DISTRIBUTIONAL INFORMATION EMBEDDING: A FRAMEWORK FOR MULTI-BIT WATERMARKING

Haiyun He[1] **Yepeng Liu** [2] **Ziqiao Wang**[3] **Yongyi Mao**[4] **Yuheng Bu**[2]*

[1]Cornell University [2]University of Florida [3]Tongji University [4]University of Ottawa

hh743@cornell.edu, {yepeng.liu,buyuheng}@ufl.edu, ziqiaowang@tongji.edu.cn, ymao@uottawa.ca

ABSTRACT

This paper introduces a novel problem, distributional information embedding, motivated by the practical demands of multi-bit watermarking for large language models (LLMs). Unlike traditional information embedding, which embeds information into a pre-existing host signal, LLM watermarking actively controls the text generation process—adjusting the token distribution—to embed a detectable signal. We develop an information-theoretic framework to analyze this distributional information embedding problem, characterizing the fundamental trade-offs among three critical performance metrics: text quality, detectability, and information rate. In the asymptotic regime, we demonstrate that the maximum achievable rate with vanishing error corresponds to the entropy of the LLM's output distribution and increases with higher allowable distortion. We also characterize the optimal watermarking scheme to achieve this rate. Extending the analysis to the finite-token case, we identify schemes that maximize detection probability while adhering to constraints on false alarm and distortion.

1 INTRODUCTION

The rapid advancement of Large Language Models (LLMs) (Touvron et al., 2023; Jiang et al., 2023) is revolutionizing numerous fields but also raises concerns about misuse, such as spreading disinformation, creating fake news, and enabling academic dishonesty. The growing prevalence and quality of AI-generated text make it challenging to *distinguish it from human-written content*.

A promising solution is to *actively* embed detectable signals into LLM-generated text, i.e., watermarks, which enable provable detection of AI-generated content. Despite recent advances in watermarking algorithms for LLM (Aaronson, 2023; Kirchenbauer et al., 2023; Kuditipudi et al., 2023; Zhao et al., 2023; Liu & Bu, 2024), they suffer from significant limitations, for example, many algorithms are heuristically designed where watermark detectability is ensured by introducing noticeable alterations to the generated content that degrade the output quality.

Additionally, most watermarking schemes are "zero-bit" schemes, designed solely to distinguish AI-generated text from human-written content without embedding any additional information. As incorporating meta-information—such as the model's name, version, and generation time—is increasingly important for forensic analysis of LLM misuse, some multi-bit watermarking algorithms (Yoo et al., 2023; 2024; Qu et al., 2024) have been developed recently. However, these approaches remain heuristic and have a low information embedding rate, with current methods unable to support messages longer than a few bits (Zhao et al., 2024).

Therefore, a *principled theoretical framework* is needed to analyze the fundamental trade-offs among key performance metrics in multi-bit LLM watermarking. These metrics include: (1) **Text quality**: ensuring that the watermarked text generated by LLMs maintains a quality comparable to unwatermarked text; (2) **Detectability**: the probability of missed detection and decoding errors; and (3) **Information rate**: the rate at which information can be embedded and reliably recovered.

Information theory has a long-standing history of guiding the design of digital watermarking, dating back to the early 00s (Moulin & O'Sullivan, 2000; Merhav, 2000; Moulin, 2001; Steinberg & Mer-

*Corresponding author.

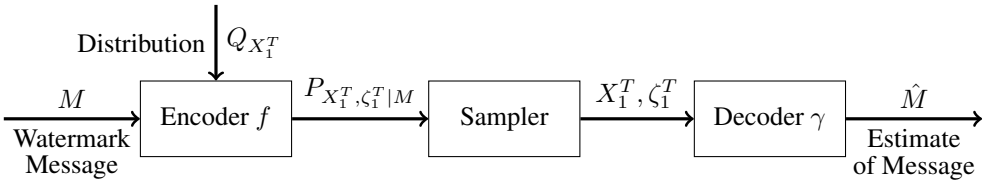

Figure 1: Illustration of multi-bit watermarking as distributional information embedding with side information.

hav, 2001; Cohen & Lapidoth, 2002), within the broader framework of the information embedding problem (Chen & Wornell, 2001; Barron et al., 2003; Moulin & O'Sullivan, 2003; Eggers et al., 2003). As we will demonstrate, watermarking in LLMs introduces a novel form of such a problem, which we term *distributional information embedding*. Unlike traditional information embedding, which focuses on reliably embedding information into a pre-existing host signal while minimizing distortion, LLM watermarking actively controls the generation process—the token distribution—to embed a detectable signal while preserving the original distribution. In other words, traditional information embedding is like writing on dirty paper (Costa, 1983), where the challenge is to convey the message clearly despite the interference from pre-existing marks. In contrast, LLM watermarking resembles generating dirty paper in real time, embedding the message into the very process that creates the marks. This fundamental difference reshapes the problem and introduces novel challenges.

In this paper, we present an information-theoretic analysis of a distributional information embedding problem motivated by multi-bit LLM watermarking. Our goal is to design the watermarking scheme by jointly optimizing the encoder and decoder. The system must distinguish human-written text from AI-generated text while ensuring reliable recovery of the embedded information. All of this must be achieved within a specified distortion constraint to preserve text quality. Our contribution includes:

- **Asymptotic analysis in the i.i.d. case**: We demonstrate that the maximum information rate with vanishing error probability corresponds to the entropy of the LLM's output distribution and increases with higher allowable distortion. Furthermore, we characterize the asymptotically optimal watermarking scheme that achieves this rate.
- **Finite token length analysis**: We extend the asymptotic analysis to a more practical scenario with finite token length, aiming to maximize the detection accuracy while satisfying both a worst-case false alarm probability constraint and a distortion constraint. We derive performance upper and lower bounds inspired by the asymptotic analysis and zero-bit watermarking.

2 Problem Formulation

Distributional Information Embedding with Side Information Consider a length-T data sequence X_1^T generated from a joint distribution $Q_{X_1^T} \in \mathcal{P}(\mathcal{X}^T)$, where $\mathcal{P}(\mathcal{X}^T)$ denotes the probability simplex in $\mathcal{X}^T$. For simplicity, we ignore the potential auto-regressive structure of $Q_{X_1^T}$ in the current analysis. In the generation process, a message M drawn from $[m] \coloneqq \{1, \ldots, m\}$ needs to be embedded in the data sequence by constructing a dependence structure between X_1^T and an auxiliary random sequence ζ_1^T with alphabet $\mathcal{Z}^T$, which serves as side information available to the decoder.

For example, the joint distribution $Q_{X_1^T}$ can be viewed as the output distribution of an LLM for a length-T token sequence X_1^T. Most LLM watermarking schemes adopt this distributional information embedding with side information framework. An example of a zero-bit watermarking scheme, where the message M simply indicates whether the content is watermarked, is provided below.

