A Free Lunch from the Noise: Provable and Practical Exploration for Representation Learning (Supplementary Materials)

Tongzheng Ren [1, 2, *] **Tianjun Zhang** [3, *] **Csaba Szepesvári** [4, 5] **Bo Dai** [2]

[1]Department of Computer Science, UT Austin
[2]Google Research, Brain Team
[3] Department of EECS, UC Berkeley
[4]DeepMind
[5]Department of Computer Science, University of Alberta

A BACKGROUNDS ON REPRODUCING KERNEL HILBERT SPACE

We briefly introduce the basic concepts of the Reproducing Kernel Hilbert Space, which is helpful on understanding our paper. To start with, we first define the inner product.

Definition 1 (Inner Product)**.** *A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is said to be an inner product on $\mathcal{H}$ if it satisfies the following conditions:*

1. *Positive Definiteness: $\forall u \in \mathcal{H}$, $\langle u, u \rangle \geq 0$, and $\langle u, u \rangle = 0 \iff u = 0$.*
2. *Symmetry: $\forall u, v \in \mathcal{H}$, $\langle u, v \rangle \in \langle v, u \rangle$.*
3. *Bilinearity: $\forall \alpha, \beta \in \mathbb{R}, u, v, w \in \mathcal{H}$, $\langle \alpha u + \beta v, w \rangle = \alpha \langle u, w \rangle + \beta \langle v, w \rangle$.*

Additionally, we can define a norm with the inner product: $\|u\| = \sqrt{\langle u, u \rangle}$.

A Hilbert space is a space equipped with an inner product and satisfies an additional technical condition of completeness. The finite-dimension vector space with the canonical inner product is an example of the Hilbert space. We remark that $\mathcal{H}$ can also be a function space, for example, the space contains all square integrable functions (i.e. $\int_{\mathbb{R}} f(x)^2 \, \mathrm{d}x < \infty$, generally denoted as L_2) is also a Hilbert space with inner product $\langle f, g \rangle = \int_{\mathbb{R}} f(x) g(x) \, \mathrm{d}x$.

We then define the kernel, and introduce the notion of positive-definite kernel [**?**].

Definition 2 ((Positive-Definite) Kernel)**.** *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a kernel on non-empty set $\mathcal{X}$ if there exists a Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$, we have*

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}.$$

Moreover, the kernel is said to be positive definite if $\forall n \geq 1$, $\forall \{a_i\}_{i \in [n]} \subset \mathbb{R}$ and mutually distinct set $\{x_i\}_{i \in [n]} \subset \mathcal{X}$, we have that

$$\sum_{i \in [n]} \sum_{j \in [n]} a_i a_j k(x_i, x_j) > 0.$$

Some well-known kernels include:

- Linear Kernel: $k(x, y) = \langle x, y \rangle$, with the canonical feature map $\phi(x) = x$.
- Polynomial Kernel: $k(x, y) = (\langle x, y \rangle + c)^m$, where $m \in \mathbb{N}^+$ and $c \in \mathbb{R}^+$.

[*] Equal Contribution

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

- Gaussian (a.k.a radial basis function, RBF) Kernel: $k(x, y) = \exp\left(\frac{\|x-y\|_2^2}{2\sigma^2}\right)$. It's known that such kernel is positive definite.

Now we can define the Reproducing Kernel Hilbert space (RKHS) [**?**].

Definition 3 (Reproducing Kernel Hilbert Space (RKHS))**.** *The Hilbert space $\mathcal{H}$ of $\mathbb{R}$-valued function defined on a non-emptry set $\mathcal{X}$ is said to be a reproducing kernel Hilbert space (RKHS) is there is a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, such that*

1. *$\forall x \in \mathcal{X}$, $k(x, \cdot) \in \mathcal{H}$.*
2. *$\forall x \in \mathcal{X}, f \in \mathcal{H}$, $\langle f, k(x, \cdot)\rangle_{\mathcal{H}} = f(x)$ (a.k.a the reproducing property), which also implies that $\langle k(x, \cdot), k(y, \cdot)\rangle = k(x, y)$.*

Here k is called a reproducing kernel of $\mathcal{H}$.

We provide an intuitive interpretation on the definition of RKHS when $\mathcal{H}$ is the space of linear function. Consider $\mathcal{X} = \mathbb{R}^d$ and $k(x, y) = \langle x, y\rangle$. With the definition of the kernel k, we can see that $k(x, \cdot) : \mathcal{X} \to \mathbb{R}$ is a linear function, and thus lies in $\mathcal{H}$. Meanwhile, $\forall f \in \mathcal{H}$, there exists θ_f such that $f(x) = \theta_f^\top x$. We define the inner product on $\mathcal{H}$ via $\langle f, g\rangle_{\mathcal{H}} = \langle \theta_f, \theta_g\rangle$, and thus $\langle f(, k(x, \cdot))\rangle_{\mathcal{H}} = \theta^\top x = f(x)$, which demonstrates the reproducing property, and shows that the space of linear function on any finite-dimensional vector space is an RKHS with linear kernel as the corresponding reproducing kernel.

We state the following theorems without the proof.

Theorem 4 (Moore-Aronszajn [**?**])**.** *Every positive definite kernel k is associated with a unique RKHS $\mathcal{H}$.*

Notice that, Moore-Aronszajn theorem guarantees that all of the positive kernel can be represented as the inner product in certain Hilbert space, hence we can have a linear representation of the Gaussian distribution induced by the reproducing property of Gaussian kernel, as we illustrated in the main text.

Theorem 5 (Bochner [**?**])**.** *A continuous, shift-invariant kernel (i.e. $k(x, y) = k(x - y)$) is positive definite if and only if $k(x - y)$ is the Fourier transform of a non-negative measure ω, i.e.*

$$k(x - y) = \int_{\mathbb{R}^d} \exp(i\omega^\top (x - y))\, \mathrm{d}\mathbb{P}(\omega) = \int_{\mathbb{R}^d \times [0,2\pi]} 2\cos(\omega^\top x + b)\cos(\omega^\top y + b)\, \mathrm{d}(\mathbb{P}(\omega) \times \mathbb{P}(b)),$$

where $\mathbb{P}(b)$ is a uniform distribution on $[0, 2\pi]$.

Bochner's theorem shows that any continuous positive definite shift-invariant kernel (e.g. Gaussian kernel, Laplacian kernel) can be represented as the inner product of random Fourier feature, which provides an additional way to provide a representation for certain distribution [see Rahimi and Recht, 2007, Dai et al., 2014].

B AN EQUIVALENT UPPER CONFIDENCE BOUND ALGORITHM

In this section, we provide a generic Upper Confidence Bound (UCB) algorithm with the OFU principle, and show the connections and differences between the UCB algorithm and the TS algorithm. The prototype for our UCB algorithm is illustrated in Algorithm 1.

Notice that, the only difference between UCB algorithm and TS algorithm is the mechanism of finding f we use to plan for each episode (highlighted in blue). For UCB algorithm, we perform an optimistic planning, which finds the $\tilde{f}_k$ that potentially has the largest cumulative reward. However, such constrained optimization problem is NP-hard even for the simplest linear bandits [Dani et al., 2008]. Instead, for TS algorithm, we only sample the f_k from the posterior distribution, which gets rid of the complicated constraint optimization. We are interested in the UCB algorithm, as the worst case regret bound of the UCB algorithm can be directly translated to the expected regret bound of the TS algorithm without the need of explicit manipulation of the prior and the posterior[Russo and Van Roy, 2013, 2014, Osband and Van Roy, 2014].

Confidence Set Construction Perhaps the most important part in OFU-style algorithm is the construction of confidence set $\mathcal{F}_k$. To enable sample-efficient learning, the confidence set should

1. contain f^* with high probability, so that we can identify f^* eventually;

Algorithm 1 Upper Confidence Bound (UCB) Algorithm

Require: Number of Episodes K, Failure Probability $\delta \in (0,1)$, Reward Function $r(s,a)$.

Initialize the history set $\mathcal{H}_0 = \emptyset$.

for episodes $k = 1, 2, \cdots$ **do**

  Compute π_k via ▷ Optimistic Planning.

$$(\pi_k, \tilde{f}_k) = \arg\max_{\pi \in \Pi, \tilde{f} \in \mathcal{F}_k} \tilde{V}_0^{\pi}(s_0).$$

  where $\mathcal{F}_k$ is defined in (2).

  for steps $h = 0, 1, \cdots, H-1$ **do** ▷ Execute π_k.

