Supplementary Material for

Feedback-Based Tree Search for Reinforcement Learning

Daniel R. Jiang, Emmanuel Ekwedike, Han Liu

A Outline

Section B contains the proofs for the results stated in the main paper.

- The proofs of Lemma 1 and Lemma 2 from the main paper are given in Sections B.1 and B.2. These two lemmas are important in that they provide the main structure for the sample complexity analysis. The bounds hold pointwise.
- In Section B.3, we provide some additional lemmas that are omitted from the main paper.
- Section B.4 gives the proof of Lemma 3 from the main paper, which makes use of Lemma 2 and the results from Section B.3.
- We prove the main result, Theorem 1, in Section B.5.

Lastly, in Section C, we provide additional implementation details regarding the neural network architecture, state features, and computation.

B Proofs

B.1 Proof of Lemma 1

This proof is a modification of arguments used in [Lazaric et al., 2016, Equation 8 and Theorem 7]. By the fixed point property of $T_{\pi_{k+1}}$ and the definition of the Bellman operator T, we have $V^{\pi_k} - V^{\pi_{k+1}} \leq T^d V^{\pi_k} - T_{\pi_{k+1}} V^{\pi_{k+1}}$. Subtracting and adding $T_{\pi_{k+1}} V^{\pi_k}$:

$$\begin{aligned} V^{\pi_k} - V^{\pi_{k+1}} &\leq T^d V^{\pi_k} - T_{\pi_{k+1}} V^{\pi_k} + T_{\pi_{k+1}} V^{\pi_k} - T_{\pi_{k+1}} V^{\pi_{k+1}} \\ &\leq T^d V^{\pi_k} - T_{\pi_{k+1}} V^{\pi_k} + (\gamma P_{\pi_{k+1}})\big(V^{\pi_k} - V^{\pi_{k+1}}\big). \end{aligned} \tag{B.1}$$

Similarly, we will bound the difference between V^* and $V^{\pi_{k+1}}$ in terms of the distances between $V^* - V^{\pi_k}$ and $V^{\pi_k} - V^{\pi_{k+1}}$:

$$V^* - V^{\pi_{k+1}} \leq (\gamma P_{\pi^*})^d \big(V^* - V^{\pi_k}\big) + T^d V^{\pi_k} - T_{\pi_{k+1}} V^{\pi_k} + (\gamma P_{\pi_{k+1}})\big(V^{\pi_k} - V^{\pi_{k+1}}\big). \tag{B.2}$$

Using the bound $V^{\pi_k} - V^{\pi_{k+1}} \leq [I - (\gamma P_{\pi_{k+1}})]^{-1} \left(T^d V^{\pi_k} - T_{\pi_{k+1}} V^{\pi_k}\right)$ from (B.1) on the last term of the right side of (B.2) along with a power series expansion on the inverse, we obtain:

$$\begin{aligned} V^* - V^{\pi_{k+1}} &\leq (\gamma P_{\pi^*})^d \left(V^* - V^{\pi_k}\right) + \left[I + (\gamma P_{\pi_{k+1}}) \sum_{j=0}^{\infty} (\gamma P_{\pi_{k+1}})^j\right] \left(T^d V^{\pi_k} - T_{\pi_{k+1}} V^{\pi_k}\right) \\ &= (\gamma P_{\pi^*})^d \left(V^* - V^{\pi_k}\right) + \left[I - (\gamma P_{\pi_{k+1}})\right]^{-1} \left(T^d V^{\pi_k} - T_{\pi_{k+1}} V^{\pi_k}\right), \end{aligned}$$

which can be iterated to show:

$$V^* - V^{\pi_K} \leq (\gamma P_{\pi^*})^{Kd} \left(V^* - V^{\pi_0}\right) + \sum_{k=1}^{K} (\gamma P_{\pi^*})^{(K-k)\,d} \left[I - (\gamma P_{\pi_k})\right]^{-1} \left(T^d V^{\pi_{k-1}} - T_{\pi_k} V^{\pi_{k-1}}\right).$$

The statement from the lemma follows from taking absolute value, bounding by the maximum norm, and integrating.

