Appendix: Proof and Simulation Details

Anonymous Author(s)
Affiliation
Address
email

This appendix consists of four sections: section 7 summarizes our improvements from D-ICRL [1], section 8 provides some basic notions and notations that will be used in the proof, section 9 presents the proofs of all the lemmas and theorem in the paper, and section 10 gives the simulation details.

7 Improvements from D-ICRL

Our algorithm improves D-ICRL in almost every aspect. In the following context, we summarize our improvements in four categories: assumption, algorithm, theoretical guarantee, and empirical performance.

Assumption. Our method has weaker assumptions than D-ICRL does: (i) we relax the linear reward assumption in D-ICRL; (ii) we do not require the learners to know the budget b while D-ICRL does; (iii) we do not require all-to-all communications among learners while D-ICRL does.

Algorithm. Our algorithm is simpler and more efficient than D-ICRL: (i) our algorithm has a simple single-loop structure where only two gradient descent steps (one for the outer problem and the other for the inner problem) are needed. D-ICRL has a double-loop structure, it needs K gradient descent steps to solve the inner problem. More importantly, we only need a simple gradient descent step to update the outer decision variable while D-ICRL needs multiple steps, including gradient tracking and successive convex approximation, to update the outer decision variables. (ii) As a result, our algorithm is more efficient in terms of computation complexity.

Theoretical guarantee. Our method achieve stronger theoretical guarantees: (i) we provide better rate of the inner problem, i.e., our rate is $O(\frac{1}{N^{1-\eta_2}} + \frac{1}{N})$ (see Subsection 9.6.3) while D-ICRL's is $O(\frac{1}{\sqrt{\log K}})$. (ii) We provide the rate of the outer problem while D-ICRL can only provides asymptotic convergence of the outer problem. (iii) we provide performance guarantee (i.e., constraint violation and cumulative reward difference between the experts and learners) when the reward and cost functions are linear, while D-ICRL does not.

Empirical performance. Our algorithm has better empirical performance. In both experiments, we extend D-ICRL to an online centralized version, called DLM. Experimental results show that our algorithm can reach the same performance with D-ICRL but is more than six times faster at each iteration and more than five times faster to reach 90% success rate.

8 Notions and notations

Define that $\mu^\pi(s,a) \triangleq \phi(s,a) + \gamma \int_{s'\in\mathcal{S}} P(s'|s,a)\mu^\pi(s')ds'$, $\mu^\pi(s) \triangleq \int_{a\in\mathcal{A}} \pi(a|s)\mu^\pi(s,a)da$, $J^\pi_{r_\theta}(s,a) \triangleq r_\theta(s,a) + \gamma \int_{s'\in\mathcal{S}} P(s'|s,a)J^\pi_{r_\theta}(s')ds'$, and $J^\pi_{r_\theta}(s) \triangleq \int_{a\in\mathcal{A}} \pi(a|s)J^\pi_{r_\theta}(s,a)da$. We define the state-action visitation frequency as $\psi^\pi(s,a) \triangleq E^\pi[\sum_{t=0}^T \gamma^t \mathbb{1}\{S_t = s\}\mathbb{1}\{A_t = a\}]$ and state visitation frequency as $\psi^\pi(s) \triangleq E^\pi[\sum_{t=0}^T \gamma^t \mathbb{1}\{S_t = s\}]$, where $\mathbb{1}\{\cdot\}$ is the indicator function. For a given vector $\bar{\omega}$, we define the cost Q-function as $Q^\pi_{\bar{\omega},\theta}(s,a) = \bar{\omega}^\top \mu^\pi(s,a)$ and the cost value-function as $V^\pi_{\bar{\omega},\theta}(s) = \bar{\omega}^\top \mu^\pi(s)$.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Lemma 5. *For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, any ω, any trajectory ζ, and any π, $||\mu^{\pi}(s)||$, $||\mu^{\pi}(s,a)||$, $||\hat{\mu}(\zeta)||$ are bounded by $\frac{1-\gamma^T}{1-\gamma}\sqrt{\sum_{i=1}^{N_E} l^{(i)}}d_1$. For any π and ζ, $||\nabla_\theta J_{r_\theta}(\pi)||$ and $||\nabla_\theta \hat{J}_{r_\theta}(\zeta)||$ are bounded by $\frac{\bar{C}(1-\gamma^T)}{1-\gamma}$.*

Proof. We know that $\mu^{\pi}(s,a) = \phi(s,a) + E^{\pi}_{S,A}[\sum_{t=1}^{T}\gamma^t\phi(S_t,A_t)|S_0 = s, A_0 = a]$. Since $||\phi(s,a)|| \leq \sqrt{\sum_{i=1}^{N_E} l^{(i)}}d_1$, then $||\mu^{\pi}(s,a)|| \leq \frac{1-\gamma^T}{1-\gamma}\sqrt{\sum_{i=1}^{N_E} l^{(i)}}d_1$. As $\mu^{\pi}(s) = \int_{a\in\mathcal{A}}\pi(a|s)\mu^{\pi}(s,a)da$, $||\mu^{\pi}(s)||$ and $||\hat{\mu}(\zeta)||$ are also bounded by $\frac{1-\gamma^T}{1-\gamma}\sqrt{\sum_{i=1}^{N_E} l^{(i)}}d_1$. Analogously, $||E^{\pi}_{S,A}[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r_\theta(S_t,A_t)|S_0 = s_0]||$ and $||\nabla_\theta \hat{J}_{r_\theta}(\zeta)||$ are bounded by $\frac{\bar{C}(1-\gamma^T)}{1-\gamma}$. □

8.1 Constrained soft Bellman policy

We provide the formula of the constrained soft Bellman policy which can be approximated through soft Q learning [2] and soft actor-critic [3]. The following formula is for discrete state-action space and the one for continuous state-action space can be found in Appendix of [1].

$$\pi_{\omega;\theta}(a|s) = \frac{\exp(Q^{\text{soft}}_{\omega;\theta}(s,a))}{\exp(V^{\text{soft}}_{\omega;\theta}(s))},$$
$$V^{\text{soft}}_{\omega;\theta}(s) = \ln(\sum_{a\in\mathcal{A}}\exp(Q^{\text{soft}}_{\omega;\theta}(s,a))),$$
$$Q^{\text{soft}}_{\omega;\theta}(s,a) = r_\theta(s,a) + \omega^\top\phi(s,a) + \gamma\sum_{s'\in\mathcal{S}}P(s'|s,a)V^{\text{soft}}_{\omega;\theta}(s').$$

It is obvious that the constrained soft Bellman policy is continuous in (θ,ω) as it is a composition of continuous functions of (θ,ω). We can regard $r_\theta + \omega^\top\phi$ as a new reward function and use soft Q-learning or soft actor-critic to approximate the constrained soft Bellman policy with this new reward function as input.

9 Proof

This section provides the proof of all the lemmas and theorem in the paper. Subsection 9.1 provides the proof of Lemmas 1 and 3, subsection 9.2 provides the proof of Lemma 2, subsection 9.3 explains why non-linear cost functions will make the problem ill-defined, subsection 9.4 provides the derivation of the gradient approximation $\bar{L}(\theta,\omega)$, subsection 9.5 provides the proof of Lemma 4, subsection 9.6 provides the proof of Theorem 1, and subsection 9.7 provides the proof of Corollary 1. All the proof is for continuous environments except the proof for Lemmas 1 and 2. The reason is that [1] has a similar proof for Lemmas 1 and 2 that proves for continuous environments and linear reward functions, for distinction, here we prove for discrete environments and non-linear reward functions.

Lemma 6. *The gradients $\nabla_\omega \ln\pi_{\omega;\theta}(a|s) = \mu^{\pi_{\omega;\theta}}(s,a) - \mu^{\pi_{\omega;\theta}}(s)$ and $\nabla_\theta \ln\pi_{\omega;\theta}(a|s) = E^{\pi_{\omega;\theta}}_{S,A}[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r_\theta(S_t,A_t)|S_0 = s, A_0 = a] - E^{\pi_{\omega;\theta}}_{S,A}[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r_\theta(S_t,A_t)|S_0 = s]$.*

Lemma 6 have been proved in [1] and thus we omit the proof.

9.1 Proof of Lemmas 1 and 3

Paper [1] has a similar Lemma where they prove for the case of linear reward functions and continuous state-action space. Here, we prove for the case of non-linear reward functions and discrete state-action space.

