Cambridge, USA. Correspondence to: Suvrit Sra <suvrit@mit.edu>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. 

*Table 1.* The settings of this work with reference to the literature

| | w/o opponents | w/ adversarial opponents |
|---|---|---|
| single-state single-horizon | constrained bandits (e.g., (Agrawal & Devanur, 2014)) | vector-valued games (e.g., (Blackwell, 1956)) |
| multi-state $H$-horizon | constrained MDPs (e.g., (Brantley et al., 2020); **this work**) | vector-valued Markov games (**this work**) |

rems 1 and 3) for both algorithms. The second algorithm has the merit of being more computationally efficient.

- When the chosen target set is not approachable, both our algorithms adapt automatically. Concretely, we describe the guarantees (Theorems 2 and 3) of the algorithms using a notion of $\delta$-approachability (Assumption 4).
- For vector-valued MDPs, via a more dedicated design of the exploration bonus, we obtain a near-optimal rate of making the average reward vector approach (Theorem 4) the target set. Moreover, under a mild assumption, we present a modified algorithm that can simultaneously minimize a convex cost function (Theorem 5). Comparing with existing results in constrained MDP, our bounds on regret and constraint violation are the sharpest with respect to their dependence on the parameters $S$, $A$, and $K$, where $S$ is the number of states, $A$ is the number of actions and $K$ is the number of episodes.

### 1.1. Related Work

**Blackwell's approachability.** Blackwell (1956) initiated the study of multi-objective learning in repeated matrix games by introducing the notion of approachability and an algorithm to approach a given set. Using a dual formulation of the distance from a point to a convex cone, Abernethy et al. (2011) show the equivalence of approachability problems and online linear optimization. Shimkin (2016) further extends the equivalence to online convex optimization (OCO) via a dual formulation of the distance from a point to a convex set. Our primal and dual algorithms generalize respectively Blackwell's algorithm (Blackwell, 1956) and the OCO-based algorithm (Shimkin, 2016) to Markov games.

**Learning in Markov games.** Markov games, also known as stochastic games (Shapley, 1953; Littman, 1994), are a general model for multi-agent reinforcement learning. In recent years, much attention has been given to learning in scalar-valued Markov games with unknown transitions. One popular aplication is to study constrained RL as a MG (Miryoosefi et al., 2019).

In the self-play setting (Bai & Jin, 2020; Xie et al., 2020; Bai et al., 2020; Liu et al., 2020), the goal is to learn a Nash equlibrium with sample complexity guarantees. Bai & Jin (2020); Xie et al. (2020); Bai et al. (2020) consider zero-sum Markov games while Liu et al. (2020) provide results for general-sum Markov games. In the online setting (Brafman & Tennenholtz, 2002; Xie et al., 2020; Tian et al., 2020b), the goal is to achieve low regret in presence of an adversarial opponent. We also study the online setting, but in contrast, we consider vector-valued returns and the goal is to make the average return approach a given set.

**Online learning with constraints.** Multi-objective RL is closely related to RL with constraints since satisfying the constraints is tantamount to having extra objectives. Badanidiyuru et al. (2013) study bandits with knapsacks, and Agrawal & Devanur (2014) study the more general setting with concave rewards and convex constraints that the method needs to approach. Beyond bandits, Jenatton et al. (2016); Yuan & Lamperski (2018) study online convex optimization with constraints given by convex functions.

**Constrained MDPs.** For MDPs with linear constraints, Efroni et al. (2020); Ding et al. (2020); Qiu et al. (2020); Brantley et al. (2020) provide algorithms with both regret and total constraint violation guarantees. As a generalization of (Agrawal & Devanur, 2014), Brantley et al. (2020) also consider MDPs with convex constraints and concave rewards and discuss as a special case MDPs with knapsacks on *all* episodes. Chen et al. (2020) formulate MDPs with knapsacks on *each* episode as factored MDPs, to which the regret bounds of factored MDPs (Osband & Van Roy, 2014; Tian et al., 2020a; Chen et al., 2020) apply. See the discussion at the end of Section 6.1 for a detailed comparison.

**Multi-objective RL with preference.** More recently, Wu et al. (2020) study single-agent multi-objective RL to accommodate potentially adversarial preference vectors. In contrast, we assume a potentially adversarial opponent that influences both the transition and the return vector. Their goal also differs from ours in that they aim to maximize the cumulative rewards defined by the observed preference vectors in each episode. The preference vector in their setting is similar to the dual variable in our algorithm. Nonetheless, our dual variable is learned by an update procedure.

All of the aforementioned works on MGs or MDPs focus on the episodic setting. See, e.g., (Cheung et al., 2019; Singh et al., 2020), for the studies of multi-objective or constrained RL in the nonepisodic setting.

## 2. Background and Problem Setup

In this section, we formulate the problem of two-player zero-sum Markov Games. We control one of the players, whom we call the *agent*. The other player is referred to as the *adversary*. We use the two-player zero-sum condition for simplicity. We can handle multi-player general-sum games by considering the product of all the opponents' actions as an augmented action (an idea also recently exploited in (Tian et al., 2020b)). Now we are ready to explain how players interact and learn in the Markov game setup.

### 2.1. Vector-valued Markov Games

**Model.** Let $[N] := \{1, 2, \dots, N\}$, and let $\Delta(\mathbb{X})$ be the set of probability distribution on set $\mathbb{X}$. Then, an episodic two-player zero-sum vector-valued MG can be denoted by the tuple $\mathrm{MG}(\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathbb{P}, \mathbf{r}, H)$, where

- $H$ is the number of steps in each episode,
- $\mathcal{S}$ is the state space,
- $\mathcal{A}$ and $\mathcal{B}$ are the action spaces of both players,
- $\mathbb{P}$ is a collection of *unknown* transition kernels $\{\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})\}_{h\in[H]}$, and
- $\mathbf{r}$ is a collection of *known* $d$-dimensional return functions $\{\mathbf{r}_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0, 1]^d\}_{h\in[H]}$, where $d \geq 2$ is the *dimensionality* of the MG. We assume known $\mathbf{r}$ only for simplicity; learning $\mathbf{r}$ poses no real difficulty–see e.g., (Azar et al., 2017; Jin et al., 2018).

Let $|\cdot|$ denote set cardinality. Then, we define the three key cardinalities $S := |\mathcal{S}|$, $A := |\mathcal{A}|$, and $B := |\mathcal{B}|$.

