

Supplementary Material

A ADDITIONAL DEFINITIONS AND NOTATION

We start with the definition of the Concentrability constant

Definition A.1. (Concentrability Constant) (Farahmand et al., 2017a)

Given $\rho, \nu \in \Delta(\mathcal{S})$, an integer $k > 0$, and an arbitrary sequence of policies $\pi_{i=1}^m$, the distribution $\rho \mathcal{P}^{\pi_1} \mathcal{P}^{\pi_2} \dots \mathcal{P}^{\pi_k}$ denotes the future state distribution obtained when the state in the first step is distributed according to ρ and the agent follows the sequence of policies π_1, \dots, π_k . Define:

$$c_{\rho, \nu}(k) = \sup_{\pi_1, \dots, \pi_k} \left\| \frac{d\rho \mathcal{P}^{\pi_1} \mathcal{P}^{\pi_2} \dots \mathcal{P}^{\pi_k}}{d\nu} \right\|_{2, \nu}$$

Here, $\|f\|_{2, \nu}^2 = \int f(s)^2 d\nu$. The derivative $\frac{d\rho \mathcal{P}^{\pi_1} \mathcal{P}^{\pi_2} \dots \mathcal{P}^{\pi_k}}{d\nu}$ is the Radon-Nikodym Derivative of two probability measures, which is well-defined up to a set of measure zero by ν if $\rho \mathcal{P}^{\pi_1} \mathcal{P}^{\pi_2} \dots \mathcal{P}^{\pi_k}$ is absolutely continuous with respect to ν . In case it's not absolutely continuous, we set it to be ∞ . Then, for a constant $0 \leq \gamma < 1$, define the discounted weighted average concentrability coefficient as

$$C(\rho, \nu) = (1 - \gamma)^2 \sum_{k=1}^{\infty} \gamma^{k-1} c_{\rho, \nu}(k)$$

Throughout the proof, we use $\|f\|_{p, \mu}$ denote the $L_p(\mu)$ -norm $1 \leq p < \infty$ of a measurable function $f : \mathcal{S} \rightarrow \mathbb{R}$ such that

$$\|f\|_{p, \mu}^p = \int_{\mathcal{S}} |f(x)|^p d\mu(x)$$

In addition, we define the empirical norm. Given a collection points $\{s_1, \dots, s_n\}$ in \mathcal{S} , define the empirical norm $L_p(s_1, \dots, s_n)$ such that

$$\|f\|_{L_p(s_1, \dots, s_n)}^p = \frac{1}{N} \sum_{i=1}^N |f(s_i)|^p$$

Finally, we define Rademacher complexity as in (Bartlett et al., 2005). Same as Farahmand et al. (2017a), we will use a local variant of Rademacher complexity to derive the rate of estimation error.

Definition A.2. (Rademacher Complexity) Let $\sigma_1, \dots, \sigma_n$ be n independent Rademacher random variables, i.e. $\mathbb{P}\{\sigma_i = -1\} = \mathbb{P}\{\sigma_i = 1\} = \frac{1}{2}$. Given a collection of measurable functions \mathcal{F} from \mathcal{X} to \mathbb{R} and a probability distribution μ over \mathcal{X} , we sample n points x_1, \dots, x_n i.i.d. from μ . Define

$$R_n \mathcal{F} = \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^n \sigma_i f(x_i)$$

Then we define the Rademacher complexity of \mathcal{F} as $\mathbb{E}[R_n \mathcal{F}]$

Besides, same as (Bartlett et al., 2005), we define the sub-root function as non-negative and non-decreasing function $\psi : [0; \infty) \rightarrow [0, \infty)$ such that $r \mapsto \frac{\psi(r)}{\sqrt{r}}$ is non-increasing for $r > 0$

B PROOF OF THE THEOREM

We begin by citing the following theorem.

