Towards Tractable Optimism in Model-Based Reinforcement Learning

Aldo Pacchiano[*1] **Philip Ball**[*2] **Jack Parker-Holder**[*2] **Krzysztof Choromanski**[3] **Stephen Roberts**[2]

[1]UC Berkeley
[2]University of Oxford
[3]Google Brain Robotics
[*]Equal Contribution.

Throughout Sections 1 and 2 we make the following assumption:

Assumption 1. *All rewards $r(s,a) \in [0,1]$ and all estimated rewards $\tilde{r}(s,a) \in [0,1]$. A simple clipping mechanism ensures the rewards $\tilde{r}(s,a)$ can be implemented to ensure this holds.*

1 OPTIMISM

1.1 AUXILIARY RESULTS

Lemma 1 (Lemma 1 of Maillard and Asadi [2018])**.** *For all $(s,a) \in \mathcal{S} \times \mathcal{A}$:*

$$\mathbb{P}\Big(\forall t \in \mathbb{N} \quad |r(s,a) - \hat{r}_k(s,a)| \geq \beta_r(N_k(s,a), \delta')\Big) \leq \delta,$$

$$\text{with } \beta_r(n,\delta') := \sqrt{\frac{\log\left(2\sqrt{n+1}/\delta'\right)}{n}}$$

$$\mathbb{P}\Big(\forall t \in \mathbb{N} \quad ||P(s,a) - \hat{P}_k(s,a)||_1 \geq \beta_P(N_k(s,a), \delta')\Big) \leq \delta,$$

$$\text{with } \beta_P(n,\delta') := \sqrt{\frac{4\log(\sqrt{n+1}\frac{2^{|S|}}{\delta'})}{n}}.$$

1.2 GAUSSIAN OPTIMISM

Lemma 2. *Let $(s,a) \in \mathcal{S} \times \mathcal{A}$. If $\tilde{r}_k^{(m)}(s,a) \sim \hat{r}_k(s,a) + \mathcal{N}(0,\sigma^2)$ for $\sigma = 2\beta_r(N_k(s,a), \frac{\delta}{2|S||A|})$ then:*

$$\mathbb{P}(\tilde{r}_k^{(m)}(s,a) \geq r(s,a)|\mathcal{E}) \geq \frac{1}{10}. \tag{1}$$

Proof. Since we are conditioning on $\mathcal{E}$, it follows that $\hat{r}_k(s,a) + \beta_r(N_k(s,a), \frac{\delta}{2|S||A|}) \geq r(s,a)$. Therefore:

$$\mathbb{P}\left(\tilde{r}_k^{(m)}(s,a) \geq \hat{r}_k(s,a) + \beta_r(N_k(s,a), \frac{\delta}{2|S||A|})\right) \geq$$

$$\frac{1}{\sqrt{2\pi}}\left(\frac{\sigma\beta_r(N_k(s,a), \frac{\delta}{2|S||A|})}{\beta_r^2(N_k(s,a), \frac{\delta}{2|S||A|}) + \sigma^2}\right) e^{-\frac{\beta_r^2(N_k(s,a), \frac{\delta}{2|S||A|})}{2\sigma^2}}$$

Setting $\sigma = 2\beta_r(N_k(s,a), \frac{\delta}{2|S||A|})$ yields:

$$\mathbb{P}\left(\tilde{r}_k^{(m)}(s,a) \geq \hat{r}_k(s,a) + \beta_r(N_k(s,a), \frac{\delta}{2|S||A|})\right) \geq \frac{1}{10}.$$

After conditioning on $\mathcal{E}$, the result follows. □

Correspondence to: pacchiano@berkeley.edu, {ball, jackph}@robots.ox.ac.uk

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

Previous results imply that with constant probability the values $\tilde{r}_k^{(m)}(s,a)$ are an overestimate of the true rewards. It is also possible to show that despite this property, $\tilde{r}_k(s,a)$ remain very close to $\hat{r}_k(s,a)$ and therefore to $r(s,a)$. Let $\sigma = 2\beta_r(N_k(s,a), \frac{\delta}{2|S||A|})$. Since $\tilde{r}_k^{(m)}(s,a) - \hat{r}_k(s,a) \sim \mathcal{N}(0,\sigma^2)$, it follows that for all t and all $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$\mathbb{P}\Big(|\tilde{r}_k^{(m)}(s,a) - \hat{r}_k(s,a)| \geq \tag{2}$$
$$2\sqrt{\log\Big(\frac{4|\mathcal{S}||\mathcal{A}|M_r}{\delta}\Big)}\beta_r\left(N_k(s,a), \frac{\delta}{2|S||A|}\right)\Big) \leq \frac{\delta}{|\mathcal{S}||\mathcal{A}|M_r}.$$

The probability of non-optimism decreases as the number of models increases, albeit at a logarithmic rate. Recall that while conditioning on $\mathcal{E}$, the confidence intervals are valid, and therefore $\tilde{r}_k(s,a)$ must also not be too far away from $\hat{r}_k(s,a)$ and therefore from $r(s,a)$. We can summarize the results of this section in the following Corollary:

Corollary 1. *The sampled rewards $\tilde{r}_k(s,a)$ are optimistic:*

$$\mathbb{P}\left(\tilde{r}_k(s,a) = \max_{m=1,\cdots,M_r} \tilde{r}_k^{(m)}(s,a) \geq r(s,a)\Big|\mathcal{E}\right) \geq 1 - \left(\frac{1}{10}\right)^{M_r} \tag{3}$$

while at the same time not being too far from the true rewards:

$$\mathbb{P}\left(|\tilde{r}_k(s,a) - r(s,a)| \geq L\beta_r\left(N_k(s,a), \frac{\delta}{2|S||A|}\right)\Big|\mathcal{E}\right) \leq \frac{\delta}{|\mathcal{S}||\mathcal{A}|}. \tag{4}$$

Where $L = \left(2\sqrt{\log\left(\frac{4|\mathcal{S}||\mathcal{A}|M_r}{\delta}\right)} + 1\right)$.

Corollary 1 shows the trade-offs when increasing the number of models in an ensemble: it increases the amount of optimism, at the expense of greater estimation error of the sample rewards.

1.3 BOOTSRAP OPTIMISM

Although our proofs are based on injecting Gaussian noise into the rewards and dynamics mean estimators, it is possible to make use of any distribution or sampling scheme that guarantees enough optimism in these mean estimators while at the same time ensuring their estimation error converges to zero. Specifically, taking inspiration from Kveton et al. [2019] we propose the following reward augmentation technique. Every time a reward sample $r_k(s,a)$ is observed, we add additional M_B fake reward samples $\{-1, 1\}$ into the buffer of rewards corresponding to state action pair (s,a). For each state action pair, the number of fake rewards up to episode k equals $2M_B N_k(s,a)$, while the number of real reward samples equals $N_k(s,a)$. This constant proportion between fake and real rewards is crucial in achieving optimism. In this case and while executing UCBVI-NARL we sample with replacement from the reward buffer corresponding to all state action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$. It can be seen, in a similar way as in Kveton et al. [2019], this mechanism provides enough optimism for the reward signals. This yields the proof of the necessary boostrap sampling anticoncentration guarantees needed for concluding the proof of the Theorem, that from this point onward follows the same argument as the proof of the guarantees for Gaussian UCBVI. The necessary concentration properties follow also immediately from noting that the buffer samples are subgaussian.

