AGENTIC TRANSFORMERS PROVABLY LEARN TO SEARCH VIA REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Tree search is a central abstraction behind many language-agent reasoning and decision-making tasks: agents must explore actions, remember failures, and backtrack toward promising alternatives. Yet, we lack a theoretical understanding of how transformer-based policies acquire such search capabilities from the training dynamics of reinforcement learning (RL). We study this question in a stochastic k-ary tree environment, where an agentic transformer observes only its trajectory history through interaction and receives a terminal reward for reaching a hidden leaf goal node. We first construct a two-head transformer that implements randomized depth-first search (DFS): one head tracks previous actions, while the other detects failure outcomes and triggers backtracking. We then analyze the training dynamics of policy gradient under a depth-wise curriculum, showing that this same DFS mechanism emerges in stages from sparse reinforcement feedback without expert demonstrations. The resulting policy exhibits depth generalization: after training only on depth-1 and depth-2 trees, it succeeds on deeper full trees. We further show that, under imbalanced goal distributions, discounting the return leads to a ranked DFS policy that prioritizes higher-probability branches. Overall, our results identify a mechanistic normal form for transformer-based search, in which attention heads specialize and cooperate to extract decision-relevant traces from context and convert them into agentic action selection via RL training.

1 INTRODUCTION

Tree search is a core computational structure in modern learning agents: they must explore different paths, memorize failed branches, and backtrack towards promising alternatives while searching for a target embedded at the leaf of the tree (Browne et al., 2012; Silver et al., 2016; 2018; Lample et al., 2022). It underlies classical systems such as Monte Carlo tree search in game playing and proof search in automated theorem proving, and has recently reappeared as an explicit test-time reasoning mechanism in language agents, such as Tree of Thoughts, planning-based reasoning, Language Agent Tree Search, and search for web agents (Yao et al., 2023a; Hao et al., 2023; Zhou et al., 2024; Koh et al., 2025; Snell et al., 2025).

In these agentic settings, the search tree is usually not handed to the agent as input. Rather, it is a hidden environment revealed through interaction: each action exposes only local information, and the reward can be highly sparse depending on the hidden goal. As a result, the agent needs to consult the entire trajectory to determine the optimal next action. Reinforcement learning (RL) has been established as an effective approach to train the agent's policy – parameterized by autoregressive transformers – to search via policy gradient methods, even in the absence of explicit supervision of when to branch or backtrack along the tree.

Despite this practical progress, the theory of how transformers (Vaswani et al., 2017) learn to search remains limited, especially from a training dynamic perspective. A growing literature shows that transformers can implement or learn graph reasoning tasks, but mostly in supervised or RL settings with static instances, where the graph is specified in the prompt or encoded as a fixed abstraction. The learned reasoning skills are therefore the retrieval, reachability, or generation of paths, not the exploration of an unknown graph through interactions (Saparov et al., 2025; Wang et al., 2024; Ye et al., 2026; Yang et al., 2025; Ran-Milo et al., 2026; Wang et al., 2026). It remains unclear how a transformer embodies an algorithmic search policy and how policy-gradient training discovers it despite nonconvexity. We therefore ask:

How can policy-gradient training make a transformer policy learn tree-search algorithms when the tree is hidden and must be discovered through stochastic interaction?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

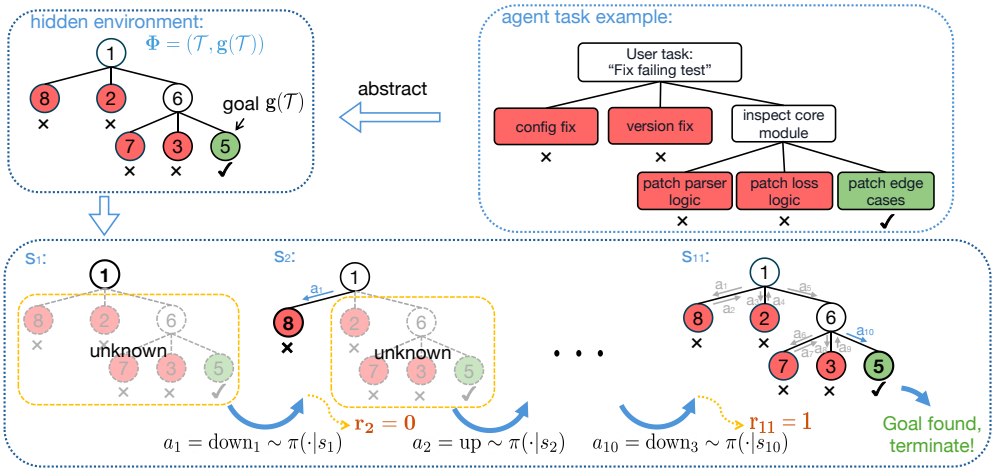


Figure 1: An illustration of the interactive tree-search task studied in this paper. The environment is a hidden stochastic k-ary tree (with $k = 3$ in the illustration), and at each step the agent chooses a search action from its interaction history rather than receiving the full tree as input.

We study this question in a stylized environment interacting with stochastic hidden k-ary trees, illustrated in Figure 1. At the start of each episode, the environment samples a tree with distinct node identities and a hidden goal leaf. The agent begins at the root. At each step, it observes the current node identity together with a local status label indicating whether the node is internal, a non-goal leaf, or the goal; it then chooses a local action, either moving to a child or backtracking to the parent. The agent receives a positive reward only upon reaching the goal.

Our analysis starts with an explicit two-head shallow transformer construction for random depth-first search (DFS): one head retrieves which child actions have already been tried at the current node, while the other uses outcome labels to decide when to backtrack. Under a depth-wise curriculum, plain policy-gradient training learns this circuit from terminal reward alone. Depth-1 training learns two primitives: failed-leaf backtracking and repeated-action avoidance. Depth-2 training composes them and learns when to backtrack from exhausted internal nodes. As a result, training only on depth-1 and depth-2 trees yields the random DFS policy that generalizes at test time to deeper trees. We also extend the analysis to imbalanced goal distributions, where maximizing discounted return yields a ranked DFS policy that prioritizes higher-probability branches.

To the best of our knowledge, our work gives the first optimization-level analysis of how policy gradient training makes a transformer policy acquire tree search algorithms in a hidden stochastic RL environment. Our stylized setting preserves the core control problem of interactive search while abstracting away the burden of opaque representation learning faced by practical LLM agents. This yields a tractable *mechanistic normal form* for search and a provable testbed for future work on, for example, memory management, context compression, or optimization algorithms for LLM agents, see Appendix A.1 for a detailed discussion.

1.1 Our Contributions

This paper gives an optimization-level account of how a transformer policy can acquire a genuine search algorithm through reinforcement learning. Rather than assuming that the search tree is provided in the prompt, we study a hidden stochastic tree that must be discovered online from trajectory history and sparse terminal reward. In this setting, we show that policy-gradient training can shape a shallow transformer into an algorithmic policy: its attention heads retrieve the traces needed for search, combine them into backtracking and action-avoidance decisions, and thereby implement depth-first search with provable generalization beyond the training depths.

Our main contributions are as follows.

- **A mechanistic transformer construction for interactive search.** We give an explicit two-head shallow transformer that implements randomized DFS. One head records which child actions have already been tried at the current node, while the other detects search outcomes and triggers backtracking. The two heads cooperate to form the search algorithm.
- **An optimization analysis of policy-gradient learning.** We prove that policy-gradient training can induce this search circuit from sparse terminal reward alone, without expert demonstrations or

action-level supervision. Under a depth-wise curriculum, the transformer first learns the primitive behaviors of failed-leaf backtracking and repeated-action avoidance, and then learns to compose them into correct internal-node backtracking.

- **Provable depth generalization from shallow training.** We show that training on depth-1 and depth-2 trees suffices for the learned policy to succeed on deeper full trees. The result demonstrates that the transformer learns compositional search logic rather than memorizing short trajectories.
- **Extension to imbalanced goal distributions.** We extend the analysis beyond the balanced case and show that, when goal probabilities differ across branches, discounted return selects a ranked DFS policy that searches more promising branches first and yields the minimum expected path length among successful policies.

These results isolate a core mechanism behind agentic reasoning: using memory of past interaction to control future exploration under partial observability and sparse feedback. By making this mechanism explicit and provable, our work provides a tractable foundation for understanding memory, credit assignment, and algorithmic generalization in agentic ML systems.

1.2 Related Work

Test-time agentic search with language models. Tree search is a central tool for sequential decision-making, from Monte Carlo tree search and AlphaGo-style systems to neural theorem proving (Browne et al., 2012; Silver et al., 2016; 2018; Lample et al., 2022). In language agents, search has appeared mainly as an inference-time mechanism: ReAct and Reflexion use interaction histories and feedback (Yao et al., 2023b; Shinn et al., 2023), while Tree-of-Thoughts, RAP, LATS, and web-agent tree search explicitly explore alternative reasoning or environment branches (Yao et al., 2023a; Hao et al., 2023; Zhou et al., 2024; Koh et al., 2025). Test-time compute scaling further shows the value of verifier- or reward-guided search (Snell et al., 2025). Together, these works establish exploration, lookahead, and backtracking as useful primitives for language agents.

Transformer mechanisms for structured search and reasoning. A growing literature of theory studies how transformers represent structured search and reasoning, including graph connectivity and path-finding (Saparov et al., 2025; Wang et al., 2024), chain-of-thought (Yang et al., 2025; Huang et al., 2025), and learned graph heuristics (Ye et al., 2026). These works do model aspects of search, but mostly through static instances encoded in the input. In contrast, our setting is interactive: the tree is hidden, observations are local, and the transformer must infer the search state from history (see also Appendix A.2).

RL training dynamics for transformer reasoning. Recent theory studies how RL improves transformer reasoning from outcome-level feedback, including planning on graph abstractions (Wang et al., 2026) and curriculum or reinforcement learning with verifiable rewards (RLVR) dynamics (Bu et al., 2025; Ran-Milo et al., 2026; Huang et al., 2026). Wang et al. (2026) are closest to ours: they analyze policy gradient and Q-learning for language-model planning, modeled as generating paths on a graph. This is a search/planning abstraction, but the environment is a fixed graph and the evolving state is the generated node prefix; process rewards can also check adjacency. We instead study a transformer policy interacting with a hidden stochastic tree, where successful search requires learning when to explore, avoid repeats, and backtrack from local observations.

Notation. Let $[n] = \{1, \ldots, n\}$. Denote $A_{:,-1}$ or $A(:, -1)$ as the last column of matrix A. I_n denotes the identity matrix of size $n \times n$, $1_{m\times n}, 0_{m\times n}$ denote the all-ones matrix and all-zeros matrix of size $m \times n$, respectively. Let $\|\cdot\|_p$ represent the ℓ_p norm of a vector. In addition, $g(x) \lesssim f(x)$ or $g(x) = O(f(x))$ (resp. $g(x) \gtrsim f(x)$ or $f(x) = O(g(x))$) means that $g(x) \leqslant c \cdot f(x)$ (resp. $g(x) \geqslant c \cdot f(x)$) for some constant $c > 0$ and all x in some range; $g(x) \asymp f(x)$ means that $g(x) \lesssim f(x)$ and $g(x) \gtrsim f(x)$. We write $g(x) = o(f(x))$ if $g(x)/f(x) \to 0$ in the relevant limit.

2 Agentic Tree Search via Reinforcement Learning

We define the following tree search task and the corresponding RL formulation.

Tree search. Throughout this work, we consider full k-ary ($k \geqslant 2$) trees $\mathcal{T}$ of depth $l(\mathcal{T}) \in [L]$ ($L \geqslant 2$) with *distinct* nodes, i.e., each node in the tree has either k children or no children. The nodes of the trees are generated by sampling randomly from the set $[N]$ without replacement, where we

assume $N \geqslant \frac{k^{L+1}-1}{k-1}$ to ensure any such k-ary tree has distinct nodes. Given a tree $\mathcal{T}$, we choose one of its leaf nodes as the goal node $g(\mathcal{T})$. The task of the agent is to find the goal node by searching over the tree without knowing its topology a priori.

Environment setup. We first fix an *environment distribution* $\mathcal{P}$ over the environment parameters $\Phi = (\mathcal{T}, g(\mathcal{T}))$. At the beginning of each episode, we sample $\Phi \sim \mathcal{P}$ independently, and the agent starts from the root node $n_1 = r(\mathcal{T})$. At the h-th timestep, the agent arrives at node $n_h \in [N]$ and takes one of the following actions from an action space $\mathcal{A}$:

$$a_h \in \mathcal{A} := \{\texttt{up}\} \cup \mathcal{A}_{\texttt{down}}, \qquad \mathcal{A}_{\texttt{down}} := \{\texttt{down}_i : i \in [k]\}, \tag{1}$$

where $\texttt{up}$ denotes moving to the parent node and $\texttt{down}_i$ denotes moving to the i-th child node. Actions are *illegal* if: (i) $\texttt{up}$ at the root, (ii) $\texttt{down}_i$ at a leaf, or (iii) the action that has already been queried at n_h. On the other hand, if a_h is legal, the agent deterministically moves to the next node n_{h+1} according to a_h. For $h \geqslant 1$, the agent also receives observation $o_h = (n_h, c_h)$, where the label c_h of n_h is defined as

$$c_h \in \{0, \times, \checkmark\}, \quad c_h = \begin{cases} 0, & \text{if } n_h \text{ is not a leaf,} \\ \times, & \text{if } n_h \text{ is a leaf and } n_h \neq g(\mathcal{T}), \\ \checkmark, & \text{if } n_h = g(\mathcal{T}). \end{cases} \tag{2}$$

Consequently, we note that $o_1 = (n_1, c_1) = (r(\mathcal{T}), 0)$. The episode terminates when $c_h = \checkmark$ or an illegal action is taken.

Markov decision process (MDP). The episode evolves deterministically given Φ and the agent's actions, and the maximum length of the episode is set as $H = 2N$.[1] We cast the interaction as a Markov decision process (MDP) across episodes by taking the *history* as the state: we set the state at timestep h as $s_h := (o_1, a_1, \ldots, o_{h-1}, a_{h-1}, o_h)$. Let $\mathcal{S}$ be the set of all possible states, and let $\mathcal{S}_\checkmark \subseteq \mathcal{S}$ denote the set of all success states, i.e., if $c_h = \checkmark$, then $s_h \in \mathcal{S}_\checkmark$. We introduce an absorbing terminal state $s_\perp$. The transition kernel is given by the posterior predictive distribution (Duff, 2002):

$$\mathbb{P}(s_{h+1} \mid s_h, a_h) = \begin{cases} \sum_{o \in \mathcal{O}} \mathbb{E}_{\Phi \sim \mathcal{P}(\cdot|s_h)}[\mathbb{1}\{o = O(s_h, a_h \mid \Phi)\}]\mathbb{1}\{s_{h+1} = (s_h, a_h, o)\}, & \\ & \text{if } s_h \notin \mathcal{S}_\checkmark,\ a_h \text{ legal at } s_h, \\ \mathbb{1}\{s_{h+1} = s_\perp\}, & \text{otherwise} \end{cases}$$

where $\mathcal{O}$ is the set of all possible observations, $O(s_h, a_h \mid \Phi)$ is the deterministic observation produced by taking action a_h in state s_h given a hidden parameter Φ, and $\mathcal{P}(\cdot|s_h)$ denotes the posterior distribution over Φ given s_h. The reward obtained after taking action a_h and transitioning to s_{h+1} is $r_{h+1} = \mathbb{1}\{s_{h+1} \in \mathcal{S}_\checkmark\}$. The objective of the agent is to learn a policy π that maximizes the expected return, i.e.,

$$J^\pi(\mathcal{P}) := \mathbb{E}_{\substack{s_1 \sim \rho,\ a_h \sim \pi(\cdot|s_h), \\ s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h),\ h \in [H]}} \left[\sum_{h=1}^{H} r_{h+1}\right] = \mathbb{E}_{\Phi \sim \mathcal{P}, \pi}\left[\sum_{h=1}^{H} r_{h+1}\right], \tag{3}$$

where $\rho(s) := \mathbb{P}_{\Phi \sim \mathcal{P}}[s_1(\Phi) = s]$ is the initial state distribution. In addition, let $R(\pi; \Phi)$ denote the expected success rate of policy π in environment $\Phi = (\mathcal{T}, g(\mathcal{T}))$, and denote

$$R(\pi; \mathcal{P}) := \mathbb{E}_{\Phi \sim \mathcal{P}}\left[R(\pi; \Phi)\right] = J^\pi(\mathcal{P}) \tag{4}$$

as the expected success rate of policy π on trees sampled from the environment distribution $\mathcal{P}$. Therefore, maximizing $J^\pi(\mathcal{P})$ is equivalent to maximizing the success rate $R(\pi; \mathcal{P})$.

3 Transformer Architecture and Construction

We set our agent as a two-layer transformer-based model, which maps its trajectory observations to a probability distribution over the action space. To formally state our model, we first describe the state and action embeddings, followed by the attention model with simplified parameters for the tractability of analysis. This simplification lets us analyze the coupled attention and output dynamics while leaving history-dependent exploration and backtracking to be learned, see Appendix A.3.

[1] Since in our setting, each edge can be visited at most twice (down/up).

State and action embeddings. For each node $i \in [N]$, let $u_i \in \mathbb{R}^{d_1}$ denote its token embedding. For the action set, let $v_i \in \mathbb{R}^{d_2}$ denote the embedding of action $\texttt{down}_i$ for $i \in [k]$, and $v_{k+1} \in \mathbb{R}^{d_2}$ denote the embedding of action $\texttt{up}$. For the label set, let $z_0, z_1, z_2 \in \mathbb{R}^{d_3}$ denote the embeddings of labels $0, \times, \checkmark$ respectively. For notational convenience, let $\overline{n}_h \in \mathbb{R}^{d_1}$, $\overline{a}_h \in \mathbb{R}^{d_2}$ and $\overline{c}_h \in \mathbb{R}^{d_3}$ denote the embedding of node n_h, action a_h, and label c_h at step h, respectively. Given any tree $\mathcal{T}$, the state embedding $E^{(h)} = E^{(h)}(\mathcal{T})$ of state s_h is given (auto-regressively) as

$$E^{(h)} := \begin{pmatrix} E_n^{(h)} \\ E_a^{(h)} \\ E_{n'}^{(h)} \\ E_c^{(h)} \end{pmatrix} := \begin{pmatrix} u_0 & \overline{n}_1 & \cdots & \overline{n}_{h-1} \\ v_0 & \overline{a}_1 & \cdots & \overline{a}_{h-1} \\ \overline{n}_1 & \overline{n}_2 & \cdots & \overline{n}_h \\ \overline{c}_1 & \overline{c}_2 & \cdots & \overline{c}_h \end{pmatrix} = \begin{pmatrix} E^{(h-1)}, & \begin{pmatrix} \overline{n}_{h-1} \\ \overline{a}_{h-1} \\ \overline{n}_h \\ \overline{c}_h \end{pmatrix} \end{pmatrix} \in \mathbb{R}^{(2d_1+d_2+d_3)\times h} \tag{5}$$

with filler embeddings $u_0 \in \mathbb{R}^{d_1}$ and $v_0 \in \mathbb{R}^{d_2}$ for the first step. Note that $E^{(h)}$ contains all the information in s_h in a matrix form, and each column i of $E^{(h)}$ represents the embedding of the transition tuple $(n_{i-1}, a_{i-1}, n_i, c_i)$ at step i. Let $\overline{\mathcal{S}}$ denote the set of all possible state embeddings.

Transformer architecture and multi-turn reasoning. Our transformer-based agent maps the state to a policy over the action space, i.e. $\pi_\theta : \overline{\mathcal{S}} \mapsto \Delta(\mathcal{A})$. Following standard simplification in theoretical analyses (Huang et al., 2024; Yang et al., 2024; Nichani et al., 2024), for an input state embedding $E = [E_n^\top, E_a^\top, E_{n'}^\top, E_c^\top]^\top$, we define a two-head transformer model $f : \overline{\mathcal{S}} \to \mathbb{R}^{k+1}$ as

$$f(E) = P \cdot \texttt{head}_1(E) + Q \cdot \texttt{head}_2(E) \tag{6}$$

where $P \in \mathbb{R}^{(k+1)\times d_2}$, $Q \in \mathbb{R}^{(k+1)\times d_3}$, and

$$\texttt{head}_1(E) = E_a \cdot \mathsf{softmax}\big(E^\top W_1^{KQ} E\big), \quad \texttt{head}_2(E) = E_c \cdot \mathsf{softmax}\big(E^\top W_2^{KQ} E\big), \tag{7}$$

with the key and query matrices merged into a single matrix W_i^{KQ}. We further simplify the parameterization by setting

$$W_1^{KQ} = \begin{pmatrix} 0_{d_1\times d_1} & 0_{d_1\times d_2} & B & 0_{d_1\times d_3} \\ 0_{d_2\times d_1} & 0_{d_2\times d_2} & 0_{d_2\times d_1} & 0_{d_2\times d_3} \\ 0_{d_1\times d_1} & 0_{d_1\times d_2} & 0_{d_1\times d_1} & 0_{d_1\times d_3} \\ 0_{d_3\times d_1} & 0_{d_3\times d_2} & 0_{d_3\times d_1} & 0_{d_3\times d_3} \end{pmatrix}, W_2^{KQ} = \begin{pmatrix} 0_{d_1\times d_1} & 0_{d_1\times d_2} & 0_{d_1\times d_1} & 0_{d_1\times d_3} \\ 0_{d_2\times d_1} & 0_{d_2\times d_2} & 0_{d_2\times d_1} & 0_{d_2\times d_3} \\ 0_{d_1\times d_1} & 0_{d_1\times d_2} & 0_{d_1\times d_1} & 0_{d_1\times d_3} \\ 0_{d_3\times d_1} & 0_{d_3\times d_2} & 0_{d_3\times d_1} & C \end{pmatrix},$$

where $B \in \mathbb{R}^{d_1\times d_1}$ and $C \in \mathbb{R}^{d_3\times d_3}$, leading to

$$f_\theta(E) := f(E) = PE_a \cdot \mathsf{softmax}\big(E_n^\top B E_{n'}\big) + QE_c \cdot \mathsf{softmax}\big(E_c^\top C E_c\big), \tag{8}$$

where $\theta := \{P, Q, B, C\}$ collects the parameters of the transformer. The policy π_θ is then given by applying softmax to the last column of $f_\theta(E)$:

$$\begin{aligned} \pi_\theta(E) := \pi_\theta(\cdot \mid E) &:= \big(\pi_\theta(\texttt{down}_1 \mid E), \ldots, \pi_\theta(\texttt{down}_k \mid E), \pi_\theta(\texttt{up} \mid E)\big)^\top \\ &:= \mathsf{softmax}\left((f_\theta(E))_{:,-1}\right) \in \Delta(\mathcal{A}). \end{aligned} \tag{9}$$

During inference, at each step h, given the state embedding $E^{(h)}$, the agent samples an action a_h according to

$$\pi_\theta(E^{(h)}) = \mathsf{softmax}\left(PE_a^{(h)}\mathsf{softmax}\big(E_n^{(h)\top} B\overline{n}_h\big) + QE_c^{(h)}\mathsf{softmax}\big(E_c^{(h)\top} C\overline{c}_h\big)\right), \tag{10}$$

and observes $o_{h+1} = (n_{h+1}, c_{h+1})$. The next state embedding $E^{(h+1)}$ is then updated auto-regressively following (5) as long as the episode is not terminated, and the agent moves to the next timestep $h+1$.

Transformer construction. We collect the embeddings in matrix forms as

$$U := (u_0, \cdots, u_N), \; V := (v_0, \cdots, v_{k+1}), \; Z := (z_0, z_1, z_2), \tag{11}$$

and define induced matrices A_M $(M \in \{B, C, P, Q\})$ as

$$\begin{aligned} A_B := U^\top B U \in \mathbb{R}^{(N+1)\times(N+1)}, \quad A_C := Z^\top C Z \in \mathbb{R}^{3\times 3}, \\ A_P := PV \in \mathbb{R}^{(k+1)\times(k+2)}, \quad A_Q := QZ \in \mathbb{R}^{(k+1)\times 3}. \end{aligned} \tag{12}$$

We first present a set of matrices $\{B, C, P, Q\}$ in Theorem 1 that enable the transformer to perform the search task, thereby clarifying its underlying mechanism. This construction is not unique; our particular choice is intended to align more closely with the training dynamics while having a simple form. See Appendix D.1 for the proof.

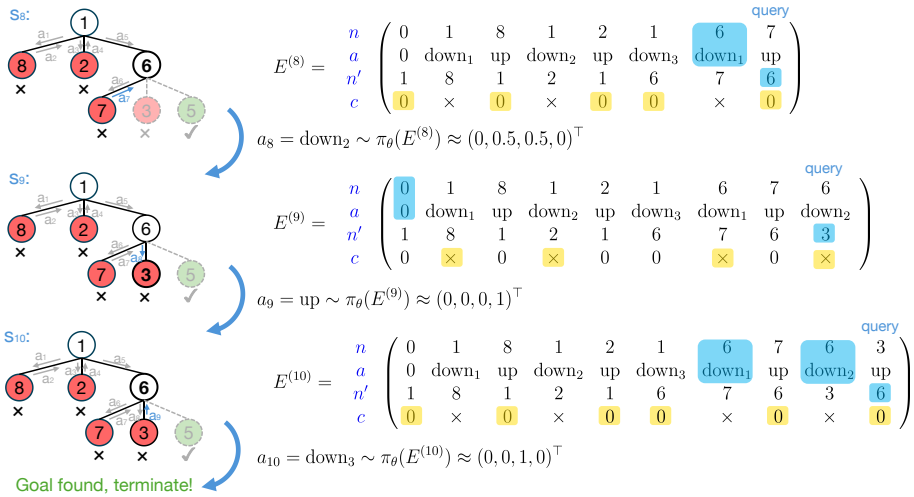


Figure 2: Transformer mechanism implementing random DFS on a full 3-ary tree. Blue and yellow shading indicate the attention associations of heads 1 and 2, respectively.

Theorem 1 (Transformer construction for random DFS)**.** *Assume U, V, Z all have full column rank. Then there exists $\theta = \{B, C, P, Q\}$ such that*

$$A_B = \begin{pmatrix} 0 & a_{b,0} 1_{1\times N} \\ 0_{N\times 1} & a_{b,1} I_N \end{pmatrix}, \quad A_C = \begin{pmatrix} a_{c,0} & -a_{c,1} & 0 \\ -a_{c,0} & a_{c,1} & 0 \\ 0 & 0 & 0 \end{pmatrix},$$
$$A_P = \begin{pmatrix} 0_{k\times 1} & -a_{p,1} I_k & 0_{k\times 1} \\ 0 & 0_{1\times k} & 0 \end{pmatrix}, \quad A_Q = \begin{pmatrix} \frac{a_{q,0}}{k} 1_{k\times 1} & -\frac{a_{q,\times}}{k} 1_{k\times 1} & 0_{k\times 1} \\ -a_{q,0} & a_{q,\times} & 0 \end{pmatrix} \tag{13}$$

for any $a_{b,0}, a_{b,1}, a_{c,0}, a_{c,1}, a_{p,1}, a_{q,0}, a_{q,\times} \in \mathbb{R}$. For any $k \geqslant 2$, $L \in \mathbb{N}_+$, $N \geqslant \frac{k^{L+1}-1}{k-1}$ and precision $\epsilon > 0$, there exists an absolute constant $C > 0$ such that when

$$e^{a_{c,0}} \geqslant \ln k, \quad e^{a_{c,1}} \geqslant \sqrt{N} \ln k, \quad e^{a_{b,0}} \geqslant N \ln k, \quad e^{a_{b,1}} \geqslant \ln k e^{a_{b,0}},$$
$$\frac{a_{p,1}}{k} \asymp a_{q,0} \asymp a_{q,\times}, \quad a_{q,0},\, a_{q,\times} \geqslant C \ln(N/\epsilon), \quad \frac{a_{p,1}}{k} \geqslant C(\ln(N/\epsilon) + a_{q,0}), \tag{14}$$

we have

$$R(\pi_\theta; \Phi) \geqslant 1 - \epsilon \tag{15}$$

for any $\Phi = (\mathcal{T}, g(\mathcal{T}))$ where $\mathcal{T}$ is a full k-ary tree with depth $l \in [L]$, and $g(\mathcal{T})$ is an arbitrary leaf of $\mathcal{T}$.

Under our construction, the agent performs *random DFS*: it selects randomly among unvisited children, and goes up only at a wrong leaf or after all children of the current node have been explored. Figure 2 illustrates how the two heads cooperate. Head 2, associated with C, Q, uses label history to promote `up` at non-goal leaves and suppress `up` at internal nodes. Head 1, associated with B, P, retrieves action history and suppresses tried children. When all children have been tried, every $\texttt{down}_i$ logit is suppressed, making `up` largest despite Head 2's internal-node bias. Thus the mechanism implements both action avoidance and backtracking.

4 Training Dynamics and Generalization of Transformers

Training with policy gradient. The construction in Theorem 1 suggests that the last row of P can be fixed to 0, so for simplicity and consistency with our construction, we let

$$P = \begin{pmatrix} \overline{P} \\ 0_{1\times d_2} \end{pmatrix} \tag{16}$$

where $\overline{P} \in \mathbb{R}^{k\times d_2}$, and set trainable parameters to be $\theta = \{B, C, \overline{P}, Q\}$. We train the policy of the form (10) with policy gradient. Specifically, we let $\theta^{(t)} = \{B^{(t)}, C^{(t)}, \overline{P}^{(t)}, Q^{(t)}\}$ denote the parameters at iteration $t \in \mathbb{N}$, and we initialize all matrices to be zero matrices, i.e.,

$$B^{(0)} = 0_{d_1\times d_1}, \quad C^{(0)} = 0_{d_3\times d_3}, \quad \overline{P}^{(0)} = 0_{k\times d_2}, \quad Q^{(0)} = 0_{(k+1)\times d_3}. \tag{17}$$

We write $J(\theta)$ as the shorthand of $J^{\pi_\theta}(\mathcal{P})$ when the training distribution $\mathcal{P}$ is clear from the context. Then under our setting, the policy gradient update at iteration t is given by

$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta J(\theta^{(t)})$$

$$= \theta^{(t)} + \eta \mathbb{E}_{\Phi \sim \mathcal{P}, \pi_\theta} \left[\sum_{h=1}^{H} \nabla_\theta \log \pi_\theta(a_h \mid E^{(h)}) \sum_{i=h}^{H} r_{i+1} \right], \tag{18}$$

where the second equality follows from policy gradient theorem (Williams, 1992; Sutton et al., 1999), and $\eta > 0$ is the learning rate.

Training distribution. We aim to demonstrate that after training on trees of depth 1 and 2 well enough, the agent learns the DFS algorithm and can perform well on deeper trees. To describe the training distribution $\mathcal{P}_l$ for any fixed $l \in [L]$, we define the training tree set $\mathcal{D}^l_{\text{train,tree}}$ as

$$\mathcal{D}^l_{\text{train,tree}} := \{\mathcal{T} \mid \mathcal{T} \text{ is a perfect } k\text{-ary tree with distinct nodes in } [N] \text{ and depth } l\}, \tag{19}$$

and define $p^l_{\text{train,tree}} = \mathcal{U}(\mathcal{D}^l_{\text{train,tree}})$ as the uniform distribution on $\mathcal{D}^l_{\text{train,tree}}$. We assume, for any given tree $\mathcal{T}$ with any depth $l \in [L]$, the environment generates the goal node $g(\mathcal{T})$ according to a latent *goal distribution* $p^a = (p^a_1, \cdots, p^a_k) \in \Delta(\mathcal{A}_{\texttt{down}})$ that is unknown to the agent. Specifically, after a tree $\mathcal{T}$ is generated from $p^l_{\text{train,tree}}$, the environment, starting from the root, at each node takes action $a^{\text{env}} \overset{i.i.d}{\sim} p^a$ until the leaf is reached. Then this leaf is assigned as a goal. We let $p^a(\cdot \mid \mathcal{T})$ denote the goal distribution induced by p^a given tree $\mathcal{T}$. We define the training distribution $\mathcal{P}_l$ of trees with depth $l \in [L]$ as

$$\mathcal{P}_l(\mathcal{T}, g) = p^l_{\text{train,tree}}(\mathcal{T}) p^a(g \mid \mathcal{T}). \tag{20}$$

Here we focus on the case where the goal distribution p^a is balanced, i.e., we assume

$$p^a_1 = p^a_2 = \cdots = p^a_k = \frac{1}{k}. \tag{21}$$

In Section 5, we also consider the case where the goal distribution p^a is imbalanced.

Depth-wise curriculum training. We use a two-stage training with curriculum schedulers $\delta_1, \delta_2 \in (0, 1)$: (i) depth-1 curriculum training (training step $t \in [T_1]$): we first train on depth-1 trees $\mathcal{T} \sim \mathcal{P}_1$; (ii) depth-2 curriculum training (training step $T_1 + 1 \leqslant t \leqslant T_2$): as soon as the success rate $R(\pi; \mathcal{P}_1)$ reaches $1 - \delta_1$, we start training on depth-2 trees $\mathcal{T} \sim \mathcal{P}_2$ until $R(\pi; \mathcal{P}_2)$ reaches $1 - \delta_2$. We show that for any $\epsilon > 0$, after training with proper choices of δ_1, δ_2, the success rate on an arbitrary full k-ary tree with depth $l \in [L]$ is at least $1 - \epsilon$.

Orthonormal token embeddings. To simplify the analysis, we make the following orthonormality assumption on the embedding matrices defined in (11):

Assumption 1 (orthonormal embeddings)**.** *U, V, Z each have orthonormal columns.*

Theorem 2 (Depth-1 curriculum training, informal)**.** *Suppose (21), Assumption 1 hold, $k, N \in \mathbb{N}_+$ are sufficiently large, and $\epsilon, \eta > 0$ are sufficiently small, $\delta_1 = \epsilon^{1.05}$. For depth-1 curriculum training, the success rate $R(\pi; \mathcal{P}_1)$ reaches $1 - \delta_1$ in $T_1 = O(1/(\eta \epsilon^{1.1}))$ iterations. After the training, we have*

$$R(\pi^{(T_1)}; \mathcal{P}_l) = O\left(1/k^{l-1}\right), \quad \forall 2 \leqslant l \leqslant L. \tag{22}$$

The formal version (for both Theorem 2 and Theorem 3) and the proof are in Appendix C. In the proof, we track the induced matrices A_B, A_C, A_P, A_Q in (12), showing that depth-1 training first learns wrong-leaf backtracking and then root-level repeat avoidance. By time T_1, A_B, A_C, A_Q are close to their DFS-construction counterparts in Theorem 1, but generalization still fails because $a^{(T_1)}_{p,2}/a^{(T_1)}_{p,1}$ remains too large: after exhausting the children of a non-root internal node, the agent does not move up with high probability. We therefore freeze B, C, Q to retain the learned attention and outcome maps while isolating how depth-2 updates to $\overline{P}$ correct this remaining imbalance. More discussions on our training settings are in Appendix A.3.

Theorem 3 (Depth-2 curriculum training and depth generalization, informal)**.** *Suppose the same assumptions in Theorem 2 hold, and fix $\delta_2 = \epsilon^{1.02}/k^L$. For depth-2 curriculum training, the success rate $R(\pi; \mathcal{P}_2)$ reaches $1 - \delta_2$ in $T_2 - T_1 = O\left(k^L/(\eta \epsilon^{1.1})\right)$ iterations. After the training, given any full k-ary tree $\mathcal{T}$ with depth $l \in [L]$ and any goal node $g(\mathcal{T})$, we have*

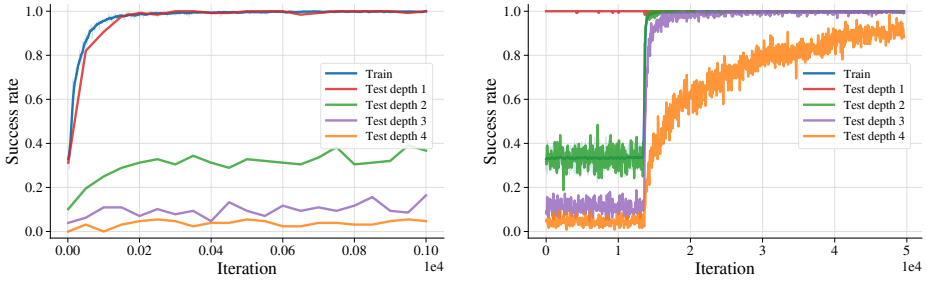
$$R(\pi^{(T_2)}; (\mathcal{T}, g(\mathcal{T}))) \geqslant 1 - \epsilon. \tag{23}$$

In the proof, we show that depth-2 training increases $a_{p,1}$ while decreasing $a_{p,2}$, keeping the remaining entries of $\overline{P}$ small, until the policy implements random DFS and generalizes as in (23). Thus, after depth-2 training, all induced matrices are close to the DFS construction in Theorem 1. Although

training uses perfect k-ary trees to exploit symmetry, the learned policy generalizes to any full k-ary tree with depth at most L, as guaranteed by Theorem 3.

Compared with analyses on prompt-revealed or fixed graphs (Saparov et al., 2025; Wang et al., 2024; 2026), or RL/RLVR over generated traces (Bu et al., 2025; Huang et al., 2026), our hidden-environment setting requires tracking a multi-phase interaction among induced matrices and return-to-go estimates; see Appendix C.1 for technical details and Appendix A.1 for the significance.

Numerical experiments. We further validate this training-dynamics analysis with numerical experiments on 3-ary trees using stochastic policy gradient. Figure 3 shows representative success-rate curves under the balanced goal distribution. We refer to Appendix E for the experiment details, heatmaps at convergence, and evolution curves of the induced matrices.



(a) Depth-1 curriculum training (b) Depth-2 curriculum training

Figure 3: Train and test rates under balanced goal distribution for depth-1 curriculum training (left) and depth-2 curriculum training (right), trained on perfect 3-ary trees with stochastic policy gradient, tested on randomly generated full 3-ary trees with depth 1 to 4.

5 Extension to Imbalanced Goal Distribution

So far we have focused on balanced goal distributions (c.f. (21)). In practice, goals can be imbalanced: for instance, in web search, desired information is usually more likely to appear among top-ranked results. This suggests searching according to the non-uniform prior, prioritizing more promising branches over the random order that is natural in the balanced case (Koh et al., 2025).

We formalize this via an averaged full-tree environment distribution $\mathcal{P}$. For any $l \in [L]$, let

$$\mathcal{D}^l_{\text{env,tree}} := \{\mathcal{T} \mid \mathcal{T} \text{ is a full } k\text{-ary tree with distinct nodes in } [N] \text{ and depth } l\}, \tag{24}$$

and let $p^l_{\text{env,tree}} = \mathcal{U}(\mathcal{D}^l_{\text{env,tree}})$. As in the balanced setting, given a sampled tree $\mathcal{T}$, the environment generates $g(\mathcal{T})$ by sampling latent actions independently from $p^a = (p^a_1, \ldots, p^a_k) \in \Delta(\mathcal{A}_{\texttt{down}})$ from the root until a leaf is reached. Denoting the induced distribution by $p^a(\cdot \mid \mathcal{T})$, define

$$\mathcal{P}(\mathcal{T}, g) = p_{\text{env,tree}}(\mathcal{T})\, p^a(g \mid \mathcal{T}), \quad \text{where} \quad p_{\text{env,tree}}(\mathcal{T}) := \sum_{l=1}^{L} \omega_l\, p^l_{\text{env,tree}}(\mathcal{T}) \tag{25}$$

where the weights $\omega_l > 0$ and $\sum_{l=1}^{L} \omega_l = 1$. Under this environment distribution, we consider both the balanced case in (21) and the imbalanced case. In the latter case, without loss of generality, we index the actions by decreasing likelihood:

$$p^a_1 > p^a_2 > \cdots > p^a_k > 0. \tag{26}$$

Here we consider maximizing the *discounted* expected return, i.e.,

$$J^\pi_\gamma(\mathcal{P}) := \mathbb{E}_{\substack{s_1 \sim \rho,\ a_h \sim \pi(\cdot|s_h),\\ s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h),\ h \in [H]}} \left[\sum_{h=1}^{H} \gamma^{h-1} r_{h+1} \right] = \mathbb{E}_{\Phi \sim \mathcal{P}, \pi} \left[\sum_{h=1}^{H} \gamma^{h-1} r_{h+1} \right], \tag{27}$$

where $\gamma \in (0, 1]$ is the discount factor.

Discounting selects among successful policies. We define

$$\Pi_{\text{success}} := \Pi_{\text{success}}(\mathcal{P}) := \{\pi : R(\pi; \mathcal{P}) = 1\}. \tag{28}$$

Thus Π_{success} is the set of policies that find the goal almost surely under $\mathcal{P}$. It is identical in the balanced and imbalanced cases, since the supports over $\mathcal{T}, g$ coincide, and $\Pi_{\text{success}} = \arg\max_\pi J_1(\pi; \mathcal{P})$. Because Π_{success} contains infinitely many policies, the undiscounted objective cannot distinguish among them; we compare them by expected path length $H^\pi(\mathcal{P})$.

In Proposition 1, we show that in the balanced case, every policy in Π_{success} has the same expected path length, so any successful DFS policy, including uniform-random DFS, is optimal.

Proposition 1 (optimal policy and objective under balanced goal distribution)**.** *Under balanced goal distribution (21), any policy $\pi \in \Pi_{success}$ has the same expected path length $H^\pi(\mathcal{P})$.*

The proof is postponed to Appendix B.1.

In contrast, the imbalanced case has a unique *ranked DFS* policy that minimizes expected path length among successful policies (up to behavior on unreachable histories). Moreover, for a suitable $\gamma < 1$, maximizing $J_\gamma^\pi(\mathcal{P})$ exactly recovers this policy, see Proposition 2 for a formal statement.

Proposition 2 (optimal policy and objective under imbalanced goal distribution)**.** *Define the* ranked DFS *policy $\pi^\star$ as: at any reachable, non-terminal state, if the current node has unvisited children, it explores them in decreasing order of p_j^a; otherwise, it takes* up*. Then under the imbalanced goal distribution (26), we have $\pi^\star \in \Pi_{\text{success}}$, and*

1. *$H^{\pi^\star}(\mathcal{P}) < H^\pi(\mathcal{P})$ for any $\pi \in \Pi_{success}$ whose reachable behavior differs from $\pi^\star$'s.*
2. *There exists an absolute constant $C > 0$ such that for any $\gamma \in \left(1 - C\,(p_k^a)^L, 1\right)$, $\pi^\star = \arg\max_\pi J_\gamma^\pi(\mathcal{P})$ up to reachable behavior.*

The proof is postponed to Appendix B.2. We also call $\pi^\star$ the *optimal policy* for the imbalanced case.

Transformer construction. Theorem 4 modifies the induced matrices in (13) by replacing the uniform row weights in A_Q with a priority vector $\lambda \in \mathbb{R}^k$, yielding ranked DFS $\pi^\star$. The proof is provided in Appendix D.2.

Theorem 4 (Transformer construction for ranked DFS)**.** *Assume U, V, Z all have full column rank. Then there exists $\theta = \{B, C, P, Q\}$ such that*

$$A_B = \begin{pmatrix} 0 & a_{b,0}1_{1\times N} \\ 0_{N\times 1} & a_{b,1}I_N \end{pmatrix}, \quad A_C = \begin{pmatrix} a_{c,0} & -a_{c,1} & 0 \\ -a_{c,0} & a_{c,1} & 0 \\ 0 & 0 & 0 \end{pmatrix},$$
$$A_P = \begin{pmatrix} 0_{k\times 1} & -a_{p,1}I_k & 0_{k\times 1} \\ 0 & 0_{1\times k} & 0 \end{pmatrix}, \quad A_Q = \begin{pmatrix} \lambda & 0_{k\times 1} & 0_{k\times 1} \\ -a_{q,0} & a_{q,\times} & 0 \end{pmatrix} \tag{29}$$

for any $a_{b,0}, a_{b,1}, a_{c,0}, a_{c,1}, a_{q,0}, a_{q,\times}, a_{p,1} \in \mathbb{R}$, any $\lambda = (\lambda_1, \ldots, \lambda_k)^\top \in \mathbb{R}^k$. For any $k \geqslant 2$, $L \in \mathbb{N}_+$, $N \geqslant \frac{k^{L+1}-1}{k-1}$ and precision $\epsilon > 0$, there exists an absolute constant $C > 0$ such that when

$$\begin{aligned} &e^{a_{c,0}} \geqslant \ln k, \quad e^{a_{c,1}} \geqslant \sqrt{Nk}\ln k, \quad e^{a_{b,0}} \geqslant Nk\ln k, \quad e^{a_{b,1}} \geqslant \ln k e^{a_{b,0}}, \\ &\lambda_i - \lambda_{i+1} \asymp a_{q,0} \asymp a_{q,\times} \gtrsim \lambda_k \geqslant 0, \quad \lambda_i - \lambda_{i+1},\, a_{q,0},\, a_{q,\times} \geqslant C\ln(N/\epsilon), \ \forall i \in [k-1], \\ &\frac{a_{p,1}}{k} \asymp \lambda_1, \quad \frac{a_{p,1}}{k} \geqslant C\left(\ln(N/\epsilon) + a_{q,0} + \lambda_1\right), \end{aligned} \tag{30}$$

for any $\Phi = (\mathcal{T}, g(\mathcal{T}))$ where $\mathcal{T}$ is a full k-ary tree with depth $l \in [L]$, and $g(\mathcal{T})$ is an arbitrary leaf, and any reachable non-terminal state s_h, we have

$$R(\pi_\theta; \Phi) \geqslant 1 - \epsilon, \quad \|\pi_\theta(\cdot \mid s_h) - \pi^\star(\cdot \mid s_h)\|_\infty \leqslant \frac{\epsilon}{2N}. \tag{31}$$

Finally, our experiments show that discounted-return training recovers $\pi^\star$ in the imbalanced setting and learns induced matrices similar to the construction above; see Appendix E.

6 Conclusion

We have shown that an interactive search algorithm can emerge from RL training in a transformer agent, even when supervision comes solely from sparse terminal reward. In our analysis, policy gradient updates shape two attention heads into complementary memory-retrieval mechanisms: one suppresses previously tried actions, while the other detects failed outcomes and triggers backtracking. Their composition yields a DFS policy that generalizes from shallow curriculum training to deep trees, and under imbalanced goal distributions, discounting further biases the learned policy toward ranked search. Our work is a first step toward understanding complex agentic behavior in stochastic RL environments, and provides a tractable testbed for future directions including principled memory management, context pruning, and scalable RL algorithms for long-horizon interactive reasoning.

AI USE STATEMENT

We did not use generative AI in the original research or manuscript writing. Its use was limited to minor proofreading, notation checks, and latex formatting. The authors are responsible for the final content.

REFERENCES

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. PC-PG: Policy cover directed exploration for provable policy gradient learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/9b3a9fb4db30fc6594ec3990cbc09932-Abstract.html.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1560–1588. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0561738a239a995c8cd2ef0e50cfa4fd-Abstract.html.

Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.

Dake Bu, Wei Huang, Andi Han, Atsushi Nitanda, Hau-San Wong, Qingfu Zhang, and Taiji Suzuki. Provable benefit of curriculum in transformer tree-reasoning post-training. *arXiv preprint arXiv:2511.07372*, 2025.

Yen-Ju Chen, Nai-Chieh Huang, Ching-Pei Lee, and Ping-Chun Hsieh. Accelerated policy gradient: On the convergence rates of the Nesterov momentum for reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6881–6949. PMLR, 2024. URL https://proceedings.mlr.press/v235/chen24t.html.

Michael O'Gordon Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst, 2002.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173, 2023.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In *International Conference on Machine Learning*, volume 235, pp. 19660–19722, 2024.

Yu Huang, Zixin Wen, Aarti Singh, Yuejie Chi, and Yuxin Chen. Transformers provably learn chain-of-thought reasoning with length generalization. *Advances in Neural Information Processing Systems*, 2025.

Yu Huang, Zixin Wen, Yuejie Chi, Yuting Wei, Aarti Singh, Yingbin Liang, and Yuxin Chen. On the learning dynamics of rlvr at the edge of competence. *arXiv preprint arXiv:2602.14872*, 2026.

Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. Tree search for language model agents. *Transactions on Machine Learning Research*, 2025.

Guillaume Lample, Marie-Anne Lachaux, Thibaut Lavril, Xavier Martinet, Amaury Hayat, Gabriel Ebner, Aurélien Rodriguez, and Timothée Lacroix. HyperTree proof search for neural theorem proving. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26337–26349, 2022.

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19689–19729. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23p.html.

Xiang Liu, Zhenheng Tang, Hong Chen, Peijie Dong, Zeyu Li, Xiuze Zhou, Bo Li, Xuming Hu, and Xiaowen Chu. Semantic integrity matters: Benchmarking and preserving high-density reasoning in KV cache compression. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2502.01941.

Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. In *International Conference on Machine Learning*, volume 235, pp. 38018–38070, 2024.

Yuval Ran-Milo, Yotam Alexander, Shahar Mendel, and Nadav Cohen. Outcome-based rl provably leads transformers to reason, but only with the right data. *arXiv preprint arXiv:2601.15158*, 2026.

Abulhair Saparov, Srushti Pawar, Shreyas Pimpalgaonkar, Nitish Joshi, Richard Yuanzhe Pang, Vishakh Padmakumar, Seyed Mehran Kazemi, Najoung Kim, and He He. Transformers struggle to learn to search. In *International Conference on Learning Representations*, 2025.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/98711dea460bdefe0e651ca23ec98ba2-Abstract-Conference.html.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419): 1140–1144, 2018.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Siwei Wang, Yifei Shen, Shi Feng, Haoran Sun, Shang-Hua Teng, and Wei Chen. ALPINE: Unveiling the planning capability of autoregressive learning in language models. In *Advances in Neural Information Processing Systems*, 2024.

Siwei Wang, Yifei Shen, Haoran Sun, Shi Feng, Shang-Hua Teng, Li Dong, Yaru Hao, and Wei Chen. Benefits and pitfalls of reinforcement learning for language model planning: a theoretical perspective. In *International Conference on Learning Representations*, 2026.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts LLM agents: An empirical study of experience-following behavior. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 623–645. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.27. URL https://aclanthology.org/2026.acl-long.27/.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. In-context learning with representations: Contextual generalization of trained transformers. *Advances in Neural Information Processing Systems*, 37:85867–85898, 2024.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. Multi-head transformers provably learn symbolic multi-step reasoning via gradient descent. *Advances in Neural Information Processing Systems*, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023b.

Qilin Ye, Deqing Fu, Robin Jia, and Vatsal Sharan. Transformers provably learn algorithmic solutions for graph connectivity, but only with the right data, 2026.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *International Conference on Machine Learning*, volume 235, pp. 62138–62160, 2024.

A Discussion of the Search Mechanism and Scope

In this section, we discuss the contribution and future directions, and explain how our abstraction and assumptions preserve the challenge of learning history-dependent search control while making policy-gradient analysis tractable.

A.1 Contribution and future directions

We emphasize that our contribution is not merely that a transformer can represent DFS, but, to our knowledge, the first optimization-level account of how sparse-reward policy gradient acquires the history-dependent control primitives of interactive search: remembering tried actions, recognizing failure, backtracking from exhausted subtrees, and composing these behaviors into a depth-generalizing policy. Unlike nearby theory where a graph is revealed in the prompt or fixed during generation (Saparov et al., 2025; Wang et al., 2026), our agent must discover a hidden instance through interaction, without demonstrations or action-level labels. The setting is deliberately tractable, but it isolates an understudied agent–environment learning problem rather than static path prediction.

The mechanism suggests three concrete research directions for LLM agents:

- **Memory management.** Similarity-based memory retrieval can cause agents to repeat mistakes from past executions (Xiong et al., 2026). Our analysis shows how retrieving tried actions at the current state and using failure outcomes enable repeat avoidance and backtracking. This motivates training policies for writing and retrieving memory using task feedback, with the goal of preventing repeated failed tool calls under a fixed memory budget.
- **Context compression.** KV-cache pruning for LLMs that performs well on NLP tasks can substantially degrade multi-step reasoning (Liu et al., 2026). Our analysis relates attention to the historical information used at each search step, motivating pruners that learn to predict future token relevance from LLM attention traces. This may allow such pruners to preserve reasoning with less context by removing exhausted-branch histories while retaining the memory needed for later decisions.
- **Optimization for agentic tasks.** Our framework provides a tractable setting for analyzing how reward or different optimization methods affect the learning of history-dependent search. Concrete questions include how to design intermediate rewards to accelerate the learning (Setlur et al., 2025), and whether momentum or other optimizer design improves convergence rates (Chen et al., 2024).
- **Sample-based learning and exploration.** Extending the analysis to stochastic policy gradient would connect our iteration bounds to the number of interactions needed to learn search. The directional margins in our phase analysis suggest a route: establish robustness to bounded gradient perturbations, then combine it with concentration for mini-batch Monte Carlo estimates to obtain high-probability guarantees with noise-dependent slack. This would also support studying how online exploration strategies (Agarwal et al., 2020) can reduce sample complexity by improving data coverage.

A.2 The search abstraction

We consider a stylized environment that abstracts away natural-language understanding and representation learning, while allowing us to focus on a different and largely unresolved difficulty: how policy-gradient training makes a transformer acquire an interactive search algorithm from sparse terminal reward.

The agent is not given the tree, a DFS trace, or action-level supervision. A new *hidden* tree and goal are sampled each episode, and the agent receives only local observations. It must use its full trajectory history to remember failed branches, avoid repeated actions, detect when an internal node is exhausted, and backtrack appropriately. Thus, our assumptions such as clean labels remove semantic ambiguity, but not the central search-control challenges of partial observability, long-horizon memory, exploration, backtracking, and credit assignment. The regular full k-ary structure also prevents structural shortcuts: unexplored branches are locally symmetric, so success requires learning a reusable search rule rather than memorizing node identities or trajectories. Variable degree of trees can be handled by padding or masking up to a maximum degree K.

More importantly, our setting provides a qualitatively more faithful abstraction of the agent–environment loop underlying real agentic search than the closest theoretical models. Prior analyses largely study supervised next-node prediction on a graph revealed in the prompt (Saparov et al., 2025), autoregressive path generation on a fixed graph (Wang et al., 2024), or RL planning where the state is a generated path prefix and the graph remains fixed (Wang et al., 2026). These formulations do not capture the combined challenge of discovering a new hidden instance through interaction, using local feedback to update the search state, remembering failed attempts, avoiding repeated actions, recognizing subtree exhaustion, and backtracking. Our model isolates precisely these control challenges while abstracting away semantic representation learning.

A.3 Role of the analytical assumptions

We make some analytical assumptions to simplify the training dynamics analysis such as structured parameterization, orthogonal embeddings, symmetric initialization, population gradients, and staged analysis. These are common proof devices in optimization-level theory of transformer learning, used to isolate mechanisms in otherwise intractable nonconvex dynamics (Li et al., 2023; Bietti et al., 2023; Huang et al., 2024). Crucially, these assumptions do not encode DFS: training must still discover from terminal reward alone how to avoid repeated actions, backtrack after failure, recognize exhausted subtrees, and compose these behaviors into DFS that generalizes beyond the training depths.

The balanced goal distribution removes any branch-priority signal and makes uniform randomized DFS the natural successful policy. Our imbalanced extension (Section 5) further shows that when this symmetry is removed, policy-gradient training learns ranked DFS instead. Likewise, zero initialization and the block parameterization select a symmetric, analyzable trajectory among many equivalent parameterizations. Orthogonality can largly simplify the computation. With learned or non-orthogonal embeddings, the same mechanism could be realized after a learned input projection or MLP, in the spirit of representation-learning analyses such as (Yang et al., 2024).

The curriculum training separates two optimization regimes: depth-1 training learns failed-leaf backtracking and repeated-action avoidance, but cannot yet backtrack from an exhausted internal node. Failure after depth 1 is also evidence that DFS is not already hard-coded. Depth-2 training must compose these primitives into the remaining DFS rule. Freezing B, C, Q there is not essential to learning, but a proof device that simplifies the depth-2 dynamics analysis. After depth-1 training, these parameters have already learned the relevant attention and backtracking primitives; freezing them isolates the remaining P-dynamics and avoids additional coupled gradient terms. Partial parameter freezing is also a standard analytical simplification in theoretical studies of transformer learning dynamics (Li et al., 2023; Bietti et al., 2023). Our experiments support this view: balanced depth-2 training also updates B (Figure 3), and imbalanced training updates all four matrices throughout (Figure 6; setup in Appendix E).

While most of the assumptions are not algorithmic necessities, removing all of them simultaneously would substantially obscure the mechanism and complicate the analysis, without necessarily yielding corresponding conceptual insight.

B Proof of Proposition 1 and Proposition 2

B.1 Proof of Proposition 1

For a fixed tree $\mathcal{T}$ and policy $\pi \in \Pi_{\text{success}}$, let $H^\pi(\mathcal{T})$ denote the expected number of steps to hit the goal when the goal is generated by the balanced latent process $p_1^a = \cdots = p_k^a = \frac{1}{k}$. Define

$$\overline{H}^\pi(l) := \mathbb{E}_{\mathcal{T}\sim p^l_{\text{env,tree}}}\left[H^\pi(\mathcal{T})\right]. \tag{32}$$

We will show that $\overline{H}^\pi(l)$ is the same for all $\pi \in \Pi_{\text{success}}$ and $l \in [L]$.

Let $\mathcal{Q}_{d,r}$ denote the collection of all posterior laws of the current unresolved subtree that can arise as follows: first sample an original tree from $p^l_{\text{env,tree}}$, then condition on a reachable state and on the event that the goal lies in the current unresolved subtree rooted at the node the agent is at, with remaining depth at most d, and with exactly r unresolved children at the current node. For $q \in \mathcal{Q}_{d,r}$ and a continuation policy $\pi \in \Pi_{\text{success}}$, let H_q^π denote the expected remaining number of steps to hit the goal.

We claim that for every $d \geqslant 1$, every $r \in [k]$, and every $q \in \mathcal{Q}_{d,r}$, the quantity H_q^π is independent of π for all $\pi \in \Pi_{\text{success}}$. We prove this by lexicographical induction on (d, r): we regulate $(d', r') < (d, r)$ if either $d' < d$ or $d' = d$ and $r' < r$.

Base case: $d = 1$**.** If $q \in \mathcal{Q}_{1,r}$, then the agent is at a node whose children are all leaves. Let $\mathcal{U}_q \subseteq [k]$ be the set of currently unresolved children and $r = |\mathcal{U}_q|$. Since the latent goal process is balanced and the prior over child positions is permutation-invariant, conditional on the current state, the correct child is uniformly distributed over $\mathcal{U}_q$. And any policy in Π_{success} must explore the children in $\mathcal{U}_q$ in some order before terminating. Therefore, the expected remaining number of steps is

$$H_q^\pi = \frac{1}{r}\sum_{i=1}^{r}\big(2(i-1)+1\big) = r,$$

which is independent of $\pi \in \Pi_{\text{success}}$.

Induction step. Fix $d \geqslant 2$, $r \in [k]$, and assume the claim holds for every $(d', r') < (d, r)$. We show it holds for (d, r). Fix any $q \in \mathcal{Q}_{d,r}$, and let $\mathcal{U}_q \subseteq [k]$ be the set of currently unresolved children and $r = |\mathcal{U}_q|$. If $r = 1$, there is only one unresolved child, so every policy in Π_{success} must choose it next. If this child is a leaf, then $H_q^\pi = 1$. Otherwise, after moving into that child, we obtain a new posterior $q' \in \mathcal{Q}_{d-1,k}$. Hence

$$H_q^\pi = 1 + H_{q'}^\pi,$$

and by induction hypothesis, the right-hand side is independent of $\pi \in \Pi_{\text{success}}$. In the following we assume $2 \leqslant r \leqslant k$.

For $\mathcal{T} \sim q$, for each $j \in \mathcal{U}_q$, write $\mathcal{T}_j$ for the child subtree rooted at the j-th child of the current node, and let $h(\mathcal{T}_j)$ be the number of steps needed to fully explore $\mathcal{T}_j$ and return to its root. If child j is chosen next and is incorrect (the child subtree doesn't contain the goal), let $q_j^\times$ denote the posterior after fully exploring $\mathcal{T}_j$ and returning to the current node, and ruling out child j. If child j is chosen next and is correct and nonterminal, let $q_j^\checkmark$ denote the posterior after moving down into that child. For any $\pi \in \Pi_{\text{success}}$, since π must explore all children in $\mathcal{U}_q$ before hitting the goal, we can let $a^\pi(q) \in \mathcal{U}_q$ denote the first child that π chooses from the current state (without loss of generality, we assume π is deterministic; if π is randomized, it can be handled by conditioning on the realized internal randomness and average afterward). There are two cases: (i) with probability $1/r$, the chosen child is the correct one; (ii) with probability $1 - 1/r$, it's not.

If the chosen child is correct, then after paying 1 step to move down into that child, there are two subcases. If $\mathcal{T}_{a^\pi(q)}$ is a leaf, then the goal is reached immediately and no continuation cost is incurred. If $\mathcal{T}_{a^\pi(q)}$ is not a leaf, then we obtain a posterior $q^\checkmark_{a^\pi(q)} \in \mathcal{Q}_{d-1,k}$. By induction hypothesis on depth, the continuation value $H_{q^\checkmark_{a^\pi(q)}} := H^\pi_{q^\checkmark_{a^\pi(q)}}$ is independent of $\pi \in \Pi_{\text{success}}$. If the chosen child is incorrect, then we pay $2 + h(\mathcal{T}_{a^\pi(q)})$ steps to fully explore that child and return, after which we obtain a posterior $q^\times_{a^\pi(q)} \in \mathcal{Q}_{d,r-1}$. By induction hypothesis, the continuation value $H_{q^\times_{a^\pi(q)}} := H^\pi_{q^\times_{a^\pi(q)}}$ is

also independent of $\pi \in \Pi_{\text{success}}$. Therefore, the expected remaining number of steps is

$$H_q^\pi = \mathbb{E}_{\mathcal{T}\sim q}\Bigg[\frac{1}{r}\,\mathbf{1}\{\mathcal{T}_{a^\pi(q)} \text{ is a leaf}\} + \frac{1}{r}\,\mathbf{1}\{\mathcal{T}_{a^\pi(q)} \text{ is not a leaf}\}\left(1 + H_{q^{\checkmark}_{a^\pi(q)}}\right) + \frac{r-1}{r}\left(2 + h(\mathcal{T}_{a^\pi(q)}) + H_{q^{\times}_{a^\pi(q)}}\right)\Bigg]. \tag{33}$$

Since by construction, every $q \in \mathcal{Q}_{d,r}$ is invariant under permutations of the children within $\mathcal{U}_q$, the random vector family $(\mathcal{T}_j, q_j^{\checkmark}, q_j^{\times})_{j\in\mathcal{U}_q}$ is exchangeable under q, i.e., for any permutation σ of $\mathcal{U}_q$,

$$(T_j, q_j^{\checkmark}, q_j^{\times})_{j\in\mathcal{U}_q} \overset{d}{=} (\mathcal{T}_{\sigma(j)}, q_{\sigma(j)}^{\checkmark}, q_{\sigma(j)}^{\times})_{j\in\mathcal{U}_q}. \tag{34}$$

Therefore the expectation on the right-hand side of (33) is the same no matter which child in $\mathcal{U}_q$ is chosen first. As a result, H_q^π is independent of $\pi \in \Pi_{\text{success}}$. This proves the induction step.

Finally, for each $l \in [L]$, let q_l be the posterior law at the root induced by $p^l_{\text{env,tree}}$. Then

$$\overline{H}^\pi(l) = H^\pi_{q_l} \tag{35}$$

is independent of $\pi \in \Pi_{\text{success}}$. And by our definition of the environment distribution $\mathcal{P}$ in (25), we have

$$\overline{H}^\pi(\mathcal{P}) = \sum_{l=1}^{L} \omega_l\, \overline{H}^\pi(l). \tag{36}$$

Because each $\overline{H}^\pi(l)$ is independent of $\pi \in \Pi_{\text{success}}$, the weighted sum $\overline{H}^\pi(\mathcal{P})$ is also independent of $\pi \in \Pi_{\text{success}}$. This proves the proposition.

B.2 Proof of Proposition 2

Proof of part 1: $\pi^\star \in \Pi_{\text{success}}$ has the minimum expected trajectory length. First of all, by definition, it's obvious that $\pi^\star \in \Pi_{\text{success}}$. For any fixed tree $\mathcal{T}$ and policy $\pi \in \Pi_{\text{success}}$, we let $H^\pi(\mathcal{T})$ denote the expected number of steps to hit the goal when the goal is generated by the imbalanced goal distribution (26). Same as (32), for the imbalanced case we also define

$$\overline{H}^\pi(l) := \mathbb{E}_{\mathcal{T}\sim p^l_{\text{env,tree}}}\left[H^\pi(\mathcal{T})\right].$$

Then by the definition of the environment distribution $\mathcal{P}$ in (25), we have

$$H^\pi(\mathcal{P}) = \sum_{l=1}^{L} \omega_l \overline{H}^\pi(l). \tag{37}$$

Therefore, to show part 1, it's enough to show that, for every $l \in [L]$, $\pi^\star$ uniquely minimizes $\overline{H}^\pi(l)$ among all policies in Π_{success}.

Fix $l \in [L]$. For $d \geqslant 1$ and nonempty $\mathcal{U} \subseteq [k]$, let $\mathcal{Q}_{d,\mathcal{U}}$ be the collection of all posterior laws q of the current unresolved subtree that can arise as follows:

1. sample the original tree from $p^l_{\text{env,tree}}$;
2. sample the goal from the goal process p^a;
3. condition on a reachable state s and on the event that the goal lies in the unresolved subtree rooted at the node the agent is currently at;
4. require that the remaining depth of that unresolved subtree is at most d, and that the currently unresolved child set is $\mathcal{U}$.

For $q \in \mathcal{Q}_{d,\mathcal{U}}$ and $\pi \in \Pi_{\text{success}}$, we let H_q^π denote the expected number of steps to hit the goal.

We claim that for every $d \geqslant 1$, every nonempty $\mathcal{U} \subseteq [k]$, and every $q \in \mathcal{Q}_{d,\mathcal{U}}$, the unique minimizer of H_q^π over $\pi \in \Pi_{\text{success}}$ is the ranked DFS policy: at the current node, explore the unresolved children

in decreasing order of p_j^a for $j \in \mathcal{U}$; when the current node has no unresolved child left, takes $\texttt{up}$; after moving to a child, apply the same rule recursively at the new node.

Let $r := |\mathcal{U}|$. We prove the above claim by lexicographic induction on the tuple (d, r): we regulate $(d', r') < (d, r)$ if either $d' < d$ or $d' = d$ and $r' < r$.

Fix any $q \in \mathcal{Q}_{d,\mathcal{U}}$, and let

$$S_{\mathcal{U}} := \sum_{j \in \mathcal{U}} p_j^a. \tag{38}$$

For each $j \in \mathcal{U}$, let G_j be the event that the goal lies in the j-th child subtree of the current node. We'll use the following two facts:

- Fact 1: posterior probability of the correct child

  $$\mathbb{P}_q(G_j) = \frac{p_j^a}{S_{\mathcal{U}}}, \ \forall j \in \mathcal{U}.$$

  This is guaranteed by our goal generation process.
- Fact 2: Under q, the unresolved child subtrees $(T_j)_{j \in \mathcal{U}}$ are exchangeable in shape across $j \in \mathcal{U}$. This is because the prior $p^l_{\text{env,tree}}$ is uniform over all full k-ary trees of depth l.

Base case: $d = 1$**.** Let $q \in \mathcal{Q}_{1,\mathcal{U}}$ and write $r := |\mathcal{U}|$. Then all unresolved children of the current node are leaves. Any policy $\pi \in \Pi_{\text{success}}$ must traverse all r child subtrees in some order before goal is found (note that taking $\texttt{up}$ now permanently exits the current unresolved subtree; under the conditioning defining q, this yields failure with probability 1; taking $\texttt{up}$ at any node when it still has unexplored children leads to a positive failure probability). Thus any $\pi \in \Pi_{\text{success}}$ is determined by an order

$$\sigma = (\sigma(1), \cdots, \sigma(r)),$$

of the elements of $\mathcal{U}$, where $\sigma(i) \in \mathcal{U}$ is the child inspected in position i. Note it suffices to consider deterministic orders of $\mathcal{U}$: a randomized policy $\pi \in \Pi_{\text{success}}$ induces a probability distribution μ over deterministic orders σ of $\mathcal{U}$. If $H_q(\sigma)$ denotes the expected remaining path length under order σ and optimal continuation thereafter, then

$$H_q(\sigma) = \sum_{\sigma} \mu(\sigma) H_q(\sigma).$$

Hence $H_q^\pi \geqslant \min_\sigma H_q(\sigma)$. Moreover, if the minimizing order is unique, then any optimal policy must put probability one on that order. If the correct child is inspected in position i, then the number of steps until success is exactly $2(i-1)+1$. Therefore, the expected number of steps until success is

$$H_q^\pi = \sum_{i=1}^{r} \mathbb{P}_q(G_{\sigma(i)}) \left(2(i-1)+1\right) = \frac{1}{S_{\mathcal{U}}} \sum_{i=1}^{r} p^a_{\sigma(i)} \left(2(i-1)+1\right),$$

where the second equality follows from Fact 1. By rearrangement inequality and our assumption that $p_1 > p_2 > \cdots > p_k$ we know that the unique optimal order is

$$\sigma(1), \cdots, \sigma(r) \quad \text{given by decreasing order of } p_j^a \text{ on } \mathcal{U}.$$

This proves the claim for all $q \in \mathcal{Q}_{1,\mathcal{U}}$.

Induction step. Fix $d \geqslant 2$, nonempty $\mathcal{U} \subseteq [k]$, and assume the claim holds for every $(d', r') < (d, r)$. We prove it for $q \in \mathcal{Q}_{d,\mathcal{U}}$.

We still let $r := |\mathcal{U}|$. If $r = 1$, i.e. $\mathcal{U} = \{j\}$ for some $j \in [k]$, then there is only one unresolved child. Any $\pi \in \Pi_{\text{success}}$ must choose $\texttt{down}_j$ next. If that child is a leaf, success occurs in one step; if not, after moving down we obtain a posterior $q' \in \mathcal{Q}_{d-1,[k]}$, and by induction hypothesis, the unique optimal continuation from there is the ranked DFS policy. So the claim holds for $r = 1$. Below we consider $2 \leqslant r \leqslant k$.

Note that under q, any $\pi \in \Pi_{\text{success}}$ must, at the current node, simply choose an order in which to inspect the children in $\mathcal{U}$. So let

$$\sigma = (\sigma(1), \cdots, \sigma(r))$$

be any deterministic order of $\mathcal{U}$. We define two constants:

- Let A_q be the expected remaining steps conditioned on the child currently chosen is the correct child and, after moving into it, one continues optimally from there.
- Let B_q be the expected steps one fully explores that child subtree and returns to the current node, conditioned on the child currently chosen is incorrect

By Fact 2, both A_q and B_q are independent of the identity of the chosen child $j \in \mathcal{U}$. Let $I \in \{1, \cdots, r\}$ by the random rank of the correct child in the order σ. By Fact 1, we have

$$\mathbb{P}(I = i) = \mathbb{P}_q(G_{\sigma(i)}) = \frac{p^a_{\sigma(i)}}{S_{\mathcal{U}}}, \ \forall i \in [r].$$

Conditional on $I = i$, the first $i - 1$ children are wrong, each costing B_q, and the i-th child is correct, costing A_q. Hence

$$H_q(\sigma) = \sum_{i=1}^{r} \frac{p^a_{\sigma(i)}}{S_{\mathcal{U}}} \left((i-1)B_q + A_q\right).$$

Since A_q and $B_q/S_{\mathcal{U}}$ do not depend on σ, $B_q > 0$, and $p^a_1 > p^a_2 > \cdots > p^a_k$, we know that $H_q(\sigma)$ is minimized only if $\sigma(1), \cdots, \sigma(r)$ is the decreasing order of p^a_j on $\mathcal{U}$. So the unique optimal first-choice order at the current node is: among unresolved children, choose the one with largest p^a_j first, then the next largest, etc.

It remains to justify the recursive part. (i) If the chosen child j is correct and non-leaf, then after moving down we obtain a posterior in $\mathcal{Q}_{d-1,[k]}$, and by induction hypothesis, the unique optimal continuation from there is again the ranked DFS policy. (ii) If the chosen child j is incorrect, then after fully exploring it and returning, the posterior belongs to $\mathcal{Q}_{d,\mathcal{U}\setminus\{j\}}$, and by induction hypothesis on $(d, r-1)$, the unique optimal continuation is also the ranked DFS policy on the remaining unresolved children. This completes the induction.

