Journal of Machine Learning Research 24 (2023) 1-43 Submitted 10/22; Revised 11/23; Published 12/23

Double Duality: Variational Primal-Dual Policy Optimization for Constrained Reinforcement Learning

Zihao Li ZIHAOLI@PRINCETON.EDU
Department of Electrical and Computer Engineering
Princeton University
Princeton, NJ 08544, USA

Boyi Liu BOYILIU2018@U.NORTHWESTERN.EDU
Department of Industrial Engineering and Management Sciences
Northwestern University
IL 60208, USA

Zhuoran Yang ZHUORANYANG.WORK@GMAIL.COM
Department of Statistics and Data Science
Yale University
CT 06511-6814, USA

Zhaoran Wang ZHAORANWANG@GMAIL.COM
Department of Industrial Engineering and Management Sciences
Northwestern University
IL 60208, USA

Mengdi Wang MENGDIW@PRINCETON.EDU
Department of Electrical and Computer Engineering
Princeton University
Princeton, NJ 08544, USA

Editor: Mingyuan Zhou

Abstract

We study the Constrained Convex Markov Decision Process (MDP), where the goal is to minimize a convex functional of the visitation measure, subject to a convex constraint. Designing algorithms for a constrained convex MDP faces several challenges, including (1) handling the large state space, (2) managing the exploration/exploitation tradeoff, and (3) solving the constrained optimization where the objective and the constraint are both nonlinear functions of the visitation measure. In this work, we present a model-based algorithm, Variational Primal-Dual Policy Optimization (VPDPO), in which Lagrangian and Fenchel duality are implemented to reformulate the original constrained problem into an unconstrained primal-dual optimization. The primal variables are updated by model-based value iteration following the principle of *Optimism in the Face of Uncertainty* (OFU), while the dual variables are updated by gradient ascent. Moreover, by embedding the visitation measure into a finite-dimensional space, we can handle large state spaces by incorporating function approximation. Two notable examples are (1) Kernelized Nonlinear Regulators and (2) Low-rank MDPs. We prove that with an optimistic planning oracle, our algorithm achieves sublinear regret and constraint violation in both cases and can attain the globally optimal policy of the original constrained problem.

Keywords: Online Learning, Function Approximation, Reinforcement Learning, Constrained Optimization, Duality Theory

1. Introduction

In recent years, constrained reinforcement learning (RL) has attracted greater research interest. In contrast to unconstrained RL, in which an agent can freely learn to maximize its cumulative reward or minimize its cost by interacting with an unknown environment, we face learning problems with various kinds of constraints in many real-world applications. For example, in autonomous driving, we want to minimize the time cost while avoiding speeding or colliding with other cars (Garcıa and Fernández, 2015). Other applications include cost-constrained RL in medical applications and business restrictions for tax collection optimization (Abe et al., 2010), in which the total budget is restricted.

However, existing works on Markov decision process (MDP) with constraints are still limited. Currently, most works consider the constrained MDP with both the objectives and constraints being linear functionals of visitation measures (Efroni et al., 2020; Ding et al., 2021). However, in many complex scenarios, we encounter problems with certain nonlinear structures. For example, in apprenticeship learning the agent aims to simulate the performance of an expert in a demonstrated task (Abbeel and Ng, 2004b). It is difficult to formulate an explicit reward function, and the learning goal is given by the ℓ_2-norm distance between the visitation measure of the agent and the expert. In multi-objective MDP, we have to consider nonlinear interaction between different objectives (Wu et al., 2021; Yu et al., 2021). Other examples include cautious MDP (Zhang et al., 2020a) and general utility MDP (Zhang et al., 2020b).

In this work, we introduce the Constrained Convex Markov Decision Process (C^{2}MDP), where we consider a constrained convex optimization over the space of visitation measures. The agent manipulates her policy over the space of visitation to minimize the objective while fulfilling the constraints. Compared to previous works, our model allows objectives and constraints to be nonlinear in visitation measure, thus significantly extending beyond Constrained MDP (Efroni et al., 2020; Ding et al., 2021). Moreover, our model covers interesting examples such as convex MDP (Zahavy et al., 2021), general utility RL (Zhang et al., 2020b), and apprenticeship learning (Abbeel and Ng, 2004b) as special cases. Challenges in designing an efficient online algorithm for constrained convex MDP are threefold:

(i) Most existing theoretical convergence guarantees for convex MDP apply only to the tabular case (Zhang et al., 2020b; Efroni et al., 2020; Zahavy et al., 2021), where the visitation measure is a vector of dimension $O(H|\mathcal{S}||\mathcal{A}|)$, making the convex MDP a convex optimization problem. However, when facing a continuous state space, the visitation measure becomes a general distribution on the state-action space. Due to the curse of dimensionality, algorithms designed for tabular MDP fail to tackle the problem.

(ii) Highly different from simple constrained MDP, which only imposes a linear constraint in the value function, the objective and constraint of C^{2}MDP can be nonlinear functionals of the visitation measure. Without knowing further structure, finding optimal solutions for such problems is much harder than Constrained MDP, which is equivalent to solving a linear programming problem (Efroni et al., 2020).

(iii) In a C^{2}MDP, the transition of the environment is unknown, and can only be learned through the transition through interacting with the environment. With limited information, designing an efficient online exploration strategy is hard.

With these coupled challenges, we ask the following question:

Can we find the globally optimal policy of constrained convex MDP in online learning?

In this work, we give an affirmative answer to this question.

- To handle (i), we incorporate function approximation and formulate the optimization in the embedded space of the visitation measures. In particular, we consider the feature map in function approximation and its expectation under the visitation measure, which is known as the kernel embedding of visitation measure (Hofmann et al., 2008; Muandet et al., 2016). We further consider the optimization with the kernel embedding of the visitation being the decision variables, which motivates us to implement online optimization techniques for solving C^{2}MDP. Such a formulation recovers the tabular setting as a special case when using the canonical embedding.

- To handle (ii), we use Lagrangian duality to transform the constrained problem to an unconstrained minimax optimization problem. In presence of Slater's condition, it is guaranteed that the original minimization shares the same optimal value with the unconstrained one. Moreover, to handle nonlinearity in the objective and the constraint, we apply Fenchel duality to introduce a linear structure. Combining the above two types of duality, we obtain a primal-dual optimization problem with a linear dependency on the kernel embedding. This allows us to construct a linear reward and adopt techniques of previous works in model-based value iteration, such as Kakade et al. (2020); Ayoub et al. (2020).

- To handle (iii), we apply the principle of *Optimism in the Face of Uncertainty* (OFU) (Jin et al., 2020; Yang et al., 2020) by an optimistic planning oracle (Jin et al., 2021; Kakade et al., 2020; Ayoub et al., 2020) which behaves as if the model parameters assume their best possible values in accordance to the observations so far.

With the above techniques, our algorithm is provably sample-efficient. In specific, we prove that our algorithm achieves $O(\sqrt{T})$ in both the regret and the constraint violation, where T is the number of the sampling episodes. To the best of our knowledge, our algorithm is the first provably sample-efficient algorithm for the constrained nonlinear optimization over visitation measures. As special cases, our method can be widely applied to multi-objective MDP, and apprenticeship learning, and lead to efficient algorithms.

1.1 Related Works

Optimization over occupancy measures/Convex MDP. Several early works (Tewari and Bartlett, 2007; Chen and Wang, 2016; Wang, 2017, 2020) studied tabular MDP via linear programming reformulation. Zahavy et al. (2021) studied convex MDP via Fenchel duality. Zhang et al. (2020a,b,b) studied for convex optimization over occupancy measures. However, while all these methods are successful in tabular MDP, they cannot (i) avoid the curse of dimensions in large state space MDP, and (ii) handle constraints.

Constrained MDP. Our work is a generalization of the constrained MDP. Efroni et al. (2020); Yu et al. (2021); Qiu et al. (2020); Brantley et al. (2021) studied tabular constrained MDP. Ding et al. (2021) studied safe reinforcement learning under function approximation setting under a linear mixture MDP model and using upper confidence bound (UCB) algorithm for exploration. Wu et al. (2021) further provided a general algorithm for Multi-objective MDP with general constraints and objective relies on multiple value functions. All of these methods assume that a given reward exists and explores the environment following the principle of optimism, and achieves great success by providing sublinear regret and constraint violation. Vaswani et al. (2022) provides a zero-constrained algorithm and provide a lower bound under such scenario. However, when there is no given reward function, these methods are no longer applicable.

Provably efficient online RL. Our work is closely related to a line of provably efficient online RL algorithms on Low-rank MDPs (Agarwal et al., 2020; Uehara et al., 2022) and kernelized nonlinear regulator (Kakade et al., 2020; Mania et al., 2020), where efficient exploration of the agent is obtained by choosing an optimistic model in the confidence set. However, these results are only designed for unconstrained problems that are linearly dependent on the occupancy measure.

1.2 Notations

We denote by $[a:b]$ the set of integers between a and b, i.e., $[a:b] = \{i \in \mathbb{Z} \mid a \leq i \leq b\}$, and write $[n] = [1:n]$. We denote by $x = (x_h)_{h\in[H]}$ the column vector obtained by concatenating the elements of $\{x_h\}_{h\in[H]}$, i.e., $x = (x_1; \cdots; x_H)$. We write $a \cdot b$ as the inner product of two finite dimensional vectors, and $\langle f, g\rangle_{\mathcal{H}}$ as the inner product of two functions f and g in the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. We also denote by $\|\cdot\|_2$ the ℓ_2-norm in Euclidean space, and $\mathcal{B}^d$ the the unit ball in $\mathbb{R}^d$, i.e., $\{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. The set of probability distribution over a space $\mathcal{X}$ is denoted by $\Delta(\mathcal{X})$. We define $\mathcal{P}(s' \mid s, a)$ as the probability for the agent transiting to state s' from s when taking action a.

2. Background

In this section, we briefly introduce the concepts of reinforcement learning, Constrained Convex MDPs, Low-rank MDPs, and Kernalized nonlinear Regulator (KNR).

2.1 MDP Setting

We consider an episodic Markov decision process problem $(\mathcal{S}, \mathcal{A}, H, c)$, where $\mathcal{S} \subset \mathbb{R}^d$ is the state space embedded in the Euclidean space, $\mathcal{A}$ is a (possibly continuous) action space, H is the horizon, and $c = \{c_h\}_{h=1}^H$ is a collection of cost functions where $c_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the cost of stage h. In each episode, we consider an agent with policy $\pi = \{\pi_h\}_{h=1}^H$, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$. At the stage h, the agent takes an action $a_h \in \mathcal{A}$ according to the policy $\pi_h(\cdot \mid s_h)$. The state then transits to s_{h+1} with probability $\mathcal{P}_h(s_{h+1} \mid s_h, a_h)$ according to the underlying transition rule. Since the choice of the initial state does not add complexity to the problem, for simplicity, we assume that the initial state is fixed, i.e., $s_1 = \overline{s}$.

We introduce the concepts of action-state value function and state value function from reinforcement learning. The action-state value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as

$$Q_h^\pi(s,a) = \mathbb{E}_\pi\biggl[\sum_{i=h}^{H} c_i(s_i,a_i) \biggm| s_h = s, a_h = a\biggr], \quad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H].$$

Correspondingly, the action-value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ is defined as

$$V_h^\pi(s) = \mathbb{E}_\pi\biggl[\sum_{i=h}^{H} c_i(s_i,a_i) \biggm| s_h = s\biggr], \quad \forall (s,h) \in \mathcal{S} \times [H]. \tag{1}$$

Here the expectation $\mathbb{E}_\pi[\cdot]$ is taken over the trajectory $\{(s_h,a_h)\}_{h\in[H]}$ induced by $\{\pi_h\}_{h\in[H]}$ and the underlying transition. For notation simplicity, we also write

$$\mathbb{P}_h f(s_h,a_h) = \mathbb{E}_{s'\sim\mathcal{P}(\cdot\,|\,s_h,a_h)}[f(s')]$$

for any integrable $f : \mathcal{S} \to \mathbb{R}$ and conditional probability $\mathcal{P}_h(\cdot \mid s_h, a_h)$.

2.2 Constrained Convex MDP (C^{2}MDP)

We generalize the problem of convex MDP (Zahavy et al., 2021), which considers a non-constrained convex optimization problem with the occupancy measure $d_\pi = (d_{\pi,h}(s,a))_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]}$ as the variable. The agent manipulates the occupancy measure by properly adjusting its policy. The aim is to find the optimal policy that minimizes the objective function. A tabular C^{2}MDP usually takes the form of

$$\min_{\pi\in\Delta(\mathcal{A}\,|\,\mathcal{S},H)} f\bigl((d_{\pi,h})_{h\in[H]}\bigr) \quad \text{s.t.} \quad \begin{aligned} &d_{\pi,h}(s,a) = \mathbb{E}_\pi\bigl[\mathbb{1}_{(s_h,a_h)=(s,a)}\bigr], \\ &g(d_{\pi,h}(s,a)) \le 0, \end{aligned} \tag{2}$$

where $f, g : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|H} \to \mathbb{R}$ are both convex functions.

In the tabular MDP, the set of all x induced by the agent's policy is represented by a polytope represented by $O(|\mathcal{S}||\mathcal{A}|)$ linear constraints (Efroni et al., 2020; Zahavy et al., 2021). However, in the continuous state space case, due to the curse of dimensionality and the shortage of memory, such an LP formulation is generally impossible. Therefore, we incorporate function approximation to handle the large state space by embedding the information of the state-action pair with a finite-dimensional feature map $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. Such a method is widely used in RL literature (Yang et al., 2020; Jin et al., 2020; Uehara et al., 2022; Kakade et al., 2020). To describe the visitation of the agent, we apply kernel embedding of probability distribution to the visitation measure (Muandet et al., 2017; Zahavy et al., 2021; Efroni et al., 2020) . By embedding the probability distribution induced by the agent's policy π on $\mathcal{S} \times \mathcal{A}$ into finite dimension linear space, the objective and constraints related to the distribution can be reformulated into a function for the kernel embedding.

Definition 1 (Kernel Embedding) *For a MDP with kernel feature mapping* $\{\psi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d\}_{h\in[H]}$*, we define its kernel embedding as*

$$\Psi^\pi = \big(\Psi_h^\pi = \mathbb{E}_\pi[\psi_h(s_h, a_h)]\big)_{h\in[H]}, \tag{3}$$

where the expectation is taken under the trajectory $\{(s_h, a_h)\}_{h\in[H]}$ *induced by policy* π *and the underlying transition.*

The kernel embedding in (3) represents the agent's visitation distribution on every state-action pair under the policy π. The kernel method is also frequently used in existing MDP literature, as it can be used to incorporate function approximation when designing learning targets. For example, The reward function is often regarded as a linear function of a kernelized feature mapping in RL literature (Yang et al., 2020; Jin et al., 2020; Ding et al., 2021; Wu et al., 2021). When the reward function is known, we can also take the reward function as the kernel feature (Kakade et al., 2020; Uehara et al., 2022). In an MDP with an underlying kernelized structure, we can evaluate the agent's policy by its initial state value function $V_1^\pi = \mathbb{E}_\pi[\sum_{h=1}^H c_h(s_h, a_h)]$. The function V_1^π can be reformulated to $V_1^\pi = \theta \cdot \Psi^\pi$ under the linear function approximation case (Yang et al., 2020; Jin et al., 2020), which is a linear mapping with respect to the kernel embedding. In the general case, when ψ_h is not given, we can learn it through a supervised learning oracle or a model-free exploration, and subsequently employ it in our downstream algorithms, e.g. see Algorithm 1 in Modi et al. (2022). Thus, in an MDP-related optimization, it is reasonable to use the kernel embedding as a measure of how a state-action pair (s, a) contributes to the objective. To this end, we aim to solve the following optimization problem defined as a constrained convex MDP,

$$\min_{\pi\in\Delta(\mathcal{A}\,|\,\mathcal{S},H)} f\big(\Psi^\pi\big) \quad \text{s.t.} \quad g(\Psi^\pi) \le 0. \tag{4}$$

To measure the efficiency of policies in the first T episodes, we introduce the following performance measures,

$$\text{Regret}(T) = T\big(f\big(\Psi^{\widehat{\pi}}\big) - f\big(\Psi^{\pi^*}\big)\big), \quad \text{Violation}(T) = Tg\big(\Psi^{\widehat{\pi}}\big). \tag{5}$$

Here $\Psi^{\widehat{\pi}} = 1/T\sum_{t=1}^T \Psi^{\pi_k}$ is the average kernel embedding corresponding to the mixed policy $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$ of the first episode. By mixed policy we mean the agent rolls out and performs a random policy of index from 1 to T in equal probability at the beginning state. The performance measures in (5) are widely adopted by previous works in RL where a convex objective function is concerned (see, e.g., Ding et al. (2021); Brantley et al. (2021); Yu et al. (2021); Wu et al. (2021)).

We remark that our model is more general than the standard RL problem. To see this, we can reduce the C^{2}MDP to standard RL by setting $\psi_h(s_h, a_h) = c_h(s_h, a_h)$, f as the linear mapping with a H-dimension one hot feature vector, and removing the constraint.

Example 1 (Multi-objective MDP, (Yu et al., 2021; Wu et al., 2021)) *A Multi-objective MDP considers the following problem,*

$$\min_{\pi\in\Delta(\mathcal{A}|\mathcal{S},H)} h_1(\boldsymbol{V}^\pi) \qquad s.t. \qquad h_2(\boldsymbol{V}^\pi) \le 0, \tag{6}$$

where

$$\boldsymbol{V}^{\pi} = \Big(\mathbb{E}_{\pi}\Big[\sum_{h=1}^{H} c_h^i(s_h, a_h)\Big]\Big)_{i\in[I]}$$

is the initial state value function vector, and $h_1, h_2 : \mathbb{R}^I \to \mathbb{R}$ are 1-Lipschitz convex functions. If we use linear function approximation for the cost function, i.e.

$$c_h^i(s,a) = \psi(s,a)\cdot\theta_h^i, \quad \forall i\in[I]$$

the Multi-objective MDP turns into a constrained convex MDP,

$$\min_{\pi\in\Delta(\mathcal{A}|\mathcal{S},H)} h_1\big(\Xi\cdot\Psi^{\pi}\big) \qquad s.t. \qquad h_2\big(\Xi\cdot\Psi^{\pi}\big)\le 0.$$

Here $\Xi = (\theta_1^{1,\top}, \cdots \theta_H^{1,\top}; \cdots \theta_1^{I,\top}, \cdots, \theta_H^{I,\top})$ is a matrix formed by concatenating by $\{\theta_h^i\}_{i\in[I]}^{h\in[H]}$. Note that when $I = 1$ and $h_1(x) = h_2(s) = x$, Multi-objective MDP reduces to the constrained MDP in Ding et al. (2021) and Efroni et al. (2020). We also claim that our model is more general than the one in (Yu et al., 2021; Wu et al., 2021), since they assume h_1, h_2 to be monotone in all components and h_2 can only take the form $d(x, \mathcal{W})$, with $\mathcal{W}$ being a convex set.

