Optimal Robust Detection for Gumbel-Max Watermarks Under Contamination

Xiang Li[1], Feng Ruan[2], Huiyuan Wang[1], Qi Long[1], Weijie J. Su[1]

[1]University of Pennsylvania, [2]Northwestern University

Introduction

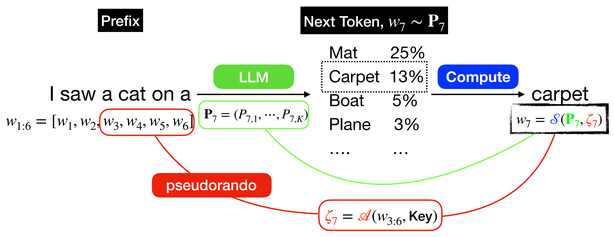
- Large language models (LLMs) have recently emerged as a groundbreaking technology for generating human-like text and other media.
- Risks arise, including misinformation, academic integrity, and data authenticity.
- Watermark serves a provable tool to detect machine-generated texts.

Autoregresive generation

- LLMs are probabilistic machines.
- Let $\mathcal{W}$ be the vocabulary and w a token therein.
- An LLM $\mathcal{M}$ generates each token sequentially by sampling from a probability distribution conditioned on previous tokens:

  $w_t \sim \boldsymbol{P}_t$ where $\boldsymbol{P}_t = \mathcal{M}(w_{1:(t-1)})$ is a dist. on $\mathcal{W}$.
- The categorical distribution $\boldsymbol{P}_t$ is referred to next-token prediction (NTP) distribution.

Watermarked generation



- Given a text $w_{1:n}$, the detector recovers $\zeta_{1:n}$ using the knowledge of $\mathcal{A}$ and $\mathtt{Key}$.
- Watermark signal is the dependence of w_t on ζ_t.
- Watermarks face significant challenges in maintaining robustness.

Questions Studied

Could we find the optimal robust detection rule even for existing watermarks?

- What robust? Which watermark?

Focus: Gumbel-max watermark [4]

- Unbiasedness: We say the decoder $\mathcal{S}$ is unbiased if for any $\boldsymbol{P}$ and $w \in \mathcal{V}$,

$$\mathbb{P}_{\zeta \sim U(\Xi)}(\mathcal{S}(\boldsymbol{P}, \zeta) = w) = P_w.$$

- Decoder for Gumbel-max watermark:

$$\mathcal{S}^{\text{gumb}}(\boldsymbol{P}, \zeta) = \arg\max_{w \in \mathcal{W}} \left\{ \frac{1}{P_w} \cdot \log U_w \right\}$$

  where $\zeta = (U_1, \ldots, U_{|\mathcal{W}|})$ with $U_k \overset{i.i.d.}{\sim} U(0,1)$.
- $\mathcal{S}^{\text{gumb}}$ is unbiased due to the Gumbel-max trick.

Previous detection framework

Find a pivotal statistic $Y_t = Y(w_t, \zeta_t)$ such that

- Under H_0, $w_t \perp \zeta_t$ so $Y_t \sim \mu_0$ for any $\boldsymbol{P}_{\text{human},t}$.
- Under H_1, $w_t = \mathcal{S}(\zeta_t, \boldsymbol{P}_t)$ so that $Y_t \sim Y(\mathcal{S}(\zeta_t, \boldsymbol{P}_t), \zeta_t)$. Hence, $Y_t | \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t}$.
- Hypothesis testing [2]:

$$H_0 : Y_t \overset{i.i.d.}{\sim} \mu_0 \text{ v.s. } H_1 : Y_t | \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t} \ \forall t \in [n].$$

- Limitation: All tokens are either human-written or LLM-generated.

Summary

- Model the robust watermark detection problem as mixture detection problem.
- GoF tests achieve the optimal detection boundary and optimal $\mathcal{P}_\Delta$-efficiency rate without any priors.
- GoF tests outperform other detection methods in detecting rates and robustness.

Goodness-of-fit (GoF) test [1]

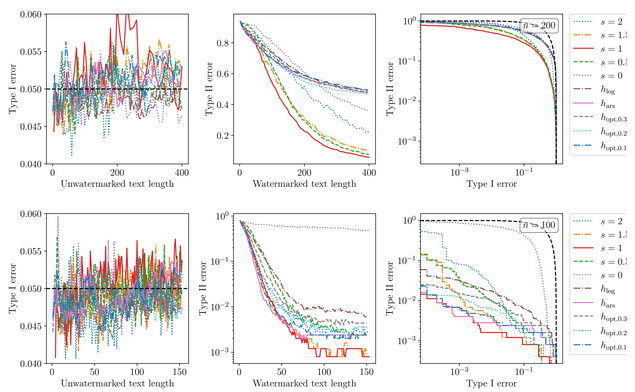
- The empirical CDF of p-values: $\mathbb{F}_n(r) = \frac{1}{n}\sum_{t=1}^n \mathbf{1}_{\mathrm{p}_t \le r}$ where $\mathrm{p}_t = 1 - Y_t$.
- Introduce a scalar convex function indexed by s:

$$\phi_s(x) = \begin{cases} x \log x - x + 1, & \text{if } s = 1, \\ \frac{1 - s + sx - x^s}{s(1-s)}, & \text{if } s \neq 0, 1, \\ -\log x + x - 1, & \text{if } s = 0. \end{cases}$$

- The ϕ_s-divergence between $\mathrm{Ber}(u)$ and $\mathrm{Ber}(v)$ is

$$K_s(u, v) = v\phi_s\left(\frac{u}{v}\right) + (1-v)\phi_s\left(\frac{1-u}{1-v}\right).$$

- For $s \in [-1, 2]$, we reject H_0 if $nS_n^+(s) = \sup_{r \in (0,1)} nK_s(\mathbb{F}_n(r), r)\mathbf{1}_{\mathbb{F}_n(r) > r}$ is larger than a given critical value $\gamma_{n,\alpha}$.



