Supplementary Material: Scaling Up Approximate Value Iteration with Options: Better Policies with Fewer Iterations

Timothy A. Mann Shie Mannor

1 Definitions and Notation

In contrast to the discounted termination state probability density $\widetilde{P}$, we denote the undiscounted probability that an option o executed from a state $x \in X$ will terminate in a subset of states $Y \subseteq X$ by

$$P^o(Y|x) = \sum_{t=1}^{\infty} P_t^o(Y|x) \ . \tag{1}$$

Notice that because (1) is undiscounted $\int \widetilde{P}^o(y|x)dy < \int P^o(y|x)dy = 1$. For an option policy $\varphi : X \to \mathcal{O}$, we will denote by $\widetilde{P}^\varphi$ discounted termination state probability distribution for executing φ once at each state (executing each option until termination) and the undiscounted termination state probability distribution P^φ analogously. Notice that for an option policy, we also have

$$P^\varphi(Y|x) = \sum_{t=1}^{\infty} P_t^\varphi(Y|x) \tag{2}$$

for all $Y \subseteq X$ and $x \in X$.

We would like to be able to express policies over primitive actions using the same sum over all timesteps used in (2). For a policy $\pi : X \to A$ defined over primitive actions, we define $P_t^\pi(Y|x) = \begin{cases} P^\pi(Y|x) & \text{if } t = 1 \\ 0 & \text{otherwise} \end{cases}$ so that

$$P^\pi(Y|x) = \sum_{t=1}^{\infty} P_t^\pi(Y|x) \tag{3}$$

for all $Y \subseteq X$ and $x \in X$.

Notice that if f is an option, an option policy, or a policy over primitive actions we can write the discounted termination state probability density by

$$\widetilde{P}^f(Y|x) = \sum_{t=1}^{\infty} \gamma^t P_t^f(Y|x)$$

for all $Y \subseteq X$ and $x \in X$. When we compose options $o_1, o_2, \ldots o_m$, we write $\widetilde{P}^{o_1 o_2 \ldots o_m} = \widetilde{P}^{o_1}\widetilde{P}^{o_2}\ldots\widetilde{P}^{o_m}$, and we can write

$$\widetilde{P}^{o_1 o_2 \ldots o_m}(Y|x) = \sum_{t=1}^{\infty} \gamma^{m+t} \left(P^{o_1}P^{o_2}\ldots P^{o_m}\right)_t (Y|x)$$

for all $Y \subseteq X$ and $x \in X$.

We will assume throughout this supplementary material that when we refer to an optimal policy π^*, it is a policy over primitive actions. Because we have assume that $\mathcal{O}$ contains the set of primitive actions A, the fixed point of the SMDP Bellman operator $\mathbb{T}$ and the MDP Bellman operator $\mathcal{T}$ is the optimal value function V^*. Thus $\mathbb{T}^{\pi^*}$ is equivalent to $\mathcal{T}^{\pi^*}$.

2 Proof of Proposition 1

Proof. (of Proposition 1) This proposition follows from Theorem 1. To see why, consider any $Z \geq 0$, there is at least one optimal policy π^* defined over primitive actions that satisfies Assumption 2 with values $\alpha = 0$, $d = 1$, $\psi = 0$, arbitrary $\rho \in M(X)$, and $j = 0$. In this case, Theorem 1 gives us the following high probability ($> 1 - \delta$) bound with $\alpha = 0$:

$$\begin{aligned} ||V^* - V^{\pi_K}||_{p,\rho} &\leq \tfrac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} \left(b_{p,\mu}(\mathbb{T}\mathcal{F}, \mathcal{F}) + \alpha\right) + \varepsilon + \left(\gamma^{K+1+(1-\psi)(d-1)\lfloor Z/\bar{j}\rfloor}\right)^{1/p} \left(\tfrac{2\|V^* - V_0\|_\infty}{(1-\gamma)}\right) \\ &\leq \tfrac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} b_{p,\mu}(\mathbb{T}\mathcal{F}, \mathcal{F}) + \varepsilon + \left(\gamma^{K+1}\right)^{1/p} \left(\tfrac{2\|V^* - V_0\|_\infty}{(1-\gamma)}\right) \ , \end{aligned}$$

where $(d-1) = 0$ and $\bar{j} = j + 1 = 1$. □

3 Proof of Theorem 1 and Supporting Lemmas

The following lemma provides sufficient values for parameters N and L to ensure that the per iteration error is less than some $\varepsilon > 0$ with probability at least $1 - \delta$. We will reuse this lemma throughout our analysis.

Lemma 1. *Let $\mathcal{M}$ be an SMDP such that the set of primitive actions A is contained by the given set of options $\mathcal{O}$, $\mathcal{F} \subset B(X; V_{\text{MAX}})$ be a bounded function space with $\left(\frac{1}{8}\left(\frac{\varepsilon}{4}\right)^p, p\right)$-covering number bounded by $\mathcal{N}$, $V \in \mathcal{F}$, and p be a fixed positive integer. For any $\varepsilon, \delta > 0$,*

$$\|V' - \mathbb{T}V\|_{p,\mu} \leq d_{p,\mu}\left(\mathbb{T}V, \mathcal{F}\right) + \varepsilon$$

holds with probability at least $1 - \delta$ provided that

$$N > 128 \left(\frac{8V_{\text{MAX}}}{\varepsilon}\right)^{2p} \left(\log(1/\delta) + \log(32\mathcal{N})\right) \tag{4}$$

and

$$L > \frac{8(R_{\text{MAX}} + \gamma V_{\text{MAX}})^2}{\varepsilon^2} \left(\log(1/\delta) + \log(8N|\mathcal{O}|)\right) \ . \tag{5}$$

The proof of Lemma 1 follows from the proof of Munos and Szepesvári [2008, Lemma 1] simply by replacing the MDP Bellman operator with the SMDP Bellman operator $\mathbb{T}$ everywhere it occurs, and noting that we must sample from $|\mathcal{O}|$ options rather than only $|A|$ primitive actions. We omit the proof here for brevity.

3.1 Bounding the Pointwise Propogation Error

We are intrested in bounding the loss due to following the policy derived by OFVI φ_K rather than following the optimal policy π^* and the optimal option policy Φ^*. However, OFVI is a value-based method. That is, performing more iterations directly attempts to improve the estimate of the optimal value function. Thus, we would like to relate the loss $\|V^{\Phi^*}(x) - V^{\varphi_K}\|_{p,\rho}$ to the quality of the final value function estimate V_K produced by the OFVI algorithm. Notice that $\pi^* \equiv \Phi^*$ in our case, because we have assumed that $\mathcal{O}$ contains all primitive actions A. The following lemma develops a pointwise relationship between the $V^{\Phi^*} - V^{\varphi_K}$ and $V^{\Phi^*} - V_K$.

Lemma 2. *Suppose OFVI is executed for K iterations with iterates V_k for $k = 0, 1, 2, \ldots, K$. Let Φ^* be the optimal policy with respect to the given options $\mathcal{O}$ and φ_K be the greedy option policy with respect to the K^{th} and final iterate V_K, then*

$$V^{\Phi^*} - V^{\varphi_K} \leq (I - \widetilde{P}^{\varphi_K})^{-1} \left(\widetilde{P}^{\Phi^*} - \widetilde{P}^{\varphi_K}\right) \left(V^{\Phi^*} - V_K\right) \ , \tag{6}$$

where I is the identity matrix.