Finally, we apply the claim at the root: for each fixed $l \in [L]$, let q_l denote the initial posterior law at the root $p^l_{\text{env,tree}}$. we have

$$\overline{H}^{\pi}(l) = H^{\pi}_{q_l}.$$

Since $q_l \in \mathcal{Q}_{l,[k]}$, the claim implies that $\pi^\star$ uniquely minimizes $\overline{H}^{\pi}(l)$ over Π_{success}. By (37) and our assumption that $\omega_l > 0$ for all $l \in [L]$ we know that $\pi^\star$ uniquely minimizes $H^{\pi}(\mathcal{P})$ over Π_{success}.

Proof of part 2: $\pi^\star = \arg\max_\pi J^\pi_\gamma(\mathcal{P})$ for all $\gamma \in \left(1 - \Theta\left((p^a_k)^L\right), 1\right)$. We show this by first showing for any $\pi \in \Pi_{\text{success}} \setminus \{\pi^\star\}$, $J^\pi_\gamma(\mathcal{P}) < J^{\pi^\star}_\gamma(\mathcal{P})$, and then showing $\arg\max_\pi J^\pi_\gamma(\mathcal{P}) \subseteq \Pi_{\text{success}}$.

- **Step 1: show for any $\pi \in \Pi_{\text{success}} \setminus \{\pi^\star\}$, $J^\pi_\gamma(\mathcal{P}) < J^{\pi^\star}_\gamma(\mathcal{P})$ for all $\gamma \in (0, 1)$.** Fix $\gamma \in (0, 1)$. We use the same posterior family $\mathcal{Q}_{d,\mathcal{U}}$ as in part 1, write $r := |\mathcal{U}|$, and consider any deterministic order $\sigma = (\sigma(1), \cdots, \sigma(r))$ in which a successful policy $\pi \in \Pi_{\text{success}}$ inspects the children in $\mathcal{U}$ before eventually going up. Let $W_q(i)$ be the discounted return obtained if the correct child is inspected in position i, and after entering that child one continues recursively in the optimal way. By Fact 2 in part 1, $W_q(i)$ depends only on the rank i, not on the label $\sigma(i)$. Moreover, since delaying the correct child by one extra wrong child inserts at least two extra steps (down and back up), and $\gamma \in (0, 1)$, we have

  $$W_q(1) > W_q(2) > \cdots > W_q(r) > 0. \tag{39}$$

  Hence the discounted value of order σ is

  $$V_q(\sigma) = \sum_{i=1}^{r} \frac{p^a_{\sigma(i)}}{S_{\mathcal{U}}} W_q(i).$$

  By (39), (26) and rearrangement inequality we know that $V_q(\sigma)$ is uniquely maximized by $\sigma(1), \cdots, \sigma(r)$ given by decreasing order of p^a_j on $\mathcal{U}$. Then similar as the proof of part 1, applying this recursively after entering the correct child proves, by lexicographic induction on $(d, |\mathcal{U}|)$, that

  $$\pi^\star = \arg \max_{\pi \in \Pi_{\text{success}}} J^\pi_\gamma(\mathcal{P}), \ \forall \gamma \in (0, 1).$$

- **Step 2: show** $\arg\max_\pi J^\pi_\gamma(\mathcal{P}) \subseteq \Pi_{\text{success}}$ **for all** $\gamma \in \left(1 - \Theta\left((p^a_k)^L\right), 1\right)$**.** We show this by fixing any policy $\pi \notin \Pi_{\text{success}}$ and showing $\pi \notin \arg\max_\pi J^\pi_\gamma(\mathcal{P})$. Since $\pi \notin \Pi_{\text{success}}$, it either (i) takes an illegal action at some (reachable, non-terminal) state at some node with positive probability, when there's still untried actions at that node, or (ii) act legally when there's still legal actions to take (with probability 1), but has positive probability to not explore all children of a node before going up at some state.

  We let $J^\star_\gamma(s)$ and $Q^\star_\gamma(s,a)$ denote the optimal value and Q-function with discount factor γ at state s and state-action pair (s,a) respectively. We let $J^\star_\gamma(\mathcal{P})$ denote the optimal expected return with discount factor γ under the training distribution $\mathcal{P}$. For case (i), consider any reachable non-terminal state s such that π assigns positive probability $p > 0$ to illegal actions at s. Let $h(s)$ be the timestep of s, and let $d^\pi_{\mathcal{P}}(s)$ be the probability that s is reached under $(\mathcal{P}, \pi)$. Define $\widetilde{\pi}$ to agree with π before reaching s, and at s choose an action $a^\star \in \arg\max_{a\in\mathcal{A}} Q^\star_\gamma(s,a)$, following an optimal continuation thereafter. Then, conditional on reaching s, policy π loses at least $pJ^\star_\gamma(s)$ in continuation value compared to $\widetilde{\pi}$. Therefore,

$$J^\pi_\gamma(\mathcal{P}) \leqslant J^{\widetilde{\pi}}_\gamma(\mathcal{P}) - d^\pi_{\mathcal{P}}(s)\gamma^{h(s)-1}pJ^\star_\gamma(s) < J^{\widetilde{\pi}}_\gamma(\mathcal{P}) \leqslant J^\star_\gamma(\mathcal{P}).$$

  Therefore, any policy that has positive probability to take illegal actions at a reachable, non-terminal state cannot be optimal.

  Thus it remains to consider case (ii), i.e., the policies that act legally everywhere but are not successful. Without loss of generality, we assume at some reachable, non-terminal state s where the agent is at node u at level $h \in [l-1]$, π, for the first time, takes action `up` with probability 1 before exhausting all currently unresolved children at the current node u, Let r be the number of still-untried children at u, and let d be the remaining depth below u.

  Construct a comparison policy π' that is identical to π except at state s: instead of going up immediately, first explores the r remaining children under u in decreasing p^a_j-order, using $\pi^\star$ recursively inside each such child subtree, and only after that returns to u and resumes π. We let $\mathcal{E}_s$ denote the event that the goal is in the untried children under u, and let $p_s \coloneqq \mathbb{P}(\mathcal{E}_s|s)$ denote the probability that $\mathcal{E}_s$ happens conditioned on the current state s. Let $R_\gamma(s)$ denote the expected return of π' conditioned on $\mathcal{E}_s$. Let $R'_\gamma(s)$ denote the expected return of π conditioned on $\mathcal{E}^c_s$. Let n_s be the number of steps taken to finish exploring the remaining r children under u and return to u under event $\mathcal{E}_s$. Then the expected return of π, π' at s, denoted by $G^\pi_\gamma(s)$ and $G^{\pi'}_\gamma(s)$ respectively, are

$$G^{\pi'}_\gamma(s) = p_s R_\gamma(s) + (1-p_s)\gamma^{n_s} R'_\gamma(s), \quad G^\pi_\gamma(s) = (1-p_s)R'_\gamma(s). \tag{40}$$

  We'll show that when $\gamma \in \left(1 - \Theta\left((p^a_k)^L\right), 1\right)$, $G^{\pi'}_{\gamma,l}(s) > G^\pi_{\gamma,l}(s)$ for all $l \in \{2, \cdots, L\}$. To do so, below we separately bounding p_s, n_s, $R_\gamma(s)$, $R'_\gamma(s)$. We let $\mathcal{U} = \{n_1, \cdots, n_r\} \subset [k]$ denote the set of untried children of u with child indices $n_1 < n_2 < \cdots < n_r$, and write

$$P_r(\mathcal{U}) \coloneqq \sum_{n_i \in \mathcal{U}} p_{n_i}. \tag{41}$$

  Bound p_s. We define

$$q_u \coloneqq \mathbb{P}\left[\text{the goal is in the subtree rooted at } u | s\right].$$

  Then we have

$$q_u \geqslant (p^a_k)^{L-d}. \tag{42}$$

  Recall we assume it's the first time for π to take action `up` before exploring all remaining children under u. (41) and (42) allow us to bound p_s as

$$p_s \geqslant \frac{q_u P_r(\mathcal{U})}{1 - q_u + q_u P_r(\mathcal{U})}. \tag{43}$$

  Bound n_s. Let h_d be the (deterministic) number of steps needed to fully explore a perfect k-ary tree of depth d and return to the root (DFS tour length). A child subtree is explored

by: go $\texttt{down}_i$ (1 step) for all $i \in [k]$, explore that subtree and return ($h_l - 1$ steps), then go $\texttt{up}$ (1 step), hence

$$h_0 = 0, \quad h_d = k(2 + h_{d-1}) = 2k\frac{k^d - 1}{k - 1}. \tag{44}$$

Each wrong child subtree under u has remaining depth at most $d-1$. The largest possible cost to fully explore one such child subtree and return to u is attained by a perfect depth-$(d-1)$ subtree, and equals $2 + h_{d-1}$. Hence

$$n_s \leqslant r(2 + h_{d-1}). \tag{45}$$

Bound $R_\gamma(s)$. Let $\tau_d^\star$ be the random hitting time of the ranked DFS policy $\pi^\star$ on a perfect k-ary subtree of depth d conditioned on the event that the goal is in that subtree, with the clock started at the subtree root. It is easy to show by induction on d that among all full k-ary subtrees of remaining depth at most d, the perfect depth-d subtree gives the smallest discounted continuation value under $\pi^\star$: let $\tau^\star(\mathcal{T})$ denote the conditional hitting time on such a subtree $\mathcal{T}$, then $\mathbb{E}[\gamma^{\tau^\star(\mathcal{T})}] \geqslant \mathbb{E}[\gamma^{\tau_d^\star}]$. Then conditioned on $\mathcal{E}_s$, recall we let π' act the same as $\pi^\star$ in every untried subtree under u, we have

$$R_\gamma(s) \geqslant \frac{1}{P_r(\mathcal{U})} \sum_{i=1}^{r} p_{n_i}^a \mathbb{E}\left[\gamma^{(2+h_{d-1})(i-1)+\tau_{d-1}^\star}\right] \geqslant \frac{p_k^a}{P_r(\mathcal{U})} \frac{1 - \gamma^{(2+h_{d-1})r}}{1 - \gamma^{2+h_{d-1}}} \mathbb{E}\left[\gamma^{\tau_{d-1}^\star}\right]. \tag{46}$$

Bound $R'_\gamma(s)$. Conditioned on $\mathcal{E}_s^c$, if π goes up immediately from u, then any eventual success can occur only after at least two more steps: one step up to the parent, and at least one step down into some other child subtree, because the goal is always a leaf. Therefore,

$$R'_\gamma(s) \leqslant \gamma^2. \tag{47}$$

Combine the bounds. From (45), (46), (47) we deduce

$$\begin{aligned} G_\gamma^{\pi'}(s) - G_\gamma^\pi(s) &\overset{(40)}{=} p_s R_\gamma(s) - (1 - p_s)\left(1 - \gamma^{n_s}\right) R'_\gamma(s) \\ &\geqslant p_s \frac{p_k^a}{P_r(\mathcal{U})} \frac{1 - \gamma^{(2+h_{d-1})r}}{1 - \gamma^{2+h_{d-1}}} \mathbb{E}\left[\gamma^{\tau_{d-1}^\star}\right] - (1 - p_s)\left(1 - \gamma^{(2+h_{d-1})r}\right)\gamma^2. \end{aligned}$$

A sufficient condition for $G_\gamma^{\pi'}(s) > G_\gamma^\pi(s)$ is

$$\frac{p_s}{1 - p_s} > \frac{P_r(\mathcal{U})}{p_k^a} \frac{(1 - \gamma^{(2+h_{d-1})})\gamma^2}{\mathbb{E}[\gamma^{\tau_{d-1}^\star}]}.$$

From (43) we know that

$$\frac{p_s}{1 - p_s} \geqslant \frac{q_u P_r(\mathcal{U})}{1 - q_u} \overset{(42)}{\geqslant} \frac{(p_k^a)^{L-d} P_r(\mathcal{U})}{1 - (p_k^a)^{L-d}}.$$

Combining the above two relations, we know that it's enough to require

$$\frac{(p_k^a)^{L-d}}{1 - (p_k^a)^{L-d}} > \frac{1}{p_k^a} \frac{(1 - \gamma^{(2+h_{d-1})})\gamma^2}{\mathbb{E}[\gamma^{\tau_{d-1}^\star}]}. \tag{48}$$

Now we bound $\mathbb{E}[\gamma^{\tau_{d-1}^\star}]$. Define

$$T_d^\star := \mathbb{E}[\tau_d^\star].$$

Then we have

$$T_0^\star = 0, \quad T_d^\star = T_{d-1}^\star + 1 + \sum_{j=1}^{k} p_j (j-1)(2 + h_{d-1}), \ \forall d \geqslant 1.$$

This further yields

$$T_d^\star = d + \sum_{j=1}^{k} p_j(j-1) \sum_{i=0}^{d-1}(2+h_i) \overset{(44)}{=} d + \frac{2\sum_{j=1}^{k} p_j(j-1)}{k-1}\left(\frac{k(k^d-1)}{k-1} - d\right). \tag{49}$$

By Bernoulli's inequality, we have

$$\gamma^h \geqslant 1 - h(1-\gamma),\ \forall h \geqslant 0,$$

so

$$\mathbb{E}[\gamma^{\tau_{d-1}^\star}] \geqslant 1 - T_{d-1}^\star(1-\gamma). \tag{50}$$

And

$$(1-\gamma^{2+h_{d-1}})\gamma^2 \leqslant 1 - \gamma^{2+h_{d-1}} \leqslant (2+h_{d-1})(1-\gamma). \tag{51}$$

Substituting (50) and (51) into (48), a sufficient condition for $G_\gamma^{\pi'}(s) > G_\gamma^{\pi}(s)$ is

$$(p_k^a)^{L-d+1}\left(1 - T_{d-1}^\star(1-\gamma)\right) > \left(1-(p_k^a)^{L-d}\right)(2+h_{d-1})(1-\gamma),$$

which is equivalent to

$$\gamma > 1 - \underbrace{\frac{(p_k^a)^{L-d+1}}{(p_k^a)^{L-d+1}T_{d-1}^\star + \left(1-(p_k^a)^{L-d}\right)(2+h_{d-1})}}_{:=\delta_d},\ \forall d \in [L].$$

Define

$$\delta := \min_{d\in[L]} \delta_d. \tag{52}$$

Then a sufficient condition for $G_\gamma^{\pi'}(s) > G_\gamma^{\pi}(s)$ is

$$\gamma > 1 - \delta. \tag{53}$$

Below we show

$$\delta = \Theta\left((p_k^a)^L\right). \tag{54}$$

When $d = 1$, since $h_0 = T_0^\star = 0$, we have

$$\delta_1 = \frac{(p_k^a)^L}{2\left(1-(p_k^a)^{L-1}\right)} = \Theta\left((p_k^a)^L\right). \tag{55}$$

For $d \geqslant 2$, by (44) and (49) we know that

$$h_{d-1} = \Theta\left(k^{d-1}\right),\ T_{d-1}^\star = O\left(k^{d-1}\right),$$

so

$$\delta_d = \Theta\left(\frac{(p_k^a)^{L-d+1}}{k^{d-1}}\right),\ \forall 2 \leqslant d \leqslant L. \tag{56}$$

From (52), (55), (56) and the fact that $p_k < 1/k$ we know that (54) holds.

C Convergence and Generalization Results

C.1 Formal statements of Theorems 2 and Theorem 3

Below we provide the formal statements of Theorems 2 and Theorem 3 in a merged version, where we use $\mathsf{poly}(k)$ to denote a polynomial function of k.

Theorem 5 (convergence and generalization results under balanced goal distribution, formal)**.** *Suppose (21), Assumption 1 hold, $k \in \mathbb{N}_+$ is larger than some absolute constant, $N = \mathsf{poly}(k) \geqslant \max\left\{k^5, \frac{k^{L+1}-1}{k-1}\right\}$, L is treated as a constant, and $\eta > 0$ is smaller than some absolute constant, $\epsilon > 0$ satisfies $\ln(1/\epsilon) \geqslant e^k$. Let $\delta_1 = \epsilon^{1.05}$, $\delta_2 = \frac{\epsilon^{1.02}}{k^L}$. Under our training setting, for depth-1 curriculum training, the success rate $R(\pi; \mathcal{P}_1)$ reaches $1-\delta_1$ in $T_1 = O\left(\frac{1}{\eta\epsilon^{1.1}}\right)$ iterations. At the end of depth-1 curriculum training, we have*

$$R(\pi^{(T_1)}; \mathcal{P}_l) = O\left(\frac{1}{k^{l-1}}\right), \quad \forall 2 \leqslant l \leqslant L. \tag{57}$$

For depth-2 curriculum training, the success rate $R(\pi; \mathcal{P}_2)$ reaches $1-\delta_2$ in $T_2 - T_1 = O\left(\frac{k^L}{\eta\epsilon^{1.1}}\right)$ iterations. After depth-2 curriculum training, given any full k-ary tree $\mathcal{T}$ with depth $l \in [L]$ and any goal node $g(\mathcal{T})$, we have

$$R(\pi^{(T_2)}; (\mathcal{T}, g(\mathcal{T}))) \geqslant 1 - \epsilon. \tag{58}$$

The proof of Theorem 5 is provided in Appendix C.2. Note that in Theorem 5, without loss of generality, we assume (the hardest and the most extreme case where) $\ln(1/\epsilon) \geqslant e^k$. If ϵ is larger, the convergence bound in Theorem 5 also holds: the proof still goes through and possibly becomes simpler with a bit of extra discussions, as depth-1 curriculum training may end before $T_{1,i}$ for some $i \in [5]$ defined in Appendix C.2 and thus we don't need to track some sub-phases.

Comparison with related training-dynamics analyses. As far as we know, ours is the first optimization-level analysis of policy-gradient training in which coupled transformer parameters $(B, C, \overline{P}, Q)$ coevolve into a two-head memory circuit for agentic search, yielding DFS and depth generalization from depth-1/2 curriculum training to deeper unseen trees. Prior analyses simplify the setting along different axes. Graph-search works typically study static instances whose graph structure is revealed in the prompt (Saparov et al., 2025; Wang et al., 2024; Yang et al., 2025). RL planning analyses operate on fixed graph abstractions, where generation amounts to extending a path by next-node tokens and process rewards may expose adjacency or target information (Wang et al., 2026). RL/RLVR reasoning analyses, while using outcome or verifier rewards, largely treat RL as post-training over self-generated reasoning traces rather than interaction with a partially observed environment (Bu et al., 2025; Huang et al., 2026); in these settings the evolving state is essentially the generated prefix, or the analysis assumes/fixes atomic operations needed for reasoning.

By contrast, our transformer is an agent in a hidden stochastic tree: it must infer the search state from trajectory history and sparse terminal reward, while learning to remember failed branches, avoid repeated actions, and backtrack from both failed leaves and exhausted internal nodes. This richer interactive setting also makes the training-dynamics analysis substantially more delicate: the proof must track a multi-phase evolution of interacting attention and output matrices, while resolving the relevant return-to-go terms under both a stochastic policy and posterior uncertainty over the hidden environment.

C.2 Proof of Theorem 5

Notation. Though we fix the last row of P to be 0 and don't update it (c.f. (16)), its gradient is still useful and we'll compute it.

$$\delta M^{(t)} := \mathbb{E}_{\Phi\sim\mathcal{P}, \pi_{\theta^{(t)}}} \left[\sum_{h=1}^{H} \nabla_M \log \pi_\theta(a_h \mid E^{(h)}) \Big|_{M=M^{(t)}} \sum_{i=h}^{H} r_{i+1} \right] \quad \text{for } M \in \{B, C, P, Q\}. \tag{59}$$

Then according to update rule (18), we have

$$M^{(t+1)} = M^{(t)} + \eta \delta M^{(t)} \quad \text{for } M \in \{B, C, \overline{P}, Q\}. \tag{60}$$

More generally, for any quantity $\{X^{(t)}\}_{t\in\mathbb{N}}$ (except for the last row of P), we define

$$\delta X^{(t)} := \frac{X^{(t+1)} - X^{(t)}}{\eta}. \tag{61}$$

We let

$$U := (u_0, \cdots, u_N) \in \mathbb{R}^{d_1\times(N+1)},\ V := (v_0, \cdots, v_{k+1}) \in \mathbb{R}^{d_2\times(k+2)},\ Z := (z_0, z_1, z_2) \in \mathbb{R}^{d_3\times 3} \tag{62}$$

denote the node, action, and label embedding matrices respectively. We define the induced matrix A_M ($M \in \{B, C, P, Q\}$) as

$$\begin{aligned} A_B := U^\top BU \in \mathbb{R}^{(N+1)\times(N+1)}, \quad A_C := Z^\top CZ \in \mathbb{R}^{3\times 3}, \\ A_P := PV \in \mathbb{R}^{(k+1)\times(k+2)}, \quad A_Q := QZ \in \mathbb{R}^{(k+1)\times 3}, \end{aligned} \tag{63}$$

and let $A_M^{(t)} := A_{M^{(t)}}$ and $\delta A_M^{(t)} := \delta A_{M^{(t)}}$ for $M \in \{B, C, P, Q\}$. Then by (60), we have

$$A_M^{(t+1)} = A_M^{(t)} + \eta\delta A_M^{(t)} \ \text{ for } M \in \{B, C, \overline{P}, Q\}. \tag{64}$$

We define

$$\begin{aligned} \pi_h := \pi_\theta(E^{(h)}), \quad \pi_h^{(t)} := \pi_{\theta^{(t)}}(E^{(h)}), \quad \widehat{G}_h := \sum_{i=h}^{H} r_{i+1}, \\ Q_h^{(t)}(a) := \mathbb{E}_{\pi^{(t)}}[\widehat{G}_h|\mathcal{F}_h, a_h = a], \quad G_h^{(t)} := \mathbb{E}_{\pi^{(t)}}[\widehat{G}_h|\mathcal{F}_h] = \sum_{a\in\mathcal{A}} \pi_h^{(t)}(a)Q_h^{(t)}(a), \end{aligned} \tag{65}$$

where $\mathcal{F}_h$ denotes the σ-field generated by the history available before choosing a_h, equivalently by $E^{(h)}$. We also write $Q_h^{(t)} \in \mathbb{R}^{k+1}$ for the vector whose a-th coordinate is $Q_h^{(t)}(a)$. Define N_h as the number of leaf nodes explored up to step h:

$$N_{h,\times} := \sum_{i=1}^{h} \mathbb{1}\{c_i = \times\}. \tag{66}$$

At any step h, the agent is at node n_h. We let

$$\mathcal{A}_{h,\times} := \begin{cases} \{i \in [k] | i \text{ is taken at } n_h\}, & \text{if } c_h = 0, \\ \varnothing, & \text{if } c_h = \times, \end{cases} \tag{67}$$

denote the set of downward actions that are taken at node n_h. Then for any action $i \in \mathcal{A}_{h,\times}$, $Q_h^{(t)}(i) = 0$. We also define

$$\forall i \in [k+1]: \quad N_h(i) := \sum_{j=1}^{h-1} \mathbb{1}\{a_j = i\} \tag{68}$$

as the number of times action i is taken up to step h. And we let $N_h(\texttt{up}) := N_h(k+1)$.

Throughout the proof, we use $o(1)$ or e.g., $O\left(\frac{1}{k}\right)$ to represent small quantities that may also be negative. We sometimes abuse the notation to use $[t_1, t_2]$ to denote $\{t_1, t_1+1, \ldots, t_2\}$ for integers $t_1 \leqslant t_2$.

We first derive the expressions of $\delta A_M^{(t)}$ and the policy π_h in Lemma 1. The expectations at step t are all taken under policy $\pi^{(t)}$. Note that histories used for gradient updates exclude the terminal observation, and thus terminal ✓ token never appears in $E_c^{(h)}$.

Lemma 1 (gradients and policy expression)**.** *Under balanced goal distribution* (21)*, Assumption 1, we have the following gradient expressions for any $t \in \mathbb{N}$:*

$$A_C^{(t)} = \begin{pmatrix} a_{c,0}^{(t)} & -a_{c,1}^{(t)} & 0 \\ -a_{c,0}^{(t)} & a_{c,1}^{(t)} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \delta A_C^{(t)} = \begin{pmatrix} \delta a_{c,0}^{(t)} & -\delta a_{c,1}^{(t)} & 0 \\ -\delta a_{c,0}^{(t)} & \delta a_{c,1}^{(t)} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{69}$$

where

$$\delta a_{c,0}^{(t)} = \left(1+\frac{1}{k}\right)\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=0\}\frac{N_{h,\times}(h-N_{h,\times})}{\left(N_{h,\times}e^{-a_{c,0}^{(t)}}+(h-N_{h,\times})e^{a_{c,0}^{(t)}}\right)^2} \times\left(a_{q,\times}^{(t)}+a_{q,0}^{(t)}\right)\pi_h^{(t)}(up)\left(G_h^{(t)}-Q_h^{(t)}(up)\right)\right], \tag{70}$$

$$\delta a_{c,1}^{(t)} = \left(1+\frac{1}{k}\right)\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=\times\}\frac{N_{h,\times}(h-N_{h,\times})}{\left(N_{h,\times}e^{a_{c,1}^{(t)}}+(h-N_{h,\times})e^{-a_{c,1}^{(t)}}\right)^2} \times\left(a_{q,\times}^{(t)}+a_{q,0}^{(t)}\right)\pi_h^{(t)}(up)(1-\pi_h^{(t)}(up))Q_h^{(t)}(up)\right]. \tag{71}$$

$$A_B^{(t)} = \begin{pmatrix} 0 & a_{b,0}^{(t)} & a_{b,0}^{(t)} & \cdots & a_{b,0}^{(t)} \\ 0 & a_{b,1}^{(t)} & a_{b,2}^{(t)} & \cdots & a_{b,2}^{(t)} \\ 0 & a_{b,2}^{(t)} & a_{b,1}^{(t)} & \cdots & a_{b,2}^{(t)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_{b,2}^{(t)} & a_{b,2}^{(t)} & \cdots & a_{b,1}^{(t)} \end{pmatrix}, \quad \delta A_B^{(t)} = \begin{pmatrix} 0 & \delta a_{b,0}^{(t)} & \delta a_{b,0}^{(t)} & \cdots & \delta a_{b,0}^{(t)} \\ 0 & \delta a_{b,1}^{(t)} & \delta a_{b,2}^{(t)} & \cdots & \delta a_{b,2}^{(t)} \\ 0 & \delta a_{b,2}^{(t)} & \delta a_{b,1}^{(t)} & \cdots & \delta a_{b,2}^{(t)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \delta a_{b,2}^{(t)} & \delta a_{b,2}^{(t)} & \cdots & \delta a_{b,1}^{(t)} \end{pmatrix}, \tag{72}$$

where

$$\delta a_{b,0}^{(t)} = \frac{1}{N}\mathbb{E}\Bigg[\sum_{h=1}^{H}\frac{e^{a_{b,0}^{(t)}}}{\left(e^{a_{b,0}^{(t)}}+|\mathcal{A}_{h,\times}|e^{a_{b,1}^{(t)}}+(h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}^{(t)}}\right)^2} \cdot\left\{\left(|\mathcal{A}_{h,\times}|e^{a_{b,1}^{(t)}}+(h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}^{(t)}}\right)\beta_{h,0}^{(t)}-e^{a_{b,1}^{(t)}}\beta_{h,1}^{(t)}-e^{a_{b,2}^{(t)}}\beta_{h,2}^{(t)}\right\}\Bigg], \tag{73}$$

$$\delta a_{b,1}^{(t)} = \frac{1}{N}\mathbb{E}\Bigg[\sum_{h=1}^{H}\frac{e^{a_{b,1}^{(t)}}}{\left(e^{a_{b,0}^{(t)}}+|\mathcal{A}_{h,\times}|e^{a_{b,1}^{(t)}}+(h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}^{(t)}}\right)^2} \cdot\left\{\left(e^{a_{b,0}^{(t)}}+(h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}^{(t)}}\right)\beta_{h,1}^{(t)}-e^{a_{b,0}^{(t)}}|\mathcal{A}_{h,\times}|\beta_{h,0}^{(t)} -e^{a_{b,2}^{(t)}}|\mathcal{A}_{h,\times}|\beta_{h,2}^{(t)}\right\}\Bigg], \tag{74}$$

$$a_{b,2}^{(t)} = \frac{-a_{b,0}^{(t)}-a_{b,1}^{(t)}}{N-1}, \quad \delta a_{b,2}^{(t)} = \frac{-\delta a_{b,0}^{(t)}-\delta a_{b,1}^{(t)}}{N-1}, \tag{75}$$

with

$$\beta_{h,0}^{(t)} := \frac{a_{p,0}^{(t)}}{k}\pi_h^{(t)}(up)\left(Q_h^{(t)}(up)-G_h^{(t)}\right), \tag{76}$$

$$\beta_{h,1}^{(t)} := (a_{p,1}^{(t)}+a_{p,2}^{(t)})\sum_{i\in\mathcal{A}_{h,\times}}\pi_h^{(t)}(i)G_h^{(t)}-a_{p,2}^{(t)}|\mathcal{A}_{h,\times}|\pi_h^{(t)}(up)(Q_h^{(t)}(up)-G_h^{(t)}), \tag{77}$$

$$\begin{aligned}\beta_{h,2}^{(t)} :=& -\left((h-1-N_h(up)-|\mathcal{A}_{h,\times}|)a_{p,2}^{(t)}+\frac{1}{k}N_h(up)a_{p,up}^{(t)}\right) \\ & \pi_h^{(t)}(up)(Q_h^{(t)}(up)-G_h^{(t)}) \\ & +(a_{p,1}^{(t)}+a_{p,2}^{(t)})\left(\sum_{i\in\mathcal{A}_{h,\times}}(N_h(i)-1)\pi_h^{(t)}(i)G_h^{(t)}+\sum_{i\in[k]\setminus\mathcal{A}_{h,\times}}N_h(i)\pi_h^{(t)}(i)(G_h^{(t)} \right. \\ & \left. -Q_h^{(t)}(i))\right).\end{aligned} \tag{78}$$

$$A_P^{(t)} = \begin{pmatrix} -\frac{a_{p,0}^{(t)}}{k} & -a_{p,1}^{(t)} & a_{p,2}^{(t)} & \cdots & a_{p,2}^{(t)} & \frac{a_{p,up}^{(t)}}{k} \\ -\frac{a_{p,0}^{(t)}}{k} & a_{p,2}^{(t)} & -a_{p,1}^{(t)} & \cdots & a_{p,2}^{(t)} & \frac{a_{p,up}^{(t)}}{k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{a_{p,0}^{(t)}}{k} & a_{p,2}^{(t)} & a_{p,2}^{(t)} & \cdots & -a_{p,1}^{(t)} & \frac{a_{p,up}^{(t)}}{k} \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

$$\delta A_P^{(t)} = \begin{pmatrix} -\frac{\delta a_{p,0}^{(t)}}{k} & -\delta a_{p,1}^{(t)} & \delta a_{p,2}^{(t)} & \cdots & \delta a_{p,2}^{(t)} & \frac{\delta a_{p,up}^{(t)}}{k} \\ -\frac{\delta a_{p,0}^{(t)}}{k} & \delta a_{p,2}^{(t)} & -\delta a_{p,1}^{(t)} & \cdots & \delta a_{p,2}^{(t)} & \frac{\delta a_{p,up}^{(t)}}{k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{\delta a_{p,0}^{(t)}}{k} & \delta a_{p,2}^{(t)} & \delta a_{p,2}^{(t)} & \cdots & -\delta a_{p,1}^{(t)} & \frac{\delta a_{p,up}^{(t)}}{k} \\ \delta a_{p,0}^{(t)} & \delta a_{p,3}^{(t)} & \delta a_{p,3}^{(t)} & \cdots & \delta a_{p,3}^{(t)} & -\delta a_{p,up}^{(t)} \end{pmatrix}, \tag{79}$$

where

$$\delta a_{p,3}^{(t)} = \delta a_{p,1}^{(t)} - (k-1)\delta a_{p,2}^{(t)}, \quad a_{p,3}^{(t)} \coloneqq \eta \sum_{t'=0}^{t-1} \delta a_{p,3}^{(t')} = a_{p,1}^{(t)} - (k-1)a_{p,2}^{(t)}, \tag{80}$$

$$\delta a_{p,0}^{(t)} = \mathbb{E}\left[\sum_{h=1}^{H} \pi_h^{(t)}(up) \frac{e^{a_{b,0}^{(t)}}}{e^{a_{b,0}^{(t)}} + |\mathcal{A}_{h,\times}| e^{a_{b,1}^{(t)}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}^{(t)}}} \Big(Q_h^{(t)}(up) - G_h^{(t)} \Big) \right], \tag{81}$$

$$\begin{aligned} \delta a_{p,1}^{(t)} = \mathbb{E}\Bigg[\sum_{h=1}^{H} \pi_h^{(t)}(j) \Bigg(& \mathbb{1}\{j \in \mathcal{A}_{h,\times}\} \frac{e^{a_{b,1}^{(t)}} + (N_h(j)-1)e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + |\mathcal{A}_{h,\times}| e^{a_{b,1}^{(t)}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}^{(t)}}} G_h^{(t)} \\ & - \mathbb{1}\{j \in [k] \backslash \mathcal{A}_{h,\times}\} \frac{N_h(j) e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + |\mathcal{A}_{h,\times}| e^{a_{b,1}^{(t)}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}^{(t)}}} \Big(Q_h^{(t)}(j) \\ & - G_h^{(t)} \Big) \Bigg) \Bigg], \end{aligned} \tag{82}$$

$$\begin{aligned} \delta a_{p,3}^{(t)} = \mathbb{E}\Bigg[\sum_{h=1}^{H} \pi_h^{(t)}(up) \Bigg(& \mathbb{1}\{j \in \mathcal{A}_{h,\times}\} \frac{e^{a_{b,1}^{(t)}} + (N_h(j)-1)e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + |\mathcal{A}_{h,\times}| e^{a_{b,1}^{(t)}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}^{(t)}}} \\ & + \mathbb{1}\{j \in [k] \backslash \mathcal{A}_{h,\times}\} \frac{N_h(j) e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + |\mathcal{A}_{h,\times}| e^{a_{b,1}^{(t)}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}^{(t)}}} \Bigg) (Q_h^{(t)}(up) \\ & - G_h^{(t)}) \Bigg], \end{aligned} \tag{83}$$

$$\begin{aligned} \delta a_{p,up}^{(t)} = -\mathbb{E}\Bigg[\sum_{h=1}^{H} \pi_h^{(t)}(up) & \frac{N_h(up) e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + |\mathcal{A}_{h,\times}| e^{a_{b,1}^{(t)}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}^{(t)}}} \\ & \Big(Q_h^{(t)}(up) - G_h^{(t)} \Big) \Bigg], \end{aligned} \tag{84}$$

$$A_Q^{(t)} = \begin{pmatrix} \frac{a_{q,0}^{(t)}}{k} & -\frac{a_{q,\times}^{(t)}}{k} & 0 \\ \vdots & \vdots & \vdots \\ \frac{a_{q,0}^{(t)}}{k} & -\frac{a_{q,\times}^{(t)}}{k} & 0 \\ -a_{q,0}^{(t)} & a_{q,\times}^{(t)} & 0 \end{pmatrix}, \quad \delta A_Q^{(t)} = \begin{pmatrix} \frac{\delta a_{q,0}^{(t)}}{k} & -\frac{\delta a_{q,\times}^{(t)}}{k} & 0 \\ \vdots & \vdots & \vdots \\ \frac{\delta a_{q,0}^{(t)}}{k} & -\frac{\delta a_{q,\times}^{(t)}}{k} & 0 \\ -\delta a_{q,0}^{(t)} & \delta a_{q,\times}^{(t)} & 0 \end{pmatrix}, \tag{85}$$

where

$$\delta a_{q,0}^{(t)} = -\mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = \times\} \frac{(h - N_{h,\times})e^{-a_{c,1}^{(t)}}}{N_{h,\times}e^{a_{c,1}^{(t)}} + (h - N_{h,\times})e^{-a_{c,1}^{(t)}}} \right.$$
$$\left. \pi_h^{(t)}(up)(1 - \pi_h^{(t)}(up))Q_h^{(t)}(up)\right]$$
$$+ \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0\} \frac{(h - N_{h,\times})e^{a_{c,0}^{(t)}}}{N_{h,\times}e^{-a_{c,0}^{(t)}} + (h - N_{h,\times})e^{a_{c,0}^{(t)}}} \pi_h^{(t)}(up)\left(G_h^{(t)} - Q_h^{(t)}(up)\right)\right], \tag{86}$$

$$\delta a_{q,\times}^{(t)} = \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = \times\} \frac{N_{h,\times}e^{a_{c,1}^{(t)}}}{N_{h,\times}e^{a_{c,1}^{(t)}} + (h - N_{h,\times})e^{-a_{c,1}^{(t)}}} \right.$$
$$\left. \pi_h^{(t)}(up)(1 - \pi_h^{(t)}(up))Q_h^{(t)}(up)\right]$$
$$- \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0\} \frac{N_{h,\times}e^{-a_{c,0}^{(t)}}}{N_{h,\times}e^{-a_{c,0}^{(t)}} + (h - N_{h,\times})e^{a_{c,0}^{(t)}}} \pi_h^{(t)}(up)\left(G_h^{(t)} - Q_h^{(t)}(up)\right)\right]. \tag{87}$$

Furthermore, for all $t \in \mathbb{N}$, $\pi_h^{(t)}$ can be expressed as follows:

$$\pi_h^{(t)}(up) = \frac{1}{Z_h^{(t)}} \exp\left(\left(1 + \frac{1}{k}\right)\xi_h^{(t)}\right),$$
$$\forall i \in [k]: \quad \pi_h^{(t)}(i) = \frac{1}{Z_h^{(t)}} \exp\left(\frac{\varphi_h^{(t)}(i)}{e^{a_{b,0}^{(t)}} + |\mathcal{A}_{h,\times}|e^{a_{b,1}^{(t)}} + (h - 1 - |\mathcal{A}_{h,\times}|)e^{a_{b,2}^{(t)}}}\right), \tag{88}$$

where

$$\varphi_h^{(t)}(i) = -\frac{a_{p,0}^{(t)}}{k}e^{a_{b,0}^{(t)}} + |\mathcal{A}_{h,\times}|a_{p,2}^{(t)}e^{a_{b,1}^{(t)}} - \mathbb{1}\{i \in \mathcal{A}_{h,\times}\}\left((a_{p,1}^{(t)} + a_{p,2}^{(t)})(e^{a_{b,1}^{(t)}} - e^{a_{b,2}^{(t)}})\right)$$
$$+ \left(N_h(up)\frac{a_{p,up}^{(t)}}{k} - N_h(i)a_{p,1}^{(t)} + (h - 1 - N_h(up) - N_h(i) - |\mathcal{A}_{h,\times}|)a_{p,2}^{(t)}\right)e^{a_{b,2}^{(t)}} \tag{89}$$

for all $i \in [k]$,

$$\xi_h^{(t)} = \begin{cases} \frac{N_{h,\times}e^{-a_{c,0}^{(t)}}a_{q,\times}^{(t)} - (h - N_{h,\times})e^{a_{c,0}^{(t)}}a_{q,0}^{(t)}}{N_{h,\times}e^{-a_{c,0}^{(t)}} + (h - N_{h,\times})e^{a_{c,0}^{(t)}}}, & \text{if } c_h = 0, \\ \frac{N_{h,\times}e^{a_{c,1}^{(t)}}a_{q,\times}^{(t)} - (h - N_{h,\times})e^{-a_{c,1}^{(t)}}a_{q,0}^{(t)}}{N_{h,\times}e^{a_{c,1}^{(t)}} + (h - N_{h,\times})e^{-a_{c,1}^{(t)}}}, & \text{if } c_h = \times, \end{cases} \tag{90}$$

and $Z_h^{(t)}$ is the normalizing factor:

$$Z_h^{(t)} := \exp\left(\left(1 + \frac{1}{k}\right)\xi_h^{(t)}\right) + \sum_{i=1}^{k} \exp$$
$$\left(\frac{\varphi_h^{(t)}(i)}{e^{a_{b,0}^{(t)}} + |\mathcal{A}_{h,\times}|e^{a_{b,1}^{(t)}} + (h - 1 - |\mathcal{A}_{h,\times}|)e^{a_{b,2}^{(t)}}}\right). \tag{91}$$

The proof of Lemma 1 is postponed to Appendix C.3.1.

C.2.1 Depth-1 Curriculum Training

To analyze the convergence of Depth-1 curriculum training, we define the following time points:

$$T_{1/2} := \sup\left\{t \in \mathbb{N} : a_{p,1}^{(t)} \leqslant 4.1k \ln k\right\}. \tag{92}$$

We divide depth-1 curriculum training into two sub-phases by $T_{1/2}$: Phase 1.1: $t \leqslant T_{1/2}$ and Phase 1.2: $T_{1/2} + 1 \leqslant t \leqslant T_1$.

Phase 1.1: $t \leqslant T_{1/2}$. Define

$$T_{1,1} := \sup\left\{t \in \mathbb{N} : a_{q,\times}^{(t)} \leqslant \ln k\right\}, \tag{93}$$

$$T_{1,2} := \sup\left\{t \in \mathbb{N} : a_{p,1}^{(t)} \leqslant k\right\}, \tag{94}$$

We will show that before $T_{1,1}$, matrix A_P is very close to 0, and the agent cannot distinguish if an action is taken or not. In this time period, the agent learns how to go up at leaf nodes that are not goal nodes, and go down at the root. Then between $T_{1,1}$ and $T_{1,2}$, $a_{p,1}$ rapidly increases to k, the agent has high confidence to go up at leaf nodes and go down at the root, but is still bad at distinguishing if an action is taken or not when the trajectory gets long. Then from $T_{1,2}$ to $T_{1/2}$, $a_{p,1}$ increases to $4k \ln k$, the agent has high confidence to follow all rules at trees of depth 1, and by the end of this period, the agent has a success rate higher than $1 - \frac{1}{k}$. During the whole Phase 1.1, matrix A_B is very close to 0 and can be neglected.

We rigorously state the training dynamics of Phase 1.1 in the following lemma:

Lemma 2. *Under balanced goal distribution (21), Assumption 1, for all $t \leqslant T_{1/2}$, for all $t' \leqslant t \leqslant T_{1/2}$, there exists an absolute constant $C > 0$ such that*

$$-\frac{\ln^2 k}{Nk} \lesssim a_{b,1}^{(t)} \lesssim \frac{k^3 \ln^2 k}{N}, \tag{95}$$

$$-\frac{k^3 \ln^2 k}{N} \lesssim a_{b,0}^{(t)} \lesssim \frac{\ln^2 k}{N}, \tag{96}$$

$$\left|a_{b,2}^{(t)}\right| \lesssim \frac{k^3 \ln^2 k}{N^2}, \tag{97}$$

$$e^{2a_{c,0}^{(t)}} - e^{2a_{c,0}^{(t')}} \geqslant 0, \tag{98}$$

$$e^{2a_{c,1}^{(t)}} - e^{2a_{c,1}^{(t')}} \geqslant \max\left\{\left(a_{q,\times}^{(t')} + a_{q,0}^{(t')}\right)\left(a_{q,\times}^{(t)} - a_{q,\times}^{(t')}\right), 0\right\}, \tag{99}$$

$$|a_{p,0}^{(t)}| \lesssim \ln k, \quad |a_{p,up}^{(t)}| \lesssim \ln k, \tag{100}$$

$$a_{p,1}^{(t)} - a_{p,1}^{(t')} \geqslant 0, \tag{101}$$

$$a_{q,0}^{(t)} \lesssim a_{q,\times}^{(t)}, \tag{102}$$

$$a_{q,\times}^{(t)} + a_{q,0}^{(t)} \asymp a_{q,\times}^{(t)}, \tag{103}$$

$$\left(a_{q,\times}^{(t)} + a_{q,0}^{(t)}\right) - \left(a_{q,\times}^{(t')} + a_{q,0}^{(t')}\right) \geqslant 0, \tag{104}$$

$$-\left(a_{q,0}^{(t)} - a_{q,0}^{(t')}\right) \lesssim \frac{1}{e^{2a_{c,1}^{(t')}}}\left(\left(a_{q,\times}^{(t)} + a_{q,0}^{(t)}\right) - \left(a_{q,\times}^{(t')} + a_{q,0}^{(t')}\right)\right), \tag{105}$$

$$\xi_{1,2}^{(t)} \lesssim a_{q,\times}^{(t)}. \tag{106}$$

Moreover, for all $t' \leqslant t \leqslant T_{1,1}$:

$$a_{p,1}^{(t)} \lesssim \frac{\ln k}{k^2}, \quad |a_{p,2}^{(t)}| \lesssim \frac{\ln k}{k^2}, \tag{107}$$

$$a_{q,\times}^{(t)} - a_{q,\times}^{(t')} \gtrsim \frac{\eta(t-t')}{k^2}, \tag{108}$$

$$\left(a_{q,\times}^{(t)} + a_{q,0}^{(t)}\right) - \left(a_{q,\times}^{(t')} + a_{q,0}^{(t')}\right) \asymp a_{q,\times}^{(t)} - a_{q,\times}^{(t')}, \tag{109}$$

$$a_{q,0}^{(t)} - a_{q,0}^{(t')} \leqslant \frac{2}{3}\left(a_{q,\times}^{(t)} - a_{q,\times}^{(t')}\right) + C. \tag{110}$$

For any t', t such that $T_{1,1} + 1 \leqslant t' \leqslant t \leqslant T_{1/2}$, we have

$$a_{p,1}^{(t)} - a_{p,1}^{(t')} \gtrsim \frac{\eta(t-t')}{k^5}, \tag{111}$$

$$e^{2a_{c,1}^{(t)}} \gtrsim \ln^2 k. \tag{112}$$

For any t', t such that $T_{1,1}+1 \leqslant t' \leqslant t \leqslant T_{1,2}$, we have

$$-\frac{\ln k}{k^2} \lesssim a_{p,2}^{(t)} \leqslant 1 + O\left(\frac{\ln k}{k^2}\right), \tag{113}$$

$$e^{2a_{c,0}^{(t)}} - e^{2a_{c,0}^{(t')}} \gtrsim \left(a_{q,\times}^{(t')} + a_{q,0}^{(t')}\right)\left(a_{q,0}^{(t)} - a_{q,0}^{(t')}\right), \tag{114}$$

$$0 \leqslant a_{q,\times}^{(t)} - a_{q,0}^{(t)} \leqslant 2.1 \ln k, \tag{115}$$

$$a_{q,0}^{(t)} + a_{q,\times}^{(t)} \asymp \ln k. \tag{116}$$

For any t', t such that $T_{1,2}+1 \leqslant t' \leqslant t \leqslant T_{1/2}$, we have

$$-C \leqslant a_{q,\times}^{(t)} - a_{q,0}^{(t)} \leqslant 6.2 \ln k, \tag{117}$$

$$a_{q,0}^{(t)} \geqslant 1.9 \ln k - 1, \tag{118}$$

$$a_{q,\times}^{(t)} - \frac{1}{2} a_{p,2}^{(t)} \geqslant 2.425 \ln k, \tag{119}$$

$$a_{p,2}^{(t)} = \frac{1+o(1)}{k} a_{p,1}^{(t)}, \tag{120}$$

$$e^{2a_{c,0}^{(t)}} \asymp \ln^2 k, \tag{121}$$

$$5.9 \ln k \leqslant a_{q,0}^{(t)} + a_{q,\times}^{(t)} \leqslant 20.4 \ln k. \tag{122}$$

When $t = T_{1/2}$, we have

$$a_{b,1}^{(T_{1/2})} \asymp \frac{k^3 \ln^2 k}{N}, \tag{123}$$

$$-\frac{k^3 \ln^2 k}{N} \lesssim a_{b,0}^{(T_{1/2})} \lesssim -\frac{k^2 \ln^2 k}{N}, \tag{124}$$

$$\frac{1}{k^{3.05}} \lesssim 1 - R\left(\pi^{(T_{1/2})}; \mathcal{P}_1\right) \lesssim \frac{1}{k^{1.01}}, \tag{125}$$

$$\forall 2 \leqslant l \leqslant L: \quad R\left(\pi^{(T_{1/2})}; \mathcal{P}_l\right) \lesssim \frac{1}{k}, \tag{126}$$

$$a_{q,\times}^{(T_{1/2})} - \frac{1}{2} a_{p,2}^{(T_{1/2})} \geqslant 3.9 \ln k. \tag{127}$$

The proof of Lemma 2 is postponed to Appendix C.3.2.

As a consequence of Lemma 2, we have

$$T_{1/2} \lesssim \frac{k^6 \ln k}{\eta}. \tag{128}$$

Phase 1.2: $T_{1/2}+1 \leqslant t \leqslant T_1$. At Phase 1.2, the effect of matrix A_B becomes non-negligible. Specifically, its main diagonal $a_{b,1}$ will monotonically increase, $a_{b,0}$ is negative at first, and then becomes positive if ϵ is small enough (smaller than $\frac{1}{k^{9/4} N \exp\left(k^{1.3}\sqrt{N}\right)}$), and $a_{b,2}$ stays small. Same as in Phase 1.1, $a_{c,0}$ and $a_{c,1}$ will monotonically increase, but $a_{c,0}$'s increase becomes much slower and $e^{2a_{c,0}}$ stays order $\ln^2 k$ after Phase 1.1. $a_{p,1}$ will monotonically increase, $a_{p,2} \approx \frac{a_{p,1}}{k-1}$ after Phase 1.1, and $a_{p,0}, a_{p,\mathrm{up}}$ are no larger than $a_{p,2}$ in magnitude. $a_{q,0}, a_{q,\times}$ increase overall, and has the same magnitude as $a_{p,2}$. The failure probability will keep decreasing until it's bounded by $\epsilon^{1.05}$ at step T_1.

To formalize the above description, we define

$$T_{1,3} := \sup\left\{t \in \mathbb{N} : a_{p,1}^{(t)} \leqslant 21 k \ln k\right\}, \tag{129}$$

$$T_{1,4} := \inf\left\{t \in \mathbb{N} : e^{a_{b,0}^{(t)}} \geqslant \frac{k}{\ln k}\right\}, \tag{130}$$

and for all $h \in [k]$, define

$$\mu_{2h-1}^{(t)} := \frac{-\frac{a_{p,0}^{(t)}}{k} e^{a_{b,0}^{(t)}} + \frac{h-1}{k} a_{p,\mathrm{up}}^{(t)} e^{a_{b,2}^{(t)}} + (h-1) a_{p,2}^{(t)} e^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + (h-1) e^{a_{b,1}^{(t)}} + (h-1) e^{a_{b,2}^{(t)}}},$$
$$\mu_{2h}^{(t)} := \frac{-\frac{a_{p,0}^{(t)}}{k} e^{a_{b,0}^{(t)}} + \frac{h-1}{k} a_{p,\mathrm{up}}^{(t)} e^{a_{b,2}^{(t)}} + h a_{p,2}^{(t)} e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1) e^{a_{b,2}^{(t)}}}, \tag{131}$$

and

$$r_{\mathrm{down}_\times}^{(t)} := \frac{e^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + (k-1) e^{a_{b,1}^{(t)}} + (k-1) e^{a_{b,2}^{(t)}}}. \tag{132}$$

We state the training dynamics of Phase 1.2 in Lemma 3.

Lemma 3. *Under balanced goal distribution (21), Assumption 1, for all $t, t' \in [T_{1/2}, T_1]$ with $t' \leqslant t$, there exists an absolute constant $C > 0$ such that we have*

$$e^{a_{b,1}^{(t)}} - e^{a_{b,1}^{(t')}} \geqslant 0, \tag{133}$$
$$a_{q,0}^{(t)} \asymp a_{q,\times}^{(t)} \asymp a_{p,2}^{(t)}, \tag{134}$$
$$e^{2a_{c,0}^{(t)}} - e^{2a_{c,0}^{(t')}} \geqslant 0, \quad e^{2a_{c,0}^{(t)}} \asymp \ln^2 k, \tag{135}$$
$$e^{2a_{c,1}^{(t)}} - e^{2a_{c,1}^{(t')}} \geqslant \max\left\{(2+o(1))\left(a_{q,\times}^{(t')} + a_{q,0}^{(t')}\right)\left(a_{q,\times}^{(t)} - a_{q,\times}^{(t')}\right), 0\right\}, \tag{136}$$
$$a_{p,1}^{(t)} - a_{p,1}^{(t')} \geqslant 0, \tag{137}$$
$$\left|a_{p,0}^{(t)} - a_{p,0}^{(t')}\right| \lesssim \left(a_{q,\times}^{(t)} + a_{q,0}^{(t)}\right) - \left(a_{q,\times}^{(t')} + a_{q,0}^{(t')}\right), \tag{138}$$
$$\left|a_{p,up}^{(t)} - a_{p,up}^{(t')}\right| \lesssim \left(a_{q,\times}^{(t)} + a_{q,0}^{(t)}\right) - \left(a_{q,\times}^{(t')} + a_{q,0}^{(t')}\right), \tag{139}$$
$$a_{p,2}^{(t)} = \left(1 + O\left(1/k^2\right)\right) \frac{a_{p,1}^{(t)}}{k-1}. \tag{140}$$

For $t, t' \in [T_{1/2}+1, T_{1,3}]$, $t' \leqslant t$, we have

$$a_{q,\times}^{(t)} - \frac{1}{2} a_{p,2}^{(t)} \geqslant 3.9 \ln k. \tag{141}$$
$$a_{q,0}^{(t)} \geqslant 1.9 \ln k - 2, \tag{142}$$
$$a_{p,1}^{(t)} - a_{p,1}^{(t')} \gtrsim \frac{\eta(t-t')}{k^{12.5}}, \tag{143}$$
$$a_{b,1}^{(t)} \asymp \frac{k^3 \ln^2 k}{N}, \tag{144}$$
$$-\frac{k^3 \ln^2 k}{N} \lesssim a_{b,0}^{(t)} \lesssim -\frac{k^2 \ln^2 k}{N}. \tag{145}$$

For $t \in [T_{1,3}+1, T_{1,4}]$, $t' \leqslant t$, we have

$$a_{p,1}^{(t)} \leqslant k^{9/4}\sqrt{N}, \tag{146}$$
$$\min\left\{a_{q,\times}^{(t)} - \mu_2^{(t)} - \ln k, a_{q,\times}^{(t)} - \mu_{2(k-1)}^{(t)} + \ln k\right\} + \frac{2}{k} a_{q,0}^{(t)} \geqslant r_{down_\times}^{(t)}\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right) + \ln k, \tag{147}$$
$$r_{down_\times}^{(t)}\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right) + 0.1\sqrt{a_{p,1}^{(t)}/k} + 5\ln k \geqslant \min\left\{a_{q,\times}^{(t)} - \mu_2^{(t)}, a_{q,\times}^{(t)} - \mu_{2(k-1)}^{(t)}\right\}, \tag{148}$$
$$r_{down_\times}^{(t)}\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right) + 2.1\ln k \geqslant \left(1 + \frac{1}{k}\right) a_{q,0}^{(t)} - \frac{a_{p,0}^{(t)}}{k}, \tag{149}$$

$$\left(1+\frac{1}{k}\right)a_{q,0}^{(t)}-\frac{a_{p,0}^{(t)}}{k}+0.1\sqrt{a_{p,1}^{(t)}/k}+3\ln k\geqslant\min\left\{a_{q,\times}^{(t)}-\mu_2^{(t)},a_{q,\times}^{(t)}-\mu_{2(k-1)}^{(t)}\right\},\tag{150}$$

$$e^{a_{b,1}^{(t)}}-1\gtrsim\frac{k\left(a_{p,1}^{(t)}\right)^2}{N}\gtrsim k^2\left(e^{a_{b,0}^{(t)}}-1\right),\tag{151}$$

$$e^{a_{b,0}^{(t)}}\gtrsim\frac{1}{\sqrt{k}},\tag{152}$$

$$e^{a_{b,1}^{(t)}}+\frac{1}{k^8N}a_{q,\times}^{(t)}-\left(e^{a_{b,1}^{(t')}}+\frac{1}{k^8N}a_{q,\times}^{(t')}\right)\geqslant\frac{1+o(1)}{N}\frac{ka_{p,1}^{(t')}}{3}\left(a_{p,1}^{(t)}-a_{p,1}^{(t')}\right),\tag{153}$$

$$e^{a_{b,1}^{(t)}}\lesssim\frac{k^{6.5}}{\ln k},\tag{154}$$

$$a_{p,1}^{(t)}-a_{p,1}^{(t')}\gtrsim\frac{\eta(t-t')}{\exp\left(k^{1.3}\sqrt{N}\right)},\tag{155}$$

For $t\in[T_{1,4}+1,T_1]$, we have

$$e^{a_{b,0}^{(t)}}\gtrsim\frac{k}{\ln k},\tag{156}$$

$$a_{p,1}^{(t)}\gtrsim\frac{k^2\sqrt{N}}{\ln k},\tag{157}$$

$$e^{a_{b,1}^{(t)}}\gtrsim\frac{k^5}{\ln^2 k},\tag{158}$$

$$\min\left\{a_{q,\times}^{(t)}-\mu_2^{(t)}-\ln k,a_{q,\times}^{(t)}-\mu_{2(k-1)}^{(t)}+\ln k\right\}+\frac{2}{k}a_{q,0}^{(t)}\geqslant r_{\mathit{dOWn}_\times}^{(t)}\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right)-0.1\sqrt{a_{p,1}^{(t)}/k}+\ln k,\tag{159}$$

$$\frac{1}{k-1}\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right)+2.1\ln k\geqslant\left(1+\frac{1}{k}\right)a_{q,0}^{(t)}-\frac{a_{p,0}^{(t)}}{k},\tag{160}$$

$$\left(1+\frac{1}{k}\right)a_{q,0}^{(t)}-\frac{a_{p,0}^{(t)}}{k}+0.1\sqrt{a_{p,1}^{(t)}/k}+3\ln k\geqslant\min\left\{a_{q,\times}^{(t)}-\mu_2^{(t)},a_{q,\times}^{(t)}-\mu_{2(k-1)}^{(t)}\right\},\tag{161}$$

$$e^{\frac{N}{N-1}a_{b,1}^{(t)}}\gtrsim\frac{k\left(a_{p,1}^{(t)}\right)^2}{N}\gtrsim k^2e^{\frac{N}{N-1}a_{b,0}^{(t)}},\tag{162}$$

$$\textit{If } a_{p,1}^{(t)}\geqslant Nk^3:\quad e^{a_{b,0}^{(t)}}\gtrsim\frac{\left(a_{p,1}^{(t)}\right)^{1-\frac{6}{N-1}}}{Nk},\tag{163}$$

$$(1+o(1))\left(a_{p,1}^{(t)}\right)^{-\frac{6}{N-1}}\leqslant e^{a_{b,2}^{(t)}}<1.\tag{164}$$

When $t=T_1$, we have

$$a_{p,1}^{(T_1)}\asymp k\ln(1/\epsilon),\tag{165}$$

$$\left(k-1+o\left(\frac{1}{\ln(1/\epsilon)}\right)\right)\ln\left(\frac{1}{k\epsilon^{1.05}}-(k-1)\right)\leqslant a_{p,1}^{(T_1)}+a_{p,2}^{(T_1)}\leqslant(k+C)\ln(k^2/\epsilon^{1.05}),\tag{166}$$

$$\left(1+\frac{1}{k}\right)a_{q,0}^{(T_1)}-\frac{a_{p,0}^{(T_1)}}{k}\geqslant\ln\left(\frac{1}{k\epsilon^{1.05}}-\frac{1}{k}\right)\tag{167}$$

$$\left(1+\frac{1}{k}\right)a_{q,\times}^{(T_1)}+\frac{a_{p,0}^{(T_1)}}{k}\geqslant\ln\left(\frac{(k-1)^2}{k\epsilon^{1.05}}-(k-1)\right)+\frac{C}{\ln(1/\epsilon)},\tag{168}$$

$$a_{c,1}^{(T_1)}\gtrsim\ln^2(1/\epsilon)\tag{169}$$

$$e^{a_{b,0}^{(T_1)}} \gtrsim \frac{1}{N}\left(\ln(1/\epsilon)\right)^{1-\frac{6}{N-1}}, \tag{170}$$

$$e^{a_{b,1}^{(T_1)}} \gtrsim \frac{k^2}{N^2}\left(\ln(1/\epsilon)\right)^{3-\frac{6}{N-1}} \gtrsim \frac{k^2}{N}\left(\ln(1/\epsilon)\right)^{1-\frac{5}{N}} e^{a_{b,0}^{(T_1)}}, \tag{171}$$

$$(1+o(1))\left(\ln(1/\epsilon)\right)^{-\frac{6}{N-1}} \lesssim e^{a_{b,2}^{(T_1)}} \leqslant \min\left\{1, (1+o(1))\left(\ln(1/\epsilon)\right)^{-\frac{4}{N-1}+\frac{12}{N(N-1)}}\right\}, \tag{172}$$

$$\forall 2 \leqslant l \leqslant L: \quad R\left(\pi^{(T_1)}; \mathcal{P}_l\right) \leqslant \frac{1+o(1)}{k^{l-1}}, \tag{173}$$

and

$$T_1 \lesssim \frac{1}{\eta\epsilon^{1.1}}. \tag{174}$$

The proof of Lemma 3 is postponed to Appendix C.3.5. The statement tracks the most stringent precision regime in Theorem 5. For a larger target precision ϵ, the depth-1 success criterion may already be met before the later Phase 1.2 thresholds, such as $T_{1,4}$, are reached. This is the regime reflected in Figure 4 and Figure 5: training can stop while $a_{b,0}$ is still slightly negative or close to zero, even though the analysis would make it positive after sufficiently long depth-1 training. The resulting policy still learns the same search mechanism. It simply realizes it with a slightly different finite-precision balance between A_B and A_P; in particular, during depth-2 curriculum training, A_P can converge to adjusted values, with $a_{p,\mathrm{up}}$ growing to the scale needed for backtracking in the search policy.

C.2.2 Depth-2 curriculum training

The main reason of poor depth generalization after depth-1 curriculum training is that $a_{p,2}$ is too large. From the proof of (173) in Lemma 3 we can see that this causes the agent not to go up with high probability at a non-leaf and non-root node, after all its children have been visited. We'll show that at depth-2 curriculum training, $a_{p,1}$ keeps monotonically increasing, $a_{p,2} = \frac{a_{p,1}-a_{p,3}}{k-1}$ (c.f. (83)) keeps decreasing until it has the right proportion to $a_{p,1}$. The changes of $a_{p,0}$ and $a_{p,\mathrm{up}}$ are relatively small during depth-2 curriculum training. Below we formally state the training dynamics of depth-2 curriculum training in Lemma 4. Recall we fix B, C, Q to be $B^{(T_1)}, C^{(T_1)}, Q^{(T_1)}$ after depth-1 curriculum training and only train $\overline{P}$ starting from $\overline{P}^{(T_1)}$.

Lemma 4. *Under balanced goal distribution (21), Assumption 1, for all $t, t' \in [T_1, T_2]$ with $t' \leqslant t$, we have*

$$|a_{p,0}^{(t)} - a_{p,0}^{(T_1)}| \lesssim \frac{a_{q,0}}{\left(\ln(1/\epsilon)\right)^{1+o(1)}}, \tag{175}$$

$$|a_{p,up}^{(t)} - a_{p,up}^{(T_1)}| \lesssim \frac{a_{q,0}}{\left(\ln(1/\epsilon)\right)^{3+o(1)}}, \tag{176}$$

$$a_{p,1}^{(t)} - a_{p,1}^{(t')} \gtrsim \frac{\epsilon^{1.1}}{k^L}(\eta(t-t')), \tag{177}$$

$$-\frac{a_{q,0}}{k} \lesssim a_{p,3}^{(t)} \lesssim k a_{q,0}, \tag{178}$$

$$a_{p,1}^{(t)} \asymp k a_{q,0}. \tag{179}$$

And there exists an absolute constant $C \in \mathbb{R}$ such that for

$$T_{2,1} := \inf_t \left\{ t \in [T_1+1, T_2] : a_{p,3}^{(t)} \geqslant \frac{a_{p,1}^{(t)}}{2} + k\left(1 + \frac{C}{\ln^2 k}\right) a_{q,0} \right\}, \tag{180}$$

for all $t, t' \in [T_1, T_{2,1}]$ with $t' \leqslant t$:

$$a_{p,3}^{(t)} - a_{p,3}^{(t')} \geqslant (1+o(1))k\left(a_{p,1}^{(t)} - a_{p,1}^{(t')}\right), \tag{181}$$

and for all $t \in [T_{2,1}+1, T_2]$:

$$a_{p,3}^{(t)} = \frac{a_{p,1}^{(t)}}{2} + k\left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}. \tag{182}$$

When $t = T_2$, there exists a constant $c_1, c_2, c_3, c_4 \in \mathbb{R}$ such that

$$\frac{a_{p,3}^{(T_2)}}{k} - \left(1 - \frac{c_1}{\ln^2 k}\right) a_{q,0} \geqslant \ln\left(\frac{k^L}{\epsilon^{1.02}}\right), \tag{183}$$

$$\frac{a_{p,1}^{(T_2)}}{k} \geqslant 2\ln\left(\frac{k^L}{\epsilon^{1.02}}\right) + \frac{c_2}{\ln^2 k} a_{q,0}, \tag{184}$$

$$\frac{a_{p,3}^{(T_2)}}{k} - \left(1 - \frac{c_3}{\ln^2 k}\right) a_{q,0} \leqslant \ln\left(\frac{k^L}{\epsilon^{1.02+o(1)}}\right), \tag{185}$$

$$\frac{a_{p,1}^{(T_2)}}{k} \leqslant 2\ln\left(\frac{k^L}{\epsilon^{1.02+o(1)}}\right) + \frac{c_4}{\ln^2 k} a_{q,0}. \tag{186}$$

Furthermore, we have

$$T_2 - T_1 \lesssim \frac{k^L}{\eta \epsilon^{1.1}}. \tag{187}$$

And for any full k-ary tree $\mathcal{T}$ (that may not be a perfect k-ary tree) with depth $l \in [L]$ and any goal node $g(\mathcal{T})$,

$$R(\pi^{(T_2)}, (\mathcal{T}, g(\mathcal{T}))) \geqslant 1 - \epsilon. \tag{188}$$

The proof of Lemma 4 is postponed to Appendix C.3.7. Theorem 5 follows directly from Lemma 2, Lemma 3 and Lemma 4.

C.3 Proof of key lemmas

C.3.1 Proof of Lemma 1

In the proof, for notation simplicity, we may drop superscript t when it's clear from the context.

Step 1: compute δB, δC, δP, δQ. We first compute δB, δC, δP, δQ by chain rule. Note that

$$d\phi_h = (dP) E_a^{(h)} \alpha_h^n,$$

and thus

$$d \log \pi_h(a_h) = \delta_h^\top d\phi_h = \delta_h^\top (dP) E_a^{(h)} \alpha_h^n = \mathrm{tr}\left(\delta_h (dP) E_a^{(h)} \alpha_h^n\right) = \mathrm{tr}\left((dP) E_a^{(h)} \alpha_h^n \delta_h^\top\right).$$

Therefore, we have

$$\nabla_P \log \pi_h(a_h) = \delta_h \left(E_a^{(h)} \alpha_h^n\right)^\top, \tag{189}$$

and thus by (18), we have

$$\delta P = \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h \delta_h \left(E_a^{(h)} \alpha_h^n\right)^\top\right], \tag{190}$$

where $\widehat{G}_h$ is defined in (65). By the same argument, we have

$$\delta Q = \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h \delta_h \left(E_c^{(h)} \alpha_h^c\right)^\top\right]. \tag{191}$$

Let the softmax Jacobians

$$J_h^n := \mathrm{diag}(\alpha_h^n) - \alpha_h^n (\alpha_h^n)^\top, \quad J_h^c := \mathrm{diag}(\alpha_h^c) - \alpha_h^c (\alpha_h^c)^\top, \tag{192}$$

then we have

$$d \log \pi_h(a_h) = \delta_h^\top d\phi_h = \delta_h^\top P E_a^{(h)} J_h^n \left(E_n^{(h)}\right)^\top (dB) \overline{n}_h = \mathrm{tr}\left(\overline{n}_h \delta_h^\top P E_a^{(h)} J_h^n\right.$$

$$\left(E_n^{(h)}\right)^\top (dB)\Big). \tag{193}$$

Therefore, we have

$$\nabla_B \log \pi_h(a_h) = E_n^{(h)} J_h^n \left(E_a^{(h)}\right)^\top P^\top \delta_h \overline{n}_h^\top, \tag{194}$$

and thus by (18), we have

$$\delta B = \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h E_n^{(h)} J_h^n \left(E_a^{(h)}\right)^\top P^\top \delta_h \overline{n}_h^\top\right]. \tag{195}$$

By the same argument, we have

$$\delta C = \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h E_c^{(h)} J_h^c \left(E_c^{(h)}\right)^\top Q^\top \delta_h \overline{c}_h^\top\right]. \tag{196}$$

Step 2: compute $\delta A_B, \delta A_C, \delta A_P, \delta A_Q$. We write the per-step selector factorizations

$$E_n^{(h)} = U S_n^{(h)}, \quad E_a^{(h)} = V S_a^{(h)}, \quad E_c^{(h)} = Z S_c^{(h)}, \tag{197}$$

where each column of $S_n^{(h)}$, $S_a^{(h)}$, $S_c^{(h)}$ is a one-hot selector vector that has 1 at the position of the corresponding node, action, or label and 0 otherwise. Then we have

$$\delta A_P = \delta P V \overset{(190)}{=} \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h \delta_h \left(E_a^{(h)} \alpha_h^n\right)^\top\right] V \overset{(197)}{=} \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h \delta_h \left(\alpha_h^n\right)^\top \left(S_a^{(h)}\right)^\top\right], \tag{198}$$

$$\delta A_Q = \delta Q Z \overset{(191)}{=} \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h \delta_h \left(E_c^{(h)} \alpha_h^c\right)^\top\right] Z \overset{(197)}{=} \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h \delta_h \left(\alpha_h^c\right)^\top \left(S_c^{(h)}\right)^\top\right], \tag{199}$$

and

$$\begin{aligned}\delta A_B = U^\top \delta B U &\overset{(195)}{=} U^\top \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h E_n^{(h)} J_h^n \left(E_a^{(h)}\right)^\top P^\top \delta_h \overline{n}_h^\top\right] U \\ &\overset{(197)}{=} \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h S_n^{(h)} J_h^n \left(S_a^{(h)}\right)^\top A_P^\top \delta_h e_{n_h}^\top\right],\end{aligned} \tag{200}$$

$$\begin{aligned}\delta A_C = Z^\top \delta C Z &\overset{(196)}{=} Z^\top \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h E_c^{(h)} J_h^c \left(E_c^{(h)}\right)^\top Q^\top \delta_h \overline{c}_h^\top\right] Z \\ &\overset{(197)}{=} \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h S_c^{(h)} J_h^c \left(S_c^{(h)}\right)^\top A_Q^\top \delta_h e_{c_h}^\top\right],\end{aligned} \tag{201}$$

where $e_{n_h} \in \mathbb{R}^{N+1}$ and $e_{c_h} \in \mathbb{R}^3$ are the one-hot vectors that have 1 at the position of n_h and c_h respectively and 0 otherwise.

We next simplify the matrix expressions and their gradients by exploiting the specific structure of the gradient matrices. To do so, we first introduce some notation. We define the (pre-softmax) logits as

$$\phi_h \coloneqq P E_a^{(h)} \alpha_h^n + Q E_c^{(h)} \alpha_h^c \tag{202}$$

with

$$\alpha_h^n \coloneqq \mathsf{softmax}\left(\left(E_n^{(h)}\right)^\top B \overline{n}_h\right), \quad \alpha_h^c \coloneqq \mathsf{softmax}\left(\left(E_c^{(h)}\right)^\top C \overline{c}_h\right). \tag{203}$$

Then by (10), we have

$$\pi_h = \mathsf{softmax}(\phi_h) = \mathsf{softmax}\left(P E_a^{(h)} \alpha_h^n + Q E_c^{(h)} \alpha_h^c\right). \tag{204}$$

We also define

$$\delta_h := e_{a_h} - \pi_h, \tag{205}$$

where e_{a_h} is the one-hot vector that has 1 at the position of a_h and 0 otherwise.

To better align with the node/action/label numbering, we let A_B, A_C be 0-indexed, A_P, A_Q's column indices be 0-indexed, and A_P, A_Q's row indices be 1-indexed, and we let $\times$ and ✓ denote index 1 and 2 respectively, e.g., $A_C(0, \times) = A_C(0, 1)$, $e_\times = e_1$. We also let up denote index $k + 1$.

