A. Sample Complexity Analysis of PC-MLP

In this section, we provide detailed analysis for PC-MLP on both KNRs and Linear MDPs.

A.1. Model Learning from MLE

We consider a fixed episode n. The derived result here can be applied to all episodes via a union bound. We derive model's generalization error under distribution d^{π_n} in terms of total variation distance. For linear MDPs, we can directly call Theorem 21 from (Agarwal et al., 2020b), which shows under realizability $P^\star \in \mathcal{P}$, the empirical maximizer of MLE directly enjoys the following generalization error.

Lemma 5 (MLE for linear MDPs (Theorem 21 from (Agarwal et al., 2020b)))**.** *Fix $\delta \in (0,1)$, and assume $|\mathcal{P}| < \infty$ and $P^\star \in \mathcal{P}$. Consider M samples with $(s_i, a_i) \sim d^{\pi_n}$, and $s'_i \sim P^\star(\cdot|s,a)$. Denote the empirical risk minimizer as $\widehat{P}_n = \operatorname{argmax}_{P\in\mathcal{P}} \sum_{i=1}^M \ln P(s'_i|s_i,a_i)$. We have that with probability at least $1-\delta$,*

$$\mathbb{E}_{s,a\sim d^{\pi_n}} \left\| \widehat{P}_n(\cdot|s,a) - P^\star(\cdot|s,a) \right\|_1^2 \le \frac{2\ln\left(|\mathcal{P}|/\delta\right)}{M}.$$

For KNRs, we do not need to rely on the exact empirical risk minimization. Instead we can use approximation optimization approach SGD here. Note that due to the Gaussian noise in the KNR model, we have that for a model P_W with W as the parameterization:

$$\ln P_W(s'|s,a) = -\|W\phi(s,a) - s'\|_2^2/\sigma^2 + \ln(1/C),$$

where C is the normalization constant for Gaussian distribution with zero mean and covariance matrix $\sigma^2 I$. Hence gradient of the log-likelihood is equivalent to the gradient of the squared ℓ_2 loss. Specifically, for approximately optimizing the empirical log-likelihood, we start with $W^0 = 0$, and perform SGD with $W^{i+1} = \prod_{W:\|W\|_2\le F}\left(W^i - \eta\left(W^i\phi(s_i,a_i) - s'_i\right)\phi(s_i,a_i)^\top\right)$ and set $\widehat{W}_n = \frac{1}{M}\sum_{i=1}^M W^i$.

To use SGD's result, we first need to bound all states s'. The following lemma indicates that with high probability, the states have bounded norm.

Lemma 6. *In each episode, we generate M data points $\{s_i, a_i, s'_i\}_{i=1}^M$ with $s_i, a_i \sim d^{\pi_n}$ and $s'_i \sim \mathcal{N}\left(W^\star\phi(s_i,a_i), \sigma^2 I\right)$ with $\|W^\star\|_2 \le F$. With probability at least $1-\delta$, we have:*

$$\|s'_i\|_2 \le F + \sigma\sqrt{d_s \ln\left(d_s M/\delta\right)}, \forall i \in [1,\dots,M].$$

Proof. Denote $s'_i = \epsilon_i + W^\star\phi(s_i,a_i)$ where $\epsilon_i \sim \mathcal{N}(0,\sigma^2 I)$. We must have that for a fixed i and dimension $j \in [d_s]$:

$$\mathbb{P}(|\epsilon_i[j]| \ge t) \le \exp(-t^2/\sigma^2).$$

Take a union bound over all i, j, we have:

$$\mathbb{P}(\exists i, j, \text{ s.t. } |\epsilon_i[j]| \ge t) \le d_s M \exp(-t^2/\sigma^2).$$

Set $d_s M \exp(-t^2/\sigma^2) = \delta$, we get:

$$t = \sigma\sqrt{\ln\left(\frac{d_s M}{\delta}\right)}.$$

Hence with probability $1-\delta$, for i, j, we have:

$$|\epsilon_i[j]| \le \sigma\sqrt{\ln\frac{d_s M}{\delta}}.$$

Hence, $\|s'_i\|_2 \le \|W^\star\phi(s_i,a_i)\|_2 + \|\epsilon_i\|_2 \le F + \sigma\sqrt{d_s\ln\left(\frac{d_s M}{\delta}\right)}$. □

Throughout this section, we assume the above event in Lemma 6 holds and we denote $F + \sigma\sqrt{d_s \ln{(d_s M/\delta)}} = B$ for notation simplicity .

Now we can call Lemma 16 (dimension-free SGD result) to conclude the following lemma.

Lemma 7 (MLE on KNRs)**.** *With probability at least* $1-\delta$*, we have that:*

$$\mathbb{E}_{s,a\sim d^{\pi_n}}\left[\left\|\widehat{W}_n\phi(s,a) - W^\star\phi(s,a)\right\|_2^2\right] \leq \frac{3(F^2+FB)\ln(1/\delta)}{\sqrt{M}},$$

where $B = F + \sigma\sqrt{d_s \ln\left(\frac{2d_sM}{\delta}\right)}$*.*

Proof. Lemma 6 states that with probability $1-\delta/2$, we have that $\|s_i'\|_2 \leq B$ for all $i \in [1,\ldots,M]$. Condition on this event, we call Lemma 16 and using the fact that $\|W^\star\|_2 \leq \|W^\star\|_F \leq F$, we get that with probability at least $1-\delta/2$,

$$\mathbb{E}_{s,a\sim d^{\pi_n}}\left[\left\|\widehat{W}_n\phi(s,a) - W^\star\phi(s,a)\right\|_2^2\right] \leq \frac{3(F^2+FB)\ln(2/\delta)}{\sqrt{M}}$$

The total probability of failure is δ.

This concludes the proof. □

A.2. Optimism at the Starting State

We denote $b_n(s,a)$ using Σ_n as follows:

$$b_n(s,a) = \min\left\{c\sqrt{\phi(s,a)^\top\Sigma_n^{-1}\phi(s,.a)},\ H\right\}.$$

Recall that the reward bonus in the algorithm is defined with respect to $\widehat{\Sigma}_n$ in Eq. 2. We will link b_n and $\widehat{b}_n$ later in the analysis.