Proof. Since $\mathbb{T}V^{\Phi^*} = V^{\Phi^*}$ and $\mathbb{T}^{\varphi_K}V^{\varphi_K} = V^{\varphi_K}$, we get

$$\begin{aligned}
V^{\Phi^*} - V^{\varphi_K} &= \mathbb{T}V^{\Phi^*} - \mathbb{T}^{\varphi_K}V^{\varphi_K} \\
&= \mathbb{T}V^{\Phi^*} - \mathbb{T}^{\Phi^*}V_K + \mathbb{T}^{\Phi^*}V_K - \mathbb{T}^{\varphi_K}V^{\varphi_K} \\
&= \widetilde{P}^{\Phi^*}\left(V^{\Phi^*} - V_K\right) + \mathbb{T}^{\Phi^*}V_K - \mathbb{T}^{\varphi_K}V^{\varphi_K} \\
&= \widetilde{P}^{\Phi^*}\left(V^{\Phi^*} - V_K\right) + \mathbb{T}^{\Phi^*}V_K - \mathbb{T}V_K + \mathbb{T}V_K - \mathbb{T}^{\varphi_K}V^{\varphi_K} \\
&\leq \widetilde{P}^{\Phi^*}\left(V^{\Phi^*} - V_K\right) + \mathbb{T}V_K - \mathbb{T}^{\varphi_K}V^{\varphi_K} \\
&= \widetilde{P}^{\Phi^*}\left(V^{\Phi^*} - V_K\right) + \mathbb{T}^{\varphi_K}V_K - \mathbb{T}^{\varphi_K}V^{\varphi_K} \\
&= \widetilde{P}^{\Phi^*}\left(V^{\Phi^*} - V_K\right) + \widetilde{P}^{\varphi_K}\left(V_K - V^{\varphi_K}\right) \\
&= \widetilde{P}^{\Phi^*}\left(V^{\Phi^*} - V_K\right) + \widetilde{P}^{\varphi_K}\left(V_K - V^{\Phi^*} + V^{\Phi^*} - V^{\varphi_K}\right) \ ,
\end{aligned}$$

where the initial equality is based on the fact that V^{Φ^*} is the unique fixed point for $\mathbb{T}$ and V^{φ_K} is the unique fixed point for $\mathbb{T}^{\varphi_K}$. The first step is obtained by inserting $(-\mathbb{T}^{\Phi^*}V_K + \mathbb{T}^{\Phi^*}V_K) = 0$. The second step pulls out the discounted transition probability kernel $\widetilde{P}^{\Phi^*}$ by subtracting $\mathbb{T}^{\Phi^*}V_K$ from $\mathbb{T}V^{\Phi^*}$. Since the backups are performed by the same policy Φ^*, the immediate reward terms are canceled, leaving only $\widetilde{P}^{\Phi^*}\left(V^{\Phi^*} - V^{\varphi_K}\right)$. The third step inserts $(-\mathbb{T}V_K + \mathbb{T}V_K) = 0$. Since $\mathbb{T}^{\Phi^*}V_K \leq \mathbb{T}V_K$, we obtain the fourth step by dropping the terms $\mathbb{T}^{\Phi^*}V_K - \mathbb{T}V_K$, which is a vector whose elements are less than zero. We obtain the fifth step by noticing that since φ_K is the greedy policy with respect to V_K, $\mathbb{T}V_K = \mathbb{T}^{\varphi_K}V_K$. The sixth step pulls out $\widetilde{P}^{\varphi_K}$ by subtracting $\mathbb{T}^{\varphi_K}V^{\varphi_K}$ from $\mathbb{T}^{\varphi_K}V_K$. The seventh step inserts $(-V^{\Phi^*} + V^{\Phi^*}) = 0$.

We can manipulate the above inequality

$$\begin{aligned}
V^{\Phi^*} - V^{\varphi_K} &\leq \widetilde{P}^{\pi^*}\left(V^{\Phi^*} - V_K\right) + \widetilde{P}^{\varphi_K}\left(V_K - V^{\Phi^*} + V^{\Phi^*} - V^{\varphi_K}\right) \\
V^{\Phi^*} - V^{\varphi_K} &\leq \left(\widetilde{P}^{\Phi^*} - \widetilde{P}^{\varphi_K}\right)(V^* - V_K) + \widetilde{P}^{\varphi_K}\left(V^{\Phi^*} - V^{\varphi_K}\right) \\
\left(V^{\Phi^*} - V^{\varphi_K}\right) - \widetilde{P}^{\varphi_K}\left(V^{\Phi^*} - V^{\varphi_K}\right) &\leq \left(\widetilde{P}^{\Phi^*} - \widetilde{P}^{\varphi_K}\right)\left(V^{\Phi^*} - V_K\right) \\
\left(I - \widetilde{P}^{\varphi_K}\right)\left(V^{\Phi^*} - V^{\varphi_K}\right) &\leq \left(\widetilde{P}^{\Phi^*} - \widetilde{P}^{\varphi_K}\right)\left(V^{\Phi^*} - V_K\right) \ ,
\end{aligned}$$

where I is the identity matrix, so that the $\left(V^{\Phi^*} - V^{\varphi_K}\right)$ terms are all on the left hand side. Since $(I - \widetilde{P}^{\varphi_K})$ is invertible and its inverse is a monotonic operator, we get

$$V^{\Phi^*} - V^{\varphi_K} \leq (I - \widetilde{P}^{\varphi_K})^{-1}\left(\widetilde{P}^{\Phi^*} - \widetilde{P}^{\varphi_K}\right)\left(V^{\Phi^*} - V_K\right) \ ,$$

which relates $(V^{\Phi^*} - V^{\varphi_K})$ to $(V^{\Phi^*} - V_K)$. □

Each iteration $k = 1, 2, \ldots, K$ of OFVI results in some error

$$\varepsilon_k = \mathbb{T}V_{k-1} - V_k \ , \tag{7}$$

which is induced by the fitting process. One of the main issues in the proof of Theorem 1 is to determine how these fitting errors propagate through the iterations.

The following lemma helps to bound the error between V^* and V_K by developing pointwise upper and lower bounds for $V^* - V_K$ that show how error propogates recursively with each iteration.

Lemma 3. *Suppose Φ^* is the optimal policy with respect to the options $\mathcal{O}$, OFVI is executed for K iterates with iterates V_k for $k = 0, 1, 2, \ldots, K$ and iteration errors ε_k for $k = 1, 2, \ldots, K$ as defined by (7), then we have the following upper bound*

$$V^{\Phi^*} - V_K \leq \sum_{k=1}^{K}\left(\widetilde{P}^{\Phi^*}\right)^{K-k}\varepsilon_k + \left(\widetilde{P}^{\Phi^*}\right)^{K}(V^* - V_0) \ , \tag{8}$$

and the following lower bound

$$V^{\Phi^*} - V_K \geq \varepsilon_K + \sum_{k=1}^{K-1}\left(\widetilde{P}^{\varphi_{K-1}}\widetilde{P}^{\varphi_{K-2}}\ldots\widetilde{P}^{\varphi_k}\right)\varepsilon_k + \left(\widetilde{P}^{\varphi_{K-1}}\widetilde{P}^{\pi_{K-2}}\ldots\widetilde{P}^{\varphi_0}\right)\left(V^{\Phi^*} - V_0\right) \ . \tag{9}$$

Proof. First we derive an upper bound for $V^{\Phi^*} - V_K$. By equation (7), we have

$$\begin{aligned} V^{\Phi^*} - V_k &= \mathbb{T}V^{\Phi^*} - \mathbb{T}V_{k-1} + \varepsilon_k \\ &= \mathbb{T}^{\Phi^*}V^{\Phi^*} - \mathbb{T}^{\pi^*}V_{k-1} + \mathbb{T}^{\pi^*}V_{k-1} - \mathbb{T}V_{k-1} + \varepsilon_k \\ &\leq \mathbb{T}V^{\Phi^*} - \mathbb{T}^{\Phi^*}V_{k-1} + \varepsilon_k \\ &= \widetilde{P}^{\Phi^*}(V^* - V_{k-1}) + \varepsilon_k \ . \end{aligned}$$

By recursing on this inequality, we obtain an upper bound

$$V^{\Phi^*} - V_K \leq \sum_{k=1}^{K} \left(\widetilde{P}^{\Phi^*}\right)^{K-k} \varepsilon_k + \left(\widetilde{P}^{\Phi^*}\right)^K (V^* - V_0) \ .$$

Now we will derive a lower bound for $V^{\Phi^*} - V_K$. Let φ_k denote the greedy policy with respect to V_k. By (7), we have

$$\begin{aligned} V^{\Phi^*} - V_k &= \mathbb{T}V^{\Phi^*} - \mathbb{T}V_{k-1} + \varepsilon_k \\ &= \mathbb{T}V^{\Phi^*} - \mathbb{T}^{\varphi_{k-1}}V^{\Phi^*} + \mathbb{T}^{\varphi_{k-1}}V^{\Phi^*} - \mathbb{T}V_{k-1} + \varepsilon_k \\ &\geq \mathbb{T}^{\varphi_{k-1}}V^{\Phi^*} - \mathbb{T}V_{k-1} + \varepsilon_k \\ &= \widetilde{P}^{\varphi_{k-1}}\left(V^{\Phi^*} - V_{k-1}\right) + \varepsilon_k \ . \end{aligned}$$

By recursing on this inequality, we obtain a lower bound

$$V^{\Phi^*} - V_K \geq \varepsilon_K + \sum_{k=1}^{K-1} \left(\widetilde{P}^{\varphi_{K-1}}\widetilde{P}^{\varphi_{K-2}} \ldots \widetilde{P}^{\varphi_k}\right)\varepsilon_k + \left(\widetilde{P}^{\varphi_{K-1}}\widetilde{P}^{\pi_{K-2}} \ldots \widetilde{P}^{\varphi_0}\right)\left(V^{\Phi^*} - V_0\right) \ .$$

□

We will make use of the following definition in defining the point-wise error bound. The lambda values are used to simplify the notation, but we also use the fact that they are carefully designed so that they sum to 1.

