Table 3: Assumption 5.1 verification.

	m	100	400	900	1600	2500
Mushroom	$g(m)$	4.55 ± 0.35	2.03 ± 0.15	1.16 ± 0.06	0.85 ± 0.05	0.65 ± 0.04
Statlog	$g(m)$	6.93 ± 0.75	3.16 ± 0.26	2.05 ± 0.12	1.48 ± 0.09	1.13 ± 0.05

A Details of Theoretical Analysis

For theoretical analysis, we define the L-layered fully-connected neural network f as follows:

$$\begin{aligned} \boldsymbol{f}_1 &= \boldsymbol{W}_1 \boldsymbol{x}, \\ \boldsymbol{f}_l &= \boldsymbol{W}_l \mathrm{ReLU}(\boldsymbol{f}_{l-1}), \quad 2 \leq l \leq L-1, \\ f(\boldsymbol{x}, \boldsymbol{w}) &= \sqrt{m} \boldsymbol{W}_L \mathrm{ReLU}(\boldsymbol{f}_{L-1}), \end{aligned} \tag{A.1}$$

where $\mathrm{ReLU}(\cdot) := \max\{\cdot, 0\}$, $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{W}_1 \in \mathbb{R}^{m \times d}$, $\boldsymbol{W}_l \in \mathbb{R}^{m \times m}$, $2 \leq l \leq L$, $\boldsymbol{W}_L \in \mathbb{R}^{1 \times m}$, and $\boldsymbol{w} = (\mathrm{vec}(\mathbf{W}_1); \cdots; \mathrm{vec}(\mathbf{W}_L)) \in \mathbb{R}^p$ is the collection vector of all the network parameters, $p = dm + m^2(L-2) + m$. The parameters are initialized as $\boldsymbol{W}_l = (\boldsymbol{W}, \boldsymbol{0}; \boldsymbol{0}, \boldsymbol{W})$ for each $1 \leq l \leq L-1$, where each entry of $\boldsymbol{W}$ is generated independently from $\mathcal{N}(0, 4/m)$, and $\boldsymbol{W}_L = ((\boldsymbol{w}')^\top, -(\boldsymbol{w}')^\top)$, where each entry of $\boldsymbol{w}'$ is generated independently from $\mathcal{N}(0, 2/m)$.

A.1 Verification of Assumption 5.1 For verifying Assumption 5.1, we define that

$$g(m) = \max_{t,k} \frac{\|\tilde{\boldsymbol{f}}_{\mathrm{repr}}(\boldsymbol{x}_{t,k}; \boldsymbol{w}_{t,\mathrm{repr}}) - \nabla_{\boldsymbol{w}} f(\boldsymbol{x}_{t,k}; \boldsymbol{w}_t)\|_2}{\sqrt{m}}. \tag{A.2}$$

Note that by the squeeze theorem if $\lim_{m \to +\infty} g(m) = 0$, then Assumption 1 holds. Therefore, we report $g(m)$ as well as the corresponding m in Table 3, where the experiment setting is the same as Section 6.1 and the experiments are repeated for 8 times independently. It shows that $g(m)$ empirically converges to 0, which plays as a rationale behind Assumption 5.1 from the empirical perspective.

A.2 Proof of Theorem 5.1 For such a neural network mentioned above, we have that $\boldsymbol{f}_{\mathrm{repr}}(\boldsymbol{x}; \boldsymbol{w}_{\mathrm{repr}}) = \sqrt{m}\mathrm{ReLU}(\boldsymbol{f}_{L-1})$ based on Equation 4.4, where $\boldsymbol{w}_{\mathrm{repr}} = (\mathrm{vec}(\mathbf{W}_1); \cdots; \mathrm{vec}(\mathbf{W}_{L-1})) \in \mathbb{R}^{p-m}$. For convenience of analysis, we denote $\lambda = m\gamma$ and $\nu = \tau^2\gamma$ in Algorithm 2.

We next present the assumptions and definitions used in the proof. Basically, Assumption A.1 requires the reward function is bounded, which is widely adopted in contextual bandit literature [31, 4]. Assumption A.2 is about the contexts and is the same as Assumption 3.4 in [31]. Definition A.1 defines constants σ, s that are used a lot in the following proof. Definition A.2 defines the error term ϵ and the probability events that are used for our high probability regret bound. Definition A.3 defines the set of saturated arms which are less likely to be selected.

Assumption A.1. *There exists an unknown reward function h such that for any $1 \leq t \leq T$ and $1 \leq k \leq K$,*

$$r_{t,k} = h(\boldsymbol{x}_{t,k}) + \xi_{t,k}, \tag{A.3}$$

where $|h(\boldsymbol{x}_{t,k})| \leq 1$, and $\{\xi_{t,k}\}$ forms an R-sub-Gaussian martingale difference sequence with constant $R > 0$, i.e., $\mathbb{E}[\exp(c\xi_{t,k}) \mid \xi_{1:t-1,k}, \boldsymbol{x}_{1:t,k}] \leq \exp(c^2R^2)$ for all $c \in \mathbb{R}$.

Assumption A.2. *There exists $\gamma_0 > 0$, such that $\boldsymbol{H} \succeq \gamma_0 \boldsymbol{I}$, where $\boldsymbol{H}$ is the neural tangent kernel matrix on the context set, defined the same with [31, 16]. In addition, for any $t \in [T]$, $k \in [K]$, $\|\boldsymbol{x}_{t,k}\|_2 = 1$ and $[\boldsymbol{x}_{t,k}]_j = [\boldsymbol{x}_{t,k}]_{j+d/2}$.*

Definition A.1. *Define $\sigma_{t,k}$, $s_{t,k}$ and $\hat{s}_{t,k}$ for $t \in [T], k \in [K]$ as*

$$\sigma_{t,k} = \sqrt{\frac{\gamma}{m}} \|\boldsymbol{f}_{repr}(\boldsymbol{x}_{t,k}; \boldsymbol{w}_{t,repr})\|_{\boldsymbol{Z}_t^{-1}}^2, \tag{A.4}$$

$$s_{t,k} = \sqrt{\frac{\gamma}{m}} \|\nabla_{\boldsymbol{w}} f(\boldsymbol{x}_{t,k}; \boldsymbol{w}_t)\|_{\boldsymbol{U}_t^{-1}}^2, \tag{A.5}$$

$$\hat{s}_{t.k} = s_{t,k} - \sigma_{t,k} \tag{A.6}$$

where $\boldsymbol{U}_t = \gamma \boldsymbol{I} + \sum_{i=1}^{t-1} \nabla_{\boldsymbol{w}} f(\boldsymbol{x}_{i,a_i}; \boldsymbol{w}_{i+1}) \nabla_{\boldsymbol{w}} f(\boldsymbol{x}_{i,a_i}; \boldsymbol{w}_{i+1})^\top / m$.

