Notes on Sample Complexity of IRL on Constrained MDPs and Bandit problems setting

Titouan Renard

June 7, 2023

Contents

1 Constrained and Unconstrained MDPs

1.1 Unconstrained Markov Decision Processes

Definition 1.1. *(**Markov Decision Process**) We define a **Markov Decision Process** (MDP) as the following 6-tuple:*

$$M = (S, A, P, \boldsymbol{\nu}, \boldsymbol{r}, \gamma),$$

*where $S = \{s_1, s_2, ..., s_n\}$ is a discrete set of **states** of size $|S| = n$, $A = \{a_1, a_2, ..., a_m\}$ a discrete set of **actions** of size $|A| = m$, is a probabilistic markovian transition law $P : S \times A \to \Delta_S$ that we can represent as a $\mathbb{R}^{nm \times n}$ matrix (where each column represents in the distribution $\in \Delta_S$), $\nu \in \Delta_s$ is the **initial distribution** of states at the start of the process, $r : S \times A \to \mathbb{R}$ is a reward function that we can conveniently represent as a vector $\boldsymbol{r} \in \mathcal{R} \subseteq \mathbb{R}^{nm}$ since S and A are discrete (we call $\mathcal{R}$ the **reward class**), finally $\gamma \in (0, 1)$ is our **discount factor**.*

Unless specified otherwise it is expected that Markov Decision Process run for an infinite time, each (infinite-time) sequence of move on the MDP is called a *trajectory* and is denoted:

$$\tau = \{s_1, a_1, s_2, a_2, ...\},$$

where the transition from one state to another is governed by the markovian transition law:

$$s_{t+1} \sim P(\cdot|s_t, a_t).$$

The MDP, doesn't specify how actions are selected, in practice, this is done by choosing a *policy function* $\pi : S \to \Delta_A$, which we use to pick the action a_t:

$$a_t \sim \pi(s_t) \in \Delta_A.$$

Definition 1.2. *(**Policy**) A **policy** is a function $\pi : S \to \Delta_A$ that maps each discrete state to a distribution over actions of an MDP M. Because we consider discrete state and action sets, we can represent policies as matrices $\pi \in \Pi \in \mathbb{R}^{n\times m}$. Where Π denotes the set of **valid policies**, i.e. :*

$$\Pi := \Big\{\pi = [\pi_{s_1}, \pi_{s_2}, ..., \pi_{s_n}], \pi_{s_i} \in \Delta_A \forall s_i \in S\Big\}.$$

Assuming actions are picked from some fixed policy function $\boldsymbol{\pi}\Pi$ the progression from state to state of the MDP becomes a Markov Chain described by the closed-loop transition law.

Definition 1.3. *(**Closed-Loop Transition Law**) We define the **closed-loop transition law** $P^\pi : S \to \Delta_S$ as a function that gives the distribution over the next state as a function of the current state, this distribution is trivially given by:*

$$P^\pi(s'|s) = \sum_a P(s'|a, s) \cdot \pi(a|s).$$

It can be represented as a square matrix $P^\pi \in \mathbb{R}^{n\times n}$, where each i-th column gives the distribution of states s' associated with the transition from the i-th state.

1.1.1 Discounted Reward, Value, Quality and Objective Functions

At each transition the agent receives reward $r_t = r(s_t, a_t)$. We want to quantify the general performance of our agents with respect to that reward function, across complete trajectories. This motivates the definition a series of functions which are used to quantify the performance of a given policy π on the MDP.

Definition 1.4. *(**Discounted Reward**) Given a policy π, and a convex **policy regularizer function** $\Omega : \Pi \to \mathbb{R}$, one can compute the discounted reward $R(\tau)$ as follows:*

$$R(\tau) = (1-\gamma) \lim_{T\to+\infty} \sum_{t=0}^{T} \gamma^t \Big(r(s_t, a_t) - \Omega(\pi(\cdot|s_t))\Big);\ s_t, a_t \in \tau.$$

Note that assuming that the regularized rewards $(r(s_t, a_t) - \Omega(\pi(\cdot|s_t)))$ are upper-bounded by some constant $C \in \mathbb{R}$, this limit is guaranteed to exist as:

$$R(\tau) = (1-\gamma) \lim_{T\to+\infty} \sum_{t=0}^{T} \gamma^t \Big(r(s_t, a_t) - \Omega(\pi(\cdot|s_t))\Big) \leq (1-\gamma) \lim_{T\to+\infty} \sum_{t=0}^{T} \gamma^t C = C.$$

Since the process is stochastic it makes more sense to think in terms of *expected discounted reward*, we define the *objective function* $J(\boldsymbol{\pi}, \boldsymbol{r})$, the *value function* $V_r^\pi(s)$ and the *quality function* $Q_r^\pi(s, a)$ (which are all associated with a policy π and a reward function r).

Definition 1.5. *(**Value Function**) For a given policy $\pi : S \to \Delta A$, a convex **policy regularizer function** $\Omega : \Pi \to \mathbb{R}$ and under the assumption that the agent starts at some state s, we define the **value function** $V_r^\pi : S \to \mathbb{R}$ as:*

$$\begin{aligned} V_r^\pi(s) &= \mathbb{E}\Big[R(\tau)\Big|s_0 = s, \pi\Big] \\ &= \mathbb{E}\left[(1-\gamma)\sum_{t=0}^{+\infty} \gamma^t \Big(r(s_t, a_t) - \Omega(\pi(\cdot|s_t))\Big)\Big|s_0 = s, \pi\right]. \end{aligned}$$

Definition 1.6. *(**Quality Function**) For a given policy $\pi : S \to \Delta A$, a convex **policy regularizer function** $\Omega : \Pi \to \mathbb{R}$ and under the assumption that the agent starts at some state s with some first action a, we define the **quality function** $Q_r^\pi : S \times A \to \mathbb{R}$ as:*

$$Q_r^\pi(s,a) = \mathbb{E}\Big[R(\tau)\Big|s_0 = s, a_0 = a, \pi\Big]$$

$$= \mathbb{E}\left[(1-\gamma)\sum_{t=0}^{+\infty}\gamma^t\Big(r(s_t,a_t) - \Omega(\pi(\cdot|s_t))\Big)\Big|s_0 = s, a_0 = a, \pi\right].$$

Definition 1.7. *(**Objective Function**) For a given policy $\pi : S \to \Delta A$, convex policy regularizer function $\Omega : \Pi \to \mathbb{R}$ and and under an initial state distribution $\boldsymbol{\nu} \in \Delta_S$ we define the **objective function** as:*

$$J(\pi, r) = \mathbb{E}\Big[R(\tau)\Big|s_0 \sim \nu, \pi\Big] = \mathbb{E}\Big[V^\pi(s)\Big|s \sim \nu\Big].$$

*Since we most often use the **objective function** for optimization in policy space, we consider that $J : \Pi \times \mathcal{R} \to \mathbb{R}$ is a function of the policy π and the reward r.*

1.1.2 The Occupancy Measure

We now discuss a descriptor of the distribution of an agent's position across all states induced by some policy π: the *occupancy measure* is defined as follows.

Definition 1.8. *(**Occupancy Measure**) The (state-action) **occupancy measure** $\boldsymbol{\mu}^\pi \in \Delta_{S\times A}$ is defined for a given state-action pair as follows:*

$$\mu^\pi(s,a) = (1-\gamma)\lim_{T\to+\infty}\sum_{t=0}^{T}\gamma^t\mathbb{P}_\nu^\pi(s_t = s, a_t = a)$$

$$= (1-\gamma)\mathbb{E}\Big[\sum_{t=0}^{+\infty}\gamma^t\mathbf{1}(s_t = s, a_t = a)\Big|s_0 \sim \nu, \pi\Big],$$

since our state and action spaces are discrete we can represent $\boldsymbol{\mu}^\pi$ as a vector in $\mathbb{R}^{nm}$.

Definition 1.9. *(**Occupancy Set**) The **set of valid occupancy measure** (or occupancy set) is given by:*

$$\mathcal{M} := \Big\{\mu \in \mathbb{R}_+^{nm} : (E - \gamma P)^T\boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}\Big\}.$$

*The equality that all points $\boldsymbol{\mu}$ in the set must satisfy are known as **Bellman Flow Constraints**.*

Definition 1.10. *(**State-Occupancy Measure**) Similarly to the (state-action) occupancy measure, one can define a **state-occupancy measure** $\boldsymbol{\mu}_S^\pi \in \Delta_S$ is defined for a given state as follows:*

$$\mu_S^\pi(s) = (1-\gamma)\lim_{T\to+\infty}\sum_{t=0}^{T}\gamma^t\mathbb{P}_\nu^\pi(s_t = s)$$

$$= (1-\gamma)\mathbb{E}\Big[\sum_{t=0}^{+\infty}\gamma^t\mathbf{1}(s_t = s)\Big|s_0 \sim \nu, \pi\Big],$$

since our state and action spaces are discrete we can represent $\boldsymbol{\mu}_S$ as a vector in $\mathbb{R}^n$.

Furthermore, when an state action pair s, a has occupancy measure $\mu^\pi(s,a) = 0$ (under policy π), we say that it is *unvisited*.

Observation 1.1. *The occupancy measure provides us with a very concise notation for writing out quantities such as the **objective function**, which can be represented in a simple scalar product form:*

$$J(\boldsymbol{\pi}, \boldsymbol{r}) = \langle \boldsymbol{r}, \boldsymbol{\mu}^\pi \rangle - \tilde{\Omega}(\boldsymbol{\mu}^\pi).$$

Which can be easily verified as follows:

$$\begin{aligned}
J^\nu(\pi) &= (1-\gamma)\mathbb{E}\Big[R(\tau)\Big|s_0 \sim \nu, \pi\Big] \\
&= (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{+\infty}\gamma^t(r(s,a) - \Omega(\pi(\cdot|s)))\Big|s_0 \sim \nu, \pi\right] \\
&= (1-\gamma)\mathbb{E}\left[\sum_{s,a}\sum_{t=0}^{+\infty}\gamma^t r(s,a)\,\boldsymbol{1}(s,a)\Big|s_0 \sim \nu, \pi\right] \\
&\quad - (1-\gamma)\mathbb{E}\left[\sum_{s}\sum_{t=0}^{+\infty}\gamma^t \Omega(\pi(\cdot|s)))\,\boldsymbol{1}(s)\Big|s_0 \sim \nu, \pi\right] \\
&= \sum_{s,a} r(s,a)\overbrace{(1-\gamma)\mathbb{E}\left[\sum_{t=0}^{+\infty}\gamma^t\,\boldsymbol{1}(s,a)\Big|s_0 \sim \nu, \pi\right]}^{=\boldsymbol{\mu}^\pi(s,a)} \\
&\quad - \sum_{s}\Omega(\pi(\cdot|s))\overbrace{(1-\gamma)\mathbb{E}\left[\sum_{t=0}^{+\infty}\gamma^t\,\boldsymbol{1}(s)\Big|s_0 \sim \nu, \pi\right]}^{=\boldsymbol{\mu}_S^\pi(s)} \\
&= \sum_{s,a} r(s,a)\boldsymbol{\mu}^\pi(s,a) - \overbrace{\sum_{s}\Omega(\pi(\cdot|s))\boldsymbol{\mu}_S^\pi(s)}^{:=\tilde{\Omega}(\boldsymbol{\mu})} \\
&= \boldsymbol{r}^T\boldsymbol{\mu}^{\boldsymbol{\pi}} - \tilde{\Omega}(\boldsymbol{\mu}^\pi).
\end{aligned}$$

Lemma 1.1. *(**Expected-Regularizer is convex in** μ) Consider the function $\tilde{\Omega} := \mathbb{E}_{s\sim\boldsymbol{\mu}}[\Omega(\pi(\cdot|s))]$,*

- *if Ω is **convex**, then so is $\tilde{\Omega}$*
- *if Ω is **strictly-convex**, then so is $\tilde{\Omega}$.*

(Proof in Schlaginhaufen 2023.)

Observation 1.1, together with lemma 1.1 is quite important as it shows that the objective function is **concave** w.r.t the occupancy measure (which opens the door for optimization approaches), moreover in the special case where $\Omega(\pi(\cdot|s)) = 0,\ \forall s \in S$ the objective is **linear** w.r.t the occupancy measure (which opens the door for linear programming approaches).

For convenience we interchangeably write the objective as a function of the policy π and of the occupancy measure $\boldsymbol{\mu}$:

$$J^{\nu}(\pi) = J^{\nu}(\boldsymbol{\mu}).$$

Observation 1.2. *There exists a (almost) well-defined bijection between any policy π and it's associated occupancy measure $\boldsymbol{\mu}^{\pi}$. With knowledge of P and π, one can compute the occupancy measure from the bellman flow constraint as follows:*

1. *compute the* ***closed loop transition law*** *P^{π},*
2. *compute the* ***state-occupancy measure*** *from the occupancy measure, this is easy as it has a matrix-geometric series form:*

$$\mu_S^{\pi} = (1-\gamma) \lim_{T \to +\infty} \sum_{t=0}^{T} (\gamma P^{\pi})^t \cdot \boldsymbol{\nu} = (1-\gamma)(I - \gamma P^{\pi})^{-1} \boldsymbol{\nu}$$

3. *finally one can compute the* ***occupancy measure*** *by*

$$\mu(s,a) = \pi(a|s)\mu_S(s).$$

Similarly, with knowledge of the occupancy mapping $\boldsymbol{\mu}$ one can recover a policy π as follows:

$$\pi^{\boldsymbol{\mu}}(a|s) = \begin{cases} \frac{\mu(s,a)}{\sum_a \mu(s,a)}, & \sum_a \mu(s,a) > 0 \\ \frac{1}{m}, & \textit{otherwise.} \end{cases}$$

Notice that the mapping is ill-defined for unvisited states, but this does not matter changes in policy on unvisited states have no impact on the discounted reward.

For convenience we define two maps that allow us to go from occupancy map to policies and from policies to occupancy maps. Here we state and prove a useful lemma which will be critical to proving the global convergence of CIRL algorithms.

Lemma 1.2 (Occupancy measure is Lipschitz with respect to the to the policy)**.** *More specifically, the occupancy measures $\boldsymbol{\mu}^{\pi}$ and $\boldsymbol{\mu}^{\bar{\pi}}$ respectively induced by policies $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ on identical MDPs (with the same state and action sets S and A and the same probabilistic markovian transition law P, rewards can be different) satisfy the following inequality:*

$$\|\boldsymbol{\mu}^{\pi} - \boldsymbol{\mu}^{\bar{\pi}}\|_2 \leq B_{\mu} \|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2,$$

where $B_{\mu} = \frac{1}{1-\gamma}$ and $\|\cdot\|_2$ denotes the l2 norm.

Proof. We start by expanding the expression from the lefthand side of the inequality we are trying to prove:

$$
\begin{aligned}
\|\boldsymbol{\mu}^{\pi} - \boldsymbol{\mu}^{\bar{\pi}}\|_2 &= \sqrt{\sum_{s,a \in S \times A} \Big(\mu^{\pi}(s,a) - \mu^{\bar{\pi}}(s,a)\Big)^2} \\
&\overset{(i)}{=} \sqrt{\sum_{s,a \in S \times A} \Big(\mu^{\pi}(s)\pi(a|s) - \mu^{\bar{\pi}}(s)\bar{\pi}(a|s)\Big)^2} \\
&\overset{(ii)}{=} \sqrt{\sum_{s,a \in S \times A} \Big(\mu^{\pi}(s)\big(\pi(a|s) - \bar{\pi}(a|s)\big) + \bar{\pi}(a|s)\big(\mu^{\pi}(s) - \mu^{\bar{\pi}}(s)\big)\Big)^2} \\
&\overset{(iii)}{\leq} \overbrace{\sqrt{\sum_{s,a \in S \times A} \Big(\mu^{\pi}(s)\big(\pi(a|s) - \bar{\pi}(a|s)\big)\Big)^2}}^{\leq \|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2} \\
&\quad + \overbrace{\sqrt{\sum_{s,a \in S \times A} \Big(\bar{\pi}(a|s)\big(\mu^{\pi}(s) - \mu^{\bar{\pi}}(s)\big)\Big)^2}}^{\leq \|\boldsymbol{\mu}_s - \bar{\boldsymbol{\mu}}_s\|_2} \\
&\overset{(iv)}{\leq} \|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2 + \|\boldsymbol{\mu}_s - \bar{\boldsymbol{\mu}}_s\|_2.
\end{aligned}
$$

Where in (i) we just plug in the defintion of the state-occupancy measure (def 1.10), in (ii) we just add $0 = \mu^{\pi}(s)\bar{\pi}(a|s) - \mu^{\pi}(s)\bar{\pi}(a|s)$ and rearrang. Next, we just use a triangle inequality (iii) and observing that both sides are upper bounded by l2 norms we are done with the first step.

We will now be concerned with bounding $\|\boldsymbol{\mu}_s - \bar{\boldsymbol{\mu}}_s\|_2$ by $\|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2$ (mutliplied by some constant term).

To do so we first show a useful result on the spectral norm of the difference between the inverse of two matrices:

$$
\begin{aligned}
\|A^{-1} + B^{-1}\| &\overset{(i)}{=} \|A^{-1}(A+B)B^{-1}\| \\
&\overset{(ii)}{\leq} \|A^{-1}\| \cdot \|(A+B)\| \cdot \|B^{-1}\| \\
&\overset{(iii)}{=} \frac{\|(A+B)\|}{\sigma_{\min}(A) \cdot \sigma_{\min}(B)}.
\end{aligned} \tag{l}
$$

Where (i) holds by the equality $A^{-1} + B^{-1} = A^{-1}(A+B)B^{-1}$ which holds for any two invertible matrices (a proof can be found in the solution handbook to Searle 1982), (ii) holds by submultiplicativity of the spectral norm and (iii) uses the definition of the spectal norm ($\sigma_{\min}(A)$ denotes the minimum eigenvalue of the matrix A).

We now get back to bounding $\|\boldsymbol{\mu}_s - \bar{\boldsymbol{\mu}}_s\|_2$:

$$\begin{aligned}
\|\boldsymbol{\mu}_s - \bar{\boldsymbol{\mu}}_s\|_2 &\overset{(i)}{=} (1-\gamma)\left\|\left[\left(I-\gamma P^{\pi}\right)^{-1} - \left(I-\gamma P^{\bar{\pi}}\right)^{-1}\right]\boldsymbol{\nu}\right\|_2 \\
&\overset{(ii)}{\leq} (1-\gamma)\left\|\left[\left(I-\gamma P^{\pi}\right)^{-1} - \left(I-\gamma P^{\bar{\pi}}\right)^{-1}\right]\right\| \cdot \|\boldsymbol{\nu}\|_2 \\
&\overset{(iii)}{\leq} (1-\gamma)\left\|\left[\left(I-\gamma P^{\pi}\right)^{-1} - \left(I-\gamma P^{\bar{\pi}}\right)^{-1}\right]\right\| \\
&\overset{(iv)}{\leq} (1-\gamma)^{-1}\left\|\left[\left(I-\gamma P^{\pi}\right) - \left(I-\gamma P^{\bar{\pi}}\right)\right]\right\| \\
&= \frac{\gamma}{1-\gamma}\left\|P^{\pi} - P^{\bar{\pi}}\right\|.
\end{aligned}$$

Where in (i) we just use the closed form computation of the state-occupancy measure from the policy (as shown in obs 1.2), in (ii) we just use the definition of the spectral norm with pulls out the $\|\boldsymbol{\nu}\|_2$ term, which we know is smaller or equal to 1 (since it is the l2 norm of a probability distribution) which gives us inequality (iii). Now plugging in the result from (l) and using that the smallest eigenvalue of $\left(I-\gamma P^{\pi}\right)^{-1}$ is greater or equal to $1-\gamma$ we get inequality (iv) which simplifies into the last line.

Now we just need to bound $\left\|P^{\pi} - P^{\bar{\pi}}\right\|$ which we do as follows:

$$\begin{aligned}
\left\|P^{\pi} - P^{\bar{\pi}}\right\| &\overset{(i)}{\leq} \left\|P^{\pi} - P^{\bar{\pi}}\right\|_F \\
&\overset{(ii)}{=} \sqrt{\sum_{s,s' \in S\times S}\left(P^{\pi}(s'|s) - P^{\bar{\pi}}(s'|s)\right)^2} \\
&\overset{(iii)}{=} \sqrt{\sum_{s,s' \in S\times S}\Big(\sum_{a\in A} P(s'|s,a)(\pi(a|s) - \bar{\pi}(a|s))\Big)^2} \\
&= \sqrt{\sum_{s,a,s' \in S\times A\times S} P(s'|s,a)^2(\pi(a|s) - \bar{\pi}(a|s))^2} \\
&= \sqrt{\sum_{s,a \in S\times A}\Big(\sum_{s'\in S} P(s'|s,a)^2\Big)(\pi(a|s) - \bar{\pi}(a|s))^2} \\
&\leq \sqrt{\sum_{s,s' \in S\times S}(\pi(a|s) - \bar{\pi}(a|s))^2} = \|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2.
\end{aligned}$$

Where (i) comes from the fact that the Frobenius norm upper bounds the spectral norm, (ii) is by the definition of the Frobenius norm, and (iii) just plugs in the definition of the closed loop transition law (def 1.3) from there we can just rearange and isolate the $\left(\sum_{s'\in S} P(s'|s,a)^2\right)$ term, which since $P(\cdot,s,a) \in \Delta_S$ we know is less than 1. From there we just observe that we have gotten to the definition of the $l2$ norm.

Putting everything back together we have:

$$\begin{aligned}\|\boldsymbol{\mu}^{\pi} - \boldsymbol{\mu}^{\bar{\pi}}\|_2 &\leq \|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2 + \|\boldsymbol{\mu}_s - \bar{\boldsymbol{\mu}}_s\|_2 \\ &\leq \|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2 + \frac{\gamma}{1-\gamma}\left\|P^{\pi} - P^{\bar{\pi}}\right\| \\ &\leq \|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2 + \frac{\gamma}{1-\gamma}\|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2 \\ &= \frac{1}{1-\gamma}\|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2\end{aligned}$$

□

Definition 1.11. *The* $OM : \Pi \to \mathcal{M}$ *returns the occupancy measure* $\boldsymbol{\mu} = OM(\boldsymbol{\pi})$ *associated with a given policy* $\boldsymbol{\pi}$*.*

Definition 1.12. *The* $PO : \mathcal{M} \to \Pi$ *returns a policy* $\boldsymbol{\pi} = PO(\boldsymbol{\mu})$ *that induces the occupancy measure* $\boldsymbol{\mu}$*.*

We expect that the policy for unvisited state is always uniform ($\pi(a|s_{\text{univisted}}) = 1/m \ \forall a \in A$), the following property is satisfied

$$\begin{aligned}(\mathrm{OM} \circ \mathrm{PO}) &= \mathrm{Id} \\ (\mathrm{PO} \circ \mathrm{OM}) &= \mathrm{Id}\end{aligned}$$

where Id is the identity operator.

1.1.3 Formulating solving MDPs as optimization problems

The goal of an MDP can be stated as follows:

$$\max_{\boldsymbol{\pi} \in \Pi} J(\boldsymbol{\pi}, \boldsymbol{r}) \tag{1.1}$$

i.e. find the policy that gives maximum expected discounted reward in the MDP M under initial state distribution ν. Solving the problem 1.1 is known as *direct policy optimization*. It is a **non-concave** optimization problem but we can find algorithms that converge globally on it Agarwal et al. 2020. Alternatively because of observation 1.2 (policy-occupancy measure bijection), one can also formulate solving 1.1 in terms of the occupancy measure as follows:

$$\begin{aligned}\max_{\boldsymbol{\mu}} \ & \langle \boldsymbol{r}, \boldsymbol{\mu} \rangle - \tilde{\Omega}(\boldsymbol{\mu}) \\ \text{s.t.} \ & (E - \gamma P)^T \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu},\end{aligned} \tag{1.2}$$

where the equality constraints $(E - \gamma P)^T \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}$ (the **Bellman Flow Constraints**) are equivalent to restricting $\boldsymbol{\mu}$ to the occupancy set $\mathcal{M}$ (see definition 1.9). Note that as the problem (1.2) has a concave cost-function with a set of linear equality constraints, it is a concave program. Observe that in the special case where $\tilde{\Omega}(\boldsymbol{\mu}) = 0$ (1.2) becomes a Linear Program that can be solved in poly-time. The problem (1.2), naturally relaxes to a an associated, unconstrained Lagrangian form:

$$\max_{\boldsymbol{\mu}} \min_{\boldsymbol{\lambda} \geq 0} \quad \langle \boldsymbol{r}, \boldsymbol{\mu} \rangle - \tilde{\Omega}(\boldsymbol{\mu}) + \langle \boldsymbol{\lambda}, \big((E - \gamma P)^T \boldsymbol{\mu} - (1-\gamma)\boldsymbol{\nu} \big) \rangle, \tag{1.3}$$

in which case the problem becomes a *saddle-point problem*. In the unregularized form ($\tilde{\Omega}(\boldsymbol{\mu}) = 0$) the problem is *bilinear* w.r.t the lagrangian multiplier $\boldsymbol{\lambda}$ and the occupancy measure (primal variable) $\boldsymbol{\mu}$, else it is concave-linear. We sometimes use notation $J(\boldsymbol{\mu}, \boldsymbol{r}) = J(\mathrm{OM}(\boldsymbol{\pi}), \boldsymbol{r})$ and $K(\boldsymbol{\mu},) = K(\mathrm{OM}(\boldsymbol{\pi}))$.

1.2 Constrained Markov Decision Processes

The *Constrained Markov Decision Process* (CMDP) is an extension of the above-defined MDP that incorporates a cost function as well as constraints on that cost function, we define the CMDP as the following 8-tuple:

$$\text{CMDP} = (S, A, P, \boldsymbol{\nu}, \boldsymbol{r}, \Psi, \boldsymbol{b}, \gamma),$$

where the terms $S, A, P, \boldsymbol{\nu}, \boldsymbol{r}, \gamma$ are defined as in an unconstrained MDP (see section 1.1) but we further incorporate a cost function $\Psi : S \times A \to \mathbb{R}^l$ which, in the discrete state-action setting, we can conveniently represent as a matrix $\Psi \in \mathbb{R}^{nm \times l}$ (here l is the constraint-dimensionality). We denote the i-th element of the output of the cost function Ψ: ψ_i. The constraint vector $\boldsymbol{b} \in \mathbb{R}^l$ allows us to define our CMDP constraints.

1.2.1 Discounted Costs, Cost-Value, Cost-Quality, Cost-Objective function

We define equivalent quantities (discounted cost, value, quality and objective) for cost to the ones we defined for the reward function.

Definition 1.13. *(**Discounted Cost**) For a given trajectory τ of the MDP M one can compute the i-th element of the **discounted cost** $U(\tau) \in \mathbb{R}^l$ as follows:*

$$U(\tau, i) = (1-\gamma) \lim_{T \to +\infty} \sum_{T}^{t=0} \gamma^t \psi_i(s_t, a_t),$$

we denote the full vector obtained by stack elements as defined above as $\boldsymbol{U}(\tau)$,

$$\boldsymbol{U}(\tau) = (1-\gamma) \lim_{T \to +\infty} \sum_{T}^{t=0} \gamma^t \Psi(s_t, a_t).$$

Definition 1.14. *(**Cost-Value Function**) For a given policy $\pi : S \to \Delta A$ and under the assumption that the agent starts at some state s_0, we define the **cost-value function** $\boldsymbol{V}_\Psi^\pi : S \to \mathbb{R}^l$ as:*

$$\boldsymbol{V}_\Psi^\pi(s) = \mathbb{E}\Big[\boldsymbol{U}(\tau)\Big| s_0 = s, \pi\Big] = \mathbb{E}\left[(1-\gamma)\sum_{t=0}^{+\infty} \gamma^t \Psi(s_t, a_t)\Big| s_0 = s, \pi\right].$$

Definition 1.15. *(**Cost-Quality Function**) For a given policy $\pi : S \to \Delta A$ and under the assumption that the agent starts at some state s_0 with first action a_0, we define the **cost-quality function** $\boldsymbol{Q}^\pi : S \times A \to \mathbb{R}^l$ as:*

$$\boldsymbol{Q}_\Psi^\pi(s, a) = \mathbb{E}\Big[\boldsymbol{U}(\tau)\Big| s_0 = s, a_0 = a, \pi\Big] = \mathbb{E}\left[(1-\gamma)\sum_{t=0}^{+\infty} \gamma^t \Psi(s_t, a_t)\Big| s_0 = s, a_0 = a, \pi\right].$$

Definition 1.16. *(**Cost-Objective Function**) For a given policy $\pi : S \to \Delta A$ and under an initial state distribution $\boldsymbol{\nu} \in \Delta_S$ we define the **cost-objective function** as:*

$$\boldsymbol{K}(\pi) = \mathbb{E}\Big[\boldsymbol{U}(\tau)\Big| s_0 \sim \nu, \pi\Big] = \mathbb{E}\Big[V_\Psi^\pi(s)\Big| s \sim \nu\Big].$$

Observation 1.3. *As in the unconstrained MDP setting, one can rewrite the cost-objective as a function of the occupancy measure:*

$$\Psi^T \boldsymbol{\mu}^\pi = \boldsymbol{K}(\pi).$$

1.2.2 Feasibility

The main difference between *MDP*s and *CMDP*s is that in a *CMDP*, not all policies are admissible, in order to be feasible, a policy needs to meet the constraint:

$$\boldsymbol{K}(\pi) \leq \boldsymbol{b},$$

or equivalently

$$\Psi^T \boldsymbol{\mu}^\pi \leq \boldsymbol{b}.$$

This restricts the set of admissible policies and occupancy measures and allows us to define the *feasible set* of the CMDP.

