Towards Optimal Statistical Watermarking

Baihe Huang
University of California, Berkeley
baihe_huang@berkeley.edu

Banghua Zhu
University of California, Berkeley
banghua@berkeley.edu

Hanlin Zhu
University of California, Berkeley
hanlinzhu@berkeley.edu

Jason D. Lee
Princeton University
jasonlee@princeton.edu

Jiantao Jiao
University of California, Berkeley
jiantao@eecs.berkeley.edu

Michael I. Jordan
University of California, Berkeley
jordan@cs.berkeley.edu

Abstract

We study statistical watermarking by formulating it as a hypothesis testing problem, a general framework which subsumes all previous statistical watermarking methods. Key to our formulation is a coupling of the output tokens and the rejection region, realized by pseudo-random generators in practice, that allows non-trivial trade-off between the Type I error and Type II error. We characterize the Uniformly Most Powerful (UMP) watermark in this context. In the most common scenario where the output is a sequence of n tokens, we establish matching upper and lower bounds on the number of i.i.d. tokens required to guarantee small Type I and Type II errors. Our rate scales as $\Theta(h^{-1}\log(1/h))$ with respect to the average entropy per token h and thus greatly improves the $O(h^{-2})$ rate in the previous works. For scenarios where the detector lacks knowledge of the model's distribution, we introduce the concept of model-agnostic watermarking and establish the minimax bounds for the resultant increase in Type II error. Moreover, we formulate the robust watermarking problem where user is allowed to perform a class of perturbation on the generated texts, and characterize the optimal type II error of robust UMP tests via a linear programming problem. To the best of our knowledge, this is the first systematic statistical treatment on the watermarking problem with near-optimal rates in the i.i.d. setting, and might be of interest for future works.

1 Introduction

The prevalence of large language models (LLMs) in recent years makes it challenging and important to detect whether a human-like text is produced by the LLM system [8, 11, 4, 23, 5, 6, 20, 21, 12, 25, 10]. On the one hand, some of the most advanced LLMs to date, such as GPT-4 [13], are good at producing human-like texts, which might be hard to distinguish from human-generated texts even for humans, in various scenarios. On the other hand, it is important to keep human-produced text datasets separated from machine-produced texts to avoid the spread of misleading information [19] and the contamination of training datasets for future language models [11].

To detect machine-generated content, a recent line of work [8, 11, 4] proposes to inject *statistical watermarks*, a signal embedded within the generated texts which reveals the generation source, into texts. As discussed in [11], an ideal watermarking scheme should satisfy three properties: 1. distortion-

Socially Responsible Language Modelling Research (SoLaR) workshop at the 37th Neural Information Processing Systems (NeurIPS 2023).

free: the watermark should not alert the distribution of the generated texts; 2. agnostic: the detector needs not to know the language model or the prompt; 3. robust: the detector should be able to detect the watermark even under slight perturbation of the generated texts. However, previously proposed methods are either heuristic or guaranteed by different, sub-optimal mathematical descriptions of the above properties, making it difficult to systematically evaluate the watermarking schemes and to draw useful statistical conclusions.

Motivated by this, we propose a unifying formulation of statistical watermarking based on hypothesis testing, and study the trade-off between the Type I error and the Type II error. More specifically, our contributions are summarized as follows:

- We formulate statistical watermarking schemes as hypothesis testing with random rejection region. The random rejection region captures the secret key and the corresponding detection rules used in practice, and thus enabling our framework to encompass previous methods such as [1, 4, 8, 25, 11].
- We explicitly characterize the Uniformly Most Powerful (UMP) watermarking scheme and find the optimal type II error among all level-α tests.
- In the context where the sample is a sequence of many i.i.d. tokens, we provide nearly-matching upper bound and lower bound of the minimum number of tokens required to guarantee type I error $\leq \alpha$ and type II error $\leq \beta$; our rate $h^{-1}\log(1/h)$ improves the previous works of h^{-2} where h is the average entropy of per generated tokens.
- We introduce the concept of model-agnostic watermarking, where the distribution of the rejection region is independent of the underlying model distribution, as a notion highly practical in real-world applications. We establishes the minimax rate of the increase in Type II error loss associated with model-agnostic watermarking in comparison to UMP watermarking schemes.
- In Appendix C, we also formulate a robust watermarking problem where the watermarking scheme is robust to a class of perturbation that the user can employ to the outputs. In this setting, we also construct the robust UMP test and characterize the type II error via linear programming.

1.1 Notations

Define $(x)_+ := \max\{x, 0\}$, $x \wedge y := \min\{x, y\}$, $x \vee y = \max\{x, y\}$. For any set A, we use A^c to denote the complement of set A, $|A|$ to denote its cardinality, and $2^A := \{B : B \subset A\}$ to denote the power set of A. The total variation (TV) distance between two probability measures μ, ν is denoted by $\mathtt{TV}(\mu\|\nu)$. Given a measureable space $(\Omega, \mathcal{F})$, let $\Delta(\Omega, \mathcal{F})$ denote the set of all Baire measures over $(\Omega, \mathcal{F})$ (we will abbreviate as $\Delta(\Omega)$ when $\mathcal{F}$ is given in the context). Let δ_x denote the Dirac measure on x, i.e., $\delta_x(A) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$. We use $\operatorname{supp}(\mu)$ denote the support of a probability measure ρ. Throughout, we use $\log$ to denote natural logarithm.

2 Watermarking as a Hypothesis Testing Problem

In the problem of statistical watermarking, a service provider (e.g., a language model system), who possesses a distribution ρ over the measurable space $(\Omega, \mathcal{F})$, aims to make the samples from ρ distinguishable by a detector. The service provider achieves this by sharing a watermark key (generated from a distribution that is *coupled with* ρ) with the detector, with the goal of controlling both the Type I error (an independent output is falsely detected as from ρ) and the Type II error (an output from ρ fails to be detected). This random key together with the detection rule can be seen as a random rejection region. In the following, we formulate this problem as hypothesis testing with random rejection regions.

Problem 2.1 (Watermarking)**.** Fix $\epsilon \geq 0$. Given a probability measure ρ over a measurable space $(\Omega, \mathcal{F})$[1], an ϵ-distorted watermarking scheme of ρ is a probability measure $\mathcal{P}$ (a joint probability of the output X and the rejection region R) over the measurable space $\left(\Omega \otimes 2^\Omega, \mathcal{F} \times 2^{2^\Omega}\right)$ such that

[1] Throughout we will assume that Ω is discrete, as in most applications.

$\mathtt{TV}(\mathcal{P}(\cdot, 2^\Omega) \| \rho) \leq \epsilon$, where $\mathcal{P}(\cdot, 2^\Omega)$ is the marginal probability of X over $(\Omega, \mathcal{F})$. In the generation phase, the service provider samples (X, R) from $\mathcal{P}$, provides the output X to the service user, and sends the rejection region R to the detector.

In the detection phase, a detector is given a tuple $(X, R) \in \Omega \otimes 2^\Omega$ where X is sampled from an unknown distribution and R, given by the service provider, is sampled from the marginal probability $\mathcal{P}(\Omega, \cdot)$ over $\left(2^\Omega, 2^{2^\Omega}\right)$. The detector is tasked with conducting a hypothesis test involving two competing hypotheses:

$$\begin{aligned} H_0 &: X \text{ is sampled independently from } R, \\ \textit{versus} \;\; H_1 &: (X, R) \text{ is sampled from the joint distribution } \mathcal{P}. \end{aligned}$$

The *Type I error of* $\mathcal{P}$, defined as $\alpha(\mathcal{P}) := \sup_{\pi \in \Delta(\Omega)} \mathbb{P}_{Y \sim \pi, (X,R) \sim \mathcal{P}}(Y \in R)$, is the maximum probability that an independent sample Y is falsely rejected. The *Type II error of* $\mathcal{P}$, defined as $\beta(\mathcal{P}) := \mathbb{P}_{(X,R) \sim \mathcal{P}}(X \notin R)$, is the probability that the sample (X, R) from the joint probability $\mathcal{P}$ is not detected.

We discuss examples of statistical watermarking in Appendix B. A few remarks are in order.

Remark 2.2 (Difference between classical hypothesis testing)**.** *In classical hypothesis testing, the rejection region is often nonrandomized or randomly sampled from a distribution prior to and independently from the test statistics. However, in watermarking problem, the service provider has the incentive to facilitate the detection. The key insight is that $\mathcal{P}$ is a coupling of the random output X and the random rejection region R, so that $X \in R$ occurs with a high probability (low Type II error), while any independent sample Y lies in R with a low probability (low Type I error).*

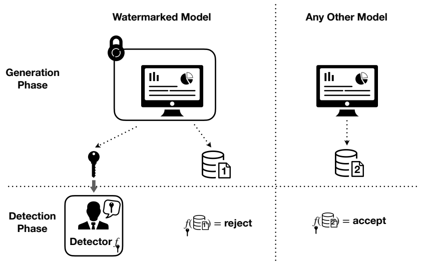

Figure 1: Illustration of watermarking in practice.

