A. Proofs for Behavioral Cloning

We prove Theorem 5.1 in this section by proving Lemma 5.2,5.3. In this section, we abuse notation and define $\ell^\mu(\phi, f) := \ell^\mu(\pi^{\phi,f})$, where ℓ^μ is defined in Equation 4. We rewrite it here for convenience.

$$\ell^\mu(\pi) = \mathop{\mathbb{E}}_{(s,a)\sim\mu} \ell(\pi(s), a) = \mathop{\mathbb{E}}_{(s,a)\sim\mu} -\log(\pi(s)_a)$$

Let $\hat{f}^\phi_{\mathbf{x}} = \arg\min_{f\in\mathcal{F}} \ell^{\mathbf{x}}(\phi, f)$ be the optimal task specific parameter for task μ by fixing representation ϕ. Thus by our definitions in Section 5, we get $\pi^{\phi,\mathbf{x}} = \pi^{\phi,\hat{f}^\phi_{\mathbf{x}}}$. We assume w.l.o.g. that $\mathcal{A} = [K]$. Remember that $\ell : \triangle(\mathcal{A}) \times \mathcal{A} \to \mathbb{R}$ is defined as $\ell(\boldsymbol{v}, a) = -\log(\boldsymbol{v}_a)$ for some $\boldsymbol{v} \in \mathbb{R}^K$ and $\boldsymbol{v}_a$ is the coordinate corresponding to action $a \in \mathcal{A} = [K]$. We define a new function class and loss function that will be useful for our proofs

$$\mathcal{F}' = \{x \to Wx \mid W \in \mathbb{R}^{K\times d}, \|W\|_F \le 1\} \tag{9}$$

$$\ell'(\boldsymbol{v}, a) = -\log(\texttt{softmax}(\boldsymbol{v})_a), \boldsymbol{v} \in \mathbb{R}^K, a \in \mathcal{A} \tag{10}$$

We basically offloaded the burden of computing $\texttt{softmax}$ from the class $\mathcal{F}$ to the loss ℓ'. We can convert any function $f' \in \mathcal{F}'$ to one in $\mathcal{F}$ by transforming it to $\texttt{softmax}(f')$. We now proceed to proving the lemmas

Proof of Lemma 5.2. We can then rewrite the various loss functions from Section 5 as follows

$$\hat{L}(\phi) = \frac{1}{T}\sum_{i=1}^{T} \min_{f'\in\mathcal{F}'} \frac{1}{n}\sum_{j=1}^{n} \ell'(f'(\phi(s)), a)$$
$$L(\phi) = \mathop{\mathbb{E}}_{\mu\sim\eta} \min_{f'\in\mathcal{F}'} \mathop{\mathbb{E}}_{(s,a)\sim\mu} \ell'(f'(\phi(s)), a)$$
$$\bar{L}(\phi) = \mathop{\mathbb{E}}_{\mu\sim\eta} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n} \mathop{\mathbb{E}}_{(s,a)\sim\mu} \ell'(\hat{f'}^\phi_{\mathbf{x}}(\phi(s)), a)$$

where $\hat{f'}^\phi_\mu \in \arg\min_{f'\in\mathcal{F}'} \ell^{\mathbf{x}}(\phi, \texttt{softmax}(f'))$. It is easy to show that both $\ell'(\cdot, a)$ $\ell'(f'(\cdot), \cdot)$ are 2-lipschitz in their arguments for every $a \in \mathcal{A}$ and $f' \in \mathcal{F}'$. Using a slightly modified version of Theorem 2(i) from Maurer et al. (2016), we get that for $\hat{\phi} \in \arg\min_{\phi\in\Phi} \hat{L}(\phi)$, with probability at least $1-\delta$ over the choice of $\mathbf{X}$

$$\bar{L}(\hat{\phi}) - \min_{\phi\in\Phi} L(\phi) \le \frac{2\sqrt{2\pi}G(\Phi(\mathbf{S}))}{T\sqrt{n}} + \sqrt{2\pi}Q' \sup_{\phi\in\Phi} \sqrt{\frac{\mathop{\mathbb{E}}_{\mu\sim\eta,(s,a)\sim\mu} \|\phi(s)\|^2}{n}} + \sqrt{\frac{8\log(4/\delta)}{T}}$$

$$\bar{L}(\hat{\phi}) - \min_{\phi\in\Phi} L(\phi) \le c\frac{G(\Phi(\mathbf{S}))}{T\sqrt{n}} + c'\frac{Q'R}{\sqrt{n}} + c''\sqrt{\frac{\log(4/\delta)}{T}} \tag{11}$$

where $Q' = \sup_{y\in\mathbb{R}^{dn}\setminus\{0\}} \frac{1}{\|y\|}\mathbb{E} \sup_{f\in\mathcal{F}'} \sum_{i=1,j=1}^{n,K} \gamma_{ij} f'(y_i)_j$. First we discuss why we need a modified version of their theorem. Our setting differs from the setting for Theorem 2 from Maurer et al. (2016) in the following ways

- $\mathcal{F}'$ is a class of vector valued function in our case, whereas in Maurer et al. (2016) it is assumed to contain scalar valued. The only place in the proof of the theorem where this shows up is in the definition of Q', which we have updated accordingly.
- Maurer et al. (2016) assumes that $\ell'(\cdot, a)$ is 1-lipschitz for every $a \in \mathcal{A}$ and that $f'(\cdot)$ is L lipschitz for every $f' \in \mathcal{F}'$. However the only properties that are used in the proof of Theorem 16 are that $\ell'(\cdot, a)$ is 1-lipschitz and that $\ell'(f'(\cdot), a)$ is L-lipschitz for every $a \in \mathcal{A}$, which is exactly the property that we have. Hence their proof follows through for our setting as well.

Lemma A.1. $Q' := \sup_{y\in\mathbb{R}^{dn}\setminus\{0\}} \frac{1}{\|y\|}\mathbb{E} \sup_{f\in\mathcal{F}'} \sum_{i=1,j=1}^{n,K} \gamma_{ij} f'(y_i)_j \le \sqrt{K}$

Proof.

$$
\begin{aligned}
Q' &:= \sup_{y\in\mathbb{R}^{dn}\setminus\{0\}} \frac{1}{\|y\|}\mathbb{E}\sup_{f\in\mathcal{F}'}\sum_{i=1,j=1}^{n,K}\gamma_{ij}f'(y_i)_j \\
&= \sup_{y\in\mathbb{R}^{dn}\setminus\{0\}} \frac{1}{\|y\|}\mathbb{E}\sup_{\|W\|_F\le 1}\sum_{i=1,j=1}^{n,K}\gamma_{ij}\langle W_j, y_i\rangle \\
&= \sup_{y\in\mathbb{R}^{dn}\setminus\{0\}} \frac{1}{\|y\|}\mathbb{E}\sup_{\|W\|_F\le 1}\sum_{j=1}^{K}\langle W_j, \sum_{i=1}^{n}\gamma_{ij}y_i\rangle \\
&=^{(a)} \sup_{y\in\mathbb{R}^{dn}\setminus\{0\}} \frac{1}{\|y\|}\mathbb{E}\sqrt{\sum_{j=1}^{K}\left\|\sum_{i=1}^{n}\gamma_{ij}y_i\right\|^2} \\
&\le^{(b)} \sup_{y\in\mathbb{R}^{dn}\setminus\{0\}} \frac{1}{\|y\|}\sqrt{\sum_{j=1}^{K}\mathbb{E}\left\|\sum_{i=1}^{n}\gamma_{ij}y_i\right\|^2} = \sup_{y\in\mathbb{R}^{dn}\setminus\{0\}} \frac{1}{\|y\|}\sqrt{\sum_{j=1}^{K}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{i'=1}^{n}\gamma_{ij}\gamma_{i'j}\langle y_i, y_{i'}\rangle\right]} \\
&=^{(c)} \sup_{y\in\mathbb{R}^{dn}\setminus\{0\}} \frac{1}{\|y\|}\sqrt{\sum_{j=1}^{K}\sum_{i=1}^{n}\|y_i\|^2} = \frac{1}{\|y\|}\sqrt{K\|y\|^2} = \sqrt{K}
\end{aligned}
$$

