
Efficient Performance Bounds for Primal-Dual Reinforcement Learning from Demonstrations

Angeliki Kamoutsi¹ Goran Banjac¹ John Lygeros¹

Abstract

We consider large-scale Markov decision processes with an unknown cost function and address the problem of learning a policy from a finite set of expert demonstrations. We assume that the learner is not allowed to interact with the expert and has no access to reinforcement signal of any kind. Existing inverse reinforcement learning methods come with strong theoretical guarantees, but are computationally expensive, while state-of-the-art policy optimization algorithms achieve significant empirical success, but are hampered by limited theoretical understanding. To bridge the gap between theory and practice, we introduce a novel bilinear saddle-point framework using Lagrangian duality. The proposed primal-dual viewpoint allows us to develop a model-free provably efficient algorithm through the lens of stochastic convex optimization. The method enjoys the advantages of simplicity of implementation, low memory requirements, and computational and sample complexities independent of the number of states. We further present an equivalent no-regret online-learning interpretation.

1. Introduction

Reinforcement learning (RL) is an area in machine learning with connections to control and optimization that has shown tremendous success in large-scale real-world applications, such as robotics, artificial intelligence, cognitive autonomy, operations research, and healthcare (Tesauro & Kephart, 2002; Mnih et al., 2015; Chen et al., 2017; Vamvoudakis & Kokolakis, 2020). It studies the problem of learning to act optimally in a sequential decision-making problem, while interacting with an unknown environment (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998).

¹Automatic Control Laboratory, ETH Zurich, Switzerland. Correspondence to: Angeliki Kamoutsi <kamoutsa@ethz.ch>.

In the standard RL setting a cost signal is given to instruct agents how to complete the desired task. However, oftentimes encoding preferences using demonstrations provided by an expert, is easier than designing a cost function (Pomerleau, 1991; Russell, 1998; Bagnell, 2015). In such cases, the goal of *learning from demonstrations* (LfD) or *imitation learning* (IL) is to learn a policy that achieves or even surpasses the performance of the expert policy.

A lot of methods have been proposed to solve the LfD problem. The most straightforward approach is *behavior cloning*, which casts the problem as a supervised learning problem, in which the goal is to learn a map from states to optimal actions (Pomerleau, 1991). Although behavior cloning is simple and easy to implement, the crucial i.i.d. assumption made in supervised learning is violated. As a result, the approach suffers from the problem of *cascading errors*, which is related to *covariate shift* (Bagnell, 2015; Ho et al., 2016). Later works (Ross et al., 2011) eliminate this distribution mismatch by formulating the problem as a no-regret online learning problem, but require interaction with the expert. On the contrary, in this paper, we consider the *batch* LfD scenario, i.e., the learner can observe only a finite set of expert demonstrations, is not allowed to query the expert for more data while training, and is not provided any reinforcement signal.

Inverse reinforcement learning (IRL) (Abbeel & Ng, 2004) is a prevalent approach to LfD. In this paradigm, the learner first infers the unknown cost function that the expert tries to minimize and then uses it to reproduce the optimal behavior. IRL algorithms do not suffer from the problem of cascading errors because the training takes place over entire expert trajectories, rather than individual actions. In addition, since the recovered cost function “explains” the expert behavior, they can easily generalize to unseen states or even new environments. Note however, that most existing IRL algorithms (Abbeel & Ng, 2004; Ratliff et al., 2006; Syed & Schapire, 2007; Neu & Szepesvári, 2007; Ziebart et al., 2008; Abbeel et al., 2008; Levine et al., 2010; 2011) are computationally expensive because they use RL as a subroutine.

On the other hand, under the assumption of a linearly parameterized cost class, one can frame the problem as a single convex program (Syed et al., 2008), bypassing the inter-

mediate step of learning the cost function. Although the associated program can be solved exactly for small-sized MDPs, the approach suffers from the *curse of dimensionality*, making it intractable for large-scale problems.

The formulations and reasoning in (Syed et al., 2008) formed the ground and inspired later state-of-the-art policy optimization algorithms (Ho et al., 2016; Ho & Ermon, 2016). In particular, the authors in (Ho et al., 2016) consider the case of linearly parameterized cost classes and present policy gradient algorithms, which are parallel to those proposed in (Williams, 1992; Schulman et al., 2015) for RL. Moreover, (Ho & Ermon, 2016) propose a generative adversarial imitation learning (GAIL) method for the case of nonlinear costs. They do so, by formulating the problem as minimax optimization and drawing a connection to generative adversarial networks (Goodfellow et al., 2014). In particular, GAIL solves IL with alternating updates of both policy and cost functions. These approaches are model-free and achieve significant empirical success in challenging benchmark tasks. However, in general the associated minimax problem is highly nonconvex-nonconcave and as a result remains hampered by limited theoretical understanding.

Indeed, the global convergence properties of alternating policy gradient schemes for the minimax formulation of GAIL have been studied only for linear (although infinite-dimensional) Markov decision processes (MDPs) and linear or linearizable costs (Cai et al., 2019; Zhang et al., 2020). It however remains unclear whether such neural policy gradient methods converge to the optimal policy or if they converge at all, for the case of nonlinear dynamics. As a result, provably efficient policy optimization schemes for the LfD problem beyond the linear setting remain largely unexplored.

Contributions. In an attempt to tackle this longstanding question, in this work we present a convex-analytic viewpoint of the LfD problem. To this end, we adopt the apprenticeship learning (AL) formalism which carries the assumption that the unknown true cost function can be represented as a weighted combination of some known basis functions, where the true unknown weights specify how different desiderata should be traded-off (Abbeel & Ng, 2004; Syed & Schapire, 2007; Syed et al., 2008; Ho & Ermon, 2016; Brown et al., 2020a). In particular, we make no restrictive assumptions on the MDP model (linearity or ergodicity).

Following the recent line of works (Chen et al., 2018; Lee & He, 2019; Wang, 2020; Bas-Serrano & Neu, 2020; Cheng et al., 2020; Jin & Sidford, 2020; Shariff & Szepesvári, 2020) on approximate linear programming (ALP) for large-scale MDPs, we formulate the LfD problem as a bilinear saddle-point problem in light of Lagrangian duality. We

study primal-dual optimality conditions and prove relations between saddle points of the Lagrangian function and optimal solutions to the LfD problem. In particular, we show that under the expert optimality assumption, the set of solutions to the dual linear program (LP) characterizes the set of solutions to the inverse problem, i.e., the set of cost functions for which the expert is optimal. Moreover, in this case, we show that the expert policy, the true cost function, and the true optimal value function form a saddle point of the proposed Lagrangian.

Analogous to ALP, we obtain a linearly-relaxed saddle-point formulation by limiting our search to a linear subspace defined by a small number of features. We exhibit a formal link between approximate saddle points of the reduced Lagrangian and the optimality gap of our problem.

The aforementioned analysis lays theoretical foundations for a provably efficient stochastic primal-dual algorithm. By using linear function approximators, we propose a mirror-descent-based algorithm with a generative model and derive explicit probabilistic performance bounds on the quality of the extracted policy. A salient feature of the algorithm is that its sample complexity does not depend on the size of the state space but instead on the number of approximation features. We note however that our algorithm degrades with the approximation error due to the linear approximation architecture. We consider both the case of weak and strong linear features (Lakshminarayanan et al., 2018; Shariff & Szepesvári, 2020). Finally, we present an equivalent no-regret online-learning interpretation of our primal-dual algorithm. This kind of reduction has been studied up to now only for the online IL setting (Ross et al., 2011), where interaction with the expert is allowed.

Related works Our work is related to the LP approach to AL (Syed et al., 2008) and the theoretical analysis of GAIL made in (Cai et al., 2019; Chen et al., 2020a; Zhang et al., 2020). Unlike (Syed et al., 2008), who consider the case of tabular MDPs and known dynamics, we consider large-scale problems and only access to a simulator for the MDP model. We revisit the LP approach in (Syed et al., 2008) and focus instead on its saddle-point formulation due to its potential for scalable algorithms with theoretical guarantees (Cheng et al., 2020; Nachum & Dai, 2020). The authors in (Cai et al., 2019) study the global convergence properties of GAIL for the linear quadratic regulator problem by extending the results in (Fazel et al., 2018), while the authors in (Zhang et al., 2020) consider the case of infinite-dimensional linear MDPs and linearizable cost functions. Similar to (Wang et al., 2020) the policies and cost functions are approximated by overparameterized ReLU neural networks. In our work, we consider general nonlinear MDPs without any restrictive assumption such as linearity or ergodicity, and employ linear function approximators. The authors in (Chen et al.,

2020a) study the convergence and generalization of GAIL for general MDPs. However they only prove convergence to a stationary point. On the contrary, our algorithm produces a nearly-optimal policy.

Our work builds upon a vast line of works on ALP for forward RL (De Farias & Van Roy, 2003; Abbasi-Yadkori et al., 2014; Chen et al., 2018; Lakshminarayanan et al., 2018; Mohajerin Esfahani et al., 2018; Wang, 2020; Lee & He, 2019; Bas-Serrano et al., 2020; Cheng et al., 2020; Jin & Sidford, 2020; Shariff & Szepesvári, 2020; Bas-Serrano & Neu, 2020). The LP approach to MDPs, which dates back to (Manne, 1960; Hernández-Lerma & Lasserre, 1996; Borkar, 1988), has recently gained traction as an alternative to dynamic programming techniques, for its advantage to lead to problem formulations that are directly amenable to modern large-scale stochastic optimization methods. Moreover, this optimization-based approach can tackle unconventional problems involving additional safety constraints or secondary costs, where traditional dynamic programming techniques are not applicable (Hernández-Lerma et al., 2003; Dufour & Prieto-Rumeau, 2013; Shafieepoofard et al., 2013). In this paper, we develop an ALP framework for the LfD problem. While the forward policy optimization problem tries to minimize the total expected cost, the LfD problem is a minimax problem that tries to match the expert across a given cost class. In our saddle-point formulation the cost function is not fixed but is itself a decision variable. Thus, the variation of the cost function during the algorithm continuously changes the Lagrangian making the analysis of the problem more challenging.

2. Preliminaries and Problem Setup

2.1. Basic Definitions and Notations

We denote by \mathbf{R}_+^n and \mathbf{R}_{++}^n the sets of n -dimensional vectors with nonnegative and positive real elements, respectively. For a matrix $\mathbf{A} \in \mathbf{R}^{m \times n}$, its p -norm is defined by $\|\mathbf{A}\|_p \triangleq \sup\{\|\mathbf{Ax}\|_p \mid \|\mathbf{x}\|_p = 1\}$. For vectors \mathbf{x} and \mathbf{y} , we denote by $\langle \mathbf{x}, \mathbf{y} \rangle$ the usual inner product. Moreover, $\mathbf{x} \leq \mathbf{y}$ denotes elementwise inequality, i.e., $x_i \leq y_i$ for all i . We use $\mathbf{1}$ and $\mathbf{0}$ to denote vectors with all elements equal to one and zero, respectively. The set of probability distributions on a finite set \mathcal{S} is denoted by $\Delta_{\mathcal{S}}$, i.e., $\Delta_{\mathcal{S}} \triangleq \{\mathbf{p} \in \mathbf{R}_+^{|\mathcal{S}|} \mid \sum_{s \in \mathcal{S}} p(s) = 1\}$, where $|\mathcal{S}|$ is the cardinality of \mathcal{S} . Sums spanning over the spaces \mathcal{X} and \mathcal{A} will be simply denoted by \sum_x and \sum_a , respectively. For $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \Delta_{\mathcal{X}}$ their *Kullback–Leibler divergence* is given by $\text{KL}(\boldsymbol{\mu} \parallel \boldsymbol{\mu}') \triangleq \sum_x \mu(x) \log \frac{\mu(x)}{\mu'(x)}$. For a nonempty closed convex set Θ , the Euclidean projection of \mathbf{x} onto Θ is given by $\Pi_{\Theta}(\mathbf{x}) \triangleq \arg \min_{\mathbf{y} \in \Theta} \|\mathbf{x} - \mathbf{y}\|_2$.

2.2. Reinforcement Learning

A finite MDP is given by a tuple $(\mathcal{X}, \mathcal{A}, P, \boldsymbol{\nu}_0, \mathbf{c}, \gamma)$, where \mathcal{X} is the state space, \mathcal{A} is the action space, $P : \mathcal{X} \times \mathcal{A} \rightarrow \Delta_{\mathcal{X}}$ is the transition law, $\boldsymbol{\nu}_0 \in \Delta_{\mathcal{X}}$ is the initial state distribution, $\mathbf{c} \in [-1, 1]^{|\mathcal{X}| \times |\mathcal{A}|}$ is the one-stage cost, and $\gamma \in (0, 1)$ is the discount factor. We focus on problems where \mathcal{X} and \mathcal{A} are too large to be enumerated.

The MDP models a controlled discrete-time stochastic system with initial state $x_0 \sim \boldsymbol{\nu}_0$. At each round t , if the system is in state $x_t = x \in \mathcal{X}$ and the action $a_t = a \in \mathcal{A}$ is taken, then a cost $c(x, a)$ is incurred, and the system transitions to the next state $x_{t+1} \sim P(\cdot | x, a)$.

A *stationary Markov policy* is a map $\pi : \mathcal{X} \rightarrow \Delta_{\mathcal{A}}$, and $\pi(a|x)$ denotes the probability of choosing action a , while being in state x . We denote the space of stationary Markov policies by Π_0 .

The *value function* $\mathbf{V}_{\mathbf{c}}^{\pi} \in \mathbf{R}^{|\mathcal{X}|}$ of π , given a cost \mathbf{c} , is defined by $V_{\mathbf{c}}^{\pi}(x) \triangleq (1-\gamma) \mathbf{E}_x^{\pi} \left[\sum_{t=0}^{\infty} \gamma^t c(x_t, a_t) \right]$, where \mathbf{E}_x^{π} denotes the expectation with respect to the trajectories generated by π starting from $x_0 = x$.

The goal of RL is to solve the following optimal control problem

$$\rho_{\mathbf{c}}^* \triangleq \min_{\pi \in \Pi_0} \rho_{\mathbf{c}}(\pi), \quad (\mathbf{RL}_{\mathbf{c}})$$

where $\rho_{\mathbf{c}}(\pi) = \langle \boldsymbol{\nu}_0, \mathbf{V}_{\mathbf{c}}^{\pi} \rangle$ is the *total expected cost* of π .

For every policy π , we define the *normalized state-action occupancy measure* $\boldsymbol{\mu}_{\pi} \in \Delta_{\mathcal{X} \times \mathcal{A}}$, by $\boldsymbol{\mu}_{\pi}(x, a) \triangleq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbf{P}_{\boldsymbol{\nu}_0}^{\pi} [x_t = x, a_t = a]$, where $\mathbf{P}_{\boldsymbol{\nu}_0}^{\pi}[\cdot]$ denotes the probability of an event when following π starting from $x_0 \sim \boldsymbol{\nu}_0$. The occupancy measure can be interpreted as the discounted visitation frequency of state-action pairs. This allows us to write $\rho_{\mathbf{c}}(\pi) = \langle \boldsymbol{\mu}_{\pi}, \mathbf{c} \rangle$.

The *optimal value function* $\mathbf{V}_{\mathbf{c}}^* \in \mathbf{R}^{|\mathcal{X}|}$ is defined by $V_{\mathbf{c}}^*(x) \triangleq \min_{\pi \in \Pi_0} V_{\mathbf{c}}^{\pi}(x)$. For clarity, when the cost \mathbf{c} is a parameterized function, written as $\mathbf{c}_{\mathbf{w}}$ for a parameter vector \mathbf{w} , we will replace \mathbf{c} by \mathbf{w} in the notation of the aforementioned quantities, e.g., $\rho_{\mathbf{w}}(\pi)$, $\mathbf{V}_{\mathbf{w}}^{\pi}$, etc.

2.3. Learning from Demonstrations

The goal of LfD is to learn a policy that outperforms the expert policy π_E for an unknown true cost function \mathbf{c}_{true} . We assume that the learner is given only a finite set of truncated expert sample trajectories and is not allowed to interact or query the expert for more data while training.

Although the MDP model is not known, we assume access to a *generative-model oracle* which, given a state-action pair (x, a) , outputs the next state $x' \sim P(\cdot | x, a)$. Moreover we can sample $x_0 \sim \boldsymbol{\nu}_0$. This is also known as the simulator-defined MDP (Szörényi et al., 2014; Taleghan et al., 2015).

To address the LfD problem, we adopt the AL formalism (Abbeel & Ng, 2004; Syed et al., 2008; Ho et al., 2016; Ho & Ermon, 2016), which carries the assumption that \mathbf{c}_{true} belongs to a class of cost functions \mathcal{C} . We then seek a policy that performs better than the expert across \mathcal{C} by solving the following minimax optimization problem

$$\alpha^* \triangleq \min_{\pi \in \Pi_0} \max_{\mathbf{c} \in \mathcal{C}} \rho_{\mathbf{c}}(\pi) - \rho_{\mathbf{c}}(\pi_E). \quad (\mathbf{LfD}_{\pi_E})$$

Equivalently, we can write $\alpha^* = \min_{\pi} \delta_{\mathcal{C}}(\pi, \pi_E)$, where $\delta_{\mathcal{C}}(\pi, \pi_E) \triangleq \max_{\mathbf{c} \in \mathcal{C}} (\rho_{\mathbf{c}}(\pi) - \rho_{\mathbf{c}}(\pi_E))$ denotes the \mathcal{C} -distance between π and π_E (Ho et al., 2016; Chen et al., 2020a; Zhang et al., 2020). An optimal solution π_A to (\mathbf{LfD}_{π_E}) is called an *apprentice policy* and satisfies $\rho_{\mathbf{c}_{\text{true}}}(\pi_A) \leq \rho_{\mathbf{c}_{\text{true}}}(\pi_E) + \alpha^*$ with α^* being always nonpositive.

Intuitively, the cost class \mathcal{C} distinguishes the expert from other policies. The maximization in (\mathbf{LfD}_{π_E}) assigns high total cost to non-expert policies and low total cost to π_E (Ho et al., 2016), while the minimization aims to find the policy that matches the expert as close as possible with respect to the \mathcal{C} -distance.

In this work, we assume that the true cost function can be represented as a convex combination of some known features, where the true unknown weights specify how different desiderata should be traded-off. In particular, we consider the following cost function class (Syed & Schapire, 2007; Syed et al., 2008; Ho et al., 2016)

$$\mathcal{C} = \mathcal{C}_{\text{conv}} \triangleq \{\mathbf{c}_{\mathbf{w}} \triangleq \sum_{i=1}^{n_{\mathbf{c}}} w_i \mathbf{c}_i \mid w_i \geq 0, \sum_{i=1}^{n_{\mathbf{c}}} w_i = 1\},$$

where $\{\mathbf{c}_i\}_{i=1}^{n_{\mathbf{c}}} \subset \mathbf{R}^{|\mathcal{X}||\mathcal{A}|}$ are fixed cost vectors, such that $\|\mathbf{c}_i\|_{\infty} \leq 1$ for all $i = 1, \dots, n_{\mathbf{c}}$. In this case, $\delta_{\mathcal{C}}(\pi, \pi_E) = \max_{i \in [n_{\mathbf{c}}]} (\rho_{\mathbf{c}_i}(\pi) - \rho_{\mathbf{c}_i}(\pi_E))$.

Note that this assumption is not necessarily restrictive as usually in practice the true cost function depends on just a few key properties, but the desirable weighting is unknown (Abbeel & Ng, 2004). Moreover, these features can be arbitrarily complex nonlinear functions and can be obtained via unsupervised learning from raw state observations (Brown et al., 2020b; Chen et al., 2020b).

We highlight that one can consider other linearly parameterized cost classes, e.g., $\mathcal{C}_{\text{lin}} = \{\sum_{i=1}^{n_{\mathbf{c}}} w_i \psi_i \mid \|w\|_2 \leq 1\}$ (Abbeel & Ng, 2004), leading to similar reasoning and analysis.

2.4. Technique Overview

Before diving into technical details, we provide a technique overview by formalizing the main building block of our analysis, as described informally in the introduction. Figure 1 illustrates the big picture behind our reduction.

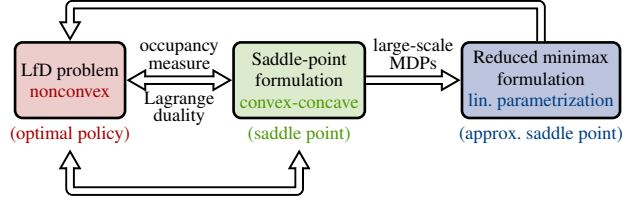


Figure 1. Main building blocks of our analysis.

The main difficulty in (\mathbf{LfD}_{π_E}) comes from its nonconvex-nonconcave structure. In Section 3, by using Lagrangian duality, we handle this difficulty by transforming the original nonconvex-nonconcave optimization program (\mathbf{LfD}_{π_E}) into the bilinear saddle-point problem (1). The new decision variables are occupancy measures μ , value functions \mathbf{u} and cost weights \mathbf{w} . The two formulations are equivalent, since the primal optimizer μ_A directly maps to that of the original problem π_A and vice versa. Moreover, under the assumption of expert optimality the triplet $(\mu_{\pi_E}, \mathbf{V}_{\mathbf{w}_{\text{true}}}^*, \mathbf{w}_{\text{true}})$ is a saddle-point of (1).

For large-scale MDPs the saddle-point formulation (1) is intractable. To mitigate this difficulty, in Section 4 we consider the reduced saddle-point problem (2) by introducing parameterized families $\{\mu_{\theta} \mid \theta \in \Theta\}$ and $\{\mathbf{u}_{\lambda} \mid \lambda \in \Lambda\}$, for appropriately chosen parameter sets Θ and Λ . We then relate the saddle-point residual $\epsilon_{\text{sad}}(\theta, \lambda, \mathbf{w})$ (3) of an approximate saddle-point $(\theta, \lambda, \mathbf{w})$ to the suboptimality gap of the extracted policy π to the original (\mathbf{LfD}_{π_E}) . In particular we show that $\delta_{\mathcal{C}}(\pi_{\theta}, \pi_E) - \delta_{\mathcal{C}}(\pi_A, \pi_E) \leq k \epsilon_{\text{sad}}(\theta, \lambda, \mathbf{w}) + \epsilon_{\text{approx}}$, where ϵ_{approx} is a measure of expressivity of the function approximators and $k \in \{1, 3\}$. In this way, it is apparent that (\mathbf{LfD}_{π_E}) is decoupled in two parts: minimization of $\epsilon_{\text{sad}}(\theta, \lambda, \mathbf{w})$ and function approximation.

In Section 5, we choose linear function approximators (LFA) to preserve the convexity-concavity of the Lagrangian and design a scalable algorithm with sublinear convergence rate to a neighborhood of π_A , where the representation power of LFA determines the convergence error. In particular, we present a variance reduced stochastic mirror descent algorithm (Nemirovski et al., 2009) under local norms and coordinate-wise gradient estimators (Carmon et al., 2019; Jin & Sidford, 2020) to solve the (\mathbf{LfD}_{π_E}) problem with a generative-model oracle.

3. The Convex Optimization View

In this section, we revisit the LP approach to the LfD problem (Syed et al., 2008) and introduce a novel linear duality framework, which will be later exploited to lay theoretical foundations for an efficient stochastic primal-dual algorithm. The details on the formulations and the proofs of this section

can be found in Appendix A.

3.1. Primal-Dual Linear Programming Formulations

Assuming that $\mathcal{C} = \mathcal{C}_{\text{conv}}$, one can frame the (\mathbf{Lfd}_{π_E}) problem as an LP over occupancy measures (Syed et al., 2008). We first review briefly the rationale behind the formulation.

The set of occupancy measures can be characterized in terms of linear constraint satisfaction. To this aim, let $\mathfrak{F} \triangleq \{\boldsymbol{\mu} \in \mathbf{R}^{|\mathcal{X}||\mathcal{A}|} \mid \mathbf{T}_\gamma \boldsymbol{\mu} = \boldsymbol{\nu}_0, \boldsymbol{\mu} \geq \mathbf{0}\}$, where $\mathbf{T}_\gamma: \mathbf{R}^{|\mathcal{X}||\mathcal{A}|} \rightarrow \mathbf{R}^{|\mathcal{X}|}$ is a linear operator given by $\mathbf{T}_\gamma \boldsymbol{\mu} = \frac{1}{1-\gamma}(\mathbf{B} - \gamma\mathbf{P})^\top \boldsymbol{\mu}$. Here, \mathbf{P} is the vector form of P , i.e., $P(x,a), x' \triangleq P(x'|x, a)$, and \mathbf{B} is a binary matrix defined by $B(x,a), x' \triangleq 1$ if $x = x'$, and $B(x,a), x' \triangleq 0$ otherwise.

Proposition 1. (Puterman, 1994) *It holds that, $\mathfrak{F} = \{\boldsymbol{\mu}_\pi \mid \pi \in \Pi_0\}$. Indeed, for every $\pi \in \Pi_0$, we have that $\boldsymbol{\mu}_\pi \in \mathfrak{F}$. Moreover, for every feasible solution $\boldsymbol{\mu} \in \mathfrak{F}$, we can obtain a stationary Markov policy $\pi_\mu \in \Pi_0$ by $\pi_\mu(a|x) \triangleq \frac{\boldsymbol{\mu}(x,a)}{\sum_{a' \in \mathcal{A}} \boldsymbol{\mu}(x,a')}$. Then, the induced occupancy measure is exactly $\boldsymbol{\mu}$.*

Using the epigraph reformulation and Proposition 1, it follows that (\mathbf{Lfd}_{π_E}) is equivalent to the following primal LP:

$$\begin{aligned} \min_{(\boldsymbol{\mu}, \varepsilon)} \quad & \varepsilon \\ \text{s.t.} \quad & \langle \boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_i \rangle \leq \varepsilon, \quad i \in [n_c], \\ & \boldsymbol{\mu} \in \mathfrak{F}. \end{aligned} \quad (\mathbf{P}_{\pi_E})$$

The linear program (\mathbf{P}_{π_E}) resembles the dual LP for solving forward MDPs (Puterman, 1994). Its optimization variables are occupancy measures and a maximum per-feature cost component.

