Supplementary Material

Combining Pessimism with Optimism for Robust and Efficient Model-based Deep Reinforcement Learning

Submitted to ICML 2021

A. Relevant Definitions and Results

Lemma 1 (Adapted from Corollary 1 in Curi et al. (2020a))**.** *Based on Assumptions 1 and 3, for every* $\mathrm{s}, \mathrm{s}' \in \mathcal{S}$*, it holds:*

$$\|f(\mathrm{s}, \pi(\mathrm{s}), \bar{\pi}(\mathrm{s})) - f(\mathrm{s}', \pi(\mathrm{s}'), \bar{\pi}(\mathrm{s}'))\|_2 \leq L_f \sqrt{1 + L_\pi^2 + L_{\bar{\pi}}^2} \|\mathrm{s} - \mathrm{s}'\|_2. \tag{14}$$

Proof.

$$\begin{aligned} \|f(\mathrm{s}, \pi(\mathrm{s}), \bar{\pi}(\mathrm{s})) - f(\mathrm{s}', \pi(\mathrm{s}'), \bar{\pi}(\mathrm{s}'))\|_2 &\leq L_f \sqrt{\|\mathrm{s} - \mathrm{s}'\|_2^2 + \|\pi(\mathrm{s}) - \pi(\mathrm{s}')\|_2^2 + \|\bar{\pi}(\mathrm{s}') - \bar{\pi}(\mathrm{s})\|_2^2} & \text{(15a)} \\ &\leq \sqrt{\|\mathrm{s} - \mathrm{s}'\|_2^2 + L_\pi^2 \|\mathrm{s} - \mathrm{s}'\|_2^2 + L_{\bar{\pi}}^2 \|\mathrm{s} - \mathrm{s}'\|_2^2} & \text{(15b)} \\ &= L_f \sqrt{1 + L_\pi^2 + L_{\bar{\pi}}^2} \|\mathrm{s} - \mathrm{s}'\|_2. & \text{(15c)} \end{aligned}$$

Eq. (15a) holds due to Lipschitz continuity of f and Eq. (15b) is due to Lipschitz continuity of π and $\bar{\pi}$, which we assume in Assumptions 1 and 3. □

Lemma 2 (Adapted from Lemma 3 in Curi et al. (2020a))**.** *Based on Assumptions 1 and 3, it holds:*

$$|J(f, \pi, \bar{\pi}) - J(\tilde{f}, \pi, \bar{\pi})| \leq L_r \sqrt{1 + L_\pi^2 + L_{\bar{\pi}}^2} \sum_{h=0}^{H} \mathbb{E}[\|\mathrm{s}_h - \tilde{\mathrm{s}}_h\|_2], \tag{16}$$

where $\tilde{\mathrm{s}}_h$ *for* $h = 0, \ldots, H$ *is the trajectory generated by the dynamics* $\tilde{f}$*, starting from* $\tilde{\mathrm{s}}_0 = \mathrm{s}_0$ *with* $\omega_h = \tilde{\omega}_h$*.*

Proof.

$$\begin{aligned} |J(f, \pi, \bar{\pi}) - J(\tilde{f}, \pi, \bar{\pi})| &= \left| \mathbb{E}\left[\sum_{h=0}^{H} r(\mathrm{s}, \mathbf{a}, \bar{\mathbf{a}}) - \sum_{h=0}^{H} r(\tilde{\mathrm{s}}, \tilde{\mathbf{a}}, \tilde{\bar{\mathbf{a}}}) \right] \right| & \text{(17a)} \\ &= \left| \sum_{h=0}^{H} \mathbb{E}\left[r(\mathrm{s}, \mathbf{a}, \bar{\mathbf{a}}) - r(\tilde{\mathrm{s}}, \tilde{\mathbf{a}}, \tilde{\bar{\mathbf{a}}}) \right] \right| & \text{(17b)} \\ &\leq L_r \sqrt{1 + L_\pi^2 + L_{\bar{\pi}}^2} \sum_{h=0}^{H} \mathbb{E}[\|\mathrm{s}_h - \tilde{\mathrm{s}}_h\|_2]. & \text{(17c)} \end{aligned}$$

Eq. (17a) follows by definition of J, Eq. (17b) from linearity of expectation, and Eq. (17c) from Lipschitzness of the policy and the reward function, which we assume in Assumption 1.

□

The following lemma bounds the deviation between the optimistic/pessimistic and the true trajectory in a single episode.

Lemma 3 (Adapted from Lemma 4 in (Curi et al., 2020a))**.** *Under Assumptions 1 to 3, for all episodes* $t \geq 1$*, any* $\eta \in [-1, 1]$*,* $h \in \{1, \ldots, H\}$*,* $\pi \in \Pi$ *and* $\bar{\pi} \in \bar{\Pi}$ *it holds:*

$$\|\mathrm{s}_{h,t} - \tilde{\mathrm{s}}_{h,t}\|_2 \leq 2\beta_{t-1} \Big(1 + (L_f + 2\beta_{t-1} L_\sigma) \sqrt{1 + L_\pi^2 + L_{\bar{\pi}}^2}\Big)^{h-1} \sum_{h'=0}^{h-1} \|\boldsymbol{\sigma}_{t-1}(\mathrm{s}_{h',t}, \pi_t(\mathrm{s}_{h',t}), \bar{\pi}_t(\mathrm{s}_{h',t}))\|_2, \tag{18}$$

where $\tilde{\mathrm{s}}_{h,t}$ *is generated by any system* $\tilde{f} \in \mathcal{M}_t := \Big\{ \tilde{f}$ *s.t.* $|\tilde{f}(\mathrm{s}, \mathbf{a}, \bar{\mathbf{a}}) - \boldsymbol{\mu}_{t-1}(\mathrm{s}, \mathbf{a}, \bar{\mathbf{a}})| \leq \beta_t \boldsymbol{\sigma}_{t-1}(\mathrm{s}, \mathbf{a}, \bar{\mathbf{a}}) \Big\}$*. We refer to* $\mathcal{M}_t$ *as the set of plausible models at the beginning of episode* t*.*

