8. Appendix

This appendix should be read as a continuation of the main paper.

8.1. Gradient of log likelihood

The gradient of (5) is

$$\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \frac{1}{N}\sum_{i=1}^{M}\left[0 + \nabla_\theta \log \zeta_\theta(\tau^{(i)})\right] - \frac{1}{\int \exp(\beta r(\tau))\zeta_\theta(\tau)d\tau}\int \exp(\beta r(\tau))\nabla_\theta \zeta_\theta(\tau)d\tau \\
&= \frac{1}{N}\sum_{i=1}^{N}\nabla_\theta \log \zeta_\theta(\tau^{(i)}) - \int \frac{\exp(\beta r(\tau))\zeta_\theta(\tau)}{\int \exp(\beta r(\tau'))\zeta_\theta(\tau')d\tau'}\nabla_\theta \log \zeta_\theta(\tau)d\tau \\
&= \frac{1}{N}\sum_{i=1}^{N}\nabla_\theta \log \zeta_\theta(\tau^{(i)}) - \int p_{\mathcal{M}^{\bar{\zeta}_\theta}}(\tau)\nabla_\theta \log \zeta_\theta(\tau)d\tau \\
&= \frac{1}{N}\sum_{i=1}^{N}\nabla_\theta \log \zeta_\theta(\tau^{(i)}) - \mathbb{E}_{\tau\sim\pi_{\mathcal{M}^{\zeta_\theta}}}\left[\nabla_\theta \log \zeta_\theta(\tau)\right],
\end{aligned} \tag{13}$$

where the second line follows from the identity $\nabla_\theta \zeta_\theta(\tau) \equiv \zeta_\theta(\tau)\nabla_\theta \log \zeta_\theta(\tau)$ and the fourth line from the MaxEnt assumption.

8.2. Derivation of importance sampling weights

Suppose that at some iteration of our training procedure we are interested in approximating the gradient of the log of the partition function $\nabla_\theta \log Z_\theta$ (where θ are the current parameters of our classifier) using an older policy $\pi_{\zeta_{\bar{\theta}}}$ (where $\bar{\theta}$ were the parameters of the classifier which induced the constraint set that this policy respects). We can do so by noting that

$$\begin{aligned}
Z_\theta &= \int \exp(r(\tau))\zeta_\theta(\tau)d\tau \\
&= \int \pi_{\zeta_{\bar{\theta}}}(\tau)\left[\frac{\exp(r(\tau))\zeta_\theta(\tau)}{\pi_{\zeta_{\bar{\theta}}}(\tau)}\right]d\tau \\
&= \mathbb{E}_{\tau\sim\pi_{\zeta_{\bar{\theta}}}}\left[\frac{\exp(r(\tau))\zeta_\theta(\tau)}{\pi_{\zeta_{\bar{\theta}}}(\tau)}\right] \\
&= Z_{\bar{\theta}} \cdot \mathbb{E}_{\tau\sim\pi_{\zeta_{\bar{\theta}}}}\left[\frac{\zeta_\theta(\tau)}{\zeta_{\bar{\theta}}(\tau)}\right].
\end{aligned} \tag{14}$$

where the fourth lines follows from our MaxEnt assumption, i.e., $\pi_{\zeta_{\bar{\theta}}}(\tau) = \exp(r(\tau))\zeta_{\bar{\theta}}(\tau)/Z_{\bar{\theta}}$.