The Lagrangian of problem (2) is $F(\pi,\omega;\theta) \triangleq H(\pi) + J_{r_\theta}(\pi) + \omega^\top(\mu(\pi) - \frac{1}{N_L}\sum_{v=1}^{N_L}\hat{\mu}(\zeta^{[v]}))$. To find the optimal solution of

$$\max_\pi F(\pi,\omega;\theta) \quad \text{s.t.} \sum_{a\in\mathcal{A}}\pi(a|s) = 1 \quad \forall s\in\mathcal{S}, \quad \pi(a|s)\geq 0 \quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}$$

. We introduce the following auxiliary problem:

$$\max_{\pi,\lambda} \bar{F}(\pi^t,\lambda,\omega;\theta) \quad \text{s.t. } \pi^t(a|s) \geq 0 \quad \forall (s,a)\in\mathcal{S}\times\mathcal{A}, \quad t\geq 0, \tag{6}$$

where $\bar{F}(\pi^t,\lambda,\omega;\theta) = F(\pi,\omega;\theta) + \sum_{s\in\mathcal{S},t\geq 0}\lambda_{s,t}(\sum_{a\in\mathcal{A}}\pi^t(a|s)-1)$. Here, the policy π^t depends on t and we force π^t to be stationary. To solve the auxiliary problem (6), we take the partial derivatives of $\bar{F}$ with respect to π and λ to 0:

$$\begin{aligned}
&\frac{\partial \bar{F}(\pi^t,\lambda,\omega;\theta)}{\partial \pi^t(a|s)} = -\gamma^t P(S_t=s)(\ln\pi^t(a|s) + 1 + r_\theta(s,a) + \omega^\top\phi(s,a) + P(S_t=s)\cdot \\
&E^{\pi}_{S,A}[\sum_{\tau=t+1}^{\infty}\gamma^\tau(-\ln\pi^\tau(A_\tau|S_\tau) + r_\theta(S_\tau,A_\tau) + \omega^\top\phi(S_\tau,A_\tau))|S_t=s,A_t=a] + \lambda_{s,t} = 0, \\
&\frac{\partial \bar{F}(\pi^t,\lambda,\omega;\theta)}{\partial \lambda_{s,t}} = \sum_{a\in\mathcal{A}}\pi^t(a|s) - 1 = 0.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\pi^t_{\omega;\theta}(a|s) = \exp(\frac{\lambda_{s,t,\omega;\theta}}{\gamma^t P(S_t=s)} - 1)\exp\Bigg\{ r_\theta(s,a) + \omega^\top\phi(s,a) \\
&+ E^{\pi_{\omega;\theta}}_{S,A}[\sum_{\tau=t+1}^{\infty}\gamma^{\tau-t}(-\ln\pi^\tau_{\omega;\theta}(A_\tau|S_\tau) + r_\theta(S_\tau,A_\tau) + \omega^\top\phi(S_\tau,A_\tau))|S_t=s,A_t=a]\Bigg\} \geq 0, \\
&\sum_{a\in\mathcal{A}}\pi^t_{\omega;\theta}(a|s) = 1,
\end{aligned}$$

where $\pi^t_{\omega;\theta}$ and $\lambda_{s,t,\omega;\theta}$ are optimal solutions of (6).Denote

$$\begin{aligned}
Q^{\text{soft}}_{\omega;\theta}(s,a) &= r_\theta(s,a) + \omega^\top\phi(s,a) \\
&\quad + \gamma E^{\pi_{\omega;\theta}}_{S,A}[\sum_{\tau=0}^{\infty}\gamma^\tau(-\ln\pi^\tau_{\omega;\theta}(A_\tau|S_\tau) + r_\theta(S_\tau,A_\tau) + \omega^\top\phi(S_\tau,A_\tau))|S_t=s,A_t=a], \\
V^{\text{soft}}_{\omega;\theta}(s) &= \ln(\frac{1}{\exp(\frac{\lambda_{s,t,\omega;\theta}}{\gamma^t P(S_t=s)} - 1)}),
\end{aligned}$$

we can verify that

$$\begin{aligned}
&1 = \sum_{a\in\mathcal{A}}\pi^t_{\omega;\theta}(a|s) = \sum_{a\in\mathcal{A}}\frac{\exp(Q^{\text{soft}}_{\omega;\theta}(s,a))}{\exp(V^{\text{soft}}_{\omega;\theta}(s))} \Rightarrow V^{\text{soft}}_{\omega;\theta}(s) = \ln(\sum_{a\in\mathcal{A}}\exp(Q^{\text{soft}}_{\omega;\theta}(s,a))), \\
&Q^{\text{soft}}_{\omega;\theta}(s,a) = r_\theta(s,a) + \omega^\top\phi(s,a) + \gamma\sum_{s'\in\mathcal{S}}P(s'|s,a)\sum_{a'\in\mathcal{A}}\pi_{\omega;\theta}(a'|s')\Bigg[-\ln\pi_{\omega;\theta}(a'|s') \\
&+ r_\theta(s',a') + \omega^\top\phi(s',a') \\
&+ E^{\pi_{\omega;\theta}}_{S,A}[\sum_{\tau=t+1}^{\infty}\gamma^{\tau-t}(-\ln\pi^\tau_{\omega;\theta}(A_\tau|S_\tau) + r_\theta(S_\tau,A_\tau) + \omega^\top\phi(S_\tau,A_\tau))|S_{t+1}=s',A_{t+1}=a']\Bigg], \\
&= r_\theta(s,a) + \omega^\top\phi(s,a) + \gamma\sum_{s'\in\mathcal{S}}P(s'|s,a)\sum_{a'\in\mathcal{A}}\pi_{\omega;\theta}(a'|s')\Big[-\ln\pi_{\omega;\theta}(a'|s') + Q^{\text{soft}}_{\omega;\theta}(s',a')\Big], \\
&= r_\theta(s,a) + \omega^\top\phi(s,a) + \gamma\sum_{s'\in\mathcal{S}}P(s'|s,a)V^{\text{soft}}_{\omega;\theta}(s').
\end{aligned}$$

Therefore, the constrained soft Bellman policy is the optimal policy of the auxiliary problem (6) and thus is the optimal policy of $\max_{\pi\in\Pi}F(\pi,\omega;\theta)$ given that $\sum_{a\in\mathcal{A}}\pi_{\omega;\theta}(a|s)=1$.

Therefore, $G(\omega;\theta) = F(\pi_{\omega;\theta},\omega;\theta)$. Because the feasible set Π is compact, according to Property 4.2.3 in [4], $G(\omega;\theta)$ is differentiable in ω and $\nabla_\omega G(\omega;\theta) = \mu(\pi_{\omega;\theta}) - \frac{1}{N_L}\sum_{v=1}^{N_L}\hat{\mu}(\zeta^{[v]})$. Similarly, we can get $\nabla_\omega G^{[v]}(\omega;\theta) = \mu(\pi_{\omega;\theta}) - \hat{\mu}(\zeta^{[v]})$.

When $G(\omega;\theta)$ reaches its optimal point $\omega^*(\theta)$, we know that $\nabla_\omega G(\omega^*(\theta);\theta) = \mu(\pi_{\omega^*(\theta);\theta}) - \frac{1}{N_L}\sum_{v=1}^{N_L}\hat{\mu}(\zeta^{[v]}) = 0$. We use p^* to denote the maximum value of problem (2) and d^* to denote the minimum value of $\min_\omega G(\omega;\theta)$. Therefore, we have that

$$H(\pi_{\omega;\theta}) + J_{r_\theta}(\pi_{\omega;\theta}) \le p^* \le d^* = G(\omega^*(\theta);\theta) = H(\pi_{\omega;\theta}) + J_{r_\theta}(\pi_{\omega;\theta}).$$

Therefore, p^* is obtained at $\pi_{\omega;\theta}$ and thus $\pi_{\omega;\theta}$ is the optimal solution of problem (2).

9.2 Proof of Lemma 2

This suffices to show that $G(\omega;\theta)$ is strictly convex in ω.

In this proof, we use the continuous version of the constrained soft Bellman policy [1] and the proof also holds for discrete version of the constrained soft Bellman policy.

We show that the Hessian of $G(\omega;\theta)$ is positive definite. From Lemma 1, we know that $\nabla_\omega G(\omega;\theta) = \mu(\pi_{\omega;\theta})$. Therefore, we have that:

$$\begin{aligned}
&\nabla^2_{\omega\omega} G(\omega;\theta) = \nabla_\omega \mu(\pi_{\omega;\theta}),\\
&= \nabla_\omega \int_{s_0\in\mathcal{S}} P_0(s_0)\mu^{\pi_{\omega;\theta}}(s_0)ds_0,\\
&= \int_{s_0\in\mathcal{S}} P_0(s_0)\nabla_\omega\mu^{\pi_{\omega;\theta}}(s_0)ds_0,\\
&= \int_{s_0\in\mathcal{S}} P_0(s_0)\nabla_\omega \int_{a_0\in\mathcal{A}} \pi_{\omega;\theta}(a_0|s_0)\mu^{\pi_{\omega;\theta}}(s_0,a_0)da_0ds_0,\\
&= \int_{s_0\in\mathcal{S}} P_0(s_0) \int_{a_0\in\mathcal{A}} \Big[\nabla_\omega\pi_{\omega;\theta}(a_0|s_0)\cdot\mu^{\pi_{\omega;\theta}}(s_0,a_0) + \pi_{\omega;\theta}(a_0|s_0)\cdot\nabla_\omega\mu^{\pi_{\omega;\theta}}(s_0,a_0)\Big]da_0ds_0,\\
&= \int_{s_0\in\mathcal{S}} P_0(s_0) \int_{a_0\in\mathcal{A}} \nabla_\omega\pi_{\omega;\theta}(a_0|s_0)\cdot\mu^{\pi_{\omega;\theta}}(s_0,a_0)da_0ds_0\\
&+ \int_{s_0\in\mathcal{S}} P_0(s_0) \int_{a_0\in\mathcal{A}} \pi_{\omega;\theta}(a_0|s_0)\cdot\nabla_\omega\mu^{\pi_{\omega;\theta}}(s_0,a_0)da_0ds_0,\\
&= \int_{s_0\in\mathcal{S}} P_0(s_0) \int_{a_0\in\mathcal{A}} \nabla_\omega\pi_{\omega;\theta}(a_0|s_0)\cdot\mu^{\pi_{\omega;\theta}}(s_0,a_0)da_0ds_0\\
&+ \int_{s_0\in\mathcal{S}} P_0(s_0) \int_{a_0\in\mathcal{A}} \pi_{\omega;\theta}(a_0|s_0)\int_{s_1\in\mathcal{S}} P(s_1|s_0,a_0)\nabla_\omega\mu^{\pi_{\omega;\theta}}(s_1)ds_1da_0ds_0.
\end{aligned}$$

Keep the expansion, we know that

$$\begin{aligned}
\nabla^2_{\omega\omega} G(\omega;\theta) &= \int_{s\in\mathcal{S}} \psi^{\pi_{\omega;\theta}}(s)\int_{a\in\mathcal{A}} \nabla_\omega\pi_{\omega;\theta}(a|s)\mu^{\pi_{\omega;\theta}}(s,a)dads,\\
&= \int_{s\in\mathcal{S}} \psi^{\pi_{\omega;\theta}}(s)\int_{a\in\mathcal{A}} \pi_{\omega;\theta}(a|s)\nabla_\omega\ln\pi_{\omega;\theta}(a|s)\mu^{\pi_{\omega;\theta}}(s,a)dads,\\
&= \int_{s\in\mathcal{S}} \psi^{\pi_{\omega;\theta}}(s)\int_{a\in\mathcal{A}} \pi_{\omega;\theta}(a|s)\big[\mu^{\pi_{\omega;\theta}}(s,a) - \mu^{\pi_{\omega;\theta}}(s)\big]\mu^{\pi_{\omega;\theta}}(s,a)dads,
\end{aligned}$$

where the last inequality follows Lemma 6.

