FUNDAMENTAL TRADE-OFFS IN MULTI-BIT WATERMARKING OF STOCHASTIC PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study multi-bit watermarking for data generated by stochastic processes, where a hidden message is embedded during sampling and must be decodable by an authorized detector that possesses side information unavailable to unauthorized observers. In high-stakes deployments, a practical watermark must simultaneously control false alarms, preserve generation quality without distorting the output distribution, and support reliable multi-bit decoding. Satisfying all three goals at once inevitably creates fundamental trade-offs. We formulate watermark embedding as a distributional information-embedding problem and watermark detection as a multiple-hypothesis testing problem under distortion and rate constraints, leading to four fundamental metrics: false-alarm probability, detection error probability, distortion, and information rate. Within this information-theoretic framework, we derive matched converse and achievability bounds that characterize the optimal trade-offs and provide scheme-agnostic benchmarks for any watermarking method. For stationary ergodic stochastic processes, we further obtain matched asymptotic limits and connect them to the finite-sample regime. Finally, we present a reference watermarking construction satisfying our assumptions and empirically illustrating the predicted trade-offs.

1 INTRODUCTION

Many modern systems generate data by sampling from stochastic processes specified by a known or learned generative distribution. This umbrella includes classical Markov (Davis, 2018) and autoregressive models, contemporary diffusion models (Song et al., 2020), and large language models (LLMs) (Achiam et al., 2023), and extends to protein/DNA sequence generators (Madani et al., 2020), randomized simulators (Tran et al., 2017), and synthetic data generation (Lu et al., 2023).

Within such stochastic systems, it is increasingly desirable to embed hidden information directly into the generation process so that each generated sample carries verifiable watermarks. The embedded watermarks can certify ownership by identifying the source model or provider, or establish provenance by recording the generation context or other traceable identifiers. It can also support integrity guarantees by making tampering or unauthorized post-processing easier to detect. This goal naturally generalizes existing zero-bit watermarking in generative models (Kirchenbauer et al., 2023; Zhao et al., 2023a; Liu and Bu, 2024; Dathathri et al., 2024; Kuditipudi et al., 2023), which only enables detection of AI-generated content, to multi-bit watermarking (Yoo et al., 2024; Cohen et al., 2025; Qu et al., 2025; Huang et al., 2026), where an authorized detector can not only detect the presence of a watermark but also decode an embedded message. For a more comprehensive literature review, please refer to Appendix A. Despite notable recent advances in watermarking for generative AI (Yang et al., 2025), much of the existing literature remains ad hoc and empirical, and lacks in-depth understanding of the tension between the various performance metrics and how they trade off at a fundamental level.

These trade-offs are especially critical in high-stakes settings, where watermarking must satisfy multiple, often competing, design requirements. Watermarking must tightly control false-alarms to avoid costly misattribution, imposing stringent false positive rate (FPR) constraints on detection (He et al., 2025a). At the same time, it should be distortion-free: watermarked outputs must preserve generation quality and remain statistically indistinguishable from unwatermarked ones (Boroujeny et al., 2024). For multi-bit watermarking, this distribution-preservation requirement must hold uniformly over the message space, meaning that for any embedded message, the output distribution

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

should match the unwatermarked distribution. However, reliable multi-bit decoding pushes in the opposite direction, because increasing payload inevitably perturbs the generative distribution. Therefore, reconciling low false-alarms, reliable multi-bit decoding, and distribution-preserving generation is what makes these trade-offs fundamental. This motivates a foundational question: *how much information can be embedded and reliably decoded from a stochastic process while keeping the resulting distribution indistinguishable from the original?*

This paper formulates multi-bit watermarking of stochastic processes as a distributional information embedding problem. We study how to embed a message into samples from a generative distribution so that an authorized detector can reliably decode it with side information (e.g., secret key), while the resulting outputs remain statistically indistinguishable from unwatermarked samples. The framework abstracts away model-specific details and applies broadly to autoregressive and Markov processes, including language generation and protein sequence generation.

This viewpoint yields a clean formulation of watermark design as constrained optimization: maximize detection and decoding performance subject to worst-case false-alarm constraint and generation distortion. We summarize performance through four facets $(\alpha, \frac{1}{m}\sum_{j=1}^{m}\beta_j, d, R)$: the false-alarm probability α, the average detection error $\frac{1}{m}\sum_{j=1}^{m}\beta_j$, the distortion level d (deviation from the unwatermarked process), and the information rate R (average payload per sample). Framing watermarking in information-theoretic terms puts these objectives in a common language and makes the unavoidable trade-offs explicit. Our results establish an information-theoretic benchmark that any multi-bit watermarking scheme must satisfy, regardless of its embedding mechanism.

Our contribution includes:

- We formulate multi-bit watermark embedding as a distributional information embedding problem and detection as an $(m+1)$-ary hypothesis test performed by an authorized detector with side information, under distortion and rate constraints. This yields four fundamental performance metrics: false-alarm probability, detection error probability, distortion, and information rate.
- We derive matched information-theoretic converse and achievability bounds that characterize the fundamental trade-offs among these metrics, establishing a performance benchmark applicable to any watermarking scheme of a stochastic process within our framework.
- For stationary ergodic stochastic processes, we develop an asymptotic characterization with matched upper and lower bounds, and show how the finite-sample trade-offs converge to the asymptotic limits.
- We propose a reference watermarking construction that satisfies the theoretical assumptions and empirically demonstrates the predicted trade-offs, serving as a concrete instantiation of the theory. Numerical results are presented to visualize the trade-offs and the achievable region.

2 Problem Formulation

2.1 Notation

For any set $\mathcal{X}$, we denote the space of all probability measures over $\mathcal{X}$ by $\mathcal{P}(\mathcal{X})$. For a random variable taking values in $\mathcal{X}$, we use P_X or Q_X to denote its distribution and use the lower-cased letter x to denote a realization of X. For a sequence of random variables $X_1, \ldots, X_n$, and any $i, j \in [n]$ with $i \le j$, we denote $X_i^j := (X_i, \ldots, X_j)$ and drop the subscript if $i = 1$. We may use distortion function, namely, a function $\mathsf{D} : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0, +\infty)$ to measure the dissimilarity between two distributions in $\mathcal{P}(\mathcal{X})$. For example, the total variation distance, as a distortion, between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is $\mathsf{D}_{\mathsf{TV}}(\mu, \nu) := \int \frac{1}{2}|\frac{d\mu}{d\nu} - 1|\, d\nu$. The KL divergence between μ, ν ($\mu \ll \nu$) is $\mathsf{D}_{\mathsf{KL}}(\mu, \nu) := \int \frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu}\, d\nu$. For any set $A \subseteq \mathcal{X}$, we use δ_A to denote its indicator function, namely, $\delta_A(x) := \mathbb{1}\{x \in A\}$. Additionally, we denote $(x)_+ := \max\{x, 0\}$ and $x \wedge y := \min\{x, y\}$.

2.2 Multi-Bit Watermark Embedding

Let $\mathcal{X}$ be a finite alphabet and let $(X_t)_{t \ge 1}$ be a discrete-time stochastic process on $\mathcal{X}$, and $X^T = (X_1, \ldots, X_T) \in \mathcal{X}^T$ be its first T random variables, characterized by a distribution $Q_{X^T} \in \mathcal{P}(\mathcal{X}^T)$. To embed a multi-bit message, let M be uniformly distributed on $[m] := \{1, \ldots, m\}$. In addition to

the generated sequence X^T, we will introduce an *auxiliary* (decoder-observable) sequence that plays the role of side information for watermark decoding[1]. Concretely, let $(\zeta_t)_{t\geq 1}$ denote an auxiliary stochastic process on a finite alphabet $\mathcal{Z}$, and ζ^T be the first T random variables. Watermark embedding is performed at the distributional level. Given M and the data-generation distribution Q_{X^T}, the encoder is a (possibly randomized) mapping $f : [m] \times \mathcal{P}(\mathcal{X}^T) \to \mathcal{P}(\mathcal{X}^T \times \mathcal{Z}^T)$, which maps (M, Q_{X^T}) to a conditional distribution $P_{X^T,\zeta^T|M}(\cdot \mid M)$. Equivalently, for each $m \in [m]$, $f(m, Q_{X^T})$ specifies the conditional law $P_{X^T,\zeta^T|M=m} \in \mathcal{P}(\mathcal{X}^T \times \mathcal{Z}^T)$, inducing a controlled statistical dependence between the generated process X^T and the side-information process ζ^T.

By contrast, conventional post-processing watermarking embeds the message M by modifying a realized data sequence x^T, typically without access to—or explicit modeling of—the underlying data-generation distribution. In comparison, modern distributional embedding operates at the level of the generative distribution, exploiting distributional information available during generation to expand the design space of watermarking schemes while limiting unnecessary perturbations to the generated data sequence. This perspective aligns with both information-theoretic watermarking models and modern generative-model watermarking methods that modify sampling distributions rather than post-process realizations.

Notably, after watermarking, the generated process X^T is jointly distributed according to $P_{X^T} = \mathbb{E}_M[\sum_{\zeta^T \in \mathcal{Z}^T} P_{X^T,\zeta^T|M}]$, which generally differs from the original distribution Q_{X^T}. Additionally, we use the statistical divergence to measure the *distortion level* of a watermarking scheme.

Definition 1 (d-Distorted Watermarking) *A watermark encoder f is said to be d-*distorted *with respect to a distortion function* D*, if for any $M \in [m]$ and $Q_{X^T} \in \mathcal{P}(\mathcal{X}^T)$, the marginal distribution on $\mathcal{X}^T$ induced by the encoder output $P_{X^T,\zeta^T|M}$ satisfies $\mathsf{D}(P_{X^T|M}, Q_{X^T}) \leq d$.*

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

Existing works have considered several other distortion measures, e.g., distance between realizations (Moulin and O'Sullivan, 2000) or divergence between per-step conditional distributions (Kuditipudi et al., 2023; Boroujeny et al., 2024). In our setup, we quantify distortion at the level of joint distributions on the stochastic process rather than individual realizations or per-step conditional distributions. This reflects the fact that watermark embedding operates by modifying the data-generation mechanism of the process rather than deterministically altering a particular realization, and realization-level distortion therefore conflates embedding effects with sampling randomness. Moreover, constraining per-step conditional distributions is unnecessarily restrictive: it reduces the design freedom and feasible region of watermarking schemes, even though generated data quality is inherently a sequence-level property. By defining distortion between joint distributions, we allow coordinated deviations across time while preserving sequence structural properties, enabling a distribution-level non-asymptotic analysis and a principled characterization of the trade-offs among detection errors, distortion, and information rate.

Moreover, to ensure *strong secrecy* of the embedded watermark message, we assume that the embedded message M cannot be decoded by simply observing either the watermarked sequence X^T or the auxiliary sequence ζ^T.

Assumption 1 (Strong Watermark Secrecy) *Watermark embedding must ensure that both X^T and ζ^T are statistically independent of the embedded message M.*

Under Assumption 1, the embedded message M cannot be decoded only with X^T or ζ^T, i.e.,

$$\mathsf{I}(M; X^T) = 0, \;\; \mathsf{I}(M; \zeta^T) = 0, \;\; \text{and } \mathsf{I}(M; X^T, \zeta^T) = \mathsf{I}(M; \zeta^T | X^T) = \mathsf{I}(M; X^T | \zeta^T).$$

To detect if X^T is watermarked, the decoder must exploit its dependency on auxiliary sequence ζ^T. This corresponds to decoding with side information. In multi-bit watermarking, an embedded message may encode traceable information, including user/model IDs, or other sensitive metadata. Requiring strong watermark secrecy ensures such information is not extractable from the watermarked output alone, nor from the side-information alone, and can only be recovered by an authorized detector who

[1] See Appendix B for concrete examples of how side information is used in practical watermarking schemes.

has access to both X^T and ζ^T with their correlation. Without this constraint, the payload may leak either through shifts in the observable output distribution, which leads to unauthorized decoding or easy detection, or trivially through the side-information, which makes decoding independent of the watermarking mechanism. Therefore, strong watermark secrecy formalizes a threat model and rules out degenerate constructions, allowing us to study fundamental trade-offs under distribution-preserving watermarking.

2.3 Statistical Detection and Message Decoding

Within our theoretical framework, we assume that the decoder $\gamma : \mathcal{X}^T \times \mathcal{Z}^T \to [0:m] = \{0, 1, \ldots, m\}$ receives a data-auxiliary sequence pair (X^T, ζ^T) drawn from $P_{X^T,\zeta^T|M}$ to recover the message M, i.e., $\hat{M} = \gamma(X^T, \zeta^T)$. If $\hat{M} = 0$, the sequence X^T is decoded as unwatermarked; if $\hat{M} \in [m]$, X^T is decoded as watermarked with message $\hat{M}$. In practical algorithm design, however, we do not assume that ζ^T is directly accessible by the detector, but extracted from the data sequence via a shared secret key. This system, consisting of the encoder f and decoder γ, defines an (m, T) watermarking scheme with information rate $R \coloneqq \frac{1}{T}\log m$. According to the watermark embedding process, if the stochastic process X^T is unwatermarked or naturally generated, it is independent of ζ^T; otherwise, (X^T, ζ^T) is jointly distributed according to one of the m distributions $\{P_{X^T,\zeta^T|M=j}\}_{j=1}^m$. Thus, detecting and decoding the message M boils down to the $(m+1)$-ary hypothesis testing:

- H_0: X^T is unwatermarked, i.e., $(X^T, \zeta^T) \sim \mathbb{P}_0 \coloneqq P_{X^T,\zeta^T|M=0} = Q_{X^T} \otimes P_{\zeta^T}$;
- $\mathrm{H}_j, \forall j \in [m]$: X^T is watermarked and embedded with message j, i.e., $(X^T, \zeta^T) \sim \mathbb{P}_j \coloneqq P_{X^T,\zeta^T|M=j}$.