1.4 REWARDS DATA AUGMENTATION.

Dealing with $N_k(s,a) = 0$. When $N_k(s,a) = 0$ we declare $\hat{r}_k(s,a) = 0$ and we let $\xi_k^{(m)}(s,a) \sim \mathcal{N}(0,1)$. In this case, and by Assumption 1, simply by considering the probability of a sample to be larger than 1, we conclude that in this edge case an equivalent version of Corollary 2 holds. This affects only mildly the constants in the definition of M_r.

We state and prove a slightly more general version of Lemma 2:

Lemma 3. *Let $(s,a) \in \mathcal{S} \times \mathcal{A}$. If $\tilde{r}_k^{(m)}(s,a) \sim \hat{r}_k(s,a) + \mathcal{N}(0,\sigma^2)$ for $\sigma = \gamma\beta_r(N_k(s,a), \frac{\delta}{2|S||A|})$ then:*

$$\mathbb{P}(\tilde{r}_k^{(m)}(s,a) \geq \hat{r}(s,a) + \beta_r(N_k(s,a), \frac{\delta}{|\mathcal{S}||\mathcal{A}|})) \geq \frac{1}{\sqrt{2\pi}}\frac{\gamma}{1+\gamma^2}e^{-\frac{1}{2\gamma^2}} \geq \frac{1}{\sqrt{2\pi}}\frac{\gamma}{1+\gamma^2}(1 - \frac{1}{2\gamma^2}) \tag{5}$$

Proof. Since we are conditioning on $\mathcal{E}$, it follows that $\hat{r}_k(s,a) + \beta_r(N_k(s,a), \frac{\delta}{2|S||A|}) \geq r(s,a)$. Hence,

$$\mathbb{P}\left(\tilde{r}_k^{(m)}(s,a) \geq \hat{r}_k(s,a) + \beta_r(N_k(s,a), \frac{\delta}{2|S||A|})\right) \geq \frac{1}{\sqrt{2\pi}}\left(\frac{\sigma\beta_r(N_k(s,a), \frac{\delta}{2|S||A|})}{\beta_r^2(N_k(s,a), \frac{\delta}{2|S||A|}) + \sigma^2}\right) e^{-\frac{\beta_r^2(N_k(s,a), \frac{\delta}{2|S||A|})}{2\sigma^2}}$$

Setting $\sigma = \gamma\beta_r(N_k(s,a), \frac{\delta}{2|S||A|})$ yields:

$$\mathbb{P}\left(\tilde{r}_k^{(m)}(s,a) \geq \hat{r}_k(s,a) + \beta_r(N_k(s,a), \frac{\delta}{2|S||A|})\right) \geq \frac{1}{\sqrt{2\pi}}\frac{\gamma}{1+\gamma^2}e^{-\frac{1}{2\gamma^2}} \geq \frac{1}{\sqrt{2\pi}}\frac{\gamma}{1+\gamma^2}(1 - \frac{1}{2\gamma^2})$$

□

And the following Corollary,

Corollary 2. *Let* $p(\gamma) = \frac{1}{\sqrt{2\pi}}\frac{\gamma}{1+\gamma^2}(1 - \frac{1}{2\gamma^2})$.

$$\mathbb{P}\left(\tilde{r}_k(s,a) = \max_{m=1,\cdots,M_r} \tilde{r}_k^{(m)}(s,a) \geq r(s,a)\right) \geq 1 - (p(\gamma))^{M_r} \tag{6}$$

And conditioned on $\mathcal{E}$ *and for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$*:*

$$\mathbb{P}\left(|\tilde{r}_k(s,a) - r(s,a)| \geq \left(\sqrt{\log\left(\frac{4|\mathcal{S}||\mathcal{A}|M_r}{\delta}\right)}\gamma + 1\right)\beta_r\left(N_k(s,a), \frac{\delta}{2|S||A|}\right)\right) \leq \frac{\delta}{|\mathcal{S}||\mathcal{A}|} \tag{7}$$

1.5 DYNAMICS DATA AUGMENTATION.

We can also show that for appropriate noise processes $\{\boldsymbol{\xi}_k^{(m)}(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, we can obtain an appropriate balance between optimism and estimation error, as we did for the rewards. In the case of a Gaussian noise process $\{\boldsymbol{\xi}_k^{(m)}(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, we can use anti concentration to show the following result.

Lemma 4. *Assume* $|\mathcal{S}| \geq 2$ *and let* $(s,a) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$*. If* $\tilde{P}_k^{(m)}(s,a,s') = \hat{P}_k(s,a,s') + \mathcal{N}(0,\sigma^2)$ *for all* s' *then and for* $\sigma = 2\beta_P\left(N_k(s,a), \frac{\delta}{|\mathcal{S}||\mathcal{A}|}\right)$*, then for any fixed vector* $\mathbf{v} \in \mathbb{R}^{\mathcal{S}}$ *:*

$$\mathbb{P}\left(\mathbb{E}_{s'\sim\tilde{P}_k^{(m)}(s,a,s')}\left[\mathbf{v}[s']\right] \geq \mathbb{E}_{s'\sim P(s,a,s')}\left[\mathbf{v}[s']\right]|\mathcal{E}\right) \geq \frac{1}{9|\mathcal{S}|}.$$

The proof is in Appendix 1.5.1.

Since $\tilde{P}_k^{(m)}(s,a) - \hat{P}_k(s,a) \sim \mathcal{N}(0, I\sigma^2)$, it follows that for all k and all $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$\mathbb{P}\left(\|\tilde{P}_k^{(m)}(s,a) - \hat{P}_k(s,a)\|_1 \geq 2\sqrt{|\mathcal{S}|\log\left(\frac{4|\mathcal{S}||\mathcal{A}|M_P}{\delta}\right)}\beta_P\left(N_k(s,a), \frac{\delta}{2|S||A|}\right)\right) \leq \frac{\delta}{|\mathcal{S}||\mathcal{A}|M_P}. \tag{8}$$

The extra $\sqrt{|\mathcal{S}|}$ not present in 2, comes from bounding the l_1 norm of an $|\mathcal{S}|$-dimensional Gaussian vector and it is akin to the extra $\sqrt{d}$ in the regret of linear Thompson sampling Abeille et al. [2017]. Let $\mathbf{v} = \mathbf{V}^{h+1}(\pi^*) \in \mathbb{R}^{|\mathcal{S}|}$, the value vector of π^* at $h+1$. Lemma 4 and Equation 8 imply:

Corollary 3. *The sampled dynamics* $\tilde{P}_k^{(m)}(s,a)$ *are optimistic:*

$$\mathbb{P}\left(\max_{m=1,\cdots,M_P} \mathbb{E}_{s'\sim\tilde{P}_k^{(m)}(s,a,s')}\left[\mathbf{V}^{h+1}(\pi^*)(s')\right] \geq \mathbb{E}_{s'\sim P(s,a,s')}\left[\mathbf{V}^{h+1}(\pi^*)(s')\right]\right) \geq 1 - \left(\frac{1}{9|\mathcal{S}|}\right)^{M_P}$$

while at the same time not being too far from the true dynamics:

$$\mathbb{P}\left(\|\tilde{P}_k(s,a) - P(s,a)\|_1 \geq \left(2\sqrt{|\mathcal{S}|\log\left(\frac{4|\mathcal{S}||\mathcal{A}|M_P}{\delta}\right)} + 1\right)\beta_P\left(N_k(s,a), \frac{\delta}{2|S||A|}\right)\Big|\mathcal{E}\right) \leq \frac{\delta}{|\mathcal{S}||\mathcal{A}|}. \tag{9}$$

Now we can proceed to show that as long as M_P is chosen appropriately, we can ensure optimism holds with enough probability:

Theorem 1 (Optimism)**.** *If* $M_r \geq \frac{\log\left(\frac{2|\mathcal{S}||\mathcal{A}|H}{\delta}\right)}{3}$ *and* $M_P \geq 3 + \frac{\log\left(\frac{2|\mathcal{A}|H}{\delta}\right)}{3}$*. Then with probability at least* $1 - 2\delta$*:*

$$\tilde{V}_k(\pi_k) \geq V(\pi^*)$$

and therefore $\mathrm{I} \leq 0$ *with probability at least* $1 - 2\delta K$*.*

The proof makes use of an inductive argument and can be found in Appendix 1.6.

Dealing with $N_k(s,a) = 0$**.** When $N_k(s,a) = 0$ we declare $\hat{P}_k(s,a) = \frac{1}{|\mathcal{S}|}\mathbf{1}$ and we let $\boldsymbol{\xi}_k^{(m)}(s,a) \sim \mathcal{N}(0, \mathbb{I}_{|\mathcal{S}|})$. We can also make use of the alternative, $\hat{P}_k(s,a) = \mathbf{0}$ if we allow for $\hat{P}_k(s,a)$ to be a signed measure that is not a probability measure. Using these definitions we can easily derive a version of Corollary 3. Taking this into account only adds a simple adjustment of the constants for the definition of M_P.

1.5.1 Proof of Lemma 4

Proof. Since conditioned on $\mathcal{E}$, $\|P(s,a) - \hat{P}_k(s,a)\|_1 \leq \beta_P(N_k(s,a), \frac{\delta}{|\mathcal{S}||\mathcal{A}|})$ and therefore:

$$\begin{aligned}
|\langle v, \hat{P}_k(s,a)\rangle - \langle v, P(s,a)\rangle| &\leq \|P(s,a) - \hat{P}_k(s,a)\|_1 \|v\|_\infty \\
&\leq \beta_P\left(N_k(s,a), \frac{\delta}{|\mathcal{S}||\mathcal{A}|}\right)\|v\|_\infty
\end{aligned}$$

Let $\tilde{P}_k^{(m)}(s,a,s') - \hat{P}_k(s,a,s') = \xi \sim \mathcal{N}(0,\sigma^2)$. Notice that $\langle \tilde{P}_k^{(m)}(s,a,s'), v\rangle = \langle \tilde{P}_k^{(m)}(s,a,s'), v\rangle + \langle v, \xi\rangle$. And therefore $\langle v, \xi\rangle \sim \mathcal{N}(0, \|v\|^2\sigma^2)$. Hence:

$$\begin{aligned}
\mathbb{P}\left(\langle v,\xi\rangle \geq \beta_P\left(N_k(s,a), \frac{\delta}{|\mathcal{S}||\mathcal{A}|}\right)\|v\|_\infty\right) &\geq \frac{1}{\sqrt{2\pi}}\frac{\gamma\|v\|_2\|v\|_\infty}{\gamma^2\|v\|_2^2 + \|v\|_\infty^2}\exp\left(-\frac{\|v\|_\infty^2}{2\gamma^2\|v\|_2^2}\right) \\
&\overset{(i)}{\geq} \frac{1}{\sqrt{2\pi}}\frac{\gamma\|v\|_2\|v\|_\infty}{\gamma^2\|v\|_2^2 + \|v\|_\infty^2}\exp\left(-\frac{1}{2\gamma^2}\right) \\
&\overset{(ii)}{\geq} \frac{1}{\sqrt{2\pi}}\frac{\gamma}{S\gamma^2+1}(1 - \frac{1}{2\gamma^2})
\end{aligned}$$

Inequality (i) holds because $\|v\|_\infty \leq \|v\|_2$. Inequality (ii) holds as a consequence of:

$$\gamma\|v\|_2\|v\|_\infty \geq \gamma\|v\|_\infty^2 = \frac{\gamma}{S\gamma^2+1}\left(S\gamma^2\|v\|_\infty^2 + \|v\|_\infty^2\right) \geq \frac{\gamma}{S\gamma^2+1}\left(\gamma^2\|v\|_2^2 + \|v\|_\infty^2\right).$$

□

1.6 PROOF OF OPTIMISM THEOREM

In this section we prove Theorem 1.

Proof. We proceed by induction. Notice that for all $s \in \mathcal{S}$:

$$\mathbf{V}^{H-1}(\pi^*)[s] = \max_{a\in\mathcal{A}} r(s,a)$$

By Corollary 2, if $M_r \geq \frac{\log\left(\frac{2|\mathcal{S}||\mathcal{A}|H}{\delta}\right)}{3}$, for each $(s,a) \in \mathcal{S}\times\mathcal{A}$ and with probability at least $1 - \frac{\delta}{2|\mathcal{S}||\mathcal{A}|H}$:

$$\tilde{r}_k(s,a) \geq r(s,a). \tag{10}$$

Therefore, for any s, with probability at least $1 - \frac{\delta}{2|\mathcal{S}|H}$:

$$\tilde{\mathbf{V}}_k^{H-1}(\pi_k)[s] = \max_{a\in\mathcal{A}} \tilde{r}_k(s,a) \geq \max_{a\in\mathcal{A}} r(s,a) = \mathbf{V}^{H-1}(\pi^*)[s]$$

And therefore it also holds with probability at least $1 - \frac{\delta}{2H}$ and for all $s \in \mathcal{S}$ simultaneously.