To show that $\nabla^2_{\omega\omega}G(\omega;\theta)$ is positive definite, for any nonzero vector $\bar{\omega}$, we have:

$$\begin{aligned}
&\bar{\omega}^\top\nabla^2_{\omega\omega}G(\omega;\theta)\bar{\omega},\\
&= \int_{s\in\mathcal{S}} \psi^{\pi_{\omega;\theta}}(s)\int_{a\in\mathcal{A}} \pi_{\theta,\omega}(a|s)\Big[\omega^\top\mu^{\pi_{\omega;\theta}}(s,a) - \omega^\top\mu^{\pi_{\omega;\theta}}(s)\Big](\mu^{\pi_{\omega;\theta}}(s,a))\top\bar{\omega}dads,\\
&= \int_{s\in\mathcal{S}} \psi^{\pi_{\omega;\theta}}(s)\int_{a\in\mathcal{A}} \pi_{\omega;\theta}(a|s)\Big[Q^{\pi_{\omega;\theta}}_{\bar{\omega}}(s,a) - V^{\pi_{\omega;\theta}}_{\bar{\omega}}(s)\Big](Q^{\pi_{\omega;\theta}}_{\bar{\omega}}(s,a))dads,\\
&= \int_{s\in\mathcal{S}} \psi^{\pi_{\omega;\theta}}(s)\mathrm{Var}(Q^{\pi_{\omega;\theta}}_{\bar{\omega}}(s,\cdot))ds,
\end{aligned}$$

where $\text{Var}(Q_{\bar{\omega}}^{\pi_{\omega;\theta}}(s,\cdot))$ is the variance of the cost Q-function $Q_{\bar{\omega}}^{\pi_{\omega;\theta}}$ at state s.

Since $\pi_{\omega;\theta}(a|s)$ has non-zero probability to choose any action a at state s, we know that $\text{Var}(Q_{\bar{\omega}}^{\pi_{\omega;\theta}}(s,\cdot)) > 0$. Therefore, $\nabla^2_{\omega\omega}G(\omega;\theta)$ is positive definite and $G(\omega;\theta)$ is strictly convex.

9.3 Ill-defined problem when the cost function is non-linear

From 9.2, we can see that $G(\omega;\theta)$ is strictly convex in ω and $\arg\min_\omega G(\omega;\theta)$ has a unique optimal solution if the cost function is linear. If the cost function is non-linear, $G(\omega;\theta)$ is not guaranteed to be strictly convex in ω and thus there may be multiple $\omega^*(\theta)$ given a θ. Therefore, the outer problem is ill-defined since $L(\theta,\omega^*(\theta))$ may have multiple different values given a certain θ.

Moreover, this is also the reason that we learn the reward functions in the outer level instead of learning them in the inner level as in [1]. Since the reward functions are non-linear, learning them in the inner level can make the problem ill-defined.

Lemma 7. *The two gradients of the global loss function* $\nabla_\omega L(\theta,\omega) = N_L\mu(\pi) - \sum_{v=1}^{N_L}\hat{\mu}(\zeta^{[v]})$ *and* $\nabla_\theta L(\theta,\omega) = N_L E_{S,A}^{\pi_{\omega;\theta}}[\sum_{t=0}^{T}\gamma^t\nabla_\theta r_\theta(S_t,A_t)] - \sum_{v=1}^{N_L}\nabla_\theta \hat{J}_{r_\theta}(\zeta^{[v]})$.

Proof. Paper [1] provides a similar proof for linear reward functions. Here, we prove the proof for non-linear reward function. The global loss function is:

$$L(\theta,\omega) = -\sum_{t=0}^{T}\gamma^t\int_{s\in\mathcal{S}}\int_{a\in\mathcal{A}} N_L P_{\mathcal{D}}(S_t=s,A_t=a)\ln\pi_{\omega;\theta}(a|s)dads,$$

where $P_{\mathcal{D}}(S_t=s,A_t=a)$ is the empirical probability of (s,a) occurring at time t in a trajectory in the demonstrations $\mathcal{D}$ presented by the experts at each online iteration:

$$P_{\mathcal{D}}(S_t=s,A_t=a) \triangleq \frac{1}{N_L}\sum_{j=1}^{N_L}(\mathbb{1}\{s_t^j=s\}\mathbb{1}\{a_t^j=a\}),$$

We can reformulate $\pi_{\omega;\theta}$ as follows:

$$\begin{aligned}
\pi_{\omega;\theta}(a|s) &= \frac{Z_{\omega;\theta}(s,a)}{Z_{\omega;\theta}(s)},\\
\ln Z_{\omega;\theta}(s,a) &= r_\theta(s,a) + \omega^\top\phi(s,a) + \gamma\int_{s'\in\mathcal{S}}P(s'|s,a)\ln Z_{\omega;\theta}(s')ds',\\
\ln Z_{\omega;\theta}(s) &= \ln\int_{a\in\mathcal{A}}Z_{\omega;\theta}(s,a)da.
\end{aligned}$$

Thus,

$$\begin{aligned}
\nabla_\omega L(\theta,\omega) &= -\sum_{t=0}^{T}\gamma^t\int_{s\in\mathcal{S}}\int_{a\in\mathcal{A}}N_L P_{\mathcal{D}}(S_t=s,A_t=a)\nabla_\omega(\ln Z_{\omega;\theta}(s,a)-\ln Z_{\omega;\theta}(s))dads,\\
&= \sum_{t=0}^{T}\gamma^t\int_{s\in\mathcal{S}}\int_{a\in\mathcal{A}}N_L P_{\mathcal{D}}(S_t=s,A_t=a)\Big\{\phi(s,a)\\
&+ E_{S,A}^{\pi_{\omega;\theta}}[\sum_{\tau=t+1}^{T}\gamma^{\tau-t}\phi(S_\tau,A_\tau)|S_t=s,A_t=a] - E_{S,A}^{\pi_{\omega;\theta}}[\sum_{\tau=t}^{T}\gamma^{\tau-t}\phi(S_\tau,A_\tau)|S_t=s]\Big\}dads,
\end{aligned}$$

where the last inequality follows from Lemma 6. Here,

$$\begin{aligned}
&\gamma\int_{s'\in\mathcal{S}}\int_{a'\in\mathcal{A}}P_{\mathcal{D}}(S_{t+1}=s',A_{t+1}=a')E_{S,A}^{\pi_{\omega;\theta}}\left[\sum_{\tau=t+1}^{T}\gamma^{\tau-t-1}\phi(S_\tau,A_\tau)|S_{t+1}=s'\right]da'ds',\\
&= \gamma\int_{s'\in\mathcal{S}}P_{\mathcal{D}}(S_{t+1}=s')E_{S,A}^{\pi_{\omega;\theta}}\left[\sum_{\tau=t+1}^{T}\gamma^{\tau-t-1}\phi(S_\tau,A_\tau)|S_{t+1}=s'\right]ds',
\end{aligned}$$

$$
\begin{aligned}
&= \gamma \int_{s\in\mathcal{S}} \int_{a\in\mathcal{A}} P_{\mathcal{D}}(S_t = s, A_t = a) \int_{s'\in\mathcal{S}} P(s'|s,a)\cdot \\
&E^{\pi_{\omega;\theta}}_{S,A}\Bigg[\sum_{\tau=t+1}^{T} \gamma^{\tau-t-1}\phi(S_\tau, A_\tau)|S_{t+1} = s'\Bigg] ds' dads, \\
&= \int_{s\in\mathcal{S}} \int_{a\in\mathcal{A}} P_{\mathcal{D}}(S_t = s, A_t = a) E^{\pi_{\omega;\theta}}_{S,A}\Bigg[\sum_{\tau=t+1}^{T} \gamma^{\tau-t}\phi(S_\tau, A_\tau)|S_t = s, A_t = a\Bigg] dads.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\nabla_\omega L(\theta,\omega) &= -\sum_{t=0}^{T}\gamma^t \int_{s\in\mathcal{S}} \int_{a\in\mathcal{A}} N_L P_{\mathcal{D}}(S_t = s, A_t = a)\phi(s,a) dads \\
&+ \int_{s\in\mathcal{S}} \int_{a\in\mathcal{A}} N_L P_{\mathcal{D}}(S_0 = s, A_0 = a) E^{\pi_{\omega;\theta}}_{S,A}\Bigg[\sum_{t=0}^{T}\gamma^t\phi(S_t, A_t)|S_0 = s\Bigg] dads, \\
&= N_L \mu(\pi_{\omega;\theta}) - \sum_{v=1}^{N_L} \hat{\mu}(\zeta^{[v]}).
\end{aligned}
$$

Similarly, we have $\nabla_\theta L(\theta,\omega) = N_L E^{\pi_{\omega;\theta}}_{S,A}[\sum_{t=0}^{T}\gamma^t \nabla_\theta r_\theta(S_t, A_t)] - \sum_{v=1}^{N_L} \nabla_\theta \hat{J}_{r_\theta}(\zeta^{[v]})$. ☐

Lemma 8. *(i) There is a positive constant $C_{\nabla_\theta L^{[v]}}$ such that for any θ and ω, it holds that $||\nabla_\theta L^{[v]}(\omega;\theta)|| \leq C_{\nabla_\theta L^{[v]}}$ and $||\nabla_\theta L(\omega;\theta)|| \leq C_{\nabla_\theta L} \triangleq \sum_{v=1}^{N_L} C_{\nabla_\theta L^{[v]}}$.*
(ii) There is a positive constant $C_{\nabla_\omega L^{[v]}}$ such that for any θ and ω, it holds that $||\nabla_\omega L^{[v]}(\omega;\theta)|| \leq C_{\nabla_\omega L^{[v]}}$ and $||\nabla_\omega L(\omega;\theta)|| \leq C_{\nabla_\omega L} \triangleq \sum_{v=1}^{N_L} C_{\nabla_\omega L^{[v]}}$.