Remark 2.3 (Implementation)**.** *Note that the detector only needs to observe the rejection region, and they do not need have access to the underlying distribution ρ (language model). In fact, it is imperative for the detector to observe the rejection region: without which the output from the service provider and another independent output from the same marginal distribution would be statistically indistinguishable.*

In practice, the process of coupling and sending the rejection region can be implemented by cryptography: the service provider could hash a secret key $\mathtt{sk}$, and use a pseudo-random function F to generate $(X, R) = F(\mathtt{sk})$. Now it suffices to send the secret key to the detector, who can then reproduce the reject region using the pseudo-random function F. This process is illustrated in Figure 1.

By introducing the coupled and random rejection region, we abstract away the minutiae of cryptographical implementations, therefore allowing us to focus solely on the statistical trade-off.

In Appendix B, we provide examples of existing watermarking schemes that seamlessly fit in our framework.

3 Statistical Limit in Watermarking

Given the formulation of statistical watermarking, it is demanding to understand its statistical limit. In particular, we study the following notion of Uniformly Most Powerful (UMP) test, i.e., the watermarking scheme that achieves the minimum achievable Type II error among all possible tests with Type I error $\leq \alpha$.

Definition 3.1 (Uniformly Most Powerful Watermark)**.** A watermarking scheme $\mathcal{P}$ is called *Uniformly Most Powerful (UMP) ϵ-distorted watermark of level α*, if it achieves the minimum achievable Type II error among all ϵ-distorted watermarking with Type I error $\leq \alpha$.

The following result gives an exact characterization of the UMP watermark and its Type II error.

Theorem 3.2. *For probability measure ρ, the Uniformly Most Powerful ϵ-distorted watermark of level α, denoted by $\mathcal{P}^*$, is given by*

$$\mathcal{P}^*(X = x, R = R_0) = \begin{cases} \rho^*(x) \cdot \left(1 \wedge \frac{\alpha}{\rho^*(x)}\right), & R_0 = \{x\} \\ \rho^*(x) \cdot \left(1 - \frac{\alpha}{\rho^*(x)}\right)_+, & R_0 = \emptyset \\ 0, & \textit{else} \end{cases}$$

where $\rho^* = \arg\min_{TV(\rho' \| \rho) \leq \epsilon} \sum_{x \in \Omega : \rho'(x) > \alpha} \left(\rho'(x) - \alpha\right)$. *Its Type II error is given by* $\min_{TV(\rho' \| \rho) \leq \epsilon} \sum_{x \in \Omega : \rho'(x) > \alpha} \left(\rho'(x) - \alpha\right)$*, and when* $|\Omega| \geq \frac{1}{\alpha}$ *simplifies to*

$$\left(\sum_{x \in \Omega : \rho(x) > \alpha} \left(\rho(x) - \alpha\right) - \epsilon \right)_+ \tag{1}$$

Remark 3.3 (Dependence on distortion parameter ϵ)**.** *As seen from the theorem, when a larger distortion parameter ϵ is allowed, the Type II error would decrease. This aligns with the intuition that adding statistical bias would make the output easier to detect [1, 8]. Among all choices of ϵ, the case $\epsilon = 0$ is of particular interest since it preserves the marginal distribution of the service provider's output. Therefore, we will focus on this distortion-free case in the following sections.*

Remark 3.4 (Intuition behind $\mathcal{P}^*$)**.** *Recall that in practice, the watermarks are implemented via pseudo-random functions. Therefore, the uniformly most powerful test in Theorem 3.2 is effectively using a pseudo-random generator to approximate the distribution ρ, combined with an α-clipping to control Type I error. This construction reveals a surprising message: simply using pseudo-random generator to approximate the distribution is optimal.*

Remark 3.5 (Dependence on the randomness of ρ)**.** *If ρ is deterministic, the Type II error* $\left(\sum_{x \in \Omega : \rho(x) > \alpha} \left(\rho(x) - \alpha\right) - \epsilon\right)_+$ *reduces to $1 - \alpha - \epsilon$ and has limited practical utility. This is expected since when the service provider deterministically outputs z, it would be nearly impossible to distinguish the watermark distribution with an independent output from δ_z. In general, Theorem 3.2 implies that the Type II error decreases when the randomness in ρ increases, matching the reasoning in previous works [1, 4].*

3.1 Rates in the i.i.d. setting

In practice, the sample space Ω is usually a Cartesian product of a set Ω_0 repeated n times. For example, in large language models, the output takes form of a sequence of tokens, each coming from the same vocabulary set V. This raises the important question of specializing Theorem 3.2 to deal with distributions in product measureable spaces, and finding the explicit rates of the Type II error in Eq. (1).

In this section, we consider the product distribution $\rho = \rho_0^{\otimes n}$ over $\left(\Omega_0^{\otimes n}, (2^{\Omega_0})^{\otimes n}\right)$ and the important setting of $\epsilon = 0$ (distortion-free watermarking). Let h denote the entropy of ρ_0, the minimum number of tokens required to achieve Type I error $\leq \alpha$ and Type II error $\leq \beta$ is defined as

$$n(h, \alpha, \beta) = \min_{n \in \mathbb{Z}_+} \max_{\rho_0 : H(\rho_0) = h} \frac{n}{\mathbb{1}\left(\exists \text{ 0-distorted watermark } \mathcal{P} \text{ of } \rho_0^{\otimes n} : \alpha(\mathcal{P}) \leq \alpha, \beta(\mathcal{P}) \leq \beta\right)}.$$

In the above definition, $\frac{n}{\mathbb{1}\left(\exists \text{ 0-distorted watermark } \mathcal{P} \text{ of } \rho_0^{\otimes n} : \alpha(\mathcal{P}) \leq \alpha, \beta(\mathcal{P}) \leq \beta\right)} = +\infty$ when no 0-distorted watermark can achieve Type I error $\leq \alpha$ and Type II error $\leq \beta$. Therefore, the quantity $n(h, \alpha, \beta)$ serves as a critical threshold beyond which the desired statistical conclusions can be drawn regarding the output, making it an essential parameter in watermarking applications.

The following result gives a nearly-matching upper bound and lower bound of $n(h, \alpha, \beta)$.

Theorem 3.6. *Suppose $\alpha, \beta < 0.1$. We have*

$$n(h, \alpha, \beta) \geq \left(\frac{\log \frac{\log 2}{h}}{2h} \cdot \left(\log \frac{1}{2\alpha} \wedge \log \frac{1}{2\beta} \right) \right) \vee \frac{\log \frac{1}{2\alpha}}{h}.$$

Furthermore, let $k = |\Omega_0|$, we have

$$n(h, \alpha, \beta) \le 200 \left(\frac{2 \log \frac{9k}{h}}{h} \cdot \left(\log \frac{1}{\alpha} \wedge \log \frac{1}{\beta} \right) \right) \vee \frac{(18 + 4 \log(9k)) \log \frac{1}{\alpha}}{h}.$$

Remark 3.7 (Tightness)**.** *Up to a constant and logarithmic factor in k, our upper bound matches the lower bound. Notice that since any model with an arbitrary token set can be reduced into a model with a binary token set [4] (i.e. $k = 2$), our bound is therefore tight up to a constant factor.*

Remark 3.8 (Comparison with previous works)**.** *As commented in Remark 3.5, the regime $h \ll 1$ is more important and challenging because it is the scenario where watermarking is difficult. In a line of works [1, 8, 25, 12, 11], the rejection regions take the form of $R = \{\sum_{i=1}^{n} s_i(x_i) \ge C\}$ where s_i's are certain score functions and C is the rejection threshold. Due to Cramér's Theorem, this type of methods requires $n \ge \frac{1}{h^2}$ tokens to achieve constant Type I and Type II errors, in the regime of $h \to 0$. Comparing to them, we improve the dependence on h from $\frac{1}{h^2}$ to $\frac{\log \frac{1}{h}}{h}$ and further show that this rate is optimal.*

4 Model-Agnostic Watermarking

For practical applications, it is additionally desirable for watermarking schemes to be model-agnostic, i.e, the marginal distribution of the rejection region is independent of the watermarked distribution. Recall from Remark 2.3 that in practice detectors usually adopts a pseudo-random function to generate the reject region from the shared secret keys. If the watermarking scheme $\mathcal{P}$ depends on the underlying distribution ρ, then the pseudo-random function, and effectively the detector, need to know ρ. Therefore, model-agnostic watermarking enables the detector to use a fixed, pre-determined pseudo-random function to generate the reject region, and hence perform hypothesis-testing *without the knowledge of the underlying model that generates the output*. This is an important property enjoyed by Example B.1 and Example B.2. In this section, we formulate model-agnostic within our hypothesis testing framework and study the most powerful test.