Definition 1.17. ***(Feasible Set)*** *Given a CMPD C we define it's* ***feasible set*** $\mathcal{F}$ *as the set of points that satisfy the bellman flow constraints without violating the constraints:*

$$\mathcal{F} := \left\{ \mu \in \mathcal{M} \middle| \Psi^T \mu \leq \boldsymbol{b} \right\}.$$

1.2.3 Goal of the CMDP, formulating optimization problems

The goal of a CMDP can be stated as follows:

$$\begin{aligned} &\max_{\pi \in \Pi} J(\boldsymbol{\pi}, \boldsymbol{r}), \\ &\quad \text{s.t } \boldsymbol{J}^\nu_\Psi(\pi) \leq \boldsymbol{b}, \end{aligned} \tag{1.4}$$

which in plain english reads as "*find the policy that maximizes the expected discounted reward while keeping the expected discounted cost below the constraints*". As we already did in section 1.1.3, we can reformulate our problems in terms of the occupancy measure:

$$\begin{aligned} \max_{\boldsymbol{\mu}} \quad & \boldsymbol{r}^T \boldsymbol{\mu} - \tilde{\Omega}(\mu), \\ \text{s.t.} \quad & (E - \gamma P)^T \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}, \\ & \Psi^T \boldsymbol{\mu} \leq \boldsymbol{b}, \end{aligned} \tag{1.5}$$

as we showed in section 1.1.3 for the unconstrained setting, this yields a concave program, with linear constraints, which becomes an LP in the special case where $\Omega(\pi) = 0$.

1.2.4 Solution Maps

We formulate the following solution maps are associated with the solutions of the optimization problems associated with the (constrained or unconstrained) MDPs.

Definition 1.18. *(**Policy Solution Map**) Given a CMDP C and it's reward* $\boldsymbol{r}$*, we define it's* ***policy solution map*** *as:*

$$CRL_{\pi}(\boldsymbol{r}) = \arg\max_{\pi \in \Pi} J(\boldsymbol{\pi}, \boldsymbol{r})$$
$$s.t.\ OM(\pi) \in \mathcal{F}$$

Definition 1.19. *(**Occupancy Solution Map**) Given a CMDP C and it's reward* $\boldsymbol{r}$*, we define it's* ***occupancy solution map*** *as:*

$$CRL_{\mu}(\boldsymbol{r}) = \arg\max_{\mu \in \mathcal{F}} J(\boldsymbol{\mu}, \boldsymbol{r})$$

Note that we also use the notations $\mathrm{RL}_{\pi/\mu}(\boldsymbol{r})$ for the unconstrained solution maps in occupancy and policy.

Observation 1.4. *The solution maps from def 1.18 and from def 1.19 are equivalent under the bijection from observation 1.2, since by observation 1.1 the optimized values are identical.*

1.3 Inverse Reinforcement Learning

We will now consider the inverse problem to the MDP and CMDP problems that we discussed before, **Constrained Inverse Reinforcement Learning** (IRL). In the IRL setting (*on a known MDP*) we are given an *MDP* or a *CMDP* without its reward:

$$CMDP \setminus \boldsymbol{r} = (S, A, P, \boldsymbol{\nu}, \Psi, \boldsymbol{b}, \gamma),$$

as well as a dataset $\mathcal{D}$ of expert example in the form of trajectories:

$$\mathcal{D} = \{\tau_1, \tau_2, ..., \tau_N\},$$

produced by some expert policy $\boldsymbol{\pi}^E$. Assuming the dataset is large enough (which we will do at first), this is equivalent to getting an approximation of the expert's occupancy measure $\boldsymbol{\mu}^E$. In the CIRL setting we also generally restrict the reward class $\mathcal{R}$ to some arbitrary convex set in $\mathbb{R}^{nm}$.

Definition 1.20. *(**CIRL Solution Map**) The **goal of CIRL** is to find some mapping $CIRL_\pi : \Pi \to \mathcal{R}$ (or alternatively $CIRL_\mu : \mathcal{M} \to \mathcal{R}$). That satisfies*

$$(CRL_\pi \circ CIRL_\pi)(\boldsymbol{\pi}^E) = \boldsymbol{\pi}^E.$$

In plain english **we want to find a method for recovering a reward for which the original expert is optimal.**

Assumption 1.1. *(**Realizability**) We assume that the expert policy is optimal w.r.t. some reward $\boldsymbol{r}^E \in \mathcal{R}$, i.e.*

$$\boldsymbol{\mu}^E = CRL_\mu(\boldsymbol{r}^E).$$

The goal of the IRL problem is to recover a reward $\hat{\boldsymbol{r}}$ s.t. the expert $\boldsymbol{\mu}^E = \mathrm{CRL}_\mu(\hat{\boldsymbol{r}})$ (the expert is optimal for that reward).

Proposition 1.1. *(**Min-Max Program to solve IRL**) If assumption 1.1 is true, then the rewards optimizing the program:*

$$\min_{\boldsymbol{r} \in \mathcal{R}} \max_{\boldsymbol{\mu} \in \mathcal{F}} \langle \boldsymbol{r}, \boldsymbol{\mu} - \boldsymbol{\mu}^E \rangle - \tilde{\Omega}(\boldsymbol{\mu}), \tag{1.6}$$

are exactly the rewards in $\mathcal{R}$ for which $\boldsymbol{\mu}^E$ is optimal.

Proof. (Of proposition 1.1) We can rewrite an equivalent problem to (1.6) as follows:

$$\min_{\boldsymbol{r} \in \mathcal{R}} \max_{\boldsymbol{\mu} \in \mathcal{F}} \langle \boldsymbol{r}, \boldsymbol{\mu} \rangle - \tilde{\Omega}(\boldsymbol{\mu}) - \langle \boldsymbol{r}, \boldsymbol{\mu}^E \rangle + \tilde{\Omega}(\boldsymbol{\mu}^E),$$

where the addition of the $\tilde{\Omega}(\boldsymbol{\mu}^E)$ term has no influence on the optimizers, as it is a function of neither of the decision variables. For convenience we write:

$$\boldsymbol{r}^T \boldsymbol{\mu} - \tilde{\Omega}(\boldsymbol{\mu}) - \boldsymbol{r}^T \boldsymbol{\mu}^E + \tilde{\Omega}(\boldsymbol{\mu}^E) = L(\boldsymbol{r}, \boldsymbol{\mu}).$$

For any fixed $\boldsymbol{r} \in \mathcal{R}$ it holds that

$$\arg\max_{\boldsymbol{\mu} \in \mathcal{F}} \langle \boldsymbol{r}, \boldsymbol{\mu} \rangle - \tilde{\Omega}(\boldsymbol{\mu}) - \overbrace{\langle \boldsymbol{r}, \boldsymbol{\mu}^E \rangle + \tilde{\Omega}(\boldsymbol{\mu}^E)}^{\text{constant w.r.t } \boldsymbol{\mu}} = \mathrm{CRL}_\mu(\boldsymbol{r}),\ \boldsymbol{r} \in \mathcal{R},$$

i.e. the optimal $\boldsymbol{\mu}$ for any fixed $\boldsymbol{r} \in \mathcal{R}$ gives the optimal policy. Moreover, for any fixed $\boldsymbol{r} \in \mathcal{R}$, we get that:

$$\max_{\boldsymbol{\mu} \in \mathcal{F}} L(\boldsymbol{r}, \boldsymbol{\mu}). \geq 0,\ \boldsymbol{r} \in \mathcal{R},$$

since:

1. either μ^E maximizes the value $\langle \boldsymbol{r}, \boldsymbol{\mu}^E \rangle - \tilde{\Omega}(\boldsymbol{\mu}^E)$, in which case, we can set the optimizer $\boldsymbol{\mu} = \boldsymbol{\mu}^E$ and get $L(\boldsymbol{r}, \boldsymbol{\mu}) = 0$,
2. or μ^E does not maximize the value $\langle \boldsymbol{r}, \boldsymbol{\mu}^E \rangle - \tilde{\Omega}(\boldsymbol{\mu}^E)$, in which case we get $L(\boldsymbol{r}, \boldsymbol{\mu}) > 0$.

Observe that the lower bound ($L(\boldsymbol{r}, \boldsymbol{\mu}) = 0$) is only achieved if $\langle \boldsymbol{r}, \boldsymbol{\mu}^E \rangle - \tilde{\Omega}(\boldsymbol{\mu}^E) = \langle \boldsymbol{r}, \boldsymbol{\mu} \rangle - \tilde{\Omega}(\boldsymbol{\mu})$, i.e. if $\boldsymbol{\mu}^E \in \arg\max_{\boldsymbol{\mu} \in \mathcal{F}} \langle \boldsymbol{r}, \boldsymbol{\mu}^E \rangle + \tilde{\Omega}(\boldsymbol{\mu}^E)$ which is equivalent to $\boldsymbol{\mu}^E \in \mathrm{CRL}_\mu(\boldsymbol{r})$. I.e. the optimal reward $\boldsymbol{r}^*$ of the min-max problem is such that the expert $\boldsymbol{\mu}^E$ is optimal w.r.t $\boldsymbol{r}^*$.

□

Proposition 1.2 (Strong Duality)**.** *Assuming that $\exists \boldsymbol{\pi}^E \in \mathcal{F}$, and that obj is strictly convex. Then dual optimum is attained for some $\boldsymbol{\lambda}^* \geq 0$ and problem 4.1 is equivalent to an unconstrained bandit problem with reward $\boldsymbol{r} - \Psi\boldsymbol{\lambda}^*$. I.e.*

$$CRL_\pi(\boldsymbol{r}) = RL_\pi(\boldsymbol{r} - \boldsymbol{\Psi}\boldsymbol{\lambda}^*).$$

Proof. Using generic Lagrangian Duality theory, we make two observations:

1. observe that the problem: $\max_{\boldsymbol{\mu} \in \mathcal{F}} \langle \boldsymbol{r}, \boldsymbol{\mu} \rangle - \tilde{\Omega}(\boldsymbol{\pi})$, is a **strictly convex optimization problem** (for a proof of the convexity of $\tilde{\Omega}$, assuming that Ω is convex refer to Schlaginhaufen 2023),
2. the primal optimum $\boldsymbol{\pi}^*$ is finite as the feasible set $\mathcal{F}$ is bounded and the objective is upper-bounded (since Ω is strictly convex).

From Slater's condition it follows that strong duality holds. □

Definition 1.21 (Diminished Reward)**.** *We call diminished reward the function $\tilde{r} : S \times A \to \mathcal{R}$ defined as:*

$$\tilde{\boldsymbol{r}} = \boldsymbol{r} - \boldsymbol{\Psi}\boldsymbol{\lambda},$$
$$\tilde{r}(s, a) = \boldsymbol{r}(s, a) - \langle \boldsymbol{\Psi}(s, a), \boldsymbol{\lambda}(s, a) \rangle,$$

this definition provides a shorthand for making use of strong duality (proposition 1.2) in the analysis of CIRL convergence.

2 Min-Max Optimization and Saddle Point Problems

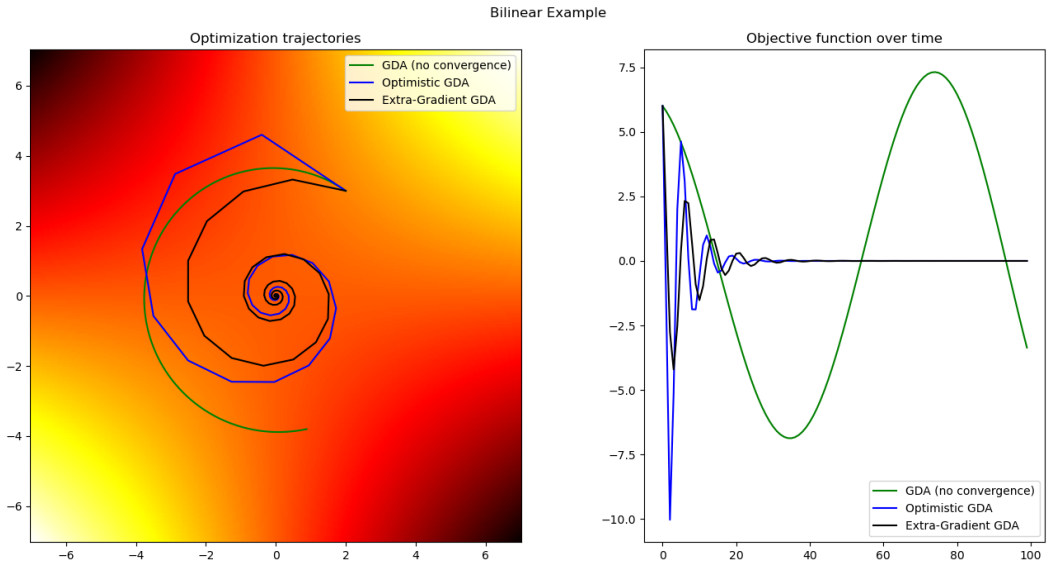

Figure 1: Iterations of three optimization methods (Gradient Descent-Ascent, Optimistic Gradient and Extra Gradient methods) on an unconstrained bilinear saddle-point optimization problem.

In the following section, we introduce, discuss and analyze algorithms for solving *saddle-point* problems of the form;

$$\min_{\boldsymbol{x}\in X}\max_{\boldsymbol{y}\in Y} f(\boldsymbol{x},\boldsymbol{y}).$$

Where $f : X \times Y \to \mathbb{R}$ is some scalar function for which we want to find a *saddle point*, i.e. a point $(\boldsymbol{x}^*, \boldsymbol{y}^*) \in X \times Y$, s.t.

$$f(\boldsymbol{x}^*, \boldsymbol{y}) \leq f(\boldsymbol{x}^*, \boldsymbol{y}^*) \leq f(\boldsymbol{x}, \boldsymbol{y}^*),$$

i.e. in plain english "$(\boldsymbol{x}^*, \boldsymbol{y}^*)$ *is a minimizer in* $\boldsymbol{x}$ *and a maximizer in* $\boldsymbol{y}$". Such problems which will prove especially relevant in our study of inverse reinforcement learning.

2.1 Preliminaries

Definition 2.1. ***(Lipschitz-continuous function)*** *A function* $f : \boldsymbol{dom}(f) \to \mathbb{R}^m$ *is* B*-Lipschitz if for any* $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{dom}(f)$*, there exists* $B \in \mathbb{R}_+$

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \leq B\|\boldsymbol{x} - \boldsymbol{y}\|.$$

Definition 2.2. ***(Smooth function)*** *A function* $f : \boldsymbol{dom}(f) \to \mathbb{R}$ *is called* L*-smooth if it has* L*-Lipschitz continuous gradients, i.e. if for any* $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{dom}(f)$*, there exists* $L \in \mathbb{R}_+$*:*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|.$$

Definition 2.3. *(**Convex set**) A set $C \subseteq \mathbb{R}^d$ is convex if for any two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, the connecting line segment is contained in C, i.e. $\forall \lambda, \lambda \in [0,1]$:*

$$\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in C.$$

Definition 2.4. *(**Convex function**) A function $f : \boldsymbol{dom}(f) \to \mathbb{R}$ is called convex on $\boldsymbol{dom}(f)$ if*

1. *$\boldsymbol{dom}(f)$ is convex,*
2. *for all $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{dom}(f)$ and $\lambda \in [0,1]$ we have:*

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}).$$

Definition 2.5. *(**Concave function**) A function $f : \boldsymbol{dom}(f) \to \mathbb{R}$ is called concave on $\boldsymbol{dom}(f)$ if the function $g : \boldsymbol{dom}(f) \to \mathbb{R}$ defined as $g(x) = -f(x)$ is convex on $\boldsymbol{dom}(f)$.*

Definition 2.6. *(**Saddle point**) The point $(\boldsymbol{x}^*, \boldsymbol{y}^*) \in X \times Y$ is a saddle point of the function $f : X, Y \to \mathbb{R}$ if for any $\boldsymbol{x} \in X$ and $\boldsymbol{y} \in Y$ we have:*

$$f(\boldsymbol{x}^*, \boldsymbol{y}) \leq f(\boldsymbol{x}^*, \boldsymbol{y}^*) \leq f(\boldsymbol{x}, \boldsymbol{y}^*).$$

Definition 2.7. *(**Saddle-Point Problem**) Consider $f : X, Y \to \mathbb{R}$ a continuously differentiable scalar function on convex domains $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$. For convenience use notation $Z = X \times Y$ and $z = (x, y)$. We call the optimization problem below:*

$$\min_{\boldsymbol{x} \in X} \max_{\boldsymbol{y} \in Y} f(\boldsymbol{x}, \boldsymbol{y}).$$

a saddle point problem. The solution set $\mathcal{Z}$ defined as:

$$\mathcal{Z} := \left\{ (x^*, y^*) \in X \times Y \,\middle|\, f(\boldsymbol{x}^*, \boldsymbol{y}) \leq f(\boldsymbol{x}^*, \boldsymbol{y}^*) \leq f(\boldsymbol{x}, \boldsymbol{y}^*),\ \forall x \in X,\ y \in Y \right\}$$

is the set of all saddle points (see def 2.6) of the function f. Such a problem can be constrained ($X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$) or unconstrained ($X = \mathbb{R}^n$, $Y = \mathbb{R}^m$).

Definition 2.8. *(**Convex-Concave Problem**) Consider a saddle point problem (see def 2.7), defined by $f : X, Y \to \mathbb{R}$ a continuously differentiable scalar function on convex domains $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$. We call the problem a convex-concave problem if f is convex in x and concave in y.*

Proposition 2.1. *In general:*

$$\max_{\boldsymbol{x} \in X} \min_{\boldsymbol{y} \in Y} f(\boldsymbol{x}, \boldsymbol{y}) \leq \min_{\boldsymbol{y} \in Y} \max_{\boldsymbol{x} \in X} f(\boldsymbol{x}, \boldsymbol{y}).$$

Proof. Let $\tilde{\boldsymbol{x}} = \arg\max_{\boldsymbol{y} \in Y} \min_{\boldsymbol{x} \in X} f(x, y)$, $\tilde{\boldsymbol{y}} = \arg\max_{\boldsymbol{y} \in Y} \min_{\boldsymbol{x} \in X} f(x, y)$, $\square$

Theorem 2.1. *(**Sion's Minimax Theorem** Sion 1958) If X and Y are convex compact sets, and if $f : X \times Y \to \mathbb{R}$ is convex concave, then:*

$$\max_{\boldsymbol{x} \in X} \min_{\boldsymbol{y} \in Y} f(\boldsymbol{x}, \boldsymbol{y}) = \min_{\boldsymbol{y} \in Y} \max_{\boldsymbol{x} \in X} f(\boldsymbol{x}, \boldsymbol{y}).$$

Definition 2.9. *(**Projection Operator**) given some convex set $S \subseteq \mathbb{R}^n$ as well a point $\boldsymbol{x} \in \mathbb{R}^n$, we define the projection $\boldsymbol{p}$ of $\boldsymbol{x}$ onto S as:*

$$\boldsymbol{p} = \arg\min_{\boldsymbol{y} \in S} \|\boldsymbol{x} - \boldsymbol{y}\|,$$

which we often simply denote as the output of the projection operator associated with the set S:

$$\Pi_S(\boldsymbol{x}) := \arg\min_{\boldsymbol{y} \in S} \|\boldsymbol{x} - \boldsymbol{y}\|.$$

When the norm used is the Euclidian norm (or L_2 norm), we call this operation the Euclidian Projection.

Definition 2.10. *(**Duality Gap**) We define the duality gap as a way of characterizing the sub-optimality of a point. Given some point $\tilde{\boldsymbol{z}} = (\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \in Z$, its duality gap is given by:*

$$\textit{Duality Gap}(\tilde{\boldsymbol{z}}) := \max_{\boldsymbol{y} \in Y} f(\tilde{\boldsymbol{x}}, \boldsymbol{y}) - \min_{\boldsymbol{x} \in X} f(\boldsymbol{x}, \tilde{\boldsymbol{y}})$$

2.2 Gradient Descent Ascent

The most simple algorithm is the natural extension of gradient descent/ascent to the saddle point problem. In the following section, we define the gradient descent-ascent (GDA) algorithm, analyze it and discuss it's performance and limitations. Note that we specifically analyze *projected* gradient descent-ascent, as results for projected GDA trivially generalize to unconstrained GDA and the projected setting will be useful for our applications in inverse reinforcement learning.

Algorithm 1: (Projected) Gradient Descent Ascent

Set the learning rate $\eta > 0$
Initialize the algorithm at some point $(\boldsymbol{x}_0, \boldsymbol{y}_0)$
foreach *iteration* $k = 0, 2, ..., K-1$ **do**
  $\boldsymbol{x}_{k+1} \leftarrow \Pi_X\big(\boldsymbol{x}_k - \eta \nabla_x f(x_k, y_k)\big)$
  $\boldsymbol{y}_{k+1} \leftarrow \Pi_Y\big(\boldsymbol{y}_k + \eta \nabla_y f(x_k, y_k)\big)$
end
return $(\boldsymbol{x}_N, \boldsymbol{y}_N)$

Proposition 2.2. *($O(\frac{1}{\sqrt{K}})$ **convergence rate**) For B-Lipschitz convex concave functions $f : X, Y \to \mathbb{R}$, defined on two convex domains X and Y with bounded diameter $D = \max\{diam(X), diam(Y)\}$, after K steps, projected gradient descent-ascent (alg 1) with learning rate $\eta_k = \frac{D}{\sqrt{2kB}}$ has gives an average point for which the duality gap is bounded by:*

$$\textit{Duality Gap}(\bar{\boldsymbol{z}}_K) \leq \frac{4DB}{\sqrt{K}},$$

where $\bar{\boldsymbol{z}}_K = \frac{1}{K} \sum_{k=1...K} (\boldsymbol{x}_k, \boldsymbol{y}_k)$.

Proof. (of proposition 2.2) Given a fixed-point $\boldsymbol{z}^* = (\boldsymbol{x}^*, \boldsymbol{y}^*)$, using the projected gradient update on $\boldsymbol{x}$ we get:

$$\begin{aligned}
\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 &= \|\Pi_X(\boldsymbol{x}_k - \eta_k \nabla_x f(\boldsymbol{x}_k, \boldsymbol{y}_k)) - \Pi(\boldsymbol{x}^*)\|^2 && \text{since } \boldsymbol{x}^* \in X \\
&\leq \|\boldsymbol{x}_k - \eta_k \nabla_x f(\boldsymbol{x}_k, \boldsymbol{y}_k) - \boldsymbol{x}^*\|^2 && \text{projections are non-expansive} \\
&= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + \eta_k^2 \|\nabla_x f(\boldsymbol{x}_k, \boldsymbol{y}_k)\|^2 \\
&\quad - 2\eta_K \langle \boldsymbol{x}_k - \boldsymbol{x}^*, \nabla_x f(\boldsymbol{x}_k, \boldsymbol{y}_k) \rangle \\
&\leq \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + \eta_k^2 B^2 && \text{since } f \text{ is } B\text{-Lipschitz} \\
&\quad - 2\eta_K [f(\boldsymbol{x}_k, \boldsymbol{y}_k) - f(\boldsymbol{x}^*, \boldsymbol{y}_k)] && \text{by convexity of } f
\end{aligned}$$

And thus:

$$\Rightarrow \quad f(\boldsymbol{x}_k, \boldsymbol{y}_k) - f(\boldsymbol{x}^*, \boldsymbol{y}_k) \leq \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2}{2\eta_K} + \frac{\eta_K}{2} B^2. \qquad (a)$$

Similarly using the projected gradient update on $\boldsymbol{y}$ we have:

$$\begin{aligned}
\|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*\|^2 &\leq \|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2 + \eta_k^2 \|\nabla_y f(\boldsymbol{x}_k, \boldsymbol{y}_k)\|^2 \\
&\quad + 2\eta_K \langle \boldsymbol{y}_k - \boldsymbol{y}^*, \nabla_y f(\boldsymbol{x}_k, \boldsymbol{y}_k) \rangle \\
&\leq \|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2 + \eta_k^2 B^2 && \text{since } f \text{ is } B\text{-Lipschitz} \\
&\quad - 2\eta_K [f(\boldsymbol{x}_k, \boldsymbol{y}_k) - f(\boldsymbol{x}_k, \boldsymbol{y}^*)] && \text{by concavity of } f
\end{aligned}$$

And thus:

$$\Rightarrow \quad f(\boldsymbol{x}_k, \boldsymbol{y}^*) - f(\boldsymbol{x}_k, \boldsymbol{y}_k) \leq \frac{\|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2 - \|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*\|^2}{2\eta_K} + \frac{\eta_K}{2} B^2. \qquad (b)$$

Adding (a) and (b) together we get:

$$\begin{aligned}
&f(\boldsymbol{x}_k, \boldsymbol{y}_k) - f(\boldsymbol{x}^*, \boldsymbol{y}_k) + f(\boldsymbol{x}_k, \boldsymbol{y}^*) - f(\boldsymbol{x}_k, \boldsymbol{y}_k) = f(\boldsymbol{x}_k, \boldsymbol{y}^*) - f(\boldsymbol{x}^*, \boldsymbol{y}_k) \\
&\leq \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2}{2\eta_K} + \frac{\|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2 - \|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*\|^2}{2\eta_K} + \eta_K B^2. \qquad (c)
\end{aligned}$$

Summing up over all optimization steps we have:

$$\begin{aligned}
&\sum_{k=1\ldots K} f(\boldsymbol{x}_k, \boldsymbol{y}^*) - f(\boldsymbol{x}^*, \boldsymbol{y}_k) \\
&\leq \sum_{k=1\ldots K} \left[\frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2}{2\eta_K} \right] \\
&+ \sum_{k=1\ldots K} \left[\frac{\|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2 - \|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*\|^2}{2\eta_K} \right] + B^2 \sum_{k=1\ldots K} \eta_K. \qquad (d)
\end{aligned}$$

Observing that the terms in the sum telescope we have:

$$\begin{aligned}
\sum_{k=1\ldots K} \left[\frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2}{2\eta_K} \right] &\leq \frac{D^2}{2\eta_K} \\
\sum_{k=1\ldots K} \left[\frac{\|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2 - \|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*\|^2}{2\eta_K} \right] &\leq \frac{D^2}{2\eta_K}.
\end{aligned}$$

Which once plugged into (d) give:

$$\sum_{k=1\ldots K} f(\boldsymbol{x}_k, \boldsymbol{y}^*) - f(\boldsymbol{x}^*, \boldsymbol{y}_k) \leq \frac{D^2}{\eta_K} + B^2 \sum_{k=1\ldots K} \eta_k = \frac{D^2\sqrt{K}}{\eta} + B^2\eta \sum_{k=1\ldots K} \frac{1}{\sqrt{k}}$$
$$\leq \frac{D^2\sqrt{K}}{\eta} + 2B^2\eta\sqrt{K}.$$

Thus we have:

$$\frac{1}{K}\Big[\sum_{k=1\ldots K} f(\boldsymbol{x}_k, \boldsymbol{y}^*) - f(\boldsymbol{x}^*, \boldsymbol{y}_k)\Big] \leq \frac{D^2}{\eta\sqrt{K}} + 2B^2\eta\frac{1}{\sqrt{K}}$$
$$\text{(Optimizing } \eta) \quad = \frac{4DB}{\sqrt{K}}$$

□

2.3 Extra-Gradient Descent Ascent

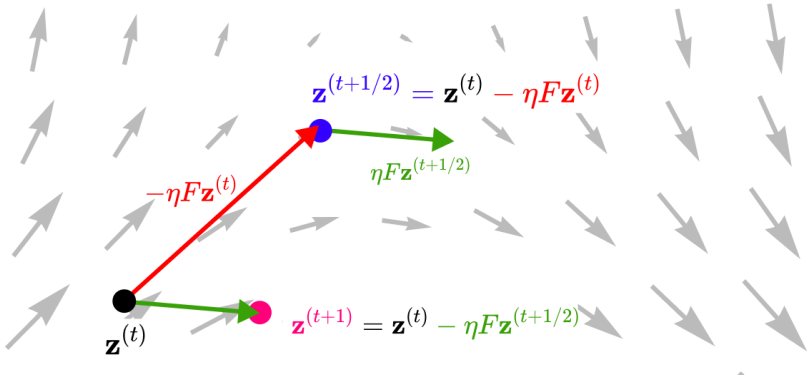

Figure 2: An illustration of an extra-gradient step.