The constraints that define the set \mathfrak{F} are also known as *Bellman flow constraints* and ensure that $\boldsymbol{\mu}$ is an occupancy measure generated by a stationary Markov policy. In particular, $(\boldsymbol{\mu}_{\pi_A}, \alpha^*)$ is a primal optimizer. Conversely, by an optimal occupancy measure $\boldsymbol{\mu}_A$, an apprentice policy can be extracted as $\pi_{\boldsymbol{\mu}_A}$.

The main objective of this subsection is to shed light to the dual of (\mathbf{P}_{π_E}) and interpret the dual optimizers.

It holds that $\langle \boldsymbol{\nu}_0, \mathbf{u} \rangle = \langle \mathbf{T}_\gamma \boldsymbol{\mu}_{\pi_E}, \mathbf{u} \rangle = \langle \boldsymbol{\mu}_{\pi_E}, \mathbf{T}_\gamma^* \mathbf{u} \rangle$, where $\mathbf{T}_\gamma^*: \mathbf{R}^{|\mathcal{X}|} \rightarrow \mathbf{R}^{|\mathcal{X}||\mathcal{A}|}$, given by $\mathbf{T}_\gamma^* \mathbf{u} = \frac{1}{1-\gamma}(\mathbf{B} - \gamma\mathbf{P})\mathbf{u}$, is the adjoint operator of \mathbf{T}_γ . We then get by standard linear duality that the dual LP is given by

$$\begin{aligned} \max_{(\mathbf{u}, \mathbf{w})} \quad & \langle \boldsymbol{\mu}_{\pi_E}, \mathbf{T}_\gamma^* \mathbf{u} - \mathbf{c}_w \rangle \\ \text{s.t.} \quad & \mathbf{c}_w - \mathbf{T}_\gamma^* \mathbf{u} \geq \mathbf{0}, \\ & \mathbf{u} \in \mathbf{R}^{|\mathcal{X}|}, \quad \mathbf{w} \in \Delta_{[n_c]}. \end{aligned} \quad (\mathbf{D}_{\pi_E})$$

The inequality constraint in (\mathbf{D}_{π_E}) is a relaxation of the Bellman optimality conditions. Therefore, (\mathbf{D}_{π_E}) resembles

the primal LP for solving forward MDPs but its optimization variables are both value functions and cost weights.

Next, we derive optimality conditions for our LP formulations. We first assume that the expert is optimal for the true cost. In this case, the expert policy π_E is an optimal solution to (\mathbf{Lfd}_{π_E}) .

Lemma 1. *Assume that π_E is optimal for $(\mathbf{RL}_{\mathbf{c}_{\text{true}}})$. Then π_E is optimal for (\mathbf{Lfd}_{π_E}) with optimal value $\alpha^* = 0$. Equivalently, $(\boldsymbol{\mu}_{\pi_E}, 0)$ is optimal for (\mathbf{P}_{π_E}) .*

The following Proposition gives necessary and sufficient conditions for dual optimality. The proof is based on Lemma 1 (if π_E is optimal, then $\alpha^* = 0$) and complementary slackness conditions for the LP approach to forward MDPs.

Proposition 2 (Optimal expert). *Assume that π_E is optimal for $(\mathbf{RL}_{\mathbf{c}_{\text{true}}})$. A pair $(\mathbf{u}_A, \mathbf{w}_A)$ is optimal for (\mathbf{D}_{π_E}) if and only if $\mathbf{w}_A \in \Delta_{[n_c]}$, π_E is optimal for $(\mathbf{RL}_{\mathbf{c}_{\mathbf{w}_A}})$ and $\mathbf{u}_A = \mathbf{V}_{\mathbf{w}_A}^*$. In particular, $(\mathbf{V}_{\mathbf{w}_{\text{true}}}^*, \mathbf{w}_{\text{true}})$ is an optimal solution to (\mathbf{D}_{π_E}) .*

Proposition 2 states that, under the expert optimality assumption, the set of dual optimizers characterizes the set of solutions to the IRL problem, i.e, the set of costs in $\mathcal{C}_{\text{conv}}$ for which the expert is optimal. Indeed, a weight vector $\mathbf{w}_A \in \Delta_{[n_c]}$ is dual optimal if and only if the expert policy π_E is optimal for the forward RL problem with cost $\mathbf{c}_{\mathbf{w}_A}$. In this case, \mathbf{u}_A coincides with the corresponding optimal value function¹. In particular, the true weights \mathbf{w}_{true} and the true optimal value function $\mathbf{V}_{\mathbf{w}_{\text{true}}}^*$ are dual optimizers.

Is is worth noting that the linear program (\mathbf{D}_{π_E}) is different from the LP formulation of IRL proposed in (Ng & Russell, 2000; Komanduru & Honorio, 2019), which is based on dynamic programming principles. In contrast, our characterization of solutions to the inverse problem is based on duality arguments in the same line as (Ahuja & Orlin, 2001; Iyengar & Kang, 2005; Pauwels et al., 2016). Moreover, the dual LP formulation (\mathbf{D}_{π_E}) is independent of the complexity of the expert policy π_E , which can be even history-dependent.

In the general case of a suboptimal expert, we have the following necessary conditions for dual optimality.

Proposition 3 (Suboptimal expert). *If $\boldsymbol{\mu}_A$ is optimal for (\mathbf{P}_{π_E}) and $(\mathbf{u}_A, \mathbf{w}_A)$ is optimal for (\mathbf{D}_{π_E}) , then any apprentice policy π_A is optimal for $(\mathbf{RL}_{\mathbf{c}_{\mathbf{w}_A}})$ and $\mathbf{u}_A = \mathbf{V}_{\mathbf{w}_A}^*$.*

Proposition 3 states that, in general, the apprentice policy π_A is an optimal solution to the forward RL problem with cost $\mathbf{c}_{\mathbf{w}_A}$, where \mathbf{w}_A is an optimal dual variable. In addition, \mathbf{u}_A coincides with the corresponding optimal value function.

A similar result to Proposition 3 has been obtained in (Ho

¹To be precise, this is the case if $\boldsymbol{\nu}_0 \in \mathbf{R}_{++}^{|\mathcal{X}|}$, otherwise they coincide $\boldsymbol{\nu}_0$ -almost surely.

& Ermon, 2016) for the convex-concave formulation of maximum causal entropy IRL, by using the Sion minimax theorem (Sion, 1958).

3.2. Saddle-Point Formulation

A serious limitation of the primal and dual LP formulations in Section 3.1 is that they have an intractable number of constraints, which in addition may be not satisfied when working with function approximators. To mitigate this difficulty, we propose a more tractable unconstrained formulation by using the Lagrangian

$$\mathcal{L}(\boldsymbol{\mu}, \mathbf{u}, \mathbf{w}) = \langle \boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_w - \mathbf{T}_\gamma^* \mathbf{u} \rangle.$$

We can then frame the problem as a bilinear saddle-point problem

$$\alpha^* = \min_{\boldsymbol{\mu} \in \Delta_{\mathcal{X} \times \mathcal{A}}} \max_{\substack{\mathbf{u} \in \mathcal{U} \\ \mathbf{w} \in \Delta_{[n_c]}}} \mathcal{L}(\boldsymbol{\mu}, \mathbf{u}, \mathbf{w}), \quad (1)$$

where $\mathcal{U} \triangleq \{\mathbf{u} \in \mathbf{R}^{|\mathcal{X}|} \mid \|\mathbf{u}\|_\infty \leq 1\}$. Note that we have added extra constraints $\boldsymbol{\mu} \in \Delta_{\mathcal{X} \times \mathcal{A}}$ and $\mathbf{u} \in \mathcal{U}$ for the sake of analysis. These constraints do not change the problem optimality, but will considerably accelerate the convergence of the algorithm by considering smaller domains.

Indeed, note that the constraint $\boldsymbol{\mu} \in \Delta_{\mathcal{X} \times \mathcal{A}}$ is redundant since it is satisfied for all primal feasible solutions. Moreover, by Proposition 3, for all dual optimizers $(\mathbf{u}_A, \mathbf{w}_A)$ it holds that $\mathbf{u}_A = \mathbf{V}_{\mathbf{w}_A}^*$. Thus, we obtain the bound $\|\mathbf{u}_A\|_\infty \leq 1$.

The next Corollary follows from Propositions 2–3.

Corollary 1. *Suppose that the expert is optimal for \mathbf{c}_{true} . Then, $(\boldsymbol{\mu}_{\pi_E}, \mathbf{V}_{\mathbf{c}_{\text{true}}}^*, \mathbf{w}_{\text{true}})$ is a saddle-point to the minimax problem (1). In the general case, for a saddle-point $(\boldsymbol{\mu}_A, \mathbf{u}_A, \mathbf{w}_A)$ it holds that: (i) $\pi_{\boldsymbol{\mu}_A}$ is optimal for (\mathbf{Lfd}_{π_E}) , and (ii) $\pi_{\boldsymbol{\mu}_A}$ is optimal for $\mathbf{c}_{\mathbf{w}_A}$ and $\mathbf{u}_A = \mathbf{V}_{\mathbf{w}_A}^*$.*

4. A Linearly-Relaxed Saddle-Point Problem

Optimizing directly over $\boldsymbol{\mu}$ and \mathbf{u} is impractical since their dimensions scale linearly with the size of the state space. We reduce the order of complexity by limiting our search to linear subspaces defined by a small number of features. In particular, we consider the feature matrices $\Phi \in \mathbf{R}^{|\mathcal{X}| \times |\mathcal{A}| \times n_\mu}$ and $\Psi \in \mathbf{R}^{|\mathcal{X}| \times n_u}$. We then assume that the decision variables are parameterized in the form $\boldsymbol{\mu}_\theta = \Phi \boldsymbol{\theta}$ and $\mathbf{u}_\lambda = \Psi \boldsymbol{\lambda}$, where $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\lambda} \in \Lambda$ are the parameters to learn. Moreover, Θ and Λ are appropriately chosen parameter sets. We denote $\mathcal{W} \triangleq \Delta_{[n_c]}$ for brevity.

The corresponding *linearly-relaxed* saddle-point formulation is

$$\min_{\boldsymbol{\theta} \in \Theta} \max_{\substack{\boldsymbol{\lambda} \in \Lambda \\ \mathbf{w} \in \mathcal{W}}} \mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}), \quad (2)$$

where $\mathcal{L}_r: \Theta \times \Lambda \times \mathcal{W} \rightarrow \mathbf{R}$ is the *reduced* bilinear Lagrangian $\mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \triangleq \langle \boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_w - \mathbf{T}_\gamma^* \mathbf{u}_\lambda \rangle$.

We measure the quality of an approximate saddle-point $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w})$ by its *saddle-point residual (SPR)* defined as

$$\epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) = \max_{\substack{\boldsymbol{\lambda}' \in \Lambda \\ \mathbf{w}' \in \mathcal{W}}} \mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}', \mathbf{w}') - \min_{\boldsymbol{\theta}' \in \Theta} \mathcal{L}_r(\boldsymbol{\theta}', \boldsymbol{\lambda}, \mathbf{w}). \quad (3)$$

We assume that every column of Φ belongs to $\Delta_{\mathcal{X} \times \mathcal{A}}$, and every column of Ψ belongs to \mathcal{U} . These conditions ensure that if we choose

$$\Theta \triangleq \Delta_{[n_\mu]}, \quad \Lambda \triangleq \{\boldsymbol{\lambda} \in \mathbf{R}^{n_u} \mid \|\boldsymbol{\lambda}\|_2 \leq \frac{\beta}{\sqrt{n_u}}\}, \quad \beta \geq 2,$$

then $\boldsymbol{\mu}_\theta \in \Delta_{\mathcal{X} \times \mathcal{A}}$ and $\|\mathbf{u}_\lambda\|_\infty \leq \beta$, for all $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\lambda} \in \Lambda$. Note that we have enlarged the original optimization domain for the \mathbf{u} variable. In addition, by the definition of $\mathcal{C}_{\text{conv}}$, we get $\|\mathbf{c}_w\|_\infty \leq 1$, for all $\mathbf{w} \in \mathcal{W}$. These bounds will be used in the sequel in the algorithm design and its theoretical analysis.

4.1. From Approximate Saddle Points to Optimal Policies

A crucial part in our setting is to connect the saddle-point residual $\epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w})$ of a feasible solution $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w})$ for the linearly-relaxed saddle-point problem (2) to the suboptimality gap $\delta_{\mathcal{C}}(\pi_{\boldsymbol{\mu}_\theta}, \pi_E) - \delta_{\mathcal{C}}(\pi_A, \pi_E)$ of the induced policy $\pi_{\boldsymbol{\mu}_\theta}$. If small SPR implies small suboptimality gap, then we can first run any stochastic primal-dual algorithm with sublinear convergence rate to the objective in (2), and then convert the obtained approximate saddle-point to a nearly optimal policy for the original (\mathbf{Lfd}_{π_E}) problem.

For the following results, let π_A be an apprentice policy, i.e., π_A is optimal for (\mathbf{Lfd}_{π_E}) . The proofs can be found in Appendix B. Our first result makes no assumptions on the choice of features.

Proposition 4. *Let $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \in \Theta \times \Lambda \times \mathcal{W}$ be a feasible solution to (2). Set $\pi_\theta = \pi_{\boldsymbol{\mu}_\theta}$. It then holds that*

$$\delta_{\mathcal{C}}(\pi_\theta, \pi_E) - \delta_{\mathcal{C}}(\pi_A, \pi_E) \leq \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) + \epsilon_{\text{approx}, \boldsymbol{\theta}},$$

where the approximation error $\epsilon_{\text{approx}, \boldsymbol{\theta}}$ is no larger than $\frac{2}{1-\gamma} (\beta \min_{\boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\mu}_{\boldsymbol{\theta}'} - \boldsymbol{\mu}_{\pi_A}\|_1 + \min_{\boldsymbol{\lambda}' \in \Lambda} \|\mathbf{u}_{\boldsymbol{\lambda}'} - \mathbf{V}_{\mathbf{c}_{i_\theta}}^{\pi_\theta}\|_\infty)$, and $i_\theta \triangleq \arg \max_{i \in [n_c]} (\rho_{\mathbf{c}_i}(\pi_\theta) - \rho_{\mathbf{c}_i}(\pi_E))$. In particular, $\rho_{\mathbf{c}_{\text{true}}}(\pi_\theta) - \rho_{\mathbf{c}_{\text{true}}}(\pi_E) \leq \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) + \epsilon_{\text{approx}, \boldsymbol{\theta}} + \alpha^*$.

By Proposition 4, if $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w})$ is an (ϵ, δ) -optimal saddle point to (2), i.e., $\epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \leq \epsilon$ with probability at least $1 - \delta$, then the associated induced policy π_θ is $(\epsilon + \epsilon_{\text{approx}, \boldsymbol{\theta}} + \alpha^*)$ -optimal for the original (\mathbf{Lfd}_{π_E}) problem with high probability. Recall, that α^* is always zero or negative. The term $\epsilon_{\text{approx}, \boldsymbol{\theta}}$ is a measure of expressiveness

of the linear function approximators. The approximation error $\varepsilon_{\text{approx},\theta}$ depends on how well $\{\boldsymbol{\mu}_\theta : \theta \in \Theta\}$ approximates the apprentice occupancy measure $\boldsymbol{\mu}_{\pi_\Lambda}$ and how well $\{\mathbf{u}_\lambda : \lambda \in \Lambda\}$ approximates the value function $\mathbf{V}_{\mathbf{c}_{i_\theta}}^{\pi_\theta}$ of the extracted policy π_θ under the cost \mathbf{c}_{i_θ} . Note, however, that $\mathbf{V}_{\mathbf{c}_{i_\theta}}^{\pi_\theta}$ is not fixed before the learning process. Therefore, in order to guarantee a priori low approximation error, we need the optimal occupancy measure to be accurately representable by $\{\boldsymbol{\mu}_\theta : \theta \in \Theta\}$, while $\{\mathbf{u}_\lambda : \lambda \in \Lambda\}$ is required to accurately represent the value functions of non-optimal policies as well, under costs in $\{\mathbf{c}_i : i \in [n_c]\}$. Borrowing the terminology from (Shariff & Szepesvári, 2020), we require the occupancy measure features to be *weak*, while the value function features to be *strong*.

An open question is whether we can relax the previously described notion of *good features* for the value function approximation (Shariff & Szepesvári, 2020). We provide the following result in this direction. However, the analysis requires stronger conditions on the choice of features. In particular, we introduce the following assumption first studied in (Bas-Serrano & Neu, 2020).

Assumption 1 (Coherence Assumption). *The columns of Ψ are well-conditioned in the following sense: for every $\theta \in \Theta$ and $\mathbf{u} \in \mathbf{R}^{|\mathcal{X}|}$ with $\|\mathbf{u}\|_\infty \leq 2$, there exists $\lambda \in \Lambda$ such that $\langle \boldsymbol{\nu}_0 - \mathbf{T}_\gamma \boldsymbol{\mu}_\theta, \mathbf{u} - \mathbf{u}_\lambda \rangle = 0$.*

Proposition 5. *Let Assumption 1 hold. Let $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \in \Theta \times \Lambda \times \mathcal{W}$ be a feasible solution to (2). Set $\pi_\theta = \pi_{\boldsymbol{\mu}_\theta}$. It then holds that*

$$\rho_{\mathbf{c}_{\text{true}}}(\pi_\theta) - \rho_{\mathbf{c}_{\text{true}}}(\pi_E) \leq 3\varepsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) + \varepsilon_{\text{approx},w} + \alpha^*,$$

where the weak approximation error $\varepsilon_{\text{approx},w}$ is no larger than $\frac{2}{1-\gamma} \left(\beta \min_{\theta' \in \Theta} \|\boldsymbol{\mu}_{\theta'} - \boldsymbol{\mu}_{\pi_\Lambda}\|_1 + 2\|\mathbf{V}_{\mathbf{w}_A}^* - \Psi \boldsymbol{\lambda}^*\|_\infty \right)$, where $(\mathbf{V}_{\mathbf{w}_A}^*, \mathbf{w}_A)$ is dual optimal for (\mathbf{D}_{π_E}) and $\boldsymbol{\lambda}^*$ is a dual optimizer of (2).

Assumption 1 is satisfied when the rows of Ψ are orthonormal. As it is apparent from the proof of Proposition 5, Assumption 1 controls the *distribution mismatch* between $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\mu}_{\pi_\theta}$. In general $\boldsymbol{\mu}_\theta$ is not necessarily an occupancy measure generated by a policy. The term $\|\mathbf{T}_\gamma \boldsymbol{\mu}_\theta - \boldsymbol{\nu}_0\|_1$, i.e., the violation degree of the Bellman flow constraints is related to the quality of the extracted policy π_θ . Assumption 1 ensures that an approximate saddle point with small violation degree of the aggregated Bellman flow constraints $\Psi^\top(\mathbf{T}_\gamma \boldsymbol{\mu}_\theta - \boldsymbol{\nu}_0) = 0$, has also small violation degree for the original Bellman flow constraints and thus produces a *good* policy.

The approximation error term $\|\mathbf{V}_{\mathbf{w}_A}^* - \Psi \boldsymbol{\lambda}^*\|_\infty$ can be bounded by approximate dynamic programming techniques (De Fariás & Van Roy, 2003; 2004; Shariff & Szepesvári, 2020). For example, under the *realizability assumption* (Chen et al., 2018; Bas-Serrano et al., 2020), we

get that $\varepsilon_{\text{approx},w} = 0$. This is formally stated in Lemma 6 in Appendix B.

Moreover, assuming that there exists a set of *core* states whose features span those of other states, the weak value function features ensure a low approximation error $\|\mathbf{V}_{\mathbf{w}_A}^* - \Psi \boldsymbol{\lambda}^*\|$ (Lakshminarayanan et al., 2018; Shariff & Szepesvári, 2020).

5. Algorithm and Finite-Sample Analysis

After the analysis of our saddle-point setup, the aim of this section is (i) to provide a computationally efficient stochastic primal-dual algorithm whose iteration and sample complexities do not grow with the size of state and action spaces, and (ii) to obtain explicit probabilistic performance bounds on the quality of the extracted policy with respect to the unknown true cost function.

5.1. Stochastic Primal-Dual LfD Algorithm

Having formulated the linearly-relaxed saddle-point problem (2) with a few variables and constraints, we will now propose an iterative stochastic approximation algorithm for the batch LfD problem. Assuming access to a generative-model oracle and a finite set of expert demonstrations, we propose a stochastic mirror descent primal-dual algorithm which keeps the advantages of simplicity of implementation, low memory requirements, and low computational complexity.

We will consider the slightly modified Lagrangian

$$\begin{aligned} \bar{\mathcal{L}}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \triangleq & \langle \boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_w - \mathbf{T}_\gamma^* \mathbf{u}_\lambda \rangle \\ & + C_{\beta,\gamma} \langle \boldsymbol{\mu}_\theta, \mathbf{1} \rangle - 2 \langle \mathbf{w}, \mathbf{1} \rangle, \end{aligned}$$

where $C_{\beta,\gamma} \triangleq 2\beta/(1-\gamma)$. Note that $\bar{\mathcal{L}}_r$ differs from \mathcal{L}_r up to a constant, since $\langle \boldsymbol{\mu}_\theta, \mathbf{1} \rangle = \langle \mathbf{w}, \mathbf{1} \rangle = 1$, for all $\theta \in \Theta$ and $\mathbf{w} \in \mathcal{W}$. Thus, using $\bar{\mathcal{L}}_r$ instead of \mathcal{L}_r does not change the saddle-point residuals. In addition, we have that $\nabla_{\boldsymbol{\theta}} \bar{\mathcal{L}}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \geq \mathbf{0}$ and $\nabla_{\mathbf{w}} \bar{\mathcal{L}}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \leq \mathbf{0}$, for all $\theta \in \Theta$, $\lambda \in \Lambda$ and $\mathbf{w} \in \mathcal{W}$. The same is true for their stochastic gradients. We refer to Sec. 2.8 in (Shalev-Shwartz, 2012) for the theory behind this technique. We will need this property in order to derive finer error bounds by using local norm arguments. For a more detailed discussion, see Section 5.2.

In the forthcoming material we use the following notation. The cost matrix is denoted by $\mathbf{C} \triangleq [\mathbf{c}_1 \ \dots \ \mathbf{c}_{n_c}]$ and the feature matrices by $\Phi \triangleq [\phi_1 \ \dots \ \phi_{n_\mu}]$ and $\Psi \triangleq [\psi_1 \ \dots \ \psi_{n_u}]$. The (x, a) -th row of \mathbf{C} is denoted by $\mathbf{c}_{(x,a)}$ and the x -th row of Ψ by ψ_x .

Note that both the expert policy π_E and the transition law P are unknown and they do appear in the reduced Lagrangians \mathcal{L}_r and $\bar{\mathcal{L}}_r$. We will now explain how to tackle this difficulty.