Definition 1. *For $t = 1, 2, \ldots, \infty$, let*

$$\lambda_{0,t} = \frac{\gamma^{K-1+t}}{1-\gamma^{K+1}} \tag{10}$$

and let

$$\lambda_{k,t} = \frac{\gamma^{K-k-2+t}}{1-\gamma^{K+1}} \tag{11}$$

for $k = 1, \ldots, K$.

Lemma 4. *The $\lambda_{\cdot,\cdot}$ values defined by (10) and (11) satisfy $\sum_{t=1}^{\infty}\sum_{k=0}^{K}\lambda_{k,t} = 1$.*

Proof.

$$\begin{aligned} \textstyle\sum_{t=1}^{\infty}\sum_{k=0}^{K}\lambda_{k,t} &= \textstyle\sum_{t=1}^{\infty}\frac{\gamma^{K-1+t}}{1-\gamma^{K+1}} + \sum_{k=1}^{K}\frac{\gamma^{K-k-2+t}}{1-\gamma^{K+1}} \\ &= \left(\frac{1}{1-\gamma^{K+1}}\right)\textstyle\sum_{t=1}^{\infty}\gamma^{K+(t-1)} + \sum_{k=1}^{K}\gamma^{K-k-1+(t-1)} \\ &= \left(\frac{1}{1-\gamma^{K+1}}\right)\left(\gamma^K + \textstyle\sum_{k=1}^{K}\gamma^{K-k-1}\right)\left(\sum_{t=0}^{\infty}\gamma^t\right) \\ &= \left(\frac{1}{1-\gamma^{K+1}}\right)\left(\gamma^K + \textstyle\sum_{k=0}^{K-1}\gamma^{K-k-1+1}\right)(1-\gamma) \\ &= \left(\frac{1}{1-\gamma^{K+1}}\right)\left(\textstyle\sum_{k=0}^{K}\gamma^k\right)(1-\gamma) \\ &= \left(\frac{1}{1-\gamma^{K+1}}\right)\left(\textstyle\sum_{k=0}^{K}\gamma^k - \gamma^{k+1}\right) \\ &= \left(\frac{1}{1-\gamma^{K+1}}\right)(1-\gamma^{K+1}) \\ &= 1 \ . \end{aligned}$$

□

Now we are ready to define and prove the point-wise error bound.

Lemma 5. *Let $Z \in \{0, 1, 2, \ldots, K\}$, φ_k be the greedy policy with respect to the k^{th} iterate V_k derived by OFVI, and Φ be an option policy such that $Q^*(x, \Phi(x)) \geq V^*(x) - \alpha$ for all $x \in X$. If $A2(\alpha, d, \psi, \rho, j)$ (Assumption 2) is true and the first Z iterates of OFVI are pessimistic (i.e., for all $x \in X$ and $k \in \{0, 1, 2, \ldots, Z\}$, $V^*(x) \geq V_k(x)$), then the difference between V^* and the value of the option policy φ_K returned by OFVI is bounded by*

$$V^* - V^{\varphi_K} \leq \left(\frac{2\gamma(1-\gamma^{K+1})}{1-\gamma}\right) \left\{\sum_{t=1}^{\infty} \sum_{k=0}^{K} \lambda_{k,t} P_{k,t} |\xi_k|\right\} \ ,$$

where the $\lambda_{k,t}$'s are defined by (10) and (11),

$$P_{k,t} = \left(\frac{1-\gamma}{2}\right)(I - \widetilde{P}^{\varphi_K})^{-1} \begin{cases} 2\left[\left(P^{\pi^*}\right)^{K-Z}\left(P^{\Phi}\right)^{Z-k}\right]_t & 0 \leq k \leq Z \\ \left[\left(P^{\pi^*}\right)_t^{K-k} + \left(P^{\varphi_{K-1}} P^{\varphi_{K-2}} \ldots P^{\varphi_k}\right)\right] & Z < k < K \end{cases}$$

for $t \geq 1$, and

$$\xi_k = \begin{cases} V^* - V_0 & k = 0 \\ \varepsilon_k + \alpha & 1 \leq k \leq Z \\ \varepsilon_k & Z < k < K \end{cases} \ .$$

Proof. We can place an upper bound (8) and a lower bound (9) on the relationship between V_K and V^*. Then we can use this information to bound the difference between V^{φ_K} and V^*. However, in this lemma, we will exploit the pessimism of the first Z iterates and the option policy Φ to achieve a more informative bound.

Let us denote by $\Phi\pi^*$ the policy that from the first encountered state selects an option according to Φ and from then on always selects an action according to π^*. When an iterate V_k is pessimistic $V^* - V_k$ is lower bounded by 0. For an upper bound, we have

$$\begin{aligned} V^* - V_k &= V^* - V^{\Phi\pi^*} + V^{\Phi\pi^*} - V_k \\ &\leq \alpha + V^{\Phi\pi^*} - V_k \\ &= \alpha + \mathbb{T}^{\Phi\pi^*} V^{\Phi\pi^*} - \mathbb{T} V_{k-1} + \varepsilon_k \\ &= \alpha + \mathbb{T}^{\Phi\pi^*} V^{\Phi\pi^*} - \mathbb{T}^{\Phi\pi^*} V_{k-1} + \mathbb{T}^{\Phi\pi^*} V_{k-1} - \mathbb{T} V_{k-1} + \varepsilon_k \\ &\leq \alpha + \mathbb{T}^{\Phi\pi^*} V^{\Phi\pi^*} - \mathbb{T}^{\Phi\pi^*} V_{k-1} + \varepsilon_k \\ &\leq \widetilde{P}^{\Phi}\left(V^* - V_{k-1}\right) + (\varepsilon_k + \alpha) \ , \end{aligned}$$

where the initial inequality inserts the term $(-V^{\Phi\pi^*} + V^{\Phi\pi^*}) = 0$. The first step follows from the fact that $\Phi\pi^*$ is an α-optimal policy, so $V^* - V^{\Phi\pi^*} \leq \alpha$. The second step is due to the definition of ε_k from (7). The third step inserts $(-\mathbb{T}^{\Phi\pi^*} V_{k-1} + \mathbb{T}^{\Phi\pi^*} V_{k-1}) = 0$. The fourth step removes $\mathbb{T}^{\Phi\pi^*} V_{k-1} - \mathbb{T} V_{k-1}$ because the sum of those two terms is less than or equal to zero (since $\mathbb{T}$ updates using the max operator, while $\mathbb{T}^{\Phi\pi^*}$ updates using the policy Φ). The fifth and final step pulls out the discounted transition probability kernel $\widetilde{P}^{\Phi}$.

By recursing on this inequality $Z \geq 0$ times we obtain

$$V^* - V_Z \leq \begin{cases} V^* - V_0 & Z = 0 \\ \left(\sum_{j=1}^{Z} \left(\widetilde{P}^{\Phi}\right)^{Z-j} (\varepsilon_j + \alpha)\right) + \left(\widetilde{P}^{\Phi}\right)^{Z} (V^* - V_0) & 1 \leq Z \leq K \end{cases} \ . \tag{12}$$

By combining our upper bound recursion from (8) with (12), we obtain terms

$$u_k \xi_k = \begin{cases} \left[\left(\widetilde{P}^{\pi^*}\right)^{K-Z}\left(\widetilde{P}^{\Phi}\right)^{Z}\right](V^* - V_0) & k = 0 \\ \left[\left(\widetilde{P}^{\pi^*}\right)^{K-Z}\left(\widetilde{P}^{\Phi}\right)^{Z-k}\right](\varepsilon_k + \alpha) & k = 1, 2, \ldots, Z \\ \left[\left(\widetilde{P}^{\pi^*}\right)^{K-k}\right](\varepsilon_k) & k = Z+1, Z+2, \ldots, K \end{cases}$$

such that

$$V^* - V_K \leq \sum_{k=0}^{K} u_k \xi_k$$

upper bounds the difference between V^* and the final iterate derived by OFVI, V_K.