Next, we consider the Extra-Gradient Descent Ascent method (EG), which provides an approximation of the *implicit* proximal point method and has better performance compared to the naïve GDA approach, specifically:

1. a better convergence rate on general convex-concave programs ($O(\frac{1}{K})$ instead of $O(\frac{1}{\sqrt{K}})$)
2. last iterate convergence for all convex-concave functions, including for bilinear functions (which isn't true for naïve GDA).

As in the naïve GDA case, we choose to analyze *projected* EG, because this will

prove useful to solve the IRL problems we are concerned about.

Algorithm 2: (Projected) Extra-Gradient Descent Ascent

Set the learning rate $\eta > 0$
Initialize the algorithm at some point $(\boldsymbol{x}_0, \boldsymbol{y}_0)$
foreach *iteration* $k = 0, 2, ..., K-1$ **do**
  $\boldsymbol{x}_{k+1/2} \leftarrow \Pi_X\big(\boldsymbol{x}_k - \eta \nabla_x f(\boldsymbol{x}_k, \boldsymbol{y}_k)\big)$
  $\boldsymbol{y}_{k+1/2} \leftarrow \Pi_Y\big(\boldsymbol{y}_k + \eta \nabla_y f(\boldsymbol{x}_k, \boldsymbol{y}_k)\big)$
  $\boldsymbol{x}_{k+1} \leftarrow \Pi_X\big(\boldsymbol{x}_k - \eta \nabla_x f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}_{k+1/2})\big)$
  $\boldsymbol{y}_{k+1} \leftarrow \Pi_Y\big(\boldsymbol{y}_k + \eta \nabla_y f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}_{k+1/2})\big)$
end
return $(\boldsymbol{x}_K, \boldsymbol{y}_K)$

In order to make the computations a bit lighter, we will use the following notation, let F denote the flow (the direction of gradients ascending and descending) of GDA/EG:

$$F(\boldsymbol{z}) = \begin{bmatrix} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}). \end{bmatrix}$$

We also define a single the feasible set $S := \{\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}); \boldsymbol{x} \in X, \boldsymbol{y} \in Y\}$, and it's associated projection operator $\Pi_S(\boldsymbol{z}) = [\Pi_X(\boldsymbol{x}), \Pi_Y(\boldsymbol{y})]^T$. In this setting we can express the EG updates, in a much more concise form as:

$$\begin{aligned} \boldsymbol{z}_{k+1/2} &= \Pi_S(\boldsymbol{z}_k - \eta F(\boldsymbol{z}_k)), \\ \boldsymbol{z}_{k+1} &= \Pi_S(\boldsymbol{z}_k - \eta F(\boldsymbol{z}_{k+1/2})). \end{aligned}$$

Given a fixed point $\boldsymbol{z}^*$, and using the first order characterization of optimality we have that $\forall (\boldsymbol{x}, \boldsymbol{y}) \in X \times Y$:

$$\begin{aligned} \langle \nabla_x f(\boldsymbol{x}^*, \boldsymbol{y}), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0, \\ \langle -\nabla_y f(\boldsymbol{x}, \boldsymbol{y}^*), \boldsymbol{y} - \boldsymbol{y}^* \rangle \geq 0. \end{aligned}$$

which can be expressed more concisely as:

$$\left\langle \begin{bmatrix} \nabla_x f(\boldsymbol{x}^*, \boldsymbol{y}) \\ -\nabla_y f(\boldsymbol{x}, \boldsymbol{y}^*) \end{bmatrix}, \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{x}^* \end{bmatrix} - \begin{bmatrix} \boldsymbol{y}^* \\ \boldsymbol{y}^* \end{bmatrix} \right\rangle \geq 0 \Rightarrow \langle F(\boldsymbol{z}^*), \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq 0.$$

Observe that for ∇f μ-Lipschitz, F is a 2μ-Lipschitz operator.

Proposition 2.3. *For μ-smooth convex-concave functions, where X and Y have max diameter D, expected gradient with $\eta_k = \frac{1}{2\mu}$ gives the following duality gap for the expected steps: the duality gap is bounded by:*

$$\textit{Duality Gap}(\bar{\boldsymbol{z}}_K) \leq \frac{2D^2\mu}{K},$$

where $\bar{\boldsymbol{z}}_K = \frac{1}{K} \sum_{k=1...K} (\boldsymbol{x}_k, \boldsymbol{y}_k)$.

Proof. For some $\boldsymbol{x} \in X$, $\boldsymbol{y} \in Y$, we have:

$$
\begin{aligned}
& f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}) - f(\boldsymbol{x}, \boldsymbol{y}_{k+1/2}) \\
&= f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}) \overbrace{-f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}_{k+1/2}) + f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}_{k+1/2})}^{=0} - f(\boldsymbol{x}, \boldsymbol{y}_{k+1/2}) \\
& \qquad \text{(by convexity/concavity)} \\
&\leq \langle \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}_{k+1/2}), \boldsymbol{y} - \boldsymbol{y}_{k+1/2} \rangle + \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}_{k+1/2}), \boldsymbol{x}_{k+1/2} - \boldsymbol{x} \rangle \\
&= \left\langle \begin{bmatrix} \nabla_x f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}_{k+1/2}) \\ -\nabla_y f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}_{k+1/2}) \end{bmatrix}, \begin{bmatrix} \boldsymbol{x}_{k+1/2} - \boldsymbol{x} \\ \boldsymbol{y}_{k+1/2} - \boldsymbol{y} \end{bmatrix} \right\rangle \\
&\Rightarrow f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}) - f(\boldsymbol{x}, \boldsymbol{y}_{k+1/2}) \leq \langle F(\boldsymbol{z}_{k+1/2}), \boldsymbol{z}_{k+1/2} - \boldsymbol{z} \rangle.
\end{aligned}
$$

This provides us with a way to bound $\langle F(\boldsymbol{z}_{k+1/2}), \boldsymbol{z}_{k+1/2} - \boldsymbol{z} \rangle$ and thus the duality gap, by finding some upper-bound for the right-hand side of the equation, $\langle F(\boldsymbol{z}_{k+1/2}), \boldsymbol{z}_{k+1/2} - \boldsymbol{z} \rangle$, which is what we will do next.

$$
\begin{aligned}
\langle F(\boldsymbol{z}_{k+1/2}), \boldsymbol{z}_{k+1/2} - \boldsymbol{z} \rangle &= \left\langle \frac{\boldsymbol{z}_k - \tilde{\boldsymbol{z}}_{k+1}}{\eta}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z} \right\rangle \\
& \qquad \text{(From the upddates, the tilde means before projection.)} \\
&= \left\langle \frac{\boldsymbol{z}_k - \tilde{\boldsymbol{z}}_{k+1}}{\eta}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1} \right\rangle + \left\langle \frac{\boldsymbol{z}_k - \tilde{\boldsymbol{z}}_{k+1}}{\eta}, \boldsymbol{z}_{k+1} - \boldsymbol{z} \right\rangle \\
& \qquad \text{Adding } 0 = \boldsymbol{z}_{k+1} - \mathbf{z}_{k+1} \\
&\leq \left\langle \frac{\boldsymbol{z}_k - \tilde{\boldsymbol{z}}_{k+1}}{\eta}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1} \right\rangle + \left\langle \frac{\boldsymbol{z}_k - \boldsymbol{z}_{k+1}}{\eta}, \boldsymbol{z}_{k+1} - \boldsymbol{z} \right\rangle \\
& \text{(Using that } \langle \tilde{\boldsymbol{z}}_{k+1} - \boldsymbol{z}_{k+1}, \boldsymbol{z} - \boldsymbol{z}_{k+1} \rangle \Rightarrow \langle -\tilde{\boldsymbol{z}}_{k+1}, \boldsymbol{z} - \boldsymbol{z}_{k+1} \rangle \leq -\boldsymbol{z}_{k+1}, \boldsymbol{z} - \boldsymbol{z}_{k+1} \rangle \\
& \qquad \text{from the properties of the projection operator.)} \\
&= \left\langle \frac{\boldsymbol{z}_k - \tilde{\boldsymbol{z}}_{k+1/2}}{\eta}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1} \right\rangle + \left\langle \frac{\tilde{\boldsymbol{z}}_{k+1/2} - \tilde{\boldsymbol{z}}_{k+1}}{\eta}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1} \right\rangle \\
& \qquad + \left\langle \frac{\boldsymbol{z}_k - \boldsymbol{z}_{k+1}}{\eta}, \boldsymbol{z}_{k+1} - \boldsymbol{z} \right\rangle \\
& \qquad \text{Adding } 0 = \tilde{\boldsymbol{z}}_{k+1/2} - \tilde{\boldsymbol{z}}_{k+1/2} \\
&= \left\langle \frac{\boldsymbol{z}_k - \boldsymbol{z}_{k+1/2}}{\eta}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1} \right\rangle + \left\langle \frac{\tilde{\boldsymbol{z}}_{k+1/2} - \tilde{\boldsymbol{z}}_{k+1}}{\eta}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1} \right\rangle \\
& \qquad + \left\langle \frac{\boldsymbol{z}_k - \boldsymbol{z}_{k+1}}{\eta}, \boldsymbol{z}_{k+1} - \boldsymbol{z} \right\rangle \\
& \qquad \text{(Using that } \langle \tilde{\boldsymbol{z}}_{k+1/2} - \boldsymbol{z}_{k+1/2}, \boldsymbol{z} - \boldsymbol{z}_{k+1/2} \rangle \\
& \qquad \Rightarrow \langle -\tilde{\boldsymbol{z}}_{k+1/2}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z} \rangle \leq -\boldsymbol{z}_{k+1/2}, \boldsymbol{z}_{k+1} - \boldsymbol{z} \rangle \\
& \qquad \text{from the properties of the projection operator.)}
\end{aligned}
$$

We thus have an expression made up of three terms:

$$\langle F(\boldsymbol{z}_{k+1/2}), \boldsymbol{z}_{k+1/2} - \boldsymbol{z}\rangle \leq \left\langle \frac{\boldsymbol{z}_k - \boldsymbol{z}_{k+1/2}}{\eta}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1} \right\rangle +$$
$$\left\langle \frac{\tilde{\boldsymbol{z}}_{k+1/2} - \tilde{\boldsymbol{z}}_{k+1}}{\eta}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1} \right\rangle + \left\langle \frac{\boldsymbol{z}_k - \boldsymbol{z}_{k+1}}{\eta}, \boldsymbol{z}_{k+1} - \boldsymbol{z} \right\rangle$$
$$\Rightarrow \eta\langle F(\boldsymbol{z}_{k+1/2}), \boldsymbol{z}_{k+1/2} - \boldsymbol{z}\rangle \leq \overbrace{\langle \boldsymbol{z}_k - \boldsymbol{z}_{k+1/2}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\rangle}^{(a)} +$$
$$\overbrace{\langle \tilde{\boldsymbol{z}}_{k+1/2} - \tilde{\boldsymbol{z}}_{k+1}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\rangle}^{(b)} + \overbrace{\langle \boldsymbol{z}_k - \boldsymbol{z}_{k+1}, \boldsymbol{z}_{k+1} - \boldsymbol{z}\rangle}^{(c)}.$$

We will bound these terms individually, starting with (b):

$$\begin{aligned}
&\langle \tilde{\boldsymbol{z}}_{k+1/2} - \tilde{\boldsymbol{z}}_{k+1}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\rangle \\
&= \langle \tilde{\boldsymbol{z}}_{k+1/2} - \tilde{\boldsymbol{z}}_k, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\rangle + \langle \boldsymbol{z}_k - \tilde{\boldsymbol{z}}_{k+1}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\rangle && \text{(add } 0 = \boldsymbol{z}_k - \boldsymbol{z}_k\text{)} \\
&= \eta\langle F(\boldsymbol{z}_{k+1/2}) - F(\boldsymbol{z}_k), \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\rangle && \text{(from the updates)} \\
&\leq \eta\|F(\boldsymbol{z}_{k+1/2}) - F(\boldsymbol{z}_k)\| \cdot \|\boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\| && \text{(Cauchy Schwartz)} \\
&\leq 2\eta\mu\|\boldsymbol{z}_{k+1/2} - \boldsymbol{z}_k\| \cdot \|\boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\| && (F \text{ is } 2\mu\text{-Lipschitz}) \\
&\leq \frac{1}{2}(4\eta^2\mu^2\|\boldsymbol{z}_{k+1/2} - \boldsymbol{z}_k\|^2 \cdot \|\boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\|^2) && \text{(Young's inequality)}
\end{aligned}$$

Next, for (a) and (c) we will use that $\langle a, b\rangle = \frac{\|a+b\|^2 - \|a\|^2 - \|b\|^2}{2}$. For (a) we get:

$$\begin{aligned}
&\langle \boldsymbol{z}_k - \boldsymbol{z}_{k+1/2}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\rangle \\
&= \frac{1}{2}\Big[\|\boldsymbol{z}_k - \boldsymbol{z}_{k+1}\|^2 - \|\boldsymbol{z}_k - \boldsymbol{z}_{k+1/2}\|^2 - \|\boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\|^2\Big],
\end{aligned}$$

and for (c) we have:

$$\begin{aligned}
&\langle \boldsymbol{z}_k - \boldsymbol{z}_{k+1}, \boldsymbol{z}_{k+1} - \boldsymbol{z}\rangle \\
&= \frac{1}{2}\Big[\|\boldsymbol{z}_k - \boldsymbol{z}\|^2 - \|\boldsymbol{z}_k - \boldsymbol{z}_{k+1}\|^2 - \|\boldsymbol{z}_{k+1} - \boldsymbol{z}\|^2\Big].
\end{aligned}$$

Putting the three simplified terms together we have:

$$\begin{aligned}
&2\langle \tilde{\boldsymbol{z}}_{k+1/2} - \tilde{\boldsymbol{z}}_{k+1}, \boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\rangle \\
&\leq \Big[\|\boldsymbol{z}_k - \boldsymbol{z}_{k+1}\|^2 - \|\boldsymbol{z}_k - \boldsymbol{z}_{k+1/2}\|^2 - \|\boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\|\Big] \\
&\quad + (4\eta^2\mu^2\|\boldsymbol{z}_{k+1/2} - \boldsymbol{z}_k\|^2 \cdot \|\boldsymbol{z}_{k+1/2} - \boldsymbol{z}_{k+1}\|^2) \\
&\quad + \Big[\|\boldsymbol{z}_k - \boldsymbol{z}\|^2 - \|\boldsymbol{z}_k - \boldsymbol{z}_{k+1}\| - \|\boldsymbol{z}_{k+1} - \boldsymbol{z}\|\Big] \\
&= \|\boldsymbol{z}_k - \boldsymbol{z}_{k+1/2}\|^2(4\mu^2\eta^2 - 1) + \|\boldsymbol{z}_k - \boldsymbol{z}\|^2 - \|\boldsymbol{z}_{k+1} - \boldsymbol{z}\|^2.
\end{aligned}$$

Which implies:

$$\begin{aligned}
&\eta\langle F(\boldsymbol{z}_{k+1/2}), \boldsymbol{z}_{k+1/2} - \boldsymbol{z}\rangle \\
&\leq \frac{1}{2}\Big[\|\boldsymbol{z}_k - \boldsymbol{z}_{k+1/2}\|^2(4\mu^2\eta^2 - 1) + \|\boldsymbol{z}_k - \boldsymbol{z}\|^2 - \|\boldsymbol{z}_{k+1} - \boldsymbol{z}\|^2.\Big]
\end{aligned}$$

Setting, $\eta = \frac{1}{2\mu}$ (optimizing η), we have:

$$\langle F(\boldsymbol{z}_{k+1/2}), \boldsymbol{z}_{k+1/2} - \boldsymbol{z} \rangle \leq \mu \Big[\|\boldsymbol{z}_k - \boldsymbol{z}\|^2 - \|\boldsymbol{z}_{k+1} - \boldsymbol{z}\|^2 \Big].$$

Summing over all ks we thus get:

$$\begin{aligned} \sum_{k=1}^{K} \langle F(\boldsymbol{z}_{k+1/2}), \boldsymbol{z}_{k+1/2} - \boldsymbol{z} \rangle &\leq \sum_{k=1}^{K} \mu \Big[\|\boldsymbol{z}_k - \boldsymbol{z}\|^2 - \|\boldsymbol{z}_{k+1} - \boldsymbol{z}\|^2 \Big] \\ &\leq \mu \|\boldsymbol{z}_1 - \boldsymbol{z}\|^2 \leq 2D^2 \mu. \end{aligned}$$

Now going back to the "top" of the proof, we get our final bound:

$$\frac{1}{K} \sum_{k=1}^{K} f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}) - f(\boldsymbol{x}, \boldsymbol{y}_{k+1/2}) \leq \frac{\mu \|\boldsymbol{z}_1 - \boldsymbol{z}\|^2 \leq 2D^2 \mu}{K}.$$

Which, since the functions are convex and we can use Jensen's inequality leads to :

$$f(\bar{\boldsymbol{x}}_K, \boldsymbol{y}) - f(\boldsymbol{x}, \bar{\boldsymbol{y}}_K) \leq \frac{\mu \|\boldsymbol{z}_1 - \boldsymbol{z}\|^2 \leq 2D^2 \mu}{K}.$$

Which bounds the duality gap and completes the proof. □

2.4 Optimistic Gradient Descent Ascent

Here we consider the Optimistic Gradient Descent-Ascent method (OG) (sometimes also refered to as Past-Extra Gradient), which provides an alternate approximation of the (implicit) proximal point method. It displays better performance compared to the naïve GDA approach (from section 2.2).

Algorithm 3: (Projected) Optimistic Gradient Descent Ascent

Set the learning rate $\eta > 0$
Initialize the algorithm at some point $(\boldsymbol{x}_0, \boldsymbol{y}_0)$
Initialize the algorithm at some point $(\boldsymbol{x}_0, \boldsymbol{y}_0)$
foreach *iteration* $k = 0, 2, ..., K-1$ **do**
  $\boldsymbol{x}_{k+1} \leftarrow \Pi_X \big(\boldsymbol{x}_k - 2\eta \nabla_x f(x_k, y_k) + \eta \nabla_x f(x_{k-1}, y_{k-1}) \big)$
  $\boldsymbol{y}_{k+1} \leftarrow \Pi_Y \big(\boldsymbol{y}_k + 2\eta \nabla_y f(x_k, y_k) - \eta \nabla_y f(x_{k-1}, y_{k-1}) \big)$
end
return $(\boldsymbol{x}_N, \boldsymbol{y}_N)$

2.5 Summary of Known Results

Below we provide a summary of known results for convergence of gradient-based methods on saddle point problems. Table 1 gives a summary of convergence rates for Lipschitz, smooth convex-concave functions under convex constraints (which roughly corresponds to the occupancy measure formulation of the IRL problem described in section 1.3).

Method	Average-iterate convergence rate	Last-iterate convergence rate
GDA (alg. 1)	$O(1/\sqrt{K})$ (proof of prop.2.2)	No conv. guarantees ($\exists$ limit cycles).
EG (alg. 2)	$O(1/K)$ (proof of prop.2.3)	$O(1/\sqrt{K})$ (Cai et al. 2022 Cai, Oikonomou, and Zheng 2022)
OG (alg. 3)	$O(1/K)$ (Mokhtari et al. 2020 Aryan, E., and Sarath 2020)	$O(1/\sqrt{K})$ (Gorbunov et al. 2022) Gorbunov, Taylor, and Gidel 2022

Table 1: Convergence results for minimax optimization methods with convex constraints in the general (smooth) convex-concave setting.

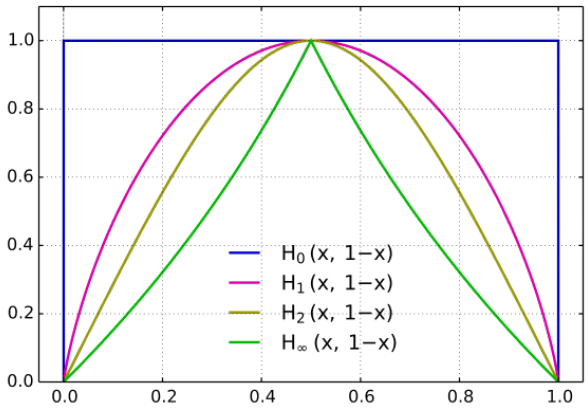

Figure 3: A comparison of Rényi entropies of different orders, plotted as they evolve on a 2-dimensional probability vector. Here we use the result that $\lim \alpha \to 1 H_\alpha = H$ and let the Shannon entropy be denoted by the "first-order Rényi entropy", which is here denoted as H_1.

3 Properties of regularizers

3.1 Shannon Entropy

First, we consider the most common measure of information (or "surpise"), the Shannon Entropy.

Definition 3.1. *(**Shannon Entropy**) For a discrete random variable X, the Shannon Entropy is defined as:*

$$H(X) = -\sum_{i=1}^{n} p_i \log(p_i).$$

Observation 3.1. *(**Properties of the Shannon Entropy**) The Shannon entropy has the nice property of being a (strictly) concave function, but regardless, it poses several problems for first-order optimization methods. One big problem is that H is not defined when a probability in the distribution goes to 0 (as $\log 0$ does not exist), as $\lim_{x\to 0} x \log x = 0$ we can for convenience re-define H as :*

$$H(X) := \begin{cases} -\sum_{i=1}^{n} p_i \log(p_i) & \text{if } p_1 > 0 \forall i \in [n], \\ 0 & \text{otherwise.} \end{cases}$$

This approach "plugs" the hole in the Shannon entropy but doesn't fix all of the problems associated with. Specifically, the gradient of H, which is expressed as:

$$[\nabla H(X)]_i = -\log p_i - 1,$$

is neither Lipschitz nor bounded, this is easily verifiable by observing that since $\lim_{x\to 0a} \log(x) = -\infty$*, when we pick a probability vector* $\boldsymbol{p}$*, s.t. for which some element* $p_i \to 0$ *we have that:* $\|\nabla H(\boldsymbol{p})\| \to +\infty$*. This makes* H *violate an essential requirement for convergence of first-order optimization methods. (The gradient does have the big advantage of being separable coordinate, by coordinate, which will prove useful later.)*

3.1.1 Shannon Shenanigans: Lipschitzness of the Shannon Entropy over $\Delta_n^{(\rho)}$

The following section is sort of a detour around the Shannon Entropy properties. The fact that H is non Lipschitz when some probability in the distribution goes to 0 motivates the study of the properties of H on a restricted domain, for which we guarantee that no probability p_i ever goes to close to 0, for this reason we define the ρ-non-vanishing simplex of dimension n: $\Delta_n^{(\rho)}$.

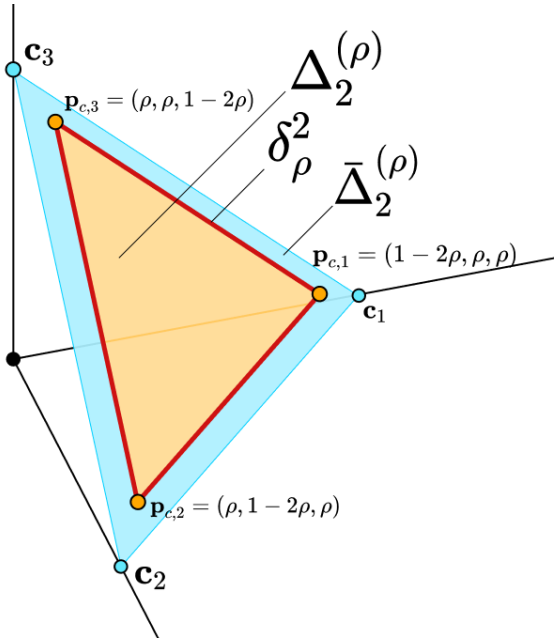

Figure 4: An illustration of the 2 dimensional ρ non-vanishing simplex $\Delta_n^{(\rho)}$ (the orange inner-surface, on the plot). We have that $\Delta_n^{(\rho)} \subset \Delta^2$ (the non-vanishing simplex is a subset of the simplex) and we denote on the plot the area of the simplex Δ^2 that is excluded by $\Delta_n^{(\rho)}$ (i.e. $\bar{\Delta}_\rho^2 = \Delta^2 \setminus \Delta_n^{(\rho)}$) in blue. The plot also labels the corners $\boldsymbol{c}_i$ of the simplex, those $\boldsymbol{p}_{c,i}$ of the non-vanishing simplex, as well as the border $\delta_2^{(\rho)}$ of $\Delta_n^{(\rho)}$ (the red line surrounding the orange area).

Definition 3.2. ***(ρ-non-vanishing Simplex)*** *We define "ρ-non-vanishing simplex"* $\Delta_n^{(\rho)}$*, for some* $0 < \rho < 1/(n+1)$ *as follows :*

$$\Delta_n^{(\rho)} := \Big\{ p \in \Delta \subset \mathbb{R}^n;\ [p]_i > \rho,\ \forall i \in [n+1] \Big\},$$

or equivalently as a convex combination over "corners" $\boldsymbol{p}_{c,i}$ of the ρ non-vanishing simplex:

$$\Delta_n^{(\rho)} = \Big\{ \sum_{i\in[n+1]} \eta_i \boldsymbol{p}_{c,i}, \quad \sum_{i\in[n+1]} \eta_i = 1\},$$

(which will be useful when considering its border). (For a better intuition of what these spaces and objects are, refer to fig. 4) Note that the corner $\boldsymbol{p}_{c,i}$ associated with the i-th axes of the $n+1$ dimensional space in which $\Delta_n^{(\rho)}$ is contained has the following coordinates:

$$[\boldsymbol{p}_{c,i}]_j \begin{cases} 1-(n+1)\cdot\rho & \text{if } j = i, \\ \rho & \text{otherwise.} \end{cases}$$

Which gives a hint of why we need to upper bound ρ by $1/(n+1)$. We define the border $\delta_n^{(\rho)}$ of $\Delta_n^{(\rho)}$ as the union of a set of $n+1$ subspaces ("edges") $\boldsymbol{E}^{(i)}$:

$$\delta_n^{(\rho)} = \bigcup_{i\in[n+1]} \boldsymbol{E}^{(i)},$$

for which each edge $\boldsymbol{E}_\rho^{(i)}$ is defined as a convex combination over all the "corners" of $\Delta_n^{(\rho)}$, excluding the $p_{c,i}$ corner:

$$\boldsymbol{E}_\rho^{(i)} = \Big\{ \sum_{j\in[n+1]\setminus i} \eta_j p_{c,j}, \quad \sum_{j\in[n+1]\setminus i} \eta_i = 1\}.$$

For convenience of notation we also write points on the edge $\boldsymbol{E}_\rho^{(i)}$ as a function of a parameter vector $\boldsymbol{\eta} \in \Delta^{n-1}$. In that setting we write a point on the edge as:

$$\boldsymbol{E}_\rho^{(i)}(\boldsymbol{\eta}) = F_E^{(i)}\boldsymbol{\eta} = \sum_{j\in[n+1]\setminus i} \eta_j \boldsymbol{p}_{c,j}.$$

Note that each edge is thus a convex hull on some set of corners $\boldsymbol{P}_c^{(i)} = \boldsymbol{P}_c \setminus \boldsymbol{p}_{c,i}$ of cardinality n (where $\boldsymbol{P}_c$ is the set of all corners). An illustration making this more intuitive can be found in fig 5.