Example 1 (Existing watermarking schemes as special cases)**.** *In the Green-Red List watermarking scheme (Kirchenbauer et al., 2023), at each position t, the token vocabulary $\mathcal{X}$ is randomly split into a green list $\mathcal{G}$ and a red list $\mathcal{R}$, with $|\mathcal{G}| = \rho|\mathcal{X}|$. This split is represented by a $|\mathcal{X}|$-dimensional*

binary auxiliary variable ζ_t, indexed by $x \in \mathcal{X}$, where $\zeta_t(x) = 1$ means $x \in \mathcal{G}$; otherwise, $x \in \mathcal{R}$. The watermarking scheme is as follows:

- *Compute a hash of the previous token X_{t-1} using a hash function $hash : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ and a shared secret key: $hash(X_{t-1}, \text{key})$.*
- *Use $hash(X_{t-1}, \text{key})$ as a seed to uniformly sample the auxiliary variable ζ_t from the set $\{\zeta \in \{0,1\}^{|\mathcal{X}|} : \|\zeta\|_1 = \rho|\mathcal{X}|\}$ to construct the green list $\mathcal{G}$.*
- *Sample X_t from the modified token-generating distribution which increases the logit of tokens in $\mathcal{G}$ by $\delta > 0$:*

$$P_{X_t|x_1^{t-1},\zeta_t}(x) = \frac{Q_{X_t|x_1^{t-1}}(x)\exp(\delta \cdot \mathbb{1}\{\zeta_t(x) = 1\})}{\sum_{x\in\mathcal{V}} Q_{X_t|x_1^{t-1}}(x)\exp(\delta \cdot \mathbb{1}\{\zeta_t(x) = 1\})}.$$

More examples of several other watermarking schemes, for example, Gumbel-Max (Aaronson, 2023), EXP-Edit (Kuditipudi et al., 2023) and text-adaptive watermark (Liu & Bu, 2024), are provided in He et al. (2024, Appendix A).

In this paper, we focus on studying one usage scenario within this framework: multi-bit watermarking. Below, we formulate the multi-bit watermarking problem during data generation as a distributional information embedding problem with side information, as illustrated in Figure 1.

Definition 1 (Multi-bit Watermarking)**.** *A watermarking system is an encoder/decoder pair (f, γ). The encoder $f : [m] \times \mathcal{P}(\mathcal{X}^T) \to \mathcal{P}(\mathcal{X}^T \times \mathcal{Z}^T|[m])$ inputs a watermark message M drawn from the index set $[m]$ and the data generation distribution $Q_{X_1^T}$, outputting a joint distribution $P_{X_1^T,\zeta_1^T|M}$ that creates dependence between the generated data and auxiliary random sequence ζ_1^T. The decoder receives (X_1^T, ζ_1^T) sampled from $P_{X_1^T,\zeta_1^T|M}$, and guesses the message M with decoder $\gamma : \mathcal{X}^T \times \mathcal{Z}^T \to [0 : m]$, i.e., $\hat{M} = \gamma(X_1^T, \zeta_1^T)$. If $\hat{M} = 0$, the sequence X_1^T is decoded as unwatermarked; if $\hat{M} \in [m]$, X_1^T is decoded as watermarked with message $\hat{M}$. This system defines an (m, T) watermarking scheme with an information rate $R := \log m/T$.*

Note that the watermarked sequence is generated from $P_{X_1^T}$ (induced by the encoder f) instead of the original $Q_{X_1^T}$. To measure the *distortion level* of a watermarking scheme, we use the divergence between these two distributions.

Definition 2 (d-Distorted Watermarking)**.** *A watermarking encoder f is d-*distorted *with respect to the distortion D, if for any $M \in [m]$ and $Q_{X_1^T} \in \mathcal{P}(\mathcal{X}^T)$, the marginal distribution of the output $P_{X_1^T,\zeta_1^T|M}$ satisfies $\mathsf{D}(P_{X_1^T|M}, Q_{X_1^T}) \le d$.*

Here, D can be any divergence. Common examples of such divergences include total variation, KL divergence, and Wasserstein distance. For $d = 0$, the watermarking scheme is called *distortion-free*.

Moreover, to ensure the secrecy of the embedded message, we assume that the watermarked sequence X_1^T should be indistinguishable for any embedded message M, provided the auxiliary sequence is unknown. A distortion-free watermarking scheme satisfies this condition, as it ensures $P_{X_1^T|M=j} = Q_{X_1^T}$, for all $j \in [m]$. Additionally, the auxiliary sequence itself should not reveal any information about the message. Otherwise, the message M can be transmitted directly via the dependence between ζ and M, bypassing the need for the generated text.

Assumption 1. *The encoder f must ensure that both X_1^T and ζ_1^T are statistically independent of the embedded message M.*

Under this assumption, the embedded message M cannot be decoded with only X_1^T or ζ_1^T and $\mathsf{I}(M; X_1^T, \zeta_1^T) = \mathsf{I}(M; X_1^T|\zeta_1^T) = \mathsf{I}(M; \zeta_1^T|X_1^T)$. To detect if X_1^T is watermarked, the decoder must exploit the auxiliary sequence ζ_1^T. This corresponds to decoding with side information.

Watermark Detection and Decoding Under our framework, if the token sequence X_1^T is unwatermarked, it is independent of the sequence ζ_1^T; otherwise, (X_1^T, ζ_1^T) is jointly distributed according to one of the m distributions $\{P_{X_1^T,\zeta_1^T|M=j}\}_{j=1}^m$. Thus, detecting and decoding the watermark message M boils down to the $(m+1)$-ary hypothesis testing:

- H_0: X_1^T is generated by a human, i.e., $(X_1^T, \zeta_1^T) \sim \mathbb{P}_0 := P_{X_1^T,\zeta_1^T|M=0} = Q_{X_1^T} \otimes P_{\zeta_1^T}$;

- $\mathrm{H}_j, \forall j \in [m]$: X_1^T is generated by a watermarked LLM and embedded with message j, $(X_1^T, \zeta_1^T) \sim \mathbb{P}_j := P_{X_1^T,\zeta_1^T|M=j}$.

Detection performance is measured by the j-th error probability: for any $j \in [0:m]$,

$$\beta_j(\gamma, P_{X_1^T,\zeta_1^T|M=j}) := \mathbb{P}_j(\gamma(X_1^T, \zeta_1^T) \neq j).$$

Note that for $j \neq 0$, $\beta_j(\gamma, P_{X_1^T,\zeta_1^T|M=j}) = \mathbb{P}_j(\gamma(X_1^T, \zeta_1^T) = 0) + \mathbb{P}_j(\gamma(X_1^T, \zeta_1^T) \in [m]\backslash j)$ is the sum of miss detection error and miss decoding error. For $j = 0$, $\beta_0(\gamma, Q_{X_1^T} \otimes P_{\zeta_1^T})$ is the false alarm error. Since human-generated texts can vary widely, we aim to control the *worst-case* false alarm error $\sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T} \otimes P_{\zeta_1^T})$ at a given $\alpha \in (0, 1)$.

