A Task Details

	Pendulum	Hopper	2D Walker	3D Walker	Reacher
Observation space dimension	3	11	17	41	21
Action space dimension	1	3	6	15	5
Number of samples per iteration	4k	16k	16k	16k	40k
Number of iterations	100	200	200	1000	500
Number of TRPO iterations for expert	50	50	100	500	100
Upper limit of number of imitation steps of LOKI	10	20	25	50	25
Truncated horizon of THOR	40	40	250	250	250

The expert value estimator $\hat{V}_{\pi^*}$ needed by SLOLS and THOR were trained on a large set of samples (50 times the number of samples used in each batch in the later policy learning), and the final average TD error are: Pendulum (0.972), Hopper (0.989), 2D Walker (0.975), 3D Walker (0.983), and Reacher (0.973), measured in terms of explained variance, which is defined as 1- (variance of error / variance of prediction).

B Proof of Section 4

B.1 Proof of Proposition 1

To prove Proposition 1, we first prove a useful Lemma 2.

Lemma 2. *Let $\mathcal{K}$ be a convex set. Let $h = \mathbb{E}[g]$. Suppose R is α-strongly convex with respect to norm $\|\cdot\|$.*

$$y = \arg\min_{z\in\mathcal{K}} \langle g, z\rangle + \frac{1}{\eta} D_R(z||x) =: P_{g,\eta}(x)$$

where η satisfies that $-\alpha\eta + \frac{\beta\eta^2}{2} \le 0$. Then it holds

$$\mathbb{E}[\langle h, y-x\rangle + \frac{\beta}{2}\|x-y\|^2] \le \frac{1}{2}\left(-\alpha\eta + \frac{\beta\eta^2}{2}\right)\mathbb{E}\left[\|H\|^2\right] + \frac{2\eta}{\alpha}\mathbb{E}\left[\|g-h\|_*^2\right]$$

where $H = \frac{1}{\eta}(x - P_{h,\eta}(x))$. In particular, if $\|\cdot\| = \|\cdot\|_W$ for some positive definite matrix W, R is quadratic, and $\mathcal{K}$ is Euclidean space,

$$\mathbb{E}[\langle h, y-x\rangle + \frac{\beta}{2}\|x-y\|^2] \le \left(-\alpha\eta + \frac{\beta\eta^2}{2}\right)\mathbb{E}\left[\|H\|^2\right] + \frac{\beta\eta^2}{2}\mathbb{E}[\|H-G\|^2]$$

Proof. Let $G = \frac{1}{\eta}(x - P_{g,\eta}(x))$. First we show for the special case (i.e. suppose $R(x) = \frac{1}{2}\langle x, Mx\rangle$ for some positive definite matrix M, and therefore $G = M^{-1}g$ and $H = M^{-1}h$).

$$\mathbb{E}[\langle h, y-x\rangle] = -\eta\mathbb{E}[\langle h, G\rangle] = -\eta\mathbb{E}[\langle h, H\rangle] \le -\alpha\eta\|H\|^2$$

and because g is unbiased,

$$\mathbb{E}\left[\frac{\beta}{2}\|x-y\|^2\right] = \mathbb{E}\left[\frac{\eta^2\beta}{2}\|H\|^2 + \frac{\eta^2\beta}{2}\|G-H\|^2\right]$$

For general setups, we first separate the term into two parts

$$\langle h, y-x\rangle = \langle g, y-x\rangle + \langle h-g, y-x\rangle$$

For the first term, we use the optimality condition

$$\langle g + \frac{1}{\eta}\nabla R(y) - \frac{1}{\eta}\nabla R(x), z-y\rangle \ge 0, \quad \forall z \in \mathcal{K}$$

which implies

$$\langle g, x-y\rangle \geq \frac{1}{\eta}\langle \nabla R(y) - \nabla R(x), y-x\rangle \geq \frac{\alpha}{\eta}\|x-y\|^2$$

Therefore, we can bound the first term by

$$\langle g, y-x\rangle \leq -\frac{\alpha}{\eta}\|x-y\|^2 = -\alpha\eta\|G\|^2$$

On the other hand, for the second term, we first write

$$\begin{aligned}\langle h-g, y-x\rangle &= -\eta\langle h-g, G\rangle \\ &= -\eta\langle h-g, H\rangle + \eta\langle h-g, H-G\rangle\end{aligned}$$

and we show that

$$\langle h-g, H-G\rangle \leq \|h-g\|_*\|H-G\| \leq \frac{\|h-g\|_*^2}{\alpha} \tag{16}$$

