NUS National University of Singapore
sea connecting the dots

BanditSpec: Adaptive Speculative Decoding via Bandit Algorithms

Yunlong Hou*[1], Fengzhuo Zhang*[1], Cunxiao Du*[2], Xuan Zhang*[3],
Jiachun Pan[1], Tianyu Pang[2], Chao Du[2], Vincent Y. F. Tan[1], Zhuoran Yang[4]
[1]National University of Singapore [2]Sea AI Lab [3]Singapore Management University [4]Yale University

SMU SINGAPORE MANAGEMENT UNIVERSITY
Yale

Main Message

Q: How can we *adaptively* select hyperparameters for speculative decoding?

A: BanditSpec views the text generation process as an *online decision-making problem*. It adaptively selects hyperparameters using *Multi-Armed Bandit* algorithms.

Training-Free with *Theoretical Guarantees* !

Scan for paper

Problem Formulation

- Accelerate the inference of LLMs while maintaining high generation quality.

Draft → Verify → Accept

$$T_{\text{spec}} = T_{\text{draft}\times L} + T_{\text{target}} + T_{\text{accept}\times n_{\text{accepted}}} \approx T_{\text{target}}$$

- Total time saved

$$T_{\text{target}} \times \mathbb{E}\Big[\tau_{\text{c}} - \tau_{\text{spec}}\Big]$$

T_{target} : the time of one forward pass of the target model.
τ_{c} : the number of canonical decoding rounds.
τ_{spec} : the number of speculative decoding rounds.

The BanditSpec Framework

Algorithm 1 Speculative Decoding with Bandits (BanditSpec)

Inputs: arm selection algorithm $\texttt{ALG}$, initial prompt $\mathrm{pt}_0 = \mathrm{pt} \in \mathcal{X}^*$, bandit configuration $\nu = (P, \mathcal{S} = \{S_i\}_{i\in[K]}, L)$.
Procedures:

$t = 0, \mathcal{H}_0 = \emptyset, I_0 = 1, x_{I_0,0} = \emptyset$.
while $\mathrm{EOS} \notin x_{I_t,t}$ **do**
  $t = t + 1$.
  Select a hyperparameter index $I_t = \texttt{ALG}(\mathcal{H}_{t-1})$.
  $x_{I_t,t} = \mathrm{SpecDecSub}(\mathrm{pt}_{t-1}, P, S_{I_t}, L)$.
  $\mathrm{pt}_t = \mathrm{concat}(\mathrm{pt}_{t-1}, x_{I_t,t})$.
  $\mathcal{H}_t = \mathrm{concat}(\mathcal{H}_{t-1}, (I_t, x_{I_t,t}))$.
end while
return $\mathrm{ST}(\texttt{ALG}, \mathrm{pt}, \nu) = t,\ \mathrm{pt}_{\mathrm{ST}(\texttt{ALG},\mathrm{pt},\nu)} = \mathrm{pt}_t$.

- **Objective:** Devise an arm selection rule $\texttt{ALG}$ to minimize the *stopping time regret*

$$\mathrm{Reg}(\texttt{ALG}, \mathrm{pt}, \nu) := \mathbb{E}\Big[\mathrm{ST}(\texttt{ALG}, \mathrm{pt}, \nu) \mid \mathrm{pt}, \nu\Big] - \mathbb{E}\Big[\mathrm{ST}(\texttt{ALG}_{i^*(\mathrm{pt},\nu)}, \mathrm{pt}, \nu) \mid \mathrm{pt}, \nu\Big]$$

- **Desired result:**

$$\mathrm{Reg}(\texttt{ALG}, \mathrm{pt}, \nu) = o\Big(\mathbb{E}[\mathrm{len}(\mathrm{pt}_{\tau_{\text{c}}})]\Big) \text{ or } o\Big(\mathbb{E}[\tau_{\text{c}}]\Big).$$

Main Results

Stationary Mean Values assumption: There exist K values $\{\mu_i\}_{i\in[K]} \subset [1, L+1]$, such that conditioned on the history $\mathcal{H}_{t-1}$ and the chosen arm I_t at time t, the expected number of the accepted tokens $\mathbb{E}[Y_{I_t,t} \mid \mathcal{H}_{t-1}, I_t] = \mu_{I_t}$.

Adversarial Mean Values assumption: Let the number of accepted tokens generated by hyperparameter S_i at time step t be $y_{i,t} = \mathrm{len}(X_{i,t})$. We assume $\{y_{i,t}\}_{i\in[K],t\in\mathbb{N}}$ is fixed by the environment before the algorithm starts.