Definition A.2. *If we denote*

$$\epsilon_1(m) = C_{\epsilon,1}(1 - \eta m\gamma)^{C_{\epsilon,2}} \sqrt{TL/\gamma}, \quad \epsilon_2(m) = \mathcal{O}(\sqrt{\log m} m^{-1/6}), \tag{A.7}$$

where $C_{\epsilon,1}$ and $C_{\epsilon,2}$ are positive constants, we could define events $\mathcal{E}_t^\sigma$ and $\mathcal{E}_t^\mu$ as

$$\mathcal{E}_t^\sigma = \left\{\omega \in \mathcal{F}_{t+1} : \forall k \in [K], \quad \left|\hat{\theta}_{t,\boldsymbol{x}_{t,k}} - f(\boldsymbol{x}_{t,k}; \boldsymbol{w}_t)\right| \leq c_t \tau \sigma_{t,k}\right\}, \tag{A.8}$$

$$\mathcal{E}_t^\mu = \{\omega \in \mathcal{F}_t : \forall k \in [K], \quad |f(\boldsymbol{x}_{t,k}; \boldsymbol{w}_t) - h(\boldsymbol{x}_{t,k})| \leq \tau s_{t,k} + \epsilon(m)\} \tag{A.9}$$

where $c_t = \sqrt{4 \log t + 2 \log K}$, $\epsilon(m) = \epsilon_1(m) + \epsilon_2(m)$, and $\mathcal{F}_t := \{a_i, r_{i,a_i}, \boldsymbol{x}_{i,k}, k \in [K], i \in [t]\}$ denotes filtration (historical information).

Definition A.3. *Define the set of saturated arms S_t as follows*

$$\begin{aligned} S_t = &\{k \mid k \in [K], h(\boldsymbol{x}_{t,a_t^*}) - h(\boldsymbol{x}_{t,k}) \\ &\geq (1 + c_t)\tau\sigma_{t,k} + 2\tau\hat{s}_{t,k} + 2\epsilon(m)\}. \end{aligned} \tag{A.10}$$

We next introduce the conditions and lemmas in our proof. Basically, Condition A.1 provides the requirement of the width m for the neural network. Based on Condition A.1, Lemma A.1 and Lemma A.2 bound the probability of the events defined above; Lemma A.3 bounds the regret of each single round.

Condition A.1. *The network width m satisfies*

$$\begin{aligned} m \geq C_m \max\{&\sqrt{\gamma} L^{-3/2} \left[\log\left(TKL^2/\delta\right)\right]^{3/2}, \\ &T^6 K^6 L^6 \log(TKL/\delta) \max\{\gamma_0^{-4}, 1\}\}, \end{aligned} \tag{A.11}$$

$$\begin{aligned} m[\log m]^{-3} \geq C_m &TL^{12}\gamma^{-1} + C_m T^7 \gamma^{-8} L^{18} (\gamma + LT)^6 \\ &+ C_m L^{21} T^7 \gamma^{-7} (1 + \sqrt{T/\gamma})^6, \end{aligned} \tag{A.12}$$

where C_m is a positive constant and L is the network depth.

Lemma A.1. *(Lemma 4.3 in [31]). Suppose the network width m satisfies Condition A.1, and set $\eta = C_\eta (m\gamma + mLT)^{-1}$. Then, it holds that $\Pr\left(\forall t \in [T], \mathcal{E}_t^\mu\right) \geq 1 - \delta$, where C_η is a positive constant.*

Lemma A.2. *For all $t \in [T]$, the probability for the chosen arm is not in saturated arm set satisfies*

$$\Pr\left(a_t \notin S_t \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right) \geq \frac{1}{4e\sqrt{\pi}} - \frac{1}{t^2}. \tag{A.13}$$

Lemma A.3. *For all $t \in [T]$, we have that*

$$\begin{aligned} &\mathbb{E}\left[h\left(\boldsymbol{x}_{t,a_t^*}\right) - h\left(\boldsymbol{x}_{t,a_t}\right) \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] \\ &\leq \min\left\{\mathbb{E}\left[C_1 \tau s_{t,a_t} + C_2 \tau |\sigma_{t,a_t} - s_{t,a_t}| \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right], 2\right\} \\ &\quad + 4\epsilon(m) + 2t^{-2}. \end{aligned} \tag{A.14}$$

where C_1 and C_2 are some positive constants.

Next, we introduce the core lemma used in the proof, Lemma A.4, which decomposes the regret of the algorithm mainly into two parts, i.e., ϵ_f and $\min\{s_{t,a_t}, 1\}$. Lemma A.5 and Lemma A.6 bound the these parts, respectively.

Lemma A.4. *Suppose the network width m satisfies Condition A.1, and set $\eta = C_\eta (m\gamma + mLT)^{-1}$. Then with the probability at least $1 - \delta$, it holds that*

$$\begin{aligned} &\sum_{t=1}^T \left(h\left(\boldsymbol{x}_{t,a_t^*}\right) - h\left(\boldsymbol{x}_{t,a_t}\right)\right) \mathbb{1}\left(\mathcal{E}_t^\mu\right) \\ &\leq 4T\epsilon(m) + \pi^2/3 + C_2 \tau T \epsilon_f(m) + C_1 \tau C_s \sqrt{L} \sum_{t=1}^T \min\{s_{t,a_t}, 1\} \\ &\quad + \left(4 + C_1 \tau C_s \sqrt{L} + C_2 \tau \epsilon_f(m) + 4\epsilon(m)\right) \sqrt{2\log(1/\delta) T}, \end{aligned} \tag{A.15}$$

where C_η, C_1, C_2, C_s are some positive constants, $\epsilon_f(m) = \max_{t\in[T]} |\sigma_{t,a_t} - s_{t,a_t}|$, and $\mathbb{1}$ denotes the indicator function for $\mathcal{E}_t^\mu$.

LEMMA A.5. *(Lemma 4.8 in [31]). Suppose the network width m satisfies Condition A.1, and set $\eta = C_\eta(m\gamma + mLT)^{-1}$. Then with the probability at least $1-\delta$, it holds that*

$$\sum_{t=1}^{T} \min\left\{s_{t,a_t}, 1\right\} \leq \sqrt{2\gamma T(C_3 \log(1+TK)+1)} + C_4 T^{13/6}\sqrt{\log m} m^{-1/6}\gamma^{-2/3}L^{9/2}, \tag{A.16}$$

where C_η, C_3, C_4 are some positive constants.

LEMMA A.6. *Under Assumption 5.1, suppose the network width m satisfies Condition A.1, and set $\eta = C_\eta(m\gamma + mLT)^{-1}$. Then with the probability at least $1-\delta$, it holds that*

$$\epsilon_f(m) \leq \gamma^{-1/2} m^{-1/2}\left(1 + C_s^2 LT\right)\zeta_T \tag{A.17}$$

where C_η, C_s are some positive constants, and $\zeta_T = o(\sqrt{m})$.

The proof of these lemmas are postponed to Appendix A.3 for the clarification purpose. With all the above lemmas, we are ready to prove Theorem 5.1.