This can be proved by Legendre transform:

$$\begin{aligned}P_{g,\eta}(x) &= \arg\min_{z\in\mathcal{K}}\langle g, z\rangle + \frac{1}{\eta}D_R(z||x) \\ &= \arg\min_{z\in\mathcal{K}}\langle g - \frac{1}{\eta}\nabla R(x), z\rangle + \frac{1}{\eta}R(z) \\ &= \nabla\left(\frac{1}{\eta}R\right)^*\left(\frac{1}{\eta}\nabla R(x) - g\right)\end{aligned}$$

Because $\frac{1}{\eta}R$ is $\frac{\alpha}{\eta}$-strongly convex with respect to norm $\|\cdot\|$, $\left(\frac{1}{\eta}R\right)^*$ is $\frac{\eta}{\alpha}$-smooth with respect to norm $\|\cdot\|_*$, we have

$$\|H-G\| \leq \frac{1}{\eta}\frac{\eta}{\alpha}\|g-h\|_* = \frac{1}{\alpha}\|g-h\|_*$$

which proves (16). Putting everything together, we have

$$\begin{aligned}&\mathbb{E}[\langle h, y-x\rangle + \frac{\beta}{2}\|x-y\|^2] \\ &\leq \mathbb{E}\left[\left(-\alpha\eta + \frac{\beta\eta^2}{2}\right)\|G\|^2\right] + \mathbb{E}\left[-\eta\langle h-g, H\rangle + \frac{\eta}{\alpha}\|g-h\|_*^2\right] \\ &= \mathbb{E}\left[\left(-\alpha\eta + \frac{\beta\eta^2}{2}\right)\|G\|^2\right] + \mathbb{E}\left[\frac{\eta}{\alpha}\|g-h\|_*^2\right]\end{aligned}$$

Because

$$\|H\|^2 \leq 2\|G\|^2 + 2\|H-G\|^2 \leq 2\|G\|^2 + \frac{2}{\alpha^2}\|h-g\|_*^2$$

it holds that

$$\begin{aligned}&\mathbb{E}[\langle h, y-x\rangle + \frac{\beta}{2}\|x-y\|^2] \\ &\leq \frac{1}{2}\left(-\alpha\eta + \frac{\beta\eta^2}{2}\right)\mathbb{E}\left[\|H\|^2\right] + \frac{2\eta}{\alpha}\mathbb{E}\left[\|g-h\|_*^2\right]\end{aligned}$$

■

Proof of Proposition 1 We apply Lemma 2: By smoothness of J,

$$\begin{aligned}\mathbb{E}\left[J(\pi_{n+1})\right] - J(\pi_n) &\leq \mathbb{E}\left[\langle \nabla J(\pi_n), \theta_{n+1} - \theta_n\rangle + \frac{\beta}{2}\|\theta_{n+1} - \theta_n\|^2\right] \\ &\leq \frac{1}{2}\left(-\alpha_n\eta_n + \frac{\beta\eta_n^2}{2}\right)\mathbb{E}\left[\|\hat{\nabla}_\theta J(\pi_n)\|^2\right] + \frac{2\eta_n}{\alpha_n}\|\nabla_\theta J(\pi_n) - g_n\|_*^2\end{aligned}$$

This proves the statement in Proposition 1. We note that, in the above step, the general result of Lemma 2. For the special case Lemma 2, we would recover the usual convergence property of stochastic smooth nonconvex optimization, which shows on average convergence to stationary points in expectation.

B.2 Proof of Proposition 2

We use a well-know result of mirror descent, whose proof can be found e.g. in (Juditsky et al., 2011).

Lemma 3. *Let $\mathcal{K}$ be a convex set. Suppose R is α-strongly convex with respect to norm $\|\cdot\|$. Let g be a vector in some Euclidean space and let*

$$y = \arg\min_{z\in\mathcal{K}} \langle g, z\rangle + \frac{1}{\eta} D_R(z||x) = P_{g,\eta}(x)$$

Then for all $z \in \mathcal{K}$

$$\eta\langle g, x - z\rangle \leq D_R(z||x) - D_R(z||y) + \frac{\eta^2}{2}\|g\|_*^2$$

Next we prove a lemma of performing online mirror descent with weighted cost. While weighting it not required in proving Proposition 2, it will be useful to prove Theorem 2 later in Appendix C.