Algorithm 1 Stochastic Primal-Dual LfD

```

1: Input: cost matrix  $\mathbf{C}$ , feature matrices  $\Phi$  and  $\Psi$ 
2: Input: number of iterations  $N$ , step-size  $\eta$ , radius  $\beta$ 
3: Input: expert demonstrations  $\mathcal{D}_E^{m,H}$ , generative-model
4: Compute  $\rho_{\mathbf{C}}(\widehat{\pi}_E)$  using expert demonstrations.
5: Set  $\theta_{1,i} = \frac{1}{n_\mu}$ ,  $i \in [n_\mu]$ ,  $\mathbf{w}_{1,i} = \frac{1}{n_c}$ ,  $i \in [n_c]$ ,  $\lambda = \mathbf{0}$ 
6: for  $n = 1, \dots, N - 1$  do
7:   //  $\theta$  gradient estimation
8:   Sample  $i \sim \text{Unif}([n_\mu])$ ,  $(x, a) \sim \phi_i$ ,  $y \sim P(\cdot|x, a)$ 
9:   Set
      
$$g_{n,\theta,j} = \begin{cases} \frac{n_\mu \left( (1-\gamma) \mathbf{c}_{(x,a)}^\top \mathbf{w}_n - (\psi_{x-\gamma\psi_y})^\top \lambda_{n+2\beta} \right)}{1-\gamma}, & j = i \\ 0, & \text{otherwise} \end{cases}$$

10:  //  $\lambda$  gradient estimation
11:  Sample  $x' \sim \nu_0$ ,  $i \sim \theta_n$ ,  $(x, a) \sim \phi_i$ ,  $y \sim P(\cdot|x, a)$ 
12:  Set  $\mathbf{g}_{n,\lambda} = \psi_{x'} - \frac{\psi_{x-\gamma\psi_y}}{1-\gamma}$ 
13:  //  $\mathbf{w}$  gradient estimation
14:  Sample  $i \sim \theta_n$ ,  $(x, a) \sim \phi_i$ 
15:  Set  $\mathbf{g}_{n,\mathbf{w}} = \mathbf{c}_{(x,a)} - \rho_{\mathbf{C}}(\widehat{\pi}_E) - 2 \cdot \mathbf{1}$ 
16:  // Stochastic mirror descent steps
17:  Update  $\theta_{n+1,j} \propto \theta_{n,j} e^{(-\eta g_{n,\theta,j})}$ ,  $j \in [n_\mu]$ 
18:  Update  $\lambda_{n+1} = \Pi_\Lambda(\lambda_n + \frac{\eta \beta^2}{n_u} \mathbf{g}_{n,\lambda})$ 
19:  Update  $w_{n+1,j} \propto w_{n,j} e^{(\eta g_{n,\mathbf{w},j})}$ ,  $j \in [n_c]$ 
20: end for
21: Set  $\hat{\theta}_N = \frac{1}{N} \sum_{n=1}^N \theta_n$ 
22: Output:  $\hat{\pi}_N = \pi_{\Phi \hat{\theta}_N}$ 

```

We define the *feature expectation* vector $\rho_{\mathbf{C}}(\pi_E) \in \mathbf{R}^{n_c}$ of the expert policy π_E by $\rho_{\mathbf{C}}(\pi_E) \triangleq (\rho_{\mathbf{c}_1}(\pi_E), \dots, \rho_{\mathbf{c}_{n_c}}(\pi_E))^\top$. Since in practice we do not have access to the whole policy π_E , but instead can observe a finite set of i.i.d. sample trajectories $\mathcal{D}_{\pi_E}^{m,H} \triangleq \{(x_0^k, a_0^k, x_1^k, a_1^k, \dots, x_H^k, a_H^k)\}_{k=1}^m \sim \pi_E$, we consider the empirical feature expectation vector $\rho_{\mathbf{C}}(\widehat{\pi}_E)$ by taking sample averages, i.e., for each $i = 1, \dots, n_c$, $\rho_{\mathbf{c}_i}(\widehat{\pi}_E) \triangleq (1-\gamma) \frac{1}{m} \sum_{t=0}^H \sum_{j=1}^m \gamma^t c_i(x_t^j, a_t^j)$. So the final empirical Lagrangian that our algorithm optimizes is given by

$$\widehat{\mathcal{L}}_r(\theta, \lambda, \mathbf{w}) \triangleq \langle \mu_\theta, \mathbf{c}_w - \mathbf{T}_\gamma^* \mathbf{u}_\lambda + C_{\beta,\gamma} \cdot \mathbf{1} \rangle + \langle \nu_0, \mathbf{u}_\lambda \rangle - \langle \mathbf{w}, \rho_{\mathbf{C}}(\widehat{\pi}_E) + 2 \cdot \mathbf{1} \rangle.$$

Our optimization variable is $\mathbf{z} = (\theta, \lambda, \mathbf{w}) \in \mathbf{Z}$, where $\mathbf{Z} \triangleq \Theta \times \Lambda \times \mathcal{W}$ is the decision space. The monotone operator $\mathbf{G}(\mathbf{z}) \triangleq \left[\nabla_\theta \widehat{\mathcal{L}}_r(\mathbf{z})^\top \quad -\nabla_\lambda \widehat{\mathcal{L}}_r(\mathbf{z})^\top \quad -\nabla_{\mathbf{w}} \widehat{\mathcal{L}}_r(\mathbf{z})^\top \right]^\top$ is given by

$$\mathbf{G}(\mathbf{z}) = \begin{pmatrix} \Phi^\top (\mathbf{c}_w - \mathbf{T}_\gamma^* \mathbf{u}_\lambda + C_{\beta,\gamma} \cdot \mathbf{1}) \\ \Psi^\top (\mathbf{T}_\gamma \mu_\theta - \nu_0) \\ \rho_{\mathbf{C}}(\widehat{\pi}_E) - \mathbf{C}^\top \mu_\theta + 2 \cdot \mathbf{1} \end{pmatrix}.$$

Note that when \mathbf{u}_λ is the value function of a policy, then the term $\mathbf{c}_w - \mathbf{T}_\gamma^* \mathbf{u}_\lambda$ is the corresponding *advantage function*.

Moreover, in general μ_θ is not necessarily an occupancy generated by a policy. The term $\|\Psi^\top (\mathbf{T}_\gamma \mu_\theta - \nu_0)\|_1$ is the violation degree of the aggregated Bellman flow constraints.

Although the dynamics are unknown, by having at our disposal the generative-model oracle we can compute cheap unbiased gradient estimates. Even in the case of known dynamics, we have expensive matrix computations and one can accelerate learning by randomization. The rationale behind the sampling procedure and the computation of the gradient estimates is similar to (Chen et al., 2018). Indeed, note that we can equivalently consider a stochastic saddle-point formulation, by writing the empirical Lagrangian $\widehat{\mathcal{L}}_r$ in the following sample-friendly form

$$\begin{aligned} \widehat{\mathcal{L}}_r(\mathbf{z}) &= \sum_{j=1}^{n_\mu} \theta_j \mathbf{E}_{\substack{(x,a) \sim \phi_j \\ y \sim P(\cdot|x,a)}} \left[\frac{(1-\gamma) \mathbf{c}_{(x,a)}^\top \mathbf{w} - (\psi_{x-\gamma\psi_y})^\top \lambda + 2\beta}{1-\gamma} \right] \\ &\quad + \mathbf{E}_{x \sim \nu_0} [\psi_x^\top \lambda] - \langle \mathbf{w}, \rho_{\mathbf{C}}(\widehat{\pi}_E) + 2 \cdot \mathbf{1} \rangle. \end{aligned}$$

In particular, in round n we make three independent calls of the generative oracle and compute the unbiased stochastic gradient $\mathbf{g}_n = [\mathbf{g}_{n,\theta}^\top \quad -\mathbf{g}_{n,\lambda}^\top \quad -\mathbf{g}_{n,\mathbf{w}}^\top]^\top$ of $\mathbf{G}(\mathbf{z}_n)$, as described in Algorithm 1. We defer the details to Appendix C.

Note that each iteration costs $\mathcal{O}(1)$ sample generation. Therefore the iteration and the sample complexity of the algorithm coincide. Moreover, the computation of gradient with respect to λ and \mathbf{w} is $\mathcal{O}(n_u)$ and $\mathcal{O}(n_c)$, respectively. On the other hand the computation of gradient with respect to μ is $\mathcal{O}(1)$ since only one coordinate is updated per iteration.

The update rule of primal-dual mirror descent is given by

$$\mathbf{z}_{n+1} = \arg \min_{\mathbf{z} \in \mathbf{Z}} \left(\langle \mathbf{g}_n, \mathbf{z} \rangle + \frac{1}{\eta} B_R(\mathbf{z} \| \mathbf{z}_n) \right),$$

where R is the following distance-generating function

$$R(\mathbf{z}) = \sum_{i=1}^{n_\mu} \theta_i \log \theta_i + \frac{n_u}{2\beta^2} \|\lambda\|_2^2 + \sum_{i=1}^{n_c} w_i \log w_i.$$

The choice of Shannon entropy for the variables μ and \mathbf{w} which live in the probability simplex mitigates the effects of dimension. Moreover, the factor $\frac{n_u}{2\beta^2}$ is chosen to make the size of Bregman divergence dimension-free (Cheng et al., 2020).

The analytical form of the updates can be seen in Algorithm 1. Note once more that the offsets considered in the modified Lagrangian have no effect since the gradients are exp-transformed and the resulting distribution normalized. We denote by $\widehat{\mathbf{z}}_N \triangleq \frac{1}{N} \sum_{n=1}^N \mathbf{z}_n$ the iterate average after N iterations. Then, the output of the algorithm is the policy $\widehat{\pi}_N \triangleq \pi_{\mu_{\widehat{\mathbf{z}}_N}}$.

5.2. Sample Complexity

The following theorem gives the sample complexity of Algorithm 1. The proof can be found in Appendix D and is based on high-confidence regret bounds of mirror descent with local norm arguments (Carmon et al., 2019; Jin & Sidford, 2020).

In particular, by combining the mirror descent bounds with Propositions 4–5 we get the following theorem.

Theorem 1. *Let $\hat{\pi}_N$ be the output of running Algorithm 1 for $N = \max \left\{ \mathcal{O} \left(\frac{\beta^2 n_\mu \log(\frac{1}{\delta})}{(1-\gamma)^2 \varepsilon^2} \right), \mathcal{O} \left(\frac{\beta \sqrt{n_\mu^3 \log(\frac{1}{\delta})}}{(1-\gamma)\varepsilon} \right) \right\}$ iterations, with $m = \frac{8 \log(\frac{4n_c}{\delta})}{\varepsilon^2}$ expert trajectories of length $H = \frac{1}{1-\gamma} \log(\frac{2}{\varepsilon})$, and learning rate $\eta = \frac{1-\gamma}{\beta \sqrt{N n_\mu}}$. Then, with probability $1 - \delta$ it holds that $\rho_{\mathbf{c}_{\text{true}}}(\hat{\pi}_N) - \rho_{\mathbf{c}_{\text{true}}}(\pi_E) \leq \varepsilon + \varepsilon_{\text{approx}, \hat{\theta}_N} + \alpha^*$. If in addition Assumption 1 is satisfied, then with probability $1 - \delta$, it holds that $\rho_{\mathbf{c}_{\text{true}}}(\hat{\pi}_N) - \rho_{\mathbf{c}_{\text{true}}}(\pi_E) \leq \varepsilon + \varepsilon_{\text{approx}, w} + \alpha^*$.*

Note that we do not have full access to the true dynamics and the expert policy but instead we can query the generative oracle and observe a batch finite set of truncated expert demonstrations. In our stochastic algorithm this is depicted to the fact that we replace the expert feature expectation vector $\rho_{\mathbf{C}}(\pi_E)$ by its empirical counterpart $\rho_{\mathbf{C}}(\hat{\pi}_E)$, and we do not consider the full gradient vector but instead we use unbiased gradient estimates. The estimation error is quantified in terms of concentration inequalities. Indeed for the sample average of the expert feature expectation vector we use a variant of the Hoeffding’s inequality (Syed & Schapire, 2007), and for the error due to the stochastic gradient we choose appropriate martingale concentration inequalities (McDiarmid, 1998).

By using the modified Lagrangian $\tilde{\mathcal{L}}_r$ and making the unbiased gradient estimates of $\theta \in \Delta_{[n_\mu]}$ and $\mathbf{w} \in \Delta_{[n_c]}$ nonnegative/nonpositive, our final bounds have a better dimension dependency. This trick, also used in (Jin & Sidford, 2020; Cheng et al., 2020), allows us to attain more refined convergence guarantees by exploiting the low variance bounds under the corresponding local norms (as opposed to the $\|\cdot\|_\infty$ -norm). We refer to Sec. 2.8 in (Shalev-Shwartz, 2012) for the theory behind this technique. In our case, by replacing $\|\mathbf{g}_{n, \theta}\|_\infty^2$ with the local norm $\|\mathbf{g}_{n, \theta}\|_{\hat{\theta}_n}^2 \triangleq \sum_{i=1}^{n_\mu} \theta_{n, i} g_{n, \theta, i}^2$, we improve the regret bounds by a n_μ -factor.

The parameter β acts as a regularization in learning. If it is too small, the projection residuals $\min_{\lambda \in \Lambda} \|\mathbf{u}_\lambda - \mathbf{V}_{\mathbf{c}_\theta}^{\pi_\theta}\|_\infty$ and $\|\mathbf{V}_{\mathbf{w}_\Lambda}^* - \Psi \lambda^*\|_\infty$ in the approximation error terms $\varepsilon_{\text{approx}, \theta}$ and $\varepsilon_{\text{approx}, w}$, respectively, are bigger. If it is too large, the learning becomes slower.

In Appendix E we provide preliminary empirical results on

a simple tabular MDP in order to illustrate our formulations and theoretical results.

5.3. A No-Regret Online Learning View

In Appendix F, we argue that solving the online learning problem with $\mathbf{Z} = \Theta \times \Lambda \times \mathcal{W}$ and per-round loss function

$$\ell_n(\mathbf{z}) \triangleq \mathcal{L}_r(\theta, \lambda_n, \mathbf{w}_n) - \mathcal{L}_r(\theta_n, \lambda, \mathbf{w})$$

is equivalent to solving the linearly-relaxed saddle-point problem (2).

It is worth noting that this online learning approach differs from the one in (Ross et al., 2011) where interaction with the expert is required. It is also different from the game-theoretic approach in (Syed & Schapire, 2007) where the forward RL problem has to be solved repeatedly.

5.4. Linear Function Approximators

We are motivated to work with occupancy measures (OMs) μ instead of policies π because of linearity and flexibility of (\mathbf{P}_{π_E}) and (\mathbf{D}_{π_E}) . We highlight that Propositions 4–5 hold even for nonlinear parameterizations $\{\mu_\theta \mid \theta \in \Theta\}$ and $\{\mathbf{u}_\lambda \mid \lambda \in \Lambda\}$, as long as $\mu_\theta \in \mathcal{M}$ and $\|\mathbf{u}_\lambda\|_\infty \leq \beta$ hold. Thus, the LfD problem is decoupled in two parts: minimization of SPR and function approximation. For the first part, we employ modern large-scale stochastic optimization methods while the second part can be quantified independently of the learning process. We choose LFA to preserve the convexity-concavity of the Lagrangian and design a scalable algorithm with theoretical guarantees (sublinear convergence rate), though at a cost of a potential approximation bias. Indeed, LFA require a careful choice of features. One can use prior knowledge of the MDP to choose appropriate basis functions for the value functions, e.g., in our case the value function basis $\{\psi_i = \mathbf{V}_{\mathbf{c}_i}^* \mid i \in [n_c]\}$ satisfies the realizability assumption. On the other hand the choice of OM features is trickier since only a restrictive class of policies can be represented (Banijamali et al., 2019). One can use OM of policies extracted from “heuristic” methods or policies provided from multiple “cheaper” suboptimal experts (where one can sample a lot of rollouts) and use Algorithm 1 to improve upon them. We hope that our techniques will be useful for future algorithm designers and will lay foundations for more exhaustive research in this direction. In Appendix G we point out a few interesting directions.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme grant agreement OCAL, No. 787845.

References

- Abbasi-Yadkori, Y., Bartlett, P. L., and Malek, A. Linear programming for large-scale Markov decision problems. In *International Conference on Machine Learning (ICML)*, 2014.
- Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2004.
- Abbeel, P., Dolgov, D., Ng, A. Y., and Thrun, S. Apprenticeship learning for motion planning with application to parking lot navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2008.
- Ahuja, R. K. and Orlin, J. B. Inverse optimization. *Operations Research*, 49(5):771–783, 2001.
- Bagnell, J. A. D. An invitation to imitation. Technical Report CMU-RI-TR-15-08, Carnegie Mellon University, Pittsburgh, PA, March 2015.
- Banijamali, E., Abbasi-Yadkori, Y., Ghavamzadeh, M., and Vlassis, N. Optimizing over a restricted policy class in MDPs. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.
- Banjac, G. and Lygeros, J. A data-driven policy iteration scheme based on linear programming. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 816–821, 2019.
- Bas-Serrano, J. and Neu, G. Faster saddle-point optimization for solving large-scale Markov decision processes. In *Conference on Learning for Dynamics and Control (LADC)*, 2020.
- Bas-Serrano, J., Curi, S., Krause, A., and Neu, G. Logistic Q -learning. *arXiv:2010.11151*, 2020.
- Bertsekas, D. P. *Dynamic Programming and Optimal Control: Approximate Dynamic Programming, Vol. II*. Athena Scientific, 4th edition, 2012.
- Bertsekas, D. P. and Tsitsiklis, J. *Neuro-Dynamic Programming*. Athena Scientific, 1996.
- Beuchat, P. N., Georghiou, A., and Lygeros, J. Performance guarantees for model-based approximate dynamic programming in continuous spaces. *IEEE Transactions on Automatic Control*, 65(1):143–158, 2020.
- Borkar, V. S. A convex analytic approach to Markov decision processes. *Probability Theory and Related Fields*, 78(4):583–602, 1988.
- Boyd, N., Schiebinger, G., and Recht, B. The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27(2):616–639, 2017.
- Brown, D., Niekum, S., and Petrik, M. Bayesian robust optimization for imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a.
- Brown, D. S., Coleman, R., Srinivasan, R., and Niekum, S. Safe imitation learning via fast Bayesian reward inference from preferences. In *International Conference on Machine Learning (ICML)*, 2020b.
- Cai, Q., Hong, M., Chen, Y., and Wang, Z. On the global convergence of imitation learning: a case for linear quadratic regulator. *arXiv:1901.03674*, 2019.
- Carmon, Y., Jin, Y., Sidford, A., and Tian, K. Variance reduction for matrix games. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- Chen, M., Wang, Y., Liu, T., Yang, Z., Li, X., Wang, Z., and Zhao, T. On computation and generalization of generative adversarial imitation learning. *International Conference on Learning Representations (ICLR)*, 2020a.
- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020b.
- Chen, Y., Li, L., and Wang, M. Scalable bilinear π learning using state and action features. In *International Conference on Machine Learning (ICML)*, 2018.
- Chen, Y. F., Everett, M., Liu, M., and How, J. P. Socially aware motion planning with deep reinforcement learning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.
- Cheng, C.-A., des Combes, R. T., Boots, B., and Gordon, G. A reduction from reinforcement learning to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.
- De Farias, D. P. and Van Roy, B. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.
- De Farias, D. P. and Van Roy, B. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, 29(3):462–478, 2004.
- Dufour, F. and Prieto-Rumeau, T. Finite linear programming approximations of constrained discounted Markov decision processes. *SIAM Journal on Control and Optimization*, 51(2):1298–1324, 2013.

- Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning (ICML)*, 2018.
- Ghasemipour, S. K. S., Zemel, R., and Gu, S. A divergence minimization perspective on imitation learning methods. In *Proceedings of the Conference on Robot Learning, Proceedings of Machine Learning Research*, pp. 1259–1277. PMLR, 2020.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
- Hazan, E. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.
- Hernández-Lerma, O. and Lasserre, J. B. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer-Verlag New York, 1996.
- Hernández-Lerma, O., González-Hernández, J., and López-Martínez, R. Constrained average cost Markov control processes in Borel spaces. *SIAM Journal on Control and Optimization*, 42(2):442–468, 2003.
- Ho, J. and Ermon, S. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- Ho, J., Gupta, J. K., and Ermon, S. Model-free imitation learning with policy optimization. In *International Conference on Machine Learning (ICML)*, 2016.
- Iyengar, G. and Kang, W. Inverse conic programming with applications. *Operations Research Letters*, 33(3):319–330, 2005.
- Jin, Y. and Sidford, A. Efficiently solving MDPs with stochastic mirror descent. In *International Conference on Machine Learning (ICML)*, 2020.
- Komanduru, A. and Honorio, J. On the correctness and sample complexity of inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- Kostrikov, I., Nachum, O., and Tompson, J. Imitation learning via off-policy distribution matching. In *International Conference on Learning Representations*, 2020.
- Lakshminarayanan, C., Bhatnagar, S., and Szepesvári, C. A linearly relaxed approximate linear program for Markov decision processes. *IEEE Transactions on Automatic Control*, 63(4):1185–1191, 2018.
- Lee, D. and He, N. Stochastic primal-dual Q-learning algorithm for discounted MDPs. In *American Control Conference (ACC)*, 2019.
- Levine, S., Popović, Z., and Koltun, V. Feature construction for inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2010.
- Levine, S., Popović, Z., and Koltun, V. Nonlinear inverse reinforcement learning with Gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.
- Manne, A. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960.
- Martinelli, A., Gargiani, M., and Lygeros, J. Data-driven optimal control with a relaxed linear program, 2020.
- McDiarmid, C. *Concentration*, pp. 195–248. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.
- Miryoosefi, S., Brantley, K., Daume III, H., Dudik, M., and Schapire, R. E. Reinforcement learning with convex constraints. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Hiedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, February 2015.
- Mohajerin Esfahani, P., Sutter, T., Kuhn, D., and Lygeros, J. From infinite to finite programs: explicit error bounds with applications to approximate dynamic programming. *SIAM Journal on Optimization*, 28(3):1968–1998, 2018.
- Nachum, O. and Dai, B. Reinforcement learning via Fenchel-Rockafellar duality. *arXiv:2001.01866*, 2020.
- Nachum, O., Chow, Y., Dai, B., and Li, L. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, volume 32, 2019a.
- Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. Algaedice: Policy gradient from arbitrary experience, 2019b.
- Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

- Neu, G. and Szepesvári, C. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.
- Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2000.
- Pauwels, E., Henrion, D., and Lasserre, J.-B. Linear conic optimization for inverse optimal control. *SIAM Journal on Control and Optimization*, 54(3):1798–1825, 2016.
- Pomerleau, D. A. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991.
- Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994.
- Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. A. Maximum margin planning. In *International Conference on Machine Learning (ICML)*, 2006.
- Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.
- Russell, S. Learning agents for uncertain environments (extended abstract). In *Annual Conference on Computational Learning Theory (COLT)*, 1998.
- Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, M. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, 2015.
- Shafieepoorfard, E., Raginsky, M., and Meyn, S. P. Rational inattention in controlled Markov processes. In *American Control Conference (ACC)*, 2013.
- Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.
- Shariff, R. and Szepesvári, C. Efficient planning in large MDPs with weak linear function approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.
- Sutter, T., Kamoutsi, A., Esfahani, P. E., and Lygeros, J. Data-driven approximate dynamic programming: A linear programming approach. In *IEEE Conference on Decision and Control (CDC)*, 2017.
- Sutton, R. S. and Barto, A. G. *Introduction to Reinforcement Learning*. MIT Press, 1st edition, 1998.
- Syed, U. and Schapire, R. E. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.
- Syed, U., Bowling, M., and Schapire, R. Apprenticeship learning using linear programming. In *International Conference on Machine Learning (ICML)*, 2008.
- Szörényi, B., Kedenburg, G., and Munos, R. Optimistic planning in Markov decision processes using a generative model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
- Taleghan, M. A., Dietterich, T. G., Crowley, M., Hall, K., and Albers, H. J. PAC optimal MDP planning with application to invasive species management. *Journal of Machine Learning Research*, 16(117):3877–3903, 2015.
- Tesauro, G. and Kephart, J. O. Pricing in agent economies using multi-agent Q-learning. *Autonomous Agents and Multi-Agent Systems*, 5(3):289–304, 2002.
- Vamvoudakis, K. G. and Kokolakis, N.-M. T. Synchronous reinforcement learning-based control for cognitive autonomy. *Foundations and Trends in Systems and Control*, 8(1–2):1–175, 2020.
- Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2020.
- Wang, M. Randomized linear programming solves the Markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*, 45(2):517–546, 2020.
- Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.
- Zhang, Y., Cai, Q., Yang, Z., and Wang, Z. Generative adversarial imitation learning with neural network parameterization: global optimality and convergence rate. In *International Conference on Machine Learning (ICML)*, 2020.
- Ziebart, B. D., Maas, A., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In *National Conference on Artificial Intelligence (AAAI)*, 2008.

A. Derivations and Proofs of Section 3

For the following results, we will assume that $\nu_0 \in \mathbf{R}_{++}^{|\mathcal{X}|}$. We also set $\mathcal{W} \triangleq \Delta_{[n_c]}$.