Proof. To avoid notational clutter, we denote the closed-loop dynamics as $f^{\pi,\bar{\pi}}(\mathbf{s}) = f(\mathbf{s}, \pi(\mathbf{s}), \bar{\pi}(\mathbf{s}))$ and the closed-loop epistemic uncertainty as $\boldsymbol{\sigma}^{\pi,\bar{\pi}}(\mathbf{s}) = \boldsymbol{\sigma}(\mathbf{s}, \pi(\mathbf{s}), \bar{\pi}(\mathbf{s}))$. Likewise, we use the following Lipschitz constants shorthands $L_{f,\pi} \equiv L_f\sqrt{1 + L_\pi^2 + L_{\bar{\pi}}^2}$ and $L_{\sigma,\pi} \equiv L_\sigma\sqrt{1 + L_\pi^2 + L_{\bar{\pi}}^2}$.

We first prove by induction that

$$\|\mathbf{s}_{h,t} - \tilde{\mathbf{s}}_{h,t}\|_2 \leq 2\beta_{t-1} \sum_{h'=0}^{h-1} \left(L_{f,\pi} + 2\beta_{t-1}L_{\sigma,\pi}\right)^{h-1-h'} \|\boldsymbol{\sigma}_{t-1}^{\pi_t,\bar{\pi}_t}(\mathbf{s}_{h',t})\| \tag{19a}$$

For $h = 0$, clearly $\mathbf{s}_{0,t} = \tilde{\mathbf{s}}_{0,t}$, while the right-hand-side of Eq. (19a) is always non-negative. We assume that for h the inductive hypothesis (19a) holds. For $h + 1$ we have:

$$\begin{aligned}
\|\mathbf{s}_{h+1,t} - \tilde{\mathbf{s}}_{h+1,t}\|_2 &= \|f^{\pi_t,\bar{\pi}_t}(\mathbf{s}_{h,t}) - \tilde{f}^{\pi_t,\bar{\pi}_t}(\tilde{\mathbf{s}}_{h,t})\|_2 && (19b)\\
&= \|f^{\pi_t,\bar{\pi}_t}(\mathbf{s}_{h,t}) - \tilde{f}^{\pi_t,\bar{\pi}_t}(\tilde{\mathbf{s}}_{h,t}) + f^{\pi_t,\bar{\pi}_t}(\tilde{\mathbf{s}}_{h,t}) - f^{\pi_t,\bar{\pi}_t}(\tilde{\mathbf{s}}_{h,t})\|_2 && (19c)\\
&\leq \|f^{\pi_t,\bar{\pi}_t}(\mathbf{s}_{h,t}) - f^{\pi_t,\bar{\pi}_t}(\tilde{\mathbf{s}}_{h,t})\|_2 + \|f^{\pi_t,\bar{\pi}_t}(\tilde{\mathbf{s}}_{h,t}) - \tilde{f}^{\pi_t,\bar{\pi}_t}(\tilde{\mathbf{s}}_{h,t})\|_2 && (19d)\\
&\leq L_{f,\pi}\|\mathbf{s}_{h,t} - \tilde{\mathbf{s}}_{h,t}\|_2 + \|f^{\pi_t,\bar{\pi}_t}(\tilde{\mathbf{s}}_{h,t}) - \tilde{f}^{\pi_t,\bar{\pi}_t}(\tilde{\mathbf{s}}_{h,t})\|_2 && (19e)\\
&\leq L_{f,\pi}\|\mathbf{s}_{h,t} - \tilde{\mathbf{s}}_{h,t}\|_2 + 2\beta_{t-1}\|\boldsymbol{\sigma}_{t-1}^{\pi_t,\bar{\pi}_t}(\tilde{\mathbf{s}}_{h,t})\|_2 && (19f)\\
&= L_{f,\pi}\|\mathbf{s}_{h,t} - \tilde{\mathbf{s}}_{h,t}\|_2 + 2\beta_{t-1}\|\boldsymbol{\sigma}_{t-1}^{\pi_t,\bar{\pi}_t}(\tilde{\mathbf{s}}_{h,t}) + \boldsymbol{\sigma}_{t-1}^{\pi_t,\bar{\pi}_t}(\mathbf{s}_{h,t}) - \boldsymbol{\sigma}_{t-1}^{\pi_t,\bar{\pi}_t}(\mathbf{s}_{h,t})\|_2 && (19g)\\
&\leq L_{f,\pi}\|\mathbf{s}_{h,t} - \tilde{\mathbf{s}}_{h,t}\|_2 + 2\beta_{t-1}\left(\|\boldsymbol{\sigma}_{t-1}^{\pi_t,\bar{\pi}_t}(\tilde{\mathbf{s}}_{h,t}) - \boldsymbol{\sigma}_{t-1}^{\pi_t,\bar{\pi}_t}(\mathbf{s}_{h,t})\|_2 + \|\boldsymbol{\sigma}_{t-1}^{\pi_t,\bar{\pi}_t}(\mathbf{s}_{h,t})\|_2\right) && (19h)\\
&\leq \left(L_{f,\pi} + 2\beta_{t-1}L_{\sigma,\pi}\right)\|\mathbf{s}_{h,t} - \tilde{\mathbf{s}}_{h,t}\|_2 + 2\beta_{t-1}\|\boldsymbol{\sigma}_{t-1}^{\pi_t,\bar{\pi}_t}(\mathbf{s}_{h,t})\|_2 && (19i)\\
&\leq 2\beta_{t-1}\sum_{h'=0}^{(h+1)-1}\left(L_{f,\pi} + 2\beta_{t-1}L_{\sigma,\pi}\right)^{(h+1)-1-h'}\|\boldsymbol{\sigma}_{t-1}^{\pi_t,\bar{\pi}_t}(\mathbf{s}_{h',t})\|_2 && (19j)
\end{aligned}$$

