Optimal Robust Detection for Gumbel-Max Watermarks Under Modification

Xiang Li

University of Pennsylvania

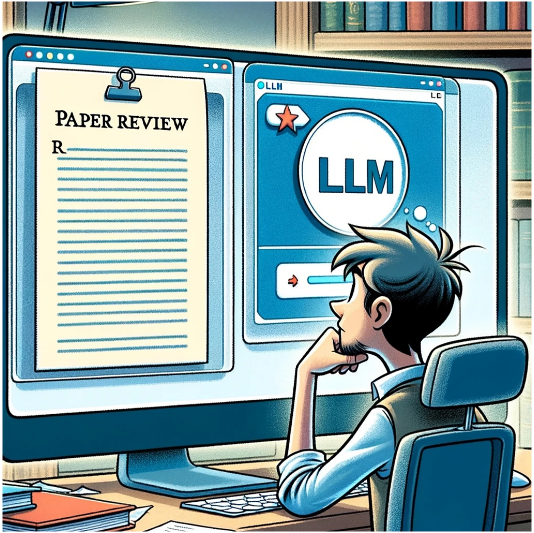
Oct. 9, 2024

Do you trust the students?

Did the student complete the homework independently, or did an LLM assist?

Peer review or LLM-assisted review?

- Liang et al. [2024]: 6.5% to 16.9% of some ML conference reviews substantially modified by LLMs.
- Is the review genuinely authored by the reviewer or significantly contributed by an LLM?

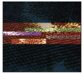


An emerging academic integrity issue

ELSEVIER

Surfaces and Interfaces

journal homepage: www.sciencedirect.com/journal/surfaces-and-interfaces



The three-dimensional porous mesh structure of Cu-based metal-organic-framework - aramid cellulose separator enhances the electrochemical performance of lithium metal anode batteries

[redacted]

ARTICLE INFO

Keywords:
Lithium metal battery
Lithium dendrites
CuMOF-ANFs separator

ABSTRACT

Lithium metal, due to its advantages of high theoretical capacity, low density and low electrochemical reaction potential, is used as a negative electrode material for batteries and brings great potential for the next generation of energy storage systems. However, the production of lithium metal dendrites makes the battery life low and poor safety, so lithium dendrites have been the biggest problem of lithium metal batteries. This study shows that the larger specific surface area and more pore structure of Cu-based metal-organic-framework - aramid cellulose (CuMOF-ANFs) composite separator can help to inhibit the formation of lithium dendrites. After 110 cycles at 1 mA/cm^2, the discharge capacity retention rate of the Li-Cu battery using the CuMOF-ANFs separator is about 96 %. Li-Li batteries can continue to maintain low hysteresis for 2000 h at the same current density. The results show that CuMOF-ANFs composite membrane can inhibit the generation of lithium dendrites and improve the cycle stability and cycle life of the battery. The three-dimensional (3D) porous mesh structure of CuMOF-ANFs separator provides a new perspective for the practical application of lithium metal battery.

1. Introduction

Certainly, here is a possible introduction for your topic:Lithium-metal batteries are promising candidates for high-energy-density rechargeable batteries due to their low electrode potentials and high

chemical stability of the separator is equally important as it ensures that the separator remains intact and does not react or degrade in the presence of the electrolyte or other battery components. A chemically stable separator helps to prevent the formation of reactive species that can further promote dendrite growth. Researchers are actively exploring

Is it possible to (reliably) detect LLM-generated text?

Applications

- Fostering original work in education and maintaining academic integrity
- Preventing fraud and deception
- Preserving the quality of data for training future AI models

Is it possible to (reliably) detect LLM-generated text?

Applications

- Fostering original work in education and maintaining academic integrity
- Preventing fraud and deception
- Preserving the quality of data for training future AI models

- Ad hoc methods leverage context, linguistic patterns, and other markers:
  - Classifiers using synthetic and human text data [GPTZero, 2023, ZeroGPT, 2023]
  - Log probability curvature [Mitchell et al., 2023, Bao et al., 2023]
  - Divergent n-gram analysis [Yang et al., 2023]

Is it possible to (reliably) detect LLM-generated text?

Applications

- Fostering original work in education and maintaining academic integrity
- Preventing fraud and deception
- Preserving the quality of data for training future AI models

- Ad hoc methods leverage context, linguistic patterns, and other markers:
  - Classifiers using synthetic and human text data [GPTZero, 2023, ZeroGPT, 2023]
  - Log probability curvature [Mitchell et al., 2023, Bao et al., 2023]
  - Divergent n-gram analysis [Yang et al., 2023]
- These methods are inaccurate, unreliable [Weber-Wulff et al., 2023], and often biased [Krishna et al., 2024, Sadasivan et al., 2023, Liang et al., 2023]

Is it possible to (reliably) detect LLM-generated text?

Applications

- Fostering original work in education and maintaining academic integrity
- Preventing fraud and deception
- Preserving the quality of data for training future AI models

- Ad hoc methods leverage context, linguistic patterns, and other markers:
  - Classifiers using synthetic and human text data [GPTZero, 2023, ZeroGPT, 2023]
  - Log probability curvature [Mitchell et al., 2023, Bao et al., 2023]
  - Divergent n-gram analysis [Yang et al., 2023]
- These methods are inaccurate, unreliable [Weber-Wulff et al., 2023], and often biased [Krishna et al., 2024, Sadasivan et al., 2023, Liang et al., 2023]
- Worse, as AI models evolve, LLM-generated text increasingly resembles human-written text!

A principled approach: Watermarking LLM-generated text

Hope: LLMs are probabilistic machines, and we control how they generate text.

A principled approach: Watermarking LLM-generated text

Hope: LLMs are probabilistic machines, and we control how they generate text.

Watermarking embeds subtle statistical signals into LLM-generated text [Kirchenbauer et al., 2023a]

- Signal: Dependence between observed text and certain hidden information for generating text.
- These signal patterns are unlikely to appear in human-written text.
- Watermarking enables provable detection of LLM-generated text.

A (very) active research area with practical importance

A Zoo of Watermarking Schemes (since January 2023):

A (very) active research area with practical importance

A Zoo of Watermarking Schemes (since January 2023):

Kirchenbauer et al. [2023a], Aaronson [2023], Kuditipudi et al. [2023], Zhao et al. [2024b], Fernandez et al. [2023], Christ et al. [2023], Wu et al. [2023], Hu et al. [2023], Kirchenbauer et al. [2023b], Zhao et al. [2024a]......

A (very) active research area with practical importance

A Zoo of Watermarking Schemes (since January 2023):
Kirchenbauer et al. [2023a], Aaronson [2023], Kuditipudi et al. [2023], Zhao et al. [2024b], Fernandez et al. [2023], Christ et al. [2023], Wu et al. [2023], Hu et al. [2023], Kirchenbauer et al. [2023b], Zhao et al. [2024a]......