Problem 4.1 (Model-Agnostic Watermarking)**.** Given a measurable space $(\Omega, \mathcal{F})$ and a set $\mathcal{Q} \subset \Delta(\Omega, \mathcal{F})$, a $\mathcal{Q}$-watermarking scheme is a tuple $(\eta, \{\mathcal{P}_\rho\}_{\rho \in \mathcal{Q}})$ where η is a probability measure over 2^Ω, such that for any probability measure ρ over $(\Omega, \mathcal{F})$, $\mathcal{P}_\rho$ is a distortion-free watermarking scheme of ρ and $\mathcal{P}_\rho(\Omega, \cdot) = \eta(\cdot)$.

A model-agnostic watermarking scheme is a $\Delta(\Omega, \mathcal{F})$-watermarking scheme.

Remark 4.2 (Information of the model)**.** *A $\mathcal{Q}$-watermarking scheme can be interpreted as a way to watermark all distributions in the set $\mathcal{Q}$ while revealing no information of the model used to generate the output other than the membership inside $\mathcal{Q}$ (i.e., observing the rejection region, one is only able to infer that the output comes from a model in $\mathcal{Q}$, but is unable to know which exactly the model is). By letting $\mathcal{Q}$ to be the set of all Baire measures over $(\Omega, \mathcal{F})$, model-agnostic watermarking thus reveals no information of the model.*

It is noticeable that for large $\mathcal{Q}$, a $\mathcal{Q}$-watermarking scheme can not perform as good as a watermarking specifically designed for ρ for any distribution $\rho \in \mathcal{Q}$. This means that Uniformly Most Powerful $\mathcal{Q}$-Watermarking might not exist in general. To evaluate model-agnostic watermarking schemes, a natural desideratum is therefore the maximum difference between its Type II error and the Type II error of the UMP watermarking of ρ over all distributions ρ, under fixed Type I error. Specifically, we introduce the following notion of minimax most powerful.

Definition 4.3 (Minimax most powerful model-agnostic watermark)**.** We say that a model-agnostic watermark $(\eta, \{\mathcal{P}_\rho\}_{\rho \in \Delta(\Omega, \mathcal{F})})$ is of level-α if the Type I error of $\mathcal{P}_\rho$ is less than or equal to α for any $\rho \in \Delta(\Omega, \mathcal{F})$. Define the *maximum Type II error loss* of $(\eta, \{\mathcal{P}_\rho\}_{\rho \in \Delta(\Omega, \mathcal{F})})$ as

$$\gamma(\eta) := \max_{\rho \in \Delta(\Omega, \mathcal{F})} \beta(\mathcal{P}_\rho) - \beta(\mathcal{P}_\rho^*)$$

where $\mathcal{P}_\rho^*$ is the UMP distortion-free watermark of ρ.

We say that a model-agnostic watermarking scheme is minimax most powerful, if it minimizes the maximum Type II error loss among all model-agnostic watermarks of level α.

It turns out that the minimax most powerful model-agnostic watermarking does not lose too much power as compared to UMP watermarking schemes.

Theorem 4.4. *Let $|\Omega| = n$ and suppose $\alpha n, \frac{1}{\alpha} \in \mathbb{Z}$*[2]*. In the minimax most powerful model-agnostic watermarking scheme of level-α, the marginal distribution of the reject region is given by*

$$\eta^*(A) = \begin{cases} \frac{1}{\binom{n}{\alpha n}}, & \textit{if } |A| = \alpha n \\ 0, & \textit{otherwise} \end{cases}.$$

The maximum Type II error loss of the minimax most powerful model-agnostic watermarking scheme of level-α is given by $\gamma(\eta^*) = \frac{\binom{n-\frac{1}{\alpha}}{\alpha n}}{\binom{n}{\alpha n}}$.

The proof is deferred to Appendix F.

Remark 4.5. *Theorem 4.4 implies that for any distribution ρ, the Type II error of model-agnostic watermark is upper bounded by $\frac{\binom{n-\frac{1}{\alpha}}{\alpha n}}{\binom{n}{\alpha n}} + \sum_{x:\rho(x)\geq\alpha}(\rho(x) - \alpha)$.*

Remark 4.6. *In practical scenarios, where the sample space is often a Cartesian product (e.g., a sequence of tokens), the regime of interest lies in where the ratio $1/(\alpha n)$ approaches zero. Under this condition, the rate displayed in Theorem 1 simplifies to:*

$$\frac{(n - \alpha n)(n - \alpha n - 1)\cdots(n - \alpha n - 1/\alpha + 1)}{n(n-1)\cdots(n - 1/\alpha + 1)} \asymp (1-\alpha)^{1/\alpha} \qquad \to e^{-1} \;\; \textit{as} \;\; \alpha \to 0_+.$$

This convergence implies that the minimax optimal model-agnostic watermark exhibits an increase in Type II error by an additive factor of e^{-1} compared to the UMP watermark in the worst-case scenario. It is noteworthy that this theoretical worst-case distribution might not align with the actual distribution of models encountered in practice, which are not necessarily adversarial. Therefore, determining the optimal Type II error in scenarios beyond the worst case presents an interesting question for future research.

Remark 4.7. *The e^{-1} maximum Type II error loss dose not contradict with the h^{-2} rates in previous works [1, 4, 11], because as $n \gtrsim h^{-2}$, the distribution of sequences of n tokens with average entropy h per token is beyond the worst case. Indeed, such distributions have higher differential entropy than the hard instances in the proof.*

5 Conclusions

This study has advanced the understanding of watermarking in the context of large language models by framing it within the paradigm of hypothesis testing. We find that using a pseudo-random generator to approximate the target distribution with certain clipping can yield the optimal Type II error among all level-α tests. Furthermore, in the context where the output is a sequence of several tokens, we develop the optimal rates regarding the number of i.i.d. tokens required to draw statistical conclusions, which improves the previous works from h^{-2} to $h^{-1}\log(1/h)$. To reflect the practical scenarios in which the detector often does not have knowledge of the model distribution, we formulate model-agnostic watermarking and establishes the minimax bound of the increase in Type II errors. In addition, we introduce a robust watermarking framework and characterize the robust UMP watermarking via a linear program.

Social Impacts. Watermarking is an essential technique to diminish the misuse of large language models. It tackles several critical social issues concerning the malicious usage of language models such as the contamination of datasets, academic misconduct of students, creation of fake news, and circulation of misinformation. By laying the theoretical foundation of statistical watermarking, our paper provides unifying and systematic approach to evaluate the statistical guarantees of existing and future watermarking schemes, elucidating the statistical limit of (robust) watermarking problems, and revealfinding the optimal rates in thone important setting of i.i.d. tokens. In the above ways, our work contributes to the research endeavours on addressing these societal issues in language modelling. In conclusion, our paper has positive social impacts.

[2] For the general case, it suffices to let $a_1 = 1/(\lceil 1/\alpha \rceil)$ and $n_1 = \lceil \alpha_1 n \rceil / \alpha_1$ and augment Ω with $n_1 - n$ dummy outcomes. Then $\alpha_1 n, 1/\alpha_1 \in \mathbb{Z}$ and hence the minimax bound for the new sample space with cardinality n_1 and the new Type I error α_1 yields a nearly-matching bound for (n, α).

References

[1] S. Aaronson. My ai safety lecture for ut effective altruism. *Shtetl-Optimized: The blog of Scott Aaronson. Retrieved on September*, 11:2023, 2022.

[2] S. Aaronson. Watermarking gpt outputs. *Scott Aaronson*, 2022.

[3] S. Abdelnabi and M. Fritz. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 121–140. IEEE, 2021.

[4] M. Christ, S. Gunn, and O. Zamir. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*, 2023.

[5] P. Fernandez, A. Chaffin, K. Tit, V. Chappelier, and T. Furon. Three bricks to consolidate watermarks for large language models. *arXiv preprint arXiv:2308.00113*, 2023.

[6] Y. Fu, D. Xiong, and Y. Dong. Watermarking conditional text generation for ai detection: Unveiling challenges and a semantic-aware watermark remedy. *arXiv preprint arXiv:2307.13808*, 2023.

[7] N. S. Kamaruddin, A. Kamsin, L. Y. Por, and H. Rahman. A review of text watermarking: theory, methods, and applications. *IEEE Access*, 6:8011–8028, 2018.

[8] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023.