The bonus is related to the uncertainty in the model. Consider any policy π, reward bonus b_n in the form of Eq. 2, and any model $\widehat{P} \in \mathcal{P}$. Denote $\widehat{V}^\pi_{r+b_n;h}$ as the value function at time step h under policy π in model $\widehat{P}$ and reward $r+b_n$. Note that $r(s,a)+b_n(s,a) \in [0, H+1]$, we have $\|\widehat{V}^\pi_{r+b_n,h}\|_\infty \leq H^2$ for any h.

Lemma 8 (Optimism in Linear MDPs)**.** *Assume the following condition hold for all* $n \in [1,\ldots,N]$*:*

$$\mathbb{E}_{s,a\sim d^{\pi_n}}\|\widehat{P}_n(\cdot|s,a) - P^\star(\cdot|s,a)\|_1^2 \leq \epsilon_{stat} \in \mathbb{R}^+, \forall n.$$

Set $c = H\sqrt{(\lambda d + N\epsilon_{stat})}$*, and assume* $b_n(s,a) \leq \widehat{b}_n(s,a) \leq 4b_n(s,a)$ *holds for all* n*. Then, we have that for any* n*:*

$$\widehat{V}^{\pi_{n+1}}_{r+\widehat{b}_n;0}(s_0) \geq \max_{\pi\in\Pi} V_0^\pi(s_0).$$

Proof. We denote $\pi^\star = \operatorname{argmax}_{\pi\in\Pi} V_0^\pi(s_0)$. We consider $\pi^\star$ specifically. Without loss of generality, we denote $\widehat{V}^\pi_{r+\hat{b}_n;H}(s) = V^\pi_H(s) = 0$ for all $s \in \mathcal{S}$. Thus, for $h = H$, we have $\widehat{V}^{\pi^\star}_{r+\hat{b}_n;H}(s) \geq V_H^{\pi^\star}(s), \forall s$.

Assume that at time step $h+1$, we have $\widehat{V}^{\pi^\star}_{r+\hat{b}_n;h+1}(s) \geq V^{\pi^\star}_{h+1}(s), \forall s$. Now we move on to prove this also holds at time step h. Denote $a^\star = \pi^\star(s)$. Also we define $\Sigma_n = \sum_{i=1}^n \Sigma_{\pi_n} + \lambda I$.

$$\begin{aligned}
\widehat{V}^{\pi^\star}_{r+\hat{b}_n;h}(s) - V_h^{\pi^\star}(s) &= r(s,a^\star) + \widehat{b}_n(s,a^\star) + \mathbb{E}_{s'\sim\widehat{P}^n(\cdot|s,a^\star)}\widehat{V}^{\pi^\star}_{r+\hat{b}_n;h+1}(s') - \Big(r(s,a^\star) + \mathbb{E}_{s'\sim P^\star(\cdot|s,a^\star)}V^{\pi^\star}_{h+1}(s')\Big)\\
&\geq b_n(s,a^\star) + \Big(\mathbb{E}_{s'\sim\widehat{P}^n(\cdot|s,a^\star)}V^{\pi^\star}_{h+1}(s') - \mathbb{E}_{s'\sim P^\star(\cdot|s,a^\star)}V^{\pi^\star}_{h+1}(s')\Big)\\
&= b_n(s,a^\star) + ((\widehat{\mu}_n - \mu^\star)\phi(s,a^\star))\cdot V^{\pi^\star}_{h+1}
\end{aligned}$$

We bound $((\widehat{\mu}_n - \mu^\star)\phi(s,a)) \cdot V^{\pi^\star}_{h+1}$ below.

$$\begin{aligned}
&\left|((\widehat{\mu}_n - \mu^\star)\phi(s,a)) \cdot V^{\pi^\star}_{h+1}\right|^2 \leq \|\phi(s,a)\|^2_{\Sigma_n^{-1}} \left\|(\widehat{\mu}_n - \mu^\star)^\top V^{\pi^\star}_{h+1}\right\|^2_{\Sigma_n} \\
&= \|\phi(s,a)\|^2_{\Sigma_n^{-1}} \left(\lambda \left\|(\widehat{\mu}_n - \mu^\star)^\top V^{\pi^\star}_{h+1}\right\|_2^2 + n\mathbb{E}_{s,a\sim d^{\pi_n}} \left(\phi(s,a)^\top (\widehat{\mu}_n - \mu^\star)^\top V^{\pi^\star}_{h+1}\right)^2\right) \\
&\leq \|\phi(s,a)\|^2_{\Sigma_n^{-1}} \left(\lambda \left\|(\widehat{\mu}_n - \mu^\star)^\top V^{\pi^\star}_{h+1}\right\|_2^2 + nH^2\mathbb{E}_{s,a\sim d^{\pi_n}} \|\widehat{P}_n(\cdot|s,a) - P^\star(\cdot|s,a)\|_1^2\right) \\
&\leq \|\phi(s,a)\|^2_{\Sigma_n^{-1}} \left(\lambda H^2 d + nH^2\mathbb{E}_{s,a\sim d^{\pi_n}} \|\widehat{P}_n(\cdot|s,a) - P^\star(\cdot|s,a)\|_1^2\right) \\
&\leq \left(\lambda H^2 d + nH^2 \epsilon_{stat}\right) \|\phi(s,a)\|^2_{\Sigma_n^{-1}} \leq c^2 \|\phi(s,a)\|^2_{\Sigma_n^{-1}}.
\end{aligned}$$

Thus, we get that:

$$\widehat{V}^{\pi^\star}_{r+\hat{b}_n;h}(s) - V^{\pi^\star}_{h+1}(s) \geq b_n(s,a^\star) - c\|\phi(s,a^\star)\|_{\Sigma_n^{-1}} = 0.$$

Note that the above holds for any s. Thus via induction, we conclude that at $h = 0$, we have $\widehat{V}^{\pi^\star}_{r+\hat{b}_n;0}(s_0) \geq V^{\pi^\star}_0(s_0)$. Using the fact that $\pi_{n+1} = \mathrm{argmax}_{\pi\in\Pi} \widehat{V}^{\pi}_{r+\hat{b}_n;0}(s_0)$, we conclude the proof. □