Proof. By Lemma A.4, with the probability at least $1-\delta$, we have

$$\begin{aligned}
R_T &= \sum_{t=1}^{T}\left(h\left(\boldsymbol{x}_{t,a_t^*}\right) - h\left(\boldsymbol{x}_{t,a_t}\right)\right)\mathbb{1}\left(\mathcal{E}_t^\mu\right) \\
&\leq 4T\epsilon(m) + \pi^2/3 + C_2\tau T\epsilon_f(m) + C_1\tau C_s\sqrt{L}\sum_{t=1}^{T}\min\left\{s_{t,a_t},1\right\} \\
&\quad + \left(4 + C_1\tau C_s\sqrt{L} + C_2\tau\epsilon_f(m) + 4\epsilon(m)\right)\sqrt{2\log(1/\delta)T} \\
&\leq C_1\tau C_s\sqrt{L}\left(\sqrt{2\gamma T(C_3\log(1+TK)+1)}\right. \\
&\quad \left. + C_2 T^{13/6}\sqrt{\log m} m^{-1/6}\gamma^{-2/3}L^{9/2}\right) \\
&\quad + \pi^2/3 + \left(4 + C_1\tau C_s\sqrt{L}\right)\sqrt{2\log(1/\delta)T} \\
&\quad + \left(4\epsilon(m) + C_2\tau\epsilon_f(m)\right)\left(T + \sqrt{2\log(1/\delta)T}\right) \\
&\leq C_1\tau C_s\sqrt{L}\sqrt{2\gamma T(C_3\log(1+TK)+1)} \\
&\quad + \left(4 + C_1\tau C_s\sqrt{L}\right)\sqrt{2\log(1/\delta)T} + \pi^2/3 \\
&\quad + C_1\tau C_s\sqrt{L}C_2 T^{13/6}\sqrt{\log m} m^{-1/6}\gamma^{-2/3}L^{9/2} \\
&\quad + 4\epsilon_2(m)\left(T + \sqrt{2\log(1/\delta)T}\right) \\
&\quad + 4C_{\epsilon,1}(1-\eta m\gamma)^{C_{\epsilon,2}}\sqrt{TL/\gamma}\left(T + \sqrt{2\log(1/\delta)T}\right) \\
&\quad + C_2\tau\gamma^{-1/2}m^{-1/2}\left(1 + C_s^2 LT\right)\zeta_T\left(T + \sqrt{2\log(1/\delta)T}\right),
\end{aligned} \tag{A.18}$$

where the first equality comes from the definition of R_T and $\mathcal{E}_t^\mu$, the first inequality comes from Lemma A.4, the second inequality comes from Lemma A.5, and the last inequality is from Lemma A.6.

By setting

$$\eta = C_4(m\gamma + mLT)^{-1}, \tag{A.19}$$

$$J = (1+LT/\gamma)\left(\log(24C_{\epsilon,1}) + \log\left(T^3 L\gamma^{-1}\log(1/\delta)\right)\right)/C_4, \tag{A.20}$$

we have

$$C_{\epsilon,1}(1-\eta m\gamma)^{C_{\epsilon,2}}\sqrt{TL/\gamma}(4 + \sqrt{2\log(1/\delta)T}) \leq 1. \tag{A.21}$$

Based on $\zeta_T = o(\sqrt{m})$ and $\epsilon_2(m) = \mathcal{O}(\sqrt{\log m} m^{-1/6})$, we then can choose m such that the following three inequalities hold:

$$C_1 C_s C_2 T^{13/6}\tau\sqrt{\log m} m^{-1/6}\gamma^{-2/3}L^5 \leq 1, \tag{A.22}$$

$$4\epsilon_2(m)\left(T + \sqrt{2\log(1/\delta)T}\right) \leq 1, \tag{A.23}$$

$$C_2\tau\gamma^{-1/2}m^{-1/2}\left(1 + C_s^2 LT\right)\zeta_T\left(T + \sqrt{2\log(1/\delta)T}\right) \leq 1. \tag{A.24}$$

We now have the bound:

$$\begin{aligned}
R_T \leq & C_1\tau C_s\sqrt{L}\sqrt{2\gamma T(C_3\log(1+TK)+1)} \\
& + \left(4 + C_1\tau C_s\sqrt{L}\right)\sqrt{2\log(1/\delta)T} + 8,
\end{aligned} \tag{A.25}$$

where C_1, C_s, C_3 are positive constants. Taking union hound over Lemmas A.3, A.4 and A.5, the above inequality holds with the probability $1-3\delta$.

By replacing δ with 3δ, it holds with the probability $1-\delta$ that

$$R_T = \mathcal{O}(\sqrt{T}), \tag{A.26}$$

which completes the proof. ∎

A.3 Proof of Lemmas in Section A.2

A.3.1 Proof of Lemma A.2

We first give the following two lemmas.

LEMMA A.7. *(Lemma 4.2 in [31]). For any $t \in [T]$, $\Pr\left(\mathcal{E}_t^\sigma \mid \mathcal{F}_t\right) \geq 1 - t^{-2}$.*

LEMMA A.8. *(Gaussian anti-concentration). For a Gaussian random variable X with mean μ and standard deviation σ, for any $\beta > 0$,*

$$\Pr\left(\frac{X-\mu}{\sigma} > \beta\right) \geq \frac{\exp\left(-\beta^2\right)}{4\sqrt{\pi}\beta}. \tag{A.27}$$

Then we start to prove our Lemma A.2.

Proof. Since $\hat{\theta}_{t,\boldsymbol{x}_{t,k}} \sim \mathcal{N}\left(f(\boldsymbol{x}_{t,k};\boldsymbol{w}_t), \tau^2\sigma_{t,k}^2\right)$ conditioned on $\mathcal{F}_t$, we have

$$\begin{aligned}
&\Pr\left(\hat{\theta}_{t,\boldsymbol{x}_{t,k}} + \tau\hat{s}_{t,k} + \epsilon(m) > h\left(\boldsymbol{x}_{t,k}\right) \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right) \\
&= \Pr\left(\frac{\hat{\theta}_{t,\boldsymbol{x}_{t,k}} - f\left(\boldsymbol{x}_{t,k};\boldsymbol{w}_t\right) + \tau\hat{s}_{t,k} + \epsilon(m)}{\tau\sigma_{t,k}}\right. \\
&\qquad \left. > \frac{h\left(\boldsymbol{x}_{t,k}\right) - f\left(\boldsymbol{x}_{t,k};\boldsymbol{w}_t\right)}{\tau\sigma_{t,k}} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right) \\
&\geq \Pr\left(\frac{\hat{\theta}_{t,\boldsymbol{x}_{t,k}} - f\left(\boldsymbol{x}_{t,k};\boldsymbol{w}_t\right) + \tau\hat{s}_{t,k}}{\tau\sigma_{t,k}}\right. \\
&\qquad \left. > \frac{\left|h\left(\boldsymbol{x}_{t,k}\right) - f\left(\boldsymbol{x}_{t,k};\boldsymbol{w}_t\right)\right| - \epsilon(m)}{\tau\sigma_{t,k}} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right) \\
&\geq \Pr\left(\frac{\hat{\theta}_{t,\boldsymbol{x}_{t,k}} - f\left(\boldsymbol{x}_{t,k};\boldsymbol{w}_t\right) + \tau\hat{s}_{t,k}}{\tau\sigma_{t,k}} > 1 \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right) \geq (4\mathrm{e}\sqrt{\pi})^{-1},
\end{aligned} \tag{A.28}$$

where the first inequality is due to $|x| \geq x$, and the second inequality is from the definition of $\mathcal{E}_t^\mu$.

Consider the following two events at round t:

$$\mathcal{A} = \left\{\forall k \in S_t, \hat{\theta}_{t,\boldsymbol{x}_{t,k}} < \hat{\theta}_{t,\boldsymbol{x}_{t,a_t^*}} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right\} \tag{A.29}$$

$$\mathcal{B} = \left\{a_t \notin S_t \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right\} \tag{A.30}$$

Clearly, $\mathcal{A}$ implies $\mathcal{B}$, since $a_t = \mathrm{argmax}_k \hat{\theta}_{t,\boldsymbol{x}_{t,k}}$. Therefore,

$$\Pr\left(a_t \notin S_t \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right) \geq \Pr\left(\forall k \in S_t, \hat{\theta}_{t,\boldsymbol{x}_{t,k}} < \hat{\theta}_{t,\boldsymbol{x}_{t,a_t^*}} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right). \tag{A.31}$$

Suppose $\mathcal{E}_t^\mu$ and $\mathcal{E}_t^\sigma$ also hold, then it is easy to show that $\forall k \in [K]$,

$$\begin{aligned}
\left|h\left(\boldsymbol{x}_{t,k}\right) - \hat{\theta}_{t,\boldsymbol{x}_{t,k}}\right| &\leq \left|h\left(\boldsymbol{x}_{t,k}\right) - f\left(\boldsymbol{x}_{t,k};\boldsymbol{w}_{t-1}\right)\right| \\
&\quad + \left|f\left(\boldsymbol{x}_{t,k};\boldsymbol{w}_{t-1}\right) - \hat{\theta}_{t,\boldsymbol{x}_{t,k}}\right| \\
&\leq \tau s_{t,k} + \epsilon(m) + c_t\tau\sigma_{t,k} \\
&= (1+c_t)\tau\sigma_{t,k} + \tau\hat{s}_{t,k} + \epsilon(m),
\end{aligned} \tag{A.32}$$

where the first inequality is due to $|x_1 + x_2| \leq |x_1| + |x_2|$, the second inequality is from Definition A.2, and the last equality is from Definition A.1.