Theorem B.1. (Bartlett et al., 2005) Let \mathcal{F} be a class of functions with values in range $[a, b]$ and assume that there are some functional $T : \mathcal{F} \rightarrow \mathbb{R}^+$ and some constant B such that for every $f \in \mathcal{F}$,

$$\text{Var}[f] \leq T(f) \leq B\mathbb{E}(f) \quad (10)$$

Let ψ be a sub-root function and let $r^* = r^*(\mathcal{F})$ be the fixed point of ψ . Assume that for any $r \leq r^*$, ψ satisfies

$$\psi(r) > B\mathbb{E}[R_n \{f \in \mathcal{F} : T(f) \leq r\}] \quad (11)$$

Then, with $c_1 = 704$ and $c_2 = 26$, for any $K > 1$ and every $x > 0$, with probability at least $1 - e^{-x}$, for any $f \in \mathcal{F}$, we have

$$\mathbb{E}[f] \leq \frac{K}{K-1} \mathbb{E}_n[f] + \frac{c_1 K}{B} r^* + \frac{x(11(b-a) + c_2 BK)}{n} \quad (12)$$

Also with a probability at least $1 - e^{-x}$, for any $f \in \mathcal{F}$, we have

$$\mathbb{E}_n[f] \leq \frac{K}{K-1} \mathbb{E}[f] + \frac{c_1 K}{B} r^* + \frac{x(11(b-a) + c_2 BK)}{n} \quad (13)$$

We will then use $r^*(\mathcal{F})$ to denote the fixed point of a sub-root function ψ that satisfies 11

Now we prove the following theorem which provides a finite sample bound on the value-aware model error.

Theorem B.2. Under the four assumptions 3.1, 3.2, 3.3, 3.5, with the probability model learned based on Equation 3, there exists a constant $\kappa(\alpha)$ which depends solely on $\alpha \in (0, 1)$, such that with probability $1 - \delta$,

$$L(\hat{\mathcal{P}}) \leq \epsilon + \frac{\kappa(\alpha) D^2 R^{\frac{2\alpha}{1+\alpha}} \ln(\frac{1}{\delta})}{N^{\frac{1}{1+\alpha}}}, \quad (14)$$

where N is the number of samples from the data-collection distribution ρ , D is the size of the state-space defined in assumption 3.2, and R is defined in the metric entropy condition of the model class in assumption 3.5.

Proof. Given a batch of state-action transition triples $\{(s_i, a_i, s'_i)\}_{i=1}^N$ with (s_i, a_i) sampled i.i.d from data distribution ρ , we denote the empirical loss

$$L_n(\mathcal{P}) = \frac{1}{N} \sum_{i=1}^N \int_{\mathcal{S}} \mathcal{P}(\hat{s}'_i | s_i, a_i) \|\hat{s}'_i - s'_i\| d\hat{s}'_i \quad (15)$$

We also denote the underlying loss over the data distribution ρ as

$$L(\mathcal{P}) = \mathbb{E}_{(s,a) \sim \rho} \left[\int_{\mathcal{S}} \mathcal{P}(\hat{s}' | s, a) \|\hat{s}' - s'\| d\hat{s}' \right] \quad (16)$$

In addition, let $\tilde{\mathcal{P}} \in \mathcal{M}$ be the best model in the transition kernel class \mathcal{M} defined in 3.5.

Now we would like to apply Theorem B.2 to bound the difference between the empirical and the true underlying loss. First, let $\mathcal{F} = \{(s \times a, s') \mapsto l(s \times a, s'; \mathcal{P}) - l(s \times a, s'; \tilde{\mathcal{P}}); \mathcal{P} \in \mathcal{M}\}$ be the class of functions in Theorem B.1, where $l(s \times a, s'; \mathcal{P})$ is the single datapoint version of the empirical loss $L_n(\mathcal{P})$. So $\mathbb{E}_{s \times a \sim \rho} [l(s \times a, s'; \mathcal{P})] = L(\mathcal{P})$. Now by assumption 3.2, $0 \leq l(s \times a, a; \mathcal{P}) \leq 2D$ for every $s \times a \in \mathcal{S} \times \mathcal{A}$, $s' \in \mathcal{S}$, and $\mathcal{P} \in \mathcal{M}$. Therefore, the value of f is bounded between $-2D$ and $2D$ for every $f \in \mathcal{F}$,