Proof. Similar to the proof of Lemma 7, we can see that $\nabla_\omega L^{[v]}(\theta,\omega) = \mu(\pi_{\omega;\theta}) - \hat{\mu}(\zeta^{[v]})$ and $\nabla_\theta L^{[v]}(\theta,\omega) = E^{\pi_{\omega;\theta}}_{S,A}[\sum_{t=0}^{T}\gamma^t\nabla_\theta r_\theta(S_t, A_t)] - \nabla_\theta \hat{J}_{r_\theta}(\zeta^{[v]})$. From Lemma 5, we can see that $||\nabla_\omega L^{[v]}(\theta,\omega)|| \leq \frac{2d_1(1-\gamma^T)}{1-\gamma}\sqrt{\sum_{i=1}^{N_E} l^{(i)}} \triangleq C_{\nabla_\omega L^{[v]}}$ and $||\nabla_\theta L^{[v]}(\theta,\omega)|| \leq \frac{2\bar{C}(1-\gamma^T)}{1-\gamma} \triangleq C_{\nabla_\theta L^{[v]}}$. The bounded gradients for the global loss function are obvious due to the fact that $||\nabla_\theta L(\theta,\omega)|| \leq \sum_{v=1}^{N_L} ||\nabla_\theta L^{[v]}(\theta,\omega)||$ and $||\nabla_\omega L(\theta,\omega)|| \leq \sum_{v=1}^{N_L}||\nabla_\omega L^{[v]}(\theta,\omega)||$. ☐

Lemma 9. *The gradient $\nabla_\theta L^{[v]}(\theta,\omega)$ is Lipschitz continuous in (θ,ω) with constant $C_\theta^{[v]}$ and $\nabla_\omega L^{[v]}(\theta,\omega)$ is Lipschitz continuous in (θ,ω) with constant $C_\omega^{[v]}$.*

Proof. From the proof of Lemma 8, we can see that $\nabla_\omega L^{[v]}(\theta,\omega) = \mu(\pi_{\omega;\theta}) - \hat{\mu}(\zeta^{[v]})$. Therefore, to show the Lipschitz continuous, we need to prove that $\nabla\mu(\pi_{\omega;\theta})$ is bounded. We show it by bounding $\nabla_\omega\mu(\pi_{\omega;\theta})$ and $\nabla_\theta\mu(\pi_{\omega;\theta})$. From Subsection 9.2, we know that $\nabla_\omega\mu(\pi_{\omega;\theta}) = \int_{s\in\mathcal{S}}\psi^{\pi_{\omega;\theta}}(s)\int_{a\in\mathcal{A}}\pi_{\omega;\theta}(a|s)\Big[\mu^{\pi_{\omega;\theta}}(s,a) - \mu^{\pi_{\omega;\theta}}(s)\Big](\mu^{\pi_{\omega;\theta}}(s,a))^\top dads$ which is bounded as each term is bounded (Lemma 5).

Similar to the proof in Subsection 9.2, we can see that

$$
\begin{aligned}
\nabla_\theta\mu(\pi_{\omega;\theta}) &= \int_{s\in\mathcal{S}}\psi^{\pi_{\omega;\theta}}(s)\int_{a\in\mathcal{A}}\pi_{\omega;\theta}(a|s)\nabla_\theta \ln\pi_{\omega;\theta}(a|s)(\mu^{\pi_{\omega;\theta}}(s,a))^\top dads, \\
&= \int_{s\in\mathcal{S}}\psi^{\pi_{\omega;\theta}}(s)\int_{a\in\mathcal{A}}\pi_{\omega;\theta}(a|s)\Bigg[E^{\pi_{\omega;\theta}}_{S,A}[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r_\theta(S_t,A_t)|S_0 = s, A_0 = a] \\
&- E^{\pi_{\omega;\theta}}_{S,A}[\sum_{t=0}^{\infty}\gamma^t\nabla_\theta r_\theta(S_t,A_t)|S_0 = s]\Bigg]\cdot(\mu^{\pi_{\omega;\theta}}(s,a))^\top dads,
\end{aligned}
$$

where each term is bounded (Lemma 5). Therefore, there exists such $C_\omega^{[v]}$. Similarly, we can see the existence of $C_\theta^{[v]}$. ☐

9.4 Derivation of the gradient approximation

From Lemma 7, we know $\nabla_\omega L(\theta,\omega) = N_L \nabla_\omega G(\omega;\theta)$. Therefore, $\nabla L(\theta,\omega^*(\theta)) = \nabla_\theta L(\theta,\omega^*(\theta)) - M(\theta,\omega^*(\theta))\nabla_\omega L(\theta,\omega^*(\theta)) = \nabla_\theta L(\theta,\omega^*(\theta))$. As in each iteration, we cannot get $\omega^*(\theta)$ but an approximation ω. Therefore, we propose the approximation gradient $\bar{\nabla} L(\theta,\omega) = \nabla_\theta L(\theta,\omega) = N_L E_{S,A}^{\pi_{\omega;\theta}}[\sum_{t=0}^{T}\gamma^t \nabla_\theta r_\theta(S_t,A_t)] - \sum_{v=1}^{N_L}\nabla_\theta \hat{J}_{r_\theta}(\zeta^{[v]})$ where the last equality follows Lemma 7. From Lemma 9, we can see that $\nabla_\theta L(\theta,\omega)$ is Lipschitz continuous in (θ,ω) with $C_\theta = \sum_{v=1}^{N_L} C_\theta^{[v]}$. Therefore, $||\nabla L(\theta,\omega^*(\theta)) - \bar{\nabla} L(\theta,\omega)|| = ||\nabla_\theta L(\theta,\omega^*(\theta)) - \nabla_\theta L(\theta,\omega)|| \leq C_\theta||\omega^*(\theta)-\omega||$.

9.5 Proof of Lemma 4

To see the existence of C_L and $\bar{C}_L$, it suffices to show that $||\nabla L(\theta,\omega)||$ and $||\nabla^2 L(\theta,\omega)||$ are bounded which can be seen in Lemmas 8 and 9.

9.6 Proof of Theorem 1

This section has three subsections where the first subsection proves the consensus, the second subsection proves the decreasing local regret, and the third subsection proves the sub-linear cumulative constraint violation.

9.6.1 Consensus

From the proof of Lemma 8, we know that $\nabla_\omega L^{[v]}(\theta,\omega) = \nabla G^{[v]}(\omega;\theta)$. For the distributed gradient descent in Algorithm 1, we know that (equation (5) in [5]):

$$\theta^{[v]}(n) = \sum_{v'=1}^{N_L}[\Phi(n-1,1)]_{v'}^{v}\theta^{[v']}(1) - \sum_{s=2}^{n-1}\alpha(s-1)\sum_{v'=1}^{N_L}[\Phi(n-1,s)]_{v'}^{v}\frac{1}{l}\sum_{i=0}^{l-1}\nabla_\theta L^{[v']}($$
$$\theta^{[v']}(s-1),\omega^{[v']}(s-1),s-1-i) - \frac{\alpha(n-1)}{l}\sum_{i=0}^{l-1}\nabla_\theta L^{[v]}(\theta^{[v]}(n-1),\omega^{[v]}(n-1),n-1-i),$$
$$\omega^{[v]}(n) = \sum_{v'=1}^{N_L}[\Phi(n-1,1)]_{v'}^{v}\omega^{[v']}(1) - \sum_{s=2}^{n-1}\beta(s-1)\sum_{v'=1}^{N_L}[\Phi(n-1,s)]_{v'}^{v}\frac{1}{l}\sum_{i=0}^{l-1}\nabla_\omega L^{[v']}($$
$$\theta^{[v']}(s-1),\omega^{[v']}(s-1),s-1-i) - \frac{\beta(n-1)}{l}\sum_{i=0}^{l-1}\nabla_\omega L^{[v]}(\theta^{[v]}(n-1),\omega^{[v]}(n-1),n-1-i),$$

where $\Phi(k,s) \triangleq W(s)W(s+1)\cdots W(k)$ is the state transition matrix and $[\Phi(k,s)]_{v'}^{v}$ is the entry at the v-th row and v'-th column.