[9] J. Kirchenbauer, J. Geiping, Y. Wen, M. Shu, K. Saifullah, K. Kong, K. Fernando, A. Saha, M. Goldblum, and T. Goldstein. On the reliability of watermarks for large language models. *arXiv preprint arXiv:2306.04634*, 2023.

[10] R. Koike, M. Kaneko, and N. Okazaki. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. *arXiv preprint arXiv:2307.11729*, 2023.

[11] R. Kuditipudi, J. Thickstun, T. Hashimoto, and P. Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

[12] A. Liu, L. Pan, X. Hu, S. Li, L. Wen, I. King, and P. S. Yu. A private watermark for large language models. *arXiv preprint arXiv:2307.16230*, 2023.

[13] OpenAI. Gpt-4 technical report, 2023.

[14] S. G. Rizzo, F. Bertini, and D. Montesi. Fine-grain watermarking for intellectual property protection. *EURASIP Journal on Information Security*, 2019:1–20, 2019.

[15] R. Sato, Y. Takezawa, H. Bao, K. Niwa, and M. Yamada. Embarrassingly simple text watermarks. *arXiv preprint arXiv:2310.08920*, 2023.

[16] V. Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

[17] F. Topsøe. Bounds for entropy and divergence for distributions over a two-element set. *J. Ineq. Pure Appl. Math*, 2(2), 2001.

[18] A. Venugopal, J. Uszkoreit, D. Talbot, F. Och, and J. Ganitkevitch. Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1363–1372, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[19] J. Vincent. AI-generated answers temporarily banned on coding q&a site stack overflow. *The Verge*, 5, 2022.

[20] L. Wang, W. Yang, D. Chen, H. Zhou, Y. Lin, F. Meng, J. Zhou, and X. Sun. Towards codable text watermarking for large language models. *arXiv preprint arXiv:2307.15992*, 2023.

[21] B. Yang, W. Li, L. Xiang, and B. Li. Towards code watermarking with dual-channel transformations. *arXiv preprint arXiv:2309.00860*, 2023.

[22] X. Yang, J. Zhang, K. Chen, W. Zhang, Z. Ma, F. Wang, and N. Yu. Tracing text provenance via context-aware lexical substitution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11613–11621, 2022.

[23] K. Yoo, W. Ahn, and N. Kwak. Advancing beyond identification: Multi-bit watermark for language models. *arXiv preprint arXiv:2308.00221*, 2023.

[24] H. Zhang, B. L. Edelman, D. Francati, D. Venturi, G. Ateniese, and B. Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. *arXiv preprint arXiv:2311.04378*, 2023.

[25] X. Zhao, P. Ananth, L. Li, and Y.-X. Wang. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023.

A Related Works

Watermarks can be injected either into a pre-existing text (edit-based watermarks) or during the text generation (generative watermarks). Although the study of generative watermarking can be dated back to [18], some older works focus on edit-based watermarking [14, 3, 22, 7], while a recent line of work studies generative watermarking [1, 8, 11, 4].

Our framework of formulating the watermark problem as a hypothesis testing is general and subsumes previous frameworks such as [8, 4]. A very recent work [11] similarly recognize the importance of correlation between the rejection region and the outputs. However, they do not study the statistical tradeoff in this paper and their rates on the required number of generated tokens is sub-optimal.

Meanwhile, various attack algorithms against watermarking schemes were also studied [8, 9, 15, 24, 11]. These attacking schemes apply quality-preserving perturbations to the watermarked outputs in delicate ways, and are therefore modelled by the perturbation graph (Definition C.1) in the robust watermark framework in Section C.

B Examples

In the following examples, we show how existing watermarking schemes fit in our framework.

Example B.1 (Text Generation with Soft Red List, [8])**.** In Algorithm 2 of [8], the watermarking scheme (over sample space $\Omega = V^*$ where V is the 'vocabulary', i.e., the set of all tokens) of ρ is given as follows:

- Fix threshold $C \in \mathbb{R}$, green list size $\gamma \in (0,1)$, and hardness parameter $\delta > 0$
- For $i = 1, 2, \ldots$
  - Randomly partition V into a green list G of size $\gamma|V|$, and a red list R of size $(1-\gamma)|V|$.
  - Sample the token X_i from the following distribution

$$\mathbb{P}(X_i = x) = \begin{cases} \frac{\rho(x|X_1,\ldots,X_{i-1})\cdot\exp(\delta)}{\sum_{x\in G}\rho(x|X_1,\ldots,X_{i-1})\cdot\exp(\delta)+\sum_{x\in R}\rho(x|X_1,\ldots,X_{i-1})}, & \text{if } x \in G \\ \frac{\rho(x|X_1,\ldots,X_{i-1})}{\sum_{x\in G}\rho(x|X_1,\ldots,X_{i-1})\cdot\exp(\delta)+\sum_{x\in R}\rho(x|X_1,\ldots,X_{i-1})}, & \text{if } x \in R \end{cases}$$

- Let the rejection region R be

$$R = \{\text{the number of green list tokens } \geq C\}.$$

The above sampling procedures as a whole define the joint distribution of the output $X = X_1 X_2 \cdots$ and the rejection region R, i.e., the $\Theta(\delta)$-distorted watermarking scheme $\mathcal{P}_{\text{SOFTREDLIST}}$. The detector observes the rejection region via the secret key that the service provider uses to generate the green and red lists.

Example B.2 (Complete watermarking algorithm Wak_{sk}, [4])**.** In Algorithm 3 of [4], the watermarking scheme (over sample space $\Omega = \{0,1\}^*$) of ρ is given as follows:

- Fix threshold $C \in \mathbb{R}$ and entropy threshold $\lambda > 0$
- Select i such that the empirical entropy of $X_1 X_2 \ldots X_i$ is greater than or equal to λ
- For $j = i+1, i+2, \ldots$
  - Sample $u_j \in [0,1]$ uniformly at random.
  - Let the binary token X_j be given by $X_j = \begin{cases} 1, & \text{if } u_j \leq \rho(1|X_1,\ldots,X_{j-1}) \\ 0, & \text{otherwise} \end{cases}$.
- Let the rejection region R be

$$R = \left\{ X : \sum_{j=i+1}^{L} \log \frac{1}{X_j u_j + (1-X_j)(1-u_j)} \geq C \right\}.$$

The above sampling procedures as a whole define the joint distribution of the output $X = X_1X_2\cdots$ and the rejection region R, i.e., the 0-distorted watermarking scheme $\mathcal{P}_{\mathrm{Wak_{sk}}}$.The detector observes the rejection region via the index i and $u_j(j > i)$.

Example B.3 (Inverse transform sampling $\mathrm{Wak}_{\mathrm{ITS}}$, [11])**.** The inverse transform sampling scheme in [4] (over sample space $\Omega = [N]^*$) of ρ is given as follows:

- Fix threshold $C \in \mathbb{R}$, resample size T, and block size k
- For $j = 1, 2, \ldots,$
    - Let $\mu \leftarrow \rho(\cdot|X_1, \ldots, X_{j-1})$.
    - Sample $\xi_j = (u_j, \pi_j), \xi_j^{(t)} = (u_j', \pi_j')$ $(t = 1, \ldots, T)$ i.i.d. according to the following distribution:
        - Sample $u \in [0, 1]$ uniformly at random;
        - Sample π uniformly at random from the space of permutations over the vocabulary $[N]$.
    - Let the token X_j be given by $X_j = \pi^{-1}\left(\min\{\pi(i) : \mu(\{j : \pi(j) \leq \pi(i)\}) \geq u\}\right)$.
- Let the rejection region R be

$$R = \left\{X : \frac{1}{T+1}\left(1 + \sum_{t=1}^{T} \mathbb{1}(\phi(X, \xi^{(t)}) \leq \phi(X, \xi)\right) \leq C\right\}$$

where $\xi = (\xi_1, \ldots, \xi_{\mathrm{len}(X)})$, $\xi^{(t)} = (\xi_1^{(t)}, \ldots, \xi_{\mathrm{len}(X)}^{(t)})$, and

$$\phi(y, \xi) = \min\left\{d\left(\{y_{i+l}\}_{l=1}^{k-1}, \{\xi_{(j+l)\%\mathrm{len}(\xi)}\}_{l=1}^{k-1}\right), i = 1, \ldots, \mathrm{len}(y) - k + 1, j = 1, \ldots, \mathrm{len}(\xi)\right\}$$

Here d is an alignment cost that is set as

$$d(y, (u, \pi)) = \sum_{i=1}^{\mathrm{len}(y)} \left|u_i - \frac{\pi_i(y_i) - 1}{N - 1}\right|$$

in [11]. Additionally, a single permutation $\pi_j^{(t)} = \pi(\forall j, t)$ is used to reduce computation overhead. The above sampling procedures as a whole define the joint distribution of the output $X = X_1X_2\cdots$ and the rejection region R in [11].The detector observes the rejection region via ξ, ξ'.