Therefore

$$\begin{aligned}
\nabla_\theta \log Z_\theta &= \frac{1}{Z_\theta}\nabla_\theta Z_\theta \\
&= \frac{1}{Z_{\bar{\theta}} \cdot \mathbb{E}_{\tau\sim\pi_{\zeta_{\bar{\theta}}}}\left[\frac{\zeta_\theta(\tau)}{\zeta_{\bar{\theta}}(\tau)}\right]}\left[Z_{\bar{\theta}} \cdot \mathbb{E}_{\tau\sim\pi_{\zeta_{\bar{\theta}}}}\left[\frac{\nabla_\theta \zeta_\theta(\tau)}{\zeta_{\bar{\theta}}(\tau)}\right]\right] \\
&= \frac{1}{\mathbb{E}_{\tau\sim\pi_{\zeta_{\bar{\theta}}}}\left[\frac{\zeta_\theta(\tau)}{\zeta_{\bar{\theta}}(\tau)}\right]}\left[\mathbb{E}_{\tau\sim\pi_{\zeta_{\bar{\theta}}}}\left[\frac{\zeta_\theta(\tau)}{\zeta_{\bar{\theta}}(\tau)}\nabla_\theta \log \zeta_\theta(\tau)\right]\right].
\end{aligned} \tag{15}$$

Note that $\mathbb{E}_{\tau\sim\pi_{\zeta_{\bar{\theta}}}}\left[\frac{\zeta_\theta(\tau)}{\zeta_{\bar{\theta}}(\tau)}\right]=\int\pi_{\zeta_\theta}(\tau)d\tau=1$. So

$$\begin{aligned}
\nabla_\theta\log Z_\theta &= \mathbb{E}_{\pi_{\zeta_{\bar{\theta}}}}\left[\frac{\zeta_\theta(\tau)}{\zeta_{\bar{\theta}}(\tau)}\nabla_\theta\log\zeta_\theta(\tau)\right]\\
&= \mathbb{E}_{\pi_{\zeta_{\bar{\theta}}}}\left[\prod_{t=1}^{T}\frac{\zeta_\theta(s_t,a_t)}{\zeta_{\bar{\theta}}(s_t,a_t)}\nabla_\theta\log\prod_{t'=1}^{T}\zeta_\theta(s_{t'},a_{t'})\right]\\
&= \mathbb{E}_{\pi_{\zeta_{\bar{\theta}}}}\left[\prod_{t=1}^{T}\frac{\zeta_\theta(s_t,a_t)}{\zeta_{\bar{\theta}}(s_t,a_t)}\sum_{t'=1}^{T}\nabla_\theta\log\zeta_\theta(s_{t'},a_{t'})\right]\\
&= \sum_{t'=1}^{T}\mathbb{E}_{\pi_{\zeta_{\bar{\theta}}}}\left[\prod_{t=1}^{T}\frac{\zeta_\theta(s_t,a_t)}{\zeta_{\bar{\theta}}(s_t,a_t)}\nabla_\theta\log\zeta_\theta(s_{t'},a_{t'})\right]\\
&= \sum_{t'=1}^{T}\mathbb{E}_{\pi_{\zeta_{\bar{\theta}}}}\left[\left(\prod_{\substack{t=1\\t\neq t'}}^{T}\frac{\zeta_\theta(s_t,a_t)}{\zeta_{\bar{\theta}}(s_t,a_t)}\right)\left(\frac{\zeta_\theta(s_{t'},a_{t'})}{\zeta_{\bar{\theta}}(s_{t'},a_{t'})}\nabla_\theta\log\zeta_\theta(s_{t'},a_{t'})\right)\right]\\
&= \sum_{t'=1}^{T}\mathbb{E}_{\tau/(s_{t'},a_{t'})\sim\pi_{\zeta_{\bar{\theta}}}}\left[\prod_{\substack{t=1\\t\neq t'}}^{T}\frac{\zeta_\theta(s_t,a_t)}{\zeta_{\bar{\theta}}(s_t,a_t)}\right]\mathbb{E}_{s_{t'},a_{t'}\sim\pi_{\zeta_{\bar{\theta}}}}\left[\frac{\zeta_\theta(s_{t'},a_{t'})}{\zeta_{\bar{\theta}}(s_{t'},a_{t'})}\nabla_\theta\log\zeta_\theta(s_{t'},a_{t'})\right]\\
&= \sum_{t'=1}^{T}\frac{Z_\theta}{Z_{\bar{\theta}}}\cdot\mathbb{E}_{\pi_{\zeta_{\bar{\theta}}}}\left[\frac{\zeta_\theta(s_{t'},a_{t'})}{\zeta_{\bar{\theta}}(s_{t'},a_{t'})}\nabla_\theta\log\zeta_\theta(s_{t'},a_{t'})\right].\\
&\approx \sum_{t'=1}^{T}\mathbb{E}_{\pi_{\zeta_{\bar{\theta}}}}\left[\frac{\zeta_\theta(s_{t'},a_{t'})}{\zeta_{\bar{\theta}}(s_{t'},a_{t'})}\nabla_\theta\log\zeta_\theta(s_{t'},a_{t'})\right],
\end{aligned}\tag{16}$$