where we use Jensen's inequality and linearity of expectation for (b) and properties of standard normal gaussian variables for (c). For (a) we observe that $\sup_{\|W\|_F\le 1}\sum_{j=1}^{K}\langle W_j, A_j\rangle = \sup_{\|W\|_F\le 1}\langle W, A\rangle = \|A\|_F = \sum_{j=1}^{K}\|A_j\|^2$. □

Plugging in Lemma A.1 into Equation 11 completes the proof. □

We now proceed to prove the next lemma.

Proof of Lemma 5.3. Suppose $\bar{L}(\phi) = \mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\ell^{\mu}(\pi^{\phi,\mathbf{x}}) \le \epsilon$. Consider a task $\mu\sim\eta$ and samples $\mathbf{x}\sim\mu^n$ and let $\epsilon_\mu(\mathbf{x}) = \ell^{\mu}(\pi^{\phi,\mathbf{x}})$ so that $\bar{L}(\phi) = \mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\epsilon_\mu(\mathbf{x})$. Since π^*_μ is deterministic, we get

$$
\begin{aligned}
\mathop{\mathbb{E}}_{s\sim\nu^*_\mu}\mathop{\mathbb{E}}_{a\sim\pi^{\phi,\mathbf{x}}}\mathbb{I}\{a\neq\pi^*_\mu(s)\} &= \mathop{\mathbb{E}}_{s\sim\nu^*_\mu}[1-\pi^{\phi,\mathbf{x}}(s)_{\pi^*_\mu(s)}] \\
&\le \mathop{\mathbb{E}}_{s\sim\nu^*_\mu}[-\log(1-(1-\pi^{\phi,\mathbf{x}}(s)_{\pi^*_\mu(s)}))] \\
&= \mathop{\mathbb{E}}_{s\sim\nu^*_\mu}[-\log(\pi^{\phi,\mathbf{x}}(s)_{\pi^*_\mu(s)})] = \epsilon_\mu(\mathbf{x})
\end{aligned}
$$

where we use the fact that $x\le -\log(1-x)$ for $x<1$. for the first inequality. Thus by using Theorem 2.1 from Ross et al. (2011), we get that $J_\mu(\pi^{\phi,\mathbf{x}}) - J_\mu(\pi^*) \le H^2\epsilon_\mu(\mathbf{x})$. Taking expectation w.r.t. $\mu\sim\eta$ and $\mathbf{x}\sim\mu^n$ completes the proof. □

Proof of Theorem 5.1. By using Assumption 5.2, we are guaranteed the existence of $\pi_\mu\in\Pi^{\phi^*}$ such that $\pi_\mu(s)_{\phi^*_\mu(s)}\ge 1-\gamma$ for every $s\in\mathcal{S}$. Thus we can get an upper bound on $L(\phi^)$

$$
\begin{aligned}
L(\phi^*) &= \mathop{\mathbb{E}}_{\mu\sim\eta}\min_{\pi\in\Pi^{\phi^*}}\mathop{\mathbb{E}}_{s\sim\nu^*_\mu} -\log(\pi(s)_{\pi^*_\mu(s)}) \\
&\le \mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{s\sim\nu^*_\mu} -\log(\pi_\mu(s)_{\pi^*_\mu(s)}) \\
&\le \mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{s\sim\nu^*_\mu} -\log(1-\gamma) \le 2\gamma
\end{aligned}
$$

where in the last step we used $-\log(1-x)\le 2x$ for $x<1/2$. Hence from Lemma 5.2 we get $\bar{L}(\hat{\phi})\le 2\gamma+\epsilon_{gen,h}$, which combining with Lemma 5.3 gives the desired result. □

B. Proofs for Observation-Alone

Before proving Theorem 6.1, we introduce the following loss functions, as we did in the proof sketch for the behavioral cloning setting. We again abuse notation and define $\ell^\mu(\phi, f) \coloneqq \ell^\mu(\pi^{\phi,f})$, where ℓ^μ is defined in Equation 6. Let $\hat{f}^\phi_{\mathbf{x}} = \arg\min_{f\in\mathcal{F}} \ell^{\mathbf{x}}(\phi, f)$ be the optimal task specific parameter for task μ by fixing representation ϕ. As before, we define the following

$$\bar{L}_h(\phi_h) = \mathop{\mathbb{E}}_{\mu\sim\eta} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu_h^n} \ell_h^\mu(\phi, \hat{f}^{\phi_h}_{\mathbf{x}})$$

We first show a guarantee on the performance of representations $(\hat{\phi}_1, \dots, \hat{\phi}_H)$ as measured by the functions $\bar{L}_1, \dots, \bar{L}_H$.

Theorem B.1. *With probability at least $1-\delta$ in the draw of $\mathbf{X} = (\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(H)})$, $\forall h \in [H]$*

$$\bar{L}_h(\hat{\phi}_h) \le \min_{\phi\in\Phi} L_h(\phi) + c\epsilon_{gen,h}(\Phi) + c'\epsilon_{gen,h}(\mathcal{F},\mathcal{G}) + c''\sqrt{\frac{\ln(H/\delta)}{T}}$$

where $\epsilon_{gen,h}(\Phi) = \frac{KG(\Phi(\mathbf{S}_h))}{T\sqrt{n}}$ *and* $\epsilon_{gen,h}(\mathcal{F},\mathcal{G}) = \mathop{\mathbb{E}}_{\mu\sim\eta} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n} \left[\frac{KG(\mathcal{G}(\bar{\mathbf{s}}_h))}{n} + \frac{G(\mathcal{G}(\bar{\mathbf{s}}_h))}{n}\right] + \frac{RK\sqrt{K}}{\sqrt{n}}$

We then connect the losses $\bar{L}_h$ to the expected cost on the tasks.