For each $h \in [H]$, we define

$$p^c_{h,\times} := \sum_{i=1}^{h} \alpha^c_h(i) \mathbb{1}\{c_i = \times\}, \quad p^c_{h,0} := \sum_{i=1}^{h} \alpha^c_h(i) \mathbb{1}\{c_i = 0\}. \tag{206}$$

Then since $c_h \in \{0, \times\}$ for all $h \in [H]$, we have

$$\forall h \in [H]: \quad p^c_{h,\times} + p^c_{h,0} = 1. \tag{207}$$

Analogously, we define

$$\begin{aligned} p^a_{h,j} &:= \sum_{i=1}^{h} \alpha^n_h(i) \mathbb{1}\{a_{i-1} = j\}, \ \forall j \in \{0, 1, \cdots, k+1\}, \\ p^n_{h,j} &:= \sum_{i=1}^{h} \alpha^n_h(i) \mathbb{1}\{n_{i-1} = j\}, \ \forall j \in \{0, 1, \cdots, N\}. \end{aligned} \tag{208}$$

Then we have

$$\forall h \in [H]: \quad \sum_{j=0}^{k+1} p^a_{h,j} = 1, \quad \sum_{j=0}^{N} p^n_{h,j} = 1. \tag{209}$$

First, since once goal is found, i.e., ✓ is observed, the episode terminates and the gradient matrices are not updated anymore. Therefore, the input matrix $E^{(h)}_c$ won't contain any z_2 (the embedding of label ✓) for any $h \in [H]$. Thus by (201) and our zero initialization, we have

$$\forall t \in \mathbb{N}: \quad \delta A^{(t)}_C(\checkmark, :) = \delta A^{(t)}_C(:, \checkmark) = 0 \quad \text{and} \quad A^{(t)}_C(\checkmark, :) = A^{(t)}_C(:, \checkmark) = 0. \tag{210}$$

For the same reason, by (199), we have

$$\forall t \in \mathbb{N}: \quad \delta A^{(t)}_Q(:, \checkmark) = 0 \quad \text{and} \quad A^{(t)}_Q(:, \checkmark) = 0. \tag{211}$$

Further, since $n_h \neq 0$ for all $h \in [H]$, by (200), we have

$$\forall t \in \mathbb{N}: \quad \delta A^{(t)}_B(:, 0) = 0 \quad \text{and} \quad A^{(t)}_B(:, 0) = 0. \tag{212}$$

Moreover, by (201), we have

$$\begin{aligned} 1^\top \delta A_C &= \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h 1^\top S^{(h)}_c J^c_h \left(S^{(h)}_c\right)^\top A^\top_Q \delta_h e^\top_{c_h}\right] \\ &= \mathbb{E}\left[\sum_{h=1}^{H} \widehat{G}_h 1^\top J^c_h \left(S^{(h)}_c\right)^\top A^\top_Q \delta_h e^\top_{c_h}\right] = 0, \end{aligned} \tag{213}$$

where the first equality uses the fact that $1^\top S^{(h)}_c = 1^\top$ and the second equality uses the fact that $1^\top J^c_h = 0$. This combined with (210) and our zero initialization suggests that

$$\forall t \in \mathbb{N}, j \in \{0, 1\}: \quad \delta A^{(t)}_C(0, j) = -\delta A^{(t)}_C(1, j) \quad \text{and} \quad A^{(t)}_C(0, j) = -A^{(t)}_C(1, j). \tag{214}$$

(210) and (214) allow us to express $A^{(t)}_C$ and $\delta A^{(t)}_C$ as in (69). Similarly, we can further compute that each column of other 3 matrices also sum up to 0 using (200), (198), (199), i.e.:

$$1^\top \delta A_M = 0 \quad \text{for } M \in \{B, P, Q\}. \tag{215}$$

This and (213) combined with our zero initialization yields

$$\forall t \in \mathbb{N}: \quad 1^\top \delta A_M^{(t)} = 0, \quad \text{for } M \in \{B, C, P, Q\}; \quad 1^\top A_M^{(t)} = 0, \quad \text{for } M \in \{B, C, Q\}. \tag{216}$$

By exploiting the symmetry of the data distribution and using (216), we could represent A_P and A_Q as in (79) and (85), and we can also obtain (80).

We next compute $\delta a_{c,0}$ and $\delta a_{c,1}$ in (69). By (201) and our notation defined at the beginning of Appendix C, we have

$$\begin{aligned}
&\delta a_{c,1} \\
&= e_1^\top \delta A_C e_1 \\
&= \mathbb{E}\Bigg[\sum_{h=1}^{H} \left((\mathbb{1}\{c_1 = \times\}, \cdots, \mathbb{1}\{c_h = \times\})\right) \left(\operatorname{diag}\{\alpha_h^c\} - \alpha_h^c (\alpha_h^c)^\top\right) \\
&\quad \left(S_c^{(h)}\right)^\top A_Q^\top (e_{a_h} - \pi_h) \mathbb{1}\{c_h = \times\} \widehat{G}_h \Bigg] \\
&= \mathbb{E}\Bigg[\sum_{h=1}^{H} \Bigg\{ \left(\sum_{i=1}^{h} \alpha_h^c(i) \mathbb{1}\{c_i = \times\}\right) A_Q(:, \times)^\top - \left(\sum_{i=1}^{h} \alpha_h^c(i) \mathbb{1}\{c_i = \times\}\right) \\
&\qquad \cdot \left(\sum_{i=1}^{h} \alpha_h^c(i) \mathbb{1}\{c_i = \times\} A_Q(:, \times)^\top + \sum_{i=1}^{h} \alpha_h^c(i) \mathbb{1}\{c_i = 0\} A_Q(:, 0)^\top\right) \\
&\quad \Bigg\} (e_{a_h} - \pi_h) \mathbb{1}\{c_h = \times\} \widehat{G}_h \Bigg] \\
&= \mathbb{E}\Bigg[\sum_{h=1}^{H} \left(\sum_{i=1}^{h} \alpha_h^c(i) \mathbb{1}\{c_i = \times\}\right) \left(\sum_{i=1}^{h} \alpha_h^c(i) \mathbb{1}\{c_i = 0\}\right) (A_Q(:, \times) \\
&\quad - A_Q(:, 0))^\top (e_{a_h} - \pi_h) \mathbb{1}\{c_h = \times\} \widehat{G}_h \Bigg].
\end{aligned} \tag{217}$$

By (217) and (206), we have

$$\begin{aligned}
\delta a_{c,1} &= \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = \times\} p_{h,\times}^c p_{h,0}^c \left(A_Q(:, \times) - A_Q(:, 0)\right)^\top (e_{a_h} - \pi_h) \widehat{G}_h\right] \\
&= \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = \times\} p_{h,\times}^c p_{h,0}^c \left(A_Q(:, \times) - A_Q(:, 0)\right)^\top (\pi_h \odot (Q_h - G_h \mathbf{1}))\right].
\end{aligned} \tag{218}$$

Similarly, we can compute that

$$\begin{aligned}
&\delta a_{c,0} \\
&= \mathbb{E}\Bigg[\sum_{h=1}^{H} \left(\sum_{i=1}^{h} \alpha_h^c(i) \mathbb{1}\{c_i = 0\}\right) \left(\sum_{i=1}^{h} \alpha_h^c(i) \mathbb{1}\{c_i = \times\}\right) (A_Q(:, 0) \\
&\quad - A_Q(:, \times))^\top (e_{a_h} - \pi_h) \mathbb{1}\{c_h = 0\} \widehat{G}_h \Bigg] \\
&= \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0\} p_{h,0}^c p_{h,\times}^c \left(A_Q(:, 0) - A_Q(:, \times)\right)^\top (e_{a_h} - \pi_h) \widehat{G}_h\right] \\
&= \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0\} p_{h,0}^c p_{h,\times}^c \left(A_Q(:, 0) - A_Q(:, \times)\right)^\top (\pi_h \odot (Q_h - G_h \mathbf{1}))\right].
\end{aligned} \tag{219}$$

From (218), (219) and (85) we further deduce

$$\delta a_{c,0}$$

$$= \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0\} p_{h,0}^c p_{h,\times}^c \left(A_Q(:,0) - A_Q(:,\times)\right)^\top \left(\pi_h \odot (Q_h - G_h \mathbf{1})\right)\right]$$
$$= \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0\} p_{h,0}^c p_{h,\times}^c \left(\frac{a_{q,\times} + a_{q,0}}{k} \sum_{i=1}^{k} \pi_h(i)(Q_h(i) - G_h) - (a_{q,\times} + a_{q,0})\pi_h(\mathtt{up})(Q_h(\mathtt{up}) - G_h)\right)\right]$$
$$= \left(1 + \frac{1}{k}\right) \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0\} p_{h,0}^c p_{h,\times}^c \left(a_{q,\times} + a_{q,0}\right) \pi_h(\mathtt{up})(G_h - Q_h(\mathtt{up}))\right], \tag{220}$$

and

$$\delta a_{c,1} = \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = \times\} p_{h,\times}^c p_{h,0}^c \left(A_Q(:,\times) - A_Q(:,0)\right)^\top \left(\pi_h \odot (Q_h - G_h \mathbf{1})\right)\right]$$
$$= \left(1 + \frac{1}{k}\right) \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = \times\} p_{h,0}^c p_{h,\times}^c \left(a_{q,\times} + a_{q,0}\right) \pi_h(\mathtt{up})(Q_h(\mathtt{up}) - G_h)\right]$$
$$= \left(1 + \frac{1}{k}\right) \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = \times\} p_{h,0}^c p_{h,\times}^c \left(a_{q,\times} + a_{q,0}\right) \pi_h(\mathtt{up})(1 - \pi_h(\mathtt{up}))Q_h(\mathtt{up})\right], \tag{221}$$

where the last relation uses the fact that $Q_h(\mathtt{down}_i) = 0$ when $c_h = \times$. Further note that

$$p_{h,\times}^c = \begin{cases} \frac{N_{h,\times} e^{a_{c,1}}}{N_{h,\times} e^{a_{c,1}} + (h - N_{h,\times}) e^{-a_{c,1}}}, & \text{if } c_h = \times, \\ \frac{N_{h,\times} e^{-a_{c,0}}}{N_{h,\times} e^{-a_{c,0}} + (h - N_{h,\times}) e^{a_{c,0}}}, & \text{if } c_h = 0, \end{cases} \quad , \quad p_{h,0}^c$$
$$= \begin{cases} \frac{(h - N_{h,\times}) e^{-a_{c,1}}}{N_{h,\times} e^{a_{c,1}} + (h - N_{h,\times}) e^{-a_{c,1}}}, & \text{if } c_h = \times, \\ \frac{(h - N_{h,\times}) e^{a_{c,0}}}{N_{h,\times} e^{-a_{c,0}} + (h - N_{h,\times}) e^{a_{c,0}}}, & \text{if } c_h = 0. \end{cases} \tag{222}$$

Plugging (222) into the above two equations about $\delta a_{c,0}$ and $\delta a_{c,1}$, we get (70) and (71).

Analogously, we can compute that for any $i, j \in \{0, \cdots, N\}$,

$$\delta A_B(i,j)$$
$$= \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{n_h = j\} \sum_{m=1}^{h} \alpha_h^n(m) \left(\mathbb{1}\{n_{m-1} = i\} - \sum_{l=1}^{h} \alpha_h^n(l) \mathbb{1}\{n_{l-1} = i\}\right) A_P(:, a_{m-1})^\top (e_{a_h} - \pi_h) \widehat{G}_h\right]. \tag{223}$$

By the above expression, the symmetry of the node distribution, our zero initialization, (215), (212) and (208), we immediately obtain (72), in which

$$\delta a_{b,0} = \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{n_h = j\} \sum_{m=1}^{h} \alpha_h^n(m) \left(\mathbb{1}\{n_{m-1} = 0\} - p_{h,0}^n\right) A_P(:, a_{m-1})^\top \left(\pi_h \odot (Q_h - G_h \mathbf{1})\right)\right], \tag{224}$$

$$\delta a_{b,1} = \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{n_h = j\} \sum_{m=1}^{h} \alpha_h^n(m) \left(\mathbb{1}\{n_{m-1} = j\} - p_{h,j}^n\right) A_P(:, a_{m-1})^\top \left(\pi_h \odot (Q_h - G_h \mathbf{1})\right)\right], \tag{225}$$

$$\delta a_{b,2} = \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{n_h = j\} \sum_{m=1}^{h} \alpha_h^n(m) \left(\mathbb{1}\{n_{m-1} = i\} - p_{h,i}^n\right) A_P(:, a_{m-1})^\top \left(\pi_h \odot (Q_h - G_h \mathbf{1})\right)\right],$$
$$= \frac{-\delta a_{b,0} - \delta a_{b,1}}{N - 1} \tag{226}$$

for any $i, j \in [N]$, $i \neq j$. (226) and (216) yield (75). And we can further simplify $\delta a_{b,0}$ as follows: Define

$$\begin{aligned}
\beta_{h,0} &:= A_P(:,0)^\top \left(\pi_h \odot (Q_h - G_h \mathbf{1})\right), \\
\beta_{h,1} &:= \sum_{i \in \mathcal{A}_{h,\times}} A_P(:,i)^\top \left(\pi_h \odot (Q_h - G_h \mathbf{1})\right), \\
\beta_{h,2} &:= \sum_{i=1}^{k} \left(N_h(i) - \mathbb{1}\{i \in \mathcal{A}_{h,\times}\}\right) A_P(:,i)^\top \left(\pi_h \odot (Q_h - G_h \mathbf{1})\right) \\
&\quad + N_h(\mathtt{up}) A_P(:,\mathtt{up})^\top \left(\pi_h \odot (Q_h - G_h \mathbf{1})\right).
\end{aligned}$$

Then

$$\begin{aligned}
\delta a_{b,0} \overset{(224)}{=} \frac{1}{N} \mathbb{E}\Bigg[\sum_{h=1}^{H} \Bigg\{ \frac{e^{a_{b,0}}}{S_h} \left(1 - \frac{e^{a_{b,0}}}{S_h}\right) \beta_{h,0} \\
- \frac{e^{a_{b,0}}}{S_h} \frac{e^{a_{b,1}}}{S_h} \beta_{h,1} - \frac{e^{a_{b,0}}}{S_h} \frac{e^{a_{b,2}}}{S_h} \beta_{h,2} \Bigg\} \Bigg],
\end{aligned} \tag{227}$$

where

$$S_h := e^{a_{b,0}} + |\mathcal{A}_{h,\times}| e^{a_{b,1}} + (h - 1 - |\mathcal{A}_{h,\times}|) e^{a_{b,2}}.$$

Similarly, we can further simplify $\delta a_{b,1}$ as follows:

$$\begin{aligned}
\delta a_{b,1} \overset{(225)}{=} \frac{1}{N} \mathbb{E}\Bigg[\sum_{h=1}^{H} \Bigg\{ \frac{e^{a_{b,1}}}{S_h} \left(1 - \frac{|\mathcal{A}_{h,\times}| e^{a_{b,1}}}{S_h}\right) \beta_{h,1} - \frac{|\mathcal{A}_{h,\times}| e^{a_{b,1}}}{S_h} \frac{e^{a_{b,0}}}{S_h} \beta_{h,0} \\
- \frac{|\mathcal{A}_{h,\times}| e^{a_{b,1}}}{S_h} \frac{e^{a_{b,2}}}{S_h} \beta_{h,2} \Bigg\} \Bigg].
\end{aligned} \tag{228}$$

By (79) we have

$$A_P(:,0)^\top \left(\pi_h \odot (Q_h - G_h \mathbf{1})\right) = \frac{1}{k} a_{p,0} \pi_h(\mathtt{up}) \left(Q_h(\mathtt{up}) - G_h\right), \tag{229}$$

$$A_P(:,\mathtt{up})^\top \left(\pi_h \odot (Q_h - G_h \mathbf{1})\right) = -\frac{1}{k} a_{p,\mathtt{up}} \pi_h(\mathtt{up}) \left(Q_h(\mathtt{up}) - G_h\right), \tag{230}$$

$$\begin{aligned}
\forall i \in [k]: \ A_P(:,i)^\top \left(\pi_h \odot (Q_h - G_h \mathbf{1})\right) &= -\left(a_{p,1} + a_{p,2}\right) \pi_h(i) \left(Q_h(i) - G_h\right) \\
&\quad - a_{p,2} \pi_h(\mathtt{up}) (Q_h(\mathtt{up}) - G_h).
\end{aligned} \tag{231}$$

Plugging them into the expressions of $\beta_{h,0}$, $\beta_{h,1}$ and $\beta_{h,2}$ and organizing the terms, we get (73), (74), (76), (77) and (78).

We next compute $\delta a_{p,i}$ for $i \in \{0, 1, 3, \mathtt{up}\}$ and $\delta a_{q,j}$ for $j \in \{0, \times\}$. First, by (198), we have for all $i \in [k+1]$, $j \in \{0, 1, \cdots, k+1\}$:

$$\begin{aligned}
\delta A_P(i,j) &= \mathbb{E}\left[\sum_{h=1}^{H} \left(\mathbb{1}\{a_h = i\} - \pi_h(i)\right) \left(\sum_{m=1}^{h} \alpha_h^n(m) \mathbb{1}\{a_{m-1} = j\} \right) \widehat{G}_h \right] \\
&= \mathbb{E}\Bigg[\sum_{h=1}^{H} \left(\sum_{m=1}^{h} \alpha_h^n(m) \mathbb{1}\{a_{m-1} = j\} \right) \Bigg(\pi_h(i) \left(1 - \pi_h(i)\right) Q_h(i) - \sum_{a \neq i} \pi_h(a) \\
&\quad \pi_h(i) Q_h(a) \Bigg) \Bigg] \\
&= \mathbb{E}\left[\sum_{h=1}^{H} \pi_h(i) \left(\sum_{m=1}^{h} \alpha_h^n(m) \mathbb{1}\{a_{m-1} = j\} \right) \left(Q_h(i) - G_h\right) \right].
\end{aligned} \tag{232}$$

This together with (208) gives

$$\delta A_P(:,j) = \mathbb{E}\left[\sum_{h=1}^{H} p_{h,j}^a \left(\pi_h \odot (Q_h - G_h \mathbf{1})\right) \right], \ \forall j \in \{0, 1, \cdots, k+1\}. \tag{233}$$

By a similar argument, we can compute that for any $i \in [k+1]$, $j \in \{0, \times\}$,

$$\begin{aligned}\delta A_Q(i,j) &= \mathbb{E}\left[\sum_{h=1}^{H}\left(\mathbb{1}\{a_h = i\} - \pi_h(i)\right)\left(\sum_{m=1}^{h}\alpha_h^c(m)\mathbb{1}\{c_m = j\}\right)\widehat{G}_h\right] \\ &= \mathbb{E}\left[\sum_{h=1}^{H}\pi_h(i)\left(\sum_{m=1}^{h}\alpha_h^c(m)\mathbb{1}\{c_m = j\}\right)(Q_h(i) - G_h)\right] \\ &\overset{(207)}{=} \mathbb{E}\left[\sum_{h=1}^{H}\pi_h(i)p_{h,j}^c\,(Q_h(i) - G_h)\right]. \end{aligned} \tag{234}$$

This together with (211) gives

$$\delta A_Q(:,j) = \begin{cases}\mathbb{E}\left[\sum_{h=1}^{H}p_{h,j}^c\,(\pi_h \odot (Q_h - G_h\mathbf{1}))\right], & j \in \{0,\times\}, \\ 0, & \text{otherwise.}\end{cases} \tag{235}$$

By (233) and (79) we have

$$\begin{aligned}\delta a_{p,0} = \delta A_P(\mathtt{up}, 0) = \mathbb{E}\Bigg[&\sum_{h=1}^{H} \\ &\pi_h(\mathtt{up})\frac{e^{a_{b,0}}}{e^{a_{b,0}} + |\mathcal{A}_{h,\times}|e^{a_{b,1}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}}}\,(Q_h(\mathtt{up}) - G_h)\Bigg],\end{aligned} \tag{236}$$

which yields (81),

$$\begin{aligned}\delta a_{p,\mathtt{up}} = -\delta A_P(\mathtt{up},\mathtt{up}) = -\mathbb{E}\Bigg[&\sum_{h=1}^{H} \\ &\pi_h(\mathtt{up})\frac{N_h(\mathtt{up})e^{a_{b,2}}}{e^{a_{b,0}} + |\mathcal{A}_{h,\times}|e^{a_{b,1}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}}}\,(Q_h(\mathtt{up}) - G_h)\Bigg],\end{aligned} \tag{237}$$

which yields (84),

$$\begin{aligned}\delta a_{p,1} = -\delta A_P(j,j) = -\mathbb{E}\Bigg[&\sum_{h=1}^{H}\pi_h(j) \\ &\Bigg(\mathbb{1}\{j \in \mathcal{A}_{h,\times}\}\frac{e^{a_{b,1}} + (N_h(j)-1)e^{a_{b,2}}}{e^{a_{b,0}} + |\mathcal{A}_{h,\times}|e^{a_{b,1}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}}}(-G_h) \\ &+ \mathbb{1}\{j \in [k]\backslash\mathcal{A}_{h,\times}\}\frac{N_h(j)e^{a_{b,2}}}{e^{a_{b,0}} + |\mathcal{A}_{h,\times}|e^{a_{b,1}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}}}\,(Q_h(j) \\ &-G_h)\Bigg)\Bigg],\end{aligned} \tag{238}$$

where $j \in [k]$. This gives (82). And similarly, we have

$$\begin{aligned}\delta a_{p,3} = \delta A_P(\mathtt{up}, j) = \mathbb{E}\Bigg[&\sum_{h=1}^{H}\pi_h(\mathtt{up}) \\ &\Bigg(\mathbb{1}\{j \in \mathcal{A}_{h,\times}\}\frac{e^{a_{b,1}} + (N_h(j)-1)e^{a_{b,2}}}{e^{a_{b,0}} + |\mathcal{A}_{h,\times}|e^{a_{b,1}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}}} \\ &+ \mathbb{1}\{j \in [k]\backslash\mathcal{A}_{h,\times}\}\frac{N_h(j)e^{a_{b,2}}}{e^{a_{b,0}} + |\mathcal{A}_{h,\times}|e^{a_{b,1}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}}}\Bigg)(Q_h(\mathtt{up}) \\ &- G_h)\Bigg],\end{aligned} \tag{239}$$

where $j \in [k]$. This gives (83). Lastly, combining (235) and (222), we have (86) and (87).

Policy computation. We let

$$\varphi_h \coloneqq A_P(:,0)e^{a_{b,0}} + \sum_{j\in\mathcal{A}_{h,\times}} A_P(:,j)e^{a_{b,1}} + \left(N_h(\texttt{up})A_P(:,\texttt{up}) + \sum_{i=1}^{k}\left(N_h(i) - \mathbb{1}\{i\in\mathcal{A}_{h,\times}\}\right)A_P(:,i)\right)e^{a_{b,2}}. \tag{240}$$

Then by (10), under Assumption 1 and the notation defined at the beginning of Appendix C, we have

$$\pi_h = \text{softmax}\Bigg(\frac{\varphi_h}{e^{a_{b,0}} + |\mathcal{A}_{h,\times}|e^{a_{b,1}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}}} + \begin{cases} \frac{N_{h,\times}e^{-a_{c,0}}A_Q(:,\times)+(h-N_{h,\times})e^{a_{c,0}}A_Q(:,0)}{N_{h,\times}e^{-a_{c,0}}+(h-N_{h,\times})e^{a_{c,0}}}, & \text{if } c_h = 0 \\ \frac{N_{h,\times}e^{a_{c,1}}A_Q(:,\times)+(h-N_{h,\times})e^{-a_{c,1}}A_Q(:,0)}{N_{h,\times}e^{a_{c,1}}+(h-N_{h,\times})e^{-a_{c,1}}}, & \text{if } c_h = \times \end{cases}\Bigg). \tag{241}$$

By (241) and (79), we have

$$\varphi_h(\texttt{up}) = 0, \tag{242}$$

$$\forall m\in[k]: \quad \varphi_h(m) = -\frac{a_{p,0}}{k}e^{a_{b,0}} + |\mathcal{A}_{h,\times}|a_{p,2}e^{a_{b,1}} - \mathbb{1}\{m\in\mathcal{A}_{h,\times}\}\left((a_{p,1}+a_{p,2})(e^{a_{b,1}} - e^{a_{b,2}})\right) + \left(N_h(\texttt{up})\frac{a_{p,\texttt{up}}}{k} - N_h(m)a_{p,1} + (h-1-N_h(\texttt{up})-N_h(m)-|\mathcal{A}_{h,\times}|)a_{p,2}\right)e^{a_{b,2}}. \tag{243}$$

The above two expressions give (89). (79) together with (241) gives (90).

C.3.2 Proof of Lemma 2

We show by induction. First, by our zero initialization, Lemma 2 holds for $t = 0$. We make the following induction hypothesis:

Induction Hypothesis 1: For all $t \leqslant s-1$ ($s\in[T_{1/2}]$), the relations in Lemma 2 hold.

Below we show Lemma 2 holds for $t = s$ under Induction Hypothesis 1.To do so, we first simplify the gradients for $l = 1$.

Notation. We define events

$$\mathcal{E}_{1,h} \coloneqq \{\text{the environment has not terminated at step } h, \text{ and the tree depth } l = 1\}. \tag{244}$$

Recall we define $\widehat{G}_h$ in (65). We let $V^{\pi}_{1,h}$ denote the expectation of $\widehat{G}_h$ conditioned on $\mathcal{E}_{1,h}$ under policy π:

$$V^{\pi}_{1,h} \coloneqq \mathbb{E}^{\pi}[\widehat{G}_h|\mathcal{E}_{1,h}]. \tag{245}$$

We let $V^{(t)}_{1,h} \coloneqq V^{\pi^{(t)}}_{1,h}$. Note that $V^{(t)}_{1,h}$ is the same for any possible s_h conditioned on $\mathcal{E}_{1,h}$ and $\pi^{(t)}_h$ by symmetry under the balanced case. We let $\pi^{(t)}_{1,h}(i)$ denote $\pi^{(t)}_h(i)$ conditioned on $\mathcal{E}_{1,h}$ and $\pi^{(t)}_h$ for any $i\in[k+1]$. Then by symmetry, for all odd steps $2h-1$ ($h\in[k]$) (where the agent is at the root node when $l=1$), we have $\pi^{(t)}_{1,2h-1}(i)$ are equal for any $i\in[k]\backslash\mathcal{A}_{2h-1,\times}$ for any $t\in\mathbb{N}$. We let $\pi^{(t)}_{1,2h-1}(\texttt{down}) \coloneqq \pi^{(t)}_{1,2h-1}(i)$ for any $i\in[k]\backslash\mathcal{A}_{2h-1,\times}$. $\pi^{(t)}_{1,2h-1}(i)$ are also equal for any $i\in\mathcal{A}_{2h-1,\times}$, and we let $\pi^{(t)}_{1,2h-1}(\texttt{down}_\times) \coloneqq \pi^{(t)}_{1,2h-1}(i)$ for any $i\in\mathcal{A}_{2h-1,\times}$. We also let

$$\pi^{(t)}_{1,2h}(\texttt{down}^0_\times) \coloneqq \pi^{(t)}_{1,2h}(i) \quad \text{for any } i\in[k]\backslash(\mathcal{A}_{2h-1,\times}\cup\{a_{2h-1}\})$$

and

$$\pi^{(t)}_{1,2h}(\texttt{down}^1_\times) \coloneqq \pi^{(t)}_{1,2h}(i) \quad \text{for any } i\in\mathcal{A}_{2h-1,\times}\cup\{a_{2h-1}\},$$

whose values are also independent of s_{2h} conditioned on $\mathcal{E}_{1,2h}$. We let $\mathbb{P}^{\pi}(\mathcal{E}_{1,h})$ denote the probability of event $\mathcal{E}_{1,h}$ happening under policy π. We let $\mathbb{P}^{(t)}(\mathcal{E}_{1,h}) \coloneqq \mathbb{P}^{\pi^{(t)}}(\mathcal{E}_{1,h})$ for notation simplicity.

Gradients/policy simplification for phase 1.1. By Induction Hypothesis 1, using the above notation we define, we could further simplify the gradients and policy expressions in Lemma 1 when $l = 1$. We summarize the results in the following lemma.

Lemma 5. *Under Induction Hypothesis 1, we have for any* $0 \leqslant t \leqslant s-1$*:*

$$\delta a_{q,0}^{(t)} = \pi_{1,1}^{(t)}(up)V_{1,1}^{(t)} + \sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(up)\Bigg(\frac{1}{1+\frac{h}{h+1}e^{-2a_{c,0}^{(t)}}}\pi_{1,2h+1}^{(t)}(up) - \frac{1}{1+e^{2a_{c,1}^{(t)}}}(1-\pi_{1,2h}^{(t)}(up))\Bigg)\Bigg)V_{1,2h+1}^{(t)}, \tag{246}$$

$$\delta a_{q,\times}^{(t)} = \sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(up)\Bigg(\frac{1}{1+e^{-2a_{c,1}^{(t)}}}\left(1-\pi_{1,2h}^{(t)}(up)\right) - \frac{1}{1+\frac{h+1}{h}e^{2a_{c,0}^{(t)}}}\pi_{1,2h+1}^{(t)}(up)\Bigg)V_{1,2h+1}^{(t)}, \tag{247}$$

$$\delta a_{c,0}^{(t)} = \left(1+\frac{1}{k}\right)\sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(up)\frac{h(h+1)}{\left(he^{-a_{c,0}^{(t)}}+(h+1)e^{a_{c,0}^{(t)}}\right)^2}\left(a_{q,\times}^{(t)} + a_{q,0}^{(t)}\right)\pi_{1,2h+1}^{(t)}(up)V_{1,2h+1}^{(t)}, \tag{248}$$

$$\delta a_{c,1}^{(t)} = \left(1+\frac{1}{k}\right)\sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(up)\frac{1}{\left(e^{a_{c,1}^{(t)}}+e^{-a_{c,1}^{(t)}}\right)^2}\left(a_{q,\times}^{(t)} + a_{q,0}^{(t)}\right)(1-\pi_{1,2h}^{(t)}(up))V_{1,2h+1}^{(t)}, \tag{249}$$

$$\delta a_{p,0}^{(t)} = (1+o(1))\left(-\pi_{1,1}^{(t)}(up)V_{1,1}^{(t)} + \sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(up)\left(\frac{1-\pi_{1,2h}^{(t)}(up)}{2h} - \frac{\pi_{1,2h+1}^{(t)}(up)}{2h+1}\right)V_{1,2h+1}^{(t)}\right), \tag{250}$$

$$\delta a_{p,1}^{(t)} = (1+o(1))\frac{1}{k}\sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(up)\left(\frac{1}{2}\pi_{1,2h}^{(t)}(\texttt{down}_\times^1) + \frac{h}{2h+1}\pi_{1,2h+1}^{(t)}(\texttt{down}_\times)\right)V_{1,2h+1}^{(t)}, \tag{251}$$

$$\delta a_{p,3}^{(t)} = (1+o(1))\frac{1}{k}\sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(up)\left(\frac{1}{2}\left(1-\pi_{1,2h}^{(t)}(up)\right) - \frac{h}{2h+1}\pi_{1,2h+1}^{(t)}(up)\right)V_{1,2h+1}^{(t)}, \tag{252}$$

$$\delta a_{p,up}^{(t)} = -(1+o(1))\sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(up)\left(\frac{h-1}{2h}\left(1-\pi_{1,2h}^{(t)}(up)\right) - \frac{h}{2h+1}\pi_{1,2h+1}^{(t)}(up)\right)V_{1,2h+1}^{(t)}, \tag{253}$$

$$\delta a_{b,0}^{(t)} = \frac{1+o(1)}{N} \sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(t)}(up) I_{h,0}^{(t)} V_{1,2h+1}^{(t)}, \tag{254}$$

where

$$I_{h,0}^{(t)} := \frac{1}{4h^2}\left(\left(\frac{2h-1}{k}a_{p,0}^{(t)} + h a_{p,2}^{(t)} + \frac{h-1}{k}a_{p,up}^{(t)}\right)\left(1-\pi_{1,2h}^{(t)}(up)\right) - \left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right) h \pi_{1,2h}^{(t)}(down_{\times}^{1})\right) - \frac{h}{(2h+1)^2}\left(\left(\frac{1}{k}\left(2a_{p,0}^{(t)} + a_{p,up}^{(t)}\right) + a_{p,2}^{(t)}\right)\pi_{1,2h+1}^{(t)}(up) + \left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right) \pi_{1,2h+1}^{(t)}(down_{\times})\right), \tag{255}$$

$$\delta a_{b,1}^{(t)} = \frac{1+o(1)}{N} \sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(t)}(up) I_{h,1}^{(t)} V_{1,2h+1}^{(t)}, \tag{256}$$

where

$$I_{h,1}^{(t)} := \frac{h}{(2h+1)^2}\left[(h+1)\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right)\pi_{1,2h+1}^{(t)}(down_{\times}) + \left(\frac{1}{k}\left(a_{p,0}^{(t)} - h a_{p,up}^{(t)}\right) + (h+1)a_{p,2}^{(t)}\right)\pi_{1,2h+1}^{(t)}(up)\right]. \tag{257}$$

For all $\forall h \in [k]$, $i \in [k]$, *we have*

$$\pi_{1,2h-1}^{(t)}(up) = (1 + o(1)) \frac{\exp\left(\xi_{2h-1}^{(t)} - \frac{h-1}{2h-1}a_{p,2}^{(t)}\right)}{(h-1)\exp\left(-\frac{a_{p,1}^{(t)}+a_{p,2}^{(t)}}{2h-1}\right) + k - h + 1 + \exp\left(\xi_{2h-1}^{(t)} - \frac{h-1}{2h-1}a_{p,2}^{(t)}\right)}, \tag{258}$$

$$\pi_{1,2h}^{(t)}(up) = (1+o(1)) \frac{\exp\left(\xi_{2h}^{(t)} - \frac{1}{2}a_{p,2}^{(t)}\right)}{h\exp\left(-\frac{a_{p,1}^{(t)}+a_{p,2}^{(t)}}{2h}\right) + k - h + \exp\left(\xi_{2h}^{(t)} - \frac{1}{2}a_{p,2}^{(t)}\right)}, \tag{259}$$

$$\pi_{1,2h-1}^{(t)}(i) = (1 + o(1)) \frac{\exp\left(-\frac{N_{2h-1}(i)}{2h-1}\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right)\right)}{(h-1)\exp\left(-\frac{a_{p,1}^{(t)}+a_{p,2}^{(t)}}{2h-1}\right) + k - h + 1 + \exp\left(\xi_{2h-1}^{(t)} - \frac{h-1}{2h-1}a_{p,2}^{(t)}\right)}, \tag{260}$$

$$\pi_{1,2h}^{(t)}(i) = (1+o(1)) \frac{\exp\left(-\frac{N_{2h}(i)}{2h}\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right)\right)}{h\exp\left(-\frac{a_{p,1}^{(t)}+a_{p,2}^{(t)}}{2h}\right) + k - h + \exp\left(\xi_{2h}^{(t)} - \frac{1}{2}a_{p,2}^{(t)}\right)}, \tag{261}$$

where

$$\xi_{2h-1}^{(t)} := -\frac{a_{q,0}^{(t)} - \frac{h-1}{h}e^{-2a_{c,0}^{(t)}}a_{q,\times}^{(t)}}{1 + \frac{h-1}{h}e^{-2a_{c,0}^{(t)}}}, \quad \xi_{2h}^{(t)} := \frac{a_{q,\times}^{(t)} - e^{-2a_{c,1}^{(t)}}a_{q,0}^{(t)}}{1 + e^{-2a_{c,1}^{(t)}}}. \tag{262}$$

The proof of Lemma 5 is postponed to Appendix C.3.3.

and define the ratios

$$\nu_h^{(t)} := \frac{\pi_{1,2h-1}^{(t)}(\mathrm{up})}{\pi_{1,2h}^{(t)}(\mathrm{up})}, \quad r_h^{(t)} := \frac{\exp\left(\xi_{2h-1}^{(t)} - \frac{h-1}{2h-1}a_{p,2}^{(t)}\right)}{\exp\left(\xi_{2h}^{(t)} - \frac{1}{2}a_{p,2}^{(t)}\right)}. \tag{263}$$

We next divide phase 1.1 into two cases: (i) $s-1 \leqslant T_{1,1}$ and (ii) $T_{1,1}+1 \leqslant s-1 \leqslant T_{1/2}$.

Case 1: $s-1 \leqslant T_{1,1}$**.** (251) and Induction Hypothesis 1 imply that (101) holds at step $s-1$. By Induction Hypothesis 1, we can simplify the policy expressions (258), (259), (260), (261) when $s-1 \leqslant T_{1,1}$ as follows:

$$\forall h \in [k], i \in [k]: \quad \pi_{1,2h-1}^{(s-1)}(\mathtt{up}) = (1+o(1)) \frac{\exp\left(\xi_{2h-1}^{(s-1)}\right)}{k+\exp\left(\xi_{2h-1}^{(s-1)}\right)}, \tag{264}$$

$$\pi_{1,2h}^{(s-1)}(\mathtt{up}) = (1+o(1)) \frac{\exp\left(\xi_{2h}^{(s-1)}\right)}{k+\exp\left(\xi_{2h}^{(s-1)}\right)}, \tag{265}$$

$$\pi_{1,2h-1}^{(s-1)}(i) = (1+o(1)) \frac{1}{k+\exp\left(\xi_{2h-1}^{(s-1)}\right)}, \tag{266}$$

$$\pi_{1,2h}^{(s-1)}(i) = (1+o(1)) \frac{1}{k+\exp\left(\xi_{2h}^{(s-1)}\right)}. \tag{267}$$

Then by (264), (265), (109) and (107) in Induction Hypothesis 1, we have

$$r_h^{(s-1)} = (1+o(1)) \exp\Bigg(\underbrace{-\frac{1-\frac{h-1}{h}e^{-2a_{c,0}^{(s-1)}}e^{-2a_{c,1}^{(s-1)}}}{\left(1+\frac{h-1}{h}e^{-2a_{c,0}^{(s-1)}}\right)\left(1+e^{-2a_{c,1}^{(s-1)}}\right)}\left(a_{q,\times}^{(s-1)}+a_{q,0}^{(s-1)}\right)}_{\leqslant 0} \Bigg) \leqslant 1+o(1). \tag{268}$$

Note that if $a_{q,\times}^{(s-1)} \gtrsim \ln k$, there exists $s' < s$ such that

$$\frac{1}{3}a_{q,\times}^{(s-1)} \leqslant a_{q,\times}^{(s')} \leqslant \frac{1}{2}a_{q,\times}^{(s-1)}, \tag{269}$$

and by (99), (103), we have

$$e^{2a_{c,1}^{(s-1)}} \geqslant e^{2a_{c,1}^{(s-1)}} - e^{2a_{c,1}^{(s')}} \gtrsim \left(a_{q,\times}^{(s-1)}\right)^2. \tag{270}$$

Thus by (102) in Induction Hypothesis 1, we have

$$\text{If } a_{q,\times}^{(s-1)} \gtrsim \ln k: \quad \xi_{2h}^{(s-1)} = \left(1+O\left(\frac{1}{\ln k}\right)\right) a_{q,\times}^{(s-1)}. \tag{271}$$

Thus by the definition of $T_{1,1}$ in (93) and (271), we have

$$\forall s-1 \leqslant T_{1,1}: \quad \pi_{1,2}^{(s-1)}(\mathtt{up}) \leqslant \frac{1+o(1)}{2}. \tag{272}$$

By (265) and (268), we have when $s-1 \leqslant T_{1,1}$:

$$\forall h \in [k]: \quad \pi_{1,2h-1}^{(s-1)}(\mathtt{up}) \leqslant (1+o(1))\pi_{1,2h}^{(s-1)}(\mathtt{up}) = (1+o(1))\pi_{1,2}^{(s-1)}(\mathtt{up}). \tag{273}$$

If $\pi_{1,2}^{(s-1)}(\mathtt{up}) \lesssim \frac{1}{\sqrt{k}}$, then by (273), we have

$$\forall h \in [k]: \quad \pi_{1,2h}^{(s-1)}(\mathtt{up}) \lesssim \frac{1}{\sqrt{k}}, \quad \pi_{1,2h-1}^{(s-1)}(\mathtt{up}) \lesssim \frac{1}{\sqrt{k}}. \tag{274}$$

Thus by (247), we have

$$\text{when } \pi_{1,2}^{(s-1)}(\mathtt{up}) \lesssim \frac{1}{\sqrt{k}}: \quad \delta a_{q,\times}^{(s-1)} = \frac{1+o(1)}{1+e^{-2a_{c,1}^{(s-1)}}} \sum_{h=1}^{k-1} \mathbb{P}^{(s-1)}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(s-1)}(\mathtt{up}) V_{1,2h+1}^{(s-1)}. \tag{275}$$

If $\pi_{1,2}^{(s-1)}(\mathtt{up}) \gtrsim \frac{1}{\sqrt{k}}$, since $\pi_{1,2}^{(s-1)}(\mathtt{up}) \leqslant \frac{1+o(1)}{2}$ (c.f. (272)), we have

$$\xi_2^{(s-1)} \overset{(262)}{=} \frac{a_{q,\times}^{(s-1)} - e^{-2a_{c,1}^{(s-1)}} a_{q,0}^{(s-1)}}{1 + e^{-2a_{c,1}^{(s-1)}}} \asymp \ln k. \tag{276}$$

By (109) and (106) Induction Hypothesis 1, we have

$$a_{q,\times}^{(s-1)} \asymp a_{q,\times}^{(s-1)} + a_{q,0}^{(s-1)} \gtrsim \xi_2^{(s-1)} \asymp \ln k. \tag{277}$$

Note that when $a_{q,0}^{(s-1)} + a_{q,\times}^{(s-1)} \gtrsim 1$, by (99) in Induction Hypothesis 1, we have

$$e^{2a_{c,1}^{(s-1)}} \geqslant C \text{ for some constant } C > 1. \tag{278}$$

Thus by (277), (278), (268), we have when $\pi_{1,2}^{(s-1)}(\mathtt{up}) \gtrsim \frac{1}{\sqrt{k}}$,

$$\nu_h^{(s-1)} \leqslant \frac{1}{k^{c_1}} \text{ for some constant } c_1 > 0. \tag{279}$$

By (274), (279) and (273), we know that

$$\forall h \in [k]: \quad \pi_{1,2h-1}^{(s-1)}(\mathtt{up}) \leqslant \frac{1}{k^c} \text{ for some constant } c > 0. \tag{280}$$

Combining (280) with (275), we have for any $s - 1 \leqslant T_{1,1}$:

$$\delta a_{q,\times}^{(s-1)} = \frac{1 + o(1)}{1 + e^{-2a_{c,1}^{(s-1)}}} \sum_{h=1}^{k-1} \mathbb{P}^{(s-1)}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(s-1)}(\mathtt{up}) \left(1 - \pi_{1,2h}^{(s-1)}(\mathtt{up})\right) V_{1,2h+1}^{(s-1)}. \tag{281}$$

By comparing (251) and (281), we have

$$\forall t \leqslant s-1: \quad 0 \leqslant \delta a_{p,1}^{(t)} \lesssim \frac{1}{k^2} \delta a_{q,\times}^{(t)}, \tag{282}$$

and thus

$$a_{p,1}^{(s)} \lesssim \frac{1}{k^2} a_{q,\times}^{(s)} \lesssim \frac{\ln k}{k^2}. \tag{283}$$

Similarly, by comparing (252) and (281), we have

$$\forall t \leqslant s-1: \quad |\delta a_{p,3}^{(t)}| \lesssim \frac{1}{k} \delta a_{q,\times}^{(t)}, \tag{284}$$

and thus by (80) we have

$$|\delta a_{p,2}^{(s)}| \lesssim \left| \frac{\delta a_{p,1}^{(s)} - \delta a_{p,3}^{(s)}}{k-1} \right| \lesssim \frac{1}{k^2} \delta a_{q,\times}^{(s)}, \tag{285}$$

which yields

$$|a_{p,2}^{(s)}| \lesssim \frac{\ln k}{k^2}. \tag{286}$$

(286) and (283) together show that (107) holds at step s. Similarly, by comparing the gradient expressions of $a_{p,0}^{(t)}$ in (250) and $a_{p,\mathtt{up}}^{(t)}$ in (253) with (281) and by our choice of $T_{1,1}$, we immediately know that (100) holds at step s.

By (246), we have

$$\begin{aligned}
\delta a_{q,0}^{(s-1)} &\leqslant \pi_{1,1}^{(s-1)}(\mathtt{up}) V_{1,1}^{(s-1)} + \frac{1}{1 + \frac{h}{h+1} e^{-2a_{c,0}^{(s-1)}}} \sum_{h=1}^{k-1} \mathbb{P}^{(s-1)}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(s-1)}(\mathtt{up}) \\
&\quad \pi_{1,2h+1}^{(s-1)}(\mathtt{up}) V_{1,2h+1}^{(s-1)} \\
&\leqslant \nu_1^{(s-1)} \pi_{1,2}^{(s-1)}(\mathtt{up}) V_{1,1}^{(s-1)} + \frac{1}{k^c} \frac{1}{1 + \frac{h}{h+1} e^{-2a_{c,0}^{(s-1)}}} \sum_{h=1}^{k-1} \mathbb{P}^{(s-1)}\left(\mathcal{E}_{1,2h}\right) \\
&\quad \pi_{1,2h}^{(s-1)}(\mathtt{up}) V_{1,2h+1}^{(s-1)}.
\end{aligned} \tag{287}$$

We introduce the following lemma which reveals the ralationships of values at different states.

Lemma 6. *Under Induction Hypothesis 1, when $l = 1$, $s-1 \leqslant T_{1,1}$, then for any $t \leqslant s-1$ we have*

$$\forall h \in [k]: \quad \frac{V_{1,2h-1}^{(t)}}{V_{1,1}^{(t)}} \asymp 1. \tag{288}$$

Moreover, if we also have $\pi_{1,2}^{(t)}(\mathtt{up}) \lesssim \frac{1}{\ln k}$, then

$$\forall h \in [k]: \quad \frac{V_{1,2h-1}^{(t)}}{V_{1,1}^{(t)}} = 1 + o(1). \tag{289}$$

The proof of Lemma 6 is postponed to Appendix C.3.4.

By (281), (287), Lemma 6 and (268) we immediately know that

$$\delta a_{q,0}^{(s-1)} \lesssim \delta a_{q,\times}^{(s-1)}. \tag{290}$$

This suggests (102) holds at step s. By Lemma 6 and (287) we also know that when $\pi_{1,2}^{(s-1)}(\mathtt{up}) \lesssim \frac{1}{\sqrt{k}}$, we have

$$\begin{aligned}\delta a_{q,0}^{(s-1)} \leqslant (1+o(1))\nu_1^{(s-1)}\pi_{1,2}^{(s-1)}(\mathtt{up})V_{1,3}^{(s-1)} + \frac{1}{k^c}\frac{1}{1+\frac{h}{h+1}e^{-2a_{c,0}^{(s-1)}}} \\ \sum_{h=1}^{k-1}\mathbb{P}^{(s-1)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(s-1)}(\mathtt{up})V_{1,2h+1}^{(s-1)}.\end{aligned} \tag{291}$$

If $\nu_1^{(s-1)} \geqslant \frac{1}{4}$, by (268), we have $a_{q,\times}^{(s-1)} + a_{q,0}^{(s-1)} \lesssim 1$, and by (103) and (102) in Induction Hypothesis 1 we have $a_{q,0}^{(s-1)} \leqslant C$ for some constant $C > 0$. And from the first two relations in (277), we know that $\xi_2^{(s-1)} \lesssim 1$, and thus by (273) we have

$$\text{When } \nu_1^{(s-1)} \geqslant \frac{1}{4}: \quad \pi_{1,2h}^{(s-1)}(\mathtt{up}) \gtrsim \frac{1}{k}. \tag{292}$$

If $\nu_1^{(s-1)} \leqslant \frac{1}{4}$, $\pi_{1,2h}^{(s-1)}(\mathtt{up}) \lesssim \frac{1}{\sqrt{k}}$, by comparing (291) and (275), we have

$$\delta a_{q,0}^{(s-1)} \leqslant \frac{2}{3}\delta a_{q,\times}^{(s-1)}. \tag{293}$$

When $\pi_{1,2}^{(s-1)}(\mathtt{up}) \gtrsim \frac{1}{\sqrt{k}}$, $\xi_2^{(s-1)} \asymp \ln k$, by (279), Lemma 6 and (291), (275), there exists constant $c_2 > 0$ such that

$$\delta a_{q,0}^{(s-1)} \leqslant \frac{1}{k^{c_2}}\delta a_{q,\times}^{(s-1)} \leqslant \frac{2}{3}\delta a_{q,\times}^{(s-1)}. \tag{294}$$

The above two inequalities suggest (110) in Induction Hypothesis 1 holds at step s.

On the other hand, for any $t \leqslant s-1$, we have

$$\begin{aligned}-\delta a_{q,0}^{(t)} &\overset{(246)}{\leqslant} \frac{1}{1+e^{2a_{c,1}^{(t)}}}\sum_{h=1}^{k-1}\mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(\mathtt{up})\left(1-\pi_{1,2h}^{(t)}(\mathtt{up})\right)V_{1,2h+1}^{(t)} \\ &\overset{(281)}{\leqslant} (1+o(1))\frac{1+e^{-2a_{c,1}^{(t)}}}{1+e^{2a_{c,1}^{(t)}}}\delta a_{q,\times}^{(t)}.\end{aligned} \tag{295}$$

This indicates that

$$-\left(a_{q,0}^{(t)} - a_{q,0}^{(t')}\right) \lesssim \frac{1}{e^{2a_{c,1}^{(t')}}}\left(a_{q,\times}^{(t)} - a_{q,\times}^{(t')}\right). \tag{296}$$

By (109) and the above inequality, we know that (105) in Induction Hypothesis 1 holds at step s.

By (110), (262) and (273), we have for any $s-1 \leqslant T_{1,1}$:

$$\forall h \in [k-1]: \quad \pi_{1,2h}^{(s-1)}(\texttt{up}) \gtrsim \frac{1}{k}. \tag{297}$$

By (260) and Induction Hypothesis 1, we have

$$\forall h \in [k]: \quad \pi_{1,2h-1}^{(s)}(\texttt{down}) = (1+o(1))\frac{1-\pi_{1,2h-1}^{(s)}(\texttt{up})}{k} \overset{(280)}{=} \frac{1+o(1)}{k}, \tag{298}$$

and thus

$$\forall h \in [k]: \quad V_{1,2h-1}^{(t)} \geqslant \pi_{1,2h-1}^{(t)}(\texttt{down}) = (1+o(1))\frac{1}{k}. \tag{299}$$

Plugging (299) and (297) into (281), we have

$$\delta a_{q,\times}^{(s-1)} \gtrsim \frac{1}{k^2}. \tag{300}$$

This combined with (108) in Induction Hypothesis 1 yields (108) holds at step s.

By (246) and (247), we have

$$\begin{aligned}\delta a_{q,0}^{(s-1)} + \delta a_{q,\times}^{(s-1)} = {} & \pi_{1,1}^{(s-1)}(\texttt{up})V_{1,1}^{(s-1)} + \sum_{h=1}^{k-1} \mathbb{P}^{(s-1)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(s-1)}(\texttt{up}) \\ & \Bigg(\frac{1-\frac{h}{h+1}e^{-2a_{c,0}^{(s-1)}}}{1+\frac{h}{h+1}e^{-2a_{c,0}^{(s-1)}}}\pi_{1,2h+1}^{(s-1)}(\texttt{up}) \\ & + \frac{1-e^{-2a_{c,1}^{(s-1)}}}{1+e^{-2a_{c,1}^{(s-1)}}}\left(1-\pi_{1,2h}^{(s-1)}(\texttt{up})\right)\Bigg)V_{1,2h+1}^{(s-1)} > 0.\end{aligned} \tag{301}$$

This indicates (104) holds at step s. Note that by (268) and (297), if $a_{q,0}^{(s-1)} + a_{q,\times}^{(s-1)} \lesssim 1$, by (277) and (268), we have

$$\pi_{1,1}^{(s-1)}(\texttt{up}) \asymp \frac{1}{k}, \quad \pi_{1,2h}^{(s-1)}(\texttt{up}) \asymp \frac{1}{k}, \tag{302}$$

and thus

$$\mathbb{P}^{(s-1)}\left(\mathcal{E}_{1,2h}\right) \leqslant \prod_{i=1}^{h-1}\pi_{1,2i}^{(s-1)}(\texttt{up}) \leqslant \left(\frac{C_1}{k}\right)^{h-1} \quad \text{for some constant } C_1 > 0, \tag{303}$$

which, combine with (301) and Lemma 6, yields

$$\delta a_{q,0}^{(s-1)} + \delta a_{q,\times}^{(s-1)} = \pi_{1,1}^{(s-1)}(\texttt{up})V_{1,1}^{(s-1)} + (1+o(1))\frac{1-e^{-2a_{c,1}^{(s-1)}}}{1+e^{-2a_{c,1}^{(s-1)}}}\pi_{1,2}^{(s-1)}(\texttt{up})V_{1,3}^{(s-1)}. \tag{304}$$

For the same reason, if $a_{q,0}^{(s-1)} + a_{q,\times}^{(s-1)} \lesssim 1$, by (281), we have

$$\delta a_{q,\times}^{(s-1)} = \frac{1+o(1)}{1+e^{-2a_{c,1}^{(s-1)}}}\pi_{1,2}^{(s-1)}(\texttt{up})V_{1,3}^{(s-1)}. \tag{305}$$

By (302) and Lemma 6, (304) and (305), we have

$$\text{If } \delta a_{q,0}^{(s-1)} + a_{q,\times}^{(s-1)} \lesssim 1: \quad \delta a_{q,0}^{(s-1)} + \delta a_{q,\times}^{(s-1)} \asymp \delta a_{q,\times}^{(s-1)}. \tag{306}$$

By (278), if $a_{q,0}^{(s-1)} + a_{q,\times}^{(s-1)} \gtrsim 1$, we have

$$\frac{1-e^{-2a_{c,1}^{(s-1)}}}{1+e^{-2a_{c,1}^{(s-1)}}} \geqslant C_2 \text{ for some constant } C_2 > 0. \tag{307}$$

Therefore, by Lemma 6, Induction Hypothesis 1, (301) and (281), we have (306) also holds when $a_{q,0}^{(s-1)} + a_{q,\times}^{(s-1)} \gtrsim 1$, i.e., for any $s-1 \leqslant T_{1,1}$:

$$\delta a_{q,0}^{(s-1)} + \delta a_{q,\times}^{(s-1)} \asymp \delta a_{q,\times}^{(s-1)}. \tag{308}$$

This shows (109) and (103) in Induction Hypothesis 1 holds at step s. By a similar argument, we can show (106) in Induction Hypothesis 1 holds at step s.

Now we bound $\delta a_{b,1}^{(s-1)}$ to show (95) holds at step s. We decompose $I_{h,1}$ defined in (257) into two parts:

$$\begin{aligned} I_{h,1}^{(t)} :=& \underbrace{\frac{h(h+1)}{(2h+1)^2}\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right)\pi_{1,2h+1}^{(t)}(\texttt{down}_\times)}_{:=I_{h,1,A}^{(t)}} \\ &+ \underbrace{\frac{h}{(2h+1)^2}\left(\frac{1}{k}\left(a_{p,0}^{(t)} - h a_{p,\texttt{up}}^{(t)}\right) + (h+1)a_{p,2}^{(t)}\right)\pi_{1,2h+1}^{(t)}(\texttt{up})}_{:=I_{h,1,B}^{(t)}}, \end{aligned} \tag{309}$$

and thus by (256) we have

$$\begin{aligned} \delta a_{b,1}^{(s-1)} =& \underbrace{\frac{1+o(1)}{N}\sum_{h=1}^{k-1}\mathbb{P}^{(s-1)}\left(\mathcal{E}_{2h}\right)\pi_{1,2h}^{(s-1)}(\texttt{up})I_{h,1,A}^{(s-1)}V_{1,2h+1}^{(s-1)}}_{:=\delta a_{b,1,A}^{(s-1)}} \\ &+ \underbrace{\frac{1+o(1)}{N}\sum_{h=1}^{k-1}\mathbb{P}^{(s-1)}\left(\mathcal{E}_{2h}\right)\pi_{1,2h}^{(s-1)}(\texttt{up})I_{h,1,B}^{(s-1)}V_{1,2h+1}^{(s-1)}}_{:=\delta a_{b,1,B}^{(s-1)}}. \end{aligned} \tag{310}$$

From (266), (267) and (268) we deduce

$$\pi_{1,2h+1}^{(t)}(\texttt{down}_\times) \gtrsim \pi_{1,2h}^{(t)}(\texttt{down}_\times^1), \tag{311}$$

and thus by (107), (309), (310) and (251) we have

$$|\delta a_{b,1,A}^{(s-1)}| \lesssim \frac{\ln k}{Nk}\delta a_{p,1}^{(s-1)}. \tag{312}$$

From (100), (107), (281), (309), (310) and (280) we have

$$|\delta a_{b,1,B}^{(s-1)}| \lesssim \frac{\ln k}{Nk^{1+c}}\delta a_{q,\times}^{(s-1)}. \tag{313}$$

The above two expressions together suggest (95) holds at step s.

Similarly, to bound (96), we decompose $I_{h,0}$ defined in (255) into two parts:

$$\begin{aligned} I_{h,0}^{(t)} &= I_{h,0,A}^{(t)} - I_{h,0,B}^{(t)}, \\ I_{h,0,A}^{(t)} &:= \frac{1}{4h^2}\left(\frac{2h-1}{k}a_{p,0}^{(t)} + h a_{p,2}^{(t)} + \frac{h-1}{k}a_{p,\texttt{up}}^{(t)}\right)\left(1 - \pi_{1,2h}^{(t)}(\texttt{up})\right) \\ &\quad - \frac{h}{(2h+1)^2}\left(\frac{2}{k}\left(a_{p,0}^{(t)} + a_{p,\texttt{up}}^{(t)}\right) + a_{p,2}^{(t)}\right)\pi_{1,2h+1}^{(t)}(\texttt{up}), \\ I_{h,0,B}^{(t)} &:= \frac{1}{4h}\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right)\pi_{1,2h}^{(t)}(\texttt{down}_\times^1) \\ &\quad + \frac{h}{(2h+1)^2}\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right)\pi_{1,2h+1}^{(t)}(\texttt{down}_\times), \end{aligned} \tag{314}$$

and thus by (254) we have

$$\delta a_{b,0}^{(s-1)} = \underbrace{\frac{1+o(1)}{N}\sum_{h=1}^{k-1}\mathbb{P}^{(s-1)}\left(\mathcal{E}_{2h}\right)\pi_{1,2h}^{(t)}(\mathtt{up})I_{h,0,A}^{(s-1)}V_{1,2h+1}^{(s-1)}}_{:=\delta a_{b,0,A}^{(s-1)}} - \underbrace{\frac{1+o(1)}{N}\sum_{h=1}^{k-1}\mathbb{P}^{(s-1)}\left(\mathcal{E}_{2h}\right)\pi_{1,2h}^{(t)}(\mathtt{up})I_{h,0,B}^{(s-1)}V_{1,2h+1}^{(s-1)}}_{:=\delta a_{b,0,B}^{(s-1)}}. \tag{315}$$

Then analogous to how we bound $\delta a_{b,1,B}^{(s-1)}$ in (313), we can bound $\delta a_{b,0,A}^{(s-1)}$ as

$$|\delta a_{b,0,A}^{(s-1)}| \lesssim \frac{\ln k}{Nk}\delta a_{q,\times}^{(s-1)}. \tag{316}$$

And analogous to how we bound $\delta a_{b,1,A}^{(s-1)}$ in (312), we can bound $\delta a_{b,0,B}^{(s-1)}$ as

$$|\delta a_{b,0,B}^{(s-1)}| \lesssim \frac{\ln k}{Nk}\delta a_{p,1}^{(s-1)}. \tag{317}$$

The above two expressions together suggest (96) holds at step s.

By (97) we know that

$$\forall t\in\mathbb{N}: \quad a_{b,2}^{(t)} = \frac{-a_{b,0}^{(t)} - a_{b,1}^{(t)}}{N-1}. \tag{318}$$

Combining (318) with the fact that (96) and (95) hold at step s we just proved, we immediately know that (97) holds at step s.

We define

$$\delta e^{2a_{c,i}^{(t)}} := \frac{e^{2a_{c,i}^{(t+1)}} - e^{2a_{c,i}^{(t)}}}{\eta} = \frac{e^{2a_{c,i}^{(t)}}}{\eta}\left(e^{2\eta\delta a_{c,i}^{(t)}} - 1\right) \geqslant \frac{e^{2a_{c,i}^{(t)}}}{\eta}\cdot 2\eta\delta a_{c,i}^{(t)} = 2e^{2a_{c,i}^{(t)}}\delta a_{c,i}^{(t)} \tag{319}$$

for $i=0,1$. Then by (249) we have for any $t\leqslant T_{1,1}$:

$$\begin{aligned}\delta e^{2a_{c,1}^{(t)}} &\geqslant \sum_{h=1}^{k-1}\mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(\mathtt{up})\frac{2}{\left(1+e^{-2a_{c,1}^{(t)}}\right)^2}\left(a_{q,\times}^{(t)}+a_{q,0}^{(t)}\right)\\ &\quad (1-\pi_{1,2h}^{(t)}(\mathtt{up}))V_{1,2h+1}^{(t)}\\ &\overset{(247)}{\geqslant} \max\left\{\left(a_{q,\times}^{(t)}+a_{q,0}^{(t)}\right)\delta a_{q,\times}^{(t)}, 0\right\}.\end{aligned} \tag{320}$$

Thus by (99) and (109) in Induction Hypothesis 1, we have

$$e^{2a_{c,1}^{(s)}} - e^{2a_{c,1}^{(s')}} \geqslant \left(a_{q,\times}^{(s')}+a_{q,0}^{(s')}\right)\left(a_{q,\times}^{(s)}-a_{q,\times}^{(s')}\right), \tag{321}$$

holds for any $s'\leqslant s$, i.e., (99) in Induction Hypothesis 1 holds at step s.

$$\delta e^{2a_{c,0}^{(t)}} \overset{(248)}{>} 0, \tag{322}$$

and thus (98) in Induction Hypothesis 1 holds at step s.

When $t=T_{1,1}$. If $t=T_{1,1}$, by (93) and (271) we have

$$\pi_{1,2}^{(T_{1,1})}(\mathtt{up}) = \frac{1+o(1)}{2}, \tag{323}$$

and by (93) and (108), we have

$$T_{1,1} \lesssim \frac{k^2\ln k}{\eta}. \tag{324}$$

Also note that by our data distribution assumption, the probability of finding the goal within $2h-1$ steps is less than $\frac{h}{k}$. And

$$\mathbb{P}\left(\mathcal{E}_{1,2h}^{(T_{1,1})}\right) \leqslant \left(\frac{1+o(1)}{2}\right)^{h-1} \leqslant \left(\frac{2}{3}\right)^{h-1}$$

where we use (273) and (323). Thus letting $h = \lfloor \frac{\ln k}{\ln(3/2)} \rfloor$, we have the success rate

$$R(\pi^{(T_{1,1})}; \mathcal{P}_1) \lesssim \frac{\ln k}{k}. \tag{325}$$

By (270) we know that

$$e^{2a_{c,1}^{(T_{1,1})}} \gtrsim \ln^2 k. \tag{326}$$

Case 2: $T_{1,1}+1 \leqslant s-1 \leqslant T_{1/2}$**.** By (248) and (251) we know that (98) and (101) hold at step s, respectively. Note that (320) still holds when $s-1 > T_{1,1}$, and thus for Case 2, (99) and (112) in Induction Hypothesis 1 holds at step s. Also note that (270) still holds here, by which and Induction Hypothesis 1 we know that when $s-1 \geqslant T_{1,1}+1$:

$$\forall h \in [k]: \quad \xi_{2h}^{(s-1)} = a_{q,\times}^{(s-1)} + O\left(\frac{1}{\ln k}\right), \tag{327}$$

and for all $h \in [k]$:

$$r_h^{(s-1)} = (1+o(1)) \exp\left(-\frac{1}{1+\frac{h-1}{h}e^{-2a_{c,0}^{(s-1)}}}\left(a_{q,\times}^{(s-1)} + a_{q,0}^{(s-1)}\right) + \frac{1}{2(2h-1)}a_{p,2}^{(s)}\right). \tag{328}$$

(327) suggests (106) in Induction Hypothesis 1 holds at step s; (328) combined with (113), (116) (for $s-1 \in [T_{1,1}+1, T_{1,2}]$) or (118), (119) (for $s-1 \in [T_{1,2}+1, T_1]$) suggests (280) still holds for Case 2, i.e.,

$$\forall h \in [k]: \quad \pi_{1,2h-1}^{(s-1)}(\texttt{up}) \leqslant \frac{1}{k^c} \text{ for some constant } c > 0. \tag{329}$$

Note that for any $s-1 \in [T_{1,1}+1, T_1]$, we have

$$\begin{aligned}\pi_{1,1}(\texttt{down}) \leqslant V_{1,1}^{(s-1)} &\leqslant \pi_{1,1}(\texttt{down}) + V_{1,3}^{(s-1)} = \frac{1-\pi_{1,1}^{(s-1)}(\texttt{up})}{k} + V_{1,3}^{(s-1)} \\ &\overset{(329)}{=} \frac{1+o(1)}{k} + V_{1,3}^{(s-1)},\end{aligned} \tag{330}$$

and by Induction Hypothesis 1 and (260), we have for any $s-1 \in [T_{1,1}+1, T_1]$,

$$\forall h \in [k]: \quad V_{1,2h-1}^{(s-1)} \geqslant \pi_{1,2h-1}^{(s-1)}(\texttt{down}) \gtrsim \frac{1}{k}. \tag{331}$$

The above two relations give

$$V_{1,1}^{(s-1)} \lesssim V_{1,3}^{(s-1)}. \tag{332}$$

Furthermore, by (258), (113) we know that

$$\text{If } T_{1,1}+1 \leqslant s-1 \leqslant T_{1,2}: \quad \pi_{1,3}^{(s-1)}(\texttt{up}) \gtrsim \pi_{1,1}^{(s-1)}(\texttt{up}). \tag{333}$$

Thus by (332), we have

$$\text{If } T_{1,1}+1 \leqslant s-1 \leqslant T_{1,2}: \quad \pi_{1,1}^{(s-1)}(\texttt{up}) V_{1,1}^{(s-1)} \lesssim \mathbb{P}\left(\mathcal{E}_{1,2}\right) \pi_{1,3}^{(s-1)}(\texttt{up}) V_{1,3}^{(s-1)}, \tag{334}$$

where we use the fact that $\mathbb{P}\left(\mathcal{E}_{1,2}\right) = 1+o(1)$ guaranteed by (329). Thus by (334), (319) and (248) and (246), we obtain

$$\text{If } T_{1,1}+1 \leqslant s-1 \leqslant T_{1,2}: \quad \delta e^{2a_{c,0}^{(s-1)}} \gtrsim (a_{q,\times}^{(s-1)} + a_{q,0}^{(s-1)}) \delta a_{q,0}^{(s-1)}. \tag{335}$$

This suggests (114) in Induction Hypothesis 1 holds at step s.

Moreover, if $s-1 \in [T_{1,2}+1, T_{1/2}]$, then by (118) in Induction Hypothesis 1 we deduce

$$a_{q,0}^{(T_{1,2})} \geqslant 1.9 \ln k - 1, \tag{336}$$

From (116) we know that

$$a_{q,0}^{(T_{1,1})} + a_{q,\times}^{(T_{1,1})} \asymp \ln k. \tag{337}$$

and by (110) and our choice of $T_{1,1}$ we know that

$$a_{q,0}^{(T_{1,1})} \leqslant 2/3 \ln k + C. \tag{338}$$

Combining the above three inequalities with (114) in Induction Hypothesis 1 we have

$$e^{2a_{c,0}^{(T_{1,2})}} \geqslant e^{2a_{c,0}^{(T_{1,2})}} - e^{2a_{c,0}^{(T_{1,1})}} \gtrsim \ln^2 k. \tag{339}$$

Thus by (98) we know that for any $T_{1,2}+1 \leqslant s-1 \leqslant T_1$, we have

$$e^{2a_{c,0}^{(s)}} \gtrsim \ln^2 k. \tag{340}$$

On the other hand, by (248), (246), (247) and Induction Hypothesis 1, we have

$$\forall t \leqslant s-1: \quad \delta e^{2a_{c,0}^{(t)}} \lesssim \left(a_{q,\times}^{(t)} + a_{q,0}^{(t)}\right)\left(\delta a_{q,0}^{(t)} + \delta a_{q,\times}^{(t)}\right), \tag{341}$$

which together with (122) suggests

$$e^{2a_{c,0}^{(s)}} \lesssim \ln^2 k. \tag{342}$$

Thus by (340) and (342) we know that (121) holds at step s.

By (301) (which holds for any $s-1 \in [T_{1,1}+1, T_{1/2}]$) we have (104) holds at step s. And by comparing (301) and (246) we know that for any $s-1 \in [T_{1,1}+1, T_{1/2}-1]$,

$$-\delta a_{q,0}^{(s-1)} \lesssim \frac{1}{e^{2a_{c,1}^{(s-1)}}}\left(\delta a_{q,\times}^{(s-1)} + \delta a_{q,0}^{(s-1)}\right). \tag{343}$$

This suggests (105) in Induction Hypothesis 1 holds at step s. Similarly, by comparing the gradient expressions of $a_{p,1}^{(t)}$ in (251) and $a_{p,\mathrm{up}}^{(t)}$ in (253) with (301) and by (116), (122), we know that (100) holds at step s for any $s-1 \in [T_{1,1}+1, T_{1/2}]$. By comparing (252) and (301), we have

$$\forall t \in [T_{1,1}+1, T_{1/2}]: \quad |\delta a_{p,3}^{(t)}| \lesssim \frac{1}{k}\left(\delta a_{q,\times}^{(t)} + \delta a_{q,0}^{(t)}\right), \tag{344}$$

and thus by (284), (344), (122) and (103) we know that

$$-\frac{\ln k}{k} \lesssim a_{p,3}^{(s)} := \eta \sum_{t=0}^{s-1} \delta a_{p,3}^{(t)} \lesssim \frac{\ln k}{k}. \tag{345}$$

and by (80) we have

$$a_{p,2}^{(s)} = \frac{a_{p,1}^{(s)} - a_{p,3}^{(s)}}{k-1}. \tag{346}$$

From the above two expressions and our choice of $T_{1,2}$ we deduce that when $T_{1,1}+1 \leqslant s-1 \leqslant T_{1,2}$, (113) holds at step s; when $T_{1,2}+1 \leqslant s-1 \leqslant T_{1/2}-1$, (120) holds at step s.

Now we show (95) in Induction Hypothesis 1 holds at step s for Case 2. Note by (261), (260) and Induction Hypothesis 1, (311) still holds for any $s-1 \in [T_{1,1}+1, T_{1/2}-1]$. By our choice of $T_{1/2}$, (101), (113) or (120) we have

$$\forall t \in [T_{1,1}+1, s-1]: \quad -\frac{\ln k}{k^2} \lesssim a_{p,1}^{(t)} + a_{p,2}^{(t)} \lesssim k \ln k, \tag{347}$$

and

$$a_{p,1}^{(t)} + a_{p,2}^{(t)} > 0 \tag{348}$$

after $a_{p,1}^{(t)} \geqslant 2$. By the above two relations and (309), (310), (311) and (251) we have

$$\forall t \in [T_{1,1}+1, s-1]: \quad -\frac{\ln k}{Nk}\delta a_{p,1}^{(t)} \lesssim \delta a_{b,1,A}^{(t)} \lesssim \frac{k^2 \ln k}{N}\delta a_{p,1}^{(t)}, \tag{349}$$

and

$$\delta a_{b,1,A}^{(t)} > 0 \tag{350}$$

after $a_{p,1}^{(t)} \geqslant 2$. Thus by the above two relations and our choice of $T_{1/2}$ we have

$$-\frac{\ln k}{Nk} \lesssim a_{b,1,A}^{(s)} - a_{b,1,A}^{(T_{1,1})} \lesssim \frac{k^3 \ln^2 k}{N}. \tag{351}$$

By (100), (113) or (120) and our choice of $T_{1/2}$ we have

$$\forall t \in [T_{1,1}+1, s-1]: \quad -\frac{\ln k}{k} \lesssim \frac{h}{(2h+1)^2}\left(\frac{1}{k}\left(a_{p,0}^{(t)} - h a_{p,\mathrm{up}}^{(t)}\right) + (h+1)a_{p,2}^{(t)}\right) \lesssim \ln k, \tag{352}$$

Note that summing up (301) and (246) gives us

$$\begin{aligned} 2\delta a_{q,0}^{(t)} + \delta a_{q,\times}^{(t)} &= 2\pi_{1,1}^{(t)}(\mathrm{up})V_{1,1}^{(t)} + \sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(\mathrm{up})\Bigg(\frac{2 - \frac{h}{h+1}e^{-2a_{c,0}^{(t)}}}{1 + \frac{h}{h+1}e^{-2a_{c,0}^{(t)}}} \\ &\quad \pi_{1,2h+1}^{(t)}(\mathrm{up}) \\ &\quad + \frac{1 - 2e^{-2a_{c,1}^{(t)}}}{1 + e^{-2a_{c,1}^{(t)}}}\left(1 - \pi_{1,2h}^{(t)}(\mathrm{up})\right)\Bigg)V_{1,2h+1}^{(t)} \\ &\overset{(112)}{\geqslant} \frac{1}{2}\sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(\mathrm{up})\pi_{1,2h+1}^{(t)}(\mathrm{up})V_{1,2h+1}^{(t)} > 0 \end{aligned} \tag{353}$$

for any $t \in [T_{1,1}+1, s-1]$. Thus by (310), (352) and the above inequality, we have

$$\forall t \in [T_{1,1}+1, s-1]: \quad -\frac{\ln k}{Nk}\left(2\delta a_{q,0}^{(t)} + \delta a_{q,\times}^{(t)}\right) \lesssim \delta a_{b,1,B}^{(t)} \lesssim \frac{\ln k}{N}\left(2\delta a_{q,0}^{(t)} + \delta a_{q,\times}^{(t)}\right), \tag{354}$$

which, combined with (102) and (103), (122) or (116), gives us

$$-\frac{\ln^2 k}{Nk} \lesssim a_{b,1,B}^{(s)} - a_{b,1,B}^{(T_{1,1})} \lesssim \frac{\ln^2 k}{N}. \tag{355}$$

By (351) and (355) and that (95) holds at step $T_{1,1}$ by Induction Hypothesis 1, we know that (95) holds at step s.