Our design objective is then three-fold: 1) maximizing the information rate R, 2) ensuring the distortion remains bounded by d, and 3) minimizing $\beta_j(\gamma, P_{X_1^T,\zeta_1^T|M=j})$ for all $j \in [m]$ while the *worst-case* false alarm error $\sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T} \otimes P_{\zeta_1^T})$ is controlled.

3 Asymptotic Results with IID Tokens

In this section, we begin with an asymptotic analysis by letting the length of tokens $T \to \infty$ for the i.i.d. case to build intuition for the optimal design of the watermarking scheme.

Suppose $X_1, \ldots, X_T$ are i.i.d. with an identical distribution P_X, and $\zeta_1, \ldots, \zeta_T$ are i.i.d. with P_ζ. Under each H_j, $(X_1, \zeta_1), \ldots, (X_T, \zeta_T)$ are conditionally i.i.d. with distribution $P_{X,\zeta|M=j}$. Specifically, $P_{X,\zeta|M=0} = Q_X \otimes P_\zeta$. Additionally, we assume a uniform prior distribution of message M on $[m]$.

3.1 Converse Result

We first analyze the maximum rate a watermarking scheme can achieve with vanishing decoding error $\Pr(\hat{M} \neq M) = \frac{1}{m}\sum_{j=1}^m \beta_j(\gamma, P_{X_1^T,\zeta_1^T|M=j})$.

Lemma 1 (Best Achievable Information Rate)**.** *Given any Q_X, P_X satisfying $\mathsf{D}(P_X^T, Q_X^T) \leq d$, and $\{P_{X,\zeta|M=i}\}_{i=0}^m$, if the decoding error $\Pr(\hat{M} \neq M) \to 0$ as $T \to \infty$, the information rate of this d-distorted (m, T) watermarking scheme is upper bounded by*

$$R \leq \mathsf{H}(P_X) \leq \sup_{P_X:\mathsf{D}(P_X^T,Q_X^T)\leq d} \mathsf{H}(P_X),$$

where the last bound holds for all watermarking schemes for the LLM $Q_{X_1^T}$.

The proof is provided in Appendix A.1. Lemma 1 shows that it is impossible for a distortion-free watermarking to embed more than approximately $2^{T\mathsf{H}(Q_X)}$ messages in a length-T i.i.d. token sequence while achieving vanishing j-th error for all $j \in [m]$, regardless of the false alarm probability. As the distortion d increases, a d-distorted watermarking can trade off text quality to achieve a higher information rate.

3.2 Achievability Result

Next, we aim to identify the asymptotically optimal watermarking scheme that can achieve vanishing j-th errors and the maximum watermarking rate, while ensuring the false alarm error below α.

To develop intuition for the optimal design, we first present an upper bound for the j-th error exponent under i.i.d. assumptions. Specifically, we extend Cover & Thomas (2006, Lemma 11.8.1) to our $(m+1)$-ary hypothesis testing setting.

Lemma 2 (Upper Bound for the j-th Error Exponent)**.** *Fix any $j \in [m]$, $\epsilon \in (0, 1/2)$ and any set of distributions $\{\mathbb{P}_i\}_{i\in[0:m]\backslash j}$. Let $\{\mathcal{B}_{T,i}\}_{i\in[0:m]\backslash j} \subset \mathcal{X}^T \times \mathcal{Z}^T$ be any collection of m disjoint sets of sequences $((x_i, \zeta_i))_{i=1}^T$ such that $\mathbb{P}_i(\mathcal{B}_{T,i}) \geq 1 - \epsilon$. Let $\mathcal{B}_{T,j} = (\cup_{i\in[0:m]\backslash j}\mathcal{B}_{T,i})^{\mathrm{c}}$. For any $\mathbb{P}_j$ such*

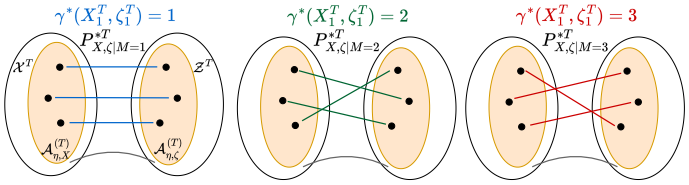

Figure 2: Illustration of the asymptotically optimal watermarking scheme when $m = 3$.

that $\max_{i\in[0:m]\setminus j} \mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i}\|P_{X,\zeta|M=j}) < \infty$,

$$-\frac{\log \mathbb{P}_j(\mathcal{B}^{\mathrm{c}}_{T,j})}{T} \le E_j^* + \epsilon - \frac{\log(m(1-2\epsilon))}{T},$$
$$\textit{where } E_j^* := \max_{P_X:\mathsf{D}(P_X^T,Q_X^T)\le d} \min_{i\in[0:m]\setminus j} \mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i}\|P_{X,\zeta|M=j}).$$

Specifically, $\mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=0}\|P_{X,\zeta|M=j}) = \mathsf{D}_{\mathsf{KL}}(P_\zeta\|P_{\zeta|X,M=j}|Q_X) + \mathsf{D}_{\mathsf{KL}}(Q_X\|P_X)$.

The proof is provided in Appendix A.2. Lemma 2 shows that given any watermarking scheme $(\mathbb{P}_0,\dots,\mathbb{P}_m)$, the minimum achievable j-th error probability for all decoders decays exponentially with the rate E_j^*, while other errors are controlled below ϵ. Furthermore, the error exponent depends on the distortion level d (cf. Definition 2), which increases as d increases. If the distortion metric is set as $\mathsf{D}_{\mathsf{KL}}(Q_X^T\|P_X^T)$, the rate is further upper bounded by $\mathsf{D}_{\mathsf{KL}}(P_\zeta\|P_{\zeta|X,M=j}|Q_X) + d$.

Inspired by Lemma 2, we can design the joint distributions $(P_{X,\zeta|M=i})_{i=0}^m$ by maximizing the error exponent E_j^*. In this way, the j-th error probability decays exponentially to 0 at the fastest rate. One solution is to make the masses of $P_{X,\zeta|M=i}$ and $P_{X,\zeta|M=j}$ concentrated at different locations for $i \neq j$, which leads to $\mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i}\|P_{X,\zeta|M=j}) \to \infty$. This hints that the optimal joint distribution produced by the encoder f should almost deterministically map (X_1^T,ζ_1^T) to a message M. Based on this intuition, we construct the asymptotically jointly optimal encoder/decoder pair in the watermarking scheme.