Using similar approaches in the above examples, we can encompass the methods of a number of works [2, 12, 25, 11] into our framework.

C Robust watermark

In the context of watermarking large language models, it's crucial to acknowledge users' capability to modify or manipulate model outputs. These modifications include cropping, paraphrasing, and translating the text, all of which may be employed to subvert watermark detection. Therefore, in this section, we introduce a graphical framework, modified from Problem 2.1, to account for potential user perturbations and investigate the optimal watermarking schemes robust to these perturbations.

Definition C.1 (Perturbation graph)**.** A perturbation graph over the discrete sample space Ω is a directed graph $G = (V, E)$ where V equals Ω and $(u, u) \in E$ for any $u \in V$. For any $v \in V$, let $in(v) = \{w \in V : (w, v) \in E\}$ denote the set of vertices with incoming edges to v, and let $out(v) = \{w \in V : (v, w) \in E\}$ denote the set of vertices with outcoming edges from v.

The perturbation graph specifies all the possible perturbations that could be made by the user: any $u \in V$ can be perturbed into $v \in V$ if and only if $(u, v) \in E$, i.e., there exists a directed edge from u to v.

Example C.2. Consider $\Omega = \Omega_0^{\otimes n}$. Let the user have the capacity to change no more than c tokens, i.e., perturb any sequence of tokens $x = x_1x_2\cdots x_n$ to another sequence $y = y_1y_2\cdots y_n$ with Hamming distance less than or equal to c. Then the perturbation graph is given by $G = (V, E)$ where $V = \Omega^n$ and $E = \{(u, v) : u, v \in V, d(u, v) \leq c\}$ (d is the Hamming distance, i.e., $d(x, y) = \sum_{i=1}^{n} \mathbb{1}(x_i \neq y_i)$).

Problem C.3 (Robust watermarking scheme)**.** A robust watermarking scheme with respect to a perturbation graph G is a watermarking scheme except that its Type II error is defined as $\mathbb{E}_{X,R\sim\mathcal{P}}\left[\max_{Y\in out(X)} \mathbb{1}(Y \notin R)\right]$, i.e., the probability of false negative given that the user adversarially perturbs the output.

Theorem C.4. *Define the shrinkage operator $\mathcal{S}_G : 2^\Omega \to 2^\Omega$ (of a perturbation graph G) by $\mathcal{S}_G(R) = \{x \in \Omega : out(x) \subset R\}$. Then the minimum Type II error of the robust, 0-distorted UMP test of level α in Problem C.3 is given by the solution of the following Linear Program*

$$\min_{x\in\mathbb{R}^{|\Omega|}} 1 - \sum_{y\in\Omega} \rho(y)x(y) \tag{2}$$
$$s.t. \sum_{y\in in(z)} \rho(y)x(y) \leq \alpha, \sum_{z\in\Omega} x(z) \leq 1, 0 \leq x(z) \leq 1, \ \forall z \in \Omega.$$

The UMP watermarking is $\mathcal{P}^*(X = y, R = R_0) = \begin{cases} \rho(y)\cdot x^*(y), & R_0 = \mathcal{S}_G^{-1}(\{y\}) \\ \rho(y)\cdot(1 - x^*(y)), & R_0 = \emptyset \\ 0, & \textit{otherwise} \end{cases}$,

where x^* is the solution of Eq. (2)*.*

Remark C.5 (Dependence on the sparsity of graph)**.** *From Eq.* (2)*, we observe that the perturbation graph influence the optimal Type II error via the constraint set. Indeed, if the graph is dense, the constraints $\sum_{y\in in(z)} \rho(y)x(y) \leq \alpha$ involve many entries of $y \in \Omega$ and thus decrease the value $\sum_{y\in\Omega} \rho(y)x(y)$, thereby increasing the Type II error. On the other extreme, when the edge set of the perturbation graph is $E = \{(u,u) : u \in v\}$, i.e., the user can not perturb the output to a different value, then optimum of Eq.* (2) *reduces to Eq.* (1) *(setting $\epsilon = 0$).*

D Proof of Theorem 3.2

Proof. Let ρ' denote the marginal probability of X and let η denote the marginal probability of R. In the bound of Type I error, choosing $\pi = \delta_y$ yields

$$\begin{aligned}\alpha &\geq \mathbb{P}_{X\sim\pi,R\sim\mathcal{P}(\Omega,\cdot)}(X \in R) \\ &= \mathbb{P}_{R\sim\eta}(y \in R) \\ &= \sum_{R\in 2^\Omega}\left(\sum_{x\in\Omega} \rho'(x)\mathcal{P}(R|x)\right)\cdot \mathbb{1}(y \in R). \end{aligned}\tag{3}$$

Now notice that

$$\begin{aligned}\mathcal{P}(X \in R) &= \mathbb{E}_{\mathcal{P}}[\mathbb{1}(X \in R)] \\ &= \sum_{y\in\Omega}\sum_{R\in 2^\Omega} \rho'(y)\mathcal{P}(R|y)\mathbb{1}(y \in R) \\ &= \sum_{y\in\Omega}\underbrace{\left(\sum_{R\in 2^\Omega} \rho'(y)\mathcal{P}(R|y)\cdot\mathbb{1}(y \in R)\right)}_{A(y)}.\end{aligned}$$

For the term $A(y)$, we first know that $A(y) \leq \rho'(y)$. Applying Eq. (3), we further have

$$\begin{aligned}A(y) &\leq \sum_{R\in 2^\Omega}\left(\sum_{x\in\Omega} \rho'(x)\mathcal{P}(R|x)\right)\cdot\mathbb{1}(y \in R) \\ &\leq \alpha.\end{aligned}$$

Combining the above two inequalies, it follows that

$$\begin{aligned}\mathcal{P}(X \in R) &\le \sum_{y\in\Omega} (\alpha \wedge \rho'(y)) \\ &= 1 - \sum_{x\in\Omega:\rho'(x)>\alpha} (\rho'(x) - \alpha) \\ &\le 1 - \min_{\texttt{TV}(\rho'\|\rho)\le\epsilon} \sum_{x\in\Omega:\rho'(x)>\alpha} (\rho'(x) - \alpha) \\ &\le 1 - \left(\sum_{x\in\Omega:\rho(x)>\alpha} (\rho(x) - \alpha) - \epsilon \right)_+\end{aligned}$$

where first equality is achieved by

$$\rho' = \arg \min_{\texttt{TV}(\rho'\|\rho)\le\epsilon} \sum_{x\in\Omega:\rho'(x)>\alpha} (\rho'(x) - \alpha)$$

and the second inequality is achieved when $\sum_{x\in\Omega:\rho(x)<\alpha} (\alpha - \rho(x)) \ge \epsilon$, a sufficient condition for which being $|\Omega| \ge 1/\alpha$. This establishes the optimal Type II error.

Finally, to verify that $\mathcal{P}^*$ satisfies the conditions, the condition $\texttt{TV}(\mathcal{P}^*(\cdot, 2^\Omega)\|\rho) \le \epsilon$ is apparently satisfied. For any $y \in \Omega$ we have

$$\begin{aligned}\mathbb{P}_{R\sim\eta}(y \in R) &= \sum_{x\in\Omega} \rho^*(x) \cdot \mathbb{P}(R = \{x\}) \cdot \mathbb{1}(y = x) \\ &= \rho^*(y) \cdot \left(1 \wedge \frac{\alpha}{\rho^*(y)}\right) \\ &\le \alpha.\end{aligned}$$

This implies the $\sup_{\pi\in\Delta(\Omega)} \mathbb{P}_{Y\sim\pi,(X,R)\sim\mathcal{P}^*}(Y \in R) \le \alpha$ because any π can be written as linear combination of δ_y. Moreover,

$$\begin{aligned}\mathcal{P}^*(X \in R) &= \sum_{x\in\Omega} \rho^*(x) \cdot \mathbb{P}(R = \{x\}) \\ &= \sum_{y\in\Omega} (\alpha \wedge \rho^*(y)) \\ &= 1 - \sum_{x\in\Omega:\rho^*(x)>\alpha} (\rho^*(x) - \alpha)\,.\end{aligned}$$

This verifies that ρ^* achieves the advertised Type II error. □

E Proof of Theorem 3.6

Proof. Throughout the proof we assume that $h < 1/4$, otherwise the bounds become trivial.

