A Dual Perspective of Reinforcement Learning for Imposing Policy Constraints

Bram De Cooman(✉), and Johan Suykens

Abstract. *Model-free reinforcement learning methods lack an inherent mechanism to impose behavioural constraints on the trained policies. While certain extensions exist, they remain limited to specific types of constraints, such as value constraints with additional reward signals or visitation density constraints. In this work we try to unify these existing techniques and bridge the gap with classical optimization and control theory, using a generic primal-dual framework for value-based and actor-critic reinforcement learning methods. The obtained dual formulations turn out to be especially useful for imposing additional constraints on the learned policy, as an intrinsic relationship between such dual constraints (or regularization terms) and reward modifications in the primal is reveiled. Furthermore, using this framework, we are able to introduce some novel types of constraints, allowing to impose bounds on the policy's action density or on costs associated with transitions between consecutive states and actions. From the adjusted primal-dual optimization problems, a practical algorithm is derived that supports various combinations of policy constraints that are automatically handled throughout training using trainable reward modifications. The resulting* `DualCRL` *method is examined in more detail and evaluated under different (combinations of) constraints on two interpretable environments. The results highlight the efficacy of the method, which ultimately provides the designer of such systems with a versatile toolbox of possible policy constraints.*

Impact Statement. *For complex control tasks it is often hard to design a suitable, scalar reward signal that leads to desirable controller behaviour for all relevant states. Manual tuning of the reward function to accommodate certain behaviour in a particular region of interest can be very time consuming and error-prone, as it can lead to unexpected or undesirable behaviour in other parts of the state-action space. A more practical and natural approach would thus be to keep the reward function as simple as possible, while imposing some additional constraints on the policy's behaviour. In this work we focus on dual formulations of popular reinforcement learning methods, where it is often easier to enforce such additional constraints. The resulting* `DualCRL` *algorithms can therefore lower the burden on designers of such systems, allowing them to specify policy constraints that are automatically taken care of throughout training.*

Keywords. *Model-Free Reinforcement Learning, Constraint Handling, Lagrange Duality, Linear Programming*

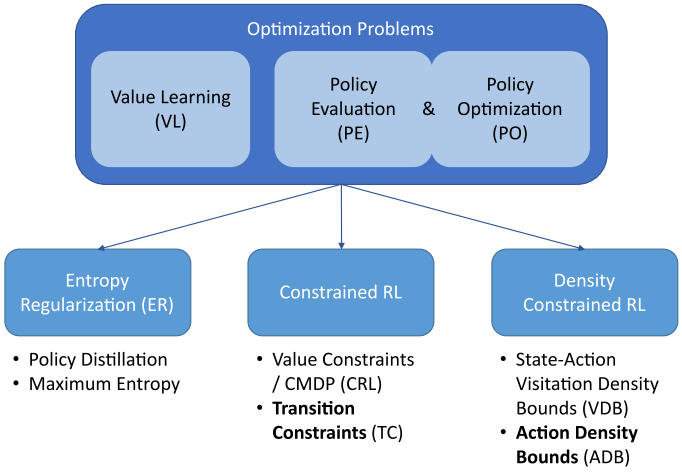

Figure 1: Overview of the discussed optimization problems in this paper and their relation to different reinforcement learning domains. Novel constraint types are indicated in bold.

1 Introduction

The field of convex optimization has a rich body of work on duality theory. Depending on the application, optimization problems can be simplified or more easily solved in either their primal or dual formulations and it is often more straightforward to extend the problem — by adding extra regularization terms to the objective function or extra constraints — in certain representations. Duality has also been explored in reinforcement learning literature, however it remains limited to applications within certain domains, such as constrained reinforcement learning [1, 2, 3, 4], offline reinforcement learning [5, 6, 7, 8] or density constrained reinforcement learning [9, 10]. In this paper, we aim at putting forward a more generic and complete primal-dual framework applicable to both value-based and actor-critic reinforcement learning methods. We show how the dual formulation is particularly useful to impose additional constraints on the learned policies and to move between seemingly different reinforcement learning fields. Figure 1 gives a general overview of the framework.

The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Advanced Grant E-DUALITY (787960). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

This research received funding from the Flemish Government (AI Research Program). Johan Suykens and Bram De Cooman are also affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium.

B. De Cooman (✉ bram.decooman@esat.kuleuven.be) and J. Suykens (✉ johan.suykens@esat.kuleuven.be) are with STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium.

Contributions Our main contributions can be summarized as follows. We provide a clear derivation of optimization problems for different reinforcement learning subdomains with policy constraints, for both value-based and actor-critic methods. Using this unifying primal-dual framework, we can provide novel insights into the intrinsic relationship between learnable reward modifications in the primal and policy constraints in the dual. Additionally, we extend the scala of policy constraints, by introducing novel transition and action density constraints. This offers the designer a broad toolbox of policy constraints, acting on different scopes in the trajectory.

Organization The necessary reinforcement learning essentials are briefly revised in Section 2, after which the basic optimization problems that are used as the building blocks of our framework are introduced in Section 3. These are then built upon and extended using various policy constraints in Section 4. This is followed by the introduction of the practical `DualCRL` algorithm in Section 5. Finally, in Section 6 the presented methods and constraints are investigated in more detail on two interpretable environments. The related work is discussed before the conclusion in Section 7.

2 Background

In this section we briefly revise the basic concepts of Reinforcement Learning (RL) used throughout this work, for a more in-depth introduction the reader is referred to Sutton and Barto [11] and Altman [12].

Reinforcement Learning We consider the Markov Decision Process (MDP) [13] $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \iota, \tau, r, \gamma)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, initial-state distribution $\iota(\boldsymbol{s}_0)$, state-transition distribution $\tau(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)$, reward function $r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})$ and discount factor $\gamma \in [0\,, 1)$. and a policy $\pi(\boldsymbol{a}|\boldsymbol{s})$ acting in this MDP. The expected discounted return of a policy $\pi(\boldsymbol{a}|\boldsymbol{s})$ acting in this MDP is given by $\nu^\pi = \mathbb{E}_{\iota,\pi,\tau}[\sum_{t=0}^{\infty} \gamma^t r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})]$, where the expectation is taken over $\boldsymbol{s}_0 \sim \iota(\cdot)$, $\boldsymbol{a}_t \sim \pi(\cdot|\boldsymbol{s}_t)$ and $\boldsymbol{s}_{t+1} \sim \tau(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)$ for all timesteps. The goal in reinforcement learning is then to find the optimal policy π^*, maximizing this expected discounted return, or (equivalently) the average reward $\bar{r}^\pi = (1-\gamma)\nu^\pi$,

$$\pi^* = \arg\max_\pi \mathop{\mathbb{E}}_{\iota,\pi,\tau}\left[\sum_{t=0}^{\infty} \gamma^t r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})\right]. \tag{1}$$

Constrained Reinforcement Learning Additional reward functions $r_k(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})$, with k an integer from the set $[K] = \{1, \dots, K\}$, can be added to enforce K additional constraints on the policy. This leads to a Constrained Markov Decision Process (CMDP) [12], where the objective is to find an optimal policy π^*, maximizing the expected discounted return, subject to the K extra constraints $\nu_k^\pi \geq V_k$, i.e.

$$\begin{aligned} &\max_\pi \mathop{\mathbb{E}}_{\iota,\pi,\tau}\left[\sum_{t=0}^{\infty} \gamma^t r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})\right] \\ &\text{s.t.} \mathop{\mathbb{E}}_{\iota,\pi,\tau}\left[\sum_{t=0}^{\infty} \gamma^t r_k(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})\right] \geq V_k \quad \forall k \end{aligned} \tag{2}$$

Entropy Regularized Reinforcement Learning The average reward objective can be extended with an extra regularization term, such as the learned (student) policy's entropy or the cross-entropy with another (teacher) policy π_{T}. The goal is still to maximize the average reward, while either acting as randomly as possible for the former *maximum entropy* methods [14], or extracting as much knowledge from the teacher as possible for the latter *policy distillation* methods [15]. Both objectives can also be combined by minimizing the KL-divergence between teacher and student policies, as $\mathrm{KL}[p \,||\, q] = \mathrm{H}(p, q) - \mathrm{H}(p)$, yielding the entropy regularized objective

$$\max_\pi \mathop{\mathbb{E}}_{\iota,\pi,\tau}\left[\sum_{t=0}^{\infty} \gamma^t r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}) - \alpha\, \mathrm{KL}\big[\pi(\cdot|\boldsymbol{s}_t) \,\big|\big|\, \pi_{\mathrm{T}}(\cdot|\boldsymbol{s}_t)\big]\right]. \tag{3}$$

Value Function The value of a policy π acting in the MDP for a certain state or state-action pair is given by $v^\pi(\boldsymbol{s}) = \mathbb{E}_{\pi,\tau}[\sum_{k=0}^{\infty} \gamma^k r(\boldsymbol{s}_{t+k}, \boldsymbol{a}_{t+k}, \boldsymbol{s}_{t+k+1}) \mid \boldsymbol{s}_t = \boldsymbol{s}]$ and $q^\pi(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}_{\pi,\tau}[\sum_{k=0}^{\infty} \gamma^k r(\boldsymbol{s}_{t+k}, \boldsymbol{a}_{t+k}, \boldsymbol{s}_{t+k+1}) \mid \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}]$. From these value functions, the expected discounted return can be easily obtained as $\nu^\pi = \mathbb{E}_{\boldsymbol{s}_0 \sim \iota(\cdot)}[v^\pi(\boldsymbol{s}_0)] = \mathbb{E}_{\boldsymbol{s}_0 \sim \iota(\cdot), \boldsymbol{a}_0 \sim \pi(\cdot|\boldsymbol{s}_0)}[q^\pi(\boldsymbol{s}_0, \boldsymbol{a}_0)]$. Both value functions satisfy recursive relationships, also known as the Bellman equations, $v^\pi(\boldsymbol{s}) = \mathbb{E}_{\pi,\tau}[r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}) + \gamma v^\pi(\boldsymbol{s}_{t+1}) \mid \boldsymbol{s}_t = \boldsymbol{s}]$ and $q^\pi(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}_{\pi,\tau}[r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}) + \gamma q^\pi(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) \mid \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}]$. The value functions for the optimal policy π^* are denoted by v^* and q^* and satisfy the Bellman optimality equations $v^*(\boldsymbol{s}) = \max_{\boldsymbol{a}} \mathbb{E}_{\pi,\tau}[r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}) + \gamma v^*(\boldsymbol{s}_{t+1}) \mid \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}]$ and $q^*(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}_{\pi,\tau}[r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}) + \gamma \max_{\boldsymbol{a}'} q^*(\boldsymbol{s}_{t+1}, \boldsymbol{a}') \mid \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}]$.

Visitation Density The visitation density (also referred to as the occupation or frequency) of a policy π acting in the MDP for a certain state is defined as $d^\pi(\boldsymbol{s}) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t P\big[\boldsymbol{s}_t = \boldsymbol{s} \,\big|\, \boldsymbol{s}_0 \sim \iota(\cdot), \forall t : \boldsymbol{a}_t \sim \pi(\cdot|\boldsymbol{s}_t), \boldsymbol{s}_{t+1} \sim \tau(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)\big]$, providing the discounted probability of visiting state $\boldsymbol{s}$ at any possible timestep. Using this definition, the average reward can be expressed as $\bar{r}^\pi = \mathbb{E}_{\boldsymbol{s} \sim d^\pi(\cdot), \boldsymbol{a} \sim \pi(\cdot|\boldsymbol{s}), \boldsymbol{s}' \sim \tau(\cdot|\boldsymbol{s}, \boldsymbol{a})}[r(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')]$. The visitation density also satisfies a recursive relation $d^\pi(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \sum_{\boldsymbol{s}' \in \mathcal{S}} \sum_{\boldsymbol{a}' \in \mathcal{A}} d^\pi(\boldsymbol{s}')\pi(\boldsymbol{a}'|\boldsymbol{s}')\tau(\boldsymbol{s}|\boldsymbol{s}', \boldsymbol{a}')$, and the visitation density of a policy for state-action pairs can be written as $p^\pi(\boldsymbol{s}, \boldsymbol{a}) = d^\pi(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s})$. To refer to the set of states visited by a policy π, we use the notation $\mathcal{S}^\pi = \{\boldsymbol{s} \in \mathcal{S} \,|\, d^\pi(\boldsymbol{s}) > 0\}$, whereas the set of actions taken by a policy in a certain state $\boldsymbol{s}$ is denoted by $\mathcal{A}^\pi(\boldsymbol{s}) = \{\boldsymbol{a} \in \mathcal{A} \mid \pi(\boldsymbol{a}|\boldsymbol{s}) > 0\}$.

Model-Free Reinforcement Learning In the model-free RL setting, several parameters of the optimization

problem (1), such as ι, τ and r are unknown, but we assume access to samples from these unknown distributions and functions. More specifically, a replay buffer $\mathcal{B}$ is available, containing experience samples $\{(\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1})\}$, where $\boldsymbol{s}_t \sim d^\beta(\cdot)$, $\boldsymbol{a}_t \sim \beta(\cdot|\boldsymbol{s}_t)$, $\boldsymbol{s}_{t+1} \sim \tau(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)$ and $r_t = r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1})$. For on-policy methods, the behavioral policy β is equal to the policy being learned π; for off-policy methods they might differ, for example to encourage sufficient exploration.

3 Optimization Perspective

In this work we analyze two optimization problems related to the various existing reinforcement learning algorithms to solve (1). More specifically, a Linear Programming (LP) formulation used for value learning and a minimax formulation used for policy learning are analyzed. The former formulation is related to the value iteration scheme and value-based RL methods, whereas the latter formulation is related to the policy iteration scheme and actor-critic RL methods. To simplify the derivations, only discrete and finite-dimensional state and action spaces are considered in this work, i.e. $\mathcal{S} = [S]$ and $\mathcal{A} = [A]$. The experiments in Section 6 illustrate the application to continuous state and action spaces.

3.1 Value Learning LP

Value-based reinforcement learning methods solve (1) indirectly by first finding the optimal value function $q^*(\boldsymbol{s}, \boldsymbol{a})$ and then obtaining an optimal (deterministic) policy as $\pi(\boldsymbol{a}|\boldsymbol{s}) = \delta_{\mathcal{A}^*(\boldsymbol{s})}(\boldsymbol{a})/|\mathcal{A}^*(\boldsymbol{s})|$ with $\mathcal{A}^*(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} q^*(\boldsymbol{s}, \boldsymbol{a})$ the set of optimal actions for state $\boldsymbol{s}$. The popular DQN method [16] (and its later extensions) follow such a strategy. This approach can be written as the following LP

$$\begin{aligned} \min_{\substack{v(\boldsymbol{s}) \\ q(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\boldsymbol{s}_0 \sim \iota(\cdot)} \big[(1-\gamma) v(\boldsymbol{s}_0)\big] \\ \text{s.t.} \quad & v(\boldsymbol{s}) \geq q(\boldsymbol{s}, \boldsymbol{a}) && \forall \boldsymbol{s}\, \forall \boldsymbol{a} \\ & q(\boldsymbol{s}, \boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}' \sim \tau(\cdot|\boldsymbol{s},\boldsymbol{a})} \big[r(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') + \gamma v(\boldsymbol{s}')\big] && \forall \boldsymbol{s}\, \forall \boldsymbol{a} \end{aligned} \tag{VL·P}$$

and has as dual

$$\begin{aligned} \max_{\substack{d(\boldsymbol{s}) \\ p(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a}) \sim p(\cdot,\cdot) \\ \boldsymbol{s}' \sim \tau(\cdot|\boldsymbol{s},\boldsymbol{a})}} \big[r(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')\big] \\ \text{s.t.} \quad & \sum_{\boldsymbol{a} \in \mathcal{A}} p(\boldsymbol{s}, \boldsymbol{a}) = d(\boldsymbol{s}) && \forall \boldsymbol{s} \\ & d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}') \sim p(\cdot,\cdot)} \big[\tau(\boldsymbol{s}|\boldsymbol{s}', \boldsymbol{a}')\big] && \forall \boldsymbol{s} \\ & p(\boldsymbol{s}, \boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\, \forall \boldsymbol{a} \end{aligned} \tag{VL·D}$$

This formulation is a slightly extended version of the primal and dual formulations commonly referred to as the V-LP [13]. The extra $q(\boldsymbol{s}, \boldsymbol{a})$ and $d(\boldsymbol{s})$ primal and dual decision variables, introduced here, could easily be eliminated but will prove to be useful when altering the optimization problems in Section 4. The following properties can be proven for this LP, using the optimality (KKT) conditions.

Theorem 1. *The policy that can be derived from the optimal dual variables, is an optimal policy for the considered MDP.*

Lemma 1. *The optimal primal variables satisfy the Bellman optimality equation and are thus equivalent to the optimal value functions v^* and q^* of the considered MDP.*

Appendix A.1 provides a proof and derivation of the Lagrangian objective $\mathcal{L}_{\text{VL}}$. The primal and dual objectives of (VL) clearly show the two equivalent formulations for the average reward (or expected discounted return) of the optimal policy

$$\bar{r}^* = (1-\gamma)\nu^* = \mathop{\mathbb{E}}_{\boldsymbol{s} \sim \iota(\cdot)} \big[(1-\gamma) v^*(\boldsymbol{s})\big] = \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a}) \sim p^*(\cdot,\cdot) \\ \boldsymbol{s}' \sim \tau(\cdot|\boldsymbol{s},\boldsymbol{a})}} \big[r(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')\big].$$

Note that it can be written as an equality because strong duality holds for LPs (there is no duality gap).

3.2 Policy Evaluation LP

A closely related linear optimization problem can be defined for the task of policy evaluation, where the goal is to find the value function q^π of a given (fixed) policy π. This can be done by solving the following LP

$$\begin{aligned} \min_{\substack{v(\boldsymbol{s}) \\ q(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\boldsymbol{s}_0 \sim \iota(\cdot)} \big[(1-\gamma) v(\boldsymbol{s}_0)\big] \\ \text{s.t.} \quad & v(\boldsymbol{s}) \geq \mathop{\mathbb{E}}_{\boldsymbol{a} \sim \pi(\cdot|\boldsymbol{s})} \big[q(\boldsymbol{s}, \boldsymbol{a})\big] && \forall \boldsymbol{s} \\ & q(\boldsymbol{s}, \boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}' \sim \tau(\cdot|\boldsymbol{s},\boldsymbol{a})} \big[r(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') + \gamma v(\boldsymbol{s}')\big] && \forall \boldsymbol{s}\, \forall \boldsymbol{a} \end{aligned} \tag{PE·P}$$

with dual

$$\begin{aligned} \max_{\substack{d(\boldsymbol{s}) \\ p(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a}) \sim p(\cdot,\cdot) \\ \boldsymbol{s}' \sim \tau(\cdot|\boldsymbol{s},\boldsymbol{a})}} \big[r(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')\big] \\ \text{s.t.} \quad & d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}') \sim p(\cdot,\cdot)} \big[\tau(\boldsymbol{s}|\boldsymbol{s}', \boldsymbol{a}')\big] \geq 0 \; \forall \boldsymbol{s} \\ & p(\boldsymbol{s}, \boldsymbol{a}) = d(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) && \forall \boldsymbol{s}\, \forall \boldsymbol{a} \end{aligned} \tag{PE·D}$$

The following properties can be proven for this LP, using the optimality (KKT) conditions.

Lemma 2. *The optimal dual variables are the visitation densities of the given policy π for the considered MDP.*

Lemma 3. *The optimal primal variables satisfy the Bellman equation and are equivalent to the value functions v^π and q^π of the given policy π for the considered MDP.*

Appendix A.2 provides a proof and derivation of the Lagrangian objective $\mathcal{L}_{\text{PE}}$. Because of strong duality (no duality gap for LPs), the average reward (or expected discounted return) of policy π can be written as the solution

of the constrained primal or dual LPs

$$\bar{r}^\pi = (1-\gamma)\nu^\pi = \mathop{\mathbb{E}}_{\boldsymbol{s}\sim\iota(\cdot)}\big[(1-\gamma)v^\pi(\boldsymbol{s})\big] = \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p^\pi(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}}[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')],$$

or as the solution of the unconstrained Lagrangian optimization problem

$$\bar{r}^\pi = (1-\gamma)\nu^\pi = \max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \min_{\substack{v(\boldsymbol{s})\\ q(\boldsymbol{s},\boldsymbol{a})}} \mathcal{L}_{\text{PE}} = \min_{\substack{v(\boldsymbol{s})\\ q(\boldsymbol{s},\boldsymbol{a})}} \max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \mathcal{L}_{\text{PE}}.$$

3.3 Policy Optimization

The optimal policy π^* for the considered MDP (1) can also be found iteratively using a two-step scheme called *generalized policy iteration* (GPI). In the first *policy evaluation* step, the value of a current policy π is determined, which is then used in the second *policy improvement* step to find a better policy. Repeatedly iterating over the two steps, eventually results in an optimal policy. Some popular methods in this category are TD3 [17] and SAC [18]. The GPI scheme can be written as a nested optimization problem, where the first policy evaluation step corresponds to an inner policy evaluation LP (PE), and the second policy improvement step corresponds to an outer policy optimization problem (PO).

$$\begin{aligned} \max_{\pi(\boldsymbol{a}|\boldsymbol{s})}\ & \bar{r}^\pi \\ \text{s.t.}\ & \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) = 1\ \forall \boldsymbol{s} \\ & \pi(\boldsymbol{a}|\boldsymbol{s}) \geq 0\ \forall \boldsymbol{s}\,\forall \boldsymbol{a} \end{aligned} \tag{PO}$$

Appendix A.3 provides more details and derivations for this nested optimization problem. The most important result is summarized below.

Theorem 2. *The optimal solution of the nested Policy Optimization* (PO) *and Policy Evaluation* (PE) *problems, is the optimal policy π^* for the considered MDP.*

4 Extending the Primal-Dual Formulations

We are now ready to alter the dual optimization problems described in the previous section and reveal the effect on the primal formulations and optimal solutions. The dual formulations (VL·D) and (PE·D) are especially useful for imposing additional constraints on the policy. For example, the addition of an extra regularization term to the dual objective, reveals a link with the policy distillation and maximum entropy RL setting. Furthermore, extra dual constraints allow us to solve constrained RL problems (CMDPs) or to bound visitation and action densities to ensure that the policy does (not) visit certain states or actions. Alternatively, we can impose transition constraints to prevent or encourage the agent to transition between two consecutive states or actions. Derivations for the various Theorems, Lemmas and Propositions introduced in this section are provided in Appendices A.4 - A.8.