Under any hypothesis H_j and any $P_{X,\zeta|M=j}$, we define the typical sets of sequences $\{(x_1^T,\zeta_1^T)\}$.

Definition 3 (Typical Sets). *For arbitrarily small $\eta > 0$, define the typical sets $\mathcal{A}^{(T)}_{\eta,X}$ and $\mathcal{A}^{(T)}_{\eta,\zeta}$ as*

$$\mathcal{A}^{(T)}_{\eta,X} := \left\{x_1^T\in\mathcal{X}^T: \left|\frac{1}{T}\frac{1}{\log P_X^T(x_1^T)} - \mathsf{H}(P_X)\right| \le \eta\right\},$$
$$\mathcal{A}^{(T)}_{\eta,\zeta} := \left\{\zeta_1^T\in\mathcal{Z}^T: \left|\frac{1}{T}\log\frac{1}{P_\zeta^T(\zeta_1^T)} - \mathsf{H}(P_\zeta)\right| \le \eta\right\}.$$

The typical sequences in $\mathcal{A}^{(T)}_{\eta,X}$ and $\mathcal{A}^{(T)}_{\eta,\zeta}$ are nearly uniformly distributed and can be mapped with almost deterministic precision. Leveraging the asymptotic equipartition property (AEP), we first present the optimal design when distortion $d = 0$ as follows. Here, we use $\doteq$ to denote equality to the first order in the exponent.

Theorem 3 (Asymptotically Optimal Distortion-Free Watermarking Scheme). *Let $P_X^* = Q_X$, $\mathcal{Z}\subset\mathbb{Z}$ and design $P_\zeta^*\in\mathcal{P}(\mathcal{Z})$ such that $\mathsf{H}(P_\zeta^*) = \mathsf{H}(P_X^*)$. Let $\eta = T^{-\frac{1}{4}}$. The class of optimal decoders is given by*

$$\Gamma_\eta^* := \left\{\gamma \middle| \gamma(x_1^T,\zeta_1^T) = \begin{cases} g(x_1^T,\zeta_1^T), & \forall x_1^T\in\mathcal{A}^{(T)}_{\eta,X},\ \forall\zeta_1^T\in\mathcal{A}^{(T)}_{\eta,\zeta}, \\ 0, & \textit{otherwise}, \end{cases} \textit{for some bijective function } g: \mathcal{A}^{(T)}_{\eta,X}\times\mathcal{A}^{(T)}_{\eta,\zeta}\to[m] \textit{ satisfying } u\neq v \Rightarrow g(x_1^T,u)\neq g(x_1^T,v), \forall x_1^T\in\mathcal{A}^{(T)}_{\eta,X}.\right\}$$

If $\frac{1}{T}(\log m - \log\alpha) \le \mathsf{H}(P_X^*)$, the corresponding asymptotically optimal encoder f^* outputs $P^*_{X,\zeta|M}$ as follows: for any $i\in[m]$,

- *for all* $x_1^T \in \mathcal{A}_{\eta,X}^{(T)}$, $P_{X,\zeta|M}^{*T}(x_1^T, \zeta_1^T|i) \doteq e^{-T\mathsf{H}(P_\zeta^*)}$ *if* $\gamma(x_1^T, \zeta_1^T) = i, \zeta_1^T \in \mathcal{A}_{\eta,\zeta}^{(T)}$;
- *for all* $x_1^T \notin \mathcal{A}_{\eta,X}^{(T)}$, *let* $P_{X,\zeta|X,M}^{*T}(x_1^T, \zeta_1^T|i)$ *take any non-negative value as long as* $\sum_{x_1^T,\zeta_1^T} P_{X,\zeta|X,M}^{*T}(x_1^T, \zeta_1^T|i) = 1$.

Thus, for any $\gamma^* \in \Gamma_\eta^*$ *and its corresponding* $P_{X,\zeta|M}^*$, *as* $T \to \infty$, *we have for all* $j \in [m]$,

$$\beta_j(\gamma^*, P_{X,\zeta|M=j}^*) \le \exp(-\Omega(T^{\frac{1}{2}})) \to 0,$$
$$\text{and } \sup_{Q_X} \beta_0(\gamma^*, Q_X^T \otimes P_\zeta^{*T}) \le \alpha + \exp(-\Omega(T^{\frac{1}{2}})) \to \alpha.$$

The proof of Theorem 3 is provided in Appendix A.3. The asymptotically optimal decoder deterministically maps a typical sequence x_1^T to a typical sequence ζ_1^T uniquely under different messages M. The corresponding optimal joint distribution output by the encoder f^* assigns probability 1 to such pair of sequences (x_1^T, ζ_1^T), making sure that the detection accuracy is high. Figure 2 illustrates the design using a toy example when $m = 3$.

Remark 1 (Existence of g function and implementations)**.** *For any* $m \doteq \exp(T\mathsf{H}(P_X^*))$, *there exists at least one valid g function. Denote the typical sequences with indices as* $\{(x_1^T)_i\}_{i=1}^m, \{(\zeta_1^T)_i\}_{i=1}^m$. *One can define* $g((x_1^T)_i, (\zeta_1^T)_{(i+M-2) \mod m+1}) = M$, *for any* $i, M \in [m]$, *which takes cyclic permutation of* $\mathcal{A}_{\eta,\zeta}^{(T)}$ *as input.*

In general, the optimal design can be implemented by lossless coding schemes where the presence of side information ζ_1^T *ensures that a codeword* X_1^T *can be uniquely decoded to one message, e.g., a conditional version of arithmetic coding.*

The information rate of this distortion-free (m, T) watermarking scheme is at most

$$R \le \mathsf{H}(Q_X) + \frac{\log \alpha}{T} \xrightarrow{T\to\infty} \mathsf{H}(Q_X),$$

which achieves the maximum rate in Lemma 1 when distortion $d = 0$.

When we allow some distortion $d > 0$ in the watermarking scheme, in Theorem 3, we can change P_X^* to any P_X satisfying $\mathsf{D}(P_X^T, Q_X^T) \le d$. Therefore, the P_X^* that maximizes the information rate is

$$P_X^* = \arg\max_{P_X:\mathsf{D}(P_X^T,Q_X^T)\le d} \mathsf{H}(P_X).$$

When the distortion metric is set as D_{KL}, the solution of P_X^* is the tilting distribution of Q_X as presented in Huang et al. (2024, Theorem 1).

Notably, the asymptotic results derived for the i.i.d. case using the classical typical set analysis can be extended to the case where X_1^T, ζ_1^T are stationary ergodic processes. In this generalization, the entropy $H(P_X^*)$ is replaced by the entropy rate of the stationary ergodic process.

4 Finite-Length Analysis

Inspired by the asymptotically optimal design, we are now ready to proceed with our analysis in the finite-length setting.