- Biden AI executive order.
- OpenAI, Google, Meta, and other tech giants have pledged to watermark AI content.

Statistical challenges/opportunities in watermark research

Control/estimation of errors

- False positive rate/Type I error: Mistakenly detecting human-written text as LLM-generated.
- False negative rate/Type II error: Incorrectly classifying LLM-generated text as human-written.

Statistical challenges/opportunities in watermark research

Control/estimation of errors

- False positive rate/Type I error: Mistakenly detecting human-written text as LLM-generated.
- False negative rate/Type II error: Incorrectly classifying LLM-generated text as human-written.

Evaluation of watermarks

1. Comparing efficiency of different watermarking schemes.
2. Finding more or (most) powerful detection rules.
3. Robust watermark detection.

Our previous work

On Goal ① and ②

https://arxiv.org/pdf/2404.01245

What we did previously

- A framework unifying different watermarks.
- Efficiency notions.
- Optimal sum-based rules.

A Statistical Framework of Watermarks for Large Language Models: Pivot, Detection Efficiency and Optimal Rules

Xiang Li[*] Feng Ruan[†] Huiyuan Wang[‡] Qi Long[§] Weijie J. Su[¶]

March 28, 2024

Abstract

Since ChatGPT was introduced in November 2022, embedding (nearly) unnoticeable statistical signals into text generated by large language models (LLMs), also known as watermarking, has been used as a principled approach to provable detection of LLM-generated text from its human-written counterpart. In this paper, we introduce a general and flexible framework for reasoning about the statistical efficiency of watermarks and designing powerful detection rules. Inspired by the hypothesis testing formulation of watermark detection, our framework starts by selecting a pivotal statistic of the text and a secret key—provided by the LLM to the verifier—to control the false positive rate (the error of mistakenly detecting human-written text as LLM-generated). Next, this framework allows one to evaluate the power of watermark detection rules by obtaining a closed-form expression of the asymptotic false negative rate (the error of incorrectly classifying LLM-generated text as human-written). Our framework further reduces the problem of determining the optimal detection rule to solving a minimax optimization program. We apply this framework to two representative watermarks—one of which has been internally implemented at OpenAI—and obtain several findings that can be instrumental in guiding the practice of implementing watermarks. In particular, we derive optimal detection rules for these watermarks under our framework. These theoretically derived detection rules are demonstrated to be competitive and sometimes enjoy a higher power than existing detection approaches through numerical experiments.

This talk (On Goal ③, coming soon)

Optimal Robust Detection for Gumbel-max Watermarks Under Modification

Xiang Li* Feng Ruan† Huiyuan Wang‡ Qi Long§ Weijie J. Su¶

October 8, 2024

Abstract

This paper examines how to robustly detect statistical language watermarks when users corrupt text generated by large language models (LLMs). We develop a statistical framework for robust watermark detection by modeling the corresponding hypothesis testing problem as a mixture detection problem. We propose using a family of goodness-of-fit (GoF) tests for this purpose, showing that they achieve optimal robustness in two ways: they not only reach the optimal detection boundary when the watermark signal diminishes, but also attain the highest detection efficiency rate in cases of constant modification. In contrast, existing sum-based detection methods for Gumbel-max watermarks fail to achieve these two optimalities without additional problem-specific information. Simulations validate our theoretical guarantees, and real-data experiments demonstrate that our method achieves superior or comparable performance in maintaining watermark detectability, especially in low-temperature settings.

Outline

Outline

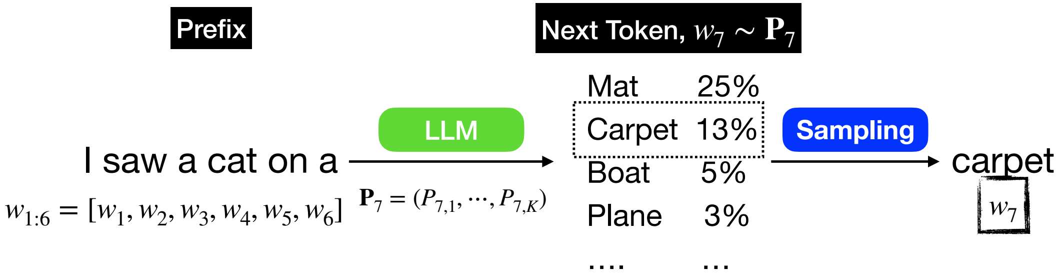
Autoregressive generation

- 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:

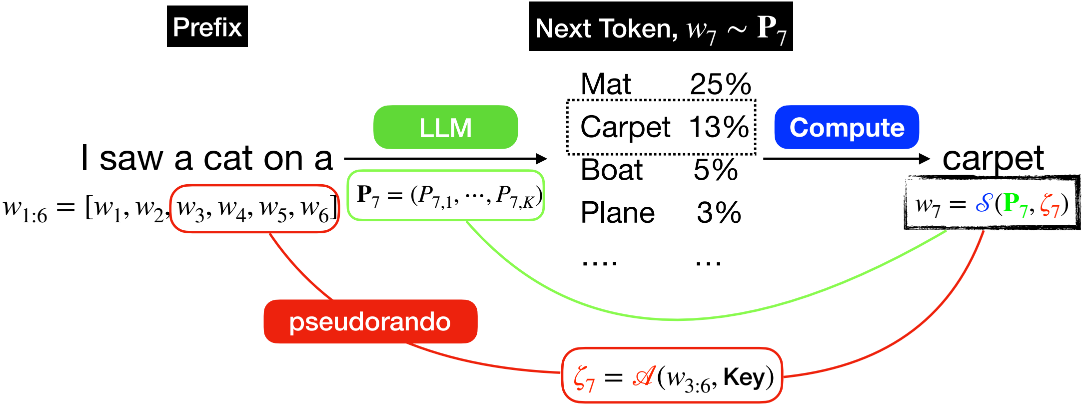
$$w_t \sim \boldsymbol{P}_t \quad \text{where} \quad \boldsymbol{P}_t = \mathcal{M}(w_{1:(t-1)}) \quad \text{is a distribution on} \quad \mathcal{W}.$$

- The categorical distribution $\boldsymbol{P}_t$ is referred to next-token prediction (NTP) distribution.

Autoregressive generation: An illustration



Watermarked generation



- Given a text $w_{1:n} = (w_1, \ldots, w_n)$, the detector recovers $\zeta_{1:n} = (\zeta_1, \ldots, \zeta_n)$ using the knowledge of $\mathcal{A}$ and Key.
- Watermark signal is the dependence of each w_t on ζ_t.