Now, since 0 lower bounds the difference between V^* and the first Z iterates of OFVI, we can use 0 as our lower bound for the first Z iterations and fill in the rest of the iterates with (9). This gives us the terms

$$l_k \xi_k = \begin{cases} 0 & 0 \leq k \leq Z \\ \left[\widetilde{P}^{\varphi_{K-1}} \widetilde{P}^{\varphi_{K-2}} \dots \widetilde{P}^{\varphi_k}\right](\varepsilon_k) & Z < k < K-1 \end{cases},$$

such that

$$V^* - V_K \geq \sum_{k=0}^{K} l_k \xi_k$$

lower bounds the difference between V^* and the final iterate V_K. This implies that $|V^* - V_K| \leq \sum_{k=0}^{K}(u_k - l_k)\xi_k$.

By Lemma 2, we have

$$\begin{aligned} V^* - V^{\varphi_K} &\leq (I - \widetilde{P}^{\varphi_K})^{-1}\left(\widetilde{P}^{\pi^*} - \widetilde{P}^{\varphi_K}\right)\left(\textstyle\sum_{k=0}^{K}(u_k - l_k)\xi_k\right) \\ &\leq (I - \widetilde{P}^{\varphi_K})^{-1}\left|\widetilde{P}^{\pi^*} - \widetilde{P}^{\varphi_K}\right|\left(\textstyle\sum_{k=0}^{K}(u_k + l_k)|\xi_k|\right), \end{aligned}$$

where we have taken the absolute value of both sides of the inequality.

For $k = 0$, we have

$$\begin{aligned} & (I - \widetilde{P}^{\varphi_K})^{-1}\left|\widetilde{P}^{\pi^*} - \widetilde{P}^{\varphi_K}\right|((u_0 + l_0)|\xi_0|) \\ \leq\ & \gamma\left(\tfrac{1-\gamma}{1-\gamma}\right)\left(\tfrac{2}{2}\right)(I - \widetilde{P}^{\varphi_K})^{-1}(u_0|\xi_0|) \\ =\ & \left(\tfrac{\gamma}{1-\gamma}\right)\left(\tfrac{1-\gamma}{2}\right)(I - \widetilde{P}^{\varphi_K})^{-1}\left(2\left[\left(\widetilde{P}^{\pi^*}\right)^{K-Z}\left(\widetilde{P}^{\Phi}\right)^{Z}\right]\right)|\xi_0| \\ =\ & \left(\tfrac{\gamma}{1-\gamma}\right)\left(\tfrac{1-\gamma}{2}\right)(I - \widetilde{P}^{\varphi_K})^{-1}\left(2\sum_{t=1}^{\infty}\gamma^{K-1+t}\left[\left(P^{\pi^*}\right)^{K-Z}\left(P^{\Phi}\right)^{Z}\right]_t\right)|\xi_0| \\ =\ & \left(\tfrac{\gamma}{1-\gamma}\right)\sum_{t=1}^{\infty}\gamma^{K-1+t}\left(\left(\tfrac{1-\gamma}{2}\right)(I - \widetilde{P}^{\varphi_K})^{-1}\left(2\left[\left(P^{\pi^*}\right)^{K-Z}\left(P^{\Phi}\right)^{Z}\right]_t\right)\right)|\xi_0| \\ =\ & \left(\tfrac{\gamma}{1-\gamma}\right)\sum_{t=1}^{\infty}\gamma^{K-1+t}P_{0,t}|\xi_0| \\ =\ & \left(\tfrac{\gamma}{1-\gamma}\right)\sum_{t=1}^{\infty}\gamma^{K-1+t}\left(\tfrac{1-\gamma^{K+1}}{1-\gamma^{K+1}}\right)P_{0,t}|\xi_0| \\ =\ & \left(\tfrac{\gamma(1-\gamma^{K+1})}{1-\gamma}\right)\sum_{t=1}^{\infty}\tfrac{\gamma^{K-1+t}}{1-\gamma^{K+1}}P_{0,t}|\xi_0| \\ =\ & \left(\tfrac{\gamma(1-\gamma^{K+1})}{1-\gamma}\right)\sum_{t=1}^{\infty}\lambda_{0,t}P_{0,t}|\xi_0| \\ \leq\ & \left(\tfrac{2\gamma(1-\gamma^{K+1})}{1-\gamma}\right)\sum_{t=1}^{\infty}\lambda_{0,t}P_{0,t}|\xi_0| \end{aligned}$$

For $k = 1, 2, \dots, Z$, we have

$$(I - \widetilde{P}^{\varphi_K})^{-1}\left|\widetilde{P}^{\pi^*} - \widetilde{P}^{\varphi_K}\right|((u_k + l_k)|\xi_k|)$$

$$
\begin{aligned}
&\leq \gamma\left(\tfrac{1-\gamma}{1-\gamma}\right)\left(\tfrac{2}{2}\right)(I-\widetilde{P}^{\varphi_K})^{-1}(u_k|\xi_k|)\\
&= \left(\tfrac{\gamma}{1-\gamma}\right)\left(\tfrac{1-\gamma}{2}\right)(I-\widetilde{P}^{\varphi_K})^{-1}\left(2\left[\left(\widetilde{P}^{\pi^*}\right)^{K-Z}\left(\widetilde{P}^{\Phi}\right)^{Z-k}\right]\right)|\xi_k|\\
&= \left(\tfrac{\gamma}{1-\gamma}\right)\left(\tfrac{1-\gamma}{2}\right)(I-\widetilde{P}^{\varphi_K})^{-1}\left(2\sum_{t=1}^{\infty}\gamma^{K-k-2+t}\left[\left(P^{\pi^*}\right)^{K-Z}\left(P^{\Phi}\right)^{Z-k}\right]_t\right)|\xi_k|\\
&= \left(\tfrac{\gamma}{1-\gamma}\right)\sum_{t=1}^{\infty}\gamma^{K-k-2+t}\left(\left(\tfrac{1-\gamma}{2}\right)(I-\widetilde{P}^{\varphi_K})^{-1}\left(2\left[\left(P^{\pi^*}\right)^{K-Z}\left(P^{\Phi}\right)^{Z-k}\right]_t\right)\right)|\xi_k|\\
&= \left(\tfrac{\gamma}{1-\gamma}\right)\sum_{t=1}^{\infty}\gamma^{K-k-2+t}P_{k,t}|\xi_k|\\
&= \left(\tfrac{\gamma(1-\gamma^{K+1})}{1-\gamma}\right)\sum_{t=1}^{\infty}\lambda_{k,t}P_{k,t}|\xi_k|\\
&\leq \left(\tfrac{2\gamma(1-\gamma^{K+1})}{1-\gamma}\right)\sum_{t=1}^{\infty}\lambda_{k,t}P_{k,t}|\xi_k|
\end{aligned}
$$

