A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules

Xiang Li
Joint work with Feng Ruan, Huiyuan Wang, Qi Long, and Weijie Su

University of Pennsylvania

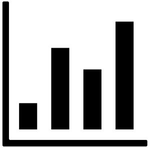
May 29, 2024

Outline

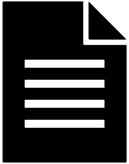
Outline

Large language models (LLMs)

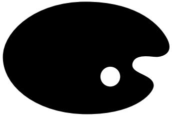
- LLMs is advanced AI systems trained to understand and generate human-like text.
- Many applications: content creation, customer service, education, code generation, healthcare, business intelligence, ...



Intelligent research assistant



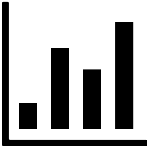
Automated document generation and editing



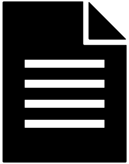
Creative content creation

Large language models (LLMs)

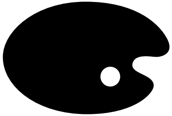
- LLMs is advanced AI systems trained to understand and generate human-like text.
- Many applications: content creation, customer service, education, code generation, healthcare, business intelligence, ...



Intelligent research assistant



Automated document generation and editing



Creative content creation

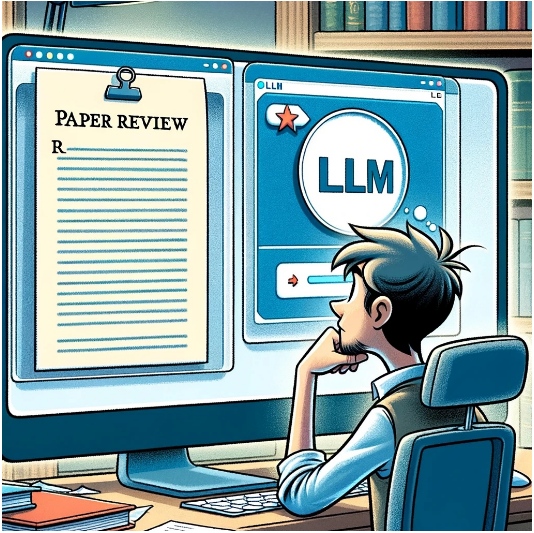
- However, there are some issues......

Academic integrity

- Did the student write this homework/paper by himself, or did an LLM lend a hand?

Peer review or LLM review?

- Liang et al. [2024] finds that between 6.5% and 16.9% reviews of some ML conferences were substantially modified by LLMs.
- Is your paper review really your own, or did an LLM lend a hand?



Document authenticity

- Is an impressively detailed patient report written by the patient or an LLM?

Central problem

Central problem

Given a text, how to determine whether it is generated by an LLM.

Central problem

Central problem

Given a text, how to determine whether it is generated by an LLM.

- Importance: misinformation, academic integrity, fair use and copyright.....

Central problem

Central problem

Given a text, how to determine whether it is generated by an LLM.

- Importance: misinformation, academic integrity, fair use and copyright.....
- Directly comparing the distribution of LLM outputs and human-written texts are neither accurate nor reliable [Weber-Wulff et al., 2023] and often biased [Krishna et al., 2024, Sadasivan et al., 2023, Liang et al., 2023].

Central problem

Central problem

Given a text, how to determine whether it is generated by an LLM.

- Importance: misinformation, academic integrity, fair use and copyright.....
- Directly comparing the distribution of LLM outputs and human-written texts are neither accurate nor reliable [Weber-Wulff et al., 2023] and often biased [Krishna et al., 2024, Sadasivan et al., 2023, Liang et al., 2023].
- More accurate detection requires us to have inner access of LLMs and thereby transition from a black-box to a white-box approach.
- Watermark is such an elegant and powerful method.

Watermark

- Watermarking enables more accurate detection of LLM-generated text by injecting subtle statistical patterns during text generation.
- Those patterns are unlikely to be replicated by a human but are constantly repeated by the watermarked LLM.
- Detecting the patterns help us detect the watermarks or equivalently the LLM-generated texts.

Preliminaries about LLM watermarks

Tokenization

- The tokenization process breaks down the text into smaller units called "tokens."
- Tokens can be words, parts of words, or even punctuation marks.[1]

GPT-3.5 & GPT-4 GPT-3 (Legacy)

```
OpenAI's large language models (sometimes referred to as GPTs) process
text using tokens, which are common sequences of characters found in a
set of text. The models learn to understand the statistical relationships
between these tokens and excel at producing the next token in a sequence
of tokens.
```

```
OpenAI's large language models (sometimes referred to as GPTs) process
text using tokens, which are common sequences of characters found in a
set of text. The models learn to understand the statistical
relationships between these tokens and excel at producing the next token
in a sequence of tokens.
```

Tokens	Characters
57	299

```
[5109, 15836, 596, 3544, 4221, 4211, 320, 57753, 14183, 311, 439, 480,
2898, 82, 8, 1920, 1495, 1701, 11460, 11, 902, 527, 4279, 24630, 315,
5885, 1766, 304, 264, 743, 315, 1495, 13, 578, 4211, 4048, 311, 3619,
279, 29564, 12135, 1990, 1521, 11460, 323, 25555, 520, 17843, 279, 1828,
4037, 304, 264, 8668, 315, 11460, 13]
```

[1] Refer to the website `https://platform.openai.com/tokenizer` for more examples of tokenization.

Autoregresive generation

- Let $\mathcal{W} = \{1, 2, \ldots, K\}$ be the vocabulary and w a token therein.

Autoregresive generation

- Let $\mathcal{W} = \{1, 2, \ldots, K\}$ be the vocabulary and w a token therein.
- The vocabulary size $K = |\mathcal{W}|$ is large and varies for difference models.
- $K = 50272$ for the OPT-1.3B model; $= 32000$ for the LLaMA-7B model.

Autoregresive generation

- Let $\mathcal{W} = \{1, 2, \ldots, K\}$ be the vocabulary and w a token therein.
- The vocabulary size $K = |\mathcal{W}|$ is large and varies for difference models.
- $K = 50272$ for the OPT-1.3B model; $= 32000$ for the LLaMA-7B model.
- 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.

Autoregresive generation

Prefix

I saw a cat on a

$w_{1:6} = [w_1, w_2, w_3, w_4, w_5, w_6]$

LLM

$\mathbf{P}_7 = (P_{7,1}, \cdots, P_{7,K})$

Next Token, $w_7 \sim \mathbf{P}_7$

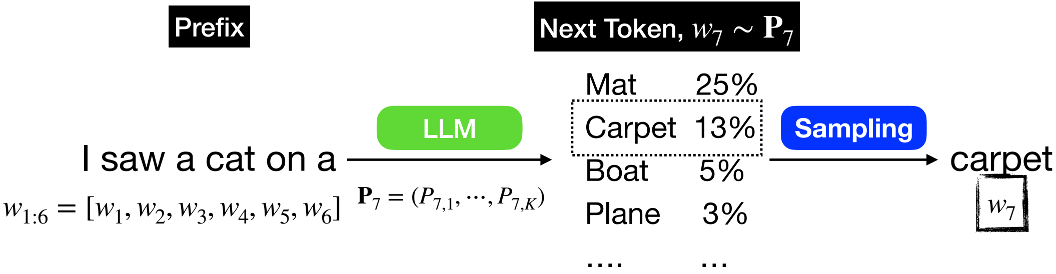
Mat	25%
Carpet	13%
Boat	5%
Plane	3%
....	...

Sampling

carpet

$\boxed{w_7}$

Autoregresive generation



- Watermarks are embedded in each next-token sampling.