We proceed by induction. Let's assume that for some $h+1 \leq H-1$ and for all s and with probability at least $1-\delta(h+1)$ simultaneously for all $s \in \mathcal{S}$ it holds that $\tilde{\mathbf{V}}_k^{h+1}(\pi_k)[s] \geq \mathbf{V}^{h+1}(\pi^*)[s]$ for some value $\delta(h+1)$ dependent on h. Recall that by NAVI:

$$\tilde{\mathbf{V}}_k^{h}(\pi_k)[s] = \max_{a\in\mathcal{A}} \left(\tilde{r}_k(s,a) + \mathbb{E}_{s'\sim\tilde{P}_k(s,a)}\left[\tilde{\mathbf{V}}_k^{h+1}(\pi_k)(s')\right]\right)$$

It follows that with probability at least $1-\delta(h+1)$ and simultaneously for all $s \in \mathcal{S}$:

$$\tilde{\mathbf{V}}_k^{h}(\pi_k)[s] \geq \max_{a\in\mathcal{A}} \left(\tilde{r}_k(s,a) + \mathbb{E}_{s'\sim\tilde{P}_k(s,a)}\left[\mathbf{V}^{h+1}(\pi^*)(s')\right]\right)$$

Call this event $\mathcal{U}(h+1)$. Notice that for all h', the value vector $\mathbf{V}^{h'}(\pi^*)$ is independent of k. If $M_P \geq \frac{\log\left(\frac{2|\mathcal{S}||\mathcal{A}|H}{\delta}\right)}{\log(9|\mathcal{S}|)}$, by Corollary 2 for each $(s,a) \in \mathcal{S}\times\mathcal{A}$ and with probability at least $1 - \frac{\delta}{|\mathcal{S}||\mathcal{A}|H}$:

$$\tilde{r}(s,a) \geq r(s,a) \qquad \text{and} \qquad \mathbb{E}_{s'\sim\tilde{P}_k(s,a)}\left[\mathbf{V}^{h+1}(\pi^*)(s')\right] \geq \mathbb{E}_{s'\sim P(s,a)}\left[\mathbf{V}^{h+1}(\pi^*)(s')\right]$$

Therefore, a union bound implies that with probability at least $1 - \frac{\delta}{H} - \mathbb{P}(\mathcal{U}^c(h+1)) = 1 - \frac{\delta}{H} - \delta(h+1)$ and simultaneously for all $s \in \mathcal{S}$:

$$\tilde{\mathbf{V}}_k^{h}(\pi_k)[s] \geq \max_{a\in\mathcal{A}} \left(r(s,a) + \mathbb{E}_{s'\sim P(s,a)}\left[\mathbf{V}^{h+1}(\pi^*)(s')\right]\right) = \mathbf{V}^{h}(\pi^*)[s]$$

The RHS equality holds by optimality of π^*. This completes the induction step. Notice that we can set $\delta(H-1) = \frac{\delta}{2H}$. And that for all other h we can define $\delta(h) = \delta(h) + \frac{\delta}{H}$. Unrolling the induction until $h = 0$ we can conclude that with probability at least $1-\delta$ and for all $s \in \mathcal{S}$:

$$\tilde{\mathbf{V}}_k^{0}(\pi_k)[s] \geq \mathbf{V}^{0}(\pi^*)[s]$$

Taking expectations w.r.t. P_0 concludes the proof by noting $\tilde{V}_k(\pi_k) = \mathbb{E}_{s\sim P_0}[\tilde{\mathbf{V}}_k^0(\pi_k)[s]]$ and $V(\pi^*) = \mathbb{E}_{s\sim P_0}[\mathbf{V}^0(\pi^*)[s]]$.

The second part of the statement follows by a simple union bound.

□

2 ESTIMATION ERROR

The goal of this section is to bound term II of the regret decomposition. Let's define an intermediate MDP $\hat{\mathcal{M}}_k$ corresponding to the MDP having $\mathcal{M}$'s true dynamics and using the rewards $\{\tilde{r}(s,a)\}_{s,a\in\mathcal{S}\times\mathcal{A}}$. Similarly let's define an approximate value function $\hat{V}_k(\pi)$ which corresponds to the expected reward of π on MDP $\hat{\mathcal{M}}_k$. Throughout this section we use the convention that whenever $N_k(s,a) = 0$, we instead use the value 1 in the definition of the relevant confidence intervals.

Term II can be written as:

$$\begin{aligned}
\text{II} &= \sum_{k=1}^{K} \tilde{V}_k(\pi_k) - V(\pi_k) \\
&= \underbrace{\sum_{k=1}^{K} \tilde{V}_k(\pi_k) - \hat{V}_k(\pi_k)}_{A} + \underbrace{\sum_{k=1}^{K} \hat{V}_k(\pi_k) - V(\pi_k)}_{B}
\end{aligned}$$

Throughout this section (bounds of terms A and B) we condition on the event $\mathcal{E}$ defined as a result of Lemma 1. Recall $\mathbb{P}(\mathcal{E}) \geq 1-\delta$.

2.1 BOUNDING TERM B

Observe that the dynamics in $\mathcal{M}$ and $\hat{\mathcal{M}}_k$ are the same. The bound proceeds in two steps.

First notice that by definition:

$$\hat{V}_k(\pi_k) = \mathbb{E}_{\pi_k}\left[\sum_{h=0}^{H-1} \tilde{r}_k(s_h, a_h)\right]$$

And therefore:

$$\hat{V}_k(\pi_k) - V(\pi_k) = \mathbb{E}_{\pi_k}\left[\tilde{r}_k(s_h, a_h) - r(s_h, a_h)\right]$$

Let $\{s_h^{(k)}, a_h^{(k)}\}_{h=1}^{H}$ be the (random) states and actions our algorithm executes at time t. Let $\mathcal{F}_{t-1}$ be the filtration corresponding to all the randomness in our process up to the beginning of episode t (before the policy is executed). It follows that:

$$\hat{V}_k(\pi_k) - V(\pi_k) = \mathbb{E}\left[\sum_{h=1}^{H-1} \tilde{r}_k(s_h^{(k)}, a_h^{(k)}) - r(s_h^{(k)}, a_h^{(k)}) | \mathcal{F}_{t-1}\right]$$

Let $X_k = \hat{V}_k(\pi_k) - V(\pi_k) - \left(\sum_{h=1}^{H} \tilde{r}_k(s_h^{(k)}, a_h^{(k)}) - r(s_h^{(k)}, a_h^{(k)})\right)$. Let $Y_k = \sum_{\ell=1}^{t} X_k$ for all $t \geq 1$ and $Y_0 = 0$ It is easy to see Y_k is a martingale satisfying a bonded differences assumption:

$$|X_k| \leq 4H$$

Which holds since r and $\tilde{r}$ are both bounded by 1. A simple application of the Azuma-Hoeffding[1] inequality for Martingales yields, for any $\delta \in (0, 1)$ with probability at least $1 - \delta$:

$$Y_K \leq H\sqrt{2K \log\left(\frac{1}{\delta}\right)}$$

This bound implies that with probability at least $1 - \delta$:

$$B \leq H\sqrt{2K \log\left(\frac{1}{\delta}\right)} + \sum_{k=1}^{K}\left(\sum_{h=0}^{H-1} \tilde{r}_k(s_h^{(k)}, a_h^{(k)}) - r(s_h^{(k)}, a_h^{(k)})\right)$$

