APPENDIX

A PROOFS

A.1 PROOF OF LEMMA 4.1 – BOUND ON THE POLICY GRADIENT VARIANCE

For any parametric policy π_θ and function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,

$$\mathrm{Var}\left(\nabla_\theta \log \pi_\theta(a|s) Q(s,a)\right) \leq \max_{s,a} \left[Q(s,a)\right]^2 \max_s \|\nabla_\theta \log \pi_\theta(\cdot|s)\|_F^2,$$

where $\nabla_\theta \log \pi_\theta(\cdot|s) \in \mathbb{R}^{A \times \dim(\theta)}$ is a matrix whose a-th row is $\nabla_\theta \log \pi_\theta(a|s)^\top$.

Proof. The variance for a parametric policy π_θ is given as follows:

$$\begin{aligned}\mathrm{Var}\left(\nabla_\theta \log \pi_\theta(a|s) Q(a,s)\right) =& \mathbb{E}_{s\sim d_{\pi_\theta}, a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s)^\top \nabla_\theta \log \pi_\theta(a|s) Q(s,a)^2\right] - \\ & \mathbb{E}_{s\sim d_{\pi_\theta}, a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) Q(s,a)\right]^\top \mathbb{E}_{s\sim d_{\pi_\theta}, a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) Q(s,a)\right],\end{aligned}$$

where $Q(s,a)$ is the currently estimated Q-function and d_{π_θ} is the discounted state visitation frequency induced by the policy π_θ. Since the second term we subtract is always positive (it is of quadratic form $v^\top v$) we can bound the variance by the first term:

$$\begin{aligned}\mathrm{Var}\left(\nabla_\theta \log \pi_\theta(a|s) Q(a,s)\right) \leq& \mathbb{E}_{s\sim d_{\pi_\theta}, a\sim\pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s)^\top \nabla_\theta \log \pi_\theta(a|s) Q(s,a)^2\right] \\ =& \sum_s d_{\pi_\theta}(s) \sum_a \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)^\top \nabla_\theta \log \pi_\theta(a|s) Q(s,a)^2 \\ \leq& \max_{s,a} \left[\left[Q(s,a)\right]^2 \pi_\theta(a|s)\right] \sum_s d_{\pi_\theta}(s) \sum_a \nabla_\theta \log \pi_\theta(a|s)^\top \nabla_\theta \log \pi_\theta(a|s) \\ \leq& \max_{s,a} \left[Q(s,a)\right]^2 \max_s \sum_a \nabla_\theta \log \pi_\theta(a|s)^\top \nabla_\theta \log \pi_\theta(a|s) \\ =& \max_{s,a} \left[Q(s,a)\right]^2 \max_s \|\nabla_\theta \log \pi_\theta(\cdot|s)\|_F^2.\end{aligned}$$

□

A.2 PROOF OF LEMMA 4.2 – VECTOR FORM OF C-SOFTTREEMAX

In vector form, (3) is given by

$$\pi^{\mathrm{C}}_{d,\theta}(\cdot|s) = \frac{\exp\left[\beta\left(C_{s,d} + \gamma^d P_s \left(P^{\pi_b}\right)^{d-1} \Theta\right)\right]}{\mathbf{1}_A^\top \exp\left[\beta\left(C_{s,d} + \gamma^d P_s \left(P^{\pi_b}\right)^{d-1} \Theta\right)\right]}, \tag{8}$$

where

$$C_{s,d} = R_s + P_s \left[\sum_{h=1}^{d-1} \gamma^h \left(P^{\pi_b}\right)^{h-1}\right] R_{\pi_b}. \tag{9}$$

Proof. Consider the vector $\ell_{s,\cdot} \in \mathbb{R}^{|\mathcal{A}|}$. Its expectation satisfies

$$\begin{aligned}\mathbb{E}^{\pi_b} \ell_{s,\cdot}(d;\theta) &= \mathbb{E}^{\pi_b}\left[\sum_{t=0}^{d-1} \gamma^t r_t + \gamma^d \theta(s_d)\right] \\ &= R_s + \sum_{t=1}^{d-1} \gamma^t P_s (P^{\pi_b})^{t-1} R_{\pi_b} + \gamma^d P_s (P^{\pi_b})^{d-1} \Theta.\end{aligned}$$

As required. □

A.3 Proof of Lemma 4.3 – Gradient of C-SoftTreeMax

The C-SoftTreeMax gradient of dimension $A \times S$ is given by

$$\nabla_\theta \log \pi^{\text{C}}_{d,\theta} = \beta\gamma^d \left[I_A - \mathbf{1}_A(\pi^{\text{C}}_{d,\theta})^\top\right] P_s \left(P^{\pi_b}\right)^{d-1},$$

where for brevity, we drop the s index in the policy above, i.e., $\pi^{\text{C}}_{d,\theta} \equiv \pi^{\text{C}}_{d,\theta}(\cdot|s)$.

Proof. The (j,k)-th entry of $\nabla_\theta \log \pi^{\text{C}}_{d,\theta}$ satisifes

$$\begin{aligned}
[\nabla_\theta \log \pi^{\text{C}}_{d,\theta}]_{j,k} &= \frac{\partial \log(\pi^{\text{C}}_{d,\theta}(a^j|s))}{\partial\theta(s^k)} \\
&= \beta\gamma^d [P_s(P^{\pi_b})^{d-1}]_{j,k} - \frac{\sum_a \left[\exp\left[\beta\left(C_{s,d} + \gamma^d P_s\left(P^{\pi_b}\right)^{d-1}\Theta\right)\right]\right]_a \beta\gamma^d \left[P_s(P^{\pi_b})^{d-1}\right]_{a,k}}{\mathbf{1}_A^\top \exp\left[\beta\left(C_{s,d} + \gamma^d P_s\left(P^{\pi_b}\right)^{d-1}\Theta\right)\right]} \\
&= \beta\gamma^d [P_s(P^{\pi_b})^{d-1}]_{j,k} - \beta\gamma^d \sum_a \pi^{\text{C}}_{d,\theta}(a|s) \left[P_s(P^{\pi_b})^{d-1}\right]_{a,k} \\
&= \beta\gamma^d [P_s(P^{\pi_b})^{d-1}]_{j,k} - \beta\gamma^d \left[(\pi^{\text{C}}_{d,\theta})^\top P_s(P^{\pi_b})^{d-1}\right]_k \\
&= \beta\gamma^d [P_s(P^{\pi_b})^{d-1}]_{j,k} - \beta\gamma^d \left[\mathbf{1}_A(\pi^{\text{C}}_{d,\theta})^\top P_s(P^{\pi_b})^{d-1}\right]_{j,k}.
\end{aligned}$$

Moving back to matrix form, we obtain the stated result. □

A.4 Proof of Theorem 4.4 – Exponential variance decay of C-SoftTreeMax

The C-SoftTreeMax policy gradient is bounded by

$$\text{Var}\left(\nabla_\theta \log \pi^{\text{C}}_{d,\theta}(a|s) Q(s,a)\right) \leq 2\frac{A^2S^2\beta^2}{(1-\gamma)^2}\gamma^{2d}|\lambda_2(P^{\pi_b})|^{2(d-1)}.$$

Proof. We use Lemma 4.1 directly. First of all, it is know that when the reward is bounded in $[0,1]$, the maximal value of the Q-function is $\frac{1}{1-\gamma}$ as the sum as infinite discounted rewards. Next, we bound the Frobenius norm of the term achieved in Lemma 4.3, by applying the eigen-decomposition on P^{π_b}:

$$P^{\pi_b} = \mathbf{1}_S\mu^\top + \sum_{i=2}^{S} \lambda_i u_i v_i^\top, \tag{10}$$

where μ is the stationary distribution of P^{π_b}, and u_i and v_i are left and right eigenvectors correspondingly.

