Beyond Greedy Exits: Improved Early Exit Decisions for Risk Control and Reliability (Supplementary)

Anonymous Author(s)
Affiliation
Address
email

1 Source Code

We provide the links to the source code in the supplementary material, where we provide two anonymized repositories for different task implementations. A unified merged code will be publicly released soon. The source code links are: `https://anonymous.4open.science/r/UAT-59AD/README.md` and `https://anonymous.4open.science/r/UAT-BERT-5D9D/README.md`.

2 Theorem 4.1 Proof (Refer this one)

Note: The proof of Theorem 1 given in the original submission appendix has some notation issues, please refer to this proof.

Theorem 4.1. *Let $(1 - C_g^i)$ approximates the value of $p_\tau(\hat{y} = y|x, \hat{y})$ with probability $(1 - \delta_1)$. Let the risk associated with τ^* be bounded by ϵ^*. Then, for sufficiently large T and given tolerance ϵ, UAT achieves $\mathbb{P}(\hat{\mathcal{R}}(\pi) \leq \epsilon^d) \geq (1 - \delta_1)(1 - \delta')$. where $\epsilon^d = \epsilon + \epsilon^*$, $\lambda \leq \frac{\epsilon}{L}$ and δ' is a constant.*

Proof:

Before moving to prove the theorem, we will first bound the single-run regret of the UAT algorithm.

At each round $t = 1, 2, \ldots, T$, the algorithm selects an arm τ_t and receives a reward $r(\tau_t)$.

Regret Definition: We define the realized (actual) regret for a single run as:

$$R(T) = \sum_{t=1}^{T} (\mu^* - \mu_{A_t}) = \sum_{i=1}^{K} \Delta_i N_i(T),$$

where $N_i(T)$ is the number of times arm i was pulled up to time T.

The UAT Algorithm: At each time t, the UAT algorithm selects the arm

$$\tau_t = \arg\max_i \left(\hat{\mu}_i(t) + \sqrt{\frac{2 \log T}{N_i(t)}} \right),$$

where: $\hat{\mu}_i(t)$ is the empirical mean reward of arm i up to time t. $N_i(t)$ is the number of times arm i has been pulled until time t.

High-Probability Bound Using Hoeffding's Inequality: By Hoeffding's inequality, for any $n \geq 1$ and any $\epsilon > 0$:

$$\mathbb{P}\left[|\hat{\mu}_i - \mu_i| \geq \epsilon\right] \leq 2 \exp(-2n\epsilon^2).$$

Define the confidence interval:

$$\mathrm{CI}_i(t) = \left[\hat{\mu}_i(t) - \sqrt{\frac{2 \log T}{N_i(t)}}, \hat{\mu}_i(t) + \sqrt{\frac{2 \log T}{N_i(t)}} \right].$$

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

In one particular round, true reward mean for ith arm lies in this interval with probability $1 - \delta$ where $\delta = 1/T^2$

Using the union bound over all arms and times, with probability at least $1 - 1/T$ that the true reward mean lies in the interval over all the rounds.

Bounding the Number of Suboptimal Pulls: Let us bound $N_i(T)$ for any suboptimal arm i with $\Delta_i > 0$.

Suppose arm i is selected at time t. For this to happen, the UAT index for arm i must be at least as large as that for the optimal arm i^*:

$$\hat{\mu}_i(t) + \sqrt{\frac{2\log T}{N_i(t)}} \geq \hat{\mu}_{i^*}(t) + \sqrt{\frac{2\log T}{N_{i^*}(t)}}.$$

If all confidence intervals are valid (which occurs with high probability), then:

$$\mu_i + 2\sqrt{\frac{2\log T}{N_i(t)}} \geq \mu^*.$$

Rearranging,

$$\sqrt{\frac{2\log T}{N_i(t)}} \geq \frac{\Delta_i}{2} \quad \Rightarrow \quad N_i(t) \leq \frac{8\log T}{\Delta_i^2}.$$