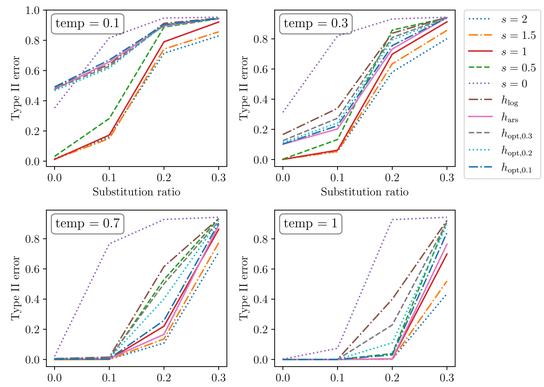
Robust detection

Example: A student modifies the LLM-generated text due to personalization or detection escape.

- Use a few tokens to compute pseudorandom numbers. For example, $\zeta_t = \mathcal{A}(w_{t-5:t-1}, \mathtt{Key})$, using the last 5 tokens.
- A modified token will turn the watermark signals in the next few 5 tokens to noise, due to $\mathcal{A}(w_{t-5:t-1}, \mathtt{Key}) \perp \mathcal{A}(w'_{t-5:t-1}, \mathtt{Key})$.
- New hypothesis testing for robustness:

$$H_0 : Y_t \sim \mu_0 \ \forall t \in [n] \quad \text{versus}$$
$$H_1^{\text{mix}} : Y_t | \boldsymbol{P}_t \sim (1-\eta_t)\mu_0 + \eta_t \mu_{1,\boldsymbol{P}_t} \ \forall t \in [n]$$

- Here $\eta_t \in \{0, 1\}$ is i.i.d. or Markovian.



Optimal detection boundary

An extreme case: sparse detection [3].

- $\mathbb{E}\,\eta_t = \varepsilon_n$ with $\varepsilon_n \asymp n^{-p}$ and $p \in (0, 1]$.
- $1 - \max_{w \in \mathcal{W}} \boldsymbol{P}_{t,w} = \Delta_n$ for all $t \in [n]$ with $\Delta_n \asymp n^{-q}$ and $q \in (0, 1)$.

If $\mathbb{E}\,\eta_t = 0$ or $\max_w \boldsymbol{P}_{t,w} = 1$, H_0 merges with H_1^{mix}.

When the robust detection is possible.

- If $q + 2p > 1$, H_0, H_1^{m} merge asym. No test is efficient.
- If $q + 2p < 1$, H_0, H_1^{m} separate asym. The likelihood-ratio test works but is not practical.

Adaptive optimality of GoF tests. If $\gamma_{n,\alpha} \asymp \log\log n$, the Type I & II errors of the GoF test $\to 0$ if $n \to \infty$ if $q + 2p < 1$ and $s \in [-1, 2]$.

Failure of sum-based tests. Consider the sum-based test that rejects H_0 if $\sum_{t=1}^n h(Y_t) \ge n \cdot \mathbb{E}_0\, h(Y) + \Theta(1) \cdot n^{\frac{1}{2}} \cdot \mathrm{poly}(\log n)$. The detection boundary for this test is $q + p = 1/2$ for all non-decreasing, $(\Delta_n, \varepsilon_n)$-free, and continuous h.

Optimal efficiency rate

$\mathcal{P}$-efficiency [2]. Defined as the rate of exponential decrease in Type II errors for a fixed significance level α and the worst-case alternative within a belief set $\mathcal{P}$.

$$\lim_{n \to \infty} \sup_{\boldsymbol{P}_t \in \mathcal{P}, \forall t \in [n]} \frac{1}{n} \log \mathbb{P}_1(S_n \le \gamma_{n,\alpha}) = -R_{\mathcal{P}}(S_n).$$

Optimal efficiency rate. Let $s \in (0, 1)$, $\varepsilon_n \equiv \varepsilon \in (0, 1]$ and $\Delta_n \equiv \Delta \in (0, 1)$.

$$R_{\mathcal{P}_\Delta}(\text{any dect. rule}) = D_{\mathrm{KL}}(\mu_0, (1-\varepsilon)\mu_0 + \varepsilon\mu_{1,\boldsymbol{P}_\Delta^\star}) = R_{\mathcal{P}_\Delta}(\text{GoF})$$

where $\boldsymbol{P}_\Delta^\star$ is the least-favorable NTP distribution whose first $\lfloor \frac{1}{1-\Delta} \rfloor$ coordinates are $1 - \Delta$.

References

[1] Leah Jager and Jon A Wellner. Goodness-of-fit tests via phi-divergences. *Annals of Statistics*, 35(5):2018–2053, 2007.

[2] Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J. Su. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *arXiv preprint arXiv:2404.01245*, 2024.

[3] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

[4] Scott Aaronson. Watermarking of large language models. `https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17`, August 2023.