Proposition 3.1. *(H **is** $\sqrt{m}(\log(\rho^{-1})-1)$-**Lipschitz over** $\Delta_n^{(\rho)}$).*

Proof. This proof is quite direct. We just bound the gradient norm by:

$$\begin{aligned} \|\nabla H(p)\|_2 &= \sqrt{[\nabla H(p)]_1^2 + [\nabla H(p)]_2^2 + [\nabla H(p)]_{n+1}^2} \\ &\leq \sqrt{m} \sup_{\mathbf{p}\in\Delta_n,\ i\in[n+1]} |[\nabla H(p)]_{n+1}| \\ &= \sqrt{m}(-\log\rho - 1) \end{aligned}$$

□

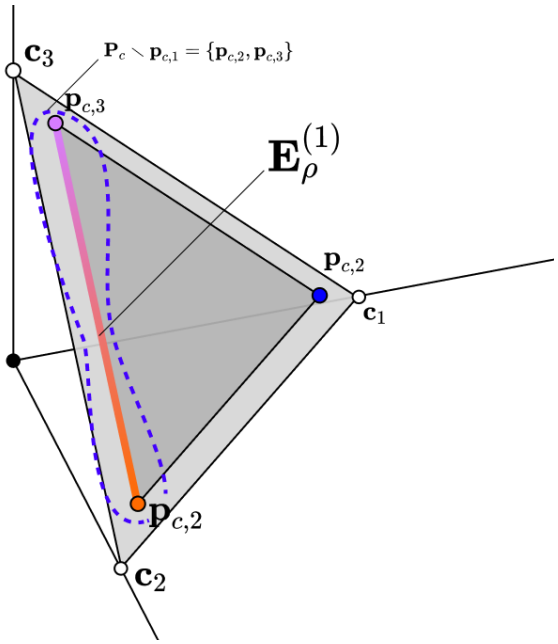

Figure 5: An illustration of how each edge of $\Delta_n^{(\rho)}$ is a convex-hull over a subset of $\Delta_n^{(\rho)}$'s corners.

3.2 Rényi Entropy

An alternative, and more general, in a sense, definition of Entropy is that of the Rényi Entropy. It defines a large class of entropy functions with various properties.

Definition 3.3. *(**Rényi Entropy**) For a discrete random variable X, the Rényi Entropy of order α is defined as:*

$$H_\alpha(X) = \frac{1}{1-\alpha} \log\Big(\sum_{i=1}^{n} p_i^\alpha\Big).$$

Conveniently, it can also be denoted as a function of the L_α norm of the vector of probabilities $\boldsymbol{p}_X$ associated with the random variable X:

$$H_\alpha(X) = H_\alpha(\boldsymbol{p}_X) = \frac{\alpha}{1-\alpha} \log \|\boldsymbol{p}_X\|_\alpha.$$

Observation 3.2. *(**Gradient Properties of the Rényi Entropy of order 2**) The gradient of the Rényi entropy of order 2 (H_2) is simply given by:*

$$\nabla_{\boldsymbol{p}_X} H_2(\boldsymbol{p}_X) = -2\frac{\boldsymbol{p}_X}{\|\boldsymbol{p}_X\|_2^2}.$$

And observing that, on the simplex Δ_n of dimensionality n we have:

$$\begin{aligned}\|\nabla_{\boldsymbol{p}_X} H_2(\boldsymbol{p}_X)\| &= \left\| -2\frac{\boldsymbol{p}_X}{\|\boldsymbol{p}_X\|_2^2} \right\| \\ &= 2\frac{1}{\|\boldsymbol{p}_X\|_2} \leq 2\sqrt{n},\ \ \boldsymbol{p}_X \in \Delta_n,\end{aligned}$$

it is clear that the Rényi entropy is $2\sqrt{n}$-Lipschitz. Furthermore, when computing the eigenvalues the Hessian of H_2, we get:

$$\lambda_{1,2} = \frac{\pm 2}{\|\boldsymbol{p}_X\|_2^2}$$

which, on the Δ_n simplex, is clearly bounded by:

$$\lambda_{1,2} = \frac{\pm 2}{\|\boldsymbol{p}_X\|_2^2} \leq \pm 2n.$$

Which shows H_2 to also be $2n$-smooth.

4 Constrained and Unconstrained Bandits, Iteration Complexity for different Regularizers

In order to work towards a better understanding of the IRL problems at hand we first restrict ourselves to the study of a single-state MDP, which is equivalent to a *K-Armed Bandit Problem*.

4.1 A few observations about Bandit problems

The "*bandit*" (a.k.a single state MDP has a few interesting properties that make the analysis of optimization algorithms on it simpler). In the bandit setting the policy $\boldsymbol{\pi} \in \Delta_A$ becomes simply a distribution on the action-set rather than a function form the states to distributions on the actions-set, we can thus represent $\boldsymbol{\pi}$ as a vector in $\mathbb{R}^n$

Observation 4.1. *(**Policy-Occupancy Measure Equality**) In the specific case of the bandit setting, we have that:*

$$\boldsymbol{\mu}^{\pi} = \boldsymbol{\pi}.$$

This can easily be verified as follows:

$$\begin{aligned}
\mu^{\pi}(a) &= (1-\gamma) \lim_{T \to +\infty} \sum_{t=0}^{T} \gamma^t \mathbb{P}_{\nu}^{\pi}(a_t = a) \\
&= (1-\gamma) \lim_{T \to +\infty} \sum_{t=0}^{T} \gamma^t \pi(a_t = a) \\
&= \pi(a).
\end{aligned}$$

Observation 4.2. *(**Objective Function in the Bandit Setting**) In the specific case of the bandit setting, we have that the objective function reduces to:*

$$\begin{aligned}
J^{\nu}(\pi) &= \langle \boldsymbol{r}, \boldsymbol{\mu}^{\pi} \rangle - \tilde{\Omega}(\boldsymbol{\pi}) \\
&= \langle \boldsymbol{r}, \boldsymbol{\pi} \rangle - \Omega(\boldsymbol{\pi}).
\end{aligned}$$

Observation 4.3. *In the bandit setting, the Bellman Flow constraints are always trivially satisfied (the probability that the agent ends up in state s at time $t+1$ is always 1).*

Acceptable policies are defined by the feasible set (as for CRL), which is here given by:

$$\mathcal{F} := \left\{ \boldsymbol{\pi} \in \Delta_A, \Psi \boldsymbol{\pi} \leq \boldsymbol{b} \right\},$$

where Ψ is the cost matrix and $\boldsymbol{b}$ the constraint vector (see section 1.2 for details).

Definition 4.1. *(**Constrained Bandit Solution Map**) we define the constrained bandit solution map $CB : \mathcal{R} \to \Pi$ as:*

$$CB(\boldsymbol{r}) = \arg\max_{\pi \in \mathcal{F}} J(\boldsymbol{\pi}, \boldsymbol{r})$$

(note that here $\mathcal{F} \subseteq \Pi = \mathcal{M} = \Delta_A$).

4.2 Constrained Inverse bandit (CIB)

Definition 4.2. *(**CIRL Solution Map**) The **goal of CIB** is to find some mapping* $CIRL_\pi : \Pi \to \mathcal{R}$.

$$(CB_\pi \circ CIB_\pi)(\boldsymbol{\pi}^E) = \boldsymbol{\pi}^E.$$

We consider the following constraint optimization problem:

$$\min_{\boldsymbol{r}\in\mathcal{R}} \max_{\boldsymbol{\pi}\in\mathcal{F}} \langle \boldsymbol{r}, \boldsymbol{\pi} - \boldsymbol{\pi}^E \rangle - \Omega(\boldsymbol{\pi}), \tag{4.1}$$

which provides a way of evaluating 4.2 (this can be shown to be true by proposition 1.1). The Lagrangian dual of 4.1, found by relaxing the feasibility constraints is given by:

$$\min_{\boldsymbol{r}\in\mathcal{R},\boldsymbol{\lambda}\geq 0} \max_{\boldsymbol{\pi}\in\Delta_A} f_{\text{CIB}}(\boldsymbol{r}, \boldsymbol{\lambda}, \boldsymbol{\pi}). \tag{4.2}$$

where $f_{\text{CIB}}(\boldsymbol{r}, \boldsymbol{\lambda}, \boldsymbol{\pi}) = \langle \boldsymbol{r}, \boldsymbol{\pi} - \boldsymbol{\pi}^E \rangle - \Omega(\boldsymbol{\pi}) + \langle \boldsymbol{\lambda}, \boldsymbol{b} - \Psi\boldsymbol{\pi} \rangle$ is our objective function. From proposition 1.2 we know that solving problem 4.2 is equivalent to solving 4.1. We will study convergence of extra-gradient-based saddle-point algorithms on 4.2.

4.2.1 Solving the Shannon Entropy regularized problem with an extra-gradient approach: EG-COP

The most common regularizer used in practice for IRL is the Shannon entropy (see definition 3.1), the use of this particular regularizer poses some problems as it does not have Lipschitz gradients on the domain Δ_A (see discussion in section 3.1). Therefore if we want to show convergence we have to use some trick to ensure gradients are indeed Lipschitz. To go around this limitation we suggest to project the gradients in the ρ-non-vanishing simplex $\Delta_A^{(\rho)} \subset \Delta_A$, on which we can show that our objective function is indeed Lipschitz. In order to more specifically define our problem let us pick a reward class $\mathcal{R}$: we choose our reward class $\mathcal{R}_{L_1}$ to be the L_1 ball centered on the origin, as it closely matches the analysis of identifiability made in Schlaginhaufen 2023. Note that reward class can be easily substituted for another convex reward class without changing the analysis very much.

We thus consider the following problem:

$$\min_{\boldsymbol{r}\in\mathcal{R},\boldsymbol{\lambda}\geq 0} \max_{\boldsymbol{\pi}\in\Delta_A^{(\rho)}} f_{\text{CIB-H}}(\boldsymbol{r}, \boldsymbol{\lambda}, \boldsymbol{\pi}). \tag{4.3}$$

where $f_{\text{CIB-H}}(\boldsymbol{r}, \boldsymbol{\lambda}, \boldsymbol{\pi}) = \langle \boldsymbol{r}, \boldsymbol{\pi} - \boldsymbol{\pi}^E \rangle - H(\boldsymbol{\pi}) + \langle \boldsymbol{\lambda}, \boldsymbol{b} - \Psi\boldsymbol{\pi} \rangle$ is our objective function (the dual Lagrangian of the problem 4.1). For our result to hold, we need the following assumption:

Assumption 4.1. *(**Sufficient Slater's Condition**) we assume that* $\exists \chi > 0$ *and* $\pi = \Delta_A$ *s.t.*

$$\boldsymbol{b} - \Psi^T\boldsymbol{\pi} \geq \chi.$$

In order to get a convergence result we need to show that, under our assumptions:

1. The decision variables exist on a domain with a finite bounded diameter (specifically we need to show this for the Lagrange multipliers $\boldsymbol{\lambda}$ which are apparently unbounded in the original problem statement).

2. $\pi^* \in \Delta_A^{(\rho)}$, i.e. the optimal policy lies in our restricted policy set,

3. $f_{\text{CIB-H}}$ is smooth over the domain $\Delta_A^{(\rho)}$,

Proposition 4.1. *Our Lagrange multiplier vector is contained in a box:* $\boldsymbol{\lambda} \in \mathcal{B}$, *where*

$$\mathcal{B} := \left\{ \boldsymbol{\lambda} \in \mathbb{R}^d : 0 \leq [\boldsymbol{\lambda}]_i \leq \frac{2(R + \beta \log m)}{\chi(1-\gamma)} \forall i \right\}.$$

Proof. Let $\boldsymbol{Z}_a(\boldsymbol{r}) := \{\boldsymbol{\lambda} \geq 0 : \max_{\boldsymbol{\pi}} f_{\text{CIB-H}}(\boldsymbol{r}, \boldsymbol{\lambda}, \boldsymbol{\pi}) \leq a\}$ be the sublevel set of the dual function $\max_{\boldsymbol{\pi}} f_{\text{CIB-H}}(\boldsymbol{r}, \boldsymbol{\lambda}, \boldsymbol{\pi})$ for any $a \in \mathbb{R}$, the for any $\boldsymbol{\lambda} \in \boldsymbol{Z}_a$, $\boldsymbol{r} \in \mathcal{R}$ we have:

$$a \geq \max_{\boldsymbol{\pi}} f_{\text{CIB-H}}(\boldsymbol{r}, \boldsymbol{\lambda}, \bar{\boldsymbol{\pi}}) \geq J(\bar{\boldsymbol{\pi}}, \boldsymbol{r}) + \lambda^T (b - \Psi^T) \geq \langle \bar{\boldsymbol{\pi}}, \boldsymbol{r} \rangle + \chi \lambda^T \mathbf{1}.$$

From which we deduce $[\boldsymbol{\lambda}]_i \leq \frac{a - J(\bar{\boldsymbol{\pi}}, \boldsymbol{r})}{\chi}$, choosing $a = J(\boldsymbol{\pi}^*, \boldsymbol{r}^*)$ and using that $J(\boldsymbol{\pi}, \boldsymbol{r}) \geq \frac{-R - \beta \log m}{1-\gamma}$ we get our result for the upper bound. For the lower bound we just use that $\boldsymbol{\lambda} \geq 0$. □

Proposition 4.2. *The optimal policy* $\boldsymbol{\pi}^*$ *lies in the* ρ*-non-vanishing simplex* $\Delta_A^{(\rho)}$ *with parameter* $\rho = \frac{1}{m} \exp\left(\frac{-2(R + \beta \log(m))(2\|\Psi\| - \chi\gamma + \chi)}{\chi(1-\gamma)^2} \right)$.

Proof. Recall that the optimal policy (for the unconstrained MDP) is given in terms of the optimal Q-values Q^* by :

$$\pi^*(a) = \frac{\exp(Q^*(a)/\beta)}{\sum_{a' \in A} \exp(Q^*(a')/\beta)}. \tag{A}$$

Using from proposition 1.2 that $\text{CRL}_\pi(\boldsymbol{r}) = \text{RL}_\pi(\boldsymbol{r} - \boldsymbol{\Psi}\boldsymbol{\lambda}^*)$ we will use that result (on the modified reward function $\tilde{r}$) to compute our bound. We will start by bounding Q^*:

$$\begin{aligned} Q^*(a) &= \tilde{r}(a) + \gamma \mathbb{E}_{s \sim s_0}\left[\tilde{V}^*(s)\right] = r(a) + [\boldsymbol{\Psi}\boldsymbol{\lambda}^*]_a + \tilde{V}^*(s) \\ &= r(a) + [\boldsymbol{\Psi}\boldsymbol{\lambda}^*]_a + \gamma \max_{\boldsymbol{\pi} \in \Delta_A} \mathbb{E}_{s \sim s_0}\Big[\sum_{t=0}^{+\infty} \gamma^t (r(a) + [\boldsymbol{\Psi}\boldsymbol{\lambda}^*]_a + \beta H(\boldsymbol{\pi})) \Big]. \end{aligned}$$

We now let $\alpha = \frac{R+\beta \log m}{1-\gamma}$ and claim that $-\alpha \leq Q(a) \leq \alpha$, this can be verified by:

$$\begin{aligned}
Q^*(a) &\leq \max_{\boldsymbol{\pi} \in \Delta_A} \left[\sum_{t=0}^{+\infty} \gamma^t \big(r(a) + [\boldsymbol{\Psi} \boldsymbol{\lambda}^*]_a + \beta H(\boldsymbol{\pi}) \big) \right] \\
&\leq \Big(\sum_{t=0}^{+\infty} \gamma^t \Big) \Big(R + \|\Psi\| \sup_{\boldsymbol{\lambda}^*} \|\boldsymbol{\lambda}^*\|_\infty + \beta \log m \Big) \\
&= \frac{(R + \beta \log(m))\,(2\|\Psi\| - \chi\gamma + \chi)}{\chi\,(1-\gamma)^2} = \alpha, \\
Q^*(a) &\geq \min_{\boldsymbol{\pi} \in \Delta_A} \left[\sum_{t=0}^{+\infty} \gamma^t \big(r(a) + [\boldsymbol{\Psi} \boldsymbol{\lambda}^*]_a + \beta H(\boldsymbol{\pi}) \big) \right] \\
&\geq \Big(\sum_{t=0}^{+\infty} \gamma^t \Big) \Big(- R \Big) \\
&\geq \Big(\sum_{t=0}^{+\infty} \gamma^t \Big) \Big(- R - \beta \log m \Big) \geq -\alpha.
\end{aligned}$$

Thus using (A) we have that:

$$\pi^*(a) \geq \frac{\exp(-\alpha)}{\sum'_a \exp(\alpha)} = \frac{\exp -2\alpha}{m} = \frac{1}{m} \exp\left(\frac{-2\,(R + \beta \log(m))\,(2\|\Psi\| - \chi\gamma + \chi)}{\chi\,(1-\gamma)^2} \right).$$

Which completes the proof (by definition of the the ρ-non-vanishing simplex). □

Proposition 4.3. *$f_{CIB\text{-}H}$ is smooth with constant $L = m \exp\left(\frac{2(R+\beta \log(m))(2\|\Psi\| - \chi\gamma + \chi)}{\chi(1-\gamma)^2} \right)$.*

Proof. We look at $\nabla^2 f_{\text{CIB-H}}$ the hessian of the objective function. Direct computation shows that it is a diagonal matrix of the form :

$$[\nabla^2 f(\boldsymbol{r}, \boldsymbol{\lambda}, \boldsymbol{\pi})]_{i,j} = \begin{cases} \frac{1}{p_i} & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

The diagonal form of $\nabla^2 f(\boldsymbol{r}, \boldsymbol{\lambda}, \boldsymbol{\pi})$ makes it trivial to bound it's spectral norm (the spectral norm is whatever element of the diagonal is maximize). It is thus just a matter of bounding $\frac{1}{p_i}$ which we can do using proposition 4.2:

$$\|\nabla^2 f\| \leq \max \operatorname{eig}\Big(\nabla^2 f \Big) \leq m \exp\left(\frac{2\,(R + \beta \log(m))\,(2\|\Psi\| - \chi\gamma + \chi)}{\chi\,(1-\gamma)^2} \right).$$

□

Now we tackle showing that the diameter is bounded, since our reward class is a ball and since our policies are constrained to the ρ-non-vanishing simplex, this is only a matter of showing that the Lagrange multipliers $\boldsymbol{\lambda}$ are bounded to some box. Which is what we do in the next proposition.

At this point we are ready to describe the algorithm that we will use to solve the CIRL problem. It is simply a specific application of projected extra-gradient GDA (see alg. 2 in section 2.3) to the objective function $f_{\text{CIB-H}}$ where we project on the domain $\mathcal{R}_{L_1} \times \mathcal{B} \times \Delta_A^{(\rho)}$. Theorem 4.1 proves that:

1. the algorithm converges in finite time, with a $O(1/\epsilon)$ rate,

2. the algorithm recovers the exact solution (up to an error ϵ) if the problem is identifiable.

Algorithm 4: EG-COP: Extra-gradient constrained inverse bandit algorithm

Set the learning rate $\eta > 0$
Initialize the algorithm at some point $(\boldsymbol{r}_0, \boldsymbol{\lambda}_0, \boldsymbol{\pi}_0)$
foreach *iteration* $k = 0, 2, ..., K-1$ **do**
$\boldsymbol{r}_{k+1/2} \leftarrow \Pi_{\mathcal{R}_{L_1}}\big(\boldsymbol{r}_k - \eta \nabla_{\boldsymbol{r}} f_{\text{CIB-H}}(\boldsymbol{r}_k, \boldsymbol{\lambda}_k, \boldsymbol{\pi}_k)\big)$
$\boldsymbol{\lambda}_{k+1/2} \leftarrow \Pi_{\boldsymbol{\lambda} \in \mathcal{B}}\big(\boldsymbol{\lambda}_k - \eta \nabla_{\boldsymbol{\lambda}} f_{\text{CIB-H}}(\boldsymbol{r}_k, \boldsymbol{\lambda}_k, \boldsymbol{\pi}_k)\big)$
$\boldsymbol{\pi}_{k+1/2} \leftarrow \Pi_{\Delta_A^{(\rho)}}\big(\boldsymbol{\pi}_k + \eta \nabla_{\boldsymbol{\pi}} f_{\text{CIB-H}}(\boldsymbol{r}_k, \boldsymbol{\lambda}_k, \boldsymbol{\pi}_k)\big)$
$\boldsymbol{r}_{k+1} \leftarrow \Pi_{\mathcal{R}_{L_1}}\big(\boldsymbol{r}_k - \eta \nabla_{\boldsymbol{r}} f_{\text{CIB-H}}(\boldsymbol{r}_{k+1/2}, \boldsymbol{\lambda}_{k+1/2}, \boldsymbol{\pi}_{k+1/2})\big)$
$\boldsymbol{\lambda}_{k+1} \leftarrow \Pi_{\boldsymbol{\lambda} \in \mathcal{B}}\big(\boldsymbol{\lambda}_k - \eta \nabla_{\boldsymbol{\lambda}} f_{\text{CIB-H}}(\boldsymbol{r}_{k+1/2}, \boldsymbol{\lambda}_{k+1/2}, \boldsymbol{\pi}_{k+1/2})\big)$
$\boldsymbol{\pi}_{k+1} \leftarrow \Pi_{\Delta_A^{(\rho)}}\big(\boldsymbol{\pi}_k + \eta \nabla_{\boldsymbol{\pi}} f_{\text{CIB-H}}(\boldsymbol{r}_{k+1/2}, \boldsymbol{\lambda}_{k+1/2}, \boldsymbol{\pi}_{k+1/2})\big)$
end
return $(\hat{\boldsymbol{r}}_K, \hat{\boldsymbol{\lambda}}_K, \hat{\boldsymbol{\pi}}_K)$, where $\hat{\cdot}$ denotes the empirical mean over a sequence.

Theorem 4.1. *Algorithm 4 recovers the optimal reward and policy (up to reward shaping transformations), up to approximation error ϵ in time $O(1/K)$ (where K is the number of iterations). More specifically, assuming we choose learning rate $\eta = \frac{1}{2m} \exp\left(\frac{-2(R+\beta \log(m))(2\|\Psi\|-\chi\gamma+\chi)}{\chi(1-\gamma)^2}\right)$, we have:*

$$DG(\boldsymbol{r}_k, \boldsymbol{\lambda}_k, \boldsymbol{\pi}_k) \leq C_{\textit{EG-COP}} \frac{1}{K},$$

where:

$$C_{\textit{EG-COP}} = 2mD^2 \exp\left(\frac{2\left(R + \beta \log(m)\right)\left(2\|\Psi\| - \chi\gamma + \chi\right)}{\chi\left(1-\gamma\right)^2}\right).$$

In which:

$$D = \max\left\{\sqrt{m}, 2R + \frac{2(R + \beta \log m)}{\chi(1-\gamma)}\right\}.$$

Proof. The proof follows the analysis of alg. 2, which we do not re-state for brevity, so we just have to verify that the assumptions are satisfied:

1. we note that the function $f_{\text{CIB-H}}$ is concave-convex in $(\boldsymbol{r}, \boldsymbol{\lambda}, \boldsymbol{\pi})$ (where we just stack $\boldsymbol{r}$ and $\boldsymbol{\lambda}$ to have a saddle-point formulation),

2. we have that the domain on which the variables are defined is bounded by $D = \max\left\{\sqrt{m}, 2R + \frac{2(R+\beta \log m)}{\chi(1-\gamma)}\right\}$ (from the definitions of the reward class $\mathcal{R}_{L_1}$, the fact that the policy in in the non-vanishing simplex is contained in the simplex and thus has diameter $< \sqrt{m}$ and using the result of proposition 1.2 to get the diameter of $\mathcal{B}$),

3. we known that the function $f_{\text{CIB-H}}$ is smooth with parameter

$$m \exp\Big(\frac{2\left(R+\beta\log\left(m\right)\right)\left(2\|\Psi\|-\chi\gamma+\chi\right)}{\chi\left(1-\gamma\right)^{2}}\Big)$$

from proposition 4.3.

Plugging those values into proposition 2.3 gives us our convergence time. To verify that what is recovered is indeed correct we refer to proposition 1.1 (which applies as the bandit case is simply a restriction of the generic IRL case). The choice of learning rate simply comes from the result of proposition 4.3 together with the required assumptions of proposition 1.1. □

4.2.2 Solving the Shannon-regularized problem with gradient descent-ascent : GDA-COP

The very large constant in front of the EG-COP algorithm's rate motivates the search for alternative methods. The most straight-forward approach to that problem is to simply take average iterates of the GDA algorithm (using the same projections as in algorithm 4).

Algorithm 5: GDA-COP: Gradient-Descent Ascent constrained inverse bandit algorithm

Set the learning rate $\eta_k > 0$
Initialize the algorithm at some point $(\boldsymbol{r}_0, \boldsymbol{\lambda}_0, \boldsymbol{\pi}_0)$
foreach *iteration* $k = 0, 2, ..., K-1$ **do**
  $\eta_k \to \xi\frac{1}{\sqrt{k}}$
  $\boldsymbol{r}_{k+1} \leftarrow \Pi_{\mathcal{R}_{L_1}}\big(\boldsymbol{r}_k - \eta_k \nabla_{\boldsymbol{r}} f_{\text{CIB-H}}(\boldsymbol{r}_k, \boldsymbol{\lambda}_k, \boldsymbol{\pi}_k)\big)$
  $\boldsymbol{\lambda}_{k+1} \leftarrow \Pi_{\boldsymbol{\lambda}\in\mathcal{B}}\big(\boldsymbol{\lambda}_k - \eta_k \nabla_{\boldsymbol{\lambda}} f_{\text{CIB-H}}(\boldsymbol{r}_k, \boldsymbol{\lambda}_k, \boldsymbol{\pi}_k)\big)$
  $\boldsymbol{\pi}_{k+1} \leftarrow \Pi_{\Delta_A^{(\rho)}}\big(\boldsymbol{\pi}_k + \eta_k \nabla_{\boldsymbol{\pi}} f_{\text{CIB-H}}(\boldsymbol{r}_k, \boldsymbol{\lambda}_k, \boldsymbol{\pi}_k)\big)$
end
return $(\hat{\boldsymbol{r}}_K, \hat{\boldsymbol{\lambda}}_K, \hat{\boldsymbol{\pi}}_K)$, where $\hat{\cdot}$ denotes the empirical mean over a sequence.

Theorem 4.2. *Algorithm 5 recovers the optimal reward and policy (up to reward shaping transformations), up to approximation error ϵ in time $O(1/\sqrt{K})$ (where K is the number of iterations). More specifically, assuming we use the decreasing learning rate $\eta_k = \xi\frac{1}{\sqrt{k}}$, we have:*

$$DG(\boldsymbol{r}_k, \boldsymbol{\lambda}_k, \boldsymbol{\pi}_k) \leq C_{\textit{GDA-COP}}\frac{1}{\sqrt{K}},$$

where:

$$C_{\textit{GDA-COP}} = 4\sqrt{m}D\Big(\log m + \frac{2\left(R+\beta\log\left(m\right)\right)\left(2\|\Psi\|-\chi\gamma+\chi\right)}{\chi\left(1-\gamma\right)^{2}} - 1\Big).$$

In which:

$$D = \max\Big\{\sqrt{m}, 2R + \frac{2(R+\beta\log m)}{\chi(1-\gamma)}\Big\},$$

is the domain diameter and:

$$\xi = \frac{D}{\sqrt{2m}\Big(\log m + \frac{2(R+\beta\log(m))(2\|\Psi\|-\chi\gamma+\chi)}{\chi(1-\gamma)^2} - 1\Big)}.$$

Proof. The proof follows the analysis of alg. 1, which we do not re-state for brevity, so we just have to verify that the assumptions are satisfied:

1. we note that the function $f_{\text{CIB-H}}$ is concave-convex in $(\boldsymbol{r}, \boldsymbol{\lambda}, \boldsymbol{\pi})$ (where we just stack $\boldsymbol{r}$ and $\boldsymbol{\lambda}$ to have a saddle-point formulation),

2. we have that the domain on which the variables are defined is bounded by $D = \max\Big\{\sqrt{m}, 2R + \frac{2(R+\beta\log m)}{\chi(1-\gamma)}\Big\}$ (from the definitions of the reward class $\mathcal{R}_{L_1}$, the fact that the policy in in the non-vanishing simplex is contained in the simplex and thus has diameter $< \sqrt{m}$ and using the result of proposition 1.2 to get the diameter of $\mathcal{B}$),

3. we known that the function $f_{\text{CIB-H}}$ is Lipschitz with parameter

$$\sqrt{m}\Big(\log m + \frac{2\left(R+\beta\log(m)\right)\left(2\|\Psi\|-\chi\gamma+\chi\right)}{\chi\left(1-\gamma\right)^2} - 1\Big).$$

   from the Lipschitz gradient result of proposition 3.1 and using the ρ bound of proposition 4.2.