    Execute $a_h^k \sim \pi_k^h(s_h^k)$.

    Observe s_{h+1}.

  end for

  Set $\mathcal{H}_k = \mathcal{H}_{k-1} \cup \{(s_h^k, a_h^k, s_{h+1}^k)\}_{h=0}^{H-1}$. ▷ Update the History.

end for

2. shrink as fast as possible, so that we can identify f^* efficiently.

In the tabular setting, $\mathcal{F}_k$ is constructed via the concentration of sub-Gaussian/sub-Gamma random variable [e.g. Azar et al., 2017], and in the linear MDP setting, $\mathcal{F}_k$ is constructed via the concentration on the linear parameters. As we don't assume any specific structures, we instead constructed $\mathcal{F}_k$ via the concentration on the ℓ_2 error, following the idea of [Russo and Van Roy, 2013, Osband and Van Roy, 2014]. Specifically, consider the least-square estimates defined by

$$\hat{f}_K = \arg\min_{f \in \mathcal{F}} L_{2,K}(f) := \sum_{k \in [K]} \sum_{h=0}^{H-1} \|f(s_h^k, a_h^k) - s_{h+1}^k\|_2^2. \tag{1}$$

As $s_{h+1}^k = f^*(s_h^k, a_h^k) + \epsilon_h^k$ where ϵ_h^k is the Gaussian noise added to the step h at the k-th episode, we know $\hat{f}_K$ will not deviate from f^* a lot. Meanwhile, as K increases, the estimation $\hat{f}_K$ should become closer to f^*. Specifically, define the empirical 2-norm $\|\cdot\|_{2,E_t}$ as

$$\|g\|_{2,E_K}^2 := \sum_{k \in [K]} \sum_{h=0}^{H-1} \|g(s_h^k, a_h^k)\|_2^2.$$

We can construct the confidence set based on the following lemma:

Lemma 6 (Confidence Set Construction [Russo and Van Roy, 2013, Osband and Van Roy, 2014]). *Define*

$$\mathcal{F}_K = \left\{ f \in \mathcal{F} : \|f - \hat{f}_K\|_{2,E_K} \le \sqrt{\beta_K^*(\mathcal{F}, \delta, \alpha)} \right\}, \tag{2}$$

then

$$\mathbb{P}_{f^*}\left(f^* \in \bigcap_{k=1}^{\infty} \mathcal{F}_k \right) \ge 1 - 2\delta, \tag{3}$$

where

$$\beta_K^*(\mathcal{F}, \delta, \alpha) = 8\sigma^2 \log(\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_2)/\delta) + 2H\alpha(12C + \sqrt{8d\sigma^2 \log(4K^2H/\delta)}). \tag{4}$$

The proof can be found in Appendix C.1. Notice that, the empirical 2-norm $\|f - \hat{f}_K\|_{2,E_K}$ scales linearly with K, and $\beta_K^*(\mathcal{F}, \delta, \alpha)$ only scales as $\log K$, so the confidence set shrinks. Meanwhile, Equation 3 guarantees that $f^* \in \mathcal{F}_k, \forall k$ with high probability. Hence, it satisfies our requirement for the confidence set.

Regret Upper Bound We have the following upper bound of the regret for the UCB algorithm:

Theorem 7 (Regret Bound)**.** *Assume Assumption 2 to 5 holds. We have that*

$$\text{Regret}(K) \leq \tilde{O}(\sqrt{H^2 T \cdot \log \mathcal{N}(\mathcal{F}, T^{-1/2}, \|\cdot\|_2) \cdot \dim_E(\mathcal{F}, T^{-1/2})}).$$

where $\tilde{O}$ represents the order up to logarithm factors.

C TECHNICAL PROOF

C.1 PROOF FOR LEMMA 6

Proof. We first show the following concentration on the ℓ_2 error:

Lemma 8 (Concentration of ℓ_2 error [Russo and Van Roy, 2013, Osband and Van Roy, 2014, **?**])**.** $\forall \delta > 0, f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, *we have*

$$\mathbb{P}_{f^*}\left(L_{2,K}(f) \geq L_{2,K}(f^*) + \frac{1}{2}\|f - f^*\|_{2,E_K}^2 - 4\sigma^2 \log(1/\delta), \quad \forall K \in \mathbb{N}\right) \geq 1 - \delta$$

Proof. Define the filtration $\mathcal{H}_{k,h} = \{(s_h^i, a_h^i)\}_{i\in[k-1],h=0,\cdots,H-1} \cup \{(s_i^k, a_i^k)\}_{h=0}^{h-1}$, and the random variable $Z_{k,h}$ adapted to the filtration $\mathcal{H}_{k,h}$ via:

$$\begin{aligned}
Z_{k,h} =& \|f^*(s_h^k, a_h^k) - s_{h+1}^k\|_2^2 - \|f(s_h^k, a_h^k) - s_{h+1}^k\|_2^2 \\
=& \|f^*(s_h^k, a_h^k) - s_{h+1}^k\|_2^2 - \|f(s_h^k, a_h^k) - f^*(s_h^k, a_h^k) + f^*(s_h^k, a_h^k) - s_{h+1}^k\|_2^2 \\
=& -\|f(s_h^k, a_h^k) - f^*(s_h^k, a_h^k)\|_2^2 + 2\langle f(s_h^k, a_h^k) - f^*(s_h^k, a_h^k), \epsilon_h^k\rangle,
\end{aligned}$$

where $\epsilon_h^k = s_{h+1}^k - f^*(s_h^k, a_h^k)$. Thus, $\mathbb{E}(Z_k^h|\mathcal{H}_{k,h}) = -\|f(s_h^k, a_h^k) - f^*(s_h^k, a_h^k)\|_2^2$, and $Z_k^h + \|f(s_h^k, a_h^k) - f^*(s_h^k, a_h^k)\|_2^2$ is a martingale w.r.t $\mathcal{H}_{k,h}$. Notice that we assume ϵ is an isotropic Gaussian noise with variance σ^2 on each of the dimension, thus the conditional moment generating function of $Z_k^h + \|f(s_h^k, a_h^k) - f^*(s_h^k, a_h^k)\|_2^2$ satisfies:

$$\begin{aligned}
M_{k,h}(\lambda) =& \log \mathbb{E}[\exp(\lambda(Z_k^h + \|f(s_h^k, a_h^k) - f^*(s_h^k, a_h^k)\|_2^2))|\mathcal{H}_{k,h}] \\
=& \log \mathbb{E}[\exp(\langle 2\lambda f(s_h^k, a_h^k) - f^*(s_h^k, a_h^k), \epsilon_h^k\rangle)|\mathcal{H}_{k,h}] \\
\leq& 2\sigma^2\lambda^2 \|f(s_h^k) - f^*(s_h^k, a_h^k)\|_2^2.
\end{aligned}$$

Applying Lemma 4 in [Russo and Van Roy, 2013], we have that, $\forall x, \lambda \geq 0$,

$$\mathbb{P}_{f^*}\left(\sum_{k\in[K]}\sum_{h=0}^{H-1} \lambda Z_{k,h} \leq x - \lambda(1 - 2\lambda\sigma^2) \sum_{k\in[K]}\sum_{h=0}^{H-1} \|f(s_h^k, a_h^k) - f^*(s_h^k, a_h^k)\|_2^2, \quad \forall k \in \mathbb{N}\right) \leq 1 - \exp(-x).$$

Take $\lambda = \frac{1}{4\sigma^2}, x = \log 1/\delta$, and notice that $\sum_{k\in[K]}\sum_{h=0}^{H-1} Z_{k,h} = L_{2,K}(f^*) - L_{2,K}(f)$, we have the desired result. □

We construct an α-cover $\mathcal{F}_\alpha$ in $\mathcal{F}$ with respect to $\|\cdot\|_2$. With a standard union bound, we know that condition on f^*, with probability at least $1 - \delta$, we have that

$$L_{2,K}(f^\alpha) - L_{2,K}(f^*) \geq \frac{1}{2}\|f^\alpha - f^*\|_{2,E_K}^2 - 4\sigma^2 \log(|F^\alpha|/\delta), \quad \forall K \in \mathbb{N}, f^\alpha \in \mathcal{F}^\alpha.$$