Thus, the number of times arm i is pulled is at most:

$$N_i(T) \leq \left\lceil \frac{8\log T}{\Delta_i^2} \right\rceil + 1.$$

Total Regret Bound: Using $R(T) = \sum_{i:\Delta_i>0} \Delta_i N_i(T)$, we obtain:

$$R(T) \leq \sum_{i:\Delta_i>0} \Delta_i \left(\frac{8\log T}{\Delta_i^2} + 1 \right) = \sum_{i:\Delta_i>0} \left(\frac{8\log T}{\Delta_i} + \Delta_i \right).$$

Finally: With probability at least $1 - \mathcal{O}(1/T)$, the realized regret of UAT satisfies:

$$R(T) \leq \sum_{i:\Delta_i>0} \left(\frac{8\log T}{\Delta_i} + \Delta_i \right) = \beta(T)$$

Let $p_\tau(\hat{y}|x)$ denote the model's confidence and $p_\tau(y = \hat{y}|x, \hat{y})$ its calibrated correctness probability. The empirical risk could be written as:

$$\hat{\mathcal{R}}(\pi) \triangleq 1 - \frac{1}{T} \sum_{t=1}^{T} p_\tau(y = \hat{y}|x, \hat{y})$$

The UAT algorithm that provides a threshold $\tau_t = \pi(x_t)$ using the policy π, for simplicity, we drop the index t. With probability atleast $\delta' = 1 - \frac{1}{T}$, the single run regret of UAT algorithm can be bounded as:

$$R(T) \leq \beta(T)$$

where the reward $r(\tau) = C_\tau^i \cdot (1 - C_g^i) - \lambda \cdot i = p_\tau(\hat{y}|x) \cdot p_\tau(y = \hat{y}|x, \hat{y}) - \lambda i$ as we consider $(1 - C_g^i)$ approximates the $p_\tau(y = \hat{y}|x, \hat{y})$ with probability δ_1. Choose $\delta = (1 - \delta_1)(1 - \delta')$ then with probability δ:

$$\sum_{t=1}^{T} [p_{\tau^*}(\hat{y}|x) p_{\tau^*}(y = \hat{y}|x, \hat{y}) - p_\tau(\hat{y}|x) p_\tau(y = \hat{y}|x, \hat{y}) - \lambda(i - i^*)] \leq \beta(T) \tag{1}$$

Rearranging terms and bounding exit differences $|i - i^*| \leq L$:

$$\sum_{t=1}^{T} (p_{\tau^*}(\cdot) - p_\tau(\cdot)) \leq \beta(T) + \lambda LT \tag{2}$$

where $p_\tau(\cdot) \triangleq p_\tau(\hat{y}|x)p_\tau(y = \hat{y}|x, \hat{y})$.
Using $p_\tau(\hat{y}|x) \leq 1$:

$$\sum_{t=1}^{T} p_\tau(y = \hat{y}|x, \hat{y}) \geq \sum_{t=1}^{T} p_\tau(\cdot) \geq \sum_{t=1}^{T} p_{\tau^*}(\cdot) - \beta(T) - \lambda LT$$

Dividing by T and substituting into the risk definition:

$$\hat{\mathcal{R}}(\pi) \leq 1 - \frac{1}{T}\sum_{t=1}^{T} p_{\tau^*}(\cdot) + \frac{\beta(T)}{T} + \lambda L$$

As we assume that $\mathcal{R}^* = 1 - \frac{1}{T}\sum_{t=1}^{T} p_{\tau^*}(\cdot) \leq \epsilon_0$

$$\hat{\mathcal{R}}(\pi) \leq \epsilon^* + \frac{\beta(T)}{T} + \lambda L$$

For large T, $\frac{\beta(T)}{T} \to 0$ (since $\beta(T) = \mathcal{O}(\log T)$). To ensure $\hat{\mathcal{R}}(\pi) \leq \epsilon$:

$$\lambda L \leq \epsilon^d - \epsilon_0 \implies \lambda \leq \frac{\epsilon^d - \epsilon^*}{L} = \frac{\epsilon}{L}$$

Therefore, if we select $\lambda \leq \frac{\epsilon}{L}$, then with probability at least $\delta = (1-\delta)(1-\delta^{'})$, the empirical risk $\hat{\mathcal{R}}(\pi)$ is bounded by ϵ. This completes the proof. $\square$