Example 2 (Feasibility/Apprenticeship Learning, (Abbeel and Ng, 2004a; Syed et al., 2008; Mi *Feasibility learning considers minimizing the distance between the kernel embedding of the probability induced by the performance policy and a convex set $\mathcal{W}$, i.e.,*

$$\min_{\pi\in\Delta(\mathcal{A}|\mathcal{S},H)} \mathrm{dist}(\Psi^{\pi}, \mathcal{W}). \tag{7}$$

Here dist *can be chosen as any sort of discrepancy measure.When $\mathcal{W}$ reduces to a singleton $\{\Psi = (\mathbb{E}_{\mathcal{P}}[\psi(s_h, a_h)])_{h\in[H]}\}$, i.e. the kernel embedding of a given probability distribution $\{\mathcal{P}_h\}_{h\in[H]}$, the optimization reduced to apprenticeship learning.*

2.3 Examples of the Underlying Transition Models

Recall that C^{2}MDP is defined for any decision problem with a given linear kernel in its objective. With additional assumptions on the underlying transition, we can define different algorithms for solving it. The transition models we discuss here are (1) Kernelized Nonlinear Regulator (KNR) setting and (2) Low-rank MDP setting, which cannot be solved by algorithms design for tabular setting.

Kernelized Nonlinear Regulator. The *Kernelized Nonlinear Regulator* setting generalizes the linear quadratic regulator (LQR) setting (Kakade et al., 2020) and is especially helpful in continuous control problems. A KNR is an MDP with the following transition model,

$$s_{h+1} = W^{*}\phi(s_h, a_h) + \epsilon, \quad \epsilon\sim\mathcal{N}\big(0, \sigma^2\mathcal{I}\big) \tag{8}$$

for all $h\in[H]$, where $\phi : \mathcal{S}\times\mathcal{A}\to\mathcal{H}$ is a given kernel feature mapping of a d-dimension euclidean space $\mathbb{R}^d$ (Kakade et al., 2020; Mania et al., 2020; Song and Sun, 2021). The transition parameterization W^* characterizes the mapping from the feature $\phi(s_h, a_h)$ to the expectation of the next state s_{h+1}. We also remark that the KNR is a general model in the sense that both the state space $\mathcal{S}$ and the action space $\mathcal{A}$ can be continuous.

Low-rank MDP. In a Low-rank MDP(Uehara and Sun, 2021; Agarwal et al., 2020; Modi et al., 2022), the underlying transition takes the form

$$\mathcal{P}_h^*(s_{h+1} \mid s_h, a_h) = \langle \phi_h^*(s_h, a_h), \mu_h^*(s_{h+1}) \rangle \tag{9}$$

for all $h \in [H]$. Here the vector $\mu_h = (\mu_h^{(1)}, \dots, \mu_h^{(d)})$ is the concatenation of d unknown (signed) measures over $\mathcal{S}$. Unlike KNR , both the feature mapping ϕ_h^* and the measure μ_h^* in Low-rank MDP are unknown to the agent and need to be learned. For Low-rank MDP, it is natural to assume the agent access to two function classes $\Theta \subset \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\Upsilon \subset \mathcal{S} \to \mathbb{R}^d$ for candidate mappings for learning the true embeddings (μ_h^*, ϕ_h^*). Thus we make the following assumption,

Assumption 2 (Realizability) *The model class* (Θ, Υ) *with* $\{\mu_h^*\}_{h\in[H]} \subset \Theta$ *and* $\{\phi_h^*\}_{h\in[H]} \subset \Upsilon$ *is known, where both* Θ *and* Υ *are finite sets.*

Uehara et al. (2022) show that the case of finite function class can be easily generalized to infinite case. When feature ϕ^* is known, such a setting degerates to the linear MDP Yang and Wang (2019, 2020) Without loss of generality, we also make the following standard assumptions (Kakade et al., 2020; Uehara et al., 2022). The choice of the upper bound will not add complexity to our analysis.

Assumption 3 *We have the following assumptions.*

1. *For the KNR case, we assume that the feature* ϕ *of the underlying RKHS is uniformly bounded, i.e.,* $\|\phi(s, a)\|_2 \leq 1$ *for all* $(s, a) \in \mathcal{S} \times \mathcal{A}$*. For simplicity, we also assume that the transition parametrization satisfies* $\|W^*\|_2 \leq 1$*, here* $\|\cdot\|_2$ *is the matrix* 2*-norm.*

2. *For the Low-rank MDPs, we assume that* $\|\phi_h(s, a)\|_2 \leq 1$ *for all* $(s, a) \in \mathcal{S} \times \mathcal{A}$*, and for any function* $g : \mathcal{S} \to [0, 1]$ *and* $\mu \in \Upsilon$*,* $\| \int_{\mathcal{S}} \mu_h(s) g(s) d\nu \|_2 \leq \sqrt{d}$*, here* $\nu(\cdot)$ *is a given abstract measure defined on the state space* $\mathcal{S}$*.*

3. *For the kernel vectors* $\{\psi_h\}_{h\in[H]}$ *in the objective and constraint, we assume* $\|\psi_h(s, a)\|_2 \leq B$*.*

4. *We assume that the objective* f *and the constraint* g *in* (4) *are convex and* 1*-Lipschitz, which further implies that* $\{\|\partial f\|_2, \|\partial g\|_2\} \leq 1$*.*

For both cases mentioned above, the underlying transition probability is unknown, and can only be estimated through stochastic interactions with the environment. Thus, directly representing the set of all kernel embedding, i.e., $\mathcal{V} = \{\Psi^\pi : \text{any } \pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)\}$ is impossible, which makes (4) a challenging problem. As a consequence, we cannot simply regard (4) as a constrained optimization problem. Instead, we have to learn the optimal policy by collecting data via interacting the environment. Moreover, with the general constraint $g(\Psi^\pi) \leq 0$ on the distribution, the simple dual optimization method for set constraint (Yu et al., 2021) becomes infeasible. To address these challenges, we introduce a primal-dual algorithm in the subsequent section.

3. Main Algorithm

In this section, we provide a primal-dual algorithm Variational Primal-Dual Policy Optimization (VPDPO) for (4), which achieves sublinear in both regret and constraint violation.

3.1 Reformulation: Double Duality

In this subsection, we reformulate (4) as an unconstrained convex-concave problem, so that we can utilize the standard MDP method to solve it. Doing so will enable us to design a provably efficient algorithm.

The convex problem (4) is nontrivial only when its feasible set is none-empty. With the set of all reachable kernel embedding $\mathcal{V}$, we assume that $\mathcal{V} \cap \{g(\Psi^{\pi}) \leq 0\}$ is not empty, so that (4) is well-posed. To verify the convexity of feasible set (4), we first present the following proposition.

Proposition 4 (Convex Problem) *The generalized optimization problem in* (4) *is a convex problem.*

Proof See Appendix E.3 for detailed proof. ■

Next, we make the following assumption on g, which is standard in convex optimization and constrained convex MDP literature (Zahavy et al. (2021), Efroni et al. (2020), Ding et al. (2021)).

Assumption 5 (Slater Point) *There exists a policy π', such that* (4) *holds with strict inequality, i.e., $g(\Psi^{\pi'}) < 0$.*

Note that in Assumption 5, we do not require a pre-knowledge for π'. From an optimization perspective, a problem-dependent Slater condition is a measure of the size of the feasible region and determines the difficulty of solving a constrained optimization. The absence of such a condition may result in the lack of constraint qualification and cause failure in even simple optimization problems, for example, see Hijazi and Liberti (2016). With Assumption 5, we can reformulate (4) to a standard Lagrangian optimization problem (Corollary 28.1.1, Rockafellar (1970)). The Lagrangian function of (4) takes the form

$$\min_{\Psi \in \mathcal{V}} \max_{\gamma \geq 0} \big(f(\Psi) + \gamma \cdot g(\Psi)\big). \tag{10}$$

Slater's condition not only justifies the application of the Lagrangian duality but also allows us to bound the optimal value of the Lagrangian dual variable γ^* from above, which will further be helpful for our algorithm for the gradient update of the dual variables.

Lemma 6 (Bounded Lagrangian Dual Variable) *With Slater's condition in* (5) *, we have*

$$0 \leq \gamma^* \leq \Gamma := -\big(f(\Psi^{\pi'}) - f(\Psi^{\pi^*})\big)/g(\Psi^{\pi'}). \tag{11}$$

Proof See Appendix E for detailed proof. ■

Lemma 6 provides an upper bound for the optimal dual variable γ^*. In order to find γ^*, we only need to focus on the interval $[0, \Gamma]$. In practice, we only need to know an upper bound of Γ, which can be easily achieved through linear search.

Since f, g are 1-Lipschitz continuous and satisfy the closed-proper function condition, we have

$$f(\Psi^\pi) = \max_{\alpha \in \mathcal{B}^{dH}} \big(\alpha^\top \Psi^\pi - f^*(\alpha)\big), \quad \gamma g(\Psi^\pi) = \max_{\beta/\gamma \in \mathcal{B}^{dH}} \big(\beta^\top \Psi^\pi - \gamma \cdot g^*(\beta/\gamma)\big), \tag{12}$$

for all $\gamma \geq 0$ (Corollary 13.3.3, Rockafellar (1970)). Here f^* and g^* are the Fenchel duals of f and g, respectively. With these relations, we linearize the objective functions in (10) by introducing the variables α, β,

$$\min_{\Psi^\pi \in \mathcal{V}} \max_{\gamma \geq 0, \alpha, \beta/\gamma \in \mathcal{B}^{dH}} \mathcal{D}(\alpha, \beta, \gamma, \pi) = \big((\alpha + \beta)^\top \Psi^\pi - f^*(\alpha) - \gamma \cdot g^*(\beta/\gamma)\big). \tag{13}$$

We now reformulate the originally non-linear minimization problem into a min-max problem that is linear in Ψ and concave in (α, β, γ). Note that $\mathcal{V}$ is a closed convex set due to our setting and Assumption 3. Meanwhile, the feasible set for the dual variables (α, β, γ) is a convex compact set. Therefore, by the minimax theorem (Rockafellar, 1970), we can reformulate (13) to

$$\max_{\gamma \geq 0, \alpha, \beta/\gamma \in \mathcal{B}^{dH}} \min_{\Psi^\pi \in \mathcal{V}} \mathcal{D}(\alpha, \beta, \gamma, \pi) = \big((\alpha + \beta)^\top \Psi^\pi - f^*(\alpha) - \gamma \cdot g^*(\beta/\gamma)\big). \tag{14}$$

In the rest of this paper, we denote by α^*, β^* and γ^* the optima of the dual variables in (14), $\pi^* = \{\pi_h^*\}_{h \in [H]}$ the optimal policy, and Ψ^* the kernel embedding corresponding to π^*. We can rewrite (14) as $\max_{\gamma, \alpha, \beta} \mathcal{L}(\gamma, \alpha, \beta)$, where $\mathcal{L}(\gamma, \alpha, \beta) = \min_{\Psi^\pi \in \mathcal{V}} \mathcal{D}(\alpha, \beta, \gamma, \pi)$. When the dual variables are fixed, it suffices to implement model-based value iteration for solving $\mathcal{L}(\gamma, \alpha, \beta)$. By simultaneously updating γ, α and β, we can reach optimality by a primal-dual method.

Remark 7 *In* (14)*, the term $\gamma g^*(\beta/\gamma)$ is a convex function composed with a perspective function, so it must be convex in (β, γ). See Boyd et al. (2004) for details.*

3.2 Solution: Primal-Dual Method

The minimax structure in (13) implies us to implement a primal-dual method. Such implementation is common when facing nonlinearity in visitation measures (e.g., Wu et al. (2021) and Efroni et al. (2020)).

Dual Update. We perform an online projected gradient ascent method for a dual update. In each iteration, we update α by moving α^k to a direction of maximizing the dual function $\mathcal{D}(\alpha, \beta, \gamma, \pi)$ and then project it to the unit ball. To represent the projection set for (β, γ), we combine the restriction imposed by Fenchel dual and Slater's condition and define

$$\mathcal{G} = \{(\beta, \gamma) : \|\beta\|_2 \leq \gamma, \gamma \in [0, \Gamma]\}.$$

When the Slater's condition holds, the optimal solution (β^*, γ^*) always lies in $\mathcal{G}$ by Lemma 6. If we know the underlying transition map W^* in priori, we can solve the outer iteration of the minimax problem in (13) by value iteration and implement Ψ^{π^t} in the gradient ascent step. However, since the transition remains obscure to us, we use $\Psi^t = (\Psi_h)_{h\in[H]}$ as a proxy, where $\Psi_h^t = \mathbb{E}_{\pi,\mathcal{P}^t}[\Psi_h(s_h,a_h)]$. In the dual update, the step size η^t is set as $O(1/\sqrt{t})$ (or $O(1/\sqrt{T})$ when T is given). In Algorithm 1, ∂_γ and ∂_β are the subgradient operator with γ and β as the variable, respectively.

Algorithm 1 Variational Primal-Dual Policy Optimization

```
Require: Step size $\{\eta^t\}_{t=1}^T$, $\alpha^1 \in \mathcal{B}^{dH}$, $\gamma^1 \in [0,\Gamma]$, $\beta^1 \in \gamma^1 \cdot \mathcal{B}^{dH}$
 1: for $t = 1,\dots,T$ do
 2:     $\alpha^{t+1} \leftarrow \Pi_{\mathcal{B}^{dH}}\{\alpha^t + \eta^t(\Psi^t - \partial f^*(\alpha^t))\}$
 3:     $\widehat{\beta}^{t+1} \leftarrow \beta^t + \eta^t(\Psi^t - \partial_\beta g^*(\beta^t/\gamma^t))$
 4:     $\widehat{\gamma}^{t+1} \leftarrow \gamma^t + \eta^t(\partial_\gamma(-\gamma^t g(\beta^t/\gamma^t)))$
 5:     $(\beta^t,\gamma^t) = \Pi_{\mathcal{G}}(\widehat{\beta}^{t+1},\widehat{\gamma}^{t+1})$
 6:     $\theta^{t+1} \leftarrow \alpha^{t+1} + \beta^{t+1}$
 7:     Update the cost function $\{c_h^{t+1}(s,a) = \theta_h^{t+1} \cdot \psi(s,a)\}_{h\in[H]}$
 8:     Update the confidence set $\mathcal{C}^{t+1}$ by Algorithm 2 or 3
 9:     $(\pi^{t+1},\mathcal{P}^{t+1}) \leftarrow \mathrm{argmin}_\pi \min_{\mathcal{P}\in\mathcal{C}^{t+1}} V_{1,\mathcal{P}}^{t+1,\pi}$.
10:     Calculate $\Psi^{t+1} = (\mathbb{E}_{\pi^{t+1},\mathcal{P}^{t+1}}[\psi(s_h,a_h)])_{h\in[H]}$
11: end for
```

Primal Update: Construct a cost. Algorithm 1 further relies on the agent's exploration to estimate the transition $\mathcal{P}^t$ with experience in the previous $t-1$ episodes. Since an explicit cost does not necessarily occur in our optimization problem, to implement value iteration, we construct a cost by introducing the dual vector $\theta^t = \alpha^t + \beta^t$ for all t, and set a temporary reward $c_h^t = \psi_h \cdot \theta_h$. Note that in the minimax problem (13), with fixed (α,β), the objective function turns into

$$\min_\pi \big((\alpha+\beta)\cdot\Psi^\pi\big) = \sum_{h=1}^H \mathbb{E}_{\pi,\mathcal{P}^*}[\Psi_h(s_h,a_h)\cdot(\alpha_h+\beta_h)],$$

which can be viewed as an accumulative cost minimization problem. This is essentially an optimal control problem. Corresponding to c_h^t, we set the value functions

$$V_{h,\mathcal{P}}^{t,\pi}(s) = \mathbb{E}_{\pi,\mathcal{P}}\bigg[\sum_{i=h}^H c_h^t(s_i,a_i)\,\bigg|\, s_h = s\bigg], \tag{15}$$

$$Q_{h,\mathcal{P}}^{t,\pi}(s,a) = \mathbb{E}_{\pi,\mathcal{P}}\bigg[\sum_{i=h}^H c_h^t(s_i,a_i)\,\bigg|\, s_h = s, a_h = a\bigg]. \tag{16}$$

for policy π. For simplicity, we denote $V_{h,\mathcal{P}}^{t,\pi_t}$ and $Q_{h,\mathcal{P}}^{t,\pi_t}$ as $V_{h,\mathcal{P}}^t$ and $Q_{h,\mathcal{P}}^t$, respectively. Here and in the rest of this paper, we denote by $\mathbb{E}_{\pi,P}[\cdot]$ the expectation taken over the trajectory $\{(s_h,a_h)\}_{h\in[H]}$ induced by $\{\pi_h\}_{h\in[H]}$ and the underlying transition kernel $\mathcal{P}$. With the confidence set $\mathcal{C}^t$ given by Algorithms 2 and 3, Line 9 in Algorithm 1 follows the principle of "*Optimism in the Face of Uncertainty*", and chooses the policy and model in the confidence set that can incur the smallest cost. We highlight that Algorithm 1 is a model-based algorithm, as it explicitly learns the underlying transition probability.