4.1 Entropy Regularization

To mimic a given teacher policy π_{T} and encourage exploration, the entropy regularized objective (3) can be used. By expressing this objective in terms of the visitation densities d^π and p^π, it can replace the policy evaluation dual objective of (PE·D). The resulting regularized dual LP is thus obtained as

$$\begin{aligned} \max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}}\ & \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}}\big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big] - \alpha \mathop{\mathbb{E}}_{\boldsymbol{s}\sim d(\cdot)}\big[\text{KL}\big[\pi(\cdot|\boldsymbol{s})\,\big|\big|\,\pi_{\text{T}}(\cdot|\boldsymbol{s})\big]\big] \\ \text{s.t.}\ & d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)}\big[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\big] \geq 0\ \forall \boldsymbol{s} \\ & p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a} \end{aligned} \tag{ER·D}$$

corresponding to the modified primal

$$\begin{aligned} \min_{\substack{v(\boldsymbol{s})\\ q(\boldsymbol{s},\boldsymbol{a})}}\ & \mathop{\mathbb{E}}_{\boldsymbol{s}_0\sim\iota(\cdot)}\big[(1-\gamma)v(\boldsymbol{s}_0)\big] \\ \text{s.t.}\ & v(\boldsymbol{s}) \geq \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s})}\left[q(\boldsymbol{s},\boldsymbol{a}) - \alpha\log\frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\text{T}}(\boldsymbol{a}|\boldsymbol{s})}\right]\ \forall \boldsymbol{s} \\ & q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}\big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}')\big] \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a} \end{aligned} \tag{ER·P}$$

Lemma 2 still holds for this regularized Policy Evaluation problem, so the optimal dual variables again correspond to the visitation densities d^π and p^π of the considered policy. Notice how the extra regularization term in the dual, leads to a modified reward

$$r_{\text{ER}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') - \alpha\log\frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\text{T}}(\boldsymbol{a}|\boldsymbol{s})} \tag{4}$$

in the primal. Lemma 3 can be modified accordingly, taking the entropy regularization into account through adjusted value functions using this modified reward.

Lemma 4. *The optimal primal variables satisfy the entropy regularized Bellman equation and are equivalent to the entropy regularized value functions v^π_{ER} and q^π_{ER} of the given policy π for the considered MDP.*

The optimal (Boltzmann) policy π^*_{ER} can then be found by plugging this regularized policy evaluation problem in the policy optimization problem (PO), resulting in a *Soft Policy Iteration* scheme [14].

Theorem 3. *The optimal solution of the nested Policy Optimization* (PO) *and Entropy-Regularized Policy Evaluation* (ER) *problems, is a conditional Boltzmann policy π^*_{ER}.*

The resulting optimal policy maximizes the reward while staying close to the teacher policy. Schulman et al. [19] showed the link between policy gradient methods and soft Q-learning in this entropy-regularized setting. Such regularized learning objectives for policy distillation were also investigated in the context of transfer and multitask learning [20, 21, 15].

For the special case of a uniform teaching policy $\pi_{\text{T}}(\boldsymbol{a}|\boldsymbol{s}) = 1/|\mathcal{A}|$, the dual's regularization term effectively

simplifies to $+\alpha\,\mathbb{E}_{\boldsymbol{s}\sim d(\cdot)}[\mathrm{H}(\pi(\cdot|\boldsymbol{s}))]$ (up to a neglected constant $\alpha \log|\mathcal{A}|$), which corresponds to the (plain) maximum entropy RL case [14]. A popular deep RL algorithm in this field is 'Soft Actor-Critic' (SAC) [18].

The temperature parameter α determines how much influence the teacher policy has on the learned student policy. High temperatures lead to policies remaining closer to the teacher policy (or more random in the special maximum entropy case), whereas lower temperatures lead to policies moving closer to the optimal policy π^* for the considered MDP. As α goes to 0, the "softmax" Boltzmann policy thus becomes a "hardmax" greedy policy, and the primal and dual formulations become equivalent to the non-regularized Policy Evaluation LP (PE). Typically the value of the temperature is decreased throughout training, ensuring the policy becomes more independent from the teacher and more deterministic at the end of the training process.

4.2 Constrained Reinforcement Learning

The K additional constraints of the CMDP (2) can be expressed in terms of the visitation density $p^\pi(\boldsymbol{s},\boldsymbol{a})$ as $(1-\gamma)^{-1}\,\mathbb{E}_{(\boldsymbol{s},\boldsymbol{a})\sim p^\pi(\cdot,\cdot),\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}[r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')] \geq V_k$. In this form, the constraints can be readily added to the dual formulations (VL·D) and (PE·D). The constrained dual LP for value learning is then obtained as

$$
\begin{aligned}
\max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}}\big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big] && \\
\text{s.t.} \quad & \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s}) && \forall \boldsymbol{s} \\
& d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) \\
& \qquad + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)}\big[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\big] && \forall \boldsymbol{s} \\
& p(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
& (1-\gamma)V_k \leq \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}}[r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')] && \forall k
\end{aligned}
\tag{CRL·D}
$$

and corresponds to the altered primal

$$
\begin{aligned}
\min_{\substack{v(\boldsymbol{s}),w_k\\ q(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\boldsymbol{s}_0\sim\iota(\cdot)}\big[(1-\gamma)v(\boldsymbol{s}_0)\big] - \sum_{k=0}^{K} w_k(1-\gamma)V_k && \\
\text{s.t.} \quad & v(\boldsymbol{s}) \geq q(\boldsymbol{s},\boldsymbol{a}) && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
& q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}\Big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \sum_{k=0}^{K} w_k r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') \\
& \qquad\qquad + \gamma v(\boldsymbol{s}')\Big] && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
& w_k \geq 0 && \forall k
\end{aligned}
\tag{CRL·P}
$$

In this primal formulation, the reward is modified to

$$
r_{\mathrm{CRL}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \sum_{k=0}^{K} w_k r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') \tag{5}
$$

with *learnable* parameters $w_k \geq 0$. If a feasible policy exists, satisfying all additional constraints, the following properties can be proven, extending Theorem 1 and Lemma 1.

Theorem 4. *The policy that can be derived from the optimal dual variables is an optimal policy for the considered CMDP.*

Lemma 5. *The optimal primal variables satisfy the Bellman optimality equation and are thus equivalent to the optimal adjusted value functions v^*_{CRL} and q^*_{CRL} for the considered MDP with optimal modified reward r^*_{CRL}.*

These optimal adjusted value functions and optimal reward modifications also satisfy following properties.

Theorem 5. *The optimal policy is greedy with respect to the optimal adjusted value functions v^*_{CRL} and q^*_{CRL}.*

Proposition 1. *The optimal modified reward is only altered by the tight constraints $V_k = \nu_k^{\pi^*}$.*

On the other hand, if there is no policy that can satisfy all additional K constraints, the dual problem effectively becomes infeasible and the primal becomes unbounded, caused by the w_k of the infeasible constraints approaching infinity.

Remark that the simplified version of this constrained value learning LP was also covered by Altman [12] in the context of CMDPs, where the primal and dual formulations are typically swapped. More recently, the Lagrangian method for solving constrained RL tasks has also been used to derive several actor-critic algorithms [1, 2, 4], such as RCPO [3]. Such approaches can also be related to an equivalent constrained formulation of the policy evaluation problem (PE) to derive constrained GPI schemes. This is illustrated for visitation density bounds in the next subsection.

4.3 Visitation Density Bounds

To ensure that the optimal policy does (not) visit certain states or state-action pairs with a certain (discounted) probability, visitation density bounds can be added to the dual formulations (VL·D) or (PE·D). Such bounds can be imposed both on the state-action visitation density, $p_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a}) \leq p^\pi(\boldsymbol{s},\boldsymbol{a}) \leq p_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a})$, and on the (marginal) state-visitation density, $d_{\mathrm{L}}(\boldsymbol{s}) \leq d^\pi(\boldsymbol{s}) \leq d_{\mathrm{U}}(\boldsymbol{s})$. As $d^\pi(\boldsymbol{s})$ and $p^\pi(\boldsymbol{s},\boldsymbol{a})$ are proper probability density functions, the lower and upper bounds should satisfy $\sum_{\boldsymbol{s}\in\mathcal{S}} d_{\mathrm{L}}(\boldsymbol{s}) \leq 1$, $\sum_{\boldsymbol{s}\in\mathcal{S}} d_{\mathrm{U}}(\boldsymbol{s}) \geq 1$, $\sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a}) \leq 1$ and $\sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a}) \geq 1$. These conditions on the bounds are however not sufficient to guarantee feasibility of the constrained optimization problems, as this depends on the dynamics of the considered MDP.

Appendix A.6 provides primal and dual formulations of (VL) with bounds on the marginal *state* visitation density, whereas here we focus on the primal and dual formulations of (PE) with *state-action* visitation density bounds. Both types of constraints can however be applied to either optimization problem. The idea of imposing bounds on the state visitation density was also explored by Chen and Ames [9], who related it to optimal control, and Qin et al. [10]. The formulations in this and the following section, where bounds on the conditional action density are

imposed, can be regarded as an extension of the density constrained RL field (Figure 1).

The primal and dual formulations of (PE) with state-action visitation density bounds are given by

$$\begin{aligned}
\min_{\substack{v(\boldsymbol{s})\\ q(\boldsymbol{s},\boldsymbol{a})\\ r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a})\\ r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\boldsymbol{s}_0\sim\iota(\cdot)}\left[(1-\gamma)v(\boldsymbol{s}_0)\right] - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a})p_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a}) \\
& \qquad + \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a})p_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a}) \\
\text{s.t.} \quad & v(\boldsymbol{s}) \geq \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s})}\left[q(\boldsymbol{s},\boldsymbol{a})\right] && \forall \boldsymbol{s} \\
& q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}\left[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a})\right. \\
& \qquad \left. - r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a}) + \gamma v(\boldsymbol{s}')\right] && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
& r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a}) \geq 0 \wedge r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned} \tag{VDB·P}$$

and

$$\begin{aligned}
\max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}}\left[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\right] \\
\text{s.t.} \quad & d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) \\
& \qquad + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)}\left[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\right] \geq 0 \;\forall \boldsymbol{s} \\
& p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
& p_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a}) \leq p(\boldsymbol{s},\boldsymbol{a}) \leq p_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a}) && \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned} \tag{VDB·D}$$

As before, the reward is modified in the primal and becomes

$$r_{\mathrm{VDB}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a}) - r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a}). \tag{6}$$

Remark the *automatic* adjustment of the reward signal, through the learnable decision variables $r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a})$ and $r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a})$. Intuitively, this modified reward can be understood as follows: to ensure a minimum visitation of a state-action pair, the reward can be increased by $r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a})$; whereas to ensure a maximum visitation of a state-action pair, the reward can be decreased by $r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a})$.

Let us first consider the case of a feasible policy, satisfying all imposed density constraints. Then, Lemma 2 still holds for this constrained Policy Evaluation problem, so the optimal dual variables again correspond to the visitation densities d^π and p^π of the considered policy. Lemma 3 can be modified to take the visitation density constraints into account through adjusted value functions using the modified reward.

Lemma 6. *The optimal primal variables satisfy the Bellman equation and are equivalent to the adjusted value functions v^π_{VDB} and q^π_{VDB} of the given policy π for the considered MDP with optimal modified reward r^*_{VDB}.*

These optimal reward modifications satisfy following property.

Proposition 2. *The optimal modified reward is only altered by the tight bounds $p_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a}) = p^\pi(\boldsymbol{s},\boldsymbol{a})$ and $p^\pi(\boldsymbol{s},\boldsymbol{a}) = p_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a})$.*

In the other case, if the policy does not satisfy all additional constraints, the dual problem effectively becomes infeasible and the primal becomes unbounded, caused by the $r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a})$ and $r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a})$ of the infeasible constraints approaching infinity. It might thus seem that this altered Policy Evaluation LP is not really useful on its own. Its utility, however, becomes clear when combining it with the Policy Optimization problem to find optimal constrained policies.

Theorem 6. *The optimal solution of the nested Policy Optimization (PO) and constrained Policy Evaluation (VDB) problems, is the optimal policy π^* for the considered MDP with visitation density bounds.*

The resulting nested optimization problem can be seen as a *Constrained* GPI scheme, in which the policy, adjusted value functions and reward modifications are jointly learned. Although the inner Policy Evaluation problem diverges for an infeasible policy, violating the visitation density bounds, the directions in which the learnable reward parameters r_{L} and r_{U} are updated, remain useful for the combined policy iteration scheme: the reward is effectively increased for *undervisited* state-action pairs, while it is decreased for *overvisited* state-action pairs. As a result, during subsequent policy optimization steps, there is an increased incentive for the policy to visit the undervisited states and actions, while staying away from the overvisited states and actions. The reward is thus modified in such a way that constraint-violating policies are no longer optimal or greedy with respect to the adjusted value functions.

4.4 Action Density Bounds

With a similar trick, one can also impose bounds on the conditional action density of the learned policy. In this case, the introduced constraints are $d^\pi(\boldsymbol{s})\pi_{\mathrm{L}}(\boldsymbol{a}|\boldsymbol{s}) \leq p^\pi(\boldsymbol{s},\boldsymbol{a}) \leq d^\pi(\boldsymbol{s})\pi_{\mathrm{U}}(\boldsymbol{a}|\boldsymbol{s})$ and the restrictions on the bounds become $\sum_{\boldsymbol{a}\in\mathcal{A}}\pi_{\mathrm{L}}(\boldsymbol{a}|\boldsymbol{s}) \leq 1$, $\sum_{\boldsymbol{a}\in\mathcal{A}}\pi_{\mathrm{U}}(\boldsymbol{a}|\boldsymbol{s}) \geq 1$. The altered dual formulation for the value learning problem (VL) is then given by

$$\begin{aligned}
\max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}}\left[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\right] \\
\text{s.t.} \quad & \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s}) && \forall \boldsymbol{s} \\
& d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) \\
& \qquad + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)}\left[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\right] \;\forall \boldsymbol{s} \\
& p(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
& d(\boldsymbol{s})\pi_{\mathrm{L}}(\boldsymbol{a}|\boldsymbol{s}) \leq p(\boldsymbol{s},\boldsymbol{a}) \leq d(\boldsymbol{s})\pi_{\mathrm{U}}(\boldsymbol{a}|\boldsymbol{s}) && \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned} \tag{ADB·D}$$

and has the corresponding primal formulation

$$\min_{\substack{v(\boldsymbol{s}),q(\boldsymbol{s},\boldsymbol{a})\\ r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a}),r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a})}} \mathop{\mathbb{E}}_{\boldsymbol{s}_0\sim\iota(\cdot)}\big[(1-\gamma)v(\boldsymbol{s}_0)\big] \tag{ADB·P}$$

$$\text{s.t.}\quad v(\boldsymbol{s}) \geq q(\boldsymbol{s},\boldsymbol{a}) \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$

$$q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}\Big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a}) - r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a}) - \sum_{\tilde{\boldsymbol{a}}\in\mathcal{A}} r_{\mathrm{L}}(\boldsymbol{s},\tilde{\boldsymbol{a}})\pi_{\mathrm{L}}(\tilde{\boldsymbol{a}}|\boldsymbol{s}) + \sum_{\tilde{\boldsymbol{a}}\in\mathcal{A}} r_{\mathrm{U}}(\boldsymbol{s},\tilde{\boldsymbol{a}})\pi_{\mathrm{U}}(\tilde{\boldsymbol{a}}|\boldsymbol{s}) + \gamma v(\boldsymbol{s}')\Big] \quad \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$

$$r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a}) \geq 0 \wedge r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a}) \geq 0 \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$

The reward modifications in the primal are more complex in this case

$$r_{\mathrm{ADB}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a}) - \sum_{\tilde{\boldsymbol{a}}\in\mathcal{A}} r_{\mathrm{L}}(\boldsymbol{s},\tilde{\boldsymbol{a}})\pi_{\mathrm{L}}(\tilde{\boldsymbol{a}}|\boldsymbol{s}) - r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a}) + \sum_{\tilde{\boldsymbol{a}}\in\mathcal{A}} r_{\mathrm{U}}(\boldsymbol{s},\tilde{\boldsymbol{a}})\pi_{\mathrm{U}}(\tilde{\boldsymbol{a}}|\boldsymbol{s}), \tag{7}$$

and have learnable decision variables $r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a})$ and $r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a})$. For feasible problems, following properties can be derived, extending Theorem 1 and Lemma 1.

Theorem 7. *The policy that can be derived from the optimal dual variables is an optimal policy for the considered MDP with action density bounds.*

Lemma 7. *The optimal primal variables satisfy the Bellman optimality equation and are thus equivalent to the optimal adjusted value functions v^*_{VDB} and q^*_{VDB} for the considered MDP with optimal modified reward r^*_{VDB}.*

Additionally, these optimal adjusted value functions and optimal reward modifications satisfy following properties.

Theorem 8. *The optimal policy is greedy with respect to the optimal adjusted value functions v^*_{ADB} and q^*_{ADB}.*

Proposition 3. *The optimal modified reward is only altered by the tight bounds $\pi_{\mathrm{L}}(\boldsymbol{a}|\boldsymbol{s}) = \pi^\star(\boldsymbol{a}|\boldsymbol{s})$ and $\pi^\star(\boldsymbol{a}|\boldsymbol{s}) = \pi_{\mathrm{U}}(\boldsymbol{a}|\boldsymbol{s})$.*

On the other hand, if there is no policy that can satisfy all additional constraints, the dual problem effectively becomes infeasible and the primal becomes unbounded, caused by the $r_{\mathrm{L}}(\boldsymbol{s},\boldsymbol{a})$ and $r_{\mathrm{U}}(\boldsymbol{s},\boldsymbol{a})$ of the infeasible constraints approaching infinity.

Similar derivations are possible for the policy evaluation problem (PE), leading to a Constrained GPI scheme that is comparable to the one for the visitation density bounds.

4.5 Transition Constraints

In control applications, one often wants to prevent abrupt changes in the applied controls and/or in the state. This can improve user comfort and lead to more smooth control. To enforce such constraints, typically a cost is associated with the variation of states and inputs, which can then be upper bounded. The dual formulations (VL·D) and (PE·D) allow us to do a similar thing for reinforcement learning policies. We denote by $c(\boldsymbol{s}_t,\boldsymbol{a}_t,\boldsymbol{s}_{t+1}) : \mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}$ the cost of transitioning to state $\boldsymbol{s}_{t+1}$ from state $\boldsymbol{s}_t$ by taking action $\boldsymbol{a}_t$. The expected transition cost of a certain state-action pair can then be written as $\bar{c}(\boldsymbol{s},\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}\big[c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big]$. To ensure that this average transition cost remains upper bounded for states and actions that are visited by the policy, constraints of the form $p^\pi(\boldsymbol{s},\boldsymbol{a})\bar{c}(\boldsymbol{s},\boldsymbol{a}) \leq 0$ can then be added to the dual. Note that this could be easily extended to multiple such constraints with different cost functions and upper bounds, as was done for the constrained RL formulation (CRL). The altered value learning dual with average transition constraints then becomes

$$\max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}}\big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big]$$

$$\text{s.t.}\quad \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s}) \qquad \forall \boldsymbol{s}$$

$$d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)}\big[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\big] \quad \forall \boldsymbol{s} \tag{ATC·D}$$

$$p(\boldsymbol{s},\boldsymbol{a}) \geq 0 \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$

$$0 \geq p(\boldsymbol{s},\boldsymbol{a}) \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}\big[c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big] \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$

with corresponding primal

$$\min_{\substack{v(\boldsymbol{s}),q(\boldsymbol{s},\boldsymbol{a})\\ r_{\mathrm{C}}(\boldsymbol{s},\boldsymbol{a})}} \mathop{\mathbb{E}}_{\boldsymbol{s}_0\sim\iota(\cdot)}\big[(1-\gamma)v(\boldsymbol{s}_0)\big] \tag{ATC·P}$$

$$\text{s.t.}\quad v(\boldsymbol{s}) \geq q(\boldsymbol{s},\boldsymbol{a}) \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$

$$q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}\big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') - r_{\mathrm{C}}(\boldsymbol{s},\boldsymbol{a})c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}')\big] \quad \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$

$$r_{\mathrm{C}}(\boldsymbol{s},\boldsymbol{a}) \geq 0 \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$

The reward modifications in the primal have a slightly different form now, combining a learnable reward weight $r_{\mathrm{C}}(\boldsymbol{s},\boldsymbol{a})$ with the given cost $c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')$ in what looks to be a mixture of the (CRL) and (VDB) formulations

$$r_{\mathrm{ATC}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') - r_{\mathrm{C}}(\boldsymbol{s},\boldsymbol{a})c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}'). \tag{8}$$

For feasible problems, following properties hold, extending Theorem 1 and Lemma 1.

Theorem 9. *The policy that can be derived from the optimal dual variables is an optimal policy for the considered MDP with average transition constraints.*

Lemma 8. *The optimal primal variables satisfy the Bellman optimality equation and are thus equivalent to the optimal adjusted value functions v^*_{ATC} and q^*_{ATC} for the considered MDP with optimal modified reward r^*_{ATC}.*

Furthermore, following properties hold for these optimal adjusted value functions and optimal reward modifications.

Theorem 10. *The optimal policy is greedy with respect to the optimal adjusted value functions v^*_{ATC} and q^*_{ATC}.*

Proposition 4. *The optimal modified reward is only altered by nonvisited state-action pairs or tight bounds* $\bar{c}(\boldsymbol{s}, \boldsymbol{a}) = 0$.

On the other hand, if there is no policy that can satisfy all additional constraints, the dual problem effectively becomes infeasible and the primal becomes unbounded, caused by the $r_C(\boldsymbol{s}, \boldsymbol{a})$ of the infeasible constraints approaching infinity.

