Improving the Trade-off Between Watermark Strength and Speculative Sampling Efficiency for Language Models

Weiqing He* Xiang Li* Li Shen Weijie Su Qi Long

University of Pennsylvania

Motivation and Background

Watermarking embeds verifiable signals into LLM outputs via pseudorandom numbers ζ, creating a statistical dependence between tokens w_t and ζ_t that enables detection.

Speculative sampling accelerates inference by using a lightweight draft model to propose tokens, verified in parallel by the target model. Efficiency depends on the acceptance rate.

The trade-off: Previous work show that combining watermarking with speculative sampling creates a seemingly unavoidable trade-off—one cannot simultaneously achieve both the highest watermark strength and the highest sampling efficiency.

🤔 **Our insight**: This impossibility relies on a **binary** definition of watermark strength. By introducing a **quantitative** measure, this trade-off may be improved.

🔍 Our Contributions

- **Quantifying watermark strength**: We define watermark strength as $\mathrm{WM}(\boldsymbol{P}_\zeta) = \mathbb{E}_\zeta[\mathrm{KL}(\boldsymbol{P}_\zeta \| \boldsymbol{P})]$, governing the decay rate of p-values in detection.
- **Characterizing the trade-off**: We formalize the Pareto frontier between watermark strength and sampling efficiency as a constrained optimization problem.
- **Improving the trade-off**: We propose pseudorandom draft-token acceptance that achieves **maximal watermark strength** while **preserving sampling efficiency**.

Watermark Strength

Definition. For a watermarking scheme sampling tokens from $\boldsymbol{P}_\zeta = \mathcal{S}(\boldsymbol{P}, \zeta)$:

$$\mathrm{WM}(\boldsymbol{P}_\zeta) = \mathbb{E}_\zeta[\mathrm{KL}(\boldsymbol{P}_\zeta \,\|\, \boldsymbol{P})] = \mathbb{E}_\zeta\left[\sum_{w\in\mathcal{V}} P_{\zeta,w} \log \frac{P_{\zeta,w}}{P_w}\right].$$

Under unbiasedness $\mathbb{E}_\zeta[\boldsymbol{P}_\zeta] = \boldsymbol{P}$, this equals the mutual information $I(w;\zeta)$.

Maximum strength. For unbiased watermarks:

$$\mathrm{WM}(\boldsymbol{P}_\zeta) = \mathrm{Ent}(\boldsymbol{P}) - \mathbb{E}_\zeta[\mathrm{Ent}(\boldsymbol{P}_\zeta)] \le \mathrm{Ent}(\boldsymbol{P}).$$

Equality holds iff $\boldsymbol{P}_\zeta$ is **degenerate** (tokens are deterministic functions of ζ).