Lemma 9 (Optimism in KNRs). *Assume the following condition hold for all $n \in [1,\ldots,N]$:*

$$\mathbb{E}_{s,a\sim d^{\pi_n}} \left\|\widehat{W}_n\phi(s,a) - W^\star\phi(s,a)\right\|_2^2 \leq \epsilon_{stat}, \forall n,$$

Set $c = \frac{1}{\sigma}H\sqrt{\lambda 4F^2 + N\epsilon_{stat}}$, and assume that $b_n(s,a) \leq \widehat{b}_n(s,a) \leq 4b_n(s,a)$ holds for all n. we have that for any n:

$$\widehat{V}^{\pi_{n+1}}_{r+\hat{b}_n;0}(s_0) \geq \max_{\pi\in\Pi} V^\pi_0(s_0).$$

Proof. For any n, the condition in the lemma implies that:

$$\sum_{i=1}^n \mathbb{E}_{s,a\sim d^{\pi_i}} \left\|\widehat{W}_n\phi(s,a) - W^\star\phi(s,a)\right\|_2^2 = \mathrm{tr}\left(\left(\widehat{W}_n - W^\star\right)\sum_{i=1}^n \Sigma_{\pi_i}\left(\widehat{W}_n - W^\star\right)^\top\right) \leq n\epsilon_{stat}$$

Note that $\Sigma_n = \sum_{i=1}^n \Sigma_{\pi_i} + \lambda I$, we have that:

$$\left\|\left(\widehat{W}_n - W^\star\right)\Sigma_n^{1/2}\right\|_2^2 \leq \mathrm{tr}\left(\left(\widehat{W}_n - W^\star\right)\Sigma_n\left(\widehat{W}_n - W^\star\right)^\top\right) \leq n\epsilon_{stat} + \lambda 4F^2,$$

where we use the norm bound that $\|\widehat{W}_n\|_F^2 \leq F^2, \|W^\star\|_F^2 \leq F^2$.

Similarly, we can use induction to prove optimism. Assume $\widehat{V}^{\pi^\star}_{r+\hat{b}_n;h+1}(s) \geq V^{\pi^\star}_{h+1}(s)$ for all s. For any s, denote $a^\star = \pi^\star(s)$, we have:

$$\begin{aligned}
\widehat{V}^{\pi^\star}_{r+\hat{b}_n;h}(s) - V^{\pi^\star}_h(s) &\geq b_n(s,a^\star) + \left(\mathbb{E}_{s'\sim\widehat{P}^n(\cdot|s,a^\star)} V^{\pi^\star}_{h+1}(s') - \mathbb{E}_{s'\sim P^\star(\cdot|s,a^\star)} V^{\pi^\star}_{h+1}(s')\right) \\
&\geq b_n(s,a^\star) - \left\|\widehat{P}_n(\cdot|s,a^\star) - P^\star(\cdot|s,a^\star)\right\|_1 \|V^{\pi^\star}_{h+1}\|_\infty \\
&\geq b_n(s,a^\star) - \frac{H}{\sigma}\left\|\left(\widehat{W}_n - W^\star\right)\phi(s,a^\star)\right\|_2 \\
&\geq b_n(s,a^\star) - \frac{H}{\sigma}\left\|\widehat{W}_n - W^\star\right\|_{\Sigma_n} \|\phi(s,a^\star)\|_{\Sigma_n^{-1}} \\
&\geq b_n(s,a^\star) - \frac{H\sqrt{n\epsilon_{stat} + \lambda 4F^2}}{\sigma}\|\phi(s,a^\star)\|_{\Sigma_n^{-1}} \geq 0,
\end{aligned}$$

due to the set up of c. Similar via induction, this concludes the proof. □

A.3. Regret Upper Bound

Below we consider bounding $\sum_{n=1}^{N}\left(J(\pi^\star; P^\star) - J(\pi_n; P^\star)\right)$ using optimism we proved in the section above.

Lemma 10 (Regret bound in linear MDPs)**.** *Assuming all conditions in Lemma 8 holds. We have:*

$$\sum_{n=1}^{N}\left(J(\pi^\star; r, P^\star) - J(\pi_n; r, P^\star)\right) \leq 6H^2 \sum_{n=1}^{N-1} \mathbb{E}_{s,a\sim d^{\pi_{n+1}}}\left[b_n(s,a)\right] + H.$$

Proof. Since the condition in Lemma 8 holds, we have that for all n, optimism holds, i.e., $J(\pi_{n+1}; r + \widehat{b}_n, \widehat{P}_n) \geq J(\pi^\star; r, P^\star)$. Hence, together with the simulation lemma (Lemma 19) we have:

$$\begin{aligned} J(\pi^\star; r, P^\star) - J(\pi_{n+1}; r, P^\star) &\leq J(\pi_{n+1}; r + \widehat{b}_n, \widehat{P}_n) - J(\pi_{n+1}; r, P^\star) \\ &= \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi_{n+1}}}\left[\widehat{b}_n(s,a) + \left(\widehat{P}_n(\cdot|s,a) - P^\star(\cdot|s,a)\right)\cdot \widehat{V}^{\pi_{n+1}}_{r+\hat{b}_n;h+1}\right] \\ &\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi_{n+1}}}\left[4b_n(s,a) + \left(\widehat{P}_n(\cdot|s,a) - P^\star(\cdot|s,a)\right)\cdot \widehat{V}^{\pi_{n+1}}_{r+\hat{b}_n;h+1}\right]. \end{aligned}$$

Note that $\|\widehat{V}^{\pi}_{r+\hat{b};h}\|_\infty \leq H^2$ for any π, h, n. Following similar derivation in the proof of Lemma 8, we have:

$$\begin{aligned} \left|((\widehat{\mu}_n - \mu^\star)\phi(s,a))\cdot \widehat{V}^{\pi_{n+1}}_{r+b_n;h+1}\right| &\leq \min\left\{\|\phi(s,a)\|_{\Sigma_n^{-1}}\sqrt{(\lambda H^4 d + nH^4\epsilon_{stat})}, 2H^2\right\} \\ &\leq 2H\min\left\{H\sqrt{\lambda d + N\epsilon_{stat}}\cdot\|\phi(s,a)\|_{\Sigma_n^{-1}},\ H\right\} = 2Hb_n(s,a). \end{aligned}$$