Watermarked generation

Prefix

I saw a cat on a

$w_{1:6} = [w_1, w_2, w_3, w_4, w_5, w_6]$

LLM $\longrightarrow$

$\mathbf{P}_7 = (P_{7,1}, \cdots, P_{7,K})$

Next Token, $w_7 \sim \mathbf{P}_7$

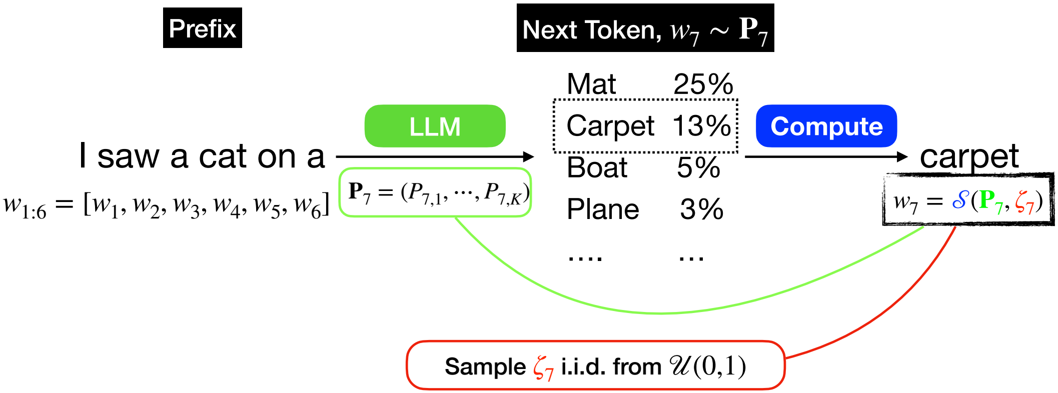
Mat	25%
Carpet	13%
Boat	5%
Plane	3%
....	...

Compute $\longrightarrow$ carpet

$\boxed{w_7 = \mathcal{S}(\mathbf{P}_7, \zeta_7)}$

Sample ζ_7 i.i.d. from $\mathcal{U}(0,1)$

Watermarked generation



- $\mathcal{S}$ is referred to as decoder.
- The watermark signal is the dependence of w_7 on ζ_7.

A baby watermark

- Let $\mathcal{W} = \{0, 1\}$, $\boldsymbol{P}_t = (P_{t,0}, P_{t,1})$, ζ_t be i.i.d. copies of $\mathcal{U}(0, 1)$.
- The decoder is

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

A baby watermark

- Let $\mathcal{W} = \{0, 1\}$, $\boldsymbol{P}_t = (P_{t,0}, P_{t,1})$, ζ_t be i.i.d. copies of $\mathcal{U}(0, 1)$.
- The decoder is

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

- This watermark is unbiased in the following sense:

Definition (Unbiased)

For any $\boldsymbol{P} = (P_{t,0}, P_{t,1})$ *and* $w \in \{0, 1\}$,

$$\mathbb{P}_{\zeta_t \sim \mathcal{U}(0,1)}(\mathcal{S}(\boldsymbol{P}_t, \zeta_t) = w_t) = P_{t,w}.$$

A baby watermark

- Let $\mathcal{W} = \{0, 1\}$, $\boldsymbol{P}_t = (P_{t,0}, P_{t,1})$, ζ_t be i.i.d. copies of $\mathcal{U}(0, 1)$.
- The decoder is

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

- This watermark is unbiased in the following sense:

Definition (Unbiased)

For any $\boldsymbol{P} = (P_{t,0}, P_{t,1})$ *and* $w \in \{0, 1\}$,

$$\mathbb{P}_{\zeta_t \sim \mathcal{U}(0,1)}(\mathcal{S}(\boldsymbol{P}_t, \zeta_t) = w_t) = P_{t,w}.$$

- If ζ_t is large, then w_t is more likely to be 1 instead of 0, and vice versa.
- Using the following statistic for detecting the watermark:

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

Watermarks are embedded in each next-token sampling: $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$

Q1 How to make these $\zeta_{1:n}$ recoverable?

Q2 Any examples of other unbiased decoder $\mathcal{S}$?

Q3 How to detect the dependence of w_t on ζ_t?

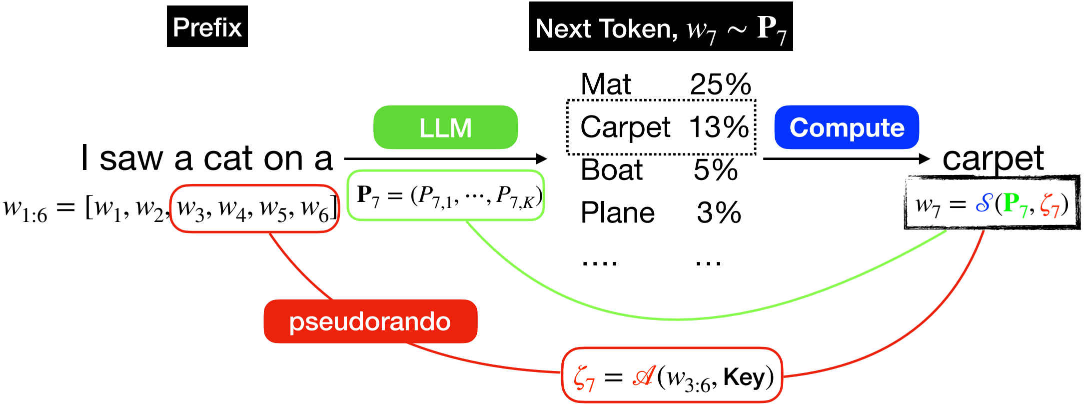
Watermarks are embedded in each next-token sampling: $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$

Q1 How to make these $\zeta_{1:n}$ recoverable?

Q2 Any examples of other unbiased decoder $\mathcal{S}$?

Q3 How to detect the dependence of w_t on ζ_t?

Q1. Use pseudorandom variables



Hash function $\mathcal{A}$

- In general, $\zeta_t = \mathcal{A}(w_{1:(t-1)}, \mathrm{Key})$ is deterministic function.
- The deterministic nature makes $\zeta_{1:n}$ recoverable so they're termed as pseudorandom.

Hash function $\mathcal{A}$

- In general, $\zeta_t = \mathcal{A}(w_{1:(t-1)}, \mathrm{Key})$ is deterministic function.
- The deterministic nature makes $\zeta_{1:n}$ recoverable so they're termed as pseudorandom.
- In analysis, we assume perfect randomness: $\zeta_t = \mathcal{A}(w_{1:(t-1)}, \mathrm{Key}) \overset{i.i.d.}{\sim} \mathcal{U}(\Xi)$.
- The distribution of ζ_t in the watermarked model is computationally indistinguishable to the unwatermarked distribution.

Hash function $\mathcal{A}$

- In general, $\zeta_t = \mathcal{A}(w_{1:(t-1)}, \mathrm{Key})$ is deterministic function.
- The deterministic nature makes $\zeta_{1:n}$ recoverable so they're termed as pseudorandom.
- In analysis, we assume perfect randomness: $\zeta_t = \mathcal{A}(w_{1:(t-1)}, \mathrm{Key}) \overset{i.i.d.}{\sim} \mathcal{U}(\Xi)$.
- The distribution of ζ_t in the watermarked model is computationally indistinguishable to the unwatermarked distribution.
- Finding a feasible $\mathcal{A}$ is a cryptography problem.

Watermarks are embedded in each next-token sampling: $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$

Q1 How to make these $\zeta_{1:n}$ recoverable? Use pseudorandom variables.