Algorithm 2 VPDPO for KNR case

```
Require: $\{(s_h^i, a_h^i)\}_{i\in[t],h\in[H]}$ , $\lambda > 0$, $C_1 > 0$, $\theta^t$,$\Lambda_0 = \lambda I$, $\pi_0 = a_0$
Execute $\pi^t$ to sample a new trajectory $\{(s_h^t, a_h^t)\}_{h\in[H]}$
$\widehat{W}^t \leftarrow \arg\min_W \sum_{\tau=1}^t \sum_{h=1}^H \left\|W\phi\left(s_h^\tau, a_h^\tau\right) - s_{h+1}^\tau\right\|_2^2 + \lambda\|W\|_F^2$.
$\Lambda^t \leftarrow \lambda I + \sum_{\tau=1}^t \sum_{h=1}^H \phi\left(s_h^\tau, a_h^\tau\right)\phi\left(s_h^\tau, a_h^\tau\right)^\top$.
Update $\mathcal{C}^t \leftarrow \left\{\mathcal{P} \mid \left\|\left(W - \widehat{W}^t\right)\left(\Lambda^t\right)^{1/2}\right\|_2^2 \le R^t, \|W\|_2 \le 1, \mathcal{P}$ parametrized by $W\right\}$
    with $R^t$ defined in (17).
return Confidence set $\mathcal{C}^t$
```

Algorithm 3 VPDPO for Low-rank MDP case

```
Require: model set $\mathcal{M} = \{(\mu, \phi) : \mu \in \Upsilon, \phi \in \Theta\}$, $\mathcal{D}_{0,h} = \varnothing$, $\pi_0 = U(\mathcal{A})$
Collect a set of tuples $\{(s_h^t, a_h^t, s_{h+1}^t)\}_{h\in[H-1]}$ by rolling out $s_h^t$ with policy $\pi_t$ and then
    select $a_h^t$ by a uniform distribution on $\mathcal{A}$, i.e. $a_h^t \sim U(\mathcal{A})$, $s_{h+1}^t \sim \mathcal{P}(\cdot \mid s_h^t, a_h^t)$, .
Update $\mathcal{D}_{h,t} = \mathcal{D}_{h,t-1} \cup \{(s_h^t, a_h^t, s_{h+1}^t)\}$.
$(\widehat{\mu}_h^t, \widehat{\phi}_h^t) \leftarrow \mathrm{argmax}_{(\mu,\phi)\in\mathcal{M}} \mathbb{E}_{\mathcal{D}_{h,t}}[\log(\mu(s_{h+1})^\top \phi(s_h, a_h))]$ .
$\widehat{\mathcal{P}}_h^t(\cdot \mid s_h, a_h) \leftarrow \widehat{\mu}_h^t(\cdot)^\top \widehat{\phi}_h^t(s_h, a_h)$.
Update $\mathcal{C}^t \leftarrow \left\{\mathcal{P} = \{\mathcal{P}_h\}_{h\in[H]} \middle| \mathbb{E}_{\mathcal{D}_h^t}\left[\|\widehat{\mathcal{P}}_h^t(\cdot \mid s_h, a_h) - P_h(\cdot \mid s_h, a_h)\|_1^2\right] \le R^t\right\}$ with $R^t$
    defined in (18)
return Confidence set $\mathcal{C}^t$
```

Algorithms 2 and 3 interact with the environment with policy $\pi^t = \{\pi_h^t\}_{h\in[H]}$ given by Algorithm 1, and then construct confidence set for possible models. In each episode, we construct a confidence set $\mathcal{C}^t$, whose center and weighted radius are designed deliberately. The center of the confidence set is chosen by the maximum likelihood estimation (MLE), and the weighted radius R^t is chosen so that the real transition mapping $\mathcal{P}^*$ lies in $\mathcal{C}^t$ for every t with a high probability. Specifically, in Algorithm 2 we set

$$R^t = c(\lambda\sigma^2 + \sigma^2(d + \log(t\det(\Lambda^t)/\delta\det(\Lambda^0))) \tag{17}$$

for the KNR case, and in Algorithm 3 for Low-rank MDP we set

$$R^t = c\log(TH|\Upsilon||\Theta|/\delta)/t \tag{18}$$

The difference between Algorithms 2 and 3 is that, in the KNR setting, the agent collects a full trajectory by performing the same policy π_t, while in the Low-rank MDP setting, for each epoch t and for $h \in [H]$, the agent performs π^t for the first h step and then augment the trajectory by a randomly choose an action and then transit to the next state, i.e., $a_h \sim U(\mathcal{A})$, $s_{h+1} \sim \mathcal{P}(\cdot \mid s_h, a_h)$. Note that this exploration manner only influences the degree of H in the sampling complexity, and does not affect the sublinear regret and violation.

We remark on the computation efficiency of Algorithms 1-3. For Algorithm 1, the projection set $\mathcal{G}$ for dual variable (β, γ) can be seen as an intersection of a second-order cone $\{(x,t) : \|x\|_2 \leq t\}$ and a half space $\{(x,t) : t \in [0, \Gamma]\}$. Projection to both sets has a closed-form solution. The projection to $\mathcal{G}$ can thus be computed via implementing the alternating projection method, which involves a sequence of gradient steps and projection (Bregman, 1967). The proxy Ψ^t can be estimated by Monte Carlo method, with W^t as a known transition. We would also like to remark that the calculation of Line 9 of Algorithm 1, known as the *optimistic planning*, is in general NP-hard (Dani et al., 2008), and we assume there is an oracle to implement it (Kakade et al., 2020; Uehara and Sun, 2021; Jin et al., 2021; Ayoub et al., 2020). Then we only focus on the statistical complexity. From that, we make the following assumption.

Assumption 8 (Black-box Computation Oracle) *We assume that there is an oracle that implements Line 9 of Algorithm 1.*

In practice, several effective heuristics may be available through gradient-based methods such as iLQG (Todorov and Li, 2005), and CIO Mordatch et al. (2012), or sampling-based methods, such as MPPI (Williams et al., 2015) and DMDMPC (Wagener et al., 2019).

In the Low-rank MDP setting, motivated by the estimation of conditional probability (Uehara et al., 2022; Agarwal et al., 2020), we use MLE for estimating the underlying transition. Unlike in the KNR case where the MLE has a closed-form solution, it is hard to find a general closed-form solution for representation learning by MLE. Correspondingly, we need an oracle for efficient MLE computation for Line 1 in Algorithm 3.

Assumption 9 (Maximum-Likelihood Estimation) *Consider the model class $\mathcal{M}$ and a dataset $\mathcal{D}$ in the form of (s, a, s'), the MLE oracle returns the maximum likelihood estimator,*

$$(\widehat{\mu}, \widehat{\phi}) = \operatorname{argmax}_{(\mu,\phi)\in\mathcal{M}} \mathbb{E}_{\mathcal{D}}\big[\log\big(\mu\left(s'\right)^\top \phi(s,a)\big)\big],$$

which implements Line 3 of Algorithm 3.

We assume there exists practical algorithms that avoid explicitly enumerating over all functions in the model space $\mathcal{M}$. In practice, such oracles can be reasonably approximated whenever optimizing over $\mathcal{M}$ is feasible, such as in neural networks.

4. Theoretical Results

In this section, we provide theoretical analysis for Algorithms 1 and 2. For the regret and the constraint violation, we make the decompositions

$$T\big(f(\Psi^{\widehat{\pi}}) - f(\Psi^*)\big) = \underbrace{T\big(f(\Psi^{\widehat{\pi}}) - f(\widehat{\Psi})\big)}_{\text{(R.i)}} + \underbrace{T\big(f(\widehat{\Psi}) - f(\Psi^*)\big)}_{\text{(R.ii)}},$$

$$T\big(g(\Psi^{\widehat{\pi}})\big) = \underbrace{T\big(g(\Psi^{\widehat{\pi}}) - g(\widehat{\Psi})\big)}_{\text{(V.i)}} + \underbrace{T \cdot g(\widehat{\Psi})}_{\text{(V.ii)}},$$

where we recall that $\Psi^{\widehat{\pi}} = 1/T\sum_{t=1}^T \Psi^{\pi_t}$, $\widehat{\Psi} = 1/T\sum_{t=1}^T \Psi^t$. Here (R.i) and (V.i) are the estimation errors incurred by the noise in the regression. With the Lipschitz condition imposed on f and g, it suffices to bound $\| \sum_{t=1}^{T}(\Psi^{\pi_t} - \Psi^t)\|_2$. We reformulate it into bounding a value difference summation $\sum_{t=1}^{T}(V_{1,\mathcal{P}^t}^t - V_{1,\mathcal{P}^*}^t)$. The gradient update for the dual variables allows us to give an upper bound for (R.ii) and (V.ii) in terms of a value difference sequence as well.

However, we first need to handle the non-linearity in (4). By implementing the online gradient ascent method in Algorithm 1, we can guarantee that the following coupling term can be bounded by the value difference of two processes and an $O(\sqrt{T})$ term.

Lemma 10 (Dual Update: Gradient Ascent) *For all $\gamma \in [0, \Gamma]$, we have*

$$T \cdot \left[f(\widehat{\Psi}) - f(\Psi^*) + \gamma \cdot g(\widehat{\Psi})\right] \leq \sum_{t=1}^{T} \theta^t \cdot (\Psi^t - \Psi^*) + CB\Gamma\sqrt{HT}, \tag{19}$$

where $C > 0$ is an absolute constant.

Proof See Appendix B for detailed proof. ■

Lemma 10 displays a coupling between the regret and the constraint violation, which is also frequently met in online algorithms using dual updates, such as CMDP and Multi-objectives (Ding et al., 2021; Yu et al., 2021). The proof of Lemma 10 incorporates the standard regret analysis of online gradient ascent and the self-dual property of Fenchel dual, which is a common technique in analyzing nonlinear function differences with gradient updates. The occurrence of the coupling term directly comes from the gradient update of the dual variables in Algorithm 1.

In the following lemma we introduce the difference of a sequence of projected kernel embedding, which can be interpreted as the performance difference of two systems in T episodes. When $\mathcal{P}^*$ falls in $\mathcal{C}^t$, by the principle of optimism implemented in Line 9 of Algorithm 1, the value difference is always negative. In this paper, we denote the event of $\mathcal{P}^* \in \mathcal{C}^t$ for all $t \in [T]$ by $\mathcal{E}_{cb}$, i.e., $\mathcal{E}_{cb} = \cup_{t=1}^{T}\{\mathcal{P}^* \in \mathcal{C}^t\}$. By the construction of the confidence set, we can further prove that $\mathcal{P}^*$ always lies in $\mathcal{C}^t$ with the probability of at least $1 - \delta$. With the construction of confidence set in Algorithms 2 and 3, we choose the transition model and policy that would incur the highest accumulative reward in expectation. Therefore, as long as the real dynamic falls in the confidence set, we can obtain optimism in the sense of the following lemma,

Lemma 11 (Optimism: Value Difference) *If the real model $\mathcal{P}^*$ falls in the confidence set $\mathcal{C}^t$ for all t, then we have the following inequality,*

$$\sum_{t=1}^{T} \theta^t \cdot (\Psi^t - \Psi^*) \leq 0. \tag{20}$$

Proof The inequality comes from the construction of the cost function in (15) and the choice of $\mathcal{P}^t$ and π^t in Line 9 in Algorithm 1. ■

Conditioning on the event that Lemma 36 holds, we actually claim that the coupling term in (19) can be bounded by $O(\sqrt{T})$. Combining this with the optimization trick of Theorem 33, we can further prove that (R.i) and (V.i) are bounded by $O(\sqrt{T})$. We leave the detailed proof in Section B.

Lemma 12 *Assume that $\mathcal{P}^* \in \mathcal{C}^t$ for all $t \in [T]$. Then for all $\gamma \in [0, \Gamma]$, we have*

$$T(f(\widehat{\Psi}) - f(\Psi^*)) \leq CB\Gamma\sqrt{HT}, \tag{21}$$

$$T \cdot g(\widehat{\Psi}) \leq CB\sqrt{HT}. \tag{22}$$

We now bound the difference of the coupling of the objective and constraint violation by $\sqrt{T}$, with the estimated feature embedding $\widehat{\Psi}$ as a self variable. But what can we say about the difference between the estimated average feature embedding $\widehat{\Psi}$ and the real average feature embedding $\Psi^{\widehat{\pi}}$, $\|\Psi^{\widehat{\pi}} - \widehat{\Psi}\|_2$? To tackle this issue, we interpret the difference of the kernel mean embedding as the supreme of a set of value differences. For a fixed $x = (x_h)_{h\in[H]} \in \mathbb{R}^{dH}$ with $\|x\|_2 \leq 1$, we can consider $\sum_{t=1}^{T}(\Psi^{\widehat{\pi}} - \widehat{\Psi}) \cdot x$ as the value difference of two processes, with cost at stage h defined as $c_h(s_h, a_h) = \psi(s_h, a_h) \cdot x_h$. For simplicity, we denote $x \cdot (\Psi^{\pi_t} - \Psi^t) = V_1^{\pi_t} - V_1^t$. As long as we can uniformly upper bound $V_1^{\pi_t} - V_1^t$ for all $\|x\|_2 \leq 1$, we can give a bound for $\|\Psi^{\widehat{\pi}} - \widehat{\Psi}\|_2$. The following lemma allows us to decompose a value difference and is useful in our analysis.

Lemma 13 (Value Difference Lemma) *Consider two MDPs $\big(\mathcal{S}, \mathcal{A}, \{\mathcal{P}_h^1\}_{h=1}^H, \{r_h\}_{h=1}^H\big)$ and $\big(\mathcal{S}, \mathcal{A}, \{\mathcal{P}_h^2\}_{h=1}^H, \{r_h\}_{h=1}^H\big)$ and a given policy $\pi = \{\pi_h\}_{h\in[H]}$. Then for all $h \in [H]$ the following relation holds,*

$$V_h^{\pi}(s) - V_h^{\pi'}(s) = \mathbb{E}_{\pi, \mathcal{P}^2}\Bigg[\sum_{i=h}^{H}(\mathbb{P}_i^1 V_{i+1}^{\pi}(s_i, a_i) - \mathbb{P}_i^2 V_{i+1}^{\pi}(s_i, a_i))\,\Bigg|\, s_h = s\Bigg]. \tag{23}$$

Proof This lemma is a direct corollary of Lemma 36 in the appendix, as the two MDP share the same reward. ■

Next, we directly give the performance guarantees for KNR and low-rank MDP cases, and give a brief proof under this value difference routine for the two cases respectively. Both results contain a $O(\sqrt{T}\log T)$ scale in the regret and violation, which shows that VODPO learns in C^{2}MDPs in a statistically efficient manner. As T grows bigger, the mixed policy $\widehat{\pi}$ would achieve an suboptimality that decreases in a $O(\log T/\sqrt{T})$ manner. To the best of our knowledge, this algorithm is the first one that achieves sublinear regret and constraint violation in C^{2}MDP.

4.1 Analysis of the KNR Case

Theorem 14 *Assume that Assumptions 3-5 and 8 hold. Set $\lambda = \max\{\sigma^2, 1\}$. For Algorithm 1 and 2, with probability at least $1 - \delta$, the regret is bounded by*

$$\mathrm{Regret}(T) \leq O\Bigg(\Gamma B\sqrt{HT} + CBHd\sqrt{T}\log\left(\frac{HT}{d\delta}\right)\Bigg),$$

and the constraint violation is bounded by

$$\mathrm{Violation}(T) \leq O\Bigg(CBHd\sqrt{T}\log\left(\frac{HT}{d\delta}\right)\Bigg).$$

4.2 Proof Sketch of Theorem 14

In this section, we sketch the proof of Theorem 14. The detailed proof is deferred to Appendix C.

Lemma 15 (Simulation Lemma) *For any policy π, feature mapping W, bounded cost c, and for any initial state s_1, with the value function defined in* (15)*(with a upper bound of $\sqrt{H}$), we have*

$$V^{\pi}_{1,\mathcal{P}^*}(s_1) - V^{\pi}_{1,\mathcal{P}}(s_1) \leq O\bigg(B\sqrt{H} \cdot \mathbb{E}_{\pi,\mathcal{P}^*}\bigg[\sum_{h=1}^{H} \big\|(W^\star - W)\phi(s_h, a_h)\big\|_2\bigg]\bigg),$$

where the state-value function is defined with underlying cost c. Here $\mathcal{P}^*$ and $\mathcal{P}$ are the conditional distribution induced by W^* and W, respectively.

Proof With Lemma 13 we have

$$\begin{aligned} V^{\pi}_{1,\mathcal{P}^*}(s_1) - V^{\pi}_{1,\mathcal{P}}(s_1) &\leq \mathbb{E}_{\pi,\mathcal{P}^*}\bigg[B\sqrt{H}\sum_{h=1}^{H} \big\|\mathcal{P}^*_h(\cdot\,|\,s_h, a_h) - \mathcal{P}_h(\cdot\,|\,s_h, a_h)\big\|_1\bigg] \\ &\lesssim B\sqrt{H}\mathbb{E}_{\pi,\mathcal{P}^*}\bigg[\sum_{h=1}^{H} \big\|(W^\star - W)\phi(s_h, a_h)\big\|_2\bigg], \end{aligned}$$

where the second inequality follows from the estimation

$$\big\|\mathcal{P}^*_h(\cdot\,|\,s_h, a_h) - \mathcal{P}_h(\cdot\,|\,s_h, a_h)\big\|_1 = O\Big(\big\|(W^\star - W)\phi(s_h, a_h)\big\|_2\Big)$$

from Devroye et al. (2018). Here we drop the constants that only depend on σ. ■

By Lemma 15 and the Elliptical Potential Lemma (Uehara and Sun, 2021), following the value decomposition routine, we give an upper bound for the estimation error in terms of the maximum information gain in the following lemma.

Lemma 16 (Estimation Error) *For Algorithms 1 and 2, with $\lambda = \max\{\sigma^2, 1\}$, we have $\mathcal{P}^* \in \mathcal{C}^t$ for all $t \in [T]$ holds with probability at least $1-\delta$, and*

$$T\|\Psi^{\widehat{\pi}} - \widehat{\Psi}\|_2 \leq CBHd\sqrt{T}\log\left(\frac{HT}{d\delta}\right) \tag{24}$$

holds with probability at least $1-\delta$, where $C > 0$ is an absolute constant that only depends on σ.