Note that the regret at the first policy π_1 is at most H. This concludes the proof. □

Lemma 11 (Regret bound in KNRs)**.** *Assuming all conditions in Lemma 9 holds. We have:*

$$\sum_{n=1}^{N}\left(J(\pi^\star; r, P^\star) - J(\pi_n; r, P^\star)\right) \leq 5H^2 \sum_{n=1}^{N-1} \mathbb{E}_{s,a\sim d^{\pi_{n+1}}}\left[b_n(s,a)\right] + H.$$

Proof. Again, via simulation lemma and optimism, we have:

$$\begin{aligned} J(\pi^\star; r, P^\star) - J(\pi_{n+1}; r, P^\star) &\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi_{n+1}}}\left[\widehat{b}_n(s,a) + \left(\widehat{P}_n(\cdot|s,a) - P^\star(\cdot|s,a)\right)\cdot \widehat{V}^{\pi_{n+1}}_{r+b_n;h+1}\right] \\ &\leq \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi_{n+1}}}\left[4b_n(s,a) + \left(\widehat{P}_n(\cdot|s,a) - P^\star(\cdot|s,a)\right)\cdot \widehat{V}^{\pi_{n+1}}_{r+b_n;h+1}\right] \end{aligned}$$

Following the derivation in the proof of Lemma 9, we have:

$$\left|\left(\widehat{P}_n(\cdot|s,a) - P^\star(\cdot|s,a)\right)\cdot \widehat{V}^{\pi_{n+1}}_{r+b_n;h+1}\right| \leq \frac{H^2\sqrt{n\epsilon_{stat} + \lambda 4F^2 d_s}}{\sigma}\|\phi(s,a)\|_{\Sigma_n^{-1}} = Hb_n(s,a).$$

Combine the above two inequalities, we conclude the proof. □

Recall the definition of information gain, $\mathcal{I}_N(\lambda) = \max_{\pi_1,\ldots,\pi_N} \ln\det\left(I + \frac{1}{\lambda}\sum_{n=1}^{N}\Sigma_{\pi_n}\right)$.

Lemma 12. *For any sequence of policies $\pi_1, \ldots, \pi_n$, with $\Sigma_n = \sum_{i=1}^n \Sigma_{\pi_n}$ and $b_n(s,a) \leq c\sqrt{\phi(s,a)^\top \Sigma_n^{-1} \phi(s,a)}$, we have that:*

$$\sum_{n=1}^{N-1} \mathbb{E}_{s,a\sim d^{\pi_{n+1}}} b_n(s,a) \leq c\sqrt{2N\mathcal{I}_N(\lambda)}.$$

When $\phi \in \mathbb{R}^d$, we have:

$$\sum_{n=1}^{N-1} \mathbb{E}_{s,a\sim d^{\pi_{n+1}}} b_n(s,a) \leq c\sqrt{2Nd\ln\left(1 + N/\lambda\right)}.$$

Proof. Starting from the definition of b_n, we have:

$$\sum_{n=1}^{N-1} \mathbb{E}_{s,a\sim d^{\pi_{n+1}}} b_n(s,a) \leq c\sum_{n=1}^{N-1} \mathbb{E}_{s,a\sim d^{\pi_{n+1}}} \sqrt{\phi(s,a)^\top \Sigma_n^{-1} \phi(s,a)} \leq c\sqrt{N}\sqrt{\sum_{n=1}^{N-1} \mathbb{E}_{s,a\sim d^{\pi_{n+1}}} \phi(s,a)^\top \Sigma_n^{-1}\phi(s,a)}$$

$$= c\sqrt{N}\sqrt{\sum_{n=1}^{N-1} \operatorname{tr}\left(\Sigma_{\pi_{n+1}}\Sigma_n^{-1}\right)} \leq c\sqrt{2N\ln\left(\det(\Sigma_N)/\det(\lambda I)\right)}$$

Here in second inequality we use Cauchy-Schwartz inequality, in the third inequality, we use Lemma 18.

If $\phi \in \mathbb{R}^d$, we have that:

$$\sqrt{N\ln(\det(\Sigma_N)/\det(\lambda I))} \leq \sqrt{Nd\ln(1+N/\lambda)},$$

where we use $\|\phi(s,a)\|_2 \leq 1$.

This concludes the proof. □

A.4. Concluding the Sample Complexity Calculation

Before concluding the final sample complexity, we need to link $\widehat{\Sigma}_n$ to Σ_n, as our reward bonus in the algorithm is defined in terms of the empirical estimate $\widehat{\Sigma}_n$.

Lemma 13 (Relating $\widehat{b}_n$ and b_n)**.** *With probability at least $1-\delta$, for all $n \in [1, \ldots, N]$, we have:*

$$b_n(s,a) \leq \widehat{b}_n(s,a) \leq 4b_n(s,a), \forall s, a.$$

Proof. The proof uses Lemma 20. Under Lemma 20, we have:

$$\min\left\{c\sqrt{\phi(s,a)^\top \widehat{\Sigma}_n^{-1}\phi(s,a)}, H\right\} \leq \min\left\{c\sqrt{2}\sqrt{\phi(s,a)^\top \Sigma_n^{-1}\phi(s,a)}, H\right\}$$

$$\leq \sqrt{2}\min\left\{c\sqrt{\phi(s,a)^\top \Sigma_n^{-1}\phi(s,a)}, H\right\} = \sqrt{2}b_n(s,a).$$

and

$$b_n(s,a)/\sqrt{2} = \min\left\{c\sqrt{\phi(s,a)^\top \Sigma_n^{-1}\phi(s,a)}, H\right\}/\sqrt{2}$$

$$\leq \min\left\{c\sqrt{(1/2)\phi(s,a)^\top \Sigma_n^{-1}\phi(s,a)}, H\right\} \leq \min\left\{c\sqrt{\phi(s,a)^\top \widehat{\Sigma}_n^{-1}\phi(s,a)}, H\right\},$$