As a consequence, for every $f \in \mathcal{F}$, $\text{Var}(f) \leq \mathbb{E}[f^2] \leq 4D^2$. So we can set

$$T(f) = 4D^2 \mathbb{E} \left[\int_{\mathcal{S}} \mathcal{P}(\hat{s}' | s, a) \|\hat{s}' - s'\| d\hat{s}' \right]$$

$$B = 4D^2$$

Now, we can apply Theorem B.2 to conclude that with probability $1 - \delta$ (let $K = 2$),

$$L(\mathcal{P}) - L(\tilde{\mathcal{P}}) \leq 2(L_n(\mathcal{P}) - L_n(\tilde{\mathcal{P}})) + \frac{2 \times 704}{4D^2} r^*(\mathcal{F}) + \frac{(11 \times 4D + 2 \times 26 \times 4D^2) \ln(\frac{1}{\delta})}{N} \quad (17)$$

Since $\hat{\mathcal{P}}$ is the minimizer of the empirical loss $L_n(\mathcal{P})$,

$$L(\hat{\mathcal{P}}) - L(\tilde{\mathcal{P}}) \leq \frac{352}{D^2} r^*(\mathcal{F}) + \frac{(44D + 208D^2) \ln(\frac{1}{\delta})}{N} \quad (18)$$

We can provide an upper bound of the local Rademacher complexity $r^*(\mathcal{F})$: there exists a finite constant $\tau > 0$ such that for a given $0 \leq \alpha \leq 1$, we have

$$r^*(\mathcal{F}) \leq \frac{c_1(\alpha)D^4R^{\frac{2\alpha}{1+\alpha}}}{N^{\frac{1}{1+\alpha}}} + \frac{\tau D^4 \ln N}{N}, \quad (19)$$

where $c(\alpha) = \frac{\tau}{(1-\alpha)^{\frac{2}{1+\alpha}}}$. The proof follows the exact same steps of Proposition 10 in Farahmand et al. (2017a). Now back to Equation 18, by the realizability assumption 3.3, the best model $\tilde{\mathcal{P}}$ in the model class satisfies that $L(\tilde{\mathcal{P}}) \leq \epsilon$. Therefore, with probability $1 - \delta$,

$$L(\hat{\mathcal{P}}) \leq \epsilon + \frac{352c_1(\alpha)D^2R^{\frac{2\alpha}{1+\alpha}}}{N^{\frac{1}{1+\alpha}}} + \frac{352\tau D^2 \ln N}{N} + \frac{(44D + 208D^2) \ln(\frac{1}{\delta})}{N} \quad (20)$$

Finally, there should exist a constant $\kappa(\alpha)$ sufficiently large such that with probability $1 - \delta$,

$$L(\hat{\mathcal{P}}) \leq \epsilon + \frac{\kappa(\alpha)D^2R^{\frac{2\alpha}{1+\alpha}} \ln(\frac{1}{\delta})}{N^{\frac{1}{1+\alpha}}} \quad (21)$$

Corollary B.3. *Under the five assumptions 3.1, 3.2, 3.3, and 3.5 with the probability model learned based on Equation 3, there exists a constant $\kappa(\alpha)$ which depends solely on $\alpha \in (0, 1)$, such that with probability $1 - \exp(-\frac{\epsilon N^{\frac{1}{1+\alpha}}}{\kappa(\alpha)D^2R^{\frac{2\alpha}{1+\alpha}}})$,*

$$L(\hat{\mathcal{P}}) \leq 2\epsilon, \quad (22)$$

Proof. This is a straightforward application of Theorem B.2, where we could just let $\epsilon = \frac{\kappa(\alpha)D^2R^{\frac{2\alpha}{1+\alpha}} \ln(\frac{1}{\delta})}{N^{\frac{1}{1+\alpha}}}$.