Here, Eq. (19b) holds by applying the transition dynamics $f^{\pi_t,\bar{\pi}_t}$ and $\tilde{f}^{\pi_t,\bar{\pi}_t}$ with the same noise realization $\omega_h = \tilde{\omega}_h$; Eq. (19c) holds by adding and subtracting $f^{\pi_t,\bar{\pi}_t}(\tilde{\mathbf{s}}_{h,t})$; Eq. (19d) follows from the triangular inequality; Eq. (19e) comes from Lemma 1; Eq. (19f) holds due to both f and $\tilde{f}$ belonging to the set of plausible models $\mathcal{M}_t$ (see Lemma 3 for its definition); Eq. (19g) holds by adding and substracting $\boldsymbol{\sigma}^{\pi_t,\bar{\pi}_t}(\mathbf{s}_{h,t})$; Eq. (19h) is by applying the triangular inequality once more; Eq. (19i) is due to the Lipschitz continuity of $\boldsymbol{\sigma}$ as per assumption 3; and (19j) holds by replacing the inductive hypothesis from (19a).

Finally, we notice that $\left(L_{f,\pi} + 2\beta_{t-1}L_{\sigma,\pi}\right)^{h-1-h'} < \left(1 + L_{f,\pi} + 2\beta_{t-1}L_{\sigma,\pi}\right)^{h-1-h'} \leq \left(1 + L_{f,\pi} + 2\beta_{t-1}L_{\sigma,\pi}\right)^{h-1}$ and the main result follows by combining this with Eq. (19j).

□

B. Proofs from Section 4.4

We start the analysis of the performance of `RH-UCRL`, by first bounding its instantaneous robust-regret by the difference between optimistic and pessimistic performance estimates.

Lemma 4. *Let $\pi^\star$ be the benchmark policy from Eq. (3), and let π_t and $\bar{\pi}_t$ be the policies selected by `RH-UCRL` at time t. Under the callibrated model Assumption 2, the following holds with probability at least $1 - \delta$:*

$$\min_{\bar{\pi}\in\bar{\Pi}} J(f, \pi^\star, \bar{\pi}) - \min_{\bar{\pi}\in\bar{\Pi}} J(f, \pi_t, \bar{\pi}) \leq J_t^{(o)}(\pi_t, \bar{\pi}_t) - J^{(p)}(\pi_t, \bar{\pi}_t). \tag{20}$$

Proof. We refer to the considered quantity $\min_{\bar{\pi}\in\bar{\Pi}} J(f, \pi^\star, \bar{\pi}) - \min_{\bar{\pi}\in\bar{\Pi}} J(f, \pi_t, \bar{\pi})$ as the robust instantaneous regret of the selected policy π_t, and we proceed by providing its upper bound:

$$\begin{aligned}
\min_{\bar{\pi}\in\bar{\Pi}} J(f,\pi^\star,\bar{\pi}) - \min_{\bar{\pi}\in\bar{\Pi}} J(f,\pi_t,\bar{\pi}) &\leq \min_{\bar{\pi}\in\bar{\Pi}} J_t^{(o)}(\pi^\star,\bar{\pi}) - \min_{\bar{\pi}\in\bar{\Pi}} J(f,\pi_t,\bar{\pi}) && \text{(21a)}\\
&\leq \min_{\bar{\pi}\in\bar{\Pi}} J_t^{(o)}(\pi_t,\bar{\pi}) - \min_{\bar{\pi}\in\bar{\Pi}} J(f,\pi_t,\bar{\pi}) && \text{(21b)}\\
&\leq J_t^{(o)}(\pi_t,\bar{\pi}_t) - \min_{\bar{\pi}\in\bar{\Pi}} J(f,\pi_t,\bar{\pi}) && \text{(21c)}\\
&\leq J_t^{(o)}(\pi_t,\bar{\pi}_t) - \min_{\bar{\pi}\in\bar{\Pi}} J^{(p)}(\pi_t,\bar{\pi}) && \text{(21d)}\\
&= J_t^{(o)}(\pi_t,\bar{\pi}_t) - J^{(p)}(\pi_t,\bar{\pi}_t). && \text{(21e)}
\end{aligned}$$

Here, inequality (21a) holds by definition of the optimistic estimate in Eq. (5a); inequality (21b) holds by definition of protagonist policy in the `RH-UCRL` algorithm (7a); and inequality (21d) holds by definition of the pessimistic estimate in Eq. (6a); finally, equality (21e) holds by definition of the antagonist policy in the `RH-UCRL` algorithm (7b). □

Lemma 5. *Under Assumptions 1 to 3, let π_t and $\bar{\pi}_t$ be the policies selected by `RH-UCRL` at episode t. Then, the following holds for the difference between its optimistic and pessimistic performance:*

$$J_t^{(o)}(\pi_t,\bar{\pi}_t) - J^{(p)}(\pi_t,\bar{\pi}_t) \leq 4L_r\beta_T^H C^H \sum_{h=0}^{H} \mathbb{E}\left[\sum_{h'=0}^{h-1} \|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t},\pi_t(\mathbf{s}_{h',t}),\bar{\pi}_t(\mathbf{s}_{h',t}))\|_2\right], \tag{22}$$

where $C = (1+L_f+L_\sigma)(1+L_\pi^2+L_{\bar{\pi}}^2)^{1/2}$.