Q2 Any examples of other unbiased decoder $\mathcal{S}$?

Q3 How to detect the dependence of w_t on ζ_t?

Watermarks are embedded in each next-token sampling: $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$

Q1 How to make these $\zeta_{1:n}$ recoverable? Use pseudorandom variables.

Q2 Any examples of other unbiased decoder $\mathcal{S}$?

Q3 How to detect the dependence of w_t on ζ_t?

Q1. Decoder $\mathcal{S}^{\mathrm{gum}}$ from Gumbel-max trick

Definition (Unbiased)

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

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

Q1. Decoder $\mathcal{S}^{\text{gum}}$ from Gumbel-max trick

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

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

Q1. Decoder $\mathcal{S}^{\mathrm{gum}}$ from Gumbel-max trick

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

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

Gumbel-max watermark [Aaronson, 2023]

$$\mathcal{S}^{\mathrm{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{k \in [K]} \left\{ \frac{1}{P_k} \cdot \log U_k \right\} \quad \text{where} \quad \zeta = (U_1, \ldots, U_K).$$

Q1. Decoder $\mathcal{S}^{\text{inv}}$ from inverse transform sampling

Inverse transform sampling

The CDF of $\boldsymbol{P}$ on $\mathcal{W}$ is

$$F_{\boldsymbol{P}}(x) = \sum_{w' \in \mathcal{W}} P_{w'} \cdot \mathbf{1}_{\{w' \leq x\}}.$$

Taking as input $U \sim U(0,1)$, the generalized inverse of this CDF is defined as

$$F_{\boldsymbol{P}}^{-1}(U) = \min \left\{ x : F_{\boldsymbol{P}}(x) \geq U \right\}.$$

The inverse transform sampling asserts that $F_{\boldsymbol{P}}^{-1}(U) \sim \boldsymbol{P} \equiv (P_w)_{w \in \mathcal{W}}$.

Q1. Decoder $\mathcal{S}^{\text{inv}}$ from inverse transform sampling

Inverse transform sampling

The CDF of $\boldsymbol{P}$ on $\mathcal{W}$ is

$$F_{\boldsymbol{P}}(x) = \sum_{w' \in \mathcal{W}} P_{w'} \cdot \mathbf{1}_{\{w' \le x\}}.$$

Taking as input $U \sim U(0,1)$, the generalized inverse of this CDF is defined as

$$F_{\boldsymbol{P}}^{-1}(U) = \min \left\{ x : F_{\boldsymbol{P}}(x) \ge U \right\}.$$

The inverse transform sampling asserts that $F_{\boldsymbol{P}}^{-1}(U) \sim \boldsymbol{P} \equiv (P_w)_{w \in \mathcal{W}}$.

Inverse transform watermark [Kuditipudi et al., 2023]

$$\mathcal{S}_{\text{inv}}(\boldsymbol{P}, \zeta) := \pi \circ (F_{\pi(\boldsymbol{P})}^{-1}(U)) \ \text{ where } \ \zeta = (U, \pi), \ U \sim \mathcal{U}(\Pi), \ U \perp \pi.$$

Watermarks are embedded in each next-token sampling: $w_t = \mathcal{S}(P_t, \zeta_t)$

Q1 How to make these $\zeta_{1:n}$ recoverable? Use pseudorandom variables.

Q2 Any examples of other unbiased decoder $\mathcal{S}$? $\mathcal{S}^{\mathrm{gum}}, \mathcal{S}^{\mathrm{inv}} \ldots$

Q3 How to detect the dependence of w_t on ζ_t?

Watermarks are embedded in each next-token sampling: $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$

Q1 How to make these $\zeta_{1:n}$ recoverable? Use pseudorandom variables.

Q2 Any examples of other unbiased decoder $\mathcal{S}$? $\mathcal{S}^{\mathrm{gum}}, \mathcal{S}^{\mathrm{inv}} \ldots$

Q3 How to detect the dependence of w_t on ζ_t?

Watermarks are embedded in each next-token sampling: $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$

Q1 How to make these $\zeta_{1:n}$ recoverable? Use pseudorandom variables.

Q2 Any examples of other unbiased decoder $\mathcal{S}$? $\mathcal{S}^{\mathrm{gum}}, \mathcal{S}^{\mathrm{inv}} \ldots$

Q3 How to detect the dependence of w_t on ζ_t? A statistical framework is needed.

Outline

Review

Two-step watermarked generation

1. Generate a pseudorandom number: $\zeta_t = \mathcal{A}(w_{1:(t-1)}, \mathrm{Key}) \sim \mathcal{U}(\Xi)$.
2. Compute the next token: $w_t = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$.

Definition (Watermark)

A watermark is defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathrm{Key})$.
The watermark signal is the dependence of each w_t on ζ_t.

Definition (Unbiased)

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

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

Detection: The difficulty

Given data $(w_{1:n}, \zeta_{1:n})$,

$$H_0 : w_{1:n} \text{ is written by human}, \quad \textit{versus} \quad H_1 : w_{1:n} \text{ is watermarked by } (\mathcal{S}, \mathcal{A}, \mathrm{Key}).$$

Detection: The difficulty

Given data $(w_{1:n}, \zeta_{1:n})$,

$$H_0 : w_{1:n} \text{ is written by human}, \quad \textit{versus} \quad H_1 : w_{1:n} \text{ is watermarked by } (\mathcal{S}, \mathcal{A}, \mathrm{Key}).$$

Working Hypothesis

- Under H_0, $(w_t, \zeta_t) \mid (w_{1:(t-1)}, \zeta_{1:(t-1)}) \overset{d}{=} \boldsymbol{P}_{\text{human},t} \times \mathcal{U}(\Xi)$.
- Under H_1, $(w_t, \zeta_t) \mid (w_{1:(t-1)}, \zeta_{1:(t-1)}) \overset{d}{=} (\mathcal{S}(\zeta_t, \boldsymbol{P}_t), \zeta_t)$ and $\zeta_t \sim \mathcal{U}(\Xi)$.

Detection: The difficulty

Given data $(w_{1:n}, \zeta_{1:n})$,

$H_0 : w_{1:n}$ is written by human, *versus* $H_1 : w_{1:n}$ is watermarked by $(\mathcal{S}, \mathcal{A}, \mathrm{Key})$.

Working Hypothesis

- Under H_0, $(w_t, \zeta_t) \mid (w_{1:(t-1)}, \zeta_{1:(t-1)}) \stackrel{d}{=} \boldsymbol{P}_{\text{human},t} \times \mathcal{U}(\Xi)$.
- Under H_1, $(w_t, \zeta_t) \mid (w_{1:(t-1)}, \zeta_{1:(t-1)}) \stackrel{d}{=} (\mathcal{S}(\zeta_t, \boldsymbol{P}_t), \zeta_t)$ and $\zeta_t \sim \mathcal{U}(\Xi)$.

But, we don't know $\boldsymbol{P}_{human,1}, \ldots, \boldsymbol{P}_{human,n}$ *and other* $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_n$.

Detection: The difficulty

Given data $(w_{1:n}, \zeta_{1:n})$,

$H_0 : w_{1:n}$ is written by human, *versus* $H_1 : w_{1:n}$ is watermarked by $(\mathcal{S}, \mathcal{A}, \mathrm{Key})$.

Working Hypothesis

- Under H_0, $(w_t, \zeta_t) \mid (w_{1:(t-1)}, \zeta_{1:(t-1)}) \stackrel{d}{=} \boldsymbol{P}_{\text{human},t} \times \mathcal{U}(\Xi)$.
- Under H_1, $(w_t, \zeta_t) \mid (w_{1:(t-1)}, \zeta_{1:(t-1)}) \stackrel{d}{=} (\mathcal{S}(\zeta_t, \boldsymbol{P}_t), \zeta_t)$ and $\zeta_t \sim \mathcal{U}(\Xi)$.