We first prove the lower bound. For this purpose, we construct the hard instance: let $q_0 = H_b^{-1}(h)$ (take the one $\ge 1/2$) where H_b is the binary entropy function defined by $H_b(x) = -x \ln x - (1 - x) \ln(1 - x)$, and set $\rho_0 = (1 - q_0)\delta_{x_1} + q_0\delta_{x_2}$ where x_1, x_2 are two different elements in Ω_0. Then

Lemma E.2 implies that $q_0 \geq 3/4$. By Theorem 3.2,

$$\begin{aligned}
\beta = 1 - \mathcal{P}(X \in R) &= \sum_{x \in \Omega : \rho(x) > \alpha} (\rho(x) - \alpha) \\
&\geq \frac{1}{2} \cdot \mathbb{P}\left(\rho(X) \geq 2\alpha\right) \\
&= \frac{1}{2} \cdot \mathbb{P}\left(\sum_{i=1}^{n} \ln \rho_0(X_i) \geq \ln(2\alpha)\right) \\
&\geq \mathbb{1}(n \ln q_0 \geq \ln(2\alpha)) \cdot \frac{1}{2} q_0^n \\
&\geq \mathbb{1}(2n(1-q_0) \leq -\ln(2\alpha)) \cdot \frac{1}{2} \exp\left(-n(1-q_0)\right) \\
&\geq \mathbb{1}\left(n \leq \frac{\ln \frac{1}{2\alpha}}{2h / \ln \frac{\ln 2}{h}}\right) \cdot \frac{1}{2} \exp\left(-\frac{nh}{2 \ln \frac{\ln 2}{h}}\right)
\end{aligned}$$

where the last inequality follows from Lemma E.3. It follows that

$$n(h, \alpha, \beta) \geq \frac{\ln \frac{\ln 2}{h}}{2h} \cdot \left(\ln \frac{1}{2\alpha} \wedge \ln \frac{1}{2\beta}\right). \tag{4}$$

Furthermore, suppose $n \leq \frac{\ln \frac{1}{2\alpha}}{(1-q_0) \ln \frac{1}{1-q_0}}$. Define $Y = \sum_{i=1}^{n} \mathbb{1}(\rho_0(X_i) = 1 - q_0)$, then notice that $Y \sim \mathrm{Binom}(n, 1-q_0)$ and if $Y \leq \frac{\ln \frac{1}{2\alpha}}{2 \ln \frac{1}{1-q_0}}$, then

$$\begin{aligned}
\sum_{i=1}^{n} \ln \rho_0(X_i) &\geq \frac{\ln \frac{1}{2\alpha}}{2 \ln \frac{1}{1-q_0}} \cdot \ln(1-q_0) + n \cdot \ln q_0 \\
&\geq \ln(2\alpha)
\end{aligned}$$

where the last inequality is due to $n \cdot \ln q_0 \geq -2(1-q_0)n \geq -\frac{2nh}{\ln \frac{1}{h}} \geq \frac{\alpha}{2}$. Applying this and Markov's inequality,

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{n} \ln \rho_0(X_i) \geq \ln(2\alpha)\right) &\geq \mathbb{P}\left(Y \leq \frac{\ln \frac{1}{2\alpha}}{2 \ln \frac{1}{1-q_0}}\right) \\
&\geq 1 - \frac{n(1-q_0)}{\frac{\ln \frac{1}{2\alpha}}{2 \ln \frac{1}{1-q_0}}} \\
&\geq \frac{1}{2}.
\end{aligned}$$

This is a contradiction. As a result,

$$\begin{aligned}
n(h, \alpha, \beta) &\geq \frac{\ln \frac{1}{2\alpha}}{(1-q_0) \ln \frac{1}{1-q_0}} \\
&\geq \frac{\ln \frac{1}{2\alpha}}{h}. \qquad (5)
\end{aligned}$$

Combining Eq. (4) and Eq. (5), we established the lower bound.

For the upper bound, we define $q = \max_{x \in \Omega_0} \rho_0(x)$, then Lemma E.2 implies that $q \geq 1/2$. Define $Y = \sum_{i=1}^{n} \mathbb{1}(\rho_0(X_i) \neq q)$ (recall that $Y \sim \mathrm{Binom}(n, 1-q)$). It suffices to show when

$$n = 200 \left(\frac{2 \ln \frac{9k}{h}}{h} \cdot \left(\ln \frac{1}{\alpha} \wedge \ln \frac{1}{\beta}\right)\right) \vee \frac{(18 + 4\ln(9k)) \ln \frac{1}{\alpha}}{h}$$

the Type II error of the UMP watermark $1 - \mathcal{P}^*(X \in R) \leq \beta$.

By Theorem 3.2 and Bennett's inequality,

$$\begin{aligned}
1-\mathcal{P}^*(X\in\mathbb{R}) &= \sum_{x\in\Omega:\rho(x)>\alpha}(\rho(x)-\alpha)\\
&\le \mathbb{P}\left(\rho(X)\ge\alpha\right)\\
&= \mathbb{P}\left(\sum_{i=1}^{n}\ln\rho_0(X_i)\ge\ln(\alpha)\right)\\
&\le \mathbb{P}\left(Y\ge\frac{\ln\frac{1}{\alpha}}{\ln\frac{1}{1-q}}\right)\\
&\le \exp\left(-nq(1-q)\theta\left(\frac{1-q-\frac{\ln\frac{1}{\alpha}}{n\ln\frac{1}{1-q}}}{q(1-q)}\right)\right) \qquad (6)
\end{aligned}$$

where $\theta(x)=(1+x)\ln(1+x)-x$.

Notice that by Lemma E.2,

$$\begin{aligned}
(1-q)\ln\frac{1}{1-q} &\ge \frac{h}{9\ln\frac{9k\ln(9k)}{h}}\cdot\ln\frac{\ln\frac{1}{h}}{h}\\
&= h\cdot\frac{\ln\ln\frac{1}{h}+\ln\frac{1}{h}}{9\left(\ln\frac{1}{h}+\ln(9k\ln(9k))\right)}\\
&\ge \frac{h}{9+\ln(9k\ln(9k))}.
\end{aligned}$$

Since $n\ge\frac{(18+2\ln(9k\ln(9k)))\ln\frac{1}{\alpha}}{h}$, we have $n\ge\frac{2}{1-q}\frac{\ln\frac{1}{\alpha}}{\ln\frac{1}{1-q}}$. Under this condition, we have the simplification

$$\begin{aligned}
\theta\left(\frac{1-q-\frac{\ln\frac{1}{\alpha}}{n\ln\frac{1}{1-q}}}{q(1-q)}\right) &\ge \theta\left(\frac{1}{2q}\right)\\
&\ge \frac{1}{50}.
\end{aligned}$$

Plugging back to Eq. (6), we have

$$\begin{aligned}
1-\mathcal{P}^*(X\in\mathbb{R}) &\le \exp\left(-nq(1-q)\theta\left(\frac{1-q-\frac{\ln\frac{1}{\alpha}}{n\ln\frac{1}{1-q}}}{q(1-q)}\right)\right)\\
&\le \exp\left(-\frac{n(1-q)}{100}\right)\\
&\le \exp\left(-\frac{nh}{200\ln\frac{9k\ln(9k)}{h}}\right)
\end{aligned}$$

where we applied Lemma E.3 in the last step. As $n\ge 200\left(\frac{\ln\frac{9k\ln(9k)}{h}}{h}\cdot\left(\ln\frac{1}{\alpha}\wedge\ln\frac{1}{\beta}\right)\right)$, we know that $1-\mathcal{P}^*(X\in\mathbb{R})\le\beta$. This establishes the upper bound. □

E.1 Supporting lemmata

Lemma E.1 ([17], Theorem 1.2)**.** *Define the binary entropy function $H_b:(0,1)\to\mathbb{R}$ as $H_b(x)=-x\ln x-(1-x)\ln(1-x)$. Then $4x(1-x)\le H_b(x)\le(4x(1-x))^{1/\ln 4}$.*

Lemma E.2. *Suppose ρ is a probability measure over Ω such that $H(\rho)=h$, define $q=\max_{x\in\Omega}\rho(x)$. If $H(\rho)\le 1/4$, then $q\ge 1/2$. Furthermore, if $H_b(q)\le 1/4$, then $q\ge 3/4$.*

Proof. Suppose $q \le 1/2$. By convexity of H,

$$H(\rho) \ge -\left\lfloor \frac{1}{q} \right\rfloor q \ln q \ge -\frac{1}{2} \ln \frac{1}{2} \ge 1/4.$$

This is a contradiction.

Suppose $q \le 3/4$, then Lemma E.1 implies that

$$H_b(q) \ge 4q(1-q) \ge 1/4.$$

This is a contradiction. □

Lemma E.3. *Suppose ρ is a probability measure over Ω such that $H(\rho) = h$ and $|\Omega| = k$, define $q = \max_{x\in\Omega} \rho(x)$. If $q \ge 1/2$, then we have*

$$\frac{h}{9 \ln \frac{9k \ln(9k)}{h}} \le 1 - q \le \frac{h}{\ln \frac{\ln 2}{h}}$$

Proof. We have

$$H(\rho) \ge -(1-q)\ln(1-q) \ge (1-q) \cdot \ln 2.$$

It follows that

$$\begin{aligned} h \ge & -(1-q)\ln(1-q) \\ & \ge (1-q) \ln \frac{\ln 2}{h}. \end{aligned}$$

Therefore $1 - q \le \frac{h}{\ln \frac{\ln 2}{h}}$.