Hence, for all $k \in S_t$, we have that

(A.33)
$$h\left(\boldsymbol{x}_{t,a_t^*}\right) - \hat{\theta}_{t,\boldsymbol{x}_{t,k}} \geq h\left(\boldsymbol{x}_{t,a_t^*}\right) - h\left(\boldsymbol{x}_{t,k}\right) - \left|h\left(\boldsymbol{x}_{t,k}\right) - \hat{\theta}_{t,\boldsymbol{x}_{t,k}}\right| \geq \tau\hat{s}_{t,k} + \epsilon(m),$$

where the first inequality is due to $|x| \geq x$, and the second inequality is from Definition A.3 and Equation A.32.

Consider the following event

(A.34)
$$\mathcal{C} = \left\{h\left(\boldsymbol{x}_{t,a_t^*}\right) - \tau\hat{s}_{t,k} - \epsilon(m) < \hat{\theta}_{t,\boldsymbol{x}_{t,a_t^*}} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right\}.$$

Since $\mathcal{E}_t^\mu$ holds, $\mathcal{E}_t^\sigma$ then implies (A.33) as shown above. We then have that if $\mathcal{C}$ and $\mathcal{E}_t^\sigma$ hold, then $\mathcal{A}$ holds, i.e. $\mathcal{E}_t^\sigma \cap \mathcal{C} \subseteq \mathcal{A}$. Taking union with $\overline{\mathcal{E}}_t^\sigma$, we have that $\mathcal{C} = \overline{\mathcal{E}}_t^\sigma \cup \mathcal{E}_t^\sigma \cap \mathcal{C} \subseteq \mathcal{A} \cup \overline{\mathcal{E}}_t^\sigma$, which implies

(A.35)
$$\Pr(\mathcal{A}) + \Pr\left(\overline{\mathcal{E}}_t^\sigma\right) \geq \Pr(\mathcal{C}).$$

Then, we have that

$$\begin{aligned}
&\Pr\left(a_t \notin S_t \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right) \\
&\geq \Pr\left(\forall k \in S_t, \hat{\theta}_{t,\boldsymbol{x}_{t,k}} < \hat{\theta}_{t,\boldsymbol{x}_{t,a_t^*}} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right) \\
&\geq \Pr\left(\hat{\theta}_{t,\boldsymbol{x}_{t,a_t^*}} + \tau\hat{s}_{t,k} + \epsilon(m) > h\left(\boldsymbol{x}_{t,a_t^*}\right) \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right) \\
&\quad - \Pr\left(\overline{\mathcal{E}}_t^\sigma \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right) \\
&\geq \frac{1}{4\mathrm{e}\sqrt{\pi}} - \frac{1}{t^2},
\end{aligned} \tag{A.36}$$

where the first equality is due to (A.31), the first inequality is due to (A.35) and (A.34), and the second inequality is from Lemmas A.7 and (A.28). □

A.3.2 Proof of Lemma A.3

Proof. Recall that given $\mathcal{F}_t$ and $\mathcal{E}_t^\mu$, the only randomness comes from sampling $\hat{\theta}_{t,\boldsymbol{x}_{t,k}}$ for $k \in [K]$.

We denote $\bar{k}_t$ as

(A.37)
$$\bar{k}_t = \operatorname*{argmin}_{k \notin S_t} \quad 2\left(1 + c_t\right)\tau\sigma_{t,k} + 3\tau\hat{s}_{t,k}.$$

Then we have

(A.38)
$$\begin{aligned}
&\mathbb{E}\left[2\left(1 + c_t\right)\tau\sigma_{t,a_t} + 3\tau\hat{s}_{t,a_t} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] \\
&\geq \mathbb{E}\left[2\left(1 + c_t\right)\tau\sigma_{t,a_t} + 3\tau\hat{s}_{t,a_t} \mid \mathcal{F}_t, \mathcal{E}_t^\mu, a_t \notin S_t\right]\Pr\left(a_t \notin S_t \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right) \\
&\geq \left(2\left(1 + c_t\right)\tau\sigma_{t,\bar{k}_t} + 3\tau\hat{s}_{t,\bar{k}_t}\right)\left(\frac{1}{4\mathrm{e}\sqrt{\pi}} - \frac{1}{t^2}\right),
\end{aligned}$$

where the first inequality is due to the property of conditional expectation, and the second inequality is from Lemma A.2 and (A.37).

If both $\mathcal{E}_t^\sigma$ and $\mathcal{E}_t^\mu$ hold, then

(A.39)
$$\forall k \in [K], \left|h\left(\mathbf{x}_{t,k}\right) - \hat{\theta}_{t,\boldsymbol{x}_{t,k}}\right| \leq \epsilon(m) + \left(1 + c_t\right)\tau\sigma_{t,k} + \tau\hat{s}_{t,k}$$

as proved in (A.32). Thus,

(A.40)
$$\begin{aligned}
&h\left(\mathbf{x}_{t,a_t^*}\right) - h\left(\mathbf{x}_{t,a_t}\right) \\
&= h\left(\mathbf{x}_{t,a_t^*}\right) - h\left(\mathbf{x}_{t,\bar{k}_t}\right) + h\left(\mathbf{x}_{t,\bar{k}_t}\right) - h\left(\mathbf{x}_{t,a_t}\right) \\
&\leq \left(1 + c_t\right)\tau\sigma_{t,\bar{k}} + 2\tau\hat{s}_{t,\bar{k}} + 2\epsilon(m) + h(\mathbf{x}_{t,\bar{k}_t}) \\
&\quad - \hat{\theta}_{t,\boldsymbol{x}_{t,\bar{k}_t}} - h\left(\mathbf{x}_{t,a_t}\right) + \hat{\theta}_{t,\boldsymbol{x}_{t,a_t}} + \hat{\theta}_{t,\boldsymbol{x}_{t,\bar{k}_t}} - \hat{\theta}_{t,\boldsymbol{x}_{t,a_t}} \\
&\leq \left(1 + c_t\right)\tau\sigma_{t,\bar{k}} + 2\tau\hat{s}_{t,\bar{k}} + 2\epsilon(m) \\
&\quad + \left|h\left(\mathbf{x}_{t,\bar{k}_t}\right) - \hat{\theta}_{t,\boldsymbol{x}_{t,\bar{k}_t}}\right| + \left|h\left(\mathbf{x}_{t,a_t}\right) - \hat{\theta}_{t,\boldsymbol{x}_{t,a_t}}\right| \\
&\leq \left(1 + c_t\right)\tau\sigma_{t,\bar{k}} + 2\tau\hat{\sigma}_{t,\bar{k}} + 2\epsilon(m) + \epsilon(m) + \left(1 + c_t\right)\tau\sigma_{t,\bar{k}} + \tau\hat{s}_{t,\bar{k}} \\
&\quad + \epsilon(m) + \left(1 + c_t\right)\tau\sigma_{t,a_t} + \tau\hat{s}_{t,a_t} \\
&= \left(1 + c_t\right)\tau(2\sigma_{t,\bar{k}} + \sigma_{t,a_t}) + 3\tau\hat{s}_{t,\bar{k}} + \tau\hat{s}_{t,a_t} + 4\epsilon(m),
\end{aligned}$$

where the first inequality is due to Definition A.3 and $\bar{k}_t \notin S_t$, the second inequality is due to $|x| \geq x$ and $\hat{\theta}_{t,\boldsymbol{x}_{t,\bar{k}_t}} \leq \hat{\theta}_{t,\boldsymbol{x}_{t,a_t}}$, and the last inequality is from (A.39).