$$\begin{aligned}
\|\beta\gamma^d \left(I_{A,A} - \mathbf{1}_A\pi^\top\right) P_s(P^{\pi_b})^{d-1}\|_F &= \beta\gamma^d \| \left(I_{A,A} - \mathbf{1}_A\pi^\top\right) P_s \left(\mathbf{1}_S\mu^\top + \sum_{i=2}^{S} \lambda_i^{d-1} u_i v_i^\top\right) \|_F \\
\textit{(P_s is stochastic)} \quad &= \beta\gamma^d \| \left(I_{A,A} - \mathbf{1}_A\pi^\top\right) \left(\mathbf{1}_A\mu^\top + \sum_{i=2}^{S} \lambda_i^{d-1} P_s u_i v_i^\top\right) \|_F \\
\textit{(projection nullifies $\mathbf{1}_A\mu^\top$)} \quad &= \beta\gamma^d \| \left(I_{A,A} - \mathbf{1}_A\pi^\top\right) \left(\sum_{i=2}^{S} \lambda_i^{d-1} P_s u_i v_i^\top\right) \|_F \\
\textit{(triangle inequality)} \quad &\leq \beta\gamma^d \sum_{i=2}^{S} \| \left(I_{A,A} - \mathbf{1}_A\pi^\top\right) \left(\lambda_i^{d-1} P_s u_i v_i^\top\right) \|_F \\
\textit{(matrix norm sub-multiplicativity)} \quad &\leq \beta\gamma^d |\lambda_2^{d-1}| \sum_{i=2}^{S} \|I_{A,A} - \mathbf{1}_A\pi^\top\|_F \|P_s\|_F \|u_i v_i^\top\|_F \\
&= \beta\gamma^d |\lambda_2^{d-1}| (S-1) \|I_{A,A} - \mathbf{1}_A\pi^\top\|_F \|P_s\|_F.
\end{aligned}$$

Now, we can bound the norm $\|I_{A,A} - \mathbf{1}_A \pi^\top\|_F$ by direct calculation:

$$\|I_{A,A} - \mathbf{1}_A \pi^\top\|_F^2 = \mathrm{Tr}\left[\left(I_{A,A} - \mathbf{1}_A \pi^\top\right)\left(I_{A,A} - \mathbf{1}_A \pi^\top\right)^\top\right] \tag{11}$$

$$= \mathrm{Tr}\left[I_{A,A} - \mathbf{1}_A \pi^\top - \pi \mathbf{1}_A^\top + \pi^\top \pi \mathbf{1}_A \mathbf{1}_A^\top\right] \tag{12}$$

$$= A - 1 - 1 + A\pi^\top \pi \tag{13}$$

$$\leq 2A. \tag{14}$$

From the Cauchy-Schwartz inequality,

$$\|P_s\|_F^2 = \sum_a \sum_s [[P_s]_{a,s}]^2 = \sum_a \|[P_s]_{a,\cdot}\|_2^2 \leq \sum_a \|[P_s]_{a,\cdot}\|_1 \|[P_s]_{a,\cdot}\|_\infty \leq A.$$

So,

$$\begin{aligned}
\mathrm{Var}\left(\nabla_\theta \log \pi^{\mathrm{C}}_{d,\theta}(a|s) Q(s,a)\right) &\leq \max_{s,a} [Q(s,a)]^2 \max_s \|\nabla_\theta \log \pi^{\mathrm{C}}_{d,\theta}(\cdot|s)\|_F^2 \\
&\leq \frac{1}{(1-\gamma)^2} \|\beta\gamma^d \left(I_{A,A} - \mathbf{1}_A \pi^\top\right) P_s (P^{\pi_b})^{d-1}\|_F^2 \\
&\leq \frac{1}{(1-\gamma)^2} \beta^2 \gamma^{2d} |\lambda_2(P^{\pi_b})|^{2(d-1)} S^2(2A^2),
\end{aligned}$$

which obtains the desired bound. □

A.5 A LOWER BOUND ON C-SOFTTREEMAX GRADIENT (RESULT NOT IN THE PAPER)

For completeness we also supply a lower bound on the Frobenius norm of the gradient. Note that this result does not translate to the a lower bound on the variance since we have no lower bound equivalence of Lemma 4.1.

Lemma A.1. *The Frobenius norm on the gradient of the policy is lower-bounded by:*

$$\|\nabla_\theta \log \pi^{C}_{d,\theta}(\cdot|s)\|_F \geq C \cdot \beta\gamma^d |\lambda_2(P^{\pi_b})|^{(d-1)}. \tag{15}$$

Proof. We begin by moving to the induced l_2 norm by norm-equivalence:

$$\|\beta\gamma^d \left(I_{A,A} - \mathbf{1}_A \pi^\top\right) P_s (P^{\pi_b})^{d-1}\|_F \geq \|\beta\gamma^d \left(I_{A,A} - \mathbf{1}_A \pi^\top\right) P_s (P^{\pi_b})^{d-1}\|_2.$$

Now, taking the vector u to be the eigenvector of the second eigenvalue of P^{π_b}:

$$\begin{aligned}
\|\beta\gamma^d \left(I_{A,A} - \mathbf{1}_A \pi^\top\right) P_s (P^{\pi_b})^{d-1}\|_2 &\geq \|\beta\gamma^d \left(I_{A,A} - \mathbf{1}_A \pi^\top\right) P_s (P^{\pi_b})^{d-1} u\|_2 \\
&= \beta\gamma^d \|\left(I_{A,A} - \mathbf{1}_A \pi^\top\right) P_s u\|_2 \\
&= \beta\gamma^d |\lambda_2(P^{\pi_b})|^{(d-1)} \|\left(I_{A,A} - \mathbf{1}_A \pi^\top\right) P_s u\|_2.
\end{aligned}$$

Note that even though $P_s u$ can be 0, that is not the common case since we can freely change π_b (and therefore the eigenvectors of P^{π_b}). □

A.6 PROOF OF LEMMA 4.5 – VECTOR FORM OF E-SOFTTREEMAX

For $d \geq 1$, (4) is given by

$$\pi^{\mathrm{E}}_{d,\theta}(\cdot|s) = \frac{E_{s,d} \exp(\beta\gamma^d \Theta)}{1_A^\top E_{s,d} \exp(\beta\gamma^d \Theta)}, \tag{16}$$

where

$$E_{s,d} = P_s \prod_{h=1}^{d-1} \left(D\left(\exp[\beta\gamma^h R]\right) P^{\pi_b}\right) \tag{17}$$

with R being the $|S|$-dimensional vector whose s-th coordinate is $r(s)$.