Plugging those values into proposition 2.2 gives us our convergence time. To verify that what is recovered is indeed correct we refer to proposition 1.1 (which applies as the bandit case is simply a restriction of the generic IRL case). The choice of learning rate simply comes from the result of proposition 2.2 together with the required assumptions of proposition 1.1. □

5 Convergence of CIRL with exact gradients

We will first study the algorithm under the assumption that we have direct access to the expert policy $\boldsymbol{\pi}^E$ (we will look at sampling later). Our analysis will focus on exploiting the fast convergence of the NPG method under Shannon entropy regularization established by Cen et al. 2021 and can be though of as an extension of the result provided by Zeng et al. 2022 for the *ML-IRL* algorithm. It is worth mentioning that it also closely resembles the analysis developed by D. Ding et al. 2020 for convergence of NPG on CMDPs, but our analysis fundamentally differs from the one provided by the authors of D. Ding et al. 2020 as we leverage a linear convergence rate in the primal (soft policy iteration, or *SPI*) which is a consequence of the entropy regularization.

5.1 Problem Setting, NPG-CIRL algorithm

Recall that from proposition 1.1 we get that:

$$\min_{\boldsymbol{r}\in\mathcal{R}}\max_{\boldsymbol{\mu}\in\mathcal{F}}\langle \boldsymbol{r}, \boldsymbol{\mu}-\boldsymbol{\mu}^E\rangle - \tilde{\Omega}(\boldsymbol{\mu}), \tag{p}$$

which, we know by proposition 1.2 is equivalent to:

$$\min_{\boldsymbol{r}\in\mathcal{R},\ \boldsymbol{\lambda}\succcurlyeq 0}\max_{\boldsymbol{\mu}\in\mathcal{M}}\langle \boldsymbol{r}, \boldsymbol{\mu}-\boldsymbol{\mu}^E\rangle - (\tilde{\Omega}(\boldsymbol{\mu}) - \tilde{\Omega}(\boldsymbol{\mu}^E)) + \langle \boldsymbol{\lambda}, \boldsymbol{b} - \Psi^T\boldsymbol{\mu}\rangle. \tag{d}$$

Problem (d) is itself equivalent the following policy-form expression:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d,\ \boldsymbol{\lambda}\succcurlyeq 0}\max_{\boldsymbol{\theta}\in\mathbb{R}^p} J(\boldsymbol{\theta},\boldsymbol{w}) - J(\boldsymbol{\theta}^E,\boldsymbol{w}) + \langle \boldsymbol{\lambda}, \boldsymbol{b} - \boldsymbol{K}(\boldsymbol{\theta})\rangle, \tag{D}$$

Where $\boldsymbol{w}$ parametrizes the reward function $\boldsymbol{r} : \mathrm{dom}(\boldsymbol{w}^d) \to \mathcal{R}$ and $\boldsymbol{\theta}$ the policy function $\pi : \mathbb{R}^p \to \boldsymbol{\Pi}$. Although $\boldsymbol{r}$ is a function of $\boldsymbol{w}$ and $\boldsymbol{\pi}$ is a function of $\boldsymbol{\theta}$ we choose for brevity to omit this from the notation. So by convention we have: $\boldsymbol{r} := \boldsymbol{r}(\boldsymbol{w})$ and $\boldsymbol{\pi} := \boldsymbol{\pi}(\boldsymbol{\theta})$. Similarly we write $\boldsymbol{\mu} := \mathrm{OM}(\boldsymbol{\pi})$ to denote the occupancy measure associated with a certain parametrization/policy. The objective function $J : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ is defined (as in definition 1.7) as:

$$J(\boldsymbol{\theta},\boldsymbol{w}) := (1-\gamma)\mathbb{E}_{s\sim\boldsymbol{\mu}}\Big[\sum_{t=0}^{+\infty}\gamma^t\big(r_{\boldsymbol{w}}(s_t,a_t) - \beta\log\pi(a_t|s_t)\big)\Big|s_0\sim\boldsymbol{\nu},\ a_t\sim\pi_{\boldsymbol{\theta}}\Big],$$

and the cost function $\boldsymbol{K} : \mathbb{R}^p \to \mathbb{R}^c$ by:

$$\boldsymbol{K}(\boldsymbol{\theta}) := (1-\gamma)\mathbb{E}_{s\sim\boldsymbol{\mu}}\Big[\sum_{t=0}^{+\infty}\gamma^t\big(\Psi(s_t,a_t)\big)\Big|s_0\sim\boldsymbol{\nu}\Big].$$

For brevity we write problem (D) as:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d,\ \boldsymbol{\lambda}\succcurlyeq 0}\max_{\boldsymbol{\theta}\in\mathbb{R}^p} f(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\theta}), \tag{D}$$

respectively.

5.2 Policy and Reward Parametrization and Update Rules

We will specifically study convergence of our algorithm when the **policy** is *softmax* parameterized, recall that the *softmax* transform $\text{softmax} : \mathbb{R}^n \to \Delta_{n-1}$ is given by:

$$\big[\text{softmax}(\boldsymbol{\theta})\big]_i := \frac{\exp \boldsymbol{\theta}_i}{\sum_{j\in[n]} \exp(\boldsymbol{\theta}_j)}.$$

Or specifically, when considering policies, we write $\pi_{\boldsymbol{\theta}}(a|s) := \big[\text{softmax}(\boldsymbol{\theta})\big]_{s,a}$.

We will first limit our study to linear reward classes of the form:

$$\boldsymbol{r}_{\boldsymbol{w}} = \Big\{\Phi \boldsymbol{w} \Big| \boldsymbol{w} \in \mathbb{R}^d;\ \|\Phi \boldsymbol{w}\|_P \leq 1\Big\},$$

where $\Phi \in \mathbb{R}^{nm\times d}$ defines a linear space and P defines a ball of l_P norms in which we know the reward is contained. The simplest possible reward class is given by taking $\Phi = I^{nm\times nm}$ together with $P = 1$.

5.2.1 Primal Step: Natural Policy Gradients reduces to Multiplicative Weights Update

Lemma 5.1 (NPG reduces to MWU update)**.** *When applying NPG updates to a tabular softmax parameterized policy $\pi(a|s) := [softmax(\boldsymbol{\omega})](s,a)$ in the Shannon-entropy regularized setting, updates in policy space satisfy the following MWU update form:*

$$\begin{aligned}\pi^{(t+1)}(a|s) &= \frac{1}{Z^{(t)}(s)}\big(\pi^{(t)}(a|s)\big)^{(1-\eta_\theta\beta)} \exp\big(\eta_\theta Q_{\tilde{r}}^{\pi^{(t)}}(s,a)\big),\\ &= \frac{1}{Z^{(t)}(s)}\big(\pi^{(t)}(a|s)\big)^{(1-\eta_\theta\beta)} \exp\big(\eta_\theta (Q^{\pi}(s,a) - \langle \boldsymbol{\lambda}, \boldsymbol{Q}_{\Psi}^{\pi}(s,a)\rangle)\big).\end{aligned}$$

Where $Z^(t)$ normalizes each state in such a way that $\boldsymbol{\pi} \in \Pi$.

Proof. Before getting into the analysis, recall that the advantage function is given by :

$$\forall (s,a) \in S \times A :\ A^{\pi}(s,a) = Q^{\pi}(s,a) - \beta \log \pi(a|s) - V^{\pi}(s).$$

We also consider a "Lagrangian" advantage function $A_{\tilde{r}}^{\pi}(s,a)$:

$$\forall (s,a) \in S \times A :\ A_{\tilde{r}}^{\pi}(s,a) = Q_{\tilde{r}}^{\pi}(s,a) - \beta \log \pi(a|s) - V_{\tilde{r}}^{\pi}(s), \tag{5.1}$$

in which we consider the diminished reward, (obtained by subtracting the Lagrangian term from the reward, see definition 1.21): $\tilde{r}(s,a) = r(s,a) - \langle \lambda, \Psi(s,a)\rangle$. Our primal (policy) step is given by :

$$\boldsymbol{\theta}^{(k+1)} \leftarrow \boldsymbol{\theta}^{(k)} + \eta_\theta (\mathfrak{F}^{\boldsymbol{\theta}})^{\dagger} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{w}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{\theta}^{(k)}), \tag{S}$$

where $(\mathfrak{F}^{\boldsymbol{\theta}})^{\dagger}$ denotes the Moore-Penrose inverse of the Fisher information matrix, itself defined as:

$$\mathfrak{F}^{\boldsymbol{\theta}} := \mathbb{E}_{s\sim \text{OM}(\boldsymbol{\pi}),\ a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\Big[(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s))(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s))^T\Big].$$

Taking gradients of f w.r.t. $\boldsymbol{\theta}$ we get:

$$\begin{aligned}\frac{\partial f(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\theta})}{\partial\theta(s,a)} &= \mu^{\pi^\theta}(s)\pi^\theta(a|s)\big(A^{\pi^\theta}(s,a) - \langle\boldsymbol{\lambda}, \boldsymbol{A}_\Psi^{\pi^\theta}(s,a)\rangle\big),\\ &= \mu^{\pi^\theta}(s)\pi^\theta(a|s)A_{\tilde{r}}^{\pi^\theta}(s,a),\end{aligned} \tag{5.2}$$

which is the same form as what Cen et al. 2021 finds except the advantage (and thus the reward) is reduced by the Lagrangian term. This is to be expected as we know that $\mathrm{CRL}_{\mathrm{MO}(\boldsymbol{\pi})\in\mathcal{F}}(\boldsymbol{r}) = \mathrm{CRL}(\boldsymbol{r} - \langle\boldsymbol{\lambda}^*, \boldsymbol{K}(\theta)\rangle)$ (from strong duality, see proposition 1.2). We denote by $A_{\tilde{r}}^{\pi^\theta}(s,a)$ the regularized advantage function computed on the diminished reward $\tilde{r}(s,a) = r(s,a) - \langle\lambda, \Psi(s,a)\rangle$. Furthermore, still following the derivation steps from Cen et al. 2021 appendix C.6, we get that:

$$\begin{aligned}\Big(\big(\tilde{\mathfrak{F}}^{\boldsymbol{\theta}}\big)^\dagger\nabla_{\boldsymbol{\theta}} f(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\theta})\Big)(s,a) &= A_{\tilde{r}}^{\pi^\theta}(s,a) + c(s),\\ &= \big(A^{\pi^\theta}(s,a) - \langle\boldsymbol{\lambda}, \boldsymbol{A}_\Psi^{\pi^\theta}(s,a)\rangle\big) + c(s),\end{aligned} \tag{5.3}$$

where $c(s)$ doesn't depend on a. Now onto showing that this update corresponds to a Multiplicative Weight Update (MWU) step. The derivation goes as follows:

$$\begin{aligned}\pi^{(t+1)}(s,a) &\propto \exp\big(\theta^{(t+1)}(s,a)\big) && \text{softmax}\\ &= \exp\big(\theta^{(t)}(s,a) + \eta_\theta(\tilde{\mathfrak{F}}^{\boldsymbol{\theta}})^\dagger\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(k)})\big) && \text{NPG step}\\ &\overset{(i)}{\propto} \exp\big(\theta^{(t)}(s,a) + \eta_\theta A_{\tilde{r}}^{\pi^\theta}(s,a)\big) && \text{from (5.3)}\\ &\overset{(ii)}{\propto} \pi^{(t)}(s,a)\exp\big(\eta_\theta Q_{\tilde{r}}^{\pi}(s,a) - \eta_\theta\beta\log\pi(a|s)\big) && \text{from (5.1)}\\ &= \big(\pi^{(t)}(s,a)\big)^{1-\beta\eta_\theta}\exp\big(\eta_\theta Q_{\tilde{r}}^{\pi}(s,a)\big)\\ &= \big(\pi^{(t)}(s,a)\big)^{1-\beta\eta_\theta}\exp\big(\eta_\theta(Q^{\pi}(s,a) - \langle\boldsymbol{\lambda}, \boldsymbol{Q}_\Psi^{\pi}(s,a)\rangle)\big)\end{aligned}$$

Note that on (i) and (ii) we drop terms which are independent on a (because of the regularization). Which concludes the proof. □

Remark 5.1 (Optimal Rate/Soft Policy Iteration)**.** *Picking the policy learning rate $\eta_\theta = \frac{1}{\beta}$ yields a the following update rule:*

$$\pi^{(t+1)}(s,a) = \frac{1}{Z^{(t)}(s)}\exp\big(\eta_\theta Q_{\tilde{r}}^{\pi}(s,a)\big),$$

which can be interpreted as a "soft" version of the classical policy iteration algorithm and is thus often called ***soft policy iteration****. Here the normalization factor $Z^{(t)}$ is given by:*

$$Z^{(t)}(s) = \sum_{a'\in A}\exp\big(\eta_\theta Q_{\tilde{r}}^{\pi}(s,a')\big).$$

5.2.2 Performance difference lemma, soft Bellman optimality operator

As in most global convergence analysis of policy-gradient related algorithm we need a performance difference/improvement lemma. Here we get the following result.

Lemma 5.2 (Performance Improvement (PI))**.**

$$\begin{aligned} & f(\boldsymbol{w}^{(t)}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{\theta}^{(t+1)}) - f(\boldsymbol{w}^{(t)}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{\theta}^{(t)}) \\ = \mathbb{E}_{s\sim OM(\pi^t)(\cdot)} & \left[\Big(\frac{1}{\eta_\theta} - \beta\Big) KL\Big(\pi^{(t+1)}(\cdot|s) \Big|\Big| \pi^{(t)}(\cdot|s)\Big) \right. \\ & \left. + \frac{1}{\eta_\theta} KL\Big(\pi^{(t)}(\cdot|s) \Big|\Big| \pi^{(t+1)}(\cdot|s)\Big) \right]. \end{aligned}$$

For proof of this result refer to Cen et al. 2021.

Definition 5.1 (Soft Bellman Optimality Operator)**.** *Consider* $\mathcal{T}_\beta : \mathbb{R}^{nm} \to \mathbb{R}^{nm}$ *defined as,* $\forall (a, s) \in A \times S$ *:*

$$\mathcal{T}_\beta(Q)(s,a) := r(s,a) + \gamma \mathbb{E}_{s'\sim P(s'|s,a)} \left[\max_{\pi \in \Delta_A} \mathbb{E}_{a'\sim\pi(\cdot|s')} \Big[Q(s',a') - \beta \log \pi(a'|s') \Big] \right].$$

For proof of this result refer to Cen et al. 2021.

Lemma 5.3 (Properties of $\mathcal{T}_\beta$)**.** *The operator* $\mathcal{T}_\beta$ *satisfies the following:*

1. *it admits the closed-form expression:*

$$\mathcal{T}_\beta(Q)(s,a) = r(s,a) + \gamma \mathbb{E}\Big[\tau \log\big(\| \exp(Q(s',\cdot)/\beta)\|_1\big)\Big],$$

2. *the optimal solution of the MDP is a fixed point:*

$$\mathcal{T}_\beta(\boldsymbol{Q}^*) = \boldsymbol{Q}^*,$$

3. *it is a* γ*-contraction in* ∞*-norm:*

$$\|\mathcal{T}_\beta(\boldsymbol{Q}_1) - \mathcal{T}_\beta(\boldsymbol{Q}_2)\|_\infty \le \gamma \|\boldsymbol{Q}_1 - \boldsymbol{Q}_2\|_\infty.$$

For proof of this result refer to Cen et al. 2021.

Claim 5.1 (SPI implements the Soft Bellman Operator)**.** *We claim that the NPG step as defined in 5.3 with learning rate* $\eta_\theta = \frac{1}{\beta}$ *(a.k.a. the SPI step as in definition 5.1) implements the soft bellman operator as defined in definition 5.1.*

Proof.

$$\begin{aligned} Q^{(t+1)}(s,a) & \overset{(i)}{=} r(s,a) + \gamma \mathbb{E}_{s'\sim P(s'|s,a)} \Big[V^{(t+1)}(s') \Big] \\ & \overset{(ii)}{=} r(s,a) + \gamma \mathbb{E}_{s'\sim P(s'|s,a), a'\sim\pi(\cdot|s')} \Big[Q^{(t+1)}(s',a') - \beta \log \pi^{(t+1)}(a'|s') \Big] \\ & \overset{(iii)}{\ge} r(s,a) + \gamma \mathbb{E}_{s'\sim P(s'|s,a), a'\sim\pi(\cdot|s')} \Big[Q^{(t)}(s',a') - \beta \log \pi^{(t+1)}(a'|s') \Big] \\ & \overset{(iv)}{=} r(s,a) + \gamma \mathbb{E}_{s'\sim P(s'|s,a), a'\sim\pi(\cdot|s')} \Big[\beta \log\Big(\big\| \exp Q^{(t)}(s',\cdot)/\beta \big\|_1 \Big) \Big] \\ & \overset{(v)}{=} \mathcal{T}\big(Q^{(t)}\big)(s,a). \end{aligned}$$

In (i) we plug in the definition of the Q-value, in (ii) we use that the V-value is the average Q (to which we subtract the regularization), in (iii) we use that Q is monotonous (which is a consequence of lemma 5.2) and in (iv) we plug in the SPI update. Finally in (v) we just observe that we now have the soft bellman operator as defined in definition 5.1. □

5.3 The NPG-CIRL algorithm

Solving the saddle point problem (D) in parameter-space, using natural gradients, naturally yields algorithm 6. Which we are now going to try to analyze.

Algorithm 6: NPG-CIRL: Natural Policy Gradient CIRL (Exact Gradients)

Set the learning rates $\eta_\theta = \frac{1}{\beta}$, $\eta_z = \frac{1}{T^u}$
Initialize the algorithm at some point $(\boldsymbol{w}^{(0)}, \boldsymbol{\lambda}^{(0)}, \boldsymbol{\theta}^{(0)})$
foreach *iteration* $k = 0, 2, ..., K-1$ **do**
  $\boldsymbol{w}^{(k)} \leftarrow \Pi_{\mathrm{dom}(\boldsymbol{w})}\big(\boldsymbol{r}_k - \eta_z \nabla_{\boldsymbol{w}} f(\boldsymbol{w}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{\theta}^{(k)})\big)$
  $\boldsymbol{\lambda}^{(k+1)} \leftarrow \Pi_{\boldsymbol{\lambda}\in\mathcal{B}}\big(\boldsymbol{\lambda}^{(k)} - \eta_z \nabla_{\boldsymbol{\lambda}} f(\boldsymbol{w}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{\theta}^{(k)})\big)$
  $\boldsymbol{\theta}^{(k+1)} \leftarrow \boldsymbol{\theta}^{(k)} + \eta_\theta (\mathfrak{F}^{\boldsymbol{\theta}})^\dagger \nabla_{\boldsymbol{\theta}} f(\boldsymbol{w}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{\theta}^{(k)})$
end
return $(\boldsymbol{w}^{(K)}, \boldsymbol{\lambda}^{(K)}, \boldsymbol{\theta}^{(K)})$.

The algorithm requires tuning two parameters: a descent learning rate η_z and a policy (ascent) learning rate θ. Note that the tuning of both learning rates is done using parameters to which we have trivially access. We will show that the following theorem is true:

Theorem 5.4. *Considering a CIRL problem (as defined in 1.20) satisfying assumption 1.1, where the reward parametrization satisfies assumption 5.1, we have that the duality gap $f(\boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)}) - f^*$ goes converges to an ϵ error in finite time with the following rate:*

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)}) - f^* \leq C_{NPG\text{-}CIRL} T^{-1/3},$$

with the constant:

$$C_{NPG\text{-}CIRL} = + \left(\frac{D_z + (B_w + \|\boldsymbol{b}\|_2 + \|\Psi\|)^2}{2} + \frac{C_\pi}{\eta_\theta (1-u)} \right) + \frac{2 D_{\boldsymbol{z}} B_a \sqrt{n C_\pi}}{(1-\gamma)(1-u/2)}.$$

Where $D^z = diam(D^{\mathcal{R}}) + diam(D^{\mathcal{B}})$ is an upper bound the diameter of the reward class and the diameter of the set of admissible Lagrange multiplier vectors (which can be bounded explicitly as in 4.1), B_w is an upper bound of the gradient of the reward class. The norm $\|\boldsymbol{b}\|_2$ is the norm of the constraint vector and $\|\Psi\|$ is the spectral norm of the cost matrix. C_π is given in lemma 5.5. Finally we have $B_a = B_w + \|\Psi\|$ and $B_b = diam(\mathcal{R}) + diam(\mathcal{B})\|\Psi\|$.

5.4 Analysis of NPG-CIRL

We now enter the analysis section of the algorithm. Our analysis hinges on two main results:

1. the algorithm converges fast to a locally optimal policy $\tilde{\pi}$, which we show in section 5.4.1,

2. this convergence results allows us to analyze the global convergence of the algorithm, which we do in section 5.4.2.

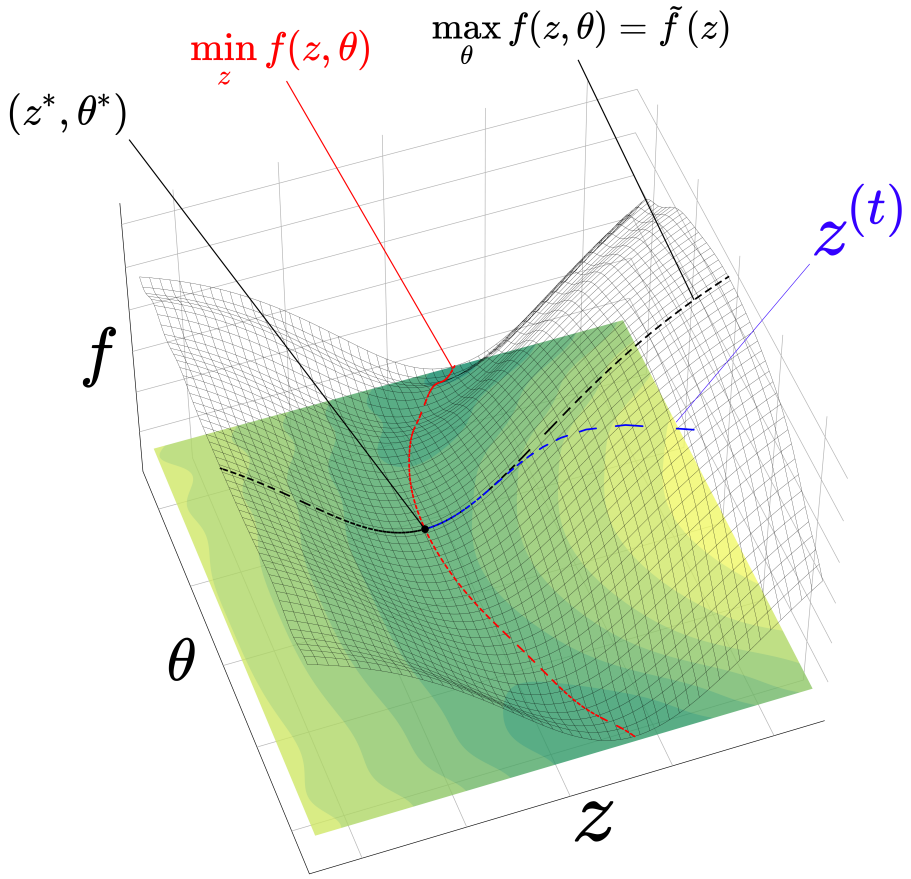

Figure 6: An intuitive illustration of the NPG-CIRL algorithms inner-working. We simultaneously maximize in the policy parameter $\boldsymbol{\theta}$ and minimize in the remaining parameters that are concatenated in the $\boldsymbol{z}$ vector. The algorithm converges to a saddle-point $(\boldsymbol{\theta}^*, \boldsymbol{z}^*)$ which lies at the intersection of manifolds $\tilde{f}$ (the set of $\boldsymbol{\theta}$ maximizing f for a fixed $\boldsymbol{z}$) and $\min_{\boldsymbol{z}} f(\boldsymbol{z}, \boldsymbol{\theta})$ (the set of minimizer of $\boldsymbol{z}$ for a fixed theta). Our algorithm converges fast of $\tilde{f}$ and we leverage the convexity of $\tilde{f}$ to show global convergence of alg. 6.

5.4.1 Fast convergence to a locally optimal policy

We consider the **locally optimal policy** $\tilde{\boldsymbol{\pi}}^{(t)}$ which is optimal for reward $\boldsymbol{r}^{(t)}$ and lagrange multiplier $\boldsymbol{\lambda}^{(t)}$, which we can write as:

$$\tilde{\boldsymbol{\pi}}^{(t)} := \mathrm{RL}_{\pi}\big(\boldsymbol{r}^{(t)} - \Psi\boldsymbol{\lambda}^{(t)}\big).$$

For brevity we will make use of $\tilde{\boldsymbol{r}}^{(t)} = \boldsymbol{r}^{(t)} - \Psi\boldsymbol{\lambda}^{(t)}$, the diminished reward (as defined in 1.21). We will prove the following result:

Lemma 5.5 (Policy convergence to local optimum)**.** *Assuming the reward-parameter learning rate is set to $\eta_w = \frac{1}{K^u}$, algorithm 6 converges in Q-values and in log policy error to the locally optimal policy $\tilde{\boldsymbol{\pi}}^{(t)}$ at a rate of:*

$$\|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}}\|_{\infty} \leq \frac{2(2B_w^2 + \|\Psi\|^2)}{(1-\gamma)T^u} + \frac{\gamma\|\boldsymbol{Q}^{(0)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}}\|_{\infty}}{T},$$

$$\|\log\boldsymbol{\pi}^{(t+1)} - \log\tilde{\boldsymbol{\pi}}^{(t)}\|_{\infty} \leq \frac{4(2B_w^2 + \|\Psi\|^2)}{(1-\gamma)T^u} + \frac{2\gamma\|\boldsymbol{Q}^{(0)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}}\|_{\infty}}{T}.$$

Note that since $u \in (0,1)$ we have that:

$$\begin{aligned}
\|\log\boldsymbol{\pi}^{(t+1)} - \log\tilde{\boldsymbol{\pi}}^{(t)}\|_{\infty} &\leq \frac{4(2B_w^2 + \|\Psi\|^2)}{(1-\gamma)T^u} + \frac{2\gamma\|\boldsymbol{Q}^{(0)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}}\|_{\infty}}{T}, \\
&\leq \left(\frac{4(2B_w^2 + \|\Psi\|^2)}{(1-\gamma)} + 2\gamma\|\boldsymbol{Q}^{(0)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}}\|_{\infty}\right)\frac{1}{T^u}, \\
&\leq C_{\pi}\frac{1}{T^u},
\end{aligned}$$

where $C_{\pi} = \frac{4(2B_w^2+\|\Psi\|^2)}{(1-\gamma)} + 2\gamma\|\boldsymbol{Q}^{(0)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}}\|_{\infty}$, which allows for a slightly more digest notation.