Since a trivial bound on $h\left(\mathbf{x}_{t,a_t^*}\right) - h\left(\mathbf{x}_{t,a_t}\right)$ could be derived by

(A.41)
$$h\left(\mathbf{x}_{t,a_t^*}\right) - h\left(\mathbf{x}_{t,a_t}\right) \leq \left|h\left(\mathbf{x}_{t,a_t^*}\right)\right| + \left|h\left(\mathbf{x}_{t,a_t}\right)\right| \leq 2,$$

then we have

(A.42)
$$\begin{aligned}
&\mathbb{E}\left[h\left(\mathbf{x}_{t,a_t^*}\right) - h\left(\mathbf{x}_{t,a_t}\right) \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] \\
&= \mathbb{E}\left[h\left(\mathbf{x}_{t,a_t^*}\right) - h\left(\mathbf{x}_{t,a_t}\right) \mid \mathcal{F}_t, \mathcal{E}_t^\mu, \mathcal{E}_t^\sigma\right]\Pr\left(\mathcal{E}_t^\sigma\right) \\
&\quad + \mathbb{E}\left[h\left(\mathbf{x}_{t,a_t^*}\right) - h\left(\mathbf{x}_{t,a_t}\right) \mid \mathcal{F}_t, \mathcal{E}_t^\mu, \overline{\mathcal{E}}_t^\sigma\right]\Pr\left(\overline{\mathcal{E}}_t^\sigma\right) \\
&\leq \left(1 + c_t\right)\tau\left(2\sigma_{t,\bar{k}_t} + \mathbb{E}\left[\sigma_{t,a_t} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right]\right) \\
&\quad + 4\epsilon(m) + \frac{2}{t^2} + 3\tau\hat{s}_{t,\bar{k}} + \tau\mathbb{E}\left[\hat{s}_{t,a_t} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] \\
&= 2\left(1 + c_t\right)\tau\sigma_{t,\bar{k}_t} + 3\tau\hat{s}_{t,\bar{k}} + 4\epsilon(m) \\
&\quad + \frac{2}{t^2} + \mathbb{E}\left[\left(1 + c_t\right)\tau\sigma_{t,a_t} + \tau\hat{s}_{t,a_t} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] \\
&\leq \frac{\mathbb{E}\left[2\left(1 + c_t\right)\tau\sigma_{t,a_t} + 3\tau\hat{s}_{t,a_t} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right]}{\frac{1}{4\mathrm{e}\sqrt{\pi}} - \frac{1}{t^2}} \\
&\quad + 4\epsilon(m) + \frac{2}{t^2} + \mathbb{E}\left[\left(1 + c_t\right)\tau\sigma_{t,a_t} + \tau\hat{s}_{t,a_t} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] \\
&\leq \mathbb{E}\left[44\mathrm{e}\sqrt{\pi}\left(1 + c_t\right)\tau\sigma_{t,a_t} + 64\mathrm{e}\sqrt{\pi}\tau\hat{s}_{t,a_t} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] + 4\epsilon(m) + 2t^{-2}
\end{aligned}$$

where the first equality is due to the property of conditional expectation, the first inequality uses the bound provide in (A.40) and the trivial bound of $h\left(\mathbf{x}_{t,a_t^*}\right) - h\left(\mathbf{x}_{t,a_t}\right)$ for the second term plus Lemma A.7, the second inequality uses the bound provided in (A.38), the last inequality is directly calculated by $1 \leq 4\mathrm{e}\sqrt{\pi}$ and

(A.43)
$$\frac{1}{\frac{1}{4\mathrm{e}\sqrt{\pi}} - \frac{1}{t^2}} \leq 20\mathrm{e}\sqrt{\pi},$$

which holds since LHS is negative when $t \leq 4$ and when $t = 5$, the LHS reach its maximum as $\approx 84.11 < 96.36 \approx$ RHS.

Noticing that $|h(\mathbf{x})| \leq 1$, it is trivial to further extend the bound as

(A.44)
$$\begin{aligned}
&\mathbb{E}\left[h\left(\mathbf{x}_{t,a_t^*}\right) - h\left(\mathbf{x}_{t,a_t}\right) \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] \\
&\leq \min\left\{\mathbb{E}\left[44\mathrm{e}\sqrt{\pi}\left(1 + c_t\right)\tau\sigma_{t,a_t} + 64\mathrm{e}\sqrt{\pi}\tau\hat{s}_{t,a_t} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right], 2\right\} \\
&\quad + 4\epsilon(m) + 2t^{-2} \\
&= \min\left\{\mathbb{E}\left[44\mathrm{e}\sqrt{\pi}\left(1 + c_t\right)\tau s_{t,a_t}\right.\right. \\
&\qquad + (44\mathrm{e}\sqrt{\pi}\left(1 + c_t\right) - 64\mathrm{e}\sqrt{\pi})\tau(\sigma_{t,a_t} - s_{t,a_t}) \\
&\qquad \left.\left. \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right], 2\right\} + 4\epsilon(m) + 2t^{-2} \\
&\leq \min\left\{\mathbb{E}\left[44\mathrm{e}\sqrt{\pi}\left(1 + c_T\right)\tau s_{t,a_t}\right.\right. \\
&\qquad + (44\mathrm{e}\sqrt{\pi}\left(1 + c_T\right) - 64\mathrm{e}\sqrt{\pi})\tau|\sigma_{t,a_t} - s_{t,a_t}| \\
&\qquad \left.\left. \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right], 2\right\} + 4\epsilon(m) + 2t^{-2}.
\end{aligned}$$

Let denote $C_1 = 44\mathrm{e}\sqrt{\pi}\left(1 + c_T\right)$ and $C_2 = 44\mathrm{e}\sqrt{\pi}\left(1 + c_T\right) - 64\mathrm{e}\sqrt{\pi}$. We have that

(A.45)
$$\begin{aligned}
&\mathbb{E}\left[h\left(\mathbf{x}_{t,a_t^*}\right) - h\left(\mathbf{x}_{t,a_t}\right) \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] \\
&\leq \min\left\{\mathbb{E}\left[C_1\tau s_{t,a_t} + C_2\tau|\sigma_{t,a_t} - s_{t,a_t}| \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right], 2\right\} + 4\epsilon(m) + 2t^{-2}.
\end{aligned}$$

□

A.3.3 Proof of Lemma A.4

We first introduce two lemmas to support our proof.

Lemma A.9. *(Lemma B.9 in [31]). For any $t \in [T]$, $k \in [K]$ and $\delta \in (0,1)$, if the network width m satisfies Condition A.1, we have, with probability at least $1 - \delta$, that*

(A.46)
$$s_{t,k} \leq C_s\sqrt{L},$$

where C_s is a positive constant.