Next, we consider the local Lipschitz condition of the value function and provide a finite sample bound of the value-aware model error.

Theorem B.4. *Under the five assumptions 3.1, 3.2, 3.3, 3.4, and 3.5, with the probability model learned based on Equation 3, there exists a constant $\kappa(\alpha)$ which depends solely on $\alpha \in (0, 1)$, such that for any $m > 1$,*

$$\int |\mathcal{T}^*Q(s, a) - \hat{\mathcal{T}}^*Q(s, a)|^2 d\rho(s, a) \leq \gamma^2 [4\epsilon^2 L^2 \xi + \frac{R_{\max}^2}{(1-\gamma)^2} (1 - \xi)], \quad (23)$$

where $\xi = 1 - \exp(-\frac{\epsilon N^{\frac{1}{1+\alpha}}}{\kappa(\alpha)D^2R^{\frac{2\alpha}{1+\alpha}}})$

Proof.

$$\begin{aligned} \|\mathcal{T}^*Q - \hat{\mathcal{T}}^*Q\|_\rho^2 &= \int \left| r(s, a) + \gamma V(s') - r(s, a) - \gamma \int \hat{\mathcal{P}}(ds'|s, a) V(s') \right|^2 d\rho(s, a) \\ &= \gamma^2 \int \left| \int \hat{\mathcal{P}}(ds'|s, a) (V(s') - V(s')) \right|^2 d\rho(s, a) \\ &\leq \gamma^2 \int \int \mathcal{P}(ds'|s, a) (V(s') - V(s'))^2 d\rho(s, a) \\ &\leq \gamma^2 \int \int \mathbb{1}\{\|s' - \hat{s}'\| \leq 2\epsilon\} (V(s') - V(s'))^2 \hat{\mathcal{P}}(ds'|s, a) d\rho(s, a) \\ &\quad + \gamma^2 \int \int \mathbb{1}\{\|s' - \hat{s}'\| > 2\epsilon\} (V(s') - V(s'))^2 \hat{\mathcal{P}}(ds'|s, a) d\rho(s, a) \\ &\leq \gamma^2 2^2 \epsilon^2 L^2 \mathbb{P}\{\|s' - \hat{s}'\| \leq 2\epsilon\} + \gamma^2 \frac{R_{\max}^2}{(1-\gamma)^2} \mathbb{P}\{\|s' - \hat{s}'\| > 2\epsilon\} \end{aligned}$$

Now apply Corollary B,

$$\begin{aligned} \|\mathcal{T}^*Q - \hat{\mathcal{T}}^*Q\|_\rho^2 &\leq \gamma^2 4\epsilon^2 L^2 \left(1 - \exp\left(-\frac{\epsilon N^{\frac{1}{1+\alpha}}}{\kappa(\alpha) D^2 R^{\frac{2\alpha}{1+\alpha}}}\right)\right) \\ &\quad + \gamma^2 \frac{R_{\max}^2}{(1-\gamma)^2} \exp\left(-\frac{\epsilon N^{\frac{1}{1+\alpha}}}{\kappa(\alpha) D^2 R^{\frac{2\alpha}{1+\alpha}}}\right) \end{aligned}$$

Finally, with the value-aware model error bounded, we could apply the error propagation results from (Munos, 2005; Farahmand et al., 2017a) and prove our main theorem, which relates the local Lipschitz constant to the sub-optimality of the approximate model-based value iteration algorithm.