Lemma 4. *Let f_n be σ-strongly convex with respect to some strongly convex function R_n, i.e.*

$$f_n(x) \geq f_n(y) + \langle \nabla f_n(y), x - y\rangle + \sigma D_{R_n}(x||y)$$

and let $\{w_n\}_{n=1}^N$ be a sequence of positive numbers. Consider the update rule

$$x_{n+1} = \arg\min_{z\in\mathcal{K}} \langle w_n g_n, x\rangle + \frac{1}{\eta_n} D_{R_n}(z||x_n)$$

where $g_n = \nabla f_n(x_n)$ and $\eta_n = \frac{1}{\hat{\sigma}\sum_{m=1}^n w_m}$. Suppose $\hat{\sigma} \leq \sigma$. Then for all $x^* \in \mathcal{K}$, $N \geq M \geq 1$, it holds that

$$\sum_{n=M}^N w_n f_n(x_n) - w_n f_n(x^*) \leq \hat{\sigma} D_{R_M}(x^*||x_M)\sum_{n=1}^{M-1} w_n + \frac{1}{2\hat{\sigma}}\sum_{n=1}^N \frac{w_n^2\|g_n\|_*^2}{\sum_{m=1}^n w_m}$$

Proof. The proof is straight forward by strong convexity of f_n and Lemma 3.

$$\begin{aligned}&\sum_{n=M}^N w_n(f_n(x_n) - f_n(x^*)) \\ &\leq \sum_{n=M}^N w_n\left(\langle g_n, x_n - x^*\rangle - \sigma D_{R_n}(x^*||x_n)\right) \qquad (\sigma\text{-strong convexity}) \\ &\leq \sum_{n=M}^N \frac{1}{\eta_n} D_{R_n}(x^*||x_n) - \frac{1}{\eta_n} D_{R_n}(x^*||x_{n+1}) - w_n\sigma D_{R_n}(x^*||x_n) + \frac{w_n^2\eta_n}{2}\|g_n\|_*^2 \qquad (\text{Lemma 3}) \\ &\leq \frac{D_{R_M}(x^*||x_M)}{\eta_{M-1}} + \sum_{n=M}^N \left(\frac{1}{\eta_n} - \frac{1}{\eta_{n-1}} - w_n\sigma\right) D_{R_n}(x^*||x_n) + \frac{w_n^2\eta_n}{2}\|g_n\|_*^2 \qquad (\text{We define } \frac{1}{\eta_0} = 0)\end{aligned}$$

$$= \hat{\sigma} D_{R_M}(x^*||x_M) \sum_{n=1}^{M-1} w_n + \sum_{n=1}^{N} \left(w_n \hat{\sigma} - w_n \sigma\right) D_{R_n}(x^*||x_n) + \frac{1}{2\hat{\sigma}} \sum_{n=1}^{N} \frac{w_n^2 \|g_n\|_*^2}{\sum_{m=1}^{n} w_m}$$
$$\leq \hat{\sigma} D_{R_M}(x^*||x_M) \sum_{n=1}^{M-1} w_n + \frac{1}{2\hat{\sigma}} \sum_{n=1}^{N} \frac{w_n^2 \|g_n\|_*^2}{\sum_{m=1}^{n} w_m}$$

■

Proof of Proposition 2 Now we use Lemma 4 to prove the final result. It's easy to see that if g_n is an unbiased stochastic estimate of $\nabla f_n(x_n)$ in Lemma 4, then the performance bound would hold in expectation since x_n does not depend on g_n. Finally, by definition of ϵ_{class}, this concludes the proof.

C Proof of Section 5

C.1 Proof of Theorem 1

Let $w_n = n^d$. The proof is similar to the proof of Proposition 2 but with weighted cost. First we use Lemma 1 and bound the series of weighted accumulated loss

$$\mathbb{E}\left[\sum_{n=N_m}^{N_M} w_n J(\pi_n)\right] - \left(\sum_{n=N_m}^{N_M} w_n\right) J(\pi^*) \leq \frac{C_{\pi^*}}{1-\gamma} \sum_{n=N_m}^{N_M} w_n l_n(\pi_n)$$

Then we bound the right-hand side by using Lemma 4,

$$\sum_{n=N_m}^{N_M} w_n l_n(\pi_n) - \min_{\pi \in \Pi} \sum_{n=N_m}^{N_M} w_n l_n(\pi) \leq \hat{\sigma} D_{\mathcal{R}} \sum_{n=1}^{N_m - 1} w_n + \frac{1}{2\hat{\sigma}} \sum_{n=N_m}^{N_M} \frac{w_n^2 \|g_n\|_*^2}{\sum_{m=1}^{n} w_m}$$
$$\leq \frac{\hat{\sigma} D_{\mathcal{R}} N_m^{d+1}}{d+1} + \frac{d+1}{2\hat{\sigma}} \sum_{n=N_m}^{N_M} \|g_n\|_*^2 n^{d-1}$$