Thus, we have that

$$\begin{aligned}
L_{2,K}(f) - L_{2,K}(f^*) \geq& \frac{1}{2}\|f - f^*\|_{2,E_K}^2 - 4\sigma^2 \log(|F^\alpha|/\delta) \\
&+ \underbrace{\min_{f^\alpha \in \mathcal{F}^\alpha}\left\{\frac{1}{2}\|f^\alpha - f^*\|_{2,E_K}^2 - \frac{1}{2}\|f - f^*\|_{2,E_K}^2 + L_{2,K}(f) - L_{2,K}(f^\alpha)\right\}}_{\text{Discretization Error}}.
\end{aligned}$$

We then deal with the discretization error. Assume $\alpha \leq 2C$ (or otherwise we only have a trivial cover) and $\|f^\alpha(s,a) - f(s,a)\|_2 \leq \alpha$, we have that

$$\begin{aligned}
&\|f^\alpha(s,a) - f^*(s,a)\|_2^2 - \|f(s,a) - f^*(s,a)\|_2^2 \\
=&\|f^\alpha(s,a)\|_2^2 - \|f(s,a)\|_2^2 + 2\langle f^*(s,a), f(s,a) - f^\alpha(s,a)\rangle \\
\leq& \max_{\|y\|_2 \leq \alpha} \{\|f(s,a) + y\|_2^2 - \|f(s,a)\|_2^2\} + 2C\alpha \\
=& \max_{\|y\|_2 \leq \alpha} \{2\langle f(s,a), y\rangle + \|y\|_2^2\} + 2C\alpha \\
\leq& 4C\alpha + \alpha^2 \leq 6C\alpha,
\end{aligned}$$

where the inequality is by Cauchy-Schwartz inequality and $\alpha \leq 2C$. Meanwhile,

$$\begin{aligned}
&\|s' - f(s,a)\|_2^2 - \|s' - f^\alpha(s,a)\|_2^2 \\
=&2\langle s', f^\alpha(s,a) - f(s,a)\rangle + \|f(s,a)\|_2^2 - \|f^\alpha(s,a)\|_2^2 \\
\leq&2\langle \epsilon, f^\alpha(s,a) - f(s,a)\rangle + 2\langle f^*(s,a), f^\alpha(s,a) - f(s,a)\rangle + 2C\alpha + \alpha^2 \\
\leq&2\|\epsilon\|_2\alpha + 6C\alpha.
\end{aligned}$$

We now consider the concentration property of $\|\epsilon\|_2$. Here we simply follow [**?**] and notice that ϵ is $\sqrt{d}\sigma$-norm-sub-Gaussian, we have that

$$\mathbb{P}(\|\epsilon\|_2 > \sqrt{2d\sigma^2 \log(2/\delta)}) \leq \delta.$$

By a union bound, we have that

$$\mathbb{P}(\exists k, \|\epsilon\|_2 > \sqrt{2d\sigma^2 \log(4k^2H/\delta)}) \leq \frac{\delta}{2}\sum_{k=1}^{\infty}\sum_{h=0}^{H-1}\frac{1}{k^2H} \leq \delta.$$

Sum all these up, we can see with probability $1-\delta$, $\forall K \in \mathbb{N}$, the discretization error is upper bounded by:

$$H\alpha(12C + \sqrt{8d\sigma^2 \log(4K^2H/\delta)}).$$

As we consider the least square estimate $\hat{f}_K$, we have that $L_{2,K}(\hat{f}_K) - L_{2,K}(f^*) \leq 0$. Substitute back, we have the desired results. □

C.2 SIMULATION LEMMA

Lemma 9 (Simulation Lemma (adapted from Lemma 3.9 in [Kakade et al., 2020]))**.** *Given $\hat{f}$, $\forall s \in \mathcal{S}$, the value function $\hat{V}^\pi$ and V^π corresponding to the model $\hat{f}$ and f^* satisfies*

$$\hat{V}_0^\pi(s) - V_0^\pi(s) \leq H^{3/2}\sqrt{\mathbb{E}\left[\sum_{h=0}^{H-1}\min\left\{\frac{2\|f^*(s_h,a_h) - \hat{f}(s_h,a_h)\|_2^2}{\sigma^2}, 1\right\}\right]}.$$

Proof. We first show the following difference lemma:

Lemma 10 (Difference Lemma)**.** *Assume the trajectory $\{(s_h,a_h)\}_{h=0}^{H-1}$ is generated via policy π and ground truth f^*, define*

$$V_h = \sum_{\tau=h}^{H-1} r(s_\tau, a_\tau)$$

then $\forall \tau \in \{1, \cdots, H-1\}$, we have:

$$\begin{aligned}
\hat{V}_0^\pi(s_0) - V_0 =& \mathbb{E}_{s'_\tau \sim \mathcal{N}(\hat{f}(s_{\tau-1},a_{\tau-1}),\sigma^2 I)}\left[\hat{V}_\tau^\pi(s'_\tau)\right] - V_\tau \\
&+ \sum_{h=1}^{\tau-1}\left[\mathbb{E}_{s'_h \sim \mathcal{N}(f(s_{h-1},a_{h-1}),\sigma^2 I)}\left[\hat{V}_h^\pi(s'_h)\right] - \hat{V}_h^\pi(s_h)\right].
\end{aligned}$$

Proof. When $\tau = 1$, we can obtain the result with $a_0 = \pi(s_0)$ and

$$\hat{V}_0^\pi(s_0) = r(s_0, \pi(s_0)) + \mathbb{E}_{s_1' \sim \mathcal{N}(f(s_0,a_0),\sigma^2 I)} \hat{V}_1^\pi(s_1').$$

We only need to show the case when $\tau = 2$, and the case when $\tau > 2$ can be derived via recursion. Notice that

$$\begin{aligned}
\hat{V}_0^\pi(s_0) - V_0 =& \mathbb{E}_{s_1' \sim \mathcal{N}(f(s_0,a_0),\sigma^2 I)} \left[\hat{V}_1^\pi(s_1')\right] - V_1 \\
=& \hat{V}_1^\pi(s_1) - V_1 + \mathbb{E}_{s_1' \sim \mathcal{N}(f(s_0,a_0),\sigma^2 I)} \left[\hat{V}_1^\pi(s_1')\right] - \hat{V}_1^\pi(s_1) \\
=& \mathbb{E}_{s_2' \sim \mathcal{N}(f(s_1,a_1),\sigma^2 I)} \left[\hat{V}_2^\pi(s_2')\right] - V_2 + \mathbb{E}_{s_1' \sim \mathcal{N}(f(s_0,a_0),\sigma^2 I)} \left[\hat{V}_1^\pi(s_1')\right] - \hat{V}_1^\pi(s_1),
\end{aligned}$$

where the last equality is due to the fact that $a_1 = \pi(s_1)$. □

We then follow the idea of "optional stopping" used in [Kakade et al., 2020] and show the following "optional stopping" simulation lemma.

Lemma 11 ("Optional Stopping" Simulation Lemma)**.** *Consider the stochastic process over the trajectories $\{(s_h, a_h)\}_{h=0}^{H-1}$ generated via policy π and ground truth f^*, where the randomness is from the Gaussian noise in the dynamics. Define a stopping time τ w.r.t this stochastic process and a given model $\hat{f}$ via:*

$$\tau := \min\{h \geq 0 : \hat{V}_h^\pi(s_h) \leq V_h^\pi(s_h)\}.$$

Furthermore, define a random variable:

$$\tilde{V}_h^\pi(s_h) = \max\{\hat{V}_h^\pi(s_h), V_h^\pi(s_h)\},$$

we have that

$$\hat{V}_0^\pi(s_0) - V_0^\pi(s_0) \leq \mathbb{E}\left[\sum_{h=0}^{H-1} \mathbf{1}_{h<\tau} \left(\mathbb{E}_{s_{h+1}' \sim \mathcal{N}(f^*(s_h,a_h),\sigma^2 I)} \tilde{V}_h^\pi(s_{h+1}') - \mathbb{E}_{s_{h+1}' \sim \mathcal{N}(\hat{f}(s_h,a_h),\sigma^2 I)} \tilde{V}_h^\pi(s_{h+1}')\right)\right],$$

where the expectation is w.r.t the stochastic process over the trajectories.