where the last step assumes that $Z_\theta\approx Z_{\bar{\theta}}$. This is justified since we restrict the extent to which ζ_θ can change via the early stopping technique.

8.3. Forward and reverse KL divergences between two policies

Consider two policies $\pi_{\bar{\theta}}$ and π_θ. Using our MaxEnt assumption, we can write the forward KL divergence as

$$\begin{aligned}
D_{KL}(\pi_{\bar{\theta}}||\pi_\theta) &= \mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}\left[\log\frac{\pi_{\bar{\theta}}(\tau)}{\pi_\theta(\tau)}\right]\\
&= \mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}\left[\log\frac{\zeta_{\bar{\theta}}(\tau)}{\zeta_\theta(\tau)}\right]+\log\frac{Z_\theta}{Z_{\bar{\theta}}}.
\end{aligned}\tag{17}$$

Let $\omega(\tau)$ denote $\zeta_{\bar{\theta}}(\tau)/\zeta_\theta(\tau)$. Plugging in the expression for Z_θ from (14) and using Jensen's inequality gives

$$\begin{aligned}
D_{KL}(\pi_{\bar{\theta}}||\pi_\theta) &= \mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}\left[\log\omega(\tau)\right]+\log\mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}\left[\omega(\tau)\right]\\
&\leq 2\log\mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}\left[\omega(\tau)\right].
\end{aligned}\tag{18}$$

Similarly, the reverse KL divergence is

$$\begin{aligned}
D_{KL}(\pi_\theta||\pi_{\bar{\theta}}) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\log\frac{\pi_\theta(\tau)}{\pi_{\bar{\theta}}(\tau)}\right]\\
&= \mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}\left[\frac{\pi_\theta(\tau)}{\pi_{\bar{\theta}}(\tau)}\log\frac{\pi_\theta(\tau)}{\pi_{\bar{\theta}}(\tau)}\right]\\
&= \mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}\left[\omega(\tau)\frac{Z_{\bar{\theta}}}{Z_\theta}\log\omega(\tau)\frac{Z_{\bar{\theta}}}{Z_\theta}\right]\\
&= \mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}\left[\omega(\tau)\log\omega(\tau)\right]\frac{Z_{\bar{\theta}}}{Z_\theta}+\mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}\left[\omega(\tau)\right]\frac{Z_{\bar{\theta}}}{Z_\theta}\log\frac{Z_{\bar{\theta}}}{Z_\theta}.
\end{aligned}\tag{19}$$

From (14) we know that $Z_{\bar{\theta}}/Z_\theta = 1/\mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}\omega(\tau)$. Using Jensen's inequality we have

$$\begin{aligned} D_{KL}(\pi_\theta||\pi_{\bar{\theta}}) &= \frac{1}{\mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}[\omega(\tau)]}\mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}[\omega(\tau)\log\omega(\tau)] - \log\mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}[\omega(\tau)] \\ &\leq \frac{1}{\mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}[\omega(\tau)]}\mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}[\omega(\tau)\log\omega(\tau)] - \mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}[\log\omega(\tau)]\,. \end{aligned} \tag{20}$$

Letting $\bar{\omega}$ denote $\mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}[\omega(\tau)]$ gives us

$$D_{KL}(\pi_\theta||\pi_{\bar{\theta}}) \leq \frac{\mathbb{E}_{\tau\sim\pi_{\bar{\theta}}}[(\omega(\tau)-\bar{\omega})\log\omega(\tau)]}{\bar{\omega}}. \tag{21}$$