**Interaction protocol.** Without loss of generality, in each episode the MG starts at a *fixed* initial state $s_1 \in \mathcal{S}_1$. At each step $h \in [H]$, the two players observe the state $s_h \in \mathcal{S}$ and simultaneously take actions $a_h \in \mathcal{A}$, $b_h \in \mathcal{B}$. This decision is specified by the players' policies $\mu_h(s_h) \in \Delta(\mathcal{A})$ and $\nu_h(s_h) \in \Delta(\mathcal{B})$. Then the environment transitions to the next state $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h, b_h)$ and outputs the return $\mathbf{r}_h(s_h, a_h, b_h)$. Let $\mathcal{F}_h^k$ be the filtration generated by all these random variables until the $k$-th episode and $i$-th step.

**Value functions.** Analogous to usual MDPs, for a policy pair $(\mu, \nu)$, step $h \in [H]$, state $s \in \mathcal{S}$ and actions $a \in \mathcal{A}, b \in \mathcal{B}$, we define the State- and Q-value functions as:

$$\mathbf{V}_h^{\mu,\nu}(s) := \mathbb{E}_{\mu,\nu}\Big[\sum_{l=h}^{H} \mathbf{r}_l(s_l, a_l, b_l)|s_h = s\Big],$$
$$\mathbf{Q}_h^{\mu,\nu}(s, a, b) :=$$
$$\mathbb{E}_{\mu,\nu}\Big[\sum_{l=h}^{H} \mathbf{r}_l(s_l, a_l, b_l)|s_h = s, a_h = a, b_h = b\Big].$$

For compactness of notation, for any $\mathbf{V} \in [0, H]^{dS}$ and $\mathbf{Q} \in [0, H]^{dSAB}$ we introduce the operators $\mathbb{P}$ and $\mathbb{D}$ by

$$\mathbb{P}_h[\mathbf{V}](s, a, b) := \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a,b)}[\mathbf{V}(s')],$$

$$\mathbb{D}_{\mu,\nu}[\mathbf{Q}](s) := \mathbb{E}_{a\sim\mu(\cdot|s),b\sim\nu(\cdot|s)}[\mathbf{Q}(s, a, b)].$$

With this notation we obtain the Bellman equations:

$$\mathbf{V}_h^{\mu,\nu}(s) = \mathbb{D}_{\mu_h,\nu_h}[\mathbf{Q}_h^{\mu,\nu}](s),$$
$$\mathbf{Q}_h^{\mu,\nu}(s, a, b) = (r_h + \mathbb{P}_h[\mathbf{V}_{h+1}^{\mu,\nu}])(s, a, b).$$

For convenience define $\mathbf{V}_{H+1}^{\mu,\nu}(s) = 0$ for any $s \in \mathcal{S}$.

**Satisfiability.** Let $\mathbb{W}^\star$ denote a desired target set. Henceforth, we assume that $\mathbb{W}^\star$ is a is closed and convex subset of $[0, H]^d$. Let $\hat{\mathbf{V}}^k$ be the cumulative return received by the agent in the $k$th episode and $\mathbf{W}^K := \frac{1}{K}\sum_{k=1}^{K} \hat{\mathbf{V}}^k$ be the average for the first $K$ episodes. The goal of the agent is to guarantee that $\mathbf{W}^K \in \mathbb{W}^\star$. This goal is achievable under the following satisfiability assumption.

**Assumption 1** (Satisfiability)**.** *Given a vector-valued MG* $\mathrm{MG}(\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathbb{P}, \mathbf{r}, H)$*, we say a closed and convex target set* $\mathbb{W}^\star$ *is satisfiable, if there exists a policy* $\mu$ *such that for any policy* $\nu$*, the vector value* $\mathbf{V}_1^{\mu,\nu}(s_1) \in \mathbb{W}^\star$.

Informally, satisfiability means that the agent can ensure the cumulative return is contained in the target set, regardless of the opponent's action. A weaker notion is if upon knowing the opponent's policy the agent can satisfy the target set. Thus, we call it *Response-satisfiability*.

**Assumption 2** (Response-satisfiability)**.** *Given a vector-valued MG* $\mathrm{MG}(\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathbb{P}, \mathbf{r}, H)$*, we say a closed and convex target set* $\mathbb{W}^\star$ *is response-satisfiable, if for any policy* $\nu$*,there exists a policy* $\mu$ *such that* $\mathbf{V}_1^{\mu,\nu}(s_1) \in \mathbb{W}^\star$.

Both notions coincide in a scalar-valued zero-sum game, as a result of von Neumann's minimax theorem. However, for vector-valued games, satisfiability is strictly stronger. Indeed, satisfiability fails even in some simple games while response-satisfiability holds. See the discussion at the end of Section 2.2 in (Abernethy et al., 2011) for a concrete example.

Without satisfiability, we cannot expect to reach the target set $\mathbb{W}^\star$. Luckily, approaching a response-satisfiable set $\mathbb{W}^\star$ on average is still possible. To that end, we can reduce the vector-valued MG to a scalar-valued one, as shown below.

### 2.2. Scalar Reduction and Minimax Theorem

We can convert a vector-valued MG to a scalar-valued one by replacing the return vector $\mathbf{r}$ by the scalar $\mathbf{r} \cdot \boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ is a fixed vector. Importantly, we will treat $\boldsymbol{\theta}$ as a dual variable in our algorithms. For the resulting MG we can define $V_h^{\mu,\nu}(\boldsymbol{\theta}, s)$ and $Q_h^{\mu,\nu}(\boldsymbol{\theta}, s, a, b)$ similarly.

We call the two players the "min-player" and the "max-player"[1]. Let $\nu$ be a policy of the max-player. There exists

[1]To accommodate conventions in Approachability, we make the agent the min-player (usually the max-player in MG literature).

a *best response* $\mu^\dagger$ to $\nu$, such that for any step $h \in [H]$ and state $s \in \mathcal{S}$ we have $V_h^{\mu^\dagger,\nu}(s) = V_h^{\dagger,\nu}(s) := \min_\mu V_h^{\mu,\nu}(s)$. A symmetric discussion applies to the best response to a min-player's policy. The following minimax equality holds: for any step $h \in [H]$ and state $s \in \mathcal{S}$,

$$\min_\mu \max_\nu V_h^{\mu,\nu}(\boldsymbol{\theta}, s) = \max_\nu \min_\mu V_h^{\mu,\nu}(\boldsymbol{\theta}, s).$$

A policy pair $(\mu^\star, \nu^\star)$ that achieves the equality is known as a *Nash equilibrium*. We use $V_h^\star(\boldsymbol{\theta}, s) := V_h^{\mu^\star,\nu^\star}(\boldsymbol{\theta}, s)$ to denote the value at the Nash equilibrium, which is unique for the MG and we call the *minimax value* of the MG.