Proof. Define the filtration $\mathcal{F}_h := \{\epsilon_i\}_{i=0}^{h-1}$, where ϵ_i is the noise that add to the dynamics at step i. Define

$$M_h = \mathbb{E}[\hat{V}_0^\pi(s_0) - V_0 | \mathcal{F}_h],$$

which is a Doob martingale with respect to $\mathcal{F}_i$ [**?**]. As $\tau \leq H$, by Doob's optional stopping theorem, we have that

$$\mathbb{E}[\hat{V}_0^\pi(s_0) - V_0] = \mathbb{E}[M_\tau] = \mathbb{E}[\mathbb{E}[\hat{V}_0^\pi(s_0) - V_0 | \mathcal{F}_\tau]].$$

We then provide a bound for M_τ. By Lemma 10, we have that

$$\begin{aligned}
M_\tau =& \mathbb{E}[\hat{V}_0^\pi(s_0) - V_0 | \mathcal{F}_\tau] \\
=& \mathbb{E}_{s_\tau' \sim \mathcal{N}(\hat{f}(s_{\tau-1},a_{\tau-1}),\sigma^2 I)} \left[\hat{V}_\tau^\pi(s_\tau')\right] - V_\tau^\pi(s_\tau) \\
&+ \mathbb{E}_{s_h' \sim \mathcal{N}(\hat{f}(s_{h-1},a_{h-1}),\sigma^2 I)} \left[\hat{V}_h^\pi(s_h')\right] - \sum_{h=1}^{\tau-1} \hat{V}_h^\pi(s_h) \\
=& \sum_{h=1}^{\tau} \left(\mathbb{E}_{s_h' \sim \mathcal{N}(\hat{f}(s_h,a_h),\sigma^2)} \left[\hat{V}_h^\pi(s_h')\right] - \tilde{V}_h^\pi(s_h)\right) \\
\leq& \sum_{h=1}^{\tau} \left(\mathbb{E}_{s_h' \sim \mathcal{N}(\hat{f}(s_h,a_h),\sigma^2 I)} \left[\tilde{V}_h^\pi(s_h)\right] - \tilde{V}_h^\pi(s_h)\right) \\
=& \sum_{h=1}^{H} \mathbf{1}_{h\leq\tau} \left(\mathbb{E}_{s_h' \sim \mathcal{N}(\hat{f}(s_h,a_h),\sigma^2 I)} \left[\tilde{V}_h^\pi(s_h)\right] - \tilde{V}_h^\pi(s_h)\right),
\end{aligned}$$

where the third inequality follows the definition of τ (and thus $V_\tau^\pi(s_\tau) = \tilde{V}_\tau^\pi(s_\tau)$ and $\hat{V}_h^\pi(s_h) = \tilde{V}_h^\pi(s_h)$ for $h < \tau$.)

The proof is then concluded via the following observation:

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{h\le\tau}\left(\mathbb{E}_{s'_h\sim\mathcal{N}(\hat{f}(s_h,a_h),\sigma^2 I)}\left[\tilde{V}_h^\pi(s_h)\right] - \tilde{V}_h^\pi(s_h)\right)\right]\\
=&\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{h\le\tau}\left(\mathbb{E}_{s'_h\sim\mathcal{N}(\hat{f}(s_h,a_h),\sigma^2 I)}\left[\tilde{V}_h^\pi(s_h)\right] - \tilde{V}_h^\pi(s_h)\right)\middle|\mathcal{F}_{h-1}\right]\right]\\
=&\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{h-1<\tau}\left(\mathbb{E}_{s'_h\sim\mathcal{N}(\hat{f}(s_h,a_h),\sigma^2 I)}\left[\tilde{V}_h^\pi(s_h)\right] - \tilde{V}_h^\pi(s_h)\right)\middle|\mathcal{F}_{h-1}\right]\right]\\
=&\mathbb{E}\left[\mathbf{1}_{h-1<\tau}\mathbb{E}\left[\left(\mathbb{E}_{s'_h\sim\mathcal{N}(\hat{f}(s_h,a_h),\sigma^2 I)}\left[\tilde{V}_h^\pi(s_h)\right] - \tilde{V}_h^\pi(s_h)\right)\middle|\mathcal{F}_{h-1}\right]\right]\\
=&\mathbb{E}\left[\mathbf{1}_{h-1<\tau}\left(\mathbb{E}_{s'_h\sim\mathcal{N}(\hat{f}(s_h,a_h),\sigma^2)}\left[\tilde{V}_h^\pi(s_h)\right] - \mathbb{E}_{s'_h\sim\mathcal{N}(f^*(s_h,a_h),\sigma^2)}\left[\tilde{V}_h^\pi(s_h)\right)\right)\right],
\end{aligned}$$

where the third equality is due to the fact that $\mathbf{1}_{h-1<\tau}$ is measurable under $\mathcal{F}_{h-1}$. □

Before we finally provide the proof of Lemma 9, we state the following lemma that bound the expectation under two isotropic Gaussian distribution with different mean:

Lemma 12 (Difference of Expectation under Different Mean Isotropic Gaussian)**.** $\forall$ *(approximately measurable) positive function g, we have that*

$$\mathbb{E}_{z\sim\mathcal{N}(\mu_1,\sigma^2 I)}[g(z)] - \mathbb{E}_{z\sim\mathcal{N}(\mu_2,\sigma^2 I)}[g(z)] \le \min\left\{\frac{\sqrt{2}\|\mu_1-\mu_2\|}{\sigma}, 1\right\}\sqrt{\mathbb{E}_{z\sim\mathcal{N}(\mu_1,\sigma^2 I)}[g(z)^2]}$$

Proof.

$$\begin{aligned}
&\mathbb{E}_{z\sim\mathcal{N}(\mu_1,\sigma^2 I)}[g(z)] - \mathbb{E}_{z\sim\mathcal{N}(\mu_2,\sigma^2 I)}[g(z)]\\
=&\mathbb{E}_{z\sim\mathcal{N}(\mu_1,\sigma^2 I)}\left[g(z)\left(1-\exp\left(\frac{2(\mu_1-\mu_2)^\top z + \|\mu_2\|^2 - \|\mu_1\|^2}{2\sigma^2}\right)\right)\right]\\
\le&\sqrt{\mathbb{E}_{z\sim\mathcal{N}(\mu_1,\sigma^2 I)}[g(z)^2]}\sqrt{\mathbb{E}_{z\sim\mathcal{N}(\mu_1,\sigma^2 I)}\left(1-\exp\left(\frac{2(\mu_2-\mu_1)^\top z - \|\mu_2\|^2 + \|\mu_1\|^2}{2\sigma^2}\right)\right)^2}
\end{aligned}$$

We then calculate

$$\begin{aligned}
&\mathbb{E}_{z\sim\mathcal{N}(\mu_1,\sigma^2 I)}\left(1-\exp\left(\frac{2(\mu_2-\mu_1)^\top z - \|\mu_2\|^2 + \|\mu_1\|^2}{2\sigma^2}\right)\right)^2\\
=&1 - \frac{2}{\sqrt{2\pi}\sigma^{d/2}}\int\exp\left(\frac{-\|z-\mu_1\|_2^2 + 2(\mu_2-\mu_1)^\top z - \|\mu_2\|^2 + \|\mu_1\|^2}{2\sigma^2}\right)dz\\
&+\frac{1}{\sqrt{2\pi}\sigma^{d/2}}\int\exp\left(\frac{-\|z-\mu_1\|_2^2 + 4(\mu_2-\mu_1)^\top z - 2\|\mu_2\|^2 + 2\|\mu_1\|^2}{2\sigma^2}\right)dz\\
=&-1 + \frac{1}{\sqrt{2\pi}\sigma^{d/2}}\int\exp\left(\frac{-\|z-(2\mu_2-\mu_1)\|_2^2 + 2\|\mu_2-\mu_1\|_2^2}{2\sigma^2}\right)dz\\
=&-1+\exp\left(\frac{\|\mu_2-\mu_1\|_2^2}{\sigma^2}\right).
\end{aligned}$$

Also notice that, as g is positive, a simple bound is that

$$\mathbb{E}_{z\sim\mathcal{N}(\mu_1,\sigma^2 I)}[g(z)] - \mathbb{E}_{z\sim\mathcal{N}(\mu_2,\sigma^2 I)}[g(z)] \le \mathbb{E}_{z\sim\mathcal{N}(\mu_1,\sigma^2 I)}[g(z)] \le \sqrt{\mathbb{E}_{z\sim\mathcal{N}(\mu_1,\sigma^2 I)}[g(z)^2]}.$$