8.4. Rationale for (9)

Consider a constrained MDP $\mathcal{M}^{\mathcal{C}}$ as defined in Section 2.2. We are interested in recovering the following policy

$$\pi_{\mathcal{M}^{\mathcal{C}}}(\tau) = \frac{\exp(\beta r(\tau))}{Z_{\mathcal{M}^{\mathcal{C}}}}\mathbb{1}^{\mathcal{C}}(\tau) \tag{22}$$

where $Z_{\mathcal{M}^{\mathcal{C}}} = \int \exp(\beta r(\tau))\mathbb{1}^{\mathcal{C}}(\tau)d\tau$ is the partition function and $\mathbb{1}^{\mathcal{C}}$ is an indicator function that is 0 if $\tau \in \mathcal{C}$ and 1 otherwise.

Lemma: The Boltzmann policy $\pi^B(\tau) = \exp(\beta r(\tau))/Z$ maximizes $\mathcal{L}(\pi) = \mathbb{E}_{\tau\sim\pi}[r(\tau)] + \frac{1}{\beta}\mathcal{H}(\pi)$, where $H(\pi)$ denotes the entropy of π.

Proof: Note that the KL-divergence, D_{KL}, between a policy π and π^B can be written as

$$\begin{aligned} \mathcal{D}_{KL}(\pi||\pi^B) &= \mathbb{E}_{\tau\sim\pi}[\log\pi(\tau) - \log\pi^B(\tau)] \\ &= \mathbb{E}_{\tau\sim\pi}[\log\pi(\tau) - \beta r(\tau) + \log Z] \\ &= -\mathbb{E}_{\tau\sim\pi}[\beta r(\tau)] - \mathcal{H}(\pi) + \log Z \\ &= -\beta\mathcal{L}(\pi) + \log Z. \end{aligned} \tag{23}$$

Since $\log Z$ is constant, minimizing $D_{KL}(\pi||\pi^B)$ is equivalent to maximizing $\mathcal{L}(\pi)$. Also, we know that $\mathcal{D}_{KL}(\pi||\pi^B)$ is minimized when $\pi = \pi^B$. Therefore, π^B maximizes $\mathcal{L}$.

Proposition: The policy in (22) is a solution of

$$\underset{\lambda\geq 0}{\text{minimize}}\,\max_\pi \mathbb{E}_{\tau\sim\pi}[r(\tau)] + \frac{1}{\beta}\mathcal{H}(\pi^\phi) - \lambda(\mathbb{E}_{\tau\sim\pi^\phi}[\bar{\zeta}_\theta(\tau)] - \alpha). \tag{24}$$

Proof: Let us rewrite the inner optimization problem as

$$\max_\pi\ \mathbb{E}_{\tau\sim\pi}[r(\tau) - \lambda(\bar{\zeta}_\theta(\tau) - \alpha)] + \frac{1}{\beta}\mathcal{H}(\pi). \tag{25}$$

From the Lemma we know that the solution to this is

$$\pi(\tau,\lambda) = \frac{g(\tau,\lambda)}{\int g(\tau',\lambda)d\tau'}, \tag{26}$$

where $g(\tau,\lambda) = \exp(\beta(r(\tau) - \lambda(\bar{\zeta}_\theta(\tau) - \alpha)))$. To find $\pi^*(\tau) = \min_\lambda \pi(\tau,\lambda)$, note that:

1. When $\bar{\zeta}_\theta(\tau) \leq \alpha$, then $\lambda^* = 0$ minimizes π. In this case $g(\tau,\lambda^*) = \exp(\beta r(\tau))$.
2. When $\bar{\zeta}_\theta(\tau) > \alpha$, then $\lambda^* \to \infty$ minimizes π. In this case $g(\tau,\lambda^*) = 0$.