Proof.

$$\begin{aligned}
J_t^{(o)}(\pi_t,\bar{\pi}_t) - J^{(p)}(\pi_t,\bar{\pi}_t) &\leq \left|J_t^{(o)}(\pi_t,\bar{\pi}_t) - J(f,\pi_t,\bar{\pi}_t)\right| + \left|J_t^{(p)}(\pi_t,\bar{\pi}_t) - J(f,\pi_t,\bar{\pi}_t)\right| && \text{(23a)}\\
&\leq L_r\sqrt{1+L_\pi^2+L_{\bar{\pi}}^2}\sum_{h=0}^{H}\left(\mathbb{E}\left[\|\mathbf{s}_{h,t}-\mathbf{s}_{h,t}^{(o)}\|_2\right] + \mathbb{E}\left[\|\mathbf{s}_{h,t}-\mathbf{s}_{h,t}^{(p)}\|_2\right]\right) && \text{(23b)}
\end{aligned}$$

Here, Eq. (23a) holds by the triangle inequality and Eq. (23b) follows from Lemma 2.

We proceed to upper bound terms $\|\mathbf{s}_{h,t}-\mathbf{s}_{h,t}^{(o)}\|_2$ and $\|\mathbf{s}_{h,t}-\mathbf{s}_{h,t}^{(p)}\|_2$. From Lemma 3, it follows that both terms can be bounded in the same way as follows:

$$\|\mathbf{s}_{h,t}-\mathbf{s}_{h,t}^{(o)}\|_2 \leq 2\beta_{t-1}\left((1+L_f+2\beta_{t-1}L_\sigma)\sqrt{1+L_\pi^2+L_{\bar{\pi}}^2}\right)^{h-1}\sum_{h'=0}^{h-1}\|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t},\pi_t(\mathbf{s}_{h',t}),\bar{\pi}_t(\mathbf{s}_{h',t}))\|_2, \tag{23c}$$

as $f^{(o)}$ and $f^{(p)}$ belong to the set of plausible models $\mathcal{M}_t$ (from Lemma 3).

By applying the previous bound twice in Eq. (23b), and by denoting

$$C := (1+L_f+2L_\sigma)(1+L_\pi^2+L_{\bar{\pi}}^2)^{1/2},$$

we arrive at:

$$J_t^{(o)}(\pi_t,\bar{\pi}_t) - J^{(p)}(\pi_t,\bar{\pi}_t) \leq 4L_r\beta_T^H C^H \sum_{h=0}^{H} \mathbb{E}\left[\sum_{h'=0}^{h-1} \|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t},\pi_t(\mathbf{s}_{h',t}),\bar{\pi}_t(\mathbf{s}_{h',t}))\|_2\right], \tag{23d}$$

where we used $t \leq T$ and $1 \leq \beta_t$ is non-decreasing in t.

□

Theorem 1. *Under Assumptions 1 to 3, let $C = (1+L_f+2L_\sigma)(1+L_\pi^2+L_{\bar{\pi}}^2)^{1/2}$ and let $\mathbf{s}_{t,h}\in\mathcal{S}$, $\mathbf{a}_{t,h}\in\mathcal{A}$, $\bar{\mathbf{a}}_{t,h}\in\bar{\mathcal{A}}$ for all $t,h>0$. Then, for any fixed $H\geq 1$, with probability at least $1-\delta$, the robust cumulative regret of `RH-UCRL` is upper bounded by:*

$$R_T = \mathcal{O}\Big(L_r C^H \beta_T^H H^{3/2}\sqrt{T\,\Gamma_T}\Big).$$

Proof of Theorem 1. We bound the robust cumulative regret as follows:

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} \underbrace{\min_{\bar{\pi}\in\bar{\Pi}} J(f,\pi^\star,\bar{\pi}) - \min_{\bar{\pi}\in\bar{\Pi}} J(f,\pi_t,\bar{\pi}_t)}_{:=r_t} && \text{(24a)}\\
&\le \sqrt{T\sum_{t=1}^{T} r_t^2} && \text{(24b)}\\
&\le \sqrt{T\sum_{t=1}^{T}(4L_r\beta_T^H C^H)^2\left(\sum_{h=0}^{H}\mathbb{E}\left[\sum_{h'=0}^{h-1}\|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t},\pi_t(\mathbf{s}_{h',t}),\bar{\pi}_t(\mathbf{s}_{h',t}))\|_2\right]\right)^2} && \text{(24c)}\\
&= 4L_r\beta_T^H C^H\sqrt{T}\sqrt{\sum_{t=1}^{T}\left(\sum_{h=0}^{H}\mathbb{E}\left[\sum_{h'=0}^{h-1}\|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t},\pi_t(\mathbf{s}_{h',t}),\bar{\pi}_t(\mathbf{s}_{h',t}))\|_2\right]\right)^2} && \text{(24d)}\\
&\le 4L_r\beta_T^H C^H H\sqrt{T}\sqrt{\sum_{t=1}^{T}\left(\mathbb{E}\left[\sum_{h'=0}^{H}\|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t},\pi_t(\mathbf{s}_{h',t}),\bar{\pi}_t(\mathbf{s}_{h',t}))\|_2\right]\right)^2} && \text{(24e)}\\
&\le 4L_r\beta_T^H C^H H\sqrt{T}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\left(\sum_{h'=0}^{H}\|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t},\pi_t(\mathbf{s}_{h',t}),\bar{\pi}_t(\mathbf{s}_{h',t}))\|_2\right)^2\right]} && \text{(24f)}\\
&\le 4L_r\beta_T^H C^H H^{3/2}\sqrt{T}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\sum_{h'=0}^{H}\|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t},\pi_t(\mathbf{s}_{h',t}),\bar{\pi}_t(\mathbf{s}_{h',t}))\|_2^2\right]} && \text{(24g)}\\
&\le 4L_r\beta_T^H C^H H^{3/2}\sqrt{T\Gamma_T}, && \text{(24h)}
\end{aligned}
$$

where Eq. (24b) is due to the Cauchy-Schwarz's inequality; Eq. (24c) is due to Lemma 4 and Lemma 5. Finally, Eq. (24f) follows from Jensen's inequality, Eq. (24g) follows from Cauchy-Schwarz's inequality, and Eq. (24h) follows from the definition of Γ_T.