Theorem B.2. *Consider representations $(\phi_1, \dots, \phi_H)$ with $\bar{L}_h(\phi_h) \le \epsilon_h$. Let $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_H)$ be samples at different levels for a newly sampled task $\mu \sim \eta$ such that $\mathbf{x}_h \sim \mu_h^n$. Let $\boldsymbol{\pi}^{\phi,\mathbf{x}} = (\pi^{\phi_1,\mathbf{x}_1}, \dots, \pi^{\phi_H,\mathbf{x}_H})$ be policies learned using the samples, then under Assumption 6.1,*

$$\mathop{\mathbb{E}}_{\mu\sim\eta} \mathop{\mathbb{E}}_{\mathbf{x}} J(\boldsymbol{\pi}^{\phi,\mathbf{x}}) - \mathop{\mathbb{E}}_{\mu\sim\eta} J(\boldsymbol{\pi}^*_\mu) \le \sum_{h=1}^H (2H-2h+1)\epsilon_h + O(H^2)\epsilon^\phi_{be}$$

where $\epsilon^\phi_{be} = \mathop{\mathbb{E}}_{\mu\sim\eta} \mathop{\mathbb{E}}_{\mathbf{x}} [\epsilon^{\boldsymbol{\pi}^{\phi,\mathbf{x}}}_{be}]$ *is the average inherent Bellman error.*

It is easy to show that under Assumption 6.2, $\min_{\phi\in\Phi} L_h(\phi) = 0$ for every $h \in [H]$. Thus from Theorem B.1, we get that $\bar{L}_h(\hat{\phi}_h) \le \epsilon_{gen,h}$, where $\epsilon_{gen,h} = \epsilon_{gen,h}(\Phi) + \epsilon_{gen,h}(\mathcal{F},\mathcal{G}) + c''\sqrt{\frac{\ln(H/\delta)}{T}}$. Invoking Theorem B.2 on the representations $\{\hat{\phi}_h\}$ completes the proof.

B.1. Proof of Theorem B.1

Before proving the theorem, we discuss important lemmas. In yet another abuse of notation, we define $\ell_h^\mu(\phi, f, g) = \mathbb{E}_{(s,a,\tilde{s},\bar{s})\sim\mu_h}[K\pi^{\phi,f}(a|s)g(\tilde{s}) - g(\bar{s})]$ and $\ell_h^{\mathbf{x}}(\phi, f, g) = \frac{1}{n}\sum_{j=1}^n [K\pi^{\phi,f}(a_j|s_j)g(\tilde{s}_j) - g(\bar{s}_j)]$.

Let $\hat{m}_{\mathbf{x}}(\phi) = \min_{f\in\mathcal{F}} \max_{g\in\mathcal{G}} \hat{\ell}_h^{\mathbf{x}}(\phi, f, g) = \hat{\ell}_h^{\mathbf{x}}(\phi, \hat{f}^\phi_{\mathbf{x}}, \hat{g}^\phi_{\mathbf{x}})$, $\bar{m}_{\mu,\mathbf{x}}(\phi) = \max_{g\in\mathcal{G}} \ell_h^\mu(\phi, \hat{f}^\phi_{\mathbf{x}}, g)$ and $m_\mu(\phi) = \min_{f\in\mathcal{F}} \max_{g\in\mathcal{G}} \ell_h^\mu(\phi, f, g)$. Note that $L_h(\phi) = \mathop{\mathbb{E}}_{\mu\sim\eta} m(\phi)$, $\bar{L}_h(\phi) = \mathop{\mathbb{E}}_{\mu\sim\eta} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n} \bar{m}_{\mu,\mathbf{x}}(\phi)$. Define the distribution ρ_h where $\mathbf{x} \sim \rho_h$ is the same as $\mu \sim \eta$ and then $\mathbf{x} \sim \mu_h^n$.

Lemma B.3. *For every $\phi \in \Phi$ and $h \in [H]$,*

$$\mathop{\mathbb{E}}_{\mu\sim\eta} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n} \sup_{f\in\mathcal{F}} \sup_{g\in\mathcal{G}} \left[\hat{\ell}_h^{\mathbf{x}}(\phi, f, g) - \ell_h^\mu(\phi, f, g)\right] \le \epsilon_{gen,h}(\mathcal{F},\mathcal{G})$$

Lemma B.4. *With probability $1-\delta$, for every $\phi \in \Phi$,*

$$\bar{L}_h(\phi) - \mathop{\mathbb{E}}_{\mathbf{x}\sim\rho_h} \hat{m}_{\mathbf{x}}(\phi) \le \epsilon_{gen,h}(\mathcal{F},\mathcal{G})$$

Lemma B.5. *With probability $1-\delta$, for every $\phi \in \Phi$,*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\rho_h} \hat{m}_{\mathbf{x}}(\phi) - \frac{1}{T}\sum_i \hat{m}_{\mathbf{x}^{(i)}}(\phi) \le \epsilon_{gen,h}(\Phi) + O\left(\sqrt{\frac{\log(\frac{1}{\delta})}{T}}\right)$$

We prove these lemmas later. First we prove Theorem B.1 using them. If $\phi_h^* = \arg\min_{\phi\in\Phi} L_h(\phi)$, then

$$\begin{aligned}
\bar{L}_h(\hat{\phi}_h) - L_h(\phi_h^*) = &\left(\bar{L}_h(\hat{\phi}_h) - \mathop{\mathbb{E}}_{\mathbf{x}\sim\rho_h} \hat{m}_{\mathbf{x}}(\phi)\right) \\
&+ \left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\rho_h} \hat{m}_{\mathbf{x}}(\phi) - \frac{1}{T}\sum_i \hat{m}_{\mathbf{x}^{(i)}}(\hat{\phi}_h)\right) \\
&+ \left(\frac{1}{T}\sum_i \hat{m}_{\mathbf{x}^{(i)}}(\hat{\phi}_h) - \frac{1}{T}\sum_i \hat{m}_{\mathbf{x}^{(i)}}(\phi_h^*)\right) \\
&+ \left(\frac{1}{T}\sum_i \hat{m}_{\mathbf{x}^{(i)}}(\phi_h^*) - \mathop{\mathbb{E}}_{\mathbf{x}\sim\rho_h} \hat{m}_{\mathbf{x}}(\phi_h^*)\right) \\
&+ \mathop{\mathbb{E}}_{\mu\sim\eta}\left[\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n} \hat{m}_{\mathbf{x}}(\phi_h^*) - m_\mu(\phi_h^*)\right] \\
&\le 2\epsilon_{gen,h}(\mathcal{F},\mathcal{G}) + \epsilon_{gen,h}(\Phi) + O\left(\sqrt{\frac{\log(\frac{1}{\delta})}{T}}\right)
\end{aligned}$$

where for the first part we use Lemma B.4, second part we use Lemma B.5, third part is upper bounded by 0 by optimality of $\hat{\phi}_h$, fourth is upper bounded by $O(\sqrt{\frac{\log(\frac{1}{\delta})}{T}})$ by Hoeffding's inequality and fifth is bounded by the following argument: let $f^\phi, g^\phi = \arg\min_{f\in\mathcal{F}} \arg\max_{g\in\mathcal{G}} \ell^\mu(\phi, f, g)$

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n} \hat{m}_{\mathbf{x}}(\phi_h^*) &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n} \min_{f\in\mathcal{F}} \max_{g\in\mathcal{G}} \hat{\ell}_h^{\mathbf{x}}(\phi_h^*, f, g) \\
&\le \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n} \max_{g\in\mathcal{G}} \hat{\ell}_h^{\mathbf{x}}(\phi_h^*, f^{\phi_h^*}, g) \\
&=^{(a)} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n} \hat{\ell}_h^{\mathbf{x}}(\phi_h^*, f^{\phi_h^*}, \tilde{g}) \\
&\le^{(b)} \ell_h^\mu(\phi_h^*, f^{\phi_h^*}, \tilde{g}) + \epsilon_{gen,h}(\mathcal{F},\mathcal{G}) \\
&\le \ell_h^\mu(\phi_h^*, f^{\phi_h^*}, g^{\phi_h^*}) + \epsilon_{gen,h}(\mathcal{F},\mathcal{G}) = m_\mu(\phi_h^*) + \epsilon_{gen,h}(\mathcal{F},\mathcal{G})
\end{aligned}$$

where in step (a) we use $\tilde{g} = \arg\max_{g\in\mathcal{G}} \hat{\ell}_h^{\mathbf{x}}(\phi_h^*, f^{\phi_h^*}, g)$, for (b) we use Lemma B.3.