Lemma A.10. *(Azuma-Hoeffding Inequality for Super Martingale). If a super-martingale Y_t,corresponding to filtration $\mathcal{F}_t$ satisfies that $|Y_t - Y_{t-1}| \leq B_t$, then for any $\delta \in (0,1)$, we have*

(A.47)
$$Y_t - Y_0 \leq \sqrt{2\log(1/\delta)\sum_{i=1}^{t} B_i^2}.$$

Then the proof for Lemma A.4 is as follows:

Proof. By setting τ such that $C_1\tau \geq 2$ and using Lemma A.3, we have

$$\begin{aligned}
&\mathbb{E}\left[h\left(\boldsymbol{x}_{t,a_t^*}\right) - h\left(\boldsymbol{x}_{t,a_t}\right) \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] \\
&\leq \min\left\{\mathbb{E}\left[C_1\tau s_{t,a_t} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right], 2\right\} \\
&\quad + \min\left\{\mathbb{E}\left[C_2\tau|\sigma_{t,a_t} - s_{t,a_t}| \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right], 2\right\} \\
&\quad + 4\epsilon(m) + 2t^{-2} \\
&\leq C_1\tau \min\left\{\mathbb{E}\left[s_{t,a_t} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right], 1\right\} + \mathbb{E}\left[C_2\tau|\sigma_{t,a_t} - s_{t,a_t}| \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] \\
&\quad + 4\epsilon(m) + 2t^{-2},
\end{aligned} \tag{A.48}$$

where the first inequality is due to $\min\{|x|+|y|, 2\} \leq \min\{|x|, 2\} + \min\{|y|, 2\}$, and the last inequality is due to $C_1\tau \geq 2$ and $\min\{x, 2\} \leq x$.

Based on Lemma A.9, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\mathbb{E}\left[h\left(\boldsymbol{x}_{t,a_t^*}\right) - h\left(\boldsymbol{x}_{t,a_t}\right) \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] \\
&\leq C_1\tau C_s\sqrt{L}\mathbb{E}\left[\min\left\{s_{t,a_t}, 1\right\} \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] \\
&\quad + \mathbb{E}\left[C_2\tau|\sigma_{t,a_t} - s_{t,a_t}| \mid \mathcal{F}_t, \mathcal{E}_t^\mu\right] \\
&\quad + 4\epsilon(m) + 2t^{-2},
\end{aligned} \tag{A.49}$$

where C_s is a positive constant.

We define $\Delta_t := \left(h\left(\boldsymbol{x}_{t,a_t^*}\right) - h\left(\boldsymbol{x}_{t,a_t}\right)\right)\mathbb{1}\left(\mathcal{E}_t^\mu\right)$, and

$$\begin{aligned}
X_t &:= \Delta_t - \Big(C_1\tau C_s\sqrt{L}\min\left\{s_{t,a_t}, 1\right\} \\
&\qquad + C_2\tau|\sigma_{t,a_t} - s_{t,a_t}| + 4\epsilon(m) + 2t^{-2}\Big), \\
Y_t &= \sum_{i=1}^{t} X_i.
\end{aligned} \tag{A.50}$$

Then, with probability at least $1-\delta$, we have

$$\begin{aligned}
|X_t| &\leq |\Delta_t| + C_1\tau C_s\sqrt{L}\min\left\{s_{t,a_t}, 1\right\} \\
&\quad + C_2\tau|\sigma_{t,a_t} - s_{t,a_t}| + 4\epsilon(m) + 2t^{-2} \\
&\leq 2 + 2t^{-2} + C_1\tau C_s\sqrt{L} + C_2\tau|\sigma_{t,a_t} - s_{t,a_t}| + 4\epsilon(m) \\
&\leq 4 + C_1\tau C_s\sqrt{L} + C_2\tau\epsilon_f(m) + 4\epsilon(m)
\end{aligned} \tag{A.51}$$

where the first inequality is due to $|x+y| \leq |x| + |y|$, the second inequality uses the fact that $h \leq 1$, and the second inequality is due to $\epsilon_f(m) := \max_{t\in[T]}|\sigma_{t,a_t} - s_{t,a_t}|$.

From Lemma A.10, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} X_t = Y_T - Y_0 \\
&\leq \left(4 + C_1\tau C_s\sqrt{L} + C_2\tau\epsilon_f(m) + 4\epsilon(m)\right)\sqrt{2\log(1/\delta)T},
\end{aligned} \tag{A.52}$$

We then can derive that

$$\begin{aligned}
\sum_{t=1}^{T}\Delta_t &\leq \left(4 + C_1\tau C_s\sqrt{L} + C_2\tau\epsilon_f(m) + 4\epsilon(m)\right)\sqrt{2\log(1/\delta)T} \\
&\quad + \sum_{t=1}^{T}\Big(C_1\tau C_s\sqrt{L}\min\left\{s_{t,a_t}, 1\right\} + C_2\tau|\sigma_{t,a_t} - s_{t,a_t}| \\
&\qquad + 4\epsilon(m) + 2t^{-2}\Big) \\
&\leq 4T\epsilon(m) + \pi^2/3 + C_2\tau T\epsilon_f(m) \\
&\quad + C_1\tau C_s\sqrt{L}\sum_{t=1}^{T}\min\left\{s_{t,a_t}, 1\right\} \\
&\quad + \left(4 + C_1\tau C_s\sqrt{L} + C_2\tau\epsilon_f(m) + 4\epsilon(m)\right)\sqrt{2\log(1/\delta)T},
\end{aligned} \tag{A.53}$$

where the first inequality is due to (A.50) and (A.52), and the lase inequality is due to the definition of $\epsilon_f(m)$ and $\sum_{i=1}^{\infty} t^{-2} = \pi^2/6$. ∎

A.3.4 Proof of Lemma A.6

We first present our Lemma A.11.

LEMMA A.11. *(Lemma B.4 in [31]). Define* $\psi\left(\mathbf{a}, \mathbf{a}_1, \cdots, \mathbf{a}_{t-1}\right)$ *as*

$$\psi\left(\mathbf{a}, \mathbf{a}_1, \cdots, \mathbf{a}_{t-1}\right) := \sqrt{\mathbf{a}^\top\left(\gamma\mathbf{I} + \sum_{i=1}^{t-1}\mathbf{a}_i\mathbf{a}_i^\top\right)^{-1}\mathbf{a}}. \tag{A.54}$$

If $\|\boldsymbol{a}\|_2 \leq C_{Lip}\sqrt{L}$, *then* $\psi\left(\mathbf{a}, \mathbf{a}_1, \cdots, \mathbf{a}_{t-1}\right)$ *is Lipschitz continued, where* $\|\nabla_{\mathbf{a}}\psi\|_2 \leq 1/\sqrt{\gamma}$ *and* $\|\nabla_{\mathbf{a}_i}\psi\|_2 \leq C_{Lip}^2 L/\sqrt{\gamma}$.

Then with the lemma above, we derive our Lemma A.6.