We define that

$$\bar{\theta}(n) \triangleq \frac{1}{N_L}\sum_{v'=1}^{N_L}\theta^{[v']}(1) - \sum_{s=2}^{n}\alpha(s-1)\sum_{v'=1}^{N_L}\frac{1}{lN_L}\sum_{i=0}^{l-1}\nabla_\theta L^{[v']}(\theta^{[v']}(s-1),\omega^{[v']}(s-1),s-1-i),$$
$$\bar{\omega}(n) \triangleq \frac{1}{N_L}\sum_{v'=1}^{N_L}\omega^{[v']}(1) - \sum_{s=2}^{n}\beta(s-1)\sum_{v'=1}^{N_L}\frac{1}{lN_L}\sum_{i=0}^{l-1}\nabla_\omega L^{[v']}(\theta^{[v']}(s-1),\omega^{[v']}(s-1),s-1-i),$$

therefore we have:

$$\bar{\theta}(n+1) = \bar{\theta}(n) - \frac{\alpha(n)}{lN_L}\sum_{v'=1}^{N_L}\sum_{i=0}^{l-1}\nabla_\theta L^{[v']}(\theta^{[v']}(n),\omega^{[v']}(n),n-i),$$
$$\bar{\omega}(n+1) = \bar{\omega}(n) - \frac{\beta(n)}{lN_L}\sum_{v'=1}^{N_L}\sum_{i=0}^{l-1}\nabla_\omega L^{[v']}(\theta^{[v']}(n),\omega^{[v']}(n),n-i).$$

Following the proof of proposition 3 in [5], we can get

$$||\bar{\theta}(n) - \theta^{[v]}(n)|| \leq 2\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{n-1}{B_0}} \sum_{v'=1}^{N_L} ||\theta^{[v']}(1)||$$

$$+ \sum_{s=2}^{n-1} 2\alpha(s-1)C_{\nabla_\theta L}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{n-1-s}{B_0}} + \frac{\alpha(n-1)}{N_L}(C_{\nabla_\theta L} + N_L C_{\nabla_\theta L^{[v]}}),$$

$$||\bar{\omega}(n) - \omega^{[v]}(n)|| \leq 2\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{n-1}{B_0}} \sum_{v'=1}^{N_L} ||\omega^{[v']}(1)||$$

$$+ \sum_{s=2}^{n-1} 2\beta(s-1)C_{\nabla_\omega L}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{n-1-s}{B_0}} + \frac{\beta(n-1)}{N_L}(C_{\nabla_\omega L} + N_L C_{\nabla_\omega L^{[v]}}),$$

where $B_0 = (N_L - 1)B$.

Therefore, the first and third terms in $||\bar{\theta}(n) - \theta^{[v]}(n)||$ are respectively $O((1-\epsilon^{B_0})^{n/B_0})$ and $O(1/n^{\eta_1})$. For the second term, we take a look at

$$\sum_{s=2}^{n-1} \alpha(s-1)(1-\epsilon^{B_0})^{\frac{n-1-s}{B_0}} = \sum_{s=2}^{\lfloor (n-1)/2 \rfloor} \alpha(s-1)(1-\epsilon^{B_0})^{\frac{n-1-s}{B_0}}$$

$$+ \sum_{\lceil (n-1)/2 \rceil}^{n-1} \alpha(s-1)(1-\epsilon^{B_0})^{\frac{n-1-s}{B_0}},$$

$$\leq \bar{\alpha} \sum_{s=2}^{\lfloor (n-1)/2 \rfloor} (1-\epsilon^{B_0})^{\frac{n-1-s}{B_0}} + \frac{\bar{\alpha}}{\lceil (n-1)/2 \rceil^{\eta_1}} \sum_{\lceil (n-1)/2 \rceil}^{n-1} (1-\epsilon^{B_0})^{\frac{n-1-s}{B_0}},$$

$$\leq \bar{\alpha}(1-\epsilon^{B_0})^{\frac{n-1-\lfloor (n-1)/2 \rfloor}{B_0}} \sum_{s=0}^{\lfloor (n-1)/2 \rfloor - 2} (1-\epsilon^{B_0})^{s/B_0}$$

$$+ \frac{\bar{\alpha}}{\lceil (n-1)/2 \rceil^{\eta_1}} \sum_{s=0}^{n-1-\lceil (n-1)/2 \rceil} (1-\epsilon^{B_0})^{s/B_0},$$

$$= O((1-\epsilon^{B_0})^{n/B_0} + \frac{1}{n^{\eta_1}}).$$

Therefore, $||\bar{\theta}(n) - \theta^{[v]}(n)|| \leq O(1/n^{\eta_1} + \bar{\epsilon}^n)$ and similarly we can get $||\bar{\omega}(n) - \omega^{[v]}(n)|| \leq O(1/n^{\eta_2} + \bar{\epsilon}^n)$ where $\bar{\epsilon} = (1-\epsilon^{B_0})^{1/B_0}$.

From Lemma 6, we can see that $||\nabla_\omega \ln \pi_{\omega;\theta}(a|s)||$ and $||\nabla_\theta \ln \pi_{\omega;\theta}(a|s)||$ are both bounded for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Therefore, $||\nabla_\omega \pi_{\omega;\theta}(a|s)|| = ||\pi_{\omega;\theta}(a|s)\nabla_\omega \ln \pi_{\omega;\theta}(a|s)|| \leq ||\nabla_\omega \ln \pi_{\omega;\theta}(a|s)||$ is bounded. Similarly, we can see that $||\nabla_\theta \pi_{\omega;\theta}(a|s)||$ is also bounded. Therefore, $\pi_{\omega;\theta}(a|s)$ is Lipschitz continuous in (ω, θ). As ω and θ reach consensus respectively, the policy reaches consensus at the rate of $O(1/n^{\eta_1} + 1/n^{\eta_2} + \bar{\epsilon}^n)$.

9.6.2 Decreasing local regret

As the trajectory demonstrated at iteration n is random, we have $E_\zeta\left[L(\theta,\omega,n)\right] = \bar{L}(\theta,\omega)$ for all n, where $\nabla_\theta \bar{L}(\theta,\omega) = E_{S,A}^{\pi_{\omega;\theta}}[\sum_{t=0}^{T} \gamma^t \nabla_\theta r_\theta(S_t, A_t)] - \nabla_\theta J_{r_\theta}(\pi_E)$ and $\nabla_\omega \bar{L}(\theta,\omega) = \mu(\pi_{\omega;\theta}) - \mu(\pi_E)$. Similarly, we can define $\bar{L}^{[v]}(\theta,\omega)$. Thus, we have:

$$E[\nabla_\theta \bar{L}^{[v]}(\theta,\omega) - \nabla_\theta L^{[v]}(\theta,\omega,n)] = 0,$$

$$E[||\nabla_\theta \bar{L}^{[v]}(\theta,\omega) - \nabla_\theta L^{[v]}(\theta,\omega,n)||^2] \leq (\frac{\bar{C}(1-\gamma^T)}{1-\gamma})^2,$$

$$E[\nabla_\omega \bar{L}^{[v]}(\theta,\omega) - \nabla_\omega L^{[v]}(\theta,\omega,n)] = 0,$$

$$E[||\nabla_\omega \bar{L}^{[v]}(\theta,\omega) - \nabla_\omega L^{[v]}(\theta,\omega,n)||^2] \le (\frac{d_1(1-\gamma^T)}{1-\gamma})^2 \sum_{i=1}^{N_E} l^{(i)}.$$

Define that $\Delta_\theta(n) \triangleq \frac{1}{N_L}\nabla_\theta \bar{L}(\bar{\theta}(n),\bar{\omega}(n)) - \frac{1}{lN_L}\sum_{v'=1}^{N_L}\sum_{i=0}^{l-1}\nabla_\theta L^{[v']}(\theta^{[v']}(n),\omega^{[v']}(n),n-i)$ and $\Delta_\omega(n) \triangleq \frac{1}{N_L}\nabla_\omega L(\bar{\theta}(n),\bar{\omega}(n)) - \frac{1}{N_L}\sum_{v'=1}^{N_L}\frac{1}{l}\sum_{i=0}^{l-1}\nabla_\omega L^{[v']}(\theta^{[v']}(n),\omega^{[v']}(n),n-i)$.

From Subsection 9.4, we know that $\bar{\nabla} L(\theta,\omega,n) = \nabla_\theta L(\theta,\omega,n)$ and $\nabla_\omega L(\theta,\omega,n) = N_L \nabla_\omega G(\omega;\theta,n)$. Then we can reformulate that $\Delta_\theta(n) \triangleq \frac{1}{N_L}\nabla_\theta \bar{L}(\bar{\theta}(n),\bar{\omega}(n)) - \frac{1}{lN_L}\sum_{v'=1}^{N_L}\sum_{i=0}^{l-1}\bar{\nabla} L^{[v']}(\theta^{[v']}(n),\omega^{[v']}(n),n-i)$ and $\Delta_\omega(n) \triangleq \frac{1}{N_L}\nabla_\omega L(\bar{\theta}(n),\bar{\omega}(n)) - \frac{1}{N_L}\sum_{v'=1}^{N_L}\frac{1}{l}\sum_{i=0}^{l-1}\nabla_\omega G^{[v']}(\omega^{[v']}(n);\theta^{[v']}(n),n-i)$. Then, we have:

$$\begin{aligned}
||E[\Delta_\theta(n)]|| &\le \frac{1}{N_L}\sum_{v=1}^{N_L}||\nabla_\theta \bar{L}^{[v]}(\bar{\theta}(n),\bar{\omega}(n)) - \nabla_\theta \bar{L}^{[v]}(\theta^{[v]}(n),\omega^{[v]}(n))||, \\
&\le \frac{1}{N_L}\sum_{v=1}^{N_L} C_\theta^{[v]}\Big[||\bar{\theta}(n)-\theta^{[v]}(n)|| + ||\bar{\omega}(n)-\omega^{[v]}(n)||\Big], \\
||E[\Delta_\omega(n)]|| &\le \frac{1}{N_L}\sum_{v=1}^{N_L} C_\omega^{[v]}\Big[||\bar{\theta}(n)-\theta^{[v]}(n)|| + ||\bar{\omega}(n)-\omega^{[v]}(n)||\Big], \\
E[||\Delta_\theta(n)||^2] &\le \frac{2}{N_L}\sum_{v=1}^{N_L} E\Big[||\nabla_\theta \bar{L}^{[v]}(\bar{\theta}(n),\bar{\omega}(n)) - \nabla_\theta \bar{L}^{[v]}(\theta^{[v]}(n),\omega^{[v]}(n))||^2 \\
&+ ||\nabla_\theta \bar{L}^{[v]}(\theta^{[v]}(n),\omega^{[v]}(n)) - \frac{1}{l}\sum_{i=0}^{l-1}\nabla_\theta L^{[v]}(\theta^{[v]}(n),\omega^{[v]}(n),n-i)||^2\Big], \\
&\le \frac{4}{N_L}\sum_{v=1}^{N_L}(C_\theta^{[v]})^2\Big[||\bar{\theta}(n)-\theta^{[v]}(n)||^2 + ||\bar{\omega}(n)-\omega^{[v]}(n)||^2\Big] + \frac{2}{l}(\frac{\bar{C}(1-\gamma^T)}{1-\gamma})^2, \\
E[||\Delta_\omega(n)||^2] &\le \frac{4}{N_L}\sum_{v=1}^{N_L}(C_\omega^{[v]})^2\Big[||\bar{\theta}(n)-\theta^{[v]}(n)||^2 + ||\bar{\omega}(n)-\omega^{[v]}(n)||^2\Big] \\
&+ \frac{2}{l}(\frac{d_1(1-\gamma^T)}{1-\gamma})^2 \sum_{i=1}^{N_E} l^{(i)}.
\end{aligned}$$

Lemma 10. *The two summations $\sum_{n=1}^{T}\alpha(n)||E[\Delta_\theta(n)]||$ and $\sum_{n=1}^{T}\beta(n)||E[\Delta_\omega(n)]||$ are bounded by C_{max}.*

Proof. We first take a look at $\sum_{n=1}^{T}\alpha(n)||\bar{\theta}(n)-\theta^{[v]}(n)||$. It has three terms and we bound each term one by one.

The first term is bounded:

$$2\sum_{n=1}^{T}\alpha(n)\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{n-1}{B_0}}\sum_{v'=1}^{N_L}||\theta^{[v']}(0)|| \le 2\bar{\alpha}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}\sum_{v'=1}^{N_L}||\theta^{[v']}(0)||\sum_{n=0}^{T}(1-\epsilon^{B_0})^{\frac{n-1}{B_0}}.$$

For the second term, let $S_n = \sum_{s=2}^{n-1}\alpha(s-1)(1-\epsilon^{B_0})^{\frac{n-1-s}{B_0}}$, then:

$$\begin{aligned}
\frac{S_{n-1}}{\alpha(n-1)} - \frac{S_n}{\alpha(n)} &= \sum_{s=2}^{n-2}\frac{\alpha(s-1)}{\alpha(n-1)}(1-\epsilon^{B_0})^{\frac{n-s-2}{B_0}} - \sum_{s=2}^{n-1}\frac{\alpha(s-1)}{\alpha(n)}(1-\epsilon^{B_0})^{\frac{n-s-1}{B_0}}, \\
&= \sum_{s=2}^{n-2}\frac{\alpha(s-1)\alpha(n)-\alpha(s)\alpha(n-1)}{\alpha(n-1)\alpha(n)}(1-\epsilon^{B_0})^{\frac{n-s-2}{B_0}} - \frac{\alpha(1)}{\alpha(n)}(1-\epsilon^{B_0})^{\frac{n-3}{B_0}},
\end{aligned}$$

$$= \sum_{s=2}^{n-3} \frac{\alpha(s-1)\alpha(n) - \alpha(s)\alpha(n-1)}{\alpha(n-1)\alpha(n)} (1-\epsilon^{B_0})^{\frac{n-s-2}{B_0}} + \frac{\alpha(n-3)\alpha(n) - \alpha(n-2)\alpha(n-1)}{\alpha(n-1)\alpha(n)}$$
$$- \frac{\alpha(1)}{\alpha(n)}(1-\epsilon^{B_0})^{\frac{n-3}{B_0}}.$$

Because exponential terms decay faster than polynomial terms, there exists $\bar{N}$ such that $\frac{\alpha(n-3)\alpha(n)-\alpha(n-2)\alpha(n-1)}{\alpha(n-1)\alpha(n)} - \frac{\alpha(1)}{\alpha(n)}(1-\epsilon^{B_0})^{\frac{n-3}{B_0}} > 0$ if $n > \bar{N}$. Moreover, $\alpha(s-1)\alpha(n) - \alpha(s)\alpha(n-1) = (sn-s)^{\eta_1} - (sn-n)^{\eta_1} > 0$, then $\frac{S_{n-1}}{\alpha(n-1)} > \frac{S_n}{\alpha(n)}$ if $n > \bar{N}$. Therefore, we can find a positive constant $\bar{M}$ such that $\frac{S_n}{\alpha(n)} < \bar{M}$ if $n > \bar{N}$. Then, $\sum_{n=1}^{T} \alpha(n) S_n \leq \sum_{n=1}^{T} (\alpha(n))^2 \bar{M}$ is bounded.

For the third term, it is easy to see that $\sum_{n=2}^{T} \alpha(n)\alpha(n-1)$ is bounded. Therefore, $\sum_{n=1}^{T} \alpha(n)||\bar{\theta}(n) - \theta^{[v]}(n)||$ is bounded.

With similar derivation, we can see that $\sum_{n=1}^{T} \alpha(n)||\bar{\omega}(n) - \omega^{[v]}(n)||$, $\sum_{n=1}^{T} \beta(n)||\bar{\theta}(n) - \theta^{[v]}(n)||$, and $\sum_{n=1}^{T} \beta(n)||\bar{\omega}(n) - \omega^{[v]}(n)||$ are bounded. Therefore, $C_{\max}$ exists. □

Lemma 11. *The summations* $\sum_{n=1}^{T} (\alpha(n))^2 E[||\Delta_\theta(n)||^2]$ *and* $\sum_{n=1}^{T} (\beta(n))^2 E[||\Delta_\omega(n)||^2]$ *are bounded by* D_{max}.

Proof. It suffices to show that $||\bar{\theta}(n) - \theta^{[v]}(n)||^2$ and $||\bar{\omega}(n) - \omega^{[v]}(n)||^2$ are bounded. First, $||\bar{\theta}(n) - \theta^{[v]}(n)||^2$ is bounded as

$$||\bar{\theta}(n) - \theta^{[v]}(n)||^2 \leq 4(\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}})^2 (1-\epsilon^{B_0})^{\frac{2n-2}{B_0}} (\sum_{v'=1}^{N_L} ||\theta^{[v']}(1)||)^2$$
$$+ (\sum_{s=2}^{n-1} 2\alpha(s-1) C_{\nabla_\theta L} \frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}} (1-\epsilon^{B_0})^{\frac{n-1-s}{B_0}})^2 + \frac{(\alpha(n-1))^2}{N_L^2} (C_{\nabla_\theta L} + N_L C_{\nabla_\theta L^{[v]}})^2,$$

where each of the three terms is bounded. Similarly, $||\bar{\omega}(n) - \omega^{[v]}(n)||^2$ is bounded. Thus $D_{\max}$ exists. □

Therefore, we have:

$$E\Big[\bar{L}(\bar{\theta}(n+1), \bar{\omega}(n+1))\Big] \leq \bar{L}(\bar{\theta}(n), \bar{\omega}(n)) + E\Big[[\nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))]^\top [\bar{\theta}(n+1) - \bar{\theta}(n)]$$
$$+ [\nabla_\omega \bar{L}(\bar{\theta}(n), \bar{\omega}(n))]^\top [\bar{\omega}(n+1) - \bar{\omega}(n)] + \bar{C}_L[||\bar{\theta}(n+1) - \bar{\theta}(n)||^2 + ||\bar{\omega}(n+1) - \bar{\omega}(n)||^2]\Big],$$
$$= \bar{L}(\bar{\theta}(n), \bar{\omega}(n)) - E\Big[\alpha(n)[\nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))]^\top [-\Delta_\theta(n) + \frac{1}{N_L}\nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))]$$
$$- \beta(n)[\nabla_\omega \bar{L}(\bar{\theta}(n), \bar{\omega}(n))]^\top \cdot [-\Delta_\omega(n) + \frac{1}{N_L}\nabla_\omega \bar{L}(\bar{\theta}(n), \bar{\omega}(n))] + \bar{C}_L[(\alpha(n))^2 \cdot$$
$$|| - \Delta_\theta(n) + \frac{1}{N_L}\nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2 + (\beta(n))^2 || - \Delta_\omega(n) + \frac{1}{N_L}\nabla_\omega \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2]\Big],$$
$$\leq \bar{L}(\bar{\theta}(n), \bar{\omega}(n)) - \frac{\alpha(n)}{N_L}||\nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2 + \alpha(n)||E[\Delta_\theta(n)]|| \cdot ||\nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||$$
$$- \frac{\beta(n)}{N_L}||\nabla_\omega \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2 + \beta(n)||E[\Delta_\omega(n)]|| \cdot ||\nabla_\omega \bar{L}(\bar{\theta}(n), \bar{\omega}(n))|| + 2\bar{C}_L \cdot E\Big[(\alpha(n))^2 \cdot$$
$$(||\Delta_\theta(n)||^2 + ||\nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2) + (\beta(n))^2(||\Delta_\omega(n)||^2 + ||\nabla_\omega \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2)\Big],$$
$$\Rightarrow E\Big[\alpha(n)(\frac{1}{N_L} - 2\bar{C}_L\alpha(n))||\nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2$$