For $k = Z+1, Z+2, \ldots, K$, we have

$$
(I-\widetilde{P}^{\varphi_K})^{-1}\left|\widetilde{P}^{\pi^*}-\widetilde{P}^{\varphi_K}\right|((u_k+l_k)|\xi_k|)
$$

$$
\begin{aligned}
&\leq \gamma\left(\tfrac{1-\gamma}{1-\gamma}\right)\left(\tfrac{2}{2}\right)(I-\widetilde{P}^{\varphi_K})^{-1}(u_k+l_k)|\xi_k|\\
&= \gamma\left(\tfrac{1-\gamma}{1-\gamma}\right)\left(\tfrac{2}{2}\right)(I-\widetilde{P}^{\varphi_K})^{-1}\left(\left(\widetilde{P}^{\pi^*}\right)^{K-k}+\left[\widetilde{P}^{\varphi_{K-1}}\widetilde{P}^{\varphi_{K-2}}\ldots\widetilde{P}^{\varphi_k}\right]\right)|\xi_k|\\
&= \left(\tfrac{2\gamma}{1-\gamma}\right)\left(\tfrac{1-\gamma}{2}\right)(I-\widetilde{P}^{\varphi_K})^{-1}\left(\left(\widetilde{P}^{\pi^*}\right)^{K-k}+\left[\widetilde{P}^{\varphi_{K-1}}\widetilde{P}^{\varphi_{K-2}}\ldots\widetilde{P}^{\varphi_k}\right]\right)|\xi_k|\\
&= \left(\tfrac{2\gamma}{1-\gamma}\right)\left(\tfrac{1-\gamma}{2}\right)(I-\widetilde{P}^{\varphi_K})^{-1}\sum_{t=1}^{\infty}\gamma^{K-k-2+t}\left(\left(P^{\pi^*}\right)_t^{K-k}+\left[P^{\varphi_{K-1}}P^{\varphi_{K-2}}\ldots P^{\varphi_k}\right]_t\right)|\xi_k|\\
&= \left(\tfrac{2\gamma}{1-\gamma}\right)\sum_{t=1}^{\infty}\gamma^{K-k-2+t}\left(\left(\tfrac{1-\gamma}{2}\right)(I-\widetilde{P}^{\varphi_K})^{-1}\left(\left(P^{\pi^*}\right)_t^{K-k}+\left[P^{\varphi_{K-1}}P^{\varphi_{K-2}}\ldots P^{\varphi_k}\right]_t\right)\right)|\xi_k|\\
&= \left(\tfrac{2\gamma}{1-\gamma}\right)\sum_{t=1}^{\infty}\gamma^{K-k-2+t}P_{k,t}|\xi_k|\\
&= \left(\tfrac{2\gamma(1-\gamma^{K+1})}{1-\gamma}\right)\sum_{t=1}^{\infty}\lambda_{k,t}P_{k,t}|\xi_k|
\end{aligned}
$$

□

3.2 From Pointwise to L_p-norm Propogation Error

To convert the point-wise error bound from Lemma 5 into a bound with respect to norms, we need to consider how each iteration affects the next state probability distribution. This is where Assumption 1 is needed, because it limits the difference between future state distributions starting from the initial state distribution ρ and the sampling distribution μ.

Lemma 6. *Suppose that $A1(\rho,\mu)$ (Assumption 1) holds, then*

$$
\rho\sum_{t=1}^{\infty}P_{k,t} \leq (1-\gamma)\sum_{t=1}^{\infty}\sum_{j=0}^{\infty}\gamma^j c(j+K-k+t-1)\mu \ , \tag{13}
$$

where $\rho,\mu \in M(X)$.

Proof. We have two cases to consider (case 1) $1 \leq k \leq Z$ and (case 2) $Z < k \leq K$.

For case 1, we have

$$\begin{aligned}
\rho \textstyle\sum_{t=1}^{\infty} P_{k,t} &= \rho \textstyle\sum_{t=K-k}^{\infty} \left(\frac{1-\gamma}{2}\right) \left(I - \widetilde{P}^{\varphi_K}\right)^{-1} 2 \left[\left(P^{\pi^*}\right)^{K-Z} \left(P^{\varphi}\right)^{Z-k}\right]_t \\
&= \rho \textstyle\sum_{t=K-k}^{\infty} \left(\frac{1-\gamma}{2}\right) \left(\sum_{i=0}^{\infty} \left(\widetilde{P}^{\varphi_K}\right)^i\right) 2 \left[\left(P^{\pi^*}\right)^{K-Z} \left(P^{\varphi}\right)^{Z-k}\right]_t \\
&= \rho \textstyle\sum_{t=K-k}^{\infty} \left(\frac{1-\gamma}{2}\right) \left(\sum_{i=0}^{\infty} \sum_{j=i}^{\infty} \gamma^j \left(P^{\varphi_K}\right)_j^i\right) 2 \left[\left(P^{\pi^*}\right)^{K-Z} \left(P^{\varphi}\right)^{Z-k}\right]_t \\
&= \textstyle\sum_{t=K-k}^{\infty} \left(\frac{1-\gamma}{2}\right) \sum_{i=0}^{\infty} \sum_{j=i}^{\infty} \gamma^j 2 \left[\rho \left(P^{\varphi_K}\right)_j^i \left(P^{\pi^*}\right)^{K-Z} \left(P^{\varphi}\right)^{Z-k}\right]_t \\
&\leq \left(\frac{1-\gamma}{2}\right) \textstyle\sum_{t=K-k}^{\infty} \sum_{i=0}^{\infty} \sum_{j=i}^{\infty} \gamma^j \left[2c(t+j)\mu\right]
\end{aligned}$$

For case 2, we have

$$\begin{aligned}
\rho \textstyle\sum_{t=1}^{\infty} P_{k,t} &= \rho \textstyle\sum_{t=K-k}^{\infty} \left(\frac{1-\gamma}{2}\right) \left(I - \widetilde{P}^{\varphi_K}\right)^{-1} \left[\left(P^{\pi^*}\right)_t^{K-k} + \left(P^{\varphi_{K-1}} P^{\varphi_{K-2}} \dots P^{\varphi_k}\right)_t\right] \\
&= \textstyle\sum_{t=K-k}^{\infty} \rho \left(\frac{1-\gamma}{2}\right) \left(\sum_{i=0}^{\infty} \left(\widetilde{P}^{\varphi_K}\right)^i\right) \left[\left(P^{\pi^*}\right)_t^{K-k} + \left(P^{\varphi_{K-1}} P^{\varphi_{K-2}} \dots P^{\varphi_k}\right)_t\right] \\
&= \textstyle\sum_{t=K-k}^{\infty} \rho \left(\frac{1-\gamma}{2}\right) \left(\sum_{i=0}^{\infty} \sum_{j=i}^{\infty} \gamma^j \left(P^{\varphi_K}\right)_j^i\right) \left[\left(P^{\pi^*}\right)_t^{K-k} + \left(P^{\varphi_{K-1}} P^{\varphi_{K-2}} \dots P^{\varphi_k}\right)_t\right] \\
&= \left(\frac{1-\gamma}{2}\right) \textstyle\sum_{t=K-k}^{\infty} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \gamma^{i+j} \left[\rho \left(P^{\varphi_K}\right)_j^i \left(P^{\pi^*}\right)_t^{K-k} + \rho \left(P^{\varphi_K}\right)_j^i \left(P^{\varphi_{K-1}} P^{\varphi_{K-2}} \dots P^{\varphi_k}\right)_t\right] \\
&\leq \left(\frac{1-\gamma}{2}\right) \textstyle\sum_{t=K-k}^{\infty} \sum_{i=0}^{\infty} \sum_{j=i}^{\infty} \gamma^j \left[2c(t+j)\mu\right]
\end{aligned}$$

Both cases are bounded by

$$\begin{aligned}
\left(\frac{1-\gamma}{2}\right) \sum_{t=K-k}^{\infty} \sum_{i=0}^{\infty} \sum_{j=i}^{\infty} \gamma^j \left[2c(t+j)\mu\right] &\leq (1-\gamma) \sum_{t=K-k}^{\infty} \sum_{j=0}^{\infty} \gamma^j c(t+j)\mu \\
&= (1-\gamma) \sum_{t=1}^{\infty} \sum_{j=0}^{\infty} \gamma^j c(j+K-k+t-1)\mu \ ,
\end{aligned}$$

which concludes our proof. □

Now the core idea behind our analysis is to find a policy defined over the option set $\mathcal{O}$ that selects temporally extended actions and we know converges quickly. Then we use Lemma 5 to obtain point-wise bounds and Lemma 6 to obtain bounds with respect to norms. The following defines the policy that we will use in our analysis.

Definition 2. *Let $d \geq 1$, $x \in X$ be a state, and $\mathcal{O}$ be a set of options. The set $\mathcal{O}_{x,d}$ denotes the subset of options $o \in \mathcal{O}$ that can be initialized from the state x, such that $\inf_{Y \subseteq X} \mathbb{E}\left[D_{x,Y}^o\right] \geq d$.*

Suppose that $A2(\alpha, d, \psi, \rho, j)$ holds for some $\alpha \geq 0$, $d \geq 1$, $j \geq 0$, $\psi \geq 0$, and $\rho \in M(X)$. We define the policy

$$\Phi(x) = \begin{cases} \arg\max_{o \in \mathcal{O}_{x,d}} Q^*(x,o) & \text{if } x \in \omega_{\alpha,d} \\ \hat{\pi} & \text{otherwise} \end{cases} \tag{14}$$

where $\hat{\pi}$ is the α-optimal "bridge" policy defined in $A2(\alpha, d, \psi, \rho, j)$.