B.2. Proof of Theorem B.2

Consider a task μ. For simplicity of notation, we use π_h instead $\pi^{\phi_h,\mathbf{x}_h}$, $\boldsymbol{\pi}$ instead of $\boldsymbol{\pi}^{\phi,\mathbf{x}}$. Let ν_h^π and ν_h^* be the state distributions at level h induced by $\boldsymbol{\pi}^{\phi,\mathbf{x}}$ and $\boldsymbol{\pi}_\mu^*$ respectively. Let

$$\epsilon_h(\mathbf{x}_h) = \max_{g\in\mathcal{G}} \mathop{\mathbb{E}}_{s\sim\nu_h^*}\left[\mathop{\mathbb{E}}_{\substack{a\sim\pi_h \\ s'\sim P_{s,a}}} g(s') - \mathop{\mathbb{E}}_{\substack{a\sim\pi_h^* \\ s'\sim P_{s,a}}} g(s')\right]$$

be the loss of policy π_h at level h. By definition, $\epsilon_h = \mathop{\mathbb{E}}_{\mu\sim\eta} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu_h^n} \epsilon_h(\mathbf{x})$. Using Lemma C.1 from (Sun et al., 2019), we have

$$J(\boldsymbol{\pi}^{\phi,\mathbf{x}}) - J(\boldsymbol{\pi}_\mu^*) = \sum_{h=1}^H \bar{\Delta}_h = \sum_{h=1}^H \mathop{\mathbb{E}}_{s\sim\nu_h^\pi}\left[\mathop{\mathbb{E}}_{a\sim\pi_h(\cdot|s),s'\sim P_{s,a}} V_{h+1}^*(s') - \mathop{\mathbb{E}}_{a\sim\pi_h^*(\cdot|s),s'\sim P_{s,a}} V_{h+1}^*(s')\right]$$

Observe that

$$\begin{aligned}
\bar{\Delta}_h = &\mathop{\mathbb{E}}_{s\sim\nu_h^\pi}\left[\mathop{\mathbb{E}}_{a\sim\pi_h(\cdot|s),s'\sim P_{s,a}} V_{h+1}^*(s') - \mathop{\mathbb{E}}_{a\sim\pi_h^*(\cdot|s),s'\sim P_{s,a}} V_{h+1}^*(s')\right] \\
\le^{(a)} &\mathop{\mathbb{E}}_{s\sim\nu_h^*} \mathop{\mathbb{E}}_{a\sim\pi_h(\cdot|s),s'\sim P_{s,a}} V_{h+1}^*(s') - \mathop{\mathbb{E}}_{s\sim\nu_h^*} \mathop{\mathbb{E}}_{a\sim\pi_h^*(\cdot|s),s'\sim P_{s,a}} V_{h+1}^*(s') +
\end{aligned}$$

$$
\begin{aligned}
& \mathop{\mathbb{E}}_{s\sim\nu_h^\pi} \mathop{\mathbb{E}}_{a\sim\pi_h(\cdot|s),s'\sim P_{s,a}} V_{h+1}^*(s) - \mathop{\mathbb{E}}_{s\sim\nu_h^*} \mathop{\mathbb{E}}_{a\sim\pi_h(\cdot|s),s'\sim P_{s,a}} V_{h+1}^*(s) + \\
& \mathop{\mathbb{E}}_{s\sim\nu_h^*} \mathop{\mathbb{E}}_{a\sim\pi_h^*(\cdot|s),s'\sim P_{s,a}} V_{h+1}^*(s') - \mathop{\mathbb{E}}_{s\sim\nu_h^\pi} \mathop{\mathbb{E}}_{a\sim\pi_h^*(\cdot|s),s'\sim P_{s,a}} V_{h+1}^*(s') \\
\leq^{(b)} \quad & \max_{g\in\mathcal{G}} \mathop{\mathbb{E}}_{s\sim\nu_h^*} [\mathop{\mathbb{E}}_{a\sim\pi_h(\cdot|s),s'\sim P_{s,a}} g(s') - \mathop{\mathbb{E}}_{a\sim\pi_h^*(\cdot|s),s'\sim P_{s,a}} g(s')] + \\
& \max_{g\in\mathcal{G}} [\mathop{\mathbb{E}}_{s\sim\nu_h^\pi} \Gamma_h^{\boldsymbol{\pi}} g(s) - \mathop{\mathbb{E}}_{s\sim\nu_h^*} \Gamma_h^{\boldsymbol{\pi}} g(s)] + [\mathop{\mathbb{E}}_{s\sim\nu_h^*} \Gamma_h^* V_{h+1}^*(s) - \mathop{\mathbb{E}}_{s\sim\nu_h^\pi} \Gamma_h^* V_{h+1}^*(s)] \\
\leq^{(c)} \quad & \epsilon_h(\mathbf{x}_h) + \max_{g\in\mathcal{G}} [\mathop{\mathbb{E}}_{s\sim\nu_h^\pi} \Gamma_h^{\boldsymbol{\pi}} g(s) - \mathop{\mathbb{E}}_{s\sim\nu_h^*} \Gamma_h^{\boldsymbol{\pi}} g(s)] + \max_{g\in\mathcal{G}} [\mathop{\mathbb{E}}_{s\sim\nu_h^\pi} g(s) - \mathop{\mathbb{E}}_{s\sim\nu_h^*} g(s)]
\end{aligned}
$$

where (a) just adds and subtracts terms, (b) uses the assumption that $V_{h+1}^* \in \mathcal{G}$ and the definitions of Γ_h^* and $\Gamma_h^{\boldsymbol{\pi}}$ from Section 3 and (c) uses the definition of $\epsilon_h(\mathbf{x}_h)$. The following lemma helps us bound the remaining two terms.

Lemma B.6. *Defining* $\Delta_h = \max_{g\in\mathcal{G}} | \mathop{\mathbb{E}}_{s\sim\nu_h^\pi} g(s) - \mathop{\mathbb{E}}_{s\sim\nu_h^*} g(s)|$, *we have*

$$
\max_{g\in\mathcal{G}} [\mathop{\mathbb{E}}_{s\sim\nu_h^\pi} \Gamma_h^{\boldsymbol{\pi}} g(s) - \mathop{\mathbb{E}}_{s\sim\nu_h^*} \Gamma_h^{\boldsymbol{\pi}} g(s)] \leq \Delta_h + 2\epsilon_{be}^{\boldsymbol{\pi}}
$$