Detection performance is measured by the *false-alarm error* and the j*-th error probability*:

$$\beta_0(\gamma, Q_{X^T} \otimes P_{\zeta^T}) \coloneqq \mathbb{P}_0(\gamma(X^T, \zeta^T) \neq 0),$$
$$\beta_j(\gamma, P_{X^T,\zeta^T|M=j}) \coloneqq \mathbb{P}_j(\gamma(X^T, \zeta^T) \neq j), \quad \forall j \in [m],$$

where the j-th error probability contains both the miss-detection error probability and the miss-decoding error probability. Let $\bar{\beta}(\gamma, (\mathbb{P}_j)_{j\in[m]}) \coloneqq \frac{1}{m}\sum_{j\in[m]} \beta_j(\gamma, P_{X^T,\zeta^T|M=j})$ denote the *average detection error probability*. There exists a natural trade-off among the false-alarm error, the average (or the max per-message) detection error probability, the distortion level and the information rate. Under Assumption 1, we aim to characterize this trade-off by solving the following optimization problems with a feasible set defined as

$$\mathcal{F}(\alpha, d) \coloneqq \Big\{(\gamma, (\mathbb{P}_k)_{k\in[m]}) : \sup_{Q_{X^T}} \beta_0(\gamma, Q_{X^T} \otimes P_{\zeta^T}) \leq \alpha,\ \mathsf{D}(P_{X^T}, Q_{X^T}) \leq d\Big\}.$$

We consider two objective criteria over the same feasible set:

$$\min_{(\gamma,(\mathbb{P}_k)_{k\in[m]})\in\mathcal{F}(\alpha,d)} \bar{\beta}(\gamma, (\mathbb{P}_k)_{k\in[m]}), \tag{P1}$$

$$\min_{(\gamma,(\mathbb{P}_k)_{k\in[m]})\in\mathcal{F}(\alpha,d)} \max_{j\in[m]} \beta_j(\gamma, \mathbb{P}_j). \tag{P2}$$

Notably, we consider controlling the *worst-case* false-alarm error $\sup_{Q_{X^T}} \beta_0(\gamma, Q_{X^T} \otimes P_{\zeta^T})$ since naturally generated data may vary widely across domains and sources. This includes human-written text, biological data sequences, synthetic data-generation mechanism and so on, whose underlying distributions are often heterogeneous and unknown.

2.4 Possible Generative Distributions

We consider in-process watermarking of stochastic processes whose generative distributions are accessible. Below, we specify the classes of discrete-alphabet processes covered by our formulation.

Arbitrary discrete-alphabet processes. In the most general setting, the generative distribution Q_{X^T} is an arbitrary probability mass function (PMF) on $\mathcal{X}^T$, with no assumptions on independence, stationarity, ergodicity, or memory. This setting encompasses general discrete stochastic processes, including non-stationary sources and distributions induced by complex generative mechanisms.

Stationary ergodic processes. As a structured special case, we consider stationary ergodic processes. A discrete-time process $(X_t)_{t\geq 1}$ on $\mathcal{X}$ is *stationary* if its finite-dimensional distributions are invariant under time shifts, and *ergodic* if every shift-invariant event has probability zero or one. For a stationary process Q, the *entropy rate* is defined as

$$\mathsf{H}_Q(\mathcal{X}) \coloneqq \lim_{T\to\infty} \frac{1}{T}\mathsf{H}(X^T),$$

provided the limit exists. For stationary ergodic processes with a finite alphabet, this limit exists and is equal to $\mathsf{H}_Q(\mathcal{X}) = \lim_{T\to\infty} \mathsf{H}(X_T \mid X^{T-1})$.

3 Fundamental Trade-Offs for Stochastic Processes with Finite Length

In this section, we characterize the fundamental trade-offs among the detection error, false-alarm error, distortion and information rate through the optimal objective values of (P1) and (P2). We begin by deriving the optimal detection error probability via matching converse and achievability results. This characterization provides a fundamental limit of multi-bit watermark detection in our problem formulation. We then present a more practically achievable decoder construction, which approach the fundamental limit up to a small (and in some cases zero) residual error.

Theorem 1 (Optimal Detection Error Probability) *Under Assumption 1 and allowing randomized decoders, the minimum average detection error probability $\bar{\beta}^*(Q_{X^T}, \alpha, d, m)$ obtained from* (P1) *and the min-max per-message error detection probability $\beta^*_{\max}(Q_{X^T}, \alpha, d, m)$ obtained from* (P2) *coincide and admit the same value*

$$\bar{\beta}^*(Q_{X^T}, \alpha, d, m) = \beta^*_{\max}(Q_{X^T}, \alpha, d, m) = \min_{P_{X^T}:\mathsf{D}(P_{X^T}, Q_{X^T})\leq d} \sum_{x^T\in\mathcal{X}^T} \left(P_{X^T}(x^T) - \tfrac{\alpha}{m}\right)_+ . \quad (1)$$

Moreover, if the distortion measure is chosen as the total variation distance $\mathsf{D}_{\mathsf{TV}}(P_{X^T}, Q_{X^T})$, the optimal value admits the closed-form expression

$$\left(\sum_{x^T} \left(Q_{X^T}(x^T) - \tfrac{\alpha}{m}\right)_+ - \left(d \wedge \sum_{x^T} \left(\tfrac{\alpha}{m} - Q_{X^T}(x^T)\right)_+\right)\right)_+ .$$

Theorem 1 characterizes a fundamental trade-off among the average (or the max per-message) detection error probability, false-alarm error, distortion, and information rate. Specifically, for a fixed blocklength T, increasing the false-alarm error α or the distortion level d, or decreasing the information rate $R = \log m/T$ (for a fixed T), allows for a smaller detection error probability. This trade-off offers a quantitative benchmark for evaluating watermarking methods in stochastic processes, e.g., watermarking for generative models, and identifies an impossibility regime determined by the underlying model and secrecy constraints. We will further demonstrate the trade-off among the four facets graphically in Section 5.

Moreover, Theorem 1 reveals an consequence of the secrecy constraint. Under Assumption 1, the observable distribution of X^T is identical across messages, preventing the decoder from statistically distinguishing among them based on the observation alone. Consequently, all messages must be symmetrically, and the optimal average detection error coincides with the optimal maximum per-message error. The resulting expression further shows that the false-alarm budget α is evenly allocated across messages, so increasing the message set size tightens the effective false-alarm constraint per message and increases the detection error probability. As a sanity check, the special case $m = 1$ recovers the zero-bit watermarking result of He et al. (2025b). However, our result is not a direct extension of that from He et al. (2025b). The multi-bit setting introduces a strong secrecy constraint and an $(m+1)$-ary hypothesis testing problem with m coupled message-dependent distributions, requiring new converse arguments and an achievability construction based on stochastic grouping and a Latin-square structure. Moreover, our asymptotic analysis in Section 4 yields a new characterization of the maximum information rate, which has no counterpart in the zero-bit setting.

The proof of Theorem 1 consists of a converse and an achievability. In the converse (proved in Appendix C), we establish a fundamental lower bound for the optimal objective values of (P1) and (P2), which holds for both randomized and deterministic decoders. In the achievability, we show that, for a feasible regime of rate, there exist watermarking schemes with randomized decoders

attaining this lower bound. When the decoder is deterministic, we propose an explicit scheme that approaches the lower bound up to a small (and in some cases zero) residual error.

Converse (Sketch) for Theorem 1. The converse establishes a lower bound on the average and maximum per-message detection error probabilities by first fixing an arbitrary decoder and exploiting the constraint that the worst-case false-alarm error is bounded above by α. We then show that this lower bound is optimized when the false-alarm budget is evenly allocated across messages and the watermarked source distribution is optimized, yielding a bound that depends only on (Q_{X^T}, d, α, m) and is independent of the decoder or watermarking joint distribution. As a result, the bound holds uniformly over the entire feasible region. See Appendix C for a complete proof.

3.1 Achievability for Theorem 1

In the converse, the optimal detection error is lower bounded by the RHS of (1) for any m. In this section, we construct the watermarking schemes with decoders using randomized grouping that achieve this lower bound, and further provide an explicit scheme with a decoder using deterministic grouping that attains the bound up to a small residual error ε_{res} for m in a feasible regime. Here, grouping refers to mapping symbols in the source alphabet to a smaller set of group labels, thereby inducing a group distribution. The proposed embedding–detection scheme is intentionally simple: it serves as a constructive realization of the theoretical benchmark and demonstrates the achievability of the derived bounds, rather than aiming for state-of-the-art empirical performance. Before stating the main results, we let $P^*_{X^T}$ be an optimizer of (1): $P^*_{X^T} \coloneqq \arg\min_{P_{X^T}:\mathsf{D}(P_{X^T},Q_{X^T})\le d} \sum_{x^T}(P_{X^T}(x^T) - \frac{\alpha}{m})_+$, and define the "heavy" set $\mathcal{S}_{\alpha/m} \coloneqq \{x^T : P^*_{X^T}(x^T) \ge \frac{\alpha}{m}\}$.

3.2 Decoding with Randomized Grouping

Theorem 2 (Optimal Watermark Embedding and Detection) *Assume Assumption 1. Fix Q_{X^T}, false-alarm constraint α, distortion level d, and message set size $m \le m^* \coloneqq \max\{m \in [|\mathcal{X}|^T] : \bar{\beta}^*(Q_{X^T}, \alpha, d, m) \le 1-\alpha\}$. Under decoding with randomized grouping of source distribution, there exist a watermark embedding distribution $P^*_{\zeta^T|X^T,M}$ and a decoder $\gamma^* : \mathcal{X}^T \times \mathcal{Z}^T \in [0:m]$ that satisfy the distortion and false-alarm constraints, and achieve the optimal value in* (1).

Theorem 2 identifies the information-theoretic limit when randomized grouping is permitted: the achieved detection performance matches the converse benchmark (1) uniformly over all Q_{X^T}. The detailed achievability construction involves stochastically grouping the source distribution into at least m equal bins and designing a decoder based on a Latin square, provided in Appendix D.1.From a constructive viewpoint, randomization effectively enables "fractional" allocation of probability mass at the threshold α/m, which is crucial for matching the converse exactly. This theorem therefore serves as the baseline against which we evaluate practical deterministic decoding.

The randomized grouping construction, however, requires an additional shared secret seed to reproduce the grouping at detection time, introducing overhead in secret-seed management and coordination between the encoder and decoder. To avoid this overhead, we develop another explicit construction based on deterministic grouping that approximates the optimal performance achieved by randomized grouping. This construction retains the core ideas of the randomized scheme while simplifying both watermark embedding and decoding. The key idea is to first *group* source sequences x^T into a finite number of bins and then carry out watermark embedding and decoding at the *group level*. Intuitively, this grouping step creates at least m "heavy" groups with mass $\ge \alpha/m$, which can reliably support the m messages; the side information ζ^T is then used to assign messages within heavy groups, while the decoder outputs 0 whenever the pair (x^T, ζ^T) falls outside the designed decodable region.

Construction (grouping → embedding → decoding). Let $K \ge m$ and let $\phi : \mathcal{X}^T \to [K]$ be a deterministic grouping function, inducing a group distribution P_G via $P_G(k) = P^*_{X^T}(\phi(X^T) = k)$. Define $\mathcal{G}_{\alpha/m} \coloneqq \{k \in [K] : P_G(k) \ge \alpha/m\}$.

Step 1 (Grouping). Choose $\phi^\star$ as a solution to

$$\min_{\phi:\mathcal{X}^T\to[K]} \quad \sum_{k=1}^{K}\Big(P_G(k) - \tfrac{\alpha}{m}\Big)_+ \quad \text{s.t.} \quad |\{k : P_G(k) \ge \tfrac{\alpha}{m}\}| \ge m,$$

with the additional constraint that every $x^T \in \mathcal{S}_{\alpha/m}$ is assigned to a singleton group.

Step 2 (Embedding). Let $\mathcal{Z}^T = \mathcal{Z}^T_{\geq} \cup \mathcal{Z}^T_{<} \cup \{\widetilde{\zeta^T}\}$ be a finite alphabet, where $|\mathcal{Z}^T_{\geq}| = |\mathcal{G}_{\alpha/m}|$, $|\mathcal{Z}^T_{<}| = K - |\mathcal{G}_{\alpha/m}|$ and $\widetilde{\zeta^T}$ is an additional redundant state. For each $j \in [m]$, define

$$P^*_{\zeta^T|X^T,M}(\zeta^T|x^T, j) = P^*_{\zeta^T|G,M}(\zeta^T|\phi^\star(x^T), j),$$

where for all $k \in [K]$,

$$P^*_{\zeta^T|G,M}(\zeta^T|k, j) = \Big(1 \wedge \tfrac{\alpha}{mP_G(k)}\Big)\mathbb{1}\{h(k, \zeta^T) = j\} + \mathbb{1}\{\zeta^T \in \mathcal{Z}^T_{<},\, g(\zeta^T) = k\} + \Big(1 - \tfrac{\alpha}{mP_G(k)}\Big)_+ \mathbb{1}\{\zeta^T = \widetilde{\zeta^T}\},$$

with $h : \mathcal{G}_{\alpha/m} \times \mathcal{Z}^T_{\geq} \to [|\mathcal{G}_{\alpha/m}|]$ bijective in either argument when the other is fixed, and $g : \mathcal{Z}^T_{<} \to [K] \setminus \mathcal{G}_{\alpha/m}$ bijective.

Step 3 (Decoding). Define the deterministic decoder $\gamma^\star : \mathcal{X}^T \times \mathcal{Z}^T \to \{0, 1, \ldots, m\}$ by

$$\gamma^\star(x^T, \zeta^T) = \begin{cases} j, & \text{if } \zeta^T \in \mathcal{Z}^T_{\geq} \text{ and } h(\phi^\star(x^T), \zeta^T) = j \leq m, \\ 0, & \text{otherwise.} \end{cases}$$

3.3 Decoding with Deterministic Grouping

We are now in a position to present our achievability result under decoding with deterministic grouping.

Theorem 3 (Decoding with Deterministic Grouping: Achievability and Residual Error) *Under Assumption 1 and fix Q_{X^T}, α, and d. For any $m \in \mathcal{M} := \Big\{m \in [|\mathcal{X}|^T] : \sum_{x^T \notin \mathcal{S}_{\alpha/m}} P^*_{X^T}(x^T) \geq \big(1 - |\mathcal{S}_{\alpha/m}|/m\big)\alpha\Big\}$, the above construction produces a deterministic decoder $\gamma^\star$ and an embedding distribution $P^*_{\zeta^T|X^T,M}$ that satisfy the distortion and false-alarm constraints. Moreover, the resulting detection error matches the randomized converse benchmark up to a residual term:*

$$\beta_{\mathrm{det}}(Q_{X^T}, \alpha, d, m) \leq \bar{\beta}^*(Q_{X^T}, \alpha, d, m) + \varepsilon_{\mathrm{res}}(Q_{X^T}, \alpha, d, m),$$

where $\varepsilon_{\mathrm{res}}$ (cf. (13) in Appendix D.2) is the residual error incurred by deterministic grouping on the discrete alphabet $\mathcal{X}^T$. In particular, $\varepsilon_{\mathrm{res}} = 0$ whenever the grouping admits exact mass balancing.