**Approachability.** Scalarizing a vector-valued MG is equivalent to considering a half-space that contains $\mathbb{W}^\star$ instead of $\mathbb{W}^\star$ itself. If we can reach $\mathbb{W}^\star$, then we can reach any half-space that contains $\mathbb{W}^\star$. Therefore, satisfiability of half-spaces that contain $\mathbb{W}^\star$ is weaker than satisfiability of $\mathbb{W}^\star$ itself. We state this condition formally below.

**Assumption 3** (Approachability)**.** *Given a vector-valued MG* $\mathrm{MG}(\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathbb{P}, \mathbf{r}, H)$*, we say a closed and convex target set* $\mathbb{W}^\star$ *is approachable, if for any vector* $\boldsymbol{\theta}$*,*

$$\max_{\mathbf{x} \in \mathbb{W}^\star} \boldsymbol{\theta} \cdot \mathbf{x} \geq V_1^\star(\boldsymbol{\theta}, s_1).$$

Assumption 3 is also known as "half-space satisfiability" in the literature (Blackwell, 1956). Indeed, it is equivalent to response-satisfiability (See Lemma 7 in (Abernethy et al., 2011). The proof therein carries over for MGs directly, since it only depends on the geometric property of $\mathbb{W}^\star$.). We will only use this approachability condition in the sequel; it results in no loss of generality, and moreover, it is easier to extend to the non-approachable case.

So far we assumed that the target set $\mathbb{W}^\star$ is approachable. In practice, this assumption may or may not hold. In both cases, we can still seek to minimize the Euclidean distance $\mathrm{dist}(\mathbf{W}^K, \mathbb{W}^\star)$ of the average return to the target set. This is analogous to the agnostic learning setting for supervised learning. Toward this end, the following condition is useful.

**Assumption 4** ($\delta$-Approachability)**.** *Given a vector-valued MG* $\mathrm{MG}(\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathbb{P}, \mathbf{r}, H)$*, we say a closed and convex target set* $\mathbb{W}^\star$ *is* $\delta$*-approachable, if for any unit vector* $\boldsymbol{\theta}$*,*

$$\max_{\mathbf{x} \in \mathbb{W}^\star} \boldsymbol{\theta} \cdot \mathbf{x} + \delta \geq V_1^\star(\boldsymbol{\theta}, s_1).$$

Equivalently, this means the $\delta$-expansion of $\mathbb{W}^\star$ is approachable. So, a larger $\delta$ means $\mathbb{W}^\star$ is harder to approach.

### 2.3. Justification of knowing $\mathbb{W}^\star$

In the whole paper, we assume $\mathbb{W}^\star$ is known. To see this is reasonable, notice the target set is usually introduced in two cases:

1. $\mathbb{W}^\star$ is defined explicitly by constraints. In the case $\mathbb{W}^\star$ is clearly known, otherwise it is impossible to measure if the current return satisfies the constraints.
2. (Usually in one dimension) We want to maximize the return, so ideally we should define $\mathbb{W}^\star$ to be $[V^\star, \infty)$. However, in general there is no way to know $V^\star$ in prior. Luckily, if we know $V^\star \in [0, H]$, then we can set $\mathbb{W}^\star$ to be $[H, \infty)$ and the non-approachability results below will guarantee we can perform almost as well as knowing $V^\star$.

## 3. Multi-objective Meta-algorithm

Equipped with the generalized concepts of approachability for vector-valued MGs, we are ready to present our algorithmic framework. To make the exposition modular, we first present **M**ulti-**O**bjective **M**eta-**A**lgorithm (MOMA), our generic learning algorithm that is displayed as Algorithm 1. Subsequently, we explain its key components.

**Algorithm 1** Multi-objective Meta-algorithm (MOMA)

1: **Initialize:** for any $(s, a, b, h, s')$, $Q_h(s, a, b) \leftarrow \sqrt{d}H$, $N_h(s, a, b) \leftarrow 0$, $N_h(s, a, b, s') \leftarrow 0$, $\mathbf{W} \leftarrow \mathbf{0}$, $\boldsymbol{\theta} \leftarrow$ any unit verctor, $\hat{\mathbb{P}} \leftarrow$ any probability distribution.
2: **for** Episode $k = 1, \ldots, K$ **do**
3: $\quad \pi \leftarrow \textsc{Planning}(\boldsymbol{\theta}, \mathbf{r}, N, \hat{\mathbb{P}})$
4: $\quad \hat{\mathbf{V}} \leftarrow \mathbf{0}$.
5: $\quad$ **for** step $h = 1, \ldots, H$ **do**
6: $\quad\quad$ take action $(a_h, \cdot) \sim \pi_h(\cdot, \cdot | s_h)$.
7: $\quad\quad$ Observe opponent's action $b_h \sim \nu_h(s_h)$ and next state $s_{h+1}$.
8: $\quad\quad \hat{\mathbf{V}} \leftarrow \hat{\mathbf{V}} + \mathbf{r}_h(s_h, a_h, b_h)$.
9: $\quad\quad N_h(s_h, a_h, b_h) \leftarrow N_h(s_h, a_h, b_h) + 1$.
10: $\quad\quad N_h(s_h, a_h, b_h, s_{h+1}) \leftarrow N_h(s_h, a_h, b_h, s_{h+1}) + 1$
11: $\quad\quad \hat{\mathbb{P}}_h(\cdot | s_h, a_h, b_h) \leftarrow \frac{N_h(s_h, a_h, b_h, \cdot)}{N_h(s_h, a_h, b_h)}$.
12: $\quad$ **end for**
13: $\quad \mathbf{W} \leftarrow ((k-1)\mathbf{W} + \hat{\mathbf{V}})/k$.
14: $\quad \boldsymbol{\theta} \leftarrow \textsc{Dual-Update}(\mathbf{W}, \mathbb{W}^\star, \hat{\mathbf{V}})$
15: **end for**

MOMA is partitioned into into three components:

- **Planning** (Line 3): In each episode, we convert the vector-valued MG into a scalar-valued one by projecting onto the direction specified by the dual variable $\boldsymbol{\theta}$ and by computing the policy $\pi$.
- **Model Update** (Line 4 to 13): We accumulate the (vector-valued) return in each episode in $\hat{\mathbf{V}}$, and $\mathbf{W}$ is the average cumulative return. Then, we update the empirical estimators of the transition kernel.
- **Dual Update** (Line 14): Finally, we need to determine which direction we want to project the vector-valued MG onto in the next episode.