□

Corollary 1. *Consider the assumptions and setup of Theorem 1, and suppose that*

$$\frac{T}{\beta_T^{2H}\Gamma_T} \ge \frac{16L_r^2H^3C^{2H}}{\epsilon^2}, \tag{10}$$

for some fixed $\epsilon > 0$ and $H \ge 1$. Then, with probability at least $1-\delta$ after T episodes, `RH-UCRL` *achieves:*

$$\min_{\bar{\pi}\in\bar{\Pi}} J(f,\hat{\pi}_T,\bar{\pi}) \ge \min_{\bar{\pi}\in\bar{\Pi}} J(f,\pi^\star,\bar{\pi}) - \epsilon, \tag{11}$$

where $\hat{\pi}_T$ is the output of `RH-UCRL`*, reported according to Eq. (8), and $\pi^\star$ is the optimal robust policy given in Eq. (3).*

Proof of Corollary 1. We start the proof by recalling some of the previously obtained results. The instantaneous regret $r_t(\pi_t)$ of a policy π_t selected at episode t in Eq. (7a) is given by:

$$r_t(\pi_t) = \min_{\bar{\pi}\in\bar{\Pi}} J(f,\pi^\star,\bar{\pi}) - \min_{\bar{\pi}\in\bar{\Pi}} J(f,\pi_t,\bar{\pi}). \tag{25}$$

From Lemma 4 and Lemma 5, it follows that

$$r_t(\pi_t) \le 4L_r\beta_T^H C^H \sum_{h=0}^{H}\mathbb{E}\left[\sum_{h'=0}^{h-1}\|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t},\pi_t(\mathbf{s}_{h',t}),\bar{\pi}_t(\mathbf{s}_{h',t}))\|_2\right]. \tag{26}$$

We also define

$$\bar{r}(\pi_t) \coloneqq \min_{\bar{\pi}\in\bar{\Pi}} J(f, \pi^{\star}, \bar{\pi}) - \min_{\bar{\pi}\in\bar{\Pi}} J^{(p)}(\pi_t, \bar{\pi}), \tag{27}$$

and note that $r(\pi_t) \leq \bar{r}(\pi_t)$ for every π_t, since $\min_{\bar{\pi}\in\bar{\Pi}} J^{(p)}(\pi_t, \bar{\pi}) \leq \min_{\bar{\pi}\in\bar{\Pi}} J(f, \pi_t, \bar{\pi})$. Another useful observation is that the same bound obtained in Equation (26) also holds in case of $\bar{r}(\pi_t)$, i.e.,

$$r(\pi_t) \leq \bar{r}(\pi_t) \leq 4L_r\beta_T^H C^H \sum_{h=0}^{H} \mathbb{E}\left[\sum_{h'=0}^{h-1} \|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t}, \pi_t(\mathbf{s}_{h',t}), \bar{\pi}_t(\mathbf{s}_{h',t}))\|_2\right]. \tag{28}$$

Recall that the reported policy $\hat{\pi}_T$ from Eq. (8) is chosen among the previously selected episodic policies $\{\pi_1, \dots, \pi_T\}$, such that

$$\hat{\pi}_T = \operatorname*{arg\,min}_{\pi\in\{\pi_1,\dots,\pi_T\}} \bar{r}(\pi). \tag{29}$$

It follows that:

$$\begin{aligned}
r(\hat{\pi}_T) &\leq \bar{r}(\hat{\pi}_T) && \text{(30a)}\\
&\leq \frac{1}{T}\sum_{t=1}^{T} \bar{r}(\pi_t) && \text{(30b)}\\
&\leq \frac{1}{T}\sum_{t=1}^{T} 4L_r\beta_T^H C^H \sum_{h=0}^{H} \mathbb{E}\left[\sum_{h'=0}^{h-1} \|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t}, \pi_t(\mathbf{s}_{h',t}), \bar{\pi}_t(\mathbf{s}_{h',t}))\|_2\right] && \text{(30c)}\\
&\leq \frac{1}{T} 4L_r\beta_T^H C^H H \sum_{t=1}^{T} \mathbb{E}\left[\sum_{h'=0}^{H} \|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t}, \pi_t(\mathbf{s}_{h',t}), \bar{\pi}_t(\mathbf{s}_{h',t}))\|_2\right] && \text{(30d)}\\
&\leq \frac{1}{T} 4L_r\beta_T^H C^H H\sqrt{T} \sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\left(\sum_{h'=0}^{H} \|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t}, \pi_t(\mathbf{s}_{h',t}), \bar{\pi}_t(\mathbf{s}_{h',t}))\|_2\right)^2\right]} && \text{(30e)}\\
&\leq \frac{1}{T} 4L_r\beta_T^H C^H H\sqrt{T} \sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\sum_{h'=0}^{H} \|\boldsymbol{\sigma}_{t-1}(\mathbf{s}_{h',t}, \pi_t(\mathbf{s}_{h',t}), \bar{\pi}_t(\mathbf{s}_{h',t}))\|_2^2\right]} && \text{(30f)}\\
&\leq \frac{4L_r\beta_T^H C^H H^{3/2}\sqrt{T\Gamma_T}}{T} && \text{(30g)}
\end{aligned}$$

where Eq. (30a) follows from Eq. (28), and Eq. (30b) follows from the policy reporting rule in Eq. (29) and by upper bounding minimum with average. Finally, Eq. (30c) is due to Eq. (28), and Eqs. (30d) to (30g) follow the same argument as in the proof of Theorem 1.