Analogous to how we bound $\delta a_{b,1,B}^{(s-1)}$ in (354), we can bound $\delta a_{b,0,A}^{(s-1)}$ as

$$\forall t \in [T_{1,1}+1, s-1]: \quad |\delta a_{b,0,A}^{(t)}| \lesssim \frac{\ln k}{N}\left(\delta a_{q,\times}^{(t)} + \delta a_{q,0}^{(t)}\right), \tag{356}$$

and thus by (122) or (116) we have

$$|a_{b,0,A}^{(s)} - a_{b,0,A}^{(T_{1,1})}| \lesssim \frac{\ln^2 k}{N}. \tag{357}$$

And analogous to how we bound $a_{b,1,A}^{(s-1)}$ in (351), here we can bound $a_{b,0,B}^{(s-1)}$ as

$$-\frac{\ln k}{Nk} \lesssim a_{b,0,B}^{(s)} - a_{b,0,B}^{(T_{1,1})} \lesssim \frac{k^3 \ln^2 k}{N}. \tag{358}$$

The above two expressions together with (315) and that (96) holds at step $T_{1,1}$ imply that (96) holds at step s. Combining (318) with the fact that (96) and (95) hold at step s we just proved, we immediately know that (97) holds at step s.

Now we show (111) in Induction Hypothesis 1 holds at step s. Note that it implies (101) in Induction Hypothesis 1 holds at step s for any $s-1 \in [T_{1,1}+1, T_{1/2}]$. We consider two sub-cases: (i) $a_{p,1}^{(s-1)} \leqslant 5\ln k$ and (ii) $a_{p,1}^{(s-1)} \geqslant 5\ln k$.

(i) $a_{p,1}^{(s-1)} \leqslant 5\ln k$. If $a_{p,1}^{(s-1)} \leqslant 5\ln k$, then by (260), (113) and (120) we have for any $s-1 \in [T_{1,1}+1, T_{1/2}]$:

$$\pi_{1,5}^{(s-1)}(\texttt{down}_\times) \geqslant \frac{1+o(1)}{k^2}, \tag{359}$$

and thus by (251), we have

$$\begin{aligned} \text{If } a_{p,1}^{(s-1)} \leqslant 5\ln k: \quad & \delta a_{p,1}^{(s-1)} \gtrsim \frac{1}{k}\mathbb{P}^{(s-1)}(\mathcal{E}_{1,4})\pi_{1,4}^{(s-1)}(\texttt{up}) \\ & \pi_{1,5}^{(s-1)}(\texttt{down}_\times)V_{1,5}^{(s-1)} \gtrsim \frac{1}{k^4}. \end{aligned} \tag{360}$$

where we use the fact that $\mathbb{P}(\mathcal{E}_{1,4}) \gtrsim 1$, $\pi_{1,4}^{(s-1)}(\texttt{up}) \gtrsim 1$, and $V_{1,5}^{(s-1)} \gtrsim \frac{1}{k}$ guaranteed by (331), Induction Hypothesis 1 and our choice of $T_{1,1}$. (360) suggests (111) holds at step s when $a_{p,1}^{(s-1)} \leqslant 5\ln k$.

(ii) $a_{p,1}^{(s-1)} \geqslant 5\ln k$. We first use contradiction to show that

$$\text{If } a_{p,1}^{(s-1)} \geqslant 5\ln k: \quad a_{q,0}^{(s-1)} + a_{q,\times}^{(s-1)} \geqslant 4\ln k. \tag{361}$$

Suppose on the contrary that (361) does not hold, then by (104) we know that for any $t \leqslant s-1$, we have

$$a_{q,0}^{(t)} + a_{q,\times}^{(t)} < 4\ln k. \tag{362}$$

Since $a_{p,1}^{(t)} \lesssim \frac{\ln k}{k^2}$ for any $t \leqslant T_{1,1}$ (c.f. (107)), there exists $t \in [T_{1,1}+1, T_{1,2}]$ such that

$$4\ln k \leqslant a_{p,1}^{(t)} \leqslant 5\ln k. \tag{363}$$

Then for any $t' \in [T_{1,1}, t]$:

- If $a_{q,\times}^{(t')} < 3\ln k$, then by (259), (270), (113) and(363) we know that we have

$$1 - \pi_{1,2h}^{(t')}(\texttt{up}) \geqslant \frac{1+o(1)}{k^2}, \ \forall h \in [k-1]. \tag{364}$$

And by (270) (which still holds here) we have

$$e^{2a_{c,1}^{(t')}} \gtrsim \ln^2 k. \tag{365}$$

Thus by (301) and the above two expressions, we have for all $t' \in [T_{1,1}, t]$:

$$\text{If } a_{q,\times}^{(t')} < 3\ln k: \quad \delta a_{q,0}^{(t')} + \delta a_{q,\times}^{(t')} \geqslant \frac{1+o(1)}{k^2}\sum_{h=1}^{k-1}\mathbb{P}^{(t')}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t')}(\texttt{up})V_{1,2h+1}^{(t')}. \tag{366}$$

- If $a_{q,\times}^{(t')} \geqslant 3\ln k$, then by (115) in Induction Hypothesis 1 and (362), we have

$$0.9\ln k \leqslant a_{q,0}^{(t')} < \ln k. \tag{367}$$

Thus by (258), (367), (113) and (363) we know that

$$\pi_{1,2h-1}^{(t')}(\texttt{up}) \geqslant \frac{1+o(1)}{k^2}, \ \forall h \in [k]. \tag{368}$$

Additionally, by (114) in Induction Hypothesis 1, (367) and (121), we have

$$e^{2a_{c,0}^{(t')}} \gtrsim \ln^2 k. \tag{369}$$

Thus by (301) and the above two expressions, we have for all $t' \in [T_{1,1}, t]$:

$$\text{If } a_{q,\times}^{(t')} \geqslant 3\ln k: \quad \delta a_{q,0}^{(t')} + \delta a_{q,\times}^{(t')} \geqslant \frac{1+o(1)}{k^2} \sum_{h=1}^{k-1} \mathbb{P}^{(t')}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(t')}(\mathtt{up}) V_{1,2h+1}^{(t')}. \tag{370}$$

On the other hand, by (251), we have for all $t' \in [T_{1,1}, t]$:

$$\begin{aligned}
&\delta a_{p,1}^{(t')} \\
&\leqslant (1+o(1))\frac{1}{k}\sum_{h=1}^{k-1} \mathbb{P}^{(t')}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(t')}(\mathtt{up}) \left(\frac{1}{2}\cdot\frac{1-\pi_{1,2h}^{(t')}(\mathtt{up})}{k} + \frac{h}{2h+1} \right. \\
&\qquad \left. \cdot\frac{1-\pi_{1,2h+1}^{(t')}(\mathtt{up})}{k}\right) V_{1,2h+1}^{(t')} \\
&\leqslant (1+o(1))\frac{1}{k}\sum_{h=1}^{k-1} \mathbb{P}^{(t')}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(t')}(\mathtt{up}) \left(\frac{1+o(1)}{4}\cdot\frac{1}{k} + \frac{h}{2h+1} \right. \\
&\qquad \left. \cdot\frac{1+o(1)}{k}\right) V_{1,2h+1}^{(t')} \\
&\leqslant \frac{3}{4}\frac{1+o(1)}{k^2}\sum_{h=1}^{k-1} \mathbb{P}^{(t')}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(t')}(\mathtt{up}) V_{1,2h+1}^{(t')} \\
&\leqslant \frac{3+o(1)}{4}\left(\delta a_{q,0}^{(t')} + \delta a_{q,\times}^{(t')}\right),
\end{aligned} \tag{371}$$

where the second inequality uses (329), $a_{p,1}^{(t')} \geqslant 0$ guaranteed by (101) in Induction Hypothesis 1, and the fact that

$$\pi_{1,2h}^{(t')}(\mathtt{up}) \geqslant \frac{1+o(1)}{2} \tag{372}$$

guaranteed by our choice of $T_{1,1}$ and (327), and the last inequality follows from (366), (370).

(371) indicates that

$$a_{p,1}^{(t)} - a_{p,1}^{(T_{1,1})} \leqslant \frac{3(1+o(1))}{4}\left(\left(a_{q,0}^{(t)} + a_{q,\times}^{(t)}\right) - \left(a_{q,0}^{(T_{1,1})} + a_{q,\times}^{(T_{1,1})}\right)\right). \tag{373}$$

By (107) and (116) in Induction Hypothesis 1, we have

$$a_{p,1}^{(T_{1,1})} \lesssim \frac{1}{k^2}\left(a_{q,0}^{(T_{1,1})} + a_{q,\times}^{(T_{1,1})}\right). \tag{374}$$

Combining the above two expressions, we have

$$a_{q,0}^{(t)} + a_{q,\times}^{(t)} > a_{p,1}^{(t)} \overset{(363)}{\geqslant} 4\ln k. \tag{375}$$

This contradicts (362). Therefore, (361) must hold.

Now we show (111) in Induction Hypothesis 1 holds at step s under (361).

First note that for all $h \in [k-1]$:

$$\mathbb{P}\left(\mathcal{E}_{1,2h}\right) = \underbrace{\frac{k-h}{k}}_{\text{(a)}} \underbrace{\prod_{i=1}^{h-1} \pi_{1,2i}^{(s-1)}(\mathtt{up})}_{\text{(b)}} \underbrace{\prod_{i=1}^{h}\left(1-(i-1)\pi_{1,2i-1}^{(s-1)}(\mathtt{down}_\times) - \pi_{1,2i-1}^{(s-1)}(\mathtt{up})\right)}_{\text{(c)}}, \tag{376}$$

where (a) stands for the probability of the event that the goal is not among the first h children visited. By (361), (115), (113) (if $s-1 \in [T_{1,1}+1, T_{1,2}-1]$) or (119), (120) (if $s-1 \in [T_{1,2}+1, T_{1/2}-1]$) we have that if $a_{p,1}^{(s-1)} \geqslant 5\ln k$, then

$$a_{q,\times}^{(s-1)} - \frac{1}{2}a_{p,2}^{(s-1)} \geqslant 2\ln k - 2, \tag{377}$$

and hence by (259) and (327) we have

$$\forall h \in [k-1]: \quad 1 - \pi_{1,2h}^{(s-1)}(\texttt{up}) \lesssim \frac{1}{k}. \tag{378}$$

From the above inequality we know that (b) in (376) satisfies

$$\forall h \in [k]: \quad (b) \asymp 1. \tag{379}$$

Note that if $T_{1,1}+1 \leqslant s-1 \leqslant T_{1,2}$, then by (361) and (115) in Induction Hypothesis 1, we have

$$a_{q,0}^{(s-1)} \geqslant 0.95\ln k, \quad a_{q,0}^{(s-1)} \asymp a_{q,\times}^{(s-1)}, \tag{380}$$

and if $s-1 \in [T_{1,1}+1, T_{1,2}]$, by (380), (114), (110) and (109) and a similar argument as how we prove (339), we have

$$e^{2a_{c,0}^{(s-1)}} \gtrsim \ln^2 k. \tag{381}$$

If $s-1 \in [T_{1,2}+1, T_{1/2}-1]$, by (121) we know that the above inequality (381) still holds. And by Induction Hypothesis 1, we have (380) holds for any t where $a_{p,1}^{(t)} \in [5\ln k, k]$. Thus by (105), (112) and (116), we have if $s-1 \in [T_{1,2}+1, T_{1/2}-1]$:

$$a_{q,0}^{(s-1)} \geqslant 0.95\ln k - C, \quad a_{q,0}^{(s-1)} \asymp a_{q,\times}^{(s-1)} \tag{382}$$

for some constant $C > 0$. Combining (258), (380) or (382) and (381) (or (121)), (113) or (120), we obtain

$$\text{If } k-h \gtrsim k: \quad \pi_{1,2h-1}^{(s-1)}(\texttt{up}) \lesssim k^{-1.95}. \tag{383}$$

Thus if

$$\forall i \leqslant h-1: \quad \pi_{1,2i-1}^{(s-1)}(\texttt{down}_\times) \leqslant \frac{4}{h^2}, \tag{384}$$

then (c) in (376) satisfies

$$\text{If } k-h \gtrsim k: \quad (c) \asymp 1. \tag{385}$$

Thus by (376),(379) and (385), when $a_{p,1}^{(s-1)} \geqslant 5\ln k$, we have

$$\text{If (384) is satisfied}: \quad \mathbb{P}\left(\mathcal{E}_{1,2h}\right) \asymp 1, \ \forall h \in \left[\lceil \frac{3k}{4} \rceil\right]. \tag{386}$$

By (260) and the definition of $\pi_{1,2h-1}^{(t)}(\texttt{down}_\times)$ at the beginning of the proof, we have for all $h \in [k]$:

$$\begin{aligned}
&\pi_{1,2h-1}^{(s-1)}(\texttt{down}_\times) \\
&= (1+o(1))\frac{\exp\left(-\frac{a_{p,1}^{(t)}+a_{p,2}^{(t)}}{2h-1}\right)}{(h-1)\exp\left(-\frac{a_{p,1}^{(t)}+a_{p,2}^{(t)}}{2h-1}\right) + k - h + 1 + \exp\left(-a_{q,0}^{(t)} - \frac{h-1}{2h-1}a_{p,2}^{(t)}\right)} \\
&\overset{(380)}{=} (1+o(1))\frac{\exp\left(-\frac{a_{p,1}^{(t)}+a_{p,2}^{(t)}}{2h-1}\right)}{(h-1)\exp\left(-\frac{a_{p,1}^{(t)}+a_{p,2}^{(t)}}{2h-1}\right) + k - h + 1} \overset{(101)}{\leqslant} \frac{1+o(1)}{k},
\end{aligned} \tag{387}$$

which implies (384) holds for $h = \lceil\sqrt{k}\rceil$. Thus we have

$$\delta a_{p,1}^{(s-1)} \overset{(251)}{\geqslant} \frac{1+o(1)}{k} \underbrace{\mathbb{P}^{(s-1)}\left(\mathcal{E}_{1,2h}\right)}_{\asymp 1 \text{ by } (386)} \underbrace{\pi_{1,2h}^{(s-1)}(\texttt{up})}_{\asymp 1 \text{ by } (372)} \frac{h}{2h+1} \pi_{1,2h+1}^{(s-1)}(\texttt{down}_\times) \underbrace{V_{1,2h+1}^{(s-1)}}_{\gtrsim \frac{1}{k} \text{ by } (331)}$$

$$\gtrsim \frac{1}{k^2} \pi_{1,2h+1}^{(s-1)}(\texttt{down}_\times), \quad \text{for } h = \lceil\sqrt{k}\rceil. \tag{388}$$

Then by (387), (388) and (113), we have

$$\text{When } a_{p,1}^{(s-1)} \leqslant 4\lceil\sqrt{k}\rceil \ln k: \quad \delta a_{p,1}^{(s-1)} \gtrsim \frac{1}{k^5}. \tag{389}$$

Note that from (387) we also know that (384) holds for all $i \leqslant h \leqslant \lceil\frac{3k}{4}\rceil$ if

$$a_{p,1}^{(s-1)} \geqslant (2h-1)\ln k. \tag{390}$$

Therefore, when $a_{p,1}^{(s-1)} \geqslant 4\lceil\sqrt{k}\rceil \ln k$, by (390) we have (384) holds for $h = 2\lceil\sqrt{k}\rceil$, and thus by a similar argument as we proved (389), we can show that

$$\text{When } 4\lceil\sqrt{k}\rceil \ln k \leqslant a_{p,1}^{(s-1)} \leqslant 8\lceil\sqrt{k}\rceil \ln k: \quad \delta a_{p,1}^{(s-1)} \gtrsim \frac{1}{k^5}. \tag{391}$$

We set $m \in \mathbb{N}_+$ such that

$$2^{m-1}\lceil\sqrt{k}\rceil \leqslant \lceil\frac{3k}{2}\rceil < 2^m\lceil\sqrt{k}\rceil. \tag{392}$$

Then by repeating the above argument, we can show that for all $m' \in \{3, \cdots, m\}$:

$$\text{When } 2^{m'-1}\lceil\sqrt{k}\rceil \ln k \leqslant a_{p,1}^{(s-1)} \leqslant 2^{m'}\lceil\sqrt{k}\rceil \ln k: \quad \delta a_{p,1}^{(s-1)} \gtrsim \frac{\gamma^{2^{m'-1}\sqrt{k}}}{k^5}. \tag{393}$$

Finally, when $a_{p,1}^{(s-1)} \geqslant 2^m\lceil\sqrt{k}\rceil \ln k \geqslant \lceil\frac{3k}{2}\rceil \ln k$, then by (390) and (386), we have

$$\mathbb{P}\left(\mathcal{E}_{1,2h}\right) \asymp 1 \text{ when } h = \lceil\frac{3k}{4}\rceil.$$

Thus similar as (388), we have

$$\delta a_{p,1}^{(s-1)} \gtrsim \frac{1}{k} \pi_{1,2\lceil\frac{3k}{4}\rceil+1}^{(s-1)}(\texttt{down}_\times) V_{1,2\lceil\frac{3k}{4}\rceil+1}^{(s-1)}. \tag{394}$$

Note that when $\lceil\frac{3k}{2}\rceil \ln k \leqslant a_{p,1}^{(s-1)} \leqslant 4.1k \ln k$, by (120) we have

$$\pi_{1,2\lceil\frac{3k}{4}\rceil+1}^{(s-1)}(\texttt{down}_\times) \overset{(387)}{\gtrsim} \frac{1}{k^{11.2/3}}, \tag{395}$$

and for all $i \in \left[\lceil\frac{3k}{4}\rceil, \lceil\frac{3k}{4}\rceil + \lceil k^{2.2/3}\rceil\right]$:

$$\pi_{1,2i+1}^{(s-1)}(\texttt{down}_\times) \overset{(387)}{\lesssim} \frac{1}{k^2}. \tag{396}$$

$$V_{1,2\lceil\frac{3k}{4}\rceil+1}^{(s-1)} \geqslant \sum_{j=\lceil\frac{3k}{4}\rceil}^{\lceil\frac{3k}{4}\rceil+\lceil k^{2.2/3}\rceil} \frac{1}{k} \underbrace{\prod_{i=\lceil\frac{3k}{4}\rceil}^{j} \left(1-(k-i)\pi_{1,2i+1}^{(s-1)}(\texttt{down}_\times) - \pi_{1,2i+1}^{(s-1)}(\texttt{up})\right)}_{\asymp 1 \text{ by } (396) \text{ and } (383)}$$

$$\cdot \underbrace{\prod_{i=\lceil\frac{3k}{4}\rceil+1}^{j} \pi_{1,2i}^{(s-1)}(\texttt{up})}_{\asymp 1 \text{ by } (378)}$$

$$\gtrsim \frac{1}{k^{0.8/3}}. \tag{397}$$

Plugging (395), (397) into (394), we have

$$\delta a_{p,1}^{(s-1)} \gtrsim \frac{1}{k^5}. \tag{398}$$

By (389), (393) and (398), we know that (111) holds at step s also for subcase (ii) (i.e., when $a_{p,1}^{(s-1)} \geqslant 5\ln k$).

When $T_{1,2}+1 \leqslant s-1 \leqslant T_{1/2}-1$, if (119), (120) and (122) hold at step s, then (102) and (103) hold at step s.

Note that when $T_{1,1}+1 \leqslant s-1 \leqslant T_{1,2}$, the first relation in (115) can indicate (102) and (103) in Case 2. Thus to show they hold at step s, we only need to show (115) holds at step s. To do so, we consider two sub-cases: (i) $1-\pi_{1,2}^{(s-1)}(\mathtt{up}) \geqslant \frac{1}{k^{C'}}$ and (ii) $1-\pi_{1,2}^{(s-1)}(\mathtt{up}) \leqslant \frac{1}{k^{C'}}$, where we choose constant (recall c is the constant in (329))

$$C' \in (0, \min\{c, 1/2\}). \tag{399}$$

(i) $1-\pi_{1,2}^{(s-1)}(\mathtt{up}) \geqslant \frac{1}{k^{C'}}$ $(T_{1,1}+1 \leqslant s-1 \leqslant T_{1,2})$.

If $1-\pi_{1,2}^{(s-1)}(\mathtt{up}) \geqslant \frac{1}{k^{C'}}$, by comparing (251) and (247) we know that

$$\forall t \leqslant s-1: \quad \delta a_{p,1}^{(s-1)} \lesssim \frac{1}{k^{2-C'}} \delta a_{q,\times}^{(s-1)}, \tag{400}$$

which indicates

$$a_{p,1}^{(s-1)} \lesssim \frac{1}{k^{2-C'}} a_{q,\times}^{(s-1)} \lesssim \frac{\ln k}{k^{2-C'}}. \tag{401}$$

Therefore, (264) and (273) still hold, and thus when $1-\pi_{1,2}^{(s-1)}(\mathtt{up}) \geqslant \frac{1}{k^{C'}}$, (281) still holds. By comparing (246) and (281), we also have

$$\text{If } 1-\pi_{1,2}^{(s-1)}(\mathtt{up}) \geqslant \frac{1}{k^{C'}}: \quad \delta a_{q,0}^{(s-1)} \leqslant \delta a_{q,\times}^{(s-1)}. \tag{402}$$

This together with (110) indicates the first relation in (115) in Induction Hypothesis 1 holds at step s when $1-\pi_{1,2}^{(s-1)}(\mathtt{up}) \geqslant \frac{1}{k^{C'}}$.

To show the second relation in (115) in Induction Hypothesis 1 holds at step s, we first show

$$-a_{q,0}^{(s-1)} < \frac{\ln k}{3} \tag{403}$$

by contradiction. Suppose

$$-a_{q,0}^{(s-1)} \geqslant \frac{\ln k}{3}, \tag{404}$$

then there exists some $s' \leqslant s-1$ such that

$$\frac{\ln k}{5} \leqslant -a_{q,0}^{(s')} \leqslant \frac{\ln k}{4}. \tag{405}$$

Then by (102) we have

$$a_{q,\times}^{(s')} \gtrsim \ln k. \tag{406}$$

And by (270) (which still holds in Case 2) and (105) we know that

$$-a_{q,0}^{(s-1)} + a_{q,0}^{(s')} \lesssim \frac{1}{a_{q,\times}^{(s')}} \lesssim \frac{1}{\ln k}, \tag{407}$$

and thus

$$-a_{q,0}^{(s-1)} \leqslant -a_{q,0}^{(s')} + O\left(\frac{1}{\ln k}\right) \leqslant \frac{\ln k}{4} + O\left(\frac{1}{\ln k}\right) < \frac{\ln k}{3}, \tag{408}$$

this contradicts (404). Thus (403) holds.

From our choice of C' in (399) and (265), (327) and Induction Hypothesis 1, we have

$$a_{q,\times}^{(s-1)} \leqslant (3/2+o(1)) \ln k. \tag{409}$$

Combining (409) and (403), we have

$$a_{q,\times}^{(s-1)} - a_{q,0}^{(s-1)} \leqslant \left(\frac{11}{6}+o(1)\right) \ln k, \tag{410}$$

and by our gradient updates we know that

$$a_{q,\times}^{(s)} - a_{q,0}^{(s)} \leqslant 2 \ln k. \tag{411}$$

This shows the second relation in (115) in Induction Hypothesis 1 holds at step s.

(ii) $1-\pi_{1,2}^{(s-1)}(\mathtt{up}) \leqslant \frac{1}{k^{C'}}$ $(T_{1,1}+1 \leqslant s-1 \leqslant T_{1,2})$.

Define ratio

$$\forall h \in [k-1]: \quad \kappa_h^{(t)} := \frac{\pi_{1,2h+1}^{(t)}(\mathtt{up})}{1-\pi_{1,2h}^{(t)}(\mathtt{up})} \tag{412}$$

If $1-\pi_{1,2}^{(s-1)}(\mathtt{up}) \leqslant \frac{1}{k^{C'}}$ and $T_{1,1}+1 \leqslant s-1 \leqslant T_{1,2}$, then by (113), (259), (258), (327) and (329), we have

$$\kappa_h^{(s-1)} = (1+o(1))\frac{\exp\left(\zeta_h^{(s-1)}\right)}{\left(h\exp\left(-\frac{a_{p,1}^{(s-1)}}{2h+1}\right)+k-h\right)\left(h\exp\left(-\frac{a_{p,1}^{(s-1)}}{2h}\right)+k-h\right)}, \tag{413}$$

where

$$\begin{aligned}\zeta_h^{(s-1)} &:= \left(1+\frac{\frac{h}{h+1}e^{-2a_{c,0}^{(s-1)}}}{\frac{h}{h+1}e^{-2a_{c,0}^{(s-1)}}+1}\right)a_{q,\times}^{(s-1)} - \frac{1}{1+\frac{h}{h+1}e^{-2a_{c,0}^{(s-1)}}}a_{q,0}^{(s-1)} \\ &\quad - \frac{4h+1}{4h+2}a_{p,2}^{(s-1)} \\ &= a_{q,\times}^{(s-1)} - a_{q,0}^{(s-1)} + \frac{\frac{h}{h+1}e^{-2a_{c,0}^{(s-1)}}}{1+\frac{h}{h+1}e^{-2a_{c,0}^{(s-1)}}}\left(a_{q,\times}^{(s-1)}+a_{q,0}^{(s-1)}\right) - \frac{4h+1}{4h+2}a_{p,2}^{(s-1)}.\end{aligned} \tag{414}$$

Below we first show by contradiction that the second relation in (115) in Induction Hypothesis 1 holds at step s. Suppose otherwise, i.e.,

$$a_{q,\times}^{(s)} - a_{q,0}^{(s)} > 2.1\ln k, \tag{415}$$

Then by our gradient updates we have

$$a_{q,\times}^{(s-1)} - a_{q,0}^{(s-1)} > 2.1\ln k - \frac{1}{2}. \tag{416}$$

This combined with the fact that $a_{q,\times}^{(t-1)} + a_{q,0}^{(t-1)} > 0$ guaranteed by Induction Hypothesis 1 and (113) gives

$$\forall h \in [k-1]: \quad \zeta_h^{(s-1)} > 2.1\ln k - 2, \tag{417}$$

Therefore, we have

$$\forall h \in [k-1]: \quad \kappa_h^{(s-1)} \overset{(101)}{\geqslant} (1+o(1))\frac{\exp(\zeta_h^{(s-1)})}{k^2} \gtrsim k^{0.1}, \tag{418}$$

which indicates

$$\forall h \in [k-1]: \quad \pi_{1,2h}^{(s-1)}(\mathtt{up}) = o\left(\pi_{1,2h+1}^{(s-1)}(\mathtt{up})\right). \tag{419}$$

Thus by comparing (246) and (247), and using (334), we know that

$$\delta a_{q,\times}^{(s-1)} - \delta a_{q,0}^{(s-1)} < 0, \tag{420}$$

which suggests

$$a_{q,\times}^{(s)} - a_{q,0}^{(s)} < a_{q,\times}^{(s-1)} - a_{q,0}^{(s-1)} \leqslant 2.1\ln k, \tag{421}$$

where the last inequality is guaranteed by Induction Hypothesis 1. This contradicts (415). Thus the second relation in (115) in Induction Hypothesis 1 holds at step s.

We now show the first relation in (115) in Induction Hypothesis 1 holds at step s also by contradiction. Suppose otherwise, i.e.,

$$a_{q,\times}^{(s)} - a_{q,0}^{(s)} < 0, \tag{422}$$

Then by our gradient updates we have

$$a_{q,\times}^{(s-1)} - a_{q,0}^{(s-1)} < \frac{1}{2}. \tag{423}$$

(423) and our choice of $T_{1,1}$ imply

$$a_{q,0}^{(s-1)} \geqslant a_{q,\times}^{(s-1)} \geqslant \ln k - \frac{1}{2}. \tag{424}$$

Note that by (110) we know that there exists $s' \in [T_{1,1}+1, s-1]$ and constants $0 < c_1 < c_2 < 1$ such that

$$c_1 a_{q,\times}^{(s-1)} \leqslant a_{q,0}^{(s')} \leqslant c_2 a_{q,\times}^{(s-1)}. \tag{425}$$

Thus by (114), (103) and the above two inequalities we have

$$e^{2a_{c,0}^{(s-1)}} \geqslant e^{2a_{c,0}^{(s-1)}} - e^{2a_{c,0}^{(s')}} \gtrsim a_{q,\times}^{(s-1)} \left(a_{q,0}^{(s')} + a_{q,\times}^{(s')}\right) \gtrsim \left(a_{q,\times}^{(s-1)} + a_{q,0}^{(s-1)}\right) \ln k. \tag{426}$$

This combined with (414) and (113) gives

$$\zeta_h^{(s-1)} \leqslant a_{q,\times}^{(s-1)} - a_{q,0}^{(s-1)} + O\left(\frac{1}{\ln k}\right) \overset{(423)}{<} 1. \tag{427}$$

Plugging this into (413), and note that $a_{p,1} \leqslant k$ by our choice of $T_{1,2}$ (c.f. (94)), thus we have

$$\forall h \in [k-1]: \quad \kappa_h^{(s-1)} \lesssim \frac{1}{k^2}, \tag{428}$$

and

$$\forall h \in k: \quad \pi_{1,2h-1}^{(s-1)}(\mathtt{up}) \lesssim \frac{1}{k^2}. \tag{429}$$

Thus by the above two expressions, (246) and (247) we have

$$\delta a_{q,\times}^{(s-1)} - \delta a_{q,0}^{(s-1)} > 0, \tag{430}$$

which suggests

$$a_{q,\times}^{(s)} - a_{q,0}^{(s)} > a_{q,\times}^{(s-1)} - a_{q,0}^{(s-1)} \geqslant 0, \tag{431}$$

where the last inequality is guaranteed by Induction Hypothesis 1. This contradicts (422). Thus the first relation in (115) in Induction Hypothesis 1 holds at step s.

By (120) and our choice of $T_{1/2}$ we know that

$$a_{p,1}^{(s-1)} \leqslant (4.1 + o(1)) \ln k. \tag{432}$$

Thus by a similar argument as we proved (115) above, we can show by contradiction that (117) holds at step s for some constant $C > 0$.

We next show by contradiction that the second relation in (122) holds at step s. Suppose otherwise, i.e.,

$$a_{q,0}^{(s)} + a_{q,\times}^{(s)} > 20.4 \ln k, \tag{433}$$

Then there exists $s' < s$ such that

$$20.3 \ln k \leqslant a_{q,0}^{(s')} + a_{q,\times}^{(s')} \leqslant 20.35 \ln k. \tag{434}$$

Then by (117) we have

$$a_{q,\times}^{(s')} \geqslant 10.05 \ln k, \quad \text{and} \quad a_{q,0}^{(s')} \geqslant 7.05 \ln k. \tag{435}$$

Then by (259), (258), (120) and (426) (which still applies here by the above expression) we have for any $t \in [s', s-1]$:

$$\forall h \in [k]: \quad \pi_{1,2h-1}^{(t)}(\mathtt{up}) \lesssim \frac{1}{k^{7.05}}, \quad \text{and} \quad \forall h \in [k-1]: \quad \pi_{1,2h}^{(t)}(\mathtt{up}) \lesssim \frac{1}{k^{7.05}}. \tag{436}$$

Thus by (301) we have

$$\forall t \in [s', s-1]: \quad \delta a_{q,0}^{(t)} + \delta a_{q,\times}^{(t)} \lesssim \frac{1}{k^{6.05}} \overset{(111)}{\lesssim} \frac{1}{k^{1.05}} \delta a_{p,1}^{(t)}. \tag{437}$$

Thus

$$\left(a_{q,0}^{(s)} + a_{q,\times}^{(s)}\right) - \left(a_{q,0}^{(s')} + a_{q,\times}^{(s')}\right) \lesssim \frac{1}{k^{1.05}} \left(a_{p,1}^{(s)} - a_{p,1}^{(s')}\right) \overset{(92)}{\lesssim} \frac{k \ln k}{k^{1.05}} = o(1). \tag{438}$$

Then by (434) and the above relation, we have

$$a_{q,0}^{(s)} + a_{q,\times}^{(s)} < 20.4 \ln k. \tag{439}$$

This contradicts (433). Thus the second relation in (122) holds at step s. Note that by (104), (103) and our choice of $T_{1,1}$ we know that $a_{q,0}^{(t)} + a_{q,\times}^{(t)} \gtrsim \ln k$ after $t \geqslant T_{1,1} + 1$. This together with the second relation in (122) indicates (116) holds at step s.

We next show the first relation in (122) holds at step s also by contradiction ($s-1 \in [T_{1,2}+1, T_{1/2}-1]$). Suppose otherwise, i.e.,

$$a_{q,0}^{(s)} + a_{q,\times}^{(s)} < 5.9 \ln k, \tag{440}$$

Then since by (104), $a_{q,0}^{(t)} + a_{q,\times}^{(t)}$ monotonically increases for $t \leqslant s-1$, thus we have

$$\forall t \in [T_{1,1}+1, T_{1,2}]: \quad a_{q,0}^{(t)} + a_{q,\times}^{(t)} < 5.9 \ln k. \tag{441}$$

Then we have either $a_{q,0}^{(t)} \leqslant 1.95 \ln k$ or $a_{q,\times}^{(t)} \geqslant 3.95 \ln k$ for any $t \in [T_{1,1}+1, T_{1,2}]$. In either case, by (259), (258), Induction Hypothesis 1 and our choice of $T_{1,2}$ we have for any $t \in [T_{1,1}+1, T_{1,2}]$:

$$\forall h \in [k-1]: \quad \pi_{1,2h}^{(t)}(\mathtt{up}) + \pi_{1,2h+1}^{(t)}(\mathtt{up}) \gtrsim \frac{1}{k^{2.95}}. \tag{442}$$

By the above expression and the second line of (371), we have

$$\forall t \in [T_{1,1}+1, T_{1,2}]: \quad \delta a_{q,0}^{(t)} + \delta a_{q,\times}^{(t)} \gtrsim \frac{1}{k^{0.95}} \delta a_{p,1}^{(t)}. \tag{443}$$

Let $t_0 \in [T_{1,1}+1, T_{1,2}]$ be the smallest t such that $a_{p,1}^{(t)} \geqslant 5 \ln k$. Then by (361) we have

$$a_{q,0}^{(t_0)} + a_{q,\times}^{(t_0)} \geqslant 4 \ln k. \tag{444}$$

Thus by (443) we have

$$\left(a_{q,0}^{(T_{1,2})} + a_{q,\times}^{(T_{1,2})}\right) - \left(a_{q,0}^{(t_0)} + a_{q,\times}^{(t_0)}\right) \gtrsim \frac{1}{k^{0.95}} \left(a_{p,1}^{(T_{1,2})} - a_{p,1}^{(t_0)}\right) \gtrsim k^{0.05} \geqslant 2 \ln k. \tag{445}$$

Thus by (444) and the above relation, we have

$$a_{q,0}^{(T_{1,2})} + a_{q,\times}^{(T_{1,2})} \geqslant 6 \ln k. \tag{446}$$

This contradicts (441). Thus the first relation in (122) holds at step s.

By the above argument we also know that

$$a_{q,0}^{(T_{1,2})} + a_{q,\times}^{(T_{1,2})} \geqslant 5.9 \ln k. \tag{447}$$

This together with (115) gives

$$a_{q,0}^{(T_{1,2})} \geqslant 1.9 \ln k. \tag{448}$$

By (326), (105), (122) we know that

$$-\left(a_{q,0}^{(s-1)} - a_{q,0}^{(T_{1,2})}\right) \lesssim \frac{1}{\ln k}, \tag{449}$$

From (447), (449) and our update rule, we know that (118) holds at step s.

We next show (119) holds at step s for $s-1 \in [T_{1,2}+1, T_{1/2}-1]$. By (113), (115) and (447) we know that

$$a_{q,\times}^{(T_{1,2})} - \frac{1}{2} a_{p,2}^{(T_{1,2})} \geqslant 2.95 \ln k - 2, \tag{450}$$

and thus (119) holds at step $T_{1,2}$. Suppose (119) does not hold at step s ($s-1 \in [T_{1,2}+1, T_{1/2}-1]$), i.e.,

$$a_{q,\times}^{(s)} - \frac{1}{2} a_{p,2}^{(s)} < 2.425 \ln k, \tag{451}$$

then by our update rule, we know that

$$a_{q,\times}^{(s-1)} - \frac{1}{2} a_{p,2}^{(s-1)} \leqslant 2.425 \ln k + 1, \tag{452}$$

By (447), (115) and a similar argument as we proved (118) holds at step s above, by using the gradient expression of $a_{q,\times}^{(t)}$ in (247) and (121) we can show that

$$a_{q,\times}^{(s-1)} \geqslant 2.95 \ln k - 1, \tag{453}$$

By the above two inequalities, we have

$$\frac{1}{2} a_{p,2}^{(s-1)} \geqslant 0.525 \ln k - 2. \tag{454}$$

This together with (118) and (258) gives

$$\forall h \in [k]: \quad \pi_{1,2h-1}^{(s-1)}(\mathtt{up}) \lesssim \frac{1}{k^{2.425}}. \tag{455}$$

By (452) and (259) we have

$$\forall h \in [k-1]: \quad 1 - \pi_{1,2h}^{(s-1)}(\mathtt{up}) \gtrsim \frac{1}{k^{2.425}}. \tag{456}$$

Plugging the above two relations into (247), and by (112), (121) we have

$$\delta a_{q,\times}^{(s-1)} \lesssim \frac{1}{2k^{2.425}} \sum_{h=1}^{k-1} \mathbb{P}^{(s-1)}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(s-1)}(\mathtt{up}) V_{1,2h+1}^{(s-1)}. \tag{457}$$

Comparing the above expression with (251) and (252) and by Induction Hypothesis 1, we deduce

$$\delta a_{p,1}^{(s-1)} \lesssim k^{0.425} \delta a_{q,\times}^{(s-1)}, \quad \delta a_{p,3}^{(s-1)} \lesssim \frac{1}{k} \delta a_{q,\times}^{(s-1)}. \tag{458}$$

Therefore, by (80) we have

$$\delta a_{p,2}^{(s-1)} = \frac{\delta a_{p,1}^{(s-1)} - \delta a_{p,3}^{(s-1)}}{k-1} \lesssim k^{-0.575} \delta a_{q,\times}^{(s-1)}, \tag{459}$$

which suggests

$$\delta a_{q,\times}^{(s-1)} - \frac{1}{2} \delta a_{p,2}^{(s-1)} > 0, \tag{460}$$

and thus

$$a_{q,\times}^{(s)} - \frac{1}{2} a_{p,2}^{(s)} > a_{q,\times}^{(s-1)} - \frac{1}{2} a_{p,2}^{(s-1)} \overset{(119)}{\geqslant} 2.425 \ln k, \tag{461}$$

which contradicts (451). Thus (119) holds at step s.

When $s = T_{1/2}$. By our choice of $T_{1/2}$, our update rule and (120), we have

$$a_{p,1}^{(T_{1/2})} \geqslant 4.1k \ln k - 1, \tag{462}$$

and

$$a_{p,2}^{(T_{1/2})} \geqslant 4.1 \ln k - 1. \tag{463}$$

We first show (127) holds. Define

$$t_1' = \inf\{t \in [T_{1,2} + 1, T_{1/2} - 1] : \quad a_{p,2}^{(t)} \geqslant 4 \ln k\}. \tag{464}$$

We first show by contradiction that there exists $t_1 \in [t_1', T_{1/2} - 1]$ such that

$$a_{q,\times}^{(t_1)} - \frac{1}{2} a_{p,2}^{(t_1)} \geqslant 3.9 \ln k. \tag{465}$$

Suppose otherwise, i.e.,

$$\forall t \in [t_1', T_{1/2} - 1] : \quad a_{q,\times}^{(t)} - \frac{1}{2} a_{p,2}^{(t)} < 3.9 \ln k, \tag{466}$$

then for all such t, by (259) we have

$$\forall h \in [k-1] : \quad 1 - \pi_{1,2h}^{(t)}(\mathtt{up}) \lesssim \frac{1}{k^{3.9}}, \tag{467}$$

and by (118), (258) and our choice of t_1' we have

$$\forall h \in [k-1] : \quad \pi_{1,2h+1}^{(t)}(\mathtt{up}) \gtrsim \frac{1}{k^{3.9}}, \tag{468}$$

Plugging the above two relations into (275), we have

$$\delta a_{q,\times}^{(t)} \lesssim \frac{1}{k^{3.9}} \sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(t)}(\mathtt{up}) V_{1,2h+1}^{(t)}. \tag{469}$$

By (261), (260), (120) and our choice of t_1' we have

$$\forall h \in [k-1] : \quad \pi_{1,2h}^{(t)}(\mathtt{down}_\times^1) = o\left(\pi_{1,2h+1}^{(t)}(\mathtt{down}_\times)\right), \quad \pi_{1,2h+1}^{(t)}(\mathtt{down}_\times) \lesssim \frac{1}{k^2}, \tag{470}$$

Thus by the above expression, (469), (251), (252) and (80) we have

$$\delta a_{p,2}^{(t)} \lesssim \frac{1}{k^4} \sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(t)}(\mathtt{up}) V_{1,2h+1}^{(t)} \lesssim \frac{1}{k^{0.1}} \delta a_{q,\times}^{(t)}, \tag{471}$$

$$a_{q,\times}^{(T_{1/2})} - a_{q,\times}^{(t_1')} \gtrsim k^{0.1} \ln k, \tag{472}$$

and thus

$$a_{q,\times}^{(T_{1/2})} - \frac{1}{2} a_{p,2}^{(T_{1/2})} \gtrsim k^{0.1} \ln k \geqslant 3.9 \ln k. \tag{473}$$

This contradicts (127). Therefore, such t_1 exists. Then we show

$$\forall t \in [t_1, T_{1/2}] : \quad a_{q,\times}^{(t)} - \frac{1}{2} a_{p,2}^{(t)} \geqslant 3.9 \ln k \tag{474}$$

again by contradiction. Suppose otherwise, then there exists the first $t_2 \in [t_1, T_{1/2}]$ such that

$$a_{q,\times}^{(t_2)} - \frac{1}{2} a_{p,2}^{(t_2)} < 3.9 \ln k, \tag{475}$$

$$a_{q,\times}^{(t_2-1)} - \frac{1}{2} a_{p,2}^{(t_2-1)} \geqslant 3.9 \ln k. \tag{476}$$

By our update rule, we have

$$a_{q,\times}^{(t_2-1)} - \frac{1}{2}a_{p,2}^{(t_2-1)} \leqslant 3.9\ln k + 1. \tag{477}$$

Then by the same argument as above, we can show that

$$\delta a_{p,\times}^{(t_2-1)} \gtrsim k^{0.1}\delta a_{p,2}^{(t_2-1)}, \tag{478}$$

which indicates

$$\delta a_{q,\times}^{(t_2-1)} - \frac{1}{2}\delta a_{p,2}^{(t_2-1)} > 0, \tag{479}$$

and thus

$$a_{q,\times}^{(t_2)} - \frac{1}{2}a_{p,2}^{(t_2)} > a_{q,\times}^{(t_2-1)} - \frac{1}{2}a_{p,2}^{(t_2-1)} \overset{(476)}{\geqslant} 3.9\ln k. \tag{480}$$

This contradicts (475). Therefore, (474), and hence (127) holds.

We now prove (125).

$$\forall h \in [k]: \quad \pi_{1,2h-1}^{(T_{1/2})}(\texttt{down}_\times) \geqslant (1+o(1))\frac{k^{-2.05}}{k-h+1}. \tag{481}$$

By (463), (118), (117), and (258), we have

$$\forall h \in [k]: \quad \pi_{1,2h-1}^{(T_{1/2})}(\texttt{up}) \lesssim \frac{1}{k^{2.9}}. \tag{482}$$

By (127), (259) we have

$$\forall h \in [k-1]: \quad 1 - \pi_{1,2h}^{(T_{1/2})}(\texttt{up}) \lesssim \frac{1}{k^{2.9}}. \tag{483}$$

Note that for any $t \in \mathbb{N}$:

$$\begin{aligned} R(\pi^{(t)};\mathcal{P}_1) = {} & \underbrace{\frac{1}{k}}_{\text{(a)}}\underbrace{\left(1-\pi_{1,1}^{(t)}(\texttt{up})\right)}_{(b)} + \sum_{h=1}^{k-1}\underbrace{\frac{1}{k}}_{\text{(c)}} \\ & \underbrace{\left(\prod_{i=1}^{h-1}\pi_{1,2i}^{(t)}(\texttt{up})\prod_{i=1}^{h}\left(1-(i-1)\pi_{1,2i-1}^{(t)}(\texttt{down}_\times)-\pi_{1,2i-1}^{(t)}(\texttt{up})\right)\right.}_{(d)} \\ & \left.\cdot\underbrace{\pi_{1,2h}^{(t)}(\texttt{up})\left(1-\pi_{1,2h+1}^{(t)}(\texttt{up})-h\pi_{1,2h+1}^{(t)}(\texttt{down}_\times)\right)}_{(e)}\right), \end{aligned} \tag{484}$$

where (a) stands for the probability of the event that the goal is the first leaf visited by the agent, and (b) stands for the probability of the agent finding the goal conditioned on this event; (c) stands for the probability of the event that the goal is the $h+1$-th child visited, (d) stands for the probability of event $\mathcal{E}_{1,2h}$ conditioned on this event, and (e) stands for the probability of the agent finding the goal conditioned on this event and $\mathcal{E}_{1,2h+1}$.

By (481), (482), (483) we know that

$$\text{(b)} \geqslant 1 - o\left(\frac{1}{k^{1.05}}\right), \quad \text{(e)} \geqslant 1 - \frac{1}{k^{1.05}}. \tag{485}$$

For (d),

$$\text{(d)} = \underbrace{\prod_{i=1}^{h-1}\pi_{1,2i}^{(t)}(\texttt{up})}_{\text{(d1)}}\underbrace{\prod_{i=1}^{h}\left(1-(i-1)\pi_{1,2i-1}^{(t)}(\texttt{down}_\times)-\pi_{1,2i-1}^{(t)}(\texttt{up})\right)}_{\text{(d2)}}. \tag{486}$$

For (d1), when $t = T_{1/2}$, we have

$$\text{(d1)} \geqslant 1 - o\left(\frac{1}{k^{1.05}}\right). \tag{487}$$

For (d2), when $t = T_{1/2}$, we have

$$\begin{aligned}
\text{(d2)} &\geqslant \prod_{i=1}^{k-1}\left(1-(i-1)\pi_{1,2i-1}^{(T_{1/2})}(\texttt{down}_\times) - \pi_{1,2i-1}^{(T_{1/2})}(\texttt{up})\right) \\
&\geqslant \prod_{i=1}^{k-1}\left(1-\frac{2k^{-1.05}}{k-i+1}\right) \\
&= \prod_{i=2}^{k}\left(1-\frac{2k^{-1.05}}{i}\right) \\
&\geqslant \prod_{i=2}^{\lfloor k^{0.04}\rfloor}\left(1-2k^{-1.05}\right)\prod_{i=\lfloor k^{0.04}\rfloor+1}^{\lfloor k^{0.08}\rfloor}\left(1-2k^{-1.09}\right)\cdots\prod_{i=\lfloor k^{0.96}\rfloor+1}^{k}\left(1-2k^{-2.01}\right) \\
&\geqslant \left(1-2k^{-1.05}\right)^{k^{0.04}}\left(1-2k^{-1.09}\right)^{k^{0.08}}\cdots\left(1-2k^{-2.01}\right)^{k}. \\
&\geqslant 1-\frac{c}{k^{1.01}}
\end{aligned} \tag{488}$$

for some constant $c > 0$. Plugging (485), (487), (488) into (484), we have

$$R(\pi^{(T_{1/2})};\mathcal{P}_1) \geqslant 1 - \frac{c'}{k^{1.01}} \tag{489}$$

where $c' > 0$ is some constant. This gives the second relation in (125).

On the other hand, note that $1 - R\left(\pi^{(T_{1/2})};\mathcal{P}_1\right)$—the failure probability of $\pi^{(T_{1/2})}$ on trees from P_1 is larger than the probability of the goal is at the k-th child the agent visits, but at step $2k-1$, the agent chooses some $\texttt{down}_i$ that's already been taken, i.e., we have

$$1 - R\left(\pi^{(T_{1/2})};\mathcal{P}_1\right) \geqslant \frac{1}{k}\pi_{1,2k-1}(\texttt{down}_\times) \overset{(481)}{\geqslant} \frac{1+o(1)}{k^{3.05}}. \tag{490}$$

This gives the first relation in (125).

When $s = T_{1/2}$, similar as how we show (95) holds for both cases, by Induction Hypothesis 1 and comparing (256) and (251), we have

$$\forall t \leqslant T_{1/2}-1,\ \text{if } a_{p,1}^{(t)} \geqslant k\ln k: \quad \delta a_{b,1}^{(t)} \gtrsim \frac{k^2\ln k}{N}\delta a_{p,1}^{(t)}. \tag{491}$$

Thus by our choice of $T_{1/2}$ and (95) we have (123) holds at step s. Similarly, by noting that

$$\begin{aligned}
\forall t \leqslant T_{1/2}-1,\ \text{if } a_{p,1}^{(t)} \geqslant k\ln k: \quad & \frac{k\ln k}{N}\delta a_{p,1}^{(t)} \lesssim \delta a_{b,0,B}^{(t)} \lesssim \frac{k^2\ln k}{N}\delta a_{p,1}^{(t)}, \\
& \delta a_{b,0,A}^{(t)} \lesssim \frac{\ln k}{N}\left(\delta a_{q,0}^{(t)} + \delta a_{q,\times}^{(t)}\right),
\end{aligned} \tag{492}$$

we have (124) holds at step s.

Below we prove (126). Write $\tau = T_{1/2}$ throughout this calculation. All probabilities here are under policy $\pi^{(\tau)}$ with a training distribution $\mathcal{P}_l$ that satisfies (21) for a fixed $l \in \{2,3,\cdots,L\}$. We define events

$$\mathcal{B}_1 := \left\{\text{goal is not at the first } k-l+1 \text{ leaves } \pi^{(\tau)} \text{ visits, and } a_i \in [k], a_i \neq a_j, \forall i,j \in [l], i \neq j\right\}, \tag{493}$$

$$\mathcal{B}_2 := \{a_{l+1} = \texttt{up}\}, \tag{494}$$

and for all $i \in \{2,3,\cdots,k-l+1\}$,

$$\mathcal{B}_{2i-1} := \left\{a_{l+2(i-1)} \in [k]\backslash\{a_1,a_2,\cdots,a_l,a_{l+2},\cdots,a_{l+2(i-2)}\}\right\}, \tag{495}$$

$$\mathcal{B}_{2i} \coloneqq \{a_{l+2i-1} = \texttt{up}\}\,. \tag{496}$$

Then we have

$$\begin{aligned}
&1 - R\left(\pi^{(\tau)}; \mathcal{P}_l\right) \\
&\geqslant \mathbb{P}(\mathcal{B}_1)\Big(\underbrace{\mathbb{P}(a_{l+1} \neq \texttt{up}|\mathcal{B}_1)}_{=1-\mathbb{P}(\mathcal{B}_2|\mathcal{B}_1)} + \mathbb{P}(\mathcal{B}_2|\mathcal{B}_1)\mathbb{P}(\{\pi^{(\tau)} \text{ fails at some step } h > l+1\}|\mathcal{B}_1\mathcal{B}_2)\Big) \\
&\geqslant \mathbb{P}(\mathcal{B}_1)\mathbb{P}(\{\pi^{(\tau)} \text{ fails at some step } h > l+1\}|\mathcal{B}_1\mathcal{B}_2) \\
&\geqslant \mathbb{P}(\mathcal{B}_1)\Big(\mathbb{P}(a_{l+2} \in \mathcal{A}_{l+2,\times}|\mathcal{B}_1\mathcal{B}_2) \\
&\qquad + \mathbb{P}(\mathcal{B}_3|\mathcal{B}_1\mathcal{B}_2)\mathbb{P}\Big(\{\pi^{(\tau)} \text{ fails at some step } h > l+2\}|\bigcap_{j=1}^{3}\mathcal{B}_j\Big)\Big) \\
&\geqslant \mathbb{P}(\mathcal{B}_1)\mathbb{P}(a_{l+2} \notin \{a_1, a_2, \cdots, a_{l-1}, \texttt{up}\}|\mathcal{B}_1\mathcal{B}_2)\mathbb{P}\Big(\{\pi^{(\tau)} \text{ fails at some step } h > l+2\}|\bigcap_{j=1}^{3}\mathcal{B}_j\Big) \\
&\geqslant \cdots \\
&\geqslant \mathbb{P}(\mathcal{B}_1)\left(\prod_{i=1}^{k-l}\mathbb{P}\Big(a_{l+2i} \notin \{a_1, a_2, \cdots, a_{l-1}, \texttt{up}\}|\bigcap_{j=1}^{2i}\mathcal{B}_j\Big)\right) \\
&\quad \cdot \left(\mathbb{P}\Big(a_{2k-l+1} \neq \texttt{up}\Big|\bigcap_{j=1}^{2(k-l)+1}\mathcal{B}_j\Big)\right. \\
&\qquad \left.+\mathbb{P}\Big(\mathcal{B}_{2(k-l+1)}\Big|\bigcap_{j=1}^{2(k-l)+1}\mathcal{B}_j\Big)\mathbb{P}\Big(a_{2k-l+2} \in \mathcal{A}_{k-l+1,\times}\Big|\bigcap_{j=1}^{2(k-l+1)}\mathcal{B}_j\Big)\right) \\
&\geqslant \mathbb{P}(\mathcal{B}_1)\left(\prod_{i=1}^{k-l}\mathbb{P}\Big(a_{l+2i} \notin \{a_1, a_2, \cdots, a_{l-1}, \texttt{up}\}\Big|\bigcap_{j=1}^{2i}\mathcal{B}_j\Big)\right) \\
&\quad \cdot \mathbb{P}\Big(a_{2k-l+2} \in \mathcal{A}_{2k-l+2,\times}\Big|\bigcap_{j=1}^{2(k-l)+1}\mathcal{B}_j\Big).
\end{aligned} \tag{497}$$

Note that conditioned on $\mathcal{B}_1$, by Induction Hypothesis 1 and (88), (89), (90), (91) in Lemma 1, we have

$$\begin{aligned}
\forall h \in [l]: \quad \varphi_h(m) &= -N_h(m)\left(a_{p,1}^{(\tau)} + a_{p,2}^{(\tau)}\right) + (h-1)a_{p,2}^{(\tau)} + o(1), \\
\xi_h &= -a_{q,0}^{(\tau)},
\end{aligned} \tag{498}$$

and thus conditioned on $\mathcal{B}_1$, for all $h \in [l]$, we have for all $m \in [k]$:

$$\begin{aligned}
&\pi_h^{(\tau)}(m) \\
&\quad = (1+o(1))\frac{\exp\left(-N_h(m)\frac{a_{p,1}^{(\tau)}+a_{p,2}^{(\tau)}}{h}\right)}{\exp\left(-a_{q,0}^{(\tau)} - \frac{h-1}{h}a_{p,2}^{(\tau)}\right) + (h-1)\exp\left(-\frac{a_{p,1}^{(\tau)}+a_{p,2}^{(\tau)}}{h}\right) + k - h + 1}.
\end{aligned} \tag{499}$$

From this, Induction Hypothesis 1, and the fact that

$$\mathbb{P}\left(\text{goal is not at the first } k-l+1 \text{ leaves } \pi^{(\tau)} \text{ visits}\right) = 1 - \frac{k-l+1}{k^l}, \tag{500}$$

we deduce

$$\mathbb{P}(\mathcal{B}_1)$$

$$
\begin{aligned}
&\geqslant \left(1-\frac{k-l+1}{k^l}\right)\prod_{h=1}^{l}(1-(1\\
&\qquad +o(1))\frac{\exp\left(-a_{q,0}^{(\tau)}-\frac{h-1}{h}a_{p,2}^{(\tau)}\right)+(h-1)\exp\left(-\frac{a_{p,1}^{(\tau)}+a_{p,2}^{(\tau)}}{h}\right)}{\exp\left(-a_{q,0}^{(\tau)}-\frac{h-1}{h}a_{p,2}^{(\tau)}\right)+(h-1)\exp\left(-\frac{a_{p,1}^{(\tau)}+a_{p,2}^{(\tau)}}{h}\right)+k-h+1}\Bigg)\\
&\geqslant \left(1-\frac{k-l+1}{k^l}\right)\left(1-(1+o(1))\frac{\exp\left(-a_{q,0}^{(\tau)}\right)}{k}\right)\\
&\overset{(118)}{\geqslant} \left(1-\frac{c_1}{k^{2.9}}\right)\left(1-\frac{k-l+1}{k^l}\right) \qquad (501)
\end{aligned}
$$

for some constant $c_1>0$.

For all $i\in[k-l+1]$, conditioned on $\bigcap_{j=1}^{2i}\mathcal{B}_j$, we have when $h=l+2i$:

$$
|\mathcal{A}_{l+2i,\times}|=N_{l+2i,\times}=N_{l+2i}(\mathtt{up})=i, \qquad (502)
$$

and thus by Induction Hypothesis 1 and (88), (89), (90), (91) in Lemma 1, we have

$$
\begin{aligned}
\varphi_{l+2i}(m)&=-N_{l+2i}(m)\left(a_{p,1}^{(\tau)}+a_{p,2}^{(\tau)}\right)+(l+i-1)a_{p,2}^{(\tau)}+o(1),\\
\xi_{l+2i}&=-a_{q,0}^{(\tau)}+O\left(\frac{1}{\ln k}\right), \qquad (503)
\end{aligned}
$$

and thus by (118), (120) in Induction Hypothesis 1 and our choice of $T_{1/2}$, for all $i\in[k-l]$, we have

$$
\begin{aligned}
&\mathbb{P}\left(a_{l+2i}\notin\{a_1,a_2,\cdots,a_{l-1},\mathtt{up}\}\Big|\bigcap_{j=1}^{2i}\mathcal{B}_j\right)\\
&=1-(1+o(1))\frac{\exp\left(-a_{q,0}^{(\tau)}-\frac{l+i-1}{l+2i}a_{p,2}^{(\tau)}\right)+(l-1)\exp\left(-\frac{a_{p,1}^{(\tau)}+a_{p,2}^{(\tau)}}{l+2i}\right)}{\exp\left(-a_{q,0}^{(\tau)}-\frac{l+i-1}{l+2i}a_{p,2}^{(\tau)}\right)+(l+i-1)\exp\left(-\frac{a_{p,1}^{(\tau)}+a_{p,2}^{(\tau)}}{l+2i}\right)+k-l-i+1}\\
&\geqslant 1-\frac{(1+o(1))l}{k^2}, \qquad (504)
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{P}\left(a_{2k-l+2}\in\mathcal{A}_{2k-l+2,\times}\Big|\bigcap_{j=1}^{2(k-l+1)}\mathcal{B}_j\right)\\
&=1-(1+o(1))\frac{(l-1)\exp\left(-\frac{a_{p,1}^{(\tau)}}{l+2(k-l+1)}\right)+\exp\left(-a_{q,0}^{(\tau)}-\frac{k}{l+2(k-l+1)}a_{p,2}^{(\tau)}\right)}{k\exp\left(-\frac{a_{p,1}^{(\tau)}}{l+2(k-l+1)}\right)+\exp\left(-a_{q,0}^{(\tau)}-\frac{k}{l+2(k-l+1)}a_{p,2}^{(\tau)}\right)}\\
&=1-(1+o(1))\frac{(l-1)\exp\left(-\frac{a_{p,1}^{(\tau)}}{2k}\right)+\exp\left(-a_{q,0}^{(\tau)}-\frac{1}{2}a_{p,2}^{(\tau)}\right)}{k\exp\left(-\frac{a_{p,1}^{(\tau)}}{2k}\right)+\exp\left(-a_{q,0}^{(\tau)}-\frac{1}{2}a_{p,2}^{(\tau)}\right)}\\
&\overset{(120)}{=}1-(1+o(1))\frac{l-1+\exp\left(-a_{q,0}^{(\tau)}\right)}{k+\exp\left(-a_{q,0}^{(\tau)}\right)}\\
&=1-\frac{(1+o(1))l}{k}. \qquad (505)
\end{aligned}
$$

Plugging (501), (504) and (505) into (497), we deduce

$$1 - R\left(\pi^{(\tau)}; \mathcal{P}_l\right) \geqslant 1 - \frac{c_2 l}{k} \tag{506}$$

for some constant $c_2 > 0$. This gives (126).

C.3.3 Proof of Lemma 5

Gradient computation. Note that when $l = 1$, for all even steps, the agent is at a leaf node. If it's not a goal node, then the agent needs to go up, or the environment terminates. Thus for any $h \in [2k]$, conditioned on $\mathcal{E}_{1,h}$, $N_{h,\times} = \lfloor h/2 \rfloor$. Therefore, we have

$$\begin{aligned}\delta a_{q,0}^{(t)} \overset{(86)}{=} & \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = \times\} \frac{e^{-a_{c,1}^{(t)}}}{e^{a_{c,1}^{(t)}} + e^{-a_{c,1}^{(t)}}} \pi_h^{(t)}(\mathrm{up})(1 - \pi_h^{(t)}(\mathrm{up})) Q_h^{(t)}(\mathrm{up}) \middle| l = 1\right] \\ & + \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0\} \frac{(h+1)e^{a_{c,0}^{(t)}}}{(h-1)e^{-a_{c,0}^{(t)}} + (h+1)e^{a_{c,0}^{(t)}}} \pi_h^{(t)}(\mathrm{up}) G_h^{(t)} \middle| l = 1\right].\end{aligned} \tag{507}$$

With the notation we define at the beginning of Appendix C.3.2, by the above equation, we have

$$\begin{aligned}& \delta a_{q,0}^{(t)} \\ & = \pi_{1,1}^{(t)}(\mathrm{up}) V_{1,1}^{(t)} + \sum_{h=1}^{k} \left(\mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h+1}\right) \frac{1}{1 + \frac{h}{h+1} e^{-2a_{c,0}^{(t)}}} \pi_{1,2h+1}^{(t)}(\mathrm{up}) \right. \\ & \quad \left. - \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right) \frac{1}{1 + e^{2a_{c,1}^{(t)}}} \pi_{1,2h}^{(t)}(\mathrm{up}) \left(1 - \pi_{1,2h}^{(t)}(\mathrm{up})\right) \right) V_{1,2h+1}^{(t)} \\ & = \pi_{1,1}^{(t)}(\mathrm{up}) V_{1,1}^{(t)} + \sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(t)}(\mathrm{up}) \left(\frac{1}{1 + \frac{h}{h+1} e^{-2a_{c,0}^{(t)}}} \pi_{1,2h+1}^{(t)}(\mathrm{up}) \right. \\ & \quad \left. - \frac{1}{1 + e^{2a_{c,1}^{(t)}}} \left(1 - \pi_{1,2h}^{(t)}(\mathrm{up})\right) \right) V_{1,2h+1}^{(t)},\end{aligned} \tag{508}$$

where the second relation uses

$$\mathbb{E}^{\pi^{(t)}}[Q_{2h}^{(t)}(\mathrm{up}) | \mathcal{E}_{1,2h}] = V_{1,2h+1}^{(t)} \tag{509}$$

by (245), and the third relation follows from the fact that

$$\mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h+1}\right) = \pi_{1,2h}^{(t)}(\mathrm{up}) \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right). \tag{510}$$

This gives (246).

Similarly, we have

$$\begin{aligned}\delta a_{q,\times}^{(t)} \overset{(87)}{=} & \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = \times\} \frac{e^{a_{c,1}^{(t)}}}{e^{a_{c,1}^{(t)}} + e^{-a_{c,1}^{(t)}}} \pi_h^{(t)}(\mathrm{up})(1 - \pi_h^{(t)}(\mathrm{up})) Q_h^{(t)}(\mathrm{up}) \middle| l = 1\right] \\ & - \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0\} \frac{(h-1)e^{-a_{c,0}^{(t)}}}{(h-1)e^{-a_{c,0}^{(t)}} + (h+1)e^{a_{c,0}^{(t)}}} \pi_h^{(t)}(\mathrm{up}) G_h^{(t)} \middle| l = 1\right],\end{aligned} \tag{511}$$

which gives (247).

$$\begin{aligned}\delta a_{c,0}^{(t)} \overset{(70)}{=} & \left(1 + \frac{1}{k}\right) \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0\} \frac{(h-1)(h+1)}{\left((h-1)e^{-a_{c,0}^{(t)}} + (h+1)e^{a_{c,0}^{(t)}}\right)^2} \left(a_{q,\times}^{(t)} \right.\right. \\ & \left.\left. + a_{q,0}^{(t)}\right) \pi_h^{(t)}(\mathrm{up}) G_h^{(t)} \middle| l = 1\right],\end{aligned} \tag{512}$$

which gives (248).

$$\delta a_{c,1}^{(t)} \overset{(71)}{=} \left(1+\frac{1}{k}\right)\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=\times\}\frac{1}{\left(e^{a_{c,1}^{(t)}}+e^{-a_{c,1}^{(t)}}\right)^2}\left(a_{q,\times}^{(t)}+a_{q,0}^{(t)}\right)\right.$$
$$\left.\pi_h^{(t)}(\mathrm{up})(1-\pi_h^{(t)}(\mathrm{up}))Q_h^{(t)}(\mathrm{up})\middle| l=1\right], \tag{513}$$

which gives (249).

$$\delta a_{p,0}^{(t)} \overset{(81)}{=} (1+o(1))\mathbb{E}\left[\sum_{h=1}^{H}\frac{\pi_h^{(t)}(\mathrm{up})}{h}\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\middle| l=1\right], \tag{514}$$

which gives (250).

$$\begin{aligned}\delta a_{p,1}^{(t)} &\overset{(82)}{=} (1+o(1))\left(\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=0\}\frac{1}{k}\frac{\lfloor h/2\rfloor}{h}\pi_h^{(t)}(j)G_h^{(t)}\middle| j\in\mathcal{A}_{h,\times}, l=1\right]\right.\\ &\left.+\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=\times\}\frac{N_h(j)}{h}\pi_h^{(t)}(j)\pi_h^{(t)}(\mathrm{up})Q_h^{(t)}(\mathrm{up})\middle| j\in[k], l=1\right]\right)\\ &=(1+o(1))\left(\mathbb{E}\left[\sum_{h=1}^{k-1}\frac{1}{k}\frac{h}{2h+1}\pi_{2h+1}^{(t)}(j)G_{2h+1}^{(t)}\middle| j\in\mathcal{A}_{2h+1,\times}, l=1\right]\right.\\ &\left.+\mathbb{E}\left[\sum_{h=1}^{k-1}\frac{1}{2k}\pi_{2h}^{(t)}(j)\pi_{2h}^{(t)}(\mathrm{up})Q_{2h}^{(t)}(\mathrm{up})\middle| j\in\mathcal{A}_{2h-1,\times}\cup\{a_{2h-1}\}, l=1\right]\right),\end{aligned} \tag{515}$$

where for the even-step term we use

$$\{N_{2h}(j)=1\}=\{j\in\mathcal{A}_{2h-1,\times}\cup\{a_{2h-1}\}\},\qquad \mathbb{P}\left(N_{2h}(j)=1\right)=\frac{h}{k}.$$

On this event,

$$\pi_{2h}^{(t)}(j)=\pi_{1,2h}^{(t)}(\mathtt{down}_\times^1).$$

For the odd-step term, we use

$$\mathbb{P}\left(j\in\mathcal{A}_{2h+1,\times}\right)=\frac{h}{k},\qquad \pi_{2h+1}^{(t)}(j)=\pi_{1,2h+1}^{(t)}(\mathtt{down}_\times).$$

Therefore, (515) gives (251). Similarly,

$$\begin{aligned}\delta a_{p,3}^{(t)} &\overset{(83)}{=} (1+o(1))\left(\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=0\}\frac{1}{k}\frac{\lfloor h/2\rfloor}{h}\pi_h^{(t)}(\mathrm{up})\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\right.\right.\\ &\left.\left.\middle| j\in\mathcal{A}_{h,\times}, l=1\right]\right.\\ &\left.+\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=\times\}\frac{1}{k}\frac{\lfloor h/2\rfloor}{h}\pi_h^{(t)}(\mathrm{up})\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\middle| N_h(j)=1, l=1\right]\right)\\ &=(1+o(1))\left(-\mathbb{E}\left[\sum_{h=1}^{k-1}\frac{1}{k}\frac{h}{2h+1}\pi_{2h+1}^{(t)}(\mathrm{up})G_{2h+1}^{(t)}\middle| j\in\mathcal{A}_{2h+1,\times}, l=1\right]\right.\\ &\left.+\mathbb{E}\left[\sum_{h=1}^{k-1}\frac{1}{2k}\pi_{2h}^{(t)}(\mathrm{up})\left(1-\pi_{2h}^{(t)}(\mathrm{up})\right)Q_{2h}^{(t)}(\mathrm{up})\middle| j\in\mathcal{A}_{2h-1,\times}\cup\{a_{2h-1}\}, l=1\right]\right),\end{aligned} \tag{516}$$

which gives (252).

$$\delta a_{p,\mathrm{up}}^{(t)} \overset{(84)}{=} -(1+o(1))\mathbb{E}\left[\sum_{h=1}^{H}\pi_h^{(t)}(\mathrm{up})\frac{\lfloor (h-1)/2\rfloor}{h}\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\middle| l=1\right], \tag{517}$$

which gives (253).

To compute the simplified form of $\delta a_{b,i}^{(t)}$ for $i \in \{0,1,2\}$, we first simplify $\beta_{h,i}^{(t)}$ for $i \in \{0,1,2\}$ defined in (76), (77) and (78) for $l = 1$ as follows for all $h \in [k]$:

$$\beta_{2h-1,0}^{(t)} = -\frac{a_{p,0}^{(t)}}{k}\pi_{2h-1}^{(t)}(\texttt{up})G_{2h-1}^{(t)}, \tag{518}$$

$$\beta_{2h,0}^{(t)} = \frac{a_{p,0}^{(t)}}{k}\pi_{2h}^{(t)}(\texttt{up})\left(1-\pi_{2h}^{(t)}(\texttt{up})\right)Q_{2h}^{(t)}(\texttt{up}), \tag{519}$$

$$\beta_{2h-1,1}^{(t)} = \left(\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right)\pi_{2h-1}^{(t)}(\texttt{down}_\times) + a_{p,2}^{(t)}\pi_{2h-1}^{(t)}(\texttt{up})\right)(h-1)G_{2h-1}^{(t)}, \tag{520}$$

$$\beta_{2h,1}^{(t)} = 0, \tag{521}$$

$$\beta_{2h-1,2}^{(t)} = \frac{h-1}{k}a_{p,\texttt{up}}^{(t)}\pi_{2h-1}^{(t)}(\texttt{up})G_{2h-1}^{(t)}, \tag{522}$$

$$\beta_{2h,2}^{(t)} = \left(-\left(ha_{p,2}^{(t)}+\frac{h-1}{k}a_{p,\texttt{up}}^{(t)}\right)\left(1-\pi_{2h}^{(t)}(\texttt{up})\right)+\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right)h \pi_{2h}^{(t)}(\texttt{down}_\times^1)\right)\pi_{2h}^{(t)}(\texttt{up})Q_{2h}^{(t)}(\texttt{up}). \tag{523}$$

By the above expressions and

$$\delta a_{b,0}^{(t)} \overset{(73)}{=} \frac{1+o(1)}{N}\mathbb{E}\left[\sum_{h=1}^{H}\frac{1}{h^2}\left((h-1)\beta_{h,0}^{(t)}-\beta_{h,1}^{(t)}-\beta_{h,2}^{(t)}\right)\right], \tag{524}$$

$$\delta a_{b,1}^{(t)} \overset{(74)}{=} \frac{1+o(1)}{N}\mathbb{E}\left[\sum_{h=1}^{k-1}\frac{1}{(2h+1)^2}\left((h+1)\beta_{2h+1,1}^{(t)}-h\beta_{2h+1,0}^{(t)}-h\beta_{2h+1,2}^{(t)}\right)\right], \tag{525}$$

we obtain (254) and (256).