Notice that $\pi$ actually defines policies for both players, but we only execute it for the agent. Let $\mu_h(\cdot|s_h)$ and $\omega_h(\cdot|s_h)$ be the marginal distributions of $\pi_h(\cdot,\cdot|s_h)$. Then action $a_h$ is indeed sampled from the marginal $\mu_h(\cdot|s_h)$, while $b_h$ is sampled from $\nu_h(\cdot|s_h)$, which is not necessarily equal to $\omega_h(\cdot|s_h)$. Using this notation, we can observe that $\hat{\mathbf{V}}$ is unbiased in the sense that $\mathbb{E}[\boldsymbol{\theta}\cdot\hat{\mathbf{V}}] = V_1^{\mu,\omega}(\boldsymbol{\theta}, s_1)$.

The idea behind Algorithm 1 is simple: In each episode we fix a direction and try to approach the target set $\mathbb{W}^\star$. In this way, we can reduce the problem to a scalar-valued MG and benefit from existing work on scalar-valued MGs (Bai & Jin, 2020; Xie et al., 2020; Bai et al., 2020; Liu et al., 2020). The implementation of model updates is described in Algorithm 1. The other two sub-procedures vary slightly in different settings as follows:

- **PLANNING**: A planning algorithm to determine the policy $\pi$ based on the current estimated transition kernel $\hat{\mathbb{P}}$. For MGs we will use VI-HOEFFDING (Algorithm 2). For MDPs, we can design a finer VI-BERNSTEIN (Algorithm 3) to achieve a sharper convergence rate. In Line 11 of VI-HOEFFDING, we use NASH to denote computing the minimax policy w.r.t. a *matrix* game, which is standard in model-based method for MGs (Bai & Jin, 2020; Xie et al., 2020; Liu et al., 2020).
- **DUAL-UPDATE**: A dual update algorithm to update the variable $\boldsymbol{\theta}$, which describes the direction to approach $\mathbb{W}^\star$ in the next episode. We propose two different candidates: (PROJECTION-BASED-DUAL-UPDATE) and (PROJECTION-FREE-DUAL-UPDATE) in the following two sections. A variant of PROJECTION-FREE-DUAL-UPDATE, DOUBLE-DUAL-UPDATE is proposed in Section 6.1 to simutaneously optimize a cost function.

**Algorithm 2** VI-Hoeffding (VI-HOEFFDING)

1: **for** step $h = H, H-1, \ldots, 1$ **do**
2:  **for** $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$ **do**
3:  $t \leftarrow N_h(s,a,b)$.
4:  **if** $t > 0$ **then**
5:  $r_h(s,a,b) = \boldsymbol{\theta}\cdot\mathbf{r}_h(s,a,b)$;
6:  $\beta \leftarrow c\sqrt{\min\{d,S\}H^2 d\iota/t}$.
7:  $Q_h(s,a,b) \leftarrow \max\{(r_h + \widehat{\mathbb{P}}_h V_{h+1})(s,a,b) - \beta, -\sqrt{d}H\}$.
8:  **end if**
9:  **end for**
10:  **for** $s \in \mathcal{S}$ **do**
11:  $\pi_h(\cdot,\cdot|s) \leftarrow \text{NASH}(Q_h(s,\cdot,\cdot))$.
12:  $V_h(s) \leftarrow (\mathbb{D}_{\pi_h} Q_h)(s)$.
13:  **end for**
14: **end for**

## 4. Projection-based Dual Update

We begin with the most intuitive way to choose the dual variable: follow the direction that minimizes the distance of a candidate vector $\mathbf{W}$ to the target set $\mathbb{W}^\star$:

$$\boldsymbol{\theta} \leftarrow \begin{cases} \frac{\mathbf{W}-\Pi_{\mathbb{W}^\star}(\mathbf{W})}{\|\mathbf{W}-\Pi_{\mathbb{W}^\star}(\mathbf{W})\|_2}, \text{if } \mathbf{W}\notin\mathbb{W}^\star, \\ \text{any unit vector, otherwise.} \end{cases}$$

(PROJECTION-BASED-DUAL-UPDATE)

To find this direction, we need to compute the orthogonal projection onto $\mathbb{W}^\star$, thus we call it PROJECTION-BASED-DUAL-UPDATE.

To give theoretical guarantees, we will prove upper bounds on the Euclidean distance from our average cumulative return in the first $K$ episodes $\mathbf{W}^K$ to the target set $\mathbb{W}^\star$. If $\mathbb{W}^\star$ is approachable, $\text{dist}(\mathbf{W}^K, \mathbb{W}^\star)$ will converge to zero.

**Theorem 1.** *Following* MOMA *with* VI-HOEFFDING *(Algorithm 2) for* PLANNING *and* PROJECTION-BASED-DUAL-UPDATE *for* DUAL-UPDATE*, if $\mathbb{W}^\star$ is approachable, with probability $1-p$,*

$$\text{dist}(\mathbf{W}^K, \mathbb{W}^\star) \le \mathcal{O}\big(\sqrt{\min\{d,S\}dH^4SAB\iota/K}\big),$$

*where $\iota = \log(SABKH/p)$.*

The approachability condition (Assumption 3) is standard in the literature (Blackwell, 1956). However in practice, the desired target set $\mathbb{W}^\star$ may rarely also happen to be approachable (since it is chosen to meet the needs of an application, not to meet our demands on approachability). In this case, one may be unable to guarantee $\text{dist}(\mathbf{W}^K, \mathbb{W}^\star)$ converges to zero, but can only minimize the distance. A natural way to model this scenario is to assume $\mathbb{W}^\star$ is $\delta$-approachable, whence the following Theorem 2 applies.

**Theorem 2.** *If we use* VI-HOEFFDING *(Algorithm 2) for* PLANNING *and (*PROJECTION-BASED-DUAL-UPDATE*) for* DUAL-UPDATE *in* MOMA*, and if $W^*$ is $\delta$-approachable, then with probability $1-p$,*

$$\text{dist}\left(\mathbf{W}^K, W^*\right) \le \delta + \mathcal{O}\left(\sqrt{\min\{d,S\}dH^4SAB\iota/K}\right)$$

*where $\iota = \log\left(SABKH/p\right)$.*

**Remark.** Although we assume $\mathbb{W}^\star$ is $\delta$-approachable, the algorithm does not need to know $\delta$. Instead, we just run the same algorithm and the guarantee is adaptive.

**Rationale behind the criterion.** When characterizing the performance of our method, we choose to compete with $\delta$, the "non-approachability gap". This choice is simple and similar to the notion of regret used in scalar-valued MGs (Xie et al., 2020; Tian et al., 2020b). One may aim to be more ambitious: compete with the best response in

hindsight, as in (Mannor et al., 2014) for the bandit (single-horizon) setting. Unfortunately, such a choice is not computationally feasible for MGs. It is computationally hard even for scalar-valued MGs; see (Bai et al., 2020) for an exponential lower bound.