$$
\begin{aligned}
&+ \beta(n)(\frac{1}{N_L} - 2\bar{C}_L\beta(n))||\nabla_\omega \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2\bigg] \\
&\leq E[\bar{L}(\bar{\theta}(n), \bar{\omega}(n))] - E[\bar{L}(\bar{\theta}(n+1), \bar{\omega}(n+1))] + \alpha(n)C_{\nabla_\theta L}||E[\Delta_\theta(n)]|| \\
&+ \beta(n)C_{\nabla_\omega L}||E[\Delta_\omega(n)]|| + 2(\alpha(n))^2\bar{C}_L E[||\Delta_\theta(n)||^2] + 2(\beta(n))^2\bar{C}_L E[||\Delta_\omega(n)||^2], \\
&\Rightarrow E\bigg[\sum_{n=1}^{N} \alpha(n)(\frac{1}{N_L} - 2\bar{C}_L\alpha(n))||\nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2 \\
&+ \beta(n)(\frac{1}{N_L} - 2\bar{C}_L\beta(n))||\nabla_\omega \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2\bigg] \\
&\leq E\bigg[L(\bar{\theta}(1), \bar{\omega}(1)) - \bar{L}(\bar{\theta}(n+1), \bar{\omega}(n+1))\bigg] + C_{\max}(C_{\nabla_\theta L} + C_{\nabla_\omega L}) + 4\bar{C}_L D_{\max}, \\
&\Rightarrow \sum_{n=1}^{N} E\bigg[\bar{\alpha}\alpha(n)||\nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2\bigg], \\
&\leq \sum_{n=1}^{N} E\bigg[\alpha(n)(\frac{1}{N_L} - 2\bar{C}_L\alpha(n))||\nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2\bigg], \\
&\Rightarrow \sum_{n=1}^{N} E\bigg[\frac{\bar{\alpha}^2}{N^{\eta_1}}||\nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2\bigg], \\
&\leq E\bigg[L(\bar{\theta}(1), \bar{\omega}(1)) - L^*\bigg] + C_{\max}(C_{\nabla_\theta L} + C_{\nabla_\omega L}) + 4\bar{C}_L D_{\max}.
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N} E\bigg[||\nabla_\omega \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2\bigg], \\
&\leq \frac{1}{\bar{\beta}^2 N^{1-\eta_2}} E\bigg[L(\bar{\theta}(1), \bar{\omega}(1)) - L^* + C_{\max}(C_{\nabla_\theta L} + C_{\nabla_\omega L}) + 4\bar{C}_L D_{\max}\bigg].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N} E\bigg[||\nabla \bar{L}(\theta^{[v]}(n), \omega^{[v]}(n))||^2\bigg], \\
&\leq \frac{2}{N}\sum_{n=1}^{N} E\bigg[||\nabla_\theta \bar{L}(\theta^{[v]}(n), \omega^{[v]}(n))||^2 + ||\nabla_\omega \bar{L}(\theta^{[v]}(n), \omega^{[v]}(n))||^2\bigg], \\
&= \frac{2}{N}\sum_{n=1}^{N} E\bigg[||\nabla_\theta \bar{L}(\theta^{[v]}(n), \omega^{[v]}(n)) - \nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n)) + \nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2 \\
&+ ||\nabla_\omega \bar{L}(\theta^{[v]}(n), \omega^{[v]}(n)) - \nabla_\omega \bar{L}(\bar{\theta}(n), \bar{\omega}(n)) + \nabla_\omega \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2\bigg], \\
&\leq \frac{4}{N}\sum_{n=1}^{N} E\bigg[||\nabla_\theta \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2 + ||\nabla_\omega \bar{L}(\bar{\theta}(n), \bar{\omega}(n))||^2 \\
&+ [\sum_{v'=1}^{N_L}(C_\theta^{[v']} + C_\omega^{[v']})(||\bar{\theta}(n) - \theta^{[v]}(n)|| + ||\bar{\omega}(n) - \omega^{[v]}(n)||)]^2\bigg].
\end{aligned}
$$

We have that

$$
\begin{aligned}
&\sum_{n=1}^{N}||\bar{\theta}(n) - \theta^{[v]}(n)||^2 \leq \sum_{n=1}^{N} 4(\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}})^2(1-\epsilon^{B_0})^{\frac{2n-2}{B_0}}(\sum_{v'=1}^{N_L}||\theta^{[v']}(1)||)^2 \\
&+ (\sum_{s=2}^{n-1} 2\alpha(s-1)C_{\nabla_\theta L}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}} \cdot (1-\epsilon^{B_0})^{\frac{n-1-s}{B_0}})^2 + \frac{(\alpha(n-1))^2}{N_L^2}(C_{\nabla_\theta L} + N_L C_{\nabla_\theta L^{[v]}})^2.
\end{aligned}
$$

It is clear that we can find a positive constant $\bar{C}_{\max}$ such that $\sum_{n=1}^{N} 4(\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}})^2(1-\epsilon^{B_0})^{\frac{2n-2}{B_0}}(\sum_{v'=1}^{N_L}||\theta^{[v']}(1)||)^2 + \frac{(\alpha(n-1))^2}{N_L^2}(C_{\nabla_\theta L} + N_L C_{\nabla_\theta L^{[v]}})^2 \leq \bar{C}_{\max}$. Now, we take a look at $(\sum_{s=2}^{n-1} 2\alpha(s-1)C_{\nabla_\theta L}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{n-1-s}{B_0}})^2$. Let $R_n = \sum_{s=2}^{n-1} 2\alpha(s-1)C_{\nabla_\theta L}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{n-1-s}{B_0}}$. Similar to Lemma 10, we can see that $\frac{R_n}{\alpha(n)} \leq \tilde{M}$ for some positive $\tilde{M}$. Then $\sum_{n=1}^{N} R_n^2 \leq \sum_{n=1}^{N}(\alpha(n))^2\tilde{M}^2$. Therefore, $\sum_{n=1}^{N}||\bar{\theta}(n) - \theta^{[v]}(n)||^2$ is bounded. With similar derivation, we can see that $\sum_{n=1}^{N}||\bar{\omega}(n) - \omega^{[v]}(n)||^2$ is bounded. We use $\tilde{C}_{\max}$ to denote $\sum_{n=1}^{N}||\bar{\theta}(n) - \theta^{[v]}(n)||^2 + ||\bar{\omega}(n) - \omega^{[v]}(n)||^2 \leq \tilde{C}_{\max}$.

Therefore,

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N} E\left[||\nabla\bar{L}(\theta^{[v]}(n),\omega^{[v]}(n))||^2\right] \\
&\leq \frac{4}{N}\sum_{n=1}^{N} E\left[||\nabla_\theta\bar{L}(\bar{\theta}(n),\bar{\omega}(n))||^2 + ||\nabla_\omega\bar{L}(\bar{\theta}(n),\bar{\omega}(n))||^2 + N_L\sum_{v'=1}^{N_L}(C_\theta^{[v']} + C_\omega^{[v']})^2\tilde{C}_{\max}\right] \\
&\leq \frac{C_1}{N^{1-\eta_1}} + \frac{C_2}{N^{1-\eta_2}} + \frac{C_3}{N},
\end{aligned} \tag{7}$$

where $C_1 = \frac{4}{\bar{\alpha}^2}E\left[L(\bar{\theta}(1),\bar{\omega}(1)) - L^* + C_{\max}(C_{\nabla_\theta L} + C_{\nabla_\omega L}) + 4\bar{C}_L D_{\max}\right]$, $C_2 = \frac{4}{\beta^2}E\left[L(\bar{\theta}(1),\bar{\omega}(1)) - L^* + C_{\max}(C_{\nabla_\theta L} + C_{\nabla_\omega L}) + 4\bar{C}_L D_{\max}\right]$, $C_3 = N_L\sum_{v'=1}^{N_L}(C_\theta^{[v']} + C_\omega^{[v']})^2\tilde{C}_{\max}$, and L^* is the optimal value of L.