Once again, a similar derivation is also possible for the policy evaluation problem (PE). Such a formulation has the added benefit that *action* transition constraints can be enforced, as shown in Appendix A.8. In this appendix an alternative *immediate* transition constraint of the form $p^\pi(\boldsymbol{s}, \boldsymbol{a})\tau(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})c(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') \leq 0$ is also examined, which is much stricter than the average transition constraints considered here.

4.6 Overview and Discussion

All altered optimization problems discussed in this section have in common that the effect of extra constraints or regularization terms in the dual leads to reward modifications (4)-(8) and sometimes objective modifications in the primal. Table 1 gives an overview of the effect on the primal for each of the discussed extensions.

Except for the entropy regularized case, the reward modifications in the primal are automatically learned, as they depend on extra decision variables (Lagrange multipliers) associated with each of the introduced constraints in the dual. These reward modifications also make sense intuitively: state-action pairs that need to be visited more frequently by the policy in order to satisfy the constraints get a (positive) reward bonus; whereas states that need to be visited less frequently get a (negative) reward penalty.

Propositions 1, 2, 3, 4 additionally show that the "smallest possible" modifications are typically optimal. More precisely, the optimal reward modifications are zero for states and actions where the corresponding inequalities are not tight. In theory, this means that the optimal policy for the problem without these constraints would also satisfy the strict inequalities. In practice, the constraints are still useful as they can guide the policy towards the desirable regions in state-action space throughout learning.

If there is no policy that can satisfy all additional constraints, the dual problems effectively become infeasible, and the primal problems become unbounded, caused by the reward modification terms associated with the infeasible constraints approaching infinity.

Remark. *In summary, the reward is modified in such a way that constraint-violating policies can no longer be optimal or greedy with respect to the adjusted value functions, whereas policies that are greedy with respect to the optimal adjusted value functions are optimal solutions to the constrained RL problems (Theorems 5, 8, 10).*

Although the modified policy evaluation LPs can thus become infeasible for certain policies, the optimal feasible policy can still be found when it is used as the inner optimization problem in a GPI scheme. The reward modifications (and hence also the adjusted value functions and objective) never reach $\pm\infty$, as the policy learns to avoid the unwanted regions in state-action space and to move towards the wanted regions in state-action space.

To conclude, the various problems summarized in Table 1 provide the designer with a versatile toolbox of policy constraints, acting on different scopes in the trajectory. In particular, there are constraints acting on individual visited states and actions (VDB), (ADB), on transitions between consecutively visited states (ATC), and on whole trajectories of visited states (CRL). While presented as individual optimization problems, the different constraints and corresponding reward modifications can be straightforwardly combined, as is also shown in the experiments (Section 6).

5 Practical Implementations

The previously introduced optimization problems cannot be solved directly, as the distributions ι and τ and the reward function r are unknown in the general model-free reinforcement learning setup. Using parameterized models and experience samples from the replay buffer, they can be approximately solved though, leading to practical algorithms resembling value-based or actor-critic methods. The introduced parameterizations are denoted as $q(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta}_q)$ for the state-action value function (or critic), $\pi(\boldsymbol{a}|\boldsymbol{s}; \boldsymbol{\theta}_\pi)$ for the policy (or actor), $d(\boldsymbol{s}; \boldsymbol{\theta}_d)$ for the state visitation density and $\tilde{r}(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}'; \boldsymbol{\theta}_r)$ for the reward modifications. For small and discrete state and action spaces, tabular models can be used, whereas for larger dimensions neural networks are more suitable. Generic pseudocode for training each of the models required for solving the various constrained RL tasks is given in Algorithm 1.

Algorithm 1 DualCRL with Learnable Reward Modifications

Initialize $\boldsymbol{\theta}_q, \boldsymbol{\theta}_\pi, \boldsymbol{\theta}_d, \boldsymbol{\theta}_r$ and collect experience in $\mathcal{B}$
for each gradient step **do**
  Sample batch $\mathcal{E} = \{(\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1})_b\}_{b\in[B]}$ from $\mathcal{B}$
  $\boldsymbol{\theta}_q \leftarrow \boldsymbol{\theta}_q - \eta_q \nabla_{\boldsymbol{\theta}_q} \widehat{L}_q(\boldsymbol{\theta}_q; \mathcal{E})$ ▷ Subsections 5.1, 5.2
  $\boldsymbol{\theta}_\pi \leftarrow \boldsymbol{\theta}_\pi - \eta_\pi \nabla_{\boldsymbol{\theta}_\pi} \widehat{L}_\pi(\boldsymbol{\theta}_\pi; \mathcal{E})$ ▷ Subsection 5.2
  $\boldsymbol{\theta}_d \leftarrow \boldsymbol{\theta}_d - \eta_d \nabla_{\boldsymbol{\theta}_d} \widehat{L}_d(\boldsymbol{\theta}_d; \mathcal{E})$ ▷ Subsection 5.4
  $\boldsymbol{\theta}_r \leftarrow \boldsymbol{\theta}_r - \eta_r \nabla_{\boldsymbol{\theta}_r} \widehat{L}_r(\boldsymbol{\theta}_r; \mathcal{E})$ ▷ Subsection 5.5
end for

The stochastic estimate of the true gradient $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ using a batch of experience samples $\mathcal{E}$ is denoted by $\nabla_{\boldsymbol{\theta}} \widehat{L}(\boldsymbol{\theta}; \mathcal{E})$. Following state-of-the-art deep RL algorithms, additional independent parameterizations and target networks (denoted with a prime, e.g. $\boldsymbol{\theta}'_q$) can be introduced as well to further improve stability and performance [22, 17].

Tuning the various learning rates η for each model can be difficult and is problem dependent. Generally, the visitation density models should be updated on a faster timescale than the actor, such that the used density estimates prop-

Table 1: Overview of the altered primal problems. Novel constraint types are indicated with an asterisk ($*$).

Problem	Extra Decision Variables	Reward Modifications	Objective Modifications
Policy Distillation & Maximum Entropy (ER) [19, 14]	/	$-\alpha \log \frac{\pi(\boldsymbol{a}\|\boldsymbol{s})}{\pi_{\mathrm{T}}(\boldsymbol{a}\|\boldsymbol{s})}$	/
Constrained RL (CRL) [12]	$w_k \geq 0$	$+\sum_{k=0}^{K} w_k r_k(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')$	$-\sum_{k=0}^{K} w_k V_k$
Visitation Density Bounds (VDB) [9, 10]	$r_{\mathrm{L}}(\boldsymbol{s}, \boldsymbol{a}) \geq 0$	$+ r_{\mathrm{L}}(\boldsymbol{s}, \boldsymbol{a})$	$-\sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{a}\in\mathcal{A}} r_{\mathrm{L}}(\boldsymbol{s}, \boldsymbol{a}) p_{\mathrm{L}}(\boldsymbol{s}, \boldsymbol{a})$
	$r_{\mathrm{U}}(\boldsymbol{s}, \boldsymbol{a}) \geq 0$	$- r_{\mathrm{U}}(\boldsymbol{s}, \boldsymbol{a})$	$+\sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{a}\in\mathcal{A}} r_{\mathrm{U}}(\boldsymbol{s}, \boldsymbol{a}) p_{\mathrm{U}}(\boldsymbol{s}, \boldsymbol{a})$
Action Density Bounds (ADB) $*$	$r_{\mathrm{L}}(\boldsymbol{s}, \boldsymbol{a}) \geq 0$	$+ r_{\mathrm{L}}(\boldsymbol{s}, \boldsymbol{a}) - \sum_{\tilde{\boldsymbol{a}}\in\mathcal{A}} r_{\mathrm{L}}(\boldsymbol{s}, \tilde{\boldsymbol{a}}) \pi_{\mathrm{L}}(\tilde{\boldsymbol{a}}\|\boldsymbol{s})$	/
	$r_{\mathrm{U}}(\boldsymbol{s}, \boldsymbol{a}) \geq 0$	$- r_{\mathrm{U}}(\boldsymbol{s}, \boldsymbol{a}) + \sum_{\tilde{\boldsymbol{a}}\in\mathcal{A}} r_{\mathrm{U}}(\boldsymbol{s}, \tilde{\boldsymbol{a}}) \pi_{\mathrm{U}}(\tilde{\boldsymbol{a}}\|\boldsymbol{s})$	
Transition Constraints (ATC) $*$	$r_{\mathrm{C}}(\boldsymbol{s}, \boldsymbol{a}) \geq 0$	$- r_{\mathrm{C}}(\boldsymbol{s}, \boldsymbol{a}) c(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}')$	/

erly reflect the learned policy's behaviour. For the reward modifications there is a certain trade-off. If η_r is too low, policies are only slowly forced to comply with constraints, leading to many constraint violating iterations. On the other hand, if η_r is too high, reward modifications might oscillate or overshoot, before being properly reflected in the value estimates, destabilizing the learning process.

5.1 Value-Based

The Value Learning LP (VL) and its extensions discussed in the previous section can be turned into value-based algorithms, resembling Q-learning. Because the optimal q^* should satisfy the Bellman optimality equation, it can be learned by minimizing the squared Bellman residual (or temporal difference error) $e_t = q(\boldsymbol{s}_t, \boldsymbol{a}_t) - \mathbb{E}_{\boldsymbol{s}_{t+1}\sim\tau(\cdot|\boldsymbol{s}_t,\boldsymbol{a}_t)}[r(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}) + \tilde{r}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}) + \gamma \max_{\boldsymbol{a}'} q(\boldsymbol{s}_{t+1}, \boldsymbol{a}')]$ for all $\boldsymbol{s}_t$ and $\boldsymbol{a}_t$. This can be approximated using samples from the replay buffer, leading to the critic loss $L_q(\boldsymbol{\theta}_q) = \mathbb{E}_{(\boldsymbol{s}_t,\boldsymbol{a}_t,r_t,\boldsymbol{s}_{t+1})\sim\mathcal{B}}\left[\left(q(\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\theta}_q) - y_t\right)^2\right]$ with $y_t = r_t + \tilde{r}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}; \boldsymbol{\theta}_r) + \gamma \max_{\boldsymbol{a}'} q(\boldsymbol{s}_{t+1}, \boldsymbol{a}'; \boldsymbol{\theta}'_q)$. There is no explicit policy model in this algorithm, but the policy can be implicitly derived as $\pi(\boldsymbol{a}|\boldsymbol{s}) = \delta_{\mathcal{A}^*(\boldsymbol{s})}(\boldsymbol{a})/|\mathcal{A}^*(\boldsymbol{s})|$ with $\mathcal{A}^*(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} q(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta}_q)$. The optimal state value function can similarly be obtained as $v^*(\boldsymbol{s}) = \max_{\boldsymbol{a}} q(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta}^*_q)$.

Note that the value-based algorithm results in a deterministic policy, which is generally not applicable to constrained RL problems, for which the optimal policy might be stochastic. Hence, if it is unknown beforehand whether the optimal policy can be deterministic, actor-critic methods are better suited, as they learn stochastic policies.

5.2 Actor-Critic

The GPI scheme, based on the outer Policy Optimization problem (PO) and inner Policy Evaluation LP (PE), and its alterations can be turned into actor-critic algorithms. The optimal q^* should satisfy the Bellman equation, and can thus be learned by minimizing the squared Bellman residual (or temporal difference error), leading to the critic loss $L_q(\boldsymbol{\theta}_q) = \mathbb{E}_{(\boldsymbol{s}_t,\boldsymbol{a}_t,r_t,\boldsymbol{s}_{t+1})\sim\mathcal{B}}\left[\left(q(\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\theta}_q) - y_t\right)^2\right]$ with $y_t = r_t + \tilde{r}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{s}_{t+1}; \boldsymbol{\theta}_r) + \gamma\, \mathbb{E}_{\boldsymbol{a}'\sim\pi(\cdot|\boldsymbol{s}_{t+1};\boldsymbol{\theta}_\pi)}[q(\boldsymbol{s}_{t+1}, \boldsymbol{a}'; \boldsymbol{\theta}'_q)]$. During Policy Evaluation steps, this loss is minimized, which moves q towards the value function of the current policy π. The state value function can be approximated as $v(\boldsymbol{s}) = \mathbb{E}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s};\boldsymbol{\theta}_\pi)}[q(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta}_q)]$.

The optimal π^* should maximize the average reward, and can thus be learned by minimizing the actor loss (derived from the Lagrangian objective) $L_\pi(\boldsymbol{\theta}_\pi) = -\mathbb{E}_{\boldsymbol{s}_t\sim\mathcal{B},\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s}_t;\boldsymbol{\theta}_\pi)}[q(\boldsymbol{s}_t, \boldsymbol{a}; \boldsymbol{\theta}_q)]$. During Policy Improvement steps, this loss is minimized, which moves π towards the greedy policy with respect to the current value function q. Note that for certain extensions, this objective contains some additional terms (originating from the Lagrangian objective). For example for the entropy regularized case the extra term $\alpha \log \frac{\pi(\boldsymbol{a}|\boldsymbol{s}_t;\boldsymbol{\theta}_\pi)}{\pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s}_t)}$ appears. Depending on the chosen parametric distribution for π, calculating the gradient of L_π typically involves using the reparametrization trick or implicit reparametrization gradients [23, 24].

5.3 Policy-Based

The actor loss expression L_π is consistent with the policy gradient theorem [11]. The same Policy Optimization problem (PO) could thus also be used to derive *policy-based* methods, in which a Monte Carlo estimate is used for determining the value $q^\pi(\boldsymbol{s}, \boldsymbol{a})$ instead of a separate critic model. This setup is however not further investigated in this work.

5.4 Visitation Density Estimation

Samples from the visitation densities can be easily obtained from the replay buffer, as is done in the evaluation of the actor and critic losses. However, for certain updates of the reward modification models, actual density probabilities need to be estimated. For discrete, finite-dimensional state spaces, (discounted) counting of the visited states in tabular density models is possible. For larger (or continuous) state spaces, other tractable density estimation techniques — such as (Gaussian) mixture models or normalizing flows [25] — can be leveraged to train the parameterized visitation density model $d(\boldsymbol{s}; \boldsymbol{\theta}_d)$ using samples from the replay buffer. Such density models are trained by maximizing the average log likelihood, leading to the density loss $L_d(\boldsymbol{\theta}_d) = -\mathbb{E}_{\boldsymbol{s}\sim\mathcal{B}}[\log d(\boldsymbol{s}; \boldsymbol{\theta}_d)]$. Alternatively, for smaller dimensional state spaces, non-parametric density estimation methods, such as Kernel Density Estimation (KDE) [26], could be applied on the most recently visited states in the replay buffer.

The state-action visitation density can either be estimated in a similar way or calculated as $p(\boldsymbol{s}, \boldsymbol{a}; \boldsymbol{\theta}_p) =$

$d(\boldsymbol{s};\boldsymbol{\theta}_d)\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta}_\pi)$.

5.5 Learnable Reward Modifications

The reward modification models $\tilde{r}$ (see also third column of Table 1) are updated by minimizing the reward loss L_r, which is derived from the Lagrangian objective and hence depends on the specific constraints that are enforced. For tabular parameterizations, a projected gradient descent step is applied to ensure that the reward modifications (or weights w_k for the constrained RL case) remain nonnegative. For neural parameterizations, this property can be enforced on the network architecture itself, for example by using a softplus activation function ($\log[1+\exp(x)]$) in the output layer.

For the value constraints of (CRL), we have $L_r(\boldsymbol{\theta}_r) = \sum_{k=0}^{K} w_k \mathbb{E}_{(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\sim\mathcal{B}}[r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') - (1-\gamma)V_k]$ with $\boldsymbol{\theta}_r = [w_1;\dots;w_K]$. The visitation density bounds of (VDB) lead to $L_r(\boldsymbol{\theta}_r) = \mathbb{E}_{(\boldsymbol{s},\boldsymbol{a})\sim\mathcal{B}}\Big[r_{\text{L}}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_{r_{\text{L}}})\Big(1 - \frac{p_{\text{L}}(\boldsymbol{s},\boldsymbol{a})}{p(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_p)}\Big) + r_{\text{U}}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_{r_{\text{U}}})\Big(\frac{p_{\text{U}}(\boldsymbol{s},\boldsymbol{a})}{p(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_p)} - 1\Big)\Big]$ with $\boldsymbol{\theta}_r = [\boldsymbol{\theta}_{r_{\text{L}}};\boldsymbol{\theta}_{r_{\text{U}}}]$. This is very similar to the reward loss for action density bounds (ADB), given by $L_r(\boldsymbol{\theta}_r) = \mathbb{E}_{(\boldsymbol{s},\boldsymbol{a})\sim\mathcal{B}}\Big[r_{\text{L}}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_{r_{\text{L}}})\Big(1 - \frac{d(\boldsymbol{s};\boldsymbol{\theta}_d)\pi_{\text{L}}(\boldsymbol{a}|\boldsymbol{s})}{p(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_p)}\Big) + r_{\text{U}}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_{r_{\text{U}}})\Big(\frac{d(\boldsymbol{s};\boldsymbol{\theta}_d)\pi_{\text{U}}(\boldsymbol{a}|\boldsymbol{s})}{p(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_p)} - 1\Big)\Big]$. Finally, for the average transition costs of (ATC), we obtain $L_r(\boldsymbol{\theta}_r) = -\mathbb{E}_{(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\sim\mathcal{B}}[r_{\text{C}}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}_r)c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')]$.

6 Experiments

The presented `DualCRL` algorithm is applied to two environments to demonstrate its utility and the effect of different types of constraints discussed in this work. Both the value-based and actor-critic implementations are first evaluated on the `CliffWalking` environment with discrete state and action spaces, using tabular models. The actor-critic implementation is afterwards used in the `Pendulum` environment with continuous state and action spaces, using neural networks. The choice for these relatively simple environments is twofold. First of all, the small dimensions of their state and action spaces, allow to visualize the effect of the various constraints and reward modifications on the learned policy's behaviour. Secondly, their simple reward structure leads to well-known controller behaviour, which can be easily extended using the various policy constraints introduced here to obtain more complex behaviour.

An implementation of the `DualCRL` method in PyTorch is made available here[1], together with some extra visualizations for the conducted experiments. Hyperparameter values were empirically determined and are summarized in Appendix B.

6.1 Cliff Walking

In the `CliffWalking` environment [11], the goal is to move from the lower left corner of a rectangular grid to the lower right corner, while avoiding the dangerous squares on the bottom row (the cliff). The state space $\mathcal{S} = [S]$ enumerates all squares of the grid and the action space $\mathcal{A} = \{\uparrow,\rightarrow,\downarrow,\leftarrow\}$ provides the possible movements to neighbouring squares. The reward is -1 for each state, except for the cliff states where it is -100.

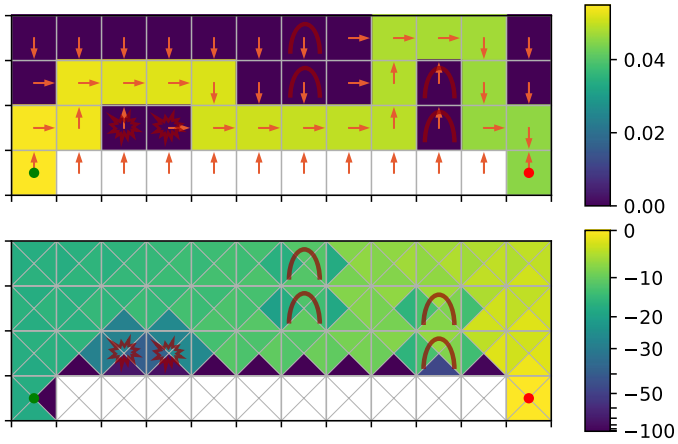

Figure 2: CliffWalking with state-visitation density bounds and transition constraints. The unstable squares and ridges are labelled with red 💥 and ∩ symbols respectively. Top: resulting policy and estimated visitation density after 20 000 timesteps. Bottom: learned adjusted value function after 50 000 timesteps.

As a first extension to this problem, we examine the effect of state-visitation density bounds and transition constraints using the value-based method. Suppose we know that two squares near the cliff edge are unstable. Then we can avoid these by imposing $d(\boldsymbol{s}) \leq \epsilon$ with ϵ a small, strictly positive number. Additionally, if there are two ridges making horizontal movements more difficult, we can assign a strictly positive transition cost $c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') > 0$ for all states $\boldsymbol{s}$ and $\boldsymbol{s}'$ on the ridge and horizontal movements $\boldsymbol{a} \in \{\rightarrow,\leftarrow\}$. After some initial exploration and training, Figure 2 shows the learned adjusted value function, the resulting policy and estimated visitation density. It is immediately clear that the policy has learned to avoid the two unstable squares and ridges by going around them. This behaviour is also clearly reflected in the learned adjusted value function, which assigns more negative values to the state-action pairs that need to be avoided. Remark that while the imposed visitation bounds lead to a reward penalty for the unwanted *states*, this also affects the value for *state-action pairs* of neighbouring states moving towards these (unstable) states. On the other hand, the imposed transition constraints only penalize the costly state-action pairs associated with horizontal movements across the ridge, while leaving the vertical movements unaffected (for this particular setup, a similar result could also be obtained using state-action visitation density bounds).