Theorem B.5. *Suppose \hat{Q}_0 is initialized such that $\hat{Q}_0(s, a) \leq \frac{R_{\max}}{1-\gamma}$ for $\forall (s, a)$. Under the assumptions of 3.1, 3.2, 3.3, 3.4, and 3.5, after K iterations of the model-based approximate value iteration algorithm, there exists a constant $\kappa(\alpha)$ which depends solely on $\alpha \in (0, 1)$ such that,*

$$\begin{aligned} \mathbb{E}_{(s,a) \sim \mathcal{P}_0} \left[\left| Q^*(s, a) - \hat{Q}_K(s, a) \right| \right] &\leq \frac{2\gamma}{(1-\gamma)^2} \left[C(\rho, \mathcal{P}_0) \left(\max_{0 \leq k \leq K} \delta_k + \gamma^2 (4\epsilon^2 L^2 \xi \right. \right. \\ &\quad \left. \left. + \frac{(1-\xi)R_{\max}^2}{(1-\gamma)^2}) \right) + 2\gamma^K R_{\max} \right] \end{aligned} \quad (24)$$

where $\delta_k^2 = \mathcal{L}_{\text{reg}}(\hat{Q}_k; \hat{Q}_{k-1}, \hat{\mathcal{P}}_k)$ is the regression error defined in Equation (4), $\xi = 1 - \exp\left(-\frac{\epsilon N^{\frac{1}{1+\alpha}}}{\kappa(\alpha) D^2 R^{\frac{2\alpha}{1+\alpha}}}\right)$, and $C(\rho, \mathcal{P}_0)$ is the concentrability constant defined in Definition A.1.

Proof. This follows directly from the Theorem B.4 and also Theorem 4 from (Farahmand et al., 2017a), where the value-aware model error $e_{\text{model}}(N) \leq \gamma^2 \left(4\epsilon^2 L^2 \xi + \frac{R_{\max}^2}{(1-\gamma)^2} (1-\xi) \right)$

C IMPLEMENTATION DETAILS

In this section, we are going to introduce the implementation details of our two proposed methods. We implement our methods based on a PyTorch implementation of MBPO (Lin, 2022). The dynamics model architecture is MLP with four hidden layers of size 200. In Ant and Humanoid, the hidden size is 400 because these two environments are more complex than others. For the probabilistic dynamics model ensemble, we set the ensemble size to 7 which is the setting used in the original paper of MBPO (Janner et al., 2019). The policy is optimized with Soft Actor-Critic (SAC) (Haarnoja et al., 2018). The actor network architecture and the critic network architecture are MLP with two hidden layers of size 256.

For robust regularization, we fix λ to 0.1 as discussed in Section 4.2, and we do a grid search of λ over $[0.01, 0.1, 1, 10]$. For spectral normalization, we add the normalization on each layer of the critic network. In particular, at every forward pass, we approximate the spectral radius of the weight matrix with one step of power iteration. The algorithm is sketched below with \mathbf{u} and \mathbf{v} being the right and left singular vector of the weight matrix W .

$$\mathbf{v} \leftarrow W\mathbf{u}^{(t-1)}; \quad \alpha \leftarrow \|\mathbf{v}\|; \quad \mathbf{v}^{(t)} \leftarrow \alpha^{-1}\mathbf{v} \quad (25)$$

$$\mathbf{u} \leftarrow W^T\mathbf{v}^{(t)}; \quad \rho \leftarrow \|\mathbf{v}\|; \quad \mathbf{u}^{(t)} \leftarrow \rho^{-1}\mathbf{u} \quad (26)$$

Table 2: Hyperparameters used in the experiments

Method	Robust Regularization λ	Spectral Normalization β
Walker	0.1	20
Ant	0.1	25
Humanoid	1.0	30
Hopper	0.01	20
HalfCheetah	0.1	25

Then we perform a projection of the parameters: $W := \max(1, \frac{\max(\alpha, \rho)}{\beta})^{-1} W$. So the spectral norm will be clipped to β if it's bigger than λ , unchanged otherwise. We do a grid search of β over [15, 20, 25, 30, 35]. Settings of λ and β across all five environments are provided in Table 2. For a fair comparison, we set the rollout horizon during model rollouts to 1 in all environments.