Theorem 3 formalizes the price of determinism. Under strong secrecy, the side information ζ^T cannot carry message information by itself and effectively acts as shared randomness; hence reliable multi-bit decoding must be supported by the observed process X^T. Decoding with randomized grouping can match the converse benchmark exactly, while decoding with deterministic grouping approximates it via a concrete grouping-based design. In general, a nonzero residual error $\varepsilon_{\mathrm{res}}$ is unavoidable: deterministic grouping partitions the finite set $\mathcal{X}^T$ and therefore cannot, in general, realize an exact "fractional" allocation of probability mass needed to match the randomized optimum for arbitrary source distribution Q_{X^T}.

We restrict attention to $m \leq |\mathcal{X}|^T$ and do not consider $m > |\mathcal{X}|^T$. Indeed, since X^T takes values in the finite alphabet $\mathcal{X}^T$, the decoder observes at most $|\mathcal{X}|^T$ distinct realizations. Under deterministic decoding and strong secrecy ($P_{\zeta^T|M=j} = P_{\zeta^T}$ for all j), the auxiliary randomness ζ^T cannot convey message information by itself; thus message distinguishability must come from X^T, which fundamentally limits reliable recovery to $m \leq |\mathcal{X}|^T$ (equivalently, $R = \frac{\log m}{T} \leq \log|\mathcal{X}|$). If $m > |\mathcal{X}|^T$, distinct messages are necessarily indistinguishable from the observation under any deterministic scheme unless one relaxes the decoding criterion or enlarges the observation model.

Finally, we note that the feasible regime of m also depends on the optimal source distribution $P^*_{X^T}$ and the false-alarm constraint α. In particular, when $P^*_{X^T}$ is uniform over $\mathcal{X}^T$, it is possible to support the maximal message set size $m = |\mathcal{X}|^T$.

4 Fundamental Trade-Offs for Stationary Ergodic Processes with Infinite Length

In this section, we present the asymptotic analysis for stationary ergodic processes by letting the block-length $T \to \infty$. Specifically, let P and Q be stationary ergodic measures on $\mathcal{X}^{\mathbb{N}}$. The resulting characterization reveals the fundamental performance limits achievable under stationarity and ergodicity, and serves as a benchmark for understanding the finite-length trade-offs developed earlier. We will explicitly compare the asymptotic and finite-length results at the end of this section. Define the relative entropy rate between two stationary ergodic processes as $\overline{\mathsf{D}}_{\mathsf{KL}}(P\|Q) \coloneqq \lim_{T\to\infty} \frac{1}{T}\mathsf{D}_{\mathsf{KL}}(P_{X^T}\|Q_{X^T})$.

Theorem 4 (Asymptotically Optimality for Stationary Ergodic Processes) *Assume Assumption 1. Fix Q be a stationary ergodic process (SEP) on the finite alphabet $\mathcal{X}$, false-alarm constraint α, distortion level d. For any SEP P satisfying $\mathsf{D}_{\mathsf{KL}}(P_{X^T}\|Q_{X^T}) \le dT$ for all T, and any (dT)-distorted (m,T) watermarking scheme with message set $[m]$ and any associated tuple $(\mathbb{P}_k)_{k\in[m]}$, if the average detection error $\bar{\beta}(\gamma,(\mathbb{P}_k)_{k\in[m]}) \to 0$ as $T\to\infty$, then the achievable information rate is upper bounded by the maximum entropy rate:*

$$R \le \max_{P:\overline{\mathsf{D}}_{\mathsf{KL}}(P\|Q)\le d} \mathsf{H}_P(\mathcal{X}).$$

When $d = 0$, for any $m \le \alpha\exp(T\mathsf{H}_Q(\mathcal{X}))$, there exist an asymptotically optimal deterministic decoder γ^*, and an optimal distortion-free watermarking scheme that achieve the detection performance: as $T\to\infty$, for all $j\in[m]$,

$$\beta_j(\gamma^*,\mathbb{P}_j^*) \le \exp(-\Omega(T^{\frac{1}{2}})) \to 0, \quad \sup_{Q_{X^T}} \beta_0(\gamma^*,\mathbb{P}_0^*) \le \alpha + \exp(-\Omega(T^{\frac{1}{2}})) \to \alpha.$$

Theorem 4 presents the asymptotically optimal information rate for watermarking on a stationary ergodic process, which is equal to its entropy rate when $d = 0$. This maximum information rate coincides with the steganography capacity characterized in classical works without adversary and noise (Merhav and Sabbag, 2008; Harmsen and Pearlman, 2009). As the distortion increases, a (dT)-distorted watermarking can trade off text quality to achieve a higher rate.

Proof Sketch for Theorem 4. Here we provide main ideas of the converse and achievability proofs.

- **Converse:** We apply Fano's inequality to obtain the maximum information rate that a dT-distorted (m,T) watermarking scheme can achieve on stochastic ergodic processes, with vanishing average detection error $\bar{\beta}(\gamma,(\mathbb{P}_k)_{k\in[m]})$. Specifically, we leverage the strong secrecy Assumption 1, where $\mathsf{H}(M) = \mathsf{H}(M|\zeta^T)$. The details are in Appendix E.
- **Achievability:** We identify the asymptotically optimal watermarking scheme that achieves vanishing j-th error probabilities and the maximum information rate, while satisfying the worst-case false-alarm error and distortion constraint. In particular, by applying asymptotic equipartition property and the techniques of typical sets in information theory, the asymptotically optimal decoder deterministically maps a typical sequence x^T to a typical sequence ζ^T uniquely under different messages M. The corresponding optimal joint distribution $P^{*T}_{X,\zeta|X,M}$ assigns probability 1 to such pair of sequences (x^T,ζ^T), making sure that the detection accuracy is high. The details are in Appendix F.

Remark 1 (Comparison with Finite-Sample Analysis) *Consider Theorem 1 and Theorem 2 under zero distortion ($d=0$). For certain source distributions Q_{X^T}, the maximum achievable message set size (in bits) is $\log m^* = \mathsf{H}(Q_{X^T})$, which coincides with the asymptotic entropy-rate characterization. For example, when Q_{X^T} is uniform over $\mathcal{X}^T$, we have $\log m^* = \log|\mathcal{X}|^T$. Specifically, as $T\to\infty$, $Q_{X^T}(\mathcal{A}^{(T)}_{\eta,X}) \to 1$ and Q_{X^T} is asymptotically uniform over this set. Then we have the matching rate $(\log|\mathcal{A}^{(T)}_{\eta,X}|)/T \to \mathsf{H}_Q(\mathcal{X})$. Under certain under appropriate scaling of the false-alarm constraint, the finite-sample optimal detection error characterized in Theorem 1 also aligns with the asymptotic result. Under the same example, the optimal detection error probability in* (1) *reduces to $\sum_{x^T}\left(\frac{1}{|\mathcal{X}|^T} - \frac{\alpha}{|\mathcal{X}|^T}\right)_+ = 1-\alpha$. If the false-alarm constraint $\alpha = 1-\exp(-\Omega(T^{\frac{1}{2}}))$ increases to 1 exponentially fast with T, then the finite-sample optimal detection error exhibits the same decay behavior as the asymptotic detection error in Theorem 4.*

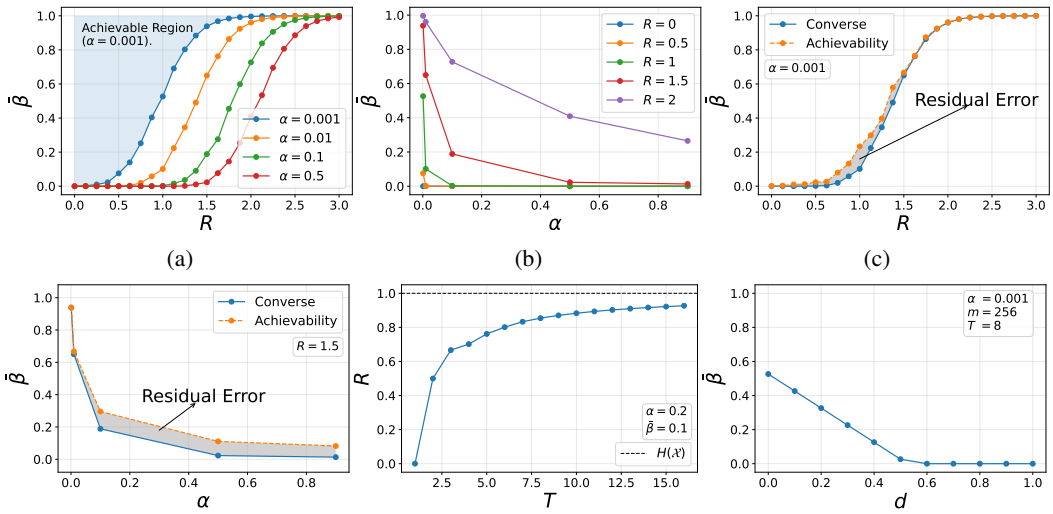


Figure 1: Numerical results on stationary ergodic Markov chain.

5 Numerical Results

In this section, we numerically compute and visualize the trade-offs characterized in Theorem 1. We also illustrate the gap between the deterministic-decoder achievability bound and the converse, and compare the finite-sample information rate to the asymptotically optimal rate for stationary ergodic stochastic processes, i.e., the entropy rate. For computational tractability, we consider a finite-state stationary ergodic Markov model.

We first present the trade-offs across three facets $(\alpha, \bar{\beta}, R)$ under zero distortion ($d = 0$) for a fixed-length process. Figure 1a presents the **rate-detection** trade-off under a fixed false-alarm error α. Clearly, as the rate R increases, the average detection error probability increases as well. The region above each curve is the achievable region for any watermarking scheme within our framework. As α increases, the achievable region enlarges. Figure 1b presents the **false alarm-detection** trade-off under a fixed information rate R. The average detection error decreases as the false-alarm error increases. Similarly, the region above each curve is the achievable region. When the rate R is small enough, almost the entire region is achievable.

Figures 1c and 1d illustrate the **gap** between the converse lower bound on the average detection error and the performance achieved by the reference watermarking scheme with a deterministic decoder (cf. Theorem 3), as quantified by the residual error. We observe that this residual error is small, indicating that the proposed reference scheme—despite its simplicity—achieves performance close to the fundamental limit.

Figure 1e shows the **length-rate** trade-off and that the finite-length achievable information rate approaches the fundamental limit, the entropy rate, as the sequence length T increases under certain generative distributions. Figure 1f further shows the **distortion-detection** trade-off when considering total variation distortion. Clearly, by allowing higher distortion level, we have more budget to optimize the watermark embedding distribution and enhance watermark detection performance.

These numerical results provide a concrete instantiation of our theoretical framework, illustrating the fundamental information-theoretic trade-offs governing multi-bit watermark embedding in stochastic processes generated from a prescribed distribution. Together with our converse and achievability results, they validate the tightness of the derived bounds and clarify the impact of false-alarm control, distortion, and information rate in finite and asymptotic regimes. Beyond serving as a proof of concept, our analysis is intended to guide the design of future practical watermarking algorithms that operate at the distributional level and approach fundamental performance limits.

6 AI USE STATEMENT

We used generative AI tools to improve the clarity and readability of the manuscript and to assist in checking the proof of Theorem 2. The authors reviewed the AI-assisted revisions and feedback and take full responsibility for the final content of the paper, including the correctness of all mathematical claims and proofs.

REFERENCES

Mark HA Davis. *Markov models & optimization*. Routledge, 2018.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. *Advances in Neural Information Processing Systems*, 30, 2017.

Yingzhou Lu, Lulu Chen, Yuanyuan Zhang, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*, 2023.

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*, pages 17061–17084. PMLR, 2023.

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

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

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 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)*, pages 4031–4055, 2024.

Aloni Cohen, Alexander Hoover, and Gabe Schoenbach. Watermarking language models for many adaptive users. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2583–2601. IEEE, 2025.

Wenjie Qu, Wengrui Zheng, Tianyang Tao, Dong Yin, Yanze Jiang, Zhihua Tian, Wei Zou, Jinyuan Jia, and Jiaheng Zhang. Provably robust multi-bit watermarking for {AI-generated} text. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 201–220, 2025.

Yu-Shin Huang, Chao Tian, and Krishna Narayanan. Optimal multi-bit generative watermarking schemes under worst-case false-alarm constraints, 2026. URL https://arxiv.org/abs/2604.08759.

Zhiguang Yang, Gejian Zhao, and Hanzhou Wu. Watermarking for large language models: A survey. *Mathematics*, 13(9):1420, 2025.

Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. Distributional information embedding: A framework for multi-bit watermarking. *arXiv preprint arXiv:2501.16558*, 2025a.

Massieh Kordi Boroujeny, Ya Jiang, Kai Zeng, and Brian Mark. Multi-bit distortion-free watermarking for large language models. *arXiv preprint arXiv:2402.16578*, 2024.

P. Moulin and J.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, pages 3630–3633 vol.6, 2000. doi: 10.1109/ICASSP.2000.860188.

Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. Theoretically grounded framework for LLM watermarking: A distribution-adaptive approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Neri Merhav and Erez Sabbag. Optimal watermark embedding and detection strategies under limited detection resources. *IEEE Transactions on Information Theory*, 54(1):255–274, 2008. doi: 10.1109/TIT.2007.911210.

Jeremiah J Harmsen and William A Pearlman. Capacity of steganographic channels. *IEEE Transactions on Information Theory*, 55(4):1775–1792, 2009.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57(2):1–36, 2024a.

Yepeng Liu, Xuandong Zhao, Dawn Song, Gregory W Wornell, and Yuheng Bu. Position: Llm watermarking should align stakeholders' incentives for practical adoption. *arXiv preprint arXiv:2510.18333*, 2025a.

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, 2024.

Mucong Ding, Tahseen Rabbani, Bang An, Souradip Chakraborty, Chenghao Deng, Mehrdad Saberi, Yuxin Wen, Xuandong Zhao, Mo Zhou, Anirudh Satheesh, et al. Erasing the invisible: A stress-test challenge for image watermarks. In *NeurIPS 2024 Competition Track*, 2024.