Policy computation. By Induction Hypothesis 1, when $l = 1$, (89) can be simplified to (recall we assume $N \geqslant k^5$)

$$\forall i \in [k]: \quad \varphi_{2h-1}^{(t)}(i) = -\frac{a_{p,0}^{(t)}}{k}+(h-1)\frac{a_{p,\texttt{up}}^{(t)}}{k}+(h-1)a_{p,2}^{(t)}-N_{2h-1}(i)\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right)+O\left(\frac{\ln^3 k}{k}\right), \tag{526}$$

$$\varphi_{2h}^{(t)}(i) = -\frac{a_{p,0}^{(t)}}{k}+(h-1)\frac{a_{p,\texttt{up}}^{(t)}}{k}+ha_{p,2}^{(t)}-N_{2h}(i)\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right)+O\left(\frac{\ln^3 k}{k}\right), \tag{527}$$

and (90) can be simplified to (262). By Induction Hypothesis 1 we know that

$$\forall h \in [k]: \quad \left|\frac{1}{k}\xi_h^{(t)}\right| \lesssim \frac{\ln k}{k}, \quad \left|\frac{a_{p,0}^{(t)}}{k}\right| \lesssim \frac{\ln k}{k}, \quad \left|\frac{a_{p,\texttt{up}}^{(t)}}{k}\right| \lesssim \frac{\ln k}{k}. \tag{528}$$

Combining (526), (527), (262), (528) with (88)-(91) in Lemma 1, we have (258), (259), (260) and (261) hold.

C.3.4 Proof of Lemma 6

We drop the superscript (t) for simplicity. By the definition of $\widehat{G}_h$ in (65) and the definition of $V_{1,h}$ in (245), we have the following recursion:

$$V_{1,2k+1} = 0, \quad \forall h \leqslant k: \; V_{1,2h-1} = \underbrace{\pi_{1,2h-1}(\texttt{down})}_{(i)} + \underbrace{(k-h)\pi_{1,2h-1}(\texttt{down})}_{(ii)} \pi_{1,2h}(\texttt{up})V_{1,2h+1}. \tag{529}$$

where (i) is the probability of the goal is reached after taking some down actions $i \in [k] \backslash \mathcal{A}_{h,\times}$, and (ii) is the probability of the goal is not reached after taking a down action $i \in [k] \backslash \mathcal{A}_{h,\times}$ (note if $\texttt{up}$ or $i \in \mathcal{A}_{h,\times}$ is taken, the goal will not be reached and the environment terminates).

To simplify notation, we define

$$y_h := V_{1,2h-1}, \quad r_h = 1 + (k-h)\pi_{1,2h}(\texttt{up})V_{1,2h+1}, \quad \beta_h = (k-h)\pi_{1,2h}(\texttt{up})\pi_{1,2h+1}(\texttt{down}), \tag{530}$$

then

$$\frac{y_h}{y_1} = \frac{\pi_{1,2h-1}(\texttt{down})r_h}{\pi_{1,1}(\texttt{down})r_1}, \tag{531}$$

and

$$r_k = 1, \quad r_h = 1 + \beta_h r_{h+1}. \tag{532}$$

If $\pi_{1,2} \lesssim \frac{1}{\ln k}$, then by (273),

$$\forall h \in [k]: \quad \pi_{1,2h}(\texttt{up}) \lesssim \frac{1}{\ln k}. \tag{533}$$

This suggests $\beta_h \lesssim \frac{1}{\ln k}$ for any $h \in [k-1]$. Thus we have

$$r_1 = 1 + \beta_1 r_2 = \cdots = 1 + \beta_1 + \beta_1\beta_2 + \cdots + \beta_1\beta_2\cdots\beta_{k-1} = 1 + O\left(\frac{1}{\ln k}\right). \tag{534}$$

Similarly, we have $r_h = 1 + O(\frac{1}{\ln k})$ for any $h \in [k]$. Thus $r_h/r_1 = 1 + O(\frac{1}{\ln k})$. Moreover, when $s - 1 \leqslant T_{1,1}$, by Induction Hypothesis 1 we have

$$\pi_{1,2h-1}(\texttt{down}) = (1+o(1))\frac{1}{k}, \quad \pi_{1,1}(\texttt{down}) = (1+o(1))\frac{1}{k}. \tag{535}$$

Thus $\pi_{1,2h-1}(\texttt{down})/\pi_{1,1}(\texttt{down}) = 1 + o(1)$. Therefore, by (531), $y_h/y_1 = 1 + o(1)$. This gives (289).

If $\pi_{1,2}^{(s-1)} \gtrsim \frac{1}{\ln k}$, then since $s - 1 \leqslant T_{1,1}$, by (273), we have

$$\forall h \in [k]: \quad \pi_{1,2h}(\texttt{up}) \leqslant \frac{1+o(1)}{2}, \tag{536}$$

which indicates $r_1 \asymp 1$, $r_h \asymp 1$ for any $h \in [k]$. Thus by (531), (288) holds.

C.3.5 Proof of Lemma 3

We show Lemma 3 by induction. First by Lemma 2 we know that Lemma 3 holds for $t = T_{1/2}$. We make the following induction hypothesis:

Induction Hypothesis 2: For all $t \leqslant s - 1$ ($s \in [T_{1/2}+1, T_1]$), the relations in Lemma 3 hold.

Below we show Lemma 3 holds for $t = s$ under Induction Hypothesis 2. We still follow the notation defined at the beginning of Appendix C.3.2. We first simplify the gradients for Phase 1.2.

Gradients/policy simplification for phase 1.2. Similar as Lemma 5, by Induction Hypothesis 2, using the notation we define at the beginning of Appendix C.3.2, we could further simplify the gradients and policy expressions in Lemma 1 when $l = 1$. We summarize the results in the following lemma.

Lemma 7. *Under Induction Hypothesis 2, we have for any $t \in [T_{1/2}, s-1]$:*

$$\delta a_{q,0}^{(t)} = (1+o(1))\left(\pi_{1,1}^{(t)}(up) + \sum_{h=1}^{k-1}\frac{k-h}{k}\left(\frac{1}{1+\frac{h}{h+1}e^{-2a_{c,0}^{(t)}}}\pi_{1,2h+1}^{(t)}(up) - \frac{1}{1+e^{2a_{c,1}^{(t)}}}\left(1-\pi_{1,2h}^{(t)}(up)\right)\right)\right), \tag{537}$$

$$\delta a_{q,\times}^{(t)} = (1+o(1)) \sum_{h=1}^{k-1} \frac{k-h}{k} \Bigg(\frac{1}{1+e^{-2a_{c,1}^{(t)}}} \left(1 - \pi_{1,2h}^{(t)}(up)\right) - \frac{1}{1+\frac{h+1}{h} e^{2a_{c,0}^{(t)}}} \pi_{1,2h+1}^{(t)}(up) \Bigg), \tag{538}$$

$$\delta a_{c,0}^{(t)} = (1+o(1)) \sum_{h=1}^{k-1} \frac{k-h}{k} \frac{h(h+1)}{\left(he^{-a_{c,0}^{(t)}} + (h+1)e^{a_{c,0}^{(t)}}\right)^2} \left(a_{q,\times}^{(t)} + a_{q,0}^{(t)}\right) \pi_{1,2h+1}^{(t)}(up), \tag{539}$$

$$\delta a_{c,1}^{(t)} = (1+o(1)) \sum_{h=1}^{k-1} \frac{k-h}{k} \frac{1}{\left(e^{a_{c,1}^{(t)}} + e^{-a_{c,1}^{(t)}}\right)^2} \left(a_{q,\times}^{(t)} + a_{q,0}^{(t)}\right) (1 - \pi_{1,2h}^{(t)}(up)), \tag{540}$$

$$\delta a_{p,0}^{(t)} = (1+o(1)) \Bigg(-\pi_{1,1}^{(t)}(up) + \sum_{h=1}^{k-1} \frac{k-h}{k} \Bigg(\frac{e^{a_{b,0}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}} (1 - \pi_{1,2h}^{(t)}(up)) - \frac{e^{a_{b,0}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(up) \Bigg) \Bigg), \tag{541}$$

$$\delta a_{p,1}^{(t)} = \frac{1+o(1)}{k} \sum_{h=1}^{k-1} \frac{k-h}{k} \Bigg(\frac{he^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}} \pi_{1,2h}^{(t)}(down_{\times}^{1}) + \frac{he^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(down_{\times}) \Bigg), \tag{542}$$

$$\delta a_{p,3}^{(t)} = \frac{1+o(1)}{k} \sum_{h=1}^{k-1} \frac{k-h}{k} \Bigg(\frac{he^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}} \left(1 - \pi_{1,2h}^{(t)}(up)\right) - \frac{he^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(up) \Bigg), \tag{543}$$

$$\delta a_{p,up}^{(t)} = (1+o(1)) \sum_{h=1}^{k-1} \frac{k-h}{k} \Bigg(\frac{he^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(up) - \frac{(h-1)e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}} \left(1 - \pi_{1,2h}^{(t)}(up)\right) \Bigg), \tag{544}$$

$$\delta a_{b,0}^{(t)} = \frac{1+o(1)}{N} \sum_{h=1}^{k-1} \frac{k-h}{k} I_{h,0}^{(t)}, \tag{545}$$

where

$$I_{h,0}^{(t)} := \frac{e^{a_{b,0}^{(t)}} e^{a_{b,2}^{(t)}}}{\left(e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}\right)^2} \Bigg(\left(\frac{2h-1}{k} a_{p,0}^{(t)} + h a_{p,2}^{(t)} + \frac{h-1}{k} a_{p,up}^{(t)} \right) (1 - \pi_{1,2h}^{(t)}(up)) - \left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right) h \pi_{1,2h}^{(t)}(down_{\times}^{1}) \Bigg)$$

$$- \frac{h e^{a_{b,0}^{(t)}}}{\left(e^{a_{b,0}^{(t)}} + h e^{a_{b,1}^{(t)}} + h e^{a_{b,2}^{(t)}}\right)^2} \Bigg[\left(e^{a_{b,1}^{(t)}} \left(a_{p,2}^{(t)} + \frac{a_{p,0}^{(t)}}{k} \right) + e^{a_{b,2}^{(t)}} \left(\frac{a_{p,up}^{(t)}}{k} + \frac{a_{p,0}^{(t)}}{k} \right) \right) \pi_{1,2h+1}^{(t)}(up) + e^{a_{b,1}^{(t)}} \left(a_{p,1}^{(t)} + a_{p,2}^{(t)} \right) \pi_{1,2h+1}^{(t)}(down_\times) \Bigg], \tag{546}$$

$$\delta a_{b,1}^{(t)} = \frac{1 + o(1)}{N} \sum_{h=1}^{k-1} \frac{k-h}{k} I_{h,1}^{(t)}, \tag{547}$$

where

$$I_{h,1}^{(t)} := \frac{h e^{a_{b,1}^{(t)}}}{\left(e^{a_{b,0}^{(t)}} + h e^{a_{b,1}^{(t)}} + h e^{a_{b,2}^{(t)}}\right)^2} \Bigg[\left(e^{a_{b,0}^{(t)}} + h e^{a_{b,2}^{(t)}} \right) \left(a_{p,1}^{(t)} + a_{p,2}^{(t)} \right) \pi_{1,2h+1}^{(t)}(down_\times) + \left(e^{a_{b,0}^{(t)}} \left(a_{p,2}^{(t)} + \frac{a_{p,0}^{(t)}}{k} \right) + h e^{a_{b,2}^{(t)}} \left(a_{p,2}^{(t)} - \frac{a_{p,up}^{(t)}}{k} \right) \right) \pi_{1,2h+1}^{(t)}(up) \Bigg]. \tag{548}$$

For all $\forall h \in [k]$, $i \in [k]$, *we have*

$$\pi_{1,h}^{(t)} := \mathsf{softmax}\left(\overline{\phi}_h^{(t)} \right), \tag{549}$$

where $\overline{\phi}_h^{(t)} \in \mathbb{R}^{k+1} = \left(\overline{\phi}_{h,1}^{(t)}, \cdots, \overline{\phi}_{h,k}^{(t)}, \overline{\phi}_{h,up}^{(t)} \right)^\top$ *is defined as*

$$\begin{aligned} \forall i \in [k]: \quad \overline{\phi}_{2h-1,i}^{(t)} &:= - \frac{N_{2h-1}(i)(a_{p,1}^{(t)} + a_{p,2}^{(t)}) e^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + (h-1) e^{a_{b,1}^{(t)}} + (h-1) e^{a_{b,2}^{(t)}}}, \\ \overline{\phi}_{2h,i}^{(t)} &:= - \frac{N_{2h}(i)(a_{p,1}^{(t)} + a_{p,2}^{(t)}) e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1) e^{a_{b,2}^{(t)}}}. \end{aligned} \tag{550}$$

$$\overline{\phi}_{h,up}^{(t)} := \left(1 + \frac{1}{k} \right) \xi_h^{(t)} - \mu_h^{(t)}, \tag{551}$$

where $\xi_h^{(t)}$ *is defined in* (262) *and* $\mu_h^{(t)}$ *is defined in* (131).

The proof of Lemma 7 is postponed to Appendix C.3.6.

From (542) and (547) we immediately know that (137) and (133) hold at step s. Also note that (319) still holds. And similar as (320) for Phase 1.1, by comparing (540) and (538), we have for any $t \leqslant s-1$:

$$\delta e^{2a_{c,1}^{(t)}} \geqslant \max\{(2 + o(1)) \left(a_{q,\times}^{(t)} + a_{q,0}^{(t)} \right) \delta a_{q,\times}^{(t)}, 0\}. \tag{552}$$

This indicates (136) holds at step s. From (539) we know that

$$\delta a_{c,0}^{(s-1)} > 0, \tag{553}$$

and thus by Induction Hypothesis 2 and (121) in Lemma 2 we know that

$$e^{2a_{c,0}^{(s)}} \gtrsim \ln^2 k. \tag{554}$$

By (134) and (103) in Lemma 2 we know that there exist constants $0 < C_1 \leqslant C_2$ such that for all $t \leqslant s-1$:

$$C_1 \left(a_{q,0}^{(t)} + a_{q,\times}^{(t)} \right) \leqslant a_{q,\times}^{(t)} \leqslant C_2 \left(a_{q,0}^{(t)} + a_{q,\times}^{(t)} \right), \tag{555}$$

and by (136) and (99) in Lemma 2 we have for any $t' \leqslant s-1$:

$$e^{2a_{c,1}^{(s-1)}} - e^{2a_{c,1}^{(t')}} \geqslant \left(a_{q,\times}^{(t')} + a_{q,0}^{(t')}\right)\left(a_{q,\times}^{(s-1)} - a_{q,\times}^{(t')}\right). \tag{556}$$

By our update rule, we can choose $t' \leqslant s-1$ such that

$$a_{q,\times}^{(t')} + a_{q,0}^{(t')} \in \left[\frac{C_1}{2C_2}\left(a_{q,\times}^{(s-1)} + a_{q,0}^{(s-1)}\right), \frac{C_1}{2C_2}\left(a_{q,\times}^{(s-1)} + a_{q,0}^{(s-1)}\right) + 1\right]. \tag{557}$$

Combining the above three relations, we have

$$\begin{aligned} e^{2a_{c,1}^{(s-1)}} &\geqslant \frac{C_1}{2C_2}\left(a_{q,\times}^{(s-1)} + a_{q,0}^{(s-1)}\right)\left(C_1\left(a_{q,\times}^{(s-1)} + a_{q,0}^{(s-1)}\right) - C_2\left(a_{q,\times}^{(t')} + a_{q,0}^{(t')}\right)\right) \\ &\geqslant \frac{C_1}{2C_2}\left(a_{q,\times}^{(s-1)} + a_{q,0}^{(s-1)}\right)\left(\frac{C_1}{2}\left(a_{q,\times}^{(s-1)} + a_{q,0}^{(s-1)}\right) - C_2\right) \\ &\asymp \left(a_{q,\times}^{(s-1)} + a_{q,0}^{(s-1)}\right)^2. \end{aligned} \tag{558}$$

By (558) and (262) we have

$$\xi_{2h}^{(s-1)} = a_{q,\times}^{(s-1)} + O\left(\frac{1}{a_{q,\times}^{(s-1)} + a_{q,0}^{(s-1)}}\right) = a_{q,\times}^{(s-1)} + O\left(\frac{1}{\ln k}\right). \tag{559}$$

By (135) and (121) in Lemma 2 we have

$$e^{2a_{c,0}^{(s-1)}} \gtrsim \ln^2 k. \tag{560}$$

Thus by (262) and (134) we have

$$\xi_{2h-1}^{(s-1)} = -\left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}^{(s-1)}. \tag{561}$$

For all $h \in [k-1]$, define

$$r_{\mathtt{up},h,0}^{(t)} := \frac{e^{a_{b,0}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}}, \quad r_{\mathtt{up},h,2}^{(t)} := \frac{e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}}, \tag{562}$$

$$\begin{aligned} r_{\mathtt{down},h,0}^{(t)} &:= \frac{e^{a_{b,0}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}}, \quad r_{\mathtt{down},h,1}^{(t)} := \frac{e^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}}, \\ &\qquad\qquad r_{\mathtt{down},h,2}^{(t)} := \frac{e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}}. \end{aligned} \tag{563}$$

Then by (132) we have for any t,

$$r_{\mathtt{down}_\times}^{(t)} = r_{\mathtt{down},k-1,1}^{(t)}. \tag{564}$$

And by (131) and (138), (139), we have

$$\mu_1^{(s-1)} = -\frac{a_{p,0}^{(s-1)}}{k}, \tag{565}$$

and for all $h \in [k-1]$:

$$\mu_{2h+1}^{(s-1)} = \left(1 + O\left(\frac{1}{k}\right)\right) h r_{\mathtt{down},h,1}^{(s-1)} a_{p,2}^{(s-1)} \geqslant \frac{h}{2h+1} a_{p,2}^{(s-1)}, \tag{566}$$

where the second inequality also uses the fact that $a_{b,1}^{(s-1)}$ is larger than $a_{b,0}^{(s-1)}$ and $a_{b,2}^{(s-1)}$ guaranteed by Induction Hypothesis 2. Thus by the gredient expressions in Lemma 7 and the above three relations we have

$$\pi_{1,1}^{(s-1)}(\mathtt{up}) = \frac{\exp\left(\overline{\phi}_{1,\mathtt{up}}^{(s-1)}\right)}{k + \exp\left(\overline{\phi}_{1,\mathtt{up}}^{(s-1)}\right)} \overset{(551)}{=} \frac{\exp\left(-\left(1+\frac{1}{k}\right)a_{q,0}^{(s-1)} + \frac{a_{p,0}^{(s-1)}}{k}\right)}{k + \exp\left(-\left(1+\frac{1}{k}\right)a_{q,0}^{(s-1)} + \frac{a_{p,0}^{(s-1)}}{k}\right)}, \tag{567}$$

$$\begin{aligned}\forall h \in [k-1]: \quad & \frac{k-h}{k}\pi_{1,2h+1}^{(s-1)}(\texttt{up}) \\ &\leqslant \frac{k-h}{k}\frac{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}^{(s-1)} - \frac{h}{2h+1}a_{p,2}^{(s-1)}\right)}{k-h} \\ &\leqslant \frac{\exp\left(-\left(1+\frac{1}{k}\right)a_{q,0}^{(s-1)} - \left(\frac{1}{3}+O\left(\frac{1}{\ln^2 k}\right)\right)a_{p,2}^{(s-1)}\right)}{k} \\ &\leqslant \frac{1}{k^{4/3}}\pi_{1,1}^{(s-1)}(\texttt{up}), \end{aligned} \tag{568}$$

where the second inequality uses (134), the third inequality uses (92). From (568) we know that

$$\pi_{1,1}^{(s-1)}(\texttt{up}) \geqslant k^{1/3}\sum_{h=1}^{k-1}\frac{k-h}{k}\pi_{1,2h+1}^{(s-1)}(\texttt{up}). \tag{569}$$

And (567), (568), (134), (138) also give the following inequality:

$$\sum_{h=1}^{k-1}\frac{k-h}{k}\pi_{1,2h+1}^{(s-1)}(\texttt{up}) \leqslant \frac{\pi_{1,1}^{(s-1)}(\texttt{up})}{\exp\left(0.3a_{p,2}^{(s-1)}\right)}. \tag{570}$$

Similarly, from Lemma 7 and Induction Hypothesis 2 we also know that

$$\begin{aligned}\forall h \in [k-1]: \quad & \pi_{1,2h+1}^{(s-1)}(\texttt{down}_\times) = (1+o \\ & \left(\frac{1}{k}\right))\frac{\exp\left(-r_{\texttt{down},h,1}^{(s-1)}(a_{p,1}^{(s-1)}+a_{p,2}^{(s-1)})\right)}{k-h}, \end{aligned} \tag{571}$$

which further indicates

$$\begin{aligned}\frac{1+o(1)}{k}\exp\left(-r_{\texttt{down}_\times}^{(s-1)}\left(a_{p,1}^{(s-1)}+a_{p,2}^{(s-1)}\right)\right) &\leqslant \sum_{h=1}^{k-1}\frac{k-h}{k}\pi_{1,2h+1}^{(s-1)}(\texttt{down}_\times) \\ &\leqslant \frac{1}{2}\exp\left(-r_{\texttt{down}_\times}^{(s-1)}\left(a_{p,1}^{(s-1)}+a_{p,2}^{(s-1)}\right)\right). \end{aligned} \tag{572}$$

Analogously, we have

$$\begin{aligned}\forall h \in [k-1]: \quad (1+o(1))\frac{k-h}{\exp\left(a_{q,\times}^{(s-1)}-\mu_{2h}^{(s-1)}\right)} &\leqslant 1-\pi_{1,2h}^{(s-1)}(\texttt{up}) \\ &\leqslant (1+o(1))\frac{k}{\exp\left(a_{q,\times}^{(s-1)}-\mu_{2h}^{(s-1)}\right)}, \end{aligned} \tag{573}$$

from which we deduce

$$\begin{aligned}&\sum_{h=1}^{k-1}\frac{k-h}{k}\left(1-\pi_{1,2h}^{(s-1)}(\texttt{up})\right) \\ &\geqslant (1+o(1))\max\left\{\frac{k}{\exp\left(a_{q,\times}^{(s-1)}-\mu_2^{(s-1)}\right)}, \frac{1}{k\exp\left(a_{q,\times}^{(s-1)}-\mu_{2(k-1)}^{(s-1)}\right)}\right\} \\ &\geqslant (1+o(1))\exp\left(-\min\left\{a_{q,\times}^{(s-1)}-\mu_2^{(s-1)}-\ln k, a_{q,\times}^{(s-1)}-\mu_{2(k-1)}^{(s-1)}+\ln k\right\}\right). \end{aligned} \tag{574}$$

Note that by (131) we have

$$\begin{aligned}\min_{h\in[k-1]}\left\{\mu_{2h}^{(s-1)}\right\} &= \min\{\mu_2^{(s-1)}, \mu_{2(k-1)}^{(s-1)}\}, \quad \max_{h\in[k-1]}\left\{\mu_{2h}^{(s-1)}\right\} \\ &= \max\{\mu_2^{(s-1)}, \mu_{2(k-1)}^{(s-1)}\}. \end{aligned} \tag{575}$$

Combining this with (573), we have

$$\sum_{h=1}^{k-1}\frac{k-h}{k}\left(1-\pi_{1,2h}^{(s-1)}(\texttt{up})\right)\leqslant(1+o(1))\sum_{h=1}^{k-1}\frac{k-h}{k}\frac{k}{\exp\left(a_{q,\times}^{(s-1)}-\max\left\{\mu_2^{(s-1)},\mu_{2(k-1)}^{(s-1)}\right\}\right)}$$
$$\leqslant\frac{1+o(1)}{2}k^2\exp\left(-\min\left\{a_{q,\times}^{(s-1)}-\mu_2^{(s-1)},a_{q,\times}^{(s-1)}-\mu_{2(k-1)}^{(s-1)}\right\}\right). \tag{576}$$

By Lemma 7 we also know that

$$\forall h\in[k-1]:\ \pi_{1,2h}^{(s-1)}(\texttt{down}_\times^1)=(1-\pi_{1,2h}^{(s-1)}(\texttt{up}))\frac{\exp\left(-r_{\texttt{up},h,2}^{(s-1)}\left(a_{p,1}^{(s-1)}+a_{p,2}^{(s-1)}\right)\right)}{k-h+h\exp\left(-r_{\texttt{up},h,2}^{(s-1)}\left(a_{p,1}^{(s-1)}+a_{p,2}^{(s-1)}\right)\right)}$$
$$\leqslant\frac{1-\pi_{1,2h}^{(s-1)}(\texttt{up})}{k}, \tag{577}$$

and thus

$$\sum_{h=1}^{k-1}\frac{k-h}{k}\pi_{1,2h}^{(s-1)}(\texttt{down}_\times^1)\leqslant\frac{1}{k}\sum_{h=1}^{k-1}\frac{k-h}{k}\left(1-\pi_{1,2h}^{(s-1)}(\texttt{up})\right). \tag{578}$$

Summing up (537) and (538) and using (569), (558), we have

$$\delta a_{q,0}^{(s-1)}+\delta a_{q,\times}^{(s-1)}=(1+o(1))\left(\pi_{1,1}^{(s-1)}(\texttt{up})+\sum_{h=1}^{k-1}\frac{k-h}{k}\left(1-\pi_{1,2h}^{(s-1)}(\texttt{up})\right)\right). \tag{579}$$

Comparing (579) with (541) and (544), we have

$$|\delta a_{p,0}^{(s-1)}|\lesssim\delta a_{q,0}^{(s-1)}+\delta a_{q,\times}^{(s-1)},\quad |\delta a_{p,\texttt{up}}^{(s-1)}|\lesssim\delta a_{q,0}^{(s-1)}+\delta a_{q,\times}^{(s-1)}. \tag{580}$$

This implies that (138) and (139) hold at step s. Similarly, we can also obtain (recall we define $a_{p,3}^{(s)}\coloneqq\eta\sum_{t=0}^{s-1}\delta a_{p,3}^{(t)}$ in (345))

$$\forall t\leqslant s:\quad |\delta a_{p,3}^{(t)}|\leqslant\frac{1+o(1)}{k}|\delta a_{q,0}^{(s-1)}+\delta a_{q,\times}^{(s-1)}|,\quad |a_{p,3}^{(t)}|\lesssim\frac{1}{k}\left(a_{q,0}^{(t)}+a_{q,\times}^{(t)}\right). \tag{581}$$

Thus by the above expresion, (80) and (134) we know that (140) holds at step s.

Further, by comparing (579), (542) and using (578), we know that

$$\delta a_{p,1}^{(s-1)}\leqslant\frac{1+o(1)}{k}\sum_{h=1}^{k-1}\frac{k-h}{k}\pi_{1,2h}^{(s-1)}(\texttt{down}_\times)+\frac{1+o(1)}{k^2}\left(\delta a_{q,0}^{(s-1)}+\delta a_{q,\times}^{(s-1)}\right), \tag{582}$$

and by (581), (80) we have

$$\left|\delta a_{p,2}^{(s-1)}-\frac{\delta a_{p,1}^{(s-1)}}{k-1}\right|=\left|\frac{\delta a_{p,3}^{(s-1)}}{k-1}\right|\leqslant\frac{1+o(1)}{k^2}\left(\delta a_{q,0}^{(s-1)}+\delta a_{q,\times}^{(s-1)}\right). \tag{583}$$

When $T_{1/2}+1\leqslant s\leqslant T_{1,3}$. By the same argument as how we show (127), (118), (123), and (124) hold in the proof of Lemma 2, we can show (141), (142), (144), (145) and (134) hold at step s. And from (542), (572) we deduce

$$\delta a_{p,1}^{(s-1)}\gtrsim\frac{\exp\left(-\frac{a_{p,1}^{(s-1)}}{2k}\right)}{k^2}\gtrsim\frac{1}{k^{12.5}}. \tag{584}$$

This and Induction Hypothesis 2 imply that (143) holds at step s.

From (539), (537) and (538), we have for any $t \in [T_{1,2}+1, s-1]$:

$$\delta e^{2a_{c,0}^{(t)}} \lesssim \left(a_{q,\times}^{(t)} + a_{q,0}^{(t)}\right)\left(\delta a_{q,0}^{(t)} + \delta a_{q,\times}^{(t)}\right). \tag{585}$$

By (585), our choice of $T_{1,3}$ and (140), (134) we know that if $s \leqslant T_{1,3}$, then

$$e^{2a_{c,0}^{(s)}} \lesssim \ln^2 k. \tag{586}$$

From this and (553), (554) we know that (135) holds at step s.

Let

$$T'_{1,3} := \sup\left\{t \in \mathbb{N} : a_{p,1}^{(t)} \leqslant 20.1k\ln k\right\}, \tag{587}$$

then $T'_{1,3} \in (T_{1/2}, T_{1,3})$. We now prove (150) holds for all $t \in [T'_{1,3}, T_{1,3}]$. We first show there exists $t_0 \in [T_{1/2}+1, T'_{1,3}]$ such that

$$\left(1+\frac{1}{k}\right)a_{q,0}^{(t)} - \frac{a_{p,0}^{(t)}}{k} + 0.05\sqrt{a_{p,1}^{(t)}/k} + 3\ln k \geqslant \min\left\{a_{q,\times}^{(t)} - \mu_2^{(t)}, a_{q,\times}^{(t)} - \mu_{2(k-1)}^{(t)}\right\} \tag{588}$$

holds at step t_0. Suppose otherwise, i.e., for all $t \in [T_{1/2}, T'_{1,3}]$, we have

$$\left(1+\frac{1}{k}\right)a_{q,0}^{(t)} - \frac{a_{p,0}^{(t)}}{k} + 0.05\sqrt{a_{p,1}^{(t)}/k} + 3\ln k < \min\left\{a_{q,\times}^{(t)} - \mu_2^{(t)}, a_{q,\times}^{(t)} - \mu_{2(k-1)}^{(t)}\right\} \tag{589}$$

Then for all $t \in [T_{1/2}, T'_{1,3}]$:

$$\begin{aligned}
\pi_{1,1}^{(t)}(\mathtt{up}) &\overset{(567)}{=} (1+o(1))\frac{1}{k}\exp\left(-\left(1+\frac{1}{k}\right)a_{q,0}^{(t)} + \frac{a_{p,0}^{(t)}}{k}\right) \\
&\geqslant (1+o(1))k^2\exp\left(-\min\left\{a_{q,\times}^{(t)} - \mu_2^{(t)}, a_{q,\times}^{(t)} - \mu_{2(k-1)}^{(t)}\right\} + 0.05\sqrt{a_{p,1}^{(t)}/k}\right) \\
&\overset{(576)}{\geqslant} 2(1+o(1))\exp\left(0.05\sqrt{a_{p,1}^{(t)}/k}\right)\sum_{h=1}^{k-1}\frac{k-h}{k}\left(1-\pi_{1,2h}^{(t)}(\mathtt{up})\right).
\end{aligned} \tag{590}$$

Thus by (590), (537) and (538), we have

$$\forall t \in [T_{1/2}, T'_{1,3}]: \quad \delta a_{q,0}^{(t)} \geqslant 2(1+o(1))\exp\left(0.05\sqrt{a_{p,1}^{(t)}/k}\right)\delta a_{q,\times}^{(t)}. \tag{591}$$

Note that from (117) and (122) we know that

$$a_{q,\times}^{(T_{1/2})} \leqslant 13.3\ln k, \tag{592}$$

and from (141) and our choice of $T_{1,3}, T'_{1,3}$ we know that

$$a_{q,\times}^{(T_{1,3})} \geqslant 14.4\ln k, \quad a_{q,\times}^{(T'_{1,3})} \geqslant 13.95\ln k. \tag{593}$$

By (592), (593) and (591), we have

$$\begin{aligned}
a_{q,0}^{(T'_{1,3})} > a_{q,0}^{(T'_{1,3})} - a_{q,0}^{(T_{1/2})} &\gtrsim \exp\left(0.05\sqrt{a_{p,1}^{(T'_{1,3})}/k}\right)\left(a_{q,\times}^{(T'_{1,3})} - a_{q,\times}^{(T_{1/2})}\right) \\
&\gtrsim \exp\left(0.05\sqrt{a_{p,1}^{(T'_{1,3})}/k}\right)\ln k \gg \ln k.
\end{aligned} \tag{594}$$

But by (134), our choice of $T_{1,3}$ (c.f. (129)) and (140), we have

$$a_{q,0}^{(T'_{1,3})} \asymp a_{p,2}^{(T'_{1,3})} \asymp \ln k. \tag{595}$$

The above two relations contradict each other, and hence there must exist $t_0 \in [T_{1/2}+1, T'_{1,3}]$ such that (588) holds at step t_0. We can choose $t_0 \in [T_{1/2}+1, T'_{1,3}]$ to be the biggest step such that (588) holds, and show that (150) holds for all $t \in [t_0, T_{1,3}]$. Suppose otherwise, i.e., there exists $t_1 \in [t_0+1, T'_{1,3}]$ such that (150) does not hold at step t_1, then by Induction Hypothesis 2 we have

$$\begin{aligned}
&\left(1+\frac{1}{k}\right) a_{q,0}^{(t_1)} + 0.1\sqrt{a_{p,1}^{(t_1)}/k} + 3\ln k + o(1) \\
&= \left(1+\frac{1}{k}\right) a_{q,0}^{(t_1)} - \frac{a_{p,0}^{(t_1)}}{k} + 0.1\sqrt{a_{p,1}^{(t_1)}/k} + 3\ln k \\
&< \min\left\{a_{q,\times}^{(t_1)} - \mu_2^{(t_1)}, a_{q,\times}^{(t_1)} - \mu_{2(k-1)}^{(t_1)}\right\} = a_{q,\times}^{(t_1)} - \frac{a_{p,1}^{(t_1)}}{2k} + o(1). \qquad (596)
\end{aligned}$$

By our choice of t_0, gradient update rule and Induction Hypothesis 2, we have for all $t \in [t_0, t_1-1]$:

$$\left(1+\frac{1}{k}\right) a_{q,0}^{(t)} + 0.05\sqrt{a_{p,1}^{(t)}/k} + 3\ln k + o(1) \leqslant a_{q,\times}^{(t)} - \frac{a_{p,1}^{(t)}}{2k}. \qquad (597)$$

Then similar as how we obtain (591), from the above relation and (537), we can compute that

$$\forall t \in [t_0, t_1-1]: \quad \delta a_{q,0}^{(t)} \gtrsim \exp\left(0.05\sqrt{a_{p,1}^{(t)}/k}\right) \delta a_{q,\times}^{(t)}. \qquad (598)$$

and thus we have

$$\begin{aligned}
&\left(1+\frac{1}{k}\right) a_{q,0}^{(t_1)} + 0.05\sqrt{a_{p,1}^{(t_1)}/k} + 3\ln k - \left(a_{q,\times}^{(t_1)} - \frac{a_{p,1}^{(t_1)}}{2k}\right) \\
&\overset{(588)}{\geqslant} o(1) + \eta \sum_{t=t_0}^{t_1-1} \left(\delta a_{q,0}^{(t)} - \delta a_{q,\times}^{(t)}\right) \\
&\overset{(598)}{\gtrsim} \eta \exp\left(0.05\sqrt{a_{p,1}^{(t_0)}/k}\right) \sum_{t=t_0}^{t_1-1} \delta a_{q,\times}^{(t)} + o(1) > o(1), \qquad (599)
\end{aligned}$$

which implies

$$\left(1+\frac{1}{k}\right) a_{q,0}^{(t_1)} + 0.1\sqrt{a_{p,1}^{(t_1)}/k} + 3\ln k \geqslant a_{q,\times}^{(t_1)} - \frac{a_{p,1}^{(t_1)}}{2k} + 0.05\sqrt{a_{p,1}^{(t_1)}/k} + o(1). \qquad (600)$$

This contradicts (596). Therefore, we have shown that (150) holds for all $t \in [T'_{1,3}, T_{1,3}]$.

Then for all $t \in [T'_{1,3}, s-1]$:

$$\begin{aligned}
&\sum_{h=1}^{k-1} \frac{k-h}{k}\left(1-\pi_{1,2h}^{(t)}(\mathtt{up})\right) \overset{(574)}{\geqslant} (1+o(1)) \exp\Big(-\min\Big\{a_{q,\times}^{(t)} - \mu_2^{(t)} \\
&\qquad -\ln k, a_{q,\times}^{(t)} - \mu_{2(k-1)}^{(t)} + \ln k\Big\}\Big) \\
&\geqslant \frac{1+o(1)}{k} \exp\left(-\min\left\{a_{q,\times}^{(t)} - \mu_2^{(t)}, a_{q,\times}^{(t)} - \mu_{2(k-1)}^{(t)}\right\}\right) \\
&\overset{(150)}{\geqslant} \frac{1+o(1)}{k^4} \exp\left(-\left(1+\frac{1}{k}\right) a_{q,0}^{(t)} + \frac{a_{p,0}^{(t)}}{k} - 0.1\sqrt{a_{p,1}^{(t)}/k}\right) \\
&\overset{(567)}{=} \frac{1+o(1)}{k^3} \exp\left(-0.1\sqrt{a_{p,1}^{(t)}/k}\right) \pi_{1,1}^{(t)}(\mathtt{up}) \\
&\overset{(570)}{\geqslant} \frac{1+o(1)}{k^3} \exp\left(0.3a_{p,2}^{(t)} - 0.1\sqrt{a_{p,1}^{(t)}/k}\right) \sum_{h=1}^{k-1} \frac{k-h}{k} \pi_{1,2h+1}^{(t)}(\mathtt{up}) \\
&\geqslant k^3 \sum_{h=1}^{k-1} \frac{k-h}{k} \pi_{1,2h+1}^{(t)}(\mathtt{up}), \qquad (601)
\end{aligned}$$

where the last inequality follows from (129) and (140), and the third line uses (150), which is proven above to be hold also for all $t \in [T'_{1,3}, T_{1,3}]$. Therefore, by (538) and (558), we have for all $t \in [T'_{1,3}, s-1]$:

$$\delta a^{(t)}_{q,\times} = (1+o(1)) \sum_{h=1}^{k-1} \frac{k-h}{k} \left(1 - \pi^{(t)}_{1,2h}(\texttt{up})\right) > 0, \tag{602}$$

and by the penultimate line of (601), (602) and our choice of $T'_{1,3}$ we know that

$$\forall t \in [T'_{1,3}, s-1]: \quad \sum_{h=1}^{k-1} \frac{k-h}{k} \pi^{(t)}_{1,2h+1}(\texttt{up}) \leqslant \exp\left(-0.1 a^{(t)}_{p,2}\right) \delta a^{(t)}_{q,\times}. \tag{603}$$

We next show (148) holds at step $T_{1,3}$. To do so, we first prove by contradiction that there exists $t_0 \in [T_{1/2}, T_{1,3}]$ such that

$$r^{(t)}_{\texttt{down}_\times}\left(a^{(t)}_{p,1} + a^{(t)}_{p,2}\right) + 0.05\sqrt{a^{(t)}_{p,1}/k} + 5\ln k \geqslant \min\left\{a^{(t)}_{q,\times} - \mu^{(t)}_2, a^{(t)}_{q,\times} - \mu^{(t)}_{2(k-1)}\right\}. \tag{604}$$

Suppose otherwise, i.e., for all $t \in [T_{1/2}, T_{1,3}]$, we have

$$r^{(t)}_{\texttt{down}_\times}\left(a^{(t)}_{p,1} + a^{(t)}_{p,2}\right) + 0.05\sqrt{a^{(t)}_{p,1}/k} + 5\ln k < \min\left\{a^{(t)}_{q,\times} - \mu^{(t)}_2, a^{(t)}_{q,\times} - \mu^{(t)}_{2(k-1)}\right\}, \tag{605}$$

then

$$\begin{aligned}
&\sum_{h=1}^{k-1} \frac{k-h}{k}\left(1 - \pi^{(t)}_{1,2h}(\texttt{up})\right) \overset{(576)}{\leqslant} \frac{1+o(1)}{2} k^3 \exp\Big(-\min\Big\{a^{(s-1)}_{q,\times} \\
&\qquad -\mu^{(s-1)}_2, a^{(s-1)}_{q,\times} - \mu^{(s-1)}_{2(k-1)}\Big\}\Big) \\
&\overset{(605)}{\leqslant} \frac{1+o(1)}{2k^2} \exp\left(-r^{(t)}_{\texttt{down}_\times}\left(a^{(t)}_{p,1} + a^{(t)}_{p,2}\right) - 0.05\sqrt{a^{(t)}_{p,1}/k}\right) \\
&\overset{(572)}{\leqslant} \frac{1+o(1)}{2k^2} \exp\left(-0.05\sqrt{a^{(t)}_{p,1}/k}\right) \sum_{h=1}^{k-1} \frac{k-h}{k} \pi^{(t)}_{1,2h+1}(\texttt{down}_\times),
\end{aligned} \tag{606}$$

which together with (542), (538) gives

$$\forall t \in [T_{1/2}, T_{1,3}]: \quad \delta a^{(t)}_{p,1} \geqslant 2(1+o(1))k \exp\left(0.05\sqrt{a^{(t)}_{p,1}/k}\right) \delta a^{(t)}_{q,\times}. \tag{607}$$

Combining (607), (592) and (592), we can see that

$$\begin{aligned}
a^{(T_{1,3})}_{p,1} &> a^{(T_{1,3})}_{p,1} - a^{(T_{1/2})}_{p,1} \geqslant 2(1+o(1))k \exp\left(0.05\sqrt{a^{(T_{1,3})}_{p,1}/k}\right)\Big(a^{(T_{1,3})}_{q,\times} \\
&\quad - a^{(T_{1/2})}_{q,\times}\Big) \gg k\ln k.
\end{aligned} \tag{608}$$

But by (134) and (140) we have

$$a^{(T_{1,3})}_{p,1} \lesssim k a^{(T_{1,3})}_{q,\times} \lesssim k\ln k. \tag{609}$$

The above two relations contradict each other. Thus there exists $t_0 \in [T_{1/2}, T_{1,3}]$ such that (604) holds at step t_0. If $t_0 = T_{1,3}$, then we are done. Otherwise, we can choose $t_0 \in [T_{1/2}, T_{1,3}]$ to be the biggest step such that (604) holds. We next show that (148) holds for all $t \in [t_0, T_{1,3}]$. Suppose otherwise, then let $t_1 \in [t_0+1, T_{1,3}]$ by the smallest step such that (148) does not hold at step t_1, then by Induction Hypothesis 2 we have

$$\begin{aligned}
&\frac{a^{(t_1)}_{p,1}}{2k} + 0.1\sqrt{a^{(t_1)}_{p,1}/k} + 5\ln k + o(1) = r^{(t_1)}_{\texttt{down}_\times}\left(a^{(t_1)}_{p,1} + a^{(t_1)}_{p,2}\right) + 0.1\sqrt{a^{(t_1)}_{p,1}/k} + 5\ln k \\
&< \min\left\{a^{(t_1)}_{q,\times} - \mu^{(t_1)}_2, a^{(t_1)}_{q,\times} - \mu^{(t_1)}_{2(k-1)}\right\} = a^{(t_1)}_{q,\times} - \frac{a^{(t_1)}_{p,1}}{2k} + o(1).
\end{aligned} \tag{610}$$

By our choice of t_0, gradient update rule and Induction Hypothesis 2, we have for all $t \in [t_0, t_1 - 1]$:

$$\frac{a_{p,1}^{(t)}}{2k} + 0.05\sqrt{a_{p,1}^{(t)}/k} + 5\ln k + o(1) \leqslant a_{q,\times}^{(t)} - \frac{a_{p,1}^{(t)}}{2k}. \tag{611}$$

and thus similar as (607), we can compute that

$$\forall t \in [t_0, t_1 - 1]: \quad \delta a_{p,1}^{(t)} \gtrsim k \exp\left(0.05\sqrt{a_{p,1}^{(t)}/k}\right) \delta a_{q,\times}^{(t)}. \tag{612}$$

Then

$$\begin{aligned}
&\frac{a_{p,1}^{(t_1)}}{2k} + 0.05\sqrt{a_{p,1}^{(t_1)}/k} + 5\ln k - \left(a_{q,\times}^{(t_1)} - \frac{a_{p,1}^{(t_1)}}{2k}\right) \\
&\overset{(604)}{\geqslant} o(1) + \eta \sum_{t=t_0}^{t_1-1} \left(\frac{1}{k}\delta a_{p,1}^{(t)} - \delta a_{q,\times}^{(t)}\right) \\
&\overset{(612)}{\gtrsim} \eta \sum_{t=t_0}^{t_1-1} \left(\exp\left(0.05\sqrt{a_{p,1}^{(t)}/k}\right) - 1\right) \delta a_{q,\times}^{(t)} + o(1) \\
&\gtrsim \eta \left(\exp\left(0.05\sqrt{a_{p,1}^{(t_0)}/k}\right) - 1\right) \sum_{t=t_0}^{t_1-1} \delta a_{q,\times}^{(t)} + o(1) \geqslant o(1).
\end{aligned} \tag{613}$$

This together with our choice of $T_{1/2}$ suggests

$$\frac{a_{p,1}^{(t_1)}}{2k} + 0.1\sqrt{a_{p,1}^{(t_1)}/k} + 5\ln k \geqslant a_{q,\times}^{(t_1)} - \frac{a_{p,1}^{(t_1)}}{2k} + 0.05\sqrt{a_{p,1}^{(t_1)}/k} + o(1), \tag{614}$$

which contradicts (610). Therefore, we have shown that (148) holds for all $t \in [t_0, T_{1,3}]$.

We next show (149) holds at step $T_{1,3}$. First, if $a_{q,0}^{(T_{1,3})} \leqslant 12\ln k$, then by our choice of $T_{1,3}$, (140), (138) and the fact that $r_{\mathtt{down}_\times}^{(t)} \geqslant \frac{1}{2k}$ for all $t \in [T_{1/2}, T_{1,3}]$ guaranteed by Induction Hypothesis 2, we have (149) holds at step $T_{1,3}$. Below we assume

$$a_{q,0}^{(T_{1,3})} > 12\ln k. \tag{615}$$

We first show that there exists $t_0 \in [T_{1/2}, T_{1,3}]$ such that

$$r_{\mathtt{down}_\times}^{(t_0)}\left(a_{p,1}^{(t_0)} + a_{p,2}^{(t_0)}\right) + 2.05\ln k \geqslant \left(1 + \frac{1}{k}\right) a_{q,0}^{(t_0)} - \frac{a_{p,0}^{(t_0)}}{k}. \tag{616}$$

Suppose for all $t \in [T_{1/2}, T_{1,3}]$, we have

$$r_{\mathtt{down}_\times}^{(t)}\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right) + 2.05\ln k < \left(1 + \frac{1}{k}\right) a_{q,0}^{(t)} - \frac{a_{p,0}^{(t)}}{k}, \tag{617}$$

then for all $t \in [T_{1/2}, T_{1,3}]$, we have

$$\begin{aligned}
\sum_{h=1}^{k-1} \frac{k-h}{k} \pi_{1,2h+1}^{(t)}(\mathtt{down}_\times) &\overset{(572)}{\geqslant} \frac{1+o(1)}{k} \exp\left(-r_{\mathtt{down}_\times}^{(t)}\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right)\right) \\
&\geqslant (1+o(1))\, k^{2.05} \frac{\exp\left(-\left(1+\frac{1}{k}\right) a_{q,0}^{(t)} + \frac{a_{p,0}^{(t)}}{k}\right)}{k} \\
&\overset{(567)}{=} k^{2.05} \pi_{1,1}^{(t)}(\mathtt{up}).
\end{aligned} \tag{618}$$

This combined with (542) and (537) suggests

$$\delta a_{p,1}^{(t)} \geqslant (1+o(1)) k^{1.05} \pi_{1,1}^{(t)}(\mathtt{up}) \geqslant (1+o(1)) k^{1.05} \delta a_{q,0}^{(t)}. \tag{619}$$

Note that by (117) and (122) we know that

$$a_{q,0}^{(T_{1/2})} \leqslant 10.2 \ln k + C/2, \tag{620}$$

Therefore, by (620), (615) and (619), we have

$$a_{p,1}^{(T_{1,3})} > a_{p,1}^{(T_{1,3})} - a_{p,1}^{(T_{1/2})} \gtrsim k^{1.05} \left(a_{q,0}^{(T_{1,3})} - a_{q,0}^{(T_{1/2})} \right) \gtrsim k^{1.05} \ln k, \tag{621}$$

which contradicts our choice of $T_{1,3}$ (c.f. (129)). Therefore, there exists $t_0 \in [T_{1/2}, T_{1,3}]$ such that (616) holds. Then similar as how we show (148) or (150) holds at step $T_{1,3}$ above, building upon (616), we can show by contradiction that (149) holds at step $T_{1,3}$.

We now prove that (147) holds at step $T_{1,3}$. We first show that there exists $t_0 \in [T'_{1,3}, T_{1,3}]$ such that

$$\begin{aligned} &\min \left\{ a_{q,\times}^{(t_0)} - \mu_2^{(t_0)} - \ln k, a_{q,\times}^{(t_0)} - \mu_{2(k-1)}^{(t_0)} + \ln k \right\} + \frac{2}{k} a_{q,0}^{(t_0)} \geqslant r_{\mathtt{down}_\times}^{(t_0)} \left(a_{p,1}^{(t_0)} \right. \\ &\left. + a_{p,2}^{(t_0)} \right) + 1.95 \ln k. \end{aligned} \tag{622}$$

Suppose for all $t \in [T'_{1,3}, T_{1,3}]$, we have

$$\begin{aligned} &\min \left\{ a_{q,\times}^{(t)} - \mu_2^{(t)} - \ln k, a_{q,\times}^{(t)} - \mu_{2(k-1)}^{(t)} + \ln k \right\} + \frac{2}{k} a_{q,0}^{(t)} < r_{\mathtt{down}_\times}^{(t)} \left(a_{p,1}^{(t)} \right. \\ &\left. + a_{p,2}^{(t)} \right) + 1.95 \ln k. \end{aligned} \tag{623}$$

Then for all $t \in [T'_{1,3}, T_{1,3}]$, we have

$$\begin{aligned} &\sum_{h=1}^{k-1} \frac{k-h}{k} \left(1 - \pi_{1,2h}^{(t)}(\mathtt{up}) \right) \\ &\overset{(574)}{\geqslant} (1 + o(1)) \exp \left(- \min \left\{ a_{q,\times}^{(t)} - \mu_2^{(t)} - \ln k, a_{q,\times}^{(t)} - \mu_{2(k-1)}^{(t)} + \ln k \right\} \right) \\ &\geqslant (1 + o(1)) \exp \left(-r_{\mathtt{down}_\times}^{(t)} \left(a_{p,1}^{(t)} + a_{p,2}^{(t)} \right) - 1.95 \ln k + \frac{2}{k} a_{q,0}^{(t)} \right) \\ &\overset{(572)}{\geqslant} \frac{2 + o(1)}{k^{1.95}} \sum_{h=1}^{k-1} \frac{k-h}{k} \pi_{1,2h+1}^{(t)}(\mathtt{down}_\times), \end{aligned} \tag{624}$$

and by (577) our choice of $T'_{1,3}$ we have

$$\sum_{h=1}^{k-1} \frac{k-h}{k} \pi_{1,2h}^{(t)}(\mathtt{down}_\times^1) = o\left(\frac{1}{k^{1.95}} \right) \sum_{h=1}^{k-1} \frac{k-h}{k} \left(1 - \pi_{1,2h}^{(t)}(\mathtt{up}) \right). \tag{625}$$

the above two expressions and (542), (602) we have for all $t \in [T'_{1,3}, T_{1,3}]$:

$$\delta a_{p,1}^{(t)} \leqslant (2 + o(1)) k^{0.95} \delta a_{q,\times}^{(t)}, \tag{626}$$

from which and our choice of $T_{1,3}, T'_{1,3}$ we deduce

$$a_{q,\times}^{(T_{1,3})} > a_{q,\times}^{(T_{1,3})} - a_{q,\times}^{(T'_{1,3})} \gtrsim k^{-0.95} \left(a_{p,1}^{(T_{1,3})} - a_{p,1}^{(T'_{1,3})} \right) \gtrsim k^{0.05} \ln k, \tag{627}$$

which contradicts that fact that $a_{q,\times}^{(T_{1,3})} \asymp \ln k$ guaranteed by (134), (140) and our choice of $T_{1,3}$. Therefore, there exists $t_0 \in [T'_{1,3}, T_{1,3}]$ such that (622) holds. Then similar as how we show (148) or (150) holds at step $T_{1,3}$ above, building upon (622), we can show by contradiction that (147) holds at step $T_{1,3}$.

When $T_{1,3} + 1 \leqslant s \leqslant T_{1,4}$. We first show relations between $a_{b,0}^{(s)}$, $a_{b,1}^{(s)}$ and $a_{b,2}^{(s)}$. First note that (601), (602) and (603) still hold here, and we'll frequently use them in the subsequent proof.

First, from (603) and (539) and (134) we know that for all $t \in [T_{1,3}, s-1]$:

$$\delta e^{2a_{c,0}^{(t)}} \lesssim \frac{a_{p,2}^{(t)}}{\exp\left(0.1 a_{p,2}^{(t)}\right)} \delta a_{q,\times}^{(t)}. \tag{628}$$

Then (i) if $a_{p,1}^{(s)} \leqslant k^2$, then by (628), (138), (134) and our choice of $T_{1,3}$ we have

$$e^{2a_{c,0}^{(s)}} - e^{2a_{c,0}^{(T_{1,3})}} \lesssim \frac{21 \ln k}{\exp\left(0.1 \times 21 \ln k\right)} k \lesssim \frac{\ln k}{k^{1.1}}. \tag{629}$$

And thus by (586) and the above relation we know that if $a_{p,1}^{(s)} \leqslant k^2$:

$$e^{2a_{c,0}^{(s)}} \lesssim \ln^2 k. \tag{630}$$

(ii) if $a_{p,1}^{(s)} > k^2$, then we can choose t_0 such that

$$t_0 := \sup\left\{t \in [T_{1,3}, s-1] : a_{p,1}^{(t)} = k^2\right\}. \tag{631}$$

Then above we have shown

$$e^{2a_{c,0}^{(t_0)}} - e^{2a_{c,0}^{(T_{1,3})}} \lesssim \frac{\ln k}{k^{1.1}}. \tag{632}$$

From (628), (146) and (140), (134) we deduce

$$e^{2a_{c,0}^{(s)}} - e^{2a_{c,0}^{(t_0)}} \lesssim \frac{k}{\exp\left(0.1k\right)} k^{5/4}\sqrt{N} = o(1). \tag{633}$$

By the two relations above and (586) we know that

$$e^{2a_{c,0}^{(s)}} \lesssim \ln^2 k. \tag{634}$$

Therefore, from (634) and (553), (554) we know that (135) holds at step s.

By (548) and Induction Hypothesis 2 we have for any $t \in [T_{1/2}, s-1]$:

$$\begin{aligned} I_{h,1}^{(t)} = &\left(1 + O\left(\frac{1}{k}\right)\right)\left(1 - h r_{\mathtt{down},h,1}^{(t)}\right) \\ &\cdot \left(k a_{p,2}^{(t)} \frac{h e^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + h e^{a_{b,1}^{(t)}} + h e^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(\mathtt{down}_\times) + h r_{\mathtt{down},h,1}^{(t)} a_{p,2}^{(t)} \pi_{1,2h+1}^{(t)}(\mathtt{up})\right). \end{aligned} \tag{635}$$

From (158) and (133) in Induction Hypothesis 2 we know that

$$\forall t \in [T_{1/2}, s-1]: \quad 0 < a_{b,1}^{(t)} \lesssim \ln k. \tag{636}$$

From our choice of $T_{1,4}$ in (130), (152) and (145) we know that

$$\forall t \in [T_{1/2}, s-1]: \quad |a_{b,0}^{(t)}| \lesssim \ln k. \tag{637}$$

Therefore, by the above two expressions and (216) we know that

$$\forall t \in [T_{1/2}, s-1]: \quad |a_{b,2}^{(t)}| \lesssim \frac{\ln k}{N}. \tag{638}$$

By our choice of $T_{1,4}$ in (130), (151) in Induction Hypothesis 2 and (638) we have

$$1 - h r_{\mathtt{down},h,1}^{(t)} \overset{(563)}{=} \frac{e^{a_{b,0}^{(t)}} + h e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + h e^{a_{b,1}^{(t)}} + h e^{a_{b,2}^{(t)}}} \lesssim \frac{N}{\left(a_{p,1}^{(t)}\right)^2 \ln k}. \tag{639}$$

Combining (635), (639), (542) and (603) we obtain that for all $t \in [T_{1,3}, s-1]$:

$$0 < \delta a_{b,1}^{(t)} \lesssim \frac{N}{\left(a_{p,1}^{(t)}\right)^2 \ln k}\left(\frac{k^2 a_{p,2}^{(t)}}{N}\delta a_{p,1}^{(t)} + \frac{a_{p,2}^{(t)}}{N\exp\left(0.1 a_{p,2}^{(t)}\right)}\delta a_{q,\times}^{(t)}\right)$$
$$\overset{(140)}{\lesssim} \frac{1}{a_{p,2}^{(t)}\ln k}\delta a_{p,1}^{(t)} + \frac{1}{k^2 \ln k a_{p,2}^{(t)}\exp\left(0.1 a_{p,2}^{(t)}\right)}\delta a_{q,\times}^{(t)}. \tag{640}$$

Similar as (319), we define

$$\forall i \in \{0,1,2\}: \quad \delta e^{a_{b,i}^{(t)}} := \frac{e^{a_{b,i}^{(t+1)}} - e^{a_{b,i}^{(t)}}}{\eta} = \frac{e^{a_{b,i}^{(t)}}}{\eta}\left(e^{\eta\delta a_{b,i}^{(t)}} - 1\right) = (1 + o(1))e^{a_{b,i}^{(t)}}\delta a_{b,i}^{(t)}. \tag{641}$$

Then by (545), (546), (638), (140) and (602) we know that for all $t \in [T_{1/2}, s-1]$:

$$\delta e^{a_{b,0}^{(t)}} \leqslant (1 + o(1))\frac{a_{p,1}^{(t)}}{N}\delta a_{q,\times}^{(t)}, \tag{642}$$

which suggests

$$e^{a_{b,0}^{(s-1)}} - e^{a_{b,0}^{(T_{1,3})}} \leqslant \frac{1 + o(1)}{N} a_{p,1}^{(s-1)} a_{q,\times}^{(s-1)} \lesssim \frac{1}{Nk}\left(a_{p,1}^{(s-1)}\right)^2, \tag{643}$$

where the last inequality follows from (140) and (134). Thus by the above expression and (145), there exists constant $c > 0$ such that

$$e^{a_{b,0}^{(s-1)}} \leqslant 1 + \frac{c}{Nk}\left(a_{p,1}^{(s-1)}\right)^2. \tag{644}$$

And by (635) we have for all $t \in [T_{1/2}, s-1]$ and $h \in [k-1]$:

$$\begin{aligned}
&e^{a_{b,1}^{(t)}} I_{h,1}^{(t)} \\
&\geqslant \left(1 + O\left(\frac{1}{k}\right)\right)\frac{e^{a_{b,1}^{(t)}}\left(e^{a_{b,0}^{(t)}} + he^{a_{b,2}^{(t)}}\right)}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} a_{p,1}^{(t)} \cdot \frac{he^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \\
&\quad \pi_{1,2h+1}^{(t)}(\mathtt{down}_\times) \\
&\overset{(638)}{\geqslant} \left(1 + O\left(\frac{1}{k}\right)\right)\frac{he^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} a_{p,1}^{(t)} \cdot \frac{he^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \\
&\quad \pi_{1,2h+1}^{(t)}(\mathtt{down}_\times) \\
&\geqslant \left(1 + O\left(\frac{1}{k}\right)\right)\frac{a_{p,1}^{(t)}}{3}\left(\frac{he^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}}\pi_{1,2h}^{(t)}(\mathtt{down}_\times^1)\right. \\
&\quad + \frac{he^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}}\pi_{1,2h+1}^{(t)}(\mathtt{down}_\times) - \frac{he^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}} \\
&\quad \left.\pi_{1,2h}^{(t)}(\mathtt{down}_\times^1)\right),
\end{aligned} \tag{645}$$

where the first inequality uses (140), and the third inequality uses (151) in Induction Hypothesis 2 and (638). By (644), the first relation in (577) and (146) we know that

$$\forall h \in [k-1]: \quad \pi_{1,2h}^{(s-1)}(\mathtt{down}_\times^1) \leqslant \exp\left(-r_{\mathtt{up},h,2}^{(s-1)}\left(a_{p,1}^{(s-1)} + a_{p,2}^{(s-1)}\right)\right)(1 - \pi_{1,2h}^{(s-1)}(\mathtt{up}))$$

$$\leqslant \exp\left(-\frac{a_{p,1}^{(s-1)}+a_{p,2}^{(s-1)}}{2h+\frac{c}{Nk}\left(a_{p,1}^{(s-1)}\right)^2}\right)\left(1-\pi_{1,2h}^{(s-1)}(\mathtt{up})\right)$$

$$\leqslant \exp\left(-\left(1+O\left(\frac{1}{k}\right)\right)\frac{1}{2\frac{k}{a_{p,1}^{(s-1)}}+\frac{c}{Nk}a_{p,1}^{(s-1)}}\right)\left(1-\pi_{1,2h}^{(s-1)}(\mathtt{up})\right). \tag{646}$$

By our choice of $T_{1,3}$, (146) and the fact that $a_{p,1}^{(t)}$ strictly increases, we have

$$\forall t\in[T_{1,3},s-1]:\quad 21k\ln k\leqslant a_{p,1}^{(t)}\leqslant k^{9/4}\sqrt{N}. \tag{647}$$

Combining the above two inequalities, we have (i) if $a_{p,1}^{(s-1)}\leqslant k^2$, then

$$\pi_{1,2h}^{(s-1)}(\mathtt{down}_\times^1)\leqslant\frac{1}{k^{10}}\left(1-\pi_{1,2h}^{(s-1)}(\mathtt{up})\right)\leqslant\frac{1-\pi_{1,2h}^{(s-1)}(\mathtt{up})}{k^8a_{p,1}^{(s-1)}}, \tag{648}$$

and (ii) if $a_{p,1}^{(s-1)}\geqslant k^2$, then

$$\pi_{1,2h}^{(s-1)}(\mathtt{down}_\times^1)\leqslant e^{-k}\left(1-\pi_{1,2h}^{(s-1)}(\mathtt{up})\right)\overset{(146)}{\leqslant}\frac{1-\pi_{1,2h}^{(t)}(\mathtt{up})}{k^8a_{p,1}^{(s-1)}}. \tag{649}$$

Thereafter, we have

$$\forall h\in[k-1],\ t\in[T_{1/2},s-1]:\quad \pi_{1,2h}^{(t)}(\mathtt{down}_\times^1)\leqslant\frac{1}{k^8a_{p,1}^{(t)}}\left(1-\pi_{1,2h}^{(t)}(\mathtt{up})\right). \tag{650}$$

By (645), (641), (542), (650) and (602) we have for all $t\in[T_{1,3},s-1]$:

$$\delta e^{a_{b,1}^{(t)}}\geqslant\frac{1+o(1)}{N}\left(\frac{ka_{p,1}^{(t)}}{3}\delta a_{p,1}^{(t)}-\frac{1}{3k^8}\delta a_{q,\times}^{(t)}\right). \tag{651}$$

This indicates that (153) holds at step s.

On the other hand, similar as how we obtain (640), from (635) we can obtain that for any $t\in[T_{1,3},s-1]$,

$$\begin{aligned}\delta e^{a_{b,1}^{(t)}}&\leqslant(1+o(1))\frac{e^{a_{b,1}^{(t)}}\left(e^{a_{b,0}^{(t)}}+he^{a_{b,2}^{(t)}}\right)}{e^{a_{b,0}^{(t)}}+he^{a_{b,1}^{(t)}}+he^{a_{b,2}^{(t)}}}\left(\frac{k^2a_{p,2}^{(t)}}{N}\delta a_{p,1}^{(t)}+\frac{a_{p,2}^{(t)}}{N\exp\left(0.1a_{p,2}^{(t)}\right)}\delta a_{q,\times}^{(t)}\right)\\&\leqslant(1+o(1))\frac{k}{\ln k}\left(\frac{k^2a_{p,2}^{(t)}}{N}\delta a_{p,1}^{(t)}+\frac{a_{p,2}^{(t)}}{Nk^2}\delta a_{q,\times}^{(t)}\right)\\&\overset{(140)}{=}(1+o(1))\frac{k}{\ln k}\left(\frac{ka_{p,1}^{(t)}}{N}\delta a_{p,1}^{(t)}+\frac{a_{p,1}^{(t)}}{Nk^3}\delta a_{q,\times}^{(t)}\right),\end{aligned} \tag{652}$$

where the second line follows from (638), and our choice of $T_{1,4}$ and $T_{1/2}$. Therefore, we have

$$\begin{aligned}e^{a_{b,1}^{(s)}}-e^{a_{b,1}^{(T_{1,3})}}&\leqslant(1+o(1))\frac{k}{N\ln k}a_{p,1}^{(s)}\left(k\left(a_{p,1}^{(s)}-a_{p,1}^{(T_{1,3})}\right)+\frac{1}{k^3}\left(a_{q,\times}^{(s)}-a_{q,\times}^{(T_{1,3})}\right)\right)\\&\leqslant(1+o(1))\frac{k^2}{N\ln k}\left(a_{p,1}^{(s)}\right)^2\overset{(146)}{\lesssim}\frac{k^{6.5}}{\ln k},\end{aligned} \tag{653}$$

where the second inequality follows from (134) and (140). This shows (154) holds at step s.

We now show (151) holds at step s. First, when $s=T_{1,3}$, by (145) and (144) we know that

$$e^{a_{b,0}^{(T_{1,3})}}-1<0<\frac{e^{a_{b,1}^{(T_{1,3})}}-1}{k^2}, \tag{654}$$

and thus (151) holds at step s. If $s > T_{1,3}$, then by setting $t = s$ in (153), and setting $t' \in [T_{1,3}, s-1]$ such that

$$\frac{1}{2}a_{p,1}^{(t)} \leqslant a_{p,1}^{(t')} \leqslant \frac{1}{2}a_{p,1}^{(t)} + 1, \tag{655}$$

we obtain that

$$e^{a_{b,1}^{(s)}} - 1 \geqslant e^{a_{b,1}^{(s)}} - e^{a_{b,1}^{(t')}} \gtrsim \frac{k}{N}\left(a_{p,1}^{(s)}\right)^2 \overset{(644)}{\gtrsim} k^2\left(e^{a_{b,0}^{(s)}} - 1\right). \tag{656}$$

This shows (151) holds at step s.

Moreover, by (650), (546) and Induction Hypothesis 2 we have

$$\begin{aligned} I_{h,0}^{(t)} \coloneqq & \left(1 + O\left(\frac{1}{k}\right)\right) r_{\mathtt{up},h,0}^{(t)} r_{\mathtt{up},h,2}^{(t)} h a_{p,2}^{(t)} \left(1 - \pi_{1,2h}^{(t)}(\mathtt{up})\right) \\ & - \left(1 + O\left(\frac{1}{k}\right)\right) h r_{\mathtt{down},h,0}^{(t)} r_{\mathtt{down},h,1}^{(t)} \left(a_{p,1}^{(t)} \pi_{1,2h+1}^{(t)}(\mathtt{down}_\times) + a_{p,2}^{(t)} \pi_{1,2h+1}^{(t)}(\mathtt{up})\right). \end{aligned} \tag{657}$$

Therefore, by (602) and (657) we have for all $t \leqslant s-1$:

$$\delta a_{b,0}^{(t)} \leqslant \frac{1 + o(1)}{4N} a_{p,2}^{(t)} \delta a_{q,\times}^{(t)}, \tag{658}$$

where we use the fact that $h r_{\mathtt{up},h,2}^{(t)} \leqslant 1 - r_{\mathtt{up},h,0}^{(t)}$, and $x(1-x) \leqslant \frac{1}{4}$ for any $x \in \mathbb{R}$. And by (542) and (603) we have for all $t \in [T_{1,3}, s-1]$:

$$\begin{aligned} -\delta a_{b,0}^{(t)} &\leqslant (1 + o(1)) r_{\mathtt{down},h,0}^{(t)} \left(\frac{k^2 a_{p,2}^{(t)}}{N} \delta a_{p,1}^{(t)} + \frac{a_{p,2}^{(t)}}{N \exp\left(0.1 a_{p,2}^{(t)}\right)} \delta a_{q,\times}^{(t)}\right) \\ &\lesssim \frac{1}{a_{p,2}^{(t)} \ln k} \delta a_{p,1}^{(t)} + \frac{1}{k^2 \ln k a_{p,2}^{(t)} \exp\left(0.1 a_{p,2}^{(t)}\right)} \delta a_{q,\times}^{(t)}, \end{aligned} \tag{659}$$

where the last relation follows from

$$r_{\mathtt{down},h,0}^{(t)} \lesssim \frac{N}{\left(a_{p,1}^{(t)}\right)^2 \ln k} \tag{660}$$

guaranteed by our choice of $T_{1,4}$ and (151).

We next show (152) holds at step s. First we have

$$\begin{aligned} \delta e^{-a_{b,0}^{(t)}} &\coloneqq \frac{e^{-a_{b,0}^{(t+1)}} - e^{-a_{b,0}^{(t)}}}{\eta} \\ &= (1 + o(1)) e^{-a_{b,0}^{(t)}} \left(-\delta a_{b,0}^{(t)}\right) \\ &\overset{(657)}{=} \frac{1 + o(1)}{N} \sum_{h=1}^{k-1} \frac{k-h}{k} \frac{h e^{a_{b,1}^{(t)}}}{\left(e^{a_{b,0}^{(t)}} + h e^{a_{b,1}^{(t)}} + h e^{a_{b,2}^{(t)}}\right)^2} \left(a_{p,1}^{(t)} \pi_{1,2h+1}^{(t)}(\mathtt{down}_\times) \right. \\ &\quad \left. + a_{p,2}^{(t)} \pi_{1,2h+1}^{(t)}(\mathtt{up})\right) \\ &\quad - \frac{1 + o(1)}{N} \sum_{h=1}^{k-1} \frac{k-h}{k} \frac{h e^{a_{b,2}^{(t)}}}{\left(e^{a_{b,0}^{(t)}} + (2h-1) e^{a_{b,2}^{(t)}}\right)^2} a_{p,2}^{(t)} \left(1 - \pi_{1,2h}^{(t)}(\mathtt{up})\right). \end{aligned} \tag{661}$$

Thus our choice of $T_{1,3}, T_{1,4}$, (638), (603), (602) and (541) we have

$$\delta e^{-a_{b,0}^{(t)}} \leqslant \frac{1 + o(1)}{N} \frac{1}{e^{a_{b,0}^{(t)}} + e^{a_{b,1}^{(t)}} + 1} \sum_{h=1}^{k-1} \frac{k-h}{k} \frac{h e^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + h e^{a_{b,1}^{(t)}} + h e^{a_{b,2}^{(t)}}} \left(a_{p,1}^{(t)}\right.$$

$$\pi^{(t)}_{1,2h+1}(\texttt{down}_\times) + a^{(t)}_{p,2}\pi^{(t)}_{1,2h+1}(\texttt{up})\Big)$$
$$- \frac{1+o(1)}{N}\left(\frac{\ln k}{k}\right)^2 \sum_{h=1}^{k-1}\frac{k-h}{k}a^{(t)}_{p,2}\left(1-\pi^{(t)}_{1,2h}(\texttt{up})\right)$$
$$\leqslant \frac{1+o(1)}{N}\frac{1}{e^{a^{(t)}_{b,0}}+e^{a^{(t)}_{b,1}}+1}\left(ka^{(t)}_{p,1}\delta a^{(t)}_{p,1} + \frac{a^{(t)}_{p,2}}{\exp\left(0.1a^{(t)}_{p,2}\right)}\delta a^{(t)}_{q,\times}\right) - \frac{1+o(1)}{N}$$
$$\left(\frac{\ln k}{k}\right)^2 a^{(t)}_{p,2}\delta a^{(t)}_{q,\times}$$
$$\overset{(140)}{\leqslant} \frac{1+o(1)}{N}\frac{ka^{(t)}_{p,1}}{e^{a^{(t)}_{b,0}}+e^{a^{(t)}_{b,1}}+1}\delta a^{(t)}_{p,1} - \frac{1+o(1)}{N}\frac{\ln^2 k}{k^3}a^{(t)}_{p,1}\delta a^{(t)}_{q,\times}. \tag{662}$$

Define

$$t_i := \inf\left\{t\in[T_{1,3}, s-1] : e^{a^{(t)}_{b,1}} \geqslant k^{i/2}\right\},\ \forall i\in\mathbb{N}. \tag{663}$$

then by (144) we know that $t_0 = T_{1,3}$. Comparing (662) and (651), we have for any $t\in[T_{1,3}, s-1]$, $i\in\mathbb{N}$:

$$\text{If } t_i \leqslant t \leqslant t_{i+1}: \quad \delta e^{-a^{(t)}_{b,0}} \lesssim k^{-i/2}\delta e^{a^{(t)}_{b,1}}, \tag{664}$$

and by (154) and the fact that $a^{(t)}_{b,1}$ monotonically increases (c.f. (133)), there exists positive integer $j\leqslant 12$ such that $k^{j/2}\leqslant e^{a^{(s)}_{b,1}}\leqslant k^{(j+1)/2}$ holds at step s. Thus by (664) we have

$$e^{-a^{(s)}_{b,0}} - e^{-a^{(T_{1,3})}_{b,0}} \lesssim k^{-j/2}\left(e^{a^{(s)}_{b,1}} - e^{a^{(t_j)}_{b,1}}\right) + \sum_{i=0}^{j-1}k^{-i/2}\left(e^{a^{(t_{i+1})}_{b,1}} - e^{a^{(t_i)}_{b,1}}\right) \overset{(663)}{\lesssim} \sqrt{k}, \tag{665}$$

which combined with (145) suggests (152) holds at step s.