B.2 Proof of Lemma 2

For part (a), we note the following:

$$\begin{aligned} \|T_\pi V - T_\pi V^\mu\|_{1,\rho_1} &= \gamma \int_{\mathcal{S}} \left|(P_\pi V)(s) - (P_\pi V^\mu)(s)\right| \rho_1(ds) \\ &\leq \gamma \int_{\mathcal{S}} \left|V(s) - V^\mu(s)\right| \frac{d(\rho_1 P_\pi)}{d\rho_0} \rho_0(ds) \\ &\leq \gamma \left\| \frac{d(\rho_1 P_\pi)}{d\rho_0} \right\|_\infty \|V - V^\mu\|_{1,\rho_0}. \end{aligned}$$

By the concentrability conditions of Assumption 5, the right-hand-side can be bounded by $\gamma\, A_1' \|V - V^\mu\|_{1,\rho_0}$. Now, we can apply the same steps with the roles of $T_\pi V$ and $T_\pi V^\mu$ reversed to see that the same inequality holds for $\|T_\pi V - T_\pi V^\mu\|_{1,\rho_1}$ and part (a) is complete.

For part (b), we partition the state space $\mathcal{S}$ into two sets:

$$\mathcal{S}^+ = \big\{s \in \mathcal{S} : (T^d J)(s) \geq (T^d V^\mu)(s)\big\} \quad \text{and} \quad \mathcal{S}^- = \big\{s \in \mathcal{S} : (T^d J)(s) < (T^d V^\mu)(s)\big\}.$$

We start with $\mathcal{S}^+$. Consider the finite-horizon d-stage MDP with terminal value J and the same dynamics as our infinite-horizon MDP of interest. Let $\pi_1^J, \pi_2^J, \ldots, \pi_d^J$ be the time-dependent optimal policy for this MDP. Thus, we have

$$T_{\pi_1^J} T_{\pi_2^J} \cdots T_{\pi_d^J} J = T^d J \quad \text{and} \quad T_{\pi_1^J} T_{\pi_2^J} \cdots T_{\pi_d^J} V^\mu \leq T^d V^\mu.$$

Using similar steps as for part (a), the following hold:

$$\begin{aligned}
\int_{\mathcal{S}^+} \big[(T^d J)(s) - (T^d V^\mu)(s)\big]\, \rho_1(ds) &\leq \int_{\mathcal{S}^+} \big[(T^d T_\mu^h V)(s) - (T_{\pi_1^J} T_{\pi_2^J} \cdots T_{\pi_d^J} T_\mu^h V^\mu)(s)\big]\, \rho_1(ds) \\
&\leq \gamma^{d+h} \int_{\mathcal{S}^+} \big|V(s) - V^\mu(s)\big| \, \frac{d(\rho_1 P_{\pi_1^J} P_{\pi_2^J} \cdots P_{\pi_d^J} P_\mu^h)}{d\rho_0}\, \rho_0(ds) \\
&\leq \gamma^{d+h}\, A'_{d+h} \int_{\mathcal{S}^+} \big|V(s) - V^\mu(s)\big|\, \rho_0(ds).
\end{aligned}$$

Now, using the optimal policy with respect to the d-stage MDP with terminal condition V^μ, we can repeat these steps to show that

$$\int_{\mathcal{S}^-} \big[(T^d V^\mu)(s) - (T^d J)(s)\big]\, \rho_1(ds) \leq \gamma^{d+h}\, A'_{d+h} \int_{\mathcal{S}^-} \big|V(s) - V^\mu(s)\big|\, \rho_0(ds).$$

Summing the two inequalities, we obtain:

$$\begin{aligned}
\|T^d J - T^d V^\mu\|_{1,\rho_1} &\leq \gamma^{d+h}\, A'_{d+h} \left[\int_{\mathcal{S}^+} \big|V(s) - V^\mu(s)\big|\, \rho_0(ds) + \int_{\mathcal{S}^-} \big|V(s) - V^\mu(s)\big|\, \rho_0(ds)\right] \\
&= \gamma^{d+h}\, A'_{d+h}\, \|V - V^\mu\|_{1,\rho_0},
\end{aligned}$$

which completes the proof.