Under the Stationary Mean Values assumption,

$\texttt{ALG} =$ UCBSpec achieves a regret upper bound as

$$\mathrm{Reg}(\texttt{ALG}, \mathrm{pt}, \nu) = O\Big(\mathrm{H}(\mathrm{pt}, \nu) \cdot L^2 \cdot \log \mathbb{E}[\mathrm{len}(\mathrm{pt}_{\tau_{\text{c}}})]\Big).$$

Additionally, for any algorithm $\texttt{ALG}$, if the acceptance token length follows the truncated geometric distribution, then

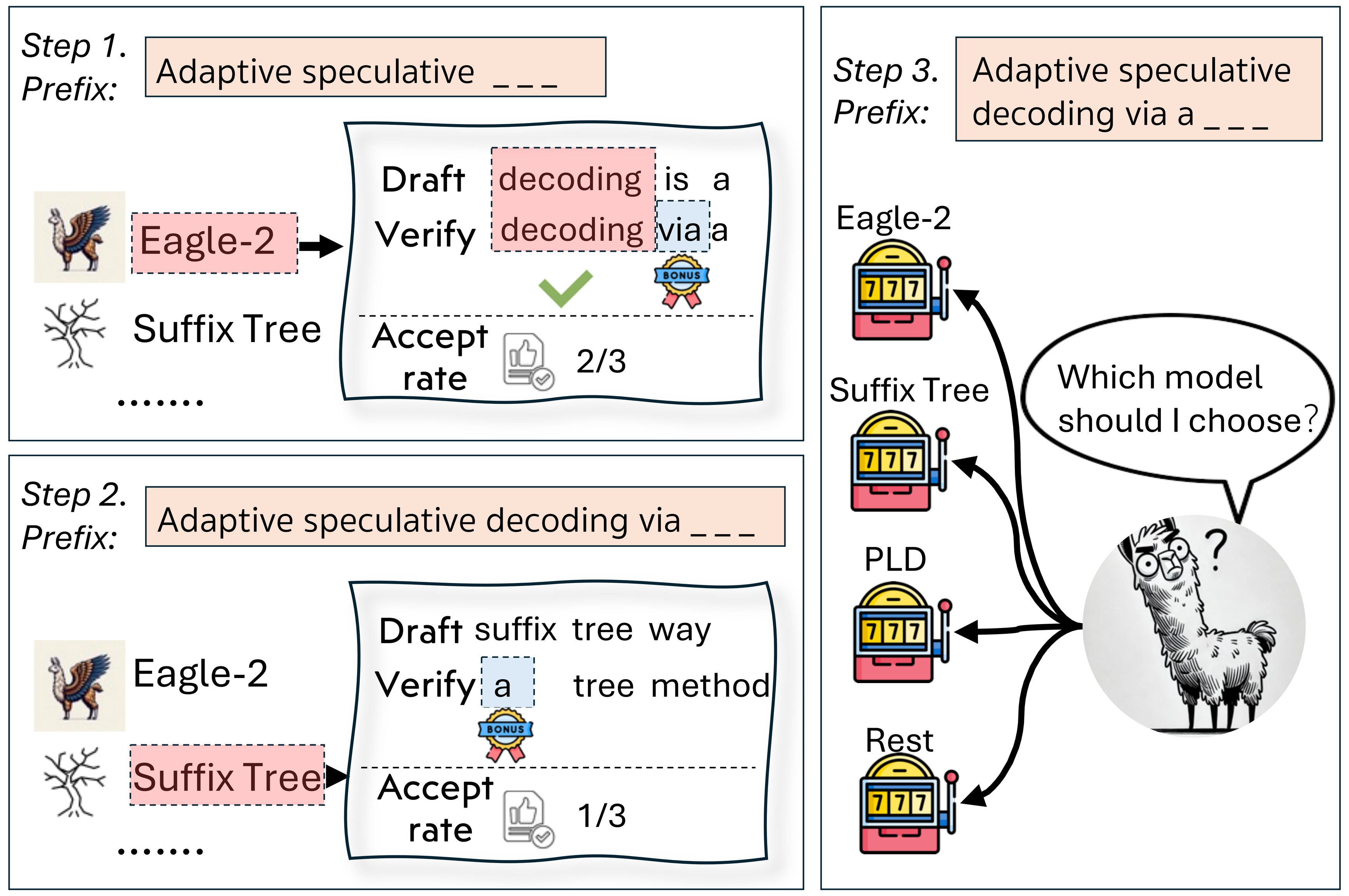
$$\liminf_{m\to\infty} \frac{\mathrm{Reg}(\texttt{ALG}, \mathrm{pt}^m, \nu)}{\log(\mathrm{len}(\mathrm{pt}^m_{\tau_{\text{c}}}))} \geq \mathrm{H}(\mathrm{pt}^m, \nu) \cdot \frac{p_{i^*}(1 - p_{i^*}^L)}{(1 - p_{i^*})}.$$