We consider the following optimization problem. For any $j \in [m]$, we aim to minimize the j-th error probability by jointly optimizing the watermarking encoder and decoder, subject to the following constraints: 1) all other errors are under control, and 2) the distortion remains bounded:

$$\inf_{\gamma, P_{X_1^T,\zeta_1^T|M=j}} \beta_j(\gamma, P_{X_1^T,\zeta_1^T|M=j}) \qquad \text{(Opt-O)}$$
$$\text{s.t.} \quad \sup_{P_{X_1^T,\zeta_1^T|M=i}} \beta_i(\gamma, P_{X_1^T,\zeta_1^T|M=i}) \le \alpha, \forall i \ne j$$
$$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le d.$$

The worst-case constraints on the error probabilities guarantee that the solutions to the m optimization problems collectively form an optimal watermarking scheme that achieves the minimum j-th error for all $j \in [m]$, with the worst-case false alarm under α and distortion under d.

Converse We present the following theorem, which characterizes a lower bound for the minimum error of this optimization problem.

Theorem 4 (Lower Bound for Minimum j-th Error)**.** *For any $j \in [m]$, the lower bound for the minimum j-th error attained from* (Opt-O) *is*

$$\beta_j^* \geq m\beta^*(\alpha, T), \quad \textit{where}$$
$$\beta^*(\alpha, T) := \min_{P_{X_1^T}: \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq d} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+,$$

for m, α satisfying $m\beta^*(\alpha, T) \leq 1$.

The proof is provided in Appendix A.4. First, we observe that $\beta^*(\alpha, T)$ represents the universally minimum Type-II error for any zero-bit watermarking scheme (He et al., 2024). In the context of a multi-bit watermarking scheme embedding m messages, the lower bound of the minimum j-th error increases by a factor of m. Second, Theorem 4 shows that no d-distorted (m, T) watermarking scheme can achieve a j-th error smaller than 1 for α too small or m too large. For any given α and T, the maximum number of messages we can embed is $m \leq 1/\beta^*(\alpha, T)$. This aligns with the i.i.d. scenario where the information rate $\frac{\log m}{T}$ should be bounded by $\sup_{P_X: \mathsf{D}(P_X^T, Q_X^T) \leq d} \mathsf{H}(P_X)$ to ensure vanishing decoding error. Third, as the distortion d increases, the lower bound for β_j^* decreases. This means that a watermarking scheme can trade off text quality for lower detection errors.

Achievability Since Theorem 4 also holds for the case where the auxiliary sequence ζ_1^T is *not* independent of the embedded message M, we present a watermarking scheme with a message-dependent auxiliary sequence design that achieves Theorem 4.

Theorem 5 (Watermarking Scheme with Message-Dependent ζ_1^T)**.** *Choose $\mathcal{Z}^T \subset \mathbb{Z}^T$ such that $|\mathcal{Z}|^T = m|\mathcal{X}|^T + 1$. Randomly pick a sequence $\tilde{\zeta_1^T} \in \mathcal{Z}^T$ and partition $\mathcal{Z}^T \backslash \{\tilde{\zeta_1^T}\}$ into m disjoint subsets $\{\mathcal{S}_j\}_{j=1}^m$ of equal size. Define a class of decoders as*

$$\Gamma_{\tilde{\zeta_1^T}} := \left\{ \gamma \;\middle|\; \gamma(x_1^T, \zeta_1^T) = \begin{cases} j, & \textit{if } \zeta_1^T \neq \tilde{\zeta_1^T} \\ & \textit{and } x_1^T = h_j(\zeta_1^T)\,, \\ 0, & \textit{otherwise}, \end{cases} \begin{array}{l} \textit{for some group of bijective functions} \\ \{h_j : \mathcal{S}_j \to \mathcal{X}^T\}_{j=1}^m. \end{array} \right\}$$

For any $\gamma \in \Gamma_{\tilde{\zeta_1^T}}$, let the corresponding encoder f outputs $(\mathbb{P}_j)_{j=0}^m$ as follows:

$$P_{X_1^T}^* = \operatorname*{arg\,min}_{P_{X_1^T}: \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq d} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+,$$

and for any $j \in [m]$,

$$\mathbb{P}_j(x_1^T, \zeta_1^T) = \begin{cases} P_{X_1^T}^* - m(P_{X_1^T}^*(x_1^T) - \alpha)_+, & \textit{if } \gamma(x_1^T, \zeta_1^T) = j; \\ (P_{X_1^T}^*(x_1^T) - \alpha)_+, & \textit{if } \gamma(x_1^T, \zeta_1^T) = i, \forall i \in [m] \backslash j, \textit{ or if } \zeta_1^T = \tilde{\zeta_1^T}; \\ 0, & \textit{otherwise}. \end{cases}$$

Specifically, $\mathbb{P}_0 = Q_{X_1^T} \otimes P_{\zeta_1^T}$, where $P_{\zeta_1^T} = \frac{1}{m} \sum_{j=1}^m P_{\zeta_1^T | M=j}$. The j-th error and false alarm error probabilities are given by: $\beta_j(\gamma, P_{X_1^T, \zeta_1^T | M=j}) = m\beta^*(\alpha, T)$ and $\sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T} \otimes P_{\zeta_1^T}) = \alpha$.

The proof of Theorem 5 is provided in Appendix A.5. This watermarking scheme is optimal in scenarios where dependence between the message and auxiliary sequence is allowed. The construction of the encoder's output distribution $P_{X_1^T, \zeta_1^T | M}$ can be regarded as an extension of He et al. (2024, Theorem 2). It is equivalent to transporting the probability mass from $\mathcal{V}^T$ to $\mathcal{Z}^T$, maximizing $P_{X_1^T, \zeta_1^T | M}^*(x_1^T, \zeta_1^T)$ for $\gamma(x_1^T, \zeta_1^T) = M$, while keeping the worst-case false alarm error below α. Moreover, the introduction of $\tilde{\zeta_1^T}$ helps to control the worst-case false alarm. If $P_{X_1^T}^*(x_1^T) > \alpha$ (i.e.,

low-entropy text), x_1^T may be mapped to $\tilde{\zeta}_1^T$ during watermarking, which makes it harder to detect as watermarked. In conclusion, the proposed scheme provides a guideline for the future design of watermarking schemes that satisfy the independence assumption in Assumption 1 and approach the lower bound in Theorem 4.

5 Discussion and Future Works

While our theoretical analysis of the distributional information embedding problem does not fully account for all aspects of LLMs (e.g., auto-regressive nature), we believe it provides valuable insights for designing multi-bit watermarking schemes. We rigorously demonstrate that the best achievable rate in the asymptotic regime is determined by the entropy of the distribution $\mathsf{H}(P_X)$, establishing a fundamental limit that serves as a benchmark for evaluating existing multi-bit watermarking schemes.