Proof. Recall that

$$\ell_{s,a}(d;\theta) = r(s) + \sum_{t=1}^{d-1} \gamma^t r(s_t) + \gamma^d \theta(s_d). \tag{18}$$

and, hence,

$$\exp[\beta \ell_{s,a}(d;\theta)] = \exp\left[\beta\left(r(s) + \sum_{t=1}^{d-1} \gamma^t r(s_t) + \gamma^d \theta(s_d)\right)\right]. \tag{19}$$

Therefore,

$$\mathbb{E}[\exp \beta \ell_{s,a}(d;\theta)] = \mathbb{E}\left[\exp\left[\beta\left(r(s) + \sum_{t=1}^{d-1} \gamma^t r(s_t)\right)\right] \mathbb{E}\left[\exp\left[\beta\left(\gamma^d \theta(s_d)\right)\right] | s_1, \ldots, s_{d-1}\right]\right] \tag{20}$$

$$= \mathbb{E}\left[\exp\left[\beta\left(r(s) + \sum_{t=1}^{d-1} \gamma^t r(s_t)\right)\right] P^{\pi_b}(\cdot|s_{d-1})\right] \exp(\beta\gamma^d \Theta) \tag{21}$$

$$= \mathbb{E}\left[\exp\left[\beta\left(r(s) + \sum_{t=1}^{d-2} \gamma^t r(s_t)\right)\right] \exp[\beta\gamma^{d-1} r(s_{d-1})] P^{\pi_b}(\cdot|s_{d-1})\right] \exp(\beta\gamma^d \Theta). \tag{22}$$

By repeatedly using iterative conditioning as above, the desired result follows. Note that $\exp(\beta r(s))$ does not depend on the action and is therefore cancelled out with the denominator. □

A.7 Proof of Lemma 4.6 – Gradient of E-SoftTreeMax

The E-SoftTreeMax gradient of dimension $A \times S$ is given by

$$\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta} = \beta\gamma^d \left[I_A - \mathbf{1}_A (\pi^{\mathrm{E}}_{d,\theta})^\top\right] \frac{D\left(\pi^{\mathrm{E}}_{d,\theta}\right)^{-1} E_{s,d} D(\exp(\beta\gamma^d \Theta))}{\mathbf{1}_A^\top E_{s,d} \exp(\beta\gamma^d \Theta)},$$

where for brevity, we drop the s index in the policy above, i.e., $\pi^{\mathrm{E}}_{d,\theta} \equiv \pi^{\mathrm{E}}_{d,\theta}(\cdot|s)$.

Proof. The (j,k)-th entry of $\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta}$ satisfies

$$\begin{aligned}
[\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta}]_{j,k} &= \frac{\partial \log(\pi^{\mathrm{E}}_{d,\theta}(a^j|s))}{\partial \theta(s^k)} \\
&= \frac{\partial}{\partial \theta(s^k)} \left(\log[(E_{s,d})_j^\top \exp(\beta\gamma^d \Theta)] - \log[\mathbf{1}_A^\top E_{s,d} \exp(\beta\gamma^d \Theta)]\right) \\
&= \frac{\beta\gamma^d (E_{s,d})_{j,k} \exp(\beta\gamma^d \theta(s^k))}{(E_{s,d})_j^\top \exp(\beta\gamma^d \Theta)} - \frac{\beta\gamma^d \mathbf{1}_A^\top E_{s,d} e_k \exp(\beta\gamma^d \theta(s^k))}{\mathbf{1}_A^\top E_{s,d} \exp(\beta\gamma^d \Theta)} \\
&= \frac{\beta\gamma^d (E_{s,d} e_k \exp(\beta\gamma^d \theta(s^k)))_j}{(E_{s,d})_j^\top \exp(\beta\gamma^d \Theta)} - \frac{\beta\gamma^d \mathbf{1}_A^\top E_{s,d} e_k \exp(\beta\gamma^d \theta(s^k))}{\mathbf{1}_A^\top E_{s,d} \exp(\beta\gamma^d \Theta)} \\
&= \beta\gamma^d \left[\frac{e_j^\top}{e_j^\top E_{s,d} \exp(\beta\gamma^d \Theta)} - \frac{\mathbf{1}_A^\top}{\mathbf{1}_A^\top E_{s,d} \exp(\beta\gamma^d \Theta)}\right] E_{s,d} e_k \exp(\beta\gamma^d \theta(s^k)).
\end{aligned}$$

Hence,

$$[\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta}]_{\cdot,k} = \beta\gamma^d \left[D(E_{s,d} \exp(\beta\gamma^d \Theta))^{-1} - (\mathbf{1}_A^\top E_{s,d} \exp(\beta\gamma^d \Theta))^{-1} \mathbf{1}_A \mathbf{1}_A^\top\right] E_{s,d} e_k \exp(\beta\gamma^d \theta(s^k))$$

From this, it follows that

$$\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta} = \beta\gamma^d \left[D\left(\pi^{\mathrm{E}}_{d,\theta}\right)^{-1} - \mathbf{1}_A \mathbf{1}_A^\top\right] \frac{E_{s,d} D(\exp(\beta\gamma^d \Theta))}{\mathbf{1}_A^\top E_{s,d} \exp(\beta\gamma^d \Theta)}. \tag{23}$$

The desired result is now easy to see. □

A.8 Proof of Theorem 4.7 — Exponential variance decay of E-SoftTreeMax

There exists $\alpha \in (0,1)$ such that, for any function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,

$$\mathrm{Var}\left(\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta}(a|s) Q(s,a)\right) \in \mathcal{O}\left(\beta^2 \gamma^{2d} \alpha^{2d}\right).$$

If all rewards are equal ($r \equiv \text{const}$), then $\alpha = |\lambda_2(P^{\pi_b})|$.

Proof outline. Recall that thanks to Lemma 4.1, we can bound the PG variance using a direct bound on the gradient norm. The definition of the induced norm is

$$\|\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta}\| = \max_{z:\|z\|=1} \|\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta} z\|,$$

with $\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta}$ given in Lemma 4.6. Let $z \in \mathbb{R}^S$ be an arbitrary vector such that $\|z\| = 1$. Then, $z = \sum_{i=1}^S c_i z_i$, where c_i are scalar coefficients and z_i are vectors spanning the S-dimensional space. In the full proof, we show our specific choice of z_i and prove they are linearly independent given that choice. We do note that $z_1 = \mathbf{1}_S$.

The first part of the proof relies on the fact that $(\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta}) z_1 = 0$. This is easy to verify using Lemma 4.6 together with (6), and because $\left[I_A - \mathbf{1}_A (\pi^{\mathrm{E}}_{d,\theta})^\top\right]$ is a projection matrix whose null-space is spanned by $\mathbf{1}_S$. Thus,

$$\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta} z = \nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta} \sum_{i=2}^S c_i z_i.$$

In the second part of the proof, we focus on $E_{s,d}$ from (6), which appears within $\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta}$. Notice that $E_{s,d}$ consists of the product $\prod_{h=1}^{d-1} \left(D\left(\exp(\beta\gamma^h R)\right) P^{\pi_b}\right)$. Even though the elements in this product are not stochastic matrices, in the full proof we show how to normalize each of them to a stochastic matrix B_h. We thus obtain that

$$E_{s,d} = P_s D(M_1) \prod_{h=1}^{d-1} B_h,$$

where $M_1 \in \mathbb{R}^S$ is some strictly positive vector. Then, we can apply a result by Mathkar and Borkar (2016), which itself builds on (Chatterjee and Seneta, 1977). The result states that the product of stochastic matrices $\prod_{h=1}^{d-1} B_h$ of our particular form converges exponentially fast to a matrix of the form $\mathbf{1}_S \mu^\top$ s.t. $\|\mathbf{1}_S \mu^\top - \prod_{h=1}^{d-1} B_h\| \le C\alpha^d$ for some constant C.