Proof See Appendix B for detailed proof. ■

With the 1-Lipschitz assumption for f, g, Lemma 16 in fact gives a uniform upper bound for (R.i) and (V.i). Combining the results on regret and constraint violation in Lemma 12 with the error estimation in Lemma 16, we finish the proof of Theorem 14.

4.3 Analysis of the Low-rank MDP case

For the Low-rank MDPs, we also prove the sublinear regret and violation under Algorithms 1 and 3.

Theorem 17 *Assume that Assumptions 3-5 and 9 hold. Set R^t as in* (18). *For Algorithms 2 and 3, with probability $1-\delta$, the regret is bounded by*

$$\mathrm{Regret}(T) \leq O\bigg(\Gamma B\sqrt{HT} + B\sqrt{TH|\mathcal{A}|d^2}\log\bigg(\frac{TH|\Theta||\Upsilon|}{\delta}\bigg)\bigg),$$

and the constraint violation is bounded by

$$\mathrm{Violation}(T) \leq O\bigg(B\sqrt{TH|\mathcal{A}|d^2}\log\bigg(\frac{TH|\Theta||\Upsilon|}{\delta}\bigg)\bigg).$$

We remark that our regret and constraint violation guarantees in Theorem 14 and 17 serve as *Probably Approximately Correct (PAC)* bounds: with probability at least $1-\delta$, we can obtain a Markov policy $\widehat{\pi} := \frac{1}{T}\sum_{t=1}^{T}\pi^t$ such that $f(\Psi^{\widehat{\pi}}) - f(\Psi^{\pi^*}) = O(1/\sqrt{T})$, and the constraint violation $g(\Psi^{\widehat{\pi}}) = O(1/\sqrt{T})$. Consequently, with a sample complexity of $O(1/\epsilon^2)$, the Markov policy $\widehat{\pi}$ such that $f(\Psi^{\widehat{\pi}}) - f(\Psi^{\pi^*}) \leq \epsilon$ and $g(\Psi^{\widehat{\pi}}) \leq \epsilon$ hold simultaneously with hight probability. From an asymptotic perspective, with T tends to infinity, $f(\Psi^{\widehat{\pi}})$ converges to the optimal value, while the violation of constraint $g(\Psi^{\widehat{\pi}})$ can be arbitrarily small with high probability. Our result is different in form from the standard definitions in online convex optimization due the existence of both optimality gap and constraint violation.

4.4 Proof Sketch of Theorem 17

In this section we briefly sketch the proofs of efficiencies of Algorithm 1 and 3 in the Low-rank MDP setting. For detailed proof, see Appendix D.
We define the state-action visitation induced by the mixed Markov policy before epoch t and the one augmented by choosing random action,

$$\rho_h^t(s_h, a_h) = \frac{1}{t-1}\sum_{i\in[t-1]} d_{\pi^i,h,\mathcal{P}^*}(s_h, a_h), \quad \widehat{\rho}_h^t(s_h, a_h) = \frac{1}{t-1}\sum_{i\in[t-1]} d_{\pi^i,h,\mathcal{P}^*}(s_h)u(a_h),$$

where $d_{\pi,h,\mathcal{P}}(s_h, a_h)$ is the visitation probability on the h-th state-action pair induced by policy π and transition kernel $\mathcal{P}$, and $u(a)$ is the uniform distribution on the action set $\mathcal{A}$. By implementing MLE in every epoch t, we claim that with high probability, the model error under the distribution of the previous policy $\mathbb{E}_{\widehat{\rho}^t}[\|\widehat{\mathcal{P}}_h^t(\cdot|s_h, a_h) - \mathcal{P}_h^*(\cdot|s_h, a_h)\|_1^2]$ is of $\widetilde{O}(1/t)$. With a standard Bernstein-type argument for martingales, we have the following lemma.

Lemma 18 (Shrinking Confidence Ball) *With probability at least $1-\delta$, we have $\mathcal{P}^* \in \mathcal{C}^t$ and*

$$\mathbb{E}_{\widehat{\rho}_h^t}\left[\|\mathcal{P}_h^*(\cdot \mid s_h, a_h) - \mathcal{P}_h(\cdot \mid s_h, a_h)\|_1^2\right] \leq \frac{c\log(TH|\Upsilon||\Theta|/\delta)}{t},$$

for all transition $\mathcal{P} \in \mathcal{C}^t$, $t \in [T]$ and $h \in [H]$, where c is an absolute constant. Here $\widehat{\mathcal{P}}_h^t$ is the transition learned by the MLE in Algorithm 3.

Proof See Appendix D for details. ■

Lemma 18 implies that with high probability, our choice of the confidence set $\mathcal{C}^t$ is good enough for the real transition to fall in. Moreover, the distance between $\mathcal{P}^*$ and other elements in $\mathcal{C}^t$ also decreases under the distribution of the mixed policy $\widehat{\rho}^t$. By the construction of the value function of (15), we obtain the following lemma. As in the KNR case, we also care for the error brought by our insufficient model estimation, $\|\Psi^t - \Psi^{\pi_t}\|_2$. To overcome this tissue, the underlying linear structure of low-rank MDPs is crucial. We introduce the following lemma, which is a modification of Lemma 16 in Uehara et al. (2022):

Lemma 19 *Take any $h \in \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that $\|h\|_\infty \le D$. Then,*

$$\mathbb{E}_\pi[h(s_h, a_h)] \le \mathbb{E}_\pi \|\phi^\star_{h-1}(s_{h-1}, a_{h-1})\|_{\Sigma^{-1}_{\rho^t,\phi^\star_{h-1}}} \sqrt{t|\mathcal{A}|\mathbb{E}_{\widehat{\rho}^t}[h^2(s,a)] + \lambda d D^2},$$

where $\Sigma_{\rho^t,\phi^\star_h} = t\mathbb{E}_{(s,a)\sim\rho^t}\left[\phi^\star_h(s,a)\phi^\star_h(s,a)^\top\right] + \lambda I$.

Proof See Appendix D for details. ■

Note that here the parameter λ and the matrix Σ_{ρ^t,ϕ^*_h} do not occur in the actual implementation. This lemma introduces an elliptical potential structure. By then, using the same method as in the KNR case, we prove the upper bound for the 2-norm estimation error.

Lemma 20 (Estimation Error) *With Assumption 3 and Algorithm 3, we have*

$$T\|\widehat{\Psi} - \Psi^{\widehat{\pi}}\|_2 \le c\sqrt{TH|\mathcal{A}|d^2} \log\left(\frac{TH|\Theta||\Upsilon|}{\delta}\right)$$

holds with probability at least $1-\delta$.

Note that f, g are both 1-Lipschitz by Assumption 4, we can thereby control the upper bound of $|f(\Psi^\pi) - f(\widehat{\Psi})|$ and $|g(\Psi^\pi) - g(\widehat{\Psi})|$ can be bounded by the same scale, combine this with Lemma 12 concludes our proof.

4.5 Applications to Concrete Examples

With the general results above, we also highlight their applications on concrete examples rise in RL. In Section 2 we introduced several settings that are well known in MDP literaturewhich can be regarded as examples of C^{2}MDP, with f being their objectives and g being their constraints. We then implement VODPO to solve them, In this section we use Multi-objective MDPs and Feasiblity Learning as examples to show the power of VODPO. First, we have the following corollary for the KNR setting,

Corollary 21 *Under Assumptions 2 - 8, we assume that $\|\theta^i_h\|_2 \le \sqrt{d}$ for all $i \in [I]$ and $h \in [H]$. Set $\lambda = \max\{\sigma^2, 1\}$, we have*

$$\mathrm{Regret}(T) \le \begin{cases} O\big(\Gamma H\sqrt{IT} + CH^{3/2}d\sqrt{IT}\log\big(\frac{HT}{d\delta}\big)\big) & \textit{for Multi-objective MDP,} \\ O\big(CHd\sqrt{T}\log\big(\frac{HT}{d\delta}\big)\big) & \textit{for Feasibility Learning,} \end{cases} \tag{25}$$

and

$$\mathrm{Violation}(T) \le O\bigg(CH^{3/2}d\sqrt{IT}\log\bigg(\frac{HT}{d\delta}\bigg)\bigg) \qquad \textit{for Multi-objective MDP,}$$

hold with probability at least $1-\delta$. *Here* $C>0$ *is an absolute constant that only depends on* σ.

Under the low-rank MDP setting, we have similar results.

Corollary 22 *Under Assumptions 2 - 9, assuming that* $\|\theta_h^i\|_2 \le \sqrt{d}$ *for all* $i \in [I]$ *and* $h \in [H]$, *we have*

$$\mathrm{Regret}(T) \le \begin{cases} O\big(\Gamma H\sqrt{IdT} + H\sqrt{|\mathcal{A}|Id^3T}\log\big(\frac{TH|\mathcal{M}|}{\delta}\big)\big) & \textit{for Multi-objective MDP,} \\ O\big(\sqrt{H|\mathcal{A}|d^2T}\log\big(\frac{TH|\mathcal{M}|}{\delta}\big)\big) & \textit{for Feasibility Learning,} \end{cases} \tag{26}$$

and

$$\mathrm{Violation}(T) \le O\bigg(H\sqrt{|\mathcal{A}|Id^3T}\log\bigg(\frac{TH|\mathcal{M}|}{\delta}\bigg)\bigg) \qquad \textit{for Multi-objective MDP,}$$

hold with probability at least $1-\delta$. *Here* $|\mathcal{M}| = |\Theta||\Upsilon|$ *is total number of the model classes.*

We claim that when f degenerates to a linear function, our results recover the regret of standard KNR in Kakade et al. (2020). Specifically, our results in regret matches Theorem 3.2Kakade et al. (2020) in terms of H and d, where they accomplish a regret of $O(\sqrt{H^3d^2T})$. When considering a low-rank MDP with a finite horizon, Uehara et al. (2022) achieves a regret of $O(\sqrt{d^3H^2|\mathcal{A}|T})$, which is also consistent with our result for low-rank MDP case. We also compare our results of low-rank MDP with existing works such as Yu et al. (2021), which focuses on the study of online Multi-objective MDP under the tabular case. Tabular MDP can be regarded as a special case of low-rank MDP with a known feature, with the dimension $d = SA$. By assuming approachability, Yu et al. (2021) propose an algorithm with regret of $O(\Gamma\sqrt{IH^3S^2A/T})$ and a constraint violation of $O(\sqrt{IH^3S^2A/T})$, where S and A are the cardinality of $\mathcal{S}$ and $\mathcal{A}$, respectively. We claim that our results have a higher-order dependence on $d = SA$ due to the error inherited from MLE and the invoke of one step back inequality. For a technical understanding, we recommend the readers to Appendix D.

5. Conclusion

In this paper, we have developed a provably efficient online algorithm, Variational Primal-Dual Policy Optimization (VPDPO) for constrained constrained convex MDP. KNR and Low-rank MDP are two examples. The algorithm extends the reward-based RL algorithm to constrained convex MDP where no explicit reward is needed and incorporates the Lagrangian primal-dual method to transform the constrained optimization into a minimax problem. To handle the balance between exploration and exploitation, we follow the principle of optimism in the face of uncertainty. We prove that that our algorithm enjoys a $\widetilde{O}(\sqrt{T})$ regret and a $\widetilde{O}(\sqrt{T})$ violation with high probability under standard optimization assumptions, where T is the total number of episodes taken by the algorithm.

Acknowledgments

Mengdi Wang acknowledges the support by National Science Foundation grants DMS-1953686, IIS-2107304, CMMI1653435, CPS2312093, ONR grant1006977, Google Research and C3.AI. Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their support. Zhuoran Yang acknowledges Simons Institute (Theory of Reinforcement Learning) for their support. We would like to thank Yufeng Zhang, Sirui Zheng, and Runzhe Wu for helpful discussions about parts of this paper. We would also like to thank the associate editor and the reviewers for many constructive comments that improved the presentation of the paper.

Appendix A. Additional Notations

We write $\mathbb{P}(A)$ as the probability of event A. For a KNR, $\mathbb{P}(\cdot \mid W, s_h, a_h)$ denotes the probability distribution over $\mathcal{S}$ when the agent is in state s_h and takes action a_h, with the transition parametrization W. For two series $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, we write $a_n \lesssim b_n$ if $a_n \leq C \cdot b_n$ holds for constant C and all sufficient large n.

Appendix B. Proof of Lemma 10

In the dual update, the cost is related to the non-stationary variable θ. With the summation of the value difference bounded, we directly prove the following lemma by adding a $O(\sqrt{T})$ scale regret which comes from the employment of online gradient ascent.

Recall that $\widehat{\Psi} = 1/T \sum_{t=1}^{T} \Psi^t$. We present the following lemma, which can be seen as a corollary of Theorem 30.

Lemma 23 *Suppose that Assumptions 3 and 4 holds. For all* $\gamma \in [0, \Gamma]$, *we have*

$$T\big(f(\widehat{\Psi}) - f(\Psi^*) + \gamma \cdot g(\widehat{\Psi})\big) \lesssim B\Gamma\sqrt{HT},$$

which further implies

$$T(f(\widehat{\Psi}) - f(\Psi^*)) \lesssim B\Gamma\sqrt{HT}, \tag{27}$$

$$T \cdot g(\widehat{\Psi}) \lesssim \sqrt{HT}. \tag{28}$$

Proof We have the following relations holds for all $\gamma \in [0, \Gamma]$,

$$\begin{aligned} &f(\widehat{\Psi}) - f(\Psi^*) + \gamma \cdot g(\widehat{\Psi}) \\ &\quad = \max_{\alpha \in \mathcal{B}, \beta \in \gamma \cdot \mathcal{B}} \big\{\alpha^\top \widehat{\Psi} - f^*(\alpha) - f(\Psi^*) + \beta^\top \widehat{\Psi} - \gamma g^*(\beta/\gamma)\big\}. \end{aligned} \tag{29}$$

Thus, the dual update is equivalent to implementing online gradient ascent on h_t, where

$$h_t(\alpha, \beta, \gamma) = \alpha^\top \Psi^t - f^*(\alpha) + \beta^\top \Psi^t - \gamma g^*(\beta/\gamma).$$

By Theorem 30, we set the step size $\eta_t = 2\Gamma/H\sqrt{t}$ (or $2\Gamma/H\sqrt{T}$ when T is pre-decided), the constants $R = 2\Gamma$ and $G = 2B\sqrt{H}$ (to verify the conditions, note that g^* is $B\sqrt{H}$-Lipschitz, see Dubovitskii and Milyutin (1965)) to get

$$\begin{aligned} &T\Big[f(\widehat{\Psi}) - f(\Psi^*) + \gamma \cdot g(\widehat{\Psi})\Big] \\ &\quad \leq \sum_{t=1}^{T} \Psi^t \cdot (\alpha^t + \beta^t) - f^*(\alpha^t) - \gamma^t g^*(\beta^t/\gamma^t) - Tf(\Psi^*) + CB\Gamma\sqrt{HT}, \end{aligned} \tag{30}$$

where C is an absolute constant. With $\gamma^t \geq 0$ and $g(\Psi^*) \leq 0$, by the definition of Fenchel dual, we have

$$0 \geq \gamma^t g(\Psi^*) \geq \beta^t \cdot \Psi^* - \gamma^t g^*(\beta^t/\gamma^t), \quad f(\Psi^*) \geq \alpha^t \cdot \Psi^* - f(\alpha^t). \tag{31}$$

Recall that $\theta^t = \alpha^t + \beta^t$. Plugging (31) back to (30), we obtain the following relation holds for all $\gamma \in [0, \Gamma]$,

$$\begin{aligned} T\Big[f(\widehat{\Psi}) - f(\Psi^*) + \gamma \cdot g(\widehat{\Psi})\Big] &\le \sum_{t=1}^{T} \theta^t \cdot (\Psi^t - \Psi^*) + cB\Gamma\sqrt{HT} \\ &\le cB\Gamma\sqrt{HT}, \end{aligned}$$

where the second inequality comes from Lemma 36 and c is an absolute constant. With $\gamma = 0$ we obtain (27). With $\gamma = \Gamma$, we have

$$T\Big[f(\widehat{\Psi}) - f(\Psi^*) + \Gamma \cdot g(\widehat{\Psi})\Big] \lesssim \Gamma\sqrt{HT}.$$

And with Theorem 33 we obtain (28). Therefore, we conclude the proof. ■

Appendix C. Proof of Theorem 14

We first show that $W^t \in \mathcal{C}^t$ with high probability if R^t is properly chosen, which ensures that Algorithm 2 induces sufficient optimism. The following lemma is frequently used to provide a sufficient trustworthy radius for a confidence set and is first proved by Kakade et al. (2020). We provide its proof for completeness.