In a second experiment, the entropy regularization and action density bounds are combined. As the teacher policy, we choose a safe policy, moving as far away as possible from the cliff. The action density bounds are chosen to "oppose"

[1] https://github.com/dcbr/dualcrl

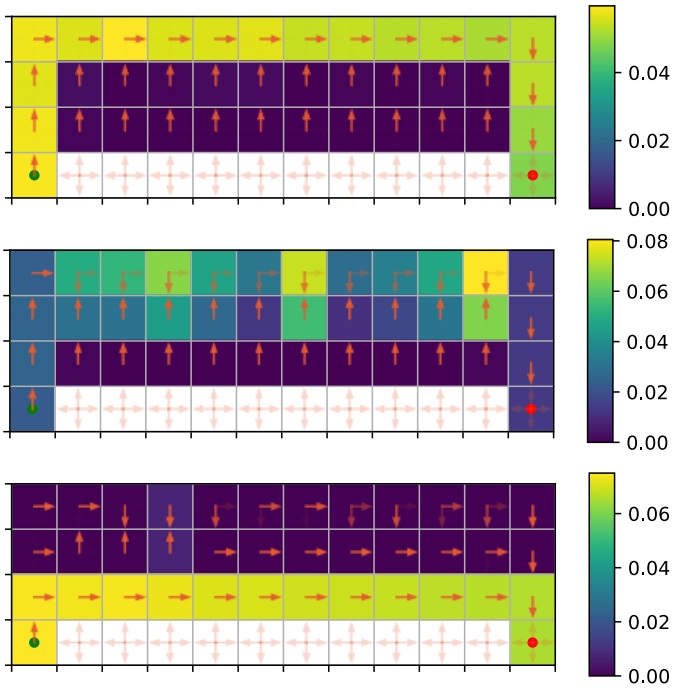

Figure 3: CliffWalking with entropy regularization and action density bounds. The learned policy and estimated visitation density are shown after 4000, 8000 and 50 000 timesteps (from top to bottom).

the teacher policy, by imposing $\pi(\downarrow | \boldsymbol{s}) \geq 0.5$ for all states in the top row (except the first and last column). The temperature parameter α is exponentially decayed throughout training. Figure 3 summarizes the results of this experiment. As can be seen, at the start of training the learned policy mimics the teacher policy in all states. As training goes on, and the temperature α is lowered, the policy gains more freedom to deviate from the teacher's behaviour, while the importance of the action density bounds also increases through updated reward modifiers. These bounds push the policy away from the top row, allowing it to further explore the state-action space and find more efficient paths towards the goal. Finally, at the end of training, the policy has found the more efficient path towards the goal, which remains closer to the cliff. This example illustrates how the learned policy behaviour can be influenced by the temperature parameter α, which controls the relative importance of the (conflicting) entropy regularization and average reward maximization objective with policy constraints.

6.2 Pendulum

In the `Pendulum` environment, the goal is to swing up and stabilize a pendulum around the upward position. The continuous state consists of the angle θ and angular velocity $\dot{\theta}$ of the pendulum and the agent receives this information through observations $\left[x, y, \dot{\theta}\right]$ with $x = \cos(\theta)$ and $y = \sin(\theta)$. The continuous action is the torque applied to the pendulum. The reward penalizes deviations from the upward position and, to a lesser degree, high angular velocity or applied torque.

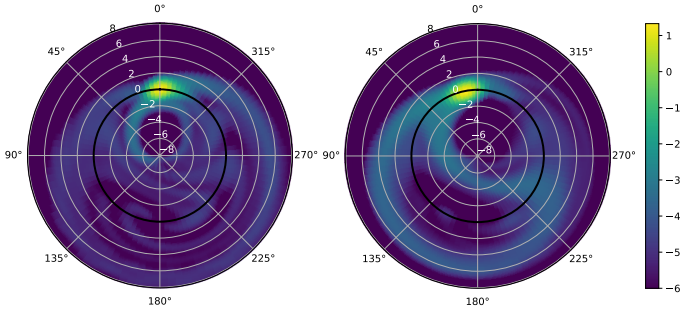

Figure 4: Visualization of the estimated state visitation density on a (natural) logarithmic scale for the Pendulum experiments after 50 000 timesteps. The angle θ and angular velocity $\dot{\theta}$ are shown on the angular and radial axis respectively. Left: Entropy regularization. Right: Entropy regularization, state-visitation density bounds and value constraints.

The experiments conducted in this environment use the actor-critic algorithm with entropy regularization to encourage exploration (similar to SAC [18]) with an exponentially decreasing temperature. The state visitation density is estimated using Gaussian Kernel Density Estimation (KDE) [26] using the last 5000 visited states of the behavioural policy.

The problem is extended by imposing state-visitation density bounds of the form $d(\boldsymbol{s}) \leq \epsilon$ for all states in the upper right quadrant, and adding a value constraint with additional reward $r_{\mathrm{V}}(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s}') = -\dot{\theta}^2$ (for all states). The aim of these additional constraints is to avoid swing-ups along the right side and avoid high angular velocities. Figure 4 visualizes the estimated visitation density of the learned policy and compares it with a baseline policy trained without such constraints. The shown visitation densities are estimated using KDE on states collected from 10 independently trained models in 4 evaluation episodes with 200 timesteps each. From those visitation plots, we can clearly see the constrained policy chooses the left side for swingups and stabilization, while avoiding the highest angular velocities. The right side is only used for gaining momentum (lower right quadrant) and stabilizing around the upward position (upper right quadrant). Note that some visitations of the upper right quadrant are still visible, as random initializations can still put the pendulum in such a state.

7 Related Work

Linear Programming (LP) formulations of Markov Decision Processes (MDP) or Reinforcement Learning (RL) have been analyzed by Puterman [13] and Altman [12]. When the environment dynamics and reward function are known, the value learning LP can be used to approximately solve MDPs [27, 28]. However, this is not easy in practice, when dealing with unknown dynamics and large state and action spaces. In this work, the LPs form the basis of our framework to derive primal-dual formulations of various

constrained RL tasks, from which practical algorithms can then be deduced.

Various primal-dual algorithms for solving different RL tasks have already been proposed. Some of those analyze the "pure" setting, without additional policy constraints [29, 30, 31], whereas others investigate one specific type of constraint. For example, there is a significant amount of research related to value constraints (CMDPs) [1, 32, 3, 2, 4, 33, 34, 35, 36]. Other work covers visitation density constraints [9, 10] and entropy regularization [37]. The intent of this paper is to provide a unifying framework for these seemingly different subdomains and derive a practical algorithm that can handle and combine different kinds of policy constraints. Moreover, by separately treating the value iteration and generalized policy iteration schemes, the impact of additional policy-dependent regularization terms or constraints can be straightforwardly analyzed in our framework. This is in contrast to other existing derivations for practical primal-dual algorithms, which introduce the policy in the dual or Lagrangian representation of (VL). As the resulting dual problem then becomes bilinear, without a clear primal representation, it is no longer suitable for such an analysis.

Primal and dual formulations of the Value Learning LP were also used to derive *off-policy* and *offline* RL algorithms by Nachum and Dai [7], such as DualDICE [5] and AlgaeDICE [6]. By replacing the dual objective with an f-divergence, suitable offline formulations can be derived. Instead of replacing the objective, we add a KL-divergence regularization term to the dual objective, which then leads to the maximum entropy and policy distillation subdomains. For the theoretical derivations in this paper, we assume on-policy experience samples are available. An extension of our proposed framework to offline RL would be an interesting future path of research, as Polosky et al. [8] have shown that LP formulations for offline policy evaluation can be combined with value constraints.

Our work can be related to *convex* RL literature [38, 39, 40], where the dual reward maximization objective is generalized to any concave utility function of the policy visitation. All of the discussed optimization problems in this work also fit in this convex RL framework. For example, our entropy regularization formulations can be seen as a hybrid combination of the standard RL objective with the imitation learning and pure exploration objectives. This is similar to the objective of Zhang et al. [41] with a convex regularizer for finding optimal caution-sensitive policies. We specifically focus on a subset of convex RL problems, for which primal and dual formulations can be easily derived and modified, with a focus on constraining the policy. For these problems, we show an intrinsic relationship between dual constraints and primal reward modifications. Our work could be further extended to other convex RL problems, or leverage some of the proposed game-theoretic solution strategies [42, 39, 40].

Finally, there are various approaches that try to learn or finetune the reward in the field of *inverse* and *automated* RL. This is however not the focus of this work, where the learned reward modifications in the primal have a particular structure that is directly related to certain policy constraints in the dual.

8 Conclusion

In this paper we outlined a generic primal-dual framework for value-based and actor-critic reinforcement learning methods. The dual formulations allowed us to easily plug in and impose additional constraints on the learned policy, unifying several reinforcement learning subdomains. Furthermore, the derivations reveiled that these dual policy constraints correspond to learnable reward modifications in the primal. A practical `DualCRL` method was derived and investigated in more detail with different combinations of constraints on two interpretable environments. Ultimately the presented algorithm aims to lower the burden on reinforcement learning designers, allowing them to naturally divide the problem at hand in a primary objective, through the reward function, and additional policy constraints. With a carefully selected reward and set of constraints, the design remains simple and easy to understand, while still supporting complex behaviour to be learned.

For future work, it would be interesting to try out the presented methods on more complex environments with higher dimensional state and action spaces. Additionally, the effect of the learning rates for the various models could be further investigated. Finally, a more thorough analysis of the introduced approximation errors for the practical algorithm (due to the introduction of parameterized models) would be useful, similar to what is done by Paternain et al. [2, 35] for the constrained RL case. Alternatively, function approximators that are convex in their parameters, such as kernel methods (with feature map representations $w^\top \phi(x)$) can be used without introducing a duality gap.

References

[1] S. Bhatnagar and K. Lakshmanan, "An Online Actor–Critic Algorithm with Function Approximation for Constrained Markov Decision Processes," *Journal of Optimization Theory and Applications*, vol. 153, no. 3, pp. 688–708, Jun. 2012.

[2] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained Reinforcement Learning Has Zero Duality Gap," in *Advances in Neural Information Processing Systems 32, NeurIPS 2019*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/c1aeb6517a1c7f33514f7ff69047e74e-Abstract.html

[3] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward Constrained Policy Optimization," in *7th International Conference on Learning Representations, ICLR 2019*, Sep. 2018.

[4] A. Stooke, J. Achiam, and P. Abbeel, "Responsive Safety in Reinforcement Learning by PID Lagrangian

Methods," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, vol. 119. PMLR, Nov. 2020, pp. 9133–9143. [Online]. Available: https://proceedings.mlr.press/v119/stooke20a.html

[5] O. Nachum, Y. Chow, B. Dai, and L. Li, "DualDICE: Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections," in *Advances in Neural Information Processing Systems 32, NeurIPS 2019*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://papers.nips.cc/paper/2019/hash/cf9a242b70f45317ffd281241fa66502-Abstract.html

[6] O. Nachum, B. Dai, I. Kostrikov, Y. Chow, L. Li, and D. Schuurmans, "AlgaeDICE: Policy Gradient from Arbitrary Experience," Dec. 2019.

[7] O. Nachum and B. Dai, "Reinforcement Learning via Fenchel-Rockafellar Duality," Jan. 2020.

[8] N. Polosky, B. C. D. Silva, M. Fiterau, and J. Jagannath, "Constrained Offline Policy Optimization," in *Proceedings of the 39th International Conference on Machine Learning, ICML 2022*, vol. 162. PMLR, Jun. 2022, pp. 17 801–17 810. [Online]. Available: https://proceedings.mlr.press/v162/polosky22a.html

[9] Y. Chen and A. D. Ames, "Duality between density function and value function with applications in constrained optimal control and Markov Decision Process," Nov. 2019. [Online]. Available: http://arxiv.org/abs/1902.09583

[10] Z. Qin, Y. Chen, and C. Fan, "Density Constrained Reinforcement Learning," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, vol. 139. PMLR, Jul. 2021, pp. 8682–8692. [Online]. Available: https://proceedings.mlr.press/v139/qin21a.html

[11] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., ser. A Bradford Book. MIT Press, Nov. 2018, vol. 258. [Online]. Available: https://books.google.be/books?id=uWV0DwAAQBAJ

[12] E. Altman, *Constrained Markov Decision Processes*, 1st ed. CRC Press, Mar. 1999. [Online]. Available: https://books.google.be/books?id=3X9S1NM2iOgC

[13] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Jan. 1994. [Online]. Available: https://books.google.com/books?id=VvBjBAAAQBAJ

[14] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, vol. 80. PMLR, Jul. 2018, pp. 1861–1870. [Online]. Available: https://proceedings.mlr.press/v80/haarnoja18b.html

[15] W. M. Czarnecki, R. Pascanu, S. Osindero, S. Jayakumar, G. Swirszcz, and M. Jaderberg, "Distilling Policy Distillation," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, vol. 89. PMLR, Apr. 2019, pp. 1331–1340. [Online]. Available: https://proceedings.mlr.press/v89/czarnecki19a.html

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[17] S. Fujimoto, H. Van Hoof, and D. Meger, "Addressing Function Approximation Error in Actor-Critic Methods," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, vol. 80. PMLR, Jul. 2018, pp. 1587–1596. [Online]. Available: https://proceedings.mlr.press/v80/fujimoto18a.html

[18] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft Actor-Critic Algorithms and Applications," Jan. 2019.

[19] J. Schulman, X. Chen, and P. Abbeel, "Equivalence Between Policy Gradients and Soft Q-Learning," Oct. 2018.

[20] E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Actor-Mimic: Deep Multitask and Transfer Reinforcement Learning," in *4th International Conference on Learning Representations, ICLR 2016*. arXiv, Feb. 2016.

[21] S. Schmitt, J. J. Hudson, A. Zidek, S. Osindero, C. Doersch, W. M. Czarnecki, J. Z. Leibo, H. Kuttler, A. Zisserman, K. Simonyan, and S. M. A. Eslami, "Kickstarting Deep Reinforcement Learning," Mar. 2018.

[22] H. Van Hasselt, A. Guez, and D. Silver, "Deep Reinforcement Learning with Double Q-Learning," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence, AAAI 2016*, vol. 30, Mar. 2016, pp. 2094–2100.

[23] D. P. Kingma, T. Salimans, and M. Welling, "Variational Dropout and the Local Reparameterization Trick," in *Advances in Neural Information Processing Systems 28, NIPS 2015*, vol. 28. Curran Associates, Inc., 2015, pp. 2575–2583. [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/bc7316929fe1545bf0b98d114ee3ecb8-Abstract.html

[24] M. Figurnov, S. Mohamed, and A. Mnih, "Implicit Reparameterization Gradients," in *Advances in Neural Information Processing Systems 31, NeurIPS 2018*, vol. 31. Curran Associates, Inc., 2018, pp. 441–452. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/hash/92c8c96e4c37100777c7190b76d28233-Abstract.html

[25] D. Rezende and S. Mohamed, "Variational Inference with Normalizing Flows," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, vol. 37. PMLR, Jun. 2015, pp. 1530–1538. [Online]. Available: https://proceedings.mlr.press/v37/rezende15.html

[26] E. Parzen, "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, Sep. 1962.

[27] D. P. de Farias and B. Van Roy, "The Linear Programming Approach to Approximate Dynamic Programming," *Operations Research*, vol. 51, no. 6, pp. 850–865, Dec. 2003.

[28] R. Cogill, "Primal-dual algorithms for discounted Markov decision processes," in *2015 European Control Conference (ECC)*, Jul. 2015, pp. 260–265.

[29] M. Yang, Y. Li, and D. Schuurmans, "Dual Temporal Difference Learning," in *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, AISTATS 2009*, vol. 5. PMLR, Apr. 2009, pp. 631–638. [Online]. Available: https://proceedings.mlr.press/v5/yang09a.html

[30] M. Wang and Y. Chen, "An online primal-dual method for discounted Markov decision processes," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, Dec. 2016, pp. 4516–4521.

[31] B. Dai, A. E. Shaw, N. He, L. Li, and L. Song, "Boosting the Actor with Dual Critic," in *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, BC, Canada, 2018. [Online]. Available: https://openreview.net/forum?id=BkUp6GZRW

[32] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained Policy Optimization," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, vol. 70. PMLR, Jul. 2017, pp. 22–31. [Online]. Available: https://proceedings.mlr.press/v70/achiam17a.html

[33] D. Ding, K. Zhang, T. Basar, and M. Jovanovic, "Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes," in *Advances in Neural Information Processing Systems 33, NeurIPS 2020*, vol. 33. Curran Associates, Inc., 2020, pp. 8378–8390. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/5f7695debd8cde8db5abcb9f161b49ea-Abstract.html

[34] Y. Chen, J. Dong, and Z. Wang, "A Primal-Dual Approach to Constrained Markov Decision Processes," Jan. 2021. [Online]. Available: http://arxiv.org/abs/2101.10895

[35] S. Paternain, M. Calvo-Fullana, L. F. O. Chamon, and A. Ribeiro, "Safe Policies for Reinforcement Learning via Primal-Dual Methods," *IEEE Transactions on Automatic Control*, vol. 68, no. 3, pp. 1321–1336, Mar. 2023.

[36] Y. Li, M. Zhao, W. Chen, and Z. Wen, "A Stochastic Composite Augmented Lagrangian Method for Reinforcement Learning," *SIAM Journal on Optimization*, vol. 33, no. 2, pp. 921–949, Jun. 2023.

[37] G. Neu, A. Jonsson, and V. Gómez, "A unified view of entropy-regularized Markov decision processes," May 2017. [Online]. Available: http://arxiv.org/abs/1705.07798

[38] J. Zhang, A. Koppel, A. S. Bedi, C. Szepesvari, and M. Wang, "Variational Policy Gradient Method for Reinforcement Learning with General Utilities," in *Advances in Neural Information Processing Systems 33, NeurIPS 2020*, vol. 33. Curran Associates, Inc., 2020, pp. 4572–4583. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/30ee748d38e21392de740e2f9dc686b6-Abstract.html

[39] T. Zahavy, B. O' Donoghue, G. Desjardins, and S. Singh, "Reward is enough for convex MDPs," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 25 746–25 759. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/hash/d7e4cdde82a894b8f633e6d61a01ef15-Abstract.html

[40] M. Geist, "Concave Utility Reinforcement Learning: The Mean-Field Game Viewpoint," in *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022*. Auckland, NZ: IFAAMAS, 2022, pp. 489–497. [Online]. Available: https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p489.pdf

[41] J. Zhang, A. S. Bedi, M. Wang, and A. Koppel, "Beyond Cumulative Returns via Reinforcement Learning over State-Action Occupancy Measures," in *Proceedings of the 2021 American Control Conference (ACC)*, May 2021, pp. 894–901.

[42] S. Miryoosefi, K. Brantley, H. Daume III, M. Dudik, and R. E. Schapire, "Reinforcement Learning with Convex Constraints," in *Advances in Neural Information Processing Systems 32, NeurIPS 2019*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/hash/873be0705c80679f2c71fbf4d872df59-Abstract.html

Appendices

A

In this appendix, the derivations of all discussed primal and dual optimization problems are given, together with optimality conditions. It is assumed the initial state distribution $\iota(\boldsymbol{s}_0)$ is strictly nonzero for all states, as commonly done in MDP literature. This means every state gets visited, i.e. $\mathcal{S}^\pi = \mathcal{S}$ for any policy π. If this is not the case, sufficient exploration is needed to ensure that every state can be visited. Remark that optimal solutions to the given optimization problems are denoted by a star $x^\star$ to distinguish them from optimal solutions to the considered reinforcement learning task, which are denoted by an asterisk x^*.

A.1 Value Learning

Starting from the primal formulation of the Value Learning problem (VL·P)

$$\begin{aligned}\min_{\substack{v(\boldsymbol{s})\\ q(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\boldsymbol{s}_0\sim\iota(\cdot)}\big[(1-\gamma)v(\boldsymbol{s}_0)\big] \\ \text{s.t.} \quad & v(\boldsymbol{s}) \geq q(\boldsymbol{s},\boldsymbol{a}) && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\ & q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}\big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}')\big] && \forall \boldsymbol{s}\,\forall \boldsymbol{a}\end{aligned}$$

the Lagrangian can be obtained as

$$\begin{aligned}&\mathcal{L}_{\text{VL}}(v(\boldsymbol{s}), q(\boldsymbol{s},\boldsymbol{a}), \mu(\boldsymbol{s},\boldsymbol{a}), \lambda(\boldsymbol{s},\boldsymbol{a})) \\ &= \sum_{\boldsymbol{s}\in\mathcal{S}} \iota(\boldsymbol{s})(1-\gamma)v(\boldsymbol{s}) - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} \mu(\boldsymbol{s},\boldsymbol{a})[v(\boldsymbol{s}) - q(\boldsymbol{s},\boldsymbol{a})] \\ &\quad - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} \lambda(\boldsymbol{s},\boldsymbol{a})\left[q(\boldsymbol{s},\boldsymbol{a}) - \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}')\big)\right] \\ &= \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} \lambda(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') \\ &\quad - \sum_{\boldsymbol{s}\in\mathcal{S}} v(\boldsymbol{s})\left[\sum_{\boldsymbol{a}\in\mathcal{A}} \mu(\boldsymbol{s},\boldsymbol{a}) - (1-\gamma)\iota(\boldsymbol{s}) \right. \\ &\qquad\qquad \left. - \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} \lambda(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\right] \\ &\quad - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} q(\boldsymbol{s},\boldsymbol{a})\big[\lambda(\boldsymbol{s},\boldsymbol{a}) - \mu(\boldsymbol{s},\boldsymbol{a})\big].\end{aligned}$$

The optimality (KKT) conditions are given by

- Stationarity of respectively q and v, introducing auxiliary d

$$\begin{aligned} p^\star(\boldsymbol{s},\boldsymbol{a}) &= \lambda^\star(\boldsymbol{s},\boldsymbol{a}) = \mu^\star(\boldsymbol{s},\boldsymbol{a}) && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\ d^\star(\boldsymbol{s}) &= \sum_{\boldsymbol{a}\in\mathcal{A}} p^\star(\boldsymbol{s},\boldsymbol{a}) \\ &= (1-\gamma)\iota(\boldsymbol{s}) + \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p^\star(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}') && \forall \boldsymbol{s}\end{aligned}$$

- Primal feasibility

$$v^\star(\boldsymbol{s}) \geq q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big) \;\forall \boldsymbol{s}\,\forall \boldsymbol{a}$$

- Dual feasibility $\qquad p^\star(\boldsymbol{s},\boldsymbol{a}) \geq 0 \;\forall \boldsymbol{s}\,\forall \boldsymbol{a}$

- Complementary slackness

$$v^\star(\boldsymbol{s}) = q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big) \vee p^\star(\boldsymbol{s},\boldsymbol{a}) = 0 \;\forall \boldsymbol{s}\,\forall \boldsymbol{a}$$

and the dual can be derived as

$$\begin{aligned}\max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}}\big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big] \\ \text{s.t.} \quad & \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s}) && \forall \boldsymbol{s} \\ & d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)}\big[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\big] && \forall \boldsymbol{s} \\ & p(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a}\end{aligned}$$

From the optimality conditions we can prove the following properties of the feasible or optimal primal and dual solutions of the Value Learning LP.