B.3 Additional Technical Lemmas

Lemma B.1 (Section 4, Corollary 2 of Haussler [1992])**.** *Let $\mathcal{G}$ be a set of functions from $\mathcal{X}$ to $[0, B]$ with pseudo-dimension $\mathrm{d}_\mathcal{G} < \infty$. Then for all $0 < \epsilon \leq B$, it holds that*

$$\mathbf{P}\left(\sup_{g \in \mathcal{G}} \left|\frac{1}{m}\sum_{i=1}^m g(X^{(i)}) - \mathbf{E}\big[g(X)\big]\right| > \epsilon\right) \leq 8\left(\frac{32eB}{\epsilon}\log\frac{32eB}{\epsilon}\right)^{\mathrm{d}_\mathcal{G}} \exp\left(-\frac{\epsilon^2 m}{64B^2}\right), \quad \text{(B.3)}$$

where $X^{(i)}$ are i.i.d. draws from the distribution of the random variable X.

Lemma B.2. *Consider a policy $\mu \in \Pi$ and suppose each s^i is sampled i.i.d. from ρ_0. Define initial states $s_0^{ij} = s^i$ for all j. Analogous to Step 5 of the algorithm and Assumption 1, let:*

$$\hat{Y}(s^i) = \frac{1}{M_0}\sum_{j=1}^{M_0}\sum_{t=0}^{\infty} \gamma^t\, r(s_t^{ij}, \mu(s_t^{ij})) \quad \textit{and} \quad V \in \arg\min_{f \in \bar{\mathcal{F}}} \frac{1}{N_0}\sum_{i=1}^{N_0} \big|f(s^i) - \hat{Y}(s^i)\big|.$$

For $\delta \in (0,1)$ and $\epsilon \in (0, V_{\max})$, if the number of sampled states N_0 satisfies the condition:

$$N_0 \geq \left(\frac{32 V_{\max}}{\epsilon}\right)^2 \left[\log \frac{32}{\delta} + 2\mathrm{d}_{\bar{\mathcal{F}}} \log \frac{64 e V_{\max}}{\epsilon}\right] =: \Gamma_{\mathrm{a}}(\epsilon, \delta),$$

and the number of rollouts performed from each state M_0 satisfies:

$$M_0 \geq 8 \left(\frac{V_{\max}}{\epsilon}\right)^2 \log \frac{8 N_0}{\delta} =: \Gamma_{\mathrm{b}}(\epsilon, \delta),$$

then we have the following bound on the error of the value function approximation:

$$\|V - V^{\mu}\|_{1,\rho_0} \leq \min_{f \in \bar{\mathcal{F}}} \|f - V^{\mu}\|_{1,\rho_0} + \epsilon,$$

with probabilty at least $1 - \delta$.

Proof. Recall that the estimated value function V satisfies

$$V \in \arg\min_{f \in \widehat{\mathcal{F}}} \frac{1}{N_0} \sum_{i=1}^{N_0} \left| f(s^i) - \frac{1}{M_0} \sum_{j=1}^{M_0} \left[V^{\pi}(s_0^i) + \xi^j(s_0^i) \right] \right|,$$

where for each i, the terms $\xi^j(s_0^i)$ are i.i.d. mean zero error. The inner summation over j is an equivalent way to write $\hat{Y}(s_0^i)$. Noting that the rollout results $V^{\mu}(s_0^i) + \xi^j(s_0^i) \in [0, V_{\max}]$, we have by Hoeffding's inequality followed by a union bound:

$$\mathbf{P}\left(\max_i \left|\hat{Y}(s_0^i) - V^{\mu}(s_0^i)\right| > \epsilon\right) \leq N_0 \, \Delta_1(\epsilon, M_0), \tag{B.4}$$

where $\Delta_1(\epsilon, M_0) = 2 \exp\left(-2 M_0 \epsilon^2 / V_{\max}^2\right)$. Define the function

$$\Delta_2(\epsilon, N_0) = 8 \left(\frac{32 e V_{\max}}{\epsilon} \log \frac{32 e V_{\max}}{\epsilon}\right)^{\mathrm{d}_{\bar{\mathcal{F}}}} \exp\left(-\frac{\epsilon^2 N_0}{64 V_{\max}^2}\right),$$

representing the right-hand-side of the bound in Lemma B.1 with $B = V_{\max}$ and $m = N_0$. Next, we define the loss minimizing function $f^* \in \arg\min_{f \in \bar{\mathcal{F}}} \|f - V^{\mu}\|_{1,\rho_0}$. By Lemma B.1, the probabilities of the events