Let's condition event $\mathcal{E}$. In this case:

$$\tilde{r}_k(s_h^{(k)}, a_h^{(k)}) - r(s_h^{(k)}, a_h^{(k)}) \leq \begin{cases} 1 & \text{if } N_k(s_h^{(k)}, a_h^{(k)}) = 0 \\ \sqrt{\frac{2\log\left(\frac{2\sqrt{N_k(s_h^{(k)}, a_h^{(k)})+1}}{\delta}\right)}{N_k(s_h^{(k)}, a_h^{(k)})}} & \text{o.w.} \end{cases}$$

As a consequence of this:

[1] We use the following version of Azuma-Hoeffding: if $Y_k, t \geq 1$ is a martingale such that $|Y_k - Y_{k-1}| \leq d_k$ for all t then for every $T \geq 1$ we have $\mathbb{P}(Y_k \geq r) \leq \exp\left(-\frac{r^2}{2\sum_{k=1}^{K} d_k^2}\right)$

$$\sum_{k=1}^{K}\left(\sum_{h=0}^{H-1}\tilde{r}_k(s_h^{(k)},a_h^{(k)})-r(s_h^{(k)},a_h^{(k)})\right)\leq\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{\ell=1}^{N_K(s,a)}\sqrt{\frac{2\log\left(\frac{2\sqrt{\ell+1}}{\delta}\right)}{\ell}}+|\mathcal{S}||\mathcal{A}|$$

$$\leq\sqrt{2\log\left(\frac{2\sqrt{KH}}{\delta}\right)}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}2\sqrt{N_K(s,a)}+|\mathcal{S}||\mathcal{A}|$$

The last inequality holds because for all (s,a), it follows that $2\sqrt{N_k(s_h^{(k)},a_h^{(k)})+1}\leq 2\sqrt{KH}$

Since $\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}N_K(s,a)=KH$, and $\sqrt{\cdot}$ is a concave function:

$$\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{N_K(s,a)}\leq\sqrt{|\mathcal{S}||\mathcal{A}|KH}$$

Assembling these pieces together we can conclude that:

$$B\leq H\sqrt{2K\log\left(\frac{1}{\delta}\right)}+2\sqrt{2\log\left(\frac{2\sqrt{KH}}{\delta}\right)\mathcal{S}\mathcal{A}KH}+|\mathcal{S}||\mathcal{A}|$$

2.2 BOUNDING TERM A

This term is more challenging since although the rewards are the same, the dynamics are different. We introduce an extra bit of notation:

Let $\tilde{\mathbf{V}}_k^h(\pi_k)\in\mathbb{R}^{|\mathcal{S}|}$ be the value function vector when running π_k in $\mathcal{M}_k$ from step h. It follows that:

$$\tilde{V}_k(\pi_k)=\langle P_0,\tilde{\mathbf{V}}_k^0(\pi_k)\rangle$$

Where $\tilde{\mathbf{V}}_k^h(\pi_k)[s]$ denotes the $s-$th entry of $\tilde{\mathbf{V}}_k^h(\pi_k)$ and P_0 is the initial state distribution.

Let's also define a family of state action value function vector as $\tilde{\mathbf{Q}}_k^h(\pi_k),\hat{\mathbf{Q}}_k^h(\pi_k)\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ where $\tilde{\mathbf{Q}}_k^h(\pi_k)$ is the state value function of π_k in $\mathcal{M}_k$ from step h. Similarly $\hat{\mathbf{Q}}_k^h(\pi_k)$ is the state value function of π_k in $\hat{\mathcal{M}}_k$ from step h.

We denote the $(s,a)-$th entry of $\tilde{\mathbf{Q}}_k^h(\pi_k)$ as $\tilde{\mathbf{Q}}_k^h(\pi_k)[s,a]$.

We condition on the following event $\mathcal{U}_k$ given by Corollary 3:

$$\mathcal{U}_k:=\{\|\tilde{P}_k(s,a)-P(s,a)\|_1\leq\left(2\sqrt{|\mathcal{S}|\log\left(\frac{4|\mathcal{S}||\mathcal{A}|K}{\delta}\right)}+1\right)\beta_P\left(N_k(s,a),\frac{\delta}{2|\mathcal{S}||A|}\right)\forall(s,a)\in\mathcal{S}\times\mathcal{A}\}$$

We define the following convenient notation for the confidence intervals:

$$\Delta_{s,a,k}=\begin{cases}1 & \text{if } N_k(s,a)=0\\ \min(\left(2\sqrt{|\mathcal{S}|\log\left(\frac{4|\mathcal{S}||\mathcal{A}|K}{\delta}\right)}+1\right)\beta_P\left(N_k(s,a),\frac{\delta}{2|\mathcal{S}||A|}\right),1) & \text{o.w.}\end{cases}$$

Notice that $\mathbb{P}(\mathcal{U}_k)\geq 1-\delta$.

Notice $\Delta_{s,a,k}\leq\min(\frac{C}{N_k(s,a)},1)$ for $C=\tilde{\mathcal{O}}(|\mathcal{S}|)$. Where $\tilde{\mathcal{O}}$ hiddes logarithmic factors in $|\mathcal{S}|,|\mathcal{A}|,\delta$ and T and independent of t.

We start by showing the following:

Lemma 5. *If* $\tilde{\mathbf{V}}_k^{h+1}(\pi_k) - \hat{\mathbf{V}}_k^{h+1}(\pi_k) = \delta_t^{h+1}(\pi_k) \in \mathbb{R}^{|\mathcal{S}|}$ *then for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$*:*

$$\tilde{\mathbf{Q}}_k^h(\pi_k)[s,a] - \hat{\mathbf{Q}}_k^h(\pi_k)[s,a] \leq \mathbb{E}_{s_{h+1} \sim P(s,a)}[\delta_k^{h+1}(\pi_k)[s_{h+1}]|(s,a)] + \min(\frac{C}{\sqrt{N_k(s,a)}}, 1)(H-h)$$

Proof. Notice that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$\tilde{\mathbf{Q}}_k^h(\pi_k)[s,a] = \tilde{r}(s,a) + \langle \tilde{P}_k(s,a), \tilde{\mathbf{V}}_k^{h+1}(\pi_k) \rangle \tag{11}$$

$$\hat{\mathbf{Q}}_k^h(\pi_k)[s,a] = \tilde{r}(s,a) + \langle P(s,a), \hat{\mathbf{V}}_k^{h+1}(\pi_k) \rangle \tag{12}$$