Now we show (146) holds at step s by contradiction. Suppose

$$a^{(s)}_{p,1} > k^{9/4}\sqrt{N}, \tag{666}$$

then we can choose $s_0\in[T_{1,3}, s-1]$ such that

$$\frac{a^{(s)}_{p,1}}{\ln k} \leqslant a^{(s_0)}_{p,1} \leqslant \frac{a^{(s)}_{p,1}}{\ln k} + 1. \tag{667}$$

Then by (151) we have

$$e^{a^{(s_0)}_{b,1}} \gtrsim \frac{k}{N}\left(a^{(s_0)}_{p,1}\right)^2 \gtrsim \frac{k^{5.5}}{\ln^2 k}. \tag{668}$$

Plugging this into (662), we obtain

$$\forall t\in[s_0, s-1]: \quad \delta e^{-a^{(t)}_{b,0}} \leqslant \frac{c}{N}\frac{a^{(s_0)}_{p,1}\ln^2 k}{k^{4.5}}\delta a^{(t)}_{p,1} - \frac{1+o(1)}{N}\frac{\ln^2 k}{k^3}a^{(s)}_{p,1}\delta a^{(t)}_{q,\times}$$
$$\overset{(667)}{\leqslant} (1+o(1))\frac{c}{N}\frac{\ln k}{k^{4.5}}a^{(s)}_{p,1}\delta a^{(t)}_{p,1} - \frac{1+o(1)}{N}\frac{\ln^2 k}{k^3}a^{(s)}_{p,1}\delta a^{(t)}_{q,\times} \tag{669}$$

with some constant $c>0$. This implies

$$e^{-a^{(s)}_{b,0}} - e^{-a^{(s_0)}_{b,0}} \leqslant -\frac{1+o(1)}{N}\frac{\ln^2 k}{k^3}a^{(s)}_{p,1}\left(a^{(s)}_{q,\times} - a^{(s_0)}_{q,\times}\right) + (1+o(1))\frac{c}{N}\frac{\ln k}{k^{4.5}}a^{(s)}_{p,1}$$
$$\left(a^{(s)}_{p,1} - a^{(s_0)}_{p,1}\right). \tag{670}$$

By (134), (140), we have

$$a^{(s)}_{q,\times} \asymp \frac{a^{(s)}_{p,1}}{k} \overset{(666)}{\gtrsim} k^{5/4}\sqrt{N}, \quad a^{(s_0)}_{q,\times} \asymp \frac{a^{(s_0)}_{p,1}}{k} \overset{(667)}{\asymp} \frac{a^{(s)}_{p,1}}{k\ln k} \asymp \frac{a^{(s)}_{q,\times}}{\ln k}. \tag{671}$$

Combining the above two expressions, we have

$$e^{-a_{b,0}^{(s)}}-e^{-a_{b,0}^{(s_0)}}\leqslant-\frac{1+o(1)}{N}\frac{\ln^2 k}{k^3}a_{p,1}^{(s)}a_{q,\times}^{(s)}\overset{(666)}{\lesssim}-\frac{1}{N}\frac{\ln^2 k}{k^4}\left(k^{9/4}\sqrt{N}\right)^2=\sqrt{k}\ln^2 k, \tag{672}$$

which together with (152) suggests there exists $c'>0$ such that

$$e^{-a_{b,0}^{(s)}}\leqslant c'\sqrt{k}-\sqrt{k}\ln^2 k<0. \tag{673}$$

This is impossible. Therefore, (146) holds at step s.

In addition, by (146), (542), (572), (140), (701) we deduce

$$\delta a_{p,1}^{(t)}\gtrsim\frac{\exp\left(-\left(1+\frac{2}{k}\right)\frac{a_{p,1}^{(t)}}{k}\right)}{k^2}\gtrsim\frac{\exp\left(-\left(1+\frac{2}{k}\right)k^{5/4}\sqrt{N}\right)}{k^2}\geqslant\exp\left(-k^{1.3}\sqrt{N}\right) \tag{674}$$

for any $t\in[T_{1,3},s-1]$, which implies (155) holds at step s.

On the other hand, if $s=T_{1,4}$, then (662) suggests

$$\forall t\in[T_{1,3},T_{1,4}-1]:\quad -\delta e^{-a_{b,0}^{(t)}}\leqslant\frac{1+o(1)}{N}\frac{\ln^2 k}{k^3}a_{p,1}^{(t)}\delta a_{q,\times}^{(t)}, \tag{675}$$

and thus

$$\begin{aligned}&\frac{1+o(1)}{N}\frac{\ln^2 k}{k^3}a_{p,1}^{(T_{1,4})}\left(a_{q,\times}^{(T_{1,4})}-a_{q,\times}^{(T_{1,3})}\right)\geqslant\eta\sum_{t=T_{1,3}}^{T_{1,4}-1}\left(-\delta e^{-a_{b,0}^{(t)}}\right)=e^{-a_{b,0}^{(T_{1,3})}}\\&\quad-e^{-a_{b,0}^{(T_{1,4})}}=1+o(1),\end{aligned} \tag{676}$$

where the last relation follows from our choice of $T_{1,4}$ and (145). From the above inequality and (134), (140) we deduce

$$\frac{1}{N}\frac{\ln^2 k}{k^4}\left(a_{p,1}^{(T_{1,4})}\right)^2\gtrsim 1, \tag{677}$$

from which we can see that

$$a_{p,1}^{(T_{1,4})}\gtrsim\frac{k^2\sqrt{N}}{\ln k}. \tag{678}$$

By (678) and (151) we have

$$e^{a_{b,1}^{(T_{1,4})}}\gtrsim\frac{k^5}{\ln^2 k}. \tag{679}$$

To show (147), (148), (149), (150) in Lemma 3 hold at step s, below we first bound the gradients of $\mu_2^{(t)}$, $\mu_{2(k-1)}^{(t)}$ and $r_{\mathtt{down},\times}^{(t)}\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right)$ that appear in those relations.

By (131),

$$\begin{aligned}&\delta\mu_{2h}^{(t)}\\&:=\frac{\mu_{2h}^{(t+1)}-\mu_{2h}^{(t)}}{\eta}\\&=\frac{1}{\eta}\left[\left(-r_{\mathtt{up},h,0}^{(t+1)}\frac{a_{p,0}^{(t+1)}}{k}+r_{\mathtt{up},h,2}^{(t+1)}\left(ha_{p,2}^{(t+1)}+\frac{h-1}{k}a_{p,\mathtt{up}}^{(t+1)}\right)\right)-\left(-r_{\mathtt{up},h,0}^{(t)}\frac{a_{p,0}^{(t)}}{k}\right.\right.\\&\quad\left.\left.+r_{\mathtt{up},h,2}^{(t)}\left(ha_{p,2}^{(t)}+\frac{h-1}{k}a_{p,\mathtt{up}}^{(t)}\right)\right)\right]\\&=-r_{\mathtt{up},h,0}^{(t+1)}\frac{\delta a_{p,0}^{(t)}}{k}-\frac{r_{\mathtt{up},h,0}^{(t+1)}-r_{\mathtt{up},h,0}^{(t)}}{\eta}\frac{a_{p,0}^{(t)}}{k}+r_{\mathtt{up},h,2}^{(t+1)}\left(h\delta a_{p,2}^{(t)}+\frac{h-1}{k}\delta a_{p,\mathtt{up}}^{(t)}\right)\end{aligned}$$

$$+\frac{r_{\mathrm{up},h,2}^{(t+1)}-r_{\mathrm{up},h,2}^{(t)}}{\eta}\left(ha_{p,2}^{(t)}+\frac{h-1}{k}a_{p,\mathrm{up}}^{(t)}\right), \tag{680}$$

and from (562) we deduce

$$\begin{aligned}\frac{r_{\mathrm{up},h,0}^{(t+1)}-r_{\mathrm{up},h,0}^{(t)}}{\eta}&=\frac{1}{\eta}\frac{(2h-1)e^{a_{b,2}^{(t)}-a_{b,0}^{(t)}}\left(1-e^{\eta\left(\delta a_{b,2}^{(t)}-\delta a_{b,0}^{(t)}\right)}\right)}{\left(1+(2h-1)e^{a_{b,2}^{(t)}-a_{b,0}^{(t)}}\right)\left(1+(2h-1)e^{a_{b,2}^{(t+1)}-a_{b,0}^{(t+1)}}\right)}\\&=r_{\mathrm{up},h,0}^{(t+1)}\left(1-r_{\mathrm{up},h,0}^{(t)}\right)\frac{1-e^{\eta\left(\delta a_{b,2}^{(t)}-\delta a_{b,0}^{(t)}\right)}}{\eta}\\&=-\left(1+o(1)\right)r_{\mathrm{up},h,0}^{(t+1)}\left(1-r_{\mathrm{up},h,0}^{(t)}\right)\frac{1-e^{\eta\left(\delta a_{b,0}^{(t)}-\delta a_{b,2}^{(t)}\right)}}{\eta}\end{aligned} \tag{681}$$

and

$$\begin{aligned}\frac{r_{\mathrm{up},h,2}^{(t+1)}-r_{\mathrm{up},h,2}^{(t)}}{\eta}&=\frac{1}{\eta}\frac{e^{a_{b,0}^{(t)}-a_{b,2}^{(t)}}\left(1-e^{\eta\left(\delta a_{b,0}^{(t)}-\delta a_{b,2}^{(t)}\right)}\right)}{\left(2h-1+e^{a_{b,0}^{(t)}-a_{b,2}^{(t)}}\right)\left(2h-1+e^{a_{b,0}^{(t+1)}-a_{b,2}^{(t+1)}}\right)}\\&=\frac{1}{2h-1}\left(1-r_{\mathrm{up},h,0}^{(t+1)}\right)r_{\mathrm{up},h,0}^{(t)}\frac{1-e^{\eta\left(\delta a_{b,0}^{(t)}-\delta a_{b,2}^{(t)}\right)}}{\eta}.\end{aligned} \tag{682}$$

Plugging the above two equations into (680), we obtain

$$\begin{aligned}\delta\mu_{2h}^{(t)}=(1+o(1))\Bigg(&-r_{\mathrm{up},h,0}^{(t+1)}\frac{\delta a_{p,0}^{(t)}}{k}+r_{\mathrm{up},h,2}^{(t+1)}h\delta a_{p,2}^{(t)}+\frac{h}{2h-1}r_{\mathrm{up},h,0}^{(t+1)}\left(1\right.\\&\left.-r_{\mathrm{up},h,0}^{(t)}\right)a_{p,2}^{(t)}\frac{1-e^{\eta\left(\delta a_{b,0}^{(t)}-\delta a_{b,2}^{(t)}\right)}}{\eta}\Bigg).\end{aligned} \tag{683}$$

where

$$\begin{aligned}-\frac{1-e^{\eta\left(\delta a_{b,0}^{(t)}-\delta a_{b,2}^{(t)}\right)}}{\eta}&=(1+o(1))\left(\delta a_{b,0}^{(t)}-\delta a_{b,2}^{(t)}\right)\overset{(75)}{=}(1+o(1))\\&\left(\frac{N}{N-1}\delta a_{b,0}^{(t)}+\frac{1}{N-1}\delta a_{b,1}^{(t)}\right),\end{aligned} \tag{684}$$

by (658) and (640), there exists a constant $c>0$ such that

$$-\frac{1-e^{\eta\left(\delta a_{b,0}^{(t)}-\delta a_{b,2}^{(t)}\right)}}{\eta}\leqslant\frac{c}{a_{p,2}^{(t)}N\ln k}\delta a_{p,1}^{(t)}+\frac{1+o(1)}{4N}a_{p,2}^{(t)}\delta a_{q,\times}^{(t)}, \tag{685}$$

and by (640), (659) we have

$$\frac{1-e^{\eta\left(\delta a_{b,0}^{(t)}-\delta a_{b,2}^{(t)}\right)}}{\eta}\lesssim\frac{1}{a_{p,2}^{(t)}\ln k}\delta a_{p,1}^{(t)}+\frac{1}{k^2\ln k a_{p,2}^{(t)}\exp\left(0.1a_{p,2}^{(t)}\right)}\delta a_{q,\times}^{(t)}. \tag{686}$$

Combining (683) and (685), we have for all $t\in[T_{1,3},s-1]$:

$$\begin{aligned}-\delta\mu_{2h}^{(t)}&\leqslant(1+o(1))\left(r_{\mathrm{up},h,0}^{(t)}\frac{\delta a_{p,0}^{(t)}}{k}-r_{\mathrm{up},h,2}^{(t)}h\delta a_{p,2}^{(t)}+\frac{c}{N\ln k}\delta a_{p,1}^{(t)}+\frac{\left(a_{p,2}^{(t)}\right)^2}{16N}\delta a_{q,\times}^{(t)}\right)\\&\overset{(80)}{=}(1+o(1))\left(r_{\mathrm{up},h,0}^{(t+1)}\frac{\delta a_{p,0}^{(t)}}{k}+r_{\mathrm{up},h,2}^{(t)}h\frac{\delta a_{p,3}^{(t)}}{k}-r_{\mathrm{up},h,2}^{(t)}h\frac{\delta a_{p,1}^{(t)}}{k}+\frac{c}{N\ln k}\delta a_{p,1}^{(t)}\right.\end{aligned}$$

$$+\frac{\left(a_{p,2}^{(t)}\right)^2}{16N}\delta a_{q,\times}^{(t)}\Bigg)$$

$$\leqslant (1+o(1))\left(\frac{1}{k}+\frac{\left(a_{p,2}^{(t)}\right)^2}{16N}\right)\delta a_{q,\times}^{(t)}-(1+o(1))\frac{\ln k}{k^2}\delta a_{p,1}^{(t)}, \tag{687}$$

where in the last line we use the fact that

$$\delta a_{p,0}^{(t)} \leqslant (1+o(1))\delta a_{q,\times}^{(t)}, \tag{688}$$

$$\delta a_{p,3}^{(t)} \leqslant \frac{1+o(1)}{k}\delta a_{q,\times}^{(t)} \tag{689}$$

indicated by (541), (543) and (602), and the fact that

$$r_{\mathrm{up},h,2}h \geqslant (1+o(1))\frac{\ln k}{k} \tag{690}$$

guaranteed by (130) and (638). Combining (683) and (686), we have for all $t \in [T_{1,3}, s-1]$:

$$\begin{aligned}
\delta\mu_{2h}^{(t)} &\leqslant (1+o(1))\left(-r_{\mathrm{up},h,0}^{(t+1)}\frac{\delta a_{p,0}^{(t)}}{k}+r_{\mathrm{up},h,2}^{(t+1)}h\delta a_{p,2}^{(t)}\right)+\frac{c}{\ln k}\delta a_{p,1}^{(t)}\\
&\quad+\frac{c}{k^2\ln k\exp\left(0.1a_{p,2}^{(t)}\right)}\delta a_{q,\times}^{(t)}\\
&\overset{(80)}{=}(1+o(1))\left(-r_{\mathrm{up},h,0}^{(t+1)}\frac{\delta a_{p,0}^{(t)}}{k}-r_{\mathrm{up},h,2}^{(t)}h\frac{\delta a_{p,3}^{(t)}}{k}+r_{\mathrm{up},h,2}^{(t)}h\frac{\delta a_{p,1}^{(t)}}{k}\right)+\frac{c}{\ln k}\delta a_{p,1}^{(t)}\\
&\quad+\frac{c}{k^2\ln k\exp\left(0.1a_{p,2}^{(t)}\right)}\delta a_{q,\times}^{(t)}\\
&=(1+o(1))\left(\frac{c}{\ln k}\delta a_{p,1}^{(t)}+\frac{1}{k}\pi_{1,1}^{(t)}(\mathrm{up})\right)
\end{aligned} \tag{691}$$

for some constant $c>0$, where the last line uses (543), (541) and (569).

Now we bound the gradient of $r_{\mathrm{down}_\times}^{(t)}\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right)$:

$$\begin{aligned}
\delta\left(r_{\mathrm{down}_\times}^{(t)}\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right)\right) &:= \frac{1}{\eta}\left(r_{\mathrm{down}_\times}^{(t+1)}\left(a_{p,1}^{(t+1)}+a_{p,2}^{(t+1)}\right)-r_{\mathrm{down}_\times}^{(t)}\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right)\right)\\
&= r_{\mathrm{down}_\times}^{(t+1)}\left(\delta a_{p,1}^{(t)}+\delta a_{p,2}^{(t)}\right)+\frac{r_{\mathrm{down}_\times}^{(t+1)}-r_{\mathrm{down}_\times}^{(t)}}{\eta}\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right),
\end{aligned} \tag{692}$$

and by (132), we have

$$\begin{aligned}
\frac{r_{\mathrm{down}_\times}^{(t+1)}-r_{\mathrm{down}_\times}^{(t)}}{\eta} &= r_{\mathrm{down},k-1,0}^{(t)}r_{\mathrm{down}_\times}^{(t+1)}\frac{1-e^{\delta a_{b,0}^{(t)}-\delta a_{b,1}^{(t)}}}{\eta}+(k\\
&\quad-1)r_{\mathrm{down},k-1,2}^{(t)}r_{\mathrm{down}_\times}^{(t+1)}\frac{1-e^{\delta a_{b,2}^{(t)}-\delta a_{b,1}^{(t)}}}{\eta},
\end{aligned} \tag{693}$$

where

$$-\frac{1-e^{\delta a_{b,0}^{(t)}-\delta a_{b,1}^{(t)}}}{\eta} = (1+o(1))\left(\delta a_{b,0}^{(t)}-\delta a_{b,1}^{(t)}\right), \tag{694}$$

and

$$-\frac{1-e^{\delta a_{b,2}^{(t)}-\delta a_{b,1}^{(t)}}}{\eta} \overset{(75)}{=} -(1+o(1))\left(\frac{N}{N-1}\delta a_{b,1}^{(t)}+\frac{1}{N-1}\delta a_{b,0}^{(t)}\right). \tag{695}$$

By (658), (640) and the above two expressions, we have

$$-\frac{1-e^{\delta a_{b,0}^{(t)}-\delta a_{b,1}^{(t)}}}{\eta} \leqslant \frac{1+o(1)}{4N} a_{p,2}^{(t)} \delta a_{q,\times}^{(t)}, \tag{696}$$

$$\frac{1-e^{\delta a_{b,0}^{(t)}-\delta a_{b,1}^{(t)}}}{\eta} \lesssim \frac{1}{a_{p,2}^{(t)} \ln k} \delta a_{p,1}^{(t)} + \frac{1}{k^2 \ln k a_{p,2}^{(t)} \exp\left(0.1 a_{p,2}^{(t)}\right)} \delta a_{q,\times}^{(t)}, \tag{697}$$

$$-\frac{1-e^{\delta a_{b,2}^{(t)}-\delta a_{b,1}^{(t)}}}{\eta} \lesssim \frac{1}{a_{p,2}^{(t)} N \ln k} \delta a_{p,1}^{(t)} + \frac{1}{N k^2 \ln k a_{p,2}^{(t)} \exp\left(0.1 a_{p,2}^{(t)}\right)} \delta a_{q,\times}^{(t)}, \tag{698}$$

and there exists a constant $c' > 0$ such that

$$\begin{aligned}\frac{1-e^{\delta a_{b,2}^{(t)}-\delta a_{b,1}^{(t)}}}{\eta} \leqslant & \frac{c'}{a_{p,2}^{(t)} \ln k} \delta a_{p,1}^{(t)} + \left(\frac{c'}{k^2 \ln k a_{p,2}^{(t)} \exp\left(0.1 a_{p,2}^{(t)}\right)} \right. \\ & \left. + \frac{1+o(1)}{4N^2} a_{p,2}^{(t)} \right) \delta a_{q,\times}^{(t)}. \end{aligned} \tag{699}$$

By (696) and (698), (693) and (140) we have

$$\begin{aligned}\frac{r_{\mathrm{down}_\times}^{(t+1)} - r_{\mathrm{down}_\times}^{(t)}}{\eta} \geqslant & -\frac{1}{k-1}\left(1-(k-1) r_{\mathrm{down},k-1,1}^{(t)}\right)\left(\frac{1+o(1)}{4N} a_{p,2}^{(t)} \delta a_{q,\times}^{(t)} \right. \\ & \left. + \frac{c_1}{a_{p,2}^{(t)} N \ln k} \delta a_{p,1}^{(t)}\right) \\ \overset{(639)}{\gtrsim} & -\frac{1}{k-1} \frac{N}{\left(a_{p,1}^{(t)}\right)^2 \ln k} \left(\frac{1}{N} a_{p,2}^{(t)} \delta a_{q,\times}^{(t)} + \frac{1}{a_{p,2}^{(t)} N \ln k} \delta a_{p,1}^{(t)} \right) \\ \asymp & -\left(\frac{1}{k^3 \ln k a_{p,2}^{(t)}} \delta a_{q,\times}^{(t)} + \frac{1}{k^3 \ln^2 k \left(a_{p,2}^{(t)}\right)^3} \delta a_{p,1}^{(t)} \right) \end{aligned} \tag{700}$$

for some constant $c_1 > 0$, where in the first relation we also use the fact that

$$r_{\mathrm{down}_\times}^{(t+1)} \geqslant \frac{1}{k-1}. \tag{701}$$

Plugging (700) into (692), and use the fact that

$$r_{\mathrm{down}_\times}^{(t+1)} \geqslant \frac{1}{2k+1} \tag{702}$$

guaranteed by (151), (638) and (144), we have

$$\delta\left(r_{\mathrm{down}_\times}^{(t)}\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right)\right) \geqslant \frac{1}{2k+1}\left(\delta a_{p,1}^{(t)} + \delta a_{p,2}^{(t)}\right) - \frac{c_2}{k^2 \ln k} \delta a_{q,\times}^{(t)}, \tag{703}$$

for some constant $c_2 > 0$. By (80) and the above inequality we have

$$\begin{aligned}\delta\left(r_{\mathrm{down}_\times}^{(t)}\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right)\right) \geqslant & (1+o(1))\left(\frac{1}{2k}\delta a_{p,1}^{(t)} - \frac{1}{2k^2}\delta a_{p,3}^{(t)} - \frac{c_2}{k^2 \ln k}\delta a_{q,\times}^{(t)}\right) \\ \overset{(689)}{\geqslant} & (1+o(1))\left(\frac{1}{2k}\delta a_{p,1}^{(t)} - \frac{c_2}{k^2 \ln k}\delta a_{q,\times}^{(t)}\right). \end{aligned} \tag{704}$$

Similar as how we compute (700), plugging (697) and (699) into (693), and using (639), (701) we deduce

$$\frac{r_{\mathrm{down}_\times}^{(t+1)} - r_{\mathrm{down}_\times}^{(t)}}{\eta}$$

$$\lesssim \frac{1}{k-1}\min\left\{\frac{N}{\left(a_{p,1}^{(t)}\right)^2\ln k},1\right\}\left(\frac{1}{a_{p,2}^{(t)}\ln k}\delta a_{p,1}^{(t)}+\left(\frac{1}{k^2\ln k a_{p,2}^{(t)}\exp\left(0.1a_{p,2}^{(t)}\right)}\right.\right.$$
$$\left.\left.+\frac{1}{N^2}a_{p,2}^{(t)}\right)\delta a_{q,\times}^{(t)}\right)$$
$$\lesssim \frac{1}{a_{p,2}^{(t)}k\ln k}\delta a_{p,1}^{(t)}+\left(\frac{1}{k^3\ln k a_{p,2}^{(t)}\exp\left(0.1a_{p,2}^{(t)}\right)}+\frac{1}{Nk^3\ln k a_{p,2}^{(t)}}\right)\delta a_{q,\times}^{(t)}. \tag{705}$$

Plugging this back into (692) and using (701), (140), we have

$$\delta\left(r_{\mathtt{down}_\times}^{(t)}\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right)\right)$$
$$\leqslant \frac{\delta a_{p,1}^{(t)}+\delta a_{p,2}^{(t)}}{k-1}+\frac{1}{\ln k}\delta a_{p,1}^{(t)}+c\left(\frac{1}{k^2\ln k\exp\left(0.1a_{p,2}^{(t)}\right)}+\frac{1}{Nk^2\ln k}\right)\delta a_{q,\times}^{(t)}$$
$$\overset{(80)}{=}(1+o(1))\left(\frac{1}{\ln k}\delta a_{p,1}^{(t)}-\frac{a_{p,3}^{(t)}}{k^2}\right)+c\left(\frac{1}{k^2\ln k\exp\left(0.1a_{p,2}^{(t)}\right)}+\frac{1}{Nk^2\ln k}\right)\delta a_{q,\times}^{(t)}$$
$$\leqslant \frac{1+o(1)}{\ln k}\delta a_{p,1}^{(t)}+c\left(\frac{1}{k^2\ln k\exp\left(0.1a_{p,2}^{(t)}\right)}+\frac{1}{Nk^2\ln k}\right)\delta a_{q,\times}^{(t)}, \tag{706}$$

for some constant $c>0$, where the last relation follows from the fact that

$$\forall t\in[T_{1,3},s-1]:\quad a_{p,3}^{(t)}>0 \tag{707}$$

guaranteed by (543) and (601).

With the above gradient bounds, we next show (147), (148), (149), (150) hold at step s when $s-1\geqslant T_{1,3}$. We first show that (148) holds at step s. Suppose otherwise, i.e., we have

$$r_{\mathtt{down}_\times}^{(s)}\left(a_{p,1}^{(s)}+a_{p,2}^{(s)}\right)+0.1\sqrt{a_{p,1}^{(s)}/k}+5\ln k<\min\left\{a_{q,\times}^{(s)}-\mu_2^{(s)},a_{q,\times}^{(s)}-\mu_{2(k-1)}^{(s)}\right\}. \tag{708}$$

By Induction Hypothesis 2 we have

$$r_{\mathtt{down}_\times}^{(s-1)}\left(a_{p,1}^{(s-1)}+a_{p,2}^{(s-1)}\right)+0.1\sqrt{a_{p,1}^{(s-1)}/k}+5\ln k\geqslant\min\left\{a_{q,\times}^{(s-1)}\right.$$
$$\left.-\mu_2^{(s-1)},a_{q,\times}^{(s-1)}-\mu_{2(k-1)}^{(s-1)}\right\}. \tag{709}$$

Meanwhile, by our gradient update rule and (708), we have

$$r_{\mathtt{down}_\times}^{(s-1)}\left(a_{p,1}^{(s-1)}+a_{p,2}^{(s-1)}\right)+0.1\sqrt{a_{p,1}^{(s-1)}/k}+5\ln k\leqslant\min\left\{a_{q,\times}^{(s-1)}\right.$$
$$\left.-\mu_2^{(s-1)},a_{q,\times}^{(s-1)}-\mu_{2(k-1)}^{(s-1)}\right\}+1, \tag{710}$$

and thus similar as (607), we can compute that

$$\delta a_{p,1}^{(s-1)}\gtrsim k\exp\left(0.1\sqrt{a_{p,1}^{(s-1)}/k}\right)\delta a_{q,\times}^{(s-1)}. \tag{711}$$

Then we have

$$r_{\mathtt{down}_\times}^{(s)}\left(a_{p,1}^{(s)}+a_{p,2}^{(s)}\right)+0.1\sqrt{a_{p,1}^{(s)}/k}+5\ln k-\min\left\{a_{q,\times}^{(s)}-\mu_2^{(s)},a_{q,\times}^{(s)}-\mu_{2(k-1)}^{(s)}\right\}$$
$$\overset{(709)}{\geqslant}\delta\left(r_{\mathtt{down}_\times}^{(s-1)}\left(a_{p,1}^{(s-1)}+a_{p,2}^{(s-1)}\right)\right)-\delta a_{q,\times}^{(s-1)}+\min\{\delta\mu_2^{(s)},\delta\mu_{2(k-1)}^{(s)}\}$$
$$\geqslant(1+o(1))\left(\frac{1}{2k}\delta a_{p,1}^{(s-1)}-\frac{c_2}{k^2\ln k}\delta a_{q,\times}^{(s-1)}\right)$$

$$
\begin{aligned}
&- \delta a_{q,\times}^{(s-1)} - (1+o(1))\left(\frac{1}{k} + \frac{\left(a_{p,2}^{(s-1)}\right)^2}{16N}\right)\delta a_{q,\times}^{(s-1)} + (1+o(1))\frac{\ln k}{k^2}\delta a_{p,1}^{(s-1)} \\
&\geqslant (1+o(1))\frac{1}{2k}\delta a_{p,1}^{(s-1)} - (1+o(1))\left(1 + \frac{\left(a_{p,2}^{(s-1)}\right)^2}{16N}\right)\delta a_{q,\times}^{(s-1)} \\
&\overset{(711)}{\gtrsim} \exp\left(0.1\sqrt{a_{p,1}^{(s-1)}/k}\right)\delta a_{q,\times}^{(s-1)} > 0,
\end{aligned} \tag{712}
$$

where the third line follows from (687) and (704). This contradicts (708). Therefore, we have shown that (148) holds at step s.

We now show (150) holds at step s. Suppose otherwise, i.e.,

$$
\left(1+\frac{1}{k}\right)a_{q,0}^{(s)} - \frac{a_{p,0}^{(s)}}{k} + 0.1\sqrt{a_{p,1}^{(s)}/k} + 3\ln k < \min\left\{a_{q,\times}^{(s)} - \mu_2^{(s)}, a_{q,\times}^{(s)} - \mu_{2(k-1)}^{(s)}\right\}. \tag{713}
$$

By Induction Hypothesis 2 we have

$$
\begin{aligned}
&\left(1+\frac{1}{k}\right)a_{q,0}^{(s-1)} - \frac{a_{p,0}^{(s-1)}}{k} + 0.1\sqrt{a_{p,1}^{(s-1)}/k} + 3\ln k \geqslant \min\left\{a_{q,\times}^{(s-1)}\right. \\
&\quad \left. -\mu_2^{(s-1)}, a_{q,\times}^{(s-1)} - \mu_{2(k-1)}^{(s-1)}\right\}.
\end{aligned} \tag{714}
$$

Meanwhile, by our gradient update rule and (713), we have

$$
\begin{aligned}
&\left(1+\frac{1}{k}\right)a_{q,0}^{(s)} - \frac{a_{p,0}^{(s)}}{k} + 0.1\sqrt{a_{p,1}^{(s)}/k} + 3\ln k \leqslant \min\left\{a_{q,\times}^{(s)} - \mu_2^{(s)}, a_{q,\times}^{(s)}\right. \\
&\quad \left. -\mu_{2(k-1)}^{(s)}\right\} + 1,
\end{aligned} \tag{715}
$$

Then similar as how we obtain (591), from the above relation and (537), we can compute that

$$
\delta a_{q,0}^{(s-1)} \gtrsim \exp\left(0.1\sqrt{a_{p,1}^{(s-1)}/k}\right)\delta a_{q,\times}^{(s-1)}. \tag{716}
$$

$$
-\delta a_{p,0}^{(s-1)} = (1+o(1))\delta a_{q,0}^{(s-1)}. \tag{717}
$$

On the other hand, by (687) we know that

$$
-\delta\mu_{2h}^{(s-1)} \leqslant (1+o(1))\left(\frac{1}{k} + \frac{\left(a_{p,2}^{(s)}\right)^2}{16N}\right)\delta a_{q,\times}^{(s)}, \tag{718}
$$

and thus we have

$$
\begin{aligned}
&\left(1+\frac{1}{k}\right)a_{q,0}^{(s)} - \frac{a_{p,0}^{(s)}}{k} + 0.1\sqrt{a_{p,1}^{(s)}/k} + 3\ln k - \min\left\{a_{q,\times}^{(s)} - \mu_2^{(s)}, a_{q,\times}^{(s)} - \mu_{2(k-1)}^{(s)}\right\} \\
&\overset{(714)}{\geqslant} \left(1+\frac{1}{k}\right)\delta a_{q,0}^{(s-1)} - \frac{\delta a_{p,0}^{(s-1)}}{k} - \delta a_{q,\times}^{(s-1)} + \min\{\delta\mu_2^{(s-1)}, \delta\mu_{2(k-1)}^{(s-1)}\} \\
&\gtrsim \exp\left(0.1\sqrt{a_{p,1}^{(s-1)}/k}\right)\delta a_{q,\times}^{(s-1)} > 0,
\end{aligned} \tag{719}
$$

where the last relation follows from (716), (717) and (718). This contradicts (713). Therefore, (150) holds at step s.

We next show (149) holds at step s. We first give a new lower bound of $\delta\left(r^{(t)}_{\mathtt{down}_\times}\left(a^{(t)}_{p,1}+a^{(t)}_{p,2}\right)\right)$ for all $t\in[T_{1,3},s-1]$. Similar as how we obtain (658), by (657) we can get a different upper bound for $\delta a^{(t)}_{b,0}$ for all $t\in[T_{1/2},s-1]$:

$$\delta a^{(t)}_{b,0}\leqslant\frac{1+o(1)}{N}a^{(t)}_{p,2}\sum_{h=1}^{k-1}r^{(t)}_{\mathtt{up},h,0}\frac{k-h}{k}\left(1-\pi^{(t)}_{1,2h}(\mathtt{up})\right),\tag{720}$$

and by (543) we have

$$\delta a^{(t)}_{p,3}\leqslant\frac{1+o(1)}{k}\sum_{h=1}^{k-1}\frac{k-h}{k}\left(1-\pi^{(t)}_{1,2h}(\mathtt{up})\right)\lesssim k^{1/2}\sum_{h=1}^{k-1}\frac{k-h}{k}r^{(t)}_{\mathtt{up},h,0}\left(1-\pi^{(t)}_{1,2h}(\mathtt{up})\right),\tag{721}$$

where we use the fact that

$$r^{(t)}_{\mathtt{up},h,0}\lesssim k^{3/2}\tag{722}$$

guaranteed by (152) and (638). Then similar as how we obtain (704), by the new bounds (720) and (721), we can compute that for all $t\in[T_{1,3},s-1]$:

$$\delta\left(r^{(t)}_{\mathtt{down}_\times}\left(a^{(t)}_{p,1}+a^{(t)}_{p,2}\right)\right)\geqslant(1+o(1))\left(\frac{1}{2k}\delta a^{(t)}_{p,1}-\frac{c_3}{k^{3/2}}\sum_{h=1}^{k-1}\frac{k-h}{k}r^{(t)}_{\mathtt{up},h,0}\left(1-\pi^{(t)}_{1,2h}(\mathtt{up})\right)\right)\tag{723}$$

for some constant $c_3>0$. Now we show that (149) holds at step s by contradiction. Suppose otherwise, i.e.,

$$r^{(s)}_{\mathtt{down}_\times}\left(a^{(s)}_{p,1}+a^{(s)}_{p,2}\right)+2.1\ln k<\left(1+\frac{1}{k}\right)a^{(s)}_{q,0}-\frac{a^{(s)}_{p,0}}{k}.\tag{724}$$

By Induction Hypothesis 2 we have

$$r^{(s-1)}_{\mathtt{down}_\times}\left(a^{(s-1)}_{p,1}+a^{(s-1)}_{p,2}\right)+2.1\ln k\geqslant\left(1+\frac{1}{k}\right)a^{(s-1)}_{q,0}-\frac{a^{(s-1)}_{p,0}}{k}.\tag{725}$$

Meanwhile, by our update rule and (724), we have

$$r^{(s-1)}_{\mathtt{down}_\times}\left(a^{(s-1)}_{p,1}+a^{(s-1)}_{p,2}\right)+2.1\ln k\leqslant\left(1+\frac{1}{k}\right)a^{(s-1)}_{q,0}-\frac{a^{(s-1)}_{p,0}}{k}+1,\tag{726}$$

and similar as how we obtain (619), we have

$$\delta a^{(s-1)}_{p,1}\gtrsim k^{1.1}\pi^{(s-1)}_{1,1}(\mathtt{up})\gtrsim k^{1.1}\delta a^{(s-1)}_{q,0}.\tag{727}$$

Therefore, we have

$$\begin{aligned}
&r^{(s)}_{\mathtt{down}_\times}\left(a^{(s)}_{p,1}+a^{(s)}_{p,2}\right)+2.1\ln k-\left(\left(1+\frac{1}{k}\right)a^{(s)}_{q,0}-\frac{a^{(s)}_{p,0}}{k}\right)\\
&\overset{(725)}{\geqslant}\delta\left(r^{(s-1)}_{\mathtt{down}_\times}\left(a^{(s-1)}_{p,1}+a^{(s-1)}_{p,2}\right)\right)-\left(1+\frac{1}{k}\right)\delta a^{(s-1)}_{q,0}+\frac{\delta a^{(s-1)}_{p,0}}{k}\\
&\geqslant(1+o(1))\left(\frac{1}{2k}\delta a^{(s-1)}_{p,1}-\frac{c_3}{k^{3/2}}\sum_{h=1}^{k-1}\frac{k-h}{k}r^{(s-1)}_{\mathtt{up},h,0}\left(1-\pi^{(s-1)}_{1,2h}(\mathtt{up})\right)\right)-\left(1+\frac{1}{k}\right)\delta a^{(s-1)}_{q,0}\\
&\quad+\frac{1+o(1)}{k}\left(-\pi^{(s-1)}_{1,1}(\mathtt{up})+\sum_{h=1}^{k-1}\frac{k-h}{k}r^{(s-1)}_{\mathtt{up},h,0}\left(1-\pi^{(s-1)}_{1,2h}(\mathtt{up})\right)\right)
\end{aligned}$$

$$\overset{(727)}{\gtrsim} k^{0.1}\delta a_{q,0}^{(s-1)} > 0, \tag{728}$$

where the third line follows from (723), (541) and (569). This contradicts (724), and thus (149) holds at step s.

We now show (147) holds at step s by contradiction. Suppose otherwise, i.e.,

$$\min\left\{a_{q,\times}^{(s)} - \mu_2^{(s)} - \ln k, a_{q,\times}^{(s)} - \mu_{2(k-1)}^{(s)} + \ln k\right\} + \frac{2}{k}a_{q,0}^{(s)} < r_{\texttt{down}_\times}^{(s)}\left(a_{p,1}^{(s)} + a_{p,2}^{(s)}\right) + \ln k, \tag{729}$$

By Induction Hypothesis 2 we have

$$\min\left\{a_{q,\times}^{(s-1)} - \mu_2^{(s-1)} - \ln k, a_{q,\times}^{(s-1)} - \mu_{2(k-1)}^{(s-1)} + \ln k\right\} + \frac{2}{k}a_{q,0}^{(s-1)} \geqslant r_{\texttt{down}_\times}^{(s-1)}\left(a_{p,1}^{(s-1)} + a_{p,2}^{(s-1)}\right) + \ln k, \tag{730}$$

Meanwhile, by our gradient update rule and (729), we have

$$\min\left\{a_{q,\times}^{(s)} - \mu_2^{(s)} - \ln k, a_{q,\times}^{(s)} - \mu_{2(k-1)}^{(s)} + \ln k\right\} + \frac{2}{k}a_{q,0}^{(s)} \leqslant r_{\texttt{down}_\times}^{(s)}\left(a_{p,1}^{(s)} + a_{p,2}^{(s)}\right) + \ln k + 1. \tag{731}$$

Then similar as how we prove (626), we can compute that

$$\delta a_{p,1}^{(s-1)} \lesssim \delta a_{q,\times}^{(s-1)}. \tag{732}$$

Then we have

$$\begin{aligned}
&\min\left\{a_{q,\times}^{(s)} - \mu_2^{(s)} - \ln k, a_{q,\times}^{(s)} - \mu_{2(k-1)}^{(s)} + \ln k\right\} + \frac{2}{k}a_{q,0}^{(s)} - \left(r_{\texttt{down}_\times}^{(s)}\left(a_{p,1}^{(s)} + a_{p,2}^{(s)}\right) + \ln k\right) \\
&\overset{(730)}{\geqslant} \delta a_{q,\times}^{(s-1)} - \max\{\mu_2^{(s-1)}, \mu_{2(k-1)}^{(s-1)}\} + \frac{2}{k}\delta a_{q,0}^{(s-1)} - \delta\left(r_{\texttt{down}_\times}^{(s-1)}\left(a_{p,1}^{(s-1)} + a_{p,2}^{(s-1)}\right)\right) \\
&\geqslant (1+o(1))\,\delta a_{q,\times}^{(s-1)} - (1+o(1))\left(\frac{c}{\ln k}\delta a_{p,1}^{(s-1)} + \frac{1}{k}\pi_{1,1}^{(s-1)}(\texttt{up})\right) + \frac{2+o(1)}{k} \\
&\quad \pi_{1,1}^{(s-1)}(\texttt{up}) - \frac{1+o(1)}{\ln k}\delta a_{p,1}^{(s-1)} \\
&\overset{(732)}{\gtrsim} \delta a_{q,\times}^{(s-1)} > 0,
\end{aligned} \tag{733}$$

where the third line follows from (537), (706) and (691). This contradicts (729). Therefore, (147) must hold at step s.

Now we have shown that (147), (148), (149), (150) and (138) hold at step s. They together guarantee that (134) holds at step s.

By (155), (146), (143) and (128) we have

$$T_{1,4} \lesssim \frac{k^{9/4}N\exp\left(k^{1.3}\sqrt{N}\right)}{\eta}. \tag{734}$$

When $T_{1,4}+1 \leqslant s \leqslant T_1$. By Induction Hypothesis 2 we have (162), (164), (158) hold for all $t \in [T_{1,4}, s-1]$. Combining them with (132), we have

$$\forall t \in [T_{1,4}, s-1]: \quad \frac{1}{k-1} - \frac{c_1}{k^4} \leqslant r_{\texttt{down}_\times}^{(t)} \leqslant \frac{1}{k-1} \tag{735}$$

for some constant $c_1 > 0$, and

$$\max_{h\in[k-1]}\{\mu_{2h}^{(t)}\} \leqslant a_{p,2}^{(t)}. \tag{736}$$

Then use a similar argument as how we show (147), (148) and (150), we can show that (159), (160) and (161) hold at step s. And from these three expressions we know that (134) holds at step s. From (678), (679) and the fact that $a_{p,1}^{(t)}, a_{b,1}^{(t)}$ monotonically increase we immediately know that (157) and (158) hold at step s. Then by the same argument as in (601), from (161) we deduce (602) still holds here, i.e.,

$$\delta a_{q,\times}^{(s-1)} = (1+o(1)) \sum_{h=1}^{k-1} \frac{k-h}{k} \left(1 - \pi_{1,2h}^{(s-1)}(\mathtt{up})\right), \tag{737}$$

And (603), (628) also still holds here, i.e., for all $t \in [T_{1,4}, s-1]$:

$$\delta e^{2a_{c,0}^{(t)}} \lesssim \frac{a_{p,2}^{(t)}}{\exp\left(0.1 a_{p,2}^{(t)}\right)} \delta a_{q,\times}^{(t)}. \tag{738}$$

From (738), (157), (140) and (134) we know that

$$e^{2a_{c,0}^{(s)}} - e^{2a_{c,0}^{(T_{1,4})}} \lesssim \sum_{i=0}^{\infty} \frac{a_{p,1}^{(T_{1,4})} k^{i-1}}{\exp\left(0.1 a_{p,1}^{(T_{1,4})} k^{i-1}\right)} a_{p,1}^{(T_{1,4})} k^i \lesssim \frac{\left(a_{p,1}^{(T_{1,4})}\right)^2 / k}{\exp\left(0.1 a_{p,1}^{(T_{1,4})} / k\right)} = o(1). \tag{739}$$

Therefore, from (739), (553) (634) we know that (135) holds at step s.

And similar as how we obtain (657), from (546) and Induction Hypothesis 2 we deduce for all $t \in [T_{1,4}, s-1]$:

$$\begin{aligned} I_{h,0}^{(t)} :=& \left(1 + O\left(\frac{1}{k}\right)\right) r_{\mathtt{up},h,0}^{(t)} r_{\mathtt{up},h,2}^{(t)} h a_{p,2}^{(t)} \left(1 - \pi_{1,2h}^{(t)}(\mathtt{up})\right) \\ &- \left(1 + O\left(\frac{1}{k}\right)\right) r_{\mathtt{up},h,0}^{(t)} r_{\mathtt{up},h,2}^{(t)} h a_{p,1}^{(t)} \pi_{1,2h}^{(t)}(\mathtt{down}_\times^1) \\ &- \left(1 + O\left(\frac{1}{k}\right)\right) h r_{\mathtt{down},h,0}^{(t)} r_{\mathtt{down},h,1}^{(t)} \left(a_{p,1}^{(t)} \pi_{1,2h+1}^{(t)}(\mathtt{down}_\times) + a_{p,2}^{(t)} \pi_{1,2h+1}^{(t)}(\mathtt{up})\right). \end{aligned} \tag{740}$$

Also note that (635) still holds here, i.e., we have for all $t \in [T_{1,4}, s-1]$:

$$\begin{aligned} I_{h,1}^{(t)} =& \left(1 + O\left(\frac{1}{k}\right)\right) \left(1 - h r_{\mathtt{down},h,1}^{(t)}\right) \\ &\cdot \left(k a_{p,2}^{(t)} \frac{h e^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + h e^{a_{b,1}^{(t)}} + h e^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(\mathtt{down}_\times) + h r_{\mathtt{down},h,1}^{(t)} a_{p,2}^{(t)} \pi_{1,2h+1}^{(t)}(\mathtt{up})\right). \end{aligned} \tag{741}$$

Consider

$$\delta e^{\frac{N}{N-1} a_{b,i}^{(t)}} := \frac{e^{\frac{N}{N-1} a_{b,i}^{(t+1)}} - e^{\frac{N}{N-1} a_{b,i}^{(t)}}}{\eta} = (1+o(1)) e^{\frac{N}{N-1} a_{b,i}^{(t)}} \delta a_{b,i}^{(t)}. \tag{742}$$

Then by (164), (740), (737) and (216) we have for all $t \in [T_{1,4}, s-1]$:

$$\delta e^{\frac{N}{N-1} a_{b,0}^{(t)}} \leqslant (1+o(1)) \frac{a_{p,1}^{(t)}}{N} \delta a_{q,\times}^{(t)}. \tag{743}$$

And same as how we obtain (644), the above expression suggests

$$e^{\frac{N}{N-1} a_{b,0}^{(t)}} \lesssim \frac{1}{Nk} \left(a_{p,1}^{(t)}\right)^2. \tag{744}$$

Moreover, similar as how we obtain (661), by (740) we have for all $t \in [T_{1,4}, s-1]$:

$$\delta e^{-a_{b,0}^{(t)}} = \frac{1+o(1)}{N} \sum_{h=1}^{k-1} \frac{k-h}{k} \left[\frac{h e^{a_{b,1}^{(t)}}}{\left(e^{a_{b,0}^{(t)}} + h e^{a_{b,1}^{(t)}} + h e^{a_{b,2}^{(t)}}\right)^2} \left(a_{p,1}^{(t)}\right.\right.$$

$$\pi^{(t)}_{1,2h+1}(\texttt{down}_\times) + a^{(t)}_{p,2}\pi^{(t)}_{1,2h+1}(\texttt{up})\Big) + \frac{he^{a^{(t)}_{b,2}}}{\left(e^{a^{(t)}_{b,0}} + (2h-1)e^{a^{(t)}_{b,2}}\right)^2} a^{(t)}_{p,1}\pi^{(t)}_{1,2h}(\texttt{down}^1_\times)\Bigg] - \frac{1+o(1)}{N}\sum_{h=1}^{k-1}\frac{k-h}{k}\frac{he^{a^{(t)}_{b,2}}}{\left(e^{a^{(t)}_{b,0}} + (2h-1)e^{a^{(t)}_{b,2}}\right)^2} a^{(t)}_{p,2}\left(1-\pi^{(t)}_{1,2h}(\texttt{up})\right). \tag{745}$$

Similar as how we show (645), from (741) we can compute that for all $t \in [T_{1,4}, s-1]$ and $h \in [k-1]$:

$$\begin{aligned}
&e^{\frac{N}{N-1}a^{(t)}_{b,1}} I^{(t)}_{h,1} \\
&\overset{(216)}{\geqslant} \left(1+O\left(\frac{1}{k}\right)\right)\frac{e^{a^{(t)}_{b,1}}\left(e^{a^{(t)}_{b,0}+\frac{1}{N-1}a^{(t)}_{b,1}} + he^{\frac{-a^{(t)}_{b,0}}{N-1}}\right)}{e^{a^{(t)}_{b,0}} + he^{a^{(t)}_{b,1}} + he^{a^{(t)}_{b,2}}} a^{(t)}_{p,1} \\
&\quad \cdot \frac{he^{a^{(t)}_{b,1}}}{e^{a^{(t)}_{b,0}} + he^{a^{(t)}_{b,1}} + he^{a^{(t)}_{b,2}}}\pi^{(t)}_{1,2h+1}(\texttt{down}_\times) \\
&\gtrsim \frac{he^{a^{(t)}_{b,1}}}{e^{a^{(t)}_{b,0}} + he^{a^{(t)}_{b,1}} + he^{a^{(t)}_{b,2}}} a^{(t)}_{p,1} \cdot \frac{he^{a^{(t)}_{b,1}}}{e^{a^{(t)}_{b,0}} + he^{a^{(t)}_{b,1}} + he^{a^{(t)}_{b,2}}}\pi^{(t)}_{1,2h+1}(\texttt{down}_\times) \\
&\gtrsim a^{(t)}_{p,1}\Bigg(\frac{he^{a^{(t)}_{b,2}}}{e^{a^{(t)}_{b,0}} + (2h-1)e^{a^{(t)}_{b,2}}}\pi^{(t)}_{1,2h}(\texttt{down}^1_\times) + \frac{he^{a^{(t)}_{b,1}}}{e^{a^{(t)}_{b,0}} + he^{a^{(t)}_{b,1}} + he^{a^{(t)}_{b,2}}} \\
&\quad \pi^{(t)}_{1,2h+1}(\texttt{down}_\times) \\
&\quad - \frac{he^{a^{(t)}_{b,2}}}{e^{a^{(t)}_{b,0}} + (2h-1)e^{a^{(t)}_{b,2}}}\pi^{(t)}_{1,2h}(\texttt{down}^1_\times)\Bigg),
\end{aligned} \tag{746}$$

where the second relation uses the following fact

$$e^{a^{(t)}_{b,0}+\frac{1}{N-1}a^{(t)}_{b,1}} + he^{\frac{-a^{(t)}_{b,0}}{N-1}} \gtrsim h. \tag{747}$$

Combining (742), (746) and (577), we obtain for all $t \in [T_{1,4}, s-1]$

$$\delta e^{\frac{N}{N-1}a^{(t)}_{b,1}} \overset{(737)}{\gtrsim} \frac{a^{(t)}_{p,1}}{N}\left(k\delta a^{(t)}_{p,1} - \frac{1}{2k}\delta a^{(t)}_{q,\times}\right). \tag{748}$$

And similar as how we obtain (656), from the above expression we deduce

$$e^{\frac{N}{N-1}a^{(s)}_{b,1}} \gtrsim \frac{k}{N}\left(a^{(s)}_{p,1}\right)^2 \overset{(744)}{\gtrsim} k^2 e^{\frac{N}{N-1}a^{(s)}_{b,0}}, \tag{749}$$

from which we know that (162) holds at step s.

Meanwhile, from (741), (162) and (164) we deduce for all $t \in [T_{1,4}, s-1]$:

$$\begin{aligned}
\delta e^{a^{(t)}_{b,1}} &\leqslant \frac{1+o(1)}{N}\sum_{h=1}^{k-1}\frac{k-h}{k}\left(e^{a^{(t)}_{b,0}} + he^{a^{(t)}_{b,2}}\right)\Bigg(ka^{(t)}_{p,2}\frac{he^{a^{(t)}_{b,1}}}{e^{a^{(t)}_{b,0}} + he^{a^{(t)}_{b,1}} + he^{a^{(t)}_{b,2}}} \\
&\quad \pi^{(t)}_{1,2h+1}(\texttt{down}_\times) \\
&\quad + hr^{(t)}_{\texttt{down},h,1}a^{(t)}_{p,2}\pi^{(t)}_{1,2h+1}(\texttt{up})\Bigg) \\
&\lesssim \frac{1}{N}\frac{\left(a^{(t)}_{p,1}\right)^{\frac{2(N-1)}{N}}}{Nk}\left(ka^{(t)}_{p,1}\delta a^{(t)}_{p,1} + \frac{ka^{(t)}_{p,2}}{\exp\left(0.1a^{(t)}_{p,2}\right)}\delta a^{(t)}_{q,\times}\right)
\end{aligned} \tag{750}$$

where the second line also use (542) and (603) (which also holds here). This together with (154), (157) suggests for all $t \in [T_{1,4}, s]$:

$$e^{a_{b,1}^{(t)}} \lesssim \frac{\left(a_{p,1}^{(t)}\right)^{4-\frac{2}{N}}}{N^2}. \tag{751}$$

Then by (751), (162) and (216) we have

$$\forall t \in [T_{1,4}, s]: \quad e^{a_{b,2}^{(t)}} \geqslant (1+o(1)) \left(a_{p,1}^{(t)}\right)^{-\frac{6}{N-1}}. \tag{752}$$

This indicates the first relation of (164) holds at step s.

Define

$$s_0 \coloneqq \sup \left\{ t \in [T_{1,4}, T_1] : a_{p,1}^{(t)} \leqslant Nk^3 \right\}. \tag{753}$$

Below we prove (156) holds for $s \in [T_{1,4}+1, s_0]$, and (163) holds for $s \in [s_0, T_1]$. We first assume $s \in [T_{1,4}+1, s_0]$, and prove (156) holds by showing

$$\forall t \in [T_{1,4}, s]: \quad e^{a_{b,0}^{(t)}} \geqslant \frac{Ck}{\ln k} \tag{754}$$

for some small enough constant $C \in (0, 1/2)$. Suppose otherwise, then from our choice of $T_{1,4}$ we know that there exists the largest $t_0', t_0 \in [T_{1,4}+1, s]$, $t_0' < t_0$ such that

$$e^{a_{b,0}^{(t_0')}} \geqslant 2C\frac{k}{\ln k}, \quad e^{a_{b,0}^{(t_0)}} \leqslant C\frac{k}{\ln k}, \tag{755}$$

and

$$\forall t \in [t_0'+1, t_0-1]: \quad C\frac{k}{\ln k} < e^{a_{b,0}^{(t)}} < 2C\frac{k}{\ln k}. \tag{756}$$

From the first line of (577), (752) (755) and (140) and (157) we know that for all $t \in [t_0', t_0]$ and $h \in [k-1]$:

$$\begin{aligned}\pi_{1,2h}^{(t)}(\mathtt{down}_\times^1) \leqslant{}& \left(1-\pi_{1,2h}^{(t)}(\mathtt{up})\right) \exp\left(-r_{\mathtt{up},h,2}^{(t)}\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right)\right) = o\left(\frac{1}{k}\right)(1 \\ &-\pi_{1,2h}^{(t)}(\mathtt{up})\Big).\end{aligned} \tag{757}$$

Note that from (752) and the fact that $s \leqslant s_0$ we know that

$$e^{a_{b,2}^{(t_0)}} \geqslant 1+o(1). \tag{758}$$

Then from (740), (758) and our choice of t_0 we deduce for all $t \in [t_0', t_0]$:

$$\begin{aligned}&\delta e^{-a_{b,0}^{(t)}} \\ &= -\frac{1+o(1)}{N}\sum_{h=1}^{k-1}\frac{k-h}{k}\frac{he^{a_{b,2}^{(t)}}}{\left(e^{a_{b,0}^{(t)}}+(2h-1)e^{a_{b,2}^{(t)}}\right)^2}a_{p,2}^{(t)}\left(1-\pi_{1,2h}^{(t)}(\mathtt{up})\right) \\ &\quad+\frac{1+o(1)}{N}\sum_{h=1}^{k-1}\frac{k-h}{k}\frac{he^{a_{b,1}^{(t)}}}{\left(e^{a_{b,0}^{(t)}}+he^{a_{b,1}^{(t)}}+he^{a_{b,2}^{(t)}}\right)^2}\Big(a_{p,1}^{(t)}\pi_{1,2h+1}^{(t)}(\mathtt{down}_\times) \\ &\quad+a_{p,2}^{(t)}\pi_{1,2h+1}^{(t)}(\mathtt{up})\Big) \\ &\leqslant -\frac{1+o(1)}{(2C)^2}\frac{\ln^2 k}{Nk^3}a_{p,1}^{(t)}\delta a_{q,\times}^{(t)}+\frac{1+o(1)}{N}\frac{1}{e^{a_{b,0}^{(t)}}+e^{a_{b,1}^{(t)}}+e^{a_{b,2}^{(t)}}}\Big(ka_{p,1}^{(t)}\delta a_{p,1}^{(t)} \\ &\quad+\frac{a_{p,2}^{(t)}}{\exp\left(0.1a_{p,2}^{(t)}\right)}\delta a_{q,\times}^{(t)}\Big)\end{aligned}$$

$$\leqslant -\frac{1+o(1)}{(2C)^2}\frac{\ln^2 k}{Nk^3}a_{p,1}^{(t)}\delta a_{q,\times}^{(t)} + \frac{c_1+o(1)}{a_{p,1}^{(t)}}\delta a_{p,1}^{(t)} \tag{759}$$

for some constant $c_1 > 0$, where we use (756), (737), (542) and (603) (which also holds here) to obtain the second relation, and use (162) to obtain the third relation. The above expression implies

$$e^{-a_{b,0}^{(t_0)}} - e^{-a_{b,0}^{(t_0')}} \leqslant -\frac{1+o(1)}{(2C)^2}\frac{\ln^2 k}{Nk^3}a_{p,1}^{(t_0')}\left(a_{q,\times}^{(t_0)} - a_{q,\times}^{(t_0')}\right) + \frac{c_1+o(1)}{a_{p,1}^{(t_0')}}\left(a_{p,1}^{(t_0)} - a_{p,1}^{(t_0')}\right). \tag{760}$$

From (756), (131), (134), (138), (139) and (140) we know that for all $t \in [t_0', t_0]$:

$$\max_{h\in[k-1]}\left\{\mu_{2h}^{(t)}\right\} = \mu_{2(k-1)}^{(t_0)}, \tag{761}$$

$$\mu_{2(k-1)}^{(t_0')} = \left(1 - \frac{C+o(1)}{\ln k}\right)\frac{a_{p,1}^{(t_0')}}{2k}, \tag{762}$$

and

$$\mu_{2(k-1)}^{(t_0)} = \left(1 - \frac{C+o(1)}{2\ln k}\right)\frac{a_{p,1}^{(t_0)}}{2k}. \tag{763}$$

Also note that by (761), (159), (160), (161), (157), (735), (134), (138), and (140) we have for all $t \in [t_0', t_0]$:

$$a_{q,\times}^{(t)} = \mu_{2(k-1)}^{(t)} + \frac{1}{k}a_{p,1}^{(t)} + O\left(\frac{1}{k^2}\right)a_{p,1}^{(t)}. \tag{764}$$

Combining (762), (763), (764) we have

$$\begin{aligned} a_{q,\times}^{(t_0)} - a_{q,\times}^{(t_0')} &= \frac{3}{2k}\left(a_{p,1}^{(t_0)} - a_{p,1}^{(t_0')}\right) - \frac{C+o(1)}{4k\ln k}a_{p,1}^{(t_0)} + \frac{C+o(1)}{2k\ln k}a_{p,1}^{(t_0')} \\ &\geqslant \frac{3}{2k}\left(a_{p,1}^{(t_0)} - a_{p,1}^{(t_0')}\right) - \frac{C+o(1)}{4k\ln k}\left(a_{p,1}^{(t_0)} - a_{p,1}^{(t_0')}\right) = \frac{3+o(1)}{2k}\left(a_{p,1}^{(t_0)} - a_{p,1}^{(t_0')}\right). \end{aligned} \tag{765}$$

Plugging this back into (760), we have

$$\begin{aligned} e^{-a_{b,0}^{(t_0)}} - e^{-a_{b,0}^{(t_0')}} &\leqslant -\frac{3+o(1)}{8C^2}\frac{\ln^2 k}{Nk^4}a_{p,1}^{(t_0')}\left(a_{p,1}^{(t_0)} - a_{p,1}^{(t_0')}\right) + \frac{c_1+o(1)}{a_{p,1}^{(t_0')}}\left(a_{p,1}^{(t_0)} - a_{p,1}^{(t_0')}\right) \\ &\overset{(157)}{\leqslant} \left(-\frac{3c_2+o(1)}{8C^2} + \frac{c_1+o(1)}{c_2}\right)\frac{\ln k}{\sqrt{N}k^2}\left(a_{p,1}^{(t_0)} - a_{p,1}^{(t_0')}\right) \end{aligned} \tag{766}$$

for some constant $c_2 > 0$. From the above inequality we can see that we can set C small enough such that

$$e^{-a_{b,0}^{(t_0)}} - e^{-a_{b,0}^{(t_0')}} < 0. \tag{767}$$

However, from (755) we have

$$e^{-a_{b,0}^{(t_0)}} - e^{-a_{b,0}^{(t_0')}} \geqslant \frac{\ln k}{2Ck}. \tag{768}$$

The above two inequalities are contradictory. Therefore, (754) must hold. This way we have shown (156).

Below we prove (163) assuming $s \geqslant s_0$. Define

$$s' = \sup\left\{t \in [T_{1,4}, s-1] : a_{p,1}^{(t)} \leqslant C_1 a_{p,1}^{(s)}\right\} \tag{769}$$

for some small enough constant $C_1 > 0$. We first show there exists $t_1 \in [s', s]$ such that

$$e^{a_{b,0}^{(t_1)}} \geqslant \frac{C\left(a_{p,1}^{(t_1)}\right)^{1-\frac{6}{N-1}}}{Nk} \tag{770}$$

for some small enough constant $C > 0$. Suppose this does not hold, i.e., for all $t \in [s', s]$, we have

$$e^{a_{b,0}^{(t)}} < \frac{C\left(a_{p,1}^{(t)}\right)^{1-\frac{6}{N-1}}}{Nk}. \tag{771}$$

Then same as for (757), we have for all $t \in [s', s]$ and $h \in [k-1]$:

$$\pi_{1,2h}^{(t)}(\texttt{down}_\times^1) \leqslant \left(1-\pi_{1,2h}^{(t)}(\texttt{up})\right)\exp\left(-r_{\texttt{up},h,2}^{(t)}\left(a_{p,1}^{(t)}+a_{p,2}^{(t)}\right)\right) = o\left(\frac{1}{k}\right)(1 - \pi_{1,2h}^{(t)}(\texttt{up})\Big). \tag{772}$$

Then by (740) we have for all $t \in [s_0', s_0]$:

$$\begin{aligned}
\delta e^{a_{b,0}^{(t)}} &= \frac{1+o(1)}{N}\sum_{h=1}^{k-1}\frac{k-h}{k}\left(r_{\texttt{up},h,0}^{(t)}\right)^2 h e^{a_{b,2}^{(t)}} a_{p,2}^{(t)}\left(1-\pi_{1,2h}^{(t)}(\texttt{up})\right)\\
&\quad - \frac{1+o(1)}{N}\sum_{h=1}^{k-1}\frac{k-h}{k}\frac{h e^{a_{b,1}^{(t)}} e^{2a_{b,0}^{(t)}}}{\left(e^{a_{b,0}^{(t)}}+h e^{a_{b,1}^{(t)}}+h e^{a_{b,2}^{(t)}}\right)^2}\Big(a_{p,1}^{(t)}\pi_{1,2h+1}^{(t)}(\texttt{down}_\times)\\
&\quad + a_{p,2}^{(t)}\pi_{1,2h+1}^{(t)}(\texttt{up})\Big)\\
&\geqslant \frac{1+o(1)}{Nk}\left(a_{p,1}^{(t)}\right)^{1-\frac{6}{N-1}}\delta a_{q,\times}^{(t)}\\
&\quad - \frac{1+o(1)}{N}\frac{e^{2a_{b,0}^{(t)}}}{e^{a_{b,0}^{(t)}}+e^{a_{b,1}^{(t)}}+e^{a_{b,2}^{(t)}}}\left(k a_{p,1}^{(t)}\delta a_{p,1}^{(t)} + \frac{a_{p,2}^{(t)}}{\exp\left(0.1 a_{p,2}^{(t)}\right)}\delta a_{q,\times}^{(t)}\right)\\
&\geqslant \frac{1+o(1)}{Nk}\left(a_{p,1}^{(t)}\right)^{1-\frac{6}{N-1}}\delta a_{q,\times}^{(t)} - \frac{C^2+o(1)}{Nk^2}\left(a_{p,1}^{(t)}\right)^{1-\frac{10}{N-1}}\delta a_{p,1}^{(t)},
\end{aligned} \tag{773}$$

where the second relation follows from (156), (752), (737), (542) and (603) (which also holds here), and the third relation follows from (162), (164) and (771). From (773) we deduce

$$\begin{aligned}
e^{a_{b,0}^{(s)}} - e^{a_{b,0}^{(s')}} &\geqslant \frac{1+o(1)}{Nk}\left(a_{p,1}^{(s')}\right)^{1-\frac{6}{N-1}}\left(a_{q,\times}^{(s)}-a_{q,\times}^{(s')}\right) - \frac{C^2+o(1)}{Nk^2}\\
&\quad \left(a_{p,1}^{(s)}\right)^{1-\frac{10}{N-1}}\left(a_{p,1}^{(s)}-a_{p,1}^{(s')}\right).
\end{aligned} \tag{774}$$

From (134), (140) we know that there exist constants $c_1, c_2 > 0$, $c_2 > c_1$ such that for all $t \in [s', s]$:

$$c_1\frac{a_{p,1}^{(t_1)}}{k} \leqslant a_{q,\times}^{(t_1)} \leqslant c_2\frac{a_{p,1}^{(t_1)}}{k}. \tag{775}$$

We can choose $C_1 \leqslant \frac{c_1}{2c_2}$ in (769), which guarantees

$$a_{q,\times}^{(s_0)} - a_{q,\times}^{(s_0')} \geqslant c_1\frac{a_{p,1}^{(s_0)}}{k} - c_2\frac{a_{p,1}^{(s_0')}}{k} \geqslant c_1\frac{a_{p,1}^{(s_0)}}{2k}. \tag{776}$$

Plugging this into (774), we have

$$\begin{aligned}
e^{a_{b,0}^{(s)}} - e^{a_{b,0}^{(s')}} &\geqslant \frac{1+o(1)}{Nk^2}\frac{c_1^2}{4c_2}\left(a_{p,1}^{(s)}\right)^{2-\frac{6}{N-1}} - \frac{C^2+o(1)}{Nk^2}\left(a_{p,1}^{(s)}\right)^{2-\frac{10}{N-1}}\\
&\gtrsim \frac{1}{Nk^2}\left(a_{p,1}^{(s)}\right)^{2-\frac{6}{N-1}} \overset{(753)}{\gg} \frac{\left(a_{p,1}^{(s)}\right)^{1-\frac{6}{N-1}}}{Nk}
\end{aligned} \tag{777}$$

if we set C small enough. This contradicts (771). Therefore, there exists $t_1 \in [s', s]$ such that (770) holds. Then by a similar argument as how we show (156), we can show (163) holds for all $t \in [s', s]$ again by contradiction.

By (158), (156) and (216) we know that

$$\forall t \in [T_{1,4}, s]: \quad a_{b,2}^{(t)} < 0. \tag{778}$$

This shows the second relation of (164) holds at step s. Therefore, at time s we have established (158), (162), and both relations in (164). Applying the same algebraic estimate from (132) at time s, we obtain

$$\frac{1}{k-1} - \frac{c_1}{k^4} \leqslant r_{\texttt{down}_\times}^{(s)} \leqslant \frac{1}{k-1},$$

possibly after enlarging the constant c_1. Combining this with (735), we conclude that

$$\forall t \in [T_{1,4}, s]: \quad \frac{1}{k-1} - \frac{c_1}{k^4} \leqslant r_{\texttt{down}_\times}^{(t)} \leqslant \frac{1}{k-1}. \tag{779}$$

When $s = T_1$: We first show (174). Same as in (484), we have for any $t \in \mathbb{N}$:

$$\begin{aligned} &R(\pi^{(t)}; \mathcal{P}_1) \\ &= \frac{1}{k}\left(1 - \pi_{1,1}^{(t)}(\texttt{up})\right) + \sum_{h=1}^{k-1} \frac{1}{k}\Bigg(\prod_{i=1}^{h-1} \pi_{1,2i}^{(t)}(\texttt{up}) \prod_{i=1}^{h}\Big(1 - (i-1)\pi_{1,2i-1}^{(t)}(\texttt{down}_\times) \\ &\quad - \pi_{1,2i-1}^{(t)}(\texttt{up})\Big) \\ &\quad \cdot \pi_{1,2h}^{(t)}(\texttt{up})\left(1 - \pi_{1,2h+1}^{(t)}(\texttt{up}) - h\pi_{1,2h+1}^{(t)}(\texttt{down}_\times)\right)\Bigg). \end{aligned} \tag{780}$$

From (125) we know that a sufficient condition for $R(\pi^{(t)}; \mathcal{P}_1) \geqslant 1 - \epsilon^{1.05}$ is

$$\begin{aligned} &\pi_{1,1}^{(t)}(\texttt{up}) \leqslant \frac{\epsilon^{1.05}}{4k}, \quad \max_{h\in[k-1]}\{1 - \pi_{1,2h}^{(t)}(\texttt{up})\} \leqslant \frac{\epsilon^{1.05}}{2k}, \quad \max_{h\in[k-1]}\{h \\ &\quad \pi_{1,2h+1}^{(t)}(\texttt{down}_\times)\} \leqslant \frac{\epsilon^{1.05}}{4k}, \end{aligned} \tag{781}$$

where we use the fact that

$$\max\{\pi_{1,2h-1}^{(t)}(\texttt{up})\}_{h\in[k]} = \pi_{1,1}^{(t)}(\texttt{up}) \tag{782}$$

guaranteed by (570). Therefore, by our curriculum design, we have

$$\max\left\{\pi_{1,1}^{(T_1-1)}(\texttt{up}), \max_{h\in[k-1]}\{1 - \pi_{1,2h}^{(T_1-1)}(\texttt{up})\}, \max_{h\in[k-1]}\{h\pi_{1,2h+1}^{(T_1-1)}(\texttt{down}_\times)\}\right\} \geqslant \frac{\epsilon^{1.05}}{4k}. \tag{783}$$

By (159), (160), (161), (571), (573), (567), (134) and (157) we know that (783) indicates

$$r_{\texttt{down}_\times}^{(T_1-1)}\left(a_{p,1}^{(T_1-1)} + a_{p,2}^{(T_1-1)}\right) \leqslant \left(1 + \frac{c_1}{k}\right)\ln(k^2/\epsilon^{1.05}), \tag{784}$$

for some constant $c_1 > 0$. This combined with (779), (140) further gives

$$a_{p,1}^{(T_1-1)} \lesssim k\ln(1/\epsilon). \tag{785}$$

By (784), (779), (140) and the fact that $a_{p,1}^{(t)}$ is increasing in t (c.f. (137)), we have for all $t \in [T_{1,4}, T_1 - 1]$:

$$\begin{aligned} \sum_{h=1}^{k-1} \frac{k-h}{k}\pi_{1,2h+1}^{(t)}(\texttt{down}_\times) &\overset{(572)}{\geqslant} \frac{1+o(1)}{k}\exp\left(-r_{\texttt{down}_\times}^{(t)}\left(a_{p,1}^{(t)} + a_{p,2}^{(t)}\right)\right) \geqslant (1 \\ &\quad + o(1))\frac{\epsilon^{1.05+\frac{c_2}{k}}}{k^3} \end{aligned} \tag{786}$$

for some constant $c_2 > 0$. By this and (542) we know that for all $t \in [T_{1,4}, T_1 - 1]$:

$$\delta a_{p,1}^{(t)} \gtrsim \frac{\epsilon^{1.05+\frac{c_2}{k}}}{k^4} \tag{787}$$

(787) together with (785) suggests

$$T_1 - T_{1,4} \lesssim \frac{k^5 \ln(1/\epsilon)}{\eta \epsilon^{1.05+\frac{c_2}{k}}} \tag{788}$$

By (734) and our assumption that $\ln(1/\epsilon) \geqslant e^k$, we have

$$T_1 \lesssim \frac{k^5 \ln(1/\epsilon)}{\eta \epsilon^{1.05+\frac{c_2}{k}}} \lesssim \frac{1}{\eta \epsilon^{1.1}}. \tag{789}$$

This gives (789).