Lemma 24 (Confidence Ball) *For all $t \in [T]$, we set $\mathcal{E}_{cb}^t$ as the event that W^* falls in $\mathcal{C}^t$, i.e.,*

$$\mathcal{E}_{cb}^t = \Big\{ \| \left(\bar{W}^t - W^\star\right) \left(\Lambda^t\right)^{1/2} \|_2^2 \le R^t \Big\},$$

and $\mathcal{E}_{cb}$ as the event that all W^t falls in $\mathcal{C}^t$, and $\mathcal{E}_{cb} = \cap_{t=1}^{T} \mathcal{E}_{cb}^t$. Let

$$R^t = 2\lambda \|W^\star\|_2^2 + 8\sigma^2\big(d\log(5) + 2\log(t) + \log(4) + \log(\det(\Lambda^t)/\det(\Lambda^0)/\delta)\big),$$

We have

$$\sum_{t=0}^{\infty} \mathbb{P}(\bar{\mathcal{E}}_{cb}^t) = \sum_{t=0}^{\infty} \mathbb{P}\left(\left\| \left(\bar{W}^t - W^\star\right) \left(\Lambda^t\right)^{1/2} \right\|_2^2 > R^t \right) \le \delta/2.$$

Proof The center of the confidence ball, $\bar{W}^t$, is the minimizer of the ridge regression objective, and its closed-form expression is

$$\bar{W}^t = \sum_{\tau=1}^{t} \sum_{h=1}^{H} s_{h+1}^{\tau} \phi(s_h^\tau, a_h^\tau)^\top (\Lambda^t)^{-1},$$

where $\Lambda^t = \lambda I + \sum_{\tau=1}^{t}\sum_{h=1}^{H}\phi(s_h^\tau, a_h^\tau)^\top\phi(s_h^\tau, a_h^\tau)^\top$. Since $s_{h+1}^\tau = W^\star\phi(s_h^\tau, a_h^\tau) + \epsilon_h^\tau$ with $\epsilon_h^\tau \sim \mathcal{N}\left(0, \sigma^2\mathcal{I}\right)$, we have

$$\begin{aligned}
\bar{W}^t - W^\star &= \sum_{\tau=1}^{t}\sum_{h=1}^{H} s_{h+1}^\tau\phi(s_h^\tau, a_h^\tau)^\top(\Lambda^t)^{-1} - W^\star \\
&= \sum_{\tau=1}^{t}\sum_{h=1}^{H}(W^\star\phi(s_h^\tau, a_h^\tau) + \epsilon_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\left(\Lambda^t\right)^{-1} - W^\star \\
&= W^\star\bigg(\sum_{\tau=1}^{t}\sum_{h=1}^{H}\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\bigg)(\Lambda^t)^{-1} - W^\star + \sum_{\tau=1}^{t}\sum_{h=1}^{H}\epsilon_h^\tau\phi(s_h^\tau, a_h^\tau)^\top(\Lambda^t)^{-1} \\
&= -\lambda W^\star(\Lambda^t)^{-1} + \sum_{\tau=1}^{t}\sum_{h=1}^{H}\epsilon_h^\tau\phi(s_h^\tau, a_h^\tau)^\top(\Lambda^t)^{-1}.
\end{aligned}$$

For any $0 < \delta_t < 1$, using Lemma 38, it holds with probability at least $1 - \delta_t$,

$$\begin{aligned}
\left\|(\bar{W}^t - W^\star)(\Lambda^t)^{1/2}\right\|_2 &\leq \left\|\lambda W^\star(\Lambda^t)^{-1/2}\right\|_2 + \left\|\sum_{\tau=1}^{t}\sum_{h=1}^{H}\epsilon_h^\tau\phi(s_h^\tau, a_h^\tau)^\top(\Lambda^t)^{-1/2}\right\|_2 \\
&\leq \sqrt{\lambda}\|W^\star\|_2 + \sigma\sqrt{8d\log(5) + 8\log\left(\det(\Lambda^t)\det(\Lambda^0)^{-1}/\delta_t\right)},
\end{aligned}$$

where the first inequality follows from the triangle inequality. Therefore, we obtain $\mathbb{P}\big(\overline{\mathcal{E}}_{cb}^t\big) \leq \delta_t$. We seek to bound $\sum_{t=0}^{\infty}\mathbb{P}(\bar{\mathcal{E}}_{cb}^t)$. Note that at $t = 0$ we have initialized $\mathcal{C}^0$ to contain $W^\star$, we have $\mathbb{P}(\bar{\mathcal{E}}_{cb}^0) = 0$. For $t \geq 1$, let us assign failure probability $\delta_t = (3\delta/\pi^2)/t^2$ for the t-th event.We obtain

$$\sum_{t=1}^{\infty}\mathbb{P}(\bar{\mathcal{E}}_{cb}^t) \leq \sum_{t=1}^{\infty}(\delta/t^2)(3/\pi^2) = \delta/2.$$

Therefore, we conclude the proof of Lemma 24. ■

For all $t \in [T]$, we set $\mathcal{E}_{cb}^t$ as the event that W^* falls in $\mathcal{C}^t$, i.e.,

$$\mathcal{E}_{cb}^t = \left\{\|\left(\bar{W}^t - W^\star\right)\left(\Lambda^t\right)^{1/2}\|_2^2 \leq R^t\right\},$$

and $\mathcal{E}_{cb}$ as the event that all W^t falls in $\mathcal{C}^t$, i.e., $\mathcal{E}_{cb} = \cap_{t=1}^T\mathcal{E}_{cb}^t$. We prove in Lemma 24 that $\sum_{t=1}^{\infty}\mathbb{P}(\bar{\mathcal{E}}_{cb}^t) \leq \delta/2$, where $\bar{\mathcal{E}}_{cb}^t$ denotes the complement of $\mathcal{E}_{cb}^t$. The following lemma shows that by efficiently implementing the principle of optimism in Algorithm 2, the summation of the expected discrepancy of two projected kernel features is bounded. The main idea is to cast the projected kernel embedding to an initial state value function. Then by the value iteration implemented in Algorithm 2, we give a general bound for regret and violation in $O(\sqrt{T})$ scales.

Lemma 25 (Optimism for KNR) *Suppose that Assumption 3 holds. For Algorithms 1 and 2, the following inequality holds with probability at least* $1 - \delta$,

$$\mathbb{E}\bigg[\sum_{t=1}^{T}\theta^t\cdot(\Psi^t - \Psi^*)\bigg] \leq (1 + \Gamma)\sqrt{H}. \tag{32}$$

Proof If W^t falls in $\mathcal{C}^t$ for all t, it holds that $\sum_{t=1}^{T} \theta^t \cdot (\Psi^t - \Psi^*) \le 0$ by optimism induced by line 9 in Algorithm 2. We condition on the event $\mathcal{E}_{cb}^t$ and the proof is done. ■

The final step is to bound the estimation error of the visitation $f(\Psi^{\widehat{\pi}}) - f(\widehat{\Psi})$ and $g(\Psi^{\widehat{\pi}}) - g(\widehat{\Psi})$. With f and g being 1-Lipschitz, it suffices to bound $T\|\Psi^{\widehat{\pi}} - \widehat{\Psi}\|_2$.

Lemma 26 (Bound for estimation error) *Suppose that Assumptions 5-3 and 8 hold. For Algorithms 1 and 2, we have*

$$T\|\Psi^{\widehat{\pi}} - \widehat{\Psi}\|_2 \le CBHd\sqrt{T}\log\left(\frac{HT}{d\delta}\right) \tag{33}$$

holds with probability at least $1-\delta$*, here* C *is an absolute constant only depends on* σ*.*

Proof For all $x = (x_h)_{h\in[H]} \in \mathbb{R}^{dH}$ with $\|x\|_2 \le 1$, we can consider $\sum_{t=1}^{T}(\Psi^{\widehat{\pi}} - \widehat{\Psi})\cdot x$ as the value difference of two processes, with cost at stage h defined as $c_h(s_h, a_h) = \psi(s_h, a_h)\cdot x_h$. For simplicity, we denote $x\cdot(\Psi^{\pi_t} - \Psi^t) = V_1^{\pi_t} - V_1^t$. In the following analysis we condition on the event $\mathcal{E}_{cb}$, further estimate the value difference. With Lemma 15, we have

$$\sum_{t=1}^{T}\big(V_1^{\pi_t}(s) - V_1^t(s)\big) \lesssim \sum_{t=1}^{T} B\sqrt{H}\mathbb{E}_{\pi_t}\Bigg[\sum_{h=1}^{H}\big\|\big(W^\star - W^t\big)\phi(s_h^t, a_h^t)\big\|_2 \big|\mathcal{H}_t\Bigg], \tag{34}$$

Here $\{\mathcal{H}_t\}_{t\in[T]}$ is the history before episode t, and the inequality holds by Lemma 15. For $W^* \in \mathcal{C}^t$, we have

$$\begin{aligned}
\big\|\big(\widehat{W}^t - W^*\big)\phi(s_h^t, a_h^t)\big\|_2 &\le \big\|\big(\widehat{W}^t - W^*\big)(\Lambda^t)^{1/2}\big\|_2 \big\|(\Lambda^t)^{-1/2}\phi(s_h^t, a_h^t)\big\|_2 \\
&\le \bigg(\big\|\big(\widehat{W}^t - \bar{W}^t\big)\big(\Lambda^t\big)^{1/2}\big\|_2 + \big\|\big(\bar{W}^t - W^*\big)(\Lambda^t)^{1/2}\big\|_2\bigg)\big\|\phi(s_h^t, a_h^t)\big\|_{(\Lambda^t)^{-1}} \\
&\le 2\sqrt{R^t}\big\|\phi(s_h^t, a_h^t)\big\|_{(\Lambda^t)^{-1}}.
\end{aligned} \tag{35}$$

Summing up (35) over $h \in [H]$, we obtain

$$\sum_{h=1}^{H}\big\|\big(W^\star - W^t\big)\phi(s_h^t, a_h^t)\big\|_2 \le 2\sqrt{R^t}\sum_{h=1}^{H}\|\phi(s_h^t, a_h^t)\|_{(\Lambda^t)^{-1}}. \tag{36}$$

Plugging (36) back to (34), we have the following holds with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{t=1}^{T}\big(V_1^{\pi_t}(s) - V_1^t(s)\big) &\lesssim B\sqrt{H}\sum_{t=1}^{T}\mathbb{E}\bigg[\sqrt{R^T}\sum_{h=1}^{H}\|\phi(s_h^t, a_h^t)\|_{(\Lambda^t)^{-1}}\,\bigg|\,\mathcal{H}_t\bigg] \\
&\lesssim B\sqrt{H\sigma^2 d\log\left(\frac{HT}{d\delta}\right)}\sum_{t=1}^{T}\mathbb{E}\bigg[\sum_{h=1}^{H}\|\phi(s_h^t, a_h^t)\|_{(\Lambda^t)^{-1}}\,\bigg|\,\mathcal{H}_t\bigg] \\
&\lesssim B\sqrt{H\sigma^2 d\log\left(\frac{HT}{d\delta}\right)}\sum_{t=1}^{T}\sum_{h=1}^{H}\|\phi(s_h^t, a_h^t)\|_{(\Lambda^t)^{-1}} + \sum_{t=1}^{T}M_t^T, \\
&\le CB\sqrt{THd\log\left(\frac{HT}{d\delta}\right)}\bigg(\sum_{t=1}^{T}\sum_{h=1}^{H}\|\phi(s_h^t, a_h^t)\|_{(\Lambda^t)^{-1}}^2\bigg)^{1/2} + H\sqrt{2T\log(4/\delta)},
\end{aligned}$$

holds with probability at least $1-\delta$. Here C is some absolute constant that only depends on σ. The second inequality comes from the fact that R^t is non-decreasing, and

$$\begin{aligned} R^T &= \max\{2\sigma^2, 2\} + 8\sigma^2\Bigg(d\log(5) + 2\log(T) + \log(4) + \log\big(\det(\Lambda^T)\det(\Lambda^0)^{-1})/\delta\Bigg) \\ &\leq C'\sigma^2\Bigg(d + \log(T) + \log\big(\det(\Lambda^T)\det\big(\Lambda^0\big)^{-1}/\delta\big)\Bigg) \lesssim \sigma^2(d\log(TH/d\delta)) \end{aligned}$$

and $\lambda = \max\{\sigma^2, 1\}$. The third inequality decompose the expectation term into a elliptical potential summation and a martingale difference series. The last inequality comes from the martingale difference is bounded by H, and with Hoeffding's inequality, we have

$$\mathbb{P}(|\sum_{t=1}^{T} M_t^T| \geq s) \leq 2\exp(\frac{-s^2}{2TH^2}),$$

set $s = H\sqrt{2T\log(4/p)}$ and we prove that $\sum_{t=1}^{T} M_t^T \leq H\sqrt{T\log(4/p)}$ with probability at least $1-\delta/2$. Since we condition on $\mathcal{E}_{cb}$, which holds with probability at least $1-\delta/2$, the inequality holds with probability at least $1-\delta$. Next, we bound the elliptical potential term. By Lemma 42, we have

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \|\phi(s_h^t, a_h^t)\|_{(\Lambda^t)^{-1}}^2 \leq 2H\log\big(\det(\Lambda^T)\det(\Lambda^0)^{-1}\big) \lesssim dH\log(\frac{TH}{d}),$$

where the third inequality comes from 41. Combining the results above we have

$$\sum_{t=1}^{T}\big(V_1^{\pi_t}(s) - V_1^t(s)\big) \lesssim CBHd\sqrt{T}\log\left(\frac{HT}{d\delta}\right),$$

here C is an absolute constant that only relates to σ. Since the argument above holds for all x with $\|x\|_2 \leq 1$, set $x = (\Psi^{\pi_t} - \Psi^t)/\|\Psi^{\pi_t} - \Psi^t\|_2$, and we conclude the proof of Lemma 26. ■

Appendix D. Proof for Theorem 17

In this section we give a detailed proof for Theorem 17. The main tool is the MLE fundamental theorem and Bernstein's inequality for martingales.

Proof First, we prove that the choice of the confidence set $\mathcal{C}^t$ is fully efficient, i.e. $\mathcal{P}^* \in \mathcal{C}^t$ with high probability.

Lemma 27 *With probability at least $1-\delta$, we have the true underlying transition kernel $\mathcal{P}_h^* : \mathcal{S}\times\mathcal{S}\times\mathcal{A} \to \mathbb{R}$ lies in the confidence set $\mathcal{C}^t$ for all $t \in [T]$ and $h \in [H]$, i.e.,*

$$\mathbb{E}_{\mathcal{D}_h^t}[\|\mathcal{P}_h^*(\cdot|s_h, a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h, a_h)\|_1^2] \leq c\frac{\log(TH|\Theta||\Upsilon|/\delta)}{t},$$

where $\mathbb{E}_{\mathcal{D}}[f(s,a)]$ takes the average of f on the dataset $\mathcal{D}$.

Proof By the construction of $\mathcal{D}_h^t$ in Algorithm 3, we have $s_h^t \sim \pi_t$ and $a_h^t \sim U(\mathcal{A})$. Recall that

$$\widehat{\rho}_h^t(s_h, a_h) = \frac{1}{t-1} \sum_{i \in [t-1]} d_{\pi^i,h}(s_h) u(a_h),$$

Therefore, $\mathbb{E}_{\mathcal{D}_h^t}[\|\mathcal{P}_h^*(\cdot|s_h,a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h,a_h)\|_1^2]$ is a empirical realization of the visitation measure for (s_h, a_h) under the Markov policy $\widehat{\rho}^t$. For notation simplicity, for each $(h,t) \in [H] \times [T]$, define $\mathcal{F}_{t,h}$ to be to the σ-algebra generated by the trajectories,

$$\mathcal{F}_{t,h} = \sigma(\{(s_i^\tau, a_i^\tau)\}_{(i,\tau)\in[H]\times[t-1]} \cup \{(s_i^t, a_i^t)\}_{i\in[h-1]}),$$

since $\pi^\tau = \pi^i(s_1^1, a_1^1, ..., s_H^{\tau-1})$ and is measurable with respect to $\mathcal{F}_{t,h}$, we have

$$\begin{aligned} t(\mathbb{E}_{\mathcal{D}_h^t}&[\|\mathcal{P}_h^*(\cdot|s_h,a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h,a_h)\|_1^2] - \mathbb{E}_{\widehat{\rho}_h^t}[\|\mathcal{P}_h^*(\cdot|s_h,a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h,a_h)\|_1^2]) \\ &= \sum_{\tau\in[t]} \|\mathcal{P}_h^*(\cdot|s_h^\tau,a_h^\tau) - \widehat{\mathcal{P}}_h^t(\cdot|s_h^\tau,a_h^\tau)\|_1^2 - \mathbb{E}_{s_h\sim\pi^\tau, a_h\sim U(\mathcal{A})}\|\mathcal{P}_h^*(\cdot|s_h,a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h,a_h)\|_1^2 \end{aligned}$$

being a martingale process with respect to the filtration $\{\mathcal{F}_{t,h}\}_{(h,t)\in[H]\times[T]}$ for all $(h,t) \in [H] \times [T]$. Therefore, by applying the Bernstein type inequality for martingale (Lemma 43), with probability at least $1-\delta$ that

$$\begin{aligned} &\big|\mathbb{E}_{\mathcal{D}_h^t}[\|\mathcal{P}_h^*(\cdot|s_h,a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h,a_h)\|_1^2] - \mathbb{E}_{\widehat{\rho}_h^t}[\|\mathcal{P}_h^*(\cdot|s_h,a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h,a_h)\|_1^2]\big| \\ &\quad \le \sqrt{\frac{2\operatorname{Var}_{\widehat{\rho}_h^t}\big[\|\mathcal{P}_h^*(\cdot|s_h,a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h,a_h)\|_1^2\big]\log(2TH/\delta)}{t}} + \frac{\log(2TH/\delta)}{3t} \\ &\quad \le \sqrt{\frac{8\mathbb{E}_{\widehat{\rho}_h^t}\big[\|\mathcal{P}_h^*(\cdot|s_h,a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h,a_h)\|_1^2\big]\log(2TH/\delta)}{t}} + \frac{\log(2TH/\delta)}{3t}, \end{aligned}$$

where the second inequality follows from $\|\mathcal{P}_h^*(\cdot|s_h,a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h,a_h)\|_1^2 \le 4$. Recall that with Lemma 39, we have

$$\mathbb{E}_{\widehat{\rho}_h^t}\big[\|\mathcal{P}_h^*(\cdot|s_h,a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h,a_h)\|_1^2 \le c\frac{\log(TH|\Theta||\Upsilon|/\delta)}{t}.$$

Therefore, we have

$$\big|\mathbb{E}_{\mathcal{D}_h^t}[\|\mathcal{P}_h^*(\cdot|s_h,a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h,a_h)\|_1^2] - \mathbb{E}_{\widehat{\rho}_h^t}[\|\mathcal{P}_h^*(\cdot|s_h,a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h,a_h)\|_1^2]\big| \le c\frac{\log(TH|\Theta||\Upsilon|/\delta)}{t}.$$

Summing this and D and we conclude the proof.

■

The next lemma ensures that as we explore and shrink the radius of the confidence set, the statistical distances between the MLE estimation and all transitions in the confidence set uniformly decrease in a $\widetilde{O}(1/t)$ manner.