A baby watermark

- Let $\mathcal{W} = \{0, 1\}$, $\boldsymbol{P}_t = (P_{t,0}, P_{t,1})$, ζ_t be iid copies of $\mathcal{U}(0, 1)$
- Decoder

$$w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t) = \begin{cases} 0 & \text{if } \zeta_t \leq P_{t,0} \\ 1 & \text{otherwise} \end{cases}$$

A baby watermark

- Let $\mathcal{W} = \{0, 1\}$, $\boldsymbol{P}_t = (P_{t,0}, P_{t,1})$, ζ_t be iid copies of $\mathcal{U}(0, 1)$
- Decoder

$$w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t) = \begin{cases} 0 & \text{if } \zeta_t \le P_{t,0} \\ 1 & \text{otherwise} \end{cases}$$

Unbiasedness

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

for $w = 0, 1$.

A baby watermark

- Let $\mathcal{W} = \{0, 1\}$, $\boldsymbol{P}_t = (P_{t,0}, P_{t,1})$, ζ_t be iid copies of $\mathcal{U}(0, 1)$
- Decoder

$$w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t) = \begin{cases} 0 & \text{if } \zeta_t \le P_{t,0} \\ 1 & \text{otherwise} \end{cases}$$

Unbiasedness

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

for $w = 0, 1$.

Embedded signal

- If ζ_t is large, w_t is more likely to be 1 instead of 0.
- Statistic for detection:

$$\sum_{t=1}^{n} (2w_t - 1)(2\zeta_t - 1).$$

A baby watermark

- Let $\mathcal{W} = \{0, 1\}$, $\boldsymbol{P}_t = (P_{t,0}, P_{t,1})$, ζ_t be iid copies of $\mathcal{U}(0, 1)$
- Decoder

$$w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t) = \begin{cases} 0 & \text{if } \zeta_t \leq P_{t,0} \\ 1 & \text{otherwise} \end{cases}$$

Unbiasedness

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

for $w = 0, 1$.

Embedded signal

- If ζ_t is large, w_t is more likely to be 1 instead of 0.
- Statistic for detection:

$$\sum_{t=1}^{n} (2w_t - 1)(2\zeta_t - 1).$$

- Statistically, a watermark = a sampling method from a multinomial distribution $\boldsymbol{P}$.

Our focus: Gumbel-max watermark

Definition (Unbiased)

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

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

Our focus: Gumbel-max watermark

Definition (Unbiased)

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

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

Gumbel-max trick [Gumbel, 1948]

Let $\Xi = [0,1]^K$ and $\zeta = (U_1, U_2, \ldots U_K) \in \Xi$ with U_k's i.i.d. copies of $\mathcal{U}(0,1)$. The Gumbel-max trick asserts that

$$\arg\max_{w \in \mathcal{W}} \frac{\log U_w}{P_w} \sim \boldsymbol{P} \equiv (P_w)_{w \in \mathcal{W}}.$$

Our focus: Gumbel-max watermark

Definition (Unbiased)

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

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

Gumbel-max trick [Gumbel, 1948]

Let $\Xi = [0,1]^K$ and $\zeta = (U_1, U_2, \dots U_K) \in \Xi$ with U_k's i.i.d. copies of $\mathcal{U}(0,1)$. The Gumbel-max trick asserts that

$$\arg\max_{w \in \mathcal{W}} \frac{\log U_w}{P_w} \sim \boldsymbol{P} \equiv (P_w)_{w \in \mathcal{W}}.$$

Gumbel-max watermark [Aaronson, 2023]

$$\mathcal{S}^{\mathrm{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{w \in \mathcal{W}} \left\{ \frac{1}{P_w} \cdot \log U_w \right\} \quad \text{where} \quad \zeta = (U_1, \dots, U_{|\mathcal{W}|}).$$

Detection framework from [Li et al., 2024]

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

- Under H_0, $w_t \perp \zeta_t$ so that $Y_t \sim \mu_0$ regardless of $\boldsymbol{P}_{\text{human},t}$.
- 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, $Y_t | \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t}$.

Detection framework from [Li et al., 2024]

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

- Under H_0, $w_t \perp \zeta_t$ so that $Y_t \sim \mu_0$ regardless of $\boldsymbol{P}_{\text{human},t}$.
- 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, $Y_t | \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t}$.

Previous formulation in [Li et al., 2024]

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

- A score function $h : \mathbb{R} \to \mathbb{R}$ introduces a detection rule $T_h = \sum_{t=1}^n h(Y_t)$ which reject H_0 if T_h is larger than a threshold.

Detection framework from [Li et al., 2024]

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

- Under H_0, $w_t \perp \zeta_t$ so that $Y_t \sim \mu_0$ regardless of $\boldsymbol{P}_{\text{human},t}$.
- 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, $Y_t | \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t}$.

Previous formulation in [Li et al., 2024]

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

- A score function $h : \mathbb{R} \to \mathbb{R}$ introduces a detection rule $T_h = \sum_{t=1}^{n} h(Y_t)$ which reject H_0 if T_h is larger than a threshold.

Limitation

Each token in the text $w_{1:n}$ are all human-written or LLM-generated.

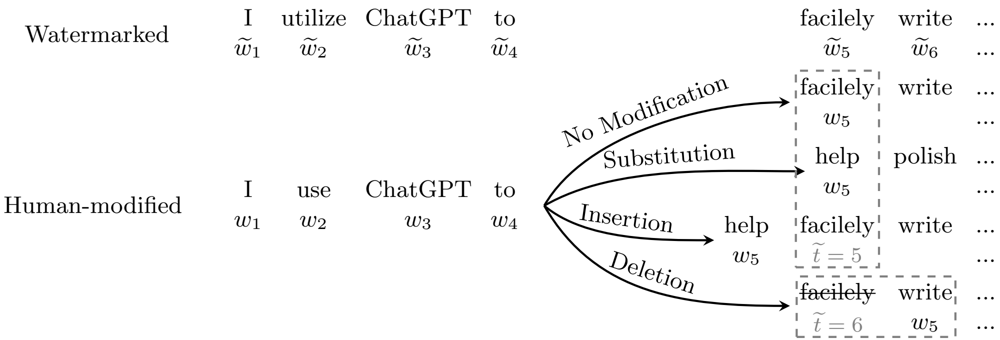
Outline

A statistical model for user modification

A student might modify the text generated from an LLM, either due to personalization or to try to escape from detection.

A statistical model for user modification

A student might modify the text generated from an LLM, either due to personalization or to try to escape from detection.