## 5. Projection-free Dual Update

The per-iteration computational bottleneck of PROJECTION-BASED-DUAL-UPDATE is to compute the projection onto $\mathbb{W}^\star$, which requires solving a convex program (a quadratic program when the constrains are linear) and can be computationally demanding. However, if we can find $\arg\max_{\mathbf{x}\in\mathbb{W}^\star} \boldsymbol{\theta}\cdot\mathbf{x}$ efficiently (e.g., when $\mathbb{W}^\star$ is a polytope), then we can develop a computation-friendly dual update based on online convex optimization (OCO) techniques (Abernethy et al., 2011; Shimkin, 2016).

To show the intuition behind PROJECTION-FREE-DUAL-UPDATE, we proceed via Fenchel duality. Consider a convex, closed, 1-Lipschitz function $f:[0,H]^d\to\mathbb{R}$. Its Fenchel conjugate is

$$f^*(\boldsymbol{\theta}) := \max_{\mathbf{x}\in X}\{\boldsymbol{\theta}\cdot\mathbf{x} - f(\mathbf{x})\}.$$

Then $f^*$ is $\sqrt{dH^2}$-Lipschitz by Corollary 13.3.3 in (Rockafellar, 1970). Fenchel duality implies

$$f(\mathbf{x}) = \max_{\|\boldsymbol{\theta}\|\le 1}\{\boldsymbol{\theta}\cdot\mathbf{x} - f^*(\boldsymbol{\theta})\}. \tag{1}$$

In particular, if $f(\mathbf{x}) = \mathrm{dist}(\mathbf{x},\mathbb{W}^\star)$, its Fenchel dual is $f^*(\boldsymbol{\theta}) = \max_{\mathbf{x}\in\mathbb{W}^\star}\boldsymbol{\theta}\cdot\mathbf{x}$ and its subdifferential is $\partial f^*(\boldsymbol{\theta}) = \arg\max_{\mathbf{x}\in\mathbb{W}^\star}\boldsymbol{\theta}\cdot\mathbf{x}$. Therefore, we can use its dual representation to "linearize" the distance. That is,

$$K\mathrm{dist}(\mathbf{W}^k,\mathbb{W}^\star) = \max_{\|\boldsymbol{\theta}\|\le 1}\left\{\boldsymbol{\theta}\cdot\sum_{k=1}^K\hat{\mathbf{V}}^k - \sum_{k=1}^K\max_{\mathbf{x}\in\mathbb{W}^\star}\boldsymbol{\theta}\cdot\mathbf{x}\right\}.$$

Ideally, if we can find the dual variable $\boldsymbol{\theta}^\star$ that maximizes the right-hand side above, minimizing the distance will be equivalent to minimizing a linear function in $\hat{\mathbf{V}}^k$, which can be handled as before if we use VI-HOEFFDING as the planning algorithm. Although we can not find $\boldsymbol{\theta}^\star$ directly, we can find a sequence of dual variables $\{\boldsymbol{\theta}\}_{k=1}^K$ such that $\sum_{k=1}^K\{\boldsymbol{\theta}^k\cdot\hat{\mathbf{V}}^k - \sum_{k=1}^K\max_{\mathbf{x}\in\mathbb{W}^\star}\boldsymbol{\theta}^k\cdot\mathbf{x}\}$ is close to $\max_{\|\boldsymbol{\theta}\|\le 1}\{\boldsymbol{\theta}\cdot\sum_{k=1}^K\hat{\mathbf{V}}^k - \sum_{k=1}^K\max_{\mathbf{x}\in\mathbb{W}^\star}\boldsymbol{\theta}\cdot\mathbf{x}\}$.

This task is precisely what online convex optimization (OCO) performs. The simplest solution is to use online subgradient method with step size $\eta^k = \sqrt{1/dH^2k}$. We define PROJECTION-FREE-DUAL-UPDATE formally below:

$$\boldsymbol{\theta}^{k+1} := \Pi_{\mathbb{B}^d}\{\boldsymbol{\theta}^k + \eta^k(\hat{\mathbf{V}}^k - \partial f^*(\boldsymbol{\theta}^k))\},$$
(PROJECTION-FREE-DUAL-UPDATE)

where $\Pi_{\mathbb{B}^d}$ denotes projection onto the $d$-dimensional unit Euclidean ball and $\partial f^*(\boldsymbol{\theta}^k)$ is a subgradient vector of $f^*$ at $\boldsymbol{\theta}^k$ (not a set).

Similarly, we provide theoretical guarantees for the new dual update rule. The proof is much simpler compared with that of Theorem 1 and Theorem 2.

**Theorem 3.** *Following* MOMA *with* VI-HOEFFDING *(Algorithm 2) for* PLANNING *and* PROJECTION-FREE-DUAL-UPDATE *for* DUAL-UPDATE*, if $\mathbb{W}^\star$ is $\delta$-approachable, with probability $1-p$,*

$$\mathrm{dist}\left(\mathbf{W}^K,\mathbb{W}^\star\right) \le \delta + \mathcal{O}\left(\sqrt{\min\{d,S\}dH^4SAB\iota/K}\right),$$

*where $\iota = \log(SABKH/p)$.*

## 6. Application to CMDPs: Near Optimal Rate

In this section, we apply our algorithmic framework to MDPs, which can be considered as a special case of MGs where the adversary cannot change the game. The stationary environment enables us to use the Bernstein-type concentration and achieve sharper dependence on the horizon $H$. The corresponding planning algorithm VI-BERNSTEIN is formalized in Algorithm 3. In Line 6 we use the empirical variance operator defined by $\widehat{\mathbb{V}}_h^k[V](s,a) := \mathrm{Var}_{s'\sim\widehat{\mathbb{P}}_h^k(\cdot|s,a)}V(s')$ for any function $V\in[-\sqrt{d}H,\sqrt{d}H]^S$. Notice that this approach does *not* work for MGs, because we need to estimate the variance of the value function $V^{\mu,\upsilon}$, a task that is impossible when the adversary's policy $\upsilon$ is unknown.