$$\begin{aligned} &\left\{ \left| \|V - V^{\mu}\|_{1,\rho_0} - \frac{1}{N_0} \sum_{i=1}^{N_0} \left|V(s^i) - V^{\mu}(s^i)\right| \right| > \frac{\epsilon}{4} \right\} \text{ and} \\ &\left\{ \left| \|f^* - V^{\mu}\|_{1,\rho_0} - \frac{1}{N_0} \sum_{i=1}^{N_0} \left|f^*(s^i) - V^{\mu}(s^i)\right| \right| > \frac{\epsilon}{4} \right\} \end{aligned} \tag{B.5}$$

are each bounded by $\Delta_2(\epsilon/4, N_0)$. Also, it follows by the definition of V that

$$\frac{1}{N_0} \sum_{i=1}^{N_0} \left|V(s^i) - \hat{Y}(s^i)\right| \leq \frac{1}{N_0} \sum_{i=1}^{N_0} \left|f^*(s^i) - \hat{Y}(s^i)\right|.$$

Therefore, using (B.4) twice and (B.5) once, we have by a union bound that the inequality

$$\|V - V^{\mu}\|_{1,\rho_0} \leq \min_{f \in \bar{\mathcal{F}}} \|f - V^{\mu}\|_{1,\rho_0} + \epsilon$$

happens with probability greater than $1 - 2\,N_0\,\Delta_1(\epsilon/4, M_0) - 2\,\Delta_2(\epsilon/4, N_0)$. We then choose N_0 so that $\Delta_2(\epsilon/4, N_0) = \delta/4$ (following Haussler [1992], we utilize the inequality $\log(a \log a) < 2\log(a/2)$ for $a \geq 5$). To conclude, we choose M_0 so that $\Delta_1(\epsilon/4, M_0) = \delta/(4N_0)$. □

Lemma B.3 (Sampling Error)**.** *Suppose* $|\mathcal{A}| = 2$ *and let* $\mathrm{d}_{\bar{\Pi}}$ *be the VC-dimension of* $\bar{\Pi}$. *Consider* $Z, V \in \mathcal{F}$ *and suppose each* s^i *is sampled i.i.d. from* ρ_1. *Also, let* w^j *be i.i.d. samples from the standard uniform distribution and* $g : \mathcal{S} \times \mathcal{A} \times [0,1] \to \mathcal{S}$ *be a transition function such that* $g(s,a,w)$ *has the same distribution as* $p(\cdot\,|\,s,a)$. *For* $\delta \in (0,1)$ *and* $\epsilon \in (0, V_{\max})$, *if the number of sampled states* N_1 *satisfies the condition:*

$$N_1 \geq 128 \left(\frac{V_{\max}}{\epsilon}\right)^2 \left[\log\frac{8}{\delta} + \mathrm{d}_{\bar{\Pi}} \log\frac{e\,N_1}{\mathrm{d}_{\bar{\Pi}}}\right] =: \Gamma_{\mathrm{c}}(\epsilon, \delta, N_1),$$

and the number of sampled transitions L_1 *satisfies:*

$$L_1 \geq 128 \left(\frac{V_{\max}}{\epsilon}\right)^2 \left[\log\frac{8}{\delta} + \mathrm{d}_{\bar{\Pi}} \log\frac{e\,L_1}{\mathrm{d}_{\bar{\Pi}}}\right] =: \Gamma_{\mathrm{d}}(\epsilon, \delta, L_1),$$

then we have the bounds:

(a) $$\sup_{\pi \in \bar{\Pi}} \left| \frac{1}{N_1} \sum_{i=1}^{N_1} |Z(s^i) - (T_\pi V)(s^i)| - \|Z - T_\pi V\|_{1,\rho_1} \right| \leq \epsilon \quad \textit{w.p. at least } 1-\delta.$$

(b) $$\sup_{\pi \in \bar{\Pi}} \left| \frac{1}{L_1} \sum_{j=1}^{L_1} \Big[r(s^i, \pi(s^i)) + \gamma V\big(g(s^i, \pi(s^i), w^j)\big) \Big] - (T_\pi V)(s^i) \right| \leq \epsilon \quad \textit{w.p. at least } 1-\delta.$$