Now we are ready to transform the point-wise bound from Lemma 5 to a bound with respect to p-norms.

Lemma 7. *Let $K, p \geq 1$, $\varepsilon > 0$, $\alpha, \psi, j \geq 0$, $\rho, \mu \in M(X)$ and $Z \in \{1, 2, \dots, K\}$. Suppose that $A1(\rho, \mu)$ and $A2(\alpha, d, \psi, \rho, j)$ hold, and the first Z iterates of OFVI are pessimistic, then*

$$\|V^* - V^{\varphi_K}\| \leq \frac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} (\varepsilon + \alpha) + \left(\gamma^{K+1+(1-\psi)(d-1)\lfloor Z/j \rfloor}\right)^{1/p} \left(\frac{2\|V^* - V_0\|_\infty}{1-\gamma}\right) \tag{15}$$

holds, provided that the approximation errors ε_k satisfy $\|\varepsilon_k\|_{p,\mu} \leq \varepsilon$ for all $k = 1, 2, \dots, K$.

Proof. First note that

$$\Phi(x) = \begin{cases} \arg\max_{o \in \mathcal{O}_{x,d}} Q^*(x,o) & \text{if } x \in \omega_{\alpha,d} \\ \pi^*(x) & \text{otherwise} \end{cases} .$$

is a policy such that $Q^*(x, \Phi(x)) \geq V^*(x) - \alpha$ for all $x \in X$. Therefore, by Lemma 5, we have

$$V^* - V^{\varphi_K} \leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)} \left[\sum_{t=1}^{\infty} \sum_{k=0}^{K} \lambda_{k,t} P_{k,t} |\xi_k| \right] .$$

Now, we have

$$\begin{aligned} \|V^* - V^{\varphi_K}\|_{p,\rho}^p &= \int \rho(x) \left| V^*(x) - V^{\pi_K}(x) \right|^p dx \\ &\leq \int \rho(x) \left(\tfrac{2\gamma(1-\gamma^{K+1})}{1-\gamma} \left[\textstyle\sum_{t=1}^{\infty} \sum_{k=1}^{K} \lambda_{k,t} P_{k,t} |\varepsilon_k + \alpha| + \lambda_{0,t} P_{0,t} |V^* - V_0| \right](x) \right)^p dx \\ &\leq \left(\tfrac{2\gamma(1-\gamma^{K+1})}{1-\gamma} \right)^p \int \rho(x) \left(\left[\textstyle\sum_{t=1}^{\infty} \sum_{k=1}^{K} \lambda_{k,t} P_{k,t} |\varepsilon_k + \alpha| + \lambda_{0,t} P_{0,t} |V^* - V_0| \right](x) \right)^p dx . \end{aligned}$$

Recall by Lemma 4 that $\sum_{t=1}^{\infty} \sum_{k=0}^{K} \lambda_{k,t} = 1$. By applying Jensen's inequality where $|\cdot|^p$ is the convex function, $\lambda_{t,k}$ for $k = 0, 1, \ldots, K$ and $t \geq 0$ are the parameters, and noticing that $\sum_{t=1}^{\infty} P_{k,t}$ are stochastic operators, we obtain

$$\|V^* - V^{\varphi_K}\|_{p,\rho}^p \leq \left(\frac{2\gamma(1-\gamma^{K+1})}{1-\gamma} \right)^p \int \rho \left[\sum_{t=1}^{\infty} \sum_{k=1}^{K} \lambda_{k,t} P_{k,t} |\varepsilon_k + \alpha|^p + \lambda_{0,t} P_{0,t} |V^* - V_0|^p \right](x) dx .$$

Noticing that $|V^* - V_0|$ is bounded by $\|V^* - V_0\|_\infty$, we obtain

$$\begin{aligned} \|V^* - V^{\varphi_K}\|_{p,\rho}^p \leq & \left(\tfrac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)} \right)^p \left[\textstyle\sum_{t=1}^{\infty} \sum_{k=1}^{K} \lambda_{k,t} P_{k,t} |\varepsilon_k + \alpha|^p + \right. \\ & \left. \int \rho(x) \lambda_{0,t} P_{0,t} \|V^* - V_0\|_\infty^p dx \right] . \end{aligned}$$

By Assumption 1 and Lemma 6, we have that

$$\rho \sum_{t=1}^{\infty} P_{k,t} \leq (1-\gamma) \sum_{t=1}^{\infty} \sum_{j=0}^{\infty} \gamma^j c(j+K-k+t-1)\mu .$$

Thus we have

$$\begin{aligned} \sum_{k=1}^{K} \sum_{t=1}^{\infty} \lambda_{k,t} \rho P_{k,t} |\varepsilon_k + \alpha|^p \leq & \sum_{k=1}^{K} \sum_{t=1}^{\infty} \lambda_{k,t} (1-\gamma) \sum_{j=0}^{\infty} \gamma^j \cdot \\ & c(j+K-k+t-1) \|\varepsilon_k + \alpha\|_{p,\mu}^p \\ \leq & \sum_{k=1}^{K} \sum_{t=1}^{\infty} \tfrac{\gamma^{K-k-2+t}}{1-\gamma^{K+1}} (1-\gamma) \sum_{j=0}^{\infty} \gamma^j \cdot \\ & c(j+K-k+t-1) \|\varepsilon_k + \alpha\|_{p,\mu}^p \\ \leq & \tfrac{1}{1-\gamma^{K+1}} (1-\gamma) \sum_{j=0}^{\infty} \sum_{k=1}^{K} \sum_{t=1}^{\infty} \gamma^{j+K-k-2+t} \cdot \\ & c(j+K-k+t-1) \|\varepsilon_k + \alpha\|_{p,\mu}^p \\ \leq & \tfrac{1}{1-\gamma^{K+1}} (1-\gamma) \sum_{j=0}^{\infty} \sum_{k=0}^{K-1} \sum_{t=1}^{\infty} \gamma^{j+K-k+t-1} \cdot \\ & c(j+K-k+t) \|\varepsilon_k + \alpha\|_{p,\mu}^p \\ \leq & \tfrac{1}{(1-\gamma)(1-\gamma^{K+1})} (1-\gamma)^2 \sum_{t=1}^{\infty} t \gamma^{t-1} c(t) \|\varepsilon_k + \alpha\|_{p,\mu}^p \\ \leq & \tfrac{1}{(1-\gamma)(1-\gamma^{K+1})} C_{\rho,\mu} (\varepsilon + \alpha)^p , \end{aligned}$$

where $C_{\rho,\mu}$ is the discounted average concentrability coefficient from Assumption 1. By replacing $\sum_{t=1}^{\infty}\sum_{k=0}^{K-1}\lambda_{k,t}P_{k,t}|\varepsilon_k+\alpha|^p$, we get

$$\begin{aligned}\|V^* - V^{\varphi_K}\|_{p,\rho}^p &\leq \left(\tfrac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)}\right)^p \left[\tfrac{1}{(1-\gamma)(1-\gamma^{K+1})}C_{\rho,\mu}\,(\varepsilon+\alpha)^p + \right. \\ &\qquad \left. \textstyle\sum_{t=1}^{\infty}\int \rho(x)\lambda_{0,t}P_{0,t}\|V^* - V_0\|_{\infty}^p dx\right] .\end{aligned} \tag{16}$$

Consider the second term in the last step of the previous inequality. By replacing $P_{0,t}$ with its definition, we get

$$\begin{aligned}\int \rho(x)\sum_{t=1}^{\infty}\lambda_{0,t}P_{0,t}dx &= \int \rho(x)\sum_{t=1}^{\infty}\tfrac{\gamma^{K+t}}{1-\gamma^{K+1}}\tfrac{1-\gamma}{2}(I-\widetilde{P}^{\varphi_k})^{-1}\left[2\left((P^{\pi^*})^{K-Z}\left(P^{\Phi}\right)^{Z}\right)_t\right]dx \\ &\leq \tfrac{1}{1-\gamma^{K+1}}\int \rho(x)\left[(\widetilde{P}^{\pi^*})^{K-Z}\left(\widetilde{P}^{\Phi}\right)^{Z}\right]dx\ .\end{aligned}$$