Moreover, this result implies an inherent connection between the problem of distributional information embedding and lossless compression, where the fundamental limit is also the entropy of the source distribution. Interestingly, Huang et al. (2024) proposes a steganography algorithm that exploits this connection by using the decoder of an arithmetic coding scheme[1] as the encoder to sample from the LLM while employing the arithmetic coding encoder as the decoder in our context. This duality between the two problems suggests that new watermarking schemes could be inspired by existing source coding techniques, presenting an intriguing direction for future exploration.

References

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

Richard J Barron, Brian Chen, and Gregory W Wornell. The duality between information embedding and source coding with side information and some applications. *IEEE Transactions on Information Theory*, 49(5):1159–1180, 2003.

Brian Chen and Gregory W Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information theory*, 47(4):1423–1443, 2001.

Aaron S Cohen and Amos Lapidoth. The gaussian watermarking game. *IEEE transactions on Information Theory*, 48(6):1639–1667, 2002.

Max Costa. Writing on dirty paper (corresp.). *IEEE transactions on information theory*, 29(3): 439–441, 1983.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

Joachim J Eggers, Robert Bauml, Roman Tzschoppe, and Bernd Girod. Scalar costa scheme for information embedding. *IEEE Transactions on signal processing*, 51(4):1003–1019, 2003.

Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. Universally optimal watermarking schemes for LLMs: from theory to practice, 2024. URL https://arxiv.org/abs/2410.02890.

Yu-Shin Huang, Peter Just, Krishna Narayanan, and Chao Tian. OD-Stega: LLM-based near-imperceptible steganography via optimized distributions, 2024. URL https://arxiv.org/abs/2410.04328.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[1]The decoder of source coding maps message bit to symbols (tokens), and the encoder maps tokens to message.

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

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

Yepeng Liu and Yuheng Bu. Adaptive text watermark for large language models. In *Forty-first International Conference on Machine Learning*, 2024.

Neri Merhav. On random coding error exponents of watermarking systems. *IEEE Transactions on Information Theory*, 46(2):420–430, 2000.

Pierre Moulin. The role of information theory in watermarking and its application to image watermarking. *Signal Processing*, 81(6):1121–1139, 2001.

Pierre Moulin and Joseph A O'Sullivan. Information-theoretic analysis of watermarking. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 6, pp. 3630–3633. IEEE, 2000.

Pierre Moulin and Joseph A O'Sullivan. Information-theoretic analysis of information hiding. *IEEE Transactions on information theory*, 49(3):563–593, 2003.

Wenjie Qu, Dong Yin, Zixin He, Wei Zou, Tianyang Tao, Jinyuan Jia, and Jiaheng Zhang. Provably robust multi-bit watermarking for AI-generated text via error correction code. *arXiv preprint arXiv:2401.16820*, 2024.

Yossef Steinberg and Neri Merhav. Identification in the presence of side information with application to watermarking. *IEEE Transactions on Information Theory*, 47(4):1410–1422, 2001.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. Robust multi-bit natural language watermarking through invariant features. *arXiv preprint arXiv:2305.01904*, 2023.

KiYoon Yoo, Wonhyuk Ahn, and Nojun Kwak. Advancing beyond identification: Multi-bit watermark for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4031–4055, 2024.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. *arXiv preprint arXiv:2306.17439*, 2023.

Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, et al. Sok: Watermarking for ai-generated content. *arXiv preprint arXiv:2411.18479*, 2024.

A APPENDIX

A.1 PROOF OF LEMMA 1

Proof. Let $P_e = \Pr(\hat{M} \neq M)$. From the Fano's inequality, we have

$$\mathsf{H}(M|\hat{M}, \zeta_1^T) \leq \mathsf{H}(M|\hat{M}) \leq 1 + P_e \log m.$$

The entropy of M is upper bounded by

$$\begin{aligned}\log m = \mathsf{H}(M) = \mathsf{H}(M|\zeta_1^T) &= \mathsf{I}(M; \hat{M}|\zeta_1^T) + \mathsf{H}(M|\hat{M}, \zeta_1^T) \\ &\leq \mathsf{I}(M; X_1^T|\zeta_1^T) + 1 + P_e \log m \\ &\leq H(X_1^T|\zeta_1^T) + 1 + P_e \log m,\end{aligned}$$

which leads to

$$\frac{\log m}{T} \leq \frac{H(X_1^T|\zeta_1^T)}{T} + \frac{1}{T} + P_e \frac{\log m}{T}.$$

If $P_e \to 0$ as $T \to \infty$, we have

$$\frac{\log m}{T} \leq \frac{H(X_1^T|\zeta_1^T)}{T} \leq \mathsf{H}(P_X) \leq \sup_{P_X : \mathsf{D}(P_X^T, Q_X^T) \leq d} \mathsf{H}(P_X).$$

□

A.2 PROOF OF LEMMA 2

Proof. For any $i \neq j$, define the relative entropy typical set

$$\mathcal{A}_{\epsilon,i,j}^{(T)}(\mathbb{P}_i \| \mathbb{P}_j) \coloneqq \left\{ (x_1^T, \zeta_1^T) : \left| \frac{1}{T} \log \frac{\mathbb{P}_i(x_1^T, \zeta_1^T)}{\mathbb{P}_j(x_1^T, \zeta_1^T)} - \mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) \right| \leq \epsilon \right\}.$$

We have $\mathbb{P}_j(\mathcal{B}_{T,j}^{\mathrm{c}}) = 1 - \mathbb{P}_j(\mathcal{B}_{T,j})$ and

$$\begin{aligned}\mathbb{P}_j(\mathcal{B}_{T,j}) &= 1 - \sum_{i:i\neq j} \mathbb{P}_j(\mathcal{B}_{T,i}) \leq 1 - \sum_{i:i\neq j} \mathbb{P}_j(\mathcal{B}_{T,i} \cap \mathcal{A}_{\epsilon,i,j}^{(T)}) \\ &\leq 1 - \sum_{i:i\neq j} \sum_{(x_1^T, \zeta_1^T) \in \mathcal{B}_{T,i} \cap \mathcal{A}_{\epsilon,i,j}^{(T)}} \mathbb{P}_i(x_1^T, \zeta_1^T) \exp(-T(\mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon)) \\ &= 1 - \sum_{i:i\neq j} \exp(-T(\mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon)) \mathbb{P}_i(\mathcal{B}_{T,i} \cap \mathcal{A}_{\epsilon,i,j}^{(T)}) \\ &\overset{\text{(a)}}{\leq} 1 - \sum_{i:i\neq j} \exp(-T(\mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon))(1 - 2\epsilon) \\ &\leq 1 - m(1-2\epsilon) \exp(-T(\min_{i:i\neq j} \mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon)) \\ &\leq 1 - m(1-2\epsilon) \exp(-T(\max_{P_X : \mathsf{D}(P_X^T, Q_X^T) \leq d} \min_{i:i\neq j} \mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon))\end{aligned}$$

where (a) follows since $\mathbb{P}_i(\mathcal{B}_{T,i} \cap \mathcal{A}_{\epsilon,i,j}^{(T)}) = 1 - \mathbb{P}_i(\mathcal{B}_{T,i}^{\mathrm{c}} \cup (\mathcal{A}_{\epsilon,i,j}^{(T)})^{\mathrm{c}}) \geq 1 - \mathbb{P}_i(\mathcal{B}_{T,i}^{\mathrm{c}}) - \mathbb{P}_i((\mathcal{A}_{\epsilon,i,j}^{(T)})^{\mathrm{c}}) \geq 1 - 2\epsilon$ for sufficiently large T. The proof is thus complete. □