**Algorithm 3** VI-BERNSTEIN

1: **for** step $h = H, H-1,\ldots,1$ **do**
2: **for** $(s,a)\in\mathcal{S}\times\mathcal{A}$ **do**
3: $t\leftarrow N_h(s,a)$.
4: **if** $t>0$ **then**
5: $r_h(s,a) = \boldsymbol{\theta}\cdot\mathbf{r}_h(s,a)$;
6: $\beta \leftarrow c(\sqrt{\hat{\mathbb{V}}_h\underline{V}_{h+1}(s,a)\min\{d,S\}\iota/t} + \hat{\mathbb{P}}_h(\overline{V}_{h+1} - \underline{V}_{h+1})(s,a)/H + \min\{d,S\}\sqrt{d}H^2\iota/t)$.
7: $\overline{Q}_h(s,a)\leftarrow\max\{(r_h+\widehat{\mathbb{P}}_h\overline{V}_{h+1})(s,a)-\beta, -\sqrt{d}H\}$.
8: $\underline{Q}_h(s,a)\leftarrow\min\{(r_h+\widehat{\mathbb{P}}_h\underline{V}_{h+1})(s,a)+\beta, \sqrt{d}H\}$.
9: **end if**
10: **end for**
11: **for** $s\in\mathcal{S}$ **do**
12: $\pi_h(s)\leftarrow\arg\min(\underline{Q}_h(s,\cdot))$.
13: $\underline{V}_h(s)\leftarrow\underline{Q}_h(s,\pi_h(s)), \overline{V}_h(s)\leftarrow\overline{Q}_h(s,\pi_h(s))$.
14: **end for**
15: **end for**

The sharper theoretical guarantee is as follows:

**Theorem 4.** *If we use* VI-BERNSTEIN *(Algorithm 3) for* PLANNING *and* (PROJECTION-BASED-DUAL-UPDATE) *or* (PROJECTION-FREE-DUAL-UPDATE) *for* DUAL-UPDATE

*in* MOMA*, and if* $\mathbb{W}^\star$ *is* $\delta$*-approachable, then with probability* $1-p$*,*

$$\operatorname{dist}(\mathbf{W}^K, \mathbb{W}^*) \leq \delta + \mathcal{O}\big(\sqrt{\min\{d,S\}dH^3SA\iota^2/K}\big),$$

*where* $\iota = \log(SAKH/p)$.

When $d \leq S$ (as is in most cases), our result is minimax optimal up to log-factors in $S, A, H, K$ according to the lower bound $\Omega\left(\sqrt{H^3SA/K}\right)$ proven in (Domingues et al., 2020). The tightness of our result in $d$ remains open. In particular, we can get a naive $\Omega\left(\sqrt{dH^3SA/K}\right)$ lower bound by duplicating the negative MDP example from (Domingues et al., 2020) $d$ times in $d$ dimensions, and the distance naturally scales up by $d$. With such a lower bound, there is still a $\sqrt{d}$ gap open. More details on the difficulty of providing a tigher lower bound are discussed in Section 7.

The upper bound in Theorem 4 allows us to find a policy that approaches the target set $\mathbb{W}^\star$ efficiently. Next, we generalize the result to the constrained MDP setting where we want to simultaneously minimize a cost function.

### 6.1. Optimizing a Cost Function Simultaneously

In this section, we show how to extend our algorithm to the constrained MDP setup (Efroni et al., 2020; Ding et al., 2020; Qiu et al., 2020; Brantley et al., 2020), in which one wants to simultaneously minimize a cost function $g : \mathbb{R}^d \to [0,1]$ defined on the return vector space. The goal is two-fold: (i) satisfy constraints defined by the target set; and (ii) minimize the cumulative cost. Note that our setup *subsumes* the canonical cost function in which the cost function is defined on the state-action pair (e.g., (Efroni et al., 2020)). Particularly, we can add an extra coordinate in the return vector space to denote the cost for each state-action pair, and pick $g$ to solely extract that cost coordinate. A more detailed comparison against constrained MDP setups from previous works can be found in Appendix C.

For our analysis, we assume that the cost function $g(\cdot)$ is 1-Lipschitz and convex. Following (Efroni et al., 2020; Ding et al., 2020; Qiu et al., 2020; Brantley et al., 2020), we also assume $\mathbb{W}^\star$ is satisfiable and that we want to compete with a policy $\mu^\star$ such that $\mathbf{V}_1^{\mu^\star}(s_1) \in \mathbb{W}^\star$. One might hope to bound the regret $\sum_{k=1}^K g(\hat{\mathbf{V}}^k) - Kg(\mathbf{V}_1^{\mu^\star}(s_1))$. But this goal is hard. Its counterpart is unknown even in the bandit setup Agrawal & Devanur (2014). Instead, we aim to upper bound both the regret $[g(\mathbf{W}^K) - g(\mathbf{V}_1^{\mu^\star}(s_1))]$ and the constraint violation $\operatorname{dist}(\mathbf{W}^K, \mathbb{W}^\star)$.

**Constraint geometry.** Toward achieving our aim, we need to impose some geometric requirements on the constraints that will help us quantify algorithmic complexity in a non-asymptotic manner. Previous works that use a primal-dual approach (e.g., (Efroni et al., 2020; Qiu et al., 2020; Ding et al., 2020)) assume knowledge of explicit structure of the constraint set, concretely by requiring $\mathbb{W}^\star = \{x \mid \forall i, g_i(x) \leq 0\}$. Subsequently, they control complexity of the constraint set by assuming Lipschitzness of the $g_i$ and a strong Slater condition, i.e., there is a strictly feasible interior point $x_0$ such that $g_i(x_0) \leq -\epsilon$ for a universal constant $\epsilon > 0$. In contrast, we do not impose explicit structure on $\mathbb{W}^\star$. Instead, we assume that we can solve linear or quadratic optimization over $\mathbb{W}^\star \subset \mathbb{R}^d$. A naive way to cast our setup into the previous form would be use the inequality $g_0(\cdot) := \operatorname{dist}(\cdot, \mathbb{W}^\star) \leq 0$. But since $g_0$ is a distance function, we cannot satisfy the strict interiority condition needed by the previous setup. Consequently, we need to limit the complexity of our constraint set through a more refined alternative.

To this end, we propose a geometric condition. In particular, we assume that the target set $\mathbb{W}^\star$ intersects with the set of achievable value vectors $\mathcal{V} = \{\mathbf{V}_1^\pi(s_1) | \text{ any policy } \pi\}$ *nonsingularly*—Figure 1 illustrates this concept. Formally, denote the set of achievable returns within the target set as $\mathcal{W} = \mathcal{V} \cap \mathbb{W}^\star$ and $\partial\mathcal{W} = \partial\mathcal{V} \cap \partial\mathbb{W}^\star$ as the intersection of the boundaries of $\mathbb{W}^\star$ and the achievable value vector set $\mathcal{V}$. Then, Assumption 5 describes nonsingular intersection.