On the other hand, by (780) and the fact that $R\left(\pi^{(T_1)}; \mathcal{P}_1\right) \geqslant 1 - \frac{\epsilon}{k^L}$ we know that

$$\pi_{1,1}^{(T_1)}(\texttt{up}) \leqslant \epsilon^{1.05}, \quad 1 - \pi_{1,2}^{(T_1)}(\texttt{up}) \leqslant \frac{k}{k-1}\epsilon^{1.05}, \quad \pi_{1,2k-1}^{(T_1)}(\texttt{down}_\times) \leqslant k\epsilon^{1.05}. \tag{790}$$

Thus by the first relation and (567) we have

$$\left(1 + \frac{1}{k}\right) a_{q,0}^{(T_1)} - \frac{a_{p,0}^{(T_1)}}{k} \geqslant \ln\left(\frac{1}{k\epsilon^{1.05}} - \frac{1}{k}\right), \tag{791}$$

which gives (167). (167) combined with (558) gives (169). (167) together with Induction Hypothesis 2 and (785) also yields (165). From (165) and (163) we know that (170) holds.

When $t \geqslant s_0$ (where s_0 is defined in (753)), similar to (746), by (163) and (741) we have

$$\begin{aligned}
&e^{a_{b,1}^{(t)}} I_{h,1}^{(t)} \\
&\geqslant \left(1 + O\left(\frac{1}{k}\right)\right) \frac{e^{a_{b,1}^{(t)}} e^{a_{b,0}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} a_{p,1}^{(t)} \cdot \frac{he^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \\
&\quad \pi_{1,2h+1}^{(t)}(\texttt{down}_\times) \\
&\geqslant \left(1 + O\left(\frac{1}{k}\right)\right) \frac{\left(a_{p,1}^{(t)}\right)^{2-\frac{6}{N-1}}}{Nk^2} \frac{he^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(\texttt{down}_\times) \\
&\geqslant \left(1 + O\left(\frac{1}{k}\right)\right) \frac{\left(a_{p,1}^{(t)}\right)^{2-\frac{6}{N-1}}}{Nk^2} \left(\frac{he^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}} \pi_{1,2h}^{(t)}(\texttt{down}_\times^1) \right. \\
&\quad + \frac{he^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(\texttt{down}_\times) - \frac{he^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}} \\
&\quad \left. \pi_{1,2h}^{(t)}(\texttt{down}_\times^1) \right),
\end{aligned} \tag{792}$$

Then similar to (748), we have

$$\delta e^{a_{b,1}^{(t)}} \geqslant \frac{1 + o(1)}{N^2k^2} \left(a_{p,1}^{(t)}\right)^{2-\frac{6}{N-1}} \left(k\delta a_{p,1}^{(t)} - \frac{1}{2k}\delta a_{q,\times}^{(t)}\right), \tag{793}$$

and same as for (749), this yields

$$e^{a_{b,1}^{(T_1)}} \gtrsim \frac{1}{N^2k} \left(a_{p,1}^{(T_1)}\right)^{3-\frac{6}{N-1}} \overset{(162)}{\gtrsim} \frac{1}{N} \left(a_{p,1}^{(T_1)}\right)^{1-\frac{5}{N}} e^{a_{b,0}^{(T_1)}}. \tag{794}$$

This combined with (165) gives (171). And by (170), (171), (164) (216) we obtain (172).

Further, by (164), (163), (140), (138) and (131) we have

$$-\frac{a_{p,0}^{(T_1)}}{k} \leqslant \mu_2^{(T_1)} \leqslant -\frac{a_{p,0}^{(T_1)}}{k} + \left(\frac{a_{p,0}^{(T_1)}}{k} + a_{p,2}^{(T_1)}\right) \frac{Nk}{\left(a_{p,1}^{(T_1)}\right)^{1-\frac{6}{N-1}}}$$

$$= -\frac{a_{p,0}^{(T_1)}}{k} + \left(1 + O\left(\frac{1}{k}\right)\right) N \left(a_{p,1}^{(T_1)}\right)^{\frac{6}{N-1}}. \tag{795}$$

By Lemma 7, the second relation in (790), (167), (559) and (795) we have (168) holds. From the third relation in (790), (167), Lemma 7 and Induction Hypothesis 2, we have

$$k\epsilon^{1.05} \geqslant \pi_{1,2k-1}^{(T_1)}(\texttt{down}_\times) = \frac{\exp\left(-r_{\texttt{down}_\times}^{(T_1)}\left(a_{p,1}^{(T_1)} + a_{p,2}^{(T_1)}\right)\right)}{(k-1)\exp\left(-r_{\texttt{down}_\times}^{(T_1)}\left(a_{p,1}^{(T_1)} + a_{p,2}^{(T_1)}\right)\right) + 1 + o(\epsilon)}. \tag{796}$$

By (171), (164) and (132) we have

$$r_{\texttt{down},\times}^{(T_1)} = \frac{1}{k-1} + O\left(\frac{1}{k^2 e^{a_{b,1}^{(T_1)}}}\right) \overset{(171)}{=} \frac{1}{k-1} + o\left(\frac{1}{k\ln(1/\epsilon)}\right). \tag{797}$$

By the above two expressions and (784), we obtain (166).

Finally, we show (173). Specifically, we'll show that given any tree with depth $l \in \{2, \cdots, L\}$ drawn from the balanced training distribution $\mathcal{P}_l$, under the event that the goal is not at the first k leaves $\pi^{(T_1)}$ visits (which happens with probability $1 - 1/k^{l-1}$), with probabillity $1 - o(1)$, the agent will go all the way down to a leaf, and visit all its siblings and go back to its parent node. At this node, with probability $1 - o(1)$, the agent will choose to go down again and visit a leaf that's already been visited, and the task fails.

Below we give the formal proof. All probabilities here are under policy $\pi^{(T_1)}$ and the balanced training distribution $\mathcal{P}_l$ that satisfies (21) for a fixed $l \in \{2, 3, \cdots, L\}$. We define events

$$\mathcal{C}_1 := \left\{\text{goal is not at the first } k \text{ leaves } \pi^{(T_1)} \text{ visits, and } a_h \in [k], \forall h \in [l]\right\}, \tag{798}$$

$$\mathcal{C}_2 := \{a_l = \texttt{up}\}, \tag{799}$$

and for all $i \in \{2, 3, \cdots, k\}$:

$$\mathcal{C}_{2i-1} := \left\{a_{l+2(i-1)} \in [k] \backslash \mathcal{A}_{l+2(i-1),\times}\right\}, \tag{800}$$

$$\mathcal{C}_{2i} := \{a_{l+2i-1} = \texttt{up}\}. \tag{801}$$

Then

$$\begin{aligned}
1 - R\left(\pi^{(T_1)}; \mathcal{P}_l\right) &\geqslant \mathbb{P}(\mathcal{C}_1)\Big(\underbrace{\mathbb{P}(a_{l+1} \neq \texttt{up}|\mathcal{C}_1)}_{=1-\mathbb{P}(\mathcal{C}_2|\mathcal{C}_1)} \\
&\quad + \mathbb{P}(\mathcal{C}_2|\mathcal{C}_1)\mathbb{P}(\{\pi^{(T_1)} \text{ fails at some step } h > l+1\}|\mathcal{C}_1\mathcal{C}_2)\Big) \\
&\geqslant \mathbb{P}(\mathcal{C}_1)\mathbb{P}(\{\pi^{(T_1)} \text{ fails at some step } h > l+1\}|\mathcal{C}_1\mathcal{C}_2) \\
&\geqslant \mathbb{P}(\mathcal{C}_1)\left(\mathbb{P}(a_{l+2} \in \mathcal{A}_{l+2,\times}|\mathcal{C}_1\mathcal{C}_2) + \mathbb{P}(\mathcal{C}_3|\mathcal{C}_1\mathcal{C}_2)\mathbb{P}\right. \\
&\quad \left.\left(\{\pi^{(T_1)} \text{ fails at some step } h > l+2\}\Big|\bigcap_{j=1}^{3}\mathcal{C}_j\right)\right) \\
&\geqslant \mathbb{P}(\mathcal{C}_1)\mathbb{P}(a_{l+2} \neq \texttt{up}|\mathcal{C}_1\mathcal{C}_2)\mathbb{P}\left(\{\pi^{(T_1)} \text{ fails at some step } h > l+2\}\Big|\bigcap_{j=1}^{3}\mathcal{C}_i\right) \\
&\geqslant \cdots \\
&\geqslant \mathbb{P}(\mathcal{C}_1)\prod_{i=1}^{k}\mathbb{P}\left(a_{l+2i} \neq \texttt{up}\Big|\bigcap_{j=1}^{2i}\mathcal{C}_{2j}\right).
\end{aligned} \tag{802}$$

Now we bound $\mathbb{P}(\mathcal{C}_1)$ and $\mathbb{P}(a_{l+2i} \neq \texttt{up}|\mathcal{C}_{2i})$ ($i \in [k]$) that appear in the last line of the above inequality. First, we have

$$\mathbb{P}\left(\text{goal is not at the first } k \text{ leaves } \pi^{(T_1)} \text{ visits}\right) = 1 - \frac{1}{k^{l-1}}. \tag{803}$$

And by Lemma 1, Induction Hypothesis 2, what we have proven to hold at time T_1 in Lemma 3, and our assumption that $\ln(1/\epsilon) \geqslant e^k$, we have for all $h \in [l]$, conditioned on $\mathcal{C}_1$:

$$\begin{aligned}
\pi_h(\texttt{up}) &= \frac{\exp\left(-\left(1+\frac{1}{k}\right)a_{q,0}^{(T_1)}\right)}{\exp\left(-\left(1+\frac{1}{k}\right)a_{q,0}^{(T_1)}\right)} \\
&\quad + \sum_{i=1}^{k} \exp\left(\frac{-\frac{a_{p,0}^{(T_1)}}{k}e^{a_{b,0}^{(T_1)}} + \left(-N_h(i)(a_{p,1}^{(T_1)}+a_{p,2}^{(T_1)}) + (h-1)a_{p,2}^{(T_1)}\right)e^{a_{b,2}^{(T_1)}}}{e^{a_{b,0}^{(T_1)}} + (h-1)e^{a_{b,2}^{(T_1)}}}\right) \\
&\leqslant \frac{\exp\left(-\left(1+\frac{1}{k}\right)a_{q,0}^{(T_1)} + \frac{a_{p,0}^{(T_1)}}{k}\right)}{\exp\left(-\left(1+\frac{1}{k}\right)a_{q,0}^{(T_1)} + \frac{a_{p,0}^{(T_1)}}{k}\right)} = \epsilon^{1.05+o(1)}. \qquad (804) \\
&\quad + \sum_{i=1}^{k} \exp\left(\frac{\left(-a_{p,1}^{(T_1)} + \frac{a_{p,0}^{(T_1)}}{k}\right)(h-1)e^{a_{b,2}^{(T_1)}}}{e^{a_{b,0}^{(T_1)}} + (h-1)e^{a_{b,2}^{(T_1)}}}\right)
\end{aligned}$$

Combining the above two facts, we have

$$\mathbb{P}(\mathcal{C}_1) \geqslant \left(1-\frac{1}{k^{l-1}}\right)\left(1-\left(\frac{\epsilon}{k^L}\right)^{1+o(1)}\right)^l = 1-\frac{1+o(1)}{k^{l-1}}. \qquad (805)$$

Similarly, by Lemma 1 and Induction Hypothesis 2, we have for all $i \in [k]$:

$$\begin{aligned}
&\mathbb{P}\left(a_{l+2i} = \texttt{up} \middle| \bigcap_{j=1}^{2i} \mathcal{C}_{2j}\right) \\
&\leqslant \frac{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}^{(T_1)}\right)}{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}^{(T_1)}\right) + (k-i)\exp\left((1+o(1))a_{p,2}^{(T_1)}\right)} \\
&\qquad + i\exp\left(\frac{1}{i}\left((i-1)a_{p,2}^{(T_1)} - a_{p,1}^{(T_1)} + o\left(\frac{1}{k^2}\right)a_{p,1}^{(T_1)}\right)\right) \\
&\overset{(140)}{\leqslant} \frac{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}^{(T_1)}\right)}{k\exp\left(O\left(\frac{1}{k^2}\right)a_{p,1}^{(T_1)}\right)} \overset{(167)}{\leqslant} \epsilon^{1.05+o(1)}. \qquad (806)
\end{aligned}$$

Combining (805), (806) and (802), we have

$$1 - R\left(\pi^{(T_1)}; \mathcal{P}_l\right) \geqslant 1 - \frac{1+o(1)}{k^{l-1}}. \qquad (807)$$

This gives (173).

C.3.6 Proof of Lemma 7

From the proof of Lemma 5 in Appendix C.3.3 we know that (246), (247), (248) and (249) still hold at Phase 1.2. Note that when $l = 1$, for all $h \in [k-1]$, we have

$$\mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right) = \underbrace{\frac{k-h}{k}}_{\text{(a)}} \prod_{i=1}^{h}\left(1-(i-1)\pi_{1,2i-1}^{(t)}(\texttt{down}_\times) - \pi_{1,2i-1}^{(t)}(\texttt{up})\right)\prod_{i=1}^{h-1}\pi_{1,2i}^{(t)}(\texttt{up}). \qquad (808)$$

where (a) represents the probability of the event that the goal is not at the first h leaves visited. By Induction Hypothesis 2 and Lemma 1 we can compute that for all $h \in [k-1]$,

$$\mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right) = \left(1+o\left(\frac{1}{k}\right)\right)\frac{k-h}{k}, \qquad (809)$$

and

$$1 - V_{1,2h+1}^{(t)} = o(1). \tag{810}$$

Combining (246), (247), (248) and (249) with (809) and (810), we obtain (537), (538), (539) and (540).

Note that by our notation, we have for all $h \in [k]$:

$$|\mathcal{A}_{2h,\times}| \mid \mathcal{E}_{2h} = 0, \quad |\mathcal{A}_{2h-1,\times}| \mid \mathcal{E}_{2h-1} = h-1, \quad N_{2h-1}(\mathtt{up}) = N_{2h}(\mathtt{up}) = h-1. \tag{811}$$

Thus we have

$$\begin{aligned}
\delta a_{p,0}^{(t)} &\overset{(81)}{=} \mathbb{E}\left[\sum_{h=1}^{k-1} \frac{e^{a_{b,0}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}} \pi_{1,2h}^{(t)}(\mathtt{up}) \left(1 - \pi_{1,2h}^{(t)}(\mathtt{up})\right) Q_{2h}^{(t)}(\mathtt{up}) \middle| l = 1\right] \\
&\quad - \mathbb{E}\left[\pi_{1,1}^{(t)}(\mathtt{up}) G_1^{(t)} | l = 1\right] - \mathbb{E}\left[\sum_{h=1}^{k-1} \frac{e^{a_{b,0}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \right. \\
&\quad \left. \pi_{1,2h+1}^{(t)}(\mathtt{up}) G_{2h+1}^{(t)} \middle| l = 1\right], \\
&\overset{(510)}{=} -\pi_{1,1}^{(t)}(\mathtt{up}) V_{1,1}^{(t)} + \sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(t)}(\mathtt{up}) \left(\frac{e^{a_{b,0}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}} (1 \right. \\
&\quad - \pi_{1,2h}^{(t)}(\mathtt{up})\Big) \\
&\quad \left. - \frac{e^{a_{b,0}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(\mathtt{up}) \right) V_{1,2h+1}^{(t)},
\end{aligned} \tag{812}$$

$$\begin{aligned}
\delta a_{p,1}^{(t)} &\overset{(82)}{=} \mathbb{E}\left[\sum_{h=1}^{k-1} \frac{h}{k} \frac{e^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(j) G_{2h+1}^{(t)} \middle| j \in \mathcal{A}_{2h+1,\times}, l = 1\right] \\
&\quad + \mathbb{E}\left[\sum_{h=1}^{k-1} \frac{N_{2h}(j) e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}} \pi_{1,2h}^{(t)}(j) \pi_{1,2h}^{(t)}(\mathtt{up}) Q_{2h}^{(t)}(\mathtt{up}) \middle| j \in [k], l = 1\right] \\
&= \mathbb{E}\left[\sum_{h=1}^{k-1} \frac{h}{k} \frac{e^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(j) G_{2h+1}^{(t)} \middle| j \in \mathcal{A}_{2h+1,\times}, l = 1\right] \\
&\quad + \mathbb{E}\left[\sum_{h=1}^{k-1} \frac{h}{k} \frac{e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}} \pi_{1,2h}^{(t)}(j) \pi_{1,2h}^{(t)}(\mathtt{up}) Q_{2h}^{(t)}(\mathtt{up}) \right. \\
&\quad \left. \middle| j \in \mathcal{A}_{2h-1,\times} \cup \{a_{2h-1}\}, l = 1\right] \\
&\overset{(510)}{=} \frac{1}{k} \sum_{h=1}^{k-1} \mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right) \pi_{1,2h}^{(t)}(\mathtt{up}) \left(\frac{he^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}} \pi_{1,2h}^{(t)}(\mathtt{down}_\times^1) \right. \\
&\quad \left. + \frac{he^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(\mathtt{down}_\times) \right) V_{1,2h+1}^{(t)},
\end{aligned} \tag{813}$$

$$\begin{aligned}
\delta a_{p,3}^{(t)} &\overset{(83)}{=} -\mathbb{E}\left[\sum_{h=1}^{k-1} \frac{h}{k} \frac{e^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(\mathtt{up}) G_{2h+1}^{(t)} \middle| j \in \mathcal{A}_{2h+1,\times}, l = 1\right] \\
&\quad + \mathbb{E}\left[\sum_{h=1}^{k-1} \frac{N_{2h}(j) e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}} + (2h-1)e^{a_{b,2}^{(t)}}} \pi_{1,2h}^{(t)}(\mathtt{up}) \left(1 - \pi_{1,2h}^{(t)}(\mathtt{up})\right) Q_{2h}^{(t)}(\mathtt{up}) \middle| j \in [k], l = 1\right] \\
&= -\mathbb{E}\left[\sum_{h=1}^{k-1} \frac{h}{k} \frac{e^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}} + he^{a_{b,1}^{(t)}} + he^{a_{b,2}^{(t)}}} \pi_{1,2h+1}^{(t)}(\mathtt{up}) G_{2h+1}^{(t)} \middle| j \in \mathcal{A}_{2h+1,\times}, l = 1\right]
\end{aligned}$$

$$+\mathbb{E}\left[\sum_{h=1}^{k-1}\frac{h}{k}\frac{e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}}+(2h-1)e^{a_{b,2}^{(t)}}}\pi_{1,2h}^{(t)}(\text{up})\left(1-\pi_{1,2h}^{(t)}(\text{up})\right)Q_{2h}^{(t)}(\text{up})\right.$$
$$\left.\Big|j\in\mathcal{A}_{2h-1,\times}\cup\{a_{2h-1}\},l=1\right]$$
$$\overset{(510)}{=}\frac{1}{k}\sum_{h=1}^{k-1}\mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(\text{up})\left(\frac{he^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}}+(2h-1)e^{a_{b,2}^{(t)}}}\left(1-\pi_{1,2h}^{(t)}(\text{up})\right)\right.$$
$$\left.-\frac{he^{a_{b,1}^{(t)}}}{e^{a_{b,0}^{(t)}}+he^{a_{b,1}^{(t)}}+he^{a_{b,2}^{(t)}}}\pi_{1,2h+1}^{(t)}(\text{up})\right)V_{1,2h+1}^{(t)},\tag{814}$$

and

$$\delta a_{p,\text{up}}^{(t)}\overset{(84)}{=}\mathbb{E}\left[\sum_{h=1}^{k-1}\frac{he^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}}+he^{a_{b,1}^{(t)}}+he^{a_{b,2}^{(t)}}}\pi_{1,2h+1}^{(t)}(\text{up})G_{2h+1}^{(t)}\Big|l=1\right]$$
$$-\mathbb{E}\left[\sum_{h=1}^{k-1}\frac{(h-1)e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}}+(2h-1)e^{a_{b,2}^{(t)}}}\pi_{1,2h}^{(t)}(\text{up})\left(1-\pi_{1,2h}^{(t)}(\text{up})\right)Q_{2h}^{(t)}(\text{up})\Big|l=1\right]$$
$$=\sum_{h=1}^{k-1}\mathbb{P}^{(t)}\left(\mathcal{E}_{1,2h}\right)\pi_{1,2h}^{(t)}(\text{up})\left(\frac{he^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}}+he^{a_{b,1}^{(t)}}+he^{a_{b,2}^{(t)}}}\pi_{1,2h+1}^{(t)}(\text{up})\right.$$
$$\left.-\frac{(h-1)e^{a_{b,2}^{(t)}}}{e^{a_{b,0}^{(t)}}+(2h-1)e^{a_{b,2}^{(t)}}}\left(1-\pi_{1,2h}^{(t)}(\text{up})\right)\right)V_{1,2h+1}^{(t)}.\tag{815}$$

Combining the above expressions with (809), (810), we obtain (541), (542), (543) and (544).

Note that (518)-(523) computed in the proof of Lemma 5 in Appendix C.3.3 still hold at Phase 1.2, and

$$\delta a_{b,0}^{(t)}\overset{(73)}{=}\frac{1}{N}\left(\mathbb{E}\left[\sum_{h=1}^{k-1}\frac{e^{a_{b,0}^{(t)}}}{(e^{a_{b,0}^{(t)}}+(2h-1)e^{a_{b,2}^{(t)}})^2}\left((2h-1)e^{a_{b,2}^{(t)}}\beta_{2h,0}^{(t)}\right.\right.\right.$$
$$\left.-e^{a_{b,1}^{(t)}}\beta_{2h,1}^{(t)}-e^{a_{b,2}^{(t)}}\beta_{2h,2}^{(t)}\right)\Big|l=1\Big]$$
$$+\mathbb{E}\left[\sum_{h=1}^{k-1}\frac{e^{a_{b,0}^{(t)}}}{(e^{a_{b,0}^{(t)}}+he^{a_{b,1}^{(t)}}+he^{a_{b,2}^{(t)}})^2}\left(\left(he^{a_{b,1}^{(t)}}+he^{a_{b,2}^{(t)}}\right)\beta_{2h+1,0}^{(t)}\right.\right.$$
$$\left.\left.\left.-e^{a_{b,1}^{(t)}}\beta_{2h+1,1}^{(t)}-e^{a_{b,2}^{(t)}}\beta_{2h+1,2}^{(t)}\right)\Big|l=1\right]\right),\tag{816}$$

$$\delta a_{b,1}^{(t)}\overset{(74)}{=}\frac{1}{N}\left(\mathbb{E}\left[\sum_{h=1}^{k-1}\frac{e^{a_{b,1}^{(t)}}}{(e^{a_{b,0}^{(t)}}+(2h-1)e^{a_{b,2}^{(t)}})^2}\left(e^{a_{b,0}^{(t)}}+(2h-1)e^{a_{b,2}^{(t)}}\right)\beta_{2h,1}^{(t)}\Big|l=1\right]\right.$$
$$+\mathbb{E}\left[\sum_{h=1}^{k-1}\frac{e^{a_{b,1}^{(t)}}}{(e^{a_{b,0}^{(t)}}+he^{a_{b,1}^{(t)}}+he^{a_{b,2}^{(t)}})^2}\left(\left(e^{a_{b,0}^{(t)}}+he^{a_{b,2}^{(t)}}\right)\beta_{2h+1,1}^{(t)}\right.\right.$$
$$\left.\left.\left.-he^{a_{b,0}^{(t)}}\beta_{2h+1,0}^{(t)}-he^{a_{b,2}^{(t)}}\beta_{2h+1,2}^{(t)}\right)\Big|l=1\right]\right).\tag{817}$$

Plugging (518)-(523) into the above two expressions, and utilizing (809), (810), we obtain (545) and (547).

Finally, plugging (811) into (88)-(91) in Lemma 1, we obtain the policy expressions in Lemma 7.

C.3.7 Proof of Lemma 4

During depth-2 curriculum training, we only train matrix P starting from $P^{(T_1)}$. We fix $B = B^{(T_1)}, C = C^{(T_1)}, Q = Q^{(T_1)}$ and drop the superscript (T_1) corresponding to these variables. All the expectations/probabilities are taken over training distribution P_2. Below we show Lemma 4 by induction.

Basic case: at step T_1: (175), (176), (177), (181) hold trivially at step T_1. By (581) and (134) we have

$$|a_{p,3}^{(T_1)}| \lesssim \frac{a_{q,0}}{k}. \tag{818}$$

Thus (178) holds at step T_1. By (165), (134) and (140) we know that (179) holds at step T_1.

We make the following induction hypothesis:

Induction Hypothesis 3: For all $t \leqslant s-1$ ($s \in [T_1+1, T_2]$), the relations in Lemma 4 hold.

Below we show Lemma 4 holds for $t = s$ under Induction Hypothesis 3.

Define

$$\overline{\pi}_2^{(t)}(\mathtt{up}) := \frac{\exp\left(\left(1+\frac{1}{k}\right)\frac{ke^{-2a_{c,0}}a_{q,\times}-(k+2)a_{q,0}}{ke^{-2a_{c,0}}+k+2}\right)}{k\exp\left(-\frac{a_{p,3}^{(t)}}{k}+\frac{1}{e^{a_{b,1}-a_{b,0}}}\left(-\frac{a_{p,0}^{(t)}}{k^2}+\frac{a_{p,3}^{(t)}}{k}\right)\right)} + \exp\left(\left(1+\frac{1}{k}\right)\frac{ke^{-2a_{c,0}}a_{q,\times}-(k+2)a_{q,0}}{ke^{-2a_{c,0}}+k+2}\right). \tag{819}$$

Let h_i denote the first timestep that the i-th child arranged in order of occurrence is visited, i.e.,

$$\forall i \in [k]: \quad h_i := \inf_{h\in\mathbb{N}_+}\{\text{at step } h,\ i \text{ chilrden of the root is visited}\}. \tag{820}$$

Then $h_i = +\infty$ if the event that i children of the root are visited never happens. Define

$$\overline{\mathcal{E}}_{2,i} := \{\text{the agent traverses all children of } n_{h_j} \text{ for all } j \in [i] \text{ at the first } h_i+2k-1 \text{ steps}, \text{ and arrives at } n_{h_i} \text{ at step } h_i+2k, l=2\} \tag{821}$$

($\overline{\mathcal{E}}_{2,k}$ is an impossible event). And define

$$Q_{2,i}^{(t)}(\mathtt{up}) := Q_{h_i+2k}^{(t)}(\mathtt{up})|\overline{\mathcal{E}}_{2,i},\ \forall i \in [k-1]. \tag{822}$$

We first simplify the gradient expressions related to P in Lemma 1 when $l=2$.

Lemma 8. *Under Induction Hypothesis 3, we have for any $t \in [T_1, s-1]$:*

$$\delta a_{p,1}^{(t)} = \frac{1+o(1)}{k^2}\sum_{i=1}^{k-1}\frac{k-i}{k}\left(\overline{\pi}_2^{(t)}(up)\right)^i\left(1-\overline{\pi}_2^{(t)}(up)\right)Q_{2,i}^{(t)}(up), \tag{823}$$

$$\delta a_{p,3}^{(t)} = \frac{1}{k}\sum_{i=1}^{k-1}\frac{k-i}{k}\left(\overline{\pi}_2^{(t)}(up)\right)^i\left((1+o(1))\left(1-\overline{\pi}_2^{(t)}(up)\right) - \frac{1}{(k-i)\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1}+\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}\right)Q_{2,i}^{(t)}(up), \tag{824}$$

$$|\delta a_{p,0}^{(t)}| \lesssim \frac{1}{(\ln(1/\epsilon))^{1+o(1)}}\left(\delta a_{p,1}^{(t)} - \frac{1}{k^2}\delta a_{p,3}^{(t)}\right), \tag{825}$$

$$|\delta a_{p,up}^{(t)}| \lesssim \frac{1}{(\ln(1/\epsilon))^{3+o(1)}}\left(\delta a_{p,1}^{(t)} - \frac{1}{k^2}\delta a_{p,3}^{(t)}\right). \tag{826}$$

The proof of Lemma 8 is postponed to Appendix C.3.8.

First, by (581) and (134) we have

$$|a_{p,3}^{(T_1)}| \lesssim \frac{a_{q,0}}{k}. \tag{827}$$

And (134), (140) guarantee

$$a_{p,1}^{(T_1)} \asymp k a_{q,0}^{(T_1)}. \tag{828}$$

Thus by the above two facts, (825), (826), our assumption that $\ln(1/\epsilon) \geqslant e^k$, which guarantees $k = \ln(1/\epsilon)^{o(1)}$, and that (178), (179) holds at step $s-1$ by Induction Hypothesis 3, we have (175), (176) hold at step s. In addition, from (177), (181), (182), (179) and (827) we know that the first inequality in (178) holds at step s. By (179), (181), (182) we deduce that the second inequality in (178) also holds at step s for any $s \in [T_1+1, T_2]$.

By the expression of $\overline{\pi}_2^{(t)}(\mathtt{up})$ in (819) we know that there exists a constant $C \in \mathbb{R}$ such that for $T_{2,1}$ defined in (180), if $s \leqslant T_{2,1}$, we have

$$\begin{aligned}\forall t \leqslant s-1: \quad & \frac{1}{\exp\left(\frac{a_{p,1}^{(t)} - a_{p,3}^{(t)}}{k-1} + \left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}\right)} = \exp\left(-c\frac{a_{q,0}}{\ln^2 k}\right)\Big(1 \\ & -\overline{\pi}_2^{(t)}(\mathtt{up})\Big) \\ & = o\left(1 - \overline{\pi}_2^{(t)}(\mathtt{up})\right)\end{aligned} \tag{829}$$

for some constant $c > 0$. By (829) and (824) we have

$$\begin{aligned}\forall t \leqslant s-1: \quad \delta a_{p,3}^{(t)} &= \frac{1+o(1)}{k} \sum_{i=1}^{k-1} \frac{k-i}{k} \left(\overline{\pi}_2^{(t)}(\mathtt{up})\right)^i \left(1 - \overline{\pi}_2^{(t)}(\mathtt{up})\right) Q_{2,i}^{(t)}(\mathtt{up}) \\ &\overset{(823)}{=} (1+o(1)) k \delta a_{p,1}^{(t)}.\end{aligned} \tag{830}$$

This proves (181) holds for $t = s$ when $s \leqslant T_{2,1}$.

When $s > T_{2,1}$, we show

$$\forall t \in [T_{2,1}, s]: \quad a_{p,3}^{(t)} \geqslant \frac{a_{p,1}^{(t)}}{2} + k\left(1 + \frac{C}{\ln^2 k}\right) a_{q,0} \tag{831}$$

by contradiction. Suppose this does not hold, then we could choose the smallest $t_0 \in [T_{2,1}+1, s]$ such that the above is violated, i.e.,

$$a_{p,3}^{(t_0)} < \frac{a_{p,1}^{(t_0)}}{2} + k\left(1 + \frac{C}{\ln^2 k}\right) a_{q,0}, \tag{832}$$

but

$$\forall t \in [T_{2,1}, t_0 - 1]: \quad a_{p,3}^{(t)} \geqslant \frac{a_{p,1}^{(t)}}{2} + k\left(1 + \frac{C}{\ln^2 k}\right) a_{q,0}. \tag{833}$$

Then by our update rule, we have

$$a_{p,3}^{(t_0-1)} = \frac{a_{p,1}^{(t_0-1)}}{2} + k\left(1 + \frac{C+o(1)}{\ln^2 k}\right) a_{q,0}, \tag{834}$$

which, together with (829) suggests

$$\begin{aligned}&\frac{1}{\exp\left(\frac{a_{p,1}^{(t_0-1)} - a_{p,3}^{(t_0-1)}}{k-1} + \left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}\right)} = \exp\left(-(c+o(1))\frac{a_{q,0}}{\ln^2 k}\right)\Big(1 \\ & \quad -\overline{\pi}_2^{(t_0-1)}(\mathtt{up})\Big)\end{aligned}$$

$$= o\left(1 - \overline{\pi}_2^{(t_0-1)}(\mathtt{up})\right), \tag{835}$$

and thus (830) also holds for $t = t_0 - 1$, i.e.,

$$\delta a_{p,3}^{(t_0-1)} = k(1 + o(1))\delta a_{p,1}^{(t_0-1)}. \tag{836}$$

This suggests

$$a_{p,3}^{(t_0)} - \frac{a_{p,1}^{(t_0)}}{2} > a_{p,3}^{(t_0-1)} - \frac{a_{p,1}^{(t_0-1)}}{2} \overset{(833)}{\geqslant} k\left(1 + \frac{C}{\ln^2 k}\right) a_{q,0}. \tag{837}$$

This contradicts (832). Thus (831) must hold. On thet other hand, there exists constant $C_1 > C$ such that for all $t \in [T_{2,1}, s]$, if

$$a_{p,3}^{(t)} \geqslant \frac{a_{p,1}^{(t)}}{2} + k\left(1 + \frac{C_1}{\ln^2 k}\right) a_{q,0}, \tag{838}$$

then we have

$$\frac{1}{\exp\left(\frac{a_{p,1}^{(t)} - a_{p,3}^{(t)}}{k-1} + \left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}\right)} = \exp\left(c\frac{a_{q,0}}{\ln^2 k}\right)\left(1 - \overline{\pi}_2^{(t)}(\mathtt{up})\right). \tag{839}$$

Then we can similarly show by contradiction that

$$\forall t \in [T_{2,1}, s]: \quad a_{p,3}^{(t)} \leqslant \frac{a_{p,1}^{(t)}}{2} + k\left(1 + \frac{C_1}{\ln^2 k}\right) a_{q,0}. \tag{840}$$

By (831) and (840) we know that (182) holds at step s.

When $s = T_2$. Note that

$$1 - R(\pi^{(T_2)}; \mathcal{P}_2) \geqslant \mathbb{P}^{\pi^{(T_2)}}\left(\overline{\mathcal{E}}_{2,1}\right)\left(1 - \pi_{2k+2}^{(T_2)}(\mathtt{up})|\overline{\mathcal{E}}_{2,1}\right), \tag{841}$$

where $\overline{\mathcal{E}}_{2,1}$ is defined in (821). Then from the policy computation in the proof of Lemma 8 (c.f. (878), (879), (882), (883), (885)) we know that

$$\mathbb{P}^{\pi^{(T_2)}}\left(\overline{\mathcal{E}}_{2,1}\right) \geqslant (1 + o\left(\epsilon^{0.5}\right))\frac{k-1}{k}, \tag{842}$$

and from (904) we know that

$$\pi_{2k+2}^{(T_2)}(\mathtt{up})|\overline{\mathcal{E}}_{2,1} = \left(1 + O\left(\frac{1}{(\ln(1/\epsilon))^{1+o(1)}}\right)\right)\overline{\pi}_2^{(T_2)}(\mathtt{up}). \tag{843}$$

Plugging (842), (843) into (841), and recall we set $\delta_2 = \frac{\epsilon^{1.02}}{k^L}$, we have

$$\left(1 + O\left(\frac{1}{(\ln(1/\epsilon))^{1+o(1)}}\right)\right)\frac{k-1}{k}\left(1 - \overline{\pi}_2^{(T_2)}(\mathtt{up})\right) \leqslant \frac{\epsilon^{1.02}}{k^L}. \tag{844}$$

This combined with (819) suggests

$$\frac{a_{p,3}^{(T_2)}}{k} - \left(1 - \frac{c_1}{\ln^2 k}\right) a_{q,0} \geqslant \ln\left(\frac{k^L}{\epsilon^{1.02}}\right) \tag{845}$$

for some constant $c_1 > 0$. This shows (183) holds. In addition, by (182) and our choice of $T_{2,1}$ (c.f. (180)) we know that when $s = T_2$, for any $t \in [T_{2,1}, T_2]$:

$$a_{p,3}^{(t)} \leqslant \frac{a_{p,1}^{(t)}}{2} + k\left(1 + \frac{C_1}{\ln^2 k}\right) a_{q,0} \tag{846}$$

for constant C_1 defined in (840). By (183) and (846) we have (184) holds.

We next show (185) and (186). For all $i, j \in [k]$, define events

$$\mathcal{C}_{2,i}^{j} := \{\text{the agent traverses all children of } n_{h_m} \text{ for all } m \in [i-1] \text{ (if } i > 1), \\ \text{then visits } j \text{ children of } n_{h_i} \text{ and reach the goal after taking } a_{h_i+2(j-1)}, \mathcal{T} \sim P_2\}, \quad (847)$$

where h_i is defined in (820). And let $[\mathcal{C}_{2,i}^{j}]_{\leqslant h}$ be the "mask of $\mathcal{C}_{2,i}^{j}$ up to step h", i.e.,

$$[\mathcal{C}_{2,i}^{j}]_{\leqslant h} := \left\{\omega \in \Omega : \exists \omega' \in \mathcal{C}_{2,i}^{j}, \text{ such that } s_h(\omega') \leqslant s_h(\omega)\right\}, \quad (848)$$

where Ω is the sample space of one full episode. And let

$$\widehat{\pi}_h := \pi_h | [\mathcal{C}_{2,i}^{j}]_{\leqslant h}. \quad (849)$$

Then we have

$$\begin{aligned}
& R(\pi^{(t)}; \mathcal{P}_2) \\
\geqslant & \sum_{i=1}^{k}\sum_{j=1}^{k} \mathbb{P}^{\pi^{(t)}}\left(\mathcal{C}_{2,i}^{j}\right) \\
= & \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}(1-\widehat{\pi}_1^{(t)}(\mathtt{up})) \\
& \left[\prod_{i'=1}^{i-1}\left(\prod_{j'=1}^{k}\left(1-\widehat{\pi}_{h_{i'}+2(j'-1)}^{(t)}(\mathtt{up})\right.\right.\right. \\
& \left.\left.\left.-\sum_{m\in\mathcal{A}_{h_{i'}+2(j'-1),\times}}\widehat{\pi}_{h_{i'}+2(j'-1)}^{(t)}(m)\right)\widehat{\pi}_{h_i+2j'-1}^{(t)}(\mathtt{up})\right)\right. \\
& \left.\cdot\,\widehat{\pi}_{h_{i'}+2k}^{(t)}(\mathtt{up})\left(1-\widehat{\pi}_{h_{i'}+2k+1}^{(t)}(\mathtt{up})-\sum_{m\in\mathcal{A}_{h_{i'}+2k+1,\times}}\widehat{\pi}_{h_{i'}+2k+1}^{(t)}(m)\right)\right] \\
& \cdot\prod_{j'=1}^{j}\left(1-\widehat{\pi}_{h_i+2(j'-1)}^{(t)}(\mathtt{up})-\sum_{m\in\mathcal{A}_{h_i+2(j'-1),\times}}\widehat{\pi}_{h_i+2(j'-1)}^{(t)}(m)\right)\prod_{j'=1}^{j-1} \\
& \pi_{h_i+2j'-1}^{(t)}(\mathtt{up}) \\
\geqslant & \frac{1-\epsilon^{1.05+o(1)}}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k} \\
& \left[\prod_{i'=1}^{i-1}\left(\prod_{j'=1}^{k-1}\left(1-\epsilon^{1.05+o(1)}-\widehat{\pi}_{h_{i'}+2j'}^{(t)}(\mathtt{up})\right)\right)\cdot\widehat{\pi}_{h_{i'}+2k}^{(t)}(\mathtt{up})\left(1\right.\right. \\
& \left.\left.-\epsilon^{1.05+o(1)}-\widehat{\pi}_{h_{i'}+2k+1}^{(t)}(\mathtt{up})\right)\right] \\
& \cdot\prod_{j'=1}^{j-1}\left(1-\epsilon^{1.05+o(1)}-\widehat{\pi}_{h_i+2j'}^{(t)}(\mathtt{up})\right). \quad (850)
\end{aligned}$$

where the last line follows from (878), (879), (885) in the proof of Lemma 8 and our assumption that ϵ is small enough. And from (882), (883) we know that there exists some constant $c > 0$ such that for all $i' \in [i-1]$, $j' \in [k-1]$:

$$\widehat{\pi}_{h_{i'}+2j'}^{(t)}(\mathtt{up}) \leqslant \frac{\exp\left(-\left(1-\frac{c}{\ln^2 k}\right)a_{q,0}\right)}{(k-j')\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1}\right)+\exp\left(-\left(1-\frac{c}{\ln^2 k}\right)a_{q,0}\right)}, \quad (851)$$

$$\widehat{\pi}_{h_{i'}+2k+1}^{(t)}(\mathtt{up}) \leqslant \frac{\exp\left(-\left(1-\frac{c}{\ln^2 k}\right)a_{q,0}\right)}{(k-i')\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1}\right)+\exp\left(-\left(1-\frac{c}{\ln^2 k}\right)a_{q,0}\right)}, \quad (852)$$

for all $j' \in [j-1]$ (if $j > 1$):

$$\widehat{\pi}^{(t)}_{h_i+2j'}(\mathtt{up}) \leqslant \frac{\exp\left(-\left(1-\frac{c}{\ln^2 k}\right)a_{q,0}\right)}{(k-j')\exp\left(\frac{a^{(t)}_{p,1}-a^{(t)}_{p,3}}{k-1}\right)+\exp\left(-\left(1-\frac{c}{\ln^2 k}\right)a_{q,0}\right)}, \tag{853}$$

for all $i' \in [i-1]$:

$$1-\widehat{\pi}^{(t)}_{h_{i'}+2k}(\mathtt{up}) \leqslant \frac{k}{k+\exp\left(-\left(1-\frac{c}{\ln^2 k}\right)a_{q,0}+\frac{a^{(t)}_{p,3}}{k}\right)}. \tag{854}$$

From (850), (851), (852), (853), (854) and our choice of δ_2 we know that for all $t \in [T_{2,1}, T_2-1]$, the following must hold:

$$\max\left\{\frac{\exp\left(-\left(1-\frac{c}{\ln^2 k}\right)a_{q,0}\right)}{\exp\left(\frac{a^{(t)}_{p,1}-a^{(t)}_{p,3}}{k-1}\right)+\exp\left(-\left(1-\frac{c}{\ln^2 k}\right)a_{q,0}\right)}, \frac{k}{k+\exp\left(-\left(1-\frac{c}{\ln^2 k}\right)a_{q,0}+\frac{a^{(t)}_{p,3}}{k}\right)}\right\}$$
$$\geqslant \frac{\epsilon^{1.02+o(1)}}{k^L} \tag{855}$$

or otherwise $R(\pi^{(t)};\mathcal{P}_2) > 1-\delta_2$ for some $t \in [T_{2,1}, T_2-1]$, which contradicts our training setting. From (182) and (855) we deduce that there eixsts constants $c_3, c_4 > 0$ such that

$$\frac{a^{(T_2)}_{p,3}}{k} - \left(1-\frac{c_3}{\ln^2 k}\right)a_{q,0} \leqslant \ln\left(\frac{k^L}{\epsilon^{1.02+o(1)}}\right), \quad \frac{a^{(T_2)}_{p,1}}{k} \leqslant 2\ln\left(\frac{k^L}{\epsilon^{1.02+o(1)}}\right)$$
$$+\frac{c_4}{\ln^2 k}a_{q,0}. \tag{856}$$

This gives (185) and (186). From (186), (177) and Lemma 3 we immediately deduce that (179) holds at step s for any $s \in [T_1+1, T_2]$.

Moreover, from (855), (846) and (819) we also know that for all $t \in [T_1, T_2-1]$:

$$1-\overline{\pi}^{(t)}_2(\mathtt{up}) \geqslant \frac{\epsilon^{1.02+o(1)}}{k^L}. \tag{857}$$

And from Lemma 3 and Induction Hypothesis 3 we also know that

$$\overline{\pi}^{(t)}_2(\mathtt{up}) \geqslant \epsilon^{1.05+o(1)}. \tag{858}$$

By (858), (857), (823) and the fact that (c.f. (921))

$$Q^{(t)}_{2,i}(\mathtt{up}) \geqslant \frac{1+o(1)}{k}, \ \forall i \in [k-1]$$

guaranteed by our policy computation in the proof of Lemma 8, we have

$$\forall t \in [T_1, T_2-1]: \quad \delta a^{(t)}_{p,1} \geqslant \min\left\{\epsilon^{1.05+o(1)}, \frac{\epsilon^{1.02+o(1)}}{k^L}\right\} \geqslant \frac{\epsilon^{1.1}}{k^L}. \tag{859}$$

(859) implies (177) holds. (859) combined with (186) and Lemma 3 also yields (187).

Finally, we show (188) holds for any full k-ary tree $\mathcal{T}$ with depth $l \in [L]$ (that may not be a perfect k-ary tree). First note that by a similar policy computation as in the proof of Lemma 8 we know that given any k-ary tree $\mathcal{T}$ with depth $l \in [L]$, (878), (879), (885) still hold, i.e.,

$$\text{When } c_h = 0 \text{ and } |\mathcal{A}_{h,\times}| = 0: \quad \pi^{(T_2)}_h(\mathtt{up}) \leqslant \epsilon^{1.05+o(1)}, \tag{860}$$
$$\text{when } c_h = \times: \quad 1-\pi^{(T_2)}_h(\mathtt{up}) \leqslant \epsilon^{1.05+o(1)}, \tag{861}$$
$$\text{when } |\mathcal{A}_{h,\times}| \in [k-1]: \quad \pi^{(T_2)}_h(j) \leqslant \epsilon^{1.05+o(1)}, \ \forall j \in \mathcal{A}_{h,\times}. \tag{862}$$

And similar as how we compute (882), (883) in the proof of Lemma 8, by (183), (184) and our choice of δ_2 we can compute that, for any k-ary tree $\mathcal{T}$ with depth $l \in [L]$, we have

$$\text{When } |\mathcal{A}_{h,\times}| \in [k-1]: \quad \pi_h^{(T_2)}(\texttt{up}) \leqslant \frac{\epsilon^{1.02+o(1)}}{k^L},$$
$$\text{When } |\mathcal{A}_{h,\times}| = k: \quad 1 - \pi_h^{(T_2)}(\texttt{up}) \leqslant \frac{\epsilon^{1.02+o(1)}}{k^L}. \tag{863}$$

Similar as how we define $\mathcal{C}_{2,i}^j$ in (847) and use it to lower bound $R(\pi^{(t)}; \mathcal{P}_2)$ in (850), here we define the event

$$\mathcal{E}(\mathcal{T}; g) := \{\text{the agent follows the DFS rule on } \mathcal{T} \text{ until it first reaches } g\}, \tag{864}$$

and let $\mathcal{V}_{\text{leaf}}(\mathcal{T})$ denote the set of leaves of $\mathcal{T}$. For any fixed goal node $g_\star \in \mathcal{V}_{\text{leaf}}(\mathcal{T})$, on the event $\mathcal{E}(\mathcal{T}; g_\star)$, the agent reaches the target goal node $g_\star$. Therefore,

$$R(\pi^{(T_2)}; (\mathcal{T}, g_\star)) \geqslant \mathbb{P}^{\pi^{(T_2)}}(\mathcal{E}(\mathcal{T}; g_\star)). \tag{865}$$

For each step h, let $\mathcal{F}_h(g_\star)$ denote the event that the trajectory has followed the DFS rule toward $g_\star$ up to step h. Equivalently, $\mathcal{F}_h(g_\star)$ is the prefix event induced by $\mathcal{E}(\mathcal{T}; g_\star)$ up to step h. Let

$$\widehat{\pi}_h^{(T_2)} := \pi_h^{(T_2)} \mid \mathcal{F}_h(g_\star). \tag{866}$$

Conditioned on $\mathcal{F}_h(g_\star)$, the one-step probability of deviating from the DFS rule is

$$\eta_h(\mathcal{T}; g_\star) := \begin{cases} \widehat{\pi}_h^{(T_2)}(\texttt{up}) + \sum_{i \in \mathcal{A}_{h,\times}} \widehat{\pi}_h^{(T_2)}(i), & \text{if } c_h = 0,\ |\mathcal{A}_{h,\times}| \leqslant k-1, \\ 1 - \widehat{\pi}_h^{(T_2)}(\texttt{up}), & \text{if } c_h = \times \text{ or } (c_h = 0 \text{ and } |\mathcal{A}_{h,\times}| = k). \end{cases} \tag{867}$$

By (860) and (863), for every h,

$$\mathbb{E}[\eta_h(\mathcal{T}; g_\star)] \leqslant \frac{\epsilon^{1.02+o(1)}}{k^L}. \tag{868}$$

Since a deviation from $\mathcal{E}(\mathcal{T}; g_\star)$ must occur at some step $h \leqslant H$, we have

$$\begin{aligned} \mathbb{P}^{\pi^{(T_2)}}(\mathcal{E}(\mathcal{T}; g_\star)^c) &\leqslant \sum_{h=1}^{H} \mathbb{E}\left[\mathbb{1}\{\mathcal{F}_h(g_\star)\}\eta_h(\mathcal{T}; g_\star)\right] \\ &\leqslant \sum_{h=1}^{H} \mathbb{E}[\eta_h(\mathcal{T}; g_\star)] \\ &\leqslant H \cdot \frac{\epsilon^{1.02+o(1)}}{k^L} \leqslant \epsilon^{1.02+o(1)} \leqslant \epsilon, \end{aligned} \tag{869}$$

where we used $H \lesssim k^L$. Therefore,

$$R(\pi^{(T_2)}; (\mathcal{T}, g_\star)) \overset{(865)}{\geqslant} \mathbb{P}^{\pi^{(T_2)}}(\mathcal{E}(\mathcal{T}; g_\star)) \geqslant 1 - \epsilon. \tag{870}$$

Since $g_\star \in \mathcal{V}_{\text{leaf}}(\mathcal{T})$ was arbitrary, (188) follows.

C.3.8 Proof of Lemma 8

Policy computation. To compute the gradients, we first compute the policies on trees with depth 2. Note that under our setting, on any perfect k-ary tree of depth l, $c_h = 0$ if $h \leqslant l$, and if $c_h = \times$ at some step h, then $c_{h-1} = 0$, and $c_{h+1} = 0$ if the environment doesn't terminate at step h. Therefore, at any step $h \in [H]$, we have

$$N_{h,\times} \leqslant \begin{cases} \max\{0, \frac{h-l}{2}\}, & \text{if } c_h = 0, \\ \max\{0, \frac{h-l+1}{2}\}, & \text{if } c_h = \times. \end{cases} \tag{871}$$

In addition, the level the agent is at a tree of depth l (root is level 0, and leaves are level l) at step h equals $\sum_{i=1}^k N_h(i) - N_h(\texttt{up})$. As a result, we have

$$0 \leqslant \sum_{i=1}^k N_h(i) - N_h(\texttt{up}) \leqslant l, \quad \sum_{i=1}^k N_h(i) + N_h(\texttt{up}) = h - 1. \tag{872}$$

The above two relations yield

$$\frac{h-1}{2} \leqslant \sum_{i=1}^{k} N_h(i) \leqslant \frac{h-1+l}{2}, \quad \max\left\{0, \frac{h-1-l}{2}\right\} \leqslant N_h(\mathtt{up}) \leqslant \frac{h-1}{2}. \tag{873}$$

By (871), (90) and Lemma 3, we have

$$\xi_h = \begin{cases} -\left(1+O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}, & \text{if } c_h = 0, \\ a_{q,\times} + O\left(\frac{k^2}{\ln(1/\epsilon)}\right), & \text{if } c_h = \times. \end{cases} \tag{874}$$

Let $\widetilde{\phi}_h := (\widetilde{\phi}_h(1), \ldots, \widetilde{\phi}_h(k), \widetilde{\phi}_h(\mathtt{up}))^\top$ with

$$\widetilde{\phi}_h(i) := \frac{\varphi_h(i)}{e^{a_{b,0}} + |\mathcal{A}_{h,\times}| e^{a_{b,1}} + (h-1-|\mathcal{A}_{h,\times}|)e^{a_{b,2}}}, \ \forall i \in [k], \quad \widetilde{\phi}_h(\mathtt{up}) := (1 + \frac{1}{k}) \xi_h. \tag{875}$$

Then by Lemma 1 we have

$$\pi_h = \mathsf{softmax}\left(\widetilde{\phi}_h\right). \tag{876}$$

In addition, when $|\mathcal{A}_{h,\times}| = 0$, for all $i \in [k]$, by Lemma 1, Lemma 3, Induction Hypothesis 3 and (873) we have

$$\begin{aligned} \widetilde{\phi}_h(i) &= -\frac{a_{p,0}^{(t)}}{k} + \frac{1}{e^{a_{b,0}-a_{b,2}}} \left(N_h(\mathtt{up}) \frac{a_{p,\mathtt{up}}^{(t)}}{k} + \frac{a_{p,0}^{(t)}}{k} + (h-1-N_h(\mathtt{up})) a_{p,2}^{(t)} \right. \\ &\quad \left. - N_h(i) \left(a_{p,1}^{(t)} + a_{p,2}^{(t)} \right) \right) \\ &\quad + O\left(\frac{1}{(\ln(1/\epsilon))^{1+o(1)}} \right) \\ &= -\frac{a_{p,0}^{(t)}}{k} + O\left(\frac{k}{(\ln(1/\epsilon))^{1+O(1/N)}} \right) a_{p,1}^{(t)}. \end{aligned} \tag{877}$$

By (876), (877), (874), (167) we have

$$\begin{aligned} &\text{When } c_h = 0 \text{ and } |\mathcal{A}_{h,\times}| = 0: \quad \pi_h(\mathtt{up}) = \frac{1}{k} \exp\left(-\left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0} \right. \\ &\quad \left. + \frac{a_{p,0}^{(t)}}{k} \right) \\ &\leqslant \epsilon^{1.05 + O\left(1/\ln^2 k\right)}. \end{aligned} \tag{878}$$

Note $|\mathcal{A}_{h,\times}| = 0$, as in (67) we define $\mathcal{A}_{h,\times} = \varnothing$ if $c_h = \times$. Then by (876), (877), (874) and (168) we have

$$\begin{aligned} &\text{When } c_h = \times: \ 1 - \pi_h(\mathtt{up}) = k \exp\left(-\left(1 + \frac{1}{k} + O\right.\right. \\ &\quad \left.\left. \left(\frac{1}{(\ln(1/\epsilon))^{1+o(1)}} \right) \right) a_{q,\times} - \frac{a_{p,0}^{(t)}}{k} \right) \\ &\leqslant \epsilon^{1.05 + O(1/k)}. \end{aligned} \tag{879}$$

Similarly, when $|\mathcal{A}_{h,\times}| \geqslant 1$, by Lemma 1, Lemma 3 and Induction Hypothesis 3, we have

$$\widetilde{\phi}_h(i) = a_{p,2}^{(t)} - \frac{1}{|\mathcal{A}_{h,\times}|} \mathbb{1}\{i \in \mathcal{A}_{h,\times}\} (a_{p,1}^{(t)} + a_{p,2}^{(t)})$$

$$+\frac{1}{e^{a_{b,1}^{(t)}-a_{b,0}^{(t)}}}\left(-\frac{a_{p,0}^{(t)}}{k|\mathcal{A}_{h,\times}|}-a_{p,2}^{(t)}+\frac{1}{|\mathcal{A}_{h,\times}|}\mathbb{1}\{i\in\mathcal{A}_{h,\times}\}(a_{p,1}^{(t)}+a_{p,2}^{(t)})\right)+O\left(\frac{1}{(\ln(1/\epsilon))^{1+o(1)}}\right)$$

$$=a_{p,2}^{(t)}-\frac{1}{|\mathcal{A}_{h,\times}|}\mathbb{1}\{i\in\mathcal{A}_{h,\times}\}(a_{p,1}^{(t)}+a_{p,2}^{(t)})+O\left(\frac{1}{(\ln(1/\epsilon))^{1+O(1/N)}}\right)a_{p,1}^{(t)}. \tag{880}$$

Especially, when $|\mathcal{A}_{h,\times}|=k$, by (880) and (80) we have

$$\forall i\in[k]:\quad \widetilde{\phi}_h(i)=-\frac{a_{p,3}^{(t)}}{k}+\frac{1}{e^{a_{b,1}^{(t)}-a_{b,0}^{(t)}}}\left(-\frac{a_{p,0}^{(t)}}{k^2}+\frac{a_{p,3}^{(t)}}{k}\right)+O\left(\frac{1}{(\ln(1/\epsilon))^{1+o(1)}}\right)$$

$$=-\frac{a_{p,3}^{(t)}}{k}+O\left(\frac{1}{k\,(\ln(1/\epsilon))^{1+O(1/N)}}\right)a_{p,1}^{(t)}. \tag{881}$$

Then by (876), (874), (80) and the above two expressions we have when $|\mathcal{A}_{h,\times}|\in[k-1]$:

$$\pi_h(\mathtt{up})=\frac{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}{(k-|\mathcal{A}_{h,\times}|)\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1}\right)+\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}$$

$$\overset{(182)}{\leqslant}\frac{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}{(k-|\mathcal{A}_{h,\times}|)\exp\left(\frac{a_{p,1}^{(t)}}{2(k-1)}-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)+\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}$$

$$\overset{(166)}{\leqslant}\epsilon^{\frac{1.05}{2}+o(1)}, \tag{882}$$

and

$$\text{When } |\mathcal{A}_{h,\times}|=k:\quad \pi_h(\mathtt{up})=\frac{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}+\frac{a_{p,3}^{(t)}}{k}\right)}{k+\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}+\frac{a_{p,3}^{(t)}}{k}\right)}. \tag{883}$$

And when $|\mathcal{A}_{h,\times}|\in[k-1]$:

$$\forall j\in\mathcal{A}_{h,\times}:\quad \pi_h(j)=\frac{1}{k-|\mathcal{A}_{h,\times}|}\exp\left(-\left(1+O\left(\frac{1}{(\ln(1/\epsilon))^{1+O(1/N)}}\right)\right)\frac{a_{p,1}^{(t)}+a_{p,2}^{(t)}}{|\mathcal{A}_{h,\times}|}\right), \tag{884}$$

which, combined with (166) and Induction Hypothesis 3, further indicates for all $j\in\mathcal{A}_{h,\times}$:

$$\text{When } |\mathcal{A}_{h,\times}|=k-1:\quad \pi_h(j)=\exp\left(-\left(1+O\left(\frac{1}{(\ln(1/\epsilon))^{1+O(1/N)}}\right)\right)\frac{a_{p,1}^{(t)}+a_{p,2}^{(t)}}{k-1}\right)$$

$$\leqslant\epsilon^{1.05+O(1/k)},$$

$$\text{when } |\mathcal{A}_{h,\times}|\in[k-2]:\quad \pi_h(j)\leqslant\epsilon^{\frac{C}{k}}\exp\left(-\left(1+O\left(\frac{1}{(\ln(1/\epsilon))^{1+O(1/N)}}\right)\right)\frac{a_{p,1}^{(t)}+a_{p,2}^{(t)}}{k-1}\right) \tag{885}$$

for some constant $C>0$. When $|\mathcal{A}_{h,\times}|=k$, from (880) we can see that

$$\forall j\in[k]:\quad \pi_h(j)=(1+o(1/k^2))\frac{1-\pi_h(\mathtt{up})}{k}$$

$$\overset{(883)}{=} \frac{1}{k + \exp\left(-\left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0} + \frac{a_{p,3}^{(t)}}{k}\right)}. \tag{886}$$

Gradient computation. For all $i \in [k]$, we define events

$$\mathcal{E}_{2,i} := \{\text{the environment has not terminated at step } h_i, l = 2\}, \tag{887}$$

where h_i is defined in (820). We also define $h_{k+1} = +\infty$. Let

$$\begin{aligned} \forall i \in [k-1]: \quad \pi_{2,i}(\mathtt{up}) &:= \pi_{h_i+2k}(\mathtt{up})|\overline{\mathcal{E}}_{2,i}, \quad Q_{2,i}^{(t)}(\mathtt{up}) := Q_{h_i+2k}^{(t)}(\mathtt{up})|\overline{\mathcal{E}}_{2,i}, \\ \pi_{2,k}(\mathtt{up}) &:= 0, \quad Q_{2,k}^{(t)}(\mathtt{up}) := 0. \end{aligned} \tag{888}$$

Note that $\pi_{2,i}(\mathtt{up})$ and $Q_{2,i}^{(t)}(\mathtt{up})$ are well-defined because they are invariant to the order of children of n_{h_j} ($j \in [i]$) and the root visited under the balanced distribution. We also define the following events for all $i', i \in [k-1]$ and $j \in \{0, 1, \ldots, k-1\}$:

$$\begin{aligned} \overline{\mathcal{E}}_{2,i}^{i,j} &:= \{\text{the agent traverses all children of } n_{h_m} \text{ for all } m \in [i-1] \text{ (if } i > 1), \\ &\qquad \text{then visits } j \text{ children of } n_{h_i} \text{ and arrives at } n_{h_i} \text{ at step } h_i + 2j, l = 2\}, \\ \forall i' < i: \quad \overline{\mathcal{E}}_{2,i}^{i',j} &:= \{\text{the agent traverses all children of } n_{h_m} \text{ for all } m \in [i] \backslash \{i'\}, \\ &\qquad \text{visits } j \text{ children of } n_{h_{i'}}, \text{ and arrives at } n_{h_i} \text{ at step } h_i + 2k, l = 2\}, \end{aligned} \tag{889}$$

and $\forall i, j \in [k]$:

$$\begin{aligned} \overline{\mathcal{E}}_{2,i}^{\times,j} := \{&\text{the agent traverses all children of } n_{h_m} \text{ for all } m \in [i-1] \text{ (if } i > 1), \\ &\text{then visits } j \text{ children of } n_{h_i} \text{ and arrives at the } j\text{-th child of } n_{h_i} \\ &\text{visited at step } h_i + 2j - 1, l = 2\}. \end{aligned} \tag{890}$$

For any sequence of random variables $\{x_h\}_{h \in [H]}$ and policy π, we have

$$\mathbb{E}_\pi\left[\sum_{h=1}^{H} x_h \middle| l = 2\right] = \mathbb{E}_\pi\left[x_1 | l = 2\right] + \sum_{i=1}^{k} \mathbb{P}^\pi\left(\mathcal{E}_{2,i}\right) \mathbb{E}_\pi\left[\sum_{h=h_i}^{h_{i+1}-1} x_h \middle| l = 2, \mathcal{E}_{2,i}\right]. \tag{891}$$

And for any $\pi = \pi^{(t)}$ ($t \in [T_1, s-1]$), we have by our policy computation (c.f. (878), (879), (882), (883), (885)) we have

$$\begin{aligned} \mathbb{P}^\pi\left(\mathcal{E}_{2,1}\right) &= 1 - \pi_1(\mathtt{up}) \overset{(878)}{\geqslant} 1 - \epsilon^{1.05 + O\left(1/\ln^2 k\right)}, \\ \forall i \in \{2, \cdots, k\}: \quad \mathbb{P}^\pi\left(\mathcal{E}_{2,i}\right) &= (1 + o(\epsilon^{0.5}))\frac{k-i+1}{k} \prod_{j=1}^{i-1} \pi_{2,j}(\mathtt{up}). \end{aligned} \tag{892}$$

With the above preparations, we are ready to simplify the gradients in Lemma 1. All expectations here are conditioned on tree depth $l = 2$ and policy $\pi^{(t)}$ unless otherwise specified, i.e., $\mathbb{E}[\cdot] = \mathbb{E}_{\pi^{(t)}}[\cdot | l = 2]$. We also use $\mathbb{P}^{(t)}(\cdot)$ as a shorthand for $\mathbb{P}^{\pi^{(t)}}(\cdot)$.