Proof. We remark that in both (a) and (b), the term within the absolute value is bounded between 0 and $V_{\max}$. A second remark is that we reformulated the problem using w^j to take advantage of the fact that these random samples do not depend on the policy π. Such a property is required to invoke [Györfi et al., 2006, Theorem 9.1], a result that [Lazaric et al., 2016, Lemma 3] depends on. With these two issues in mind, an argument similar to the proof of [Lazaric et al., 2016, Lemma 3] gives the conclusion for both (a) and (b). □

B.4 Proof of Lemma 3

On each iteration of the the algorithm, two random samples are used: $\mathcal{S}_{0,k}$ and $\mathcal{S}_{1,k}$. From $\mathcal{S}_{0,k}$, we obtain V_k and from $\mathcal{S}_{1,k}$ we obtain π_{k+1}. Let $\mathcal{S}_k = (\mathcal{S}_{0,k}, \mathcal{S}_{1,k})$ represent both of the samples at iteration k. We define:

$$\mathcal{G}_{k-1} = \sigma\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{k-1}\} \quad \text{and} \quad \mathcal{G}'_{k-1} = \sigma\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{k-1}, \mathcal{S}_{0,k}\}.$$

Due to the progression of the algorithm with two random samples per iteration, we will analyze each iteration in two steps. We first separate the two random samples by noting

that:

$$\begin{aligned}\|T^d V^{\pi_k} - T_{\pi_{k+1}} V^{\pi_k}\|_{1,\rho_1} \leq\ & \|T^d V^{\pi_k} - T^d J_k\|_{1,\rho_1} + \|T_{\pi_{k+1}} V_k - T_{\pi_{k+1}} V^{\pi_k}\|_{1,\rho_1} \\ & + \|T^d J_k - T_{\pi_{k+1}} V_k\|_{1,\rho_1} \\ \leq\ & (\gamma A'_1 + \gamma^{d+h} A'_{d+h}) \|V_k - V^{\pi_k}\|_{1,\rho_0} + \|T^d J_k - T_{\pi_{k+1}} V_k\|_{1,\rho_1},\end{aligned} \tag{B.6}$$

where the first inequality follows by adding and subtracting terms and the triangle inequality while the second inequality follows by Lemma 2. Now, we may analyze the first term on the right-hand-side conditional on $\mathcal{G}_{k-1}$ and the second term conditional on $\mathcal{G}'_{k-1}$.

As it is currently stated, Lemma B.2 gives an unconditional probability for a fixed policy μ. However, since $\mathcal{S}_{0,k}$ is independent from $\mathcal{G}_{k-1}$ and π_k is $\mathcal{G}_{k-1}$-measurable, we can utilize Lemma B.2 in a conditional setting using a well-known property of conditional expectations [Resnick, 2013, Property 12, Section 10.3]. This property will be repeatedly used in this proof (without further mention). We obtain that for a sample size $N_0 \geq \Gamma_{\mathrm{a}}(\epsilon'/(\gamma A'_1 + \gamma^{d+h} A'_{d+h}), \delta')$,

$$\mathbf{P}\Big(\|V_k - V^{\pi_k}\|_{1,\rho_0} > \min_{f \in \bar{\mathcal{F}}} \|f - V^{\pi_k}\|_{1,\rho_0} + \epsilon'/(\gamma A'_1 + \gamma^{d+h} A'_{d+h}) \,\Big|\, \mathcal{G}_{k-1}\Big) \leq \delta'. \tag{B.7}$$

It remains for us to analyze the error of the second term $\|T^d J_k - T_{\pi_{k+1}} V_k\|_{1,\rho_1}$. By part (a) of Lemma B.3 with $Z = T^d J_k$ and $V = V_k$, if $N_1 \geq \Gamma_{\mathrm{c}}(\epsilon', \delta', N_1)$ and s^i are sampled i.i.d. from ρ_1, we have

$$\mathbf{P}\left(\left|\frac{1}{N_1}\sum_{i=1}^{N_1} \left|(T^d J_k)(s^i) - (T_{\pi_{k+1}} V_k)(s^i)\right| - \left\|T^d J_k - T_{\pi_{k+1}} V_k\right\|_{1,\rho_1}\right| > \epsilon' \,\middle|\, \mathcal{G}'_{k-1}\right) \leq \delta'. \tag{B.8}$$