A.1. Primal LP formulation (\mathbf{P}_{π_E})

We have

$$\begin{aligned}
 \alpha^* &= \min_{\pi} \max_{c \in \mathcal{C}} \rho_c(\pi) - \rho_c(\pi_E) && \text{(definition of } (\mathbf{Lfd}_{\pi_E}) \text{)} \\
 &= \min_{\pi} \max_{i \in [n_c]} \rho_{c_i}(\pi) - \rho_{c_i}(\pi_E) && \text{(since } \mathcal{C} = \mathcal{C}_{\text{conv}} \text{)} \\
 &= \min_{\pi} \max_{i \in [n_c]} \langle \boldsymbol{\mu}_{\pi} - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_i \rangle \\
 &= \min_{\boldsymbol{\mu} \in \mathfrak{F}} \max_{i \in [n_c]} \langle \boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_i \rangle && \text{(by Proposition 1)} \\
 &= \min_{\boldsymbol{\mu}, \varepsilon} \{ \varepsilon \mid \langle \boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_i \rangle \leq \varepsilon, i \in [n_c], \boldsymbol{\mu} \in \mathfrak{F} \}. && \text{(by an epigraphic transformation)}
 \end{aligned}$$

A.2. Dual LP formulation (\mathbf{D}_{π_E})

We have

$$\begin{aligned}
 \alpha^* &= \min_{\boldsymbol{\mu}, \varepsilon} \{ \varepsilon \mid \langle \boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_i \rangle \leq \varepsilon, i \in [n_c], \boldsymbol{\mu} \in \mathfrak{F} \} && \text{(definition of } (\mathbf{P}_{\pi_E}) \text{)} \\
 &= \min_{\boldsymbol{\mu} \geq \mathbf{0}, \varepsilon, \mathbf{w} \geq \mathbf{0}, \mathbf{u}} \varepsilon + \sum_{i=1}^{n_c} w_i (\langle \boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_i \rangle - \varepsilon) + \langle \boldsymbol{\nu}_0 - \mathbf{T}_{\gamma} \boldsymbol{\mu}, \mathbf{u} \rangle && \text{(introduce Lagrange multipliers } \mathbf{w} \text{ and } \mathbf{u} \text{)} \\
 &= \min_{\boldsymbol{\mu} \geq \mathbf{0}, \varepsilon, \mathbf{w} \geq \mathbf{0}, \mathbf{u}} \varepsilon \left(1 - \sum_{i=1}^{n_c} w_i \right) + \langle \boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_{\mathbf{w}} - \mathbf{T}_{\gamma}^* \mathbf{u} \rangle && \text{(since } \mathbf{T}_{\gamma} \boldsymbol{\mu}_{\pi_E} = \boldsymbol{\nu}_0 \text{)} \\
 &= \max_{\mathbf{w} \geq \mathbf{0}, \mathbf{u}} \min_{\boldsymbol{\mu} \geq \mathbf{0}, \varepsilon} \varepsilon \left(1 - \sum_{i=1}^{n_c} w_i \right) + \langle \boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_{\mathbf{w}} - \mathbf{T}_{\gamma}^* \mathbf{u} \rangle && \text{(linear duality)} \\
 &= \max_{\mathbf{w} \geq \mathbf{0}, \mathbf{u}} \left\{ \langle \boldsymbol{\mu}_{\pi_E}, \mathbf{T}_{\gamma}^* \mathbf{u} - \mathbf{c}_{\mathbf{w}} \rangle \mid \mathbf{c}_{\mathbf{w}} - \mathbf{T}_{\gamma}^* \mathbf{u} \geq \mathbf{0}, \sum_{i=1}^{n_c} w_i = 1 \right\}.
 \end{aligned}$$

A.3. Proof of Proposition 2

Proposition 2 (Optimal expert). *Assume that π_E is optimal for $(\mathbf{RL}_{\mathbf{c}_{\text{true}}})$. A pair $(\mathbf{u}_A, \mathbf{w}_A)$ is optimal for (\mathbf{D}_{π_E}) if and only if $\mathbf{w}_A \in \Delta_{[n_c]}$, π_E is optimal for $(\mathbf{RL}_{\mathbf{c}_{\mathbf{w}_A}})$ and $\mathbf{u}_A = \mathbf{V}_{\mathbf{w}_A}^*$. In particular, $(\mathbf{V}_{\mathbf{w}_{\text{true}}}^*, \mathbf{w}_{\text{true}})$ is an optimal solution to (\mathbf{D}_{π_E}) .*

Lemma 1. *Assume that π_E is optimal for $(\mathbf{RL}_{\mathbf{c}_{\text{true}}})$. Then π_E is optimal for (\mathbf{Lfd}_{π_E}) with optimal value $\alpha^* = 0$. Equivalently, $(\boldsymbol{\mu}_{\pi_E}, 0)$ is optimal for (\mathbf{P}_{π_E}) .*

Proof. It holds that $\alpha^* \leq \delta_{\mathcal{C}}(\pi_E, \pi_E) = 0$. Assume for the sake of contradiction that $\alpha^* < 0$. Then $\rho_{\mathbf{c}_{\text{true}}}(\pi_A) - \rho_{\mathbf{c}_{\text{true}}}(\pi_E) \leq \delta_{\mathcal{C}}(\pi_A, \pi_E) = \alpha^* < 0$, which contradicts the assumption of optimality of π_E with respect to \mathbf{c}_{true} . Thus, $\alpha^* = 0$ and in particular π_E is optimal for (\mathbf{Lfd}_{π_E}) . \square

We recall the LP approach to MDPs (Puterman, 1994; Bertsekas, 2012). Let $\mathbf{c} \in \mathbf{R}^{|\mathcal{X}| \times |\mathcal{A}|}$ be a cost function. The forward RL problem $(\mathbf{RL}_{\mathbf{c}})$ is equivalent to the following linear programs²

$$\rho_{\mathbf{c}}^* = \min_{\boldsymbol{\mu} \in \mathbf{R}^{|\mathcal{X}| \times |\mathcal{A}|}} \{ \langle \boldsymbol{\mu}, \mathbf{c} \rangle \mid \mathbf{T}_{\gamma} \boldsymbol{\mu} = \boldsymbol{\nu}_0, \boldsymbol{\mu} \geq \mathbf{0} \} \quad \text{(MDP-P}_{\mathbf{c}} \text{)}$$

$$= \max_{\mathbf{u} \in \mathbf{R}^{|\mathcal{X}|}} \{ \langle \boldsymbol{\nu}_0, \mathbf{u} \rangle \mid \mathbf{c} - \mathbf{T}_{\gamma}^* \mathbf{u} \geq \mathbf{0} \} \quad \text{(MDP-D}_{\mathbf{c}} \text{)}$$

²Note that usually in the literature the primal LP is (MDP-D_c).

We have that if π^* is an optimal policy for (\mathbf{RL}_c) , then μ_{π^*} is optimal for $(\mathbf{MDP-P}_c)$ and conversely if μ^* is optimal for $(\mathbf{MDP-P}_c)$, then π_{μ^*} is an optimal policy for (\mathbf{RL}_c) . Moreover, the optimal value function V_c^* is an optimal solution to $(\mathbf{MDP-D}_c)$ and it is the unique optimizer when $\nu_0 \in \mathbf{R}_{++}^{|\mathcal{X}|}$.

Proof of Proposition 2. The proof is based on linear duality and complementary slackness conditions of optimality. Assume first that $(\mathbf{u}_A, \mathbf{w}_A)$ is optimal for (\mathbf{D}_{π_E}) . Then,

$$\mathbf{c}_{\mathbf{w}_A} - \mathbf{T}_\gamma^* \mathbf{u}_A \geq \mathbf{0}, \quad \mathbf{w}_A \in \mathcal{W}, \quad (4)$$

$$\langle \mu_{\pi_E}, \mathbf{T}_\gamma^* \mathbf{u}_A - \mathbf{c}_{\mathbf{w}_A} \rangle = \alpha^* = 0, \quad (5)$$

where (4) holds because $(\mathbf{u}_A, \mathbf{w}_A)$ is feasible to (\mathbf{D}_{π_E}) , and (5) holds by optimality and Lemma 1. Equation (4) states that \mathbf{u}_A is feasible for $(\mathbf{MDP-D}_c)$ with cost $\mathbf{c} = \mathbf{c}_{\mathbf{w}_A}$. Moreover, μ_{π_E} is feasible for $(\mathbf{MDP-P}_c)$ with cost $\mathbf{c} = \mathbf{c}_{\mathbf{w}_A}$. Therefore,

$$\langle \nu_0, \mathbf{u}_A \rangle \leq \rho_{\mathbf{w}_A}^* \leq \langle \mu_{\pi_E}, \mathbf{c}_{\mathbf{w}_A} \rangle. \quad (6)$$

However, by (5) we get that $\langle \nu_0, \mathbf{u}_A \rangle = \langle \mathbf{T}_\gamma \mu_{\pi_E}, \mathbf{u}_A \rangle = \langle \mu_{\pi_E}, \mathbf{T}_\gamma^* \mathbf{u}_A \rangle = \langle \mu_{\pi_E}, \mathbf{c}_{\mathbf{w}_A} \rangle$. Thus, by (6) we conclude that μ_{π_E} is optimal for $(\mathbf{MDP-P}_c)$ with cost $\mathbf{c} = \mathbf{c}_{\mathbf{w}_A}$ and \mathbf{u}_A is optimal for $(\mathbf{MDP-D}_c)$ with cost $\mathbf{c} = \mathbf{c}_{\mathbf{w}_A}$. Thus π_E is optimal for $(\mathbf{RL}_{\mathbf{c}_{\mathbf{w}_A}})$ and $\mathbf{u}_A = \mathbf{V}_{\mathbf{c}_{\mathbf{w}_A}}^*$.

Conversely, assume that $\mathbf{w}_A \in \mathcal{W}$, π_E is optimal for $(\mathbf{RL}_{\mathbf{c}_{\mathbf{w}_A}})$, and $\mathbf{u}_A = \mathbf{V}_{\mathbf{w}_A}^*$. Then, we have that μ_{π_E} is optimal for $(\mathbf{MDP-P}_c)$ with cost $\mathbf{c} = \mathbf{c}_{\mathbf{w}_A}$, and \mathbf{u}_A is optimal for $(\mathbf{MDP-D}_c)$ with $\mathbf{c} = \mathbf{c}_{\mathbf{w}_A}$. Therefore,

$$\mathbf{c}_{\mathbf{w}_A} - \mathbf{T}_\gamma^* \mathbf{u}_A \geq \mathbf{0}, \quad (7)$$

$$\langle \nu_0, \mathbf{u}_A \rangle = \langle \mu_{\pi_E}, \mathbf{c}_{\mathbf{w}_A} \rangle, \quad (8)$$

where (7) holds because \mathbf{u}_A is feasible for $(\mathbf{MDP-D}_c)$ with cost $\mathbf{c} = \mathbf{c}_{\mathbf{w}_A}$, and (8) holds by primal-dual optimality of $(\mu_{\pi_E}, \mathbf{u}_A)$. From (7), we get that $(\mathbf{u}_A, \mathbf{w}_A)$ is feasible to (\mathbf{D}_{π_E}) and from (8), we get that $\langle \mu_{\pi_E}, \mathbf{T}_\gamma^* \mathbf{u}_A - \mathbf{c}_{\mathbf{w}_A} \rangle = 0$. Since by Lemma 1, $\alpha^* = 0$ we conclude that $(\mathbf{u}_A, \mathbf{w}_A)$ is optimal for (\mathbf{D}_{π_E}) . \square

A.4. Proof of Proposition 3

Proposition 3 (Suboptimal expert). *If μ_A is optimal for (\mathbf{P}_{π_E}) and $(\mathbf{u}_A, \mathbf{w}_A)$ is optimal for (\mathbf{D}_{π_E}) , then any apprentice policy π_A is optimal for $(\mathbf{RL}_{\mathbf{c}_{\mathbf{w}_A}})$ and $\mathbf{u}_A = \mathbf{V}_{\mathbf{w}_A}^*$.*

Proof. Similarly to the reasoning of the proof of Proposition 2, we need to show that

$$\begin{aligned} \mathbf{c}_{\mathbf{w}_A} - \mathbf{T}_\gamma^* \mathbf{u}_A &\geq \mathbf{0}, && \text{(dual feasibility to } (\mathbf{MDP-D}_c) \text{ with } \mathbf{c} = \mathbf{c}_{\mathbf{w}_A}) \\ \langle \mu_{\pi_A}, \mathbf{c}_{\mathbf{w}_A} - \mathbf{T}_\gamma^* \mathbf{u}_A \rangle &= 0, && \text{(complementarity)} \end{aligned}$$

The dual feasibility follows from the fact that $(\mathbf{u}_A, \mathbf{w}_A)$ is feasible to (\mathbf{D}_{π_E}) . Moreover, we have

$$\begin{aligned} 0 &\leq \langle \mu_{\pi_A}, \mathbf{c}_{\mathbf{w}_A} - \mathbf{T}_\gamma^* \mathbf{u}_A \rangle && \text{(since } \mathbf{c}_{\mathbf{w}_A} - \mathbf{T}_\gamma^* \mathbf{u}_A \geq \mathbf{0}) \\ &= \langle \mu_{\pi_A}, \mathbf{c}_{\mathbf{w}_A} \rangle - \langle \mu_{\pi_A}, \mathbf{T}_\gamma^* \mathbf{u}_A \rangle && \text{(since } \mathbf{T}_\gamma \mu_{\pi_A} = \mathbf{T}_\gamma \mu_{\pi_E} = \nu_0) \\ &\leq \langle \mu_{\pi_E}, \mathbf{c}_{\mathbf{w}_A} \rangle + \alpha^* - \langle \mu_{\pi_E}, \mathbf{T}_\gamma^* \mathbf{u}_A \rangle && \text{(since } \pi_A \text{ is optimal for } (\mathbf{LFD}_{\pi_E})) \\ &= \alpha^* - \alpha^* = 0. && \text{(since } (\mathbf{u}_A, \mathbf{w}_A) \text{ is optimal for } (\mathbf{D}_{\pi_E})) \end{aligned}$$

So complementarity is satisfied. \square

A.5. Proof of Corollary 1

Corollary 1. *Suppose that the expert is optimal for \mathbf{c}_{true} . Then, $(\mu_{\pi_E}, \mathbf{V}_{\mathbf{c}_{\text{true}}}^*, \mathbf{w}_{\text{true}})$ is a saddle-point to the minimax problem (1). In the general case, for a saddle-point $(\mu_A, \mathbf{u}_A, \mathbf{w}_A)$ it holds that: (i) π_{μ_A} is optimal for (\mathbf{LFD}_{π_E}) , and (ii) π_{μ_A} is optimal for $\mathbf{c}_{\mathbf{w}_A}$ and $\mathbf{u}_A = \mathbf{V}_{\mathbf{w}_A}^*$.*

Proof. The result follows from Propositions 2–3 by noting that $(\mu_A, \mathbf{u}_A, \mathbf{w}_A)$ is a saddle point of the full minimax problem (1) if and only if (μ_A, α^*) is optimal for the primal linear program (\mathbf{P}_{π_E}) and $(\mathbf{u}_A, \mathbf{w}_A)$ is optimal for the dual linear program (\mathbf{D}_{π_E}) . \square

B. Proofs of Section 4.1 and Related Results

For the following results, let π_A be an apprentice policy, i.e., π_A is optimal for (\mathbf{Lfd}_{π_E}) .

B.1. Proof of Proposition 4

Proposition 4. *Let $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \in \Theta \times \Lambda \times \mathcal{W}$ be a feasible solution to (2). Set $\pi_{\boldsymbol{\theta}} = \pi_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}$. It then holds that*

$$\delta_{\mathcal{C}}(\pi_{\boldsymbol{\theta}}, \pi_E) - \delta_{\mathcal{C}}(\pi_A, \pi_E) \leq \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) + \epsilon_{\text{approx}, \boldsymbol{\theta}},$$

where the approximation error $\epsilon_{\text{approx}, \boldsymbol{\theta}}$ is no larger than $\frac{2}{1-\gamma}(\beta \min_{\boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\mu}_{\boldsymbol{\theta}'} - \boldsymbol{\mu}_{\pi_A}\|_1 + \min_{\boldsymbol{\lambda}' \in \Lambda} \|\mathbf{u}_{\boldsymbol{\lambda}'} - \mathbf{V}_{\mathbf{c}_{i_{\boldsymbol{\theta}}}}^{\pi_{\boldsymbol{\theta}}}\|_{\infty})$, and $i_{\boldsymbol{\theta}} \triangleq \arg \max_{i \in [n_c]} (\rho_{\mathbf{c}_i}(\pi_{\boldsymbol{\theta}}) - \rho_{\mathbf{c}_i}(\pi_E))$. In particular, $\rho_{\mathbf{c}_{i_{\text{true}}}}(\pi_{\boldsymbol{\theta}}) - \rho_{\mathbf{c}_{i_{\text{true}}}}(\pi_E) \leq \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) + \epsilon_{\text{approx}, \boldsymbol{\theta}} + \alpha^*$.

In the remainder of the paper, for any policy $\pi \in \Pi_0$, we define the matrix $\mathbf{M}_{\pi} \in \mathbf{R}^{|\mathcal{X}| \times |\mathcal{X}| \times |\mathcal{A}|}$ that encodes π by $M_{\pi, x', (x, a)} \triangleq \pi(a|x)$, if $x = x'$ and $M_{\pi, x', (x, a)} \triangleq 0$, otherwise. Moreover, for any cost function \mathbf{c} , we denote by $\mathbf{c}_{\pi} \in \mathbf{R}^{|\mathcal{X}|}$ the average cost under policy π , i.e., $c_{\pi}(x) \triangleq \sum_a \pi(a|x)c(x, a)$. Finally, let $\mathbf{P}_{\pi} \in \mathbf{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ be the state-transition of running policy π , i.e., $P_{\pi, x, x'} \triangleq \sum_a \pi(a|x)P(x'|x, a)$.

We will need the following Lemma.

Lemma 2. *Let $\boldsymbol{\mu} \in \Delta_{\mathcal{X} \times \mathcal{A}}$ (not necessarily an occupancy measure) and let $\pi_{\boldsymbol{\mu}} \in \Pi_0$ be the induced policy given by $\pi_{\boldsymbol{\mu}}(a|x) \triangleq \frac{\boldsymbol{\mu}(x, a)}{\sum_{a'} \boldsymbol{\mu}(x, a')}$. For any cost vector $\mathbf{c} \in \mathbf{R}^{|\mathcal{X}| \times |\mathcal{A}|}$ it holds that $\langle \boldsymbol{\mu}, \mathbf{c} - \mathbf{T}_{\gamma}^* \mathbf{V}_{\mathbf{c}}^{\pi_{\boldsymbol{\mu}}} \rangle = 0$.*

Proof. We have that $\mathbf{V}_{\mathbf{c}}^{\pi_{\boldsymbol{\mu}}}$ is the unique fixed point of the following Bellman equation (Puterman, 1994),

$$V_{\mathbf{c}}^{\pi_{\boldsymbol{\mu}}}(x) = \sum_a \pi_{\boldsymbol{\mu}}(a|x) \left((1-\gamma)c(x, a) + \gamma \sum_{x'} P(x'|x, a) V_{\mathbf{c}}^{\pi_{\boldsymbol{\mu}}}(x') \right). \quad (9)$$

Note that (9) can be written equivalently in vector form as $\mathbf{c}_{\pi_{\boldsymbol{\mu}}} - \frac{1}{1-\gamma}(\mathbf{I} - \gamma \mathbf{P}_{\pi_{\boldsymbol{\mu}}}) \mathbf{V}_{\mathbf{c}}^{\pi_{\boldsymbol{\mu}}} = 0$. We then get $\langle \mathbf{B}^{\top} \boldsymbol{\mu}, \mathbf{M}_{\pi_{\boldsymbol{\mu}}}(\mathbf{c} - \mathbf{T}_{\gamma}^* \mathbf{V}_{\mathbf{c}}^{\pi_{\boldsymbol{\mu}}}) \rangle = 0$, since $\mathbf{P}_{\pi_{\boldsymbol{\mu}}} = \mathbf{M}_{\pi_{\boldsymbol{\mu}}} \mathbf{P}$, $\mathbf{I} = \mathbf{M}_{\pi_{\boldsymbol{\mu}}} \mathbf{B}$ and $\mathbf{c}_{\pi_{\boldsymbol{\mu}}} = \mathbf{M}_{\pi_{\boldsymbol{\mu}}} \mathbf{c}$. Finally the result follows by using that $\mathbf{M}_{\pi_{\boldsymbol{\mu}}}^{\top} \mathbf{B}^{\top} \boldsymbol{\mu} = \boldsymbol{\mu}$. □

We are now ready to prove Proposition 4.

Proof of Proposition 4. Let $\mathbf{e}_{i_{\boldsymbol{\theta}}} \in \mathbf{R}^{n_c}$ be the $i_{\boldsymbol{\theta}}$ -th basis vector. We then have,

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{V}_{\mathbf{c}_{i_{\boldsymbol{\theta}}}}^{\pi_{\boldsymbol{\theta}}}, \mathbf{e}_{i_{\boldsymbol{\theta}}}) - \mathcal{L}(\boldsymbol{\mu}_{\pi_A}, \mathbf{u}_{\boldsymbol{\lambda}}, \mathbf{w}) &= \langle \boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_{i_{\boldsymbol{\theta}}} - \mathbf{T}_{\gamma}^* \mathbf{V}_{\mathbf{c}_{i_{\boldsymbol{\theta}}}}^{\pi_{\boldsymbol{\theta}}} \rangle - \langle \boldsymbol{\mu}_{\pi_A} - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_{\mathbf{w}} - \mathbf{T}_{\gamma}^* \mathbf{u}_{\boldsymbol{\lambda}} \rangle \\ &= \langle \boldsymbol{\nu}_0, \mathbf{V}_{\mathbf{c}_{i_{\boldsymbol{\theta}}}}^{\pi_{\boldsymbol{\theta}}} \rangle - \langle \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_{i_{\boldsymbol{\theta}}} \rangle - \langle \boldsymbol{\mu}_{\pi_A} - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}_{\mathbf{w}} \rangle \\ &= \underbrace{(\rho_{\mathbf{c}_{i_{\boldsymbol{\theta}}}}(\pi_{\boldsymbol{\theta}}) - \rho_{\mathbf{c}_{i_{\boldsymbol{\theta}}}}(\pi_E))}_{=\delta_{\mathcal{C}}(\pi_{\boldsymbol{\theta}}, \pi_E)} - \underbrace{(\rho_{\mathbf{c}_{\mathbf{w}}}(\pi_A) - \rho_{\mathbf{c}_{\mathbf{w}}}(\pi_E))}_{\leq \delta_{\mathcal{C}}(\pi_A, \pi_E)} \\ &\geq \delta_{\mathcal{C}}(\pi_{\boldsymbol{\theta}}, \pi_E) - \delta_{\mathcal{C}}(\pi_A, \pi_E), \end{aligned}$$

where the second line follows because $\langle \boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{c}_{i_{\boldsymbol{\theta}}} - \mathbf{T}_{\gamma}^* \mathbf{V}_{\mathbf{c}_{i_{\boldsymbol{\theta}}}}^{\pi_{\boldsymbol{\theta}}} \rangle = 0$ by Lemma 2, and $\langle \boldsymbol{\mu}_{\pi_A} - \boldsymbol{\mu}_{\pi_E}, \mathbf{T}_{\gamma}^* \mathbf{u}_{\boldsymbol{\lambda}} \rangle = 0$, since $\mathbf{T}_{\gamma} \boldsymbol{\mu}_{\pi_A} = \mathbf{T}_{\gamma} \boldsymbol{\mu}_{\pi_E} = \boldsymbol{\nu}_0$.

All in all, we have that

$$\delta_{\mathcal{C}}(\pi_{\boldsymbol{\theta}}, \pi_E) - \delta_{\mathcal{C}}(\pi_A, \pi_E) \leq \mathcal{L}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{V}_{\mathbf{c}_{i_{\boldsymbol{\theta}}}}^{\pi_{\boldsymbol{\theta}}}, \mathbf{e}_{i_{\boldsymbol{\theta}}}) - \mathcal{L}(\boldsymbol{\mu}_{\pi_A}, \mathbf{u}_{\boldsymbol{\lambda}}, \mathbf{w}) \quad (10)$$

By using a simple rearrangement trick we get

$$\delta_{\mathcal{C}}(\pi_{\theta}, \pi_{\mathbb{E}}) - \delta_{\mathcal{C}}(\pi_{\Lambda}, \pi_{\mathbb{E}}) \leq \underbrace{\mathcal{L}(\boldsymbol{\mu}_{\theta}, \mathbf{V}_{\mathbf{c}_{i_{\theta}}}^{\pi_{\theta}}, \mathbf{e}_{i_{\theta}}) - \mathcal{L}(\boldsymbol{\mu}_{\pi_{\Lambda}}, \mathbf{u}_{\lambda}, \mathbf{w}) - \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w})}_{\triangleq \varepsilon_{\text{approx}, \boldsymbol{\theta}}} + \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \quad (11)$$

It remains to bound the term $\varepsilon_{\text{approx}, \boldsymbol{\theta}}$. We set $\mathbf{z}' = (\boldsymbol{\theta}', \boldsymbol{\lambda}', \mathbf{w}')$ and $\mathbf{Z} = \Theta \times \Lambda \times \mathcal{W}$ for brevity. We then have

$$\begin{aligned} \varepsilon_{\text{approx}, \boldsymbol{\theta}} &= \min_{\mathbf{z}' \in \mathbf{Z}} \mathcal{L}(\boldsymbol{\mu}_{\theta}, \mathbf{V}_{\mathbf{c}_{i_{\theta}}}^{\pi_{\theta}}, \mathbf{e}_{i_{\theta}}) - \mathcal{L}(\boldsymbol{\mu}_{\pi_{\Lambda}}, \mathbf{u}_{\lambda}, \mathbf{w}) - \mathcal{L}(\boldsymbol{\mu}_{\theta'}, \mathbf{u}_{\lambda'}, \mathbf{w}') + \mathcal{L}(\boldsymbol{\mu}_{\theta'}, \mathbf{u}_{\lambda}, \mathbf{w}) \\ &= \min_{\mathbf{z}' \in \mathbf{Z}} \langle \boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_{\pi_{\Lambda}}, \mathbf{c}_{i_{\theta}} - \mathbf{c}_{\mathbf{w}'} \rangle + \left\langle \mathbf{T}_{\gamma} \boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0, \mathbf{u}_{\lambda'} - \mathbf{V}_{\mathbf{c}_{i_{\theta}}}^{\pi_{\theta}} \right\rangle + \langle \boldsymbol{\mu}_{\theta'} - \boldsymbol{\mu}_{\pi_{\Lambda}}, \mathbf{c}_{i_{\theta}} - \mathbf{T}_{\gamma}^* \mathbf{u}_{\lambda} \rangle \\ &\leq \min_{\mathbf{z}' \in \mathbf{Z}} \underbrace{\|\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_{\pi_{\Lambda}}\|_1}_{\leq 2} \|\mathbf{c}_{i_{\theta}} - \mathbf{c}_{\mathbf{w}'}\|_{\infty} + \underbrace{\|\mathbf{T}_{\gamma} \boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0\|_1}_{\leq \frac{2}{1-\gamma}} \|\mathbf{u}_{\lambda'} - \mathbf{V}_{\mathbf{c}_{i_{\theta}}}^{\pi_{\theta}}\|_{\infty} + \|\boldsymbol{\mu}_{\theta'} - \boldsymbol{\mu}_{\pi_{\Lambda}}\|_1 \underbrace{\|\mathbf{c}_{i_{\theta}} - \mathbf{T}_{\gamma}^* \mathbf{u}_{\lambda}\|_{\infty}}_{\leq \frac{2\beta}{1-\gamma}} \\ &\leq \frac{2\beta}{1-\gamma} \min_{\boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\mu}_{\theta'} - \boldsymbol{\mu}_{\pi_{\Lambda}}\|_1 + \frac{2}{1-\gamma} \min_{\boldsymbol{\lambda}' \in \Lambda} \|\mathbf{u}_{\lambda'} - \mathbf{V}_{\mathbf{c}_{i_{\theta}}}^{\pi_{\theta}}\|_{\infty}. \end{aligned} \quad (12)$$

The bounds on the norms involved in the first inequality have been obtained by using the triangle inequality and that $\|\boldsymbol{\mu}\|_1 = 1$, for all $\boldsymbol{\mu} \in \Delta_{\mathcal{X} \times \mathcal{A}}$, $\boldsymbol{\mu}_{\theta'} \in \Delta_{\mathcal{X} \times \mathcal{A}}$, for all $\boldsymbol{\theta}' \in \Theta$, $\|\boldsymbol{\nu}_0\|_1 = 1$, $\|\mathbf{B}\|_{\infty} = \|\mathbf{P}\|_{\infty} = 1$, $\|\mathbf{u}_{\lambda'}\|_{\infty} \leq \beta$, for all $\boldsymbol{\lambda}' \in \Lambda$, and $\|\mathbf{c}_{\mathbf{w}'}\|_{\infty} \leq 1$, for all $\mathbf{w}' \in \mathcal{W}$. The detailed computations can be found in Lemmas 7–8. For the last inequality, note that obviously $\min_{\mathbf{w}' \in \mathcal{W}} \|\mathbf{c}_{i_{\theta}} - \mathbf{c}_{\mathbf{w}'}\|_{\infty} = 0$.