where we use the fact that $d \geq 0$,

$$\frac{n^{d+1} - (m-1)^{d+1}}{d+1} \leq \sum_{k=m}^{n} k^d \leq \frac{(n+1)^{d+1} - m^{d+1}}{d+1}$$

which implies $\frac{w_n^2}{\sum_{m=1}^{n} w_m} \leq \frac{(d+1)n^{2d}}{n^{d+1}} \leq (d+1)n^{d-1}$. Combining these two steps, we see that the weighted accumulated loss on average can be bounded by

$$\mathbb{E}\left[\frac{\sum_{n=N_m}^{N_M} w_n J(\pi_n)}{\sum_{n=N_m}^{N_M} w_n}\right] \leq J(\pi^*) + \frac{C_{\pi^*}}{1-\gamma}\left(\epsilon_{\text{class}}^w + \frac{\hat{\sigma} D_{\mathcal{R}} N_m^{d+1}}{(d+1)\sum_{n=N_m}^{N_M} w_n} + \frac{d+1}{2\hat{\sigma}\sum_{n=N_m}^{N_M} w_n} \sum_{n=N_m}^{N_M} \|g_n\|_*^2 n^{d-1}\right)$$

Because $N_M \geq 2N_m$ and $\frac{x}{1-x} \leq 2x$ for $x \leq \frac{1}{2}$, we have

$$\frac{N_m^{d+1}}{(d+1)\sum_{n=N_m}^{N_M} w_n} \leq \frac{N_m^{d+1}}{N_M^{d+1} - (N_m - 1)^{d+1}} \leq \frac{N_m^{d+1}}{N_M^{d+1} - N_m^{d+1}} = \frac{\left(\frac{N_m}{N_M}\right)^{d+1}}{1 - \left(\frac{N_m}{N_M}\right)^{d+1}} \leq 2\left(\frac{N_m}{N_M}\right)^{d+1} \leq 2^{-d}$$

and, for $d \geq 1$,

$$\begin{aligned}
\frac{d+1}{\sum_{n=N_m}^{N_M} w_n} \sum_{n=N_m}^{N_M} n^{d-1} &\leq \frac{d+1}{N_M^{d+1} - (N_m - 1)^{d+1}} \frac{d+1}{d} \left((N_M + 1)^d - N_m^d \right) \\
&\leq \frac{(d+1)^2}{d} \frac{(N_M+1)^d}{N_M^{d+1} - N_m^{d+1}} \\
&\leq \frac{(d+1)^2}{d} \frac{\frac{1}{N_M}(1 + \frac{1}{N_M})^d}{1 - \left(\frac{N_m}{N_M}\right)^{d+1}} \\
&\leq \frac{16d}{3N_M} \left(1 + \frac{1}{N_M}\right)^d \qquad (N_M \geq 2N_m \text{ and } d \geq 1) \\
&\leq \frac{16d}{3N_M} \exp\left(\frac{d}{N_M}\right)
\end{aligned}$$

and for $d = 0$,

$$\frac{d+1}{\sum_{n=N_m}^{N_M} w_n} \sum_{n=N_m}^{N_M} n^{d-1} = \frac{1}{\sum_{n=N_m}^{N_M} 1} \sum_{n=N_m}^{N_M} \frac{1}{n} \leq \frac{\log(N_M) + 1}{N_M - N_m} \leq \frac{2\left(\log(N_M) + 1\right)}{N_M}$$

Thus, by the assumption that $\|g_n\|_* \leq G$ almost surely, the weighted accumulated loss on average has an upper bound

$$\mathbb{E}\left[\frac{\sum_{n=N_m}^{N_M} w_n J(\pi_n)}{\sum_{n=N_m}^{N_M} w_n}\right] \leq J(\pi^*) + \frac{C_{\pi^*}}{1-\gamma}\left(\epsilon_{\text{class}}^w + 2^{-d}\hat{\sigma} D_{\mathcal{R}} + \frac{G^2 C_{N_M}/\hat{\sigma}}{N_M}\right)$$

By sampling K according to w_s, this bound directly translates into the the bound on $J(\pi_K)$.