Under the Adversarial Mean Values assumption,

$\texttt{ALG} =$ EXP3Spec achieves a regret upper bound as

$$\mathrm{Reg}(\texttt{ALG}, \mathrm{pt}, \nu) \leq 2L \cdot O\left(\sqrt{\min_{i\in[K]} \mathrm{ST}(\texttt{ALG}_i) K \log K}\right).$$

Algorithm



Algorithm 2 UCBSpec

Inputs: number of hyperparameter specifications K, history $\mathcal{H}_t = \big((I_s, X_{I_s,s})\big)_{s=1}^{t}$, confidence parameter δ.
Procedures:

if $t \leq K - 1$ **then return** $I_{t+1} = t + 1$.
Compute the lengths $Y_{I_s,s} = \mathrm{len}(X_{I_s,s})$ for all $s \in [t]$.
Set the statistics $\{\hat{\mu}_{i,t}\}_{i\in[K]}, \{\mathrm{UCB}_{i,t} = \hat{\mu}_{i,t} + \mathrm{cr}_{i,t}\}_{i\in[K]}$, where

$$\mathrm{cr}_{i,t} = \frac{L}{2}\sqrt{\frac{1 + n_{i,t}}{n_{i,t}^2}\left(1 + 2\log\frac{Kt^2(1 + n_{i,t})^{\frac{1}{2}}}{\delta}\right)},$$

return index $I_{t+1} = \mathrm{argmax}_{i\in[K]} \mathrm{UCB}_{i,t}$.

Experiments

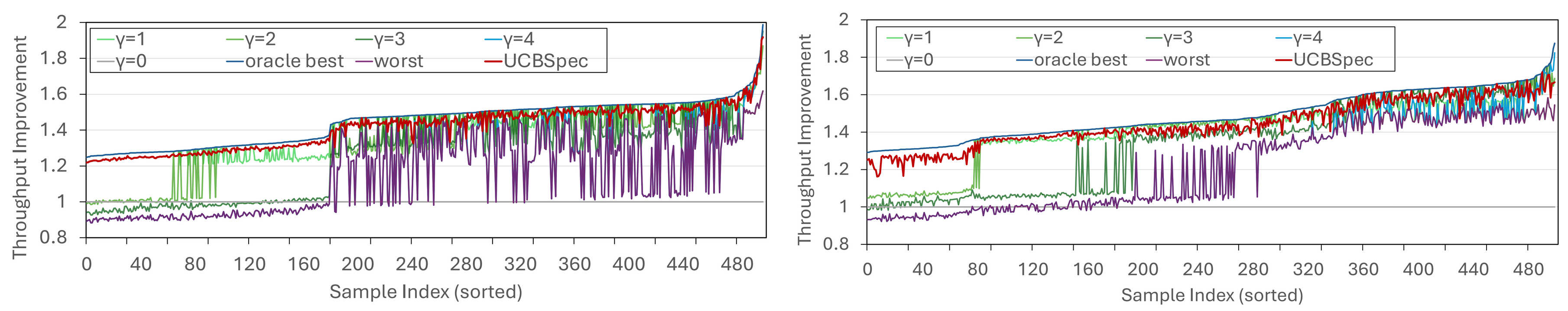


Figure 1. Throughtput comparison of speculation lengths $\gamma \in [4]$ and the canonical decoding ($\gamma = 0$).

Table 1. Empirical Comparisons measured by Mean Accepted Tokens (MAT) (↑) and Tokens/s (↑).

Methods	Spec Bench		Alpaca		Code Editor		Debug Bench	
	MAT(↑)	Tokens/s(↑)	MAT(↑)	Tokens/s(↑)	MAT(↑)	Tokens/s(↑)	MAT(↑)	Tokens/s(↑)
LLaMA3-8B-Instruct								
Vanilla	1.00	35.73	1.00	35.92	1.00	36.32	1.00	36.89
PLD	1.46	43.96	1.53	53.06	2.13	82.61	1.67	82.76
Rest	1.29	40.67	1.48	52.40	1.33	51.32	1.29	48.49
Suffix Tree	1.83	55.10	1.71	64.02	2.30	90.21	2.13	77.56
Eagle-2	3.94	98.15	4.04	110.00	4.79	128.76	**4.78**	119.12
EXP3Spec	3.65	102.10	4.23	120.38	4.36	137.29	4.50	132.25
UCBSpec	**3.98**	**105.72**	**4.35**	**125.78**	**4.83**	**138.27**	4.60	**135.34**