Illustration of Trade-off Curves

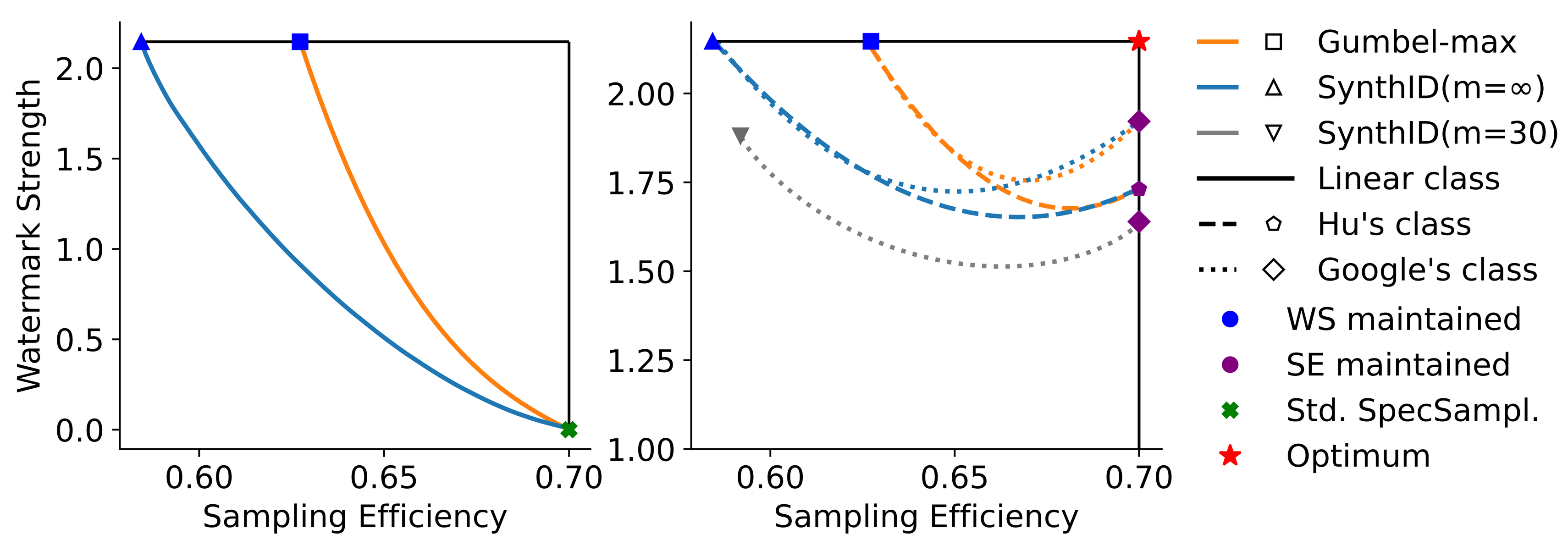

Figure 1. Trade-off curves for simulated $(\boldsymbol{Q}, \boldsymbol{P})$. **Left**: Linearly watermarked classes. **Right**: Hu's class vs. Google's class. The red star is the optimal point achieved by our algorithm.

Algorithm: Fast Watermarked Speculative Sampling with Pseudorandom Acceptance

Given: lookahead K, output length N, target model $\boldsymbol{P}$, draft model $\boldsymbol{Q}$, initial prompt $w_{1:n}$, watermarked models $\boldsymbol{Q}_{\zeta^D}$ and $\boldsymbol{P}_{\zeta^T}$, residual sampler $(\boldsymbol{P}-\boldsymbol{Q})_{+,\zeta^T}$, and pseudorandom generator G.
while $n < N$ **do**
  for $s = 1$ **to** K **do**
    Sample draft token $\tilde{w}_s \sim \boldsymbol{Q}_{\zeta^D_{n+s}}(\cdot \mid w_{1:n}, \tilde{w}_{1:s-1})$.
  end for
  In parallel: compute $K+1$ sets of target logits from draft tokens.
  for $s = 1$ **to** K **do** ▷ Sequentially try to accept each draft token.
    Compute pseudorandom $U(0,1)$ variable: $u_{n+s} \leftarrow G(\zeta^R_{n+s}) \in (0,1)$.
    if $u_{n+s} < \min\left\{1, \frac{P(\tilde{w}_s|w_{1:n})}{Q(\tilde{w}_s|w_{1:n})}\right\}$ **then**
      Accept: set $w_{n+1} \leftarrow \tilde{w}_s$; $n \leftarrow n+1$.
    else
      Reject: sample $w_{n+1} \sim (\boldsymbol{P}-\boldsymbol{Q})_{+,\zeta^T_n}(\cdot \mid w_{1:n})$; $n \leftarrow n+1$; **break**.
    end if
  end for
  if all $\tilde{w}_1, \ldots, \tilde{w}_K$ were accepted **then** ▷ Bonus step.
    Sample one extra token $w_{n+1} \sim \boldsymbol{P}_{\zeta^T_n}(\cdot \mid w_{1:n})$; $n \leftarrow n+1$.
  end if
end while

Key idea: The acceptance decision u is **pseudorandom** (not truly random), making the entire generation a deterministic function of pseudorandom variables.

Theoretical Guarantee

Theorem 4. Under the pseudorandom acceptance mechanism, with $\zeta^D, \zeta^T, \zeta^R$ independent:

- **Unbiasedness**: $\mathbb{E}_\zeta[\boldsymbol{P}'_\zeta(w)] = \boldsymbol{P}(w)$ for all $w \in \mathcal{V}$.
- **Maximum sampling efficiency**: $\mathrm{SE}(\boldsymbol{Q}_{\zeta^D}, \mathcal{A}_\zeta) = 1 - \mathrm{TV}(\boldsymbol{Q}, \boldsymbol{P})$.
- **Maximum watermark strength**: $\mathrm{WM}(\boldsymbol{P}'_\zeta) = \mathrm{Ent}(\boldsymbol{P})$.

Our algorithm simultaneously attains the **optimal point** (red star) in the trade-off figure 👈!