Proposition 5. *All feasible dual variables d and p are probability distributions over $\mathcal{S}$ and $\mathcal{S}\times\mathcal{A}$ respectively. Furthermore, they are the visitation densities of the policy $\pi(\boldsymbol{a}|\boldsymbol{s}) = \frac{p(\boldsymbol{s},\boldsymbol{a})}{d(\boldsymbol{s})}$ for the considered MDP.*

Proof. From the equality constraints (or stationarity conditions), we can derive that $\sum_{\boldsymbol{s}\in\mathcal{S}} d(\boldsymbol{s}) = \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) = 1$. Combined with the inequality constraints (or dual feasibility conditions), which imply $p(\boldsymbol{s},\boldsymbol{a}) \geq 0$ and $d(\boldsymbol{s}) \geq 0$ for all $\boldsymbol{s}$ and $\boldsymbol{a}$, this means $p(\boldsymbol{s},\boldsymbol{a})$ is a probability distribution over $\mathcal{S}\times\mathcal{A}$, with marginal distribution $d(\boldsymbol{s})$. This allows to define a policy π as the conditional distribution $\pi(\boldsymbol{a}|\boldsymbol{s}) = p(\boldsymbol{s},\boldsymbol{a})/d(\boldsymbol{s})$. The equality constraints (or stationarity conditions) can then be rewritten as $d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} d(\boldsymbol{s}')\pi(\boldsymbol{a}'|\boldsymbol{s}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')$, which is equivalent to the definition of the visitation density d^π. From any feasible solution, we can thus extract a policy π with corresponding visitation densities $d(\boldsymbol{s}) = d^\pi(\boldsymbol{s})$ and $p(\boldsymbol{s},\boldsymbol{a}) = p^\pi(\boldsymbol{s},\boldsymbol{a})$. ■

Theorem 1. *The policy $\pi^\star$ that can be derived from the optimal dual variables $d^\star$ and $p^\star$ is an optimal policy for the considered MDP.*

Proof. From Proposition 5, we know that the optimal dual variables correspond to the visitation densities of a policy $\pi^\star$ with $p^\star(\boldsymbol{s},\boldsymbol{a}) = d^{\pi^\star}(\boldsymbol{s})\pi^\star(\boldsymbol{a}|\boldsymbol{s})$. Furthermore, the optimal dual variables maximize the dual objective, which can be rewritten in terms of the policy $\pi^\star$ as $\mathbb{E}_{\boldsymbol{s}\sim d^{\pi^\star}(\cdot),\boldsymbol{a}\sim\pi^\star(\cdot|\boldsymbol{s}),\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')] = \bar{r}^{\pi^\star}$. The optimal dual variables $d^\star$ and $p^\star$ are thus the visitation densities of the optimal policy π^* for the considered MDP, maximizing the average reward (or expected discounted return). ■

Lemma 1. *The optimal primal variables $v^\star$ and $q^\star$ satisfy the Bellman optimality equation and are thus equivalent to the optimal value functions v^* and q^* of the considered MDP.*

Proof. The complementary slackness conditions imply that for each visited state $\boldsymbol{s} \in \mathcal{S}^{\pi^\star}$, the primal feasibility constraints become tight for all actions $\boldsymbol{a} \in \mathcal{A}^{\pi^\star}(\boldsymbol{s})$ taken by the policy, as $p^\star(\boldsymbol{s},\boldsymbol{a}) > 0$ for such states and actions. Combined with the primal feasibility conditions, we thus have $v^\star(\boldsymbol{s}) = \max_{\boldsymbol{a}} \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$ or $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \mathcal{A}^*(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$. Satisfying the Bellman optimality equation, $v^\star$ is thus equivalent to the optimal value function v^* for the

considered MDP. From the primal feasibility conditions $q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$, we can then conclude that $q^\star$ is also equivalent to the optimal value function q^* for the considered MDP. ■

A.2 Policy Evaluation

The primal formulation of the Policy Evaluation problem (PE·P) is given by

$$\begin{aligned}\min_{\substack{v(\boldsymbol{s})\\ q(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\boldsymbol{s}_0\sim\iota(\cdot)}\big[(1-\gamma)v(\boldsymbol{s}_0)\big] \\ \text{s.t.} \quad & v(\boldsymbol{s}) \geq \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s})}\big[q(\boldsymbol{s},\boldsymbol{a})\big] \qquad \forall \boldsymbol{s} \\ & q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}\big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}')\big] \ \forall \boldsymbol{s}\,\forall \boldsymbol{a}\end{aligned}$$

from which the Lagrangian can be written as

$$\begin{aligned}&\mathcal{L}_{\mathrm{PE}}(v(\boldsymbol{s}), q(\boldsymbol{s},\boldsymbol{a}), d(\boldsymbol{s}), p(\boldsymbol{s},\boldsymbol{a})) \\ &= \sum_{\boldsymbol{s}\in\mathcal{S}} \iota(\boldsymbol{s})(1-\gamma)v(\boldsymbol{s}) - \sum_{\boldsymbol{s}\in\mathcal{S}} d(\boldsymbol{s})\left[v(\boldsymbol{s}) - \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) q(\boldsymbol{s},\boldsymbol{a})\right] \\ &\quad - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a})\left[q(\boldsymbol{s},\boldsymbol{a}) - \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}')\big)\right] \\ &= \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') \\ &\quad - \sum_{\boldsymbol{s}\in\mathcal{S}} v(\boldsymbol{s})\left[d(\boldsymbol{s}) - (1-\gamma)\iota(\boldsymbol{s}) - \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\right] \\ &\quad - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} q(\boldsymbol{s},\boldsymbol{a})\big[p(\boldsymbol{s},\boldsymbol{a}) - d(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s})\big].\end{aligned}$$

The optimality (KKT) conditions can be derived as

- Stationarity of respectively v and q
$$\begin{aligned} d^\star(\boldsymbol{s}) &= (1-\gamma)\iota(\boldsymbol{s}) + \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p^\star(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}') \ \forall \boldsymbol{s} \\ p^\star(\boldsymbol{s},\boldsymbol{a}) &= d^\star(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}\end{aligned}$$
- Primal feasibility
$$\begin{aligned} v^\star(\boldsymbol{s}) &\geq \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) q^\star(\boldsymbol{s},\boldsymbol{a}) \qquad \forall \boldsymbol{s} \\ q^\star(\boldsymbol{s},\boldsymbol{a}) &= \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big) \ \forall \boldsymbol{s}\,\forall \boldsymbol{a}\end{aligned}$$
- Dual feasibility $\quad d^\star(\boldsymbol{s}) \geq 0 \ \forall \boldsymbol{s}$
- Complementary slackness
$$v^\star(\boldsymbol{s}) = \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) q^\star(\boldsymbol{s},\boldsymbol{a}) \vee d^\star(\boldsymbol{s}) = 0 \ \forall \boldsymbol{s}$$

and the dual is given by

$$\begin{aligned}\max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}}\big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big] \\ \text{s.t.} \quad & d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)}\big[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\big] \geq 0 \ \forall \boldsymbol{s} \\ & p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}\end{aligned}$$

From the optimality conditions, the following properties of the optimal primal and dual solutions of the Policy Evaluation LP can be proven.

Lemma 2. *The optimal dual variables $d^\star$ and $p^\star$ are probability distributions over $\mathcal{S}$ and $\mathcal{S}\times\mathcal{A}$ respectively. Furthermore, they are the visitation densities of the given policy π for the considered MDP.*

Proof. From the stationarity conditions, we can derive that $\sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}\in\mathcal{S}} d^\star(\boldsymbol{s}) = 1$. Combined with the dual feasibility conditions, which imply $d^\star(\boldsymbol{s}) \geq 0$ and $p^\star(\boldsymbol{s},\boldsymbol{a}) \geq 0$ for all $\boldsymbol{s}$ and $\boldsymbol{a}$, this means the optimal $p^\star(\boldsymbol{s},\boldsymbol{a})$ is a probability distribution over $\mathcal{S}\times\mathcal{A}$, with marginal distribution $d^\star(\boldsymbol{s})$. Substituting $p^\star$, the stationarity conditions can be rewritten as $d^\star(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma\sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} d^\star(\boldsymbol{s}')\pi(\boldsymbol{a}'|\boldsymbol{s}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')$, which is equivalent to the definition of the visitation density d^π. It follows that $p^\star(\boldsymbol{s},\boldsymbol{a}) = d^\star(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s})$ is also equivalent to the visitation density p^π. ■

Lemma 3. *The optimal primal variables $v^\star$ and $q^\star$ satisfy the Bellman equation and are equivalent to the value functions of the given policy π for the considered MDP.*

Proof. The complementary slackness conditions imply that for each visited state $\boldsymbol{s} \in \mathcal{S}^{\pi^\star}$, the primal feasibility constraints become tight, as $d^\star(\boldsymbol{s}) > 0$ for such states. Using the primal feasibility conditions to substitute $q^\star$, we then have $v^\star(\boldsymbol{s}) = \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$. Satisfying the Bellman equation, $v^\star$ is thus equivalent to the value function v^π of the given policy for the considered MDP. From the primal feasibility conditions $q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$, we can then conclude that $q^\star$ is also equivalent to the value function q^π of the given policy for the considered MDP. ■

A.3 Policy Optimization

The outer Policy Optimization problem (PO) of the Generalized Policy Iteration (GPI) scheme can be written as

$$\begin{aligned}\max_{\pi(\boldsymbol{a}|\boldsymbol{s})} \quad & \bar{r}^\pi \\ \text{s.t.} \quad & \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) = 1 \ \forall \boldsymbol{s} \\ & \pi(\boldsymbol{a}|\boldsymbol{s}) \geq 0 \ \forall \boldsymbol{s}\,\forall \boldsymbol{a}\end{aligned}$$

and has as Lagrangian

$$\begin{aligned}&\mathcal{L}_{\mathrm{PO}}(\pi(\boldsymbol{a}|\boldsymbol{s}), \lambda(\boldsymbol{s}), \mu(\boldsymbol{s},\boldsymbol{a})) \\ &\quad = -\bar{r}^\pi - \sum_{\boldsymbol{s}\in\mathcal{S}} \lambda(\boldsymbol{s})\left[1 - \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s})\right] - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} \mu(\boldsymbol{s},\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{s}) \\ &\quad = -\sum_{\boldsymbol{s}\in\mathcal{S}} \lambda(\boldsymbol{s}) - \bar{r}^\pi - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s})\big[\mu(\boldsymbol{s},\boldsymbol{a}) - \lambda(\boldsymbol{s})\big].\end{aligned}$$

The optimality (KKT) conditions are given by

- Stationarity of π
$$\lambda^\star(\boldsymbol{s}) = \mu^\star(\boldsymbol{s},\boldsymbol{a}) + \left(\frac{\partial \bar{r}^\pi}{\partial \pi(\boldsymbol{a}|\boldsymbol{s})}\right)_{\pi^\star} \ \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$
- Primal feasibility
$$\sum_{\boldsymbol{a}\in\mathcal{A}} \pi^\star(\boldsymbol{a}|\boldsymbol{s}) = 1 \wedge \pi^\star(\boldsymbol{a}|\boldsymbol{s}) \geq 0 \ \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$
- Dual feasibility $\quad \mu^\star(\boldsymbol{s},\boldsymbol{a}) \geq 0 \ \forall \boldsymbol{s}\,\forall \boldsymbol{a}$

- Complementary slackness

$$\mu^\star(\boldsymbol{s},\boldsymbol{a}) = 0 \vee \pi^\star(\boldsymbol{a}|\boldsymbol{s}) = 0 \;\forall \boldsymbol{s}\,\forall \boldsymbol{a}$$

From these optimality conditions, we can derive following property of the optimal solution to the Policy Optimization problem.

Proposition 6. *The optimal solution $\pi^\star$ is a conditional probability distribution over $\mathcal{A}$ given $\mathcal{S}$, and actions taken by this policy maximize the partial derivative of the average reward $\bar{r}^\pi$ with respect to the action sampling probability.*

Proof. From the primal feasibility conditions, it immediately follows that $\pi^\star$ is a conditional probability distribution over $\mathcal{A}$ given $\mathcal{S}$. Combining the dual feasibility and stationarity conditions, we can derive that $\lambda^\star(\boldsymbol{s}) \geq \left(\frac{\partial \bar{r}^\pi}{\partial \pi(\boldsymbol{a}|\boldsymbol{s})}\right)_{\pi^\star}$ for all $\boldsymbol{s}$ and $\boldsymbol{a}$. The complementary slackness conditions then imply that this inequality becomes tight for all actions $\boldsymbol{a} \in \mathcal{A}^{\pi^\star}(\boldsymbol{s})$ taken by the policy, i.e. $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} \left(\frac{\partial \bar{r}^\pi}{\partial \pi(\boldsymbol{a}|\boldsymbol{s})}\right)_{\pi^\star}$. ∎

More interesting properties can be derived when integrating an inner Policy Evaluation problem, resulting in a GPI scheme. The partial derivative of the average reward $\bar{r}^\pi$ can then typically be evaluated using Danskin's theorem.

Theorem 2. *The optimal solution $\pi^\star$ of the nested Policy Optimization (PO) and Policy Evaluation (PE) problems, is the optimal policy π^* for the considered MDP.*

Proof. The partial derivative of the expected discounted return $\bar{r}^\pi$ can be evaluated using Danskin's theorem as $\left(\frac{\partial \bar{r}^\pi}{\partial \pi(\boldsymbol{a}|\boldsymbol{s})}\right)_{\pi^\star} = \left(\frac{\partial \mathcal{L}_{\text{PE}}}{\partial \pi(\boldsymbol{a}|\boldsymbol{s})}\right)_{\pi^\star,d^\star,q^\star} = d^{\pi^\star}(\boldsymbol{s}) q^{\pi^\star}(\boldsymbol{s},\boldsymbol{a})$. From the optimality conditions and Proposition 6, we then have $\lambda^\star(\boldsymbol{s}) \geq d^{\pi^\star}(\boldsymbol{s}) q^{\pi^\star}(\boldsymbol{s},\boldsymbol{a})$ for all $\boldsymbol{s}$ and $\boldsymbol{a}$, and $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} d^{\pi^\star}(\boldsymbol{s}) q^{\pi^\star}(\boldsymbol{s},\boldsymbol{a})$. Hence, $\pi^\star$ is a greedy policy with respect to its own value function $q^{\pi^\star}$. This means $q^{\pi^\star}$ satisfies the Bellman optimality equation and $\pi^\star$ is an optimal policy for the considered MDP. ∎

A.4 Entropy Regularization

The altered policy evaluation dual with added KL-divergence regularization term is given by

$$\begin{aligned}
\max_{\substack{d(\boldsymbol{s}) \\ p(\boldsymbol{s},\boldsymbol{a})}} \; & \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot) \\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}} \left[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\right] - \alpha \mathop{\mathbb{E}}_{\boldsymbol{s}\sim d(\cdot)} \left[\mathrm{KL}\left[\pi(\cdot|\boldsymbol{s}) \,\middle|\middle|\, \pi_{\mathrm{T}}(\cdot|\boldsymbol{s})\right]\right] \\
\text{s.t.} \quad & d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)} \left[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\right] \geq 0 \;\forall \boldsymbol{s} \\
& p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) \qquad\qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned}$$

The Lagrangian of this regularized LP is given by

$$\begin{aligned}
&\mathcal{L}_{\text{ER}}(d(\boldsymbol{s}), p(\boldsymbol{s},\boldsymbol{a}), v(\boldsymbol{s}), q(\boldsymbol{s},\boldsymbol{a}), \mu(\boldsymbol{s})) \\
&\quad = -\sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} d(\boldsymbol{s}) \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) \alpha \log \frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s})} \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} v(\boldsymbol{s}) \left[d(\boldsymbol{s}) - (1-\gamma)\iota(\boldsymbol{s}) - \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}} \sum_{\boldsymbol{a}'\in\mathcal{A}} p(\boldsymbol{s}',\boldsymbol{a}') \tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}') \right] \\
&\qquad - \sum_{\boldsymbol{s}\in\mathcal{S}} \mu(\boldsymbol{s}) d(\boldsymbol{s}) + \sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{a}\in\mathcal{A}} q(\boldsymbol{s},\boldsymbol{a}) \left[p(\boldsymbol{s},\boldsymbol{a}) - d(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) \right] \\
&\quad = -\sum_{\boldsymbol{s}\in\mathcal{S}} \iota(\boldsymbol{s})(1-\gamma) v(\boldsymbol{s}) \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} d(\boldsymbol{s}) \left[v(\boldsymbol{s}) - \mu(\boldsymbol{s}) - \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) \left(q(\boldsymbol{s},\boldsymbol{a}) - \alpha \log \frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s})} \right) \right] \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \left[q(\boldsymbol{s},\boldsymbol{a}) - \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \left(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}') \right) \right].
\end{aligned}$$

The optimality (KKT) conditions can then be written as

- Stationarity of respectively d and p

$$\begin{aligned}
\mu^\star(\boldsymbol{s}) &= v^\star(\boldsymbol{s}) - \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) \left(q^\star(\boldsymbol{s},\boldsymbol{a}) - \alpha \log \frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s})} \right) \;\forall \boldsymbol{s} \\
q^\star(\boldsymbol{s},\boldsymbol{a}) &= \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \left(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}') \right) \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned}$$

- Dual feasibility

$$\begin{aligned}
d^\star(\boldsymbol{s}) &= (1-\gamma)\iota(\boldsymbol{s}) \\
&\qquad + \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}} \sum_{\boldsymbol{a}'\in\mathcal{A}} p^\star(\boldsymbol{s}',\boldsymbol{a}') \tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}') \geq 0 \;\forall \boldsymbol{s} \\
p^\star(\boldsymbol{s},\boldsymbol{a}) &= d^\star(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) \qquad\qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned}$$

- Primal feasibility

$$v^\star(\boldsymbol{s}) \geq \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) \left(q^\star(\boldsymbol{s},\boldsymbol{a}) - \alpha \log \frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s})} \right) \;\forall \boldsymbol{s}$$

- Complementary slackness

$$\begin{aligned}
v^\star(\boldsymbol{s}) = \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) \left(q^\star(\boldsymbol{s},\boldsymbol{a}) - \alpha \log \frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s})} \right) \\
\vee\, d^\star(\boldsymbol{s}) = 0 \;\forall \boldsymbol{s}
\end{aligned}$$

and the primal formulation is given by

$$\begin{aligned}
\min_{\substack{v(\boldsymbol{s}) \\ q(\boldsymbol{s},\boldsymbol{a})}} \; & \mathop{\mathbb{E}}_{\boldsymbol{s}_0\sim\iota(\cdot)} \left[(1-\gamma) v(\boldsymbol{s}_0)\right] \\
\text{s.t.} \quad & v(\boldsymbol{s}) \geq \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s})} \left[q(\boldsymbol{s},\boldsymbol{a}) - \alpha \log \frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s})} \right] \;\forall \boldsymbol{s} \\
& q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})} \left[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}') \right] \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned}$$

Lemma 2 can be proven for this regularized Policy Evaluation problem without major changes. Lemma 3 also holds, albeit with a modified definition of the reward (4) and value functions, taking the entropy regularization into account.

Lemma 4. *The optimal primal variables $v^\star$ and $q^\star$ satisfy the entropy regularized Bellman equation and are equivalent to the entropy regularized value functions of the given policy π for the considered MDP.*

Proof. Following similar arguments as in Lemma 3, we obtain $v^\star(\boldsymbol{s}) = \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') - \alpha \log \frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s})} + \gamma v^\star(\boldsymbol{s}')\big)$. Denoting by $r_{\mathrm{ER}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') - \alpha \log \frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s})}$ the entropy-regularized reward, we see that $v^\star$ satisfies the Bellman equation using this modified reward. We can thus conclude that $v^\star$ (and $q^\star$) are equivalent to the entropy-regularized value function v^π_{ER} (and q^π_{ER}) of the given policy for the considered MDP. ■

When plugging this entropy regularized policy evaluation LP in the Policy Optimization problem, we can prove the following property.

Theorem 3. *The optimal solution $\pi^\star$ of the nested Policy Optimization (PO) and Entropy-Regularized Policy Evaluation (ER) problems, is a conditional Boltzmann policy with as energy function the negation of the sum of the entropy-regularized value function $q^\star_{\mathrm{ER}}$ and the teacher policy's log probability.*

Proof. The partial derivative of the average reward $\bar{r}^\pi$ can be evaluated using Danskin's theorem as $\left(\frac{\partial \bar{r}^\pi}{\partial \pi(\boldsymbol{a}|\boldsymbol{s})}\right)_{\pi^\star} = -\left(\frac{\partial \mathcal{L}_{\mathrm{ER}}}{\partial \pi(\boldsymbol{a}|\boldsymbol{s})}\right)_{\pi^\star,d^\star,q^\star} = d^{\pi^\star}(\boldsymbol{s})\Big[q^{\pi^\star}_{\mathrm{ER}}(\boldsymbol{s},\boldsymbol{a}) - \alpha - \alpha \log \frac{\pi^\star(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s})}\Big]$. From the optimality conditions and Proposition 6, we then have $\lambda^\star(\boldsymbol{s}) \geq d^{\pi^\star}(\boldsymbol{s})\Big[q^{\pi^\star}_{\mathrm{ER}}(\boldsymbol{s},\boldsymbol{a}) - \alpha - \alpha \log \frac{\pi^\star(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s})}\Big]$ for all $\boldsymbol{s}$ and $\boldsymbol{a}$, and the equality holds for all actions taken by the policy $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}}\Big[q^{\pi^\star}_{\mathrm{ER}}(\boldsymbol{s},\boldsymbol{a}) - \alpha \log \frac{\pi^\star(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s})}\Big]$. This means that $\pi^\star(\boldsymbol{a}|\boldsymbol{s}) > 0$ for all actions; as otherwise the maximum would become unbounded and be reached for the actions that are not taken. Hence $\alpha \log \frac{\pi^\star(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s})} = q^{\pi^\star}_{\mathrm{ER}}(\boldsymbol{s},\boldsymbol{a}) - \frac{\lambda^\star(\boldsymbol{s})}{d^{\pi^\star}(\boldsymbol{s})} - \alpha$ holds for all states and actions, implying that $\pi^\star(\boldsymbol{a}|\boldsymbol{s}) \propto \pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s}) \exp\left(\frac{q^{\pi^\star}_{\mathrm{ER}}(\boldsymbol{s},\boldsymbol{a})}{\alpha}\right)$. The optimal policy $\pi^\star$ is thus a conditional Boltzmann distribution with energy function $-q^{\pi^\star}_{\mathrm{ER}}(\boldsymbol{s},\boldsymbol{a}) - \alpha \log \pi_{\mathrm{T}}(\boldsymbol{a}|\boldsymbol{s})$ and temperature α. ■

The resulting nested optimization problem can be seen as a *Soft Policy Iteration* scheme, for which convergence proofs can be derived, e.g. by Haarnoja et al. [14] for the case of maximum entropy RL (or a uniform teacher policy).