By the convexity of H and $-q \ln q \le 2(1-q)$,

$$\begin{aligned} H(\rho) \le & -q\ln q - (1-q)\ln \frac{1-q}{k} \\ & \le (1-q) \ln \frac{9k}{1-q}. \end{aligned}$$

This means that

$$\begin{aligned} h^2 & \le (1-q)^2 \left(\ln \frac{9k}{1-q}\right)^2 \\ & \le (1-q) \cdot (\ln^2(9k) + 18). \end{aligned}$$

It follows that

$$\begin{aligned} h & \le (1-q) \ln \frac{9k}{1-q} \\ & \le 9(1-q) \ln \frac{9k\ln(9k)}{h}. \end{aligned}$$

This establishes $1 - q \ge \frac{h}{9 \ln \frac{9k\ln(9k)}{h}}$. □

F Proof of Theorem 4.4

Proof. **Lower bound.** Let $m = \frac{1}{\alpha}$. Notice that for any level-α model-agnostic watermarking $(\eta, \{\mathcal{P}_\rho\}_{\rho\in\Delta(\Omega,\mathcal{F})})$, the following holds

$$\sum_{A\in 2^\Omega} \eta(A)\mathbb{1}(x \in A) \le \alpha, \; \forall x \in \Omega.$$

Furthermore, for any $\rho_0 = \mathrm{Unif}(i_1, i_2, \ldots, i_m)$, we have $\beta(\mathcal{P}^*_{\rho_0}) = 0$ and

$$\begin{aligned} \beta(\mathcal{P}_{\rho_0}) & \ge \mathbb{P}_{A\sim\eta}\left(\{i_1, \ldots, i_m\} \cap A = \emptyset\right) \\ & \ge \sum_A \eta(A) \cdot \prod_{j=1}^m \mathbb{1}(i_j \notin A). \end{aligned}$$

By probabilistic method,

$$\begin{aligned}\beta(\mathcal{P}_{\rho_0}) &\geq \max_{i_1<\cdots<i_m} \sum_A \eta(A)\cdot\prod_{j=1}^m \mathbb{1}(i_j \notin A)\\ &\geq \frac{1}{\binom{n}{m}} \sum_{i_1<\cdots<i_m} \sum_A \eta(A)\cdot\prod_{j=1}^m \mathbb{1}(i_j \notin A).\end{aligned}$$

It follows that the maximum Type II error loss is lower bounded by the following linear program

$$\begin{aligned} v^* = \min_\eta\ & \frac{1}{\binom{n}{m}} \sum_{i_1<\cdots<i_m} \sum_A \eta(A)\cdot\prod_{j=1}^m \mathbb{1}(i_j \notin A)\\ \text{s.t.}\ & \sum_{A\in 2^\Omega} \eta(A)\mathbb{1}(x\in A) \leq \alpha,\ \forall x\in\Omega,\\ & \sum_{A\in 2^\Omega} \eta(A) \leq 1,\ \eta(A)\geq 0,\ \forall A\in 2^\Omega.\end{aligned}$$

By duality, this is bounded by

$$\begin{aligned} &\min_{\eta\geq 0}\max_{\xi,\zeta\geq 0} \frac{1}{\binom{n}{m}}\left(\sum_{i_1<\cdots<i_m}\sum_A \eta(A)\cdot\prod_{j=1}^m \mathbb{1}(i_j\notin A) + \sum_x \xi(x)\left(\sum_{A\in 2^\Omega}\eta(A)\mathbb{1}(x\in A) - \alpha\right)\right.\\ &\left.\quad + \zeta\cdot\left(\sum_{A\in 2^\Omega}\eta(A) - 1\right)\right)\\ &= \max_{\xi,\zeta\geq 0}\min_{\eta\geq 0}\frac{1}{\binom{n}{m}}\left(\sum_A \eta(A)\cdot\left(\sum_{i_1<\cdots<i_m}\prod_{j=1}^m \mathbb{1}(i_j\notin A) + \sum_x \xi(x)\mathbb{1}(x\in A) + \zeta\right) - \alpha\cdot\sum_x \xi(x) - \zeta\right)\\ &\geq \min_{\eta\geq 0}\frac{1}{\binom{n}{m}}\sum_{l=1}^n\sum_{|A|=l}\eta(A)\cdot\left(\binom{n-l}{m} + l\cdot\xi^* + \zeta^*\right) - \frac{\alpha n\xi^* + \zeta^*}{\binom{n}{m}}\end{aligned}$$

where $\xi^* = \binom{n-\alpha n}{m}\cdot\frac{m}{n-m}$ and $\zeta^* = 0$.

Since $f(l) := \binom{n-l}{m} + l\cdot\xi^* + \zeta^*$ is a convex function and equals minimum zero at $l^* = \alpha n$, we have $\binom{n-l}{m} + l\cdot\xi^* + \zeta^* \geq \binom{n-\alpha n}{m} + \alpha n\cdot\binom{n-\alpha n}{m}\cdot\frac{m}{n-m}$ for all $l\in[n]$ and thus

$$\textbf{RHS} \geq \frac{\binom{n-\alpha n}{m}}{\binom{n}{m}} = \frac{\binom{n-m}{\alpha n}}{\binom{n}{\alpha n}} = \frac{\binom{n-\frac{1}{\alpha}}{\alpha n}}{\binom{n}{\alpha n}}.$$

Upper bound. Notice that the marginal distribution of reject region

$$\eta^*(A) = \begin{cases}\frac{1}{\binom{n}{\alpha n}}, & \text{if } |A| = \alpha n\\ 0, & \text{otherwise}\end{cases}.$$

already guarantees Type I error $\leq \alpha$. It suffices to show **(*)**: for any $\rho\in\Delta(\Omega,\mathcal{F})$, there exists a coupling $\mathcal{P}_\rho$ of η^* and ρ such that $\mathbb{P}_{(x,A)\sim\mathcal{P}_\rho}(x\notin A) \leq \frac{\binom{n-\frac{1}{\alpha}}{\alpha n}}{\binom{n}{\alpha n}} + \sum_{x:\rho(x)\geq\alpha}(\rho(x)-\alpha)$.

Define p as the projection from $\Omega\times 2^\Omega$ to 2^Ω, i.e. $p(V) = \{A\in 2^\Omega : \exists x\in\Omega,\ s.t. (x,A)\in V\}$. Let $W := \{(x,A)\in\Omega\times 2^\Omega : x\in A\}$. To show the above, we check the Strassen's condition

$$\rho(U) - \eta^*\left(p\left(W\cap(U\times 2^\Omega)\right)\right) \leq \frac{\binom{n-\frac{1}{\alpha}}{\alpha n}}{\binom{n}{\alpha n}} + \sum_{x:\rho(x)\geq\alpha}(\rho(x)-\alpha), \forall U\subset\Omega. \tag{7}$$

Indeed, given Eq. (7), Theorem 11 in [16] establishes **(*)**.

In the rest of the proof, we show Eq. (7). Fix U with cardinality k. First notice that $\rho(U) - \sum_{x:\rho(x)\geq\alpha}(\rho(x)-\alpha) \leq (\alpha k\wedge 1)$. Since $p\left(W\cap(U\times 2^\Omega)\right) = \{A\in 2^\Omega : \exists i\in U,\ s.t.\ i\in A\}$, we have

$$\eta^*\left(p\left(W \cap (U \times 2^\Omega)\right)\right) \geq 1 - \frac{\binom{n-k}{\alpha n}}{\binom{n}{\alpha n}} = 1 - \frac{\binom{n-\alpha n}{k}}{\binom{n}{k}}.$$

If $k \leq \frac{1}{\alpha}$, then because $g(k) := \alpha k - 1 + \frac{\binom{n-\alpha n}{k}}{\binom{n}{k}}$ is convex and takes maximum $\frac{\binom{n-\alpha n}{\frac{1}{\alpha}}}{\binom{n}{\frac{1}{\alpha}}} = \frac{\binom{n-\frac{1}{\alpha}}{\alpha n}}{\binom{n}{\alpha n}}$ at $k^* = \frac{1}{\alpha}$, we have

$$\begin{aligned}\rho(U) - \eta^*\left(p\left(W \cap (U \times 2^\Omega)\right)\right) &\leq \alpha k - 1 + \frac{\binom{n-\alpha n}{k}}{\binom{n}{k}} + \sum_{x:\rho(x)\geq\alpha} (\rho(x) - \alpha) \\ &= \frac{\binom{n-\frac{1}{\alpha}}{\alpha n}}{\binom{n}{\alpha n}} + \sum_{x:\rho(x)\geq\alpha} (\rho(x) - \alpha).\end{aligned}$$