D ADDITIONAL EXPERIMENT RESULTS

D.1 ADDITIONAL EXPERIMENTAL RESULTS OF VALUE-AWARE MODEL ERROR

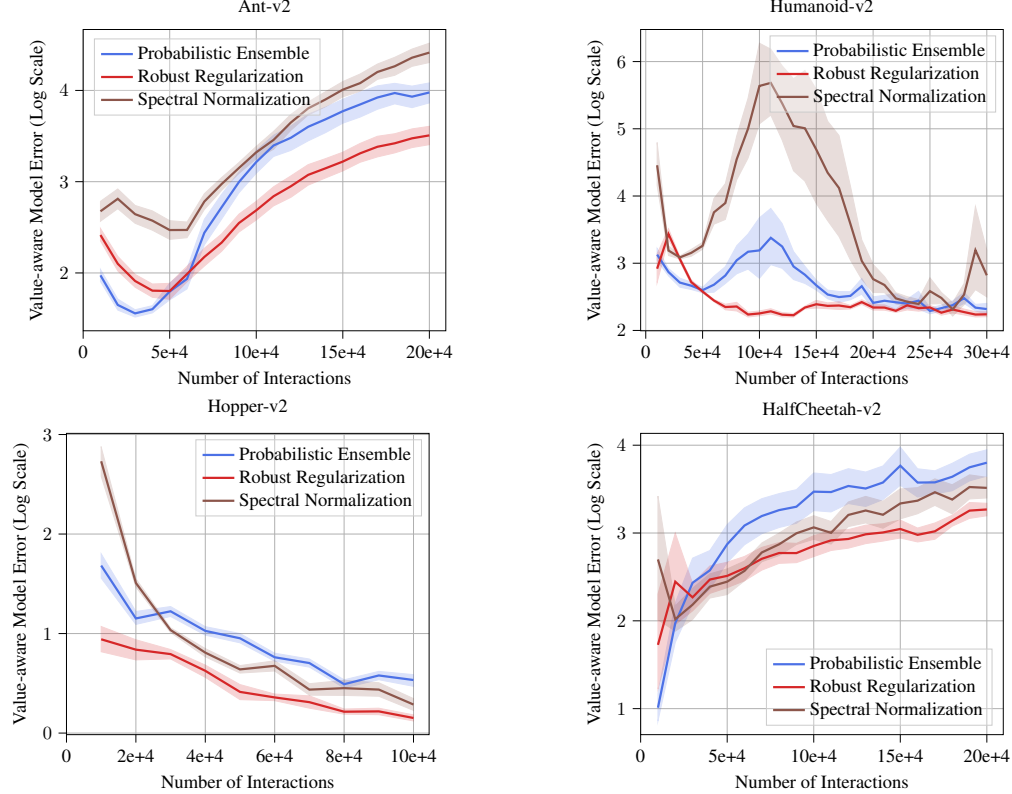


Figure 5: Comparison between the proposed mechanisms and the probabilistic ensemble baseline in terms of Value-aware Model Error (Log Scale)

In Section 5.2 and 5.3, we demonstrate the effectiveness of robust regularization in controlling the value-aware model error on Walker and also show the limitation of constraining the global Lipschitz constant by spectral normalization. Here, we provide the results of value-aware model error in the rest of the four environments. We used the same hyperparameters reported in Figure 3 and Table 2, which have the best performance under grid search. Once again, we observe that robust regularization effectively reduces the value-aware model error by constraining the local Lipschitz condition with computing adversarial perturbation. In addition, we also see that global Lipschitz constraints are too strong for spectral normalization. In two more complicated environments, Ant and Humanoid, it has to sacrifice value-aware model error for the expressive power of the value function. Therefore, spectral normalization does not achieve a good performance in these two environments. However, in two easier environments, Hopper and HalfCheetah, spectral normalization could still effectively reduce the value-aware model error and has good empirical performance.