Therefore:

$$\begin{aligned}
\tilde{\mathbf{Q}}_k^h(\pi_k)[s,a] - \hat{\mathbf{Q}}_k^h(\pi_k)[s,a] &= \langle \tilde{P}_k(s,a), \tilde{\mathbf{V}}_k^{h+1}(\pi_k) \rangle - \langle P(s,a), \hat{\mathbf{V}}_k^{h+1}(\pi_k) \rangle \\
&= \langle \tilde{P}_k(s,a), \tilde{\mathbf{V}}_k^{h+1}(\pi_k) \rangle - \langle P(s,a), \tilde{\mathbf{V}}_k^{h+1}(\pi_k) \rangle \\
&\quad + \langle P(s,a), \tilde{\mathbf{V}}_k^{h+1}(\pi_k) \rangle - \langle P(s,a), \hat{\mathbf{V}}_k^{h+1}(\pi_k) \rangle \\
&= \langle \tilde{P}_k(s,a) - P(s,a), \tilde{\mathbf{V}}_k^{h+1}(\pi_k) \rangle + \langle P(s,a), \tilde{\mathbf{V}}_k^{h+1}(\pi_k) - \hat{\mathbf{V}}_k^{h+1}(\pi_k) \rangle \\
&\leq \|\tilde{P}_k(s,a) - P(s,a)\|_1 \|\tilde{\mathbf{V}}_k^{h+1}(\pi_k)\|_\infty + \mathbb{E}_{s_{h+1} \sim P(s,a)}[\delta_k^{h+1}(\pi_k)[s_{h+1}]|(s,a)]
\end{aligned}$$

The last inequality follows by conditioning on the event the concentration bounds hold, since in this case $\|\tilde{P}_k(s,a) - P(s,a)\|_1 \leq \min(\frac{C}{\sqrt{N_k(a,s)}}, 1)$ and $\|\hat{\mathbf{V}}_k^{h+1}(\pi_k)\|_\infty \leq H - h$. The result follows. □

Notice that $\tilde{\mathbf{V}}_k^h(\pi_k)[s] = \tilde{\mathbf{Q}}_k^h(\pi_k)[s, \pi_k(s)]$. This together with Lemma 5 yields:

$$\begin{aligned}
\delta_k^h(\pi_k)[s] &= \tilde{\mathbf{V}}_k^h(\pi_k)[s] - \hat{\mathbf{V}}_k^h(\pi_k)[s] \\
&\leq \mathbb{E}_{s_{h+1} \sim P(s,\pi_k(s))}[\delta_k^{h+1}(\pi_k)[s_{h+1}]|(s,\pi_k(s))] + \min(\frac{C}{\sqrt{N_k(s,\pi_k(s))}}, 1)(H-h) \quad \forall s \in \mathcal{S}.
\end{aligned} \tag{13}$$

We further define:

$$\tilde{V}_k^h(\pi_k) = \mathbb{E}_{\pi_k}[\tilde{\mathbf{V}}_k^h(\pi_k)[s_h^{(k)}]] \tag{14}$$

$$\hat{V}_k^h(\pi_k) = \mathbb{E}_{\pi_k}[\hat{\mathbf{V}}_k^h(\pi_k)[s_h^{(k)}]] \tag{15}$$

Where the expectation is taken over the distribution of $s_h^{(k)}$ as encountered by policy π_k when ran on $\mathcal{M}$. Combining the inequality in 13 and equations 15 and 14 we obtain the following inequality:

$$\tilde{V}_k^h(\pi_k) - \hat{V}_k^h(\pi_k) \leq \tilde{V}_k^{h+1}(\pi_k) - \hat{V}_k^{h+1}(\pi_k) + \mathbb{E}_{\pi_k}\left[\min(\frac{C}{\sqrt{N_k(s_h^{(k)}, \pi_k(s_h^{(k)}))}}, 1)(H-h)\right].$$

Applying this formula recursively from $h = H-1$ down to $h = 0$ yields:

$$
\begin{aligned}
A &= \tilde{V}_k(\pi_k) - \hat{V}_k(\pi_k) \\
&= \tilde{V}_k^0(\pi_k) - \hat{V}_k^0(\pi_k) \\
&\leq \sum_{h=0}^{H-1} \mathbb{E}_{\pi_k}\left[\min\left(\frac{C}{\sqrt{N_k(s_h^{(k)}, \pi_k(s_h^{(k)}))}}, 1\right)(H-h) \right] \\
&= \underbrace{\mathbb{E}_{\pi_k}\left[\sum_{h=0}^{H-1} \min\left(\frac{C}{\sqrt{N_k(s_h^{(k)}, \pi_k(s_h^{(k)}))}}, 1\right)(H-h) \right]}_{\spadesuit}.
\end{aligned} \tag{16}
$$

We proceed to bound term ♠. Let:

$$
Z_k = \mathbb{E}_{\pi_k}\left[\sum_{h=0}^{H-1} \min\left(\frac{C}{\sqrt{N_k(s_h^{(k)}, \pi_k(s_h^{(k)}))}}, 1\right)(H-h) \right] - \sum_{h=0}^{H-1} \min\left(\frac{C}{\sqrt{N_k(s_h^{(k)}, \pi_k(s_h^{(k)}))}}, 1\right)(H-h).
$$

Let $W_k = \sum_{\ell=1}^{k} Z_\ell$. In order to get rid of the uncommon terms that achieve high values in Z_k, we define X_k as:

$$
X_k = \mathbb{E}_{\pi_k}\left[\sum_{h=0}^{H-1} \min\left(\Delta_{s_h^{(k)}, \pi_k(s_h^{(k)}), k}, \frac{1}{H}\right)(H-h) \right] - \sum_{h=0}^{H-1} \min\left(\Delta_{s_h^{(k)}, \pi_k(s_h^{(k)}), k}, \frac{1}{H}\right)(H-h).
$$

And define $Y_k = \sum_{\ell=1}^{k} X_\ell$ for all $k \geq 1$ with $Y_0 = 0$. Notice that:

$$
W_k - Y_k \leq \mathcal{O}(|\mathcal{S}|^2 |\mathcal{A}| H) \tag{17}
$$

This bound follows from the observation that whenever a pair $(s_h^{(k)}, \pi_k(s_h^{(k)})$ is encountered during the execution of policy π_k, its counter $N_k(s_h^{(k)}, \pi_k(s_h^{(k)})$ is incremented, thus inside the expectation $\mathbb{E}_{\pi_1, \ldots, \pi_K}$ the occurrence of $N_k(s_h^{(k)}, \pi_k(s_h^{(k)})) = j$ can be thought of as happening only for one k. Therefore, at most for each state action pair we have a difference of $\sum_{\ell=1}^{|\mathcal{S}|^2 H^2} \frac{1}{\ell} \approx |\mathcal{S}| H$.