Thus,

$$\mathbb{E}_{z\sim\mathcal{N}(\mu_1,\sigma^2 I)}[g(z)] - \mathbb{E}_{z\sim\mathcal{N}(\mu_2,\sigma^2 I)}[g(z)] \le \sqrt{\mathbb{E}_{z\sim\mathcal{N}(\mu_1,\sigma^2 I)}[g(z)^2]}\sqrt{\min\left\{\exp\left(\frac{\|\mu_2-\mu_1\|_2^2}{\sigma^2}\right)-1, 1\right\}}.$$

Notice that, if $\|\mu_2 - \mu_1\| \geq \sigma$, then $\exp\left(\frac{\|\mu_2-\mu_1\|_2^2}{\sigma^2}\right) - 1 \geq 1$. Meanwhile, when $x \in [0,1]$, $\exp(x) \leq 1 + 2x$. Thus,

$$\sqrt{\min\left\{\exp\left(\frac{\|\mu_2-\mu_1\|_2^2}{\sigma^2}\right) - 1, 1\right\}} \leq \sqrt{\min\left\{1 + \frac{2\|\mu_2-\mu_1\|_2^2}{\sigma^2} - 1, 1\right\}} = \min\left\{\frac{2\|\mu_2-\mu_1\|^2}{\sigma^2}, 1\right\},$$

which finishes the proof. □

With Lemma 11, we have that

$$\begin{aligned}
&\hat{V}_0^\pi(s_0) - V_0^\pi(s_0) \\
\leq &\mathbb{E}\left[\mathbf{1}_{h-1<\tau}\left(\mathbb{E}_{s'_h\sim\mathcal{N}(\hat{f}(s_h,a_h),\sigma^2)}\left[\tilde{V}_h^\pi(s_h)\right] - \mathbb{E}_{s'_h\sim\mathcal{N}(f^*(s_h,a_h),\sigma^2)}\left[\tilde{V}_h^\pi(s_h)\right)\right]\right] \\
\leq &\sum_{h=0}^{H-1}\mathbb{E}\left[\sqrt{\mathbb{E}_{s'_{h+1}\sim\mathcal{N}(\hat{f}(s_h,a_h),\sigma^2)}\left[\tilde{V}_h^\pi(s'_{h+1})^2\right]}\min\left\{\frac{\sqrt{2}\|f^*(s_h,a_h) - \hat{f}(s_h,a_h)_2\|}{\sigma}, 1\right\}\right] \\
\leq &\sum_{h=0}^{H-1}\sqrt{\mathbb{E}\left[\mathbb{E}_{s'_{h+1}\sim\mathcal{N}(\hat{f}(s_h,a_h),\sigma^2)}\left[\tilde{V}_h^\pi(s'_{h+1})^2\right]\right]}\sqrt{\mathbb{E}\left[\min\left\{\frac{2\|f^*(s_h,a_h) - \hat{f}(s_h,a_h)_2^2\|}{\sigma^2}, 1\right\}\right]} \\
\leq &\sqrt{\mathbb{E}\left[\sum_{h=0}^{H-1}\mathbb{E}_{s'_{h+1}\sim P(\cdot|f^*(s_h,a_h))}\left[\tilde{V}_h^\pi(s'_{h+1})^2\right]\right]}\sqrt{\mathbb{E}\left[\sum_{h=0}^{H-1}\min\left\{\frac{2\|f^*(s_h,a_h) - \hat{f}(s_h,a_h)_2^2\|}{\sigma^2}, 1\right\}\right]} \\
\leq &H^{3/2}\sqrt{\mathbb{E}\left[\sum_{h=0}^{H-1}\min\left\{\frac{2\|f^*(s_h,a_h) - \hat{f}(s_h,a_h)_2^2\|}{\sigma^2}, 1\right\}\right]}
\end{aligned}$$

where the second inequality is due to Lemma 12, and the last inequality is due to the fact that $\tilde{V}_h^\pi(s'_{h+1}) \leq H, \forall h$. □

C.3 SUM OF WIDTH SQUARE

Lemma 13 (Bound on the Sum of Width Square)**.** *Define*

$$w_{\mathcal{F}}(s,a) := \sup_{\bar{f},\underline{f}\in\mathcal{F}} \|\bar{f}(s,a) - \underline{f}(s,a)\|_2.$$

If $\{\beta_k^*\}_{k\in[K]}$ is a non-decreasing sequence, and $\|f\|_2 < C, \forall f \in \mathcal{F}$, then:

$$\sum_{k\in[K]}\sum_{h=0}^{H-1} w_{\mathcal{F}_t}^2(s_h^k, a_h^k) \leq 1 + 4C^2 H \dim_E\left(\mathcal{F}, T^{-1/2}\right) + 4\beta_K \dim_E\left(\mathcal{F}, T^{-1/2}\right)(1 + \log T)$$

Proof. We first show the following lemma, which will be helpful in our proof.

Lemma 14 (Lemma 1 in [Osband and Van Roy, 2014])**.** *If $\{\beta_k\}_{k\in[K]}$ is a non-decreasing sequence, we have*

$$\sum_{k\in[K]}\sum_{h=0}^{H-1}\mathbf{1}_{w_{\mathcal{F}_k}(s_h^k,a_h^k)>\epsilon} \leq \left(\frac{4\beta_K}{\epsilon^2} + H\right)\dim_E(\mathcal{F},\epsilon).$$

Proof. We first consider when $w_{\mathcal{F}_k}(s_h^k, a_h^k) > \epsilon$ and is ϵ-dependent on n disjoint sub-sequences of $\{(s_h^i, a_h^i)\}_{i\in[k-1]}$. By the definition of ϵ-dependent, we know $\|\bar{f} - \underline{f}\|_{2,E_k} > n\epsilon^2$. On the other hand, by triangle inequality, we know $\|\bar{f} - \underline{f}\|_{2,E_k} \leq 2\sqrt{\beta_k} \leq 2\sqrt{\beta_K}$, thus $n < \frac{4\beta_K}{\epsilon^2}$. Hence we know when $w_{\mathcal{F}_k}(s_h^k, a_h^k) > \epsilon$, then (s_h, a_h) is at most ϵ-dependent on $\frac{4\beta_K}{\epsilon^2}$ disjoint sub-sequences of $\{(s_h^i, a_h^i)\}^{i\in[k-1]}$.

We then show that, for any sequence $\{(s_i, a_i)\}_{i\in[N]}$, there is some element (s_j, a_j) that is ϵ-dependent on at least $\frac{n}{\dim_E(\mathcal{F},\epsilon)} - H$ disjoint sub-sequences of $\{(s_i, a_i)\}_{i\in[j-1]}$. Let n satisfies that $n\dim_E(\mathcal{F},\epsilon) + 1 \leq N \leq (n+1)\dim_E(\mathcal{F},\epsilon)$, and we

will construct n disjoint sub-sequences $\{B_i\}_{i\in[n]}$. We first let $B_i = \{(s_i, a_i)\}, \forall i \in [n]$. If (s_{k+1}, a_{k+1}) is ϵ-dependent on each $B_i, i \in [n]$, we have the desired results. Otherwise, we append (s_{k+1}, a_{k+1}) to the sub-sequence that it is ϵ-independent with. Repeat this process until some $j > n+1$ is ϵ-dependent on each sub-sequence or we have reached N. In the latter case we have $\sum_{i\in[n]} |B_i| \geq n\dim_E(\mathcal{F}, \epsilon)$ (here we can add at most $H-1$ data to avoid the case we need a new episode of data), and since each element of a sub-sequence is ϵ-independent with its predecessors, $|B_i| \leq \dim_E(\mathcal{F}, \epsilon), \forall i$ by the definition of eluder dimension. Thus $|B_i| = \dim_E(\mathcal{F}, \epsilon), \forall i$. And in this case, (s_N, a_N) must be ϵ-dependent on each sub-sequence by the definition of eluder dimension. Notice that, as our data is collected in an episodic pattern, there are at most $H-1$ sub-sequences that contains "imaginary" final episode data introduced to the construction. In this case, we know that there are at least $\frac{n}{\dim_E(\mathcal{F},\epsilon)} - H$ disjoint sub-sequences that (s_N, a_N) is ϵ-dependent, which finishes our claim.