Lastly, $\mathbf{1}_S \mu_{\pi_b}^\top$ gets canceled due to our choice of z_i, $i = 2, \ldots, S$. This observation along with the above fact that the remainder decays then shows that $\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta} \sum_{i=2}^S z_i = \mathcal{O}(\alpha^d)$, which gives the desired result. □

Full technical proof. Let $d \ge 2$. Recall that

$$E_{s,d} = P_s \prod_{h=1}^{d-1} \left(D\left(\exp[\beta\gamma^h R]\right) P^{\pi_b}\right), \tag{24}$$

and that R refers to the S-dimensional vector whose s-th coordinate is $r(s)$. Define

$$B_i = \begin{cases} P^{\pi_b} & \text{if } i = d-1, \\ D^{-1}(P^{\pi_b} M_{i+1}) P^{\pi_b} D(M_{i+1}) & \text{if } i = 1, \ldots, d-2, \end{cases} \tag{25}$$

and the vector

$$M_i = \begin{cases} \exp(\beta\gamma^{d-1} R) & \text{if } i = d, \\ \exp(\beta\gamma^i R) \circ P^{\pi_b} M_{i+1} & \text{if } i = 1, \ldots, d-2, \end{cases} \tag{26}$$

where $\circ$ denotes the element-wise product. Then,

$$E_{s,d} = P_s D(M_1) \prod_{i=1}^{d-1} B_i. \tag{27}$$

It is easy to see that each B_i is a row-stochastic matrix, i.e., all entries are non-negative and $B_i \mathbf{1}_S = \mathbf{1}_S$.

Next, we prove that all non-zeros entries of B_i are bounded away from 0 by a constant. This is necessary to apply the next result from Chatterjee and Seneta (1977). The j-th coordinate of M_i satisfies

$$(M_i)_j = \exp[\beta\gamma^i R_j] \sum_k [P^{\pi_b}]_{j,k} (M_{i+1})_k \leq \| \exp[\beta\gamma^i R] \|_\infty \| M_{i+1} \|_\infty. \tag{28}$$

Separately, observe that $\|M_{d-1}\|_\infty \leq \| \exp(\beta\gamma^{d-1} R) \|_\infty$. Plugging these relations in (26) gives

$$\|M_1\|_\infty \leq \prod_{h=1}^{d-1} \| \exp[\beta\gamma^h R] \|_\infty = \prod_{h=1}^{d-1} \| \exp[\beta R] \|_\infty^{\gamma^h} = \| \exp[\beta R] \|_\infty^{\sum_{h=1}^{d-1} \gamma^h} \leq \| \exp[\beta R] \|_\infty^{\frac{1}{1-\gamma}}. \tag{29}$$

Similarly, for every $1 \leq i \leq d-1$, we have that

$$\|M_i\|_\infty \leq \prod_{h=i}^{d-1} \| \exp[\beta R] \|_\infty^{\gamma^h} \leq \| \exp[\beta R] \|_\infty^{\frac{1}{1-\gamma}}. \tag{30}$$

The jk-th entry of $B_i = D^{-1}(P^{\pi_b} M_{i+1}) P^{\pi_b} D(M_{i+1})$ is

$$(B_i)_{jk} = \frac{P^{\pi_b}_{jk} [M_{i+1}]_k}{\sum_{\ell=1}^{|S|} P^{\pi_b}_{j\ell} [M_{i+1}]_\ell} \geq \frac{P^{\pi_b}_{jk}}{\sum_{\ell=1}^{|S|} P^{\pi_b}_{j\ell} [M_{i+1}]_\ell} \geq \frac{P^{\pi_b}_{jk}}{\| \exp[\beta R] \|_\infty^{\frac{1}{1-\gamma}}}. \tag{31}$$

Hence, for non-zero $P^{\pi_b}_{jk}$, the entries are bounded away from zero by the same. We can now proceed with applying the following result.

Now, by (Chatterjee and Seneta, 1977, Theorem 5) (see also (14) in (Mathkar and Borkar, 2016)), $\lim_{d\to\infty} \prod_{i=1}^{d-1} B_i$ exists and is of the form $\mathbf{1}_S \mu^\top$ for some probability vector μ. Furthermore, there is some $\alpha \in (0,1)$ such that $\varepsilon(d) := \left(\prod_{i=1}^{d-1} B_i \right) - \mathbf{1}_S \, \mu^\top$ satisfies

$$\|\varepsilon(d)\| = O(\alpha^d). \tag{32}$$

Pick linearly independent vectors $w_2, \ldots, w_S$ such that

$$\mu^\top w_i = 0 \text{ for } i = 2, \ldots, d. \tag{33}$$

Since $\sum_{i=2}^S \alpha_i w_i$ is perpendicular to μ for any $\alpha_2, \ldots \alpha_S$ and because $\mu^\top \exp(\beta\gamma^d \Theta) > 0$, there exists no choice of $\alpha_2, \ldots, \alpha_S$ such that $\sum_{i=2}^S \alpha_i w_i = \exp(\beta\gamma^d \Theta)$. Hence, if we let $z_1 = \mathbf{1}_S$ and $z_i = D(\exp(\beta\gamma^d \Theta))^{-1} w_i$ for $i = 2, \ldots, S$, then it follows that $\{z_1, \ldots, z_S\}$ is linearly independent. In particular, it implies that $\{z_1, \ldots, z_S\}$ spans $\mathbb{R}^S$.

Now consider an arbitrary unit norm vector $z := \sum_{i=1}^{S} c_i z_i \in \mathbb{R}^S$ s.t. $\|z\|_2 = 1$. Then,

$$\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta} z = \nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta} \sum_{i=2}^{S} c_i z_i \tag{34}$$

$$= \beta\gamma^d \left[I_A - \mathbf{1}_A(\pi^{\mathrm{E}}_{d,\theta})^\top\right] \frac{D\left(\pi^{\mathrm{E}}_{d,\theta}\right)^{-1} E_{s,d} D(\exp(\beta\gamma^d\Theta))}{\mathbf{1}_A^\top E_{s,d}\exp(\beta\gamma^d\Theta)} \sum_{i=2}^{S} c_i z_i \tag{35}$$

$$= \beta\gamma^d \left[I_A - \mathbf{1}_A(\pi^{\mathrm{E}}_{d,\theta})^\top\right] \frac{D\left(\pi^{\mathrm{E}}_{d,\theta}\right)^{-1} E_{s,d}}{\mathbf{1}_A^\top E_{s,d}\exp(\beta\gamma^d\Theta)} \sum_{i=2}^{S} c_i w_i \tag{36}$$

$$= \beta\gamma^d \left[I_A - \mathbf{1}_A(\pi^{\mathrm{E}}_{d,\theta})^\top\right] \frac{D\left(\pi^{\mathrm{E}}_{d,\theta}\right)^{-1} \left[\mathbf{1}_S\mu^\top + \varepsilon(d)\right]}{\mathbf{1}_A^\top E_{s,d}\exp(\beta\gamma^d\Theta)} \sum_{i=2}^{S} c_i w_i \tag{37}$$

$$= \beta\gamma^d \left[I_A - \mathbf{1}_A(\pi^{\mathrm{E}}_{d,\theta})^\top\right] \frac{D\left(\pi^{\mathrm{E}}_{d,\theta}\right)^{-1} \varepsilon(d)}{\mathbf{1}_A^\top E_{s,d}\exp(\beta\gamma^d\Theta)} \sum_{i=2}^{S} c_i w_i \tag{38}$$

$$= \beta\gamma^d \left[I_A - \mathbf{1}_A(\pi^{\mathrm{E}}_{d,\theta})^\top\right] \frac{D\left(\pi^{\mathrm{E}}_{d,\theta}\right)^{-1} \varepsilon(d) D(\exp(\beta\gamma^d\Theta))}{\mathbf{1}_A^\top E_{s,d}\exp(\beta\gamma^d\Theta)} (z - c_1\mathbf{1}_S), \tag{39}$$