Experimental Results

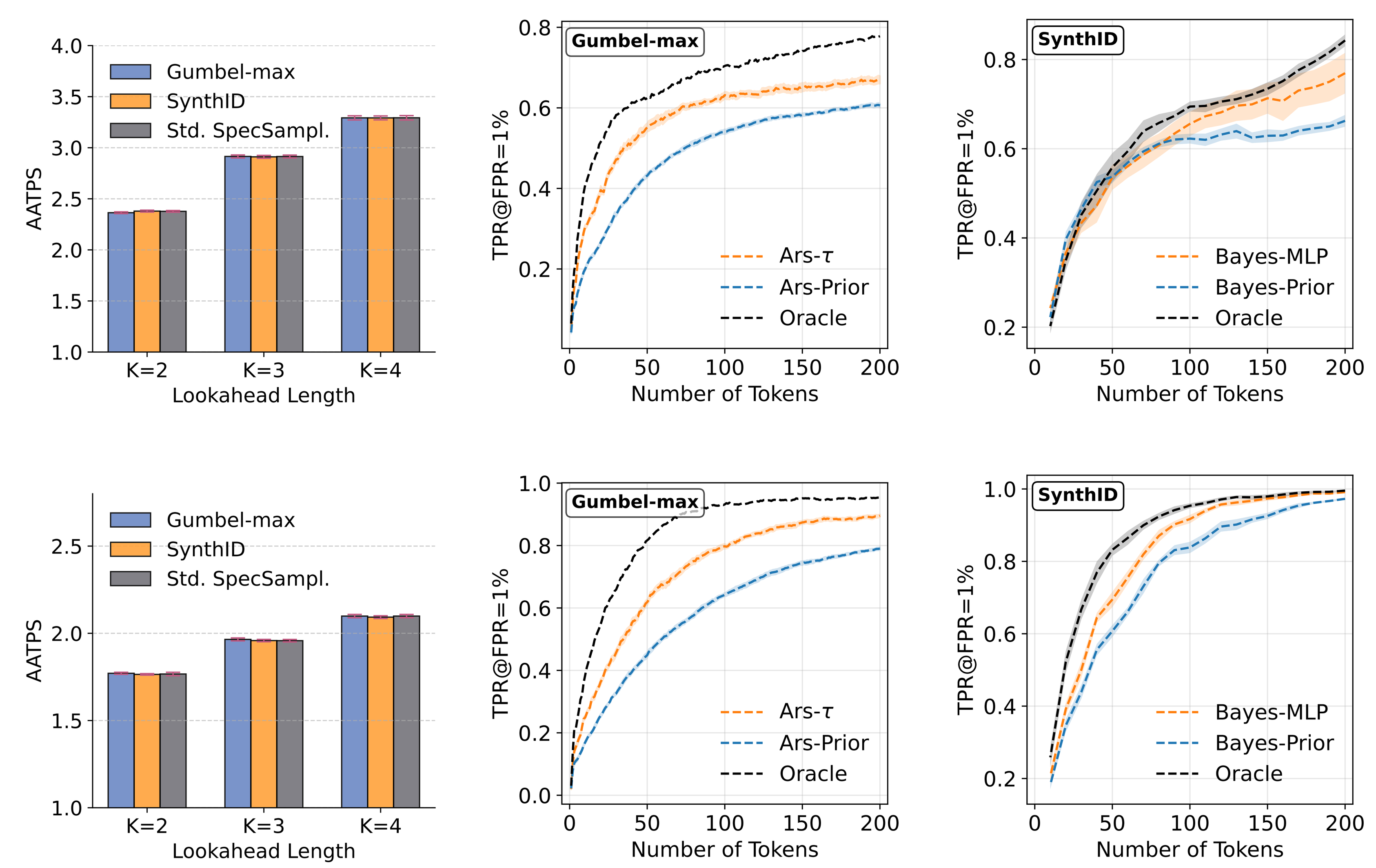

Figure 2. Results on Gemma models (upper) and Llama models (lower). **Left**: Average Accepted Tokens Per Step (AATPS). **Middle**: TPR at FPR=1% for Gumbel-max. **Right**: TPR at FPR=1% for SynthID. Orange = our method; Blue = prior-based; Black = oracle.

- **Efficiency preserved**: AATPS closely matches standard speculative sampling baseline for $K \in \{2,3,4\}$.
- **Improved detectability**: Our method (Ars-τ, Bayes-MLP) achieves higher TPR with fewer tokens.
- Our method approaches oracle performance at $\sim$200 tokens.