Proof. It is easy to verify that

$$\begin{aligned}
\sigma_{t,k} &= \psi\Bigg(\frac{\tilde{\boldsymbol{f}}_{\text{repr}}(\boldsymbol{x}_{t,k}; \boldsymbol{w}_{t,\text{repr}})}{\sqrt{m}}, \\
&\qquad \frac{\tilde{\boldsymbol{f}}_{\text{repr}}(\boldsymbol{x}_{1,a_1}; \boldsymbol{w}_{2,\text{repr}})}{\sqrt{m}}, \cdots, \frac{\tilde{\boldsymbol{f}}_{\text{repr}}(\boldsymbol{x}_{t-1,a_{t-1}}; \boldsymbol{w}_{t,\text{repr}})}{\sqrt{m}}\Bigg) \\
s_{t,k} &= \psi\Bigg(\frac{\nabla_{\boldsymbol{w}} f(\boldsymbol{x}_{t,k}; \boldsymbol{w}_t)}{\sqrt{m}}, \\
&\qquad \frac{\nabla_{\boldsymbol{w}} f(\boldsymbol{x}_{1,a_1}; \boldsymbol{w}_2)}{\sqrt{m}}, \cdots, \frac{\nabla_{\boldsymbol{w}} f(\boldsymbol{x}_{t-1,a_{t-1}}; \boldsymbol{w}_t)}{\sqrt{m}}\Bigg).
\end{aligned} \tag{A.55}$$

where $\tilde{\boldsymbol{f}}_{\text{repr}}$ is expanded from $\boldsymbol{f}_{\text{repr}}$ by aligning its elements with $\text{vec}\left(\mathbf{W}_L\right)$ in $\boldsymbol{w}_t$ and adding 0 in other positions.

By Lemma A.11, Lemma A.9 and Equation A.55, we have that:

$$\begin{aligned}
|\sigma_{t,a_t} - s_{t,a_t}| &\leq \frac{1}{\sqrt{\gamma}}\left\|\frac{\tilde{\boldsymbol{f}}_{\text{repr}}(\boldsymbol{x}_{t,k}; \boldsymbol{w}_{t,\text{repr}})}{\sqrt{m}} - \frac{\nabla_{\boldsymbol{w}} f(\boldsymbol{x}_{t,k}; \boldsymbol{w}_t)}{\sqrt{m}}\right\|_2 \\
&\quad + \frac{C_s^2 L}{\sqrt{\gamma}}\sum_{i=1}^{t-1}\left\|\frac{\tilde{\boldsymbol{f}}_{\text{repr}}(\boldsymbol{x}_{i,a_i}; \boldsymbol{w}_{i+1,\text{repr}})}{\sqrt{m}} - \frac{\nabla_{\boldsymbol{w}} f(\boldsymbol{x}_{i,a_i}; \boldsymbol{w}_{i+1})}{\sqrt{m}}\right\|_2.
\end{aligned} \tag{A.56}$$

Then, we have

$$\epsilon_f(m) \leq \gamma^{-1/2} m^{-1/2}\left(T^{-1} + C_s^2 L\right) T\zeta_T \tag{A.57}$$

where $\zeta_T = \max\{\zeta_{T,1}, \zeta_{T,2}\}$ with

$$\zeta_{T,1} = \max_{t\in[T], k\in[K]}\left\|\tilde{\boldsymbol{f}}_{\text{repr}}(\boldsymbol{x}_{t,k}; \boldsymbol{w}_{t,\text{repr}}) - \nabla_{\boldsymbol{w}} f(\boldsymbol{x}_{t,k}; \boldsymbol{w}_t)\right\|_2, \tag{A.58}$$

$$\zeta_{T,2} = \max_{t\in[T]}\left\|\tilde{\boldsymbol{f}}_{\text{repr}}(\boldsymbol{x}_{t,a_t}; \boldsymbol{w}_{t+1,\text{repr}}) - \nabla_{\boldsymbol{w}} f(\boldsymbol{x}_{t,a_t}; \boldsymbol{w}_{t+1})\right\|_2. \tag{A.59}$$

By Assumption 5.1, we know that $\zeta_T = o(\sqrt{m})$. ∎

B Details of Algorithm Variants

In this section, we introduce two variants of PlugTS in detail, Practical PlugTS (PlugTS-P) and Generalized PlugTS (PlugTS-G).

In PlugTS-P, we do not maintain the historical context information in the matrix $\boldsymbol{M}$ explicitly, which further reduces computational overhead. Similar with Algorithm 2, we formally present PlugTS-P in Algorithm 3.

In PlugTS-G, we utilize gradients of parameters for multiple layers instead of only the gradient with respect to parameters of the last layer. Note that $\nabla_{\text{vec}(\mathbf{W}_L)} f(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{f}_{\text{repr}}(\boldsymbol{x}; \boldsymbol{w}_{\text{repr}})$ for the neural network defined in (A.1). We then replace $\nabla_{\text{vec}(\mathbf{W}_L)} f(\boldsymbol{x}; \boldsymbol{w})$ with $\nabla_{\boldsymbol{w}_{\text{subset}}} f(\boldsymbol{x}; \boldsymbol{w})$, where $\boldsymbol{w}_{\text{subset}} = \left(\text{vec}\left(\mathbf{W}_{s_1}\right); \cdots; \text{vec}\left(\mathbf{W}_{s_l}\right)\right)$, $s_1, \cdots, s_l \in [L]$, i.e., $\boldsymbol{w}_{\text{subset}}$ is the collection vector of parameters for some layers.

PlugTS-G will reduce to PlugTS when $\boldsymbol{w}_{\text{subset}} = \text{vec}\left(\mathbf{W}_L\right)$, and reduce to NeuralTS [31] when $\boldsymbol{w}_{\text{subset}} = \boldsymbol{w}$. We present PlugTS-G in Algorithm 4 with denoting $\boldsymbol{f}_{\text{subset}}(\boldsymbol{x}; \boldsymbol{w}_{\text{subset}}) := \nabla_{\boldsymbol{w}_{\text{subset}}} f(\boldsymbol{x}; \boldsymbol{w})$.

We then generalize Assumption 5.1 to fit PlugTS-G.

ASSUMPTION B.1. *Define that* $\tilde{\boldsymbol{f}}_{subset}$ *is expanded from* $\boldsymbol{f}_{subset}$ *by aligning its elements with* $\boldsymbol{w}_{t,subset}$ *in* $\boldsymbol{w}_t$ *and padding with* 0

Algorithm 3: PlugTS-P (Practical PlugTS)

Input: Initialization parameter of neural network $\boldsymbol{w}_1$, hyper-parameter for exploration strength ν, network width m, regularization parameter $\lambda > 0$, positive definite matrix $\boldsymbol{M}_1 = \lambda \boldsymbol{I}$

for $t = 1, \ldots, T$ **do**
  Receive context vectors $\{\boldsymbol{x}_{t,k}\}_{k=1}^K$;
  for $k = 1, \ldots, K$ **do**
    Sample $\hat{\theta}_{t,\boldsymbol{x}_{t,k}} \sim \mathcal{N}(f(\boldsymbol{x}_{t,k}; \boldsymbol{w}_t)), \nu \| \boldsymbol{f}_{\text{repr}}(\boldsymbol{x}_{t,k}; \boldsymbol{w}_{t,\text{repr}}) \|^2_{\boldsymbol{M}_t^{-1}})$;
  end
  $a_t \leftarrow \arg\max_{a \in [K]} \hat{\theta}_{t,\boldsymbol{x}_{t,a}}$;
  Apply a_t and observe r_{t,a_t};
  $\boldsymbol{w}_{t+1} \leftarrow \arg\min_{\boldsymbol{w}} \sum_{i=1}^t (f(\boldsymbol{x}_{t,a_t}) - r_{t,a_t})^2/2 + \lambda \|\boldsymbol{w} - \boldsymbol{w}_1\|^2/2$;
  $\boldsymbol{M}_{t+1} \leftarrow \boldsymbol{M}_t + \boldsymbol{I}$;
end