By $A2(\alpha,d,\psi,\rho,j)$, the policy Φ reaches a state in $\omega_{\alpha,d}$ at least once every j timesteps with probability at least $1-\psi$. Thus during the first Z iterations, we have

$$\int \rho(x)\sum_{t=1}^{\infty}\lambda_{0,t}P_{0,t}dx \leq \tfrac{1}{1-\gamma^{K+1}}\gamma^{K-Z}\gamma^{Z+(1-\psi)(d-1)\lfloor Z/j\rfloor}\ .$$

By replacing the second term from the inequality above, we get

$$\begin{aligned}\|V^* - V^{\varphi_K}\|_{p,\rho}^p &\leq \left(\tfrac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)}\right)^p \left[\tfrac{1}{(1-\gamma)(1-\gamma^{K+1})}C_{\rho,\mu}\,(\varepsilon+\alpha)^p + \right. \\ &\qquad \left. \tfrac{1}{1-\gamma^{K+1}}\gamma^{K-Z}\gamma^{Z+(1-\psi)(d-1)\lfloor Z/j\rfloor}\|V^* - V_0\|_{\infty}^p\right] .\end{aligned}$$

Since $\left(1-\gamma^{K+1}\right)^p\left(\frac{1}{1-\gamma^{K+1}}\right)\leq 1$, then

$$\|V^* - V^{\varphi_K}\|_{p,\rho}^p \leq \left(\tfrac{2\gamma}{(1-\gamma)}\right)^p\left[\tfrac{1}{1-\gamma}C_{\rho,\mu}\,(\varepsilon+\alpha)^p + \gamma^{K-Z}\gamma^{Z+(1-\psi)(d-1)\lfloor Z/j\rfloor}\|V^* - V_0\|_{\infty}^p\right] .$$

Thus, we have

$$\|V^* - V^{\varphi_K}\|_{p,\rho} \leq \tfrac{2\gamma}{(1-\gamma)^2}C_{\rho,\mu}^{1/p}(\varepsilon+\alpha) + \left(\gamma^{(K+1-Z)}\gamma^{Z+(1-\psi)(d-1)\lfloor Z/j\rfloor}\right)^{1/p}\left(\tfrac{2\|V^*-V_0\|_{\infty}}{(1-\gamma)}\right) .$$

□

3.3 Proof of Theorem 1

Proof. (of Theorem 1)

We use Lemma 1 to select appropriate values for N and L, such that $\varepsilon' = \epsilon(1-\gamma)^2/(2\gamma C_{\rho,\mu}^{1/p})$ and $\delta' \leftarrow \frac{\delta}{K}$.

Since the iterates $V_1, V_2, \ldots, V_K$ are random objects, we cannot directly apply Lemma 1 to bound the error at each iteration. However, this problem was resolved in the proof of Munos and Szepesvári [2008, Theorem 2] by using the fact that the algorithm collects independent samples at each iteration.

The iterate V_{k+1} depends on the random variable V_k and the random samples S_k containing the $N\times L\times|\mathcal{O}|$ next states, rewards, and trajectory lengths. Let the function

$$f(S_k,V_k) = \mathbb{I}\left\{\|V_{k+1}(V_k,S_k) - \mathbb{T}V_k\|_{p,\mu} \leq b_{p,\mu}(\mathbb{T}V_k,\mathcal{F}) + \varepsilon'\right\} - (1-\delta')\ ,$$

where we have written $V_{k+1}(V_k,S_k)$ to emphasize V_{k+1}'s dependence on both random variables V_k and S_k. Notice that V_k and S_k are independent because S_k was not used to generate V_k and the simulator $\mathbb{S}$ generates independent samples. Because V_k and S_k are independent random variables, we can apply [Munos and Szepesvári, 2008, Lemma 5]. This lemma tells us that $\mathbb{E}\left[f(S_k,V_k)\mid V_k\right]\geq 0$ provided

that $\mathbb{E}[f(S_k, v)] \geq 0$ for all $v \in \mathcal{F}$. For any $v \in \mathcal{F}$, by Lemma 1, and by our choice of N and L, we have that $P\left(\|V_{k+1}(v, S_k) - \mathbb{T}v\|_{p,\mu} \leq b_{p,\mu}(\mathbb{T}v, \mathcal{F}) + \varepsilon'\right) \geq 1 - \delta'$. This implies that $\mathbb{E}[f(S_k, v)] \geq 0$. By Munos and Szepesvári [2008, Lemma 5], we have that $\mathbb{E}[f(S_k, V_k) \mid V_k] \geq 0$. Thus we have $P\left(\|V_{k+1}(V_k, S_k) - \mathbb{T}V_k\|_{p,\mu} \leq b_{p,\mu}(\mathbb{T}v, \mathcal{F}) + \varepsilon'\right) \geq 1 - \delta'$. By the union bound, this ensures that $\|\varepsilon\|_{p,\mu} \leq \varepsilon$ for all K iterations with probability at least $1 - K\delta' = 1 - K(\delta/K) = 1 - \delta$.

The result follows by applying Lemma 7 with $\|\epsilon_k\|_{p,\mu} \leq b_{p,\mu}(\mathbb{T}V_k, \mathcal{F}) + \varepsilon'$.

$$\begin{aligned}
\|V^* - V^{\varphi_K}\|_{p,\rho} &\leq \tfrac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} \left(b_{p,\mu}(\mathbb{T}\mathcal{F}, \mathcal{F}) + \alpha + \varepsilon'\right) + \left(\gamma^{K+1+(1-\psi)(d-1)\lfloor Z/\bar{j}\rfloor}\right)^{1/p} \left(\tfrac{2\|V^* - V_0\|_\infty}{(1-\gamma)}\right) \\
&= \tfrac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} \left(b_{p,\mu}(\mathbb{T}\mathcal{F}, \mathcal{F}) + \alpha + \epsilon(1-\gamma)^2/(2\gamma C_{\rho,\mu}^{1/p})\right) \\
&\quad + \left(\gamma^{K+1+(1-\psi)(d-1)\lfloor Z/\bar{j}\rfloor}\right)^{1/p} \left(\tfrac{2\|V^* - V_0\|_\infty}{(1-\gamma)}\right) \\
&= \tfrac{2\gamma}{(1-\gamma)^2} C_{\rho,\mu}^{1/p} \left(b_{p,\mu}(\mathbb{T}\mathcal{F}, \mathcal{F}) + \alpha\right) + \epsilon + \left(\gamma^{K+1+(1-\psi)(d-1)\lfloor Z/\bar{j}\rfloor}\right)^{1/p} \left(\tfrac{2\|V^* - V_0\|_\infty}{(1-\gamma)}\right) .
\end{aligned}$$

□

4 Experiment Details

In this section, we specify additional details and parameters from our experiments.

4.1 Optimal Replacement Task

In the optimal replacement problem introduced by Munos and Szepesvári [2008] there are two primitive action $A = \{K, R\}$ and the state is a 1-dimensional value in the interval $X = [0, 10]$. The dynamics of the system while executing the action K, representing maintaining the current product, are defined by

$$\Pr(y|x, K) = \begin{cases} \beta e^{-\beta(y-x)} & \text{if } y \geq x \\ 0 & \text{if } y < x \end{cases} ,$$

while the dynamics of the system while executing the action R, representing replacing the current product, are defined by

$$\Pr(y|x, R) = \begin{cases} \beta e^{-\beta y} & \text{if } y \geq 0 \\ 0 & \text{if } y < 0 \end{cases} ,$$

where $\beta = 0.5$ in our experiments. The reward function was defined by

$$r(x, a) = \begin{cases} -c(x) = -4x & \text{if } a = K \\ -c(x) - C = -4x - 30 & \text{if } a = R \end{cases} ,$$

where the choice of our parameters $\beta = 0.5$, $C = 30$, $c(x) = 4x$, and discount factor $\gamma = 0.6$ were chosen to match the experiments from Munos and Szepesvári [2008]. This allowed us to directly compare our experimental results to the experimental results from Munos and Szepesvári [2008].

The optimal value function for this problem can be computed in closed form by

$$V^*(x) = \begin{cases} -10x + 30\left(e^{0.2(x-\bar{x})} - 1\right) & \text{if } x \leq \bar{x} \\ -10\bar{x} & \text{if } x > \bar{x} \end{cases}$$

where $\bar{x} \approx 4.8665$, as derived in Munos and Szepesvári [2008].