The formal procedure

Input: The watermarked text $\widetilde{w}_{1:n_0}$ generated by $\widetilde{w}_t = \mathcal{S}(\widetilde{\boldsymbol{P}}_t, \widetilde{\xi}_t)$ and $\widetilde{\boldsymbol{P}}_t = \mathcal{M}(\widetilde{w}_{1:(t-1)})$.
Initialize: $w_{1:0} = \emptyset$, $t = t_0 = 1$, and π is the distribution that makes $\mathcal{S}$ unbiased.
while the modification is not complete **do** one of the feasible operators:
    Try to determine w_t by inspecting the referenced token $\widetilde{w}_{t_0}$.
    if the user approves $\widetilde{w}_{t_0}$ **then**
        No modification: Set $w_t = \widetilde{w}_{t_0}$ and update $(t, t_0) \leftarrow (t+1, t_0+1)$.
    else if the user prefers to generate w_t themselves **then**
        Generate a new token: $w_t = \mathcal{S}(\boldsymbol{P}_t^{\mathrm{h}}, \xi_t^{\mathrm{h}})$ where $\boldsymbol{P}_t^{\mathrm{h}} = \mathcal{H}(w_{1:(t-1)})$ and $\xi_t^{\mathrm{h}} \overset{\text{i.i.d.}}{\sim} \pi$.
        Substitution: Update $(t, t_0) \leftarrow (t+1, t_0+1)$.
        Insertion: Update $(t, t_0) \leftarrow (t+1, t_0)$.
    else if the user searches for a better alternative in the watermarked text **then**
        Deletion: Update $(t, t_0) \leftarrow (t, t_0+1)$. Note that w_t remains undetermined at this stage.
    end if
end while
Return: The modified text $w_{1:t}$.

How token modification changes the distribution of Y_t?

A key fact

- $\zeta_t = \mathcal{A}(w_{t-m:t-1}, \mathrm{Key})$ uses previous m tokens and $Y_t = Y(\mathbf{w}_t, \zeta_t)$ uses the nearest $m+1$ tokens.
- If consecutive m (or $m+1$) tokens remain unchanged, the value of ζ_t (or Y_t) is preserved.

How token modification changes the distribution of Y_t?

A key fact

- $\zeta_t = \mathcal{A}(w_{t-m:t-1}, \mathrm{Key})$ uses previous m tokens and $Y_t = Y(\mathbf{w}_t, \zeta_t)$ uses the nearest $m+1$ tokens.
- If consecutive m (or $m+1$) tokens remain unchanged, the value of ζ_t (or Y_t) is preserved.

Consider the watermarked text $\widetilde{w}_{1:n_0}$. Suppose for some t and $\widetilde{t}$,

Case A If w_t is watermarked and ζ_t are preserved, i.e., $w_{(t-m):t} = \widetilde{w}_{(\widetilde{t}-m):\widetilde{t}}$,

$$Y_t = Y(w_t, \xi_t) = Y(\widetilde{w}_{\widetilde{t}}, \widetilde{\xi}_{\widetilde{t}}) = \widetilde{Y}_{\widetilde{t}} \implies Y_t \mid \widetilde{\boldsymbol{P}}_{\widetilde{t}} \sim \mu_{1,\widetilde{\boldsymbol{P}}_{\widetilde{t}}}.$$

How token modification changes the distribution of Y_t?

A key fact

- $\zeta_t = \mathcal{A}(w_{t-m:t-1}, \mathrm{Key})$ uses previous m tokens and $Y_t = Y(\mathbf{w}_t, \zeta_t)$ uses the nearest $m+1$ tokens.
- If consecutive m (or $m+1$) tokens remain unchanged, the value of ζ_t (or Y_t) is preserved.

Consider the watermarked text $\widetilde{w}_{1:n_0}$. Suppose for some t and $\widetilde{t}$,

Case A If w_t is watermarked and ζ_t are preserved, i.e., $w_{(t-m):t} = \widetilde{w}_{(\widetilde{t}-m):\widetilde{t}}$,

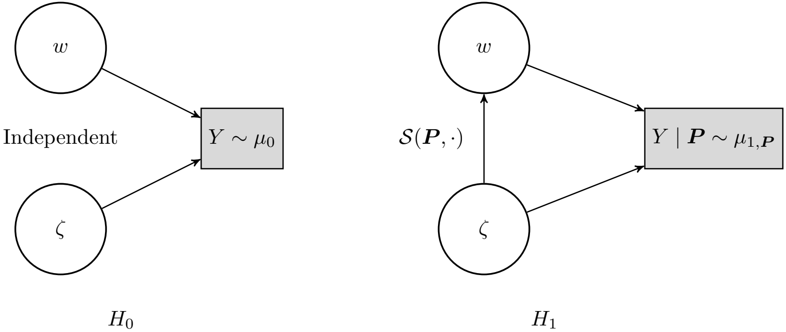
$$Y_t = Y(w_t, \xi_t) = Y(\widetilde{w}_{\widetilde{t}}, \widetilde{\xi}_{\widetilde{t}}) = \widetilde{Y}_{\widetilde{t}} \implies Y_t \mid \widetilde{\boldsymbol{P}}_{\widetilde{t}} \sim \mu_{1,\widetilde{\boldsymbol{P}}_{\widetilde{t}}}.$$

Case B If w_t is human-written, no matter whether ζ_t is preserved, $w_t \perp \zeta_t$.

Case C If w_t is watermarked but ζ_t is changed, $w_t \perp \zeta_t$.

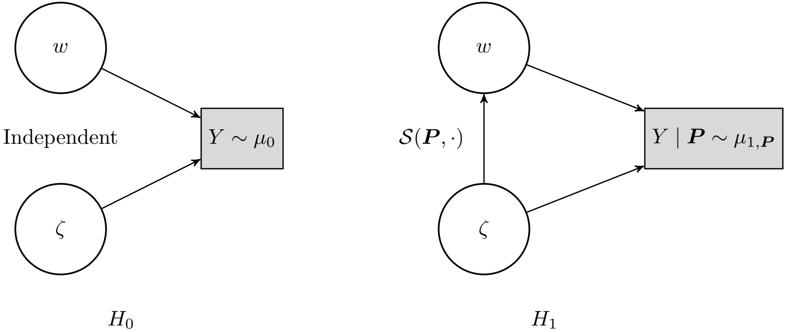
How token modification changes the distribution of Y_t?

We always have $w_t \perp \zeta_t$ or $Y_t \mid \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t}$ for some $\boldsymbol{P}_t$.



How token modification changes the distribution of Y_t?

We always have $w_t \perp \zeta_t$ or $Y_t \mid \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t}$ for some $\boldsymbol{P}_t$.



Hypothesis testing under mixtures

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

where $\eta_t \in \{0, 1\}$ is a binary random process due to user modifications.