Proof. We will track the log-error of the policy $\boldsymbol{\pi}^{(t+1)}$ generated by algorithm 6 w.r.t

to $\tilde{\boldsymbol{\pi}}^{(t)}$:

$$
\begin{aligned}
&\left|\log \pi^{(t+1)}(a|s) - \log \tilde{\pi}^{(t)}(a|s)\right| \\
&\overset{(i)}{=} \left| \log\left(\frac{\exp(Q^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a))}{\sum_{a'\in A}\exp(Q^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a'))}\right) - \log\left(\frac{\exp(Q^{*}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a))}{\sum_{a'\in A}\exp(Q^{*}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a'))}\right)\right| \\
&= \Bigg| \log\exp(Q^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a)) - \log\exp(Q^{*}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a)) \\
&+ \log\sum_{a'\in A}\exp(Q^{*}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a')) - \log\sum_{a'\in A}\exp(Q^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a'))\Bigg| \\
&\overset{(ii)}{\leq} \left|\log\exp(Q^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a)) - \log\exp(Q^{*}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a))\right| \\
&+ \left|\log\sum_{a'\in A}\exp(Q^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a')) - \log\sum_{a'\in A}\exp(Q^{*}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a'))\right| \\
&\overset{(iii)}{\leq} \left|Q^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a) - Q^{*}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a)\right| + \max_{a'\in A}\left|Q^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a') - Q^{*}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a')\right|.
\end{aligned}
$$

Where (i) comes from substituting the SPI step (as in remark 5.1) in place of $\pi^{(t+1)}(a|s)$ (claim 5.1) and by inserting the closed form expression of the optimal policy into $\tilde{\pi}^{(t)}(a|s)$, then we rearrange the expression and (ii) use a triangle inequality. From there the $\log\exp(\cdot)$ expressions simplify and we can use lemma 5.3 on the sum term to get our result. Since what we just computed is true component-wise we have that the log of our policies is bounded as follows:

$$\|\log \boldsymbol{\pi}^{(t+1)} - \log \tilde{\boldsymbol{\pi}}^{(t)}\|_\infty \leq 2\|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}}\|_\infty.$$

We will thus study the convergence of the Q values to their local optimum:

$$
\begin{aligned}
\|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}}\|_\infty &= \|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}} + \overbrace{\boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t-1)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t-1)}}}^{=0} + \overbrace{\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t-1)}} - \boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t-1)}}}^{=0}\|_\infty \\
&\overset{(i)}{\leq} \underbrace{\|\boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t-1)}}\|_\infty}_{(A)} + \underbrace{\|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t-1)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t-1)}}\|_\infty}_{(B)} \\
&\quad + \underbrace{\|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t-1)}}\|_\infty}_{(C)}
\end{aligned}
$$

here we just add two well chosen 0 in the first line and then (i) rearrange and take a triangle inequality. We get three terms (A), (B) and (C) that we need to bound separately.
We start with (A), using lemma 5.6 we get:

$$\|\boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t-1)}}\|_\infty \leq \frac{B_w}{1-\gamma}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + \frac{\|\Psi\|}{1-\gamma}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2.$$

Moving on to (C) we can similarily use lemma 5.8 and get:

$$\|\boldsymbol{Q}^{t}_{\tilde{\boldsymbol{r}}^{(t-1)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t-1)}}\|_\infty \leq \frac{B_w}{1-\gamma}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + \frac{\|\Psi\|}{1-\gamma}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2.$$

This leaves (B), for which we will use the contraction property of the soft Bellman operator. Recall that SPI gives:

$$\mathcal{T}_\beta^{\tilde{\boldsymbol{r}}^k}(\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}})(s,a) = \boldsymbol{Q}^{(t+1)}_{\tilde{\boldsymbol{r}}^{(t)}}(s,a). \tag{SPI-T}$$

Starting from (B) we have:

$$\begin{aligned}\|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t-1)}} - \boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}^{(t-1)}}\|_\infty &\overset{(i)}{=} \|\mathcal{T}_\beta^{\tilde{\boldsymbol{r}}^k}\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t-1)}} - \boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}^{(t-1)}}\|_\infty \\ &\overset{(ii)}{=} \|\mathcal{T}_\beta^{\tilde{\boldsymbol{r}}^k}\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t-1)}} - \mathcal{T}_\beta^{\tilde{\boldsymbol{r}}^k}\boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}^{(t-1)}}\|_\infty \\ &\overset{(iii)}{\leq} \gamma\|\boldsymbol{Q}^{(t-1)}_{\tilde{\boldsymbol{r}}^{(t-1)}} - \boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}^{(t-1)}}\|_\infty.\end{aligned}$$

Where (i) stems from (SPI-T), (ii) from the fact that the optimal is a fixed point of the soft Bellman operator and (iii) by the contraction property of the operator. Both (ii) and (iii) are results from lemma 5.3. Using the bounds for (A), (B) and (C) we get that:

$$\begin{aligned}\|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}^{(t)}}\|_\infty \leq \frac{2B_w}{1-\gamma}\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^{(t-1)}\|_2 + \frac{2\|\Psi\|}{1-\gamma}\|\boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}^{(t-1)}\|_2 \\ +\gamma\|\boldsymbol{Q}^{(t-1)}_{\tilde{\boldsymbol{r}}^{(t-1)}} - \boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}^{(t-1)}}\|_\infty.\end{aligned}$$

By claim 5.4 we know that :

$$\frac{2B_w}{1-\gamma}\overbrace{\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^{(t-1)}\|_2}^{\leq \eta_z 2B_w} + \frac{2\|\Psi\|}{1-\gamma}\overbrace{\|\boldsymbol{\lambda}^{(t)} - \boldsymbol{\lambda}^{(t-1)}\|_2}^{\leq \eta_z\|\Psi\|} \leq \frac{2\eta_z(2B_w^2 + \|\Psi\|^2)}{1-\gamma}$$

Thus we are equipped to complete our analysis of local convergence, for convenience we pose $H^{(t)} := \|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}^{(t)}}\|_\infty$, our descent inequality is of the form:

$$H^{(t)} \leq \frac{2\eta_z(2B_w^2 + \|\Psi\|^2)}{1-\gamma} + \gamma H^{(t-1)}.$$

Averaging we get:

$$\frac{1}{T}\sum_{t=1}^{T} H^{(t)} \leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{2\eta_z(2B_w^2 + \|\Psi\|^2)}{1-\gamma} + \gamma H^{(t-1)}.$$

Which can be rearranged into :

$$\begin{aligned}\frac{1-\gamma}{T}\sum_{t=1}^{T} H^{(t)} &\leq \frac{2\eta_z(2B_w^2 + \|\Psi\|^2)}{1-\gamma} + \frac{\gamma}{T}(H^{(0)} - H^{(t)}) \\ &\leq \frac{2\eta_z(2B_w^2 + \|\Psi\|^2)}{1-\gamma} + \frac{\gamma H^{(0)}}{T}.\end{aligned}$$

Picking $\eta_z = \frac{1}{T^u}$ where $u \in (0,1)$ we get the following convergence rate in Q values:

$$\begin{aligned}\|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}^{(t)}}\|_\infty \leq \frac{1-\gamma}{T}\sum_{t=1}^{T}\|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}^{(t)}}\|_\infty \leq \\ \frac{2(2B_w^2 + \|\Psi\|^2)}{(1-\gamma)T^u} + \frac{\gamma\|\boldsymbol{Q}^{(0)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}^{(t)}}\|_\infty}{T}.\end{aligned}$$

which immediately gives an upper bound on the log-policy error by:

$$\begin{aligned}\|\log \boldsymbol{\pi}^{(t+1)} - \log \tilde{\boldsymbol{\pi}}^{(t)}\|_\infty &\leq 2\|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}}\|_\infty \\ &\leq \frac{4(2B_w^2 + \|\Psi\|^2)}{(1-\gamma)T^u} + \frac{2\gamma\|\boldsymbol{Q}^{(0)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}}\|_\infty}{T}.\end{aligned}$$

The proof is complete.

□

5.4.2 Global convergence

Proof. (of theorem 5.4) In order to study the convergence of $f(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\theta})$ we will look at the convergence of gradient descent on the function $\tilde{f}(\boldsymbol{w}, \boldsymbol{\lambda}) := f(\boldsymbol{w}, \boldsymbol{\lambda}, \tilde{\boldsymbol{\theta}})$, where $\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} f(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\theta})$ for readability we define $\boldsymbol{z} := [\boldsymbol{\lambda}, \boldsymbol{w}]^T$, $\boldsymbol{Z} = \mathcal{B} \times \text{dom}(w)$ and write $\tilde{f}(\boldsymbol{z}) = \tilde{f}(\boldsymbol{\lambda}, \boldsymbol{w})$ and $Z = \text{dom}(\boldsymbol{w}) \times \mathcal{B}$. The main idea behind our analysis is to study convergence of the problem:

$$\begin{aligned}&\min_{\boldsymbol{z} \in \boldsymbol{Z}} \tilde{f}(\boldsymbol{z}) \\ =&\min_{\boldsymbol{z} \in \boldsymbol{Z}} \Big(\max_{\boldsymbol{\theta}} f(\boldsymbol{z}, \boldsymbol{\theta}) \Big),\end{aligned} \tag{R}$$

and see that *NPG-CIRL*'s convergence "tracks" the convergence of (R) (thanks to fast convergence to the locally optimal policy). We are now equipped to start an analysis of gradient descent, we use the following notation for the projection steps:

$$\begin{aligned}\boldsymbol{z}^{(t+1)} &:= \Pi_Z\Big(\boldsymbol{z}^{(t+1/2)}\Big), \\ \boldsymbol{z}^{(t+1/2)} &:= \boldsymbol{z}^{(t)} - \eta \boldsymbol{g}^{(t)}.\end{aligned}$$

We start our analysis by looking at the gradients w.r.t. $\boldsymbol{w}$ and $\boldsymbol{\lambda}$:

$$\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\theta}) = (\boldsymbol{\mu} - \boldsymbol{\mu}^E)\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{w}} = \overbrace{(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^E)\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{w}}}^{\nabla_{\boldsymbol{w}} \tilde{f}(\boldsymbol{w})} + \overbrace{(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}})\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{w}}}^{\boldsymbol{\sigma}_{\boldsymbol{w}}(\boldsymbol{\theta}, \boldsymbol{w})}, \tag{5.4}$$

$$\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\theta}) = (\boldsymbol{b} - \boldsymbol{K}(\boldsymbol{\theta})) = \overbrace{(\boldsymbol{b} - \boldsymbol{K}(\tilde{\boldsymbol{\theta}}))}^{\nabla_{\boldsymbol{\lambda}} \tilde{f}(\boldsymbol{w})} + \overbrace{(\boldsymbol{K}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{K}(\boldsymbol{\theta}))}^{\boldsymbol{\sigma}_{\boldsymbol{\lambda}}(\boldsymbol{\theta}, \boldsymbol{w})}. \tag{5.5}$$

We thus can rewrite the *NPG-CIRL* gradients (for convenience we write the gradient at the k-th step as $\boldsymbol{g}^{(t)} := \nabla_{\boldsymbol{z}} f(\boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)}))$ (in Lagrange multiplier and in reward parameters) as:

$$\boldsymbol{g}^{(t)} = \nabla \tilde{f}_{\boldsymbol{z}}(\boldsymbol{z}^{(t)}) + \boldsymbol{\sigma}_{\boldsymbol{z}}(\boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)}) = \begin{bmatrix} \nabla_{\boldsymbol{w}} \tilde{f}(\boldsymbol{w}^{(t)}) \\ \nabla_{\boldsymbol{\lambda}} \tilde{f}(\boldsymbol{w}^{(t)}) \end{bmatrix} + \begin{bmatrix} \boldsymbol{\sigma}_{\boldsymbol{w}}(\boldsymbol{\theta}^{(t)}, \boldsymbol{w}^{(t)}) \\ \boldsymbol{\sigma}_{\boldsymbol{\lambda}}(\boldsymbol{\theta}^{(t)}, \boldsymbol{w}^{(t)}) \end{bmatrix}.$$

For brevity we write $\boldsymbol{\sigma}^{(t)} := \boldsymbol{\sigma}_{\boldsymbol{z}}(\boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)})$.

Now let's consider the suboptimality of the function $f(\boldsymbol{z}^t, \boldsymbol{\theta}^t)$ w.r.t the optimum objective $f^* = f(\boldsymbol{z}^*, \boldsymbol{\theta}*) = \tilde{f}(\boldsymbol{z}^*)$, we have that:

$$f(\boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)}) - f^* = \overbrace{\tilde{f}(\boldsymbol{z}^{(t)}) - f^*}^{(a)} + \overbrace{f(\boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)}) - \tilde{f}(\boldsymbol{z}^{(t)})}^{(b)}.$$

We will study (a) and (b) separately and try to upper-bound both. Let's start with (a), we know from lemma 5.9 that $\tilde{f}$ is convex, hence we have:

$$\begin{aligned}
\tilde{f}(\boldsymbol{z}^{(t)}) - f^* &\overset{(i)}{\leq} \langle \nabla \tilde{f}_{\boldsymbol{z}}(\boldsymbol{z}^{(t)}), \boldsymbol{z}^{(t)} - \boldsymbol{z}^* \rangle \\
&\overset{(ii)}{=} \langle \boldsymbol{g}^{(t)} - \boldsymbol{\sigma}_{\boldsymbol{z}}^t, \boldsymbol{z}^{(t)} - \boldsymbol{z}^* \rangle = \langle \boldsymbol{g}^{(t)}, \boldsymbol{z}^{(t)} - \boldsymbol{z}^* \rangle - \langle \boldsymbol{\sigma}_{\boldsymbol{z}}^t, \boldsymbol{z}^{(t)} - \boldsymbol{z}^* \rangle \\
&\overset{(iii)}{=} \frac{1}{2\eta_z} \Big(\overbrace{\|\boldsymbol{z}^{(t)} - \boldsymbol{z}^{(t+1/2)}\|^2}^{\eta_z^2 \|\boldsymbol{g}^{(t)}\|^2} + \|\boldsymbol{z}^{(t)} - \boldsymbol{z}^*\|^2 - \overbrace{\|\boldsymbol{z}^{(t+1/2)} - \boldsymbol{z}^*\|^2}^{\geq \|\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^*\|^2} \Big) \\
&\quad - \langle \boldsymbol{\sigma}_{\boldsymbol{z}}^t, \boldsymbol{z}^{(t)} - \boldsymbol{z}^* \rangle \\
&\overset{(iv)}{\leq} \frac{1}{2\eta_z} \Big(\eta_z^2 \|\boldsymbol{g}^{(t)}\|^2 + \|\boldsymbol{z}^{(t)} - \boldsymbol{z}^*\|^2 - \|\boldsymbol{z}^{(t+1)} - \boldsymbol{z}^*\|^2 \Big) - \langle \boldsymbol{\sigma}_{\boldsymbol{z}}^t, \boldsymbol{z}^{(t)} - \boldsymbol{z}^* \rangle.
\end{aligned}$$

We start (i) by convexity of $\tilde{f}$, then (ii) we separate the terms associated with $\tilde{f}$ and with the "error-term" $\boldsymbol{\sigma}^{(t)}$, applying (iii) the parallelogram law (as is usually the case in the vanilla analysis of gradient descent), then we use (iv) that projection is non-expansive. From now on to keep the notation readable, we denote $D^t := \|\boldsymbol{z}^{(t)} - \boldsymbol{z}^*\|^2$, we thus have:

$$\tilde{f}(\boldsymbol{z}^{(t)}) - f^* \leq \frac{1}{2\eta_z} \Big(\eta_z^2 \|\boldsymbol{g}^{(t)}\|^2 + \overbrace{D^{(t)} - D^{(t+1)}}^{\text{will telescope}} \Big) - \langle \boldsymbol{\sigma}_{\boldsymbol{z}}^t, D^{(t)} \rangle.$$

We need to compute a bound for the error term $\langle \boldsymbol{\sigma}_{\boldsymbol{z}}^t, D^{(t)} \rangle$, to do so we expand from it's definition:

$$\begin{aligned}
\Big| \langle \boldsymbol{\sigma}_{\boldsymbol{z}}^t, D^{(t)} \rangle \Big| &\overset{(i)}{\leq} \|\boldsymbol{\sigma}_{\boldsymbol{z}}^t\|_2 \cdot \|D^{(t)}\|_2 \\
&\overset{(ii)}{\leq} D_{\boldsymbol{z}} \big(\|\boldsymbol{\sigma}_{\boldsymbol{w}}^t\|_2 + \|\boldsymbol{\sigma}_{\boldsymbol{\lambda}}^t\|_2 \big) \\
&\overset{(iii)}{=} D_{\boldsymbol{z}} \Big(\Big\| (\boldsymbol{\mu}^{(t)} - \tilde{\boldsymbol{\mu}}) \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{w}} \Big\| + \Big\| \Psi^T (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(t)}) \Big\|_2 \Big) \\
&\overset{(iv)}{\leq} D_{\boldsymbol{z}} \overbrace{\Big(B_w + \big\| \Psi \big\| \Big)}^{\leq B_a} \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(t)}\|_2 \\
&\overset{(v)}{\leq} \frac{D_{\boldsymbol{z}} B_a}{1-\gamma} \|\tilde{\boldsymbol{\pi}} - \boldsymbol{\pi}^{(t)}\|_2 \\
&\overset{(vi)}{\leq} \frac{2 D_{\boldsymbol{z}} B_a \sqrt{n}}{1-\gamma} \sqrt{\|\log \tilde{\boldsymbol{\pi}} - \log \boldsymbol{\pi}^{(t)}\|_\infty} \\
&\overset{(vii)}{\leq} \frac{2 D_{\boldsymbol{z}} B_a}{1-\gamma} \sqrt{\frac{n \cdot C_\pi}{t^u}}.
\end{aligned}$$

Where (i) and (ii) are Cauchy-Schwartz and triangle inequalities ($D_{\boldsymbol{z}}$ is just the diameter of the Z domain), (iii) comes from plugging in the definitions of the error terms, (iv) also is a triangle inequality (on spectral norms) and introduces B_a, a bound on the sum of spectral norms of the reward parametrization hessian and the cost matrix (which in the linear reward setting is just $\|\Phi\|$), (v) stems from lemma

1.2, (vi) uses claim 5.2, and (vii) uses the locally optimal convergence result from lemma 5.5.

We are now done with bounding (a) let's move on to (b), from the definition of the objective function we have:

$$
\begin{aligned}
|f(\boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)}) - \tilde{f}(\boldsymbol{z}^{(t)})| &\overset{(i)}{\leq} \frac{1}{\eta_\theta} \sum_{s \in S} \mu_s(s) \mathrm{D}_{\mathrm{KL}}\big(\boldsymbol{\pi}^{(t)}(\cdot|s) || \boldsymbol{\pi}^{(t+1)}(\cdot|s)\big) \\
&\overset{(ii)}{\leq} \frac{1}{\eta_\theta} \sup_{s \in S} \left[\mathrm{D}_{\mathrm{KL}}\big(\boldsymbol{\pi}^{(t)}(\cdot|s) || \boldsymbol{\pi}^{(t+1)}(\cdot|s)\big) \right] \\
&\leq \frac{1}{\eta_\theta} \sup_{s \in S} \left| \left\langle \boldsymbol{\pi}^{(t)}(\cdot|s), \log \boldsymbol{\pi}^{(t)}(\cdot|s) - \log \boldsymbol{\pi}^{(t+1)}(\cdot|s) \right\rangle \right| \\
&\overset{(iii)}{\leq} \frac{1}{\eta_\theta} \sup_{s \in S} \|\boldsymbol{\pi}^{(t)}(\cdot|s)\|_1 \cdot \| \log \boldsymbol{\pi}^{(t)}(\cdot|s) - \log \boldsymbol{\pi}^{(t+1)}(\cdot|s)\|_\infty \\
&\leq \frac{1}{\eta_\theta} \| \log \boldsymbol{\pi}^{(t)} - \log \boldsymbol{\pi}^{(t+1)} \|_\infty \\
&\overset{(iv)}{\leq} \frac{C_\pi}{\eta_\theta} \frac{1}{t^u}
\end{aligned}
$$

Where (i) stems from the soft sub-optimality lemma (lemma 5.7), (ii) upper bounds the expectation by it's max value, (iii) uses Holder's inequality on the KL divergence and (iv) upper bounds the log difference by the local convergence rate from lemma 5.5. Bringing (a) and (b) together we have:

$$
\begin{aligned}
f(\boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)}) - f^* &= \overbrace{\tilde{f}(\boldsymbol{z}^{(t)}) - f^*}^{(a)} + \overbrace{f(\boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)}) - \tilde{f}(\boldsymbol{z}^{(t)})}^{(b)} \\
&\leq \frac{1}{2\eta_z} \Big(\eta_z^2 \|\boldsymbol{g}^{(t)}\|^2 + D^{(t)} - D^{(t+1)} \Big) + \frac{2 D_{\boldsymbol{z}} B_a}{1-\gamma} \sqrt{\frac{n \cdot C_\pi}{t^u}} + \frac{C_\pi}{\eta_\theta} \frac{1}{t^u}.
\end{aligned}
$$

Summing and dividing by $\frac{1}{K}$ on both sides we have:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} f(\boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)}) - f^* \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{1}{2\eta_z}\Big(\eta_z^2\|\boldsymbol{g}^{(t)}\|^2 + D^{(t)} - D^{(t+1)}\Big) + \frac{2D_{\boldsymbol{z}}B_a}{1-\gamma}\sqrt{\frac{n\cdot C_\pi}{t^u}} + \frac{C_\pi}{\eta_\theta}\frac{1}{t^u}\right] \\
&= \frac{D^{(0)} - D^{(T)}}{2\eta_z T} + \frac{\eta_z}{2T}\sum_{t=0}^{T-1}\|\boldsymbol{g}^{(t)}\|^2 + \frac{2D_{\boldsymbol{z}}B_a\sqrt{nC_\pi}}{1-\gamma}\frac{1}{T}\sum_{t=1}^{T} t^{-u/2} + \frac{C_\pi}{\eta_\theta}\frac{1}{T}\sum_{t=1}^{T} t^{-u} \\
&\overset{(i)}{\leq} \frac{D^{(0)}}{2\eta_z T} + \frac{\eta_z(B_w + \|\boldsymbol{b}\|_2 + \|\Psi\|)^2}{2} + \frac{2D_{\boldsymbol{z}}B_a\sqrt{nC_\pi}}{1-\gamma}\frac{1}{T}\sum_{t=1}^{T} t^{-u/2} + \frac{C_\pi}{\eta_\theta}\frac{1}{T}\sum_{t=1}^{T} t^{-u} \\
&\overset{(ii)}{\leq} \frac{D^{(0)}}{2\eta_z T} + \frac{\eta_z(B_w + \|\boldsymbol{b}\|_2 + \|\Psi\|)^2}{2} + \frac{2D_{\boldsymbol{z}}B_a\sqrt{nC_\pi}}{(1-\gamma)(1-u/2)}\cdot\frac{1}{T^{u/2}} + \frac{C_\pi}{\eta_\theta(1-u)}\frac{1}{T^u} \\
&\overset{(iii)}{\leq} \frac{D^z}{2T^{1-u}} + \left(\frac{(B_w + \|\boldsymbol{b}\|_2 + \|\Psi\|)^2}{2T^u} + \frac{C_\pi}{\eta_\theta(1-u)}\right)\frac{1}{T^u} + \frac{2D_{\boldsymbol{z}}B_a\sqrt{nC_\pi}}{(1-\gamma)(1-u/2)}\cdot\frac{1}{T^{u/2}}
\end{aligned}
$$

A well chosen u finishes the proof. In (i) we bound the total gradient $\boldsymbol{g}^{(t)} = \|\nabla_z \tilde{f}\|_2$ by the sum of the bound on the gradient components, the reward gradient is bounded by B_w (assumption 5.1), the Lagrange multiplier gradient is bounded by the sum of the constraint cost $\boldsymbol{b}$ norm and the spectral norm of the cost matrix $\|\Psi\|$ (see gradient expression (5.5)). In (i) we upper-bound the sums by integrals:

$$
\frac{1}{T}\sum_{t=1}^{T} t^{-u/2} \leq \frac{1}{T}\int_{t=0}^{T} x^{-u/2}dx = \frac{1}{T}\left[\frac{x^{1-u/2}}{1-u/2}\right]_0^T = \frac{1}{1-u/2}\cdot\frac{1}{T^{u/2}},
$$

$$
\frac{1}{T}\sum_{t=1}^{T} t^{-u} \leq \frac{1}{T}\int_{t=0}^{T} x^{-u}dx = \frac{1}{T}\left[\frac{x^{1-u}}{1-u}\right]_0^T = \frac{1}{1-u}\cdot\frac{1}{T^u}.
$$

In (iii) we plug in the descent rate $\eta_z = \frac{1}{T^u}$ and upper bound the diameter $D^{(0)} \leq D^z = \operatorname{diam}(\mathcal{R}) + \operatorname{diam}(\mathcal{B})$ by the diameters of the reward class and of the Lagrange multiplier box. This leaves us with a convergence rate of the form:

$$
C_1 T^{u-1} + C_2 T^{-u} + C_3 T^{-u/2},
$$

to get the fastest convergence rate we need to pick:

$$
u^* = \arg\min_u \Big(\max\big\{u-1, -u, -u/22\big\}\Big) = 2/3,
$$

hence we get the final rate:

$$
\frac{1}{T}\sum_{t=0}^{T-1} f(\boldsymbol{z}^{(t)}, \boldsymbol{\theta}^{(t)}) - f^* \leq C_{\text{NPG-CIRL}} T^{-1/3},
$$

where:

$$
C_{\text{NPG-CIRL}} = +\left(\frac{D_z + (B_w + \|\boldsymbol{b}\|_2 + \|\Psi\|)^2}{2} + \frac{C_\pi}{\eta_\theta(1-u)}\right) + \frac{2D_{\boldsymbol{z}}B_a\sqrt{nC_\pi}}{(1-\gamma)(1-u/2)}
$$

□

5.5 Auxiliary lemmas and useful claims for the proof of CIRL Convergence

Assumption 5.1 (Reward gradients are bounded)**.** *We require that* $\forall \boldsymbol{w} \in dom(\boldsymbol{w})$ *the following is true:*

$$\|\frac{\partial \boldsymbol{r}_w}{\partial w}\|_2 \leq B_w,$$

this is trivially satisfied when consider linear reward classes, specifically we have $\boldsymbol{r}_w = \Phi \boldsymbol{w}$ *and thus:*

$$\frac{\partial \boldsymbol{r}_w}{\partial w} = \Phi,$$

and thus $B_w = \|\Phi\|$ *where* $\|\cdot\|$ *denotes the spectral norm.*

Claim 5.2. *For any two policies* $\boldsymbol{\pi}, \bar{\boldsymbol{\pi}} \in \Pi$ *we have that:*

$$\frac{1}{4}\|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2^2 \leq n \cdot \|\log \boldsymbol{\pi} - \log \bar{\boldsymbol{\pi}}\|_\infty,$$

where log *are vectorized and* $n = |S|$ *is the number of states.*

Proof.

$$\begin{aligned}
\frac{1}{4}\|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_2^2 &\leq n\frac{1}{4}\|\boldsymbol{\pi} - \bar{\boldsymbol{\pi}}\|_1^2 && \|\cdot\|_2 \leq \|\cdot\|_1 \\
&= n\frac{1}{4}\left\| \begin{bmatrix} \boldsymbol{\pi}(\cdot|s_1) \\ \vdots \\ \boldsymbol{\pi}(\cdot|s_n) \end{bmatrix} - \begin{bmatrix} \bar{\boldsymbol{\pi}}(\cdot|s_1) \\ \vdots \\ \bar{\boldsymbol{\pi}}(\cdot|s_n) \end{bmatrix} \right\|_1^2 && \text{isolate distribs.} \\
&= \frac{1}{4}\Big(\sum_{s\in S} \|\boldsymbol{\pi}(\cdot|s) - \bar{\boldsymbol{\pi}}(\cdot|s)\|_1 \Big)^2 && \text{def of } \|\cdot\|_1 \\
&= \sum_{s\in S} \frac{1}{2}\|\boldsymbol{\pi}(\cdot|s) - \bar{\boldsymbol{\pi}}(\cdot|s)\|_1^2 && (a+b)^2 \leq 2(a^2+b^2) \\
&\leq \sum_{s\in S} D_{\mathrm{KL}}(\boldsymbol{\pi}(\cdot|s)||\bar{\boldsymbol{\pi}}(\cdot|s)) && \frac{1}{2}\|\boldsymbol{p}_1 - \boldsymbol{p}_2\|_1^2 \leq D(\boldsymbol{p}_1||\boldsymbol{p}_2) \\
&= \sum_{s\in S} |\boldsymbol{\pi}(\cdot|s)^T(\log \boldsymbol{\pi}(\cdot|s) - \log \bar{\boldsymbol{\pi}}(\cdot|s))| && \text{def. of KL} \\
&\leq \sum_{s\in S} \overbrace{\|\boldsymbol{\pi}(\cdot|s)\|_1}^{=1} \cdot \|\log \boldsymbol{\pi}(\cdot|s) - \log \bar{\boldsymbol{\pi}}(\cdot|s)\|_\infty && \text{Hödler's ineq.} \\
&\leq \sum_{s\in S} \|\log \boldsymbol{\pi}(\cdot|s) - \log \bar{\boldsymbol{\pi}}(\cdot|s)\|_\infty \\
&\leq n \cdot \|\log \boldsymbol{\pi} - \log \bar{\boldsymbol{\pi}}\|_\infty
\end{aligned}$$