To achieve $r(\hat{\pi}_T) \leq \epsilon$ for some given $\epsilon > 0$, we require that

$$\frac{4L_r\beta_T^H C^H H^{3/2}\sqrt{T\Gamma_T}}{T} \leq \epsilon.$$

By simple inversion it follows that we require the following number of episodes T:

$$\frac{T}{\beta_T^{2H}\Gamma_T} \geq \frac{16L_r^2 H^3 C^{2H}}{\epsilon^2}$$

to achieve $r(\hat{\pi}_T) \leq \epsilon$. □

C. Gaussian Process Dynamical Models

In this section, we formalize the setting in which the true dynamics f in Eq. (1) has bounded norm in an RKHS induced by a continuous, symmetric positive definite kernel function $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, with $\mathcal{Z} = \mathcal{S} \times \mathcal{A} \times \bar{\mathcal{A}}$. We denote by $\mathcal{K}$ the corresponding RKHS. Having a norm $\|f\|_{\mathcal{K}} \leq B_f$ for some finite $B_f > 0$ means that the RKHS is well-suited for capturing f (Durand et al., 2018).

Due to the episodic nature of the problem, we follow the batch analysis from Desautels et al. (2014) and generalize it to the MDP setting with multiple outputs. In particular, we observe H transitions per episode and at the beginning of each episode we use the model to make decisions for other H steps. To extend to multiple outputs we build p copies of the dataset such that $\mathcal{D}_{1:t,i} = \{(\mathbf{s}_{t',h}, \mathbf{a}_{t',h}, \bar{\mathbf{a}}_{t',h}), \mathbf{s}_{t',h+1,i}\}_{h=0,t'=1}^{H-1,t}$, each with tH transitions. I.e., the i-th dataset has as covariates the state-action-adversarial action and as target the i-th coordinate of the next-state. We denote the covariates $z_{t,h} \equiv (\mathbf{s}_{t,h}, \mathbf{a}_{t,h}, \bar{\mathbf{a}}_{t,h})$ and the targets as $y_{t,h,i} \equiv \mathbf{s}_{t',h+1,i}$. Finally, we build p models as

$$\mu_t(z,i) = k_t(z)^\top (K_t + \lambda I)^{-1} y_{1:Ht,i}, \tag{31a}$$

$$k_t(z,z',i) = k(z,z') - k_t(z)^\top (K_t + \lambda I)^{-1} k_t(z'), \tag{31b}$$

$$\sigma_t^2(z,i) = k_t(z,z), \tag{31c}$$

where $\mathbf{s}'_{1:Ht,i}$ is the column vector of the i-th coordinate of all the next-states in the dataset, K_t is the kernel matrix, I is the identity matrix of appropriate dimensions and we use $\lambda = pH$ as the same data is used in all the p models.

Stacking together the posterior mean and variance into column vectors we get:

$$\boldsymbol{\mu}_t(z) = \left[\mu_t(z,1), \ldots, \mu_t(z,p)\right]^\top, \tag{31d}$$

$$\boldsymbol{\sigma}_t(z) = \left[\sigma_t^2(z,1), \ldots, \sigma_t^2(z,p)\right]^\top. \tag{31e}$$

A key quantity that we consider in this work is the *(maximum) information gain* that measures the information about the true dynamics f by observing n transitions.

Definition 1 (Information Gain (Cover & Thomas, 1991; Srinivas et al., 2010; Durand et al., 2018))**.** The information gain is the mutual information between the true function f and a set of observations at locations Z and is the difference between the entropy of such observations and the conditional entropy of the observations given function values i.e.,

$$I(f_Z; y_Z) = H(y_Z) - H(y_Z | f_Z), \tag{32a}$$

where f_Z is the noise-free evaluation of f at Z and y_Z is the noisy observation. In the case of GP models as in Eq. (31), the information gain is:

$$I(f_Z; y_Z) = \frac{1}{2} \sum_{k=1}^{n} \ln(1 + \lambda^{-1} \sigma_{k-1}^2(z_k)). \tag{32b}$$

Next, we introduce the maximum information gain, which is a parameter that quantifies how hard the learning problem is and tightly upper bounds the *effective-dimensionality* of the problem (Valko et al., 2013).

Definition 2 (Maximum Information Gain (Srinivas et al., 2010))**.** The maximum information gain is the maximum of the information gain, taken over all datasets with a fixed size n, i.e.,

$$\gamma_n(k; Z) := \max_{Z \subset \mathcal{Z}, |Z| = n} I(f_z; y_z). \tag{33a}$$

In the particular case of GP models, this reduces to:

$$\gamma_n(f; Z) = \max_{\{z_1, \ldots, z_n\} \subset \mathcal{Z}} \frac{1}{2} \sum_{k=1}^{n} \ln(1 + \lambda^{-1} \sigma_{k-1}^2(z_k)). \tag{33b}$$

(Srinivas et al., 2010) show that the Maximum Information Gain (MIG) is sub-linear in the number of observations for commonly used kernels. The main idea now is to bound the complexity measure Γ_T defined in Equation (9) in terms of the MIG and, for commonly used kernels, we achieve no-regret algorithms. Towards this end, we recall two results related to GP-models in Lemma 6.