Algorithm 4: PlugTS-G (Generalized PlugTS)

Input: Initialization parameter of neural network $\boldsymbol{w}_1$, hyper-parameter for exploration strength τ, network width m, regularization parameter $\gamma > 0$, positive definite matrix $\boldsymbol{Z}_1 = \gamma \boldsymbol{I}$

for $t = 1, \ldots, T$ **do**
  Receive context vectors $\{\boldsymbol{x}_{t,k}\}_{k=1}^K$;
  for $k = 1, \ldots, K$ **do**
    Sample $\hat{\theta}_{t,\boldsymbol{x}_{t,k}} \sim \mathcal{N}(f(\boldsymbol{x}_{t,k}; \boldsymbol{w}_t), \tau^2 \gamma \| \boldsymbol{f}_{\text{subset}}(\boldsymbol{x}_{t,k}; \boldsymbol{w}_{t,\text{subset}}) \|^2_{\boldsymbol{Z}_t^{-1}/m})$;
  end
  $a_t \leftarrow \arg\max_{a \in [K]} \hat{\theta}_{t,\boldsymbol{x}_{t,a}}$;
  Apply a_t and observe r_{t,a_t};
  $\boldsymbol{w}_{t+1} \leftarrow \arg\min_{\boldsymbol{w}} \sum_{i=1}^t (f(\boldsymbol{x}_{t,a_t}) - r_{t,a_t})^2/2 + m\gamma \|\boldsymbol{w} - \boldsymbol{w}_1\|^2/2$;
  $\boldsymbol{Z}_{t+1} \leftarrow \boldsymbol{Z}_t + \boldsymbol{f}_{\text{subset}}(\boldsymbol{x}_{t,a_t}; \boldsymbol{w}_{t+1,\text{subset}}) \boldsymbol{f}_{\text{subset}}(\boldsymbol{x}_{t,a_t}; \boldsymbol{w}_{t+1,\text{subset}})^\top / m$;
end

Table 4: Total reward averaged over 8 repeated experiments.

	PlugTS	PlugTS-G-2	PlugTS-G-3	NeuralTS
Yahoo!R6B	52381.8 ± 540.2	52346.2 ± 660.4	52237.0 ± 602.9	52235.1 ± 1809.3

for the other elements. We assume it satisfies

(B.60) $\|\tilde{\boldsymbol{f}}_{subset}(\boldsymbol{x}_{t,k}; \boldsymbol{w}_{t,subset}) - \nabla_{\boldsymbol{w}} f(\boldsymbol{x}_{t,k}; \boldsymbol{w}_t)\| = o(\sqrt{m})$.

Under Assumption B.1, it is easy to verify that the proof in Appendix A.2 still holds with replacing $\boldsymbol{f}_{\text{repr}}(\boldsymbol{x}; \boldsymbol{w}_{\text{repr}})$ with $\boldsymbol{f}_{\text{subset}}(\boldsymbol{x}; \boldsymbol{w}_{\text{subset}})$. Therefore, PlugTS-G also achieves the $\mathcal{O}(\sqrt{T})$ regret bound.

Moreover, we have realized two versions of PlugTS-G, named PlugTS-G-2 and PlugTS-G-3, where PlugTS-G-2 uses the gradients of the last two layers and PlugTS-G-3 uses the gradients of the last three layers. The experiment setting are the same as Section 6.2 and the results over 8 repeated experiments on the Yahoo!R6B dataset are reported in Table 4. The results show that there is no significant difference in the performance of the four algorithms, which furthermore verifies our theoretical finding that the full gradient is not necessary for the sublinear regret bound.

Algorithm 5: Online Recommender System Simulator

Input: randomized impression dataset $\mathcal{D}$, initialization dataset $\mathcal{D}_{\text{init}}$, selecting strategy $s : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$ and corresponding neural network f, size of partition $N > 0$

$\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{init}}$;
Train f with $\mathcal{D}_{\text{train}}$;
$r_{\text{all}} \leftarrow 0$ // `r_all:cumulative reward (number of clicks);`
for *each N-size log entries partition* $\mathcal{D}_{part} \subseteq \mathcal{D}$ **do**
  Initialize $\mathcal{D}_{\text{simulator}} \leftarrow \emptyset$;
  `// x:user feature; a:displayed item; r:binary label;`
  `A:candidate set` **for** $(x, a, r, A) \in \mathcal{D}_{part}$ **do**
    if $\arg\max_{a' \in A} s(x, a') == a$ **then**
      $\mathcal{D}_{\text{simulator}} \leftarrow \mathcal{D}_{\text{simulator}} \cup \{(x, a, r)\}$;
      $r_{\text{all}} \leftarrow r_{\text{all}} + r$;
    end
  end
  Remove the oldest $|\mathcal{D}_{\text{simulator}}|$ samples from $\mathcal{D}_{\text{train}}$;
  $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{simulator}}$;
  Train f with $\mathcal{D}_{\text{train}}$;
end
Output: r_{all}

Table 5: Total regret on classification datasets averaged over 8 repeated experiments.

	Greedy	MCDropout	NeuralTS	PlugTS	PlugTS-P
MNIST	1885.9 ± 261.8	1537.8 ± 99.9	1564.4 ± 218.2	1451.0 ± 58.2	$\mathbf{1419.1 \pm 45.8}$
Mushroom	138.0 ± 31.2	122.5 ± 15.7	119.1 ± 26.4	113.0 ± 21.4	$\mathbf{107.4 \pm 19.0}$
Statlog	204.0 ± 32.0	180.9 ± 26.8	179.1 ± 24.3	163.8 ± 41.9	$\mathbf{150.8 \pm 20.5}$
Adult	2151.2 ± 5.6	2150.6 ± 7.3	2138.8 ± 10.7	2140.4 ± 7.7	$\mathbf{2136.6 \pm 6.9}$
Covertype	2955.0 ± 37.4	2941.8 ± 21.3	2947.0 ± 55.9	2935.9 ± 47.0	$\mathbf{2925.2 \pm 41.2}$
Magic Telescope	2002.1 ± 22.1	1995.5 ± 10.2	1997.9 ± 17.6	1977.4 ± 23.2	$\mathbf{1969.9 \pm 23.9}$

Table 6: Total reward averaged over 8 repeated experiments.

	Greedy	MCDropout	NeuralTS	PlugTS	PlugTS-P
Yahoo!R6B	39632.0 ± 1122.4	47929.1 ± 476.6	52235.1 ± 1809.3	52381.8 ± 540.2	$\mathbf{53897.8 \pm 1012.1}$

Table 7: Overall statistics of classification datasets.

	Mushroom	Statlog	Adult	Covertype	Magic Telescope
# Instances	8,124	58,000	48,842	581,012	19,020
# Classes	2	7	2	7	2
# Attributes	22	9	14	54	22

C Details of Experiments

C.1 Description of Recommender Simulator The online recommender system simulator is described in Algorithm 5.

C.2 Experimental Results For MCDropout, we set the dropout probability as 0.1 and repeat the forward process three times while predicting. For NeuralTS, PlugTS and PlugTS-P, we use a grid search on $\lambda \in \{1, 0.1, 0.01\}$ and $\nu \in \{10^{-3}, 10^{-4}, 10^{-5}\}$. All experiments are repeated 8 times, and the average and standard error are reported in Table 6 and Table 5. The **Bold Faced** data is the top performance over 8 experiments.

C.3 Introduction of Classification Datasets The overall statistics of the classification datasets is shown in Table 7. In the bandit problem setting, the agent obtains reward 1 if the correct class is selected, and 0 otherwise.