**Assumption 5.** *If* $\partial\mathcal{W}$ *is not empty, then for each vector* $\mathbf{W} \in \partial\mathcal{W}$*, denote the maximum angle* $\alpha \in [0, \pi]$ *between the support vectors* $\vec{a}$ *of* $\mathbb{W}^\star$ *at* $\mathbf{W}$ *and the support vectors* $\vec{b}$ *of* $\mathcal{V}$ *at* $\mathbf{W}$ *as*

$$\alpha(\mathbf{W}) := \min\{\angle(\vec{a}, \vec{b}) \mid \vec{a}, \vec{b} \text{ are support vectors of sets } \mathbb{W}^\star \text{ and } \mathcal{V} \text{ at } \mathbf{W}\}.$$

*We assume there exists a constant* $\alpha_{\max} \in [\pi/2, \pi)$ *such that* $\max_{w \in \partial\mathcal{W}} \alpha(w) < \alpha_{\max}$*. With this upper bound on* $\alpha$*, we denote* $\gamma_{\min} = \sin(\pi - \alpha_{\max}) > 0$.

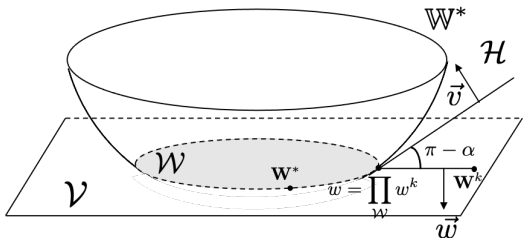


*Figure 1.* The target set intersects with the achievable return vectors nonsingularly. The angle $\alpha(\prod_{\mathcal{W}} \mathbf{W}^k)$ is upper bounded.

Assumption 5 excludes the case where the sets $\mathcal{V}$ and $\mathbb{W}^\star$ intersect tangentially (i.e., share the same supporting hyperplane) resulting in $\alpha = \pi$. The necessity of such a geometric assumption is discussed in Appendix E.1. At a high level, Assumption 5 is a geometric analog of the previously noted strict interiority condition that excludes a singular intersection of the constraint functions $g_i$. Our assumption provides

| Algorithms | Regret | Constraint Violation | Nonlinear Cost and Constraints | Computionally Efficient |
|---|---|---|---|---|
| OptCMDP-bonus (Efroni et al., 2020) | $\tilde{\mathcal{O}}\left(\sqrt{H^4S^2AK}\right)$ | $\tilde{\mathcal{O}}\left(\sqrt{dH^4S^2AK}\right)$ | | ✓ |
| (Brantley et al., 2020) | $\tilde{\mathcal{O}}\left(\sqrt{H^3S^2AK}\right)$ | $\tilde{\mathcal{O}}\left(\sqrt{d^3H^3S^2AK}\right)$ | ✓ | |
| OptPD-CMDP (Efroni et al., 2020) | $\tilde{\mathcal{O}}\left(\sqrt{(S^2A+d^2)H^4K}\right)$ | $\tilde{\mathcal{O}}\left(\sqrt{(S^2Ad^2+d^3)H^4K}\right)$ | | ✓ |
| OPDOP (Ding et al., 2020) | $\tilde{\mathcal{O}}\left(\sqrt{H^5S^4A^2K}\right)$ | $\tilde{\mathcal{O}}\left(\sqrt{H^5S^4A^2K}\right)$ | | ✓ |
| UCPD (Qiu et al., 2020) | $\tilde{\mathcal{O}}\left(\sqrt{H^5S^2AK}\right)$ | $\tilde{\mathcal{O}}\left(\sqrt{H^5S^2AK}\right)$ | | ✓ |
| **This Paper** | $\tilde{\mathcal{O}}\left(\sqrt{\min\{d,S\}dH^3SAK}\right)$ | $\tilde{\mathcal{O}}\left(\sqrt{\min\{d,S\}dH^3SAK}\right)$ | ✓ | ✓ |

*Table 2.* Comparison with constrained MDP literature.

a way to lower-bound the distance to the target set $\mathbb{W}^\star$ by the distance to the actual constraint set $\mathcal{W} = \mathcal{V} \cap \mathbb{W}^\star$, and thus prevent an algorithm from trading off too much constraint violation in exchange for a lower cost value $g(\mathbf{V}^k)$.

To minimize cost and avoid constraint violation simultaneously we need a "double" version of dual variable update. This idea is formalized in DOUBLE-DUAL-UPDATE below:

$$\begin{aligned}
\boldsymbol{\varphi}^{k+1} &= \Pi_{\mathbb{B}^d}\left\{\boldsymbol{\varphi}^k + \eta^k\left(\hat{\mathbf{V}}^k - \arg\max_{\mathbf{x}\in W^\star} \boldsymbol{\varphi}^k \cdot \mathbf{x}\right)\right\},\\
\boldsymbol{\phi}^{k+1} &= \Pi_{\mathbb{B}^d}\left\{\boldsymbol{\phi}^k + \eta^k\left(\hat{\mathbf{V}}^k - \partial g^*(\boldsymbol{\phi}^k)\right)\right\},\\
\boldsymbol{\theta}^{k+1} &= \rho\boldsymbol{\varphi}^{k+1} + \boldsymbol{\phi}^{k+1} \qquad \text{(DOUBLE-DUAL-UPDATE)}
\end{aligned}$$

where $\Pi_{\mathbb{B}^d}$ denotes projection onto the $d$-dimensional unit Euclidean ball and $\partial g^*\left(\boldsymbol{\varphi}^k\right)$ is a subgradient vector of $g^*$ at $\boldsymbol{\varphi}^k$ (not a set).

Here comes our theoretical guarantee for both constraint violation and regret.

**Theorem 5.** *Following* MOMA *with* VI-BERNSTEIN *(Algorithm 3) for* PLANNING *and* DOUBLE-DUAL-UPDATE *for* DUAL-UPDATE*, if $\mathbb{W}^\star$ is approachable and $\mu^\star$ is a policy s.t. $\mathbf{V}_1^{\mu^\star}(s_1) \in \mathbb{W}^\star$, with probability $1-p$ we can bound the constraint violation and the regret respectively as follows:*

$$\begin{aligned}
\operatorname{dist}(\mathbf{W}^K, W^\star) &\le \mathcal{O}\left(\sqrt{\min\{d,S\}dH^3SA\iota/K}\right),\\
g(\mathbf{W}^K) - g(\mathbf{V}_1^{\mu^\star}(s_1)) &\le \mathcal{O}\left(\rho\sqrt{\min\{d,S\}dH^3SA\iota/K}\right),
\end{aligned}$$

*where $\iota = \log(SAKH/p)$, $\rho = 2/\gamma_{\min}$.*

Known results on constrained MDP problems do not share a common setup and hence make a precise comparison tricky. In short, our result aims to provide a *computationally efficient* algorithm for non-linear constraints (target set) and a convex cost function (see Table 2). Please see Appendix C for a more detailed discussion of the subtleties among different constrained MDP setups, and some minor modifications needed to unify the exposition. With the existing results, our result is significant in the following aspects:

- First, our algorithm is the most general in terms of being able to handle non-linearity in the cost and constraints. The constrained MDP setting we study in Section 6.1 is a direct generalization of (Agrawal & Devanur, 2014), and is closest to (Brantley et al., 2020). While our constraint assumption is equivalent to the one in (Brantley et al., 2020), our cost functions are more general. The domain of Brantley et al.'s cost function is scalars, while that of ours is vectors.
- Furthermore, our proposed algorithm is *computationally efficient* because we do not require solving a large-scale convex optimization sub-problem with the number of variables and constraints scaling as $\mathcal{O}(SAH)$ per iteration (see Table 2). Indeed, our algorithms only comprise planning and model update procedures with a total of $\mathcal{O}(S^2AH)$ basic algebraic updates in each episode, along with a dual space optimization procedure whose computational complexity is free of $S$, $A$ and $H$.
- Our bounds on regret and constraint violation are also the sharpest with respect to their dependence on the parameters $S$, $A$, and $K$.