By combining (11) and (12) we conclude the proof. \square

B.2. Corollaries Related to Proposition 4

By modifying analogously the proof of Proposition 4, we get the following tighter bound.

Corollary 2. *Let $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \in \Theta \times \Lambda \times \mathcal{W}$ be a feasible solution to (2). Set $\pi_{\theta} = \pi_{\boldsymbol{\mu}_{\theta}}$. It then holds that*

$$\delta_{\mathcal{C}}(\pi_{\theta}, \pi_{\mathbb{E}}) - \delta_{\mathcal{C}}(\pi_{\Lambda}, \pi_{\mathbb{E}}) \leq \mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}_{\theta}, \mathbf{e}_{i_{\theta}}) - \mathcal{L}_r(\boldsymbol{\theta}^*, \boldsymbol{\lambda}, \mathbf{w}) + \varepsilon_{\text{approx}, \boldsymbol{\theta}},$$

where the approximation error $\varepsilon_{\text{approx}, \boldsymbol{\theta}}$ is no larger than $\frac{2}{1-\gamma} (\beta \min_{\boldsymbol{\theta} \in \Theta} \|\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_{\pi_{\Lambda}}\|_1 + \min_{\boldsymbol{\lambda} \in \Lambda} \|\mathbf{u}_{\lambda} - \mathbf{V}_{\mathbf{c}_{i_{\theta}}}^{\pi_{\theta}}\|_{\infty})$, and

$$\boldsymbol{\theta}^* \triangleq \arg \min_{\boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\mu}_{\theta'} - \boldsymbol{\mu}_{\pi_{\Lambda}}\|_1, \quad \boldsymbol{\lambda} \triangleq \arg \min_{\boldsymbol{\lambda}' \in \Lambda} \|\mathbf{u}_{\lambda'} - \mathbf{V}_{\mathbf{c}_{i_{\theta}}}^{\pi_{\theta}}\|_{\infty}, \quad i_{\theta} \triangleq \arg \max_{i \in [n_c]} (\rho_{\mathbf{c}_i}(\pi_{\theta}) - \rho_{\mathbf{c}_i}(\pi_{\mathbb{E}})).$$

In particular, $\rho_{\mathbf{c}_{\text{true}}}(\pi_{\theta}) - \rho_{\mathbf{c}_{\text{true}}}(\pi_{\mathbb{E}}) \leq \mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}_{\theta}, \mathbf{e}_{i_{\theta}}) - \mathcal{L}_r(\boldsymbol{\theta}^*, \boldsymbol{\lambda}, \mathbf{w}) + \varepsilon_{\text{approx}, \boldsymbol{\theta}} + \alpha^*$.

When the occupancy measure features can accurately represent the expert occupancy measure $\boldsymbol{\mu}_{\pi_{\mathbb{E}}}$ instead of the apprentice occupancy measure $\boldsymbol{\mu}_{\pi_{\Lambda}}$, we may use alternatively the following result.

Corollary 3. *Let $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \in \Theta \times \Lambda \times \mathcal{W}$ be a feasible solution to (2). Set $\pi_{\theta} = \pi_{\boldsymbol{\mu}_{\theta}}$. It then holds that*

$$\delta_{\mathcal{C}}(\pi_{\theta}, \pi_{\mathbb{E}}) \leq \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) + \varepsilon_{\text{approx}, \mathbb{E}, \boldsymbol{\theta}},$$

where the approximation error $\varepsilon_{\text{approx}, \mathbb{E}, \boldsymbol{\theta}}$ is no larger than $\frac{2}{1-\gamma} (\beta \min_{\boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\mu}_{\theta'} - \boldsymbol{\mu}_{\pi_{\mathbb{E}}}\|_1 + \min_{\boldsymbol{\lambda}' \in \Lambda} \|\mathbf{u}_{\lambda'} - \mathbf{V}_{\mathbf{c}_{i_{\theta}}}^{\pi_{\theta}}\|_{\infty})$, and $i_{\theta} \triangleq \arg \max_{i \in [n_c]} (\rho_{\mathbf{c}_i}(\pi_{\theta}) - \rho_{\mathbf{c}_i}(\pi_{\mathbb{E}}))$. In particular, $\rho_{\mathbf{c}_{\text{true}}}(\pi_{\theta}) - \rho_{\mathbf{c}_{\text{true}}}(\pi_{\mathbb{E}}) \leq \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) + \varepsilon_{\text{approx}, \mathbb{E}, \boldsymbol{\theta}}$.

Proof. The proof is the same as the proof of Proposition 4 by replacing π_{Λ} with $\pi_{\mathbb{E}}$. \square

B.3. Proof of Proposition 5

Proposition 5. *Let Assumption 1 hold. Let $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \in \Theta \times \Lambda \times \mathcal{W}$ be a feasible solution to (2). Set $\pi_{\theta} = \pi_{\boldsymbol{\mu}_{\theta}}$. It then holds that*

$$\rho_{\mathbf{c}_{\text{true}}}(\pi_{\theta}) - \rho_{\mathbf{c}_{\text{true}}}(\pi_{\mathbb{E}}) \leq 3\epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) + \varepsilon_{\text{approx}, w} + \alpha^*,$$

where the weak approximation error $\varepsilon_{\text{approx}, w}$ is no larger than $\frac{2}{1-\gamma} (\beta \min_{\boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\mu}_{\theta'} - \boldsymbol{\mu}_{\pi_{\Lambda}}\|_1 + 2\|\mathbf{V}_{\mathbf{w}_{\Lambda}}^* - \boldsymbol{\Psi}\boldsymbol{\lambda}^*\|_{\infty})$, where $(\mathbf{V}_{\mathbf{w}_{\Lambda}}^*, \mathbf{w}_{\Lambda})$ is dual optimal for $(\mathbf{D}_{\pi_{\mathbb{E}}})$ and $\boldsymbol{\lambda}^*$ is a dual optimizer of (2).

The proof of Proposition 5 is divided in several parts. We first note that even when a probability distribution $\boldsymbol{\mu} \in \Delta_{\mathcal{X} \times \mathcal{A}}$ is not an occupancy measure, i.e., it does not satisfy the Bellman flow constraints, it can still generate a policy $\pi_{\boldsymbol{\mu}}$. The following lemma states that the *distribution mismatch* between $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\pi_{\boldsymbol{\mu}}}$ depends on the violation degree of the Bellman flow constraints.

Lemma 3. *Let $\boldsymbol{\mu} \in \Delta_{\mathcal{X} \times \mathcal{A}}$ and let $\pi_{\boldsymbol{\mu}}$ be the induced policy. It holds that $\|\boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_{\boldsymbol{\mu}}}\|_1 \leq \|\boldsymbol{\nu}_0 - \mathbf{T}_{\gamma}\boldsymbol{\mu}\|_1$.*

Proof. We have that

$$\boldsymbol{\mu}_{\pi_{\boldsymbol{\mu}}}^{\top} = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \boldsymbol{\nu}_0^{\top} \mathbf{M}_{\pi_{\boldsymbol{\mu}}} (\mathbf{P} \mathbf{M}_{\pi_{\boldsymbol{\mu}}})^t \quad (13)$$

$$= (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t (\boldsymbol{\nu}_0 - \mathbf{T}_{\gamma}\boldsymbol{\mu} + \mathbf{T}_{\gamma}\boldsymbol{\mu})^{\top} \mathbf{M}_{\pi_{\boldsymbol{\mu}}} (\mathbf{P} \mathbf{M}_{\pi_{\boldsymbol{\mu}}})^t \quad (14)$$

$$= (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t (\boldsymbol{\nu}_0 - \mathbf{T}_{\gamma}\boldsymbol{\mu})^{\top} \mathbf{M}_{\pi_{\boldsymbol{\mu}}} (\mathbf{P} \mathbf{M}_{\pi_{\boldsymbol{\mu}}})^t + \sum_{t=0}^{\infty} \gamma^t \boldsymbol{\mu}^{\top} (\mathbf{P} \mathbf{M}_{\pi_{\boldsymbol{\mu}}})^t - \sum_{t=0}^{\infty} \gamma^{t+1} \boldsymbol{\mu}^{\top} (\mathbf{P} \mathbf{M}_{\pi_{\boldsymbol{\mu}}})^{t+1} \quad (15)$$

$$= (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t (\boldsymbol{\nu}_0 - \mathbf{T}_{\gamma}\boldsymbol{\mu})^{\top} \mathbf{M}_{\pi_{\boldsymbol{\mu}}} (\mathbf{P} \mathbf{M}_{\pi_{\boldsymbol{\mu}}})^t + \boldsymbol{\mu}^{\top}, \quad (16)$$

where in the third equality we used that $\mathbf{T}_{\gamma}\boldsymbol{\mu} = \frac{1}{1-\gamma}(\mathbf{B} - \gamma\mathbf{P})^{\top}\boldsymbol{\mu}$ and that $\mathbf{M}_{\pi_{\boldsymbol{\mu}}}^{\top}\mathbf{B}^{\top}\boldsymbol{\mu} = \boldsymbol{\mu}$. Therefore,

$$\|\boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_{\boldsymbol{\mu}}}\|_1 \leq (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t (\mathbf{M}_{\pi_{\boldsymbol{\mu}}}^{\top} \mathbf{P}^{\top})^t \mathbf{M}_{\pi_{\boldsymbol{\mu}}}^{\top} (\boldsymbol{\nu}_0 - \mathbf{T}_{\gamma}\boldsymbol{\mu}) \|_1 \quad (17)$$

$$\leq \|\mathbf{M}_{\pi_{\boldsymbol{\mu}}}\|_{\infty}^t \|\mathbf{P}\|_{\infty} \|\mathbf{M}_{\pi_{\boldsymbol{\mu}}}\|_{\infty} \|\boldsymbol{\nu}_0 - \mathbf{T}_{\gamma}\boldsymbol{\mu}\|_1 \quad (18)$$

$$= \|\boldsymbol{\nu}_0 - \mathbf{T}_{\gamma}\boldsymbol{\mu}\|_1, \quad (19)$$

where we used that $\|\mathbf{P}\|_{\infty} = \|\mathbf{M}_{\pi_{\boldsymbol{\mu}}}\|_{\infty} = 1$. \square

We will also need the following lemma.

Lemma 4. *Let $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \in \Theta \times \Lambda \times \mathcal{W}$ be a feasible solution to (2) and let $\pi_{\boldsymbol{\theta}} = \pi_{\boldsymbol{\mu}_{\boldsymbol{\theta}}}$. It holds that*

$$\rho_{\mathbf{c}_{\text{true}}}(\pi_{\boldsymbol{\theta}}) - \rho_{\mathbf{c}_{\text{true}}}(\pi_{\mathbf{E}}) \leq \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) + 2\|\mathbf{T}_{\gamma}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\nu}_0\|_1 + \frac{2\beta}{1-\gamma} \min_{\boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\mu}_{\boldsymbol{\theta}'} - \boldsymbol{\mu}_{\pi_{\Lambda}}\|_1 + \alpha^*.$$

Proof. We fix a $\boldsymbol{\lambda}_0 \in \Lambda$ such that $\|\boldsymbol{\lambda}_0\|_2 \leq \frac{1}{\sqrt{n_u}}$. By assumption the columns of $\boldsymbol{\Psi}$ belong to \mathcal{U} . Thus, $\|\mathbf{u}_{\boldsymbol{\lambda}_0}\|_{\infty} \leq 1$. Moreover,

$$\mathcal{L}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{u}_{\boldsymbol{\lambda}_0}, \mathbf{w}_{\text{true}}) - \mathcal{L}(\boldsymbol{\mu}_{\pi_{\Lambda}}, \mathbf{u}_{\boldsymbol{\lambda}}, \mathbf{w}) = \langle \boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\pi_{\mathbf{E}}}, \mathbf{c}_{\text{true}} - \mathbf{T}_{\gamma}^* \mathbf{u}_{\boldsymbol{\lambda}_0} \rangle - \langle \boldsymbol{\mu}_{\pi_{\Lambda}} - \boldsymbol{\mu}_{\pi_{\mathbf{E}}}, \mathbf{c}_{\mathbf{w}} - \mathbf{T}_{\gamma}^* \mathbf{u}_{\boldsymbol{\lambda}} \rangle \quad (20)$$

$$\geq \langle \boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{c}_{\text{true}} \rangle - \rho_{\mathbf{c}_{\text{true}}}(\pi_{\mathbf{E}}) - \langle \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\nu}_0, \mathbf{u}_{\boldsymbol{\lambda}_0} \rangle - \alpha^*, \quad (21)$$

where in the last inequality we used that $\langle \boldsymbol{\mu}_{\pi_{\Lambda}} - \boldsymbol{\mu}_{\pi_{\mathbf{E}}}, \mathbf{c}_{\mathbf{w}} \rangle \leq \delta_{\mathcal{C}}(\pi_{\Lambda}, \pi_{\mathbf{E}}) = \alpha^*$, and that $\mathbf{T}_{\gamma}\boldsymbol{\mu}_{\pi_{\Lambda}} = \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\pi_{\mathbf{E}}} = \boldsymbol{\nu}_0$. Now since $\langle \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\nu}_0, \mathbf{u}_{\boldsymbol{\lambda}_0} \rangle \leq \|\mathbf{T}_{\gamma}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\nu}_0\|_1 \|\mathbf{u}_{\boldsymbol{\lambda}_0}\|_{\infty}$ and $\|\mathbf{u}_{\boldsymbol{\lambda}_0}\|_{\infty} \leq 1$, we get by (21) that

$$\langle \boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{c}_{\text{true}} \rangle - \rho_{\mathbf{c}_{\text{true}}}(\pi_{\mathbf{E}}) \leq \mathcal{L}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{u}_{\boldsymbol{\lambda}_0}, \mathbf{w}_{\text{true}}) - \mathcal{L}(\boldsymbol{\mu}_{\pi_{\Lambda}}, \mathbf{u}_{\boldsymbol{\lambda}}, \mathbf{w}) + \|\mathbf{T}_{\gamma}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\nu}_0\|_1 + \alpha^*. \quad (22)$$

Moreover, we have the bound

$$\langle \boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{c}_{\text{true}} \rangle - \rho_{\mathbf{c}_{\text{true}}}(\pi_{\boldsymbol{\theta}}) = \langle \boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\pi_{\boldsymbol{\theta}}}, \mathbf{c}_{\text{true}} \rangle \quad (23)$$

$$\leq \|\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\pi_{\boldsymbol{\theta}}}\|_1 \|\mathbf{c}_{\text{true}}\|_{\infty} \quad (24)$$

$$\leq \|\mathbf{T}_{\gamma}\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\nu}_0\|_1, \quad (25)$$

where the last inequality follows by Lemma 3 and the bound $\|\mathbf{c}_{\text{true}}\|_{\infty} \leq 1$.

So by (22) and (25) we get

$$\rho_{\mathbf{c}_{\text{true}}}(\pi_{\theta}) - \rho_{\mathbf{c}_{\text{true}}}(\pi_E) \leq \mathcal{L}(\boldsymbol{\mu}_{\theta}, \mathbf{u}_{\lambda_0}, \mathbf{w}_{\text{true}}) - \mathcal{L}(\boldsymbol{\mu}_{\pi_A}, \mathbf{u}_{\lambda}, \mathbf{w}) + 2\|\mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0\|_1 + \alpha^*. \quad (26)$$

Next, we apply once more the following rearrangement trick

$$\mathcal{L}(\boldsymbol{\mu}_{\theta}, \mathbf{u}_{\lambda_0}, \mathbf{w}_{\text{true}}) - \mathcal{L}(\boldsymbol{\mu}_{\pi_A}, \mathbf{u}_{\lambda}, \mathbf{w}) = \underbrace{\mathcal{L}(\boldsymbol{\mu}_{\theta}, \mathbf{u}_{\lambda_0}, \mathbf{w}_{\text{true}}) - \mathcal{L}(\boldsymbol{\mu}_{\pi_A}, \mathbf{u}_{\lambda}, \mathbf{w})}_{\triangleq (i)} - \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) + \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}). \quad (27)$$

It remains to bound the term (i). By similar calculations to the ones in the proof of Proposition 4, we get

$$(i) \leq \frac{2\beta}{1-\gamma} \min_{\boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\mu}_{\boldsymbol{\theta}'} - \boldsymbol{\mu}_{\pi_A}\|_1 + 2 \underbrace{\min_{\mathbf{w}' \in \mathcal{W}} \|\mathbf{c}_{\mathbf{w}'} - \mathbf{c}_{\text{true}}\|_{\infty}}_{=0} + \frac{1}{1-\gamma} \min_{\boldsymbol{\lambda}' \in \Lambda} \underbrace{\|\mathbf{u}_{\boldsymbol{\lambda}'} - \mathbf{u}_{\lambda_0}\|_{\infty}}_{=0}. \quad (28)$$

By combining (26), (27) and (28), we conclude the proof. \square

Finally the following lemma bounds the term $\|\mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0\|_1$.

Lemma 5. *Let Assumption 1 hold. It holds that $\|\mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0\|_1 \leq \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) + \frac{2}{1-\gamma} \|\mathbf{V}_{\mathbf{w}_A}^* - \Psi\boldsymbol{\lambda}^*\|_{\infty}$, where $\boldsymbol{\lambda}^*$ is a dual optimizer for (2).*

Proof. Let $\tilde{\mathbf{u}} = \arg \max_{\|\mathbf{u}\|_{\infty} \leq 2} \langle \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0, -\mathbf{u} \rangle$. By the Coherence Assumption 1, there exists $\tilde{\boldsymbol{\lambda}} \in \Lambda$, such that $\langle \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0, \mathbf{u}_{\tilde{\boldsymbol{\lambda}}} - \tilde{\mathbf{u}} \rangle = 0$. We then have,

$$\begin{aligned} 2\|\mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0\|_1 &= \max_{\|\mathbf{u}\|_{\infty} \leq 2} \langle \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0, -\mathbf{u} \rangle \\ &= \langle \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0, -\tilde{\mathbf{u}} \rangle \\ &= \langle \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0, -\mathbf{u}_{\tilde{\boldsymbol{\lambda}}} \rangle \\ &= \langle \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0, \mathbf{V}_{\mathbf{w}_A}^* - \mathbf{u}_{\tilde{\boldsymbol{\lambda}}} \rangle + \langle \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0, \mathbf{V}_{\mathbf{w}_A}^* \rangle \\ &\leq \langle \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0, \mathbf{V}_{\mathbf{w}_A}^* - \mathbf{u}_{\tilde{\boldsymbol{\lambda}}} \rangle + \|\mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0\|_1, \end{aligned}$$

where in the last inequality we used that $\|\mathbf{V}_{\mathbf{w}_A}^*\|_{\infty} \leq 1$. Therefore,

$$\|\mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0\|_1 \leq \langle \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0, \mathbf{V}_{\mathbf{w}_A}^* - \mathbf{u}_{\tilde{\boldsymbol{\lambda}}} \rangle. \quad (29)$$

Now let $(\boldsymbol{\lambda}^*, \mathbf{w}^*)$ be a dual optimal solution to the linearly-relaxed saddle-point formulation (2). We can then write,

$$\langle \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0, \mathbf{V}_{\mathbf{w}_A}^* - \mathbf{u}_{\tilde{\boldsymbol{\lambda}}} \rangle = \mathcal{L}_r(\boldsymbol{\theta}, \tilde{\boldsymbol{\lambda}}, \mathbf{w}^*) - \mathcal{L}(\boldsymbol{\mu}_{\theta}, \mathbf{V}_{\mathbf{w}_A}^*, \mathbf{w}^*) \quad (30)$$

$$= \underbrace{\mathcal{L}_r(\boldsymbol{\theta}, \tilde{\boldsymbol{\lambda}}, \mathbf{w}^*) - \mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}^*, \mathbf{w}^*)}_{\leq \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w})} + \underbrace{\mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}^*, \mathbf{w}^*) - \mathcal{L}(\boldsymbol{\mu}_{\theta}, \mathbf{V}_{\mathbf{w}_A}^*, \mathbf{w}^*)}_{\leq \frac{2}{1-\gamma} \|\mathbf{V}_{\mathbf{w}_A}^* - \Psi\boldsymbol{\lambda}^*\|_{\infty}}. \quad (31)$$

We will now explain the bounds in (31). For the first term, we have

$$\begin{aligned} \mathcal{L}_r(\boldsymbol{\theta}, \tilde{\boldsymbol{\lambda}}, \mathbf{w}^*) - \mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}^*, \mathbf{w}^*) &\leq \max_{(\boldsymbol{\lambda}', \mathbf{w}') \in \Lambda \times \mathcal{W}} \mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}', \mathbf{w}') - \min_{\boldsymbol{\theta}' \in \Theta} \mathcal{L}_r(\boldsymbol{\theta}', \boldsymbol{\lambda}^*, \mathbf{w}^*) \\ &\leq \epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}). \end{aligned}$$

For the second term we have

$$\begin{aligned} \mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}^*, \mathbf{w}^*) - \mathcal{L}(\boldsymbol{\mu}_{\theta}, \mathbf{V}_{\mathbf{w}_A}^*, \mathbf{w}^*) &= \langle \mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0, \mathbf{V}_{\mathbf{w}_A}^* - \Psi\boldsymbol{\lambda}^* \rangle \\ &\leq \underbrace{\|\mathbf{T}_{\gamma}\boldsymbol{\mu}_{\theta} - \boldsymbol{\nu}_0\|_1}_{\leq \frac{2}{1-\gamma}} \|\mathbf{V}_{\mathbf{w}_A}^* - \Psi\boldsymbol{\lambda}^*\|_{\infty}. \end{aligned}$$

Putting together the bounds (29) and (31) concludes the proof. \square

Proof of Proposition 5. It follows by Lemma 4 and 5. \square

B.4. Corollaries Related to Proposition 5

When the occupancy measure features can accurately represent the expert occupancy measure μ_{π_E} instead of the apprentice occupancy measure μ_{π_A} , we may use alternatively the following result.

Corollary 4. *Let Assumption 1 hold. Let $(\theta, \lambda, \mathbf{w}) \in \Theta \times \Lambda \times \mathcal{W}$ be a feasible solution to (2). Set $\pi_\theta = \pi_{\mu_\theta}$. It then holds that*

$$\rho_{\mathbf{c}_{\text{true}}}(\pi_\theta) - \rho_{\mathbf{c}_{\text{true}}}(\pi_E) \leq 3\epsilon_{\text{sad}}(\theta, \lambda, \mathbf{w}) + \epsilon_{\text{approx}, E, w},$$

where the weak approximation error $\epsilon_{\text{approx}, E, w}$ is no larger than $\frac{2}{1-\gamma} \left(\beta \min_{\theta' \in \Theta} \|\mu_{\theta'} - \mu_{\pi_E}\|_1 + 2\|\mathbf{V}_{\mathbf{w}_A}^* - \Psi\lambda^*\|_\infty \right)$, where $(\mathbf{V}_{\mathbf{w}_A}^*, \mathbf{w}_A)$ is dual optimal for (\mathbf{D}_{π_E}) and λ^* is a dual optimizer of (2).

Proof. In the proof of Lemma 4 replace π_A with π_E . □

The following Lemma shows that the realizability assumption implies that the weak approximation error $\epsilon_{\text{approx}, w}$ is zero.