Yepeng Liu, Yiren Song, Hai Ci, Yu Zhang, Haofan Wang, Mike Zheng Shou, and Yuheng Bu. Image watermarks are removable using controllable regeneration from clean noise. *arXiv preprint arXiv:2410.05470*, 2024b.

Siqi Hui, Yiren Song, Sanping Zhou, Ye Deng, Wenli Huang, and Jinjun Wang. Autoregressive images watermarking through lexical biasing: An approach resistant to regeneration attack. *arXiv preprint arXiv:2506.01011*, 2025.

Yanshuo Chen, Zhengmian Hu, Yihan Wu, Ruibo Chen, Yongrui Jin, Marcus Zhan, Chengjin Xie, Wei Chen, and Heng Huang. Enhancing privacy in biosecurity with watermarked protein design. *Bioinformatics*, page btaf141, 2025.

Yepeng Liu, Xuandong Zhao, Dawn Song, and Yuheng Bu. Dataset protection via watermarked canaries in retrieval-augmented llms. *arXiv preprint arXiv:2502.10673*, 2025b.

Xiaojun Xu, Yuanshun Yao, and Yang Liu. Learning to watermark llm-generated text via reinforcement learning. *arXiv preprint arXiv:2403.10553*, 2024a.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. Protecting language generation models via invisible watermarking. In *International Conference on Machine Learning*, pages 42187–42199. PMLR, 2023b.

Yepeng Liu, Xuandong Zhao, Christopher Kruegel, Dawn Song, and Yuheng Bu. In-context watermarks for large language models. *arXiv preprint arXiv:2505.16934*, 2025c.

Agnibh Dasgupta, Abdullah Tanvir, and Xin Zhong. Watermarking language models through language models. *arXiv preprint arXiv:2411.05091*, 2024.

Li An, Yujian Liu, Yepeng Liu, Yang Zhang, Yuheng Bu, and Shiyu Chang. Defending llm watermarking against spoofing attacks with contrastive representation learning. *arXiv preprint arXiv:2504.06575*, 2025a.

Li An, Yujian Liu, Yepeng Liu, Yuheng Bu, Yang Zhang, and Shiyu Chang. A reinforcement learning framework for robust and secure llm watermarking. *arXiv preprint arXiv:2510.21053*, 2025b.

Yapei Chang, Kalpesh Krishna, Amir Houmansadr, John Frederick Wieting, and Mohit Iyyer. Postmark: A robust blackbox watermark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8969–8987, 2024.

Yizhou Zhao, Xiang Li, Peter Song, Qi Long, and Weijie Su. Tab-drw: A dft-based robust watermark for generative tabular data. *arXiv preprint arXiv:2511.21600*, 2025.

T Tony Cai, Xiang Li, Qi Long, Weijie J Su, and Garrett G Wen. Optimal detection for language watermarks with pseudorandom collision. *arXiv preprint arXiv:2510.22007*, 2025.

Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. Robust detection of watermarks for large language models under human edits. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf056, 2025a.

Weiqing He, Xiang Li, Tianqi Shang, Li Shen, Weijie Su, and Qi Long. On the empirical power of goodness-of-fit tests in watermark detection. *arXiv preprint arXiv:2510.03944*, 2025c.

Xiang Li, Garrett Wen, Weiqing He, Jiayuan Wu, Qi Long, and Weijie J Su. Optimal estimation of watermark proportions in hybrid ai-human texts. *arXiv preprint arXiv:2506.22343*, 2025b.

Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics*, 53(1):322–351, 2025c.

Dor Tsur, Carol Xuan Long, Claudio Mayrink Verdun, Sajani Vithana, Hsiang Hsu, Chun-Fu Chen, Haim H Permuter, and Flavio Calmon. Heavywater and simplexwater: Distortion-free llm watermarks for low-entropy distributions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Baihe Huang, Banghua Zhu, Hanlin Zhu, Jason D Lee, Jiantao Jiao, and Michael I Jordan. Towards optimal statistical watermarking. *arXiv preprint arXiv:2312.07930*, 2023.

Baihe Huang, Hanlin Zhu, Julien Piet, Banghua Zhu, Jason D. Lee, Kannan Ramchandran, Michael Jordan, and Jiantao Jiao. Watermarking using semantic-aware speculative sampling: from theory to practice, 2025. URL `https://openreview.net/forum?id=LdIlnsePNt`.

Miranda Christ and Sam Gunn. Pseudorandom error-correcting codes. In *Annual International Cryptology Conference*, pages 325–347. Springer, 2024.

Danilo Francati, Yevin Nikhel Goonatilake, Shubham Pawar, Daniele Venturi, and Giuseppe Ateniese. The coding limits of robust watermarking for generative models. *arXiv preprint arXiv:2509.10577*, 2025.

Liying Li, Yihan Bai, and Minhao Cheng. Where am i from? identifying origin of llm-generated content. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12218–12229, 2024.

Haoyu Jiang, Xuhong Wang, Ping Yi, Shanzhe Lei, and Yilun Lin. Credid: Credible multi-bit watermark for large language models identification. *arXiv preprint arXiv:2412.03107*, 2024.

Xiaojun Xu, Jinghan Jia, Yuanshun Yao, Yang Liu, and Hang Li. Robust multi-bit text watermark with llm-based paraphrasers. *arXiv preprint arXiv:2412.03123*, 2024b.

Or Zamir. Undetectable steganography for language models. *Transactions on Machine Learning Research*, 2024.

Sam Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. *arXiv preprint arXiv:2410.07369*, 2024.

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

E.J. Gumbel. *Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures*. Applied mathematics series. U.S. Government Printing Office, 1954. URL https://books.google.com/books?id=SNpJAAAAMAAJ.

A Related Works

Watermarking techniques have been widely explored across various stochastic processes, including text (Liu et al., 2024a; 2025a; Zhao et al., 2024), image (Ding et al., 2024; Liu et al., 2024b; Hui et al., 2025), and protein sequence generation (Chen et al., 2025) to support traceability and detection of synthetic content. In general, existing approaches fall into two main categories: zero-bit watermarking for mere detection and multi-bit watermarking for conveying auxiliary messages.

Zero-bit Watermark. A zero-bit watermark is a type of watermarking scheme that only indicates whether a watermark is present or not, without encoding any additional information. One line of work embeds a watermark into a sequence during the decoding process, for example, by manipulating the decoding strategy (Kirchenbauer et al., 2023; Zhao et al., 2023a; Liu and Bu, 2024; Liu et al., 2025b; Dathathri et al., 2024; Kuditipudi et al., 2023), applying post-training watermarking methods (Xu et al., 2024a; Zhao et al., 2023b), or setting a watermarking instruction (Liu et al., 2025c; Dasgupta et al., 2024). Another line of work post-processes a generated sequence to embed a watermark (An et al., 2025a;b; Chang et al., 2024; Zhao et al., 2025). Moreover, some works provide a theoretical standpoint instead of a heuristic design of the watermark (Cai et al., 2025; Li et al., 2025a; He et al., 2025c; Li et al., 2025b;c; Tsur et al., 2025; Huang et al., 2023; 2025). Specifically, Christ and Gunn (2024) applies pseudorandom error-correcting codes for watermarking to make watermarked text computationally indistinguishable from unwatermarked ones to efficient adversaries. Francati et al. (2025) establishes optimal robustness guarantees for watermarking schemes based on pseudorandom error-correcting codes. He et al. (2025b) proposes a theoretical framework that jointly optimizes the watermarking scheme and detector, showing that watermarks should be adaptive to the LLM's output distribution to achieve the best detectability–distortion trade-off. Li et al. (2025c) casts LLM watermark detection as hypothesis testing with pivotal statistics, providing provable Type-I error control and minimax-optimal detection under distributional uncertainty.

Multi-bit Watermark. Multi-bit watermarking embeds multiple bits to enable both detection and message transmission (He et al., 2025a; Li et al., 2024). Compared to the extensive literature on zero-bit watermark, research on multi-bit watermark remains sparse, with most existing methods relying on heuristic designs rather than rigorous theoretical analysis. Some works (Yoo et al., 2024; Jiang et al., 2024; Cohen et al., 2025; Qu et al., 2025) divide the generated text into blocks and embeds a watermark bit in each by using different red/green lists to represent the binary values 0 and 1. While some studies (Xu et al., 2024b; Zamir, 2024) predefine a mapping rule or use cryptographic techniques to determine which bit carries the watermark, and the work of Gunn et al. (2024) further extends this idea to image generation model. For theoretical analysis, Boroujeny et al. (2024) studies the tradeoff between embedding information rate and decoding error probability, by enlarging the message set to increase the embedding rate while designing a distortion-free scheme that preserves the model's token distribution. However, a unified information-theoretic analysis of multi-bit watermark remains under explored. We propose the first general theoretical framework to study the fundamental limits of distribution information embedding with side information.

B More Discussion on Side Information-based Formulation

• **Green-Red List watermarking** (Kirchenbauer et al., 2023). In this scheme, at each time t, the vocabulary $\mathcal{V}$ is randomly partitioned into a green set $\mathcal{G}$ and a red set $\mathcal{R}$, with $|\mathcal{G}| = \rho|\mathcal{V}|$ for some $\rho \in (0,1)$. This partition can be encoded by a binary side-information vector $\zeta_t \in \{0,1\}^{|\mathcal{V}|}$ indexed by $x \in \mathcal{V}$, where $\zeta_t(x) = 1$ indicates $x \in \mathcal{G}$ (and $\zeta_t(x) = 0$ indicates $x \in \mathcal{R}$).

The procedure can be summarized as:

- Compute a hash of the previous token using a shared secret: $h(X_{t-1}, \texttt{key})$, where $h : \mathcal{V} \times \mathbb{R} \to \mathbb{R}$.
- Use $h(X_{t-1}, \texttt{key})$ as a seed to draw ζ_t uniformly from $\{\zeta \in \{0,1\}^{|\mathcal{V}|} : \|\zeta\|_1 = \rho|\mathcal{V}|\}$, which determines $\mathcal{G}$.
- Draw X_t from a modified next-token distribution that upweights green tokens by a logit shift $\delta > 0$:

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

The same side-information perspective also captures several other watermarking schemes in the literature.

• **Gumbel-Max watermarking** (Aaronson, 2023). This method uses the Gumbel-Max trick (Gumbel, 1954), where the randomness injected by the Gumbel variables plays the role of side information. Here, ζ_t is a $|\mathcal{V}|$-dimensional vector indexed by $x \in \mathcal{V}$. For $t = 1, 2, \ldots$:

– Compute a seed from the previous n tokens and the secret key: $h(X_{t-1}^{t-n}, \texttt{key})$, with $h : \mathcal{V}^n \times \mathbb{R} \to \mathbb{R}$.
– Use this seed to sample ζ_t uniformly from $[0,1]^{|\mathcal{V}|}$.
– Generate the next token via

$$X_t = \arg\max_{x \in \mathcal{V}} \ \log Q_{X_t|x_1^{t-1}}(x) - \log\big(-\log \zeta_t(x)\big).$$

• **Inverse-transform watermarking** (Kuditipudi et al., 2023). This scheme treats $\mathcal{V}$ as $[|\mathcal{V}|]$ and uses side information consisting of a uniform random variable together with a random permutation. Specifically, let $\zeta_t = (U_t, \pi_t)$, where $U_t \sim \mathrm{Unif}[0,1]$ and π_t is a uniform permutation of $[|\mathcal{V}|]$, both generated from the shared key.

– Using key as a seed, sample $\{U_t\}_{t=1}^T$ i.i.d. from $[0,1]$ and $\{\pi_t\}_{t=1}^T$ i.i.d. from the permutation group over $[|\mathcal{V}|]$.
– Sample X_t by inverse transform with respect to the permuted ordering:

$$X_t = \pi_t^{-1}\left(\min\left\{ i \in [|\mathcal{V}|] : \sum_{x \in [|\mathcal{V}|]} Q_{X_t|x_1^{t-1}}(x)\, \mathbb{1}\{\pi_t(x) \le i\} \ge U_t \right\}\right),$$

where π_t^{-1} denotes the inverse permutation.

• **Adaptive watermarking** (Liu and Bu, 2024). This approach is structurally similar to Green–Red List watermarking, but replaces the hash with a pretrained neural network. Concretely, a network h takes a semantic representation $\phi(X_1^{t-1})$ of the generated prefix together with the secret key and outputs a seed $h(\phi(X_1^{t-1}), \texttt{key})$. Using this seed, the scheme samples ζ_t uniformly from $\{\mathbf{v} \in \{0,1\}^{|\mathcal{V}|} : \|\mathbf{v}\|_1 = \rho|\mathcal{V}|\}$ to determine the green set, and then draws X_t using the same logit-shifted sampling rule as in the Green–Red List method.

C Proof of Converse for Theorem 1

The converse proof holds for both deterministic and randomized decoders. In the following, for the sake of generality, we consider randomized decoders $\gamma(\cdot|x^T, \zeta^T) \in \mathcal{P}([0:m])$. Under this decoding strategy, the corresponding false-alarm error and j-th error probability are defined as

$$\beta_0(\gamma, Q_{X^T} \otimes P_{\zeta^T}) \coloneqq \mathbb{E}_{\mathbb{P}_0}[1 - \gamma(0|X^T, \zeta^T)], \quad \beta_j(\gamma, P_{X^T,\zeta^T|M=j}) \coloneqq \mathbb{E}_{\mathbb{P}_j}[1 - \gamma(j|X^T, \zeta^T)].$$

When γ is restricted to output one-hot vectors (i.e., $\gamma(j \mid x^T, \zeta^T) \in \{0,1\}$ for all j), this formulation reduces to the deterministic decoder defined in the problem formulation.