C.2 Proof of Theorem 3

For simplicity, we prove the result of deterministic problems. For stochastic problems, the result can be extended to expected performance, similar to the proof of Proposition 2. We first define the online learning problem of applying $g_n = \nabla_\theta l_n^\lambda(\pi)|_{\pi=\pi_n}$ to update the policy. In the nth iteration, we define the per-round cost as

$$l_n^\lambda(\pi) = \mathbb{E}_{d_{\pi_n}} \mathbb{E}_\pi[(1-\lambda) A_{\pi_n} + \lambda A_{\pi^*}] \tag{17}$$

With the strongly convexity assumption and large enough step size, similar to the proof for Proposition 2, we can show

$$\begin{aligned}
\sum_{n=1}^N l_n^\lambda(\pi_n) &\leq \min_{\pi \in \Pi} \sum_{n=1}^N (l_n^\lambda(\pi) + \epsilon_{\text{regret}}^\lambda) \\
&= \min_{\pi \in \Pi} \sum_{n=1}^N \mathbb{E}_{d_{\pi_n}} \mathbb{E}_\pi[(1-\lambda) A_{\pi_n} + \lambda A_{\pi^*}] + N\epsilon_{\text{regret}}^\lambda
\end{aligned}$$

where $\epsilon_{\text{regret}}^\lambda = \tilde{O}\left(\frac{1}{T}\right)$. Note by definition of A_{π_n}, the left-hand-side in the above bound can be written as

$$\frac{1}{N} \sum_{n=1}^N l_n^\lambda(\pi_n) = \sum_{n=1}^N \mathbb{E}_{d_{\pi_n}} \mathbb{E}_{\pi_n}[(1-\lambda) A_{\pi_n} + \lambda A_{\pi^*}] = \sum_{n=1}^N \lambda \mathbb{E}_{d_{\pi_n}} \mathbb{E}_{\pi_n}[A_{\pi^*}] \tag{18}$$

To relate this to the performance bound, we invoke Lemma 1 and write

$$
\begin{aligned}
&\sum_{n=1}^{N} J(\pi_n) - ((1-\lambda)J_n^* + \lambda J(\pi^*)) \\
&= \sum_{n=1}^{N} (1-\lambda)\left(J(\pi_n) - J_n^*\right) + \frac{1}{1-\gamma}\sum_{n=1}^{N} \lambda \mathbb{E}_{d_{\pi_n}} \mathbb{E}_{\pi_n}[A_{\pi^*}] \\
&\leq \sum_{n=1}^{N} (1-\lambda)\left(J(\pi_n) - J_n^*\right) + \min_{\pi\in\Pi} \frac{1}{1-\gamma}\sum_{n=1}^{N} \mathbb{E}_{d_{\pi_n}} \mathbb{E}_{\pi}[(1-\lambda)A_{\pi_n} + \lambda A_{\pi^*}] + \frac{N}{1-\gamma}\epsilon_{\text{regret}}^{\lambda} \\
&= \min_{\pi\in\Pi} \frac{1}{1-\gamma}\sum_{n=1}^{N} \mathbb{E}_{d_{\pi_n}} \mathbb{E}_{\pi}[(1-\lambda)Q_{\pi_n} + \lambda A_{\pi^*}] + \frac{N}{1-\gamma}\epsilon_{\text{regret}}^{\lambda} + \sum_{n=1}^{N}(1-\lambda)\left(-\frac{\mathbb{E}_{d_{\pi_n}} \mathbb{E}_{\pi_n}[Q_{\pi_n}]}{1-\gamma} + J(\pi_n) - J_n^*\right) \\
&= \min_{\pi\in\Pi} \frac{1}{1-\gamma}\sum_{n=1}^{N} \mathbb{E}_{d_{\pi_n}} \mathbb{E}_{\pi}[(1-\lambda)(Q_{\pi_n} - V_{\pi_n}^*) + \lambda(Q_{\pi^*} - V_{\pi^*})] + \frac{N}{1-\gamma}\epsilon_{\text{regret}}^{\lambda} \qquad \text{(Since } \mathbb{E}_{d_{\pi_n}} \mathbb{E}_{\pi_n}[Q_{\pi_n}] = (1-\gamma)J(\pi_n)\text{)} \\
&= \min_{\pi\in\Pi} \frac{1}{1-\gamma}\sum_{n=1}^{N} \mathbb{E}_{d_{\pi_n}} \mathbb{E}_{\pi}[(1-\lambda)Q_{\pi_n} + \lambda Q_{\pi^*}] - \frac{1}{1-\gamma}\sum_{n=1}^{N} \mathbb{E}_{d_{\pi_n}}\left[(1-\lambda)V_{\pi_n}^* + \lambda V_{\pi^*}\right] + \frac{N}{1-\gamma}\epsilon_{\text{regret}}^{\lambda} \\
&\leq \frac{N}{1-\gamma}\epsilon_{\text{class}}^{\lambda} + \frac{N}{1-\gamma}\epsilon_{\text{regret}}^{\lambda}
\end{aligned}
$$

This concludes the proof.