where (34) follows from the fact that $\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta} z_1 = \nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta} \mathbf{1}_S = 0$, (35) follows from Lemma 4.6, (36) holds since $z_i = D(\exp(\beta\gamma^d\Theta))^{-1} w_i$, (38) because μ is perpendicular w_i for each i, while (39) follows by reusing $z_i = D(\exp(\beta\gamma^d\Theta))^{-1} w_i$ relation along with the fact that $z_1 = \mathbf{1}_S$. From (39), it follows that

$$\|\nabla_\theta \log \pi^{\mathrm{E}}_{d,\theta} z\| \le \beta\gamma^d \|\varepsilon(d)\| \left\| \left[I_A - \mathbf{1}_A(\pi^{\mathrm{E}}_{d,\theta})^\top\right] \frac{D\left(\pi^{\mathrm{E}}_{d,\theta}\right)^{-1}}{\mathbf{1}_A^\top E_{s,d}\exp(\beta\gamma^d\Theta)} \right\| \|D(\exp(\beta\gamma^d\Theta))\| \, \|z - c_1\mathbf{1}_S\| \tag{40}$$

$$\le \beta\gamma^d\alpha^d(\|I_A\| + \|\mathbf{1}_A(\pi^{\mathrm{E}}_{d,\theta})^\top\|) \left\| \frac{D\left(\pi^{\mathrm{E}}_{d,\theta}\right)^{-1}}{\mathbf{1}_A^\top E_{s,d}\exp(\beta\gamma^d\Theta)} \right\| \exp(\beta\gamma^d \max_s \theta(s)) \|z - c_1\mathbf{1}_S\| \tag{41}$$

$$\le \beta\gamma^d\alpha^d(1 + \sqrt{A}) \left\| \frac{D\left(\pi^{\mathrm{E}}_{d,\theta}\right)^{-1}}{\mathbf{1}_A^\top E_{s,d}\exp(\beta\gamma^d\Theta)} \right\| \exp(\beta\gamma^d \max_s \theta(s)) \|z - c_1\mathbf{1}_S\| \tag{42}$$

$$\le \beta\gamma^d\alpha^d(1 + \sqrt{A}) \left\| D^{-1}(E_{s,d}\exp(\beta\gamma^d\Theta)) \right\| \exp(\beta\gamma^d \max_s \theta(s)) \|z - c_1\mathbf{1}_S\| \tag{43}$$

$$\le \beta\gamma^d\alpha^d(1 + \sqrt{A}) \frac{1}{\min_s [E_{s,d}\exp(\beta\gamma^d\Theta]_s} \exp(\beta\gamma^d \max_s \theta(s)) \|z - c_1\mathbf{1}_S\| \tag{44}$$

$$\le \beta\gamma^d\alpha^d(1 + \sqrt{A}) \frac{\exp(\beta\gamma^d \max_s \theta(s))}{\exp(\beta\gamma^d \min_s \theta(s)) \min_s |M_1|} \|z - c_1\mathbf{1}_S\| \tag{45}$$

$$\le \beta\gamma^d\alpha^d(1 + \sqrt{A}) \frac{\exp(\beta\gamma^d \max_s \theta(s))}{\exp(\beta\gamma^d \min_s \theta(s)) \exp(\beta \min_s r(s))} \|z - c_1\mathbf{1}_S\| \tag{46}$$

$$\le \beta\gamma^d\alpha^d(1 + \sqrt{A}) \exp(\beta[\max_s \theta(s) - \min_s \theta(s) - \min_s r(s)]) \|z - c_1\mathbf{1}_S\|. \tag{47}$$

Lastly, we prove that $\|z - c_1 \mathbf{1}_S\|$ is bounded independently of d. First, denote by $c = (c_1, \ldots, c_S)^\top$ and $\tilde{c} = (0, c_2, \ldots, c_S)^\top$. Also, denote by Z the matrix with z_i as its i-th column. Now,

$$\|z - c_1 \mathbf{1}_S\| = \|\sum_{i=2}^{S} c_i z_i\| \tag{48}$$

$$= \|Z\tilde{c}\| \tag{49}$$

$$\leq \|Z\|\|\tilde{c}\| \tag{50}$$

$$\leq \|Z\|\|c\| \tag{51}$$

$$= \|Z\|\|Z^{-1}z\| \tag{52}$$

$$\leq \|Z\|\|Z^{-1}\|, \tag{53}$$

where the last relation is due to z being a unit vector. All matrix norms here are l_2-induced norms.

Next, denote by W the matrix with w_i in its i-th column. Recall that in (33) we only defined $w_2, \ldots, w_S$. We now set $w_1 = \exp(\beta\gamma^d\Theta)$. Note that w_1 is linearly independent of $\{w_2, \ldots, w_S\}$ because of (33) together with the fact that $\mu^\top w_1 > 0$. We can now express the relation between Z and W by $Z = D^{-1}(\exp(\beta\gamma^d\Theta))W$. Substituting this in (53), we have

$$\|z - c_1 \mathbf{1}_S\| \leq \|D^{-1}(\exp(\beta\gamma^d\Theta))W\|\|W^{-1}D(\exp(\beta\gamma^d\Theta))\| \tag{54}$$

$$\leq \|W\|\|W^{-1}\|\|D(\exp(\beta\gamma^d\Theta))\|\|D^{-1}(\exp(\beta\gamma^d\Theta))\|. \tag{55}$$

It further holds that

$$\|D(\exp(\beta\gamma^d\Theta))\| \leq \max_s \exp\left(\beta\gamma^d\theta(s)\right) \leq \max\{1, \exp[\beta \max_s \theta(s)])\}, \tag{56}$$

where the last relation equals 1 if $\theta(s) < 0$ for all s. Similarly,

$$\|D^{-1}(\exp(\beta\gamma^d\Theta))\| \leq \frac{1}{\min_s \exp\left(\beta\gamma^d\theta(s)\right)} \leq \frac{1}{\min\{1, \exp[\beta \min_s \theta(s)])\}}. \tag{57}$$

Furthermore, by the properties of the l_2-induced norm,

$$\|W\|_2 \leq \sqrt{S}\|W\|_1 \tag{58}$$

$$= \sqrt{S} \max_{1 \leq i \leq S} \|w_i\|_1 \tag{59}$$

$$= \sqrt{S} \max\{\exp(\beta\gamma^d\Theta), \max_{2 \leq i \leq S} \|w_i\|_1\} \tag{60}$$

$$\leq \sqrt{S} \max\{1, \exp[\beta \max_s \theta(s)], \max_{2 \leq i \leq S} \|w_i\|_1)\}. \tag{61}$$

Lastly,

$$\|W^{-1}\| = \frac{1}{\sigma_{\min}(W)} \tag{62}$$

$$\leq \left(\prod_{i=1}^{S-1} \frac{\sigma_{\max}(W)}{\sigma_i(W)}\right) \frac{1}{\sigma_{\min}(W)} \tag{63}$$

$$= \frac{(\sigma_{\max}(W))^{S-1}}{\prod_{i=1}^{S} \sigma_i(W)} \tag{64}$$

$$= \frac{\|W\|^{S-1}}{|\det(W)|}. \tag{65}$$

The determinant of W is a sum of products involving its entries. To upper bound (65) independently of d, we lower bound its denominator by upper and lower bounds on the entries $[W]_{i,1}$ that are independent of d, depending on their sign:

$$\min\{1, \exp[\beta \min_s \theta(s)])\} \leq [W]_{i,1} \leq \max\{1, \exp[\beta \max_s \theta(s)])\}. \tag{66}$$