First, let us fix a decoder $\gamma \in \mathcal{P}([0:m] \mid \mathcal{X}^T \times \mathcal{Z}^T)$. From the worst-case false-alarm constraint, we have for all $x^T \in \mathcal{X}^T$,

$$\begin{aligned}\alpha \ge \sup_{Q_{X^T}} \beta_0(\gamma, Q_{X^T} \otimes P_{\zeta^T}) &\overset{\text{(a)}}{\ge} \sum_{\zeta^T} P_{\zeta^T}(\zeta^T)(1 - \gamma(0|x^T, \zeta^T)) \\ &= \sum_{i \in [m]} \sum_{\zeta^T} P_{\zeta^T}(\zeta^T)\gamma(i|x^T, \zeta^T),\end{aligned}$$

where (a) follows by choosing $Q_{X^T} = \delta_{x^T}$. Therefore, there exists a sequence $(\alpha_1(x^T), \ldots, \alpha_m(x^T))$ such that

$$\sum_{\zeta^T} P_{\zeta^T}(\zeta^T)\gamma(i|x^T, \zeta^T) \le \alpha_i(x^T), \quad \forall i, x^T, \quad \text{and} \sum_{i \in [m]} \alpha_i(x^T) = \alpha. \tag{2}$$

The j-th error probability can be rewritten as follows:

$$\beta_j(\gamma, P_{X^T,\zeta^T|M=j}) = 1 - \mathbb{E}_{\mathbb{P}_j}[\gamma(j|X^T, \zeta^T)]$$

$$= 1 - \sum_{x^T, \zeta^T} P_{\zeta^T}(\zeta^T) P_{X^T|\zeta^T, M}(x^T|\zeta^T, j)\gamma(j|x^T, \zeta^T),$$

and thus the average detection error probability is given by

$$\bar{\beta}(\gamma, (\mathbb{P}_j)_{j\in[m]}) = 1 - \frac{1}{m}\sum_{j\in[m]}\sum_{x^T, \zeta^T} P_{\zeta^T}(\zeta^T) P_{X^T|\zeta^T, M}(x^T|\zeta^T, j)\gamma(j|x^T, \zeta^T). \tag{3}$$

To derive a lower bound of $\bar{\beta}(\gamma, (\mathbb{P}_j)_{j\in[m]})$, we first bound the second term of (3) as follows: for all $j \in [m]$ and $x^T \in \mathcal{X}^T$,

$$\begin{aligned}
&\sum_{\zeta^T} P_{\zeta^T}(\zeta^T) \underbrace{P_{X^T|\zeta^T, M}(x^T|\zeta^T, j)}_{\leq 1} \gamma(j|x^T, \zeta^T) \\
&\leq \sum_{\zeta^T} P_{\zeta^T}(\zeta^T)\gamma(j|x^T, \zeta^T) \overset{\text{(a)}}{\leq} \alpha_j(x^T),
\end{aligned} \tag{4}$$

where (a) follows from (2), and then

$$\frac{1}{m}\sum_{j\in[m]}\sum_{\zeta^T} P_{\zeta^T}(\zeta^T)\gamma(j|x^T, \zeta^T) \leq \frac{1}{m}\sum_{j\in[m]} \alpha_j(x^T) = \frac{\alpha}{m}. \tag{5}$$

On the other hand, since $\mathbb{1}\{\gamma(x^T, \zeta^T) = j\} \leq 1$, we also have: for all $j \in [m]$ and $x^T \in \mathcal{X}^T$,

$$\begin{aligned}
&\sum_{\zeta^T} P_{\zeta^T}(\zeta^T) P_{X^T|\zeta^T, M}(x^T|\zeta^T, j)\gamma(j|x^T, \zeta^T) \\
&\leq \sum_{\zeta^T} P_{\zeta^T}(\zeta^T) P_{X^T|\zeta^T, M}(x^T|\zeta^T, j) = P_{X^T|M}(x^T|j) \overset{\text{(a)}}{=} P_{X^T}(x^T),
\end{aligned} \tag{6}$$

where (a) follows from Assumption 1. Hence, for all $x^T \in \mathcal{X}^T$, the following inequality also holds:

$$\frac{1}{m}\sum_{j\in[m]}\sum_{\zeta^T} P_{\zeta^T}(\zeta^T) P_{X^T|\zeta^T, M}(x^T|\zeta^T, j)\gamma(j|x^T, \zeta^T) \leq P_{X^T}(x^T). \tag{7}$$

Combining (5) and (7), we lower bound $\bar{\beta}$ as follows:

$$\bar{\beta}(\gamma, (\mathbb{P}_j)_{j\in[m]}) \geq 1 - \sum_{x^T\in\mathcal{X}^T}\left(P_{X^T}(x^T) \wedge \frac{\alpha}{m}\right) = \sum_{x^T\in\mathcal{X}^T}\left(P_{X^T}(x^T) - \frac{\alpha}{m}\right)_+.$$

As this lower bound holds for all P_{X^T} satisfying the distortion constraint $\mathsf{D}(P_{X^T}, Q_{X^T}) \leq d$, it is minimized when P_{X^T} is chosen as follows

$$P^*_{X^T} = \underset{P_{X^T}:\mathsf{D}(P_{X^T}, Q_{X^T})\leq d}{\arg\min} \sum_{x^T\in\mathcal{X}^T}\left(P_{X^T}(x^T) - \frac{\alpha}{m}\right)_+,$$

and eventually

$$\bar{\beta}(\gamma, (\mathbb{P}_j)_{j\in[m]}) \geq \min_{P_{X^T}:\mathsf{D}(P_{X^T}, Q_{X^T})\leq d} \sum_{x^T\in\mathcal{X}^T}\left(P_{X^T}(x^T) - \frac{\alpha}{m}\right)_+. \tag{8}$$

We observe that the lower bound (8) is independent of γ. Thus, the lower bound also holds for the optimal value of the optimization problem (P1).

For the optimization problem (P2), we apply a similar method to derive a lower bound for β_j. By combining (4) and (6), for all $j \in [m]$, we have

$$\begin{aligned}
\beta_j(\gamma, P_{X^T, \zeta^T|M=j}) &\geq 1 - \sum_{x^T}\left(P_{X^T}(x^T) \wedge \alpha_j(x^T)\right) \\
&= \sum_{x^T}\left(P_{X^T}(x^T) - \alpha_j(x^T)\right)_+ \\
&\geq \min_{P_{X^T}:\mathsf{D}(P_{X^T}, Q_{X^T})\leq d} \sum_{x^T}\left(P_{X^T}(x^T) - \alpha_j(x^T)\right)_+.
\end{aligned}$$

Among all possible sequences $(\alpha_1(x^T), \ldots, \alpha_m(x^T))$ that sum up to α, the one that minimizes the lower bound for $\max_{j\in[m]} \beta_j(\gamma, P_{X^T,\zeta^T|M=j})$ is $(\frac{\alpha}{m}, \ldots, \frac{\alpha}{m})$. The proof is as follows:

$$\begin{aligned}\max_{j\in[m]} \beta_j(\gamma, P_{X^T,\zeta^T|M=j}) &\geq \max_{j\in[m]} \min_{P_{X^T}:\mathsf{D}(P_{X^T},Q_{X^T})\leq d} \sum_{x^T} \left(P_{X^T}(x^T) - \alpha_j(x^T)\right)_+ \\ &\overset{(a)}{\geq} \min_{P_{X^T}:\mathsf{D}(P_{X^T},Q_{X^T})\leq d} \sum_{x^T} \left(P_{X^T}(x^T) - \frac{\alpha}{m}\right)_+, \end{aligned} \tag{9}$$

where (a) holds with equality when $\alpha_j(x^T) = \frac{\alpha}{m}$ for all $j \in [m]$. Similarly, the lower bound (9) is independent of γ. Thus, the lower bound also holds for the optimal value of the optimization problem (P2).

In particular, if we choose the distortion function to be the total variation distance D_{TV}, the lower bound can be further simplified as follows.

For any P_{X^T} satisfying the distortion constraint, define

$$F(P_{X^T}) \coloneqq \sum_{x^T} \left(P_{X^T}(x^T) - \frac{\alpha}{m}\right)_+.$$

Define

$$E(Q_{X^T}; \frac{\alpha}{m}) \coloneqq \sum_{x^T} (Q_{X^T}(x^T) - \frac{\alpha}{m})_+, \quad S(Q_{X^T}; \frac{\alpha}{m}) \coloneqq \sum_{x^T} (\frac{\alpha}{m} - Q_{X^T}(x^T))_+.$$

Let $\mathcal{A} \coloneqq \{x^T : Q_{X^T}(x^T) > \frac{\alpha}{m}\}$. We have

$$\begin{aligned} F(P_{X^T}) \geq \sum_{x^T:x^T\in\mathcal{A}} \left(P_{X^T}(x^T) - \frac{\alpha}{m}\right)_+ &\geq \sum_{x^T:x^T\in\mathcal{A}} \left(P_{X^T}(x^T) - \frac{\alpha}{m}\right) \\ &= \sum_{x^T:x^T\in\mathcal{A}} \left(P_{X^T}(x^T) - Q_{X^T}(x^T) + Q_{X^T}(x^T) - \frac{\alpha}{m}\right) \\ &= E(Q_{X^T}; \frac{\alpha}{m}) + \sum_{x^T:x^T\in\mathcal{A}} (P_{X^T}(x^T) - Q_{X^T}(x^T)), \end{aligned}$$

which leads to

$$\begin{aligned} E(Q_{X^T}; \frac{\alpha}{m}) - F(P_{X^T}) &\leq \sum_{x^T:x^T\in\mathcal{A}} (Q_{X^T}(x^T) - P_{X^T}(x^T)) \\ &\leq \sum_{x^T:x^T\in\mathcal{A}} (Q_{X^T}(x^T) - P_{X^T}(x^T))_+ \\ &\leq \sum_{x^T} (Q_{X^T}(x^T) - P_{X^T}(x^T))_+ \leq d. \end{aligned}$$

Thus, we have one lower bound for $F(P_{X^T})$, i.e.,

$$F(P_{X^T}) \geq E(Q_{X^T}; \frac{\alpha}{m}) - d. \tag{10}$$

On the other hand, we can derive another lower bound for $F(P_{X^T})$ using the fact that $F(P_{X^T}) = 1 - \sum_{x^T} (P_{X^T}(x^T) \wedge \frac{\alpha}{m})$. For each $x^T \in \mathcal{X}^T$,

- If $Q_{X^T}(x^T) \geq \frac{\alpha}{m}$, then $Q_{X^T}(x^T) \wedge \frac{\alpha}{m} = \frac{\alpha}{m}$ and $P_{X^T}(x^T) \wedge \frac{\alpha}{m} \leq \frac{\alpha}{m} = Q_{X^T}(x^T) \wedge \frac{\alpha}{m}$;
- If $Q_{X^T}(x^T) < \frac{\alpha}{m}$, then $Q_{X^T}(x^T) \wedge \frac{\alpha}{m} = Q_{X^T}(x^T)$ and $P_{X^T}(x^T) \wedge \frac{\alpha}{m} \leq \frac{\alpha}{m} = Q_{X^T}(x^T) + (\frac{\alpha}{m} - Q_{X^T}(x^T)) = Q_{X^T}(x^T) \wedge \frac{\alpha}{m} + (\frac{\alpha}{m} - Q_{X^T}(x^T))_+$.

As a result, we have

$$\begin{aligned} P_{X^T}(x^T) \wedge \frac{\alpha}{m} &\leq Q_{X^T}(x^T) \wedge \frac{\alpha}{m} + (\frac{\alpha}{m} - Q_{X^T}(x^T))_+, \\ \text{and } \ 1 - F(P_{X^T}) = \sum_{x^T} (P_{X^T}(x^T) \wedge \frac{\alpha}{m}) &\leq \sum_{x^T} (Q_{X^T}(x^T) \wedge \frac{\alpha}{m}) + S(Q_{X^T}; \frac{\alpha}{m}). \end{aligned}$$

Therefore, $F(P_{X^T})$ has another lower bound as follows:

$$F(P_{X^T}) \geq 1 - \sum_{x^T}(Q_{X^T}(x^T) \wedge \frac{\alpha}{m}) - S(Q_{X^T}; \frac{\alpha}{m})$$
$$= E(Q_{X^T}; \frac{\alpha}{m}) - S(Q_{X^T}; \frac{\alpha}{m}).$$

From (10) and (10) and the fact that $F(P_{X^T}) \geq 0$, we finally have

$$F(P_{X^T}) \geq \left(E(Q_{X^T}; \frac{\alpha}{m}) - \left(d \wedge S(Q_{X^T}; \frac{\alpha}{m})\right)\right)_+. \tag{11}$$

There exists a distribution $P^*_{X^T}$ that can achieve this lower bound in (11). Intuitively, constructing $P^*_{X^T}$ is equivalent to the process of moving the mass of Q_{X^T} above α to below α, under some constraints.

Define

$$\delta \coloneqq \min\left\{d,\, S(Q_{X^T}; \frac{\alpha}{m}),\, E(Q_{X^T}; \frac{\alpha}{m})\right\},$$
$$\mathcal{A} \coloneqq \{x^T : Q_{X^T}(x^T) > \frac{\alpha}{m}\}, \quad \mathcal{B} \coloneqq \{x^T : Q_{X^T}(x^T) \leq \frac{\alpha}{m}\}.$$

Choose nonnegative transfers $\tau(x^T, y^T) \geq 0$ satisfying

$$\sum_{y^T \in \mathcal{B}} \tau(x^T, y^T) \leq Q_{X^T}(x^T) - \frac{\alpha}{m}, \quad \sum_{x^T \in \mathcal{A}} \tau(x^T, y^T) \leq \frac{\alpha}{m} - Q_{X^T}(x^T), \quad \sum_{x^T \in \mathcal{A}} \sum_{y^T \in \mathcal{B}} \tau(x^T, y^T) = \delta.$$

Such τ can always be constructed using greedy algorithms. Then the optimal $P^*_{X^T}$ that achieves the lower bound in (11) is given by

$$P^*_{X^T}(x^T) = \begin{cases} Q_{X^T}(x^T) - \sum_{y^T \in \mathcal{B}} \tau(x^T, y^T), & x^T \in \mathcal{A}, \\ Q_{X^T}(x^T) + \sum_{y^T \in \mathcal{A}} \tau(y^T, x^T), & x^T \in \mathcal{B}. \end{cases}$$

Note that $P^*_{X^T}$ is a valid probability mass function and satisfies $\mathsf{D}_{\mathsf{TV}}(P^*_{X^T}, Q_{X^T}) = \delta \leq d$. By such construction, $P^*_{X^T}(x^T) \leq \frac{\alpha}{m}$ for all $x^T \in \mathcal{B}$ and $P^*_{X^T}(x^T) > \frac{\alpha}{m}$ for all $x^T \in \mathcal{A}$. Then the induced $F(P^*_{X^T})$ is

$$F(P^*_{X^T}) = \sum_{x^T}(P^*_{X^T}(x^T) - \frac{\alpha}{m})_+$$
$$= \sum_{x^T : x^T \in \mathcal{A}} (P^*_{X^T}(x^T) - \frac{\alpha}{m}) = \sum_{x^T : x^T \in \mathcal{A}} (Q_{X^T}(x^T) - \sum_{y^T \in \mathcal{B}} \tau(x^T, y^T) - \frac{\alpha}{m})$$
$$= E(Q_{X^T}; \frac{\alpha}{m}) - \delta.$$

If $\delta = E(Q_{X^T}; \frac{\alpha}{m})$, then $F(P^*_{X^T}) = 0$; otherwise, $F(P^*_{X^T}) = E(Q_{X^T}; \frac{\alpha}{m}) - \left(d \wedge S(Q_{X^T}; \frac{\alpha}{m})\right)$.