We finally consider the sub-sequence $B = \{(s_h^k, a_h^k)\}$ with $w_{\mathcal{F}_k}(s_h^k, a_h^k) > \epsilon$. We know that each element in B is ϵ-dependent on at most $\frac{4\beta_K}{\epsilon^2}$ disjoint sub-sequence of B, but at least ϵ-dependent on $\frac{|B|}{\dim_E(\mathcal{F},\epsilon)} - H$ sub-sequence of B. Thus we know $|B| \leq \left(\frac{4\beta_K}{\epsilon^2} + H\right)\dim_E(\mathcal{F}, \epsilon)$, which concludes the proof. □

For notation simplicity, we define $w_{t,h} := w_{\mathcal{F}_t}(s_h^t, a_h^t)$. We first reorder the sequence $\{w_{t,h}\}_{k\in[K],0\leq h\leq H-1} \to \{w_i\}_{i\in[KH]}$, such that $w_1 \geq \cdots w_{TH}$. Then we have

$$\sum_{k\in[K]}\sum_{h=0}^{H-1} w^2_{\mathcal{F}_t}(s_h^k, a_h^k) = \sum_{i\in[KH]} w_i^2 \leq \sum_{i\in[KH]} w_i^2 \mathbf{1}_{w_i<T^{-1/2}} + \sum_{i\in[KH]} w_i^2 \mathbf{1}_{w_i\geq T^{-1/2}} \leq 1 + \sum_{i\in[KH]} w_i^2 \mathbf{1}_{w_i\geq T^{-1/2}}.$$

As we order the sequence, $w_j \geq \epsilon$ means

$$\sum_{k\in[K]}\sum_{h=0}^{H-1} \mathbf{1}_{w_{\mathcal{F}_t}(s_h^k,a_h^k)>\epsilon} \geq j.$$

Hence we know

$$\epsilon \leq \sqrt{\frac{4\beta_K}{\frac{j}{\dim_E(\mathcal{F},\epsilon)} - H}} = \sqrt{\frac{4\beta_K \dim_E(\mathcal{F}, \epsilon)}{j - H\dim_E(\mathcal{F}, \epsilon)}},$$

which means if $w_i \geq T^{-1/2}$, then $w_i < \min\left\{2C, \sqrt{\frac{4\beta_K \dim_E(\mathcal{F},T^{-1/2})}{k-H\dim_E(\mathcal{F},T^{-1/2})}}\right\}$. Hence,

$$\begin{aligned}\sum_{i\in[KH]} w_i^2 \mathbf{1}_{w_i\geq T^{-1/2}} \leq& 4C^2 H\dim_E\left(\mathcal{F}, T^{-1/2}\right) + \sum_{j=H\dim_E(\mathcal{F},T^{-1/2})+1}^{T} \frac{4\beta_K \dim_E(\mathcal{F}, T^{-1/2})}{j - H\dim_E(\mathcal{F}, T^{-1/2})} \\ \leq& 4C^2 H\dim_E\left(\mathcal{F}, T^{-1/2}\right) + 4\beta_K \dim_E\left(\mathcal{F}, T^{-1/2}\right)(1 + \log T),\end{aligned}$$

which finishes the proof. □

C.4 PROOF FOR THEOREM 5 AND THEOREM 7

Proof. Define $\mathcal{E}_k = \mathbb{P}_{f^*}(f^* \in \mathcal{F}_k)$. When constructing the confidence set, take $\alpha = T^{-1/2}$ and $\delta = 0.25$ in Lemma 6, which leads to

$$\beta_k^* := 8\sigma^2 \log(4\mathcal{N}(\mathcal{F}, T^{-1/2}, \|\cdot\|_2)) + HT^{-1/2}(12C + \sqrt{8d\sigma^2 \log(16k^2H)}).$$

With our confidence set construction, we know that $\sum_{k\in[K]} P(\bar{\mathcal{E}}_k) \leq 0.5$. Notice that

$$\begin{aligned}\text{Regret}(K) =& \sum_{k\in[K]} \left[V_0^*(s_0^k) - V_0^{\pi_k}(s_0^k)\right] \\ \leq& \mathbb{E}\left[\sum_{k\in[K]} \mathbb{E}\left[\mathbb{P}(\mathcal{E}_k)[V^*(s_0^k) - V_0^{\pi_k}(s_0^k)]\right]\right] + H\sum_{k\in[K]} \mathbb{P}(\bar{\mathcal{E}}_k)\end{aligned}$$

$$
\begin{aligned}
\leq & \mathbb{E}\left[\sum_{k\in[K]} \mathbb{E}\left[\tilde{V}_{0,k}^{\pi_k}(s_0^k) - V_0^{\pi_k}(s_0^k)\right]\right] + 0.5H \\
\leq & H^{3/2}\sum_{k\in K}\sqrt{\mathbb{E}\left[\sum_{h=0}^{H-1}\min\left\{\frac{2\|\tilde{f}_k(s_h^k,a_h^k) - f^*(s_h^k,a_h^k)\|_2^2}{\sigma^2}, 1\right\}\right]} + 0.5H \\
\leq & \sqrt{H^2T\mathbb{E}\left[\sum_{k\in[K]}\sum_{h=0}^{H-1}\min\left\{\frac{2\|\tilde{f}_k(s_h^k,a_h^k) - \hat{f}^*(s_h^k,a_h^k)\|_2^2}{\sigma^2}, 1\right\}\right]} + 0.5H \\
\leq & \sqrt{\frac{2H^2T}{\sigma^2}\left(1 + 4C^2H\dim_E\left(\mathcal{F}, T^{-1/2}\right) + 4\beta_K^*\dim_E\left(\mathcal{F}, T^{-1/2}\right)(1+\log T)\right)} + 0.5H,
\end{aligned}
$$

where the first equality is due to the fact that the total reward for each episode is bounded in $[0, H]$, the second inequality is due to the optimism and our confidence set construction, the third inequality is due to Lemma 9, the fourth inequality is due to Cauchy-Schwartz inequality and the final inequality is due to Lemma 13 , which concludes the proof of Theorem 7. Following the idea of [Russo and Van Roy, 2013, 2014, Osband and Van Roy, 2014], we can translate the worst-case regret bound for UCB algorithm into the expected regret bound for TS algorithm, that conclude the proof of Theorem 5. □

Remark It can be undesirable that our regret bound scale with σ^{-1}, which means our algorithm can perform pretty bad when the noise level is extremely low. It is also more or less counter-intuitive. We want to remark that, such phenomenon is only an artifact introduced by our proof strategy. The simulation lemma (Lemma 9) works well when $f(s,a) - \tilde{f}(s,a)$ is small. However, we need to tolerate some bad episodes to collect sufficient samples, that can eventually make the error small. Fortunately, the regret of such bad episode is at most H. Hence, we can use the following strategy to get rid of the dependency on σ^{-1}.

Definition 15 (Bad and Good Episodes)**.** *Define episode k as a bad episode, if $\exists h \in \{0, 1, \cdots, H-1\}$, such that $w_{k,h} := w_{\mathcal{F}_k}(s_h^k, a_h^k)$ is the largest $H\dim_E(\mathcal{F}, \sigma^2T^{1/2})$ elements in the set $\{w_{k,h}\}_{k\in[K], 0\leq h\leq H-1}$. Define episode k as a good episode, if it is not a bad episode.*

By the definition, we know there are at most $H\dim_E(\mathcal{F}, \sigma^2T^{-1/2})$ bad episodes. We then show the following lemma, that can be directly generalized from Lemma 13, by setting $\epsilon = \sigma^2T^{-1/2}$ and remove the terms from bad episodes.

Lemma 16. *If $\{\beta_k^*\}_{k\in[K]}$ is a non-decreasing sequence, and $\|f\|_2 < C, \forall f \in \mathcal{F}$, then:*

$$
\sum_{k\in[K], k \text{ is good}}\sum_{h=0}^{H-1} w_{\mathcal{F}_t}^2(s_h^k, a_h^k) \leq \sigma^2 + 4\beta_K\dim_E\left(\mathcal{F}, \sigma^2T^{-1/2}\right)(1+\log T)
$$

Eventually, we can obtain the following regret bound, by setting the regret of bad episodes as H, and bounding the regret of good episodes with Lemma 16.