The features used by are function approximation architecture were fourth order polynomials. For state x, we used features $(1, x, x^2, x^3, x^4)$. This choice in features was selected so that our results could be directly compared with the results from Munos and Szepesvári [2008].

For the OFVI condition, we augmented the primitive action set $A = \{K, R\}$ by a single option $o = \langle I_o, \pi_o, \beta_o \rangle$ where the initial state set $I_o = \{x \in X | x < \tilde{x}\}$, $\forall_{x \in X} \pi_o(x) = K$, and $\beta_o(x) = \begin{cases} 0 & \text{if } x < \tilde{x} \\ 1 & \text{if } x \geq \tilde{x} \end{cases}$. Here $\tilde{x} = \bar{x} + \Delta$. For the experiments shown in this paper, we set $\Delta = 0$. However, in further experiments we found that increasing Δ decreased the convergence rate gained from including the option o.

We repeated all experiments 100 times for each condition.

4.2 Inventory Management Task

In a basic inventory management task, the objective is to maintain stock of one or more commodities to meet customer demand while at the same time minimizing ordering costs and storage costs [Scarf, 1959, Sethi and Cheng, 1997]. At each time period, the agent is given the opportunity to order shipments of commodities to resupply its warehouse.

We created an inventory management problem where the agent resupplies a warehouse with $n = 8$ different commodities. The warehouse has limited storage (500 units in our experiments). Demand for each commodity is stochastic and depends on the time of year. Figure 1 shows the expected demand $\pm$ one standard deviation over the course of twelve months. The agent is given the opportunity to place an order twice each month for a total of 24 order periods per demand cycle.

The state $\{\tau, x\}$ of the inventory management problem was a vector specifying the time of year τ and the number of each commodity stored in the warehouse x. We will denote the quantity of each commodity by x_i for $i = 1, 2, \ldots, n$. During each timestep, a demand vector ξ was drawn by sampling the demand for each commodity independently according to

$$\xi_i \sim [\mathcal{N}(\mu_i(\tau), \sigma_i)]_+$$

where $\mathcal{N}$ is the normal distribution with mean $\mu_i(\tau)$ and standard deviation σ_i. The mean

$$\mu_i(\tau) = \frac{g_i}{2}\left(\cos\left(\frac{(2\tau + 4\zeta_i)\pi}{24}\right) + 1\right) \quad ,$$

where π refers to the mathematical constant and the values of g_i, ζ_i, and σ_i can be found in Table 1. The demand vector was then subtracted from the number of each commodity stored in the warehouse. If any of the commodities were negative after subtracting the demand vector, the agent received an unmet demand penalty

$$p_{\text{ud}}(x - \xi) = \begin{cases} u_b + u_s \sum_{i=1}^{n} [x_i - \xi_i]_- & \text{if } \sum_{i=1}^{n} [x_i - \xi_i]_- < 0 \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

where $u_b = 2$, $u_s = 10$, and $[x]_- = \begin{cases} x & \text{if } x < 0 \\ 0 & \text{otherwise} \end{cases}$.

Then the agent was given the opportunity to either resupply it's warehouse or order nothing. The primitive actions available to the agent were the ability to order nothing or to order any single commodity in quantities of 25 up to the maximum size of the warehouse. An action $a = \langle i, q \rangle$ was defined by a commodity index i and a quantity q. The cost of an order was defined by

$$p_{\text{oc}}(i, q) = \begin{cases} 0 & \text{if } q = 0 \\ o_b + o_{s,i} q & \text{otherwise} \end{cases} \tag{18}$$

where $o_b = 8$ is the base ordering cost and $o_{s,i}$ (see Table 1) is the commodity dependent unit cost. The new state steps forward half a month into the future and the quantities in the inventory are updated to remove the purchased commodities and add the ordered commodities (if any). If the agent orders more than will fit in the inventory, then only the portion of the order the fits in the warehouse will be kept (but the agent will be charged for the complete order). At the end of each decision step, the agent receives a cost which is the sum of the unmet demands and the order cost.

We repeated all experiments 8 times for each condition.

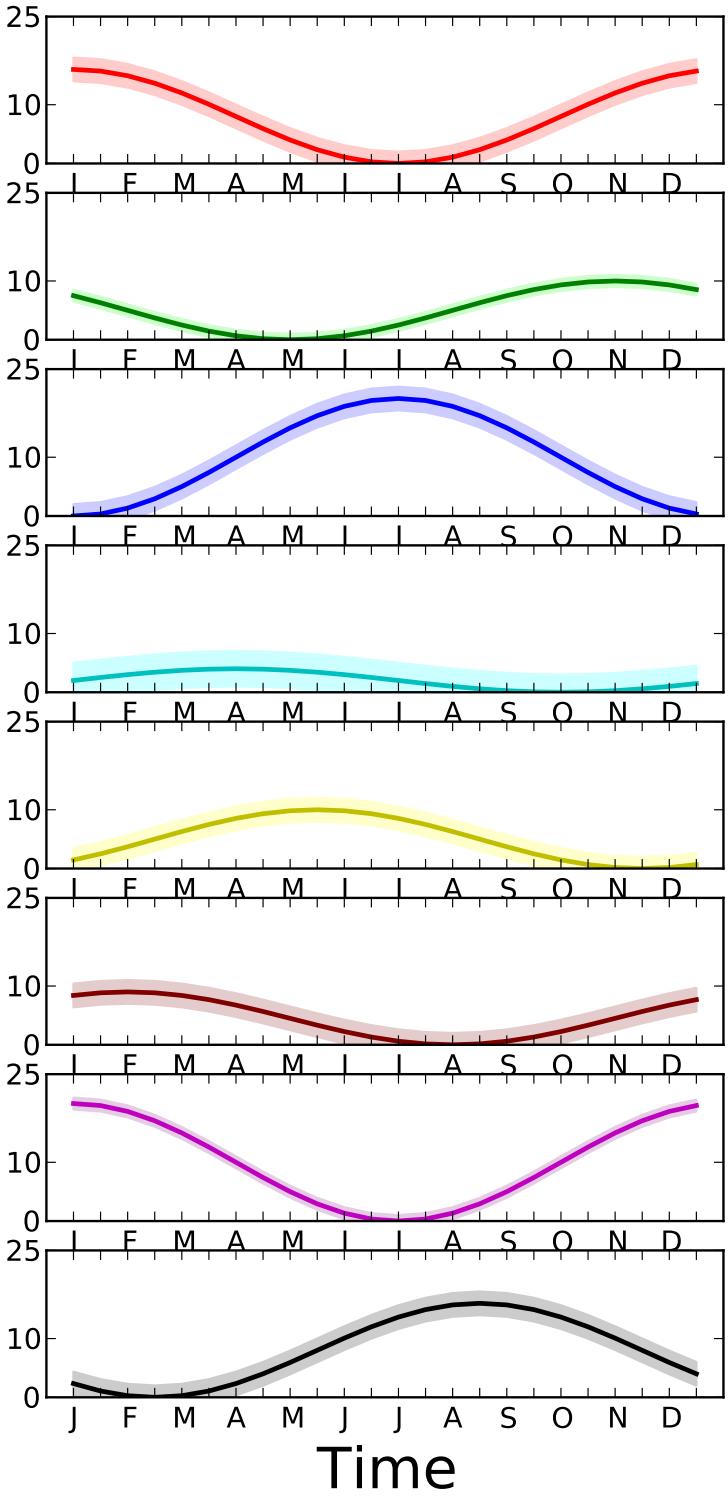

Figure 1: Expected demand (± standard deviation) for eight commodities.

Table 1: Commodity Properties

Commodity Index	1	2	3	4	5	6	7	8
Unit Cost ($o_{s,i}$)	1	3	1	2	0.5	1	1	1
Demand Peak (Month, ζ_i)	1	3	7	10	8.5	12	1	5.5
Demand Std. Deviation (σ_i)	2	1	2	3	2	2	1	2
Max. Expected Demand (g_i)	16	10	20	4	10	9	20	16

References

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.

Herbert Scarf. The optimality of (s,s) policies in the dynamic inventory problem. Technical Report NR-047-019, Office of Naval Research, April 1959.

Suresh P. Sethi and Feng Cheng. Optoptimal of (s,s) policies in inventory models with markovian demand. *Operations Research*, 45(6):931–939, 1997.