It is easy to see that Y_k is a martingale satisfying the following bounded differences condition:

$$
|X_k| \leq 2H
$$

A simple use of Azuma-Hoeffding yields that for any $\delta \in (0, 1)$ and with probability at least $1 - \delta$:

$$
Y_K \leq \frac{H}{2} \sqrt{2K \log\left(\frac{1}{\delta}\right)}
$$

Consequently:

$$
\begin{aligned}
W_K = \sum_{k=1}^{K} \left(\mathbb{E}_{\pi_k}\left[\sum_{h=0}^{H-1} \min\left(\frac{C}{\sqrt{N_k(s_h^{(k)}, \pi_k(s_h^{(k)}))}}, 1\right)(H-h) \right] - \sum_{h=0}^{H-1} \min\left(\frac{C}{\sqrt{N_k(s_h^{(k)}, \pi_k(s_h^{(k)}))}}, 1\right)(H-h) \right) \\
\leq \frac{H}{2} \sqrt{2K \log\left(\frac{1}{\delta}\right)} + \mathcal{O}(|\mathcal{S}|^2 |\mathcal{A}| H).
\end{aligned}
$$

And therefore with probability at least $1 - \delta$:

$$\sum_{k=1}^{T} \mathbb{E}_{\pi_k}\left[\sum_{h=0}^{H-1} \min\left(\frac{C}{\sqrt{N_k(s_h^{(k)}, \pi_k(s_h^{(k)}))}}, 1\right)(H-h)\right] \leq \sum_{k=1}^{K}\sum_{h=0}^{H-1} \min\left(\frac{C}{\sqrt{N_k(s_h^{(k)}, \pi_k(s_h^{(k)}))}}, 1\right)(H-h)$$
$$+ \frac{H}{2}\sqrt{2K\log\left(\frac{1}{\delta}\right)} + \mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|H)$$
$$\leq H\sum_{k=1}^{K}\sum_{h=0}^{H-1} \min\left(\frac{C}{\sqrt{N_k(s_h^{(k)}, \pi_k(s_h^{(k)}))}}, 1\right)$$
$$+ \frac{H}{2}\sqrt{2K\log\left(\frac{1}{\delta}\right)} + \mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|H)$$
$$\leq CH\sum_{s,a\in\mathcal{S}\times\mathcal{A}}\sum_{\ell=1}^{N_K(s,a)}\frac{2}{\sqrt{\ell}} + \frac{H}{2}\sqrt{2K\log\left(\frac{1}{\delta}\right)}$$
$$+ \mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|H)$$
$$\leq CH\sqrt{|\mathcal{S}||\mathcal{A}|KH} + \frac{H}{2}\sqrt{2K\log\left(\frac{1}{\delta}\right)}$$
$$+ \mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|H). \tag{18}$$

Combining inequalities 16 and 18 with all these results yields the following bound for term A:

$$A = \sum_{k=1}^{K} \tilde{V}_k(\pi_k) - \hat{V}_k(\pi_k)$$
$$\leq CH\sqrt{|\mathcal{S}||\mathcal{A}|KH} + \frac{H}{2}\sqrt{2K\log\left(\frac{1}{\delta}\right)} + \mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|H)$$
$$= \tilde{\mathcal{O}}(|\mathcal{S}|H\sqrt{|\mathcal{S}||\mathcal{A}|HK} + H\sqrt{K} + |\mathcal{S}|^2|\mathcal{A}|H)$$

2.3 ESTIMATION ERROR MAIN

The general idea behind bounding the error in term II of the regret decomposition, is similar to what UCRL does. See the details in Appendix 2. We show that with probability at least $1 - 2\delta K$ term II admits the following bound:

$$\text{II} \leq \tilde{\mathcal{O}}(|\mathcal{S}|H\sqrt{|\mathcal{S}||\mathcal{A}|T})$$

where $\tilde{O}$ hides logarithmic factors in $|\mathcal{A}|, |\mathcal{S}|, \delta$ and K. This concludes the proof of the main Gaussian augmented UCBVI Theorem.

2.4 UCRL NARL MAIN THEOREM

Putting our bounds together for term A and B yields:

$$R(T) \leq \tilde{\mathcal{O}}(|\mathcal{S}|H\sqrt{|\mathcal{S}||\mathcal{A}|HK} + |\mathcal{S}|^2|\mathcal{A}|H)$$
$$= \tilde{\mathcal{O}}(|\mathcal{S}|H\sqrt{|\mathcal{S}||\mathcal{A}|T} + |\mathcal{S}|^2|\mathcal{A}|H)$$

2.5 UCBVI NARL MAIN THEOREM

Recall that Noise Augmented Value Iteration (NAVI) proceeds as follows: at the beginning of episode k we compute a $Q-$function $\tilde{\mathbf{Q}}_k$ as:

$$\tilde{\mathbf{Q}}_{k,h}(s,a) = \min\left(\tilde{\mathbf{Q}}_{k-1,h}(s,a), H, \tilde{r}_k(s,a) + \quad \mathbb{E}_{s'\sim\hat{P}_k(s,a)}\left[\tilde{V}_{k,h+1}(s,a)\right]\right)$$
$$\tilde{\mathbf{V}}_{k,h}(s,a) = \max_{a\in\mathcal{A}} \tilde{\mathbf{Q}}_{k,h}(s,a)$$

We make use of the following theorem for Gaussian reward augmentation:

Lemma 6. *Let* $(s,a) \in \mathcal{S}\times\mathcal{A}$. *If* $\tilde{r}_k^{(m)}(s,a) \sim \hat{r}_k(s,a) + \mathcal{N}(0,\sigma^2)$ *for* $\sigma = 2H\beta_r(N_k(s,a), \frac{\delta}{2|S||A|})$ *then:*

$$\mathbb{P}(\tilde{r}_k^{(m)}(s,a) \geq r(s,a) + b_k(s,a)|\mathcal{E}) \geq \frac{1}{10}. \tag{19}$$

Where $b_k(s,a) = 7HL\sqrt{\frac{1}{N_k(s,a)}}$ *where* $L = \ln(5SAKH/\delta)$.

The same proof as in Lemma 2 yields the result.

As a consequence of this result it is easy to see that making use of multiple samples provides concentration and optimisim with enough probability. Following the logic in the proof of UCBVI guarantee with bonus 1 in Azar et al. [2017] yields the proof of the main Gaussian UCBVI guarantees.

3 ADDITIONAL EXPERIMENTAL RESULTS

Delusional World Models In Fig. 1 we show the learning curves from the nine different configurations tested on the InvertedPendulum task, shown in the main paper. In green we show the mean reward in the environment, and we see that some settings fail to learn. In red we show the reward for the policy inside the model. This clearly shows that the models are being exploited, as in some cases the model shows a high reward yet the policy is getting close to zero in the true environment (e.g. $K = 10, \epsilon_K = \text{None}$).