Using the above lemma, we get $\bar{\Delta}_h \leq \epsilon_h(\mathbf{x}_h) + 2\Delta_h + 2\epsilon_{be}^{\boldsymbol{\pi}}$. We now bound Δ_h

$$
\begin{aligned}
\Delta_h &= \max_{g\in\mathcal{G}} \left| \mathop{\mathbb{E}}_{s\sim\nu_{h-1}^\pi} \mathop{\mathbb{E}}_{\substack{a\sim\pi_{h-1}\\ s'\sim P_{s,a}}} g(s') - \mathop{\mathbb{E}}_{s\sim\nu_h^*} g(s) \right| \\
&\leq^{(a)} \max_{g\in\mathcal{G}} \left| \mathop{\mathbb{E}}_{s\sim\nu_{h-1}^\pi} \mathop{\mathbb{E}}_{\substack{a\sim\pi_{h-1}\\ s'\sim P_{s,a}}} g(s') - \mathop{\mathbb{E}}_{s\sim\nu_{h-1}^*} \mathop{\mathbb{E}}_{\substack{a\sim\pi_{h-1}\\ s'\sim P_{s,a}}} g(s) \right| + \max_{g\in\mathcal{G}} \left| \mathop{\mathbb{E}}_{s\sim\nu_{h-1}^*} \mathop{\mathbb{E}}_{\substack{a\sim\pi_{h-1}\\ s'\sim P_{s,a}}} g(s') - \mathop{\mathbb{E}}_{s\sim\nu_h^*} g(s) \right| \\
&= \max_{g\in\mathcal{G}} \left| \mathop{\mathbb{E}}_{s\sim\nu_{h-1}^\pi} \Gamma_{h-1}^{\boldsymbol{\pi}} g(s') - \mathop{\mathbb{E}}_{s\sim\nu_{h-1}^*} \Gamma_{h-1}^{\boldsymbol{\pi}} g(s) \right| + \epsilon_{h-1}(\mathbf{x}_{h-1}) \\
&\leq \Delta_{h-1} + 2\epsilon_{be}^{\boldsymbol{\pi}} + \epsilon_{h-1}(\mathbf{x}_{h-1})
\end{aligned}
$$

where (a) uses triangle inequality. Thus $\Delta_h \leq 2(h-1)\epsilon_{be}^{\boldsymbol{\pi}} + \epsilon_{1:h-1}(\mathbf{x}_{1:h-1})$ and so $\bar{\Delta}_h \leq \epsilon_{1:h}(\mathbf{x}_{1:h}) + \epsilon_{1:h-1}(\mathbf{x}_{1:h-1}) + (4h-2)\epsilon_{be}^{\boldsymbol{\pi}}$. This implies that

$$
J(\boldsymbol{\pi}^{\phi,\mathbf{x}}) - J(\boldsymbol{\pi}^*) = \sum_{h=1}^H \bar{\Delta}_h \leq \sum_{h=1}^H (2H - 2h + 1)\epsilon_h(\mathbf{x}_h) + O(H^2)\epsilon_{be}^{\boldsymbol{\pi}^{\phi,\mathbf{x}}}
$$

Taking expectation wrt $\mu \sim \eta$ and $\mathbf{x} \sim \mu^n$ completes the proof.

B.3. Proofs of Lemmas

In the following proofs, we will require the well known Slepian's lemma which lets us exploit lipschitzness of functions in gaussian averages

Lemma B.7 (Slepian's lemma)**.** *Let* $\{X\}_{s\in S}$ *and* $\{Y\}_{s\in S}$ *be zero mean Gaussian processes such that*

$$
\mathbb{E}(X_s - X_t)^2 \leq \mathbb{E}(Y_s - Y_t)^2, \forall s,t \in S
$$

Then

$$
\mathbb{E} \sup_{s\in S} X_s \leq \mathbb{E} \sup_{s\in S} Y_s
$$

We now move on to proving earlier lemmas.

Proof of Lemma B.3. Again we define $\mathcal{F}'$ as in Equation 9. Let $\ell(\boldsymbol{v}, \alpha, \beta, a) = K\texttt{softmax}(\boldsymbol{v})_a\alpha - \beta$, and let $\ell'^{\mu}_h(\phi, f', g) = \ell^{\mu}_h(\phi, \texttt{softmax}(f'), g) = \mathop{\mathbb{E}}_{(s,a,\tilde{s},\bar{s})\sim\mu_h} \ell(f'(\phi(s)), g(\tilde{s}), g(\bar{s}), a)$ for $f' \in \mathcal{F}'$ and similarly define $\hat{\ell}'^{\mathbf{x}}_h(\phi, f', g) = \hat{\ell}^{\mathbf{x}}_h(\phi, \texttt{softmax}(f'), g)$. Notice that $\ell(\cdot, \alpha, \beta, a)$ is $2K$-lipschitz, $\ell(\boldsymbol{v}, \cdot, \beta, a)$ is K-lipschitz and $\ell(\boldsymbol{v}, \alpha, \cdot, a)$ is 1-lipschitz, Using Theorem 8(i) from (Maurer et al., 2016), we get that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\sup_{f\in\mathcal{F}}\sup_{g\in\mathcal{G}}\left[\hat{\ell}^{\mathbf{x}}_h(\phi, f, g) - \ell^{\mu}_h(\phi, f, g)\right] \\
&\quad= \mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\sup_{f'\in\mathcal{F}'}\sup_{g\in\mathcal{G}}\left[\hat{\ell}'^{\mathbf{x}}_h(\phi, f', g) - \ell'^{\mu}_h(\phi, f', g)\right] \\
&\quad\le \frac{\sqrt{2\pi}\,\mathbb{E}_{\mathbf{x}}\, G(\ell(\mathcal{F}'(\phi(\mathbf{s}_h)), \mathcal{G}(\bar{\mathbf{s}}_h), \mathcal{G}(\tilde{\mathbf{s}}_h), \mathbf{a}))}{n}
\end{aligned}$$

where the gaussian average is defined as

$$G(\ell(\mathcal{F}'(\phi(\mathbf{s}_h)), \mathcal{G}(\bar{\mathbf{s}}_h), \mathcal{G}(\tilde{\mathbf{s}}_h), \mathbf{a})) = \mathop{\mathbb{E}}_{\gamma_i}\left[\sup_{f'\in\mathcal{F}, g\in\mathcal{G}}\sum_{i=1}^{n}\gamma_i\ell(f'(\phi(s_i)), g(\tilde{s}_i), g(\bar{s}_i), a_i)\right]$$

where $\mathbf{s}_h = \{s_i\}_{i=1}^n$, $\bar{\mathbf{s}}_h = \{\bar{s}_i\}_{i=1}^n$, $\tilde{\mathbf{s}}_h = \{\tilde{s}_i\}_{i=1}^n$. We will now use the lipschitzness of ℓ to get the following.

Claim B.8.

$$\begin{aligned}
(\ell(f'_1(\phi(s_i)), g_1(\tilde{s}_i), g_1(\bar{s}_i), a_i) - \ell(f'_2(\phi(s_i)), g_2(\tilde{s}_i), g_2(\bar{s}_i), a_i))^2 &\le 12K^2\|f'_1(\phi(s)) - f'_2(\phi(s))\|^2 \\
&\quad + 3K^2(g_1(\phi(\tilde{s})) - g_2(\phi(\tilde{s})))^2 \\
&\quad + 3(g_1(\phi(\bar{s})) - g_2(\phi(\bar{s})))^2
\end{aligned}$$

This follows by writing

$$\begin{aligned}
\ell(f'_1(\phi(s_i)), g_1(\tilde{s}_i), g_1(\bar{s}_i), a_i) - \ell(f'_2(\phi(s_i)), g_2(\tilde{s}_i), g_2(\bar{s}_i), a_i) = \\
\ell(f'_1(\phi(s_i)), g_1(\tilde{s}_i), g_1(\bar{s}_i), a_i) - \ell(f'_2(\phi(s_i)), g_1(\tilde{s}_i), g_1(\bar{s}_i), a_i) + \\
\ell(f'_2(\phi(s_i)), g_1(\tilde{s}_i), g_1(\bar{s}_i), a_i) - \ell(f'_2(\phi(s_i)), g_2(\tilde{s}_i), g_1(\bar{s}_i), a_i) + \\
\ell(f'_2(\phi(s_i)), g_2(\tilde{s}_i), g_1(\bar{s}_i), a_i) - \ell(f'_2(\phi(s_i)), g_2(\tilde{s}_i), g_2(\bar{s}_i), a_i)
\end{aligned}$$

and then using the per argument lipschitzness of ℓ described earlier and AM-RMS inequality proves the claim. We move on to decoupling the gaussian average using Slepian's lemma