Lemma 6. *Posterior variance bound (Chowdhury & Gopalan, 2019) Let $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a symmetric positive semi-definite kernel with bounded variance, i.e., $k(z, z) \leq 1, \forall z \in \mathcal{Z}$ and $f \sim GP_{\mathcal{Z}}(0, k)$ be a sample from the associated Gaussian process, then for all $n \geq 1$ and $z \in \mathcal{Z}$:*

$$\sigma^2_{n-1}(z) \leq (1 + \lambda^{-1})\sigma^2_n(z), \tag{34a}$$

$$\sum_{k=1}^{n} \sigma^2_{k-1}(z_k) \leq (1 + 2\lambda) \sum_{k=1}^{n} \frac{1}{2} \ln \left[1 + \lambda^{-1}\sigma_{k-1}(z_k)\right] = (1 + 2\lambda)I(f_Z; y_Z). \tag{34b}$$

Proof. See (Chowdhury & Gopalan, 2019, Lemma 2) □

Although the left-hand-side in Eq. (34b) has the flavor of the complexity measure Γ_T defined in Equation (9) it is not exactly the same as we only update the posterior once every Hp observations. This is related to the batch setting analyzed in Desautels et al. (2014). The next lemma bounds the sum of posterior variances in terms of the information gain.

Lemma 7. *Complexity measure Γ_T is upper bounded by MIG Let $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a symmetric positive semi-definite kernel with bounded variance, i.e., $k(z, z) \leq 1, \forall z \in \mathcal{Z}$ and $f \sim GP_{\mathcal{Z}}(0, k)$ be a sample from the associated Gaussian process, then for all $t \geq 1$ and $z \in \mathcal{Z}$ for the GP-model given in Eq. (31) with $\lambda = Hp$ we have that:*

$$\Gamma_t \leq 2epH\gamma_{pHt}(k, \mathcal{Z}) \tag{35}$$

Proof. The proof is based on Chowdhury & Gopalan (2019, Lemma 11) and adapted to our setting.

$$\begin{aligned}
\sum_{t=1}^{T} \sum_{(\mathbf{s},\mathbf{a},\bar{\mathbf{a}})\in\tilde{\mathcal{D}}_t} \|\boldsymbol{\sigma}_{t-1}(\mathbf{s}, \mathbf{a}, \bar{\mathbf{a}})\|_2^2 &= \sum_{t'=1}^{t} \sum_{h=0}^{H-1} \sum_{i=1}^{p} \sigma^2_{(t'-1)Hp}(z_{t',h,i}) && (36a)\\
&\leq \sum_{t'=1}^{t} \sum_{h=0}^{H-1} \sum_{i=1}^{p} (1 + \lambda^{-1})^{ph+i-1} \sigma^2_{(t'-1)Hp+hp+i}(z_{t',h,i}) && (36b)\\
&\leq (1 + \lambda^{-1})^{p(H-1)+p-1} \sum_{t'=1}^{t} \sum_{h=0}^{H-1} \sum_{i=1}^{p} \sigma^2_{(t'-1)Hp+hp+i}(z_{t',h,i}) && (36c)\\
&\leq (1 + \lambda^{-1})^{pH-1}(2\lambda + 1)I(f_Z; y_Z) && (36d)\\
&\leq 2epHI(f_Z; y_Z) && (36e)
\end{aligned}$$

Here, equality (36a) is the definition of the 2-norm; inequality (36b) is due to Eq. (34a) in Lemma 6; inequality (36c) is due to $1 + \lambda^{-1} \geq 1$; inequality (36d) is due to Eq. (34b) in Lemma 6; finally the last inequality (36e) is due to $(1 + \lambda^{-1})^{\lambda} \leq e$ and $(1 + \lambda^{-1})^{-1}(2\lambda + 1) \leq 2\lambda$. The statement follows by taking the maximum over data sets. □

Next, we will show that GP models are calibrated and satisfy Assumption 2.

Lemma 8. *Concentration of an RKHS member (Durand et al., 2018, Theorem 1) Given Assumption 1, $\|f\|_{\mathcal{K}} \leq B_f$, and $k(\cdot, \cdot) \leq 1$, then, for all $\delta \in [0, 1]$, which probability at least $1 - \delta$, it holds simultaneously over all $z \in Z$ and $t \geq 0$,*

$$|f(z) - \mu_t(z)| \leq \left(B_f + \frac{\sigma}{\lambda}\sqrt{2\ln(1/\delta) + 2\gamma_t}\right)\sigma_t(z), \tag{37}$$

where $\mu_t(z)$ and $\sigma_t(z)$ are given by Eq. (31a) and Eq. (31c).

Thus we know that, using $\beta_t = \left(B_f + \frac{\sigma}{\lambda}\sqrt{2\ln(1/\delta) + 2\gamma_t}\right)$, Assumption 1 holds for a single dimension. The extension to multiple dimensions is straightforward and has been done by (Chowdhury & Gopalan, 2019, Lemma 10) and (Curi et al., 2020a, Lemma 11), using $\lambda \leftarrow Hp$ and $t \leftarrow tHp$.

Putting together results of the previous sections, we know by Lemma 8 that, under Assumption 1 and $\|f\|_{\mathcal{K}}$, GP models satisfy Assumption 2. Furthermore, by Lemma 13 in Appendix G of Curi et al. (2020a), we know that such models also satisfy Assumption 3. The remaining condition is that the results in previous sections assume that the domain is bounded.

However, under Assumption 1 this is not true. Fortunately, Curi et al. (2020a) prove in Appendix I that the domain is bounded with high-probability. Furthermore, they prove that the MIG of kernels only increases poly-logarithmically, which does not affect the regret bounds in this paper.

D. Extended Experimental Results

In this section, we detail the experimental procedures for completeness. We first detail how we learn the dynamical model using an ensemble of neural networks in Appendix D.1 as it is common to all experiments. We then detail the inverted pendulum experiment from Section 1 in Appendix D.2. We describe the adversarial-robust experiment in Appendix D.3, the action-robust experiment in Appendix D.4, and the parameter-robust experiment in Appendix D.5. All experiments where run with 18-core Intel Xeon E5-2697v4 processors.