Examples of the binary process η_t

- $\widetilde{t}$ = the longest scanned length in $\widetilde{w}_{1:n_0}$ before w_t is finalized.
- $X_t := \mathbf{1}_{w_t = \widetilde{w}_{\widetilde{t}}}$ indicate whether the user changed the latest watermarked token $\widetilde{w}_{\widetilde{t}}$ when determining w_t.
- Following the above procedure, one can show that

$$\eta_t = \mathbf{1}_{w_{(t-m):t} = \widetilde{w}_{(\widetilde{t}-m):\widetilde{t}}} = \prod_{j=(t-m)}^{t} X_j.$$

Examples of the binary process η_t

- $\widetilde{t}$ = the longest scanned length in $\widetilde{w}_{1:n_0}$ before w_t is finalized.
- $X_t := \mathbf{1}_{w_t = \widetilde{w}_{\widetilde{t}}}$ indicate whether the user changed the latest watermarked token $\widetilde{w}_{\widetilde{t}}$ when determining w_t.
- Following the above procedure, one can show that

$$\eta_t = \mathbf{1}_{w_{(t-m):t} = \widetilde{w}_{\widetilde{(t-m)}:\widetilde{t}}} = \prod_{j=(t-m)}^{t} X_j.$$

Two types of modifications

- I.i.d.: $X_i \overset{i.i.d.}{\sim} \mathrm{Ber}(a)$, $\mathbb{P}(\eta_t = 1) = (1-a)^{m+1}$.
- Markov stationary: $\mathbb{P}(\eta_t = 1) = \mathbb{E}\eta_t = \mathbb{E}\prod_{i=1}^{m+1} X_i$.
- Leave room for further modeling.

Outline

How could we solve the new problem?

Hypothesis testing under mixtures

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

where $\eta_t \in \{0, 1\}$ is a binary random process due to user modifications.

- Difficulties: We know nothing about η_t or $\boldsymbol{P}_t$.

How could we solve the new problem?

Hypothesis testing under mixtures

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

where $\eta_t \in \{0, 1\}$ is a binary random process due to user modifications.

- Difficulties: We know nothing about η_t or $\boldsymbol{P}_t$.
- Hope: We know everything about the null H_0.

How could we solve the new problem?

Hypothesis testing under mixtures

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

where $\eta_t \in \{0, 1\}$ is a binary random process due to user modifications.

- Difficulties: We know nothing about η_t or $\boldsymbol{P}_t$.
- Hope: We know everything about the null H_0.
- Focus to determine whether the observed $Y_1, \ldots, Y_n$ follows μ_0.

Goodness-of-fit (GoF) test [Jager and Wellner, 2007]

- 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.

Formal detection procedure

1: **Input:** Modified text $w_{1:n}$, hash function $\mathcal{A}$, secret key Key, pivot statistic function Y.

2: For $t = 1, 2, \ldots, n$, compute pseudorandom $\xi_t = \mathcal{A}(w_{(t-m):(t-1)}, \mathrm{Key})$.

3: For $t = 1, 2, \ldots, n$, compute the pivot statistic $Y_t = Y(w_t, \xi_t)$.

4: For $t = 1, 2, \ldots, n$, calculate the p-value as $\mathrm{p}_t = 1 - Y_t$.

5: Sort the p-values: $\mathrm{p}_{(1)} < \mathrm{p}_{(2)} < \ldots < \mathrm{p}_{(n)}$ and set $\mathrm{p}_{(n+1)} = 1$.

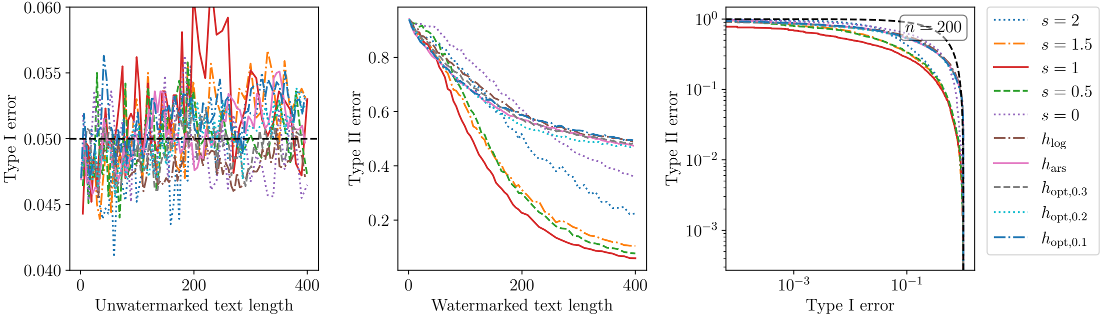
6: Compute the test statistic by

$$S_n^+(s) = \sup_t K_s(t/n, \mathrm{p}_{(t)}) \mathbf{1}_{t/n \geq \mathrm{p}_{(t)}}.$$

7: **Claim:** Text $w_{1:n}$ is modified by LLM if $nS_n^+(s)$ is too large; otherwise, it is human-written.

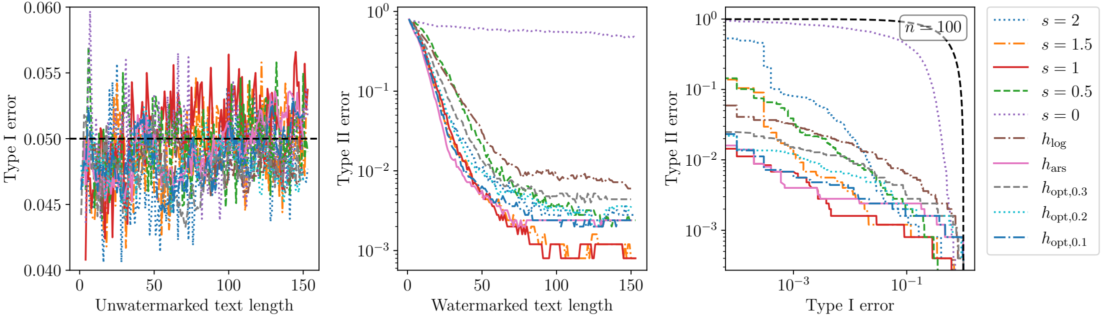
Performance on real-dataset

- OPT-1.3B [Zhang et al., 2022], newslike C4-dataset [Raffel et al., 2020].
- 0.1 (low) temperature.