Claim B.9. *The gaussian average satisfies the following*

$$G(\ell(\mathcal{F}'(\phi(\mathbf{s}_h)), \mathcal{G}(\bar{\mathbf{s}}_h), \mathcal{G}(\tilde{\mathbf{s}}_h), \mathbf{a}) \le 2\sqrt{3}KG(\mathcal{F}'(\phi(\mathbf{s}_h))) + \sqrt{3}KG(\mathcal{G}(\tilde{\mathbf{s}}_h)) + \sqrt{3}G(\mathcal{G}(\bar{\mathbf{s}}_h))$$

where the gaussian average for a class of functions is defined in Equation 2.

This can be shown by defining two gaussian processes $X_{f',g} = \sum_{i=1}^{n}\gamma_i\ell(f'(\phi(s_i)), g(\tilde{s}_i), g(\bar{s}_i), a_i)$ and $Y_{f',g} = \sum_{i=1,j=1}^{n,d}\alpha_{i,j}2\sqrt{3}Kf'(\phi(s_i))_j + \sum_{i=1}^{n}\beta_i\sqrt{3}Kg(\tilde{s}_i) + \sum_{i=1}^{n}\delta_i\sqrt{3}g(\tilde{s}_i)$. It is easy to see the following using expectation of independent gaussian variables

$$\begin{aligned}
\mathop{\mathbb{E}}_{\gamma}(X_{f'_1,g_1} - X_{f'_2,g_2})^2 &= \sum_{i=1}^{n}(\ell(f'_1(\phi(s_i)), g_1(\tilde{s}_i), g_1(\bar{s}_i), a_i) - \ell(f'_2(\phi(s_i)), g_2(\tilde{s}_i), g_2(\bar{s}_i), a_i))^2 \\
\mathop{\mathbb{E}}_{\alpha,\beta,\delta}(Y_{f'_1,g_1} - Y_{f'_2,g_2})^2 &= 12K^2\|f'_1(\phi(s)) - f'_2(\phi(s))\|^2 + 3K^2(g_1(\phi(\tilde{s})) - g_2(\phi(\tilde{s})))^2 + (g_1(\phi(\bar{s})) - g_2(\phi(\bar{s})))^2
\end{aligned}$$

Claim B.8 gives us that $\mathop{\mathbb{E}}_{\gamma}(X_{f_1',g_1} - X_{f_2',g_2})^2 \leq \mathop{\mathbb{E}}_{\alpha,\beta,\delta}(Y_{f_1',g_1} - Y_{f_2',g_2})^2$ and then Slepian's lemma will then give us that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\gamma}\sup_{f',g}\sum_{i=1}^{n}\gamma_i\ell(f'(\phi(s_i)),g(\tilde{s}_i),g(\bar{s}_i),a_i) \\
&\qquad\leq \mathop{\mathbb{E}}_{\alpha,\beta,\delta}\sup_{f',g}\left[\sum_{i=1,j=1}^{n,d}\alpha_{i,j}2\sqrt{3}Kf'(\phi(s_i))_j + \sum_{i=1}^{n}\beta_i\sqrt{3}Kg(\tilde{s}_i) + \sum_{i=1}^{n}\delta_i\sqrt{3}g(\tilde{s}_i)\right] \\
&\qquad\leq \mathop{\mathbb{E}}_{\alpha}\sup_{f'}\left[\sum_{i=1,j=1}^{n,d}\alpha_{i,j}2\sqrt{3}Kf'(\phi(s_i))_j\right] + \mathop{\mathbb{E}}_{\beta}\sup_{g}\left[\sum_{i=1}^{n}\beta_i\sqrt{3}Kg(\tilde{s}_i)\right] + \mathop{\mathbb{E}}_{\delta}\sup_{g}\left[\sum_{i=1}^{n}\delta_i\sqrt{3}g(\tilde{s}_i)\right] \\
&\qquad= 2\sqrt{3}KG(\mathcal{F}'(\phi(\mathbf{s}_h))) + \sqrt{3}KG(\mathcal{G}(\tilde{\mathbf{s}}_h)) + \sqrt{3}G(\mathcal{G}(\bar{\mathbf{s}}_h))
\end{aligned}$$

thus proving the claim. Furthermore, we notice that $G(\mathcal{F}'(\phi(\mathbf{s}_h))) \leq Q'$, where Q' is defined in Lemma A.1. Thus combining all of this, we get

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\sup_{f\in\mathcal{F}}\sup_{g\in\mathcal{G}}\left[\hat{\ell}_h^{\mathbf{x}}(\phi,f,g) - \ell_h^{\mu}(\phi,f,g)\right] \\
&\qquad\leq \mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\frac{2\sqrt{6\pi}KG(\mathcal{F}'(\phi(\mathbf{s}_h)))}{n} + \mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\left[\frac{\sqrt{6\pi}KG(\mathcal{G}(\bar{\mathbf{s}}_h))}{n} + \frac{\sqrt{6\pi}G(\mathcal{G}(\tilde{\mathbf{s}}_h))}{n}\right] \\
&\qquad\leq \frac{2\sqrt{6\pi}KQ'}{n} + \mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\left[\frac{\sqrt{6\pi}KG(\mathcal{G}(\bar{\mathbf{s}}_h))}{n} + \frac{\sqrt{6\pi}G(\mathcal{G}(\tilde{\mathbf{s}}_h))}{n}\right] \\
&\qquad\leq c\frac{RK\sqrt{K}}{\sqrt{n}} + c'\mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\left[\frac{KG(\mathcal{G}(\bar{\mathbf{s}}_h))}{n} + \frac{G(\mathcal{G}(\tilde{\mathbf{s}}_h))}{n}\right] \leq \epsilon_{gen,h}(\mathcal{F},\mathcal{G})
\end{aligned}$$

where we used Lemma A.1 for the last inequality. This completes the proof □

Proof of Lemma B.4.

$$\begin{aligned}
\bar{L}_h(\phi) - \mathop{\mathbb{E}}_{\mathbf{x}\sim\rho_h}\hat{m}_{\mathbf{x}}(\phi) &= \mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\bar{m}_{\mu,\mathbf{x}}(\phi) - \mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\hat{m}_{\mathbf{x}}(\phi) \\
&= \mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\max_{g\in\mathcal{G}}\ell_h^{\mu}(\phi,\hat{f}_{\mathbf{x}}^{\phi},g) - \max_{g\in\mathcal{G}}\hat{\ell}_h^{\mathbf{x}}(\phi,\hat{f}_{\mathbf{x}}^{\phi},g) \\
&\leq^{(a)} \mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\max_{g\in\mathcal{G}}[\ell_h^{\mu}(\phi,\hat{f}_{\mathbf{x}}^{\phi},g) - \hat{\ell}_h^{\mathbf{x}}(\phi,\hat{f}_{\mathbf{x}}^{\phi},g)] \\
&\leq \mathop{\mathbb{E}}_{\mu\sim\eta}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu^n}\max_{f\in\mathcal{F}}\max_{g\in\mathcal{G}}[\ell_h^{\mu}(\phi,f,g) - \hat{\ell}_h^{\mathbf{x}}(\phi,f,g)] \\
&\leq^{(b)} \epsilon_{gen,h}(\mathcal{F},\mathcal{G})
\end{aligned}$$

where (a) follows by observing that $\max_g[\theta(g) - \theta'(g)] \leq \max_g\theta(g) - \max_g\theta'(g)$ for any functions θ,θ', for the first inequality and (b) follows from Lemma B.3. □