We first compute $\delta a_{p,3}^{(t)}$. From (170), (171), (172) and (83) we know that

$$\begin{aligned} \delta a_{p,3}^{(t)} = &\frac{1 + o(1/k^2)}{k} \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0, |\mathcal{A}_{h,\times}| \geqslant 1\} \pi_h^{(t)}(\mathtt{up}) \left(Q_h^{(t)}(\mathtt{up}) - G_h^{(t)}\right)\right] \\ &+ (1 + o(1/k^2)) \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0, |\mathcal{A}_{h,\times}| = 0\} \pi_h^{(t)}(\mathtt{up}) \frac{N_h(j)}{e^{a_{b,0} - a_{b,2}}} \left(Q_h^{(t)}(\mathtt{up}) \right.\right. \\ &\left.\left. - G_h^{(t)}\right) \middle| j \in [k]\right] \\ &+ (1 + o(1/k^2)) \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = \times\} \pi_h^{(t)}(\mathtt{up}) \left(1\right.\right. \end{aligned}$$

$$-\pi_h^{(t)}(\mathrm{up})\Big)\frac{N_h(j)}{e^{a_{b,0}-a_{b,2}}}Q_h^{(t)}(\mathrm{up})\Big|j\in[k]\Big]. \tag{893}$$

Applying (891) to the first term of (893), and using Lemma 3 and Induction Hypothesis 3, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|\geqslant 1\}\pi_h^{(t)}(\mathrm{up})\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\right]\\&=(1+o(1/k^2))\sum_{i=1}^{k}\mathbb{P}\left(\mathcal{E}_{2,i}\right)\sum_{j=1}^{k}\frac{(k-i+1)k-j}{(k-i+1)k}\mathbb{E}\Big[\pi_{h_i+2j}^{(t)}(\mathrm{up})\Big(Q_{h_i+2j}^{(t)}(\mathrm{up})\\&\quad-G_{h_i+2j}^{(t)}\Big)\Big|\overline{\mathcal{E}}_{2,i}^{i,j}\Big].\end{aligned} \tag{894}$$

By our policy computation (c.f. (878), (879), (882), (883), (885)) and the definition of $\pi_{2,i}(\mathrm{up})$ and $Q_{2,i}(\mathrm{up})$ (c.f. (888)), we have for all $i\in[k]$ (note $\overline{\mathcal{E}}_{2,i}=\overline{\mathcal{E}}_{2,i}^{i,k}$):

$$\mathbb{E}\left[\pi_{h_i+2k}^{(t)}(\mathrm{up})\left(Q_{h_i+2k}^{(t)}(\mathrm{up})-G_{h_i+2k}^{(t)}\right)\Big|\overline{\mathcal{E}}_{2,i}\right]=\pi_{2,i}^{(t)}(\mathrm{up})\left(1-\pi_{2,i}^{(t)}(\mathrm{up})\right)Q_{2,i}^{(t)}(\mathrm{up}). \tag{895}$$

By our policy computation we can also obtain (recall $\prod_{i=a}^{b}x_i=1$ if $a>b$):

$$\forall i\in[k-1]:\quad Q_{2,i}^{(t)}(\mathrm{up})=\frac{1+o(\epsilon^{0.5})}{k-i}\sum_{j=1}^{k-i}\left(\prod_{m=1}^{j-1}\pi_{2,i+m}^{(t)}(\mathrm{up})\right), \tag{896}$$

for all $i\in[k],j\in\{0,1,\ldots,k-1\}$:

$$\begin{aligned}\mathbb{E}\left[G_{h_i+2j}^{(t)}\Big|\overline{\mathcal{E}}_{2,i}^{i,j}\right]&=(1+o(\epsilon^{0.5}))\left(\sum_{m=1}^{k-j}\frac{1}{(k-i+1)k-j}+\frac{(k-i)k}{(k-i+1)k-j}\right.\\&\quad\left.\pi_{2,i}^{(t)}(\mathrm{up})Q_{2,i}^{(t)}(\mathrm{up})\right),\end{aligned} \tag{897}$$

and

$$\forall j\in[k-1]:\quad\mathbb{E}\left[Q_{h_k+2j}^{(t)}(\mathrm{up})\Big|\overline{\mathcal{E}}_{2,k}^{k,j}\right]=0, \tag{898}$$

for all $i\in[k],j\in\{0,1,\ldots,k-1\}$:

$$\mathbb{E}\left[Q_{h_i+2j}(\mathrm{up})\Big|\overline{\mathcal{E}}_{2,i}^{i,j}\right]=\frac{(1+o(\epsilon^{0.5}))k}{(k-i+1)k-j}\sum_{j'=1}^{k-i}\left(\prod_{m=1}^{j'-1}\pi_{h_{i+m}+2k}^{(t)}(\mathrm{up})|\overline{\mathcal{E}}_{2,i+m}^{i,j}\right). \tag{899}$$

Note that

$$\forall i\in[k-1]:\quad N_{h_i+2k,\times}|\overline{\mathcal{E}}_{2,i}=ik,\quad h_i+2k|\overline{\mathcal{E}}_{2,i}=i(2k+2); \tag{900}$$

and for all $i'\in[k-1]$, $2\leqslant i\leqslant k-1$, $j\in\{0,1,\ldots,k-1\}$, $i'<i$:

$$N_{h_i+2k,\times}|\overline{\mathcal{E}}_{2,i}^{i',j}=(i-1)k+j,\quad h_i+2k|\overline{\mathcal{E}}_{2,i}^{i',j}=i(2k+2)-2(k-j). \tag{901}$$

By (900), (901) and (90) we have for all $i\in[k-1]$:

$$\xi_{h_i+2k}|\overline{\mathcal{E}}_{2,i}=\frac{ke^{-2a_{c,0}}a_{q,\times}-(k+2)a_{q,0}}{ke^{-2a_{c,0}}+k+2}, \tag{902}$$

and for all $2\leqslant i\leqslant k-1$, $i'<i$ and $j\in\{0,1,\ldots,k-1\}$:

$$\begin{aligned}\xi_{h_i+2k}|\overline{\mathcal{E}}_{2,i}^{i',j}&=\frac{\left(k-\frac{k-j}{i}\right)e^{-2a_{c,0}}a_{q,\times}-\left(k+2-\frac{k-j}{i}\right)a_{q,0}}{\left(k-\frac{k-j}{i}\right)e^{-2a_{c,0}}+k+2-\frac{k-j}{i}}\\&=\xi_{h_i+2k}|\overline{\mathcal{E}}_{2,i}-\frac{2\frac{k-j}{i}e^{-2a_{c,0}}}{(ke^{-2a_{c,0}}+k+2)\left(\left(k-\frac{k-j}{i}\right)e^{-2a_{c,0}}+k+2-\frac{k-j}{i}\right)}(a_{q,0}\end{aligned}$$

$$+a_{q,\times}\big)\,. \tag{903}$$

From (900), (881), (876) we know that

$$\forall i\in[k-1]:\quad \pi_{2,i}^{(t)}(\mathrm{up})=\left(1+O\left(\frac{1}{(\ln(1/\epsilon))^{1+o(1)}}\right)\right)\overline{\pi}_2^{(t)}(\mathrm{up}), \tag{904}$$

where $\overline{\pi}_2^{(t)}(\mathrm{up})$ is defined in (819). Then from (904), (896) we deduce

$$\forall i\in[k-1]:\quad Q_{2,i}^{(t)}(\mathrm{up})=\frac{1+o(1/k)}{k-i}\frac{1-\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^{k-i}}{1-\overline{\pi}_2^{(t)}(\mathrm{up})}, \tag{905}$$

and from (904), (892) we deduce

$$\forall i\in[k]:\quad \mathbb{P}\left(\mathcal{E}_{2,i}\right)=(1+o(1/k^2))\frac{k-i+1}{k}\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^{i-1}. \tag{906}$$

Further, by (903), (881), (876), Lemma 3 and Induction Hypothesis 3 we have for all $2\leqslant i\leqslant k-1$, $i'<i$ and $j\in\{0,1,\ldots,k-1\}$:

$$\pi_{h_i+2k}^{(t)}(\mathrm{up})|\overline{\mathcal{E}}_{2,i}^{i',j}<\overline{\pi}_2^{(t)}(\mathrm{up}), \tag{907}$$

and there exists a (large enough) constant $C>0$ such that if

$$-\left(1-\frac{1}{(\ln(1/\epsilon))^{1-C/N}}\right)\frac{a_{p,3}^{(t)}}{k}+4.1\ln k\leqslant-\left(1+\frac{1}{k}\right)a_{q,0}, \tag{908}$$

then for all $2\leqslant i\leqslant k-1$, $i'<i$ and $j\in\{0,1,\ldots,k-1\}$:

$$\pi_{h_i+2k}^{(t)}(\mathrm{up})|\overline{\mathcal{E}}_{2,i}^{i',j}=(1+o(1/k^3))\pi_{2,i}^{(t)}(\mathrm{up}). \tag{909}$$

And from (907), (909), (896) and (899) we know that for all $i\in[k-1],j\in\{0,1,\ldots,k-1\}$:

$$\mathbb{E}\left[Q_{h_i+2j}^{(t)}(\mathrm{up})\middle|\overline{\mathcal{E}}_{2,i}^{i,j}\right]<\frac{(k-i)k}{(k-i+1)k-j}Q_{2,i}^{(t)}(\mathrm{up}), \tag{910}$$

and if (908) holds, then for all $i\in[k-1]$, $j\in\{0,1,\ldots,k-1\}$:

$$\mathbb{E}\left[Q_{h_i+2j}^{(t)}(\mathrm{up})\middle|\overline{\mathcal{E}}_{2,i}^{i,j}\right]=(1+o(1/k^2))\frac{(k-i)k}{(k-i+1)k-j}Q_{2,i}^{(t)}(\mathrm{up}). \tag{911}$$

Therefore, by (910) and (897), (898) we have for all $i\in[k],j\in\{0,1,\ldots,k-1\}$ (recall we define $Q_{2,k}^{(t)}(\mathrm{up})=0$ in (888)):

$$\begin{aligned}
&\mathbb{E}\left[Q_{h_i+2j}^{(t)}(\mathrm{up})-G_{h_i+2j}^{(t)}\middle|\overline{\mathcal{E}}_{2,i}^{i,j}\right]\\
&\leqslant-(1+o(1/k))\frac{k-j}{(k-i+1)k-j}+(1+o(1/k))\frac{(k-i)k}{(k-i+1)k-j}\,(1\\
&\quad-\pi_{2,i}^{(t)}(\mathrm{up})\Big)\,Q_{2,i}^{(t)}(\mathrm{up})\\
&=-(1+o(1/k))\frac{k-j}{(k-i+1)k-j}+(1+o(1/k))\frac{k}{(k-i+1)k-j}\,(1-(\\
&\quad\pi_{2,i}^{(t)}(\mathrm{up})\Big)^{k-i}\Big),
\end{aligned} \tag{912}$$

where the last line follows from (905) and (904). We also have for all $i\in[k],j\in\{0,1,\ldots,k-1\}$:

$$\mathbb{E}\left[Q_{h_i+2j}^{(t)}(\mathrm{up})-G_{h_i+2j}^{(t)}\middle|\overline{\mathcal{E}}_{2,i}^{i,j}\right]\geqslant-1, \tag{913}$$

and when (908) holds, by (911) and (897) we have for all $i\in[k-1],j\in\{0,1,\ldots,k-1\}$:

$$\mathbb{E}\left[Q_{h_i+2j}^{(t)}(\mathrm{up})-G_{h_i+2j}^{(t)}\middle|\overline{\mathcal{E}}_{2,i}^{i,j}\right]=-(1+o(1))\frac{k-j}{(k-i+1)k-j}. \tag{914}$$

Also note that if (908) holds, then

$$\forall i \in [k-1]: \quad 1-\pi_{2,i}^{(t)}(\mathrm{up}) = o(1/k^3), \quad Q_{2,i}^{(t)}(\mathrm{up}) \overset{(896)}{=} 1+o(1), \tag{915}$$

and if (908) does not hold, then there exists a constant $c>0$ such that

$$\forall i \in [k-1]: \quad 1-\pi_{2,i}^{(t)}(\mathrm{up}) \geqslant \frac{1}{\exp\left(\frac{c}{\ln^2 k} a_{q,0}\right)}. \tag{916}$$

Therefore,

- if (908) does not hold, then by (912), (913), (916), (882), (883) and (895) we know that for all $i \in [k-1]$:

$$\begin{aligned}
&\sum_{j=1}^{k} \frac{(k-i+1)k-j}{(k-i+1)k} \mathbb{E}\left[\pi_{h_i+2j}^{(t)}(\mathrm{up})\left(Q_{h_i+2j}^{(t)}(\mathrm{up}) - G_{h_i+2j}^{(t)}\right) \middle| \overline{\mathcal{E}}_{2,i}^{i,j}\right] \\
&= (1+o(1/k^2)) \frac{k-i}{k-i+1} \pi_{2,i}^{(t)}(\mathrm{up})\left(1-\pi_{2,i}^{(t)}(\mathrm{up})\right) Q_{2,i}^{(t)}(\mathrm{up});
\end{aligned} \tag{917}$$

- if (908) holds, then by (912), (913), (914), (915), (882), (883) and (895) we know that for all $i \in [k-1]$:

$$\begin{aligned}
&\sum_{j=1}^{k} \frac{(k-i+1)k-j}{(k-i+1)k} \mathbb{E}\left[\pi_{h_i+2j}^{(t)}(\mathrm{up})\left(Q_{h_i+2j}^{(t)}(\mathrm{up}) - G_{h_i+2j}^{(t)}\right) \middle| \overline{\mathcal{E}}_{2,i}^{i,j}\right] \\
&= (1+o(1)) \frac{k-i}{k-i+1}\left(1-\pi_{2,i}^{(t)}(\mathrm{up})\right) \\
&\quad - \sum_{j=1}^{k-1} \frac{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}\right)}{(k-j)\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1}\right)} \frac{k-j}{(k-i+1)k} \\
&= \frac{k-i}{k-i+1}\left((1+o(1))\pi_{2,i}^{(t)}(\mathrm{up})\right. \\
&\quad \left. - \frac{k-1}{k(k-i)\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1} + \left(1+O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}\right)}\right).
\end{aligned} \tag{918}$$

By Lemma 3 and Induction Hypothesis 3, we could also uniformly write the two cases above as ($i \in [k-1]$):

$$\begin{aligned}
&\sum_{j=1}^{k} \frac{(k-i+1)k-j}{(k-i+1)k} \mathbb{E}\left[\pi_{h_i+2j}^{(t)}(\mathrm{up})\left(Q_{h_i+2j}^{(t)}(\mathrm{up}) - G_{h_i+2j}^{(t)}\right) \middle| \overline{\mathcal{E}}_{2,i}^{i,j}\right] \\
&= \frac{k-i}{k-i+1}\pi_{2,i}^{(t)}(\mathrm{up})\left((1+o(1))\left(1-\pi_{2,i}^{(t)}(\mathrm{up})\right)\right. \\
&\quad \left. - \frac{k-1}{k(k-i)\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1} + \left(1+O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}\right)}\right) Q_{2,i}^{(t)}(\mathrm{up}).
\end{aligned} \tag{919}$$

When $i=k$, by our gradient computation, (888), (898), (897) we have

$$\begin{aligned}
&\sum_{j=1}^{k} \frac{k-j}{k} \mathbb{E}\left[\pi_{h_k+2j}^{(t)}(\mathrm{up})\left(Q_{h_k+2j}^{(t)}(\mathrm{up}) - G_{h_k+2j}\right) \middle| \overline{\mathcal{E}}_{2,k}^{k,j}\right] \\
&= -\sum_{j=1}^{k-1} \frac{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}\right)}{(k-j)\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1}\right)} \frac{k-j}{k}
\end{aligned}$$

$$= -\frac{k-1}{k\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1}+\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}, \tag{920}$$

and

$$\forall i\in[k-1]: \quad Q_{2,i}^{(t)}(\mathrm{up}) \geqslant \frac{1+o(1)}{k}. \tag{921}$$

Therefore, by (894), (918), (917), (904), (920) and (921) we can rewrite the first term in (893) as follows:

$$\begin{aligned}
&\frac{1}{k}\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|\geqslant 1\}\pi_h^{(t)}(\mathrm{up})\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\right]\\
&=\frac{1}{k}\Bigg[\sum_{i=1}^{k-1}\mathbb{P}\left(\mathcal{E}_{2,i}\right)\frac{k-i}{k-i+1}\pi_{2,i}^{(t)}(\mathrm{up})\Bigg((1+o(1))\left(1-\pi_{2,i}^{(t)}(\mathrm{up})\right)\\
&-\frac{k-1}{k(k-i)\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1}+\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}\Bigg)Q_{2,i}^{(t)}(\mathrm{up})\Bigg]\\
&\quad-\frac{(k-1)\mathbb{P}\left(\mathcal{E}_{2,k}\right)}{k^2\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1}+\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}\\
&\overset{(906)}{=}\frac{1}{k}\sum_{i=1}^{k-1}\frac{k-i}{k}\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^i\Bigg((1+o(1))\left(1-\overline{\pi}_2^{(t)}(\mathrm{up})\right)\\
&\quad-\frac{1}{(k-i)\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1}+\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}\Bigg)Q_{2,i}^{(t)}(\mathrm{up})
\end{aligned} \tag{922}$$

where the last line follows from absorbing the last term in the second line into the other terms.

For the second term in (893), by symetry of the balanced data distribution we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|=0\}\pi_h^{(t)}(\mathrm{up})\frac{N_h(j)}{e^{a_{b,0}-a_{b,2}}}\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\Big|j\in[k]\right]\\
&=\frac{1}{e^{a_{b,0}-a_{b,2}}}\sum_{i=1}^{k}\frac{(i-1)k+i}{k}\mathbb{P}\left(\mathcal{E}_{2,i}\right)\left(\pi_{h_i}^{(t)}(\mathrm{up})|\mathcal{E}_{2,i}\right)\mathbb{E}\left[Q_{h_i}^{(t)}(\mathrm{up})-G_{h_i}^{(t)}\Big|\mathcal{E}_{2,i}\right]\\
&\overset{(906)}{=}\frac{1}{e^{a_{b,0}-a_{b,2}}}\sum_{i=1}^{k}\frac{(i-1)k+i}{k}\frac{k-i+1}{k}\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^{i-1}\left(\pi_{h_i}^{(t)}(\mathrm{up})|\mathcal{E}_{2,i}\right)\\
&\quad\mathbb{E}\left[Q_{h_i}^{(t)}(\mathrm{up})-G_{h_i}^{(t)}\Big|\mathcal{E}_{2,i}\right].
\end{aligned} \tag{923}$$

Similar as how we compute (878), we can compute that

$$\pi_{h_1}^{(t)}(\mathrm{up})|\mathcal{E}_{2,1}=\frac{\exp\left(-\left(1+\frac{1}{k}\right)a_{q,0}+\frac{a_{p,0}^{(t)}}{k}+O\left(\frac{k}{(\ln(1/\epsilon))^{1+O(1/N)}}\right)a_{p,1}^{(t)}\right)}{k+\exp\left(-\left(1+\frac{1}{k}\right)a_{q,0}+\frac{a_{p,0}^{(t)}}{k}+O\left(\frac{k}{(\ln(1/\epsilon))^{1+O(1/N)}}\right)a_{p,1}^{(t)}\right)}, \tag{924}$$

By (819), (135) in Lemma 3 and Induction Hypothesis 3 we have

$$\overline{\pi}_2^{(t)}(\mathrm{up})\geqslant\frac{\exp\left(-\left(1+\frac{1}{k}\right)a_{q,0}+\frac{a_{p,3}^{(t)}}{k}+c_1\frac{a_{q,\times}}{\ln^2 k}\right)}{k+\exp\left(-\left(1+\frac{1}{k}\right)a_{q,0}+\frac{a_{p,3}^{(t)}}{k}+c_1\frac{a_{q,\times}}{\ln^2 k}\right)} \tag{925}$$

for some constant $c_1 > 0$. Note that by (581) and (134) we have

$$|a_{p,3}^{(T_1)}| \lesssim \frac{a_{q,\times}}{k}. \tag{926}$$

Thus by (924), (925), (926) and Induction Hypothesis 3 we have

$$\overline{\pi}_2^{(t)}(\mathrm{up}) \geqslant \exp\left(c_1' \frac{a_{q,\times}}{\ln^2 k}\right) \left(\pi_{h_1}^{(t)}(\mathrm{up}) | \mathcal{E}_{2,1}\right) \tag{927}$$

for some constant $c_1' > 0$. Therefore, from (927), (878), (923) and Lemma 3 we know that

$$\begin{aligned}
&\left| \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0, |\mathcal{A}_{h,\times}| = 0\} \pi_h^{(t)}(\mathrm{up}) \frac{N_h(j)}{e^{a_{b,0}-a_{b,2}}} \left(Q_h^{(t)}(\mathrm{up}) - G_h^{(t)}\right) \middle| j \in [k] \right] \right| \\
&= o\left(\overline{\pi}_2^{(t)}(\mathrm{up}) \left(1 - \overline{\pi}_2^{(t)}(\mathrm{up})\right) Q_{2,1}^{(t)}(\mathrm{up})\right).
\end{aligned} \tag{928}$$

Using (159), (160) in Lemma 3, and by a similar argument as how we obtain (927), we can compute that

$$\overline{\pi}_2^{(t)}(\mathrm{up}) \geqslant \exp\left(c_2 \frac{a_{q,\times}}{\ln^2 k}\right) \left(\left(1 - \pi_{h_i+2j-1}^{(t)}(\mathrm{up})\right) | \overline{\mathcal{E}}_{2,i}^{\times,j}\right) \tag{929}$$

for some constant $c_2 > 0$. And similar as for (928), by (929), (879) we can bound the last term in (893) as follows:

$$\begin{aligned}
&\left| \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = \times\} \pi_h^{(t)}(\mathrm{up}) \left(1 - \pi_h^{(t)}(\mathrm{up})\right) \frac{N_h(j)}{e^{a_{b,0}-a_{b,2}}} Q_h^{(t)}(\mathrm{up}) \middle| j \in [k] \right] \right| \\
&= o\left(\overline{\pi}_2^{(t)}(\mathrm{up}) \left(1 - \overline{\pi}_2^{(t)}(\mathrm{up})\right) Q_{2,1}^{(t)}(\mathrm{up})\right).
\end{aligned} \tag{930}$$

Plugging (922), (928), and (930) into (893), we obtain

$$\begin{aligned}
\delta a_{p,3}^{(t)} = \frac{1}{k} \sum_{i=1}^{k-1} \frac{k-i}{k} \left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^i \Bigg(&(1 + o(1)) \left(1 - \overline{\pi}_2^{(t)}(\mathrm{up})\right) \\
&- \frac{1}{(k-i) \exp\left(\frac{a_{p,1}^{(t)} - a_{p,3}^{(t)}}{k-1} + \left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}\right)} \Bigg) Q_{2,i}^{(t)}(\mathrm{up}).
\end{aligned} \tag{931}$$

This gives (824).

Now we bound $\delta a_{p,1}^{(t)}$. By (82) and Lemma 3 we have

$$\begin{aligned}
\delta a_{p,1}^{(t)} = &(1 + o(1/k^2)) \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0, |\mathcal{A}_{h,\times}| \geqslant 1, j \in \mathcal{A}_{h,\times}\} \frac{1}{|\mathcal{A}_{h,\times}|} \pi_h^{(t)}(j) G_h^{(t)} \right] \\
&+ \left(1 + o(1/k^2)\right) \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = \times\} \frac{N_h(j)}{e^{a_{b,0}-a_{b,2}}} \pi_h^{(t)}(j) \pi_h^{(t)}(\mathrm{up}) Q_h^{(t)}(\mathrm{up}) \middle| j \in [k] \right] \\
&- \left(1 + o(1/k^2)\right) \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0, |\mathcal{A}_{h,\times}| = 0\} \frac{N_h(j)}{e^{a_{b,0}-a_{b,2}}} \pi_h^{(t)}(j) \left(Q_h^{(t)}(j) \right.\right. \\
&\left.\left. - G_h^{(t)}\right) \middle| j \in [k] \right] \\
&- \left(1 + o(1/k^2)\right) \mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0, |\mathcal{A}_{h,\times}| \geqslant 1, j \in [k] \backslash \mathcal{A}_{h,\times}\} \frac{N_h(j)}{|\mathcal{A}_{h,\times}| e^{a_{b,1}-a_{b,2}}} \right. \\
&\left. \pi_h^{(t)}(j) \left(Q_h^{(t)}(j) - G_h^{(t)}\right) \right].
\end{aligned} \tag{932}$$

Similar as in (894), for the first term we have

$$\mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{c_h = 0, |\mathcal{A}_{h,\times}| \geqslant 1, j \in \mathcal{A}_{h,\times}\} \frac{1}{|\mathcal{A}_{h,\times}|} \pi_h^{(t)}(j) G_h^{(t)} \right]$$

$$= (1+o(1/k^2))\sum_{i=1}^{k}\mathbb{P}\left(\mathcal{E}_{2,i}\right)\sum_{j=1}^{k}\frac{(k-i+1)k-j}{(k-i+1)k}\cdot\frac{1}{j}\mathbb{E}\left[\mathbb{1}\{m\in\mathcal{A}_{h_i+2j,\times}\}\right.$$
$$\left.\pi^{(t)}_{h_i+2j}(m)G^{(t)}_{h_i+2j}\middle|\overline{\mathcal{E}}^{i,j}_{2,i}\right]$$
$$= \frac{1+o(1/k^2)}{k}\sum_{i=1}^{k}\mathbb{P}\left(\mathcal{E}_{2,i}\right)\sum_{j=1}^{k}\frac{(k-i+1)k-j}{(k-i+1)k}\mathbb{E}\left[\right.$$
$$\left.\pi^{(t)}_{h_i+2j}(m)G^{(t)}_{h_i+2j}\middle|\overline{\mathcal{E}}^{i,j}_{2,i}, m\in\mathcal{A}_{h_i+2j,\times}\right]. \tag{933}$$

By (885), (897) we have for all $i\in[k]$ (recall ϵ is small enough, $Q^{(t)}_{2,k}(\mathtt{up})=0$):

$$\sum_{j=1}^{k-1}\frac{(k-i+1)k-j}{(k-i+1)k}\mathbb{E}\left[\pi^{(t)}_{h_i+2j}(m)G^{(t)}_{h_i+2j}\middle|\overline{\mathcal{E}}^{i,j}_{2,i}, m\in\mathcal{A}_{h_i+2j,\times}\right]$$
$$= (1+o(1))\frac{(k-i+1)k-k+1}{(k-i+1)k}\mathbb{E}\left[\right.$$
$$\left.\pi^{(t)}_{h_i+2(k-1)}(m)G^{(t)}_{h_i+2(k-1)}\middle|\overline{\mathcal{E}}^{i,k-1}_{2,i}, m\in\mathcal{A}_{h_i+2(k-1),\times}\right]$$
$$= \exp\left(-\left(1+O\left(\frac{1}{(\ln(1/\epsilon))^{1+O(1/N)}}\right)\right)\frac{a^{(t)}_{p,1}+a^{(t)}_{p,2}}{k-1}\right)\left(\frac{1}{(k-i+1)k}\right.$$
$$\left.+\frac{k-i}{k-i+1}\pi^{(t)}_{2,i}(\mathtt{up})Q^{(t)}_{2,i}(\mathtt{up})\right). \tag{934}$$

And by (886) we have for all $i\in[k]$:

$$\mathbb{E}\left[\pi^{(t)}_{h_i+2k}(m)G^{(t)}_{h_i+2k}\middle|\overline{\mathcal{E}}_{2,i}, m\in\mathcal{A}_{h_i+2k,\times}\right] = (1$$
$$+o(1/k^2))\frac{\pi^{(t)}_{2,i}(\mathtt{up})\left(1-\pi^{(t)}_{2,i}(\mathtt{up})\right)}{k}Q^{(t)}_{2,i}(\mathtt{up}). \tag{935}$$

Therefore, by (933), (934), (935), (904), we can rewrite the first term in (932) as follows:

$$\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|\geqslant 1, j\in\mathcal{A}_{h,\times}\}\frac{1}{|\mathcal{A}_{h,\times}|}\pi^{(t)}_h(j)G^{(t)}_h\right]$$
$$= \frac{1+o(1)}{k}\sum_{i=1}^{k}\mathbb{P}\left(\mathcal{E}_{2,i}\right)\left[\exp\left(-\left(1+O\left(\frac{1}{(\ln(1/\epsilon))^{1+O(1/N)}}\right)\right)\frac{a^{(t)}_{p,1}+a^{(t)}_{p,2}}{k-1}\right)\right.$$
$$\cdot\left(\frac{1}{(k-i+1)k}+\frac{k-i}{k-i+1}\pi^{(t)}_{2,i}(\mathtt{up})Q^{(t)}_{2,i}(\mathtt{up})\right)$$
$$\left.+\frac{k-i}{k-i+1}\frac{\pi^{(t)}_{2,i}(\mathtt{up})\left(1-\pi^{(t)}_{2,i}(\mathtt{up})\right)}{k}Q^{(t)}_{2,i}(\mathtt{up})\right]$$
$$\overset{(906)}{=} \frac{1+o(1)}{k^2}\sum_{i=1}^{k}\left(\overline{\pi}^{(t)}_2(\mathtt{up})\right)^{i-1}\left[\exp\left(-\left(1+O\right.\right.\right.$$
$$\left.\left.\left.\left(\frac{1}{(\ln(1/\epsilon))^{1+O(1/N)}}\right)\right)\frac{a^{(t)}_{p,1}+a^{(t)}_{p,2}}{k-1}\right)\right.$$
$$\cdot\left(\frac{1}{k}+(k-i)\overline{\pi}^{(t)}_2(\mathtt{up})Q^{(t)}_{2,i}(\mathtt{up})\right)+\frac{k-i}{k}\overline{\pi}^{(t)}_2(\mathtt{up})\left(1\right.$$
$$\left.\left.-\overline{\pi}^{(t)}_2(\mathtt{up})\right)Q^{(t)}_{2,i}(\mathtt{up})\right]. \tag{936}$$

Similar as how we bound the last two terms in (893), By Induction Hypothesis 3 and Lemma 3 we can compute that

$$\frac{1}{k^2}\sum_{i=1}^{k}\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^{i-1}\exp\left(-\left(1+O\left(\frac{1}{(\ln(1/\epsilon))^{1+O(1/N)}}\right)\right)\frac{a_{p,1}^{(t)}+a_{p,2}^{(t)}}{k-1}\right)\left(\frac{1}{k}+(k-i)\overline{\pi}_2^{(t)}(\mathrm{up})Q_{2,i}^{(t)}(\mathrm{up})\right)$$
$$=o\left(\frac{1}{k^2}\sum_{i=1}^{k-1}\frac{k-i}{k}\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^{i}\left(1-\overline{\pi}_2^{(t)}(\mathrm{up})\right)Q_{2,i}^{(t)}(\mathrm{up})\right), \tag{937}$$

and the last three terms in (932) can also be bounded by

$$o\left(\frac{1}{k^2}\sum_{i=1}^{k-1}\frac{k-i}{k}\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^{i}\left(1-\overline{\pi}_2^{(t)}(\mathrm{up})\right)Q_{2,i}^{(t)}(\mathrm{up})\right). \tag{938}$$

By the above two relations we have

$$\delta a_{p,1}^{(t)}=\frac{1+o(1)}{k^2}\sum_{i=1}^{k-1}\frac{k-i}{k}\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^{i}\left(1-\overline{\pi}_2^{(t)}(\mathrm{up})\right)Q_{2,i}^{(t)}(\mathrm{up}). \tag{939}$$

This gives (823).

Now we compute $\delta a_{p,0}^{(t)}$. From (81) we know that

$$\delta a_{p,0}^{(t)}=(1+o(1/k^2))\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|=0\}\pi_h^{(t)}(\mathrm{up})\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\right]$$
$$+(1+o(1/k^2))\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=\times\}\pi_h^{(t)}(\mathrm{up})\left(1-\pi_h^{(t)}(\mathrm{up})\right)Q_h^{(t)}(\mathrm{up})\right]$$
$$+(1+o(1/k^2))\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|\geqslant 1\}\frac{1}{|\mathcal{A}_{h,\times}|e^{a_{b,1}-a_{b,0}}}\pi_h^{(t)}(\mathrm{up})\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\right]. \tag{940}$$

Applying (891) to the first term of (940), we have

$$\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|=0\}\pi_h^{(t)}(\mathrm{up})\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\right]$$
$$=-\pi_1^{(t)}(\mathrm{up})G_1^{(t)}+\sum_{i=1}^{k}\mathbb{P}\left(\mathcal{E}_{2,i}\right)\left(\pi_{h_i}^{(t)}(\mathrm{up})|\mathcal{E}_{2,i}\right)\mathbb{E}\left[Q_{h_i}^{(t)}(\mathrm{up})-G_{h_i}^{(t)}\middle|\mathcal{E}_{2,i}\right]$$
$$\overset{(906)}{=}-\pi_1^{(t)}(\mathrm{up})G_1^{(t)}+(1+o(1))\sum_{i=1}^{k}\frac{k-i+1}{k}\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^{i-1}\left(\pi_{h_i}^{(t)}(\mathrm{up})|\mathcal{E}_{2,i}\right)\mathbb{E}\left[Q_{h_i}^{(t)}(\mathrm{up})-G_{h_i}^{(t)}\middle|\mathcal{E}_{2,i}\right]. \tag{941}$$

Same as how we compute (927), we can compute that

$$\overline{\pi}_2^{(t)}(\mathrm{up})\geqslant\exp\left(c_3\frac{a_{q,\times}}{\ln^2 k}\right)\pi_1^{(t)}(\mathrm{up}) \tag{942}$$

for some constant $c_3>0$. Therefore, similar as how we bound the first term of $\delta a_{p,3}^{(t)}$ in (928), by (942), (878), Induction Hypothesis 3 and Lemma 3 we can bound the first term in (940) as

$$\left|\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|=0\}\pi_h^{(t)}(\mathrm{up})\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\right]\right|$$

$$\leqslant \exp\left(-c_4\frac{\ln(1/\epsilon)}{\ln^2 k}\right)\overline{\pi}_2^{(t)}(\mathrm{up})\left(1-\overline{\pi}_2^{(t)}(\mathrm{up})\right)Q_{2,1}^{(t)}(\mathrm{up}) \tag{943}$$

for some constant $c_4 > 0$.

Applying (891) to the second term of (940), and using Lemma 3, Induction Hypothesis 3, we have (recall we define $\overline{\mathcal{E}}_{2,i}^{\times,j}$ in (890))

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=\times\}\pi_h^{(t)}(\mathrm{up})\left(1-\pi_h^{(t)}(\mathrm{up})\right)Q_h^{(t)}(\mathrm{up})\right]\\
&=(1+o(1))\sum_{i=1}^{k}\mathbb{P}\left(\mathcal{E}_{2,i}\right)\sum_{j=1}^{k}\frac{(k-i+1)k-j}{(k-i+1)k}\mathbb{E}\Big[\pi_{h_i+2j-1}^{(t)}(\mathrm{up})\,(1\\
&\quad-\pi_{h_i+2j-1}^{(t)}(\mathrm{up})\Big)Q_{h_i+2j-1}^{(t)}(\mathrm{up})\Big|\overline{\mathcal{E}}_{2,i}^{\times,j}\Big]\\
&=(1+o(1))\sum_{i=1}^{k}\mathbb{P}\left(\mathcal{E}_{2,i}\right)\sum_{j=1}^{k}\frac{(k-i+1)k-j}{(k-i+1)k}\left(1-\pi_{h_i+2j-1}^{(t)}(\mathrm{up})|\overline{\mathcal{E}}_{2,i}^{\times,j}\right)\\
&\quad\mathbb{E}\left[G_{h_i+2j}^{(t)}\Big|\overline{\mathcal{E}}_{2,i}^{i,j}\right]\\
&=(1+o(1))\sum_{i=1}^{k}\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^{i-1}\sum_{j=1}^{k}\left(1-\pi_{h_i+2j-1}^{(t)}(\mathrm{up})|\overline{\mathcal{E}}_{2,i}^{\times,j}\right)\left(\frac{k-j}{k^2}\right.\\
&\quad\left.+\frac{k-i}{k}\overline{\pi}_2^{(t)}(\mathrm{up})Q_{2,i}^{(t)}(\mathrm{up})\right),
\end{aligned}\tag{944}$$

where the last line follows from (897) and (906). Then similar as how we compute the last term of $\delta a_{p,3}^{(t)}$ in (930), by (929) and (879) we can obtain the following bound:

$$\begin{aligned}
&\left|\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=\times\}\pi_h^{(t)}(\mathrm{up})\left(1-\pi_h^{(t)}(\mathrm{up})\right)Q_h^{(t)}(\mathrm{up})\right]\right|\leqslant\exp\Big(\\
&\quad-c_5\frac{\ln(1/\epsilon)}{\ln^2 k}\Big)\overline{\pi}_2^{(t)}(\mathrm{up})\left(1-\overline{\pi}_2^{(t)}(\mathrm{up})\right)Q_{2,1}^{(t)}(\mathrm{up})
\end{aligned}\tag{945}$$

for some constant $c_5 > 0$.

And similar as how we compute the first term of $\delta a_{p,3}^{(t)}$ in (922), we can rewrite the last term in (940) as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|\geqslant 1\}\frac{1}{|\mathcal{A}_{h,\times}|e^{a_{b,1}-a_{b,0}}}\pi_h^{(t)}(\mathrm{up})\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\right]\\
&=\frac{1}{ke^{a_{b,1}-a_{b,0}}}\sum_{i=1}^{k-1}\frac{k-i}{k}\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^{i}\left((1+o(1))\left(1-\overline{\pi}_2^{(t)}(\mathrm{up})\right)\right.\\
&\quad\left.-\frac{\ln k}{(k-i)\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1}+\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}\right)Q_{2,i}^{(t)}(\mathrm{up}).
\end{aligned}\tag{946}$$

Combining (940) with (943), (945) and (946), and by (171) we have (recall we assume $\ln(1/\epsilon)\geqslant e^k$, and thus $k=\ln(1/\epsilon)^{o(1)}$)

$$\begin{aligned}
|\delta a_{p,0}^{(t)}|&\lesssim\frac{1}{(\ln(1/\epsilon))^{1+o(1)}}\cdot\frac{1}{k^2}\sum_{i=1}^{k}\frac{k-i}{k}\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^{i}\left(1-\overline{\pi}_2^{(t)}(\mathrm{up})\right)Q_{2,i}^{(t)}(\mathrm{up})\\
&+\frac{1}{(\ln(1/\epsilon))^{1+o(1)}}\cdot\frac{1}{k^3}\sum_{i=1}^{k-1}\frac{k-i}{k}
\end{aligned}$$

$$\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^i \frac{1}{(k-i)\exp\left(\frac{a_{p,1}^{(t)}-a_{p,3}^{(t)}}{k-1}+\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)} Q_{2,i}^{(t)}(\mathrm{up})$$
$$\lesssim \frac{1}{(\ln(1/\epsilon))^{1+o(1)}}\left(\delta a_{p,1}^{(t)}-\frac{1}{k^2}\delta a_{p,3}^{(t)}\right). \tag{947}$$

This gives (825).

Finally, we compute $\delta a_{p,\mathrm{up}}^{(t)}$. From (84) we know that

$$\begin{aligned}\delta a_{p,\mathrm{up}}^{(t)} = &-(1+o(1/k^2))\mathbb{E}\left[\sum_{h=1}^H \mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|=0\}\frac{N_h(\mathrm{up})}{e^{a_{b,0}-a_{b,2}}}\pi_h^{(t)}(\mathrm{up})\right.\\ &\left.\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\right]\\ &-(1+o(1/k^2))\mathbb{E}\left[\sum_{h=1}^H \mathbb{1}\{c_h=\times\}\frac{N_h(\mathrm{up})}{e^{a_{b,0}-a_{b,2}}}\pi_h^{(t)}(\mathrm{up})\left(1-\pi_h^{(t)}(\mathrm{up})\right)Q_h^{(t)}(\mathrm{up})\right]\\ &-(1+o(1/k^2))\mathbb{E}\left[\sum_{h=1}^H \mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|\geqslant 1\}\frac{N_h(\mathrm{up})}{|\mathcal{A}_{h,\times}|e^{a_{b,1}-a_{b,2}}}\pi_h^{(t)}(\mathrm{up})\right.\\ &\left.\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\right].\end{aligned} \tag{948}$$

Notice that for all $i,j\in[k]$,

$$\begin{aligned}N_{h_i}(\mathrm{up})|\mathcal{E}_{2,i} = (i-1)(k+1),\ N_{h_i+2j-1}(\mathrm{up})|\mathcal{E}_{2,i}^{\times,j} = (i-1)(k+1)+j-1,\\ N_{h_i+2j}(\mathrm{up})|\mathcal{E}_{2,i}^{i,j} = (i-1)(k+1)+j.\end{aligned} \tag{949}$$

Thus same as how we rewrite the first term of (940) in (941), we can rewrite the first term in (948) as

$$\begin{aligned}&\mathbb{E}\left[\sum_{h=1}^H \mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|=0\}\frac{N_h(\mathrm{up})}{e^{a_{b,0}-a_{b,2}}}\pi_h^{(t)}(\mathrm{up})\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\right]\\ &=\frac{1}{e^{a_{b,0}-a_{b,2}}}\sum_{i=1}^k (i-1)(k+1)\mathbb{P}\left(\mathcal{E}_{2,i}\right)\left(\pi_{h_i}^{(t)}(\mathrm{up})|\mathcal{E}_{2,i}\right)\mathbb{E}\left[Q_{h_i}^{(t)}(\mathrm{up})-G_{h_i}^{(t)}\middle|\mathcal{E}_{2,i}\right]\\ &\overset{(906)}{=}\frac{(1+o(1))(k+1)}{e^{a_{b,0}-a_{b,2}}}\sum_{i=1}^k (i-1)\frac{k-i+1}{k}\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^{i-1}\left(\pi_{h_i}^{(t)}(\mathrm{up})|\mathcal{E}_{2,i}\right)\\ &\quad\mathbb{E}\left[Q_{h_i}^{(t)}(\mathrm{up})-G_{h_i}^{(t)}\middle|\mathcal{E}_{2,i}\right].\end{aligned} \tag{950}$$

And same as in (943), from the above expression we can compute that

$$\begin{aligned}&\left|\mathbb{E}\left[\sum_{h=1}^H \mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|=0\}\frac{N_h(\mathrm{up})}{e^{a_{b,0}-a_{b,2}}}\pi_h^{(t)}(\mathrm{up})\left(Q_h^{(t)}(\mathrm{up})-G_h^{(t)}\right)\right]\right|\\ &\leqslant \exp\left(-c_6\frac{\ln(1/\epsilon)}{\ln^2 k}\right)\overline{\pi}_2^{(t)}(\mathrm{up})\left(1-\overline{\pi}_2^{(t)}(\mathrm{up})\right)Q_{2,1}^{(t)}(\mathrm{up})\end{aligned} \tag{951}$$

for some constant $c_6>0$.

Same as how we compute the second term of (940) in (944), we can rewrite the second term in (948) as

$$\begin{aligned}&\mathbb{E}\left[\sum_{h=1}^H \mathbb{1}\{c_h=\times\}\frac{N_h(\mathrm{up})}{e^{a_{b,0}-a_{b,2}}}\pi_h^{(t)}(\mathrm{up})\left(1-\pi_h^{(t)}(\mathrm{up})\right)Q_h^{(t)}(\mathrm{up})\right]\\ &=\frac{1+o(1)}{e^{a_{b,0}-a_{b,2}}}\sum_{i=1}^k\left(\overline{\pi}_2^{(t)}(\mathrm{up})\right)^{i-1}\sum_{j=1}^k\left((i-1)(k+1)+j-1\right)(1\end{aligned}$$

$$-\pi^{(t)}_{h_i+2j-1}(\mathrm{up})|\overline{\mathcal{E}}^{\times,j}_{2,i}\Big)$$
$$\cdot\left(\frac{k-j}{k^2}+\frac{k-i}{k}\overline{\pi}^{(t)}_2(\mathrm{up})Q^{(t)}_{2,i}(\mathrm{up})\right). \tag{952}$$

And same as in (945), we can bound the above term as

$$\left|\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=\times\}\frac{N_h(\mathrm{up})}{e^{a_{b,0}-a_{b,2}}}\pi^{(t)}_h(\mathrm{up})\left(1-\pi^{(t)}_h(\mathrm{up})\right)Q^{(t)}_h(\mathrm{up})\right]\right|$$
$$\leqslant \exp\left(-c_7\frac{\ln(1/\epsilon)}{\ln^2 k}\right)\overline{\pi}^{(t)}_2(\mathrm{up})\left(1-\overline{\pi}^{(t)}_2(\mathrm{up})\right)Q^{(t)}_{2,1}(\mathrm{up}) \tag{953}$$

for some constant $c_7>0$.

Same as how we rewrite the last term of (940) in (946), we can rewrite the last term in (948) as

$$\mathbb{E}\left[\sum_{h=1}^{H}\mathbb{1}\{c_h=0,|\mathcal{A}_{h,\times}|\geqslant 1\}\frac{N_h(\mathrm{up})}{|\mathcal{A}_{h,\times}|e^{a_{b,1}-a_{b,2}}}\pi^{(t)}_h(\mathrm{up})\left(Q^{(t)}_h(\mathrm{up})-G^{(t)}_h\right)\right]$$
$$=\frac{1}{ke^{a_{b,1}-a_{b,2}}}\sum_{i=1}^{k-1}\frac{k-i}{k}\left(\overline{\pi}^{(t)}_2(\mathrm{up})\right)^i\Bigg((1+o(1))(i(k+1)-1)\left(1-\overline{\pi}^{(t)}_2(\mathrm{up})\right)$$
$$-\frac{(i-1)(k+1)\ln k+k-1}{(k-i)\exp\left(\frac{a^{(t)}_{p,1}-a^{(t)}_{p,3}}{k-1}+\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}\Bigg)Q^{(t)}_{2,i}(\mathrm{up}). \tag{954}$$

Then similar as in (947), combining (951), (953), (954) with (948), and using (171), (172) we can bound $\delta a^{(t)}_{p,\mathrm{up}}$ as

$$|\delta a^{(t)}_{p,\mathrm{up}}|\lesssim\frac{1}{(\ln(1/\epsilon))^{3+o(1)}}\cdot\frac{1}{k^2}\sum_{i=1}^{k}\frac{k-i}{k}\left(\overline{\pi}^{(t)}_2(\mathrm{up})\right)^i\left(1-\overline{\pi}^{(t)}_2(\mathrm{up})\right)Q^{(t)}_{2,i}(\mathrm{up})$$
$$+\frac{1}{(\ln(1/\epsilon))^{3+o(1)}}\cdot\frac{1}{k^3}\sum_{i=1}^{k-1}\frac{k-i}{k}$$
$$\left(\overline{\pi}^{(t)}_2(\mathrm{up})\right)^i\frac{1}{(k-i)\exp\left(\frac{a^{(t)}_{p,1}-a^{(t)}_{p,3}}{k-1}+\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}Q^{(t)}_{2,i}(\mathrm{up})$$
$$\lesssim\frac{1}{(\ln(1/\epsilon))^{3+o(1)}}\left(\delta a^{(t)}_{p,1}-\frac{1}{k^2}\delta a^{(t)}_{p,3}\right). \tag{955}$$

This gives (826).

D Proof of Construction Results

Notation. We'll reuse some notations defined at the beginning of Appendix C.2. We copy them here for reader's convenience. We define

$$\pi_h:=\pi_\theta(E^{(h)}),\ \forall h\in[H].$$

Define N_h as the number of leaf nodes explored up to step h:

$$N_{h,\times}:=\sum_{i=1}^{h}\mathbb{1}\{c_i=\times\}. \tag{956}$$

At any step h, the agent is at node n_h, and we let

$$\mathcal{A}_{h,\times}:=\begin{cases}\{i\in[k]|i\text{ is taken at }n_h\}, & \text{if } c_h=0,\\ \varnothing, & \text{if } c_h=\times,\end{cases} \tag{957}$$

denote the set of downward actions that are taken at node n_h. Define

$$\forall i \in [k+1]: \quad N_h(i) \coloneqq \sum_{j=1}^{h-1} \mathbb{1}\{a_j = i\} \tag{958}$$

as the number of times action i is taken up to step h. And we let $N_h(\mathtt{up}) \coloneqq N_h(k+1)$.

We'll make use of the following lemma from Yang et al. (2025) to show the existence of our matrix constructions.

Lemma 9 (Restatement of Lemma 1 in Yang et al. (2025))**.** *Suppose matrices $X \in \mathbb{R}^{m_1 \times n_1}, Y \in \mathbb{R}^{m_2 \times n_2}$ have full column rank. Then for any matrix $A \in \mathbb{R}^{n_1 \times n_2}$, there exists a matrix $M \in \mathbb{R}^{m_1 \times m_2}$ such that*

$$X^\top M Y = A.$$

D.1 Proof of Theorem 1

First of all, from Lemma 9, the definition of A_B, A_C, A_P, A_Q in (12) and the full column rank assumption on U, V, Z, we know that our construction (13) exists. We use $O(f(k))$ to denote a quantity that is bounded above by a constant multiple of $f(k)$ for any $k \geqslant 2$ (we don't require k to be large).

Fix any full k-ary tree $\mathcal{T}$ of depth $l \in [L]$ and any goal leaf $g(\mathcal{T})$. Consider any legal trajectory generated by policy π on $\mathcal{T}$. From the proof of Lemma 1 we know that the policy expressions in Lemma 1 still hold here with $a_{b,2} = a_{p,\mathrm{up}} = a_{p,0} = a_{p,2} = 0$, i.e., we have

$$\begin{aligned} \pi_h(\mathtt{up}) &= \frac{1}{Z_h} \exp\left(\left(1+\frac{1}{k}\right)\xi_h\right), \\ \forall i \in [k]: \quad \pi_h(i) &= \frac{1}{Z_h} \exp\left(\frac{-\mathbb{1}\{i \in \mathcal{A}_{h,\times}\} a_{p,1}(e^{a_{b,1}}-1) - N_h(i) a_{p,1}}{e^{a_{b,0}} + |\mathcal{A}_{h,\times}| e^{a_{b,1}} + h - 1 - |\mathcal{A}_{h,\times}|}\right), \end{aligned} \tag{959}$$

where

$$\xi_h = \begin{cases} \frac{N_{h,\times} e^{-a_{c,0}} a_{q,\times} - (h - N_{h,\times}) e^{a_{c,0}} a_{q,0}}{N_{h,\times} e^{-a_{c,0}} + (h - N_{h,\times}) e^{a_{c,0}}}, & \text{if } c_h = 0, \\ \frac{N_{h,\times} e^{a_{c,1}} a_{q,\times} - (h - N_{h,\times}) e^{-a_{c,1}} a_{q,0}}{N_{h,\times} e^{a_{c,1}} + (h - N_{h,\times}) e^{-a_{c,1}}}, & \text{if } c_h = \times, \end{cases} \tag{960}$$

and Z_h is the normalizing factor:

$$Z_h \coloneqq \exp\left(\left(1+\frac{1}{k}\right)\xi_h\right) + \sum_{i=1}^{k} \exp\left(\frac{\varphi_h(i)}{e^{a_{b,0}} + |\mathcal{A}_{h,\times}| e^{a_{b,1}} + h - 1 - |\mathcal{A}_{h,\times}|}\right). \tag{961}$$

Note that the trajectory length H is upper bounded by $2N$, and $N_{h,\times} \leqslant \lfloor \frac{h}{2} \rfloor$, as under our setting, $c_1 = 0$, and two wrong-leaf observations cannot occur in consecutive timesteps. Therefore, by our choice of $a_{q,0}, a_{q,\times}, a_{c,0}$ in (14) and (960), we have

$$\xi_h = \begin{cases} -\left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}, & \text{if } c_h = 0, \\ \left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,\times}, & \text{if } c_h = \times. \end{cases} \tag{962}$$

For each step $h \in [H]$, if the trajectory has not terminated yet, only four cases would occur: (i) $c_h = 0, |\mathcal{A}_{h,\times}| = 0$: the agent enters a non-leaf node for the first time; (ii) $c_h = \times, |\mathcal{A}_{h,\times}| = 0$: the agent is at a non-goal leaf; (iii) $c_h = 0, |\mathcal{A}_{h,\times}| \in [k-1]$: the agent is at a non-leaf node and has already explored part of its children; (iv) $c_h = 0, |\mathcal{A}_{h,\times}| = k$: the agent is at a non-leaf node and has explored all k children.

For case (i), from (959), (962) we know that when $c_h = 0, |\mathcal{A}_{h,\times}| = 0$,

$$\pi_h(\mathtt{up}) = \frac{\exp\left(-\left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}\right)}{\exp\left(-\left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}\right) + \sum_{i=1}^{k} \exp\left(\frac{-N_h(i) a_{p,1}}{e^{a_{b,0}} + h - 1}\right)}$$

$$\leqslant \frac{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)+\exp\left(\frac{-\min_{i\in[k]}\{N_h(i)\}a_{p,1}}{e^{a_{b,0}}+h-1}\right)}. \tag{963}$$

Note that

$$\min_{i\in[k]}\{N_h(i)\} \leqslant \frac{h-1-N_h(\texttt{up})}{k}. \tag{964}$$

Thus by our choice of $a_{p,1}, a_{q,0}, a_{b,0}$ in (14) and (963), we have

$$\pi_h(\texttt{up}) \leqslant \exp\left(-\left(1+O\left(\frac{1}{\ln k}\right)\right)a_{q,0}\right) \leqslant \frac{\epsilon}{2N} \tag{965}$$

when constant C is large enough in (14). This suggests under our construction, when the agents enter a new non-leaf node, it will explore its children instead of going up with probability at least $1-\frac{\epsilon}{2N}$.

Similarly, for case (ii), all actions other than $\texttt{up}$ are illegal, and from (959), (962) and our choice of $a_{q,\times}, a_{p,1}$ in (14) we know that when $c_h = \times$,

$$1-\pi_h(\texttt{up}) = \frac{\sum_{i=1}^k \exp\left(\frac{-N_h(i)a_{p,1}}{e^{a_{b,0}}+h-1}\right)}{\exp\left(\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,\times}\right)+\sum_{i=1}^k \exp\left(\frac{-N_h(i)a_{p,1}}{e^{a_{b,0}}+h-1}\right)} \leqslant \frac{\epsilon}{2N} \tag{966}$$

when constant C is large enough in (14). This suggests the agent will go up at any non-goal leaf with probability at least $1-\frac{\epsilon}{2N}$.

For case (iii), $|\mathcal{A}_{h,\times}| \in [k-1]$, from (959) and our choice of $a_{p,1}, a_{b,0}, a_{b,1}$ in (14) we know that when $|\mathcal{A}_{h,\times}| \in [k-1]$,

$$\begin{aligned}
\forall i \in \mathcal{A}_{h,\times}: \quad \pi_h(i) &= \frac{1}{Z_h}\exp\left(\frac{-a_{p,1}\left(e^{a_{b,1}}-1+N_h(i)\right)}{e^{a_{b,0}}+|\mathcal{A}_{h,\times}|e^{a_{b,1}}+h-1-|\mathcal{A}_{h,\times}|}\right) \\
&\leqslant \frac{\exp\left(\frac{-a_{p,1}(e^{a_{b,1}}-1+N_h(i))}{e^{a_{b,0}}+|\mathcal{A}_{h,\times}|e^{a_{b,1}}+h-1-|\mathcal{A}_{h,\times}|}\right)}{\sum_{j\in\mathcal{A}_{h,\times}}\exp\left(\frac{-a_{p,1}N_h(j)}{e^{a_{b,0}}+|\mathcal{A}_{h,\times}|e^{a_{b,1}}+h-1-|\mathcal{A}_{h,\times}|}\right)} \\
&\leqslant \exp\left(-\frac{1}{|\mathcal{A}_{h,\times}|}\left(1+O\left(\frac{1}{\ln k}\right)\right)a_{p,1}\right) \\
&\leqslant \exp\left(-\frac{1}{k-1}\left(1+O\left(\frac{1}{\ln k}\right)\right)a_{p,1}\right) \leqslant \frac{\epsilon}{2kN}
\end{aligned} \tag{967}$$

when constant C is large enough in (14). Analogously, we have

$$\begin{aligned}
\pi_h(\texttt{up}) &\leqslant \frac{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)}{\sum_{j\in\mathcal{A}_{h,\times}}\exp\left(\frac{-a_{p,1}N_h(j)}{e^{a_{b,0}}+|\mathcal{A}_{h,\times}|e^{a_{b,1}}+h-1-|\mathcal{A}_{h,\times}|}\right)} \\
&= \exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right) \leqslant \frac{\epsilon}{2kN},
\end{aligned} \tag{968}$$

where the second line uses our choice of $a_{q,0}, a_{p,1}, a_{b,0}, a_{b,1}$ in (14) as well as the fact that

$$N_h(j) \leqslant \frac{N}{k}, \ \forall j \in [k]. \tag{969}$$

(967) and (968) together indicate in case (iii), the agent will choose to visit the unexplored children with probability at least $1-\frac{\epsilon}{2N}$.

For case (iv), from (959) and our choice of $a_{p,1}, a_{b,0}, a_{b,1}$ in (14) we deduce when $|\mathcal{A}_{h,\times}| = k$,

$$\begin{aligned}
1-\pi_h(\texttt{up}) &= \frac{\sum_{i=1}^k \exp\left(\frac{-a_{p,1}(e^{a_{b,1}}-1+N_h(i))}{e^{a_{b,0}}+ke^{a_{b,1}}+h-1-k}\right)}{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)+\sum_{i=1}^k \exp\left(\frac{-a_{p,1}(e^{a_{b,1}}-1+N_h(i))}{e^{a_{b,0}}+ke^{a_{b,1}}+h-1-k}\right)} \\
&\leqslant \frac{k\exp\left(-\left(1+O\left(\frac{1}{\ln k}\right)\right)\frac{a_{p,1}}{k}\right)}{\exp\left(-\left(1+O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right)} \leqslant \frac{\epsilon}{2N}
\end{aligned} \tag{970}$$

when constant C is large enough in (14). This suggests the agent will go up at any non-leaf node after fully exploring its children with probability at least $1 - \frac{\epsilon}{2N}$.

Note that the trajectory length can be bounded by $H = 2N$. From the above discussion on the four cases, we know that given any $\Phi = (\mathcal{T}, g(\mathcal{T}))$ where $\mathcal{T}$ is a full k-ary tree with depth $l \in [L]$, the agent will perform DFS on $\mathcal{T}$ with probability at least $1 - \epsilon$ until the goal is found. This completes the proof of Theorem 1.

D.2 Proof of Theorem 4

First, from Lemma 9, the definition of A_B, A_C, A_P, A_Q in (12) and the full column rank assumption on U, V, Z, we know that our construction (29) exists.

Fix any full k-ary tree $\mathcal{T}$ of depth $l \in [L]$ and any goal leaf $g(\mathcal{T})$. Consider any legal trajectory generated by policy π on $\mathcal{T}$. To better align with the node/action/label numbering, we let A_P, A_Q's column indices be 0-indexed, and A_P, A_Q's row indices be 1-indexed, and we let $\times$ and $\checkmark$ denote index 1 and 2 respectively, e.g., $A_Q(:, 0)$ and $A_Q(:, \times)$ denote the first and second columns of A_Q respectively. We let $\texttt{up}$ denote the action index $k + 1$.

By (10), (12) and (29) we have

$$\pi_h = \operatorname{softmax}(\mu_h + \nu_h), \tag{971}$$

where

$$\mu_h := \frac{A_P(:,0)e^{a_{b,0}} + \sum_{j\in\mathcal{A}_{h,\times}} A_P(:,j)e^{a_{b,1}} + \sum_{i=1}^{k}(N_h(i) - \mathbb{1}\{i \in \mathcal{A}_{h,\times}\})A_P(:,i)}{e^{a_{b,0}} + |\mathcal{A}_{h,\times}|e^{a_{b,1}} + h - 1 - |\mathcal{A}_{h,\times}|}, \tag{972}$$

and

$$\nu_h := \begin{cases} \frac{N_{h,\times}e^{-a_{c,0}}A_Q(:,\times) + (h-N_{h,\times})e^{a_{c,0}}A_Q(:,0)}{N_{h,\times}e^{-a_{c,0}} + (h-N_{h,\times})e^{a_{c,0}}}, & \text{if } c_h = 0 \\ \frac{N_{h,\times}e^{a_{c,1}}A_Q(:,\times) + (h-N_{h,\times})e^{-a_{c,1}}A_Q(:,0)}{N_{h,\times}e^{a_{c,1}} + (h-N_{h,\times})e^{-a_{c,1}}}, & \text{if } c_h = \times \end{cases}. \tag{973}$$

Combining (30) with the above two expressions, we can further simplify μ_h and ν_h as

$$\forall i \in [k]: \quad \mu_h(i) = \frac{-\mathbb{1}\{i \in \mathcal{A}_{h,\times}\}a_{p,1}(e^{a_{b,1}} - 1) - N_h(i)a_{p,1}}{e^{a_{b,0}} + |\mathcal{A}_{h,\times}|e^{a_{b,1}} + h - 1 - |\mathcal{A}_{h,\times}|}, \quad \mu_h(\texttt{up}) = 0, \tag{974}$$

and

$$\begin{aligned} \forall i \in [k]: \quad \nu_h(i) &= \begin{cases} \left(1 + O\left(\frac{1}{\ln^2 k}\right)\right)\lambda_i, & \text{if } c_h = 0 \\ O\left(\frac{1}{k\ln^2 k}\right)\lambda_i, & \text{if } c_h = \times \end{cases}, \\ \nu_h(\texttt{up}) &= \begin{cases} -\left(1 + O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}, & \text{if } c_h = 0 \\ \left(1 + O\left(\frac{1}{k\ln^2 k}\right)\right)a_{q,\times}, & \text{if } c_h = \times \end{cases}. \end{aligned} \tag{975}$$

Same as in the proof of Theorem 1 in Appendix D.1, for each step $h \in [H]$, if the trajectory has not terminated yet, only four cases would occur: (i) $c_h = 0, |\mathcal{A}_{h,\times}| = 0$: the agent enters a non-leaf node for the first time; (ii) $c_h = \times, |\mathcal{A}_{h,\times}| = 0$: the agent is at a non-goal leaf; (iii) $c_h = 0, |\mathcal{A}_{h,\times}| \in [k-1]$: the agent is at a non-leaf node and has already explored part of its children; (iv) $c_h = 0, |\mathcal{A}_{h,\times}| = k$: the agent is at a non-leaf node and has explored all k children.

For case (i), from (971), (974), (975), our choice of $a_{p,1}, a_{b,0}, a_{q,0}, \lambda_i$ in (30) and the fact that

$$N_h(i) \leqslant N/k \ \forall i \in [k]$$

we know that when $c_h = 0, |\mathcal{A}_{h,\times}| = 0$,

$$\begin{aligned} 1 - \pi_h(1) &\leqslant \frac{\exp(\mu_h(\texttt{up}) + \nu_h(\texttt{up})) + \sum_{i=2}^{k}\exp(\mu_h(i) + \nu_h(i))}{\exp(\mu_h(1) + \nu_h(1))} \\ &\leqslant \frac{\exp\left(-\left(1 + O\left(\frac{1}{\ln^2 k}\right)\right)a_{q,0}\right) + \sum_{i=2}^{k}\exp\left(\left(1 + O\left(\frac{1}{\ln^2 k}\right)\right)\lambda_i\right)}{\exp\left(\left(1 + O\left(\frac{1}{\ln k}\right)\right)\lambda_1\right)} \leqslant \frac{\epsilon}{2N} \end{aligned} \tag{976}$$

when constant C is large enough in (30). This suggests when entering a non-leaf node for the first time, the agent will act the same as the optimal policy $\pi^\star$ and explore its first child with probability at least $1 - \frac{\epsilon}{2N}$.

Similarly, For case (ii), from (971),(974) (975), our choice of $a_{p,1}, a_{b,0}, \lambda_i, a_{q,\times}$ in (30) we know that when $c_h = \times$,

$$\begin{aligned} 1 - \pi_h(\mathtt{up}) &\leqslant \frac{\sum_{i=1}^k \exp\left(\mu_h(i) + \nu_h(i)\right)}{\exp\left(\mu_h(\mathtt{up}) + \nu_h(\mathtt{up})\right)} \\ &\leqslant \frac{k}{\exp\left(\left(1 + O\left(\frac{1}{\ln k}\right)\right) a_{q,\times}\right)} \leqslant \frac{\epsilon}{2N} \end{aligned} \tag{977}$$

given constant C large enough in (30). This suggests when at a non-goal leaf, the agent will act the same as the optimal policy $\pi^\star$ and go up with probability at least $1 - \frac{\epsilon}{2N}$.

For case (iii) where the agent is at a node whose $i := |\mathcal{A}_{h,\times}| \in [k-1]$ children have been explored, we let m be the smallest index among legal action set $[k] \backslash \mathcal{A}_{h,\times}$. Then from (971) we deduce

$$\begin{aligned} &1 - \pi_h(m) \\ &\quad \exp\left(\mu_h(\mathtt{up}) + \nu_h(\mathtt{up})\right) + \sum_{j \in \mathcal{A}_{h,\times}} \exp\left(\mu_h(j) + \nu_h(j)\right) \\ &\quad + \sum_{j \in [k] \backslash (\mathcal{A}_{h,\times} \cup \{m\})} \exp\left(\mu_h(j) + \nu_h(j)\right) \\ &\leqslant \frac{}{\exp\left(\mu_h(m) + \nu_h(m)\right)}. \end{aligned} \tag{978}$$

From (974), (975), our choice of $a_{p,1}, a_{b,0}, a_{b,1}, \lambda_i, a_{q,0}$ in (30) we know that

$$\begin{aligned} \forall j \in \mathcal{A}_{h,\times}: \quad & \mu_h(j) + \nu_h(j) = -\frac{1}{i}\left(1 + O\left(\frac{1}{\ln k}\right)\right) a_{p,1} + \lambda_j, \\ \forall j \in [k] \backslash \mathcal{A}_{h,\times}: \quad & \mu_h(j) + \nu_h(j) = O\left(\frac{1}{k \ln^2 k}\right) a_{p,1} + \lambda_j, \\ & \mu_h(\mathtt{up}) + \nu_h(\mathtt{up}) = -\left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}. \end{aligned} \tag{979}$$

Substituting (979) into (978), and by our choice of index m and $a_{p,1}, a_{q,0}, \lambda_i$ in (30) we get

$$1 - \pi_h(m) \leqslant \frac{\epsilon}{2N} \tag{980}$$

when constant C is large enough in (30). This implies when at a node whose children are only partially visited, the agent will choose the child m with the smallest index among the unexplored children (i.e., $m = \arg\max_{j \in [k] \backslash \mathcal{A}_{h,\times}} p_j^a$), same as what $\pi^\star$ does, with probability at least $1 - \frac{\epsilon}{2N}$.

For case (iv) where the agent is at a node whose k children all have been explored ($|\mathcal{A}_{h,\times}| = k$), from (974), (975), (30) we know that

$$\begin{aligned} \forall i \in [k]: \quad & \mu_h(i) + \nu_h(i) = -\frac{1}{k}\left(1 + O\left(\frac{1}{\ln k}\right)\right) a_{p,1} + \lambda_i, \\ & \mu_h(\mathtt{up}) + \nu_h(\mathtt{up}) = -\left(1 + O\left(\frac{1}{\ln^2 k}\right)\right) a_{q,0}. \end{aligned} \tag{981}$$

By the above relation, (971) and our choice of $a_{p,1}, a_{q,0}, \lambda_i$ in (30) we have

$$1 - \pi_h(\mathtt{up}) \leqslant \frac{\sum_{i=1}^k \exp\left(\mu_h(i) + \nu_h(i)\right)}{\exp\left(\mu_h(\mathtt{up}) + \nu_h(\mathtt{up})\right)} \leqslant \frac{\epsilon}{2N} \tag{982}$$

when constant C is large enough in (30). This implies when at a node whose children all have been explored, the agent will again act the same as $\pi^\star$ and go up with probability at least $1 - \frac{\epsilon}{2N}$.

Putting the results of the above four cases together, and using the fact that the trajectory length is upper bounded by $H = 2N$, we know that (31) in Theorem 4 holds under our construction.

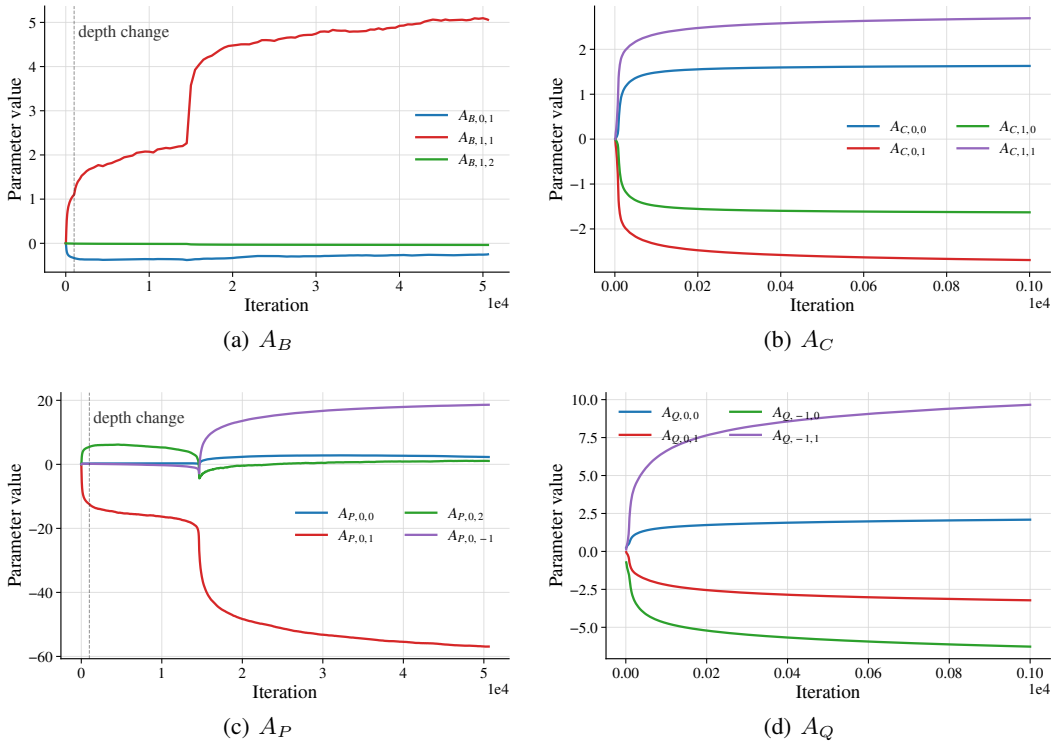
E EXPERIMENTS

Setup. To validate our training analysis, we conduct simulations using the transformer defined in (10) with zero initialization of the parameters. We set $k = 3$, $N = \frac{k^{L+1}-1}{k-1} + 10$, $d_1 = N + 1$, $d_2 = k + 2$, $d_3 = 3$, and choose $u_i \in \mathbb{R}^{d_1}$, $v_j \in \mathbb{R}^{d_2}$, $z_p \in \mathbb{R}^{d_3}$ to be the one-hot vectors with respect to the index i, j, and p respectively (assuming the index starts from 0 here). For both balanced and imbalanced cases, we follow the two-depth curriculum training procedure, and use stochastic policy gradient with batch size $b = 256$ to train the model, i.e., we have

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{b} \sum_{j=1}^{b} \sum_{h=1}^{H} \nabla_\theta \log \pi_\theta \left(a_h^{(j)} \mid E^{(h,j)} \right) \sum_{i=h}^{H} \gamma^{i-1} r_{i+1}^{(j)}, \tag{983}$$

where $E^{(h,j)}$ ($j \in [b]$) is the state embedding of the state $s_h^{(j)}$ at the h-th step in the j-th trajectory in the batch. Each trajectory $(s_1^{(j)}, a_1^{(j)}, \ldots, s_H^{(j)}, a_H^{(j)})$ is obtained by first sampling a tree $\mathcal{T} \overset{\text{iid}}{\sim} p^l_{\text{train,tree}}$ ($l = 1, 2$) from the training distribution and then executing π_θ starting from the root node, until the episode ends. We set the learning rate $\eta = 10$. For the balanced case, we set $\gamma = 1$ and train $\theta = \{B, C, \overline{P}, Q\}$ at depth-1 stage, and train $\theta = \{B, \overline{P}\}$ at depth-2 stage (for empirical acceleration, we also train B at depth-2 stage in the experiments); for the imbalanced case, we set $\gamma = 0.9$, $p^a_{i+1} = 0.9 p^a_i, i \in [k-1]$, and train $\theta = \{B, C, P, Q\}$ at both depths. For every 50 iterations, we also evaluate on L test sets, each of size 128, consisting of i.i.d. random full k-ary trees with depth $l = 1, \cdots, L$ respectively.

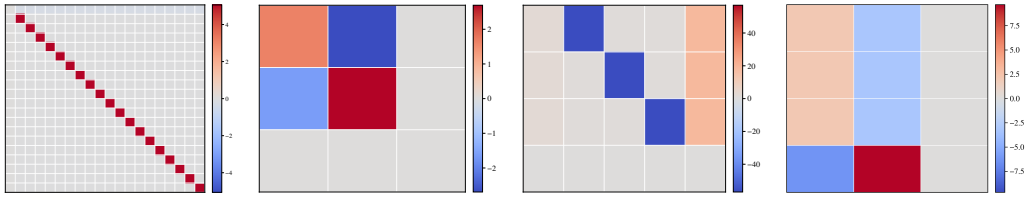
Results. For the balanced case, we set $L = 4$ and train on depth-1 and depth-2 perfect trees for 1000 and 49600 iterations, respectively. Figure 3 reports the success rate curves at depths 1 and 2. Figure 4 shows trajectories of selected matrix entries in A_B, A_C, A_P, and A_Q, and Figure 5 shows heatmaps of A_C, A_P, A_Q and the first 20 rows and 20 columns of A_B under the balanced goal distribution. We can see that the experiment results overall agree with our theoretical analysis.



(a) A_B (b) A_C

(c) A_P (d) A_Q

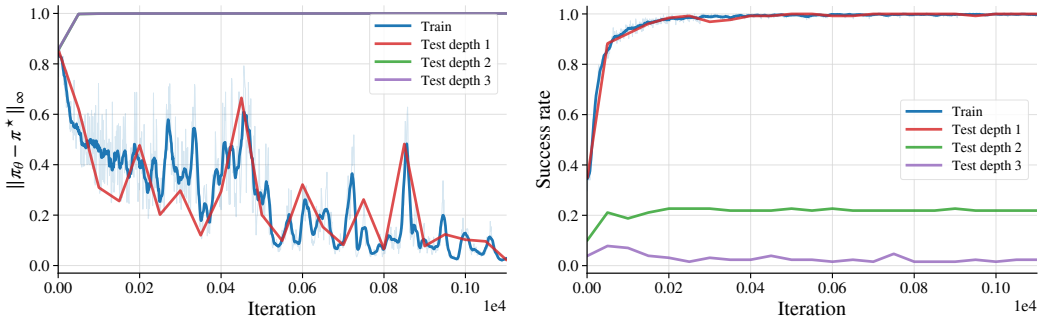
Figure 4: Selected entries of A_B, A_C, A_P, and A_Q under balanced goal distribution: A_B and A_P through depth-1+depth-2 training; A_C and A_Q through depth-1 training.



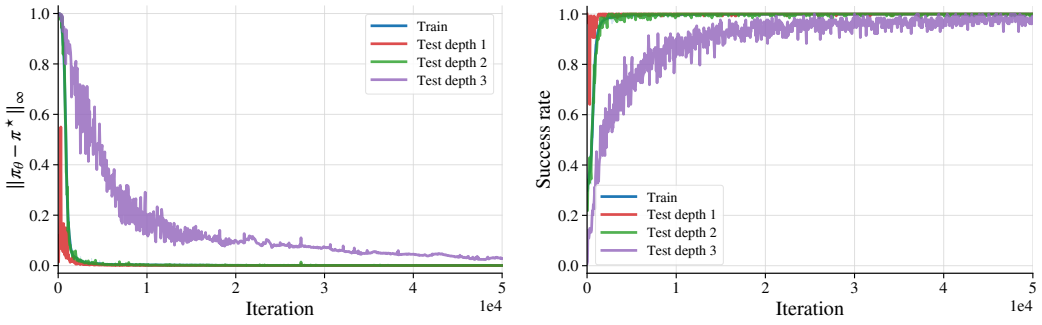
(a) $A_B[:20,:20]$ (b) A_C (c) A_P (d) A_Q

Figure 5: Heatmaps of $A_B[:20,:20]$, A_C, A_P, and A_Q under the balanced goal distribution: A_B, A_P at the end of Depth-2 curriculum training, A_C, A_Q at the end of Depth-1 curriculum training.

For the imbalanced case, we set $L = 3$ and train on depth-1 and depth-2 perfect trees for 1100 and 50000 iterations, respectively. Figure 6 reports the ℓ_∞ error and success rate for both depth-1 and depth-2 curriculum training. Figure 7 shows the corresponding trajectories of selected entries in A_B, A_C, A_P, and A_Q. Figure 8 shows the heatmaps of $A_B[:20,:20]$, A_C, A_P, and A_Q after depth-2 curriculum training.

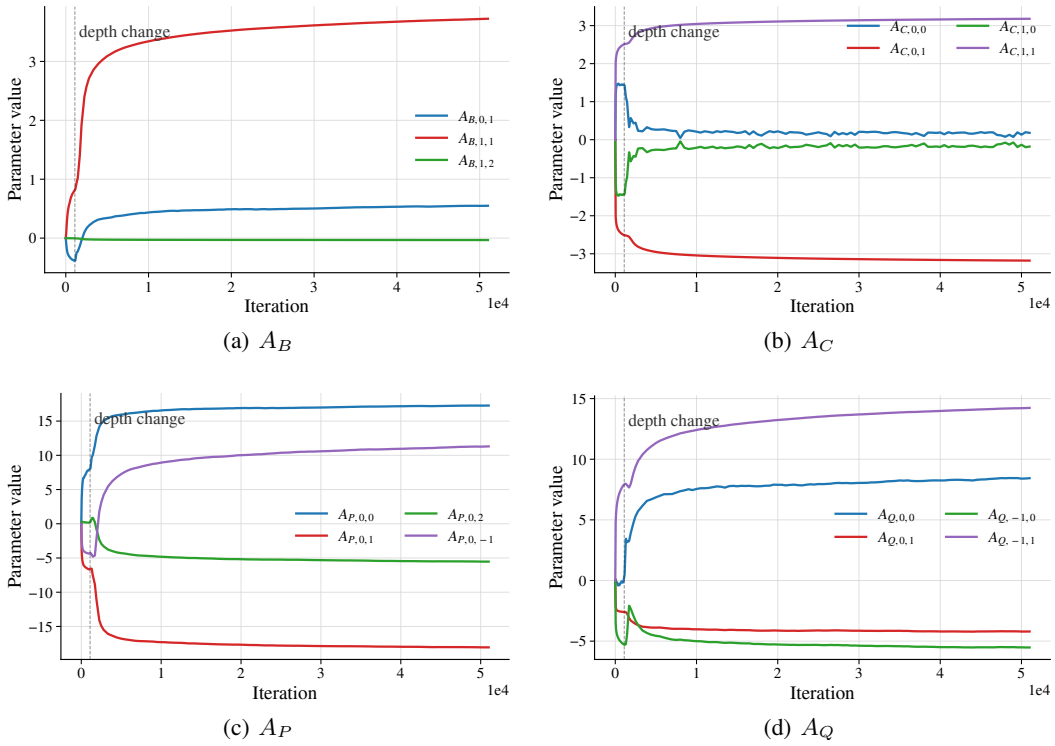


(a) Depth-1 ℓ_∞ error (b) Depth-1 success rate



(c) Depth-2 ℓ_∞ error (d) Depth-2 success rate

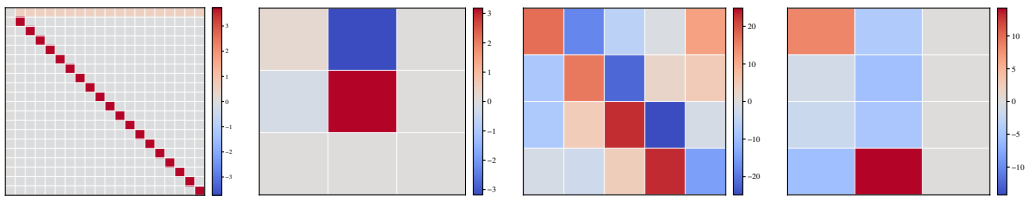
Figure 6: ℓ_∞ error to the optimal policy $\pi^\star$ and success rate curves under the imbalanced goal distribution. Top: depth-1 curriculum training for 1100 iterations. Bottom: depth-2 curriculum training for 50000 iterations.



(a) A_B (b) A_C

(c) A_P (d) A_Q

Figure 7: Selected entries of A_B, A_C, A_P, and A_Q under imbalanced training through depth-1 and depth-2 curriculum training.



(a) A_B (b) A_C (c) A_P (d) A_Q

Figure 8: Heatmaps of $A_B[:20,:20]$, A_C, A_P, and A_Q under the imbalanced goal distribution after depth-2 curriculum training.