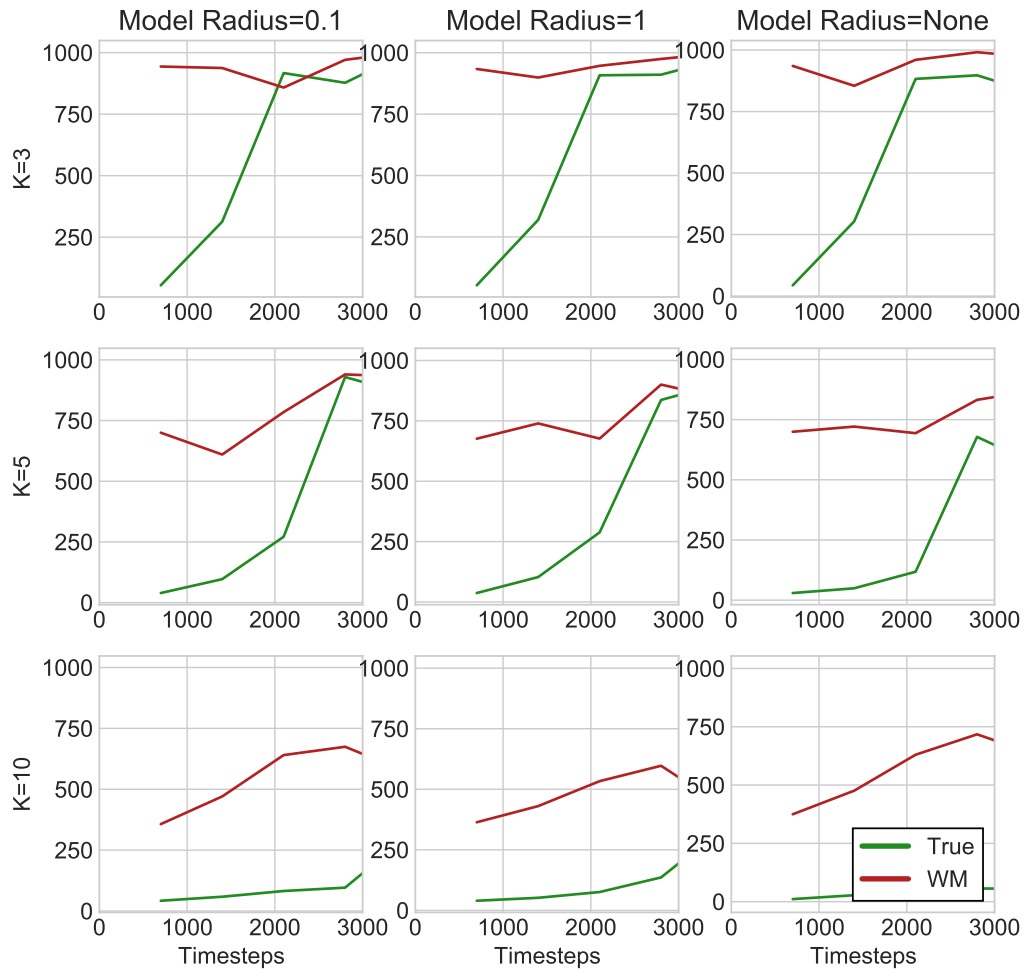

Figure 1: Mean policy performance inside the world model (WM, green), and in the true environment (red), for 3000 timesteps in the InvertedPendulum task.

4 IMPLEMENTATION DETAILS

We adapt our code from our own PyTorch implementation of MBPO Janner et al. [2019]. Most hyperparameters were chosen as in their paper. The main new hyperparameter is ϵ_M. When optimizing the policy inside the model, we use the procedure in 5.

5 EXTENDED SOFT POLICY ITERATION

For the Deep RL experiments, we need to consider that in MBPO-based approaches, a policy is learned using soft policy iteration, which aims to maximise not pure cumulative return, but cumulative return and some total entropy per timestep:

$$\pi^* = \arg\max \sum_t \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[r(\mathbf{s}_t, \mathbf{a}_t) + \alpha \mathcal{H}(\pi(\cdot|\mathbf{s}_t)) \right].$$

We must therefore integrate the policy iteration used in UCRL2 with this new entropy-dependency in order to accurately assess optimism over value and entropy.

We first write out soft-policy iteration:

$$\begin{aligned}\mathcal{T}^{\pi}Q(s,a) =& r(s,a) + \gamma \mathbb{E}_{s' \sim p}[V(s')] \\ V(s) =& \mathbb{E}_{a \sim \pi}[Q(s,a) - \alpha \pi(a|s)]\end{aligned}$$

where $\mathcal{T}^{\pi}$ is the modified Bellman operator.

We then note it is possible to write extended policy iteration from UCRL2 using Bellman operator notation:

$$\mathcal{T}^{\pi}V(s) = \max_{a \in A} \left\{ \tilde{r}(s,a) + \max_{p \in P(s,a)} \left\{ \sum_{s' \in S} p(s')V(s') \right\} \right\}$$

Combining these two approaches, one possibility is as follows, which we will term Extended Soft Policy Iteration (ESPI):

$$\begin{aligned}\mathcal{T}^{\pi}Q(s,a) =& \tilde{r}(s,a) + \gamma \max_{p \in P(s,a)} \left\{ \mathbb{E}_{s' \sim p}[V(s')] \right\} \\ V(s) =& \max_{a \in A} \{Q(s,a) - \alpha \log \pi(a|s)]\}\end{aligned}$$

However this is intractable in a continuous control; we relax this to the tractable SPI approach:

$$V(s) = \mathbb{E}_{a \sim \pi}[Q(s,a) - \alpha \log \pi(a|s)].$$

Observing the Bellman operator term in ESPI, we note that there are two key differences to standard SPI: 1) we take some optimistic reward per state-action pair; 2) we take the most optimistic dynamics model which maximises the expected return from the next state.

In order to implement the former, we apply the max over the reward predictions from the models subject to the model radius parameter ϵ_M.

In order to implement the max over models, we generate next states across all models, pass those next states through the actor to generate the stochastic action, then calculate the expected 'soft' return across off models by passing the resultant next states and actions through the critic, subject to each action's respective entropy. Since action selection is stochastic, it is possible to generate multiple samples over the policy next actions to acquire a more accurate estimation of the soft expected return, but we found sampling the actor once was sufficient in practice.

References

Marc Abeille, Alessandro Lazaric, et al. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2): 5165–5197, 2017.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Neural Information Processing Systems*. 2019.

Branislav Kveton, Csaba Szepesvári, Sharan Vaswani, Zheng Wen, Tor Lattimore, and Mohammad Ghavamzadeh. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, 2019.

Odalric-Ambrym Maillard and Mahsa Asadi. Upper confidence reinforcement learning exploiting state-action equivalence. 2018.