A.5 Constrained Reinforcement Learning

The constrained dual value learning problem can be written as

$$
\begin{aligned}
\max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \; & \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}}\left[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\right] \\
\text{s.t.} \; & \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s}) && \forall \boldsymbol{s} \\
& d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)}\left[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\right] && \forall \boldsymbol{s} \\
& p(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
& (1-\gamma)V_k \leq \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}}[r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')] && \forall k
\end{aligned}
$$

The Lagrangian of this altered dual problem can be obtained as

$$
\begin{aligned}
&\mathcal{L}_{\mathrm{CRL}}(d(\boldsymbol{s}), p(\boldsymbol{s},\boldsymbol{a}), \lambda(\boldsymbol{s}), v(\boldsymbol{s}), \mu(\boldsymbol{s},\boldsymbol{a}), w_k) \\
&= -\sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') \\
&\quad + \sum_{\boldsymbol{s}\in\mathcal{S}} \lambda(\boldsymbol{s}) \left[\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) - d(\boldsymbol{s})\right] \\
&\quad + \sum_{\boldsymbol{s}\in\mathcal{S}} v(\boldsymbol{s}) \left[d(\boldsymbol{s}) - (1-\gamma)\iota(\boldsymbol{s}) - \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\right] \\
&\quad - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} \mu(\boldsymbol{s},\boldsymbol{a}) p(\boldsymbol{s},\boldsymbol{a}) \\
&\quad - \sum_{k=0}^{K} w_k \left[\sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') - (1-\gamma)V_k\right] \\
&= -\sum_{\boldsymbol{s}\in\mathcal{S}} \iota(\boldsymbol{s})(1-\gamma)v(\boldsymbol{s}) + \sum_{k=0}^{K} w_k (1-\gamma) V_k \\
&\quad + \sum_{\boldsymbol{s}\in\mathcal{S}} d(\boldsymbol{s})[v(\boldsymbol{s}) - \lambda(\boldsymbol{s})] \\
&\quad + \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \Bigg[\lambda(\boldsymbol{s}) - \mu(\boldsymbol{s},\boldsymbol{a}) \\
&\qquad\qquad - \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \Big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}') \\
&\qquad\qquad\qquad + \sum_{k=0}^{K} w_k r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\Big)\Bigg],
\end{aligned}
$$

resulting in following optimality (KKT) conditions

- Stationarity of respectively d and p

$$
\begin{aligned}
&\lambda^\star(\boldsymbol{s}) = v^\star(\boldsymbol{s}) && \forall \boldsymbol{s} \\
&\mu^\star(\boldsymbol{s},\boldsymbol{a}) = v^\star(\boldsymbol{s}) - \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \Big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}') \\
&\qquad\qquad + \sum_{k=0}^{K} w_k^\star r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\Big] && \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned}
$$

- Dual feasibility

$$
\begin{aligned}
&\sum_{\boldsymbol{a}\in\mathcal{A}} p^\star(\boldsymbol{s},\boldsymbol{a}) = d^\star(\boldsymbol{s}) && \forall \boldsymbol{s} \\
&d^\star(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p^\star(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}') && \forall \boldsymbol{s} \\
&p^\star(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
&(1-\gamma)V_k \leq \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p^\star(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') && \forall k
\end{aligned}
$$

- Primal feasibility, introducing auxiliary q

$$
\begin{aligned}
&v^\star(\boldsymbol{s}) \geq q^\star(\boldsymbol{s},\boldsymbol{a}) && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
&q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \Big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \sum_{k=0}^{K} w_k^\star r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') \\
&\qquad\qquad + \gamma v^\star(\boldsymbol{s}')\Big] && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
&w_k^\star \geq 0 && \forall k
\end{aligned}
$$

- Complementary slackness

$$
\begin{aligned}
&v^\star(\boldsymbol{s}) = q^\star(\boldsymbol{s},\boldsymbol{a}) \vee p^\star(\boldsymbol{s},\boldsymbol{a}) = 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
&(1-\gamma)V_k = \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p^\star(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') \\
&\qquad\qquad \vee w_k^\star = 0 && \forall k
\end{aligned}
$$

The corresponding primal formulation is thus derived as

$$\begin{aligned}
\min_{\substack{v(\boldsymbol{s}),w_k\\ q(\boldsymbol{s},\boldsymbol{a})}} \; & \mathop{\mathbb{E}}_{\boldsymbol{s}_0\sim\iota(\cdot)}\big[(1-\gamma)v(\boldsymbol{s}_0)\big] - \sum_{k=0}^{K} w_k(1-\gamma)V_k \\
\text{s.t.} \quad & v(\boldsymbol{s}) \geq q(\boldsymbol{s},\boldsymbol{a}) && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
& q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}\left[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \sum_{k=0}^{K} w_k r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}')\right] && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
& w_k \geq 0 && \forall k
\end{aligned}$$

From the optimality conditions, it is immediately clear Proposition 5 still holds for this constrained variant and it can be extended as follows.

Proposition 7. *The policy $\pi(\boldsymbol{a}|\boldsymbol{s})$ that can be derived from any feasible dual variables d and p is a feasible policy for the considered CMDP.*

Proof. From Proposition 5, we know that the dual variables correspond to the visitation densities of a policy π with $p(\boldsymbol{s},\boldsymbol{a}) = d^\pi(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s})$. The K additional dual constraints can hence be rewritten in terms of the policy π as $V_k \leq (1-\gamma)^{-1}\,\mathbb{E}_{\boldsymbol{s}\sim d^\pi(\cdot),\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s}),\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}[r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')] = \nu_k^\pi$. The dual variables d and p are thus the visitation densities of a policy satisfying the CMDP constraints. ■

As a result, Theorem 1 can also be straightforwardly adapted as follows.

Theorem 4. *The policy $\pi^\star$ that can be derived from the optimal dual variables $d^\star$ and $p^\star$ is an optimal policy for the considered CMDP.*

Proof. Combine the results from the previous Proposition with the proof of Theorem 1. ■

Lemma 1 also holds, albeit for the modified reward function r_{CRL} (5), with learnable parameters w_k, and corresponding value functions.

Lemma 5. *The optimal primal variables $v^\star$ and $q^\star$ satisfy the Bellman optimality equation and are thus equivalent to the optimal adjusted value functions v^*_{CRL} and q^*_{CRL} for the considered MDP with optimal modified reward $r^\star_{\text{CRL}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \sum_{k=0}^{K} w_k^\star r_k(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')$.*

Proof. Following the same arguments as in Lemma 1, we obtain $v^\star(\boldsymbol{s}) = \max_{\boldsymbol{a}} \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\text{CRL}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$. Satisfying the Bellman optimality equation, $v^\star$ is thus equivalent to the optimal value function v^*_{CRL} for the considered MDP with altered reward $r^\star_{\text{CRL}}$. From the primal feasibility conditions $q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\text{CRL}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$, we then also obtain the equivalency between $q^\star$ and q^*_{CRL}. ■

Finally, these optimal adjusted value functions and optimal reward modifications satisfy following properties.

Theorem 5. *The optimal policy $\pi^\star$ is greedy with respect to the optimal adjusted value functions v^*_{CRL} and q^*_{CRL}.*

Proof. Following the same arguments as in the proof for Lemma 1, we obtain $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \mathcal{A}^*(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\text{CRL}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$. From the previous Lemma, it then follows that $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} q^*_{\text{CRL}}(\boldsymbol{s},\boldsymbol{a})$, showing the greediness of $\pi^\star$ with respect to the optimal adjusted value function. ■

Proposition 1. *The optimal modified reward $r^\star_{\text{CRL}}$ is only altered by the tight constraints $V_k = \nu_k^{\pi^\star}$.*

Proof. This follows immediately from the complementary slackness conditions, as $w_k^\star = 0$ (and hence the corresponding r_k does not influence the modified reward r_{CRL}) for the strictly satisfied constraints $V_k < \nu_k^{\pi^\star}$. ■

A.6 Visitation Density Bounds

A.6.1 Value Learning

The altered dual value learning problem with state-visitation density bounds is given by

$$\begin{aligned}
\max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \; & \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}}[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')] \\
\text{s.t.} \quad & \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s}) && \forall \boldsymbol{s} \\
& d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)}[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')] && \forall \boldsymbol{s} \\
& p(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
& d_{\text{L}}(\boldsymbol{s}) \leq d(\boldsymbol{s}) \leq d_{\text{U}}(\boldsymbol{s}) && \forall \boldsymbol{s}
\end{aligned}$$

for which the Lagrangian can be derived as

$$\begin{aligned}
&\mathcal{L}_{VDB}(d(\boldsymbol{s}), p(\boldsymbol{s},\boldsymbol{a}), \lambda(\boldsymbol{s}), v(\boldsymbol{s}), \mu(\boldsymbol{s},\boldsymbol{a}), r_{\text{L}}(\boldsymbol{s}), r_{\text{U}}(\boldsymbol{s})) \\
&= -\sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') \\
&\quad + \sum_{\boldsymbol{s}\in\mathcal{S}} \lambda(\boldsymbol{s}) \left[\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) - d(\boldsymbol{s})\right] \\
&\quad + \sum_{\boldsymbol{s}\in\mathcal{S}} v(\boldsymbol{s}) \left[d(\boldsymbol{s}) - (1-\gamma)\iota(\boldsymbol{s}) - \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\right] \\
&\quad - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} \mu(\boldsymbol{s},\boldsymbol{a}) p(\boldsymbol{s},\boldsymbol{a}) \\
&\quad - \sum_{\boldsymbol{s}\in\mathcal{S}} r_{\text{L}}(\boldsymbol{s})[d(\boldsymbol{s}) - d_{\text{L}}(\boldsymbol{s})] - \sum_{\boldsymbol{s}\in\mathcal{S}} r_{\text{U}}(\boldsymbol{s})[d_{\text{U}}(\boldsymbol{s}) - d(\boldsymbol{s})] \\
&= -\sum_{\boldsymbol{s}\in\mathcal{S}} \iota(\boldsymbol{s})(1-\gamma)v(\boldsymbol{s}) + \sum_{\boldsymbol{s}\in\mathcal{S}} r_{\text{L}}(\boldsymbol{s}) d_{\text{L}}(\boldsymbol{s}) - \sum_{\boldsymbol{s}\in\mathcal{S}} r_{\text{U}}(\boldsymbol{s}) d_{\text{U}}(\boldsymbol{s}) \\
&\quad + \sum_{\boldsymbol{s}\in\mathcal{S}} d(\boldsymbol{s})[v(\boldsymbol{s}) - \lambda(\boldsymbol{s}) - r_{\text{L}}(\boldsymbol{s}) + r_{\text{U}}(\boldsymbol{s})] \\
&\quad + \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \Bigg[\lambda(\boldsymbol{s}) - \mu(\boldsymbol{s},\boldsymbol{a}) \\
&\qquad\qquad - \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}')\big)\Bigg],
\end{aligned}$$

resulting in following optimality (KKT) conditions

- Stationarity of respectively d and p

$$\begin{aligned}
\lambda^\star(\boldsymbol{s}) &= v^\star(\boldsymbol{s}) - r^\star_{\text{L}}(\boldsymbol{s}) + r^\star_{\text{U}}(\boldsymbol{s}) && \forall \boldsymbol{s} \\
\mu^\star(\boldsymbol{s},\boldsymbol{a}) &= v^\star(\boldsymbol{s}) - r^\star_{\text{L}}(\boldsymbol{s}) + r^\star_{\text{U}}(\boldsymbol{s}) \\
&\quad - \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big) && \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned}$$

- Dual feasibility

$$\sum_{\boldsymbol{a}\in\mathcal{A}} p^\star(\boldsymbol{s},\boldsymbol{a}) = d^\star(\boldsymbol{s}) \qquad \forall \boldsymbol{s}$$
$$d^\star(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p^\star(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}') \;\forall \boldsymbol{s}$$
$$p^\star(\boldsymbol{s},\boldsymbol{a}) \geq 0 \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$
$$d_{\text{L}}(\boldsymbol{s}) \leq d^\star(\boldsymbol{s}) \leq d_{\text{U}}(\boldsymbol{s}) \qquad \forall \boldsymbol{s}$$

- Primal feasibility, introducing auxiliary q

$$v^\star(\boldsymbol{s}) \geq q^\star(\boldsymbol{s},\boldsymbol{a}) \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$
$$q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + r_{\text{L}}^\star(\boldsymbol{s}) - r_{\text{U}}^\star(\boldsymbol{s}) + \gamma v^\star(\boldsymbol{s}')\big) \;\forall \boldsymbol{s}\,\forall \boldsymbol{a}$$
$$r_{\text{L}}^\star(\boldsymbol{s}) \geq 0 \wedge r_{\text{U}}^\star(\boldsymbol{s}) \geq 0 \qquad \forall \boldsymbol{s}$$

- Complementary slackness

$$v^\star(\boldsymbol{s}) = q^\star(\boldsymbol{s},\boldsymbol{a}) \vee p^\star(\boldsymbol{s},\boldsymbol{a}) = 0 \;\forall \boldsymbol{s}\,\forall \boldsymbol{a}$$
$$d^\star(\boldsymbol{s}) = d_{\text{L}}(\boldsymbol{s}) \vee r_{\text{L}}^\star(\boldsymbol{s}) = 0 \qquad \forall \boldsymbol{s}$$
$$d^\star(\boldsymbol{s}) = d_{\text{U}}(\boldsymbol{s}) \vee r_{\text{U}}^\star(\boldsymbol{s}) = 0 \qquad \forall \boldsymbol{s}$$

and primal formulation

$$\min_{\substack{v(\boldsymbol{s}),q(\boldsymbol{s},\boldsymbol{a})\\ r_{\text{L}}(\boldsymbol{s}),r_{\text{U}}(\boldsymbol{s})}} \mathop{\mathbb{E}}_{\boldsymbol{s}_0\sim\iota(\cdot)}\big[(1-\gamma)v(\boldsymbol{s}_0)\big] - \sum_{\boldsymbol{s}\in\mathcal{S}} r_{\text{L}}(\boldsymbol{s})d_{\text{L}}(\boldsymbol{s}) + \sum_{\boldsymbol{s}\in\mathcal{S}} r_{\text{U}}(\boldsymbol{s})d_{\text{U}}(\boldsymbol{s})$$
$$\text{s.t.} \quad v(\boldsymbol{s}) \geq q(\boldsymbol{s},\boldsymbol{a}) \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$
$$q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}\big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + r_{\text{L}}(\boldsymbol{s}) - r_{\text{U}}(\boldsymbol{s}) + \gamma v(\boldsymbol{s}')\big] \;\forall \boldsymbol{s}\,\forall \boldsymbol{a}$$
$$r_{\text{L}}(\boldsymbol{s}) \geq 0 \wedge r_{\text{U}}(\boldsymbol{s}) \geq 0 \qquad \forall \boldsymbol{s}$$

From the optimality conditions, it is immediately clear Proposition 5 still holds for this constrained variant and it can be extended as follows.

Proposition 8. *The policy $\pi(\boldsymbol{a}|\boldsymbol{s})$ that can be derived from any feasible dual variables d and p is a feasible policy for the considered MDP with visitation density bounds.*

Proof. From Proposition 5, we know that the dual variables correspond to the visitation densities of a policy π. The additional constraints can thus be rewritten as $d_{\text{L}}(\boldsymbol{s}) \leq d^\pi(\boldsymbol{s}) \leq d_{\text{U}}(\boldsymbol{s})$. The dual variables d and p are thus the visitation densities of a policy satisfying those visitation density bounds. ■

Consequently, Theorem 1 can also be straightforwardly adapted as follows.

Lemma 9. *The policy $\pi^\star$ that can be derived from the optimal dual variables $d^\star$ and $p^\star$ is an optimal policy for the MDP with visitation density bounds.*

Proof. Combine the results from the previous Proposition with the proof of Theorem 1. ■

Lemma 1 also holds, albeit for the modified reward function $r_{\text{VDB}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + r_{\text{L}}(\boldsymbol{s}) - r_{\text{U}}(\boldsymbol{s})$, with learnable parameters r_{L} and r_{U}, and corresponding value functions.

Lemma 10. *The optimal primal variables $v^\star$ and $q^\star$ satisfy the Bellman optimality equation and are thus equivalent to the optimal adjusted value functions v^*_{VDB} and q^*_{VDB} for the considered MDP with optimal modified reward $r^\star_{\text{VDB}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + r_{\text{L}}^\star(\boldsymbol{s}) - r_{\text{U}}^\star(\boldsymbol{s})$.*

Proof. Following the same arguments as in Lemma 1, we obtain $v^\star(\boldsymbol{s}) = \max_{\boldsymbol{a}} \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\text{VDB}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$. Satisfying the Bellman optimality equation, $v^\star$ is thus equivalent to the optimal value function v^*_{VDB} for the considered MDP with altered reward $r^\star_{\text{VDB}}$. From the primal feasibility conditions $q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\text{VDB}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$, we then also obtain the equivalency between $q^\star$ and q^*_{VDB}. ■

Finally, these optimal adjusted value functions and optimal reward modifications satisfy following properties.

Lemma 11. *The optimal policy $\pi^\star$ is greedy with respect to the optimal adjusted value functions v^*_{VDB} and q^*_{VDB}.*

Proof. Following the same arguments as in the proof for Lemma 1, we obtain $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \mathcal{A}^*(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\text{VDB}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$. From the previous Lemma, it then follows that $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} q^*_{\text{VDB}}(\boldsymbol{s},\boldsymbol{a})$, showing the greediness of $\pi^\star$ with respect to the optimal adjusted value functions. ■

Lemma 12. *The optimal modified reward $r^\star_{\text{VDB}}$ is only altered by the tight bounds $d_{\text{L}}(\boldsymbol{s}) = d^{\pi^\star}(\boldsymbol{s})$ and $d^{\pi^\star}(\boldsymbol{s}) = d_{\text{U}}(\boldsymbol{s})$.*

Proof. This follows immediately from the complementary slackness conditions, as $r_{\text{L}}^\star(\boldsymbol{s}) = 0$ for the strictly satisfied constraints $d_{\text{L}}(\boldsymbol{s}) < d^{\pi^\star}(\boldsymbol{s})$, while $r_{\text{U}}^\star(\boldsymbol{s}) = 0$ for the strictly satisfied constraints $d^{\pi^\star}(\boldsymbol{s}) < d_{\text{U}}(\boldsymbol{s})$. ■

In case there is no policy that satisfies all additional constraints, the dual problem effectively becomes infeasible and the primal becomes unbounded, caused by the $r_{\text{L}}(\boldsymbol{s},\boldsymbol{a})$ and $r_{\text{U}}(\boldsymbol{s},\boldsymbol{a})$ of the infeasible constraints approaching infinity.