Therefore,

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N} E\left[||\frac{1}{l}\sum_{i=0}^{l-1}\nabla L(\theta^{[v]}(n),\omega^{[v]}(n),n-i)||^2\right], \\
&\leq \frac{2}{N}\sum_{n=1}^{N} E\left[||\nabla L(\theta^{[v]}(n),\omega^{[v]}(n))||^2 + ||\nabla L(\theta^{[v]}(n),\omega^{[v]}(n))\right. \\
&\left.- \frac{1}{l}\sum_{i=0}^{l-1}\nabla L(\theta^{[v]}(n),\omega^{[v]}(n),n-i)||^2\right], \\
&\leq \frac{2C_1}{N^{1-\eta_1}} + \frac{2C_2}{N^{1-\eta_2}} + \frac{2C_3}{N} + \frac{2(C_L)^2}{l}.
\end{aligned}$$

9.6.3 Sub-linear cumulative constraint violation

From (7), we know that $\sum_{n=1}^{N} E\left[||\nabla_\omega\bar{L}(\theta^{[v]}(n),\omega^{[v]}(n))||^2\right] = O(N^{\eta_2}+1)$. Note that $\nabla_\omega\bar{L} = N_L\nabla_\omega G$ (proved in 9.4), therefore, the rate for the inner problem $\min_\omega G(\omega;\theta)$ is $O(N^{\eta_2-1} + 1/N)$. From Lemma 7, we know that $\nabla_\omega L(\theta,\omega) = N_L\mu(\pi) - \sum_{v=1}^{N_L}\hat{\mu}(\zeta^{[v]})$, thus $\nabla_\omega\bar{L}(\theta,\omega) = N_L(\mu(\pi_{\omega;\theta}) - \mu(\pi_E))$. We know that $J_{c_E}(\pi_E) = 0$ as the experts will not violate the hard constraint. Therefore,

$$\begin{aligned}
&\sum_{n=1}^{N} E\left[J_{c_E}^2(\pi_{\omega^{[v]}(n);\theta^{[v]}(n)})\right] = \sum_{n=1}^{N} E\left[(J_{c_E}(\pi_{\omega^{[v]}(n);\theta^{[v]}(n)}) - J_{c_E}(\pi_E))^2\right], \\
&= \sum_{n=1}^{N} E\left[(\omega_E^\top\mu(\pi_{\omega^{[v]}(n);\theta^{[v]}(n)}) - \omega_E^\top\mu(\pi_E))^2\right] \leq \sum_{n=1}^{N} E\left[||\omega_E||^2||\mu(\pi_{\omega^{[v]}(n);\theta^{[v]}(n)}) - \mu(\pi_E)||^2\right], \\
&= \frac{||\omega_E||^2}{N_L^2}\sum_{n=1}^{N} E\left[||\nabla_\omega\bar{L}(\theta^{[v]}(n),\omega^{[v]}(n))||^2\right] = O(N^{\eta_2}+1).
\end{aligned}$$

9.7 Proof of Corollary 1

When the reward function is a linear combination similar to the cost function, this proof is similar to the proof of sub-linear cumulative constraint violation in Subsection 9.6.3.

10 Simulation details

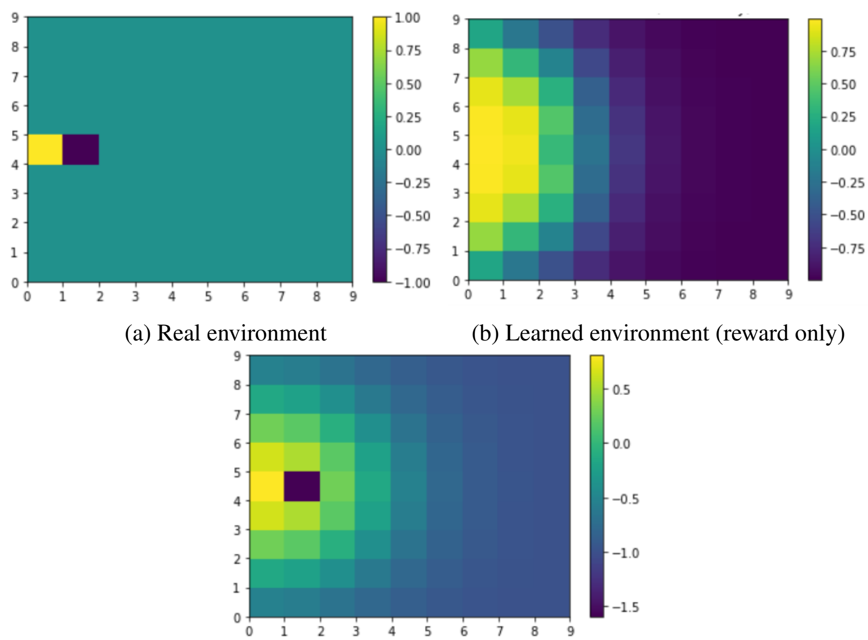
The Python3 code was run on a laptop with one Intel Core i7-9750H 2.60GHz CPU and 16 GB of RAM under Ubuntu 18.04 operating system.

10.1 The benefit of learning both reward and cost functions

While learning a well-structured reward function can prevent some “bad" movements by assigning negative reward to those movements, we provide the benefits of learning both reward and cost functions as follows:

(i) Learning both reward and cost functions can make it clear that how much a state-action pair is rewarded and penalized. For example, consider a state-action pair (s, a) that has ground truth reward $r_E(s, a) = 1$ and ground truth cost $c_E(s, a) = 0.5$. Suppose we only use a single neural network $r_\theta(s, a)$ to learn a well-structured reward function, even if we can have very good performance (say, $r_\theta(s, a) = 0.5$), we do not know whether (s, a) violates the constraints or how much it violates the constraints since the single reward function outputs positive value at (s, a). However, if we learn reward and cost function separately, we can clearly solve this problem. While learning a well-structured reward function can help discourage “bad" movements in some cases (e.g., when each state-action pair is either rewarded or penalized but not both), learning both reward and cost functions can give us more information.

(ii) Even in the cases where each state-action pair is either rewarded or penalized, it is hard for a single reward function to recover constraints that are close to the highly-rewarded areas (e.g., goals). Here, we use a single agent example to illustrate this in detail:



(a) Real environment (b) Learned environment (reward only)

(c) Learned environment (reward & cost)

Figure 1: An example where the goal and obstacle are close

Figure 1 shows a scenario where the goal and obstacle are close to each other where the yellow block is the goal and the dark block is the obstacle. In the real environment, the reward of the goal state is

1, the reward of the obstacle state is -1, and the reward of other states is 0. We can see that if we only use a neural network to learn a reward function, we cannot learn the obstacle near the goal. In contrast, if we learn both reward and cost functions, we can learn the obstacle close to the goal. The reason is that we use neural networks to learn the reward function and neural networks are continuous. Therefore, the states near the goal state will also have relatively high reward even if they may be the obstacle states. The benefit of learning an extra cost function is that now the outcome of reaching an obstacle state is its reward minus cost. Even if the reward neural network still assigns relatively high reward to the obstacle state near the goal, the extra cost function will heavily penalize the obstacle state. Then, visiting the obstacle state has low outcome (i.e., reward minus cost) even if it is close to the goal.

In conclusion, while learning a well-structured reward function may replace learning both reward and cost functions in some cases, it is not general and it does not provide enough information we want, especially for the cases where we are more interested in constraints [6]. Moreover, there are some other works that support this conclusion. For example, [7] points out that it is often the case that the recovered reward function fails to capture the implicit constraints. In [8], the authors augment some constraint signals to the reward neural network but the learned behaviors still have unsatisfying constraint violation performance.

10.2 Evasion from patrolled area

Due to the well-known curse of dimensionality, the reinforcement learning (or dynamic programming) of multiple experts is hard to compute. To alleviate this issue, we model the experts as separate MDPs for most of the time and only model them as a Markov game (MG) when they are close to each other.

At each iteration, the experts demonstrate N_L trajectories and each of the N_L learners observe one of them. We design the cost function such that it is positive at obstacles and zero elsewhere. Following [6, 8], we study hard constraints and choose $b = 0$. The adjacency matrix of the communication network is $\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$ and the reward function of each expert is a neural network with four hidden layers. The activation functions are relu and the number of neurons in each layer is respectively 512, 256, 256, 128.

10.3 Drone motion planning with obstacles

The simulator is built in Gazebo based on a package called hector_quadrotor [9]. The total demonstrations we provide are 80 pairs of trajectories. The neural network structure is same to the one in the last experiment. The state of each drone is its 2-D coordinates and the action of each drone is its moving direction which is characterized by a 2-D vector. For example, the action $[1, 1]^\top$ means that the moving direction is 45 degrees upper right and the action $[-1, 1]^\top$ means that the moving direction is 45 degrees upper left. We restrict the length of each moving step as 0.1. The state space of each drone is the set of all the 2-D coordinates in the room and the action space of each drone is all the directions.

The communication network for the four learners has two stages and the adjacency matrix in stage 1 is $\begin{bmatrix} 0.24 & 0.24 & 0.26 & 0.26 \\ 0.24 & 0.24 & 0.26 & 0.26 \\ 0.26 & 0.26 & 0.24 & 0.24 \\ 0.26 & 0.26 & 0.24 & 0.24 \end{bmatrix}$ and in stage 2 is $\begin{bmatrix} 0.26 & 0.26 & 0.24 & 0.24 \\ 0.26 & 0.26 & 0.24 & 0.24 \\ 0.24 & 0.24 & 0.26 & 0.26 \\ 0.24 & 0.24 & 0.26 & 0.26 \end{bmatrix}$.

References

[1] S. Liu and M. Zhu, “Distributed inverse constrained reinforcement learning for multi-agent systems,” in *Advances in Neural Information Processing Systems*, 2022.

[2] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, “Reinforcement learning with deep energy-based policies,” in *International Conference on Machine Learning*, pp. 1352–1361, 2017.

[3] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *International Conference on Machine Learning*, pp. 1861–1870, 2018.

[4] C. A. Floudas, *Nonlinear and mixed-integer optimization: fundamentals and applications*. Oxford University Press, 1995.

[5] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[6] D. R. Scobee and S. S. Sastry, "Maximum likelihood constraint inference for inverse reinforcement learning," in *International Conference on Learning Representations*, 2019.

[7] D. Park, M. Noseworthy, R. Paul, S. Roy, and N. Roy, "Inferring task goals and constraints using bayesian nonparametric inverse reinforcement learning," in *Conference on robot learning*, pp. 1005–1014, 2020.

[8] S. Malik, U. Anwar, A. Aghasi, and A. Ahmed, "Inverse constrained reinforcement learning," in *International Conference on Machine Learning*, pp. 7390–7399, 2021.

[9] J. Meyer, A. Sendobry, S. Kohlbrecher, U. Klingauf, and O. von Stryk, "Comprehensive simulation of quadrotor uavs using ros and gazebo," in *International Conference on Simulation, Modeling and Programming for Autonomous Robots*, pp. 400–411, 2012.