Note that $\widehat{b}_n(s,a) = 2\min\{c\sqrt{(1/2)\phi(s,a)^\top \widehat{\Sigma}_n^{-1}\phi(s,a)}, H\}$, we have that:

$$b_n(s,a)/\sqrt{2} \le \widehat{b}_n(s,a)/2 \le \sqrt{2}b_n(s,a),$$

which concludes the proof. □

In high level, from Lemma 10, Lemma 11, and Lemma 12 we know that after N iterations, we have:

$$\max_{i\in[1,\dots,N]} J(\pi_i; r, P^\star) \ge J(\pi^\star; r, P^\star) - \frac{10H^2c}{\sqrt{N}}\sqrt{d\ln(1+N/\lambda)}.$$

Hence, to ensure we get an ϵ near optimal policy, we just need to set N large enough such that $\frac{10H^2c}{\sqrt{N}}\sqrt{d\ln(1+N/\lambda)\ln} \approx \epsilon$ and to do so, we need to control M in order to make c scale as a constant.

A.4.1. Concluding for Linear MDPs

Theorem 14 (Sample Complexity for Linear MDPs). *Set $\delta \in (0, 0.5)$ and $\epsilon \in (0,1)$. There exists a set of hyper-parameters,*

$$N = \frac{80H^6d^2}{\epsilon^2}\ln\left(\frac{40H^6d^2}{\epsilon^2}\right), \quad M = 2N\ln\left(|\mathcal{P}|N/\delta\right), \quad c = H\sqrt{d+1}, \quad K = 32N^2\ln\left(8Nd/\delta\right),$$

such that with probability at least $1-2\delta$, PC-MLP returns a policy $\widehat{\pi}$ such that:

$$J(\hat{\pi}; r, P^\star) \ge \max_{\pi\in\Pi} J(\pi; r, P^\star) - \epsilon,$$

with number of samples

$$O\left(\frac{H^{18}d^6}{\epsilon^6}\cdot\ln\left(\frac{|\mathcal{P}|H^6d^2}{\epsilon^2\delta}\ln\left(\frac{H^6d^2}{\epsilon^2}\right)\right)\ln^3\left(\frac{H^6d^2}{\epsilon^2}\right)\right).$$

Ignoring log terms, we get the sample complexity scales in the order of $\widetilde{O}\left(\frac{H^{18}d^6}{\epsilon^6}\right)$.

The above theorem verifies Theorem 4

Proof. From Lemma 8, we know that:

$$c = H\sqrt{d + N\epsilon_{stat}},$$

where we have set $\lambda = 1$ explicitly. Also from Lemma 5, we know that with probability at least $1-\delta$,

$$\epsilon_{stat} = \frac{2\ln(|\mathcal{P}|N/\delta)}{M}.$$

We set M large enough such that $N\epsilon_{stat} = 1$. To achieve this, it is easy to verify that it is enough to set $M = 2N\ln(|\mathcal{P}|N/\delta)$. As $d \ge 1$, we immediately have that $c \le 2H\sqrt{d}$ in this case.

To achieve ϵ approximation error, we set N big enough such that:

$$\frac{10H^2c}{\sqrt{N}}\sqrt{d\ln(1+N/\lambda)\ln} \le \epsilon.$$

Using $c = 2H\sqrt{d}$, we get:

$$\frac{20H^3d}{\sqrt{N}}\sqrt{\ln(1+N)} \le \epsilon.$$

We can verify that the above inequality holds when we set:

$$N = \frac{80H^6 d^2}{\epsilon^2} \ln\left(\frac{40H^6 d^2}{\epsilon^2}\right).$$

Hence, the total number of samples we use for estimating models during N epsilons is bounded as:

$$N \times M = 2N^2 \ln(|\mathcal{P}|N/\delta) \leq \frac{H^{12}d^4}{\epsilon^4} \cdot 12800 \ln^2\left(\frac{40H^6d^2}{\epsilon^2}\right) \ln\left(\frac{80H^6d^2|\mathcal{P}|}{\epsilon^2\delta} \ln\left(\frac{40H^6d^2}{\epsilon^2}\right)\right).$$

We also need to count the total number of samples used to estimate the covariance matrix $\widehat{\Sigma}_n$ for all n. From Lemma 20. The number is bounded as:

$$K \cdot N = 32N^3 \ln\left(8Nd/\delta\right) = \frac{H^{18}d^6}{\epsilon^6} \cdot (32 \times 80^3) \ln^3\left(\frac{40H^6d^2}{\epsilon^2}\right) \ln\left(\frac{640H^6d^3}{\epsilon^2\delta} \ln\left(\frac{40H^6d^2}{\epsilon^2}\right)\right).$$

The total number of samples are $N \times M + N \times K$, which after rearranging terms, we get:

$$NM + NK \leq \frac{H^{18}d^6}{\epsilon^6} \cdot \ln\left(\frac{640H^6d^2|\mathcal{P}|}{\epsilon^2\delta} \ln\left(\frac{40H^6d^2}{\epsilon^2}\right)\right) \ln^3\left(\frac{40H^6d^2}{\epsilon^2}\right) \cdot (60 \times 80^3).$$