Theorem 17 (Improved Regret Bound)**.** *Assume Assumption 2 to 5 holds. Take $\alpha = \sigma^2T^{-1/2}$ and $\delta = 0.25$ in Lemma 6, which leads to*

$$
\beta_k^* := 8\sigma^2\log(4\mathcal{N}(\mathcal{F}, \sigma^2T^{-1/2}, \|\cdot\|_2)) + H\sigma^2T^{-1/2}(12C + \sqrt{8d\sigma^2\log(16k^2H)}).
$$

We have that

$$
\text{Regret}(K) \leq \sqrt{H^2T(\tfrac{8\beta_K}{\sigma^2} + 1)\dim_E(\mathcal{F}, \sigma^2T^{-1/2})(1+\log T)} + 0.5H + H^2\dim_E(\mathcal{F}, \sigma^2T^{-1/2})
$$

We would like to remark, that the definition of bad and good episodes is only used for the proof. We don't need to make any modification on the algorithm. Notice that, as $\beta_k^* \propto \sigma^2$, our upper bound in Theorem 17 can only scale with σ^{-1} through the logarithm covering number $\log(4\mathcal{N}(\mathcal{F}, \sigma^2T^{-1/2}, \|\cdot\|_2))$ and eluder dimension $\dim_E(\mathcal{F}, \sigma^2T^{-1/2})$. When $\mathcal{F}$ is a linear function class, both term should scale with $\text{polylog}(\sigma)$, that matches the result from [Kakade et al., 2020].

D BOUNDS ON THE COMPLEXITY TERM UNDER LINEAR REALIZABILITY

We provide the upper bound on the covering number and the eluder dimension of $\mathcal{F}$ when $\mathcal{F} := \{\theta^\top \varphi : \theta \in \mathbb{R}^{d_\varphi \times d}, \|\theta\|_2 \leq W\}$ where $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_\varphi}$ is some known feature map. We first make the following standard assumption:

Assumption 1 (Bounded Feature)**.**

$$\|\varphi(s,a)\|_2 \leq B, \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

D.1 COVERING NUMBER

Theorem 18 (Covering Number Bound)**.** *We have that*

$$\mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|_2) \leq \left(1 + \frac{2BW}{\epsilon}\right)^{d_\varphi}.$$

Proof. Notice that, by Cauchy-Schwartz inequality, we have that

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\varepsilon_i^\top \varphi(s,a)\|_2 \leq B\|\varepsilon_i\|_2, \quad \forall \varepsilon_i \in \mathbb{R}^{d_\varphi}.$$

Thus, denote $\varepsilon = [\varepsilon_i]_{i\in[d]}$, we have that

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\varepsilon^\top \varphi(s,a)\|_2^2 = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{i\in[d]} \|\varepsilon_i^\top \varphi(s,a)\|_2^2 \leq B^2 \sum_{i\in[d]} \|\varepsilon_i\|_2^2 = B^2\|\varepsilon\|_2^2.$$

Hence, to find an ϵ-cover for $\mathcal{F}$, we just need to find an ϵ/B-cover of $\{\theta : \theta \in \mathbb{R}^{d_\varphi \times d}, \|\theta\|_2 \leq W\}$. By standard argument on the covering number of Euclidean space (e.g. Lemma 5.7 in [Wainwright, 2019]), we can conclude the desired result. □

D.2 ELUDER DIMENSION

Theorem 19 (Eluder Dimension Bound)**.** *We have that*

$$\dim_E(\mathcal{F}, \epsilon) \leq \frac{3d_\varphi e}{e-1} \log\left(3 + \frac{12W^2B^2}{\epsilon^2}\right) + 1.$$

Proof. Our proof follows the idea in [Russo and Van Roy, 2013]. Define

$$w_k := \sup\left\{(\theta_1 - \theta_2)^\top \varphi(s,a) : \sqrt{\sum_{i\in[k-1]} \left((\theta_1-\theta_2)^\top \varphi_i(s_i,a_i)\right)^2} \leq \epsilon', \theta_1, \theta_2 \in \mathbb{R}^{d_\varphi \times d}, \|\theta_1\| \leq W, \|\theta_2\| \leq W\right\}.$$

For notation simplicity, define $\varphi_k := \varphi(s_i, a_i)$, $\theta := \theta_1 - \theta_2$, and $\Phi_k := \sum_{i\in[k-1]} \varphi_i \varphi_i^\top$. Obviously, we have that $\|\theta\| \leq 2W$. Moreover, by straightforward calculation, we know

$$\sum_{i\in[k-1]} \left((\theta_1-\theta_2)^\top \varphi_i(s_i,a_i)\right)^2 = \mathrm{Trace}(\theta^\top \varphi_k \theta).$$

Define $V_k := \Phi_k + \frac{(\epsilon')^2}{4W^2} I$, we start from considering the problem

$$\max_\theta \mathrm{Trace}(\theta^\top \varphi_k \varphi_k^\top \theta), \quad \text{subject to} \quad \mathrm{Trace}(\theta^\top V_k \theta) \leq 2\epsilon^2.$$

The Lagrangian can be formed as

$$\mathcal{L}(\theta, \gamma) = -\mathrm{Trace}(\theta^\top \varphi_k \varphi_k^\top \theta) + \lambda(\mathrm{Trace}(\theta^\top V_k \theta) - 2\epsilon^2), \quad \lambda \geq 0.$$

The optimality condition of θ is

$$(\lambda V_k - \varphi_k \varphi_k^\top)\theta = 0.$$

As V_k is of full rank, $\lambda V_k - \varphi_k \varphi_k^\top$ has rank at least $d_\varphi - 1$ (as $\varphi_k \varphi_k^\top$ is of rank 1). So the equation

$$(\lambda V_k - \varphi_k \varphi_k^\top)\theta_i = 0, \quad \theta_i \in \mathbb{R}^{d_\varphi}$$

only has one non-zero solution. Substitute back, we know that (define $\|x\|_A := \sqrt{x^\top A x}$):

$$\sup\{\mathrm{Trace}(\theta^\top \varphi_k \varphi_k^\top \theta) : \mathrm{Trace}(\theta^\top V_k \theta) \le \epsilon^2\} = \sqrt{2}\epsilon' \|\varphi_k\|_{V_k^{-1}}.$$

With the conclusion above, we have that

$$w_k \le \sup\{\theta^\top \varphi_k : \mathrm{Trace}(\theta^\top \Phi_k \theta) \le \epsilon^2, \|\theta\| \le 2W\} \le \sup\{\theta^\top \varphi_k : \mathrm{Trace}(\theta^\top V_k \theta) \le 2\epsilon^2\} = \sqrt{2}\epsilon' \|\varphi_k\|_{V_k^{-1}}.$$

Hence, if $w_k \ge \epsilon'$, then $\varphi_k V_k^{-1} \varphi_k \ge 0.5$. Moreover, with Matrix Determinant Lemma, if $w_i \ge \epsilon', \forall i < k$, we have

$$\det(V_k) = \det(V_{k-1})(1 + \varphi_k^\top V_k^{-1} \varphi_k) \ge \det(V_{k-1}) \left(\frac{3}{2}\right) \ge \cdots \ge \det\left(\frac{(\epsilon')^2}{4W^2} I\right) \left(\frac{3}{2}\right)^{k-1} = \frac{(\epsilon')^{2d}}{4W^{2d}} \left(\frac{3}{2}\right)^{k-1}.$$

Meanwhile,

$$\det(V_k) \le \left(\frac{\mathrm{Trace}(V_k)}{d}\right)^d \le \left(\frac{B^2(k-1)}{d} + \frac{(\epsilon')^2}{4W^2}\right)^d.$$

Hence, we know

$$\left(\frac{3}{2}\right)^{(k-1)/d} \le \frac{4W^2 B^2}{(\epsilon')^2} \cdot \frac{k-1}{d} + 1.$$

Now we only need to find the largest k that can make this inequality hold. For notation simplicity, define $\alpha := \frac{4W^2B^2}{(\epsilon')^2}$, $n = \frac{k-1}{d}$. As $\log(1+x) \ge \frac{x}{1+x}$ and $\log x \le x/e$, we have

$$\frac{n}{3} \le n \log 3/2 \le \log(\alpha + 1) + \log n \le \log(\alpha + 1) + \log 3 + \log(n/3) \le \log(\alpha + 1) + \log 3 + \frac{n}{3e}.$$

Substitute back, we can obtain the desired result. □

E EXPERIMENTAL DETAILS

E.1 ALGORITHM SUMMARY

Our algorithm is easily built on SAC. The only difference we make is we decouple the critic network into a representation network $\phi(\cdot)$ and a linear layer $l(\cdot)$ on top of the representation. The representation network is governed by the model dynamics loss in SPEDE, and we train a linear layer to predict the Q-value as it lies in the linear space of the representation guaranteed by our analysis. We update the representation by a momentum factor and keep the policy update the same procedure as SAC.