Proof of Lemma B.5. We will be using Slepian's lemma Using Theorem 8(ii) from (Maurer et al., 2016), we get that

$$\sup_{\phi\in\Phi}\left[\mathop{\mathbb{E}}_{\mathbf{x}\sim\rho_h}\hat{m}_{\mathbf{x}}(\phi) - \frac{1}{T}\sum_i\hat{m}_{\mathbf{x}^{(i)}}(\phi)\right] \leq \frac{\sqrt{2\pi}}{T}G(S) + \sqrt{\frac{9\ln(2/\delta)}{2T}} \tag{12}$$

where $S = \{(\hat{m}(\phi)_{\mathbf{x}_1},\ldots,\hat{m}(\phi)_{\mathbf{x}_T}) : \phi\in\Phi\}$. We bound the Gaussian average of S using Slepian's lemma. Define two Gaussian processes indexed by Φ as

$$X_\phi = \sum_i\gamma_i\hat{m}(\phi)_{\mathbf{x}^{(i)}} \text{ and } Y_\phi = \frac{2K}{\sqrt{n}}\sum_i\gamma_{ijk}\phi(s_j^i)_k$$

For $\mathbf{x} = \{(s_j, a_j, \tilde{s}_j, \bar{s}_j)\}$, consider 2 representations ϕ and ϕ',

$$\begin{aligned}
(\hat{m}(\phi)_{\mathbf{x}} - \hat{m}(\phi')_{\mathbf{x}})^2 &= (\min_{f\in\mathcal{F}}\max_{g\in\mathcal{G}} \hat{\ell}_h^{\mathbf{x}}(\phi, f, g) - \min_{f\in\mathcal{F}}\max_{g\in\mathcal{G}} \hat{\ell}_h^{\mathbf{x}}(\phi', f, g))^2 \\
&\le (\sup_{f\in\mathcal{F},g\in\mathcal{G}} |\hat{\ell}_h^{\mathbf{x}}(\phi, f, g) - \hat{\ell}_h^{\mathbf{x}}(\phi', f, g)|)^2 \\
&= \left(\sup_{f\in\mathcal{F},g\in\mathcal{G}} \left|\frac{1}{n}\sum_j [K\pi^{\phi,f}(a_j|s_j)g(\tilde{s}_j) - K\pi^{\phi',f}(a_j|s_j)g(\tilde{s}_j)]\right|\right)^2 \\
&= K^2\left(\sup_{f\in\mathcal{F},g\in\mathcal{G}} \left|\frac{1}{n}\sum_j \left(f(\phi(s_j))_{a_j} - f(\phi'(s_j))_{a_j}\right) g(\tilde{s}_j)\right|\right)^2 \\
&\le \frac{K^2}{n}\sup_{f\in\mathcal{F}}\sum_j \left(f(\phi(s_j))_{a_j} - f(\phi'(s_j))_{a_j}\right)^2 \\
&\le \frac{4K^2}{n}\sum_j |\phi(s_j) - \phi'(s_j)|^2 = \frac{4K^2}{n}\sum_{j,k}(\phi(s_j)_k - \phi'(s_j)_k)^2
\end{aligned}$$

where we prove the first inequality later, second inequality comes from g being upper bounded by 1 and by Cauchy-Schwartz inequality, third inequality comes from the 2-lipschitzness of f.

$$\begin{aligned}
\mathbb{E}(X_\phi - X_{\phi'}) &= \sum_i (\hat{m}(\phi)_{\mathbf{x}^{(i)}} - \hat{m}(\phi')_{\mathbf{x}^{(i)}})^2 \\
&\le \frac{4K^2}{n}\sum_{i,j,k}(\phi(s_j^i)_k - \phi'(s_j^i)_k)^2 = \mathbb{E}(Y_\phi - Y_{\phi'})^2
\end{aligned}$$

Thus by Slepian's lemma, we get

$$G(S) = \mathbb{E}\sup_{\phi\in\Phi} X_\phi \le \mathbb{E}\sup_{\phi\in\Phi} Y_\phi = \frac{2K}{\sqrt{n}} G(\Phi(\{s_j^i\}))$$

Plugging this into Equation 12 completes the proof. To prove the first inequality above, notice that

$$\begin{aligned}
\min_{f\in\mathcal{F}}\max_{g\in\mathcal{G}} \hat{\ell}_h^{\mathbf{x}}(\phi, f, g) - \min_{f\in\mathcal{F}}\max_{g\in\mathcal{G}} \hat{\ell}_h^{\mathbf{x}}(\phi', f, g) &= \hat{\ell}_h^{\mathbf{x}}(\phi, f, g) - \hat{\ell}_h^{\mathbf{x}}(\phi', f', g') \\
&\le \hat{\ell}_h^{\mathbf{x}}(\phi, f', g'') - \hat{\ell}_h^{\mathbf{x}}(\phi', f', g') \\
&\le \hat{\ell}_h^{\mathbf{x}}(\phi, f', g'') - \hat{\ell}_h^{\mathbf{x}}(\phi', f', g'') \\
&\le \sup_{f\in\mathcal{F},g\in\mathcal{G}} |\hat{\ell}_h^{\mathbf{x}}(\phi, f, g) - \hat{\ell}_h^{\mathbf{x}}(\phi', f, g)|
\end{aligned}$$

By symmetry, we also get that $\min_{f\in\mathcal{F}}\max_{g\in\mathcal{G}} \hat{\ell}_h^{\mathbf{x}}(\phi, f, g) - \min_{f\in\mathcal{F}}\max_{g\in\mathcal{G}} \hat{\ell}_h^{\mathbf{x}}(\phi', f, g) \le \sup_{f\in\mathcal{F},g\in\mathcal{G}} |\hat{\ell}_h^{\mathbf{x}}(\phi, f, g) - \hat{\ell}_h^{\mathbf{x}}(\phi', f, g)|$. □

Proof of Lemma B.6. Let $\bar{g} = \arg\max_{g\in\mathcal{G}} \left(\mathbb{E}_{s\sim\nu_h^\pi} \Gamma_h^{\boldsymbol{\pi}} g(s) - \mathbb{E}_{s\sim\nu_h^*} \Gamma_h^{\boldsymbol{\pi}} g(s)\right)$ and $g' = \arg\min_{g\in\mathcal{G}} |g - \Gamma_h^{\boldsymbol{\pi}}\bar{g}|_{(\nu_h^\pi+\nu_h^*)/2}$.