Lemma 6 (Realizability). *Let $(\mathbf{u}_A, \mathbf{w}_A)$ be an optimizer for the dual linear program (\mathbf{D}_{π_E}) . Assume that the realizability assumption holds, i.e., there exist $\theta_A \in \Theta$ and $\lambda_A \in \Lambda$ such that, $\mu_{\pi_A} = \Phi\theta_A$ and $\mathbf{u}_A = \Psi\lambda_A$. Then, $(\theta_A, \lambda_A, \mathbf{w}_A)$ is a saddle point of (2). In particular $\epsilon_{\text{approx}, w} = 0$.*

Proof. We need to show that

$$\max_{\lambda \in \Lambda} \max_{\mathbf{w} \in \mathcal{W}} \mathcal{L}_r(\theta_A, \lambda, \mathbf{w}) = \mathcal{L}_r(\theta_A, \lambda_A, \mathbf{w}_A) = \min_{\theta \in \Theta} \mathcal{L}_r(\theta, \lambda_A, \mathbf{w}_A). \quad (32)$$

By the optimality of $(\mu_{\pi_A}, \mathbf{u}_A, \mathbf{w}_A)$ for the full minimax problem (1), we get that

$$\mathcal{L}_r(\theta_A, \lambda_A, \mathbf{w}_A) = \mathcal{L}(\mu_{\pi_A}, \mathbf{u}_A, \mathbf{w}_A) = \alpha^*. \quad (33)$$

Moreover, for all $\theta \in \Theta$

$$\begin{aligned} \mathcal{L}_r(\theta, \lambda_A, \mathbf{w}_A) &= \langle \mu_\theta - \mu_{\pi_E}, \mathbf{c}_{\mathbf{w}_A} - \mathbf{T}_\gamma^* \mathbf{u}_A \rangle \\ &= \underbrace{\langle \mu_\theta, \mathbf{c}_{\mathbf{w}_A} - \mathbf{T}_\gamma^* \mathbf{u}_A \rangle}_{\geq 0} + \underbrace{\langle \mu_{\pi_E}, \mathbf{T}_\gamma^* \mathbf{u}_A - \mathbf{c}_{\mathbf{w}_A} \rangle}_{=\alpha^*} \geq \alpha^*. \end{aligned} \quad (34)$$

Note that by Proposition 3, \mathbf{u}_A is feasible for (MDP-D_c) with cost function $\mathbf{c} = \mathbf{c}_{\mathbf{w}_A}$. Therefore, $\mathbf{c}_{\mathbf{w}_A} - \mathbf{T}_\gamma^* \mathbf{u}_A \geq 0$. Moreover, $\langle \mu_{\pi_E}, \mathbf{T}_\gamma^* \mathbf{u}_A - \mathbf{c}_{\mathbf{w}_A} \rangle = \alpha^*$, since $(\mathbf{u}_A, \mathbf{w}_A)$ is optimal for (\mathbf{D}_{π_E}) .

In addition, for all $\lambda \in \Lambda$ and $\mathbf{w} \in \mathcal{W}$

$$\mathcal{L}_r(\theta_A, \lambda, \mathbf{w}) = \langle \mu_{\pi_A} - \mu_{\pi_E}, \mathbf{c}_{\mathbf{w}} - \mathbf{T}_\gamma^* \mathbf{u}_\lambda \rangle \quad (35)$$

$$= \langle \mu_{\pi_A} - \mu_{\pi_E}, \mathbf{c}_{\mathbf{w}} \rangle \quad (36)$$

$$\leq \max_{\mathbf{w}' \in \mathcal{W}} \langle \mu_{\pi_A} - \mu_{\pi_E}, \mathbf{c}_{\mathbf{w}'} \rangle \quad (37)$$

$$= \delta_{\mathcal{C}}(\pi_A, \pi_E) = \alpha^*, \quad (38)$$

where (36) holds because $\mathbf{T}_\gamma \mu_{\pi_A} = \mathbf{T}_\gamma \mu_{\pi_E} = \nu_0$, and (38) holds because π_A is optimal for (Lfd_{π_E}) .

Finally, by (33), (34) and (38) we conclude that (32) holds. □

C. Gradient Estimators and their Properties

We will describe formally the unbiased gradient estimates used in Algorithm 1 and study their properties. Our optimization variable is $\mathbf{z} = (\theta, \lambda, \mathbf{w}) \in \mathbf{Z}$, where $\mathbf{Z} \triangleq \Theta \times \Lambda \times \mathcal{W}$ is the decision space. The empirical Lagrangian that we will optimize is $\widehat{\mathcal{L}}_r(\theta, \lambda, \mathbf{w}) \triangleq \langle \mu_\theta, \mathbf{c}_{\mathbf{w}} - \mathbf{T}_\gamma^* \mathbf{u}_\lambda + C_{\beta, \gamma} \cdot \mathbf{1} \rangle + \langle \nu_0, \mathbf{u}_\lambda \rangle - \langle \mathbf{w}, \rho_{\mathbf{C}}(\widehat{\pi_E}) + 2 \cdot \mathbf{1} \rangle$.

Definition 1 (Gradient estimators). *We define $\mathbf{g} \triangleq [\mathbf{g}_\theta^\top \quad -\mathbf{g}_\lambda^\top \quad -\mathbf{g}_w^\top]^\top$ as follows.*

1. Let $\hat{\xi} \triangleq (\hat{l}, \hat{x}, \hat{a}, \hat{y}) \in \hat{\Xi} \triangleq [n_\mu] \times \mathcal{X} \times \mathcal{A} \times \mathcal{X}$ such that $\hat{l} \sim \text{Unif}([n_\mu])$, $(\hat{x}, \hat{a}) \sim \phi_{\hat{l}}$, $\hat{y} \sim P(\cdot | \hat{x}, \hat{a})$.

We define the θ -gradient estimator $\mathbf{g}_\theta : \mathbf{Z} \times \hat{\Xi} \rightarrow \mathbf{R}^{n_\mu}$ by

$$\mathbf{g}_{\theta,i}(\mathbf{z}, \hat{\xi}) \triangleq \begin{cases} \frac{n_\mu \left((1-\gamma) \mathbf{c}_{(\hat{x}, \hat{a})}^\top \mathbf{w} - (\psi_{\hat{x}} - \gamma \psi_{\hat{y}})^\top \boldsymbol{\lambda} + 2\beta \right)}{1-\gamma}, & i = \hat{l}, \\ 0, & \text{otherwise.} \end{cases}$$

2. Let $\tilde{\xi} \triangleq (\tilde{l}, \tilde{x}, \tilde{a}, \tilde{y}, \tilde{x}') \in \tilde{\Xi} \triangleq [n_\mu] \times \mathcal{X} \times \mathcal{A} \times \mathcal{X} \times \mathcal{X}$ such that $\tilde{l} \sim \boldsymbol{\theta}$, $(\tilde{x}, \tilde{a}) \sim \phi_{\tilde{l}}$, $\tilde{y} \sim P(\cdot | \tilde{x}, \tilde{a})$, $\tilde{x}' \sim \nu_0$. We define the λ -gradient estimator $\mathbf{g}_\lambda : \mathbf{Z} \times \tilde{\Xi} \rightarrow \mathbf{R}^{n_u}$ by $\mathbf{g}_\lambda(\mathbf{z}, \tilde{\xi}) \triangleq \psi_{\tilde{x}'} - \frac{\psi_{\tilde{x}} - \gamma \psi_{\tilde{y}}}{1-\gamma}$.

3. Let $\bar{\xi} \triangleq (\bar{l}, \bar{x}, \bar{a}) \in \bar{\Xi} \triangleq [n_\mu] \times \mathcal{X} \times \mathcal{A}$ such that $\bar{l} \sim \boldsymbol{\theta}$, $(\bar{x}, \bar{a}) \sim \phi_{\bar{l}}$. We define the \mathbf{w} -gradient estimator $\mathbf{g}_\mathbf{w} : \mathbf{Z} \times \bar{\Xi} \rightarrow \mathbf{R}^{n_c}$ by $\mathbf{g}_\mathbf{w}(\mathbf{z}, \bar{\xi}) \triangleq \mathbf{c}_{(\bar{x}, \bar{a})} - \rho_{\mathbf{C}}(\widehat{\pi}_{\mathbf{E}}) - 2 \cdot \mathbf{1}$.

Lemma 7. For all $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \in \Theta \times \Lambda \times \mathcal{W}$ it holds that $\boldsymbol{\mu}_\theta \in \Delta_{\mathcal{X} \times \mathcal{A}}$, $\|\mathbf{u}_\lambda\|_\infty \leq \beta$ and $\|\mathbf{c}_\mathbf{w}\|_\infty \leq 1$.

Proof. We will show the first inequality. By assumption $\|\psi_i\|_\infty \leq 1$, for all $i \in [n_u]$. Thus, $\|\mathbf{u}_\lambda\|_\infty = \|\sum_{i=1}^{n_u} \lambda_i \psi_i\|_\infty \leq \|\boldsymbol{\lambda}\|_1 \max_{i \in [n_u]} \|\psi_i\|_\infty \leq \beta$, where we used that $\|\boldsymbol{\lambda}\|_1 \leq \sqrt{n_u} \|\boldsymbol{\lambda}\|_2 \leq \beta$. \square

Lemma 8. For all $(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \in \Theta \times \Lambda \times \mathcal{W}$ it holds that $\|\mathbf{c}_\mathbf{w} - \mathbf{T}_\gamma^* \mathbf{u}_\lambda\|_\infty \leq C_{\beta, \gamma} \triangleq \frac{2\beta}{1-\gamma}$ and $\|\mathbf{T}_\gamma \boldsymbol{\mu}_\theta - \nu_0\|_1 \leq \frac{2}{1-\gamma}$.

Proof. Indeed, for all $\boldsymbol{\lambda} \in \Lambda$ and $\mathbf{w} \in \mathcal{W}$, we have that

$$\|\mathbf{c}_\mathbf{w} - \mathbf{T}_\gamma^* \mathbf{u}_\lambda\|_\infty = \|\mathbf{c}_\mathbf{w} - \frac{1}{1-\gamma} (\mathbf{B} - \gamma \mathbf{P}) \mathbf{u}_\lambda\|_\infty \leq \|\mathbf{c}_\mathbf{w}\|_\infty + \frac{1}{1-\gamma} (\|\mathbf{B}\|_\infty + \gamma \|\mathbf{P}\|_\infty) \|\mathbf{u}_\lambda\|_\infty \leq C_{\beta, \gamma},$$

where we used the triangle inequality, Lemma 7 and that $\|\mathbf{B}\|_\infty = \|\mathbf{P}\|_\infty = 1$. Similarly, for all $\boldsymbol{\theta} \in \Theta$, we have

$$\|\mathbf{T}_\gamma \boldsymbol{\mu}_\theta - \nu_0\|_1 = \|\nu_0 - \frac{1}{1-\gamma} (\mathbf{B} - \gamma \mathbf{P})^\top \boldsymbol{\mu}_\theta\|_1 \leq \|\nu_0\|_1 + \frac{1}{1-\gamma} (\|\mathbf{B}\|_\infty + \gamma \|\mathbf{P}\|_\infty) \|\boldsymbol{\mu}_\theta\|_1 \leq \frac{2}{1-\gamma}.$$

\square

Lemma 9. It holds that $\nabla_{\boldsymbol{\theta}} \widehat{\mathcal{L}}_r(\mathbf{z}) \geq \mathbf{0}$ and $\nabla_{\mathbf{w}} \widehat{\mathcal{L}}_r(\mathbf{z}) \leq \mathbf{0}$, for all $\mathbf{z} \in \mathbf{Z}$. Moreover, $\mathbf{g}_\theta(\mathbf{z}, \hat{\xi}) \geq \mathbf{0}$ and $\mathbf{g}_\mathbf{w}(\mathbf{z}, \bar{\xi}) \leq \mathbf{0}$ pointwise.

Proof. By Lemma 8, we have that $\mathbf{c}_\mathbf{w} - \mathbf{T}_\gamma^* \mathbf{u}_\lambda + C_{\beta, \gamma} \cdot \mathbf{1} \geq \mathbf{0}$. Since in addition all the entries of Φ are nonnegative, we conclude that $\nabla_{\boldsymbol{\theta}} \widehat{\mathcal{L}}_r(\mathbf{z}) \geq \mathbf{0}$. Moreover, $\frac{\partial}{\partial w_i} \widehat{\mathcal{L}}_r(\mathbf{z}) = \langle \mathbf{c}_i, \boldsymbol{\mu}_\theta \rangle - \rho_{\mathbf{c}_i}(\widehat{\pi}_{\mathbf{E}}) - 2 \leq \|\mathbf{c}_i\|_\infty \|\boldsymbol{\mu}_\theta\|_1 + 1 - 2 \leq 0$, for each $i \in [n_c]$, where we used Hölder's inequality and Lemma 7. Similarly, we can show that $\mathbf{g}_\theta(\mathbf{z}, \hat{\xi}) \geq \mathbf{0}$ and $\mathbf{g}_\mathbf{w}(\mathbf{z}, \bar{\xi}) \leq \mathbf{0}$. \square

The following lemma shows that $\mathbf{g} = [\mathbf{g}_\theta^\top \quad -\mathbf{g}_\lambda^\top \quad -\mathbf{g}_\mathbf{w}^\top]^\top$ is an unbiased estimator of the monotone operator $\mathbf{G}(\cdot)$ and establishes useful bounds for the error analysis of Algorithm 1.

Lemma 10. Let $\Xi \triangleq \hat{\Xi} \times \tilde{\Xi} \times \bar{\Xi}$ and let $\mathbf{g} : \mathbf{Z} \times \Xi \rightarrow \mathbf{R}^{n_\mu + n_u + n_c}$ given by $\mathbf{g}(\mathbf{z}, \xi) \triangleq [\mathbf{g}_\theta(\mathbf{z}, \hat{\xi})^\top \quad -\mathbf{g}_\lambda(\mathbf{z}, \tilde{\xi})^\top \quad -\mathbf{g}_\mathbf{w}(\mathbf{z}, \bar{\xi})^\top]^\top$ as in Definition 1. The following statements are true.

1. **[unbiased gradient estimates]** It holds that $\mathbf{E}_\xi [\mathbf{g}_\theta(\mathbf{z}, \hat{\xi})] = \nabla_{\boldsymbol{\theta}} \widehat{\mathcal{L}}_r(\mathbf{z})$, $\mathbf{E}_\xi [\mathbf{g}_\lambda(\mathbf{z}, \tilde{\xi})] = \nabla_{\boldsymbol{\lambda}} \widehat{\mathcal{L}}_r(\mathbf{z})$, and $\mathbf{E}_{\bar{\xi}} [\mathbf{g}_\mathbf{w}(\mathbf{z}, \bar{\xi})] = \nabla_{\mathbf{w}} \widehat{\mathcal{L}}_r(\mathbf{z})$, for all $\mathbf{z} \in \mathbf{Z}$.
2. **[$\|\cdot\|_\infty$ -norm bounds]** It holds that $\|\mathbf{g}_\theta(\mathbf{z}, \hat{\xi})\|_\infty \leq \frac{4\beta n_\mu}{1-\gamma}$, $\|\mathbf{g}_\lambda(\mathbf{z}, \tilde{\xi})\|_\infty \leq \frac{4\beta}{1-\gamma}$, and $\|\mathbf{g}_\mathbf{w}(\mathbf{z}, \bar{\xi})\|_\infty \leq 4$, for all $\mathbf{z} \in \mathbf{Z}$, and for all $\xi \in \Xi$.
3. **[dual and local-norm bounds]** It holds that $\|\mathbf{g}_\theta(\mathbf{z}, \hat{\xi})\|_{\theta'}^2 \leq \frac{16\beta^2 n_\mu^2}{(1-\gamma)^2}$, $\|\mathbf{g}_\lambda(\mathbf{z}, \tilde{\xi})\|_2^2 \leq \frac{4n_u}{(1-\gamma)^2}$, and $\|\mathbf{g}_\mathbf{w}(\mathbf{z}, \bar{\xi})\|_{\mathbf{w}'}^2 \leq 16$, for all $\mathbf{z} \in \mathbf{Z}$, for all $\xi \in \Xi$, for all $\theta' \in \Theta$, and for all $\mathbf{w}' \in \mathcal{W}$.

4. **[dual and local-norm second-moment bounds]** It holds that $\mathbf{E}_{\hat{\xi}} \left[\|\mathbf{g}_{\theta}(\mathbf{z}, \hat{\xi})\|_{\theta'}^2 \right] \leq \frac{16\beta^2 n_{\mu}}{(1-\gamma)^2}$, $\mathbf{E}_{\tilde{\xi}} \left[\|\mathbf{g}_{\lambda}(\mathbf{z}, \tilde{\xi})\|_2^2 \right] \leq \frac{4n_u}{(1-\gamma)^2}$, and $\mathbf{E}_{\tilde{\xi}} \left[\|\mathbf{g}_{\mathbf{w}}(\mathbf{z}, \tilde{\xi})\|_{\mathbf{w}'}^2 \right] \leq 16$, for all $\mathbf{z} \in \mathbf{Z}$, for all $\theta' \in \Theta$, and for all $\mathbf{w}' \in \mathcal{W}$.

Proof. 1. For all $i \in [n_{\mu}]$, we have that

$$\begin{aligned} \mathbf{E}_{\hat{\xi}} \left[\mathbf{g}_{\theta, i}(\mathbf{z}, \hat{\xi}) \right] &= \sum_{\hat{i}=1}^{n_{\mu}} \sum_{\hat{x}, \hat{a}, \hat{y}} \frac{1}{n_{\mu}} \phi_{\hat{i}}(\hat{x}, \hat{a}) P(\hat{y}|\hat{x}, \hat{a}) \mathbf{g}_{\theta, i}(\mathbf{z}, \hat{\xi}) \\ &= \sum_{\hat{x}, \hat{a}, \hat{y}} \frac{1}{n_{\mu}} \phi_{\hat{i}}(\hat{x}, \hat{a}) P(\hat{y}|\hat{x}, \hat{a}) \frac{n_{\mu} \left((1-\gamma) \mathbf{c}_{(\hat{x}, \hat{a})}^{\top} \mathbf{w} - (\psi_{\hat{x}} - \gamma \psi_{\hat{y}})^{\top} \boldsymbol{\lambda} + 2\beta \right)}{1-\gamma} \\ &= \langle \phi_{\hat{i}}, \mathbf{c}_{\mathbf{w}} - \mathbf{T}_{\gamma}^* \mathbf{u}_{\boldsymbol{\lambda}} + C_{\beta, \gamma} \cdot \mathbf{1} \rangle = \frac{\partial}{\partial \theta_i} \widehat{\mathcal{L}}_r(\mathbf{z}). \end{aligned}$$

Moreover, we have that

$$\begin{aligned} \mathbf{E}_{\tilde{\xi}} \left[\mathbf{g}_{\lambda}(\mathbf{z}, \tilde{\xi}) \right] &= \sum_{i=1}^{n_{\mu}} \sum_{x, a, y, x'} \theta_i \phi_i(x, a) P(y|x, a) \nu_0(x') \left(\psi_{x'} - \frac{\psi_x - \gamma \psi_y}{1-\gamma} \right) \\ &= \underbrace{\sum_{x'} \nu_0(x') \psi_{x'}}_{=\Psi^{\top} \nu_0} - \frac{1}{1-\gamma} \underbrace{\left(\sum_{x, a} \mu_{\theta}(x, a) \psi_x - \gamma \sum_{x, a} \mu_{\theta}(x, a) \sum_y P(y|x, a) \psi_y \right)}_{=\Psi^{\top} \mathbf{B}^{\top} \mu_{\theta}} \\ &= \Psi^{\top} (\nu_0 - \mathbf{T}_{\gamma} \mu_{\theta}) = \nabla_{\boldsymbol{\lambda}} \widehat{\mathcal{L}}_r(\mathbf{z}). \end{aligned}$$

Finally, we have that

$$\mathbf{E}_{\tilde{\xi}} \left[\mathbf{g}_{\mathbf{w}}(\mathbf{z}, \tilde{\xi}) \right] = \sum_{i=1}^{n_{\mu}} \sum_{x, a} \theta_i \phi_i(x, a) \mathbf{c}_{(x, a)} - \boldsymbol{\rho}_{\mathbf{C}}(\widehat{\pi}_{\mathbf{E}}) - 2 \cdot \mathbf{1} = \mathbf{C}^{\top} \mu_{\theta} - \boldsymbol{\rho}_{\mathbf{C}}(\widehat{\pi}_{\mathbf{E}}) - 2 \cdot \mathbf{1} = \nabla_{\mathbf{w}} \widehat{\mathcal{L}}_r(\mathbf{z}).$$

2. It holds that

$$\|\mathbf{g}_{\theta}(\mathbf{z}, \hat{\xi})\|_{\infty} = \left| n_{\mu} \left(\mathbf{c}_{\mathbf{w}}(\hat{x}, \hat{a}) - \frac{\mathbf{u}_{\boldsymbol{\lambda}}(\hat{x}) - \gamma \mathbf{u}_{\boldsymbol{\lambda}}(\hat{y}) - 2\beta}{1-\gamma} \right) \right| \leq \frac{4\beta n_{\mu}}{1-\gamma},$$

where we used the triangle inequality and the bounds in Lemma 7. Moreover, it holds that

$$\|\mathbf{g}_{\lambda}(\mathbf{z}, \tilde{\xi})\|_{\infty} = \left\| \psi_{\hat{x}'} - \frac{\psi_{\hat{x}} - \gamma \psi_{\hat{y}}}{1-\gamma} \right\|_{\infty} \leq \frac{2}{1-\gamma},$$

where we used the triangle inequality and that $\|\psi_x\|_{\infty} \leq 1$, for all $x \in \mathcal{X}$. Finally, it holds that

$$\|\mathbf{g}_{\mathbf{w}}(\mathbf{z}, \tilde{\xi})\|_{\infty} = \|\mathbf{c}_{(\bar{x}, \bar{a})} - \boldsymbol{\rho}_{\mathbf{C}}(\widehat{\pi}_{\mathbf{E}}) - 2 \cdot \mathbf{1}\|_{\infty} \leq 4,$$

where we used the triangle inequality, that $\|\mathbf{c}_{(\bar{x}, \bar{a})}\|_{\infty} \leq 1$, and that $\|\boldsymbol{\rho}_{\mathbf{C}}(\widehat{\pi}_{\mathbf{E}})\|_{\infty} \leq 1$.

3. Since $\theta' \in \Theta = \Delta_{[n_{\mu}]}$, we have that

$$\|\mathbf{g}_{\theta}(\mathbf{z}, \hat{\xi})\|_{\theta'}^2 \leq \|\mathbf{g}_{\theta}(\mathbf{z}, \hat{\xi})\|_{\infty}^2 \leq \frac{16\beta^2 n_{\mu}^2}{(1-\gamma)^2}.$$

Moreover,

$$\|\mathbf{g}_{\lambda}(\mathbf{z}, \tilde{\xi})\|_2^2 = \left\| \psi_{\hat{x}'} - \frac{\psi_{\hat{x}} - \gamma \psi_{\hat{y}}}{1-\gamma} \right\|_2^2 \leq \frac{4n_u}{(1-\gamma)^2},$$

where we used the triangle inequality and that $\|\psi_x\|_2 \leq \sqrt{n_u}$, for all $x \in \mathcal{X}$. Finally, since $\mathbf{w}' \in \mathcal{W} = \Delta_{[n_c]}$, we have

$$\|\mathbf{g}_{\mathbf{w}}(\mathbf{z}, \tilde{\xi})\|_{\mathbf{w}'}^2 \leq \|\mathbf{g}_{\mathbf{w}}(\mathbf{z}, \tilde{\xi})\|_{\infty}^2 \leq 16.$$

4. We need only to prove the first inequality. We have,

$$\begin{aligned} \mathbf{E}_{\hat{\xi}} \left[\|\mathbf{g}_{\theta}(\mathbf{z}, \hat{\xi})\|_{\theta'}^2 \right] &= \mathbf{E}_{\hat{\xi}} \left[\mathbf{g}_{\theta, i}^2(\mathbf{z}, \hat{\xi}) \theta'_i \right] \\ &= \sum_{i=1}^{n_{\mu}} \sum_{\hat{x}, \hat{a}, \hat{y}} \frac{1}{n_{\mu}} \phi_i(\hat{x}, \hat{a}) P(\hat{y}|\hat{x}, \hat{a}) \mathbf{g}_{\theta, i}^2(\mathbf{z}, \hat{\xi}) \theta'_i \\ &\leq \frac{1}{n_{\mu}} \|\mathbf{g}_{\theta}(\mathbf{z}, \hat{\xi})\|_{\infty}^2 \sum_{i=1}^n \theta'_i \leq \frac{16\beta^2 n_{\mu}}{(1-\gamma)^2}. \end{aligned}$$

□

D. Proof of Theorem 1

We are now going to prove Theorem 1. Our optimization variable is $\mathbf{z} = (\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \in \mathbf{Z}$, where $\mathbf{Z} \triangleq \Theta \times \Lambda \times \mathcal{W}$ is the decision space. The update rule of stochastic primal-dual mirror descent in Algorithm 1 is

$$\mathbf{z}_{n+1} = \arg \min_{\mathbf{z} \in \mathbf{Z}} \left(\langle \mathbf{g}(\mathbf{z}_n, \xi_n), \mathbf{z} \rangle + \frac{1}{\eta} B_R(\mathbf{z} \| \mathbf{z}_n) \right), \quad (39)$$

where $\mathbf{g}(\mathbf{z}, \xi) \triangleq [\mathbf{g}_{\theta}(\mathbf{z}, \hat{\xi})^{\top} \quad -\mathbf{g}_{\lambda}(\mathbf{z}, \tilde{\xi})^{\top} \quad -\mathbf{g}_{\mathbf{w}}(\mathbf{z}, \bar{\xi})^{\top}]^{\top}$ is given by Definition 1, and B_R is the following *Bregman divergence*

$$B_R(\mathbf{z} \| \mathbf{z}') \triangleq \text{KL}(\boldsymbol{\theta} \| \boldsymbol{\theta}') + \frac{n_{\mu}}{2\beta^2} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2^2 + \text{KL}(\mathbf{w} \| \mathbf{w}'). \quad (40)$$

Theorem 1. Let $\hat{\pi}_N$ be the output of running Algorithm 1 for $N = \max \left\{ \mathcal{O} \left(\frac{\beta^2 n_{\mu} \log(\frac{1}{\delta})}{(1-\gamma)^2 \varepsilon^2} \right), \mathcal{O} \left(\frac{\beta \sqrt{n_{\mu}^3 \log(\frac{1}{\delta})}}{(1-\gamma) \varepsilon} \right) \right\}$ iterations, with $m = \frac{8 \log(\frac{4n_{\mu}}{\delta})}{\varepsilon^2}$ expert trajectories of length $H = \frac{1}{1-\gamma} \log(\frac{2}{\varepsilon})$, and learning rate $\eta = \frac{1-\gamma}{\beta \sqrt{N n_{\mu}}}$. Then, with probability $1 - \delta$ it holds that $\rho_{\mathbf{c}_{\text{true}}}(\hat{\pi}_N) - \rho_{\mathbf{c}_{\text{true}}}(\pi_{\mathbf{E}}) \leq \varepsilon + \varepsilon_{\text{approx}, \hat{\boldsymbol{\theta}}_N} + \alpha^*$. If in addition Assumption 1 is satisfied, then with probability $1 - \delta$, it holds that $\rho_{\mathbf{c}_{\text{true}}}(\hat{\pi}_N) - \rho_{\mathbf{c}_{\text{true}}}(\pi_{\mathbf{E}}) \leq \varepsilon + \varepsilon_{\text{approx}, \mathbf{w}} + \alpha^*$.