If $k \geq \frac{1}{\alpha}$, then since $\frac{\binom{n-\alpha n}{k}}{\binom{n}{k}} = \frac{\binom{n-k}{\alpha n}}{\binom{n}{\alpha n}}$ is monotonously decreasing in k,

$$\begin{aligned}\rho(U) - \eta^*\left(p\left(W \cap (U \times 2^\Omega)\right)\right) &\leq \frac{\binom{n-\alpha n}{k}}{\binom{n}{k}} + \sum_{x:\rho(x)\geq\alpha} (\rho(x) - \alpha) \\ &= \frac{\binom{n-\frac{1}{\alpha}}{\alpha n}}{\binom{n}{\alpha n}} + \sum_{x:\rho(x)\geq\alpha} (\rho(x) - \alpha).\end{aligned}$$

Combining, we establishes Eq. (7). □

G Proof of Theorem C.4

Proof. Throughout the proof we omit the subscript in the shrinkage operator $\mathcal{S}$, as G is fixed. First notice that

$$\begin{aligned}\mathbb{E}_{X,R\sim\mathcal{P}}\left[\min_{Y\in out(X)} \mathbb{1}(Y \in R)\right] &= \mathcal{P}(X \in \mathcal{S}(R)) \\ &= \sum_{y\in\Omega}\sum_{R\in 2^\Omega} \rho(y)\mathcal{P}(R|y)\mathbb{1}(y \in \mathcal{S}(R)).\end{aligned}$$

Further, notice that $y \in in(z)$ and $y \in \mathcal{S}(R)$ implies that $z \in R$, thus

$$\begin{aligned}\sum_{y\in in(z)}\sum_{R\in 2^\Omega} \rho(y)\mathcal{P}(R|y)\mathbb{1}(y \in \mathcal{S}(R)) &\leq \sum_{y\in in(z)}\sum_{R\in 2^\Omega} \rho(y)\mathcal{P}(R|y)\mathbb{1}(z \in R) \\ &= \mathbb{P}_{X\sim\delta_z, R\sim\mathcal{P}(\Omega,\cdot)}(X \in R) \\ &\leq \alpha.\end{aligned}$$

It follows that the optimum Type II error is lower bounded by the optimum of the following Linear Program

$$\min_{\mathcal{P}} 1 - \sum_{y\in\Omega}\sum_{R\in 2^\Omega} \rho(y)\mathcal{P}(R|y)\mathbb{1}(y \in \mathcal{S}(R)) \tag{8}$$

$$s.t. \sum_{y\in in(z)}\sum_{R\in 2^\Omega} \rho(y)\mathcal{P}(R|y)\mathbb{1}(y \in \mathcal{S}(R)) \leq \alpha, \sum_{R\in 2^\Omega} \mathcal{P}(R|z) = 1, 0 \leq \mathcal{P}(R|z) \leq 1,\ \forall z \in \Omega, R \in 2^\Omega.$$

We claim that the minimum in Eq. (8) is equal to the minimum of Eq. (2). Indeed, it suffices to show that Eq. (8) is optimized when $\mathcal{P}(\cdot|y_0)$ is supported on $\left\{\emptyset, \mathcal{S}^{-1}(\{y_0\})\right\}$ (then setting $x(y) \equiv \mathcal{P}(\mathcal{S}^{-1}(\{y\})|y)$ reduces Eq. (8) to Eq. (2)). To see this, consider any optimizer $\widetilde{\mathcal{P}}$ such that there exists $y_0 \in \Omega$ and $R_0 \notin \left\{\emptyset, \mathcal{S}^{-1}(\{y_0\})\right\}$, with $\mathcal{P}(R_0|y_0) = 0$. We will show that there exists $\bar{\mathcal{P}}$ such that it achieves the no greater objective value, and satisfies $|\text{supp}(\bar{\mathcal{P}}(\cdot|y_0)) \cap \left\{\emptyset, \mathcal{S}^{-1}(\{y_0\})\right\}^c| = |\text{supp}(\widetilde{\mathcal{P}}(\cdot|y_0)) \cap \left\{\emptyset, \mathcal{S}^{-1}(\{y_0\})\right\}^c| - 1$ and $|\text{supp}(\bar{\mathcal{P}}(\cdot|y))| = |\text{supp}(\widetilde{\mathcal{P}}(\cdot|y))|$ for all other $y \in \Omega$. Iteratively applying this argument, we reduce $\text{supp}(\widetilde{\mathcal{P}}(\cdot|y)) \cap \left\{\emptyset, \mathcal{S}^{-1}(\{y\})\right\}^c$ to 0 for any $y \in \Omega$ and thereby prove the claim.

Consider the following two cases.

Case 1: $y_0 \notin \mathcal{S}(R_0)$. Then letting $\bar{\mathcal{P}}(R|y) = \begin{cases} \widetilde{\mathcal{P}}(R|y), & y \neq y_0, R \neq R_0 \\ \widetilde{\mathcal{P}}(R_0|y) + \widetilde{\mathcal{P}}(\emptyset|y), & y = y_0, R = \emptyset \\ 0, & y \neq y_0, R = R_0 \end{cases}$, we observe that

$$\sum_{y\in\Omega}\sum_{R\in 2^\Omega} \rho(y)\widetilde{\mathcal{P}}(R|y)\mathbb{1}(y \in \mathcal{S}(R)) = \sum_{y\in\Omega}\sum_{R\in 2^\Omega} \rho(y)\bar{\mathcal{P}}(R|y)\mathbb{1}(y \in \mathcal{S}(R))$$

and $\bar{\mathcal{P}}$ satisfies all the constraints in Eq. (8). It is obvious from the construction of $\bar{\mathcal{P}}$ that $|\mathrm{supp}(\bar{\mathcal{P}}(\cdot|y_0)) \cap \left\{\emptyset, \mathcal{S}^{-1}(\{y_0\})\right\}^c| = |\mathrm{supp}(\widetilde{\mathcal{P}}(\cdot|y_0)) \cap \left\{\emptyset, \mathcal{S}^{-1}(\{y_0\})\right\}^c| - 1$ and $|\mathrm{supp}(\bar{\mathcal{P}}(\cdot|y))| = |\mathrm{supp}(\widetilde{\mathcal{P}}(\cdot|y))|$ for all other $y \in \Omega$.

Case 2: $y_0 \in \mathcal{S}(R_0)$. Then letting $\bar{\mathcal{P}}(R|y) = \begin{cases} \widetilde{\mathcal{P}}(R|y), & y \neq y_0, R \neq R_0 \\ \widetilde{\mathcal{P}}(R_0|y) + \widetilde{\mathcal{P}}(R|y), & y = y_0, R = \{y_0\} \\ 0, & y \neq y_0, R = R_0 \end{cases}$, we observe that

$$\sum_{y\in\Omega}\sum_{R\in 2^\Omega} \rho(y)\widetilde{\mathcal{P}}(R|y)\mathbb{1}(y \in \mathcal{S}(R)) = \sum_{y\in\Omega}\sum_{R\in 2^\Omega} \rho(y)\bar{\mathcal{P}}(R|y)\mathbb{1}(y \in \mathcal{S}(R))$$

and $\bar{\mathcal{P}}$ satisfies all the constraints in Eq. (8) due to $\mathbb{1}(y \in \mathcal{S}(R_0)) \geq \mathbb{1}(y \in \mathcal{S}(\{y_0\}))$ for any $y \in \Omega$. From the construction of $\bar{\mathcal{P}}$, we know that $|\mathrm{supp}(\bar{\mathcal{P}}(\cdot|y_0)) \cap \left\{\emptyset, \mathcal{S}^{-1}(\{y_0\})\right\}^c| = |\mathrm{supp}(\widetilde{\mathcal{P}}(\cdot|y_0)) \cap \left\{\emptyset, \mathcal{S}^{-1}(\{y_0\})\right\}^c| - 1$ and $|\mathrm{supp}(\bar{\mathcal{P}}(\cdot|y))| = |\mathrm{supp}(\widetilde{\mathcal{P}}(\cdot|y))|$ for all other $y \in \Omega$.

Combining the above cases, we established our claim.

Finally, letting $\mathcal{P}^*(\cdot|y) = x^*(y) \cdot \delta_{\mathcal{S}^{-1}(\{y\})}$ for all $y \in \omega$, where x^* is the solution of Eq. (2), achieves the optimum value in Eq. (2). □