$$\begin{aligned}
&\max_{g\in\mathcal{G}} \left(\mathbb{E}_{s\sim\nu_h^\pi} \Gamma_h^{\boldsymbol{\pi}} g(s) - \mathbb{E}_{s\sim\nu_h^*} \Gamma_h^{\boldsymbol{\pi}} g(s)\right) = \mathbb{E}_{s\sim\nu_h^\pi} \Gamma_h^{\boldsymbol{\pi}}\bar{g}(s) - \mathbb{E}_{s\sim\nu_h^*} \Gamma_h^{\boldsymbol{\pi}}\bar{g}(s) \\
&= \mathbb{E}_{s\sim\nu_h^\pi} g'(s) - \mathbb{E}_{s\sim\nu_h^*} g'(s) + \mathbb{E}_{s\sim\nu_h^\pi}[\Gamma_h^{\boldsymbol{\pi}}\bar{g}(s) - g'(s)] + \mathbb{E}_{s\sim\nu_h^*}[g'(s) - \Gamma_h^{\boldsymbol{\pi}}\bar{g}(s)] \\
&\le |\mathbb{E}_{s\sim\nu_h^\pi} g'(s) - \mathbb{E}_{s\sim\nu_h^*} g'(s)| + \mathbb{E}_{s\sim\nu_h^\pi}[|g'(s) - \Gamma_h^{\boldsymbol{\pi}}\bar{g}(s)|] + \mathbb{E}_{s\sim\nu_h^*}[|g'(s) - \Gamma_h^{\boldsymbol{\pi}}\bar{g}(s)|] \\
&\le \max_{g\in\mathcal{G}} |\mathbb{E}_{s\sim\nu_h^\pi} g(s) - \mathbb{E}_{s\sim\nu_h^*} g(s)| + 2\mathbb{E}_{s\sim(\nu_h^\pi+\nu_h^*)/2}[|g'(s) - \Gamma_h^{\boldsymbol{\pi}}\bar{g}(s)|] \\
&\le \Delta_h + 2\epsilon_{be}^\pi
\end{aligned}$$

□

C. Data Set Collection Details

C.1. Dataset from trajectories

Given n expert trajectories for a task μ, for each trajectory $\tau = (s_1, a_1, \ldots, s_H, a_H)$ we can sample an $h \sim \mathcal{U}([H])$ and select the pair (s_h, a_h) from that trajectory[8]. This gives us n i.i.d. pairs $\{(s_j, a_j)\}_{j=1}^n$ for the task μ. We collect this for T tasks and get datasets $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}$.

C.2. Dataset from trajectories and interaction

Given $2n$ expert trajectories for a task μ, we use first n trajectories to get independent samples from the distributions $\nu^*_{1,\mu}, \ldots, \nu^*_{H,\mu}$ respectively for the $\bar{s}$ states in the dataset. Using the next n trajectories, we get samples from $\nu^*_{0,\mu}, \ldots, \nu^*_{H-1,\mu}$ for the s states in the dataset, and for each such state we uniformly sample an action a from $\mathcal{A}$ and then get a state $\tilde{s}$ from $P_{s,a}$ by resetting the environment to s and playing action a. We collect this for T tasks and get datasets $\mathbf{X}^{(i)} = \{\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_H^{(i)}\}$ for every $i \in [T]$, where each dataset $\mathbf{x}_h^{(i)}$ a set of n tuples obtained level h. Rearranging, we can construct the datasets $\mathbf{X}_h = \{\mathbf{x}_h^{(1)}, \ldots, \mathbf{x}_h^{(T)}\}$.

D. Experiment Details

For the policy optimization experiments, we use 5 random seeds to evaluate our algorithm. We show the results for 1 test environment as the results for other test environments are also showing the algorithm works but the magnitude of reward might be different, so we do not average the numbers over different test environments.

Environment Setup We first describe the construction of the NoisyCombinationLock environment. The state space is $\mathbb{R}^{50}$. Each state s is in the form of $[s_{\text{noise}}, s_{\text{index}}, s_{\text{real}}]$, where $s_{\text{noise}} \in \mathbb{R}^{10}$ is sampled from $\mathcal{N}(\mathbf{0}, w\mathbf{I})$, $s_{\text{index}} \in \mathbb{R}^{10}$ is an one-hot vector indicating the current step and $s_{\text{real}} \in \mathbb{R}^{10}$ is sampled from $\mathcal{N}(\mathbf{0}, w\mathbf{I})$. The constant w is set to $\sqrt{0.05}$ such that $\|s_{\text{real}}\|$ has expected norm of 1. The action space is $\{-1, 1\}$. Each MDP is parametrized by a vector $\mathbf{c}^* \in \{-1, 1\}^{20}$, which determines the optimal action sequence. We use different $\mathbf{a}^*$ to define different environments. The transition model is that: Let $s = [s_{\text{noise}}, s_{\text{index}}, s_{\text{real}}]$ be the current state and a be the action. If $s_{\text{index}} = e_i$ for some i and $\mathbf{c}^*_i a s_{\text{real}i} > 0$, then $s'_{\text{index}} = e_{i+1}$. Otherwise s'_{index} will be all zero. s'_{noise} and s'_{real} will always be sampled from the Gaussian distribution. The reward is 1 if and only if s_{index} is not a zero vector, otherwise it's 0. Note that once s_{index} is all zero, it will not change and the reward will always be 0. The maximum horiozn is set to 20 and therefore, the optimal policy has return 20. The initial s_{index} is always e_1.

The SwimmerVelocity environment is similar to goal velocity experiments in (Finn et al., 2017a), and is based on the Swimmer environment in OpenAI Gym (Brockman et al., 2016). The only difference is the reward function, which is now defined by $r(s) = |v - v_{\text{goal}}|$, where v is the current velocity of the agent and v_{goal} is the goal velocity. The state space is still $\mathbb{R}^8$. The original action space in Swimmer is $\mathbb{R}^2$, and we discretize the action space, such that each entry can be only one of $\{-1, -0.5, 0, 0.5, 1\}$. We also reduce the maximum horizon from 1000 to 50.

Experts For NoisyCombinationLock, the demonstrations are generated by the optimal policy, which has access to the hidden vector $\mathbf{c}^*$. For SwimmerVelocity, we trained the experts for 1 million steps by PPO (Schulman et al., 2017) to make sure it converges with code from Dhariwal et al. (2017).

Architecture For all of our experiments in Figure 1 and 2, the function ϕ is parametrized by $\phi(x) = \sigma(Wx + b)$ where W, b are learnable parameters of a linear layer and $\sigma(\cdot)$ is the ReLU activation function. However, the number of hidden units might vary. Note that in the experiments of verifying our theory (Figure 1), we train a policy (and the representation) at each step so the dimension of representation is smaller. See Table 1 for our choice of hyperparameters.

Optimization All optimization, including training ϕ, π and behavior cloning baseline, is done by Adam (Kingma & Ba, 2014) with learning rate 0.001 until it converges, except NoisyCombinationLock in policy optimization experiments in Figure 2 where we use learning rate 0.01 for faster convergence. To solve Equation equation 5 and equation 7, we build a

[8] In practice one can use all pairs from all trajectories, even though the samples are not strictly i.i.d.

	BC (Figure 1, left two)	OA (Figure 1, right two)	RL (Figure 2)
NoisyCombinationLock	5	5	40
SwimmerVelocity	100	20	100

Table 1. Number of hidden units for different experiments.

joint loss over ϕ and all f's in each task,

$$\mathcal{L}(\phi, f_1, \ldots, f_T) = \frac{1}{nT} \sum_{t=1}^{T} \sum_{j=1}^{n} -\log(\pi^{\phi, f_t}(s_j^t)_{a_j^t}). \tag{13}$$

Then we minimize $\mathcal{L}(\phi, f_1, \ldots, f_T)$ and obtain the optimal ϕ.