## 7. Conclusion and Future Work

In this paper, we formulate online learning in vector-valued Markov games through the lens of approaching a fixed convex target set within which the vector-valued objective should lie. We provide efficient model-based algorithms as instances of a generic meta-algorithm. Two key ideas contribute to our algorithmic design: (i) reduction of the vector-valued Markov game to a scalar-valued one, where the scalarization is iteratively updated; and (ii) strategic exploration of the environment. For vector-valued MDPs, our algorithms, after some modifications, achieve a tight rate in approaching the target set (in terms of $S, A, H, K$),

while simultaneously minimizing a convex cost function. Moreover, when the given target set is non-approachable, our algorithms automatically adapt to the degree of non-approachability.

Several problems are left open. Currently, there is still a $\sqrt{d}$ gap ($d$ is the dimensionality of the vector-valued cost) between our upper bound and the lower bound. How to close this gap to achieve the minimax rate remains unknown. The challenge in providing a tighter lower bound is that estimating a discrete distribution under the $L^2$ distance does not get harder as the dimensionality increases. Since we use the Euclidean distance to measure the performance of our algorithms, we cannot get stronger dependence on $d$. Lower bounds such as the one in (Jin et al., 2020) use a multiple hypothesis testing approach successfully because they work with an $L^1$ loss, whereas we study the standard Euclidean loss. A second question is that our result in Section 6.1 has somewhat worse dependence on $d$ and $\rho$ compared to previous results. We leave improving the dimension dependency as a future direction.

Another future direction that is worth pursuing is that of redefining the notion of regret and error. Our work measures approachability error using the Euclidean distance. In practice, this choice may not be the only useful measure. Can we develop provably efficient algorithms under other geometries and measures of approachability? Answering this question might help exploit the geometry of the target set better, and potentially lead to tighter complexity analyses.

## Acknowledgement

TY and YT acknowledge partial support as a graduate research assistant from the NSF BIGDATA grant (number 1741341). JZ acknowledge support from NSF CAREER grant Number 1846088.

## References

Abernethy, J., Bartlett, P. L., and Hazan, E. Blackwell approachability and no-regret learning are equivalent. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 27–46, 2011.

Agrawal, S. and Devanur, N. R. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 989–1006, 2014.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 263–272. JMLR. org, 2017.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 207–216. IEEE, 2013.

Bai, Y. and Jin, C. Provable self-play algorithms for competitive reinforcement learning. *arXiv preprint arXiv:2002.04017*, 2020.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. *arXiv preprint arXiv:2006.12007*, 2020.

Blackwell, D. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

Brafman, R. I. and Tennenholtz, M. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct): 213–231, 2002.

Brantley, K., Dudik, M., Lykouris, T., Miryoosefi, S., Simchowitz, M., Slivkins, A., and Sun, W. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *arXiv preprint arXiv:2006.05051*, 2020.

Chen, X., Hu, J., Li, L., and Wang, L. Efficient reinforcement learning in factored mdps with application to constrained rl. *arXiv preprint arXiv:2008.13319*, 2020.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Non-stationary reinforcement learning: The blessing of (more) optimism. *Available at SSRN 3397818*, 2019.

Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanović, M. R. Provably efficient safe exploration via primal-dual policy optimization. *arXiv preprint arXiv:2003.00534*, 2020.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. *arXiv preprint arXiv:2010.03531*, 2020.

Efroni, Y., Mannor, S., and Pirotta, M. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Jenatton, R., Huang, J., and Archambeau, C. Adaptive algorithms for online convex optimization with long-term constraints. In *International Conference on Machine Learning*, pp. 402–411. PMLR, 2016.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4863–4873, 2018.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. *arXiv preprint arXiv:2002.02794*, 2020.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604*, 2020.

Mannor, S., Perchet, V., and Stoltz, G. Approachability in unknown games: Online learning meets multi-objective optimization. In *Conference on Learning Theory*, pp. 339–355, 2014.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Miryoosefi, S., Brantley, K., Daumé III, H., Dudík, M., and Schapire, R. Reinforcement learning with convex constraints. *arXiv preprint arXiv:1906.09323*, 2019.

Neumann, J. v. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

Osband, I. and Van Roy, B. Near-optimal reinforcement learning in factored mdps. *Advances in Neural Information Processing Systems*, 27:604–612, 2014.

Qiu, S., Wei, X., Yang, Z., Ye, J., and Wang, Z. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. *Advances in Neural Information Processing Systems*, 33, 2020.

Rockafellar, R. T. *Convex analysis*. Number 28. Princeton university press, 1970.

Shapley, L. S. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

Shimkin, N. An online convex optimization approach to blackwell's approachability. *The Journal of Machine Learning Research*, 17(1):4434–4456, 2016.

Singh, R., Gupta, A., and Shroff, N. B. Learning in markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

Tian, Y., Qian, J., and Sra, S. Towards minimax optimal reinforcement learning in factored markov decision processes. *Advances in Neural Information Processing Systems*, 33, 2020a.

Tian, Y., Wang, Y., Yu, T., and Sra, S. Provably efficient online agnostic learning in markov games. *arXiv preprint arXiv:2010.15020*, 2020b.

Wu, J., Braverman, V., and Yang, L. F. Accommodating picky customers: Regret bound and exploration complexity for multi-objective reinforcement learning. *arXiv preprint arXiv:2011.13034*, 2020.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning Zero-Sum Simultaneous-Move Markov Games Using Function Approximation and Correlated Equilibrium. *arXiv preprint arXiv:2002.07066*, 2020.

Yuan, J. and Lamperski, A. Online convex optimization for cumulative constraints. In *Advances in Neural Information Processing Systems*, pp. 6137–6146, 2018.