Lemma 28 *For all $\mathcal{P} = \{\mathcal{P}_h\}_{h\in[H]} \in \mathcal{C}^t$ and all $t \in [T]$, with probability at least $1-\delta$, we have*

$$\mathbb{E}_{\widehat{\rho}^t}[\|\mathcal{P}_h(\cdot|s_h,a_h) - \mathcal{P}_h^*(\cdot|s_h,a_h)\|_1^2] \le c'\frac{\log(TH|\Theta||\Upsilon|/\delta)}{t}.$$

Proof By the construction of $\mathcal{C}^t$, we have the following inequality holds for all $\mathcal{P} = \{\mathcal{P}_h\}_{h\in[H]}$ in $\mathcal{C}^t$ with high probability,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_h^t}[\|\mathcal{P}_h(\cdot|s_h,a_h) - \mathcal{P}_h^*(\cdot|s_h,a_h)\|_1^2] \\
&\quad \le 2(\mathbb{E}_{\mathcal{D}_h^t}[\|\mathcal{P}_h(\cdot|s_h,a_h) - \mathcal{P}_h^t(\cdot|s_h,a_h)\|_1^2] + \mathbb{E}_{\mathcal{D}_h^t}[\|\mathcal{P}_h^t(\cdot|s_h,a_h) - \mathcal{P}_h^*(\cdot|s_h,a_h)\|_1^2]) \\
&\quad \le 2c\frac{\log(TH|\Theta||\Upsilon|/\delta)}{t},
\end{aligned}$$

The first inequality comes from $(a+b)^2 \le 2(a^2+b^2)$. Define

$$A(\mathcal{P}_h) = \mathbb{E}_{\widehat{\rho}_h^t}[\|\mathcal{P}_h(\cdot|s_h,a_h) - \mathcal{P}_h^*(\cdot|s_h,a_h)\|_1^2] - \mathbb{E}_{\mathcal{D}_h^t}[\|\mathcal{P}_h(\cdot|s_h,a_h) - \mathcal{P}_h^*(\cdot|s_h,a_h)\|_1^2].$$

We have

$$\mathbb{E}_{\widehat{\rho}_h^t}[\|\mathcal{P}_h(\cdot|s_h,a_h) - \mathcal{P}_h^*(\cdot|s_h,a_h)\|_1^2] \le A(\mathcal{P}_h) + 2c\frac{\log(TH|\Theta||\Upsilon|/\delta)}{t},$$

for all $(h,t) \in [H]\times[T]$. Applying Bernstein's inequality again, for any $\{\mathcal{P}_h\}_{h\in[H]} \in \mathcal{C}^t$ and $t \in [T]$, we have with probability $1-\delta$ that

$$\begin{aligned}
A(\mathcal{P}_h) &\le \sqrt{\frac{c_1 \operatorname{Var}_{\widehat{\rho}^t}\left[\|\mathcal{P}_h^*(\cdot|s_h,a_h) - \widehat{\mathcal{P}}_h^t(\cdot|s_h,a_h)\|_1^2\right]\log(TH|\Theta||\Upsilon|/\delta)}{t}} + \frac{c_2\log(TH|\Theta||\Upsilon|/\delta)}{t} \\
&\le \sqrt{\frac{c_1 \mathrm{E}_{\widehat{\rho}_h^t}\left[\|\mathcal{P}_h(\cdot|s_h,a_h) - \mathcal{P}_h^*(\cdot|s_h,a_h)\|_1^4\right]\log(TH|\Theta||\Upsilon|/\delta)}{t}} + \frac{c_2\log(TH|\Theta||\Upsilon|/\delta)}{t} \\
&\le \sqrt{\frac{4c_1 \mathrm{E}_{\widehat{\rho}_h^t}\left[\|\mathcal{P}_h(\cdot|s_h,a_h) - \mathcal{P}_h^*(\cdot|s_h,a_h)\|_1^2\right]\log(H|\Theta||\Upsilon|/\delta)}{t}} + \frac{c_2\log(H|\Theta||\Upsilon|/\delta)}{t},
\end{aligned}$$

where the first inequality comes from Bernstein's inequality, and the second inequality comes from the fact that $\|\mathcal{P}(\cdot|s,a) - \mathcal{P}'(\cdot|s,a)\|_1^2 \le 4$ for two probability distributions. Denote $\xi = \log(|\Theta||\Upsilon|TH/\delta)/t$ and taking square in both side of the (D), we have

$$\begin{aligned}
A^2(\mathcal{P}_h) &\lesssim \left(\sqrt{\frac{c(A(\mathcal{P}_h)+\xi)\log(TH|\Theta||\Upsilon|/\delta)}{t}} + \frac{c\log(|\Theta||\Upsilon|/\delta)}{t}\right)^2 \\
&\lesssim \frac{(A(\mathcal{P}_h)+\xi)\log(TH|\Theta||\Upsilon|/\delta)}{t} + \left\{\frac{c\ln(TH|\Theta||\Upsilon|/\delta)}{t}\right\}^2 \\
&\lesssim \frac{(A(\mathcal{P}_h)+\xi)\log(TH|\Theta||\Upsilon|/\delta)}{t} + \frac{c_2\log(TH|\Theta||\Upsilon|/\delta)}{t}. \\
&\lesssim \frac{(A(\mathcal{P}_h)+1/t\log(TH|\Theta||\Upsilon|/\delta))\log(TH|\Theta||\Upsilon|/\delta)}{t}.
\end{aligned}$$

Then, we have

$$A^2(\mathcal{P}_h) - B_1A(\mathcal{P}_h) - B_2 \le 0, \quad B_1 = c\log(TH|\Theta||\Upsilon|/\delta)/t, \quad B_2 = c(1/t)^2\log(TH|\Theta||\Upsilon|/\delta)^2$$

holds for all $\{\mathcal{P}_h\}_{h\in[H]} \in \mathcal{C}^t$ for all $t \in [T]$. This concludes

$$0 \le A(\mathcal{P}_h) \le \frac{B_1 + \sqrt{B_1^2 + 4B_2}}{2} \le c\left(B_1 + \sqrt{B_2}\right) \le c\frac{\log(TH|\Theta||\Upsilon|/\delta)}{t} \lesssim \xi.$$

Thus, by using the above $A(\mathcal{P}) \lesssim \xi \left(\mathcal{P} \in \mathcal{C}^t\right)$ and , with probability $1-\delta$, we have

$$\mathbb{E}_{\widehat{\rho}_h^t}\left[\|\mathcal{P}_h(\cdot|s_h,a_h) - \mathcal{P}_h^*(\cdot|s_h,a_h)\|_1^2\right] \le A(P) + c\xi \lesssim \xi, \quad \{\mathcal{P}_h\}_{h\in[H]} \in \mathcal{C}^t.$$

■

To conclude our proof, the final step is to bound $T\|\widehat{\Psi} - \Psi^{\widehat{\pi}}\|_2 = \|\sum_{t=1}^T (\Psi^t - \Psi^{\pi^t})\|_2$. To this end, we still consider to find an uniform upper bound for $\sum_{t=1}^T (\Psi^t - \Psi^{\pi^t}) \cdot \theta$, $\theta = (\theta_h)_{h\in[H]}$ with $\|\theta\|_2 \le 1$. As in the case of KNR, we define

$$\begin{aligned}
&c_h^t(s,a) = \psi_h(s,a)\cdot\theta_h,\\
&\Psi^t \cdot \theta = \mathbb{E}_{\pi^t,\mathcal{P}^t}\left[\sum_{h=1}^H c_h^t(s_h,a_h)\right] = V_1^t,\\
&\Psi^{\pi^t} \cdot \theta = \mathbb{E}_{\pi^t,\mathcal{P}^*}\left[\sum_{h=1}^H r_h^t(s_h,a_h)\right] = V_1^{\pi^t},
\end{aligned}$$

With standard notations in reinforcement learning, we can define value function $V_h(s,a)$ for all stage $h \in [H]$. Using the value-decomposition lemma, we decompose the value difference $V_1^t - V_1^{\pi^t}$,

$$\begin{aligned}
\sum_{t=1}^T (\Psi^t - \Psi^{\pi^t})\cdot\theta &= \sum_{t=1}^T V_1^t - V_1^{\pi^t}\\
&= \sum_{h=1}^H \sum_{t=1}^T \mathbb{E}_{\pi^t}\left[\mathbb{P}_h^* V_{h+1}^{\pi^t}(s_h,a_h) - \mathbb{P}_h^t V_{h+1}^{\pi^t}(s_h,a_h)\right] \qquad (37)\\
&\le \sum_{h=1}^H \sum_{t=1}^T B\sqrt{H}\mathbb{E}_{\pi^t}\left[\|\mathcal{P}_h^t(\cdot|s_h,a_h) - \mathcal{P}_h^*(\cdot|s_h,a_h)\|_1\right],
\end{aligned}$$

here the second equation comes from the value difference lemma, and the third inequality comes from the fact that $\|V_1^\pi\|_\infty \le B\sqrt{H}$, since $\|\theta\|_2 \le 1$ and $\|\psi_h(s,a)\|_2 \le B$. The next lemma shows that we can upper bound $\mathbb{E}_{\pi^t}[H(s_h,a_h)]$ using $\mathbb{E}_{\pi^t}\|\phi_h^*(s_h,a_h)\|_{\Sigma^{-1}_{\rho^t,\phi^*_{h-1}}}$ once we can upper bound $\mathbb{E}_{\widehat{\rho}^t}[H^2(s_h,a_h)]$.

Lemma 29 (One step back inequality) *Take any $H \in \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ such that $\|H\|_\infty \le B$. Then,*

$$\mathbb{E}_\pi[H(s_h,a_h)] \le \mathbb{E}_\pi\|\phi^\star_{h-1}(s_{h-1},a_{h-1})\|_{\Sigma^{-1}_{\rho^t_{h-1},\phi^\star_{h-1}}} \sqrt{t|\mathcal{A}|\mathbb{E}_{\widehat{\rho}_h^t}[H^2(s,a)] + \lambda d B^2},$$

where $\Sigma_{\rho_h^t,\phi_h^\star} = t\mathbb{E}_{\rho_h^t}\left[\phi_h^\star(s_h,a_h)\phi_h^\star(s_h,a_h)^\top\right] + \lambda I$. Note that here the parameter λ and the matrix Σ_{ρ^t,ϕ_h^*} doesn't occur in the actual implementation.

Proof First, we have

$$\begin{aligned}\mathbb{E}_\pi[H(s_h,a_h)] &= \mathbb{E}_{(s_{h-1},a_{h-1})\sim\pi, s_h\sim\mathcal{P}(s_{h-1},a_{h-1}), a_h\sim\pi}[H(s_h,a_h)]\\ &= \mathbb{E}_\pi\left[\phi^*_{h-1}(s_{h-1},a_{h-1})^\top \int\sum_{a_h}\mu^*_{h-1}(s_h)\pi_h(a_h|s_h)H(s_h,a_h)d\nu\right]\\ &\le \mathbb{E}_\pi\left[\|\phi^*_{h-1}(s_{h-1},a_{h-1})\|_{\Sigma^{-1}_{\rho^t_h,\phi^\star_{h-1}}}\left\|\int\sum_{a_h}\mu^*_{h-1}(s_h)\pi_h(a_h|s_h)H(s_h,a_h)d\nu\right\|_{\Sigma_{\rho^t_h,\phi^\star_{h-1}}}\right],\end{aligned}$$

where the third inequality comes from Cauchy's inequality. Here, we have

$$\begin{aligned}&\left\|\int\sum_{a_h}\mu^\star_{h-1}(s)\pi_h(a_h|s_h)H(s_h,a_h)d\nu(s)\right\|^2_{\Sigma_{\rho^t_{h-1},\phi^\star_{h-1}}}\\ &\le \left\{\int\sum_{a_h}\mu^\star_{h-1}(s_h)\pi_h(a_h|s_h)H(s_h,a_h)d\nu(s)\right\}^\top\left\{t\mathbb{E}_{\rho^t_{h-1}}\left[\phi^\star_{h-1}(s_{h-1},a_{h-1})\left\{\phi^\star_{h-1}(s_h,a_h)\right\}^\top\right]+\lambda I\right\}\left\{\int\sum_{a}\mu^\star_h\right.\\ &\le t\mathbb{E}_{\rho^t_{h-1}}\left\{\left[\int\sum_a\mu^\star_{h-1}(s_h)^\top\phi^\star_{h-1}(s_{h-1},a_{h-1})\pi_h(a_h\mid s_h)H(s_h,a_h)d\nu(s)\right]^2\right\}+\lambda dB^2\\ &\le t\left\{\mathbb{E}_{(s_{h-1},a_{h-1})\sim\rho^t_{h-1},s_h\sim P^\star(s_{h-1},a_{h-1}),a_h\sim\pi(s)}\left[H^2(s_h,a_h)\right]\right\}+\lambda dB^2,\end{aligned}$$

where the last inequality comes from Jensen's inequality. Further, we have that

$$\begin{aligned}\mathbb{E}_{(s_{h-1},a_{h-1})\sim\rho^t,s_h\sim P^\star(s_{h-1},a_{h-1}),a_h\sim\pi(s)}\left[H^2(s_h,a_h)\right] &\le |\mathcal{A}|\mathbb{E}_{(s_{h-1},a_{h-1})\sim\rho^t,s_h\sim P^\star(s_{h-1},a_{h-1}),a_h\sim U(\mathcal{A})}\left[H^2(s_h,a_h)\right]\\ &= |\mathcal{A}|\mathbb{E}_{\widehat{\rho}^t_h}[H^2(s_h,a_h)]\end{aligned}$$

which concludes the proof. ■

We then condition on the event

$$\mathbb{E}_{\widehat{\rho}^t_h}[\|\mathcal{P}^t_h(\cdot|s_h,a_h)-\mathcal{P}^*_h(\cdot|s_h,a_h)\|^2_1]\le c\frac{\log(TH|\Upsilon||\Theta|)}{t}, \forall(h,t)\in[H]\times[T],$$

which holds with probability at least $1-\delta$, and use Lemma 29 on (37) by setting $\pi=\pi^t$ for $\|\mathcal{P}^t_h(\cdot|s_h,a_h)-\mathcal{P}^*_h(\cdot|s_h,a_h)\|_1$, we have

$$\begin{aligned}&\sum_{h=1}^H\sum_{t=1}^T\mathbb{E}_{\pi^t}\left[\|\mathcal{P}^t_h(\cdot|s_h,a_h)-\mathcal{P}^*_h(\cdot|s_h,a_h)\|_1\right]\\ &\qquad\le\sum_{h=1}^H\sum_{t=1}^T\mathbb{E}_{\pi^t}\|\phi^*_h(s_h,a_h)\|_{\Sigma^{-1}_{\rho^t_h,\phi^*_h}}\sqrt{t|\mathcal{A}|\mathbb{E}_{\widehat{\rho}^t_h}[\|\mathcal{P}^t_h-\mathcal{P}^*_h\|^2_1]+4\lambda d}\\ &\qquad\lesssim\sum_{h=1}^H\sum_{t=1}^T\mathbb{E}_{\pi^t}\|\phi^*_h(s_h,a_h)\|_{\Sigma^{-1}_{\rho^t_h,\phi^*_h}}\sqrt{|\mathcal{A}|\log(TH|\Theta||\Upsilon|/\delta)+\lambda d},\end{aligned}$$

here the first inequality comes from the one-step back inequality and the fact that every term in the summation is positive, the second inequality comes from our condition event. We also have

$$\sqrt{|\mathcal{A}|\log(TH|\Theta||\Upsilon|/\delta)+\lambda d} \lesssim \sqrt{|\mathcal{A}|\lambda d\log(TH|\Theta||\Upsilon|/\delta)} = \xi_T,$$

therefore

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{t=1}^{T}\mathbb{E}_{\pi^t}\left[\|\mathcal{P}_h^t(\cdot|s_h,a_h)-\mathcal{P}_h^*(\cdot|s_h,a_h)\|_1\right] \\
&\quad\lesssim \sum_{h=1}^{H}\xi_T\big(\sum_{t=1}^{T}\mathbb{E}_{\pi^t}\|\phi_h^*(s_h,a_h)\|_{\Sigma^{-1}_{\rho_h^t,\phi_h^*}}\big) \\
&\quad\le \sum_{h=1}^{H}\xi_T\cdot\sqrt{T\sum_{t=1}^{T}\mathbb{E}_{\pi^t}\left[\phi_h^*(s_h,a_h)^\top\Sigma^{-1}_{\rho_h^t,\phi_h^*}\phi_h^*(s_h,a_h)\right]},
\end{aligned}$$

where the second inequality comes from Jensen's inequality. By Lemma 40 and Lemma 41, we have

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}_{\pi^t}\left[\phi_h^*(s_h,a_h)^\top\Sigma^{-1}_{\rho_h^t,\phi_h^*}\phi_h^*(s_h,a_h)\right] &= \sum_{t=1}^{T}\mathrm{Tr}\left(\Sigma^{-1}_{\rho_h^t,\phi_h^*}\cdot\mathbb{E}_{\pi_t}\left[\phi_h^*(s_h,a_h)\phi_h^*(s_h,a_h)^\top\right]\right) \\
&\le 2\Big(\log\det\left(\Sigma_{\rho_h^T,\phi_h^*}\right)-2\log\det\left(\lambda I\right)\Big) \\
&\le d\log\left(1+\frac{T}{d\lambda}\right)
\end{aligned}$$

holds for all $h\in[H]$. By then we have

$$\sum_{h=1}^{H}\sum_{t=1}^{T}\mathbb{E}_{\pi^t}\left[\|\mathcal{P}_h^t(\cdot|s_h,a_h)-\mathcal{P}_h^*(\cdot|s_h,a_h)\|_1\right] \le \sqrt{T\log\left(1+\frac{T}{d\lambda}\right)}|\mathcal{A}|^{1/2}d\lambda^{1/2}\log(TH|\Theta||\Upsilon|/\delta),$$

combine with (37) and set $\lambda=1$, $\theta=(\widehat{\Psi}-\Psi^{\widehat{\pi}})/\|\widehat{\Psi}-\Psi^{\widehat{\pi}}\|_2$, we conclude the proof of Lemma 20

Combine Lemma 12 and 20 we finish the proof of Theorem 17. ■

Appendix E. Lemmas for Optimization

E.1 Online learning

Online learning involves two players: the adversary and the player. The online learning protocol is shown in Algorithm 4.