The proof hinges on the inequality:

$$\frac{1}{2}\|\boldsymbol{p}_1 - \boldsymbol{p}_2\|_1^2 \leq D(\boldsymbol{p}_1||\boldsymbol{p}_2),$$

which can be found in the Theorem 11.6 of Cover and Thomas 2006. □

Claim 5.3. *For any two vectors $v_1, v_2 \in \mathbb{R}^n$ we have that:*

$$|\log\|\exp v_1\|_1 - \log\|\exp v_2\|_1| \leq \|v_1 - v_2\|_\infty,$$

where $\exp\cdot$ *is applied element-wise.*

Proof. We start by using that the mean-value theorem states that there exists a vector v_c which is a convex combination of v_1 and v_2 s.t. the following equality is verified:

$$|\log\big(\|\exp v_1\|_1\big) - \log\big(\|\exp v_2\|_1\big)| = |\big\langle v_1 - v_2, \nabla_v \log(\exp v)|_{v=v_c}\big\rangle|,$$

and then it's just a matter of bounding the right terms:

$$\begin{aligned}
&|\log\big(\|\exp v_1\|_1\big) - \log\big(\|\exp v_2\|_1\big)| \\
=&|\big\langle v_1 - v_2, \nabla_v \log(\exp v)|_{v=v_c}\big\rangle| && \text{by the mean value theorem} \\
\leq&\|v_1 - v_2\|_\infty \|\nabla_v \log(\exp v)|_{v=v_c}\|_1 && \text{Hodler's inequality} \\
=&\|v_1 - v_2\|_\infty && \|\nabla_v \log(\exp v)|_{v=v_c}\|_1 = \left\|\frac{\exp(v_c)}{\|v_c\|_1}\right\|_1 = 1.
\end{aligned}$$

□

Claim 5.4 (Bounded gradients of the objective function with respect to the Lagrande and reward parameters)**.** *For any Shannon regularized CMDP, the following holds:*

$$\begin{aligned}
\|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\theta})\|_2 &\leq 2B_{\boldsymbol{w}} \\
\|\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\theta})\|_2 &\leq \|\Psi\|.
\end{aligned}$$

Proof. This is just a matter of calculating the gradients explicitly and using the bounds that we know are true:

$$\begin{aligned}
\|\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\theta})\|_2 &= \Big\|\nabla_{\boldsymbol{w}}\Big(\langle \boldsymbol{r}, \boldsymbol{\mu} - \boldsymbol{\mu}^E\rangle \\
&\quad - (\tilde{\Omega}(\boldsymbol{\mu}) - \tilde{\Omega}(\boldsymbol{\mu}^E)) + \langle \boldsymbol{\lambda}, \boldsymbol{b} - \Psi^T \boldsymbol{\mu}\rangle\Big)\Big\|_2 \\
&= \left\|(\boldsymbol{\mu} - \boldsymbol{\mu}^E)\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{w}}\right\|_2 \\
&\leq \|(\boldsymbol{\mu} - \boldsymbol{\mu}^E)\|_2 \cdot \|\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{w}}\|_2 && \text{triangle ineq} \\
&\leq \|(\boldsymbol{\mu} - \boldsymbol{\mu}^E)\|_2 \cdot B_w && \text{assumption 5.1} \\
&\leq 2 \sup_{\boldsymbol{\mu}} \|\boldsymbol{\mu}\|_2 \cdot B_w = 2B_w. && \mu \in \Delta_{A\times S}
\end{aligned}$$

$$\begin{aligned}
\|\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\theta})\|_2 &= \Big\|\nabla_{\boldsymbol{\lambda}}\Big(\langle \boldsymbol{r}, \boldsymbol{\mu} - \boldsymbol{\mu}^E\rangle \\
&\quad - (\tilde{\Omega}(\boldsymbol{\mu}) - \tilde{\Omega}(\boldsymbol{\mu}^E)) + \langle \boldsymbol{\lambda}, \boldsymbol{b} - \Psi^T \boldsymbol{\mu}\rangle\Big)\Big\|_2 \\
&= \big\|\Psi^T \boldsymbol{\mu}\big\|_2 \\
&\leq \|\Psi\|. && \text{spectral norm def.}
\end{aligned}$$

□

Lemma 5.6 (Optimal Q-values are Lipschitz in reward and Lagrange parameters)**.** *Consider the optimal Q values induced by two different rewards in a constrained, Shannon-regularized Markov decision process, we have that, assuming assumption 5.1 is satisfied, for any $\boldsymbol{w}_1, \boldsymbol{w_2} \in dom(\boldsymbol{w})$ and any $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathcal{B}$, the inequality below holds:*

$$\|\boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}_1} - \boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}_2}\|_\infty \leq \frac{B_w}{1-\gamma}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + \frac{\|\Psi\|}{1-\gamma}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2.$$

Proof. We follow a derivation very close to the one of Lemma 5.3 in Zeng et al. 2022. We show lipschitzness by bounding the gradient with respect to parameters $\boldsymbol{\lambda}$ and $\boldsymbol{w}$, since the derivation is the same for both parameters, we write $\nabla_\lozenge$ for the gradient with respect to either $\boldsymbol{\lambda}$ or $\boldsymbol{w}$. We start from the definition of the optimal Q value associated with diminished reward $\tilde{\boldsymbol{r}}$ for some $(s, a) \in S \times A$s state action pair:

$$\begin{aligned}
\nabla_\lozenge Q^*_{\tilde{\boldsymbol{r}}}(s_0, a_0) &= \nabla_\lozenge \Big[\tilde{r}_{w,\lambda}(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P^\pi(s_0,a_0)}\big[V^*_{\tilde{\boldsymbol{r}}}(s)(s_1)\big]\Big] \\
&\overset{(i)}{=} \nabla_\lozenge \tilde{r}_{w,\lambda}(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P^\pi(s_0,a_0)}\Big[\nabla_\lozenge \log \big\| \exp \boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}}(s_1, \cdot)\big\|_1\Big] \\
&= \nabla_\lozenge \tilde{r}_{w,\lambda}(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P^\pi(s_0,a_0)}\left[\sum_a \frac{Q^*_{\tilde{\boldsymbol{r}}}}{\sum_{a'} Q^*_{\tilde{\boldsymbol{r}}}(s_1, a')} \nabla_\lozenge Q^*_{\tilde{\boldsymbol{r}}}(s_1, a)\right] \\
&= \nabla_\lozenge \tilde{r}_{w,\lambda}(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P^\pi(s_0,a_0)}\Big[\sum_a \tilde{\pi}(a|s_1) \nabla_\lozenge Q^*_{\tilde{\boldsymbol{r}}}(s_1, a)\Big] \\
&= \nabla_\lozenge \tilde{r}_{w,\lambda}(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P^\pi(s_0,a_0)\ , a_1 \sim \tilde{\pi}}\Big[\nabla_\lozenge Q^*_{\tilde{\boldsymbol{r}}}(s_1, a_1)\Big] \\
&\overset{(ii)}{=} \mathbb{E}_{\tau \sim \tilde{\pi}}\Big[\sum_{t=0}^{+\infty} \gamma^t \nabla_\lozenge \tilde{\boldsymbol{r}}^{(t)}_{w,\lambda}\Big].
\end{aligned}$$

We use (ii) that the V value is the average across Q values under policy distribution and simplify until we (ii) identify a recursion. Since we only care about gradient norms we have that:

$$\begin{aligned}
\big\|\nabla_\lozenge Q^*_{\tilde{\boldsymbol{r}}}(s, a)\big\|_2 &= \left\|\nabla_\lozenge \mathbb{E}_{\tau \sim \tilde{\pi}}\left[\sum_{t=0}^{+\infty} \gamma^t \nabla_\lozenge \tilde{\boldsymbol{r}}^{(t)}_{w,\lambda} \middle| s_0 = s, a_0 = a\right]\right\|_2 \\
&= \mathbb{E}_{\tau \sim \tilde{\pi}}\left[\sum_{t=0}^{+\infty} \gamma^t \overbrace{\|\nabla_\lozenge \tilde{\boldsymbol{r}}^{(t)}_{w,\lambda}\|_2}^{\leq B_\lozenge} \middle| s_0 = s, a_0 = a\right] && \text{Jensen's ineq.} \\
&\leq \mathbb{E}_{\tau \sim \tilde{\pi}}\left[\sum_{t=0}^{+\infty} \gamma^t B_\lozenge \middle| s_0 = s, a_0 = a\right] && \text{Bouded } \tilde{r} \text{ gradients.} \\
&= \frac{B_\lozenge}{1-\gamma}
\end{aligned}$$

Now using assumption 5.1 together with the discounted reward we clearly see that the diminished reward $\tilde{\boldsymbol{r}} = \boldsymbol{r}_w - \Psi\boldsymbol{\lambda}$ has bounded gradients in $\boldsymbol{\lambda}$ and $\boldsymbol{w}$:

$$\begin{aligned}
\|\nabla_w \tilde{\boldsymbol{r}}\| = \|\nabla_w \boldsymbol{r}\| \leq B_w && \text{assumption 5.1} \\
\|\nabla_\lambda \tilde{\boldsymbol{r}}\| \leq \|\Psi\| && \text{by def of } \tilde{r}.
\end{aligned}$$

Which completes the proof. (The bound established for any state action pair trivially bounds the $\|\cdot\|_\infty$ norm.) $\square$

Lemma 5.7 (Soft suboptimality)**.** *Consider an SPI step ($\eta_\theta = \frac{1}{\beta}$), and one has:*

$$\tilde{f}(\boldsymbol{w}, \boldsymbol{\lambda}) - f(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\theta}^{(t)}) \leq \frac{1}{\eta_\theta} \mathbb{E}_{s\sim\tilde{\boldsymbol{\mu}}} \left[D_{KL}\Big(\boldsymbol{\pi}^{(t)}(\cdot|s) || \boldsymbol{\pi}^{(t+1)}(\cdot|s)\Big) \right].$$

Refer to Mei et al. 2020 for proof.

Lemma 5.8 (Q-values are Lipschitz in reward and Lagrange parameters for fixed policy)**.** *Consider the Q values induced by the same policy on two different rewards in a constrained, Shannon-regularized Markov decision process, we have that, assuming assumption 5.1 is satisfied, for any $\boldsymbol{w}_1, \boldsymbol{w_2} \in dom(\boldsymbol{w})$ and any $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathcal{B}$, the inequality below holds:*

$$\|\boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}_1} - \boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}_2}\|_\infty \leq \frac{B_w}{1-\gamma}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + \frac{\|\Psi\|}{1-\gamma}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2.$$

Proof. We follow a derivation very close to the one of Lemma B.3 in Zeng et al. 2022. We show lipschitzness by bounding the gradient with respect to parameters $\boldsymbol{\lambda}$ and $\boldsymbol{w}$, since the derivation is the same for both parameters, we write $\nabla_\lozenge$ for the gradient with respect to either $\boldsymbol{\lambda}$ or $\boldsymbol{w}$. We start by writing out the expectation form of the Q-value, for some policy π:

$$Q^\pi_{\tilde{r}}(s,a) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{+\infty} \gamma^t (\tilde{r}(s_t,a_t) + H(\boldsymbol{\pi}(\cdot|s_t))) \middle| s_0 = s, a_0 = a \right].$$

Taking the difference for two different diminished rewards $\tilde{\boldsymbol{r}}_1$ and $\tilde{\boldsymbol{r}}_2$ we have:

$$\begin{aligned}
|Q^\pi_{\tilde{r}_1}(s,a) - Q^\pi_{\tilde{r}_2}(s,a)| = \Bigg| & \mathbb{E}_{\tau\sim\pi}\Bigg[\sum_{t=0}^{+\infty} \gamma^t (\tilde{r}_1(s_t,a_t) - \tilde{r}_2(s_t,a_t) \\
& + H(\boldsymbol{\pi}(\cdot|s_t)) - H(\boldsymbol{\pi}(\cdot|s_t))) \Big| s_0 = s, a_0 = a \Bigg] \Bigg| \\
\overset{(i)}{=} & \left| \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{+\infty} \gamma^t (\tilde{r}_1(s_t,a_t) - \tilde{r}_2(s_t,a_t)) \middle| s_0 = s, a_0 = a \right] \right| \\
\overset{(ii)}{\leq} & \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{+\infty} \gamma^t \big|\tilde{r}_1(s_t,a_t) - \tilde{r}_2(s_t,a_t)\big| \middle| s_0 = s, a_0 = a \right] \\
\overset{(iii)}{\leq} & \mathbb{E}_{\tau\sim\pi}\Bigg[\sum_{t=0}^{+\infty} \gamma^t B_w \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 \\
& + \|\Psi\|\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \Big| s_0 = s, a_0 = a \Bigg] \\
\overset{(iv)}{=} & \frac{B_w}{1-\gamma}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + \frac{\|\Psi\|}{1-\gamma}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2
\end{aligned}$$

In (i) the regularizers cancel out, then in (ii) we use Jensen's inequality, in (iii) we just plug in the lipschitzness of the reward parametrization (assumption 5.1 and by definition of the diminished reward). Finally in (iv) we observe that what is under the expectation is deterministic and that the sum is a geometric series. Since the bound established for any state action pair trivially bounds the $\|\cdot\|_\infty$ norm this completes the proof. □

Lemma 5.9 ($\tilde{f}$ is convex)**.** *Proof.* This is a classical convex analysis result that can be found in Boyd and Vandenberghe 2004. Consider $f(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\theta})$ convex in $\boldsymbol{z} = [\boldsymbol{w}, \boldsymbol{\lambda}]$, assuming that $\tilde{\boldsymbol{\theta}} := \arg\max_\theta f(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\theta})$ for fixed $\boldsymbol{z}$, and $\tilde{f}(\boldsymbol{z}) := f(\boldsymbol{w}, \boldsymbol{\lambda}, \tilde{\boldsymbol{\theta}})$ Starting with the definition of convexity we have that, for $\gamma \in [0, 1]$ and any $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathrm{dom}(\boldsymbol{z})$:

$$\begin{aligned}
\gamma\tilde{f}(\boldsymbol{z}_1) + (1-\gamma)\tilde{f}(\boldsymbol{z}_2) &= \max_\theta f(\gamma\boldsymbol{z}_1, \theta) + \max_\theta f((1-\gamma)\boldsymbol{z}_2, \theta) \\
&\geq \max_\theta \Big(f(\gamma\boldsymbol{z}_1, \theta) + f((1-\gamma)\boldsymbol{z}_2, \theta) \Big) \\
&\geq \max_\theta f(\gamma\boldsymbol{z}_1 + (1-\gamma)\boldsymbol{z}_2, \theta) \\
&= \max_\theta \tilde{f}(\gamma\boldsymbol{z}_1 + (1-\gamma)\boldsymbol{z}_2).
\end{aligned}$$

Which verifies by the convexity of f, $\tilde{f}$ indeed satisfies the definition of convexity. □

6 Sample complexity - the stochastic gradient setting

In the following section we discuss the generalization of results to the situation in which we only have access to inexact gradients. More specifically,

discuss the dataset and simulation access

The stochastic generalization of our algorithm thus takes the form of alg 7.

Algorithm 7: NPG-CIRL: Natural Policy Gradient CIRL (Sampled Gradients)

Set the learning rates $\eta_\theta = \frac{1}{\beta}$, $\eta_z = \frac{1}{T^u}$
Initialize the algorithm at some point $(\boldsymbol{w}^{(0)}, \boldsymbol{\lambda}^{(0)}, \boldsymbol{\theta}^{(0)})$
Compute the feature expectation $\hat{\boldsymbol{\Upsilon}}_H$ from the expert dataset D^E.
for *iteration* $t = 1, 2, ..., T$ **do**
  Sample a batch of trajectories.
  for *batch* $i = 1, 2, ..., B$ **do**
    draw $s_0^{(i)} \sim \nu_0$
    for *step* $k = 1, ...K$ **do**
      pick $a_{k-1}^{(i)} \sim \pi(\cdot|s_{k-1}^{(i)}, \boldsymbol{\theta}_t)$
      draw $s_k^{(i)} \sim P(s_k^{(i)}|s_{k-1}^{(i)}, a_{k-1}^{(i)})$
      $r_k^{(i)} \leftarrow r(s_k^i, a_k^i)$
      $\boldsymbol{k}_k^{(i)} \leftarrow \boldsymbol{k}(s_k^i, a_k^i)$
    end
  end
  Update the gradient estimates using the trajectories $\left\{\{s_k^i, a_k^i, r_k^i, \boldsymbol{k}_k^{(i)}\}_{k=1}^K\right\}_{i=1}^B$.
  Compute the reward gradient estimate $\hat{\boldsymbol{g}}_w \leftarrow \hat{\nabla}_{\boldsymbol{w}} f(\boldsymbol{w}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{\theta}^{(k)})$
  Compute the Lagrangian gradient estimate $\hat{\boldsymbol{g}}_\lambda \leftarrow \nabla_{\boldsymbol{\lambda}} f(\boldsymbol{w}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{\theta}^{(k)})$
  Compute the policy gradient estimate $\hat{\boldsymbol{g}}_\theta \leftarrow \hat{\nabla}_{\boldsymbol{\theta}} f(\boldsymbol{w}^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{\theta}^{(k)})$
  $\boldsymbol{w}^{(k)} \leftarrow \Pi_{\mathrm{dom}(\boldsymbol{w})}\left(\boldsymbol{r}_k - \eta_z \hat{\boldsymbol{g}}_w\right)$
  $\boldsymbol{\lambda}^{(k+1)} \leftarrow \Pi_{\boldsymbol{\lambda} \in \mathcal{B}}\left(\boldsymbol{\lambda}^{(k)} - \eta_z \hat{\boldsymbol{g}}_\lambda\right)$
  $\boldsymbol{\theta}^{(k+1)} \leftarrow \boldsymbol{\theta}^{(k)} + \eta_\theta (\mathfrak{F}^{\boldsymbol{\theta}})^\dagger \hat{\boldsymbol{g}}_\theta$
end
return $(\boldsymbol{w}^{(K)}, \boldsymbol{\lambda}^{(K)}, \boldsymbol{\theta}^{(K)})$.

1. What gradient estimator do we use for the NPG steps?
   (a) Monte-Carlo (GOMDP) ?
   (b) TD (Actor-Critic)?
2. What gradient estimator do we use for the reward gradient estimate ? $\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\theta}) = (\boldsymbol{\mu} - \boldsymbol{\mu}^E)\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{w}}$ using a feature expectation Monte-Carlo estimator.
3. What gradient estimator do we use for the Lagrangian gradient estimate ? $\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\theta}) = (\boldsymbol{b} - \boldsymbol{K}(\boldsymbol{\theta}))$ the fastest convergence probably would be TD on $\boldsymbol{K}$.

And most importantly, how do we tie all of this together? Probably the same way as in the non-noisy case, but the proof will be long.....

6.1 Dealing with a biased gradient estimator, convergence in expectation

Most practical policy gradient estimators introduce some kind of bias, Most often as a consequence of the truncation of the infinite sum in the expected discounted reward that necessarily occurs when sampling a finite time horizon:

$$\hat{J} - J = \frac{1}{B}\sum_{t=1}^{B}\sum_{t=0}^{H-1}\gamma^t r(s_t, a_t) - \mathbb{E}\left[\sum_{t=0}^{+\infty}\gamma^t r(s_t, a_t)\right],$$

$$\mathbb{E}\big[\hat{J} - J\big] = \mathbb{E}\left[\frac{1}{B}\sum_{t=1}^{B}\sum_{t=0}^{H-1}\gamma^t r(s_t, a_t) - \mathbb{E}\Big[\sum_{t=0}^{+\infty}\gamma^t r(s_t, a_t)\Big]\right],$$

$$= \mathbb{E}\left[\sum_{t=0}^{H-1}\gamma^t(r(s_t, a_t) - r(s_t, a_t)) - \sum_{t=H}^{+\infty}\gamma^t r(s_t, a_t)\right] = -\mathbb{E}\left[\sum_{t=H}^{+\infty}\gamma^t r(s_t, a_t)\right].$$

This bias can actually be eliminated (by choosing randomly sampled horizon length, as proposed and analyzed in Y. Ding, Zhang, and Lavaei 2021), but since getting an unbiased estimator comes at a high sample complexity cost, it makes more sense (if possible) to show that the optimization procedure is robust to small biases, which is what we will show next.

Showing our result will take a few assumptions:

Assumption 6.1 (Bounded variance gradient estimator)**.**

$$\mathbb{E}\Big[\|\hat{\nabla}_\theta f - \nabla_\theta f\|_2^2\Big] \leq \sigma^2$$

Assumption 6.2 (Bounded bias gradient estimator)**.**

$$\|\mathbb{E}\big[\hat{\nabla}_\theta f\big] - \nabla_\theta f\|_2 \leq \delta$$

Lemma 6.1 (Impact of perturbed gradient on policy updates)**.** *Using perturbed gradients of the form $\hat{\nabla}_\theta f = \nabla_\theta f + \delta$ yields updates of the form:*

$$\pi^{(t+1)}(a|s) = \big(\pi^{(t)}(a|s)\big)^{1-\eta\beta}\exp\Big(\eta\hat{Q}^{\pi^\theta}(s, a)\Big),$$

where $\hat{Q}^{\pi^\theta}(s, a) = Q^{\pi^\theta}(s, a) + \frac{\delta(s,a)}{\mu^\theta(s,a)}$.

Proof. The proof is deferred to section 6.2.1. □

6.1.1 Local convergence in expectation of sampled NPG-CIRL

We will show the following lemma :

Lemma 6.2 (Local convergence of stochastic NPG-CIRL)**.** *The NPG-CIRL error decreases at a rate of:*

$$\begin{aligned}\mathbb{E}\Big[\|\log\boldsymbol{\pi}^{(t+1)} - \log\tilde{\boldsymbol{\pi}}^{(t)}\|_\infty\Big] &\leq 2\mathbb{E}\Big[\|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^{*}_{\tilde{\boldsymbol{r}}^{(t)}}\|_\infty\Big] \\ &\leq C_1(1-(1-\gamma)\eta\beta)^t + C_2\delta + C_3\eta_z,\end{aligned} \tag{6.1}$$

where:

$$C_1 = \gamma \frac{(\alpha-1)x_1^{(0)} + x_2^{(0)} + x_3^{(0)}}{\alpha + \alpha\gamma - \gamma}$$
$$C_2 = \gamma \frac{2 + 2\gamma/(1-\gamma) - 2\alpha}{(\alpha-1)(\gamma-1)}$$
$$C_3 = \gamma \frac{2C\alpha + \delta(2-2\alpha)}{(\alpha-1)(\gamma-1)}$$

Proof. Inspired by the linear inequality system analysis developed for the entropy regularized NPG algorithm in Cen et al. 2021 we write out a set of inequalities of our own to analyze the robustness of our algorithm. We let:

$$\alpha = 1 - \eta\beta$$

and define a well-chose auxiliary sequence $\left\{\boldsymbol{\xi}^{(t)}\right\}_{t=0}^{T-1}$ recursively as follows,:

$$\boldsymbol{\xi}^{(0)}(s,a) := \|\boldsymbol{Q}^*_{\tilde{r}^{(t)}}(s,\cdot)/\beta\|_1 \cdot \pi^0(a|s) \tag{6.2}$$
$$\boldsymbol{\xi}^{(t+1)}(s,a) := \left(\boldsymbol{\xi}^{(t)}(s,a)\right)^{\alpha} \exp\left(\frac{1-\alpha}{\beta}\hat{Q}^{(t)}_{\tilde{r}^{(t)}}(s,a)\right). \tag{6.3}$$

We will consider three terms:

$$x_1^{(t)} := \|\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}}\|_\infty, \tag{6.4}$$
$$x_2^{(t)} := \|\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \beta\log\boldsymbol{\xi}^{(t)}\|_\infty, \tag{6.5}$$
$$x_3^{(t)} := -\min_{s,a}\Big(\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}}(s,a) - \beta\log\boldsymbol{\xi}^{(t)}(s,a)\Big). \tag{6.6}$$

We will will show interdependence between the sequences $\{x_1^{(t)}\}_{t=0}^T$, $\{x_2^{(t)}\}_{t=0}^T$ and $\{x_3^{(t)}\}_{t=0}^T$ and then show that they all converge to 0. To do so, will have to study the evolution of all three sequences, we start with the term $x_2^{(t+1)} = \|\boldsymbol{Q}^*_{\tilde{r}^{(t+1)}} - \beta\log\boldsymbol{\xi}^{(t+1)}\|_\infty$. To do so we look at the vector inside of the norm and study it's value a $t+1$:

$$\begin{aligned}
\boldsymbol{Q}^*_{\tilde{r}^{(t+1)}} - \beta\log\boldsymbol{\xi}^{(t+1)} &\overset{(i)}{=} \boldsymbol{Q}^*_{\tilde{r}^{(t+1)}} - \alpha\beta\log\boldsymbol{\xi}^{(t)} - (1-\alpha)\hat{\boldsymbol{Q}}^{(t)}_{\tilde{r}^{(t)}} \\
&= \alpha\Big(\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \beta\log\boldsymbol{\xi}^{(t)}\Big) - \alpha\boldsymbol{Q}^*_{\tilde{r}^{(t)}} \\
&\quad - (1-\alpha)\hat{\boldsymbol{Q}}^{(t)}_{\tilde{r}^{(t)}} + \boldsymbol{Q}^*_{\tilde{r}^{(t+1)}} \\
&= \alpha\Big(\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \beta\log\boldsymbol{\xi}^{(t)}\Big) \\
&\quad + (1-\alpha)\Big(\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \hat{\boldsymbol{Q}}^{(t)}_{\tilde{r}^{(t)}}\Big) + \Big(\boldsymbol{Q}^*_{\tilde{r}^{(t+1)}} - \boldsymbol{Q}^*_{\tilde{r}^{(t)}}\Big) \\
&= \alpha\Big(\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \beta\log\boldsymbol{\xi}^{(t)}\Big) \\
&\quad + (1-\alpha)\Big(\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}}\Big) + (1-\alpha)\Big(\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \hat{\boldsymbol{Q}}^{(t)}_{\tilde{r}^{(t)}}\Big) \\
&\quad + \Big(\boldsymbol{Q}^*_{\tilde{r}^{(t+1)}} - \boldsymbol{Q}^*_{\tilde{r}^{(t)}}\Big)
\end{aligned}$$

Where (i) stems from the definition of the auxiliary sequence recursion (6.3) and the following steps are all obtained by adding terms that cancel out each other until the quantities we care about appear. Taking expectations on both sides (here we take expectations with respect to the stochasticity in the step, variables at time $(t-1)$ are considered deterministic) we get:

$$\begin{aligned}\mathbb{E}\Big[\boldsymbol{Q}^*_{\tilde{r}^{(t+1)}} - \beta \log \boldsymbol{\xi}^{(t+1)}\Big] &= \alpha \boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \beta \log \boldsymbol{\xi}^{(t)} \\ &+ (1-\alpha)\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} \\ &+ (1-\alpha)\mathbb{E}\Big[\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \hat{\boldsymbol{Q}}^{(t)}_{\tilde{r}^{(t)}}\Big] \\ &+ \mathbb{E}\Big[\boldsymbol{Q}^*_{\tilde{r}^{(t+1)}} - \boldsymbol{Q}^*_{\tilde{r}^{(t)}}\Big].\end{aligned}$$

Using the triangle inequality on infinity norms we get:

$$\begin{aligned}\mathbb{E}\Big[\|\boldsymbol{Q}^*_{\tilde{r}^{(t+1)}} - \beta \log \boldsymbol{\xi}^{(t+1)}\|_\infty\Big] &\leq \alpha \overbrace{\|\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \beta \log \boldsymbol{\xi}^{(t)}\|_\infty}^{:=x_2^{(t)}} \\ &+ (1-\alpha)\overbrace{\|\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}}\|_\infty}^{:=x_1^{(t)}} \\ &+ (1-\alpha)\mathbb{E}\Big[\|\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \hat{\boldsymbol{Q}}^{(t)}_{\tilde{r}^{(t)}}\|_\infty\Big] \\ &+ \mathbb{E}\Big[\|\boldsymbol{Q}^*_{\tilde{r}^{(t+1)}} - \boldsymbol{Q}^*_{\tilde{r}^{(t)}}\|_\infty\Big].\end{aligned}$$

Which, plugging in the definition of the terms from (6.4), can be written as:

$$x_2^{(t+1)} \leq \alpha x_2^{(t)} + (1-\alpha)x_1^{(t)} + (1-\alpha)\mathbb{E}\Big[\|\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \hat{\boldsymbol{Q}}^{(t)}_{\tilde{r}^{(t)}}\|_\infty\Big] + \mathbb{E}\Big[\|\boldsymbol{Q}^*_{\tilde{r}^{(t+1)}} - \boldsymbol{Q}^*_{\tilde{r}^{(t)}}\|_\infty\Big].$$