We conclude here by noting that the total failure probability is at most 2δ. □

A.4.2. Concluding for KNRs

Theorem 15 (Sample Complexity for KNRs)**.** *Set* $\delta \in (0, 0.5)$ *and* $\epsilon \in (0, 1)$. *There exists a set of hyper-parameters,*

$$N = \Theta\left(\frac{H^6F^2d_sd}{\sigma^2\epsilon^2} \ln\left(\frac{H^6F^2d_sd}{\sigma^2\epsilon^2}\right)\right), \quad M = \Theta\left(N^2(F^2 + FB)^2 \ln^2\left(\frac{N}{\delta}\right)\right),$$
$$c = \Theta\left(\frac{H}{\sigma}\sqrt{d+1}\right), \quad K = 32N^2 \ln\left(8Nd/\delta\right),$$

such that with probability at least $1 - 2\delta$, *PC-MLP returns a policy* $\widehat{\pi}$ *such that:*

$$J(\hat{\pi}; r, P^\star) \geq \max_{\pi \in \Pi} J(\pi; r, P^\star) - \epsilon,$$

with number of samples

$$O\left(\frac{H^{18}d^3d_s^3}{\sigma^6\epsilon^6} \cdot \left(F^{10} + \sigma^2F^8d_s\right)\nu\right),$$

where ν *only contains log terms,*

$$\nu = \ln^2\left(\frac{N}{\delta}\right) \ln\left(\frac{2d_sN}{\delta}\right) + \ln\left(\left(8F^4 + 9\sigma^2F^2d_s \ln\left(\frac{2d_sN}{\delta}\right)\right) \ln^2\left(\frac{N}{\delta}\right) N^2 + 18\sigma^2F^2d_s\right)$$
$$+ \ln^3\left(\frac{6400H^6F^2d_sd}{\sigma^2\epsilon^2}\right) \ln\left(Nd/\delta\right).$$

Ignoring log terms, we get sample complexity scales in the order $\widetilde{O}\left(\frac{H^{18}d^3d_s^3}{\sigma^6\epsilon^6} \cdot \left(F^{10} + \sigma^2F^8d_s\right)\right)$.

The above theorem verifies Theorem 3

Proof. The proof is similar to the proof of Theorem 14. From Lemma 9, we know that $c \leq \frac{4H}{\sigma}\sqrt{\lambda F^2d_s/\sigma^2 + N\epsilon_{stat}/\sigma^2}$. Also from Lemma 7, we know that

$$\epsilon_{stat} = \frac{3(F^2 + FB)\ln(N/\delta)}{\sqrt{M}},$$

with $B = F + \sigma\sqrt{d_s \ln(2d_s NM/\delta)}$, for all n. We set M such that $N\epsilon_{stat} = 1$. This gives us that:

$$M \geq 9N^2(F^2 + FB)^2 \ln^2\left(\frac{N}{\delta}\right).$$

Solve for M, we can verify that it suffices to set M as:

$$\begin{aligned} M = &\left(8F^4 + 9\sigma^2 F^2 d_s \ln\left(\frac{2d_s N}{\delta}\right)\right)\ln^2\left(\frac{N}{\delta}\right)N^2 \\ &+ 18\sigma^2 F^2 d_s \ln\left(\left(8F^4 + 9\sigma^2 F^2 d_s \ln\left(\frac{2d_s N}{\delta}\right)\right)\ln^2\left(\frac{N}{\delta}\right)N^2 + 18\sigma^2 F^2 d_s\right) \end{aligned}$$

This gives us $c = \frac{8H}{\sigma}\sqrt{F^2 d_s}$.

Similarly, we will set N such that $\frac{10H^2 c}{\sqrt{N}}\sqrt{d\ln(1 + N/\lambda)} \leq \epsilon$. With $c = \frac{8H}{\sigma}\sqrt{F^2 d_s}$, we get:

$$\frac{80H^3\sqrt{F^2 d_s}}{\sigma\sqrt{N}}\sqrt{d\ln(1 + N/\lambda)} \leq \epsilon$$

We can verify that it suffices to set N as:

$$N = \frac{12800H^6 F^2 d_s d}{\sigma^2\epsilon^2}\ln\left(\frac{6400H^6 F^2 d_s d}{\sigma^2\epsilon^2}\right)$$

Thus the total number of samples used for model learning is upper bounded as:

$$\begin{aligned} NM &= O\left(N^3\left(F^4 + \sigma^2 F^2 d_s\right)\ln^2\left(\frac{N}{\delta}\right)\ln\left(\frac{2d_s N}{\delta}\right)\right) \\ &\quad + O\left(N\sigma^2 F^2 d_s \cdot \ln\left(\left(8F^4 + 9\sigma^2 F^2 d_s \ln\left(\frac{2d_s N}{\delta}\right)\right)\ln^2\left(\frac{N}{\delta}\right)N^2 + 18\sigma^2 F^2 d_s\right)\right) \\ &= O\left(\nu\frac{H^{18} d^3 d_s^3 F^6 (F^4 + \sigma^2 F^2 d_s)}{\sigma^6\epsilon^6}\right), \end{aligned}$$

where v only contains log terms, i.e.,

$$\nu = \ln^2\left(\frac{N}{\delta}\right)\ln\left(\frac{2d_s N}{\delta}\right) + \ln\left(\left(8F^4 + 9\sigma^2 F^2 d_s \ln\left(\frac{2d_s N}{\delta}\right)\right)\ln^2\left(\frac{N}{\delta}\right)N^2 + 18\sigma^2 F^2 d_s\right).$$

We also need to count the total number of samples used to estimate the covariance matrix $\widehat{\Sigma}_n$ for all n. From Lemma 20. The number is bounded as:

$$K \cdot N = O\left(N^3 \ln\left(Nd/\delta\right)\right) = O\left(\nu_1 \cdot \frac{H^{18} F^6 d_s^3 d^3}{\sigma^6\epsilon^6}\right)$$

where ν_1 only contains log terms, i.e.,

$$\nu_1 = \ln^3\left(\frac{6400H^6 F^2 d_s d}{\sigma^2\epsilon^2}\right)\ln\left(Nd/\delta\right).$$

Combine the two terms, we can conclude that:

$$KN + KM = O\left(\frac{H^{18} d^3 d_s^3}{\sigma^6\epsilon^6} \cdot \left(F^{10} + \sigma^2 F^8 d_s\right)(\nu + \nu_1)\right).$$