E.2 FULL EXPERIMENTS

Table 1: Performance of SPEDE on various MuJoCo control suite tasks. Our method achieve strong performance even comparing to pure empirical baselines. To be specific, in hard tasks like Humanoid-ET and Ant-ET, SPEDE outperforms the baselines significantly. Results with * are directly adopted from MBBL [Wang et al., 2019]. We also provide the SoTA model-free RL method SAC as a reference.

	Swimmer	Ant-ET	Hopper-ET	Pendulum
ME-TRPO*	30.1±9.7	42.6±21.1	4.9±4.0	177.3±1.9
PETS-RS*	42.1±20.2	130.0±148.1	205.8±36.5	167.9±35.8
PETS-CEM*	22.1±25.2	81.6±145.8	129.3±36.0	167.4±53.0
DeepSF	25.5±13.5	768.1±44.1	548.9±253.3	168.6±5.1
SPEDE	42.6±4.2	806.2±60.2	732.2±263.9	169.5±0.6
SAC*	41.2±4.6	2012.7±571.3	1815.5±655.1	168.2±9.5
	Reacher	Cartpole	I-pendulum	Walker-ET
ME-TRPO*	-13.4±5.2	160.1±69.1	-126.2±86.6	-9.5±4.6
PETS-RS*	-40.1±6.9	195.0±28.0	-12.1±25.1	-0.8±3.2
PETS-CEM*	-12.3±5.2	199.5±3.0	-20.5±28.9	-2.5±6.8
DeepSF	-16.8±3.6	194.5±5.8	-0.2±0.3	165.6±127.9
SPEDE	-7.2±1.1	138.2±39.5	0.0±0.0	501.58±204.0
SAC*	-6.4±0.5	199.4±0.4	-0.2±0.1	2216.4±678.7
	MountainCar	Acrobot	SlimHumanoid-ET	Humanoid-ET
ME-TRPO*	-42.5±26.6	68.1±6.7	76.1±8.8	776.8±62.9
PETS-RS*	-78.5±2.1	-71.5±44.6	320.7±182.2	106.9±102.6
PETS-CEM*	-57.9±3.6	12.5±29.0	355.1±157.1	110.8±91.0
DeepSF	-17.0±23.4	-74.4±3.2	533.8±154.9	241.1±116.6
SPEDE	50.3±1.1	-69.0±3.3	986.4±154.7	886.9±95.2
SAC*	52.6±0.6	-52.9±2.0	843.6±313.1	1794.4±458.3

E.3 ABLATIONS

Table 2: Ablation Suty of SPEDE on MuJoCo tasks. We see that a small momentum factor help stabilize the performance, especially in environments like Huamoid and Hopper-ET.

	Hopper-ET	Ant-ET	S-Humanoid-ET	Humanoid-ET
SPEDE-0.9	593.2±37.4	**877.7±45.9**	881.6±385.2	232.9±63.4
SPEDE-0.99	305.9±13.4	707.9±51.1	629.3±106.9	818.1±130.6
SPEDE-0.999	**732.2±263.9**	806.2±60.2	**986.4±154.7**	**886.9±95.2**

Momentum Update Our ablation experiments are trying to study an important design choice of the practical algorithm: the momentum used to update the critic function. We summarize the results in Table 2. We can see that using a small large momentum factor such as 0.999 shows better performance. This is intuitively understandable: large momentum factor slows down the update speed of the representation of the critic function and thus stabilize the training. Such phenomenon illustrates the importance of slowly update the representation.

Random Feature Dimension We also conduct the experiments on how does the random feature dimension affect the final performance of the algorithm. We plot the results in HalfCheetah environment in Fig. 1. We can see that when we increasing the random feature dimension, we see a performance gain on the final return. This suggests that using a larger number of feature dimension would help the performance.

MLP Network for Critic Network We also conduct an experiment to study whether adding a MLP network on top of our representation could work. We show such ablation in Tab. 3. From the results, we see that the performance of MLP network is in generally better than the Linear network.

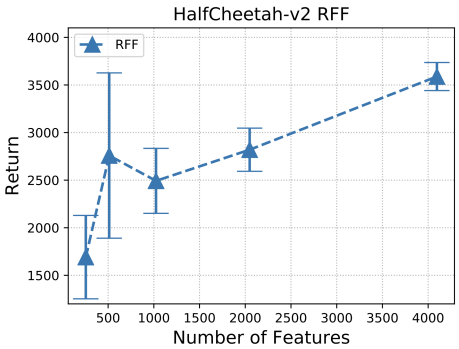

Figure 1: Increasing the number of random features can also lead to a performance gain.

Table 3: Comparison of SPEDE linear critic network and critic network. Results show that in general MLP network will further improve the performance.

	Reacher	MountainCar	Cartpole	Acrobot
SPEDE-Linear	-7.2±1.1	50.3±1.1	138.2±39.5	-69.0±3.3
SPEDE-MLP	-6.8±0.4	**53.8±1.1**	171.9±31.0	**-15.6±1.9**
SAC	**-6.4±0.5**	52.6±0.6	**199.4±0.4**	-52.9±2.0
	Pendulum	I-Pendulum	Walker-ET	S-Humanoid-ET
SPEDE-Linear	**169.5±0.6**	**0.0±0.0**	501.6±204.0	986.4±154.7
SPEDE-MLP	165.9±4.2	**0.0±0.0**	1005.7±458.4	**2521.1±420.8**
SAC	168.2±9.5	-0.2±0.1	**2216.4±678.7**	843.6±313.1

E.4 COMPARISON TO LC3

We provide a comparison of empirical results with LC3 [Kakade et al., 2020], which is also an algorithm with rigorous theoretical guarantees. Despite the major difference that we are learning the representation while LC3 assumes a given feature, the performance of SPEDE is much better than LC3 in tasks like Mountain Car and Hopper.

Table 4: Comparison of SPEDE with LC3 on MuJoCo tasks. LC3 only achieves good performance on relatively easy tasks like Reacher. However, their performance on Hopper and Mountain-Car is much worse than SPEDE.

	Reacher	MountainCar	Hopper
SPEDE	-7.2±1.1	**50.3±1.1**	**732.2±263.9**
LC3	**-4.1±1.6**	27.3±8.1	-1016.5±607.4

E.5 PERFORMANCE CURVES

We provide an additional performance curve including ME-TRPO in Figure 2 for a reference.

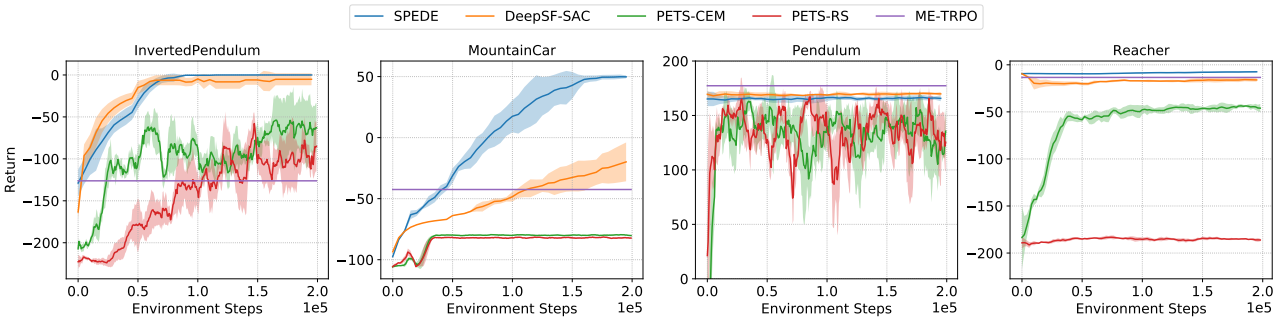

Figure 2: **Experiments on MuJoCo:** We show curves of the return versus the training steps for SPEDE and model-based RL baselines. We also include the final performance of ME-TRPO from [Wang et al., 2019] for reference.

E.6 HYPERPARAMETERS

We conclude the hyperparameter we use in our experiments in the following.

Table 5: Hyperparameters used for SPEDE in all the environments in MuJoCo.

	Hyperparameter Value
Actor lr	0.0003
Model lr	0.0001
Actor Network Size	(1024, 1024, 1024)
Fourier Feature Size	1024
Discount	0.99
Target Update Tau	0.005
Model Update Tau	0.001
Batch Size	256