Using this, together with (53), (55), (56), (57), and (61), we showed that $\|z - c_1 \mathbf{1}_S\|$ is upper bounded by a constant independent of d. This concludes the proof. □

A.9 BIAS ESTIMATES

Lemma A.2. *For any matrix A and $\hat{A}$,*

$$\hat{A}^k - A^k = \sum_{h=1}^{k} \hat{A}^{h-1}(\hat{A} - A)A^{k-h}.$$

Proof. The proof follows from first principles:

$$\sum_{h=1}^{k} \hat{A}^{h-1}(\hat{A} - A)A^{k-h} = \sum_{h=1}^{k} \hat{A}^{h-1}\hat{A}A^{k-h} - \sum_{h=1}^{k} \hat{A}^{h-1}AA^{k-h} \tag{67}$$

$$= \sum_{h=1}^{k} \hat{A}^{h}A^{k-h} - \sum_{h=1}^{k} \hat{A}^{h-1}A^{k-h+1} \tag{68}$$

$$= \hat{A}^k - A^k + \sum_{h=1}^{k-1} \hat{A}^{h}A^{k-h} - \sum_{h=2}^{k} \hat{A}^{h-1}A^{k-h+1} \tag{69}$$

$$= \hat{A}^k - A^k. \tag{70}$$

□

Henceforth, $\|\cdot\|$ will refer to $\|\cdot\|_\infty$, i.e. the induced infinity norm. Also, for brevity, we denote $\pi^{\mathrm{C}}_{d,\theta}$ and $\hat{\pi}^{\mathrm{C}}_{d,\theta}$ by π_θ and $\hat{\pi}_\theta$, respectively. Similarly, we use d_{π_θ} and $d_{\hat{\pi}_\theta}$ to denote $d_{\pi^{\mathrm{C}}_{d,\theta}}$ and $d_{\hat{\pi}^{\mathrm{C}}_{d,\theta}}$. As for the induced norm of the matrix P and its perturbed counterpart $\hat{P}$, which are of size $S \times A \times S$, we slightly abuse notation and denote $\|P - \hat{P}\| = \max_s\{\|P_s - \hat{P}_s\|\}$, where P_s is as defined in Section 2.

Definition A.3. Let ϵ be the maximal model mis-specification, i.e., $\max\{\|P - \hat{P}\|, \|r - \hat{r}\|\} = \epsilon$.

Lemma A.4. *Recall the definitions of R_s, P_s, R_{π_b} and P^{π_b} from Section 2, and respectively denote their perturbed counterparts by $\hat{R}_s, \hat{P}_s, \hat{R}_{\pi_b}$ and $\hat{P}^{\pi_b}$. Then, for ϵ defined in Definition A.3,*

$$\max\{\|R_s - \hat{R}_s\|, \|P_s - \hat{P}_s\|, \|R_{\pi_b} - \hat{R}_{\pi_b}\|, \|P^{\pi_b} - \hat{P}^{\pi_b}\|\} = O(\epsilon). \tag{71}$$

Proof. The proof follows easily from the fact that the differences above are convex combinations of $P - \hat{P}$ and $r - \hat{r}$. □

Lemma A.5. *Let π_θ be as in (5), and let $\hat{\pi}_\theta$ also be defined as in (5), but with R_s, P_s, P^{π_b} replaced by their perturbed counterparts $\hat{R}_s, \hat{P}_s, \hat{P}^{\pi_b}$ throughout. Then,*

$$\|\pi^{C}_{d,\theta} - \hat{\pi}^{C}_{d,\theta}\| = O(\beta d\epsilon). \tag{72}$$

Proof. To prove the desired result, we work with (5) to bound the error between $R_s, P_s, P^{\pi_b}, R_{\pi_b}$ and their perturbed versions.

First, we apply Lemma A.2 together with Lemma A.4 to obtain that $\|(P^{\pi_b})^k - (\hat{P}^{\pi_b})^k\| = O(k\epsilon)$. Next, denote by M the argument in the exponent in (5), i.e.

$$M := \beta[C_{s,d} + \gamma^d P_s (P^{\pi_b})^{d-1}\Theta].$$

Similarly, let $\hat{M}$ be the corresponding perturbed sum that relies on $\hat{P}$ and $\hat{r}$. Combining the bounds from Lemma A.4, and using the triangle inequality, we have that $\|\hat{M} - M\| = O(\beta d\epsilon)$.

Eq. (5) states that the C-SoftTreeMax policy in the true environment is $\pi_\theta = \exp(M)/(1^\top \exp(M))$. Similarly define $\hat{\pi}_\theta$ using $\hat{M}$ for the approximate model. Then,

$$\hat{\pi}_\theta = (\pi_\theta \circ \exp(M - \hat{M}))1^\top \exp(M)/(1^\top \exp(\hat{M})),$$

where $\circ$ denotes element-wise multiplication. Using the above relation, we have that $\|\hat{\pi}_\theta - \pi_\theta\| = \|\pi_\theta\| \| \frac{\exp(M-\hat{M})1^\top \exp(M)}{1^\top \exp(\hat{M})} - 1\|$. Using the relation $|e^x - 1| = O(x)$ as $x \to 0$, the desired result follows.

□

Theorem A.6. *Let ϵ be as in Definition A.3. Further let $\hat{\pi}^C_{d,\theta}$ being the corresponding approximate policy as given in Lemma 4.2. Then, the policy gradient bias is bounded by*

$$\left\| \frac{\partial}{\partial \theta} \left(\nu^\top V^{\pi_\theta} \right) - \frac{\partial}{\partial \theta} \left(\nu^\top V^{\hat{\pi}_\theta} \right) \right\| = \mathcal{O}\left(\frac{\gamma^d}{(1-\gamma)^2} S \beta^2 d \epsilon \right). \tag{73}$$

Proof. We have

$$\frac{\partial}{\partial \theta} \left(\nu^\top V^{\pi_\theta} \right) - \frac{\partial}{\partial \theta} \left(\nu^\top V^{\pi'_\theta} \right) \tag{74}$$

$$= \mathbb{E}_{s \sim d_{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \right] - \mathbb{E}_{s \sim d_{\hat{\pi}_\theta}, a \sim \hat{\pi}_\theta(\cdot|s)} \left[\nabla_\theta \log \hat{\pi}_\theta(a|s) Q^{\hat{\pi}_\theta}(s,a) \right] \tag{75}$$

$$= \sum_{s,a} \left(d_{\pi_\theta}(s) \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a) - d_{\hat{\pi}_\theta}(s) \hat{\pi}_\theta(a|s) \nabla_\theta \log \hat{\pi}_\theta(a|s) Q^{\hat{\pi}_\theta}(s,a) \right) \tag{76}$$

$$= \sum_{s} \Big(d_{\pi_\theta}(s) (\nabla_\theta \log \pi_\theta(\cdot|s))^\top D(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot) \tag{77}$$

$$- d_{\hat{\pi}_\theta}(s) (\nabla_\theta \log \hat{\pi}_\theta(\cdot|s))^\top D(\hat{\pi}_\theta(\cdot|s)) Q^{\hat{\pi}_\theta}(s,\cdot) \Big) \tag{78}$$

$$= \sum_s \left(\prod_{i=1}^4 X_i(s) - \prod_{i=1}^4 \hat{X}_i(s) \right) \tag{79}$$

$$= \sum_s \sum_{i=1}^4 \hat{X}_1(s) \cdots \hat{X}_{i-1}(s) \left(X_i(s) - \hat{X}_i(s) \right) X_{i+1}(s) \cdots X_4(s), \tag{80}$$

where $X_1(s) = d_{\pi_\theta}(s) \in \mathbb{R}$, $X_2(s) = (\nabla_\theta \log \pi_\theta(\cdot|s))^\top \in \mathbb{R}^{S \times A}$, $X_3(s) = D(\pi_\theta(\cdot|s)) \in \mathbb{R}^{A \times A}$, $X_4(s) = Q^{\pi_\theta}(s,\cdot) \in \mathbb{R}^{A \times A}$, and $\hat{X}_1(s), \ldots, \hat{X}_4(s)$ are similarly defined with π_θ replaced by $\hat{\pi}_\theta$.