A.6.2 Policy Iteration

The altered dual policy evaluation problem with state-action-visitation density bounds can be written as

$$\max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}}\big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big]$$
$$\text{s.t.} \quad d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)}\big[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\big] \geq 0 \;\forall \boldsymbol{s}$$
$$p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$
$$p_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) \leq p(\boldsymbol{s},\boldsymbol{a}) \leq p_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) \qquad \forall \boldsymbol{s}\,\forall \boldsymbol{a}$$

The Lagrangian of this altered LP is given by

$$
\begin{aligned}
&\mathcal{L}_{\text{VDB}}(d(\boldsymbol{s}), p(\boldsymbol{s},\boldsymbol{a}), v(\boldsymbol{s}), q(\boldsymbol{s},\boldsymbol{a}), \mu(\boldsymbol{s}), r_{\text{L}}(\boldsymbol{s},\boldsymbol{a}), r_{\text{U}}(\boldsymbol{s},\boldsymbol{a})) \\
&\quad = -\sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} v(\boldsymbol{s}) \left[d(\boldsymbol{s}) - (1-\gamma)\iota(\boldsymbol{s}) - \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}') \right] \\
&\qquad - \sum_{\boldsymbol{s}\in\mathcal{S}} \mu(\boldsymbol{s}) d(\boldsymbol{s}) + \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} q(\boldsymbol{s},\boldsymbol{a}) \big[p(\boldsymbol{s},\boldsymbol{a}) - d(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) \big] \\
&\qquad - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} r_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) [p(\boldsymbol{s},\boldsymbol{a}) - p_{\text{L}}(\boldsymbol{s},\boldsymbol{a})] \\
&\qquad - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} r_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) [p_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) - p(\boldsymbol{s},\boldsymbol{a})] \\
&\quad = -\sum_{\boldsymbol{s}\in\mathcal{S}} \iota(\boldsymbol{s})(1-\gamma) v(\boldsymbol{s}) \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} r_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) p_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} r_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) p_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} d(\boldsymbol{s}) \left[v(\boldsymbol{s}) - \mu(\boldsymbol{s}) - \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) q(\boldsymbol{s},\boldsymbol{a}) \right] \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \Bigg[q(\boldsymbol{s},\boldsymbol{a}) - r_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) + r_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) \\
&\qquad\qquad - \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}') \big) \Bigg],
\end{aligned}
$$

from which following optimality (KKT) conditions can be derived

- Stationarity of respectively d and p

$$
\begin{aligned}
&\mu^\star(\boldsymbol{s}) = v^\star(\boldsymbol{s}) - \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) q^\star(\boldsymbol{s},\boldsymbol{a}) && \forall \boldsymbol{s} \\
&q^\star(\boldsymbol{s},\boldsymbol{a}) = r_{\text{L}}^\star(\boldsymbol{s},\boldsymbol{a}) - r_{\text{U}}^\star(\boldsymbol{s},\boldsymbol{a}) \\
&\qquad\qquad + \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}') \big) && \forall \boldsymbol{s}\, \forall \boldsymbol{a}
\end{aligned}
$$

- Dual feasibility

$$
\begin{aligned}
&d^\star(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}') \geq 0 && \forall \boldsymbol{s} \\
&p^\star(\boldsymbol{s},\boldsymbol{a}) = d^\star(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) && \forall \boldsymbol{s}\, \forall \boldsymbol{a} \\
&p_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) \leq p^\star(\boldsymbol{s},\boldsymbol{a}) \leq p_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) && \forall \boldsymbol{s}\, \forall \boldsymbol{a}
\end{aligned}
$$

- Primal feasibility

$$
\begin{aligned}
&v^\star(\boldsymbol{s}) \geq \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) q^\star(\boldsymbol{s},\boldsymbol{a}) \;\forall \boldsymbol{s} \\
&r_{\text{L}}^\star(\boldsymbol{s},\boldsymbol{a}) \geq 0 \wedge r_{\text{U}}^\star(\boldsymbol{s},\boldsymbol{a}) \geq 0 \qquad \forall \boldsymbol{s}\, \forall \boldsymbol{a}
\end{aligned}
$$

- Complementary slackness

$$
\begin{aligned}
&v^\star(\boldsymbol{s}) = \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) q^\star(\boldsymbol{s},\boldsymbol{a}) \vee d^\star(\boldsymbol{s}) = 0 \;\forall \boldsymbol{s} \\
&p^\star(\boldsymbol{s},\boldsymbol{a}) = p_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) \vee r_{\text{L}}^\star(\boldsymbol{s},\boldsymbol{a}) = 0 && \forall \boldsymbol{s}\, \forall \boldsymbol{a} \\
&p^\star(\boldsymbol{s},\boldsymbol{a}) = p_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) \vee r_{\text{U}}^\star(\boldsymbol{s},\boldsymbol{a}) = 0 && \forall \boldsymbol{s}\, \forall \boldsymbol{a}
\end{aligned}
$$

The corresponding primal formulation can be written as

$$
\begin{aligned}
\min_{\substack{v(\boldsymbol{s}),q(\boldsymbol{s},\boldsymbol{a}) \\ r_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) \\ r_{\text{U}}(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathbb{E}_{\boldsymbol{s}_0\sim\iota(\cdot)} \big[(1-\gamma) v(\boldsymbol{s}_0) \big] - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} r_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) p_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) \\
& \qquad + \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} r_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) p_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) \\
\text{s.t.} \quad & v(\boldsymbol{s}) \geq \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s})} \big[q(\boldsymbol{s},\boldsymbol{a}) \big] && \forall \boldsymbol{s} \\
& q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})} \big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + r_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) - r_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) + \gamma v(\boldsymbol{s}') \big] && \forall \boldsymbol{s}\, \forall \boldsymbol{a} \\
& r_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) \geq 0 \wedge r_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\, \forall \boldsymbol{a}
\end{aligned}
$$

Lemma 2 can be proven for this constrained Policy Evaluation problem without major changes. Lemma 3 also holds, albeit for the modified reward function r_{VDB} (6), with learnable parameters r_{L} and r_{U}, and corresponding value functions.

Lemma 6. *The optimal primal variables $v^\star$ and $q^\star$ satisfy the Bellman equation and are equivalent to the adjusted value functions of the given policy π for the considered MDP with optimal modified reward $r_{\text{VDB}}^\star(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + r_{\text{L}}^\star(\boldsymbol{s},\boldsymbol{a}) - r_{\text{U}}^\star(\boldsymbol{s},\boldsymbol{a})$.*

Proof. Following similar arguments as in Lemma 3, we obtain $v^\star(\boldsymbol{s}) = \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \big(r_{\text{VDB}}^\star(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}') \big)$ and $q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \big(r_{\text{VDB}}^\star(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}') \big)$. From the former equality it is clear that $v^\star$ satisfies the Bellman equation and is thus equivalent to the adjusted value function v_{VDB}^π of the given policy for the considered MDP with altered reward $r_{\text{VDB}}^\star$. From the latter equality we can then obtain the equivalency between $q^\star$ and q_{VDB}^π. ∎

Furthermore, these optimal reward modifications satisfy following property.

Proposition 2. *The optimal modified reward $r_{\text{VDB}}^\star$ is only altered by the tight bounds $p_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) = p^\pi(\boldsymbol{s},\boldsymbol{a})$ and $p^\pi(\boldsymbol{s},\boldsymbol{a}) = p_{\text{U}}(\boldsymbol{s},\boldsymbol{a})$.*

Proof. This follows immediately from the complementary slackness conditions, as $r_{\text{L}}^\star(\boldsymbol{s},\boldsymbol{a}) = 0$ for the strictly satisfied constraints $p_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) < p^\pi(\boldsymbol{s},\boldsymbol{a})$, while $r_{\text{U}}^\star(\boldsymbol{s},\boldsymbol{a}) = 0$ for the strictly satisfied constraints $p^\pi(\boldsymbol{s},\boldsymbol{a}) < p_{\text{U}}(\boldsymbol{s},\boldsymbol{a})$. ∎

By plugging this altered policy evaluation LP in the Policy Optimization problem, we can prove the following property.

Theorem 6. *The optimal solution $\pi^\star$ of the nested Policy Optimization (PO) and constrained Policy Evaluation (VDB) problems, is the optimal policy π^* for the considered MDP with visitation density bounds.*

Proof. The partial derivative of the average reward $\bar{r}^\pi$ can be evaluated using Danskin's theorem as $\left(\frac{\partial \bar{r}^\pi}{\partial \pi(\boldsymbol{a}|\boldsymbol{s})} \right)_{\pi^\star} = -\left(\frac{\partial \mathcal{L}_{\text{VDB}}}{\partial \pi(\boldsymbol{a}|\boldsymbol{s})} \right)_{\pi^\star,d^\star,q^\star} = d^{\pi^\star}(\boldsymbol{s}) q_{\text{VDB}}^{\pi^\star}(\boldsymbol{s},\boldsymbol{a})$. From the optimality conditions and Proposition 6, we then have $\lambda^\star(\boldsymbol{s}) \geq d^{\pi^\star}(\boldsymbol{s}) q_{\text{VDB}}^{\pi^\star}(\boldsymbol{s},\boldsymbol{a})$ for all $\boldsymbol{s}$ and $\boldsymbol{a}$, and the equality holds for all actions taken by the policy $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} q_{\text{VDB}}^{\pi^\star}(\boldsymbol{s},\boldsymbol{a})$. Hence, $\pi^\star$ is a greedy policy with respect to its own adjusted value function $q_{\text{VDB}}^{\pi^\star}$. This means $v_{\text{VDB}}^{\pi^\star}$ and $q_{\text{VDB}}^{\pi^\star}$ satisfy the Bellman optimality equation and are equivalent to the optimal adjusted value functions v_{VDB}^* and q_{VDB}^*. Combining the results from Lemma 9 and Lemma 11, we can then conclude that $\pi^\star$ is an optimal policy for the considered MDP with visitation density bounds. ∎

A.7 Action Density Bounds

The altered dual value learning problem with action density bounds can be written as

$$
\begin{aligned}
\max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \; & \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}} \big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big] \\
\text{s.t.} \quad & \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s}) && \forall \boldsymbol{s} \\
& d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)} \big[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\big] && \forall \boldsymbol{s} \\
& p(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\, \forall \boldsymbol{a} \\
& d(\boldsymbol{s})\pi_{\text{L}}(\boldsymbol{a}|\boldsymbol{s}) \leq p(\boldsymbol{s},\boldsymbol{a}) \leq d(\boldsymbol{s})\pi_{\text{U}}(\boldsymbol{a}|\boldsymbol{s}) && \forall \boldsymbol{s}\, \forall \boldsymbol{a}
\end{aligned}
$$

The Lagrangian of this altered dual problem is given by

$$
\begin{aligned}
&\mathcal{L}_{ADB}(d(\boldsymbol{s}), p(\boldsymbol{s},\boldsymbol{a}), \lambda(\boldsymbol{s}), v(\boldsymbol{s}), \mu(\boldsymbol{s},\boldsymbol{a}), r_{\text{L}}(\boldsymbol{s},\boldsymbol{a}), r_{\text{U}}(\boldsymbol{s},\boldsymbol{a})) \\
&\quad = -\sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} \lambda(\boldsymbol{s}) \left[\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) - d(\boldsymbol{s})\right] \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} v(\boldsymbol{s}) \left[d(\boldsymbol{s}) - (1-\gamma)\iota(\boldsymbol{s}) - \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\right] \\
&\qquad - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} \mu(\boldsymbol{s},\boldsymbol{a}) p(\boldsymbol{s},\boldsymbol{a}) \\
&\qquad - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} r_{\text{L}}(\boldsymbol{s},\boldsymbol{a})[p(\boldsymbol{s},\boldsymbol{a}) - d(\boldsymbol{s})\pi_{\text{L}}(\boldsymbol{a}|\boldsymbol{s})] \\
&\qquad - \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} r_{\text{U}}(\boldsymbol{s},\boldsymbol{a})[d(\boldsymbol{s})\pi_{\text{U}}(\boldsymbol{a}|\boldsymbol{s}) - p(\boldsymbol{s},\boldsymbol{a})] \\
&\quad = -\sum_{\boldsymbol{s}\in\mathcal{S}} \iota(\boldsymbol{s})(1-\gamma)v(\boldsymbol{s}) \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} d(\boldsymbol{s}) \left[v(\boldsymbol{s}) - \lambda(\boldsymbol{s}) + \sum_{\boldsymbol{a}\in\mathcal{A}} \big(r_{\text{L}}(\boldsymbol{s},\boldsymbol{a})\pi_{\text{L}}(\boldsymbol{a}|\boldsymbol{s}) - r_{\text{U}}(\boldsymbol{s},\boldsymbol{a})\pi_{\text{U}}(\boldsymbol{a}|\boldsymbol{s})\big)\right] \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \left[\lambda(\boldsymbol{s}) - \mu(\boldsymbol{s},\boldsymbol{a}) - r_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) + r_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) - \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}')\big)\right],
\end{aligned}
$$

resulting in following optimality (KKT) conditions

- Stationarity of respectively d and p

$$
\begin{aligned}
\lambda^\star(\boldsymbol{s}) &= v^\star(\boldsymbol{s}) + \sum_{\boldsymbol{a}\in\mathcal{A}} \big(r^\star_{\text{L}}(\boldsymbol{s},\boldsymbol{a})\pi_{\text{L}}(\boldsymbol{a}|\boldsymbol{s}) - r^\star_{\text{U}}(\boldsymbol{s},\boldsymbol{a})\pi_{\text{U}}(\boldsymbol{a}|\boldsymbol{s})\big) && \forall \boldsymbol{s} \\
\mu^\star(\boldsymbol{s},\boldsymbol{a}) &= v^\star(\boldsymbol{s}) + \sum_{\tilde{\boldsymbol{a}}\in\mathcal{A}} \big(r^\star_{\text{L}}(\boldsymbol{s},\tilde{\boldsymbol{a}})\pi_{\text{L}}(\tilde{\boldsymbol{a}}|\boldsymbol{s}) - r^\star_{\text{U}}(\boldsymbol{s},\tilde{\boldsymbol{a}})\pi_{\text{U}}(\tilde{\boldsymbol{a}}|\boldsymbol{s})\big) \\
&\quad - r^\star_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) + r^\star_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) \\
&\quad - \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big) && \forall \boldsymbol{s}\, \forall \boldsymbol{a}
\end{aligned}
$$

- Dual feasibility

$$
\begin{aligned}
&\sum_{\boldsymbol{a}\in\mathcal{A}} p^\star(\boldsymbol{s},\boldsymbol{a}) = d^\star(\boldsymbol{s}) && \forall \boldsymbol{s} \\
&\quad d^\star(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p^\star(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}') && \forall \boldsymbol{s} \\
&\quad p^\star(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\, \forall \boldsymbol{a} \\
&d(\boldsymbol{s})\pi_{\text{L}}(\boldsymbol{a}|\boldsymbol{s}) \leq p^\star(\boldsymbol{s},\boldsymbol{a}) \leq d(\boldsymbol{s})\pi_{\text{U}}(\boldsymbol{a}|\boldsymbol{s}) && \forall \boldsymbol{s}
\end{aligned}
$$

- Primal feasibility, introducing auxiliary q

$$
\begin{aligned}
&v^\star(\boldsymbol{s}) \geq q^\star(\boldsymbol{s},\boldsymbol{a}) \\
&q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \Big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + r^\star_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) - r^\star_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) \\
&\qquad\qquad - \sum_{\tilde{\boldsymbol{a}}\in\mathcal{A}} r^\star_{\text{L}}(\boldsymbol{s},\tilde{\boldsymbol{a}})\pi_{\text{L}}(\tilde{\boldsymbol{a}}|\boldsymbol{s}) \\
&\qquad\qquad + \sum_{\tilde{\boldsymbol{a}}\in\mathcal{A}} r^\star_{\text{U}}(\boldsymbol{s},\tilde{\boldsymbol{a}})\pi_{\text{U}}(\tilde{\boldsymbol{a}}|\boldsymbol{s}) \\
&\qquad\qquad + \gamma v^\star(\boldsymbol{s}')\Big] && \forall \boldsymbol{s}\, \forall \boldsymbol{a} \\
&r^\star_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) \geq 0 \wedge r^\star_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\, \forall \boldsymbol{a}
\end{aligned}
$$

- Complementary slackness

$$
\begin{aligned}
v^\star(\boldsymbol{s}) = q^\star(\boldsymbol{s},\boldsymbol{a}) \vee p^\star(\boldsymbol{s},\boldsymbol{a}) = 0 \qquad & \forall \boldsymbol{s}\, \forall \boldsymbol{a} \\
p^\star(\boldsymbol{s},\boldsymbol{a}) = d^\star(\boldsymbol{s})\pi_{\text{L}}(\boldsymbol{a}|\boldsymbol{s}) \vee r^\star_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) = 0 \; & \forall \boldsymbol{s}\, \forall \boldsymbol{a} \\
p^\star(\boldsymbol{s},\boldsymbol{a}) = d^\star(\boldsymbol{s})\pi_{\text{U}}(\boldsymbol{a}|\boldsymbol{s}) \vee r^\star_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) = 0 \; & \forall \boldsymbol{s}\, \forall \boldsymbol{a}
\end{aligned}
$$

The primal formulation is obtained as

$$
\begin{aligned}
\min_{\substack{v(\boldsymbol{s}),q(\boldsymbol{s},\boldsymbol{a})\\ r_{\text{L}}(\boldsymbol{s},\boldsymbol{a}),r_{\text{U}}(\boldsymbol{s},\boldsymbol{a})}} \; & \mathop{\mathbb{E}}_{\boldsymbol{s}_0\sim\iota(\cdot)} \big[(1-\gamma)v(\boldsymbol{s}_0)\big] \\
\text{s.t.} \quad & v(\boldsymbol{s}) \geq q(\boldsymbol{s},\boldsymbol{a}) && \forall \boldsymbol{s}\, \forall \boldsymbol{a} \\
& q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})} \Big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + r_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) - \sum_{\tilde{\boldsymbol{a}}\in\mathcal{A}} r_{\text{L}}(\boldsymbol{s},\tilde{\boldsymbol{a}})\pi_{\text{L}}(\tilde{\boldsymbol{a}}|\boldsymbol{s}) \\
&\qquad\qquad - r_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) + \sum_{\tilde{\boldsymbol{a}}\in\mathcal{A}} r_{\text{U}}(\boldsymbol{s},\tilde{\boldsymbol{a}})\pi_{\text{U}}(\tilde{\boldsymbol{a}}|\boldsymbol{s}) \\
&\qquad\qquad + \gamma v(\boldsymbol{s}')\Big] && \forall \boldsymbol{s}\, \forall \boldsymbol{a} \\
& r_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) \geq 0 \wedge r_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\, \forall \boldsymbol{a}
\end{aligned}
$$

From the optimality conditions, it is immediately clear Proposition 5 still holds for this constrained variant and it can be extended as follows.

Proposition 9. *The policy $\pi(\boldsymbol{a}|\boldsymbol{s})$ that can be derived from any feasible dual variables d and p is a feasible policy for the considered MDP with action density bounds.*

Proof. From Proposition 5, we know that the dual variables correspond to the visitation densities of a policy π, i.e. $d(\boldsymbol{s}) = d^\pi(\boldsymbol{s})$ and $p(\boldsymbol{s},\boldsymbol{a}) = d^\pi(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s})$. The additional constraints can thus be rewritten as $d^\pi(\boldsymbol{s})\pi_{\text{L}}(\boldsymbol{a}|\boldsymbol{s}) \leq d^\pi(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) \leq d^\pi(\boldsymbol{s})\pi_{\text{U}}(\boldsymbol{a}|\boldsymbol{s})$, implying $\pi_{\text{L}}(\boldsymbol{a}|\boldsymbol{s}) \leq \pi(\boldsymbol{a}|\boldsymbol{s}) \leq \pi_{\text{U}}(\boldsymbol{a}|\boldsymbol{s})$ for all visited states. The dual variables d and p are thus the visitation densities of a policy satisfying those action density bounds. ■

Consequently, Theorem 1 can also be straightforwardly adapted as follows.

Theorem 7. *The policy $\pi^\star$ that can be derived from the optimal dual variables $d^\star$ and $p^\star$ is an optimal policy for the considered MDP with action density bounds.*

Proof. Combine the results from the previous Proposition with the proof of Theorem 1. ■

Lemma 1 also holds, albeit for the modified reward function r_{ADB} (7), with learnable parameters r_{L} and r_{U}, and corresponding value functions.

Lemma 7. *The optimal primal variables $v^\star$ and $q^\star$ satisfy the Bellman optimality equation and are thus equivalent to the optimal adjusted value functions v^*_{VDB} and q^*_{VDB} for the considered MDP with optimal modified reward $r^\star_{\text{VDB}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + r^\star_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) - \sum_{\tilde{\boldsymbol{a}}\in\mathcal{A}} r_{\text{L}}(\boldsymbol{s},\tilde{\boldsymbol{a}})\pi_{\text{L}}(\tilde{\boldsymbol{a}}|\boldsymbol{s}) - r^\star_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) + \sum_{\tilde{\boldsymbol{a}}\in\mathcal{A}} r_{\text{U}}(\boldsymbol{s},\tilde{\boldsymbol{a}})\pi_{\text{U}}(\tilde{\boldsymbol{a}}|\boldsymbol{s})$.*

Proof. Following the same arguments as in Lemma 1, we obtain $v^\star(\boldsymbol{s}) = \max_{\boldsymbol{a}} \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\text{ADB}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$. Satisfying the Bellman optimality equation, $v^\star$ is thus equivalent to the optimal value function v^*_{ADB} for the considered MDP with altered reward $r^\star_{\text{ADB}}$. From the primal feasibility conditions $q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\text{ADB}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$, we then also obtain the equivalency between $q^\star$ and q^*_{ADB}. ■

Finally, these optimal adjusted value functions and optimal reward modifications satisfy following properties.