But, we don't know $\boldsymbol{P}_{human,1}, \ldots, \boldsymbol{P}_{human,n}$ and other $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_n$.
Impossible to estimate these unknown $\boldsymbol{P}_t$'s.

Detection: The solution

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

- Under H_0, $Y_t \sim \mu_0$ regardless of $\boldsymbol{P}_{\text{human},t}$.
- Under H_1, $Y_t \sim Y(\mathcal{S}(\zeta_t, \boldsymbol{P}_t), \zeta_t)$. Hence, $Y_t | \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t}$.

Detection: The solution

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

- Under H_0, $Y_t \sim \mu_0$ regardless of $\boldsymbol{P}_{\text{human},t}$.
- Under H_1, $Y_t \sim Y(\mathcal{S}(\zeta_t, \boldsymbol{P}_t), \zeta_t)$. Hence, $Y_t | \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t}$.

Finall formulation

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

Good Detect distributional difference rather than independence.

Bad Use less information and sacrifice some power.

Detection: The solution

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

- Under H_0, $Y_t \sim \mu_0$ regardless of $\boldsymbol{P}_{\text{human},t}$.
- Under H_1, $Y_t \sim Y(\mathcal{S}(\zeta_t, \boldsymbol{P}_t), \zeta_t)$. Hence, $Y_t | \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t}$.

Finall formulation

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

Good Detect distributional difference rather than independence.

Bad Use less information and sacrifice some power.

- 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.

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

Let's see some detection examples.

Detection for Gumbel-max watermark

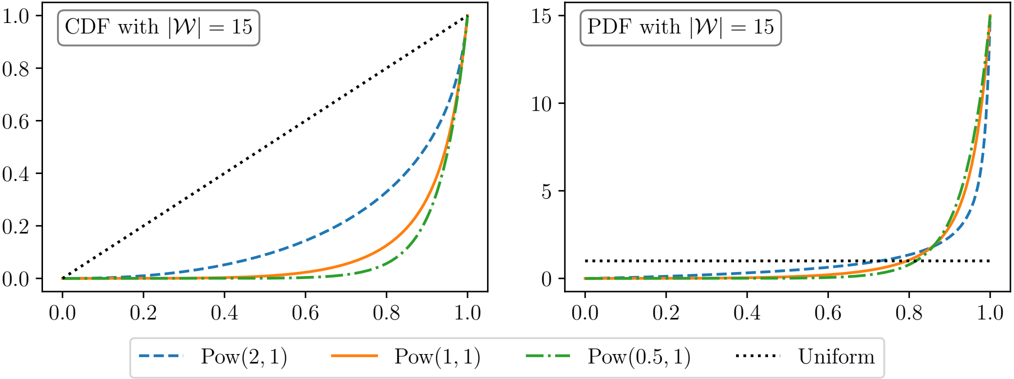
- Recall $\mathcal{S}^{\mathrm{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{k \in [K]} \frac{\log U_k}{P_k}$ and $\zeta_t = (U_{t,1}, \ldots, U_{t,K}) \in [0,1]^K$.
- The pivotal statistic is $Y_t^{\mathrm{ars}} = U_{t, w_t}$.

Detection for Gumbel-max watermark

- Recall $\mathcal{S}^{\mathrm{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{k \in [K]} \frac{\log U_k}{P_k}$ and $\zeta_t = (U_{t,1}, \ldots, U_{t,K}) \in [0,1]^K$.
- The pivotal statistic is $Y_t^{\mathrm{ars}} = U_{t,w_t}$.
- Under H_0, $Y_t^{\mathrm{ars}} \overset{i.i.d.}{\sim} \mu_0 = \mathcal{U}(0,1)$.
- Under H_1, the CDF of $\mu_{1,\boldsymbol{P}_t}$ is $\mathbb{P}_1(Y_t^{\mathrm{ars}} \le y | \boldsymbol{P}_t) = \sum_{k \in [K]} P_{t,k} y^{1/P_{t,k}}$.

Detection for Gumbel-max watermark

- Recall $\mathcal{S}^{\mathrm{gum}}(\boldsymbol{P}, \zeta) = \arg\max_{k \in [K]} \frac{\log U_k}{P_k}$ and $\zeta_t = (U_{t,1}, \ldots, U_{t,K}) \in [0,1]^K$.
- The pivotal statistic is $Y_t^{\mathrm{ars}} = U_{t,w_t}$.
- Under H_0, $Y_t^{\mathrm{ars}} \overset{i.i.d.}{\sim} \mu_0 = \mathcal{U}(0,1)$.
- Under H_1, the CDF of $\mu_{1,\boldsymbol{P}_t}$ is $\mathbb{P}_1(Y_t^{\mathrm{ars}} \leq y | \boldsymbol{P}_t) = \sum_{k \in [K]} P_{t,k} y^{1/P_{t,k}}$.



Detection for Gumbel-max watermark

Default detection for Gumbel-max watermark [Aaronson, 2023]

Aaronson proposes to reject H_0 if the following quantity is larger than a given threshold:

$$T_{h_{\mathrm{ars}}} = \sum_{t=1}^{n} h_{\mathrm{ars}}(Y_t^{\mathrm{ars}}) \quad \text{where} \quad h_{\mathrm{ars}}(y) = -\log(1-y).$$

Detection for Gumbel-max watermark

Default detection for Gumbel-max watermark [Aaronson, 2023]

Aaronson proposes to reject H_0 if the following quantity is larger than a given threshold:

$$T_{h_{\mathrm{ars}}} = \sum_{t=1}^{n} h_{\mathrm{ars}}(Y_t^{\mathrm{ars}}) \quad \text{where} \quad h_{\mathrm{ars}}(y) = -\log(1-y).$$

- Under H_0, $h_{\mathrm{ars}}(Y_t^{\mathrm{ars}}) \overset{i.i.d.}{\sim} \mathrm{Exp}(1)$ so that $\mathbb{E}_0 T_{\mathrm{ars}} = n$.
- Under H_1, $\mathbb{E}_1 T_{\mathrm{ars}} \geq n + \left(\frac{\pi^2}{6} - 1\right) \sum_{t=1}^{n} \mathbb{E}_1 \mathrm{Ent}(\boldsymbol{P}_t)$ where $\mathrm{Ent}(\boldsymbol{P}_t)$ is Shannon entropy defined by $\sum_{k=1}^{K} P_{t,k} \log \frac{1}{P_{t,k}}$.

Detection for Gumbel-max watermark

Default detection for Gumbel-max watermark [Aaronson, 2023]

Aaronson proposes to reject H_0 if the following quantity is larger than a given threshold:

$$T_{h_{\mathrm{ars}}} = \sum_{t=1}^{n} h_{\mathrm{ars}}(Y_t^{\mathrm{ars}}) \quad \text{where} \quad h_{\mathrm{ars}}(y) = -\log(1-y).$$

- Under H_0, $h_{\mathrm{ars}}(Y_t^{\mathrm{ars}}) \overset{i.i.d.}{\sim} \mathrm{Exp}(1)$ so that $\mathbb{E}_0 T_{\mathrm{ars}} = n$.
- Under H_1, $\mathbb{E}_1 T_{\mathrm{ars}} \geq n + \left(\frac{\pi^2}{6} - 1\right) \sum_{t=1}^{n} \mathbb{E}_1 \mathrm{Ent}(\boldsymbol{P}_t)$ where $\mathrm{Ent}(\boldsymbol{P}_t)$ is Shannon entropy defined by $\sum_{k=1}^{K} P_{t,k} \log \frac{1}{P_{t,k}}$.
- Other h? Using the same Y_t^{ars}, Fernandez et al. [2023] finds that $-\log(1-y)$ works better than the variant $\log y$.