D.1. Model Learning and Calibration

To learn the dynamics, we use a probabilistic ensemble with five heads as in Chua et al. (2018). The model predicts the change in state, i.e, $\delta_h = \mathbf{s}_{h+1} - \mathbf{s}_h$ and we normalize the states, actions and change in next-states, with the running mean and standard deviation, similar in nature to van Hasselt et al. (2016). After each episode, we split the data into a train and validation set with a 0.9/0.1 ratio. For each ensemble member, we also sample a weight to simulate bootstrapping as in Osband et al. (2016). Finally, each model is trained minimizing the negative log-likelihood of a Gaussian distribution. We train for 20 epochs and early stop if the prediction mean-squared-error when the epoch-loss on a validation set is 10% larger than the minimum epoch-loss in the same validation set. After training, we recalibrate on the validation set using temperature scaling. In particular, we use binary search to find the best parameter in the interval $[0.01, 100]$ that minimizes the expected calibration error (Malik et al., 2019).

D.2. Inverted Pendulum Swing-Up Task

The pendulum swing up task has a reward function given by $r(\mathbf{s}, \mathbf{a}) = -(\theta^2 + 0.1 * \dot{\theta}^2)$, where θ is the angle and $\dot{\theta}$ is the angular velocity. The Pendulum always starts from $\theta = \pi$ in the bottom down position and the goal is to swing the pendulum to the top-up position at $\theta = 0$. Crucially, the initial distribution is a dirac-distribution located at $\theta = \pi$, i.e., it does not have enough coverage for algorithms to explore with it.

Adversarial-Robust. In this setting, the adversary can change the relative gravity and the relative mass of the environment at every episode between $[1 - \alpha, 1 + \alpha]$, for varying α. We train RH-UCRL for 200 episodes, H-UCRL with the nominal gravity and mass for 200 episodes, and the baseline in this setting is RARL, which we train for 1000 episodes. To evaluate the robust performance, we train SAC for 200 episodes, fixing the agent policy of the algorithms.

Action-Robust. In this setting, the action is a mixture sampled with probability α of the learner and the adversary, i.e., the adversary only affects the input torque to the pendulum. The training and evaluation procedure is the same as in the adversarial robust setting. The baseline in this setting is AR-DDPG, which we train for 200 episodes.

Parameter-Robust. In this setting, we consider robustness to mass change. Compared to the adversarial-robust setting, here the adversary is only allowed to change the mass once per episode. In this setting, H-UCRL is trained for 200 episodes with the nominal mass, and then it is evaluated for varying masses. RH-UCRL and the baseline, EP-Opt are allowed to change the mass also during training. In this setting, there is no worst-case adversary during evaluation.

D.3. Adversarial-Robust RL

Next we detail the environments, the training and evaluation procedure, and the hyper-parameters in different paragraphs.

Environments. For the Half-Cheetah environment, the adversary acts on the torso, the front foot and the back foot. For the Hopper environment, the adversary acts on the torso. For the Inverted Pendulum, the adversary acts on the pole. The Inverted Pendulum task is different here as it starts from a perturbation of the top-up position and the task is to stabilize the pendulum. For the Reacher2d environment, the adversary acts on the body0 link. For the Swimmer, the adversary acts on the torso. For the Walker, the adversary acts on the torso. For all environments, we use the adversarial input magnitude $\bar{\mathcal{A}} = [-10, 10]^{\bar{q}}$, where $\bar{q}$ is environment dependent.

Training and Evaluation. We train RH-UCRL, BestResponse, MaxiMin-MB, MaxiMin-MF with in an adversarial environment for 200 episodes. We train RARL and RAP for 1000 episodes in an adversarial environment. Finally, we train H-UCRL for 200 episodes in a standard environment.To evaluate the robust performance of each algorithm, we freeze the output policy of the training step and train an adversary using SAC for 200 episodes. We perform five independent runs and report the mean and standard deviation over the runs.

Algorithm Hyper-Parameters. For RH-UCRL and its variants, we fix $\beta = 1.0$, we train every time step and do two gradient steps with Adam (Kingma & Ba, 2014) with learning rate $= 3 \times 10^{-4}$. To compute a policy gradient, we take pathwise derivatives of a learned critic using the learned model for 3 time steps and weight each estimates using td-λ, with $\lambda = 0.1$ (Sutton & Barto, 2018). We also add entropy regularization with parameter 0.2. We did not do hyper parameter search, but rather use the software default values. For RARL and RAP, we use the PPO algorithm from Schulman et al. (2017) as this performed better than TRPO from Schulman et al. (2015). We train PPO after collecting a batch of 4 episodes, for 80 gradient steps, using early stopping once the KL divergence between the initial and the current policy is more than 0.0075.

D.4. Action-Robust RL

We use the noisy robust setting from Tessler et al. (2019) with mixture parameter $\alpha = 0.3$. The training and evaluation procedures, as well as the hyperparameters, are identical to the adversarial-robust experiment. The baseline is AR-DDPG that Tessler et al. (2019) propose.

D.5. Parameter-Robust

In this setting, we are robust to a relative mass change, i.e., when the relative mass equals one, then the environment has the nominal mass. For all environments except for the Swimmer, the relative mass is bounded in the interval $[0.001, 2]$. For the Swimmer, there were numerical errors due to ill-conditioned mass in the simulator, so we limit the range to $[0.5, 1.5]$.

We train RH-UCRL, BestResponse, MaxiMin-MB, MaxiMin-MF with in an adversary that is allowed to select the worst-case mass from within the range for 200 episodes. We train the baselines, DomainRandomization and EPOpt for 1000 episodes as they were also based on PPO. Finally, we train H-UCRL for 200 episodes in a standard environment. To evaluate the performance of each algorithm, we evaluate the different relative masses in the interval and do five independent runs. We report the mean and standard deviation over the runs.