We can combine both of these cases by writing

$$\pi^*(\tau) = \frac{\exp(r(\tau))}{\int \exp(r(\tau'))\mathbb{1}^{\bar{\zeta}_\theta}(\tau')d\tau'}\mathbb{1}^{\bar{\zeta}_\theta}(\tau), \tag{27}$$

where $\mathbb{1}^{\bar{\zeta}_\theta}(\tau)$ is 1 if $\bar{\zeta}_\theta(\tau) \leq \alpha$ and 0 otherwise. (Note that the denominator is greater than 0 as long as we have at least one τ for which $\bar{\zeta}_\theta(\tau) \leq \alpha$, i.e., we have at least one feasible solution.)

8.5. Experimental settings

Our codebase is built on top of the stable-baselines codebase (Hill et al., 2018). We used W&B (Biewald, 2020) to manage our experiments and conduct sweeps on hyperparameters. We used Adam (Kingma & Ba, 2015) to optimize all of our networks. All important hyperparameters are listed in Table 1. For the ablation studies we used the same parameters as listed in the table for HalfCheetah. Details on the environments can be found below.

8.5.1. LAPGRIDWORLD

Here, agents move on a 11×11 grid by taking either clockwise or anti-clockwise actions. The agent is awarded a reward 3 each time it moves onto a bridge with a dollar (see Figure 2). The agent's state is the number of the grid it is on.

8.5.2. HALFCHEETAH, ANT AND ANT-BROKEN

The original reward schemes for HalfCheetah and Ant in OpenAI Gym (Brockman et al., 2016), reward the agents proportional to the distance they cover in the forward direction. We modify this and instead simply reward the agents according to the amount of distance they cover (irrespective of the direction they move in). For Ant-Broken we simply disable two of the legs of Ant by hard-coding a torque of 0 on their motors.

8.5.3. POINT

For the Point agent, the reward function at each timestep is defined as follows

$$r := \frac{ydx - xdx}{\left(1 + |\sqrt{x^2 + y^2} - 10)|\right)} \tag{28}$$

where x, y are the position coordinates of the agent and dx and dy are the distances that the agent has moved in x and y directions respectively in that timestep.

Table 1. List of hyperparameters. For neural network architectures we report the number of hidden units in each layer. All hidden layers use the tanh activation function.

PARAMETER	LAPGRIDWORLD	HALFCHEETAH	ANT	POINT	ANTBROKEN
POLICY, π_ϕ					
ARCHITECTURE					
POLICY NETWORK	64, 64	64, 64	64, 64	64, 64	64, 64
VALUE NETWORK	64, 64	64, 64	64, 64	64, 64	64, 64
COST VALUE NETWORK	64, 64	64, 64	64, 64	64, 64	64, 64
BATCH SIZE	64	64	128	64	128
PPO TARGET KL	0.01	0.01	0.01	0.01	0.01
LEARNING RATE	3×10^{-4}	3×10^{-4}	3×10^{-5}	3×10^{-4}	3×10^{-5}
REWARD-GAE-γ	0.99	0.99	0.99	0.99	0.99
REWARD-GAE-λ	0.95	0.95	0.90	0.95	0.90
COST-GAE-γ	0.99	0.99	0.99	0.99	0.99
COST-GAE-λ	0.95	0.95	0.95	0.95	0.95
ENTROPY BONUS, $1/\beta$	0.0	0.0	0.0	0.0	0.0
LAGRANGIAN, λ					
INITIAL VALUE	1.0	1.0	0.1	1.0	0.1
LEARNING RATE	0.1	0.1	1.0	0.1	1.0
BUDGET	0.0	0.0	0.0	0.0	0.0
CONSTRAINT FUNCTION, ζ_θ					
ARCHITECTURE	20	20	40,40	-	-
LEARNING RATE	0.01	0.01	0.01	-	-
BACKWARD ITERATIONS	10	10	10	-	-
REGULARIZER WEIGHT	0.5	0.5	0.6	-	-
MAX FORWARD KL, ϵ_F	10	10	10	-	-
MAX BACKWARD KL, ϵ_B	2.5	2.5	2.5	-	-
MISCELLANEOUS					
EXPERT ROLLOUTS	1	10	45	-	-
ROLLOUT LENGTH	200	1000	500	150	500