Detection for inverse transform watermark

- Recall that $\zeta_t = (\pi_t, U_t) \in \Pi \times [0, 1]$. Define $\eta(k) = (k-1)/(K-1)$.
- The pivotal statistic used by Kuditipudi et al. [2023] is $Y_t^{\mathrm{dif}} = |U_t - \eta(\pi_t(w_t))|$.

Detection for inverse transform watermark

- Recall that $\zeta_t = (\pi_t, U_t) \in \Pi \times [0,1]$. Define $\eta(k) = (k-1)/(K-1)$.
- The pivotal statistic used by Kuditipudi et al. [2023] is $Y_t^{\mathrm{dif}} = |U_t - \eta(\pi_t(w_t))|$.

Default detection for inverse transform watermark [Kuditipudi et al., 2023]

Kuditipudi proposes to reject H_0 if the following quantity is larger than a given threshold:

$$T_{h_{\mathrm{neg}}} = \sum_{t=1}^{n} h_{\mathrm{neg}}(Y_t^{\mathrm{dif}}) \quad \text{where} \quad h_{\mathrm{neg}}(y) = -y.$$

- Y_t^{dif} should be smaller under H_1 than under H_0 due to the correlation between U_t and $\eta(\pi_t(w_t))$.
- No analysis for μ_0 and $\mu_{1,\boldsymbol{P}_t}$ as well as the detection rationale.

Questions studied

- Multiple detection rules for Gumbel-max watermark. No theoretical justification.
- No theoretical analysis for the inverse transform watermark.
- What is the "optimal" score function h?

Questions studied

- Multiple detection rules for Gumbel-max watermark. No theoretical justification.
- No theoretical analysis for the inverse transform watermark.
- What is the "optimal" score function h?

Main questions

- Could we propose an efficiency measure to rank different detection rule?
- Could we find the optimal detection rule according to the efficiency measure?

Outline

Efficiency measure

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

Efficiency measure

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

- Under H_0, $Y_t \overset{i.i.d.}{\sim} \mu_0$ so that Type I error can be controlled.
- Under H_1, $Y_t | \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t}$. Unknown and varies $\boldsymbol{P}_t$'s make analysis hard.

Efficiency measure

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

- Under H_0, $Y_t \overset{i.i.d.}{\sim} \mu_0$ so that Type I error can be controlled.
- Under H_1, $Y_t | \boldsymbol{P}_t \sim \mu_{1,\boldsymbol{P}_t}$. Unknown and varies $\boldsymbol{P}_t$'s make analysis hard.

Key idea

Given a prior set $\mathcal{P}$, we consider the least-favorable type II error over $\mathcal{P}$.

- We then believe all the LLM-generated $\boldsymbol{P}_t$'s fall into $\mathcal{P}$.

Efficiency measure

Theorem

Fix a pivot scalar Y and a score function h. Let the Type I error of T_h be α. Then

$$\lim_{n\to\infty}\sup \sup_{\forall \boldsymbol{P}_t\in\mathcal{P}} [1-\mathbb{E}_1 T_h(Y_{1:n})]^{1/n} \le \exp(-R_{\mathcal{P}}(h)),$$

where $R_{\mathcal{P}}(h)$ is $\mathcal{P}$-dependent efficiency rate defined by

$$R_{\mathcal{P}}(h) = -\inf_{\theta\ge 0}\{\theta\mathbb{E}_0 h(Y) + \log\phi_{\mathcal{P},h}(\theta)\} \quad \textit{with} \quad \phi_{\mathcal{P},h}(\theta) = \sup_{\boldsymbol{P}\in\mathcal{P}} \mathbb{E}_1[e^{-\theta h(Y)}|\boldsymbol{P}].$$

Efficiency measure

Theorem

Fix a pivot scalar Y and a score function h. Let the Type I error of T_h be α. Then

$$\lim_{n\to\infty} \sup_{\forall \boldsymbol{P}_t \in \mathcal{P}} \sup [1 - \mathbb{E}_1 T_h(Y_{1:n})]^{1/n} \le \exp(-R_{\mathcal{P}}(h)),$$

where $R_{\mathcal{P}}(h)$ is $\mathcal{P}$-dependent efficiency rate defined by

$$R_{\mathcal{P}}(h) = -\inf_{\theta \ge 0} \{\theta \mathbb{E}_0 h(Y) + \log \phi_{\mathcal{P},h}(\theta)\} \quad \textit{with} \quad \phi_{\mathcal{P},h}(\theta) = \sup_{\boldsymbol{P} \in \mathcal{P}} \mathbb{E}_1[e^{-\theta h(Y)} | \boldsymbol{P}].$$

- Motivated by large deviation theory.
- $R_{\mathcal{P}}(h) \ge 0$ by definition.
- The upper bound is tight under some regularities.

Optimal detection rule

- Maximize to find the optimal score function $h^\star$:

$$h^\star = \arg\max_h R_{\mathcal{P}}(h)$$

where we need to solve the following minimax problem:

$$\inf_h \left\{ \mathbb{E}_0 h(Y) + \sup_{\boldsymbol{P}\in\mathcal{P}} \log\left(\mathbb{E}_1[e^{-h(Y)} | \boldsymbol{P}]\right) \right\}.$$

Optimal detection rule

- Maximize to find the optimal score function $h^\star$:

$$h^\star = \arg\max_h R_{\mathcal{P}}(h)$$

where we need to solve the following minimax problem:

$$\inf_h \left\{ \mathbb{E}_0 h(Y) + \sup_{P \in \mathcal{P}} \log \left(\mathbb{E}_1[e^{-h(Y)} | \boldsymbol{P}] \right) \right\}.$$

- It is generally not jointly convex-concave.

How to maximize $R_{\mathcal{P}}(h)$

$$\inf_{h} \left\{ \mathbb{E}_0 h(Y) + \sup_{\boldsymbol{P} \in \mathcal{P}} \log \left(\mathbb{E}_1[e^{-h(Y)} | \boldsymbol{P}] \right) \right\}$$

Theorem

If there exists an $\boldsymbol{P}^\star \in \mathcal{P}$ and a score function class $\mathcal{H}$ such that for all $h \in \mathcal{H}$,

$$\sup_{\boldsymbol{P} \in \mathcal{P}} \mathbb{E}_1[e^{-h(Y)} | \boldsymbol{P}] = \mathbb{E}_1[e^{-h(Y)} | \boldsymbol{P}^\star], \tag{$\boldsymbol{P}^\star$}$$

$$h^\star := \log \frac{\mathrm{d}\mu_{1,\boldsymbol{P}^\star}}{\mathrm{d}\mu_0} \in \mathcal{H}, \tag{$h^\star$}$$

we then have

$$\max_{h} R_{\mathcal{P}}(h) = D_{\mathrm{KL}}(\mu_0, \mu_{1,\boldsymbol{P}^\star}),$$

where the maximum is obtained at $h^\star$.

Our choice of the prior set $\mathcal{P}$

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

Our choice of the prior set $\mathcal{P}$

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

- Singleton: $\mathcal{P} = \{\boldsymbol{P}\}$. Classic focus [Chernoff, 1952, 1956, Bahadur, 1960].