Algorithm 4 Protocol of Online Learning

```
for $t = 1, \dots, T$ do
    The player chooses an action $x_t$.
    The adversary picks a function $f_t$.
    The player obtains reward $f_t(x_t)$.
    The player learns via $f_t$.
end for
```

Note that there is no assumption on how the adversary will pick the function f_t, and it may be adversarially chosen. The player aims to minimize the regret:

$$\text{Regret} = \max_x \sum_{t=1}^T f_t(x) - \sum_{t=1}^T f_t(x_t), \tag{38}$$

which measures the quality of the player's strategy $x_1, \dots, x_T$ compared with the single best decision in hindsight.

Projected Subgradient Method. The projected subgradient method is a particular case of mirror descent/ascent with Euclidean distance. Applying this method to online learning produces a regret bound of the order $O(\sqrt{T})$.

Suppose that the actions x_t are required to be contained in some convex set $\mathcal{X}$, i.e., $x_t \in \mathcal{X}$. Denote a subgradient of f_t at x_t by $g_t \in \partial f_t(x_t)$, G and R are two constants such that $\max_{x,y\in\mathcal{X}} \|x - y\|_2 \leq R$ and $\max_{t\in[T]} \|\partial f_t(x_t)\|_2 \leq G$. We set the step length η_t at the t-th iteration to $R/G\sqrt{t}$ if we do not know the number of iterations T in advance and to $R/G\sqrt{T}$ if we have the knowledge of T. The latter case will leads to an upper bound with a smaller constant multiplicative factor. With these notations, the update rule of projected subgradient method can be expressed as

$$x_{t+1} \leftarrow \arg\max_{x\in\mathcal{X}} \big\{ f_t(x_t) + \langle \eta_t g_t, x - x_t \rangle - \|x - x_t\|_2^2/2 \big\}.$$

We describe the complete method in Algorithm 5.

Algorithm 5 projected subgradient method

```
Arbitrarily initialize $x_1 \in \mathcal{X}$.
for $t = 1, \dots, T-1$ do
    Update $x_{t+1} \leftarrow \arg\max_{x\in\mathcal{X}} \big\{ f_t(x_t) + \langle \eta_t g_t, x - x_t \rangle - \|x - x_t\|_2^2/2 \big\}$
end for
```

By this method, the regret is guaranteed to increase sublinearly as stated in the following theorem.

Theorem 30 *Using projected subgradient method mentioned in Algorithm 5, it holds that for all x in the convex set $\mathcal{X}$ we have*

$$\sum_{t=1}^T f_t(x) - \sum_{t=1}^T f_t(x_t) \leq CRG\sqrt{T},$$

where C is an absolute constant.

Proof See Zinkevich (2003) for a detailed proof. Note that the choice of x is irrelevant in the proof. ■

E.2 Constrained Optimization

In this subsection we consider a general constrained optimization and discuss its properties. We consider

$$f_{\text{opt}} = \min_{x \in X} \{f(x) : g(x) \leq 0, Ax + b = 0\}, \tag{39}$$

where and $f, g : \mathbb{R} \to (-\infty, \infty)$ are convex real-valued functions, $A \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p$. We define a value function associated with (39),

$$v(u, t) = \min_{x \in X} \{f(x) : g(x) \leq u, Ax + b = t\}.$$

Furthermore, we define the dual problem to (39). The dual function is

$$q(\lambda, \gamma) = \min_{x \in X} \left\{ L(x, \lambda, \gamma) = f(x) + \lambda^T g(x) + \gamma^T (Ax + b) \right\},$$

where $\lambda \in \mathbb{R}^m_+, \gamma \in \mathbb{R}^p$. The corresponding dual problem is

$$q_{\text{opt}} = \max_{\lambda \in \mathbb{R}^m_+, \gamma \in \mathbb{R}^p} \{q(\lambda, \gamma) : (\lambda, \gamma) \in \text{dom}(-q)\}.$$

Where $\text{dom}(-q) = \left\{ (\lambda, \gamma) \in \mathbb{R}^m_+, \gamma \in \mathbb{R}^p : q(\lambda, \gamma) > -\infty \right\}$. Furthermore, we denote an optimal solution of (E.2) by λ^*, γ^*.

We make the following assumption which will be verified to hold. The assumption implies strong duality, i.e., $q_{\text{opt}} = f_{\text{opt}}$.

Assumption 31 *The optimal value of* (39) *is finite and exists a Slater point* $\overline{\mathrm{x}}$ *such that* $g(\overline{x} < 0$ *and exists a point* $\widehat{x} \in \text{ri}(X)$ *satisfying* $A\widehat{x} + b = 0$*, where* $\text{ri}(X)$ *is the relative interior of* X*.*

The following theorem is proved in Beck (2017).

Theorem 32 *The dual variable* (λ^*, γ^*) *is an optimal solution of* (39) *if and only if*

$$-(\lambda^*, \gamma^*) \in \partial v(0, 0),$$

where $\partial f(x)$ *denotes the set of all sub-gradients of* f *at* $\mathbf{x}$*.*

Proof See Theorem 3.59, Beck (2017). ■

Using this result we arrive at the following theorem, which is a variant of Beck (2017) , Theorem 3.60.

Theorem 33 *Let λ^* be an optimal solution of the dual* (39) *and assume that $2\|\lambda^*\|_1 \leq \rho$. Let $\widetilde{x}$ satisfy $A\widetilde{x} + b = 0$ and*

$$f(\widetilde{x}) - f_{opt} \ + \rho\|[g(\widetilde{x})]_+\|_\infty \leq \delta$$

Then we have

$$\|[g(\widetilde{x})]_+\|_\infty \leq \frac{\delta}{\rho}.$$

Proof Let

$$v(u,t) = \min_{x\in X}\{f(x) : g(x) \leq u, Ax + b = t\}.$$

Since $(-\lambda^*, \gamma^*)$ is an optimal solution of the dual problem it follows by Theorem 32 that $(-\lambda^*, \gamma^*) \in \partial v(0,0)$. Therefore, for any $(u,0) \in \mathrm{dom}(v)$,

$$v(u,0) - v(0,0) \geq \langle -\lambda^*, u\rangle . \tag{40}$$

Set $u = \widetilde{u} = [g(\widetilde{x})]_+$. Since $\overline{u} \geq 0$, we have

$$v(\widetilde{u}, 0) \leq v(0,0) = f_{\mathrm{opt}} \leq f(\widetilde{x}).$$

Thus, (40) implies that

$$f(\widetilde{x}) - f_{\mathrm{opt}} \geq \langle -\lambda^*, \widetilde{u}\rangle. \tag{41}$$

Thus, we obtain

$$\begin{aligned}(\rho - \|\lambda^*\|_1)\|\widetilde{u}\|_\infty &= -\|\lambda^*\|_1\|\widetilde{u}\|_\infty + \rho\|\widetilde{u}\|_\infty \\ &\leq \langle -\lambda^*, \widetilde{u}\rangle + \rho\|\widetilde{u}\|_\infty \\ &= f(\widetilde{x}) - f_{\mathrm{opt}} + \rho\|\overline{u}\|_\infty \leq \delta,\end{aligned}$$

where the last relation follows from (41). Rearranging the terms and using the assumption $2\|\lambda^*\|_1 \leq \rho$, we obtain

$$\|[g(\widetilde{x})]_+\|_\infty = \|\overline{u}\|_\infty \leq \frac{\delta}{\rho - \|\lambda^*\|_1} \leq \frac{2}{\rho}\delta.$$

Therefore, we conclude the proof of Theorem 32. ■

For the solution of the dual function, the following lemma is an adjustment of Beck (2017).

Theorem 34 *Let $\overline{x} \in X$ be a point satisfying $g(\bar{x}) < 0$ and $A\overline{x} + b = 0$. Then, for any $\lambda, \gamma \in \{\lambda \in \mathbb{R}_+^m, \gamma \in \mathbb{R}_+^p : q(\lambda, \gamma) \geq M\}$, we have*

$$\|\lambda\|_1 \leq \frac{f(\bar{x}) - M}{\min_{j\in[m]}\{-g_j(\bar{x})\}}.$$

Proof Let

$$S_M = \{\lambda \in \mathbb{R}_+^m, \gamma \in \mathbb{R}_+^p : q(\lambda, \gamma) \geq M\}.$$

By the definition of S_M, for any $\lambda, \gamma \in S_M$ we have

$$\begin{aligned} M &\leq q(\lambda, \gamma) \\ &= \min_{x \in X} \{f(x) + \lambda^T g(x) + \gamma^T (Ax + b)\} \\ &\leq f(\overline{x}) + \lambda^T g(\overline{x}) + \gamma^T (A\overline{x} + b) \\ &= f(\overline{x}) + \sum_{j=1}^m \lambda_j g_j(\overline{x}). \end{aligned}$$

Therefore, we obtain

$$-\sum_{j=1}^m \lambda_j g_j(\overline{x}) \leq f(\overline{x}) - M,$$

which implies that for any $(\lambda, \gamma) \in S_M$,

$$\sum_{j=1}^m \lambda_j = \|\lambda\|_1 \leq \frac{f(\overline{x}) - M}{\min_{j \in [m]}\{-g_j(\overline{x})\}}.$$

Therefore, we conclude the proof of Theorem 34. ■

A simple corollary gives an estimation of the optimal dual solution λ^*.

Corollary 35 *Let $\overline{x} \in X$ be a point satisfying $g(\bar{x}) < 0$ and $A\overline{x} + b = 0$, and λ^* be an optimal dual solution. Then, it holds that*

$$\|\lambda^*\|_1 \leq \frac{f(\bar{x}) - M}{\min_{j \in [m]}\{-g_j(\bar{x})\}}.$$

Proof Since $(\lambda^*, \gamma^*) \in S_{f_{\text{opt}}}$ be an optimal solution of the dual problem equation 39, we finish the proof by Theorem 34. ■

E.3 Proof of Proposition 4

Proof To prove the convexity of (4), it suffices to show that $\mathcal{V}$ is convex. We allow some initial randomizing mechanisms such that the policy $\{\pi_h\}_{h \in [H]}$ not only rely on h, but also depends on a randomizing mechanism. We may have a set of policies $\mathcal{U}$ and a distribution $q \in \Delta(\mathcal{U})$. Then the mixed policy $\widehat{\pi}$ of $\mathcal{U}$, is defined such that we choose some policy $\pi \in \mathcal{U}$ using q and then the agent proceeds executing with only that policy (Altman, 1999). We have the following equality,

$$\Psi^{\widehat{\pi}} = \mathbb{E}_q[\Psi^{\pi}],$$

where the expectation is taken with respect to the underlying distribution q and all policy $\pi \in \mathcal{U}$. When q is set as the uniform distribution on set $\{\pi_k\}_{k \in [K]}$, we have

$$\Psi^{\widehat{\pi}} = \frac{1}{K} \sum_{k=1}^K \Psi^{\pi_k}.$$

Since $\Psi^{\widehat{\pi}} \in \mathcal{V}$ with our definition, $\mathcal{V}$ is a convex set. The optima of (4) over the mixed policy will remain the same, and $\mathcal{V}$ is proved to be a convex set. The feasible set for (4) is thus convex and the problem is indeed a convex optimization. ■

Appendix F. Auxiliary Results

The difference of value functions between two MDPs has the following general decomposition, which is rather useful in our analysis.

Lemma 36 (Value Difference Lemma) *Consider two MDPs* $\big(\mathcal{S}, \mathcal{A}, \{\mathcal{P}_h^1\}_{h=1}^H, \{r_h^1\}_{h=1}^H\big)$ *and* $\big(\mathcal{S}, \mathcal{A}, \{\mathcal{P}_h^2\}_{h=1}^H, \{r_h^2\}_{h=1}^H\big)$ *and a given policy* $\pi = \{\pi_h\}_{h\in[H]}$*. Their corresponding value functions in the h-th horizon are* V_h^π *and* $V_h^{\pi'}$ *respectively. Then for all* $h \in [H]$ *the following relation holds,*

$$V_h^\pi(s) - V_h^{\pi'}(s) = \mathbb{E}_{\pi,\mathcal{P}}[\sum_{i=h}^{H}(r_i(s_i,a_i) - r_i^{'}(s_i,a_i)) \mid s_h = s] \tag{42}$$

$$+ \mathbb{E}_{\pi,\mathcal{P}'}[\sum_{i=h}^{H}(\mathbb{P}_i V_{i+1}^\pi(s_i,a_i) - \mathbb{P}_i^{'} V_{i+1}^\pi(s_i,a_i)) \mid s_h = s] \tag{43}$$

Proof See Lemma E.15 in Dann et al. (2017) for details. ■

We introduce the following lemma, which gives a self-normalized bound for vector value martingales(Abbasi-Yadkori et al., 2011).

Lemma 37 (Self-Normalized Bound for Vector-Valued Martingales) *Let* $\{\varepsilon_i\}_{i=1}^\infty$ *be a real-valued stochastic process with corresponding filtration* $\{\mathcal{F}_i\}_{i=1}^\infty$ *such that* ε_i *is* $\mathcal{F}_i$ *measurable,* $\mathbb{E}[\varepsilon_i \mid \mathcal{F}_{i-1}] = 0$*, and* ε_i *is conditionally* σ*-sub-Gaussian with* $\sigma \in \mathbb{R}^+$*. Let* $\{X_i\}_{i=1}^\infty$ *be a stochastic process with* $X_i \in \mathcal{H}$ *(some Hilbert space) and* X_i *being* $\mathcal{F}_t$*-measurable. Assume that a linear operator* $V : \mathcal{H} \to \mathcal{H}$ *is positive definite, i.e.,* $x^\top V x > 0$ *for any* $x \in \mathcal{H}$*. For any* t*, define the linear operator* $V_t = V + \sum_{i=1}^t X_i X_i^\top$ *(here* $xx^\top$ *denotes outer-product in* $\mathcal{H}$ *). With probability at least* $1-\delta$*, we have for all* $t \geq 1$

$$\bigg\| \sum_{i=1}^t X_i \varepsilon_i \bigg\|_{V_t^{-1}}^2 \leq 2\sigma^2 \log\left(\frac{\det(V_t)^{1/2}\det(V)^{-1/2}}{\delta}\right).$$

Proof For a detailed proof, see Abbasi-Yadkori et al. (2011). ■

Lemma 38 can be generalized to the case of matrix-valued martingales.

Lemma 38 (Self-Normalized Bound for Matrix-Valued Martingales) *Let* $\{\varepsilon_i\}_{i=1}^\infty$ *be a d-dimensional vector-valued stochastic process with corresponding filtration* $\{\mathcal{F}_i\}_{i=1}^\infty$ *such that* ε_i *is* $\mathcal{F}_i$ *measurable,* $\mathbb{E}\left[\varepsilon_i \mid \mathcal{F}_{i-1}\right] = 0$*, and* ε_i *is conditionally* σ*-sub-Gaussian with* $\sigma \in \mathbb{R}^d$ *Let* $\{X_i\}_{i=1}^\infty$ *be a stochastic process with* $X_i \in \mathcal{H}$ *(some Hilbert space) and* X_i *being* $\mathcal{F}_t$ *measurable. Assume that a linear operator* $V : \mathcal{H} \to \mathcal{H}$ *is positive definite. For any* t*,* *define the linear operator $V_t = V + \sum_{i=1}^{t} X_i X_i^\top$ Then, with probability at least $1-\delta$, we have for all t, we have:*

$$\left\| \sum_{i=1}^{t} \epsilon_i X_i^\top V_t^{-1/2} \right\|_2^2 \le 8\sigma^2 d \log(5) + 8\sigma^2 \log\left(\frac{\det(V_t)^{1/2} \det(V)^{-1/2}}{\delta} \right).$$

Proof Denote $S = \sum_{i=1}^{t} \epsilon_i X_i^\top$. Let us form an ϵ-net, in ℓ_2 distance, $\mathcal{C}$ over the unit ball $\{w : \|w\|_2 \le 1, w \in \mathbb{R}^d\}$. Via a standard covering argument, we can choose $\mathcal{C}$ such that $\log(|\mathcal{C}|) \le d \log(1 + 2/\epsilon)$.

Consider a fixed $w \in \mathcal{C}$ and $w^\top S = \sum_{i=1}^{t} w^\top \epsilon_i X_i^T$. Note that $w^\top \epsilon_i$ is a σ-sub Gaussian due to $\|w\|_2 \le 1$. Hence, Lemma 38 implies that with probability at least $1-\delta$, for all t

$$\left\| V_t^{-1/2} \sum_{i=1}^{t} X_i \left(w^\top \epsilon_i \right) \right\|_2 \le \sqrt{2}\sigma \sqrt{ \log\left(\frac{\det\left(V_t\right)^{1/2} \det(V)^{-1/2}}{\delta} \right)}.$$

Now apply a union bound over $\mathcal{C}$, we get that with probability at least $1-\delta$,

$$\forall w \in \mathcal{C} : \left\| V_t^{-1/2} \sum_{i=1}^{t} X_i \left(w^\top \epsilon_i \right) \right\|_2 \le \sqrt{2}\sigma \sqrt{ d\log(1+2/\epsilon) + \log\left(\frac{\det\left(V_t\right)^{1/2} \det(V)^{-1/2}}{\delta} \right)}.$$

For any w with $\|w\|_2 \le 1$, there exists a $w' \in \mathcal{C}$ such that $\|w - w'\|_2 \le \epsilon$. Hence, for all w such that $\|w\|_2 \le 1$,

$$\begin{aligned} \left\| V_t^{-1/2} \sum_{i=1}^{t} X_i \left(w^\top \epsilon_i \right) \right\|_2 \le & \sqrt{2}\sigma \sqrt{ d\log(1+2/\epsilon) + \log\left(\frac{\det\left(V_t\right)^{1/2} \det(V)^{-1/2}}{\delta} \right)} \\ & + \epsilon \left\| \sum_{i=1}^{t} \epsilon_i X_i^\top V_t^{-1/2} \right\|_2 . \end{aligned}$$

By the definition of the spectral norm, this implies that,

$$\left\| \sum_{i=1}^{t} \epsilon_i X_i^\top V_t^{-1/2} \right\|_2 \le \frac{1}{1-\epsilon} \sqrt{2}\sigma \sqrt{ d\log(1+2/\epsilon) + \log\left(\frac{\det\left(V_t\right)^{1/2} \det(V)^{-1/2}}{\delta} \right)}.$$

Taking $\epsilon = 1/2$ concludes the proof. ■

We introduce the following lemma, which guarantees the MLE convergence refer to Agarwal et al. (2020).