The term $(T_{\pi_{k+1}} V_k)(s^i)$ is approximated using L_1 samples. Part (b) of Lemma B.3 along with a union bound shows that if $L_1 \geq \Gamma_{\mathrm{d}}(\epsilon', \delta'/N_1, L_1)$, then

$$\mathbf{P}\Big(\max_i \left|\hat{Q}_k(s^i, \pi_{k+1}(s^i)) - (T_{\pi_{k+1}} V_k)(s^i)\right| > \epsilon' \,\Big|\, \mathcal{G}'_{k-1}\Big) \leq \delta'. \tag{B.9}$$

Similarly, by Assumption 3, if the number of iterations of `MCTS` M_1 exceeds $m(\epsilon', \delta'/N_1)$, we can take a union bound to arrive at

$$\mathbf{P}\Big(\max_i \left|\hat{U}_k(s^i) - (T^d J_k)(s^i)\right| > \epsilon' \,\Big|\, \mathcal{G}'_{k-1}\Big) \leq \delta'. \tag{B.10}$$

The maximum over i can be replaced with an average over the N_1 samples and the conclusion of the last two bounds would remain unchanged. Since π_{k+1} is assumed to optimize a quantity involving $\hat{U}_k$ and $\hat{Q}_k$, we want to relate this back to $\|T^d J_k - T_{\pi_{k+1}} V_k\|_{1,\rho_1}$. Indeed, taking expectation of both sides of inequalities (B.8)–(B.10) and then combining, we obtain that with probability at least $1 - 3\delta'$,

$$\left\|T^d J_k - T_{\pi_{k+1}} V_k\right\|_{1,\rho_1} \leq \frac{1}{N_1}\sum_{i=1}^{N_1} \left|\hat{U}_k(s^i) - \hat{Q}_k(s^i, \pi_{k+1}(s^i))\right| + 3\epsilon'$$

$$\leq \frac{1}{N_1} \sum_{i=1}^{N_1} \left| \hat{U}_k(s^i) - \hat{Q}_k(s^i, \tilde{\pi}(s^i)) \right| + 3\epsilon'$$

where $\tilde{\pi} \in \arg\min_{\pi \in \bar{\Pi}} \|T^d V^{\pi_k} - T_\pi V^{\pi_k}\|_{1,\rho_1}$. Following the same steps in reverse , we have:

$$\left\| T^d J_k - T_{\pi_{k+1}} V_k \right\|_{1,\rho_1} \leq \min_{\pi \in \bar{\Pi}} \|T^d V^{\pi_k} - T_\pi V^{\pi_k}\|_{1,\rho_1} + 6\epsilon', \tag{B.11}$$

with probability at least $1 - 6\,\delta'$. Finally, we take expectation of both sides of (B.7) and then combine with (B.6) and (B.11) while setting $\epsilon' = \epsilon/7$ and $\delta' = \delta/7$ to obtain

$$\begin{aligned} \|T^d V^{\pi_k} - T_{\pi_{k+1}} V^{\pi_k}\|_{1,\rho_1} \leq (\gamma A'_1 + \gamma^{d+h} A'_{d+h}) \min_{f \in \bar{\mathcal{F}}} \|f - V^{\pi_k}\|_{1,\rho_0} \\ + \min_{\pi \in \bar{\Pi}} \|T^d V^{\pi_k} - T_\pi V^{\pi_k}\|_{1,\rho_1} + \epsilon \end{aligned}$$

with probability at least $1 - \delta$.

B.5 Proof of Theorem 1

This proof synthesizes the previous lemmas. From the definition of $D_0(\bar{\Pi}, \bar{\mathcal{F}})$ and $\mathbb{D}_1^d(\bar{\Pi})$ from the main paper, we note that if the sample size assumptions of Lemma 3 are satisfied,

$$\|T^d V^{\pi_k} - T_{\pi_{k+1}} V^{\pi_k}\|_{1,\rho_1} \leq B'_\gamma \, \mathbb{D}_0(\bar{\Pi}, \bar{\mathcal{F}}) + \mathbb{D}_1^d(\bar{\Pi}) + \epsilon, \tag{B.12}$$

with probability at least $1 - \delta$. This removes any dependence on the iteration k from the right-hand-side. We now integrate all results with Lemma 1 in order to find a bound on the suboptimality $\|V^* - V^{\pi_K}\|_{1,\nu}$. Consider the distribution $\Lambda_{\nu,k}$, as defined in Lemma 1, which needs to be related to ν. We can use the power series expansion to write:

$$\Lambda_{\nu,k} = \nu \, (P_{\pi^*})^{K-k} \sum_{i=0}^{\infty} (\gamma P_{\pi_k})^i.$$