Our choice of the prior set $\mathcal{P}$

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

- Singleton: $\mathcal{P} = \{\boldsymbol{P}\}$. Classic focus [Chernoff, 1952, 1956, Bahadur, 1960].
- Our focus: Δ-regular set, i.e.,

$$\mathcal{P}_\Delta := \{\boldsymbol{P} = (P_1, \cdots, P_K) : \max_k P_k \leq 1 - \Delta\}.$$

Our choice of the prior set $\mathcal{P}$

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

- Singleton: $\mathcal{P} = \{\boldsymbol{P}\}$. Classic focus [Chernoff, 1952, 1956, Bahadur, 1960].
- Our focus: Δ-regular set, i.e.,

$$\mathcal{P}_\Delta := \{\boldsymbol{P} = (P_1, \cdots, P_K) : \max_k P_k \leq 1 - \Delta\}.$$

- Why called regular? $\mu_0 \in \mathcal{P}_\Delta|_{\Delta=0}$ so $R_{\mathcal{P}}(h) = 0$ for any h.

Outline

Application to Gumbel-max watermark

$$\sup_{\boldsymbol{P}\in\mathcal{P}_\Delta} \phi_h(\boldsymbol{P}) := \mathbb{E}_1[e^{-h(Y)}|\boldsymbol{P}] \quad \text{where} \quad \mathcal{P}_\Delta = \{\boldsymbol{P} = (P_1, \cdots, P_K) : \max_k P_k \le 1 - \Delta\}.$$

To verify the $(\boldsymbol{P}^\star)$ condition, we show that

Application to Gumbel-max watermark

$$\sup_{\boldsymbol{P}\in\mathcal{P}_\Delta} \phi_h(\boldsymbol{P}) := \mathbb{E}_1[e^{-h(Y)}|\boldsymbol{P}] \;\text{ where }\; \mathcal{P}_\Delta = \{\boldsymbol{P} = (P_1, \cdots, P_K) : \max_k P_k \le 1 - \Delta\}.$$

To verify the $(\boldsymbol{P}^\star)$ condition, we show that

- For any non-decreasing h, $\boldsymbol{P} \mapsto \phi_h(\boldsymbol{P})$ is convex in $\boldsymbol{P}$.

Application to Gumbel-max watermark

$$\sup_{\boldsymbol{P}\in\mathcal{P}_\Delta} \phi_h(\boldsymbol{P}) := \mathbb{E}_1[e^{-h(Y)}|\boldsymbol{P}] \quad \text{where} \quad \mathcal{P}_\Delta = \{\boldsymbol{P} = (P_1, \cdots, P_K) : \max_k P_k \le 1-\Delta\}.$$

To verify the $(\boldsymbol{P}^\star)$ condition, we show that

- For any non-decreasing h, $\boldsymbol{P} \mapsto \phi_h(\boldsymbol{P})$ is convex in $\boldsymbol{P}$.
- Let $\pi(\boldsymbol{P})$ denote the distribution whose kth coordinate is $P_{\pi(k)}$.

$$\text{Extreme points of } \mathcal{P}_\Delta = \{\pi(\boldsymbol{P}^\star_\Delta) : \pi \text{ is a permutation on } \{1, 2, \ldots, |\mathcal{W}|\}\},$$

where

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

Application to Gumbel-max watermark

$$\sup_{\boldsymbol{P}\in\mathcal{P}_\Delta} \phi_h(\boldsymbol{P}) := \mathbb{E}_1[e^{-h(Y)}|\boldsymbol{P}] \quad \text{where} \quad \mathcal{P}_\Delta = \{\boldsymbol{P} = (P_1, \cdots, P_K) : \max_k P_k \le 1-\Delta\}.$$

To verify the ($\boldsymbol{P}^\star$) condition, we show that

- For any non-decreasing h, $\boldsymbol{P} \mapsto \phi_h(\boldsymbol{P})$ is convex in $\boldsymbol{P}$.
- Let $\pi(\boldsymbol{P})$ denote the distribution whose kth coordinate is $P_{\pi(k)}$.

$$\text{Extreme points of } \mathcal{P}_\Delta = \{\pi(\boldsymbol{P}^\star_\Delta) : \pi \text{ is a permutation on } \{1, 2, \ldots, |\mathcal{W}|\}\},$$

where

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

$\Longrightarrow$ $\sup_{\boldsymbol{P}\in\mathcal{P}_\Delta} \phi_h(\boldsymbol{P}) := \mathbb{E}_1[e^{-h(Y)}|\boldsymbol{P}^\star_\Delta]$ for all any non-decreasing h.

Efficiency comparison for Gumbel watermark

To verify the $(h^{\star})$ condition, we find that

Efficiency comparison for Gumbel watermark

To verify the $(h^\star)$ condition, we find that

- The score function $h^\star_{\mathrm{gum},\Delta}$ is increasing:

$$h^\star_{\mathrm{gum},\Delta}(y) = \log \frac{\mathrm{d}\mu_{1,P^\star_\Delta}}{\mathrm{d}\mu_0}(y) = \log\left(\left\lfloor \frac{1}{1-\Delta} \right\rfloor y^{\frac{\Delta}{1-\Delta}} + y^{\frac{\widetilde{\Delta}}{1-\widetilde{\Delta}}} \right).$$

  where

$$\widetilde{\Delta} = 1 - (1-\Delta) \cdot \left\lfloor \frac{1}{1-\Delta} \right\rfloor.$$

Efficiency comparison for Gumbel watermark

To verify the $(h^{\star})$ condition, we find that

- The score function $h^{\star}_{\mathrm{gum},\Delta}$ is increasing:

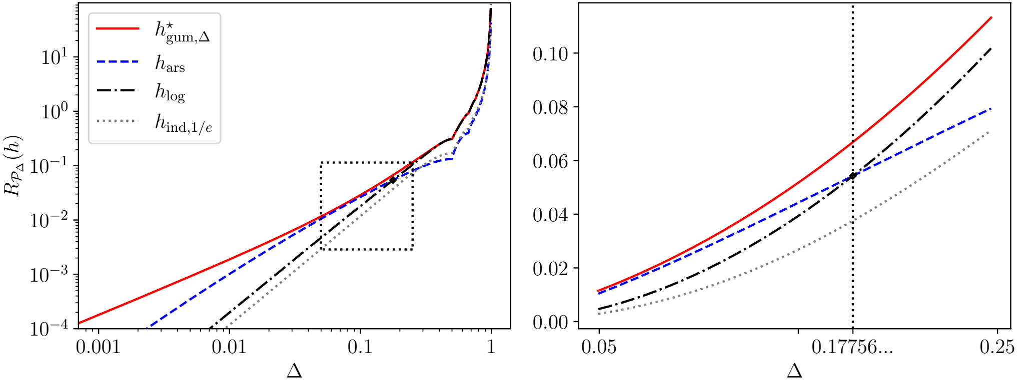
$$h^{\star}_{\mathrm{gum},\Delta}(y) = \log \frac{\mathrm{d}\mu_{1,P^{\star}_{\Delta}}}{\mathrm{d}\mu_0}(y) = \log\left(\left\lfloor \frac{1}{1-\Delta} \right\rfloor y^{\frac{\Delta}{1-\Delta}} + y^{\frac{\widetilde{\Delta}}{1-\widetilde{\Delta}}} \right).$$