Because of lemma 5.6 we can bound $\mathbb{E}\big[\|\boldsymbol{Q}^*_{\tilde{r}^{(t+1)}} - \boldsymbol{Q}^*_{\tilde{r}^{(t)}}\|_\infty\big] \leq \eta_z C_z$ (where C_z can be computed from the parameter set bounded diameter and from the lipschitz constants of the dual) and using lemma 6.1 we have: $\mathbb{E}\big[\|\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \hat{\boldsymbol{Q}}^{(t)}_{\tilde{r}^{(t)}}\|_\infty\big] \leq \delta_{\mathcal{F}}$. We thus have:

$$\mathbb{E}\big[x_2^{(t+1)}\big] \leq (1-\alpha)x_1^{(t)} + \alpha x_2^{(t)} + (1-\alpha)\delta_{\mathcal{F}} + C_z\eta_z. \tag{6.7}$$

Now in an effort to study $x_3^{(t)}$, we consider the vector $-\Big(\boldsymbol{Q}^{(t+1)}_{\tilde{r}^{(t+1)}} - \beta \log \boldsymbol{\xi}^{(t+1)}\Big)$:

$$\begin{aligned}-\Big(\boldsymbol{Q}^{(t+1)}_{\tilde{r}^{(t+1)}} - \beta \log \boldsymbol{\xi}^{(t+1)}\Big) &= -\boldsymbol{Q}^{(t+1)}_{\tilde{r}^{(t+1)}} + \alpha\beta \log \boldsymbol{\xi}^{(t)} + (1-\alpha)\hat{\boldsymbol{Q}}^{(t)}_{\tilde{r}^{(t)}} \\ &= -\boldsymbol{Q}^{(t+1)}_{\tilde{r}^{(t+1)}} - \alpha\Big(\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \beta \log \boldsymbol{\xi}^{(t)}\Big) + (1-\alpha)\hat{\boldsymbol{Q}}^{(t)}_{\tilde{r}^{(t)}} + \alpha\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} \\ &= -\alpha\Big(\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \beta \log \boldsymbol{\xi}^{(t)}\Big) + (1-\alpha)\Big(\hat{\boldsymbol{Q}}^{(t)}_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}}\Big) \\ &+ \Big(\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t+1)}_{\tilde{r}^{(t+1)}}\Big) \\ &= -\alpha\Big(\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \beta \log \boldsymbol{\xi}^{(t)}\Big) + (1-\alpha)\Big(\hat{\boldsymbol{Q}}^{(t)}_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}}\Big) \\ &+ \Big(\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t+1)}_{\tilde{r}^{(t)}}\Big) + \Big(\boldsymbol{Q}^{(t+1)}_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t+1)}_{\tilde{r}^{(t+1)}}\Big)\end{aligned}$$

taking maximums and expectations on both sides we get:

$$\begin{aligned}
&\mathbb{E}\Big[-\min_{s,a}\Big(\boldsymbol{Q}^{(t+1)}_{\tilde{r}^{(t+1)}}(s,a) - \beta\log\boldsymbol{\xi}^{(t+1)}(s,a)\Big)\Big] \\
&\leq \mathbb{E}\Big[-\alpha\min_{s,a}\Big(\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}}(s,a) - \beta\log\boldsymbol{\xi}^{(t)}(s,a)\Big) \\
&\qquad +(1-\alpha)\overbrace{\mathbb{E}\Big[\Big\|\hat{\boldsymbol{Q}}^{(t)}_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}}\Big\|\Big]}^{\leq\delta_{\mathcal{F}}} \\
&\qquad +\overbrace{\mathbb{E}\Big[\Big\|\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t+1)}_{\tilde{r}^{(t)}}\Big\|_\infty\Big]}^{\leq\rho \text{ by approx pd lemmma}} \\
&\qquad +\overbrace{\mathbb{E}\Big[\Big\|\boldsymbol{Q}^{(t+1)}_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t+1)}_{\tilde{r}^{(t+1)}}\Big\|_\infty\Big]}^{\leq C_z\eta_z} \\
&\overset{(i)}{\leq} -\alpha\min_{s,a}\Big(\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \beta\log\boldsymbol{\xi}^{(t)}\Big) \\
&\qquad +(1-\alpha)\delta_{\mathcal{F}} + \rho + C_z\eta_z,
\end{aligned}$$

where (i) uses lemma 6.1 as well as lemma 5.8 we thus have:

$$\mathbb{E}\big[x_3^{(t+1)}\big] \leq \alpha x_3^{(t)} + (1-\alpha)\delta_{\mathcal{F}} + \rho + \eta_z C_z. \tag{6.8}$$

Finally, in order to study the evolution of $x_1^{(t)}$, we first consider the policy step:

$$\pi^{(t+1)}(a|s) := \frac{1}{Z^{(t)}(s)}\big(\pi^{(t)}(a|s)\big)^{1-\eta\beta}\exp\big(\eta\hat{Q}^{(t)}_{\tilde{r}^{(t)}}(s,a)\big), \tag{6.9}$$

together with the auxiliary sequence:

$$\pi^{(t)}(a|s) := \frac{\xi^{(t)}(a|s)}{\|\xi^{(t)}(\cdot|s)\|_1}. \tag{6.10}$$

We can derivate the following inequality:

$$\begin{aligned}
\log\pi^{(t+1)}(a|s) &= \log\xi^{(t+1)}(s,a) - \log\|\xi^{(t+1)}(s,\cdot)\|_1 &(6.11)\\
&= \alpha\log\xi^{(t)}(s,a) + \frac{1-\alpha}{\beta}\hat{Q}^{(t)}_{(t)}(s,a) - \log\|\xi^{(t+1)}(s,\cdot)\|_1. &(6.12)
\end{aligned}$$

Finally in order to study the $x_1^{(t)}$ step, which is the distance from the current Q

iterate to the optimal Q-value one can compute:

$$\begin{aligned}
Q^*_{\tilde{r}^{(t+1)}}(s,a) - Q^{(t+1)}_{\tilde{r}^{(t+1)}}(s,a) &= r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}\Big[V^*_{\tilde{r}^{(t+1)}}(s')\Big] \\
&\quad - r(s,a) - \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}\Big[V^{t+1}_{\tilde{r}^{(t+1)}}(s')\Big] \\
&= \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\beta \log \left\| \exp \frac{Q^*_{\tilde{r}^{(t+1)}}(s',\cdot)}{\beta}\right\|_1\right] \\
&\quad - \gamma \mathbb{E}_{\substack{s'\sim P(\cdot|s,a)\\ a'\sim \pi^{(t+1)}(\cdot|s')}}\left[Q^{(t+1)}_{\tilde{r}^{(t+1)}}(s',a') - \beta \log \pi^{(t+1)}(a'|s')\right] \\
&= \overbrace{\gamma\, \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\beta \log \left\| \exp \frac{Q^*_{\tilde{r}^{(t+1)}}(s',\cdot)}{\beta}\right\|_1 - \beta \log \|\xi^{(t+1)}(s,\cdot)\|_1\right]}^{\text{UB with claim 5.3}} \\
&\quad - \gamma \mathbb{E}_{\substack{s'\sim P(\cdot|s,a)\\ a'\sim \pi^{(t+1)}(\cdot|s')}}\left[Q^{(t+1)}_{\tilde{r}^{(t+1)}}(s',a') - \beta\Big(\alpha \log \xi^{(t)}(s,a) + \frac{1-\alpha}{\beta}\hat{Q}^{(t)}_{(t)}(s,a)\Big)\right].
\end{aligned}$$

And hence (taking expectation over the algorithm step on both sides, we have):

$$\begin{aligned}
&\mathbb{E}\Big[\|Q^*_{\tilde{r}^{(t+1)}}(s,a) - Q^{(t+1)}_{\tilde{r}^{(t+1)}}(s,a)\|_\infty\Big] \\
&\leq \gamma \mathbb{E}\Big[\overbrace{\|\boldsymbol{Q}^*_{\tilde{r}^{(t+1)}} - \beta \log \boldsymbol{\xi}^{(t+1)}\|_\infty}^{\text{Upper bound with (6.7)}} - \gamma \overbrace{\min_{s,a}\Big(Q^{(t+1)}_{\tilde{r}^{t+1}}(s,a) - \beta \log \xi^{(t+1)}(s,a)\Big)}^{\text{Upper bound with (6.8)}}\Big] \\
&\leq \gamma\left(\alpha\|\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \beta \log \boldsymbol{\xi}^{(t)}\|_\infty + (1-\alpha)\|\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}}\|_\infty + (1-\alpha)\delta_{\mathcal{F}} + C_z\eta_z\right) \\
&\quad + \gamma\left(\alpha \min_{s,a}\Big(\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \beta \log \boldsymbol{\xi}^{(t)}\Big) + (1-\alpha)\delta_{\mathcal{F}} + \rho + C_z\eta_z\right) \\
&= \gamma\alpha\|\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \beta \log \boldsymbol{\xi}^{(t)}\|_\infty + \gamma(1-\alpha)\|\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}}\|_\infty \\
&\quad \gamma\alpha - \min_{s,a}\Big(\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \beta \log \boldsymbol{\xi}^{(t)}\Big) + \gamma\Big(2(1-1\alpha)\delta_{\mathcal{F}} + 2C_z\eta_z + \rho\Big).
\end{aligned}$$

Which leaves us with:

$$\begin{aligned}
\mathbb{E}\Big[\|Q^*_{\tilde{r}^{(t+1)}}(s,a) - Q^{(t+1)}_{\tilde{r}^{(t+1)}}(s,a)\|_\infty\Big] &\leq \gamma\alpha\|\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \beta \log \boldsymbol{\xi}^{(t)}\|_\infty \\
&\quad + \gamma(1-\alpha)\|\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}}\|_\infty \\
&\quad - \gamma\alpha \min_{s,a}\Big(\boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}} - \beta \log \boldsymbol{\xi}^{(t)}\Big) \\
&\quad + \gamma\Big(2(1-\alpha)\delta_{\mathcal{F}} + 2C_z + \rho\Big).
\end{aligned} \tag{6.13}$$

Which once we plug in our identified terms, gives us:

$$\mathbb{E}\big[x_1^{(t+1)}\big] \leq \gamma(1-\alpha)x_1^{(t)} + \gamma\alpha x_2^{(t)} + \gamma\alpha x_3^{(t)} + \gamma\Big(2(1-\alpha)\delta_{\mathcal{F}} + 2C_z + \rho\Big), \tag{6.14}$$

We have thus computed three bounds, (6.7), (6.8) and (6.14), which are expressed recursively as functions of the terms, this enables us to rewrite the inequality system

as the following affine system:

$$\overbrace{\mathbb{E}\left[\begin{bmatrix} x_1^{(t+1)} \\ x_2^{(t+1)} \\ x_3^{(t+1)} \end{bmatrix}\right]}^{:=\mathbb{E}[\boldsymbol{x}^{(t+1)}]} \leq \overbrace{\begin{bmatrix} \gamma(1-\alpha) & \gamma\alpha & \gamma\alpha \\ (1-\alpha) & \alpha & 0 \\ 0 & 0 & \alpha \end{bmatrix}}^{:=A} \overbrace{\begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \\ x_3^{(t)} \end{bmatrix}}^{:=\boldsymbol{x}^{(t)}} + \overbrace{\begin{bmatrix} \gamma\Big(2(1-\alpha)\delta_{\mathcal{F}} + 2C_z + \rho\Big) \\ \big((1-\alpha)\delta_{\mathcal{F}} + C_z\big) \\ \big((1-\alpha)\delta_{\mathcal{F}} + \rho + C_z\eta_z\big) \end{bmatrix}}^{:=b}. \tag{6.15}$$

Which we can study using linear system theory. The eigenvalues of the matrix A are the following:

$$\lambda_1 = \alpha + \gamma(1-\alpha) = 1 - (1-\gamma)\eta\beta, \qquad \lambda_2 = \alpha = 1 - \eta\beta, \qquad \lambda_3 = 0,$$

and they are associated with the following eigenvectors:

$$\boldsymbol{v}_1 = \begin{bmatrix} \gamma \\ 1 \\ 0 \end{bmatrix}, \qquad \boldsymbol{v}_2 = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}, \qquad \boldsymbol{v}_3 = \begin{bmatrix} \frac{\alpha}{1-\alpha} \\ 1 \\ 0 \end{bmatrix}.$$

Assume that our system starts at some point $\boldsymbol{x}^{(0)}$, and consider it's representation into the basis formed by the eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$:

$$\boldsymbol{x}_0 = \sum_{i\in[3]} \boldsymbol{v_i} a_i.$$

The iterations of the affine system (6.15) can thus be bounded by:

$$\begin{aligned}
\boldsymbol{x}_t = A\boldsymbol{x}_{t-1} + \boldsymbol{b} &= (A)^t \boldsymbol{x}_t + \sum_{t=0}^{t-1} (A^t)\boldsymbol{b} \\
&= \sum_{i\in[3]} \lambda^t \boldsymbol{v_i} a_i + \sum_{t=0}^{t-1} (A^t)\boldsymbol{b} \\
&\overset{(i)}{\leq} \sum_{i\in[3]} \lambda^t \boldsymbol{v_i} a_i + \sum_{t=0}^{+\infty} (A^t)\boldsymbol{b} \\
&\overset{(ii)}{=} \sum_{i\in[3]} \lambda_i^t \boldsymbol{v_i} a_i + (I-A)^{-1}\boldsymbol{b} \\
&\overset{(iii)}{=} \lambda_1^t \begin{bmatrix} \gamma \\ 1 \\ 0 \end{bmatrix} a_1 + \lambda_2^t \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} a_2 + (I-A)^{-1}\boldsymbol{b}
\end{aligned}$$

where the inequality in (i) is element-wise true because all elements of the vector $\boldsymbol{b}$ are positive and all eigenvalues of A are positive, equality (ii) is just a matrix geometric series. Note that in (iii) we omit the third eigenvector as it's eigenvalues is 0.

Explicit computation tells us that the coefficients a_1, a_2, a_3 are given by:

$$\begin{aligned}
a_1 &= \frac{1}{\alpha + \alpha\gamma - \gamma}\Big((\alpha-1)x_1^{(0)} + x_2^{(0)} + x_3^{(0)}\Big), \\
a_2 &= x_3^{(0)}, \\
a_3 &= \frac{1-\alpha}{\alpha + \alpha\gamma - \gamma}(x_1^{(0)} - \gamma x_2^{(0)} - \gamma x_3^{(0)}).
\end{aligned}$$

Introducing the constants into the linear system, we get:

$$\overbrace{\mathbb{E}\left[\begin{bmatrix} x_1^{(t)} \\ x_2^{(t)} \\ x_3^{(t)} \end{bmatrix}\right]}^{:=\boldsymbol{x}^{(t)}} \leq \frac{(\alpha-1)x_1^{(0)} + x_2^{(0)} + x_3^{(0)}}{\alpha + \alpha\gamma - \gamma}\lambda_1^t \begin{bmatrix} \gamma \\ 1 \\ 0 \end{bmatrix} + x_3^{(0)}\lambda_2^t \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$
$$+\frac{1}{(\alpha-1)(\gamma-1)} \begin{bmatrix} \gamma\Big[2\big(\delta(1-\alpha)(\eta_z+1)\alpha C_z\eta_z\big) + \rho\Big] \\ 2C_z\eta_z\alpha\gamma - C_z\eta_z\gamma + C_z\eta_z - 2\alpha\delta_{\mathbb{F}}\gamma - \alpha\delta_{\mathcal{F}}\gamma - \alpha\delta_{\mathcal{F}} + 2\delta_{\mathcal{F}}\eta_z\gamma + \delta_{\mathcal{F}} + \gamma\rho \\ -(\gamma-1)\cdot(C_z\eta_z - \alpha\delta_{\mathcal{F}} + \delta_{\mathcal{F}} + \rho) \end{bmatrix}$$
$$\leq \frac{(\alpha-1)x_1^{(0)} + x_2^{(0)} + x_3^{(0)}}{\alpha + \alpha\gamma - \gamma}\lambda_1^t \begin{bmatrix} \gamma \\ 1 \\ 0 \end{bmatrix} + x_3^{(0)}\lambda_2^t \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$
$$+\frac{1}{(\alpha-1)(\gamma-1)} \begin{bmatrix} \gamma\Big[2\big(\delta(1-\alpha)(\eta_z+1)\alpha C_z\eta_z\big) + \rho\Big] \\ 2C_z\eta_z\alpha\gamma + C_z\eta_z + 2\delta_{\mathcal{F}}\eta_z\gamma + \delta_{\mathcal{F}} + \gamma\rho \\ 0 \end{bmatrix}$$

We mostly care about the $x_1^{(t)}$ term, so we only explicitly compute this one:

$$\mathbb{E}[x_1^{(t)}] \leq \gamma\frac{(\alpha-1)x_1^{(0)} + x_2^{(0)} + x_3^{(0)}}{\alpha + \alpha\gamma - \gamma}(1-(1-\gamma)\eta\beta)^t + \frac{\gamma\Big[2\big(\delta(1-\alpha)(\eta_z+1)\alpha C_z\eta_z\big) + \rho\Big]}{(\alpha-1)(\gamma-1)}$$
$$\mathbb{E}\Big[\|\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(t)}_{\tilde{r}^{(t)}}\|_\infty\Big] \leq \gamma\frac{(\alpha-1)\|\boldsymbol{Q}^*_{\tilde{r}^{(t)}} - \boldsymbol{Q}^{(0)}_{\tilde{r}^{(0)}}\|_\infty + x_2^{(0)} + x_3^{(0)}}{\alpha + \alpha\gamma - \gamma}(1-(1-\gamma)\eta\beta)^t + \frac{\gamma\Big[2\big(\delta(1-\alpha)(\eta_z+1)\alpha C_z\eta_z\big) + \rho\Big]}{(\alpha-1)(\gamma-1)}.$$

Using that log policy suboptimality is bounded by Q-value sub-optimality and rearranging the terms we get that:

$$\begin{aligned}\mathbb{E}\Big[\|\log\boldsymbol{\pi}^{(t+1)} - \log\tilde{\boldsymbol{\pi}}^{(t)}\|_\infty\Big] &\leq 2\mathbb{E}\Big[\|\boldsymbol{Q}^{(t)}_{\tilde{\boldsymbol{r}}^{(t)}} - \boldsymbol{Q}^*_{\tilde{\boldsymbol{r}}^{(t)}}\|_\infty\Big] \\ &\leq C_1(1-(1-\gamma)\eta\beta)^t + C_2\delta + C_3\eta_z,\end{aligned} \tag{6.16}$$

where:

$$C_1 = \gamma\frac{(\alpha-1)x_1^{(0)} + x_2^{(0)} + x_3^{(0)}}{\alpha + \alpha\gamma - \gamma}$$
$$C_2 = \gamma\frac{2 + 2\gamma/(1-\gamma) - 2\alpha}{(\alpha-1)(\gamma-1)}$$
$$C_3 = \gamma\frac{2C\alpha + \delta(2-2\alpha)}{(\alpha-1)(\gamma-1)}$$

□

6.2 Auxiliary lemmas and useful claims for the proof of Stochastic CIRL Convergence

6.2.1 Proof of lemma 6.1

Proof. Recall that:

$$\boldsymbol{w}_\theta = \big(\mathcal{F}^\theta\big)^\dagger \nabla_\theta J(\theta) = \arg\min_{\boldsymbol{w}\in\mathbb{R}^{S\times A}} \Big\| \mathcal{F}^\theta \mathbf{w} - \nabla_\theta J(\theta) \Big\|_2^2, \tag{6.17}$$

where $\mathcal{F}^\theta$ is the Fisher Information Matrix (FIM), given by:

$$\mathcal{F}^\theta = \mathbb{E}_{(s,a)\sim\mu(\theta)} \Bigg[\big(\nabla_\theta \log \pi^\theta(a|s)\big) \big(\nabla_\theta \log \pi^\theta(a|s)\big)^T \Bigg].$$

To show our result, we consider the perturbed gradient step:

$$\hat{\boldsymbol{w}}_\theta = \big(\mathcal{F}^\theta\big)^\dagger \hat{\nabla}_\theta J(\theta) = \arg\min_{\boldsymbol{w}\in\mathbb{R}^{S\times A}} \Big\| \mathcal{F}^\theta \mathbf{w} - \hat{\nabla}_\theta J(\theta) \Big\|_2^2.$$

Let us look at the matrix-vector product $\mathcal{F}^\theta \boldsymbol{w}$:

$$\begin{aligned}
\big(\mathcal{F}^\theta \boldsymbol{w}\big)(s,a) &= \mathbb{E}_{\substack{s'\sim\mu_s(\theta) \\ a'\sim\pi^\theta(\cdot|s)}} \Bigg[\frac{\partial \log \pi^\theta(a'|s')}{\partial\theta(s,a)} \Big(\sum_{\tilde{a},\tilde{s}} \frac{\partial \log \pi^\theta(a'|s')}{\partial\theta(\tilde{s},\tilde{a})} w_{\tilde{s},\tilde{a}} \Big) \Bigg] \\
&= \mu_s^\theta(s)\pi^\theta(a|s) \Bigg[w_{s,a} - \sum_a \pi^\theta(a|s) w_{s,a} \Bigg].
\end{aligned}$$

Plugging the matrix vector product above into (6.17), we get the following form:

$$\begin{aligned}
\boldsymbol{w}_\theta &= \arg\min_{\boldsymbol{w}\in\mathbb{R}^{S\times A}} \Big\| \mathcal{F}^\theta \mathbf{w} - \nabla_\theta J(\theta) \Big\|_2^2 \\
&\overset{(i)}{=} \arg\min_{\boldsymbol{w}\in\mathbb{R}^{S\times A}} \sum_{s,a} \Bigg(\mu_s^\theta(s)\pi^\theta(a|s) \Big[w_{s,a} - \overbrace{\sum_a \pi^\theta(a|s) w_{s,a}}^{:=c(s)} \Big] - \frac{\partial \hat{J}(\theta)}{\theta(s,a)} \Bigg)^2 \\
&= \arg\min_{\boldsymbol{w}\in\mathbb{R}^{S\times A}} \sum_{s,a} \Bigg(\mu_s^\theta(s)\pi^\theta(a|s) \Big[w_{s,a} - c(s) \Big] - \frac{\partial J(\theta)}{\theta(s,a)} - \delta(s,a) \Bigg)^2 \\
&\overset{(ii)}{=} \arg\min_{\boldsymbol{w}\in\mathbb{R}^{S\times A}} \sum_{s,a} \Bigg(\mu_s^\theta(s)\pi^\theta(a|s) \Big[w_{s,a} - c(s) - A^{\pi^\theta}(s,a) \Big] - \delta(s,a) \Bigg)^2 \\
&\overset{(iii)}{=} \arg\min_{\boldsymbol{w}\in\mathbb{R}^{S\times A}} \sum_{s,a} \Bigg(\mu_s^\theta(s)\pi^\theta(a|s) \Big[w_{s,a} - c(s) - A^{\pi^\theta}(s,a) - \frac{\delta(s,a)}{\mu^\theta(s,a)} \Big] \Bigg)^2.
\end{aligned}$$

In which in (i) we introduce our matrix vector product into the norm, in two (ii) we plugin the policy gradient theorem and then in (iii) we rearrange to deduce the impact of the perturbation. Choosing $w_{s,a}^\theta$ as:

$$w_{s,a}^\theta = c(s) + A^{\pi^\theta}(s,a) + \frac{\delta(s,a)}{\mu^\theta(s,a)} \tag{6.18}$$

minimizes the function. Hence we have:

$$\Big[\big(\mathcal{F}^{\theta}\big)^{\dagger}\nabla_{\theta}J(\theta)\Big](s,a) = c(s) + A^{\pi^{\theta}}(s,a) + \frac{\delta(s,a)}{\mu^{\theta}(s,a)}. \tag{6.19}$$

Now we can follow the normal MWU derivation steps and get to the policy-update:

$$\begin{aligned}
\pi^{(t+1)}(a|s) &= \frac{\exp\big(\theta^{t+1}(s,a)\big)}{\sum_{a'}\exp\theta^{t+1}(s,a')} \\
&= \frac{\exp\left(\theta^{t}(s,a) + \eta\Big[\big(\mathcal{F}^{\theta^t}\big)^{\dagger}\hat{\nabla}_{\theta}J(\theta^t)\Big](s,a)\right)}{\sum_{a'}\exp\left(\theta^{t}(s,a') + \eta\Big[\big(\mathcal{F}^{\theta^t}\big)^{\dagger}\hat{\nabla}_{\theta}J(\theta^t)\Big](s,a)\right)} \\
&= \frac{\exp\left(\theta^{t}(s,a) + \eta\Big[c(s) + A^{\pi^{\theta}}(s,a) + \frac{\delta(s,a)}{\mu^{\theta}(s,a)}\Big]\right)}{\left(\sum_{a'}\exp\theta^{t}(s,a') + \eta\Big[c(s) + A^{\pi^{\theta}}(s,a') + \frac{\delta(s,a')}{\mu^{\theta}(s,a')}\Big]\right)} \\
&\propto \pi^{t}(s,a)\cdot\eta\Big[Q^{\pi^{\theta}}(s,a) - \log\pi^{\theta}(a|s) - V^{\pi^{\theta}}(s) + \frac{\delta(s,a)}{\mu^{\theta}(s,a)}\Big] \\
&\propto \pi^{t}(s,a)\cdot\exp\left(\eta\Big[Q^{\pi^{\theta}}(s,a) - \log\pi^{\theta}(a|s) + \frac{\delta(s,a)}{\mu^{\theta}(s,a)}\Big]\right) \\
&= \Big(\pi^{t}(s,a)\Big)^{1-\eta\beta}\exp\left(\eta\Big[\overbrace{Q^{\pi^{\theta}}(s,a) + \frac{\delta(s,a)}{\mu^{\theta}(s,a)}}^{=\hat{Q}^{\pi^{\theta}}(s,a)}\Big]\right).
\end{aligned}$$

Where $\hat{Q}^{\pi^{\theta}}(s,a)$ is given by:

$$\hat{Q}^{\pi^{\theta}}(s,a) = Q^{\pi^{\theta}}(s,a) + \frac{\delta(s,a)}{\mu^{\theta}(s,a)}.$$

The proof is complete. □

References

Sion, Maurice (Mar. 1958). "On general minimax theorems". en. In: *Pacific Journal of Mathematics* 8.1, pp. 171–176. ISSN: 0030-8730, 0030-8730. DOI: 10.2140/pjm.1958.8.171. URL: http://msp.org/pjm/1958/8-1/p14.xhtml (visited on 03/20/2023).

Searle, Shayle R. (1982). *Matrix Algebra Useful for Statistics*. John Wiley and Sons.

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

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience. ISBN: 0471241954.

Agarwal, Alekh et al. (2020). "On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift". In: *Proceedings of Machine Learning Research*.

Aryan, Mokhtari, Ozdaglar Asuman E., and Pattathil Sarath (2020). "Convergence Rate of O(1/k) for Optimistic Gradient and Extragradient Methods in Smooth Convex-Concave Saddle Point Problems". In: *SIAM Journal on Optimization*. DOI: 10.1137/19M127375X.

Ding, Dongsheng et al. (2020). "Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 8378–8390. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/5f7695debd8cde8db5abcb9f161b49ea-Paper.pdf.

Mei, Jincheng et al. (2020). "On the Global Convergence Rates of Softmax Policy Gradient Methods". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org.

Cen, Shicong et al. (2021). "Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization". In: *Operations Research*.

Ding, Yuhao, Junzi Zhang, and Javad Lavaei (2021). "Beyond Exact Gradients: Convergence of Stochastic Soft-Max Policy Gradient Methods with Entropy Regularization". In: *ArXiv* abs/2110.10117.

Cai, Yang, Argyris Oikonomou, and Weiqiang Zheng (2022). "Tight Last-Iterate Convergence of the Extragradient and the Optimistic Gradient Descent-Ascent Algorithm for Constrained Monotone Variational Inequalities". In: *Advances in Neural Information Processing Systems*.

Gorbunov, Eduard, Adrien Taylor, and Gauthier Gidel (2022). "Last-Iterate Convergence of Optimistic Gradient Method for Monotone Variational Inequalities". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al.

Zeng, Siliang et al. (2022). "Maximum-Likelihood Inverse Reinforcement Learning with Finite-Time Guarantees". In: *arxiv*.

Schlaginhaufen, Andreas (2023). "Identifiability and generalizability in constrained inverse reinforcement learning". In: *preprint*.