Thus, $P^*_{X^T}$ achieves the lower bound in (11), proving optimality. In conclusion, under the distortion constraint defined with total variation distance, the lower bound for the detection error probability is given by the lower bound in (11).

D Proof of Achievability for Theorem 1

D.1 Decoding with Randomized Grouping (Proof of Theorem 2)

Note that the minimum detection error probability $\bar{\beta}^*(Q_{X^T}, \alpha, d, m)$ (cf. (1)) is non-decreasing in m. If the message set size m does not exceed

$$m^* \coloneqq \max\{m \in [|\mathcal{X}|^T] : \bar{\beta}^*(Q_{X^T}, \alpha, d, m) \leq 1 - \alpha\},$$

then the total probability mass of $P^*_{X^T}$ below $\frac{\alpha}{m}$ exceeds α. This ensures that the total probability mass of $P^*_{X^T}$ can be redistributed via some stochastic grouping kernel to form at least m groups each with probability at least $\frac{\alpha}{m}$, while the overflow mass above $\frac{\alpha}{m}$ is exactly equal to $\bar{\beta}^*(Q_{X^T}, \alpha, d, m)$. Therefore, m^* depends on $P^*_{X^T}$ and α. As a simple illustration, if $P^*_{X^T} = \mathrm{Unif}(\mathcal{X}^T)$ (i.e., the

maximum-entropy distribution on $\mathcal{X}^T$), then $m^* = 2^{\mathsf{H}(P^*_{X^T})} = |\mathcal{X}|^T$, which reaches the maximum information rate in the asymptotic case.

We define a stochastic grouping kernel $\phi : \mathcal{X}^T \to [K]$ (with $K \geq m$) satisfying $\sum_{k\in[K]} \phi(k|x^T) = 1$ for all x^T, which induces group probabilities

$$P_G(k) \;:=\; \Pr(G = k) = \sum_{x^T} P^*_{X^T}(x^T)\phi(k|x^T).$$

We aim to construct a kernel ϕ such that there are at least m groups, denoted by $k_1, k_2, \ldots, k_m$, with probability $P_G(k_l) \geq \frac{\alpha}{m}$ for all $l \in [m]$, while the overflow mass above $\frac{\alpha}{m}$ matches the lower bound in (1):

$$\sum_{k\in[K]} (P_G(k) - \frac{\alpha}{m})_+ = \sum_{x^T} (P^*_{X^T}(x^T) - \frac{\alpha}{m})_+ \triangleq \bar{\beta}^*(Q_{X^T}, \alpha, d, m).$$

For any $m \leq m^*$, there exists a stochastic grouping kernel ϕ achieving the desired properties. The randomized decoder is then constructed based on the kernel ϕ. In particular, we can choose

$$K = \left\lceil \frac{1 - \bar{\beta}^*(Q_{X^T}, \alpha, d, m)}{\alpha/m} \right\rceil, \tag{12}$$

which represents the maximum number of groups each with probability at least $\frac{\alpha}{m}$ (with a possible additional group accounting for any residual mass) that can be constructed by a kernel ϕ satisfying the above properties. It is straightforward to verify that $K \geq m$.

After constructing the groups, suppose $K \geq m+1$ there remains a residual group k_0 with probability mass $P_G(k_0) < \frac{\alpha}{m}$. We further split this residual mass into m groups of equal probability. Specifically, introduce m new groups indexed by $(k_0, 1), \ldots, (k_0, m)$, and define the refined stochastic kernel $\psi : \mathcal{X}^T \to ([K] \setminus \{k_0\}) \cup \{(k_0, l) : l \in [m]\}$ by

$$\begin{cases} \psi(k|x^T) = \phi(k|x^T) & \forall k \in [K] \setminus \{k_0\}; \\ \psi((k_0, l)|x^T) = \frac{1}{m}\phi(k_0|x^T) & \forall l \in [m]. \end{cases}$$

Then ψ is a valid stochastic kernel, since

$$\sum_{k\in[K]\setminus\{k_0\}} \psi(k|x^T) + \sum_{l=1}^{m} \psi((k_0, l)|x^T) = \sum_{k\in[K]} \phi(k|x^T) = 1.$$

The induced probabilities of the new residual groups are

$$\tilde{P}_G((k_0, l)) = \sum_{x^T} P^*_{X^T}(x^T)\psi((k_0, l)|x^T) = \frac{P_G(k_0)}{m}, \qquad l \in [m].$$

Thus, if a residual mass exists, the refined kernel outputs $K - 1 + m$ groups: the original $K - 1$ non-residual groups together with m equally weighted residual subgroups.

Example construction of P_G, ϕ and ψ With the chosen K in (12), define a target group PMF $(p_k)_{k=1}^K$ by

$$p_1 = \frac{\alpha}{m} + \bar{\beta}^*, \quad p_2 = \cdots = p_{K-1} = \frac{\alpha}{m}, \quad p_K = 1 - \frac{(K-1)\alpha}{m} - \bar{\beta}^* \leq \frac{\alpha}{m}.$$

Once we have the target PMF $(p_k)_{k=1}^K$, there are many ϕ's that induce it. The simplest is the "independent" kernel:

$$\phi(k \mid x^T) \;\triangleq\; p_k, \qquad \forall x^T,$$

which means moving p_k fraction of each $P^*_{X^T}$ to the group k. Then

$$P_G(k) = \sum_{x^T} P^*_{X^T}(x^T)\phi(k \mid x^T) = p_k \sum_{x^T} P^*_{X^T}(x^T) = p_k, \quad \forall k \in [K].$$

After constructing ϕ and the induced group distribution P_G, we further refine the residual group, if it exists, by constructing the stochastic kernel ψ by definition. The the residual group is group K and ψ further equally splits it to m new subgroups indexed by $K, K+1, \ldots, K-1+m$. This yields the refined group distribution $\tilde{P}_G$, in which the residual mass is split equally into m new residual groups.

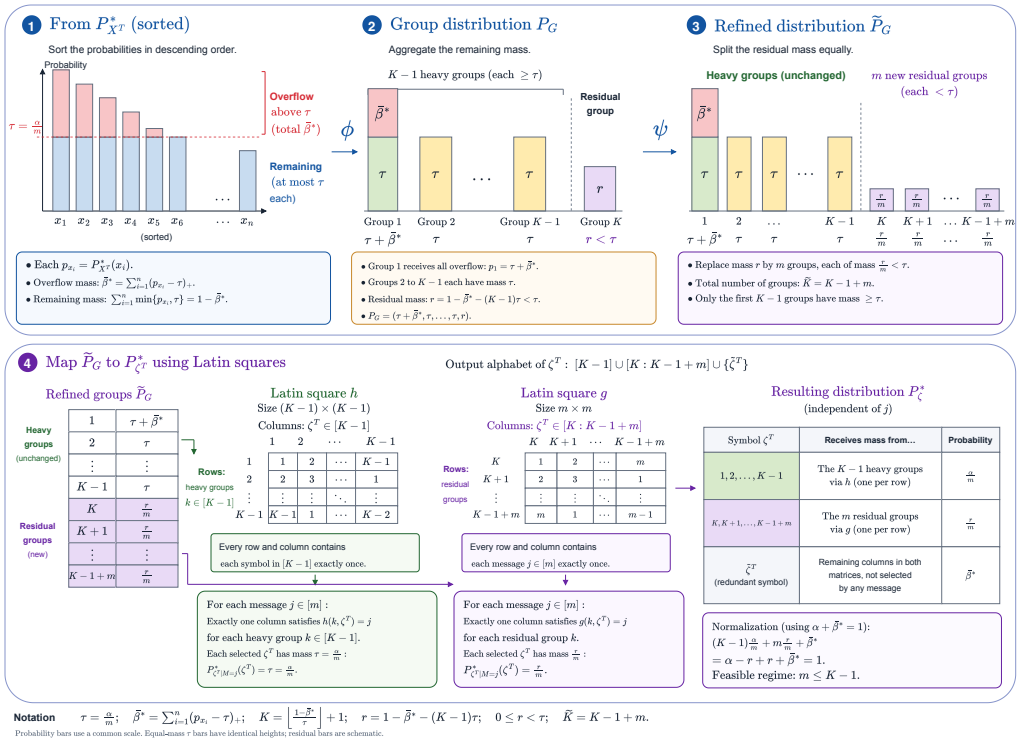


Figure 2: Grouping process and decoder design for the case when there is a residual group.

Existence of optimal decoder Without loss of generality, we consider the case where there is a residual group k_0 with $P_G(k_0) < \frac{\alpha}{m}$. Let $\tilde{K} = \mathrm{supp}(\tilde{P}_G)$ be the size of the support of the refined group distribution. Then $\tilde{K} = K - 1 + m \geq 2m$ in this case.

For simplicity, we let the auxiliary random variable ζ^T take values in a finite alphabet $\mathcal{Z}^T = [\tilde{K}] \cup \{\widetilde{\zeta^T}\}$, whose cardinality may depend on T but does not represent a length-T Cartesian product.

The "heavy" group set is the set of the first $K-1$ indices:

$$\{k \in [K] : P_G(k) \geq \frac{\alpha}{m}\} = [K-1].$$

Let $h : [K-1] \times [K-1] \to [K-1]$ be a function such that for every $k \in [K-1]$, $h(k, \cdot)$ is a bijection, and for every $\zeta^T \in [K-1]$, $h(\cdot, \zeta^T)$ is a bijection. A Latin square matrix with size $(K-1) \times (K-1)$ satisfies such structure. Recall that $K - 1 \geq m$ under the setting.

Furthermore, let $g : [m] \times [m] \to [m]$ be a function such that for every $l \in [m]$, $g(l, \cdot)$ is a bijection, and for every $j \in [m]$, $g(\cdot, j)$ is a bijection. A Latin square matrix with size $m \times m$ satisfies such structure.

Define a group-level decoder $\gamma_G : [\tilde{K}] \times \mathcal{Z}^T \to [0 : m]$ as follows. For $j \in [m]$,

$$\gamma_G(j|k, \zeta^T) = \mathbb{1}\{k, \zeta^T \in [K-1], h(k, \zeta^T) = j\} + \mathbb{1}\{k, \zeta^T \in [K : K-1+m], g(k-K+1, \zeta^T - K + 1) = j\}.$$

For the rejection output 0, define

$$\gamma_G(0 \mid k, \zeta^T) = 1 - \sum_{j=1}^{m} \gamma_G(j \mid k, \zeta^T) \in \{0, 1\},$$

and for the redundant symbol $\widetilde{\zeta^T}$,

$$\gamma_G(0|k, \widetilde{\zeta^T}) = 1, \quad \forall k \in [\tilde{K}].$$

Finally, based on the refined stochastic group kernel ψ and a shared secret key, the decoder obtains the grouping $G_{\text{key}}(x^T)$ from $\psi(\cdot|x^T)$ and outputs

$$\gamma^*(x^T, \zeta^T) = j, \quad \text{if } \gamma_G(j|G_{\text{key}}(x^T), \zeta^T) = 1, \quad j \in [0:m].$$

Existence of optimal watermark embedding: The corresponding conditional distribution $P^*_{\zeta^T|X^T,M}$ is constructed as follows: for all $j \in [m]$,

$$P^*_{\zeta^T|X^T,M}(\zeta^T|x^T, j) = \sum_{k\in[\tilde{K}]} P^*_{\zeta^T,G|X^T,M}(\zeta^T, k|x^T, j) = \sum_{k\in[\tilde{K}]} P^*_{\zeta^T|G,M}(\zeta^T|k, j)\psi(k|x^T),$$

where for all $k \in [\tilde{K}]$,

$$P^*_{\zeta^T|G,M}(\zeta^T|k,j) = \begin{cases} 1 \wedge \frac{\alpha}{m\tilde{P}_G(k)}, & \text{if } k, \zeta^T \in [K-1], h(k,\zeta^T) = j, \\ & \quad \text{or} \quad k, \zeta^T \in [K : K-1+m], g(k-K+1, \zeta^T - K + 1) = j \\ (1 - \frac{\alpha}{m\tilde{P}_G(k)})_+, & \text{if } \zeta^T = \widetilde{\zeta^T}, \\ 0, & \text{otherwise.} \end{cases}$$

We can verify that $P^*_{\zeta^T|M=j}$ is the same for all $j \in [m]$ as follows:

$$\begin{aligned} P^*_{\zeta^T|M}(\zeta^T|j) &= \sum_{k\in[\tilde{K}]} P^*_{\zeta^T|G,M}(\zeta^T|k,j)\tilde{P}_G(k) \\ &= \begin{cases} \frac{\alpha}{m}, & \text{if } \zeta^T \in [K-1], \\ \tilde{P}_G(\zeta^T), & \text{if } \zeta^T \in [K : K-1+m], \\ \sum_{k\in[\tilde{K}]}(\tilde{P}_G(k) - \frac{\alpha}{m})_+, & \text{if } \zeta^T = \widetilde{\zeta^T}, \end{cases} \\ &\triangleq P^*_{\zeta^T}(\zeta^T). \end{aligned}$$