  where

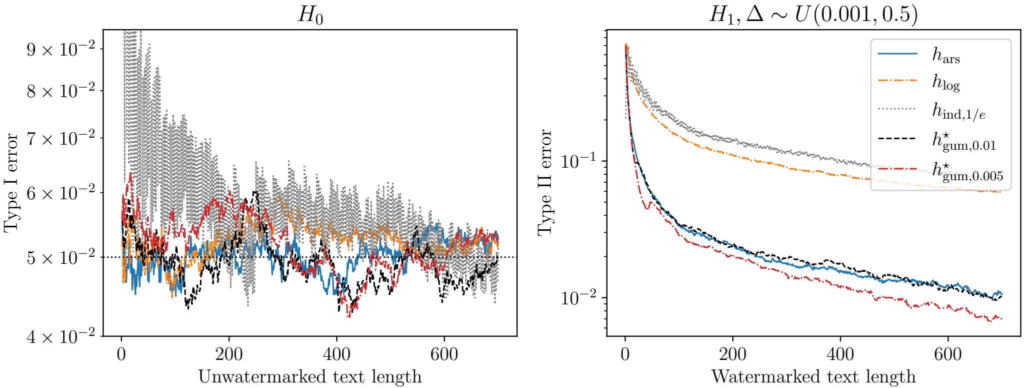
$$\widetilde{\Delta} = 1 - (1-\Delta) \cdot \left\lfloor \frac{1}{1-\Delta} \right\rfloor.$$

$\Longrightarrow$ The optimal score function is $h^{\star}_{\mathrm{gum},\Delta}$.

Efficiency comparison for Gumbel watermark



Simulation results



Outline

Analysis of μ_0 and $\mu_{1,\boldsymbol{P}}$ for inverse transform watermark

Lemma

- *Recall* $\eta(i) := \frac{i-1}{|\mathcal{W}|-1}$. *Under* H_0, *the CDF of* Y_t^{dif} *is*

$$\mathbb{P}_0(Y_t^{\mathrm{dif}} \leq r) = \frac{1}{|\mathcal{W}|} \sum_{i=1}^{|\mathcal{W}|} \left[\min\{\eta(i) + r, 1\} - \max\{\eta(i) - r, 0\}\right]. \qquad (\mu_0)$$

- *Under* H_1, *the conditional CDF of* Y_t^{dif} *is*

$$\mathbb{P}_1(Y_t^{\mathrm{dif}} \leq r | \boldsymbol{P}_t) = \frac{1}{|\mathcal{W}|!} \sum_{\pi} \sum_{i=1}^{|\mathcal{W}|} |(a_{\pi,i-1}, a_{\pi,i}] \cap B(\eta(i), r)|, \qquad (\mu_{1,\boldsymbol{P}_t})$$

where $a_{\pi,i} = \sum_{j=1}^{i} P_{t,\pi(j)}$, $B(v,r) = \{x \in [0,1] : |x - v| \leq r\}$ *and* $|\cdot|$ *represents the length of an interval.*

Analysis of μ_0 and $\mu_{1,\boldsymbol{P}}$ for inverse transform watermark

Lemma

- *Recall* $\eta(i) := \frac{i-1}{|\mathcal{W}|-1}$. *Under* H_0, *the CDF of* Y_t^{dif} *is*

$$\mathbb{P}_0(Y_t^{\mathrm{dif}} \le r) = \frac{1}{|\mathcal{W}|} \sum_{i=1}^{|\mathcal{W}|} \left[\min\{\eta(i) + r, 1\} - \max\{\eta(i) - r, 0\}\right]. \qquad (\mu_0)$$

- *Under* H_1, *the conditional CDF of* Y_t^{dif} *is*

$$\mathbb{P}_1(Y_t^{\mathrm{dif}} \le r | \boldsymbol{P}_t) = \frac{1}{|\mathcal{W}|!} \sum_{\pi} \sum_{i=1}^{|\mathcal{W}|} |(a_{\pi,i-1}, a_{\pi,i}] \cap B(\eta(i), r)|, \qquad (\mu_{1,\boldsymbol{P}_t})$$

where $a_{\pi,i} = \sum_{j=1}^{i} P_{t,\pi(j)}$, $B(v, r) = \{x \in [0,1] : |x - v| \le r\}$ *and* $|\cdot|$ *represents the length of an interval.*

- A complicated combinatorial structure due to the permutation.

Weak convergence of CDFs

Key observation

Only dominant probabilities in $\boldsymbol{P}$ matters. Simplify formulation by asymptotic analysis.

Weak convergence of CDFs

Key observation

Only dominant probabilities in $\boldsymbol{P}$ matters. Simplify formulation by asymptotic analysis.

Theorem

Under H_0,

$$\lim_{|\mathcal{W}|\to\infty} \mathbb{P}_0(Y_t^{\mathrm{dif}} \leq y) = 1 - (1-y)^2 \quad \textit{for any} \quad y \in [0,1]. \tag{μ_0}$$

Under H_1, *assuming that* $\lim_{|\mathcal{W}|\to\infty} P_{t,(1)} = 1 - \Delta$ *and* $\lim_{|\mathcal{W}|\to\infty} \log|\mathcal{W}| \cdot P_{t,(2)} = 0$ *hold,*

$$\lim_{|\mathcal{W}|\to\infty} \mathbb{P}_1(Y_t^{\mathrm{dif}} \leq y \mid \boldsymbol{P}_t) = 1 - \left(1 - \frac{y}{1-\Delta}\right)^2 \quad \textit{for any} \quad y \in [0, 1-\Delta]. \tag{$\mu_{1,\Delta}$}$$

Limit efficiency measure

- Let $K = |\mathcal{W}|$. For any $\log K \cdot \varepsilon_K \to 0$, we define

$$\overline{\mathcal{P}}_{\Delta} = \left\{ \boldsymbol{P} : P_{(1)} \leq 1 - \Delta,\ P_{(2)} \leq \varepsilon_K \right\}.$$

- The asymptotic perspective leads to the following limit efficiency measure:

$$\overline{R}_{\Delta}(h) = \liminf_{K \to \infty} R_{\overline{\mathcal{P}}_{\Delta}}(h).$$

Limit efficiency measure

- Let $K = |\mathcal{W}|$. For any $\log K \cdot \varepsilon_K \to 0$, we define

$$\overline{\mathcal{P}}_{\Delta} = \left\{ \boldsymbol{P} : P_{(1)} \leq 1 - \Delta,\ P_{(2)} \leq \varepsilon_K \right\}.$$

- The asymptotic perspective leads to the following limit efficiency measure:

$$\overline{R}_{\Delta}(h) = \liminf_{K \to \infty} R_{\overline{\mathcal{P}}_{\Delta}}(h).$$

- Under some technical conditions, we can exchange the order of $\liminf, \sup, \inf$, so

$$\overline{R}_{\Delta}(h) = -\inf_{\theta \geq 0} \left\{ \theta \mathbb{E}_{\mu_0} h(Y^{\mathrm{dif}}) + \sup_{\Delta_0 \geq \Delta} \log \mathbb{E}_{\mu_{1,\Delta_0}} [e^{-\theta h(Y^{\mathrm{dif}})}] \right\}.$$

Optimal detection rule

Finding optimal detection rule is reduced to solve the minimax optimization problem:

$$\sup_h \overline{R}_\Delta(h) = -\inf_h \left\{ \mathbb{E}_{\mu_0} h(Y^{\mathrm{dif}}) + \sup_{\Delta_0 \geq \Delta} \log \phi_h(\Delta_0) \right\} \text{ where } \phi_h(\Delta_0) := \mathbb{E}_{\mu_{1,\Delta_0}}[e^{-h(Y^{\mathrm{dif}})}].$$