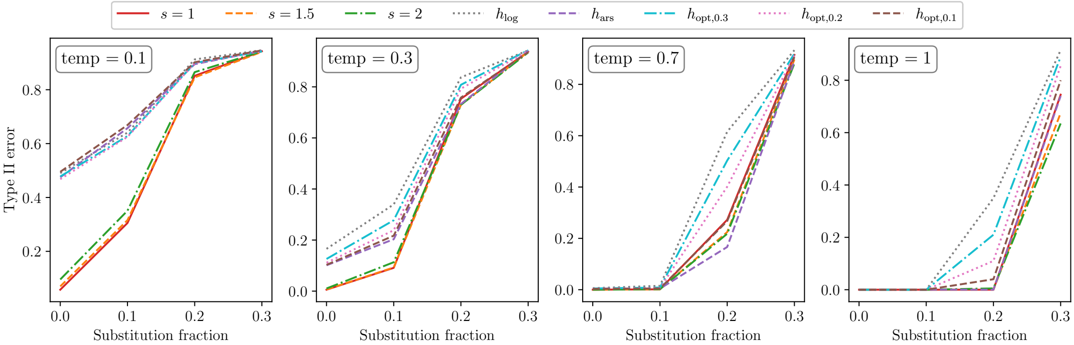


Performance on real-dataset

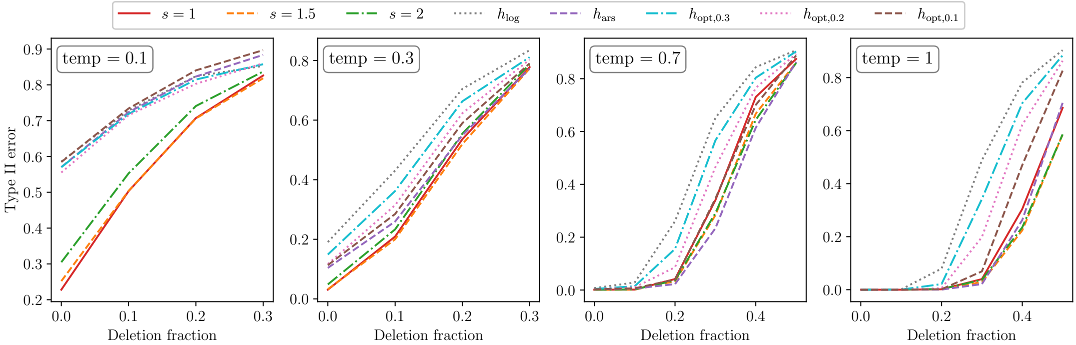
- OPT-1.3B [Zhang et al., 2022], newslike C4-dataset [Raffel et al., 2020].
- 0.7 (high) temperature.



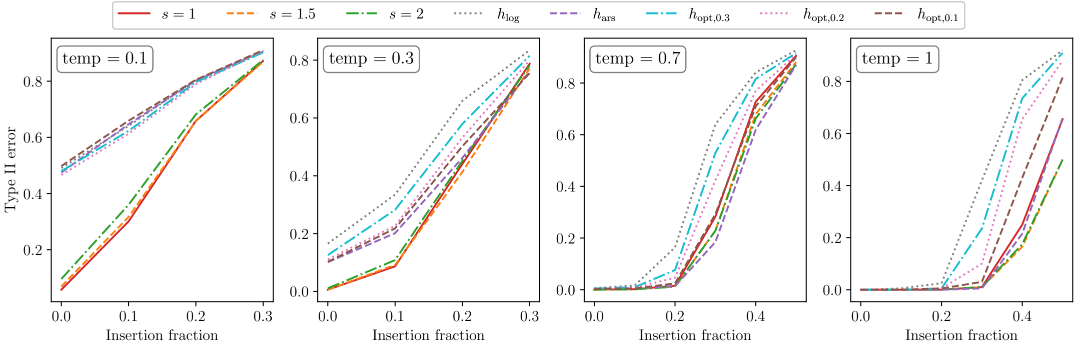
Under controllable random substitution



Under controllable random deletion



Under controllable random insertion

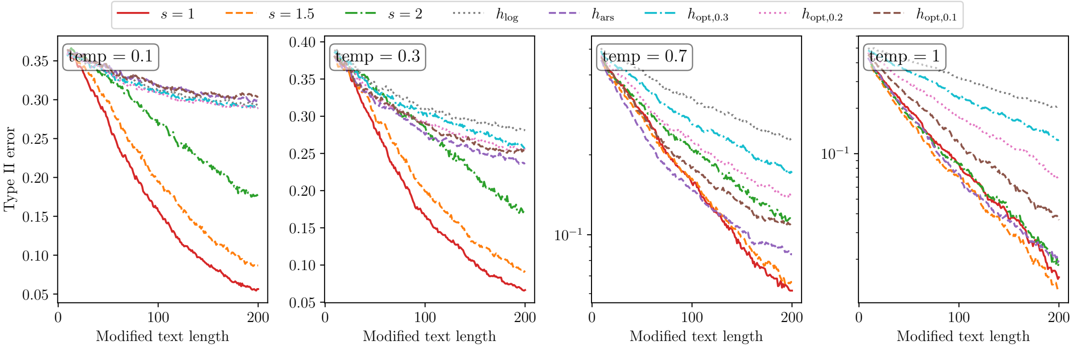


Under controllable random modifications

Task	Modification	$s=1$	$s=1.5$	$s=2$	h_{log}	h_{ars}	$h_{\mathrm{opt},0.3}$	$h_{\mathrm{opt},0.2}$	$h_{\mathrm{opt},0.1}$
Poem Recitation	Substitution	30.6	**31.96**	31.23	24.62	26.59	26.39	26.86	23.72
	Insertion	33.14	35.22	**35.92**	26.28	27.71	27.57	27.98	23.51
	Deletion	46.14	47.93	**49.42**	39.24	29.08	40.89	42.69	22.36
Poem Generation	Substitution	40.08	41.98	**42.58**	29.51	41.19	32.44	33.9	35.74
	Insertion	44.51	46.95	**48.7**	30.7	45.44	33.19	35.52	39.68
	Deletion	45.5	47.76	**48.66**	32.02	47.36	34.85	37.47	39.95

Table: The modification tolerance limits (%) for detection methods on the OPT-1.3B model.

Under non-controllable round-trip translation attack



Why the GoF test performs so well?

A question

Why the GoF test performs so well in the robust detection problem?

- We focus on the Gumbel-max watermark. Similar analysis could be paralleled to other watermarks.

Why the GoF test performs so well?

A question

Why the GoF test performs so well in the robust detection problem?

- We focus on the Gumbel-max watermark. Similar analysis could be paralleled to other watermarks.

High-level answers

The GoF test achieves optimal robustness in two senses:

1. Optimal detection boundary in a decaying watermark-signal case.
2. Optimal detection efficiency rate in a constant corruption case.

!!! The GoF test doesn't require any prior knowledge.

Outline

When the robust detection is possible?

Hypothesis testing under mixtures

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

A difficulty case

We consider an extreme case where

- $\mathbb{E}\eta_t = \varepsilon_n$ for all $t \in [n]$ 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 $q \in (0,1)$.

- Motivated by sparse detection problem [Donoho and Jin, 2004, 2015].
- If $\mathbb{E}\eta_t = 0$ or $1 - \max_{w \in \mathcal{W}} \boldsymbol{P}_{t,w} = 0$, $(1-\eta_t)\mu_0 + \eta_t \mu_{1,\boldsymbol{P}_t} = \mu_0$, i.e., H_0 merges with H_1^{mix}.

When the robust detection is possible?