A.3 PROOF OF THEOREM 3

Let $\eta = T^{-\frac{1}{4}}$ and define the set $\mathcal{A}_{\eta,j}^{(T)}$ of jointly typical sequences $\{(x_1^T, \zeta_1^T)\}$ w.r.t. the distribution $P_{X,\zeta|M=j}$ as

$$\begin{aligned}\mathcal{A}_{\eta,j}^{(T)} \coloneqq \Bigg\{ (x_1^T, \zeta_1^T) \in \mathcal{X}^T \times \mathcal{Z}^T : &\left| -\frac{1}{T} \log P_X^T(x_1^T) - \mathsf{H}(P_X) \right| \leq \eta, \left| -\frac{1}{T} \log P_\zeta^T(\zeta_1^T) - \mathsf{H}(P_\zeta) \right| \leq \eta, \\ &\left| -\frac{1}{T} \log P_{X,\zeta|M=j}^T(x_1^T, \zeta_1^T) - \mathsf{H}(P_{X,\zeta|M=j}) \right| \leq \eta \Bigg\}.\end{aligned}$$

First, we bound the probability of the atypical sets $(\mathcal{A}_{\eta,X}^{(T)})^{\mathrm{c}}, (\mathcal{A}_{\eta,\eta}^{(T)})^{\mathrm{c}}, (\mathcal{A}_{\eta,j}^{(T)})^{\mathrm{c}}$. From the union bound, we have

$$\mathbb{P}_j((X_1^T, \zeta_1^T) \notin \mathcal{A}_{\eta,j}^{(T)}) \leq \mathbb{P}_j\bigg(\bigg| -\frac{1}{T}\log P_X^T(x_1^T) - \mathsf{H}(P_X)\bigg| \geq \eta\bigg) + \mathbb{P}_j\bigg(\bigg| -\frac{1}{T}\log P_\zeta^T(\zeta_1^T) - \mathsf{H}(P_\zeta)\bigg| \geq \eta\bigg)$$

$$+ \mathbb{P}_j\bigg(\bigg| -\frac{1}{T}\log P_{X,\zeta|M=j}^T(x_1^T, \zeta_1^T) - \mathsf{H}(P_{X,\zeta|M=j})\bigg| \geq \eta\bigg).$$

Then, by the Chernoff bound, we have

$$\begin{aligned}
&\mathbb{P}_j\bigg(\bigg| -\frac{1}{T}\log P_X^T(x_1^T) - \mathsf{H}(P_X)\bigg| \geq \eta\bigg) \\
&\leq 2\mathbb{P}_j\bigg(-\frac{1}{T}\log P_X^T(x_1^T) - \mathsf{H}(P_X) \geq \eta\bigg) \\
&\leq 2\exp\bigg(-T\sup_{s\geq 0}(s\eta - \log\mathbb{E}[\exp(-s\log P_{X_1^T}(X_1^T))])\bigg) \\
&\overset{(a)}{\approx} 2\exp\bigg(-T\sup_{s\geq 0}(s\eta - \big(-s\mathbb{E}[\log P_{X_1^T}(X_1^T)] + s^2\mathbb{E}[(\log P_{X_1^T}(X_1^T))^2]\big)\bigg) \\
&\overset{(b)}{=} 2\exp(-\Omega(T\eta^2)) = \exp(-\Omega(T^{\frac{1}{2}})),
\end{aligned}$$

where (a) follows from the Taylor expansion of $\exp(\cdot)$ and $\log(\cdot)$ and (b) follows since the maximum is achieved by $s = O(\eta)$. The rest of the terms in the union bound can be similarly proved.

Thus, the probability of the atypical set is upper bounded by

$$\mathbb{P}_j((X_1^T, \zeta_1^T) \notin \mathcal{A}_{\eta,j}^{(T)}) \leq 3\exp(-\Omega(T^{\frac{1}{2}})) = \exp(-\Omega(T^{\frac{1}{2}})).$$

Let $P_X^* = Q_X$, $\mathcal{Z} \subset \mathbb{Z}$ and design $P_\zeta^* \in \mathcal{P}(\mathcal{Z})$ such that $\mathsf{H}(P_\zeta^*) = \mathsf{H}(P_X^*)$.

For any $\gamma^* \in \Gamma^*$, any $j \in [m]$, the j-th error probability is given by

$$\begin{aligned}
\beta_j(\gamma^*, P_{X_1^T,\zeta_1^T|M=j}^*) &= \sum_{x_1^T,\zeta_1^T} P_{X_1^T,\zeta_1^T|M}^*(x_1^T, \zeta_1^T|j)\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T) \neq j\} \\
&\leq \sum_{(x_1^T,\zeta_1^T)\in\mathcal{A}_{\eta,j}^{(T)}} P_{X_1^T,\zeta_1^T|M}^*(x_1^T, \zeta_1^T|j)\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T) \neq j\} + \exp(-\Omega(T^{\frac{1}{2}})) \\
&= \exp(-\Omega(T^{\frac{1}{2}})) \to 0 \text{ as } T \to \infty.
\end{aligned}$$

For $j = 0$, the worst-case false alarm error probability is upper bounded as follows. For any $x_1^T \in \mathcal{X}^T$,

$$\begin{aligned}
\sum_{\zeta_1^T} P_\zeta^*(\zeta_1^T)\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T) \neq 0\} &\leq \sum_{\zeta_1^T\in\mathcal{A}_{n,\zeta}^{(T)}} P_\zeta^*(\zeta_1^T)\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T) \neq 0\} + \exp(-\Omega(T^{\frac{1}{2}})) \\
&\doteq \sum_{i\in[m]}\sum_{\zeta_1^T\in\mathcal{A}_{n,\zeta}^{(T)}} e^{-T\mathsf{H}(\zeta)}\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T) = i\} + \exp(-\Omega(T^{\frac{1}{2}})) \\
&= me^{-T\mathsf{H}(\zeta)} + \exp(-\Omega(T^{\frac{1}{2}})) \\
&= \alpha + \exp(-\Omega(T^{\frac{1}{2}})) \\
&\xrightarrow{T\to\infty} \alpha.
\end{aligned}$$