Achievable Errors: For the optimal watermark embedding $P^*_{\zeta^T|X^T,M}$ and optimal decoder $\gamma^*(x^T, \zeta)$, let $\gamma^*(j|x^T, \zeta) = \mathbb{1}\{\gamma^*(x^T, \zeta) = 1\}$ and then the j-th error probability is

$$\begin{aligned} \beta_j(\gamma, \mathbb{P}_j) &= 1 - \underbrace{\sum_{x^T,\zeta^T} P^*_{X^T,\zeta|M=j}(x^T, \zeta^T)\,\gamma^*(j|x^T, \zeta^T)}_{=\Pr(\gamma^*=j|M=j)} \\ &= 1 - \mathbb{E}_{\text{key}}\Big[\sum_{x^T,\zeta^T} P^*_{X^T,\zeta|M=j}(x^T, \zeta^T)\,\gamma^*(j|x^T, \zeta^T) \mid \text{key}\Big] \\ &= 1 - \mathbb{E}_{\text{key}}\Big[\sum_{x^T,\zeta^T} P^*_{X^T}(x^T)P^*_{\zeta|X^T,M=j}(\zeta^T|x^T)\gamma_G(j|G_{\text{key}}(x^T), \zeta^T)\Big] \\ &= 1 - \mathbb{E}_{\text{key}}\Big[\sum_{x^T,\zeta^T} P^*_{X^T}(x^T)\sum_{k\in[\tilde{K}]} P^*_{\zeta^T|G,M}(\zeta^T|k,j)\psi(k|x^T)\gamma_G(j|G_{\text{key}}(x^T), \zeta^T)\Big] \\ &= 1 - \mathbb{E}_{\text{key}}\Big[\sum_{k\in[\tilde{K}]}\sum_{\zeta^T} P^*_{\zeta^T|G,M}(\zeta^T|k,j)\sum_{x^T:G_{\text{key}}(x^T)=k} P^*_{X^T}(x^T)\psi(k|x^T)\gamma_G(j|k,\zeta^T)\Big] \\ &= 1 - \mathbb{E}_{\text{key}}\Big[\sum_{k\in[\tilde{K}]}\sum_{\zeta^T} P^*_{\zeta^T|G,M}(\zeta^T|k,j)\tilde{P}^*_G(k)\gamma_G(j|k,\zeta^T)\Big] \\ &= 1 - \sum_{k\in[\tilde{K}]}\sum_{\zeta^T} P^*_{\zeta^T|G,M}(\zeta^T|k,j)\tilde{P}^*_G(k)\gamma_G(j|k,\zeta^T) \\ &\overset{\text{(a)}}{=} 1 - \sum_{k\in[\tilde{K}]}(P_G(k) \wedge \frac{\alpha}{m}) \\ &= 1 - \sum_{x^T}(P_{X^T}(x^T) \wedge \frac{\alpha}{m}) \end{aligned}$$

$$= \underbrace{\sum_{x^T \in \mathcal{X}^T} (P^*_{X^T}(x^T) - \frac{\alpha}{m})_+}_{\text{lower bound in Theorem 1}},$$

where (a) follows since

$$\begin{aligned}
&\sum_{k\in[\tilde{K}]} \tilde{P}^*_G(k) \sum_{\zeta^T \in \mathcal{Z}^T} P^*_{\zeta^T|G,M}(\zeta^T \mid k, j)\, \gamma_G(j \mid k, \zeta^T) \\
&= \sum_{k\in[\tilde{K}]} \tilde{P}^*_G(k) \sum_{\zeta^T \in \mathcal{Z}^T} P^*_{\zeta^T|G,M}(\zeta^T \mid k, j) \Big[\mathbb{1}\{k, \zeta^T \in [K-1],\, h(k,\zeta^T) = j\} \\
&\qquad + \mathbb{1}\{k, \zeta^T \in [K : K-1+m],\, g(k-K+1, \zeta^T - K + 1) = j\} \Big] \\
&= \sum_{k\in[\tilde{K}]} \tilde{P}^*_G(k) \Bigg[\mathbb{1}\{k \in [K-1]\} \sum_{\substack{\zeta^T \in [K-1]: \\ h(k,\zeta^T)=j}} \left(1 \wedge \frac{\alpha}{m\tilde{P}_G(k)}\right) \\
&\qquad + \mathbb{1}\{k \in [K : K-1+m]\} \sum_{\substack{\zeta^T \in [K:K-1+m]: \\ g(k-K+1,\zeta^T-K+1)=j}} \left(1 \wedge \frac{\alpha}{m\tilde{P}^*_G(k)}\right) \Bigg] \\
&= \sum_{k\in[\tilde{K}]} \tilde{P}_G(k) \left(1 \wedge \frac{\alpha}{m\tilde{P}^*_G(k)}\right) \\
&= \sum_{k\in[\tilde{K}]} \left(\tilde{P}^*_G(k) \wedge \frac{\alpha}{m}\right).
\end{aligned}$$

The induced false-alarm error is derived as follows. For all $x^T \in \mathcal{X}^T$, we have

$$\begin{aligned}
\sum_{j\in[m]} \sum_{\zeta^T} P^*_{\zeta^T}(\zeta^T)\gamma^*(j|x^T,\zeta^T) &= \sum_{j\in[m]} \sum_{\zeta^T} P^*_{\zeta^T}(\zeta^T) \sum_{k\in[\tilde{K}]} \psi(k|x^T)\gamma_G(j|k,\zeta^T) \\
&= \sum_{k\in[\tilde{K}]} \psi(k|x^T) \sum_{j\in[m]} \sum_{\zeta^T} P^*_{\zeta^T}(\zeta^T)\gamma_G(j|k,\zeta^T).
\end{aligned}$$

For a heavy group $k \in [K-1]$, by the Latin-square property of h, exactly m values of $\zeta^T \in [K-1]$ decode to messages in $[m]$. Since each such ζ^T has probability α/m,

$$\sum_{j\in[m]} \sum_{\zeta^T} P^*_{\zeta^T}(\zeta^T)\gamma_G(j|k,\zeta^T) = m \cdot \frac{\alpha}{m} = \alpha.$$

For a refined residual group $k \in [K : K-1+m]$, all m residual auxiliary symbols decode to some message in $[m]$, so

$$\sum_{j\in[m]} \sum_{\zeta^T} P^*_{\zeta^T}(\zeta^T)\gamma_G(j|k,\zeta^T) = \sum_{\zeta^T=K}^{K-1+m} \tilde{P}_G(\zeta^T) = P_G(k_0) < \alpha/m,$$

where $P_G(k_0)$ is the original residual mass.

Therefore, for all $x^T \in \mathcal{X}^T$,

$$\sum_{j\in[m]} \sum_{\zeta^T} P^*_{\zeta^T}(\zeta^T)\gamma^*(j|x^T,\zeta^T) = \alpha \sum_{k=1}^{K-1} \psi(k|x^T) + P_G(k_0) \sum_{k=K}^{K-1+m} \psi(k|x^T) \le \alpha.$$

Thus, as any Q_{X^T} can be represented by a linear combination of $\{\delta_{x^T}\}_{x^T \in \mathcal{X}^T}$, the worst-case false-alarm error is given by

$$\beta_0(\gamma, Q_{X^T} \otimes P_{\zeta^T}) = \sup_{Q_{X^T}} \sum_{j\in[m]} \sum_{x^T,\zeta^T} Q_{X^T}(x^T) P^*_{\zeta^T}(\zeta^T)\gamma^*(j|x^T,\zeta^T) \le \alpha.$$

The optimality of the construction is thus proved.

D.2 Decoding with Deterministic Grouping (Proof of Theorem 3)

Given any source distribution Q_{X^T}, any message set size $m \le m^*$, and any false-alarm constraint $\alpha \in (0,1)$, the optimal joint distribution of the watermarked sequence X^T is given by

$$P^*_{X^T} \coloneqq \underset{P_{X^T}:\mathsf{D}(P_{X^T},Q_{X^T})\le d}{\arg\min} \sum_{x^T} \left(P_{X^T}(x^T) - \frac{\alpha}{m}\right)_+.$$

Then the next step is to find the conditional distribution $P^*_{\zeta^T|X^T,M}$ and the detector γ^* such that detection error lower bound is achieved and the worst-case false-alarm constraint is satisfied.

We define a deterministic grouping function $\phi : \mathcal{X}^T \to [K]$ (with $K \ge m$) that assigns each symbol $x^T \in \mathcal{X}^T$ to a group index $k \in [K]$, thereby inducing group probabilities

$$P_G(k) \;\coloneqq\; \Pr(G = k) = \sum_{x^T:\,\phi(x^T)=k} P^*_{X^T}(x^T).$$

We impose the additional constraint that any symbol x^T with $P^*_{X^T}(x^T) \ge \frac{\alpha}{m}$ forms a singleton group, i.e.,

$$P^*_{X^T}(x^T) \ge \frac{\alpha}{m} \;\;\Rightarrow\;\; \phi(x^T) \ne \phi(\tilde{x}^T) \text{ for all } \tilde{x}^T \ne x^T.$$

Let $\mathcal{S}_{\alpha/m} \coloneqq \{x^T : P^*_{X^T}(x^T) \ge \frac{\alpha}{m}\}$ and $\mathcal{G}_{\alpha/m} \coloneqq \{k \in [K] : P_G(k) \ge \frac{\alpha}{m}\}$. We define $\phi^\star$ as a solution to the optimization problem

$$\min_{\phi:\mathcal{X}^T\to[K]} \quad \sum_{k=1}^{K} \big(P_G(k) - \frac{\alpha}{m}\big)_+ \tag{P3}$$
$$\text{s.t.} \quad \big|\mathcal{G}_{\alpha/m}\big| \ge m, \quad \text{every } x^T \in \mathcal{S}_{\alpha/m} \text{ is assigned to a singleton group.}$$

The feasible set is non-empty if $m \in \mathcal{M} \coloneqq \{m \in [|\mathcal{X}|^T] : \sum_{x^T \notin \mathcal{S}_{\alpha/m}} P^*_{X^T}(x^T) \ge (1 - \frac{|\mathcal{S}_{\alpha/m}|}{m})\alpha\}$. This means it is possible to redistribute the mass below $\frac{\alpha}{m}$ to at least $m - |\mathcal{S}_{\alpha/m}|$ groups with probability exceeding $\frac{\alpha}{m}$. Moreover, $\phi^\star$ exists because the set of deterministic groupings of a finite alphabet is finite. Intuitively, this optimization seeks to partition the probability mass of $P^*_{X^T}$ into at least m groups of mass at least $\frac{\alpha}{m}$ while minimizing the total overflow above $\frac{\alpha}{m}$ across all groups. A feasible grouping can be constructed via a greedy procedure, while $\phi^\star$ denotes an overflow-minimizing solution.

Optimal Decoder: Let the auxiliary alphabet be the union of three disjoint sets: $\mathcal{Z}^T = \mathcal{Z}^T_{\ge} \cup \mathcal{Z}^T_{<} \cup \{\widetilde{\zeta^T}\}$, where $|\mathcal{Z}^T_{\ge}| = |\mathcal{G}_{\alpha/m}|$ and $|\mathcal{Z}^T_{<}| = K - |\mathcal{G}_{\alpha/m}|$. Based on the optimized function ϕ^*, we choose the decoder as

$$\gamma^*(x^T, \zeta^T) = \begin{cases} j, & \text{if } \zeta^T \in \mathcal{Z}^T_{\ge} \text{ and } h(\phi^*(x^T), \zeta^T) = j \le m \\ 0, & \text{otherwise.} \end{cases}$$

where $h : \mathcal{G}_{\alpha/m} \times \mathcal{Z}^T_{\ge} \to [|\mathcal{G}_{\alpha/m}|]$ is a function such that for every $k \in \mathcal{G}_{\alpha/m}$, $h(k,\cdot)$ is a bijection, and for every $\zeta^T \in \mathcal{Z}^T_{\ge}$, $h(\cdot, \zeta^T)$ is a bijection. A Latin square matrix satisfies such structure.

Optimal Watermark Embedding: The corresponding conditional distribution $P^*_{\zeta^T|X^T,M}$ is constructed as follows: for all $j \in [m]$,

$$P^*_{\zeta^T|X^T,M}(\zeta^T|x^T, j) = \sum_{k\in[K]} P^*_{\zeta^T,G|X^T,M}(\zeta^T, k|x^T, j) = \sum_{k\in[K]} P^*_{\zeta^T|G,M}(\zeta^T|k, j) P_{G|X^T}(k|x^T)$$

$$= P^*_{\zeta^T|G,M}(\zeta^T|\phi^*(x^T), j),$$

where for all $k \in [K]$

$$P^*_{\zeta^T|G,M}(\zeta^T|k, j) = \begin{cases} 1 \wedge \frac{\alpha}{mP_G(k)}, & \text{if } h(k, \zeta^T) = j, \\ 1, & \text{if } \zeta^T \in \mathcal{Z}^T_{<} \text{ and } g(\zeta^T) = k \\ (1 - \frac{\alpha}{mP_G(k)})_+, & \text{if } \zeta^T = \widetilde{\zeta^T}, \\ 0, & \text{otherwise,} \end{cases}$$

and $g : \mathcal{Z}_<^T \to [K] \setminus \mathcal{G}_{\alpha/m}$ is a bijective function. We can verify that $P^*_{\zeta^T|M=j} = P^*_{\zeta^T}$ for all $j \in [m]$ as follows:

$$\begin{aligned} P^*_{\zeta^T|M}(\zeta^T|j) &= \sum_{k\in[K]} P^*_{\zeta^T|G,M}(\zeta^T|k,j)P_G(k) \\ &= \begin{cases} \frac{\alpha}{m}, & \text{if } \zeta^T \in \mathcal{Z}_>^T, \\ P_G(g(\zeta^T)), & \text{if } \zeta^T \in \mathcal{Z}_<^T, \\ \sum_{k\in[K]}(P_G(k) - \frac{\alpha}{m})_+, & \text{if } \zeta^T = \widetilde{\zeta^T}, \end{cases} \\ &\triangleq P^*_{\zeta^T}(\zeta^T), \end{aligned}$$

which is the same for all $j \in [m]$.