Theorem 8. *The optimal policy $\pi^\star$ is greedy with respect to the optimal adjusted value functions v^*_{ADB} and q^*_{ADB}.*

Proof. Following the same arguments as in the proof for Lemma 1, we obtain $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \mathcal{A}^*(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\text{ADB}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$. From the previous Lemma, it then follows that $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} q^*_{\text{ADB}}(\boldsymbol{s},\boldsymbol{a})$, showing the greediness of $\pi^\star$ with respect to the optimal adjusted value functions. ■

Proposition 3. *The optimal modified reward $r^\star_{\text{ADB}}$ is only altered by the tight bounds $\pi_{\text{L}}(\boldsymbol{a}|\boldsymbol{s}) = \pi^\star(\boldsymbol{a}|\boldsymbol{s})$ and $\pi^\star(\boldsymbol{a}|\boldsymbol{s}) = \pi_{\text{U}}(\boldsymbol{a}|\boldsymbol{s})$.*

Proof. This follows from the complementary slackness conditions, as $r^\star_{\text{L}}(\boldsymbol{s},\boldsymbol{a}) = 0$ for the strictly satisfied constraints $d^{\pi^\star}(\boldsymbol{s})\pi_{\text{L}}(\boldsymbol{a}|\boldsymbol{s}) < p^{\pi^\star}(\boldsymbol{s},\boldsymbol{a})$, while $r^\star_{\text{U}}(\boldsymbol{s},\boldsymbol{a}) = 0$ for the strictly satisfied constraints $p^{\pi^\star}(\boldsymbol{s},\boldsymbol{a}) < d^{\pi^\star}(\boldsymbol{s})\pi_{\text{U}}(\boldsymbol{a}|\boldsymbol{s})$. ■

A.8 Transition Constraints

A.8.1 Value Learning

The altered dual value learning problem with an average transition constraint is denoted as

$$
\begin{aligned}
\max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}} \big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big] \\
\text{s.t.} \quad & \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s}) && \forall \boldsymbol{s} \\
& d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)} \big[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\big] && \forall \boldsymbol{s} \\
& p(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
& 0 \geq p(\boldsymbol{s},\boldsymbol{a}) \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})} \big[c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big] && \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned}
$$

The Lagrangian of this altered dual problem is then obtained as

$$
\begin{aligned}
&\mathcal{L}_{ATC}(d(\boldsymbol{s}), p(\boldsymbol{s},\boldsymbol{a}), \lambda(\boldsymbol{s}), v(\boldsymbol{s}), \mu(\boldsymbol{s},\boldsymbol{a}), r_{\text{C}}(\boldsymbol{s},\boldsymbol{a})) \\
&\quad = -\sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} \lambda(\boldsymbol{s}) \left[\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) - d(\boldsymbol{s})\right] \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} v(\boldsymbol{s}) \left[d(\boldsymbol{s}) - (1-\gamma)\iota(\boldsymbol{s}) - \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}} \sum_{\boldsymbol{a}'\in\mathcal{A}} p(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\right] \\
&\qquad - \sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{a}\in\mathcal{A}} \mu(\boldsymbol{s},\boldsymbol{a}) p(\boldsymbol{s},\boldsymbol{a}) \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{a}\in\mathcal{A}} r_{\text{C}}(\boldsymbol{s},\boldsymbol{a}) p(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') \\
&\quad = -\sum_{\boldsymbol{s}\in\mathcal{S}} \iota(\boldsymbol{s})(1-\gamma) v(\boldsymbol{s}) \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} d(\boldsymbol{s})\big[v(\boldsymbol{s}) - \lambda(\boldsymbol{s})\big] \\
&\qquad + \sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \Bigg[\lambda(\boldsymbol{s}) - \mu(\boldsymbol{s},\boldsymbol{a}) \\
&\qquad\qquad - \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}') - r_{\text{C}}(\boldsymbol{s},\boldsymbol{a}) c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big)\Bigg],
\end{aligned}
$$

leading to following optimality (KKT) conditions

- Stationarity of respectively d and p

$$
\begin{aligned}
\lambda^\star(\boldsymbol{s}) &= v^\star(\boldsymbol{s}) && \forall \boldsymbol{s} \\
\mu^\star(\boldsymbol{s},\boldsymbol{a}) &= v^\star(\boldsymbol{s}) - \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}') - r^\star_{\text{C}}(\boldsymbol{s},\boldsymbol{a}) c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big) && \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned}
$$

- Dual feasibility

$$
\begin{aligned}
\sum_{\boldsymbol{a}\in\mathcal{A}} p^\star(\boldsymbol{s},\boldsymbol{a}) &= d^\star(\boldsymbol{s}) && \forall \boldsymbol{s} \\
d^\star(\boldsymbol{s}) &= (1-\gamma)\iota(\boldsymbol{s}) + \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}} \sum_{\boldsymbol{a}'\in\mathcal{A}} p^\star(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}') && \forall \boldsymbol{s} \\
p^\star(\boldsymbol{s},\boldsymbol{a}) &\geq 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
0 &\geq p^\star(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') && \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned}
$$

- Primal feasibility, introducing auxiliary q

$$
\begin{aligned}
v^\star(\boldsymbol{s}) &\geq q^\star(\boldsymbol{s},\boldsymbol{a}) && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
q^\star(\boldsymbol{s},\boldsymbol{a}) &= \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') - r^\star_{\text{C}}(\boldsymbol{s},\boldsymbol{a}) c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big) && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
r^\star_{\text{C}}(\boldsymbol{s},\boldsymbol{a}) &\geq 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned}
$$

- Complementary slackness

$$
\begin{aligned}
v^\star(\boldsymbol{s}) = q^\star(\boldsymbol{s},\boldsymbol{a}) \vee p^\star(\boldsymbol{s},\boldsymbol{a}) = 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
p^\star(\boldsymbol{s},\boldsymbol{a}) = 0 \vee \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = 0 \vee r^\star_{\text{C}}(\boldsymbol{s},\boldsymbol{a}) = 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned}
$$

The corresponding primal formulation is given by

$$
\begin{aligned}
\min_{\substack{v(\boldsymbol{s}), q(\boldsymbol{s},\boldsymbol{a})\\ r_{\text{C}}(\boldsymbol{s},\boldsymbol{a})}} \quad & \mathop{\mathbb{E}}_{\boldsymbol{s}_0\sim\iota(\cdot)} \big[(1-\gamma) v(\boldsymbol{s}_0)\big] \\
\text{s.t.} \quad & v(\boldsymbol{s}) \geq q(\boldsymbol{s},\boldsymbol{a}) && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
& q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})} \big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') - r_{\text{C}}(\boldsymbol{s},\boldsymbol{a}) c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}')\big] && \forall \boldsymbol{s}\,\forall \boldsymbol{a} \\
& r_{\text{C}}(\boldsymbol{s},\boldsymbol{a}) \geq 0 && \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned}
$$

From the optimality conditions, it is immediately clear Proposition 5 still holds for this constrained variant and it can be extended as follows.

Proposition 10. *The policy $\pi(\boldsymbol{a}|\boldsymbol{s})$ that can be derived from any feasible dual variables d and p is a feasible policy for the considered MDP with average transition constraints.*

Proof. From Proposition 5, we know that the dual variables correspond to the visitation densities of a policy π with $p(\boldsymbol{s},\boldsymbol{a}) = d^\pi(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s})$. The additional constraints can thus be rewritten as $d^\pi(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s})\,\mathbb{E}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}[c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')] \leq 0$, implying $\bar{c}(\boldsymbol{s},\boldsymbol{a}) \leq 0$ for all visited states and actions. The dual variables d and p are thus the visitation densities of a policy satisfying those average transition constraints. ■

Consequently, Theorem 1 can also be straightforwardly adapted as follows.

Theorem 9. *The policy $\pi^\star$ that can be derived from the optimal dual variables $d^\star$ and $p^\star$ is an optimal policy for the considered MDP with average transition constraints.*

Proof. Combine the results from the previous Proposition with the proof of Theorem 1. ■

Lemma 1 also holds, albeit for the modified reward function r_{ATC} (8), with learnable parameters r_{C}, and corresponding value functions.

Lemma 8. *The optimal primal variables $v^\star$ and $q^\star$ satisfy the Bellman optimality equation and are thus equivalent to the optimal adjusted value functions v^*_{ATC} and q^*_{ATC} for the considered MDP with optimal modified reward $r^\star_{\text{ATC}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') - r^\star_{\text{C}}(\boldsymbol{s},\boldsymbol{a})c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')$.*

Proof. Following the same arguments as in Lemma 1, we obtain $v^\star(\boldsymbol{s}) = \max_{\boldsymbol{a}} \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\text{ATC}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$. Satisfying the Bellman optimality equation, $v^\star$ is thus equivalent to the optimal value function v^*_{ATC} for the considered MDP with altered reward $r^\star_{\text{ATC}}$. From the primal feasibility conditions $q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\text{ATC}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$, we then also obtain the equivalency between $q^\star$ and q^*_{ATC}. ■

Finally, these optimal adjusted value functions and optimal reward modifications satisfy following properties.

Theorem 10. *The optimal policy $\pi^\star$ is greedy with respect to the optimal adjusted value functions v^*_{ATC} and q^*_{ATC}.*

Proof. Following the same arguments as in the proof for Lemma 1, we obtain $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \mathcal{A}^*(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\text{ATC}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$. From the previous Lemma, it then follows that $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} q^*_{\text{ATC}}(\boldsymbol{s},\boldsymbol{a})$, showing the greediness of $\pi^\star$ with respect to the optimal adjusted value functions. ■

Proposition 4. *The optimal modified reward $r^\star_{\text{ATC}}$ is only altered by nonvisited state-action pairs or tight bounds $\bar{c}(\boldsymbol{s},\boldsymbol{a}) = 0$.*

Proof. This follows from the complementary slackness conditions, as $r^\star_{\text{C}}(\boldsymbol{s},\boldsymbol{a}) = 0$ for the strictly satisfied constraints $\mathbb{E}_{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}[c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')] < 0$ and $p^\star(\boldsymbol{s},\boldsymbol{a}) > 0$. ■

The derived properties are especially useful to obtain similar Constrained GPI schemes for solving RL problems with transition constraints, as shown in the next part.

A.8.2 Policy Iteration

The altered dual policy evaluation problem with immediate state-action transition constraints can be written as

$$\max_{\substack{d(\boldsymbol{s})\\ p(\boldsymbol{s},\boldsymbol{a})}} \mathop{\mathbb{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a})\sim p(\cdot,\cdot)\\ \boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})}} \big[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\big] \tag{ITC·D}$$

$$\begin{aligned}
\text{s.t.}\quad & d(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s}) + \gamma \mathop{\mathbb{E}}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p(\cdot,\cdot)} \big[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')\big] \geq 0 \;\forall \boldsymbol{s}\\
& p(\boldsymbol{s},\boldsymbol{a}) = d(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) && \forall \boldsymbol{s}\,\forall \boldsymbol{a}\\
& 0 \geq p(\boldsymbol{s},\boldsymbol{a})\tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\pi(\boldsymbol{a}'|\boldsymbol{s}')c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') \;\forall \boldsymbol{s}\,\forall \boldsymbol{a}\,\forall \boldsymbol{s}'\,\forall \boldsymbol{a}'
\end{aligned}$$

and has as Lagrangian

$$\begin{aligned}
&\mathcal{L}_{\text{ITC}}(d(\boldsymbol{s}), p(\boldsymbol{s},\boldsymbol{a}), v(\boldsymbol{s}), q(\boldsymbol{s},\boldsymbol{a}), \mu(\boldsymbol{s}), r_{\text{C}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}'))\\
&= -\sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\\
&\quad + \sum_{\boldsymbol{s}\in\mathcal{S}} v(\boldsymbol{s}) \left[d(\boldsymbol{s}) - (1-\gamma)\iota(\boldsymbol{s}) - \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}') \right]\\
&\quad + \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} q(\boldsymbol{s},\boldsymbol{a}) \big[p(\boldsymbol{s},\boldsymbol{a}) - d(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) \big]\\
&\quad - \sum_{\boldsymbol{s}\in\mathcal{S}} \mu(\boldsymbol{s}) d(\boldsymbol{s})\\
&\quad + \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}}\sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} \big[r_{\text{C}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') p(\boldsymbol{s},\boldsymbol{a})\tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\pi(\boldsymbol{a}'|\boldsymbol{s}')\\
&\qquad\qquad c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') \big]\\
&= -\sum_{\boldsymbol{s}\in\mathcal{S}} \iota(\boldsymbol{s})(1-\gamma) v(\boldsymbol{s})\\
&\quad + \sum_{\boldsymbol{s}\in\mathcal{S}} d(\boldsymbol{s}) \left[v(\boldsymbol{s}) - \mu(\boldsymbol{s}) - \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) q(\boldsymbol{s},\boldsymbol{a}) \right]\\
&\quad + \sum_{\boldsymbol{s}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{s},\boldsymbol{a}) \left[q(\boldsymbol{s},\boldsymbol{a}) - \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \Big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v(\boldsymbol{s}')\right.\right.\\
&\qquad\qquad \left.\left. - \sum_{\boldsymbol{a}'\in\mathcal{A}} \pi(\boldsymbol{a}'|\boldsymbol{s}') r_{\text{C}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') \Big) \right],
\end{aligned}$$

The optimality (KKT) conditions can be derived as

- Stationarity of respectively d and p

$$\begin{aligned}
&\mu^\star(\boldsymbol{s}) = v^\star(\boldsymbol{s}) - \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) q^\star(\boldsymbol{s},\boldsymbol{a}) && \forall \boldsymbol{s}\\
&q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}} \tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) \Big(r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\\
&\qquad\qquad - \sum_{\boldsymbol{a}'\in\mathcal{A}} \pi(\boldsymbol{a}'|\boldsymbol{s}') r^\star_{\text{C}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') \Big) && \forall \boldsymbol{s}\,\forall \boldsymbol{a}
\end{aligned}$$

- Dual feasibility

$$\begin{aligned}
&d^\star(\boldsymbol{s}) = (1-\gamma)\iota(\boldsymbol{s})\\
&\qquad + \gamma \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p^\star(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}') \geq 0 \;\forall \boldsymbol{s}\\
&p^\star(\boldsymbol{s},\boldsymbol{a}) = d^\star(\boldsymbol{s})\pi(\boldsymbol{a}|\boldsymbol{s}) && \forall \boldsymbol{s}\,\forall \boldsymbol{a}\\
&0 \geq p^\star(\boldsymbol{s},\boldsymbol{a})\tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\pi(\boldsymbol{a}'|\boldsymbol{s}')c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') && \forall \boldsymbol{s}\,\forall \boldsymbol{a}\,\forall \boldsymbol{s}'\,\forall \boldsymbol{a}'
\end{aligned}$$

- Primal feasibility

$$\begin{aligned}
&v^\star(\boldsymbol{s}) \geq \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) q^\star(\boldsymbol{s},\boldsymbol{a}) \;\forall \boldsymbol{s}\\
&0 \leq r^\star_{\text{C}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') && \forall \boldsymbol{s}\,\forall \boldsymbol{a}\,\forall \boldsymbol{s}'\,\forall \boldsymbol{a}'
\end{aligned}$$

- Complementary slackness

$$\begin{aligned}
&v^\star(\boldsymbol{s}) = \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}|\boldsymbol{s}) q^\star(\boldsymbol{s},\boldsymbol{a}) \vee d^\star(\boldsymbol{s}) = 0 && \forall \boldsymbol{s}\\
&p^\star(\boldsymbol{s},\boldsymbol{a})\tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\pi(\boldsymbol{a}'|\boldsymbol{s}') = 0 \vee c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') = 0\\
&\qquad\qquad \vee r^\star_{\text{C}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') = 0 \;\forall \boldsymbol{s}\,\forall \boldsymbol{a}\,\forall \boldsymbol{s}'\,\forall \boldsymbol{a}'
\end{aligned}$$

The corresponding primal LP is given by

$$\begin{aligned}
\min_{\substack{v(\boldsymbol{s}),q(\boldsymbol{s},\boldsymbol{a})\\ r_{\mathrm{C}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}')}} \quad & \mathop{\mathbb{E}}_{\boldsymbol{s}_0\sim\iota(\cdot)}\left[(1-\gamma)v(\boldsymbol{s}_0)\right] && \text{(ITC·P)}\\
\text{s.t.} \quad & v(\boldsymbol{s}) \geq \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{s})}\left[q(\boldsymbol{s},\boldsymbol{a})\right] && \forall \boldsymbol{s}\\
& q(\boldsymbol{s},\boldsymbol{a}) = \mathop{\mathbb{E}}_{\substack{\boldsymbol{s}'\sim\tau(\cdot|\boldsymbol{s},\boldsymbol{a})\\ \boldsymbol{a}'\sim\pi(\cdot|\boldsymbol{s}')}}\left[r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') - r_{\mathrm{C}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}')c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') + \gamma v(\boldsymbol{s}')\right] && \forall \boldsymbol{s}\,\forall \boldsymbol{a}\\
& 0 \leq r_{\mathrm{C}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') && \forall \boldsymbol{s}\forall \boldsymbol{a}\forall \boldsymbol{s}'\,\forall \boldsymbol{a}'
\end{aligned}$$

For feasible policies, satisfying all imposed transition constraints, Lemma 2 can be proven without major changes. Lemma 3 also holds, albeit for the modified reward function r_{ITC}, with learnable parameters r_{C}, and corresponding value functions.

Lemma 13. *The optimal primal variables $v^\star$ and $q^\star$ satisfy the Bellman equation and are equivalent to the adjusted value functions of the given policy π for the considered MDP with optimal modified reward $r^\star_{\mathrm{ITC}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') = r(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') - \mathbb{E}_{\boldsymbol{a}'\sim\pi(\cdot|\boldsymbol{s}')}[r^\star_{\mathrm{C}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}')c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}')]$.*

Proof. Following similar arguments as in Lemma 3, we obtain $v^\star(\boldsymbol{s}) = \sum_{\boldsymbol{a}\in\mathcal{A}}\pi(\boldsymbol{a}|\boldsymbol{s})\sum_{\boldsymbol{s}'\in\mathcal{S}}\tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\mathrm{ITC}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$ and $q^\star(\boldsymbol{s},\boldsymbol{a}) = \sum_{\boldsymbol{s}'\in\mathcal{S}}\tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a})\big(r^\star_{\mathrm{ITC}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}') + \gamma v^\star(\boldsymbol{s}')\big)$. From the former equality it is clear that $v^\star$ satisfies the Bellman equation and is thus equivalent to the adjusted value function v^π_{ITC} of the given policy for the considered MDP with altered reward $r^\star_{\mathrm{ITC}}$. From the latter equality we can then obtain the equivalency between $q^\star$ and q^π_{ITC}. ■

Furthermore, these optimal reward modifications satisfy following property.

Lemma 14. *The optimal modified reward $r^\star_{\mathrm{ITC}}$ is only altered by nonvisited transitions or transitions with tight bounds $c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') = 0$.*

Proof. This follows from the complementary slackness conditions, as $r^\star_{\mathrm{C}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') = 0$ for the visited transitions with negative cost, i.e. $p^\star(\boldsymbol{s},\boldsymbol{a}) > 0$ and $\tau(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}) > 0$ and $\pi(\boldsymbol{a}'|\boldsymbol{s}') > 0$ and $c(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}') < 0$. ■

On the other hand, if the policy does not satisfy all additional constraints, the dual problem effectively becomes infeasible and the primal becomes unbounded, caused by the $r_{\mathrm{C}}(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}',\boldsymbol{a}')$ of the infeasible constraints approaching infinity. As before, this altered Policy Evaluation LP is plugged in the Policy Optimization problem to find optimal constrained policies, using following property.

Lemma 15. *The optimal solution $\pi^\star$ of the nested Policy Optimization (PO) and constrained Policy Evaluation (ITC) problems, is the optimal policy π^* for the considered MDP with transition constraints.*

Proof. The partial derivative of the average reward $\bar{r}^\pi$ can be evaluated using Danskin's theorem as $\left(\frac{\partial \bar{r}^\pi}{\partial \pi(\boldsymbol{a}|\boldsymbol{s})}\right)_{\pi^\star} = -\left(\frac{\partial \mathcal{L}_{\mathrm{ITC}}}{\partial \pi(\boldsymbol{a}|\boldsymbol{s})}\right)_{\pi^\star,d^\star,p^\star,q^\star,r^\star_{\mathrm{C}}} = d^{\pi^\star}(\boldsymbol{s})q^{\pi^\star}_{\mathrm{ITC}}(\boldsymbol{s},\boldsymbol{a}) - \sum_{\boldsymbol{s}'\in\mathcal{S}}\sum_{\boldsymbol{a}'\in\mathcal{A}} p^{\pi^\star}(\boldsymbol{s}',\boldsymbol{a}')\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')r^\star_{\mathrm{C}}(\boldsymbol{s}',\boldsymbol{a}',\boldsymbol{s},\boldsymbol{a})c(\boldsymbol{s}',\boldsymbol{a}',\boldsymbol{s},\boldsymbol{a})$.

Notice that the second term evaluates to zero for all actions taken by the policy $\boldsymbol{a}\in\mathcal{A}^{\pi^\star}(\boldsymbol{s})$, because of the complementary slackness conditions. From the optimality conditions and Proposition 6, we then have $\lambda^\star(\boldsymbol{s}) \geq d^{\pi^\star}(\boldsymbol{s})q^{\pi^\star}_{\mathrm{ITC}}(\boldsymbol{s},\boldsymbol{a}) - \mathbb{E}_{(\boldsymbol{s}',\boldsymbol{a}')\sim p^{\pi^\star}(\cdot,\cdot)}[\tau(\boldsymbol{s}|\boldsymbol{s}',\boldsymbol{a}')r^\star_{\mathrm{C}}(\boldsymbol{s}',\boldsymbol{a}',\boldsymbol{s},\boldsymbol{a})c(\boldsymbol{s}',\boldsymbol{a}',\boldsymbol{s},\boldsymbol{a})]$ for all $\boldsymbol{s}$ and $\boldsymbol{a}$, and the equality holds for all actions taken by the policy $\mathcal{A}^{\pi^\star}(\boldsymbol{s}) = \arg\max_{\boldsymbol{a}} q^{\pi^\star}_{\mathrm{ITC}}(\boldsymbol{s},\boldsymbol{a})$. Hence, $\pi^\star$ is a greedy policy with respect to its own adjusted value function $q^{\pi^\star}_{\mathrm{ITC}}$. This means $v^{\pi^\star}_{\mathrm{ITC}}$ and $q^{\pi^\star}_{\mathrm{ITC}}$ satisfy the Bellman optimality equation and are equivalent to the optimal adjusted value functions v^*_{ITC} and q^*_{ITC}. Combining the results from Theorem 9 and Theorem 10 after modifying them for immediate transition constraints, we can then conclude that $\pi^\star$ is an optimal policy for the considered MDP with transition constraints. ■

The resulting nested optimization problem can be seen as a *Constrained* GPI scheme, in which the policy, adjusted value functions and reward modifications are jointly learned. Note that for an infeasible policy, violating the transition constraints, the inner Policy Evaluation problem diverges. However, the directions in which the learnable reward parameters r_{C} are updated, remain useful for the combined policy iteration scheme: the reward is effectively decreased for transitions with positive costs. As a result, during subsequent policy optimization steps, there is an increased incentive for the policy to stay away from such transitions. The reward is thus modified in such a way that constraint-violating policies are no longer optimal or greedy with respect to the adjusted value functions.

B Hyperparameters

Table 2 below shows the used hyperparameters for the `DualCRL` algorithm on the two used gymnasium environments.

Table 2: Overview of the used hyperparameters for both environments.

Hyperparameters		**CliffWalking**	**Pendulum**
Total training timesteps	k_M	50 000	50 000
Warmup timesteps		640	1000
Replay buffer size	$\lvert\mathcal{B}\rvert$	10 000	50 000
Batch size	B	32	64
Discount factor	γ	0.99	0.99
Learning rates (strides)	η_q	$2\cdot10^{-2}$ (1)	$4\cdot10^{-4}$ (1)
	η_π	$4\cdot10^{-3}$ (2)	$4\cdot10^{-4}$ (2)
	η_d	$4\cdot10^{-3}$ (1)	-
	η_r	$1\cdot10^{-2}$ (10)	$4\cdot10^{-4}$ (10)
Target networks averaging constant (stride)	τ	$1\cdot10^{-3}$ (2)	$1\cdot10^{-3}$ (2)
Network architecture – hidden dimensions		-	50×10
Entropy temperature	α	$1\rightarrow10^{-2}$	$10^{-1}\rightarrow10^{-3}$