Note that because the definition of the total expected cost is scaled by a factor $(1-\gamma)$, we finally have sample complexity $\max \left\{ \mathcal{O} \left(\frac{\beta^2 n_{\mu} \log(\frac{1}{\delta})}{(1-\gamma)^4 \varepsilon^2} \right), \mathcal{O} \left(\frac{\beta \sqrt{n_{\mu}^3 \log(\frac{1}{\delta})}}{(1-\gamma)^2 \varepsilon} \right) \right\}$ for the conventional discounted MDP setting. Similarly, the expert sample complexity is $m = \frac{8 \log(\frac{4n_{\mu}}{\delta})}{(1-\gamma)^2 \varepsilon^2}$. The sample complexity scales linearly with the number of features and does not depend on the number of state action spaces.

Proof of Theorem 1. By Proposition 4 applied to $\hat{\mathbf{z}}_N = (\hat{\boldsymbol{\theta}}_N, \hat{\boldsymbol{\lambda}}_N, \hat{\mathbf{w}}_N)$ and $\hat{\pi}_N \triangleq \pi_{\boldsymbol{\mu}_{\hat{\boldsymbol{\theta}}_N}}$ we get

$$\rho_{\mathbf{c}_{\text{true}}}(\hat{\pi}_N) - \rho_{\mathbf{c}_{\text{true}}}(\pi_{\mathbf{E}}) \leq \epsilon_{\text{sad}}(\hat{\mathbf{z}}_N) + \varepsilon_{\text{approx}, \hat{\boldsymbol{\theta}}_N} + \alpha^*.$$

where the approximation error $\varepsilon_{\text{approx}, \hat{\boldsymbol{\theta}}_N}$ is no larger than $\frac{2}{1-\gamma} (\beta \min_{\boldsymbol{\theta} \in \Theta} \|\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\pi_{\Lambda}}\|_1 + \min_{\boldsymbol{\lambda} \in \Lambda} \|\mathbf{u}_{\boldsymbol{\lambda}} - \mathbf{V}_{\mathbf{w}_N^*}^{\hat{\pi}_N}\|_{\infty})$, and $\mathbf{w}_N^* \triangleq \arg \max_{\mathbf{w} \in \mathcal{W}} [\rho_{\mathbf{w}}(\hat{\pi}_N) - \rho_{\mathbf{w}}(\pi_{\mathbf{E}})]$.

Moreover, under Assumption 1, it holds that

$$\rho_{\mathbf{c}_{\text{true}}}(\pi_{\boldsymbol{\theta}}) - \rho_{\mathbf{c}_{\text{true}}}(\pi_{\mathbf{E}}) \leq 3\epsilon_{\text{sad}}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) + \varepsilon_{\text{approx}, \mathbf{w}} + \alpha^*.$$

So, in both cases we need to determine how many samples N and expert trajectories m are needed so that $\epsilon_{\text{sad}}(\hat{\mathbf{z}}_N) \leq \varepsilon$

with probability at least $1 - \delta$. We have,

$$\epsilon_{\text{sad}}(\widehat{\mathbf{z}}_N) = \max_{(\boldsymbol{\lambda}, \mathbf{w}) \in \Lambda \times \mathcal{W}} \mathcal{L}_r(\widehat{\boldsymbol{\theta}}_N, \boldsymbol{\lambda}, \mathbf{w}) - \min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}_r(\boldsymbol{\theta}, \widehat{\boldsymbol{\lambda}}_N, \widehat{\mathbf{w}}_N) \quad (41)$$

$$= \max_{(\boldsymbol{\lambda}, \mathbf{w}) \in \Lambda \times \mathcal{W}} \bar{\mathcal{L}}_r(\widehat{\boldsymbol{\theta}}_N, \boldsymbol{\lambda}, \mathbf{w}) - \min_{\boldsymbol{\theta} \in \Theta} \bar{\mathcal{L}}_r(\boldsymbol{\theta}, \widehat{\boldsymbol{\lambda}}_N, \widehat{\mathbf{w}}_N) \quad (42)$$

$$= \max_{(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \in \mathbf{Z}} \left[\widehat{\mathcal{L}}_r(\widehat{\boldsymbol{\theta}}_N, \boldsymbol{\lambda}, \mathbf{w}) - \widehat{\mathcal{L}}_r(\boldsymbol{\theta}, \widehat{\boldsymbol{\lambda}}_N, \widehat{\mathbf{w}}_N) + \langle \boldsymbol{\rho}_{\mathbf{C}}(\widehat{\pi}_{\mathbf{E}}) - \boldsymbol{\rho}_{\mathbf{C}}(\pi_{\mathbf{E}}), \mathbf{w} - \widehat{\mathbf{w}}_N \rangle \right] \quad (43)$$

$$\leq \max_{(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \in \mathbf{Z}} \left[\widehat{\mathcal{L}}_r(\widehat{\boldsymbol{\theta}}_N, \boldsymbol{\lambda}, \mathbf{w}) - \widehat{\mathcal{L}}_r(\boldsymbol{\theta}, \widehat{\boldsymbol{\lambda}}_N, \widehat{\mathbf{w}}_N) + \|\boldsymbol{\rho}_{\mathbf{C}}(\widehat{\pi}_{\mathbf{E}}) - \boldsymbol{\rho}_{\mathbf{C}}(\pi_{\mathbf{E}})\|_{\infty} \|\mathbf{w} - \widehat{\mathbf{w}}_N\|_1 \right] \quad (44)$$

$$\leq \max_{(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w}) \in \mathbf{Z}} \left[\widehat{\mathcal{L}}_r(\widehat{\boldsymbol{\theta}}_N, \boldsymbol{\lambda}, \mathbf{w}) - \widehat{\mathcal{L}}_r(\boldsymbol{\theta}, \widehat{\boldsymbol{\lambda}}_N, \widehat{\mathbf{w}}_N) \right] + 2\|\boldsymbol{\rho}_{\mathbf{C}}(\widehat{\pi}_{\mathbf{E}}) - \boldsymbol{\rho}_{\mathbf{C}}(\pi_{\mathbf{E}})\|_{\infty}, \quad (45)$$

where we used Hölder's inequality in (44) and the triangle inequality in (45).

By virtue of Lemma 1 in (Syed & Schapire, 2007) we have that for $\|\boldsymbol{\rho}_{\mathbf{C}}(\widehat{\pi}_{\mathbf{E}}) - \boldsymbol{\rho}_{\mathbf{C}}(\pi_{\mathbf{E}})\|_{\infty} \leq \frac{\varepsilon}{4}$ to hold with probability at least $1 - \frac{\delta}{2}$, it suffices that $m \geq \frac{8 \log(\frac{4n\varepsilon}{\delta})}{\varepsilon^2}$ and $H \geq \frac{1}{1-\gamma} \log(\frac{2}{\varepsilon})$. Note that Lemma 1 in (Syed & Schapire, 2007) is a variant of the Hoeffding's inequality (Theorem 2.5 in (McDiarmid, 1998)), since we have to take into account that $\boldsymbol{\rho}_{\mathbf{C}}(\widehat{\pi}_{\mathbf{E}})$ is a biased estimate of $\boldsymbol{\rho}_{\mathbf{C}}(\pi_{\mathbf{E}})$ because the trajectories are truncated. By combining this last bound with (45), we get that if $m \geq \frac{8 \log(\frac{4n\varepsilon}{\delta})}{\varepsilon^2}$ and $H \geq \frac{1}{1-\gamma} \log(\frac{2}{\varepsilon})$, then with probability at least $1 - \frac{\delta}{2}$ it holds that

$$\epsilon_{\text{sad}}(\widehat{\mathbf{z}}_N) \leq \widehat{\epsilon}_{\text{sad}}(\widehat{\mathbf{z}}_N) + \frac{\varepsilon}{2}, \quad (46)$$

where

$$\widehat{\epsilon}_{\text{sad}}(\widehat{\mathbf{z}}_N) \triangleq \max_{(\boldsymbol{\lambda}, \mathbf{w}) \in \Lambda \times \mathcal{W}} \widehat{\mathcal{L}}_r(\widehat{\boldsymbol{\theta}}_N, \boldsymbol{\lambda}, \mathbf{w}) - \min_{\boldsymbol{\theta} \in \Theta} \widehat{\mathcal{L}}_r(\boldsymbol{\theta}, \widehat{\boldsymbol{\lambda}}_N, \widehat{\mathbf{w}}_N)$$

is the empirical saddle-point residual for $\widehat{\mathbf{z}}_N$. All in all, we need $\widehat{\epsilon}_{\text{sad}}(\widehat{\mathbf{z}}_N) \leq \frac{\varepsilon}{2}$ with confidence $1 - \frac{\delta}{2}$.

By linearity of $\widehat{\mathcal{L}}_r(\cdot, \boldsymbol{\lambda}, \mathbf{w})$ we have that for all $n \in \mathbb{N}$, and for all $\boldsymbol{\theta} \in \Theta$,

$$\widehat{\mathcal{L}}_r(\boldsymbol{\theta}_n, \boldsymbol{\lambda}_n, \mathbf{w}_n) - \widehat{\mathcal{L}}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}_n, \mathbf{w}_n) = \left\langle \nabla_{\boldsymbol{\theta}} \widehat{\mathcal{L}}_r(\boldsymbol{\theta}_n, \boldsymbol{\lambda}_n, \mathbf{w}_n), \boldsymbol{\theta}_n - \boldsymbol{\theta} \right\rangle. \quad (47)$$

Similarly, by linearity of $\widehat{\mathcal{L}}_r(\boldsymbol{\theta}, \cdot, \cdot)$ we have that for all $n \in \mathbb{N}$, and for all $(\boldsymbol{\lambda}, \mathbf{w}) \in \Lambda \times \mathcal{W}$,

$$\widehat{\mathcal{L}}_r(\boldsymbol{\theta}_n, \boldsymbol{\lambda}, \mathbf{w}) - \widehat{\mathcal{L}}_r(\boldsymbol{\theta}_n, \boldsymbol{\lambda}_n, \mathbf{w}_n) = \left\langle \nabla_{\boldsymbol{\lambda}} \widehat{\mathcal{L}}_r(\boldsymbol{\theta}_n, \boldsymbol{\lambda}_n, \mathbf{w}_n), \boldsymbol{\lambda} - \boldsymbol{\lambda}_n \right\rangle + \left\langle \nabla_{\mathbf{w}} \widehat{\mathcal{L}}_r(\boldsymbol{\theta}_n, \boldsymbol{\lambda}_n, \mathbf{w}_n), \mathbf{w} - \mathbf{w}_n \right\rangle. \quad (48)$$

Therefore,

$$\widehat{\epsilon}_{\text{sad}}(\widehat{\mathbf{z}}_N) = \frac{1}{N} \left(\max_{(\boldsymbol{\lambda}, \mathbf{w}) \in \Lambda \times \mathcal{W}} \sum_{n=1}^N \widehat{\mathcal{L}}_r(\boldsymbol{\theta}_n, \boldsymbol{\lambda}, \mathbf{w}) - \min_{\boldsymbol{\theta} \in \Theta} \sum_{n=1}^N \widehat{\mathcal{L}}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}_n, \mathbf{w}_n) \right) = \frac{1}{N} \max_{\mathbf{z} \in \mathbf{Z}} \sum_{n=1}^N \langle \mathbf{G}(\mathbf{z}_n), \mathbf{z}_n - \mathbf{z} \rangle,$$

where the first equality follows once more by linearity of $\widehat{\mathcal{L}}_r(\cdot, \boldsymbol{\lambda}, \mathbf{w})$ and $\widehat{\mathcal{L}}_r(\boldsymbol{\theta}, \cdot, \cdot)$, and the second equality by (47) and (48).

Consider now the sequence of *ghost iterates* $(\bar{\mathbf{z}}_n)_{n=1}^N$ in \mathbf{Z} generated by taking gradient step with $\mathbf{G}(\mathbf{z}_n) - \mathbf{g}(\mathbf{z}_n, \xi_n)$ coupled with each iteration, i.e.,

$$\bar{\mathbf{z}}_1 = \mathbf{z}_1, \quad \bar{\mathbf{z}}_{n+1} = \arg \min_{\bar{\mathbf{z}} \in \mathbf{Z}} \left(\langle \mathbf{G}(\mathbf{z}_n) - \mathbf{g}(\mathbf{z}_n, \xi_n), \bar{\mathbf{z}} \rangle + \frac{1}{\eta} B_R(\bar{\mathbf{z}} | \bar{\mathbf{z}}_n) \right),$$

where B_R is the same Bregman divergence as in (40).

We then have

$$\max_{\mathbf{z} \in \mathbf{Z}} \sum_{n=1}^N \langle \mathbf{G}(\mathbf{z}_n), \mathbf{z}_n - \mathbf{z} \rangle \leq \underbrace{\sum_{n=1}^N \langle \Delta_n, \mathbf{z}_n - \bar{\mathbf{z}}_n \rangle}_{\triangleq T_1} + \underbrace{\max_{\mathbf{z} \in \mathbf{Z}} \sum_{n=1}^N \langle \mathbf{g}_n, \mathbf{z}_n - \mathbf{z} \rangle}_{\triangleq T_2} + \underbrace{\max_{\bar{\mathbf{z}} \in \mathbf{Z}} \sum_{n=1}^N \langle \Delta_n, \bar{\mathbf{z}}_n - \bar{\mathbf{z}} \rangle}_{\triangleq T_3}, \quad (49)$$

where

$$\mathbf{g}_n \triangleq \mathbf{g}(\mathbf{z}_n, \xi_n) \quad \text{and} \quad \Delta_n \triangleq \mathbf{G}(\mathbf{z}_n) - \mathbf{g}(\mathbf{z}_n, \xi_n).$$

Moreover, we set for brevity $\mathbf{g}_{n,\theta} \triangleq \mathbf{g}_\theta(\mathbf{z}_n, \hat{\xi}_n)$, $\mathbf{g}_{n,\lambda} \triangleq \mathbf{g}_\lambda(\mathbf{z}_n, \tilde{\xi}_n)$, $\mathbf{g}_{n,\mathbf{w}} \triangleq \mathbf{g}_\mathbf{w}(\mathbf{z}_n, \bar{\xi}_n)$, $\Delta_{n,\theta} \triangleq \nabla_{\theta} \widehat{\mathcal{L}}_r(\mathbf{z}_n) - \mathbf{g}_\theta(\mathbf{z}_n, \hat{\xi}_n)$, $\Delta_{n,\lambda} \triangleq \nabla_{\lambda} \widehat{\mathcal{L}}_r(\mathbf{z}_n) - \mathbf{g}_\lambda(\mathbf{z}_n, \tilde{\xi}_n)$, and $\Delta_{n,\mathbf{w}} \triangleq \nabla_{\mathbf{w}} \widehat{\mathcal{L}}_r(\mathbf{z}_n) - \mathbf{g}_\mathbf{w}(\mathbf{z}_n, \bar{\xi}_n)$.

We are now going to bound each term in (49) separately.

Bounding the term T_1 . Let $\xi_n = (\hat{\xi}_n, \tilde{\xi}_n, \bar{\xi}_n)$ be the generated sample at round n and let $\mathcal{F}_n = \sigma(\xi_1, \dots, \xi_n)$ be the σ -algebra generated by the history process. The iterates $\mathbf{z}_n = \mathbf{z}_n(\xi_1, \dots, \xi_{n-1})$ and $\bar{\mathbf{z}}_n = \bar{\mathbf{z}}_n(\xi_1, \dots, \xi_{n-1})$ are functions of ξ_1, \dots, ξ_{n-1} (and thus they are random). Since in addition, \mathbf{z}_n and $\bar{\mathbf{z}}_n$ are independent of the randomness of ξ_n (we sample ξ_n after the definition of \mathbf{z}_n and $\bar{\mathbf{z}}_n$), we get by Lemma 10 that the sequence of random variables $Y_n = \langle \Delta_n, \mathbf{z}_n - \bar{\mathbf{z}}_n \rangle$ forms a *martingale difference*, i.e., Y_n is \mathcal{F}_n -measurable and $\mathbf{E}[Y_n | \mathcal{F}_{n-1}] = 0$. In order to bound the term $T_1 = \sum_{n=1}^N Y_n$, we will use a Bernstein martingale concentration inequality (Theorem 3.15 in (McDiarmid, 1998)). In particular, for any $\varepsilon \geq 0$,

$$\mathbf{P} \left[\sum_{n=1}^N Y_n \geq \varepsilon \right] \leq \exp \left(\frac{-\varepsilon^2}{2N\sigma^2 \left(1 + \frac{b\varepsilon}{3N\sigma^2}\right)} \right), \quad (50)$$

where $|Y_n| \leq b$, for all $n \in \mathbb{N}$ and $\mathbf{Var}[Y_n | \mathcal{F}_{n-1}] \leq \sigma^2$, for all $n \in \mathbb{N}$. We think of the term $\frac{b\varepsilon}{3N\sigma^2}$ as a negligible error term. In order to use (50) we need to compute the constants b and σ^2 . We have

$$Y_n = \langle \Delta_n, \mathbf{z}_n - \bar{\mathbf{z}}_n \rangle = \langle \Delta_{n,\theta}, \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n \rangle - \langle \Delta_{n,\lambda}, \boldsymbol{\lambda}_n - \bar{\boldsymbol{\lambda}}_n \rangle - \langle \Delta_{n,\mathbf{w}}, \mathbf{w}_n - \bar{\mathbf{w}}_n \rangle.$$

For the term $\langle \Delta_{n,\theta}, \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n \rangle$ we have

$$\begin{aligned} |\langle \Delta_{n,\theta}, \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n \rangle| &\leq |\langle \mathbf{c}_{\mathbf{w}_n} - \mathbf{T}_\gamma^* \mathbf{u}_{\boldsymbol{\lambda}_n} + C_{\beta,\gamma} \cdot \mathbf{1}, \boldsymbol{\Phi} \boldsymbol{\theta}_n - \boldsymbol{\Phi} \bar{\boldsymbol{\theta}}_n \rangle| + |\langle \mathbf{g}_{n,\theta}, \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n \rangle| && \text{(By triangle inequality)} \\ &\leq \|\mathbf{c}_{\mathbf{w}_n} - \mathbf{T}_\gamma^* \mathbf{u}_{\boldsymbol{\lambda}_n} + C_{\beta,\gamma} \cdot \mathbf{1}\|_\infty \|\boldsymbol{\mu}_{\boldsymbol{\theta}_n} - \boldsymbol{\mu}_{\bar{\boldsymbol{\theta}}_n}\|_1 + \|\mathbf{g}_{n,\theta}\|_\infty \|\boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n\|_1 && \text{(By Hölder's inequality)} \\ &\leq \frac{8\beta}{1-\gamma} + \frac{8\beta n_\mu}{1-\gamma} = \frac{8\beta(n_\mu + 1)}{1-\gamma}. && \text{(By Lemmas 7, 8 and 10)} \end{aligned}$$

Moreover, for the term $\langle \Delta_{n,\lambda}, \boldsymbol{\lambda}_n - \bar{\boldsymbol{\lambda}}_n \rangle$, we have

$$\begin{aligned} |-\langle \Delta_{n,\lambda}, \boldsymbol{\lambda}_n - \bar{\boldsymbol{\lambda}}_n \rangle| &\leq |\langle \boldsymbol{\nu}_0 - \mathbf{T}_\gamma \boldsymbol{\mu}_{\boldsymbol{\theta}_n}, \boldsymbol{\Psi} \boldsymbol{\lambda}_n - \boldsymbol{\Psi} \bar{\boldsymbol{\lambda}}_n \rangle| + |\langle \mathbf{g}_{n,\lambda}, \boldsymbol{\lambda}_n - \bar{\boldsymbol{\lambda}}_n \rangle| && \text{(By triangle inequality)} \\ &\leq \|\boldsymbol{\nu}_0 - \mathbf{T}_\gamma \boldsymbol{\mu}_{\boldsymbol{\theta}_n}\|_1 \|\mathbf{u}_{\boldsymbol{\lambda}_n} - \mathbf{u}_{\bar{\boldsymbol{\lambda}}_n}\|_\infty + \|\mathbf{g}_{n,\lambda}\|_\infty \|\boldsymbol{\lambda}_n - \bar{\boldsymbol{\lambda}}_n\|_1 && \text{(By Hölder's inequality)} \\ &\leq \frac{8\beta}{1-\gamma}. && \text{(By Lemmas 7, 8 and 10)} \end{aligned}$$

Finally, for the term $\langle \Delta_{n,\mathbf{w}}, \mathbf{w}_n - \bar{\mathbf{w}}_n \rangle$, we have

$$\begin{aligned} |-\langle \Delta_{n,\mathbf{w}}, \mathbf{w}_n - \bar{\mathbf{w}}_n \rangle| &= |\langle \mathbf{c}_{(\bar{x}, \bar{a})}, \mathbf{w}_n - \bar{\mathbf{w}}_n \rangle + \langle \boldsymbol{\mu}_{\boldsymbol{\theta}_n}, \mathbf{C} \mathbf{w}_n - \mathbf{C} \bar{\mathbf{w}}_n \rangle| && \text{(By Definition 1)} \\ &\leq \|\mathbf{c}_{(\bar{x}, \bar{a})}\|_\infty \|\mathbf{w}_n - \bar{\mathbf{w}}_n\|_1 + \|\boldsymbol{\mu}_{\boldsymbol{\theta}_n}\|_1 \|\mathbf{c}_{\mathbf{w}_n} - \mathbf{c}_{\bar{\mathbf{w}}_n}\|_\infty && \text{(By Hölder's inequality)} \\ &\leq 4. && \text{(By Lemma 7)} \end{aligned}$$

All in all, $|Y_n| \leq b \triangleq \frac{8\beta(n_\mu+2)}{1-\gamma} + 4$.

We are now going to bound the conditional variance of Y_n . We have

$$\text{Var} [\langle \Delta_{n,\theta}, \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n \rangle | \mathcal{F}_{n-1}] = \mathbf{E} \left[\langle \Delta_{n,\theta}, \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n \rangle^2 | \mathcal{F}_{n-1} \right] \quad (51)$$

$$= \mathbf{E} \left[\langle \mathbf{g}_{n,\theta}, \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n \rangle^2 | \mathcal{F}_{n-1} \right] - \left\langle \nabla_{\boldsymbol{\theta}} \hat{\mathcal{L}}_r(\mathbf{z}_n), \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n \right\rangle^2 \quad (52)$$

$$\leq \mathbf{E} \left[\langle \mathbf{g}_{n,\theta}, \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n \rangle^2 | \mathcal{F}_{n-1} \right] \quad (53)$$

$$= \mathbf{E} \left[\mathbf{g}_{n,\theta,\hat{\iota}_n}^2 (\theta_{n,\hat{\iota}_n} - \bar{\theta}_{n,\hat{\iota}_n})^2 | \mathcal{F}_{n-1} \right] \quad (54)$$

$$= \sum_{i=1}^{n_\mu} \sum_{x,a,y} \frac{1}{n_\mu} \phi_i(x,a) P(y|x,a) \mathbf{g}_{n,\theta,i}^2 (\theta_{n,i} - \bar{\theta}_{n,i})^2 \quad (55)$$

$$\leq \frac{1}{n_\mu} \|\mathbf{g}_{n,\theta}\|_\infty^2 \|\boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n\|_2^2 \quad (56)$$

$$\leq \frac{1}{n_\mu} \|\mathbf{g}_{n,\theta}\|_\infty^2 \|\boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n\|_1^2 \quad (57)$$

$$\leq \frac{64\beta^2 n_\mu}{(1-\gamma)^2}, \quad (58)$$

where in (51) we used that $\mathbf{E}[\langle \Delta_{n,\theta}, \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n \rangle | \mathcal{F}_{n-1}] = 0$, in (52) we used that the random variables $\mathbf{z}_n - \bar{\mathbf{z}}_n$ and $\boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n$ are \mathcal{F}_{n-1} -measurable and that $\mathbf{E}[\mathbf{g}_{n,\theta} | \mathcal{F}_{n-1}] = \nabla_{\boldsymbol{\theta}} \hat{\mathcal{L}}_r(\mathbf{z}_n)$, in (55) and (56) we used Definition 1, in (56) we used that $\phi_i \in \Delta_{\mathcal{X} \times \mathcal{A}}$ and $P(\cdot|x,a) \in \Delta_{\mathcal{X}}$, in (57) we used that $\|\boldsymbol{\theta}\|_2 \leq \|\boldsymbol{\theta}\|_1$, and in (58) we used the triangle inequality and Lemma 10.