Theorem

- *If $q + 2p > 1$, H_0 and H_1^{m} merge asymptotically. For any test, the sum of Type I and Type II error probabilities is 1 as $n \to \infty$.*
- *If $q + 2p < 1$, H_0 and H_1^{m} separate asymptotically. Furthermore, for the likelihood-ratio test that rejects H_0 if the log-likelihood ratio is positive, the sum of Type I and Type II error probabilities tends to 0 as $n \to \infty$.*

When the robust detection is possible?

Theorem

- *If* $q + 2p > 1$, H_0 *and* H_1^{m} *merge asymptotically. For any test, the sum of Type I and Type II error probabilities is 1 as* $n \to \infty$.
- *If* $q + 2p < 1$, H_0 *and* H_1^{m} *separate asymptotically. Furthermore, for the likelihood-ratio test that rejects* H_0 *if the log-likelihood ratio is positive, the sum of Type I and Type II error probabilities tends to 0 as* $n \to \infty$.

$\Longrightarrow$ Robust detection is impossible for small watermark signal, i.e., $q + 2p > 1$.

When the robust detection is possible?

Theorem

- *If* $q + 2p > 1$, H_0 *and* H_1^{m} *merge asymptotically. For any test, the sum of Type I and Type II error probabilities is 1 as* $n \to \infty$.
- *If* $q + 2p < 1$, H_0 *and* H_1^{m} *separate asymptotically. Furthermore, for the likelihood-ratio test that rejects* H_0 *if the log-likelihood ratio is positive, the sum of Type I and Type II error probabilities tends to 0 as* $n \to \infty$.

$\Longrightarrow$ Robust detection is impossible for small watermark signal, i.e., $q + 2p > 1$.

$\Longrightarrow$ With sufficient watermark signal, detection is possible with the likelihood-ratio test an optimal rule, i.e., $q + 2p < 1$.

!!! The likelihood-ratio test is impractical as it needs to know $\boldsymbol{P}_t$'s and ε_n.

Optimal detection boundary

Target

An ideal optimal detection method should work as long as $q + 2p < 1$ and don't requires the knowledge of $\boldsymbol{P}_t$'s and ε_n.

Optimal detection boundary

Target

An ideal optimal detection method should work as long as $q + 2p < 1$ and don't requires the knowledge of $\boldsymbol{P}_t$'s and ε_n.

Our finding

The GoF test achieves this optimal detection boundary.

Optimal detection boundary

Target

An ideal optimal detection method should work as long as $q + 2p < 1$ and don't requires the knowledge of $\boldsymbol{P}_t$'s and ε_n.

Our finding

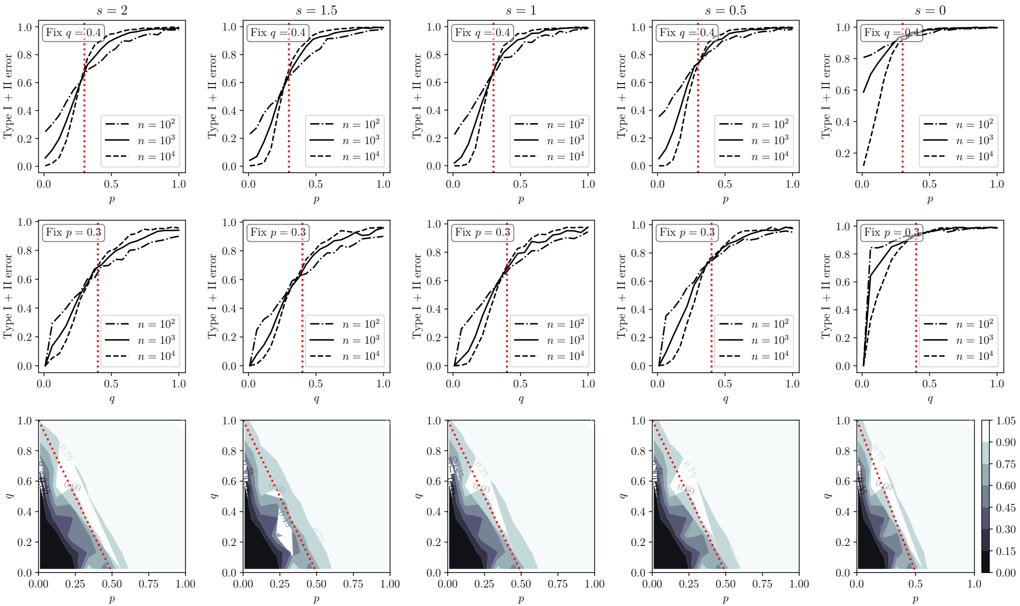
The GoF test achieves this optimal detection boundary.

Theorem (Adaptive optimality)

If the critical value $\asymp \log\log n$, *the Type I and II errors of the GoF test* $\to 0$ *if* $n \to \infty$ *as long as* $q + 2p < 1$ *and* $s \in [-1, 2]$.

- Optimal adaptivity without any prior knowledge.

Empirical detection boundaries v.s. theoretical $q + 2p = 1$



Failure of all sum-based tests

- Consider the sum-based test in the form that rejects H_0 if

$$\sum_{t=1}^{n} h(Y_t) \geq n \cdot \mathbb{E}_0 h(Y) + \Theta(1) \cdot n^{\frac{1}{2}} \cdot \mathrm{poly}(\log n).$$

Failure of all sum-based tests

- Consider the sum-based test in the form that rejects H_0 if

$$\sum_{t=1}^{n} h(Y_t) \geq n \cdot \mathbb{E}_0 h(Y) + \Theta(1) \cdot n^{\frac{1}{2}} \cdot \mathrm{poly}(\log n).$$

Theorem

The detection boundary for sum-based tests is $q + p = 1/2$ for all non-decreasing, $(\Delta_n, \varepsilon_n)$-free, and continuous h.

Failure of all sum-based tests

- Consider the sum-based test in the form that rejects H_0 if

$$\sum_{t=1}^{n} h(Y_t) \geq n \cdot \mathbb{E}_0 h(Y) + \Theta(1) \cdot n^{\frac{1}{2}} \cdot \mathrm{poly}(\log n).$$

Theorem

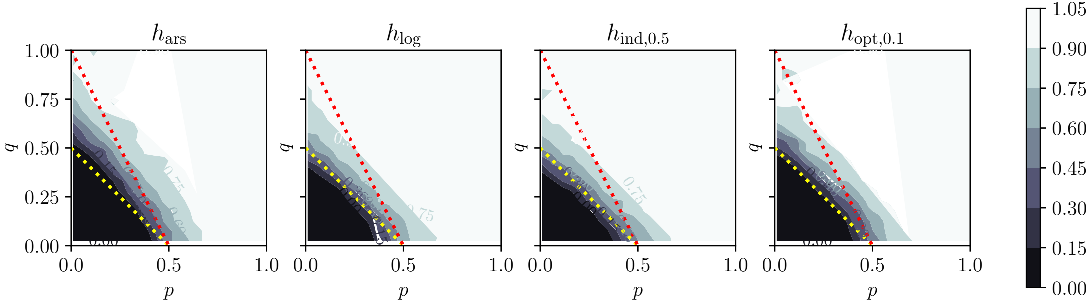
The detection boundary for sum-based tests is $q + p = 1/2$ for all non-decreasing, $(\Delta_n, \varepsilon_n)$-free, and continuous h.