For a fixed i, the measure ν is transformed by applying π^* a total of $K - k$ times and then π_k a total of i times. We see that the summation term on the right-hand-side of Lemma 1 can be upper-bounded in the following way:

$$\begin{aligned} \sum_{k=1}^{K} \gamma^{K-k} \left\| T^d V^{\pi_{k-1}} - T_{\pi_k} V^{\pi_{k-1}} \right\|_{1,\Lambda_{\nu,k}} \\ \leq \left(\sum_{j=0}^{K-1} \sum_{i=0}^{\infty} \gamma^{j+i} \, A_{j+i} \right) \max_{k \leq K} \left\| T^d V^{\pi_{k-1}} - T_{\pi_k} V^{\pi_{k-1}} \right\|_{1,\rho_1}, \end{aligned}$$

where we use Assumption 5 with $m = K - k + i$, maximize over k for the loss term, and then re-index with $j = K - k$. The coefficient in parentheses can be upper-bounded by B_γ (since all A_{j+i} are nonnegative). Finally, we use (B.12) and then a union bound over the K iterations to conclude the statement of the theorem.

C Implementation Details

C.1 Neural Network Architecture

The policy and value function approximations use fully-connected neural networks with five and two hidden layers, respectively, and SELU (scaled exponential linear unit) activation [Klambauer et al., 2017]. The policy network contains two sets of outputs: (1) one of seven actions (no action, normal attack, move, skill 1, skill 2, skill 3, and heal) and (2) a two-dimensional direction parameter used for the action. The first two hidden layers are shared and have 120 and 100 hidden units, while each of the two outputs corresponds to a set of three hidden layers with 80, 70, and 50 hidden units. The value function approximation uses a fully-connected network with 128 hidden units in the first layer and 96 hidden units in the second layer. As mentioned in the main paper, this architecture is consistent across all agents whenever policy and/or value networks are needed.

C.2 Features of the State

As shown in Table 1, the state of the game is represented by 41-dimensional feature vector, which was constructed using the output from the game engine and API. The features consists of basic attributes of the two heroes, the computer-controlled units, and structures. The feature lists also have information on the relative positions of the other units and structures with respect to the hero controlled by algorithm.

Table 1: State Feature List

No.	Feature	Dimensions
1	Location of Hero 1	2
2	Location of Hero 2	2
3	HP of Hero 1	1
4	HP of Hero 2	1
5	Hero 1 skill cooldowns	5
6	Hero 2 skill cooldowns	5
7	Direction to enemy hero	3
8	Direction to enemy tower	4
9	Direction to enemy minion	3
10	Enemy tower HP	1
11	Enemy minion HP	1
12	Direction to the spring	3
13	Total HP of allied minions	1
14	Enemy's tower attacking Hero 1	3
15	Hero 1 in range of enemy towers	3
16	Hero 2 in range of enemy towers	3

C.3 Tree Search Details

We provide some more information regarding the implementation of feedback-based tree search. A major challenge in implementing in *King of Glory* is that the game engine can only move forward, meaning that our sampled states are not i.i.d. and instead follow the trajectory of the policy induced by `MCTS`. However, to decrease the correlation between visited states, we inject random movements and random switches to the internal AI policy in order to move to a "more random" next state. Rollouts are performed on separate processors to enable tree search in a game engine that cannot rewind. All experiments use the `c4.2xlarge` instances on Amazon Web Services, and we utilized parallelization across four cores within each call to `MCTS`.

References

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer Science & Business Media, 2006.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pages 972–981, 2017.

A. Lazaric, M. Ghavamzadeh, and R. Munos. Analysis of classification-based policy iteration algorithms. *Journal of Machine Learning Research*, 17(19):1–30, 2016.

S. I. Resnick. *A Probability Path*. Springer Science & Business Media, 2013.