□

B. Auxiliary Lemmas

Lemma 16 (Dimension-free SGD (Lemma G.1 from (Agarwal et al., 2020a)))**.** *Consider the following learning process. Initialize* $W_1 = \mathbf{0}$*. For* $i = 1,\dots,N$*, draw* $x_i, y_i \sim \nu$*,* $\|y_i\|_2 \le B$*,* $\|x_i\| \le 1$*; Set* $W_{i+1} = \prod_{\mathcal{W}:=\{W:\|W\|_2\le F\}}\left(W_i - \eta_i\left(W_i x_i - y_i\right)x_i^\top\right)$ *with* $\eta_i = (F^2)/((F+B)\sqrt{N})$*. Set* $\widehat{W} = \frac{1}{N}\sum_{i=1}^N W_i$*, we have that with probability at least* $1-\delta$*:*

$$\mathbb{E}_{x\sim\nu}\left[\left\|\widehat{W}\cdot x - \mathbb{E}\left[y|x\right]\right\|_2^2\right] \le \mathbb{E}_{x\sim\nu}\left[\left\|W^\star\cdot x - \mathbb{E}\left[y|x\right]\right\|_2^2\right] + \frac{R\sqrt{\ln(1/\delta)}}{\sqrt{N}},$$

with any $W^\star$ *such that* $\|W^\star\|_2 \le F$ *and* $R = 3(F^2 + FB)$*.*

Lemma 17 (Total Variation Distance between Two Gaussians)**.** *Given two Gaussian distributions* $P_1 = \mathcal{N}(\mu_1, \sigma^2 I)$ *and* $P_2 = \mathcal{N}(\mu_2, \sigma^2 I)$*, we have* $\|P_1 - P_2\|_{tv} \le \min\{\frac{1}{\sigma}\|\mu_1 - \mu_2\|_2, 1\}$*.*

The above lemma can be verified using the KL divergence between two Gaussians and the application of Pinsker's inequality (Devroye et al., 2018).

Lemma 18. *Consider the following process. For* $n = 1,\dots,N$*,* $M_n = M_{n-1} + \Sigma_n$ *with* $M_0 = \lambda\mathbf{I}$ *and* Σ_n *being PSD matrix with eigenvalues upper bounded by* 1*. We have that:*

$$2\log\det(M_N) - 2\log\det(\lambda\mathbf{I}) \ge \sum_{n=1}^{N}\mathrm{tr}\left(\Sigma_i M_{i-1}^{-1}\right).$$

The proof of the above lemma is standard and can be found in Lemma G.2 from (Agarwal et al., 2020a) for instance.

Lemma 19 (Simulation Lemma)**.** *Consider a MDPs* $\mathcal{M}_1 = \{\hat{r}, \widehat{P}\}$ *where* $\hat{r}$ *and* $\widehat{P}$ *represent reward and transition. For any policy* $\pi : \mathcal{S}\times\mathcal{A} \mapsto \Delta(\mathcal{A})$*, we have:*

$$J(\pi;\hat{r},\widehat{P}) - J(\pi;r,P^\star) = \sum_{h=0}^{H-1}\mathbb{E}_{s,a\sim d_h^\pi}\left[r'(s,a) - r(s,a) + \mathbb{E}_{s'\sim\widehat{P}(\cdot|s,a)}\widehat{V}_h^\pi(s') - \mathbb{E}_{s'\sim P^\star(\cdot|s,a)}\widehat{V}_h^\pi(s')\right].$$

Simulation lemma is widely used in proving sample complexity for RL algorithms. For proof, see Lemma 10 in (Sun et al., 2019) for instance.

The following lemma studies the concentration related to the empirical covariance matrix $\widehat{\Sigma}_n$ and Σ_n.

Lemma 20 (Concentration with the Inverse of Covariance Matrix (Lemma G.4 from (Agarwal et al., 2020a)))**.** *Consider a fixed* N*. Assume* $\phi \in \mathbb{R}^d$*. Given* N *distributions* $\nu_1,\dots,\nu_N$ *with* $\nu_i \in \Delta(\mathcal{S}\times\mathcal{A})$*, assume we draw* K *i.i.d samples from* ν_i *and form* $\widehat{\Sigma}^i = \sum_{j=1}^K \phi_j\phi_j^\top/K$ *for all i. Denote* $\Sigma_n = \sum_{i=1}^n \mathbb{E}_{(s,a)\sim\nu_i}\phi(s,a)\phi(s,a)^\top + \lambda I$ *and* $\widehat{\Sigma}_n = \sum_{i=1}^n \widehat{\Sigma}^i + \lambda I$ *with* $\lambda \in (0,1]$*. Setting* $K = 32N^2\log\left(8Nd/\delta\right)/\lambda^2$*, with probability at least* $1-\delta$*, we have that for any* $n \in [1,\dots,N]$*,*

$$\frac{1}{2}x^T\left(\Sigma_n\right)^{-1}x \le x^T\left(\widehat{\Sigma}_n\right)^{-1}x \le 2x^T\left(\Sigma_n\right)^{-1}x,$$

for all x *with* $\|x\|_2 \le 1$*.*

C. Additional Experiments

C.1. MPPI vs TRPO

In this section we compare two of our practical implementation of Deep PC-MLP: a) using MPPI as the planner and use random RFF feature as ϕ, b) using TRPO as the planner and use the fully connected layer of a random network as ϕ. We plot the learning curves of the two in Fig. 5. The settings of the experiments follow Sec. 7.1. We observe that both implementations achieve the optimal performance while all the other baselines completely fail. In terms of stability, TRPO outperforms MPPI since the performance of MPPI still oscillates before fully converges.

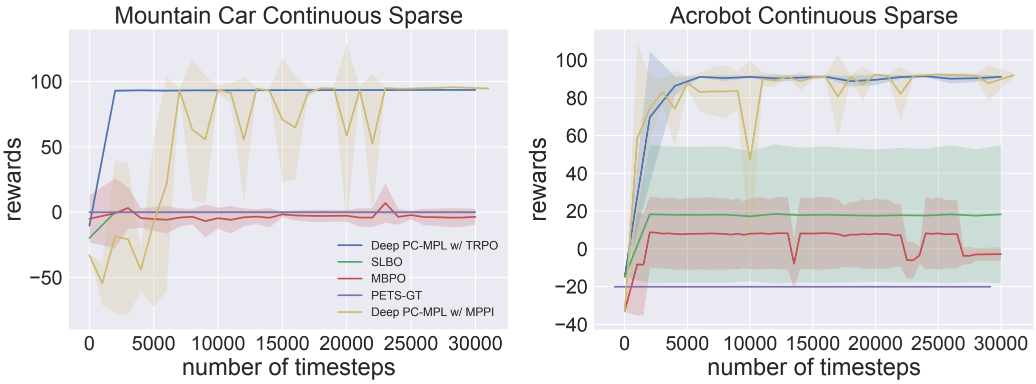

Figure 5. Performance comparison between our two practical implementations.