Corollary

The detection boundary for the existing score function $h \in \{h_{\mathrm{ars}}, h_{\mathrm{log}}, h_{\mathrm{ind}}, h^{\star}_{\mathrm{gum},\Delta}\}$ with both $\delta, \Delta_0 \in (0,1)$ is $q + p = 1/2$.

- Sum-based tests fail to achieve adaptivity.

Failure of sum-based tests



What about constant corruption?

- The optimal detection boundary cares about the diminishing region where the watermark signal decays with the text length n.
- Practical settings meet with the constant corruption case, i.e., $\varepsilon_n \equiv \varepsilon$.
- The problem is detectable because $p = q = 0$ (within $q + 2p < 1$).

What about constant corruption?

- The optimal detection boundary cares about the diminishing region where the watermark signal decays with the text length n.
- Practical settings meet with the constant corruption case, i.e., $\varepsilon_n \equiv \varepsilon$.
- The problem is detectable because $p = q = 0$ (within $q + 2p < 1$).
- Use $\mathcal{P}$-efficiency: the rate of exponential decrease in Type II errors for a fixed significance level α and the worst-case alternative within a belief set $\mathcal{P}$.

Definition ($\mathcal{P}$-efficiency [Li et al., 2024])

Let $\gamma_{n,\alpha}$ satisfy $\mathbb{P}_0(S_n \geq \gamma_{n,\alpha}) = \alpha$ for $n \geq 1$. For a given belief set $\mathcal{P}$, we define the following limit (if exists) as the $\mathcal{P}$-efficiency of S_n and denote it by $R_{\mathcal{P}}(S_n)$:

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

What about constant corruption?

Theorem (Optimal $\mathcal{P}_\Delta$-efficiency)

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

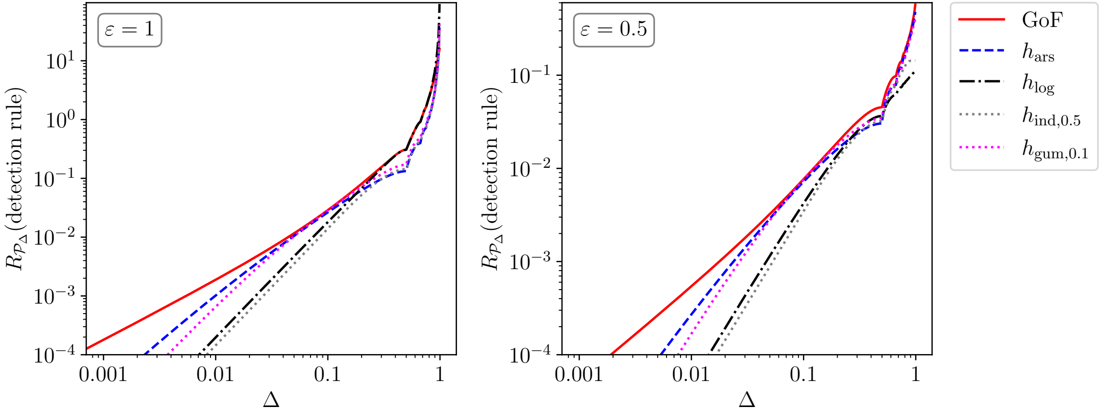
$$R_{\mathcal{P}_\Delta}(\text{any detection rule}) \leq D_{\mathrm{KL}}(\mu_0, (1-\varepsilon)\mu_0 + \varepsilon\mu_{1,\boldsymbol{P}_\Delta^\star}) \leq R_{\mathcal{P}_\Delta}(\text{GoF})$$

where $\boldsymbol{P}_\Delta^\star$ is the least-favorable NTP distribution defined by

$$\boldsymbol{P}_\Delta^\star = \Big(\underbrace{1-\Delta, \ldots, 1-\Delta}_{\lfloor \frac{1}{1-\Delta} \rfloor \text{ times}}, 1-(1-\Delta)\cdot\left\lfloor \frac{1}{1-\Delta} \right\rfloor, 0, \ldots\Big).$$

- Upper and lower bounds.
- When $\varepsilon = 1$, this rate is obtained by the sum-based test defined by $h^\star_{\mathrm{gum},\Delta}$.
- Optimal efficiency without any prior knowledge.

Theoretical $\mathcal{P}_\Delta$-efficiency comparison



Outline

Summary

- Model the robust watermark detection problem as mixture detection problem.
- GoF tests achieve the optimal detection boundary and the optimal $\mathcal{P}_{\Delta}$-efficiency without any prior knowledge.
- GoF tests outperform other detection methods in low-temperature cases and perform comparably in high-temperature cases.

Future directions

- Other optimal detection rule?
- Optimal for other watermarks?
- Estimate the non-null fraction ε.

References I

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

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*, 2023.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*, 2023.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures1. *The Annals of Statistics*, 32(3):962–994, 2004.

David L Donoho and Jiashun Jin. Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical science*, 30(1):1–25, 2015.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. Three bricks to consolidate watermarks for large language models. *arXiv preprint arXiv:2308.00113*, 2023.

GPTZero. GPTZero: More than an AI detector preserve what's human. https://gptzero.me/, 2023.

Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: A series of lectures*, volume 33. US Government Printing Office, 1948.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. *arXiv preprint arXiv:2310.10669*, 2023.

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

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, volume 202, pages 17061–17084, 2023a.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. *arXiv preprint arXiv:2306.04634*, 2023b.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

References II

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.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. GPT detectors are biased against non-native english writers. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. Monitoring AI-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*, 2024.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomá**vs** Foltỳnek, Jean Guerrero-Dib, Olumide Popoola, Petr **v**Sigut, and Lorna Waddington. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1):26, 2023.

Yihan Wu, Zhengmian Hu, Hongyang Zhang, and Heng Huang. DiPmark: A stealthy, efficient and resilient watermark for large language models. *arXiv preprint arXiv:2310.07710*, 2023.

Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. DNA-GPT: Divergent n-gram analysis for training-free detection of GPT-generated text. *arXiv preprint arXiv:2305.17359*, 2023.

ZeroGPT. ZeroGPT: Trusted GPT-4, ChatGPT and AI detector tool by ZeroGPT. `https://www.zerogpt.com/`, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. In *International Conference on Learning Representations*, 2024a. URL `https://openreview.net/forum?id=SsmT8aO45L`.

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. Permute-and-Flip: An optimally robust and watermarkable decoder for llms. *arXiv preprint arXiv:2402.05864*, 2024b.