D.2 PROPOSED MECHANISMS WITH SINGLE PROBABILISTIC MODEL

In the experiment section, we combine our proposed mechanisms with a single deterministic model and compare it against MBPO using an ensemble of probabilistic models. The purpose is to verify

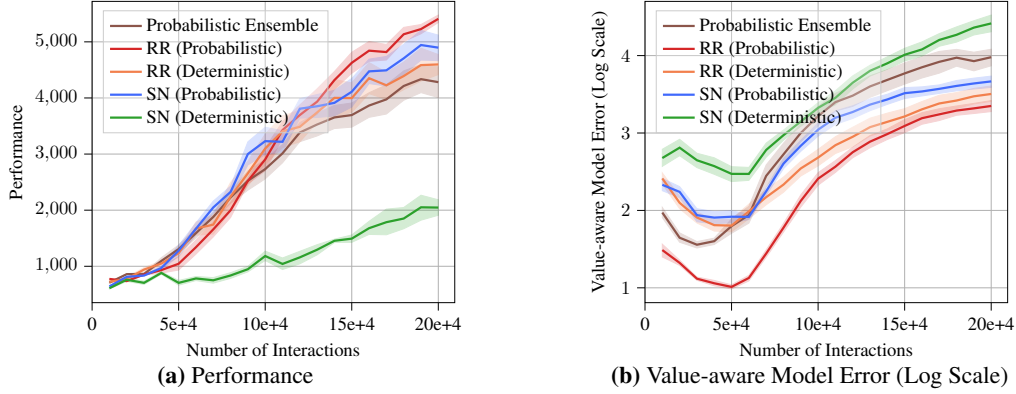


Figure 6: Spectral Normalization and Robust Regularization with a single probabilistic model on Ant. RR is short for Robust Regularization, and SN is short for Spectral Normalization

that regularization of the local Lipschitz constant is critical in MBRL algorithms and propose a computationally efficient MBRL algorithm without a model ensemble. In practice, complementary to our proposed Lipschitz regularization mechanisms, we can also use a single probabilistic model to further regularize the local Lipschitz condition of the value function. In addition, training a probabilistic environment model would be better suited for environments with stochastic transitions.

Here in Figure 6, we combine our two proposed mechanisms with both a probabilistic and deterministic model on Ant, comparing them with the probabilistic ensemble baseline. From Figure 6b, we see that although spectral normalization with a single deterministic model has a large value-aware model error, it is significantly reduced when combining it with a probabilistic dynamics model. Therefore, we find that spectral normalization with a probabilistic model achieves much better performance and even outperforms MBPO with an ensemble of probabilistic models. For robust regularization, using a probabilistic model also helps improve the algorithm’s value-aware model error and performance. This observation suggests that the two Lipschitz regularization approaches, explicit regularization by spectral normalization or robust regularization and implicit regularization by probabilistic models, are complementary. In practice, we can combine the two approaches to get the best performance of the MBRL algorithm.

D.3 ROBUST REGULARIZATION WITH FGSM vs. PGD

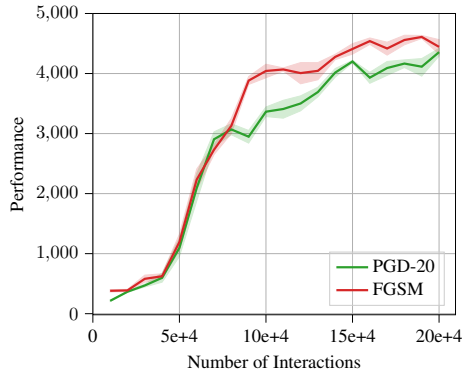


Figure 7: Robust Regularization with 20 steps of Project Gradient Descent (PGD-20) against Fast Gradient Sign Method (FGSM).

In Figure 7, we compare the performance of robust regularization with 20 steps of Project Gradient Descent (PGD-20) against the Fast Gradient Sign Method (FGSM) on Walker. In particular, although PGD-20 is much more computationally expensive, we do not observe the improvement with this more powerful constrained optimization solver.