C.2. Computation Efficiency of PC-MLP

In this section we investigate the wall-clock running time of Deep PC-MLP comparing with other baselines. We test on the running time of each algorithm on the MountainCar environment. We summarize the wall-clock running time in table 3. We see that our practical implementation can run as fast as other baselines. Taking the computation of exploration bonus into consideration, the results show that our algorithm is indeed computationally efficient.

	Time
PC-MLP	148m43s
PETS-GT	132m50s
MBPO	153m30s

Table 3. Wall-clock running time comparison

D. Experimental Details

D.1. MPPI Pseudocode

Here we present the pseudocode for MPPI in Alg. 3:

D.2. Hyperparameters of Deep PC-MLP

D.2.1. MountainCar and Acrobot

We provide the hyperparameters we considered and finally adopted for MountainCar and Acrobot environments in Table 4 (using TRPO as planner) and Table 5 (using MPPI as planner).

D.2.2. Hand Egg and Dense Reward Environments

We provide the hyperparameters we considered and finally adopted for Hand Egg and dense reward environments in Table 6.

Algorithm 3 MPPI

```
Require: Learned dynamics $\hat{P}$, reward $r$, number of samples $K$, shooting horizon $H$, noise convariance $\Sigma$, temperature $\lambda$, initial state $s_0$
if First time planning then
    Initialize $\{a_0, a_1, \dots, a_{T-1}\}$
end if
for $k = 1, \dots, K$ do
    Sample $\mathcal{E}^k = \{\varepsilon^k_0, \varepsilon^k_1, \dots, \varepsilon^k_{T-1}\}$
    for $t = 1, \dots, T$ do
        $S(\mathcal{E}^k) \mathrel{+}= -r(s_{t-1}, a_{t-1} + \varepsilon^k_{t-1}) + \lambda a^T_{t-1}\Sigma^{-1}\epsilon^k_{t-1}$
        $s_t = \hat{P}(s_{t-1}, a_{t-1} + \varepsilon^k_{t-1})$
    end for
    $\beta = \min_k S(\mathcal{E}^k)$
    $\eta = \sum_{k=0}^{K-1} \exp\left(-\frac{1}{\lambda}\left(S(\mathcal{E}^k) - \beta\right)\right)$
    for $k = 1, \dots, K$ do
        $w(\mathcal{E}^k) = \frac{1}{\eta}\exp\left(-\frac{1}{\lambda}\left(S(\mathcal{E}^k) - \beta\right)\right)$
    end for
    for $t = 1, \dots, T$ do
        $a_t \mathrel{+}= \sum_{k=1}^{K} w(\mathcal{E}^k)\epsilon^k_t$
    end for
end for
$a = a_0$
for $t = 1, \dots, T-1$ do
    $a_{t-1} = a_t$
end for
Initialize $a_{T-1}$
```

	Value Considered	Final Value
Model Learning Rate	{1e-3, 5e-3, 1e-4}	1e-3
Dynamics Model Hidden Layer Size	{[500,500]}	[500,500]
Policy Learning Rate	{3e-4}	3e-4
Policy Hidden Layer Size	{[32,32]}	[32,32]
Number of Model Updates	{100}	100
Number of Policy Updates	{40}	40
Number of Iterations	{30,60}	15
Multistep Loss L	{2}	2
Sample Size per Iteration (K)	{1000,2000,3000}	2000
Covariance Regulerazation Coefficient (λ)	{1,0.1,0.01,0.001}	0.01
Replay Buffer Size	{30000}	30000
Bonus Scale (C)	{1,2,5}	5

Table 4. Hyperparameter for MountainCarContinuous environment using TRPO as planner.

	Value Considered	Final Value
Learning Rate	$\{$1e-3, 5e-3, 1e-4$\}$	5e-3
Hidden Layer Size	$\{64\}$	64
Number of Iterations	$\{30,60\}$	30
Sample Size per Iteration (K)	$\{1000,2000,3000\}$	1000
Covariance Regulerazation Coefficient (λ)	$\{1,0.1,0.01,0.001\}$	0.01
Replay Buffer Size	$\{10000\}$	10000
Bonus Scale (C)	$\{1,2\}$	1
Dimention of RFF Feature ($\lvert\phi\rvert$)	$\{10,15,20,30\}$	20
MPPI Sampling Size (K)	$\{15,100,200,300\}$	200
MPPI Shooting Horizon (H)	$\{15,30,45,60\}$	30
MPPI Temperature (λ)	$\{1,0.2,0.1,0.001\}$	0.2
MPPI Noise Covariance (Σ)	$\{0.2,0.3,0.5,1\}I$	$0.3I$

Table 5. Hyperparameter for MountainCarContinuous environment using MPPI as planner.

	Value Considered	Final Value
Model Learning Rate	$\{$1e-3, 5e-3, 1e-4$\}$	1e-3
Dynamics Model Hidden Layer Size	$\{[500,500]\}$	[500,500]
Policy Learning Rate	$\{$3e-4$\}$	3e-4
Policy Hidden Layer Size	$\{[32,32]\}$	[32,32]
Number of Model Updates	$\{100\}$	100
Number of Policy Updates	$\{40\}$	40
Number of Iterations	$\{200,100,50\}$	50
Multistep Loss L	$\{2\}$	2
Sample Size per Iteration (K)	$\{1000,2000,4000\}$	4000
Covariance Regulerazation Coefficient (λ)	$\{1,0.1,0.01,0.001\}$	0.01
Replay Buffer Size	$\{100000\}$	100000
Bonus Scale (C)	$\{0.1,1\}$	0.1

Table 6. Hyperparameter for HandEgg and dense reward environment.