Achievable Errors: The induced j-th error probability is given by

$$\begin{aligned} \beta_j(\gamma, \mathbb{P}_j) &= 1 - \sum_{x^T,\zeta^T} P^*_{X^T}(x^T)P^*_{\zeta^T|X^T,M}(\zeta^T|x^T,j)\mathbb{1}\{\gamma^*(x^T,\zeta^T) = j\} \\ &= 1 - \sum_{x^T,\zeta^T} P^*_{X^T}(x^T)P^*_{\zeta^T|G,M}(\zeta^T|\phi^*(x^T),j)\mathbb{1}\{\gamma^*(x^T,\zeta^T) = j\} \\ &= 1 - \sum_{k\in[K]}\sum_{\zeta^T}\sum_{x^T:\phi^*(x^T)} P^*_{X^T}(x^T)P^*_{\zeta^T|G,M}(\zeta^T|k,j)\mathbb{1}\{\gamma^*(x^T,\zeta^T) = j\} \\ &= 1 - \sum_{k\in[K]}\sum_{\zeta^T} P_G(k)P^*_{\zeta^T|G,M}(\zeta^T|k,j)\mathbb{1}\{h(k,\zeta^T) = j\} \\ &= 1 - \sum_{k\in[K]} (P_G(k) \wedge \frac{\alpha}{m}) \\ &= \sum_{k\in[K]} (P_G(k) - \frac{\alpha}{m})_+ \\ &= \underbrace{\sum_{x^T\in\mathcal{S}_{\alpha/m}} (P^*_{X^T}(x^T) - \frac{\alpha}{m})_+}_{\text{lower bound in Theorem 1}} + \underbrace{\sum_{k\in\mathcal{G}_{\alpha/m}\setminus\phi^*(\mathcal{S}_{\alpha/m})} (P_G(k) - \frac{\alpha}{m})_+}_{:=\varepsilon_{\text{res}}\ \text{(residual error)}}, \end{aligned} \tag{13}$$

where $\phi^*(\mathcal{S}_{\alpha/m})$ denote the collection of the singleton groups. Since ϕ^* is optimized as an optimizer of (P3), the residual error is minimized and may vanish for certain generative distributions Q_{X^T}. For example, when the distortion level $d = 0$ so that $P^*_{X^T} = Q_{X^T}$, and when all probability masses satisfy $Q_{X^T}(x^T) \geq \frac{\alpha}{m}$, then the residual error $\varepsilon_{\text{res}} = 0$. In this case, j-th error probability $\beta_j(\gamma, \mathbb{P}_j)$ reaches the lower bound in Theorem 1.

The induced false-alarm error is derived as follows. For all $x^T \in \mathcal{X}^T$, we have

$$\sum_{\zeta^T} P^*_{\zeta^T}(\zeta^T)\mathbb{1}\{\gamma^*(x^T,\zeta^T) \neq 0\} = \sum_{j\in[m]}\sum_{\zeta^T} P^*_{\zeta^T}(\zeta^T)\mathbb{1}\{\gamma^*(x^T,\zeta^T) = j\} = m \cdot \frac{\alpha}{m} = \alpha.$$

Thus, as any Q_{X^T} can be represented by a linear combination of $\{\delta_{x^T}\}_{x^T\in\mathcal{X}^T}$, the worst-case false-alarm error is given by

$$\beta_0(\gamma, Q_{X^T} \otimes P_{\zeta^T}) = \sup_{Q_{X^T}} Q_{X^T} \otimes P_{\zeta^T}(\gamma^*(x^T,\zeta^T) \neq 0) = \alpha.$$

E Proof of Converse for Theorem 4

Let $P_e = \Pr(\hat{M} \neq M) = \bar{\beta}(\gamma, (\mathbb{P}_k)_{k\in[m]})$. From the Fano's inequality, we have

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

With Assumption 1, the entropy of M is upper bounded by

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

$$\leq H(X^T|\zeta^T) + 1 + P_e \log m,$$

which leads to

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

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

$$\frac{\log m}{T} \leq \frac{H(X^T|\zeta^T)}{T} \leq \frac{\mathsf{H}(X^T)}{T} \leq \max_{P_{X^T}: \frac{1}{T}\mathsf{D}_{\mathsf{KL}}(P_{X^T}\|Q_{X^T})\leq d} \frac{\mathsf{H}(X^T)}{T}$$

and

$$\limsup_{T\to\infty} \frac{\log m}{T} \leq \limsup_{T\to\infty} \max_{P_{X^T}: \frac{1}{T}\mathsf{D}_{\mathsf{KL}}(P_{X^T}\|Q_{X^T})\leq d} \frac{\mathsf{H}(X^T)}{T} = \max_{P:\overline{\mathsf{D}}_{\mathsf{KL}}(P\|Q)\leq d} \mathsf{H}_P(\mathcal{X}),$$

where the last inequality holds as the limit exists for a stationary ergodic process.

F Proof of Achievability for Theorem 4

Embedding and decoding construction In the following, we identify the asymptotically optimal watermarking scheme that achieves vanishing j-th error probabilities and the maximum information rate, while satisfying the worst-case false-alarm error and distortion constraint. Under any hypothesis H_j and any $\mathbb{P}_j$, we define the typical sets of stationary ergodic processes $x^T \in \mathcal{X}^T$ and $\zeta^T \in \mathcal{Z}^T$.

Definition 2 (Typical Sets) *For arbitrarily small $\eta > 0$, define the typical sets $\mathcal{A}_{\eta,X}^{(T)}$ and $\mathcal{A}_{\eta,\zeta}^{(T)}$ as*

$$\mathcal{A}_{\eta,X}^{(T)} := \left\{x^T: \left|\frac{1}{T}\log\frac{1}{P_{X^T}(x^T)} - \mathsf{H}_P(\mathcal{X})\right| \leq \eta\right\}, \mathcal{A}_{\eta,\zeta}^{(T)} := \left\{\zeta^T: \left|\frac{1}{T}\log\frac{1}{P_{\zeta^T}(\zeta^T)} - \mathsf{H}_P(\mathcal{Z})\right| \leq \eta\right\}.$$

The typical sequences in $\mathcal{A}_{\eta,X}^{(T)}$ and $\mathcal{A}_{\eta,\zeta}^{(T)}$ 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 $d = 0$, i.e., $P_{X^T} = Q_{X^T}$, as follows. Here, we use $\doteq$ to denote equality to the first order in the exponent.

Step 1 (Embedding). Let $\mathcal{Z} \subset \mathbb{Z}$ and design a stationary ergodic process P_ζ^* on $\mathcal{Z}$ such that $\mathsf{H}_P(\mathcal{Z}) = \mathsf{H}_Q(\mathcal{X})$. Let $\eta = T^{-\frac{1}{4}}$. The asymptotically optimal joint distribution $P_{X^T,\zeta^T|M}^*$ is constructed as follows: for any $i \in [m]$,

- for all $x^T \in \mathcal{A}_{\eta,X}^{(T)}$, $P_{X,\zeta|M}^{*T}(x^T,\zeta^T|i) = \begin{cases} Q_{X^T}(x^T) \doteq e^{-T\mathsf{H}_Q(\mathcal{X})}, & \text{if } \zeta^T \in \mathcal{A}_{\eta,\zeta}^{(T)} \text{ and } g(x^T,\zeta^T) = i; \\ 0, & \text{otherwise,} \end{cases}$

  for some function $g : \mathcal{A}_{\eta,X}^{(T)} \times \mathcal{A}_{\eta,\zeta}^{(T)} \to \left[|\mathcal{A}_{\eta,X}^{(T)}|\right]$ such that it is bijective in either argument when the other is fixed.
- for all $x^T \notin \mathcal{A}_{\eta,X}^{(T)}$, let $P_{X,\zeta|X,M}^{*T}(x^T,\zeta^T|i)$ take any non-negative value as long as $P_{X,\zeta|X,M}^{*T}$ is a valid probability distribution.

Step 2 (Decoding). The optimal decoder accepts the form:

$$\gamma^*(x^T,\zeta^T) = \begin{cases} g(x^T,\zeta^T), & \forall x^T \in \mathcal{A}_{\eta,X}^{(T)}, \zeta^T \in \mathcal{A}_{\eta,\zeta}^{(T)}, \text{ and } g(x^T,\zeta^T) \leq m; \\ 0, & \text{otherwise.} \end{cases}$$

Figure 3 illustrates the design using a toy example when $m = 3$.

Existence of asymptotically optimal decoders First, the function g proposed in Theorem 4 always exists. For example, a Latin square matrix satisfies the structure. If the message set size m satisfies $\frac{1}{T}(\log m - \log\alpha) \leq \mathsf{H}_Q(\mathcal{X})$, then we have

$$m \leq e^{T\mathsf{H}_Q(\mathcal{X})} \doteq |\mathcal{A}_{\eta,X}^{(T)}|,$$

and the output space of g contains $[m]$. Thus, any optimal decoder γ^* can decode messages drawn from $[m]$.

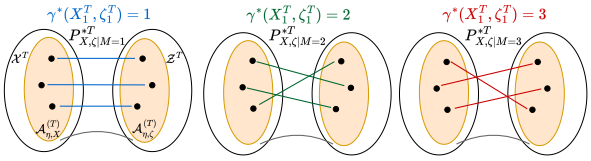


Figure 3: Illustration of the asymptotically optimal watermarking scheme when $m = 3$.

Asymptotic optimality For any optimal decoder γ^*, one can always construct the corresponding joint distribution $P^*_{X,\zeta|M}$ in Theorem 4. In the following, we first show that the probability of the atypical set decays exponentially with T. We then prove that the j-th error probability vanishes to 0 while the worst-case false-alarm error is upper bounded by α as $T \to \infty$.

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

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

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

$$\begin{aligned}\mathbb{P}_j((X^T,\zeta^T) \notin \mathcal{A}^{(T)}_{\eta,j}) \le\ &\mathbb{P}_j\left(\left| -\frac{1}{T}\log P_{X^T}(x^T) - \mathsf{H}_P(\mathcal{X})\right| \ge \eta\right) \\ &+ \mathbb{P}_j\left(\left| -\frac{1}{T}\log P_{\zeta^T}(\zeta^T) - \mathsf{H}_P(\mathcal{Z})\right| \ge \eta\right) \\ &+ \mathbb{P}_j\left(\left| -\frac{1}{T}\log P_{X^T,\zeta^T|M=j}(x^T,\zeta^T) - \mathsf{H}_P(\mathcal{X}\times\mathcal{Z})\right| \ge \eta\right). \quad (14)\end{aligned}$$

Then, by the Chernoff bound, we have

$$\begin{aligned}&\mathbb{P}_j\left(\left| -\frac{1}{T}\log P_{X^T}(x^T) - \mathsf{H}_P(\mathcal{X})\right| \ge \eta\right) \le 2\mathbb{P}_j\left(\left| -\frac{1}{T}\log P_{X^T}(x^T) - \mathsf{H}_P(\mathcal{X}) \ge \eta\right)\right. \\ &\le 2\exp\left(-T\sup_{s\ge 0}(s\eta - \log\mathbb{E}[\exp(-s\log P_{X^T}(X^T))])\right) \\ &\overset{(a)}{\approx} 2\exp\left(-T\sup_{s\ge 0}(s\eta - (-s\mathbb{E}[\log P_{X^T}(X^T)] + s^2\mathbb{E}[(\log P_{X^T}(X^T))^2])\right) \\ &\overset{(b)}{=} 2\exp(-\Omega(T\eta^2)) = \exp(-\Omega(T^{\frac{1}{2}})),\end{aligned}$$

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 (14) can be similarly proved.

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

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

Next, we prove that the proposed watermarking scheme in Theorem 4 achieves the asymptotic optimality.

Given any optimal γ^* and the watermarking scheme in Theorem 4, for any $j \in [m]$, the j-th error probability is given by

$$\beta_j(\gamma^*, P^*_{X^T,\zeta^T|M=j}) = \sum_{x^T,\zeta^T} P^*_{X^T,\zeta^T|M}(x^T,\zeta^T|j)\mathbb{1}\{\gamma^*(x^T,\zeta^T) \ne j\}$$

$$\leq \sum_{(x^T,\zeta^T)\in\mathcal{A}_{\eta,j}^{(T)}} P^*_{X^T,\zeta^T|M}(x^T,\zeta^T|j)\mathbb{1}\{\gamma^*(x^T,\zeta^T)\neq j\} + \exp(-\Omega(T^{\frac{1}{2}}))$$

$$= \exp(-\Omega(T^{\frac{1}{2}})) \to 0 \text{ as } T\to\infty.$$

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

$$\begin{aligned}
&\sum_{\zeta^T} P^*_\zeta(\zeta^T)\mathbb{1}\{\gamma^*(x^T,\zeta^T)\neq 0\}\\
&\leq \sum_{\zeta^T\in\mathcal{A}_{n,\zeta}^{(T)}} P^*_\zeta(\zeta^T)\mathbb{1}\{\gamma^*(x^T,\zeta^T)\neq 0\} + \exp(-\Omega(T^{\frac{1}{2}}))\\
&= \sum_{i\in[m]}\sum_{\zeta^T\in\mathcal{A}_{n,\zeta}^{(T)}} P^*_\zeta(\zeta^T)\mathbb{1}\{\gamma^*(x^T,\zeta^T)= i\} + \exp(-\Omega(T^{\frac{1}{2}}))\\
&= \sum_{i\in[m]}\sum_{\zeta^T\in\mathcal{A}_{n,\zeta}^{(T)}} \left(\frac{1}{m}\sum_{j\in[m]}\sum_{x^T} P^*_{X^T,\zeta^T|M}(x^T,\zeta^T|j)\right)\mathbb{1}\{\gamma^*(x^T,\zeta^T)= i\} + \exp(-\Omega(T^{\frac{1}{2}}))\\
&\doteq \sum_{i\in[m]}\sum_{\zeta^T\in\mathcal{A}_{n,\zeta}^{(T)}} e^{-T\mathsf{H}(\mathcal{Z})}\mathbb{1}\{\gamma^*(x^T,\zeta^T)= i\} + \exp(-\Omega(T^{\frac{1}{2}}))\\
&= me^{-T\mathsf{H}(\mathcal{Z})} + \exp(-\Omega(T^{\frac{1}{2}}))\\
&\overset{\text{(a)}}{\leq} \alpha + \exp(-\Omega(T^{\frac{1}{2}}))\\
&\xrightarrow{T\to\infty} \alpha,
\end{aligned}$$

where (a) follows from the condition $\log m \leq \log\alpha + T\mathsf{H}(\mathcal{Z}) = \log\alpha + T\mathsf{H}_Q(\mathcal{X})$ in Theorem 4.

For any $x^T \in (\mathcal{A}_{\eta,X}^{(T)})^{\mathrm{c}}$,

$$\sum_{\zeta^T} P^*_\zeta(\zeta^T)\mathbb{1}\{\gamma^*(x^T,\zeta^T)\neq 0\} = 0.$$

Since any distribution Q_X^T can be written as a linear combinations of $\{\delta_{x^T}\}_{x^T\in\mathcal{X}^T}$, we have

$$\sup_{Q_X}\beta_0(\gamma^*, Q_X\otimes P^*_\zeta) = \sup_{Q_X}\sum_{x^T,\zeta^T} Q_X^T(x^T)P^*_\zeta(\zeta^T)\mathbb{1}\{\gamma^*(x^T,\zeta^T)\neq 0\} \to \alpha, \text{ as } T\to\infty.$$