Therefore,

$$\left\| \frac{\partial}{\partial \theta} \left(\nu^\top V^{\pi_\theta} \right) - \frac{\partial}{\partial \theta} \left(\nu^\top V^{\pi'_\theta} \right) \right\| \leq \left(\max_s \Gamma(s) \right) S, \tag{81}$$

where

$$\Gamma(s) = \| \sum_s \sum_{i=1}^4 \hat{X}_1(s) \cdots \hat{X}_{i-1}(s) \left(X_i(s) - \hat{X}_i(s) \right) X_{i+1}(s) \cdots X_4(s) \|. \tag{82}$$

Next, since $d_{\pi_\theta}, d_{\hat{\pi}_\theta}, \pi_\theta$, and $\hat{\pi}_\theta$ are all distributions, we have

$$\max\{|X_1(s)|, |\hat{X}_1(s)|, |X_3(s,a)|, |\hat{X}_3(s,a)|\} \leq 1. \tag{83}$$

Separately, using Lemma 4.3, we have

$$\|X_2\| = \|\nabla_\theta \log \pi_\theta(a|s)\| \leq \beta \gamma^d (\|I_A\| + \|\mathbf{1}_A \pi_\theta^\top\|) \|P_s\| \|(P^{\pi_b})^{d-1}\|. \tag{84}$$

Since all rows of the above matrices have non-negative entries that add up to 1, we get

$$\|Y\| \leq 2\beta\gamma^d. \tag{85}$$

In the rest of the proof, we bound each of $\|X_1 - \hat{X}_1\|, \ldots, \|X_4 - \hat{X}_4\|$.

Finally,

$$\|X_4\| \leq \frac{1}{1-\gamma}. \tag{86}$$

Similarly, the same bounds hold for $\hat{X}_1, \hat{X}_2, \hat{X}_3$ and $\hat{X}_4$.

From, we have

$$\|X_1 - \hat{X}_1\| \leq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\|\nu^\top(P^{\pi_\theta})^t - \nu^\top(P^{\hat{\pi}_\theta})^t\| \tag{87}$$

$$\leq (1-\gamma)\|\nu\|\sum_{t=0}^{\infty}\gamma^t t d\epsilon \tag{88}$$

$$\leq (1-\gamma)d\epsilon\sum_{t=0}^{\infty}\gamma^t t \tag{89}$$

$$= \frac{\gamma d\epsilon}{1-\gamma}. \tag{90}$$

The last relation follows from the fact that $(1-\gamma)^{-1} = \sum_{t=0}^{\infty}\gamma^t$, which in turn implies

$$\gamma\frac{\partial}{\partial\gamma}\left(\frac{1}{1-\gamma}\right) = \sum_{t=0}^{\infty}t\gamma^t. \tag{91}$$

From Lemma A.5, it follows that

$$\|X_3 - \hat{X}_3\| = O(\beta d\epsilon). \tag{92}$$

Next, recall that from Lemma 4.3 that

$$X_2(s,\cdot) = \beta\gamma^d\left[I_A - \mathbf{1}_A(\pi_\theta)^\top\right]P_s\left(P^{\pi_b}\right)^{d-1}.$$

Then,

$$\|X_2(s,\cdot) - \hat{X}_2(s,\cdot)\| \leq \|\beta\gamma^d\left[I_A - \mathbf{1}_A(\pi_\theta)^\top\right]P_s\|\|\left(P^{\pi_b}\right)^{d-1} - \left(\hat{P}^{\pi_b}\right)^{d-1}\| \tag{93}$$

$$+ \|\beta\gamma^d\left[I_A - \mathbf{1}_A(\pi_\theta)^\top\right]\|\|P_s - \hat{P}_s\|\|\left(\hat{P}^{\pi_b}\right)^{d-1}\| \tag{94}$$

$$+ \beta\gamma^d\|\mathbf{1}_A(\pi_\theta)^\top - \mathbf{1}_A(\hat{\pi}_\theta)^\top\|\|\hat{P}_s\left(\hat{P}^{\pi_b}\right)^{d-1}\|. \tag{95}$$

Following the same argument as in (85) and applying Lemma A.2, we have that (93) is $O(\beta\gamma^d d\epsilon)$. Similarly, from the argument of (85), Eq. (94) is $O(\beta\gamma^d\epsilon)$. Lastly, (95) is $O(\beta\gamma^d d\epsilon)$ due to Lemma A.5. Putting the above three terms together, we have that

$$\|X_2(s,\cdot) - \hat{X}_2(s,\cdot)\| = O(\beta\gamma^d d\epsilon). \tag{96}$$

Since the state-action value function satisfies the Bellman equation, we have

$$Q^{\pi_\theta} = r + \gamma P Q^{\pi_\theta} \tag{97}$$

and

$$Q^{\hat{\pi}_\theta} = \hat{r} + \gamma\hat{P}Q^{\hat{\pi}_\theta}. \tag{98}$$

Consequently,

$$\|Q^{\pi_\theta} - Q^{\hat{\pi}_\theta}\| \leq \|r - \hat{r}\| + \gamma\|PQ^{\pi_\theta} - PQ^{\hat{\pi}_\theta}\| + \gamma\|PQ^{\hat{\pi}_\theta} - \hat{P}Q^{\hat{\pi}_\theta}\| \tag{99}$$

$$\leq \epsilon + \gamma\|P\|\|Q^{\pi_\theta} - Q^{\hat{\pi}_\theta}\| + \gamma\|P - \hat{P}\|\|Q^{\hat{\pi}_\theta}\| \tag{100}$$

$$\leq \epsilon + \gamma\|Q^{\pi_\theta} - Q^{\hat{\pi}_\theta}\| + \frac{\gamma}{1-\gamma}\epsilon, \tag{101}$$

which finally shows that

$$\|X_4 - \hat{X}_4\| = \|Q^{\pi_\theta} - Q^{\hat{\pi}_\theta}\| \leq \frac{\epsilon}{(1-\gamma)^2}. \tag{102}$$

□

B EXPERIMENTS

B.1 IMPLEMENTATION DETAILS

The environment engine is the highly efficient Atari-CuLE (Dalton et al., 2020), a CUDA-based version of Atari that runs on GPU. Similarly, we use Atari-CuLE for the GPU-based breadth-first TS as done in Dalal et al. (2021): In every tree expansion, the state S_t is duplicated and concatenated with all possible actions. The resulting tensor is fed into the GPU forward model to generate the tensor of next states $(S_{t+1}^0, \ldots, S_{t+1}^{A-1})$. The next-state tensor is then duplicated and concatenated again with all possible actions, fed into the forward model, etc. This procedure is repeated until the final depth is reached, for which $W_\theta(s)$ is applied per state.

We train SoftTreeMax for depths $d = 1 \ldots 8$, with a single worker. We use five seeds for each experiment.

For the implementation, we extend Stable-Baselines3 (Raffin et al., 2019) with all parameters taken as default from the original PPO paper (Schulman et al., 2017). For depths $d \geq 3$, we limited the tree to a maximum width of 1024 nodes and pruned non-promising trajectories in terms of estimated weights. Since the distributed PPO baseline advances significantly faster in terms of environment steps, for a fair comparison, we ran all experiments for one week on the same machine and use the wall-clock time as the x-axis. We use Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz equipped with one NVIDIA Tesla V100 32GB.