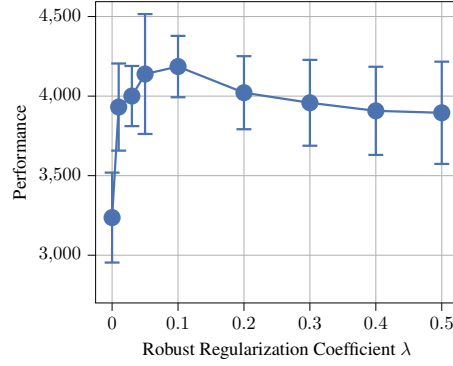


Figure 8: Robust Regularization with different regularization weights λ 's on Walker. Experiments are all with 8 random seeds.

D.4 ROBUST REGULARIZATION WITH DIFFERENT REGULARIZATION WEIGHTS

Figure 8 further visualizes how the regularization weight λ of robust regularization influences the algorithm performance. Similar to the findings of the experiments on spectral normalization, we see that the algorithm's performance first increases and drops as the regularization weight gets larger. This verifies our theoretical insights from Theorem B.5 that with a small λ , the algorithm gets less regularization and thus has a big value-aware model error. But meanwhile, it also has a small regression error since the regularization has little effect on the expressive power of the value function. When λ goes up, the regularization will have a stronger negative effect on the expressive power of the value function, but the value-aware model error will also get smaller. We observe that the algorithm performs the best with $\lambda = 0.1$, achieving the balance between value-aware model error and the value function's expressive power.

E ADDITIONAL DETAILS ON THE INVERTED-PENDULUM EXPERIMENT

In Section 3.2, we provide an experiment of model-based value iteration on the Inverted Pendulum to further verify the validity of our theorem. Below we provide the pseudocode for it.

Algorithm 1 Model-based Approximate Value Iteration on Inverted Pendulum

K : Total number of iterations for the algorithm
 H : Number of gradient steps to solve the inner optimization
 \mathcal{M}, \mathcal{G} : Space of transition probability kernels and reward functions
 \mathcal{F} : Space of value function
 $\mu \in \Delta(\mathcal{S} \times \mathcal{A})$: a data-collecting state-action distribution
 Sample i.i.d from μ to generate the training dataset $\mathcal{D} = \{(s_i, a_i, s'_i, r_i)\}_{i=1}^N$

$$\hat{\mathcal{P}} \leftarrow \arg \min_{\mathcal{P} \in \mathcal{M}} \sum_{i=1}^N \|s'_i - \int \hat{\mathcal{P}}(ds'|s, a)s'\|^2$$

$$\hat{r} \leftarrow \arg \min_{\hat{r} \in \mathcal{G}} \sum_{i=1}^N (r_i - \hat{r}(s_i, a_i))^2$$

Initialize the value function \hat{Q}_0 .

repeat

for $k = 0$ **to** $K - 1$ **do**

 Sample i.i.d N state-action pairs from $\rho : \{(s_i, a_i)\}_{i=1}^N$

 Compute $\hat{s}'_i \sim \mathcal{P}(\cdot|s_i, a_i)$, $\hat{r}_i = \hat{r}(s_i, a_i)$

for $t = 0$ **to** $H - 1$ **do**

 Update policy using gradient ascent with $\frac{1}{N} \sum_{i=1}^N \nabla_{\theta} \hat{Q}_{\phi}(s_i, \pi_{\theta}(s_i))$

end for

for $t = 0$ **to** $H - 1$ **do**

 Update value function using gradient descent with

$$\frac{1}{N} \sum_{i=1}^N \nabla_{\phi} \left(\hat{r}_i + \gamma \hat{Q}_{\phi}(\hat{s}'_i, \pi_{\theta}(\hat{s}'_i)) - \hat{Q}_{\phi}(s_i, a_i) \right)^2$$

end for

end for

until end of training

Output: \hat{Q}_{ϕ}