Optimal detection rule

Finding optimal detection rule is reduced to solve the minimax optimization problem:

$$\sup_h \overline{R}_\Delta(h) = -\inf_h \left\{ \mathbb{E}_{\mu_0} h(Y^{\mathrm{dif}}) + \sup_{\Delta_0 \geq \Delta} \log \phi_h(\Delta_0) \right\} \text{ where } \phi_h(\Delta_0) := \mathbb{E}_{\mu_{1,\Delta_0}}[e^{-h(Y^{\mathrm{dif}})}].$$

- For $(\boldsymbol{P}^\star)$, $\sup_{\Delta_0 \geq \Delta} \phi_h(\Delta_0) := \phi_h(\Delta)$ for all any non-increasing and Lipschitz h.
- For $(h^\star)$, the score function $h^\star_{\mathrm{dif},\Delta}$ is decreasing and local Lipschitz:

$$h^\star_{\mathrm{dif},\Delta}(y) = \log \frac{f_{\mathrm{dif},\Delta}(y)}{f_{\mathrm{dif},0}(y)} \quad \text{where} \quad f_{\mathrm{dif},\Delta}(y) = \frac{2}{1-\Delta} \cdot \max\left\{1 - \frac{y}{1-\Delta}, 0\right\}.$$

Optimal detection rule

Finding optimal detection rule is reduced to solve the minimax optimization problem:

$$\sup_h \overline{R}_\Delta(h) = -\inf_h \left\{ \mathbb{E}_{\mu_0} h(Y^{\mathrm{dif}}) + \sup_{\Delta_0 \geq \Delta} \log \phi_h(\Delta_0) \right\} \text{ where } \phi_h(\Delta_0) := \mathbb{E}_{\mu_{1,\Delta_0}}[e^{-h(Y^{\mathrm{dif}})}].$$

- For $(\boldsymbol{P}^\star)$, $\sup_{\Delta_0 \geq \Delta} \phi_h(\Delta_0) := \phi_h(\Delta)$ for all any non-increasing and Lipschitz h.
- For $(h^\star)$, the score function $h^\star_{\mathrm{dif},\Delta}$ is decreasing and local Lipschitz:

$$h^\star_{\mathrm{dif},\Delta}(y) = \log \frac{f_{\mathrm{dif},\Delta}(y)}{f_{\mathrm{dif},0}(y)} \text{ where } f_{\mathrm{dif},\Delta}(y) = \frac{2}{1-\Delta} \cdot \max\left\{1 - \frac{y}{1-\Delta}, 0\right\}.$$

$\Longrightarrow$ The optimal score function is $h^\star_{\mathrm{dif},\Delta}$.

Optimal detection rule

Finding optimal detection rule is reduced to solve the minimax optimization problem:

$$\sup_h \overline{R}_\Delta(h) = -\inf_h \left\{ \mathbb{E}_{\mu_0} h(Y^{\mathrm{dif}}) + \sup_{\Delta_0 \geq \Delta} \log \phi_h(\Delta_0) \right\} \text{ where } \phi_h(\Delta_0) := \mathbb{E}_{\mu_{1,\Delta_0}}[e^{-h(Y^{\mathrm{dif}})}].$$

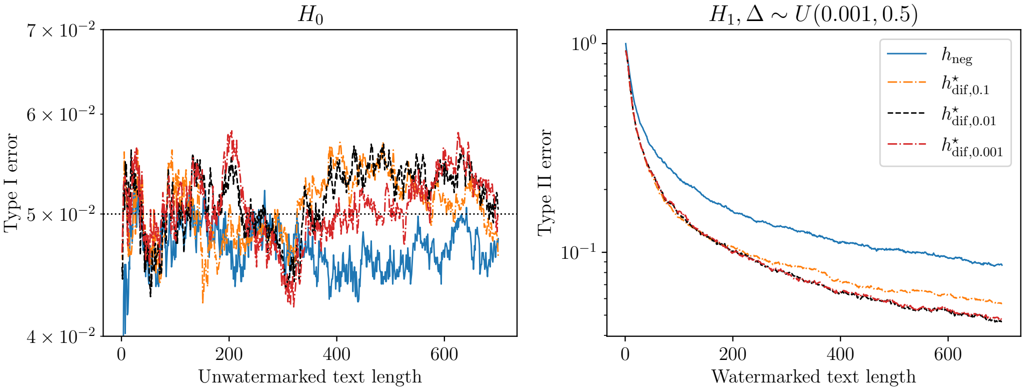
- For $(\boldsymbol{P}^\star)$, $\sup_{\Delta_0 \geq \Delta} \phi_h(\Delta_0) := \phi_h(\Delta)$ for all any non-increasing and Lipschitz h.
- For $(h^\star)$, the score function $h^\star_{\mathrm{dif},\Delta}$ is decreasing and local Lipschitz:

$$h^\star_{\mathrm{dif},\Delta}(y) = \log \frac{f_{\mathrm{dif},\Delta}(y)}{f_{\mathrm{dif},0}(y)} \quad \text{where} \quad f_{\mathrm{dif},\Delta}(y) = \frac{2}{1-\Delta} \cdot \max\left\{1 - \frac{y}{1-\Delta}, 0\right\}.$$

$\Longrightarrow$ The optimal score function is $h^\star_{\mathrm{dif},\Delta}$.

!! $\sup_h \overline{R}_\Delta(h) = \infty$. Be cautious to interpret this result.

Simulation results



Efficiency comparison on C4 newslike dataset and OPT-1.3B

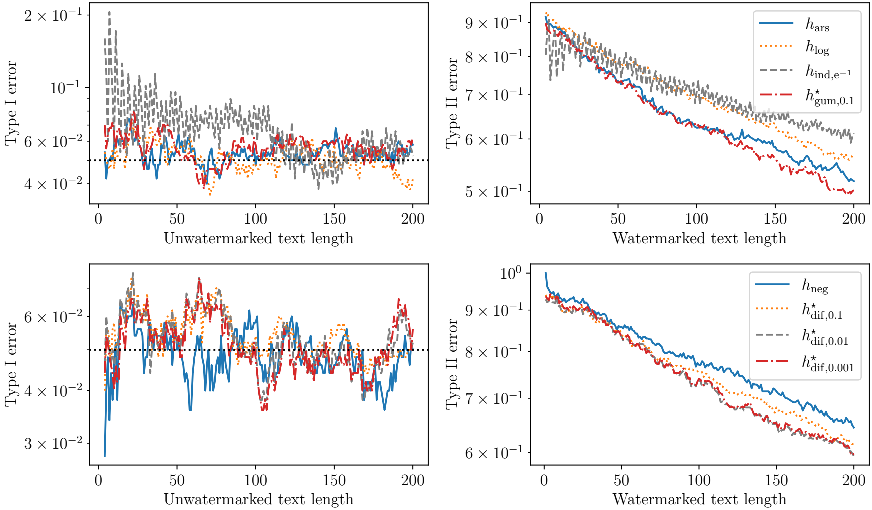


Figure: Left: Type I. Right: Type II. Top: Gumbel-max. Bottom: Inverse transform.

Outline

Summary

- A statistical framework for detecting unbiased watermarks.
- Define the least-favorable efficiency measure to compare different detection rules.
- Identify the optimal detection rule according to the efficiency measure.

Future work

- Robust detection methods.
- Compare different watermarks or pivots.
- Analysis for biased watermarks.
- Failure of working hypothesis.

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`.

Raghu R Bahadur. On the asymptotic efficiency of tests and estimates. *Sankhyā: The Indian Journal of Statistics*, pages 229–252, 1960.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

Herman Chernoff. Large-sample theory: Parametric case. *The Annals of Mathematical Statistics*, 27(1):1–22, 1956.

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.

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

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.

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.

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**v**s 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.