Since any distribution Q_X^T can be written as a linear combinations of $\delta_{x_1^T}$, we have

$$\sup_{Q_X}\beta_0(\gamma^*, Q_X \otimes P_\zeta^*) = \sup_{Q_X}\sum_{x_1^T,\zeta_1^T} Q_X^T(x_1^T)P_\zeta^*(\zeta_1^T)\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T) \neq 0\} \leq \alpha$$

A.4 PROOF OF THEOREM 4

First, we have

$$\beta_j(\gamma, P_{X_1^T,\zeta_1^T|M=j}) = \sum_{i:i\neq j} \mathbb{P}_j(\gamma(X_1^T,\zeta_1^T)=i).$$

For any $i \neq j$, the optimization constraints imply that for any $y_1^T \in \mathcal{X}^T$,

$$\alpha \geq \sup_{P_{X_1^T,\zeta_1^T|M=i}} \beta_i(\gamma, P_{X_1^T,\zeta_1^T|M=i}) \geq \sum_{\zeta_1^T} P_{\zeta_1^T}(\zeta_1^T)\mathbb{1}\{\gamma(y_1^T,\zeta_1^T)\neq i\}.$$

Then we have

$$\begin{aligned}\mathbb{P}_j(\gamma(X_1^T,\zeta_1^T)\neq i) &= \sum_{x_1^T,\zeta_1^T} P_{\zeta_1^T}(\zeta_1^T)P_{X_1^T|\zeta_1^T,M=j}(x_1^T|\zeta_1^T,M=j)\mathbb{1}\{\gamma(x_1^T,\zeta_1^T)\neq i\}\\ &\overset{(a)}{\leq} \sum_{x_1^T}(P_{X_1^T}(x_1^T)\wedge\alpha),\end{aligned}$$

where (a) follows since $\sum_{\zeta_1^T} P_{\zeta_1^T}(\zeta_1^T)\mathbb{1}\{\gamma(x_1^T,\zeta_1^T)\neq i\} \leq \alpha$ and $\sum_{\zeta_1^T} P_{\zeta_1^T}(\zeta_1^T)P_{X_1^T|\zeta_1^T,M=j}(x_1^T|\zeta_1^T,M=j)\mathbb{1}\{\gamma(x_1^T,\zeta_1^T)\neq i\} \leq \sum_{\zeta_1^T} P_{\zeta_1^T}(\zeta_1^T)P_{X_1^T|\zeta_1^T,M=j}(x_1^T|\zeta_1^T,M=j) = P_{X_1^T}(x_1^T)$ for all x_1^T.

Consequently,

$$\begin{aligned}\beta_j(\gamma, P_{X_1^T,\zeta_1^T|M=j}) &= \sum_{i:i\neq j}\mathbb{P}_j(\gamma(X_1^T,\zeta_1^T)=i)\\ &\geq \sum_{i:i\neq j}(1-\sum_{x_1^T}(P_{X_1^T}(x_1^T)\wedge\alpha))\\ &= m\sum_{x_1^T}(P_{X_1^T}(x_1^T)-\alpha)_+\\ &\geq \min_{P_{X_1^T}:\mathsf{D}(P_{X_1^T},Q_{X_1^T})\leq d} m\sum_{x_1^T}(P_{X_1^T}(x_1^T)-\alpha)_+,\end{aligned}$$

where m, α should satisfy $m\sum_{x_1^T}(P_{X_1^T}(x_1^T)-\alpha)_+ \leq 1$ and the lower bound holds for all γ and $P_{X_1^T,\zeta_1^T|M}$.

Additionally, the analyses still hold when $P_{\zeta_1^T|M=j}$ are not the same for all j.

A.5 PROOF OF THEOREM 5

Choose $\mathcal{Z} \subset \mathbb{Z}^T$ such that $|\mathcal{Z}|^T = m|\mathcal{X}|^T + 1$.Randomly pick a sequence $\tilde{\zeta_1^T} \in \mathcal{Z}^T$ and partition $\mathcal{Z}^T\backslash\{\tilde{\zeta_1^T}\}$ into m disjoint subsets $\{\mathcal{S}_j\}_{j=1}^m$ of equal size. Define a set of decoders as

$$\Gamma_{\tilde{\zeta_1^T}} := \left\{ \gamma \;\middle|\; \gamma(x_1^T,\zeta_1^T) = \begin{cases} j, & \text{if } \zeta_1^T \neq \tilde{\zeta_1^T} \\ & \text{and } x_1^T = h_j(\zeta_1^T),, \\ 0, & \text{otherwise,}\end{cases} \text{for some group of bijective functions } \{h_j : \mathcal{S}_j \to \mathcal{X}^T\}_{j=1}^m \right\}$$

For any $\gamma \in \Gamma_{\tilde{\zeta_1^T}}$, under the watermarking scheme presented in Theorem 5, we have:

– For any $j \in [m]$, the j-th error probability is give by

$$\begin{aligned}\beta_j(\gamma, P_{X_1^T,\zeta_1^T|M=j}) &= \sum_{i\in[0:m]\backslash j}\mathbb{P}_j(\gamma(X_1^T,\zeta_1^T)=i)\\ &= m\min_{P_{X_1^T}:\mathsf{D}(P_{X_1^T},Q_{X_1^T})\leq d}\sum_{x_1^T}(P_{X_1^T}(x_1^T)-\alpha)_+.\end{aligned}$$

– False alarm error: for any $x_1^T \in \mathcal{X}^T$,

$$\begin{aligned}\sum_{\zeta_1^T} P_\zeta(\zeta_1^T)\mathbb{1}\{\gamma(x_1^T,\zeta_1^T)\neq 0\} &= \sum_{i=1}^{m}\sum_{\zeta_1^T} P_\zeta(\zeta_1^T)\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T)=i\}\\ &= (P^*_{X_1^T}(x_1^T) - m(P^*_{X_1^T}(x_1^T)-\alpha)_+) + (m-1)(P^*_{X_1^T}(x_1^T)-\alpha)_+\\ &= P^*_{X_1^T}(x_1^T) - (P^*_{X_1^T}(x_1^T)-\alpha)_+\\ &= P^*_{X_1^T}(x_1^T)\wedge\alpha \le \alpha.\end{aligned}$$

Since any distribution $Q_{X_1^T}$ can be represented by a linear combination of $\delta_{x_1^T}$, the worst-case false alarm error is upper bounded by

$$\sup_{Q_{X_1^T}} \beta_0(\gamma, P_{X_1^T,\zeta_1^T|M=j}) \le \alpha.$$