B.2 TIME-BASED TRAINING CURVES

We provide the training curves in Figure 4. For brevity, we exclude a few of the depths from the plots. As seen, there is a clear benefit for SoftTreeMax over distributed PPO with the standard softmax policy. In most games, PPO with the SoftTreeMax policy shows very high sample efficiency: it achieves higher episodic reward although it observes much less episodes, for the same running time.

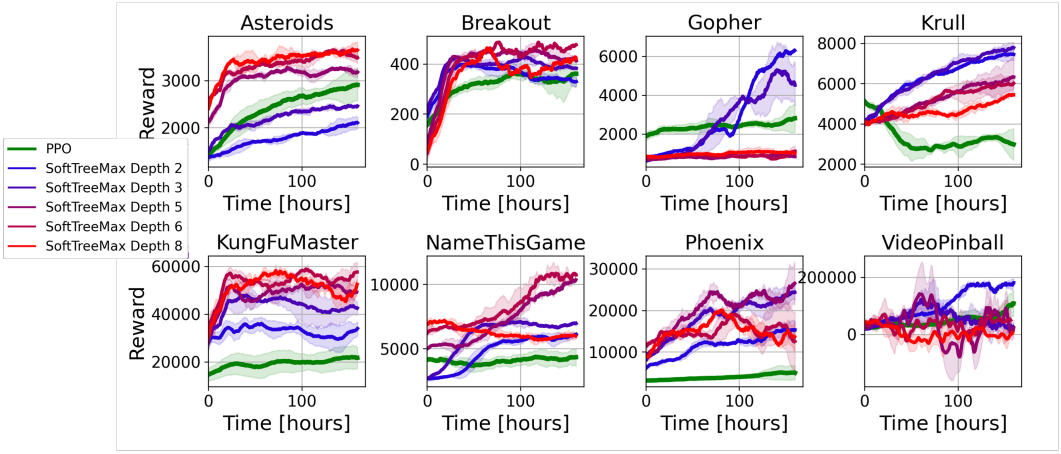

Figure 4: **Training curves: GPU SoftTreeMax (single worker) vs PPO (256 GPU workers).** The plots show average reward and standard deviation over 5 seeds. The x-axis is the wall-clock time. The runs ended after one week with varying number of time-steps. The training curves correspond to the evaluation runs in Figure 3.

B.3 STEP-BASED TRAINING CURVES

In Figure 5 we also provide the same convergence plots where the x-axis is now the number of online interactions with the environment, thus excluding the tree expansion complexity. As seen, due to the complexity of the tree expansion, less steps are conducted during training (limited to one week) as the depth increases. In this plot, the monotone improvement of the reward with increasing tree depth is noticeable in most games.

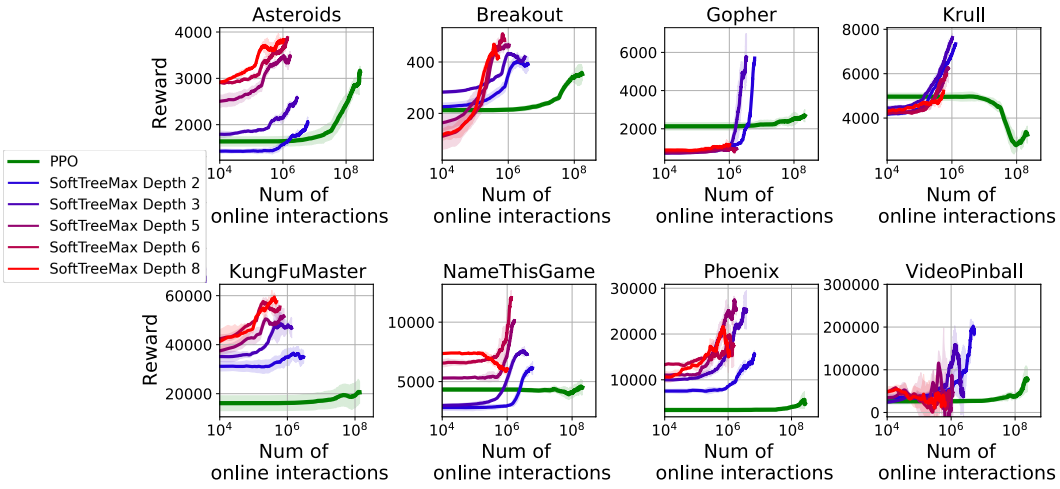

Figure 5: **Training curves: GPU SoftTreeMax (single worker) vs PPO (256 GPU workers).** The plots show average reward and standard deviation over 5 seeds. The x-axis is the number of online interactions with the environment. The runs ended after one week with varying number of time-steps. The training curves correspond to the evaluation runs in Figure 3.

We note that not for all games we see monotonicity. Our explanation for this phenomenon relates to how immediate reward contributes to performance compared to the value. Different games benefit differently from long-term as opposed to short-term planning. Games that require longer-term planning need a better value estimate. A good value estimate takes longer to obtain with larger depths, in which we apply the network to states that are very different from the ones observed so far in the buffer (recall that as in any deep RL algorithm, we train the model only on states in the buffer). If the model hasn't learned a good enough value function yet, and there is no guiding dense reward along the trajectory, the policy becomes noisier, and can take more steps to converge – even more than those we run in our week-long experiment.

For a concrete example, let us compare Breakout to Gopher. Inspecting Fig. 5, we observe that Breakout quickly (and monotonically) gains from large depths since it relies on the short term goal of simply keeping the paddle below the moving ball. In Gopher, however, for large depths (>=5), learning barely started even by the end of the training run. Presumably, this is because the task in Gopher involves multiple considerations and steps: the agent needs to move to the right spot and then hit the mallet the right amount of times, while balancing different locations. This task requires long-term planning and thus depends more strongly on the accuracy of the value function estimate. In that case, for depth 5 or more, we would require more train steps for the value to "kick in" and become beneficial beyond the gain from the reward in the tree.

The figures above convey two key observations that occur for at least some non-zero depth: (1) The final performance with the tree is better than PPO (Fig. 3); and (2) the intermediate step-based results with the tree are better than PPO (Fig. 5). This leads to our main takeaway from this work — there is no reason to believe that the vanilla policy gradient algorithm should be better than a multi-step variant. Indeed, we show that this is not the case.

C Further Discussion

C.1 The case of $\lambda_2(P^{\pi_b}) = 0$

When P^{π_b} is rank one, it is not only its variance that becomes 0, but also the norm of the gradient itself (similarly to the case of $d \to \infty$). Note that such a situation will happen rarely, in degenerate MDPs. This is a local minimum for SoftTreeMax and it would cause the PG iteration to get stuck, and to the optimum in the (desired but impractical) case where π_b is the optimal policy. However, a similar phenomenon was also discovered in the standard softmax with deterministic policies:

$\theta(s,a) \to \infty$ for one a per s. PG with softmax would suffer very slow convergence near these local equilibria, as observed in Mei et al. (2020a). To see this, note that the softmax gradient is $\nabla_\theta \log \pi_\theta(a|s) = e_a - \pi_\theta(\cdot|s)$, where $e_a \in [0,1]^A$ is the vector with 0 everywhere except for the a-th coordinate. I.e., it will be zero for a deterministic policy. SoftTreeMax avoids these local optima by integrating the reward into the policy itself (but may get stuck in another, as discussed above).