By similar arguments, we have

$$\text{Var}[\langle \Delta_{n,\lambda}, \boldsymbol{\lambda}_n - \bar{\boldsymbol{\lambda}}_n \rangle | \mathcal{F}_{n-1}] \leq \mathbf{E} \left[\langle \mathbf{g}_{n,\lambda}, \boldsymbol{\lambda}_n - \bar{\boldsymbol{\lambda}}_n \rangle^2 | \mathcal{F}_{n-1} \right] \leq \|\mathbf{g}_{n,\lambda}\|_\infty^2 \|\boldsymbol{\lambda}_n - \bar{\boldsymbol{\lambda}}_n\|_1^2 \leq \frac{16\beta^2}{(1-\gamma^2)}. \quad (59)$$

Finally,

$$\text{Var}[\langle \Delta_{n,\mathbf{w}}, \mathbf{w}_n - \bar{\mathbf{w}}_n \rangle | \mathcal{F}_{n-1}] \leq \mathbf{E} \left[\langle \mathbf{g}_{n,\mathbf{w}}, \mathbf{w}_n - \bar{\mathbf{w}}_n \rangle^2 | \mathcal{F}_{n-1} \right] \leq \|\mathbf{g}_{n,\mathbf{w}}\|_\infty^2 \|\mathbf{w}_n - \bar{\mathbf{w}}_n\|_1^2 \leq 64. \quad (60)$$

Now, since $\hat{\xi}_n, \tilde{\xi}_n, \bar{\xi}_n$ are independent and both \mathbf{z}_n and $\bar{\mathbf{z}}_n$ are \mathcal{F}_{n-1} -measurable, we get that $\langle \Delta_{n,\theta}, \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n \rangle$, $\langle \Delta_{n,\lambda}, \boldsymbol{\lambda}_n - \bar{\boldsymbol{\lambda}}_n \rangle$, and $\langle \Delta_{n,\mathbf{w}}, \mathbf{w}_n - \bar{\mathbf{w}}_n \rangle$ are independent given \mathcal{F}_{n-1} . Therefore,

$$\begin{aligned} \text{Var}[Y_n | \mathcal{F}_{n-1}] &= \text{Var}[\langle \Delta_{n,\theta}, \boldsymbol{\theta}_n - \bar{\boldsymbol{\theta}}_n \rangle | \mathcal{F}_{n-1}] + \text{Var}[\langle \Delta_{n,\lambda}, \boldsymbol{\lambda}_n - \bar{\boldsymbol{\lambda}}_n \rangle | \mathcal{F}_{n-1}] + \text{Var}[\langle \Delta_{n,\mathbf{w}}, \mathbf{w}_n - \bar{\mathbf{w}}_n \rangle | \mathcal{F}_{n-1}] \\ &\leq \sigma^2 \triangleq \frac{16\beta^2(4n_\mu + 1)}{(1-\gamma)^2} + 64. \end{aligned}$$

Therefore by the Bernstein martingale concentration inequality (50) we have with probability at least $1 - \delta$

$$\sum_{n=1}^N Y_n \leq \sqrt{2N\sigma^2(1+o(1)) \log\left(\frac{1}{\delta}\right)} = \mathcal{O}\left(\frac{\beta\sqrt{Nn_\mu \log\left(\frac{1}{\delta}\right)}}{1-\gamma}\right). \quad (61)$$

Bounding the term T_2 . By Lemma 9, we know that $\mathbf{g}_{n,\theta} \geq \mathbf{0}$ and $-\mathbf{g}_{n,\mathbf{w}} \leq \mathbf{0}$, for all $n \in \mathbb{N}$. Therefore, by Lemma 10 and Lemma 11 in (Jin & Sidford, 2020)³ for a fixed $\mathbf{z} \in \mathbf{Z}$, we have that

$$\sum_{n=1}^N \langle \mathbf{g}_n, \mathbf{z}_n - \mathbf{z} \rangle \leq \frac{1}{\eta} B_R(\mathbf{z}|\mathbf{z}_1) + \sum_{n=1}^N \left[\langle \mathbf{g}_n, \mathbf{z}_n - \mathbf{z}_{n+1} \rangle - \frac{1}{\eta} B_R(\mathbf{z}_{n+1}|\mathbf{z}_n) \right] \quad (62)$$

$$\leq \frac{1}{\eta} B_R(\mathbf{z}|\mathbf{z}_1) + \frac{\eta}{2} \sum_{n=1}^N \left[\|\mathbf{g}_{n,\theta}\|_{\boldsymbol{\theta}_n}^2 + \frac{\beta^2}{n_u} \|\mathbf{g}_{n,\lambda}\|_2^2 + \|\mathbf{g}_{n,\mathbf{w}}\|_{\mathbf{w}_n}^2 \right]. \quad (63)$$

³Note that Lemma 11 in (Jin & Sidford, 2020) still holds for $\gamma_t \geq -1$ since the inequality $e^{-a} \leq 1 - a + a^2$ holds for all $a > -1$

Now, we define

$$Z_n = \left[\|\mathbf{g}_{n,\theta}\|_{\boldsymbol{\theta}_n}^2 + \frac{\beta^2}{n_u} \|\mathbf{g}_{n,\lambda}\|_2^2 + \|\mathbf{g}_{n,\mathbf{w}}\|_{\mathbf{w}_n}^2 \right] - \mathbf{E} \left[\|\mathbf{g}_{n,\theta}\|_{\boldsymbol{\theta}_n}^2 + \frac{\beta^2}{n_u} \|\mathbf{g}_{n,\lambda}\|_2^2 + \|\mathbf{g}_{n,\mathbf{w}}\|_{\mathbf{w}_n}^2 \mid \mathcal{F}_{n-1} \right].$$

Clearly Z_n is an \mathcal{F}_n -martingale difference. Therefore, by the Azuma-Hoeffding inequality (Theorem 3.14 in (McDiarmid, 1998)), with probability at least $1 - \delta$, it holds that

$$\sum_{n=1}^N Z_n \leq \sqrt{\frac{N \log\left(\frac{1}{\delta}\right)}{2}} \cdot M_1, \quad (64)$$

where $|Z_n| \leq M_1$, for all $n \in \mathbb{N}$.

Putting them all together we get that with probability at least $1 - \delta$

$$\max_{\mathbf{z} \in \mathbf{Z}} \sum_{n=1}^N \langle \mathbf{g}_n, \mathbf{z}_n - \mathbf{z} \rangle \leq \underbrace{\frac{1}{\eta} \max_{\mathbf{z} \in \mathbf{Z}} B_R(\mathbf{z} \mid \mathbf{z}_1)}_{\leq M_2} + \underbrace{\frac{\eta}{2} \sum_{n=1}^N \mathbf{E} \left[\|\mathbf{g}_{n,\theta}\|_{\boldsymbol{\theta}_n}^2 + \frac{\beta^2}{n_u} \|\mathbf{g}_{n,\lambda}\|_2^2 + \|\mathbf{g}_{n,\mathbf{w}}\|_{\mathbf{w}_n}^2 \mid \mathcal{F}_{n-1} \right]}_{\leq M_3} + \frac{\eta M_1}{2} \sqrt{\frac{N \log\left(\frac{1}{\delta}\right)}{2}}. \quad (65)$$

By using the gradient bounds in Lemma 10, we get

$$M_3 \triangleq N \left(\frac{16\beta^2 n_\mu}{(1-\gamma)^2} + \frac{\beta^2}{n_u} \frac{4n_u}{(1-\gamma)^2} + 16 \right) = \mathcal{O} \left(\frac{\beta^2 n_\mu N}{(1-\gamma)^2} \right)$$

and

$$M_1 \triangleq 2 \left(\frac{16\beta^2 n_\mu^2}{(1-\gamma)^2} + \frac{\beta^2}{n_u} \frac{4n_u}{(1-\gamma)^2} + 16 \right) = \mathcal{O} \left(\frac{\beta^2 n_\mu^2}{(1-\gamma)^2} \right).$$

Moreover, by considering the Bregman divergence given in (40) and that $\mathbf{z}_1 = (\frac{1}{n_\mu} \mathbf{1}, \mathbf{0}, \frac{1}{n_c} \mathbf{1})$, we get that

$$M_2 \triangleq \left(\log n_\mu + \log n_c + \frac{1}{2} \right).$$

Therefore, by setting $\eta \triangleq \frac{1-\gamma}{\beta \sqrt{N n_\mu}}$ we get by (65) that with probability at least $1 - \delta$

$$\max_{\mathbf{z} \in \mathbf{Z}} \sum_{n=1}^N \langle \mathbf{g}_n, \mathbf{z}_n - \mathbf{z} \rangle \leq \mathcal{O} \left(\frac{\beta \sqrt{N n_\mu}}{1-\gamma} \right) + \mathcal{O} \left(\frac{\beta \sqrt{N n_\mu}}{1-\gamma} \right) + \mathcal{O} \left(\frac{\beta \sqrt{n_\mu^3 \log\left(\frac{1}{\delta}\right)}}{1-\gamma} \right). \quad (66)$$

Bounding the term T_3 . By the choice of $\eta \triangleq \frac{1-\gamma}{\beta \sqrt{N n_\mu}}$ and Lemma 10 we get that for all n , it holds that $\|\eta \Delta_{n,\theta}\|_\infty \leq 1$ and $\|\eta \Delta_{n,\mathbf{w}}\|_\infty \leq 1$. Indeed, note for example that by Jensen's inequality for conditional expectations and Lemma 10, we have $\|\eta \Delta_{n,\theta}\|_\infty \leq 2\eta \|\mathbf{g}_{n,\theta}\|_\infty \leq 8\sqrt{\frac{n_\mu}{N}} \leq 1$, for sufficiently large N . Therefore, once more we can combine regret bounds of stochastic mirror descent with local norms (Lemma 10 and Lemma 11 in (Jin & Sidford, 2020)) and the Azuma-Hoeffding inequality (Theorem 3.14 in (McDiarmid, 1998)) to end up that with probability at least $1 - \delta$

$$\max_{\bar{\mathbf{z}} \in \bar{\mathbf{Z}}} \sum_{n=1}^N \langle \Delta_n, \bar{\mathbf{z}}_n - \bar{\mathbf{z}} \rangle \leq \underbrace{\frac{1}{\eta} \max_{\bar{\mathbf{z}} \in \bar{\mathbf{Z}}} B_R(\bar{\mathbf{z}} \mid \bar{\mathbf{z}}_1)}_{\leq m_2} + \underbrace{\frac{\eta}{2} \sum_{n=1}^N \mathbf{E} \left[\|\Delta_{n,\theta}\|_{\boldsymbol{\theta}_n}^2 + \frac{\beta^2}{n_u} \|\Delta_{n,\lambda}\|_2^2 + \|\Delta_{n,\mathbf{w}}\|_{\mathbf{w}_n}^2 \mid \mathcal{F}_{n-1} \right]}_{\leq m_3} + \frac{\eta m_1}{2} \sqrt{\frac{N \log\left(\frac{1}{\delta}\right)}{2}}, \quad (67)$$

where

$$\left[\|\Delta_{n,\theta}\|_{\boldsymbol{\theta}_n}^2 + \frac{\beta^2}{n_u} \|\Delta_{n,\lambda}\|_2^2 + \|\Delta_{n,\mathbf{w}}\|_{\mathbf{w}_n}^2 \right] - \mathbf{E} \left[\|\Delta_{n,\theta}\|_{\boldsymbol{\theta}_n}^2 + \frac{\beta^2}{n_u} \|\Delta_{n,\lambda}\|_2^2 + \|\Delta_{n,\mathbf{w}}\|_{\mathbf{w}_n}^2 \mid \mathcal{F}_{n-1} \right] \leq m_1.$$

By using similar arguments as the ones regarding the term T_2 , we get that with probability at least $1 - \delta$

$$\max_{\bar{\mathbf{z}} \in \mathbf{Z}} \sum_{n=1}^N \langle \Delta_n, \bar{\mathbf{z}}_n - \bar{\mathbf{z}} \rangle \leq \mathcal{O} \left(\frac{\beta \sqrt{N n_\mu}}{1 - \gamma} \right) + \mathcal{O} \left(\frac{\beta \sqrt{N n_\mu}}{1 - \gamma} \right) + \mathcal{O} \left(\frac{\beta \sqrt{n_\mu^3 \log(\frac{1}{\delta})}}{1 - \gamma} \right). \quad (68)$$

Putting them all together. Combining the bounds for the three terms T_1 , T_2 and T_3 above and using a union bound, we get that with probability at least $1 - \delta$

$$\hat{\epsilon}_{\text{sad}}(\hat{\mathbf{z}}_N) \leq \mathcal{O} \left(\frac{\beta \sqrt{n_\mu \log(\frac{1}{\delta})}}{(1 - \gamma) \sqrt{N}} \right) + \mathcal{O} \left(\frac{\beta \sqrt{n_\mu^3 \log(\frac{1}{\delta})}}{(1 - \gamma) N} \right). \quad (69)$$

Thus the sample complexity of mirror descent is $N = \max \left\{ \mathcal{O} \left(\frac{\beta^2 n_\mu \log(\frac{1}{\delta})}{(1 - \gamma)^2 \varepsilon^2} \right), \mathcal{O} \left(\frac{\beta \sqrt{n_\mu^3 \log(\frac{1}{\delta})}}{(1 - \gamma) \varepsilon} \right) \right\}$.

□

E. Numerical Experiments

In this section we analyze performance of the policy extracted from a saddle point of the Lagrangian

$$\mathcal{L}(\boldsymbol{\mu}, \mathbf{u}, \mathbf{w}) = \langle \boldsymbol{\mu}, \mathbf{c}_\mathbf{w} - \mathbf{T}_\gamma^* \mathbf{u} \rangle - \langle \mathbf{w}, \boldsymbol{\rho}_\mathbf{C}(\widehat{\pi}_\mathbf{E}) \rangle + \langle \boldsymbol{\nu}_0, \mathbf{u} \rangle$$

as a function of the number of trajectories m used to compute the empirical feature expectation vector $\boldsymbol{\rho}_\mathbf{C}(\widehat{\pi}_\mathbf{E})$. The length of each trajectory is set to $H = 200$. These preliminary results do not consider approximation of $\boldsymbol{\mu}$ and \mathbf{u} by linear combinations of features.

Our environment is a 16-by-16 gridworld with state space $\mathcal{X} = \{(i, j) \mid i \in [16], j \in [16]\}$ and action space $\mathcal{A} = \{\text{'up'}, \text{'down'}, \text{'left'}, \text{'right'}\}$. The initial state distribution is uniform over \mathcal{X} . When an action is played, the executed direction will be the same as the intended with probability 0.7 and any of the remaining directions with probability 0.1. If the executed direction does not take the agent outside the grid, then the agent will move in the intended direction; otherwise, it will remain at the same cell.

We generate 5 cost vectors as $\mathbf{c}_i = \bar{\mathbf{c}}_i \otimes \mathbf{1}$ with $\mathbf{1} \in \mathbf{R}^{|\mathcal{A}|}$, where \otimes denotes the Kronecker product and elements of $\bar{\mathbf{c}}_i \in \mathbf{R}^{|\mathcal{X}|}$ are drawn from the set $\{0, -1, 1\}$ with probabilities $\{0.8, 0.1, 0.1\}$, respectively. The expert policy is an optimal one for cost vector $\mathbf{c}_{\text{true}} = \mathbf{c}_1$.

Figure E shows statistics for the performance of the learned policies depending on the number of sample trajectories m used to compute $\boldsymbol{\rho}_\mathbf{C}(\widehat{\pi}_\mathbf{E})$. For each m we performed 20 runs. The figure shows that, even for a modest number of sample trajectories, the performance of the extracted policy is close to that of the expert.

F. A No-Regret Online Learning View

An online learning problem (Shalev-Shwartz, 2012; Hazan, 2016) describes the iterative interactions between a learner and an opponent. In round n , the learner chooses a *decision* \mathbf{z}_n in the decision set \mathbf{Z} , the opponent chooses a *per-round loss function* $\ell_n : \mathbf{Z} \rightarrow \mathbf{R}$, and information about $\nabla \ell_n(\mathbf{z}_n)$ is revealed to the learner.

After N rounds, the goal is to minimize the *regret*

$$\text{Regret}_N(\mathbf{Z}) \triangleq \sum_{n=1}^N \ell_n(\mathbf{z}_n) - \min_{\mathbf{z} \in \mathbf{Z}} \sum_{n=1}^N \ell_n(\mathbf{z}).$$

In Appendix, we argue that solving the online learning problem with $\mathbf{Z} = \Theta \times \Lambda \times \mathcal{W}$ and per-round loss function

$$\begin{aligned} \ell_n(\mathbf{z}) &\triangleq \mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}_n, \mathbf{w}_n) - \mathcal{L}_r(\boldsymbol{\theta}_n, \boldsymbol{\lambda}, \mathbf{w}) \\ &= \langle \boldsymbol{\mu}_\boldsymbol{\theta} - \boldsymbol{\mu}_{\pi_\mathbf{E}}, \mathbf{c}_{\mathbf{w}_n} - \mathbf{T}_\gamma^* \mathbf{u}_{\boldsymbol{\lambda}_n} \rangle - \langle \boldsymbol{\mu}_{\boldsymbol{\theta}_n} - \boldsymbol{\mu}_{\pi_\mathbf{E}}, \mathbf{c}_\mathbf{w} - \mathbf{T}_\gamma^* \mathbf{u}_\boldsymbol{\lambda} \rangle, \end{aligned}$$

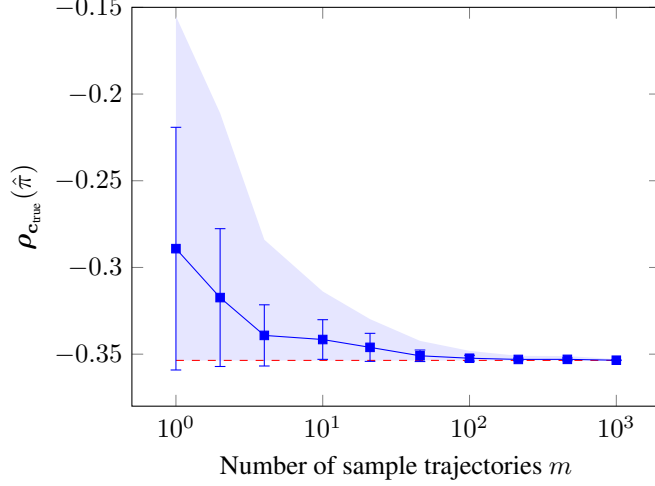


Figure 2. Performance of learned policies for a 16-by-16 gridworld environment. The dashed line shows $\rho_{c_{\text{true}}}(\pi_E)$. The error bars show the mean and standard deviation over 20 runs and the shaded area provides bounds for the computed policies.

is equivalent to solving the linearly-relaxed saddle-point problem (2).

Note that the per-round loss ℓ_n is linear in $\mathbf{z} = (\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{w})$. Moreover, the corresponding online learning problem is *continuous* and *predictable*. In other words, the per-round loss ℓ_n changes continuously with respect to the played decisions of the learner, and the environment is not completely adversarial since the transition matrix \mathbf{P} and the expert policy π_E remain the same across different rounds of online learning. A similar reduction of the forward RL problem to no-regret online learning has been investigated in (Cheng et al., 2020).

To see why the two problems are equivalent, first of all note that $\nabla_{\mathbf{z}} \ell_n(\mathbf{z}_n) = \mathbf{G}(\mathbf{z}_n)$. Moreover, by the bilinearity of the Lagrangian \mathcal{L} we get that

$$\begin{aligned}
 & \epsilon_{\text{sad}}(\hat{\boldsymbol{\theta}}_N, \hat{\boldsymbol{\lambda}}_N, \hat{\mathbf{w}}_N) \\
 &= \max_{\mathbf{z} \in \mathbf{Z}} \left[\mathcal{L}_r(\hat{\boldsymbol{\theta}}_N, \boldsymbol{\lambda}, \mathbf{w}) - \mathcal{L}_r(\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}_N, \hat{\mathbf{w}}_N) \right] \\
 &= \frac{1}{N} \max_{\mathbf{z} \in \mathbf{Z}} \sum_{i=1}^N [\mathcal{L}_r(\boldsymbol{\theta}_n, \boldsymbol{\lambda}, \mathbf{w}) - \mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{\lambda}_n, \mathbf{w}_n)] \\
 &= \frac{1}{N} \left(\sum_{n=1}^N \ell_n(\mathbf{z}_n) - \min_{\mathbf{z} \in \mathbf{Z}} \sum_{n=1}^N \ell_n(\mathbf{z}) \right) = \frac{1}{N} \text{Regret}_N(\mathbf{Z}).
 \end{aligned}$$

Therefore, any stochastic first-order primal-dual algorithm for the bilinear Lagrangian \mathcal{L} is equivalent to the corresponding online learning algorithm with first-order noisy feedback for the linear losses ℓ_n . Moreover, the average $\hat{\mathbf{z}}_N$ is an (ϵ, δ) -optimal saddle point to (2) if and only if the average regret $\frac{1}{N} \text{Regret}_N(\mathbf{Z}) < \epsilon$ with probability $1 - \delta$. In particular, Algorithm 1 can be seen as online mirror descent for the empirical per-round loss function $\hat{\ell}_n$.

It is worth noting that this online learning approach differs from the one in (Ross et al., 2011) where interaction with the expert is required. It is also different from the game-theoretic approach in (Syed & Schapire, 2007) where the forward RL problem has to be solved repeatedly.

G. Future Directions

There is an emerging body of literature (De Farias & Van Roy, 2003; Abbasi-Yadkori et al., 2014; Chen et al., 2018; Sutter et al., 2017; Lakshminarayanan et al., 2018; Mohajerin Esfahani et al., 2018; Wang, 2020; Lee & He, 2019; Banjac & Lygeros, 2019; Beuchat et al., 2020; Martinelli et al., 2020; Cheng et al., 2020; Jin & Sidford, 2020; Shariff & Szepesvári, 2020; Bas-Serrano et al., 2020; Bas-Serrano & Neu, 2020) that studies ALP for the forward RL. In this paper we analyzed

applied ALP techniques for the design of a provably efficient IRL algorithm. We hope that our techniques will be useful for future algorithm designers and will lay foundations for more exhaustive research in this direction. We discuss a few interesting directions.

- *Extension to continuous MDPs.* In this case the primal-dual LPs are infinite-dimensional. Indeed, the occupancy measures μ live in a measure space, the \mathbf{u} variables live in a function space, and the constraints are uncountably infinite. Also the vector norms are replaced by the total variation norm of measures and the sup-norm of functions. We have to impose appropriate regularity assumptions on the MDP model, so that the linear operators \mathbf{T}_γ and \mathbf{T}_γ^* are well-defined and continuous. Moreover, we have to show that strong duality holds, since this is not anymore straightforward. We expect, that these technical issues can be tackled by similar tools as for forward MDPs (Ch.6 in (Hernández-Lerma & Lasserre, 1996)). A more crucial point is that while one can adopt our approach, there is no analytical expression for the extracted policy $\pi_{\hat{\theta}_N}$ as for finite MDPs. A remedy for this difficulty could be to add a sparse solution promoting regularizer for the μ -variable as proposed e.g., in (Boyd et al., 2017).
- *Extension more general and other IRL settings.* Our methodology is naturally extensible to the more general case of RL with convex constraints (Miryoosefi et al., 2019). A less straightforward extension is IRL with risk-sensitive objectives (Brown et al., 2020a). We consider this as an exciting future direction.
- *Imitation learning as f -divergence minimization.* In this work we utilize Lagrangian duality and propose a primal-dual algorithm for the resulting bilinear minimax formulation of the LfD problem. On the other hand, (Nachum et al., 2019a; Nachum & Dai, 2020) regularize the objective of the LP formulation for the policy evaluation problem with an f -divergence, so that the Fenchel-Rockafellar dual is an unconstrained convex program over Q -functions. In this way, one avoids possible complications due to the nested min-max optimization in the Lagrangian (e.g., in our case the need of strong value function features), but is not able to use standard stochastic optimization techniques because the gradient estimates become biased. Moreover, one can derive an actor-critic-type method (Nachum et al., 2019b), in which a Q -value is learned to minimize the *squared Bellman error* and a policy π is learned to maximize the Q -values. The same reasoning can be applied for imitation learning with f -divergences (Ghasemipour et al., 2020; Kostrikov et al., 2020), which is a generalization of the maximum causal entropy IRL framework in (Ho & Ermon, 2016). While these algorithms are theoretically justified up to a point, they come without convergence guarantees. Indeed, the squared Bellman error is not convex and can lead to instabilities. A possible remedy is to use the convex and smooth *logistic Bellman error* introduced in (Bas-Serrano et al., 2020) which arises from a different regularized Q -LP formulation. This approach is quite intriguing, requires careful design and analysis, and will be the subject of future work.