Lemma 39 (MLE bound, Agarwal et al. (2020)) *By Algorithm 3, for a fixed $t \ge 0$ and $h \in [H]$, with probability at least $1-\delta$, we have*

$$\mathbb{E}_{\widehat{\rho}^t}[\|\mathcal{P}_h^*(\cdot|s,a) - \widehat{\mathcal{P}}^t(\cdot|s,a)\|_1^2] \le \frac{2\log(|\Theta||\Upsilon|/\delta)}{t}.$$

As a straightforward corollary, we have with probability at least $1-\delta$,

$$\mathbb{E}_{\widehat{\rho}^t}[\|\mathcal{P}_h^*(\cdot|s,a) - \widehat{\mathcal{P}}^t(\cdot|s,a)\|_1^2] \le \frac{2\log(TH|\Theta||\Upsilon|/\delta)}{t},$$

for all $t \in [T]$ and $h \in [H]$.

The following is a standard inequality to prove regret bounds for online learning in linear models.

Lemma 40 (Agarwal et al. (2020)) *Consider the following process. For* $t = 1, \cdots, T$, $M_t = M_{t-1} + G_t$ *with* $M_0 = \lambda_0 I$ *and* G_t *being a positive semidefinite matrix with eigenvalues upper-bounded by* 1. *We have that*

$$2 \ln \det (M_T) - 2 \ln \det (\lambda_0 I) \geq \sum_{n=1}^{T} \operatorname{Tr} \left(G_t M_{t-1}^{-1} \right).$$

The next lemma provides an upper bound for the potential elliptical lemma and was first proved in Lemma 20 of Uehara et al. (2022). For completeness, we provide its proof.

Lemma 41 ((Uehara et al., 2022)) *Suppose* $\operatorname{Tr}(G_n) \leq B^2$.

$$2 \ln \det (M_N) - 2 \ln \det (\lambda_0 I) \leq d \ln \left(1 + \frac{NB^2}{d\lambda_0} \right).$$

Proof Let $\sigma_1, \cdots, \sigma_d$ be the set of singular values of M_N recalling M_N is a positive semidefinite matrix. Then, by the AM-GM inequality,

$$\ln \det (M_N) / \det (\lambda_0 I) = \ln \prod_{i=1}^{d} (\sigma_i / \lambda_0) \leq d \ln \left(\frac{1}{d} \sum_{i=1}^{d} (\sigma_i / \lambda_0) \right)$$

Since we have $\sum_i \sigma_i = \operatorname{Tr}(M_N) \leq d\lambda_0 + NB^2$, the statement is concluded. ■

The next lemma provides an upper bound for the summation of potential function and is a simple generalization of the elliptical potential lemma(Abbasi-Yadkori et al., 2011). In fact, it is a special case of Lemma 40.

Lemma 42 (Elliptical Potential Lemma) *For any sequence of* $\{\phi_h(s_h^t, a_h^t)\}_{t \in [T], h \in [H]}$, *we have*

$$\sum_{t=1}^{T} \sum_{h=1}^{H} \left\| \phi_h(s_h^t, a_h^t) \right\|_{(\Lambda^t)^{-1}}^2 \leq 2H \log \Bigg(\det(\Lambda^T) \det(\Lambda^0)^{-1} \Bigg).$$

Proof Denote $\phi_h(s_h^t, a_h^t)$ by ϕ_h^t. Recall that $\Lambda^{t+1} = \Lambda^t + \sum_{h=0}^{H-1} \phi_h^t \left(\phi_h^t \right)^\top$ and $\Lambda^0 = \lambda I$. Since $\lambda \geq 1$ and $\|\phi\|_2 \leq 1$, $\|\phi_h^t\|_{(\Lambda^t)^{-1}} \leq 1$ for all $(t, h) \in [T] \times [H]$. Use $x \leq 2H \log(1 + x)$ for $x \in [0, H]$, we have

$$\sum_{h=1}^{H} \left\| \phi_h^t \right\|_{(\Lambda^t)^{-1}}^2 \leq 2H \log \left(1 + \sum_{h=1}^{H} \left\| \phi_h^t \right\|_{(\Lambda^t)^{-1}}^2 \right).$$

For Λ^{t+1}, using its recursive formulation, we have:

$$\log \det(\Lambda^{t+1}) = \log \det(\Lambda^t) + \log \det \left(I + (\Lambda^t)^{-1/2} \sum_{h=1}^{H} \phi_h^t (\phi_h^t)^\top (\Lambda^t)^{-1/2} \right).$$

Denote the eigenvalues of $(\Lambda^t)^{-1/2}\sum_{h=1}^{H}\phi_h^t(\phi_h^t)^\top(\Lambda^t)^{-1/2}$ as σ_i for $i \geq 1$. We have

$$\log\det\left(I + (\Lambda^t)^{-1/2}\sum_{h=1}^{H}\phi_h^t(\phi_h^t)^\top(\Lambda^t)^{-1/2}\right) = \log\prod_{i\geq 1}(1+\sigma_i) \geq \log\big(1+\sum_{i\geq 1}\sigma_i\big),$$

where the last inequality uses that $\sigma_i \geq 0$ for all i. Using the above and the definition of the trace,

$$\begin{aligned}\log\det\left(I + (\Lambda^t)^{-1/2}\sum_{h=1}^{H}\phi_h^t(\phi_h^t)^\top(\Lambda^t)^{-1/2}\right) &\geq \log\left(1+\operatorname{tr}\left((\Lambda^t)^{-1/2}\sum_{h=1}^{H}\phi_h^t(\phi_h^t)^\top(\Lambda^t)^{-1/2}\right)\right)\\ &= \log\left(1+\sum_{h=1}^{H}(\phi_h^t)^\top(\Lambda^t)^{-1}\phi_h^t\right).\end{aligned} \tag{44}$$

Telescoping over $t \in [T]$, we have

$$\begin{aligned}2H\sum_{t=1}^{T}\log\left(1+\sum_{h=1}^{H}(\phi_h^t)^\top(\Lambda^t)^{-1}\phi_h^t\right) &\leq 2H\sum_{t=1}^{T}\big(\log\det(\Lambda^{t+1}) - \log\det(\Lambda^t)\big)\\ &= 2H\log\big(\det(\Lambda^T)\det(\Lambda^0)^{-1}\big),\end{aligned}$$

Therefore, we conclude the proof of Lemma 42. ■

The following lemma was proved in Freedman (1975) and generalizes Bernstein's inequality for independent variables to martingale case.

Lemma 43 (Bernstein's inequality for martingales) *Suppose $X_1,\ldots,X_n$ is a sequence of random variables such that $0 \leq X_i \leq 1$. Define the martingale difference sequence $\{Y_n = \mathbb{E}[X_n \mid X_1,\ldots,X_{n-1}] - X_n\}$ and note K_n the sum of the conditional variances*

$$K_n = \sum_{t=1}^{n}\operatorname{Var}[X_n \mid X_1,\ldots,X_{n-1}].$$

Let $S_n = \sum_{i=1}^{n} X_i$, then for all $\epsilon, v \geq 0$,

$$\mathbb{P}\left(\textstyle\sum_{i=1}^{n}\mathbb{E}[X_n \mid X_1,\ldots,X_{n-1}] - S_n \geq \epsilon, K_n \leq k\right) \leq \exp\left(-\frac{\epsilon^2}{2k+2\epsilon/3}\right)$$

Lemma 44 (χ^2-Distance Between Two Gaussians) *For Gaussian distributions $\mathcal{N}\left(\mu_1,\sigma^2\mathcal{I}\right)$ and $\mathcal{N}\left(\mu_2,\sigma^2\mathcal{I}\right)$, the (squared) chi-squared distance between $\mathcal{N}_1$ and $\mathcal{N}_2$ is,*

$$\int\frac{(\mathcal{N}_1(z)-\mathcal{N}_2(z))^2}{\mathcal{N}_1(z)}dz = \exp\left(\frac{\|\mu_1-\mu_2\|^2}{2\sigma^2}\right) - 1.$$

Proof Note that,

$$\int\frac{(\mathcal{N}_1(z)-\mathcal{N}_2(z))^2}{\mathcal{N}_1(z)}dz = \int\mathcal{N}_1(z) - 2\mathcal{N}_2(z) + \frac{\mathcal{N}_2(z)^2}{\mathcal{N}_1(z)}dz = -1 + \int\frac{\mathcal{N}_2(z)^2}{\mathcal{N}_1(z)}dz.$$

Also note that for $\mathcal{N}_2^2(z)/\mathcal{N}_1(z)$, we have

$$\mathcal{N}_2^2(z)/\mathcal{N}_1(z) = \frac{1}{Z}\exp\left(-\frac{1}{2\sigma^2}\left(2\left\|z-\mu_2\right\|_2^2 - \left\|z-\mu_1\right\|_2^2\right)\right),$$

where Z is the normalization constant for $\mathcal{N}\left(0,\sigma^2\mathcal{I}\right)$, i.e. $Z = \int \exp\left(-\frac{1}{2\sigma^2}\|z\|_2^2\right)dz$. Thus, for $2\left\|z-\mu_2\right\|_2^2 - \left\|z-\mu_1\right\|_2^2$, we can verify that

$$2\left\|z-\mu_2\right\|_2^2 - \left\|z-\mu_1\right\|_2^2 = \left\|z+(\mu_1-2\mu_2)\right\|_2^2 - 2\left\|\mu_1-\mu_2\right\|_2^2.$$

which implies,

$$\begin{aligned}\int \frac{\mathcal{N}_2(z)^2}{\mathcal{N}_1(z)}dz &= \frac{1}{Z}\int \exp\left(-\frac{1}{2\sigma^2}\left(\left\|z-(2\mu_2-\mu_1)\right\|_2^2 - 2\left\|\mu_1-\mu_2\right\|\right)\right)dz \\ &= \frac{1}{Z}\exp\left(\frac{\left\|\mu_1-\mu_2\right\|_2^2}{\sigma^2}\right)\int \exp\left(-\frac{1}{2\sigma^2}\left\|z-(2\mu_2-\mu_1)\right\|_2^2\right)dz \\ &= \exp\left(\frac{\left\|\mu_1-\mu_2\right\|_2^2}{\sigma^2}\right).\end{aligned}$$

Therefore, we conclude the proof. ■

Lemma 45 (Expectation Difference Under Two Gaussians) *For Gaussian distribution $\mathcal{N}(\mu_1,\sigma^2\mathcal{I})$ and $\mathcal{N}(\mu_2,\sigma^2\mathcal{I})$, suppose that $\{\|\mu_1\|_2, \|\mu_2\|_2\} \le B$, then for any (appropriately measurable) positive function g, it holds that:*

$$\mathbb{E}_{z\sim\mathcal{N}_1}[g(z)] - \mathbb{E}_{z\sim\mathcal{N}_2}[g(z)] \le C(\sigma,B)\cdot\frac{\|\mu_1-\mu_2\|_2}{\sigma}\sqrt{\mathbb{E}_{z\sim\mathcal{N}_1}[g(z)^2]},$$

where $C(\sigma,B) = \exp(B^2/\sigma^2)$.

Proof Define $m_i = \mathbb{E}_{z\sim\mathcal{N}_1}[g(z)]$ for $i \in \{0,1\}$. We have:

$$\begin{aligned} m_1 - m_2 &= \mathbb{E}_{z\sim\mathcal{N}_1}\left[g(z)\left(1-\frac{N_2(z)}{N_1(z)}\right)\right] \\ &\le \sqrt{\mathbb{E}_{z\sim\mathcal{N}_1}\left[g(z)^2\right]}\sqrt{\int\frac{\left(N_1(z)-N_2(z)\right)^2}{N_1(z)}dz} \\ &= \sqrt{\mathbb{E}_{z\sim\mathcal{N}_1}[g(z)^2]}\sqrt{\exp\left(\frac{\|\mu_1-\mu_2\|_2^2}{2\sigma^2}\right)-1}. \end{aligned}$$

By convexity we have $\exp(x) \le 1 + x\exp(x)$ for all x, we have

$$\begin{aligned}\exp\left(\frac{\|\mu_1-\mu_2\|_2^2}{2\sigma^2}\right) - 1 &\le \frac{\|\mu_1-\mu_2\|_2^2}{2\sigma^2}\cdot\exp\left(\frac{\|\mu_1-\mu_2\|_2^2}{2\sigma^2}\right) \\ &\le \frac{\|\mu_1-\mu_2\|_2^2}{2\sigma^2}\cdot\exp\left(\frac{2B^2}{\sigma^2}\right).\end{aligned}$$

Therefore, we have

$$m_1 - m_2 \leq \exp\left(\frac{B^2}{\sigma^2}\right) \cdot \frac{\|\mu_1 - \mu_2\|_2}{\sigma} \sqrt{\mathbb{E}_{z \sim \mathcal{N}_1}[g(z)^2]}.$$

■

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004a.

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 1, New York, NY, USA, 2004b. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015430. URL `https://doi.org/10.1145/1015330.1015430`.

Naoki Abe, Prem Melville, Cezar Pendus, Chandan K. Reddy, David L. Jensen, Vince P. Thomas, James J. Bennett, Gary F. Anderson, Brent R. Cooley, Melissa Kowalczyk, Mark Domick, and Timothy Gardinier. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 75–84, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300551. doi: 10.1145/1835804.1835817. URL `https://doi.org/10.1145/1835804.1835817`.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *arXiv preprint arXiv:2006.10814*, 2020.

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

Amir Beck. *First-order methods in optimization.* SIAM, 2017.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

Kianté Brantley, Miroslav Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings, 2021.

L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(67)90040-7.

Yichen Chen and Mengdi Wang. Stochastic primal-dual methods and sample complexity of reinforcement learning. *arXiv preprint arXiv:1612.02516*, 2016.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. *arXiv preprint*, 2008.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710*, 2017.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021.

A.Ya. Dubovitskii and A.A. Milyutin. Extremum problems in the presence of restrictions. *USSR Computational Mathematics and Mathematical Physics*, 5(3):1–80, 1965. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(65)90148-5. URL https://www.sciencedirect.com/science/article/pii/0041555365901485.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

David A. Freedman. On Tail Probabilities for Martingales. *The Annals of Probability*, 3(1):100 – 118, 1975. doi: 10.1214/aop/1176996452. URL https://doi.org/10.1214/aop/1176996452.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Hassan Hijazi and Leo Liberti. Constraint qualification failure in action. *Operations Research Letters*, 44(4):503–506, 2016.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3), Jun 2008. ISSN 0090-5364. doi: 10.1214/009053607000000677. URL http://dx.doi.org/10.1214/009053607000000677.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *arXiv preprint arXiv:2006.12466*, 2020.

Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *arXiv preprint arXiv:2006.10277*, 2020.

Sobhan Miryoosefi, Kianté Brantley, Hal Daume III, Miro Dudik, and Robert E Schapire. Reinforcement learning with convex constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps, 2022.

Igor Mordatch, Zoran Popović, and Emanuel Todorov. Contact-invariant optimization for hand manipulation. In *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 137–144, 2012.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*, 2016.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. *Advances in Neural Information Processing Systems*, 33:15277–15287, 2020.

R Tyrrell Rockafellar. *Convex Analysis*. Citeseer, 1970.

Yuda Song and Wen Sun. Pc-mlp: Model-based reinforcement learning with policy cover guided exploration, 2021.

Umar Syed, Michael Bowling, and Robert E Schapire. Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on Machine learning*, pages 1032–1039, 2008.

Ambuj Tewari and Peter Bartlett. Optimistic linear programming gives logarithmic regret for irreducible mdps. *Advances in Neural Information Processing Systems*, 20, 2007.

Emanuel Todorov and Weiwei Li. A generalized iterative lqg method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *Proceedings of the 2005, American Control Conference, 2005.*, pages 300–306. IEEE, 2005.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline rl: Pac bounds and posterior sampling under partial coverage. *arXiv e-prints*, pages arXiv–2107, 2021.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps, 2022.

Sharan Vaswani, Lin F. Yang, and Csaba Szepesvári. Near-optimal sample complexity bounds for constrained mdps, 2022.

Nolan Wagener, Ching-An Cheng, Jacob Sacks, and Byron Boots. An online learning approach to model predictive control, 2019.

Mengdi Wang. Primal-dual π learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017.

Mengdi Wang. Randomized linear programming solves the markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*, 45(2): 517–546, 2020.

Grady Williams, Andrew Aldrich, and Evangelos Theodorou. Model predictive path integral control using covariance variable importance sampling, 2015.

Runzhe Wu, Yufeng Zhang, Zhuoran Yang, and Zhaoran Wang. Offline constrained multi-objective reinforcement learning via pessimistic dual value iteration. *Advances in Neural Information Processing Systems*, 34:25439–25451, 2021.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael Jordan. Provably efficient reinforcement learning with kernel and neural function approximations. *Advances in Neural Information Processing Systems*, 33, 2020.

Tiancheng Yu, Yi Tian, Jingzhao Zhang, and Suvrit Sra. Provably efficient algorithms for multi-objective competitive rl. *arXiv preprint arXiv:2102.03192*, 2021.

Tom Zahavy, Alon Cohen, Haim Kaplan, and Yishay Mansour. Apprenticeship learning via frank-wolfe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6720–6728, 2020.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. *arXiv preprint arXiv:2106.00661*, 2021.

Junyu Zhang, Amrit Singh Bedi, Mengdi Wang, and Alec Koppel. Cautious reinforcement learning via distributional risk in the dual domain. *arXiv preprint arXiv:2002.12475*, 2020a.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities, 2020b.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.