OPTIMIZED COUPLINGS FOR WATERMARKING LARGE LANGUAGE MODELS

Carol Xuan Long* Harvard University
Dor Tsur* Ben Gurion University
Claudio Mayrink Verdun Harvard University
Hsiang Hsu Harvard University
Haim Permuter Ben Gurion University
Flavio P. Calmon Harvard University

ABSTRACT

Large language models (LLMs) are now able to produce text that is indistinguishable from human-generated content. This has fueled the development of watermarks that imprint a "signal" in LLM-generated text with minimal perturbation of an LLM's output. This paper provides an analysis of text watermarking in a one-shot setting. Through the lens of hypothesis testing with side information, we formulate and analyze the fundamental trade-off between watermark detection power and distortion in generated textual quality. We argue that a key component in watermark design is generating a coupling between the side information shared with the watermark detector and a random partition of the LLM vocabulary. Our analysis identifies the optimal coupling and randomization strategy under the worst-case LLM next-token distribution that satisfies a min-entropy constraint. We provide a closed-form expression of the resulting detection rate under the proposed scheme and quantify the cost in a max-min sense. Finally, we numerically compare the proposed scheme with the theoretical optimum.

1 INTRODUCTION

A large language model (LLM) is a generative model that, given a string of input tokens, outputs a probability distribution Q_X for the next token X in the sequence. The emergence of LLMs that generate text that is largely indistinguishable from humans has led to the creation of trustworthy text generation algorithms Huang et al. (2024) that create safe Bai et al. (2022), interpretable Geva et al. (2021), and authentic Lin et al. (2022) content. This work focuses on *watermarking*: the process of embedding a "signal" at the token level in LLM-generated text. The goal of a watermark is to enable automated detection of AI-generated content, providing proof of its authenticity (or lack thereof) and potentially of its origin. The past two years have witnessed the creation of increasingly sophisticated LLM watermarking schemes Kirchenbauer et al. (2023); Christ et al. (2024); Kuditipudi et al. (2023); Zhao et al. (2024a); Aaronson (2023); He et al. (2024); Bahri et al. (2024); Dathathri et al. (2024); Yang et al. (2023); Ren et al. (2024); Hu et al. (2024); Zhao et al. (2024c); Chao et al. (2024); Qu et al. (2024); Xie et al. (2024); Liu & Bu (2024); Fernandez et al. (2023).

A hallmark of existing LLM watermarks is their reliance on either distorting or coupling the next-token distribution Q_X with a random variable S drawn from a known distribution P_S. Here, S represents shared randomness known both by the watermark generator and detector. For instance, Kirchenbauer et al. (2023) – which ignited the recent interest in LLM watermarking in the machine learning community – distorts Q_X by randomly choosing a set of tokens (as determined by S) to be on a "green list," i.e., a subset of tokens that are favored during generation, and increasing the mass of those tokens accordingly. The detector then counts the number of tokens in a sequence that appears on the green list and declares the text watermarked (i.e., AI-generated) if this count exceeds a threshold. However, such a distortion of the LLM distribution may impair the textual quality. Alternative approaches include Aaronson (2023); Kuditipudi et al. (2023); He et al. (2024); Chao et al. (2024), which instead couple Q_X with the distribution P_S. Such couplings enable "distortion-free" watermarks that (averaged over P_S) do not change the expected next-token distribution, yet are still detectable.

The exact nature of the shared randomness S between the model and the detector varies across watermark implementations. S can be, for example, generated from the hash of previous tokens in a sequence Kirchenbauer et al. (2023) (where a hash function converts the token history into a fixed-size value that deterministically produces pseudo-random bits) or sophisticated tournament-like sampling strategies Dathathri et al. (2024). For our theoretical analysis, we abstract away the exact generation process of the shared randomness S.

At a high level, existing LLM watermarks perform two steps when generating a sequence of tokens $\{X_i\}_{i=1}^n$ given shared randomness $\{S_i\}_{i=1}^n$:

1. *Watermark Generation:* For the i-th generated token and given S_i and the predicted next token distribution Q_X, draw the next token by sampling from $X_i \sim \tilde{Q}_{X|S_i}$.
2. *Detection:* Given a sequence $\{(X_i, S_i)\}_{i=1}^n$, compute the statistic $T_n = \frac{1}{n}\sum_{i=1}^n f(X_i, S_i)$ for some function $f : \mathcal{X} \times \mathcal{S} \to [0, 1]$, and declare that the sequence $\{X_i\}_{i=1}^n$ is watermarked if $T_n \geq \tau$.

Importantly, a crucial assumption of current LLM watermarking schemes is that the function f *does not* assume knowledge of the token distribution Q_{X^n}. This allows watermarks that are directly detectable from the sequence $\{(X_i, S_i)\}_{i=1}^n$, i.e., directly from generated text, without accessing the underlying LLM. If the distribution of the generated tokens Q_{X^n} was known, then a standard likelihood ratio test (LRT) would suffice for watermark detection. What makes LLM watermarking distinct from existing information-theoretic watermarking schemes (e.g., Gel'Fand & Pinsker (1980); Willems (2000); Chen (2000); Moulin & O'Sullivan (2003); Martinian et al. (2005); Villán et al. (2006)) are the assumptions that (i) the source distribution is unknown to the watermark detector and (ii) watermarking is performed on a per-token (vs. sequence) level.

1.1 Main Contributions

Motivated by the success of token-level schemes for LLM watermarking, we provide an in-depth analysis of a single-token watermarking process, i.e., when $n = 1$. Specifically, we study how to generate a coupling $\tilde{Q}_{X,S}$ and the corresponding detection function f that maximizes the probability of detection of the watermark, while controlling the quality of the text. The latter is controlled through the distortion relative to Q_X – a quantity we call *perception*, following recent trends in the information theory literature on the source coding problem Blau & Michaeli (2019); Theis & Wagner (2021); Chen et al. (2022). We refer to this setting as *one-shot watermarking*. We jointly optimize $\tilde{Q}_{X,S}$ and f given a perception constraint, with the case $\bar{Q}_X = Q_X$ corresponding to the *perfect perception* setting. We focus on one-shot watermarking since, as mentioned above, existing schemes are constrained to watermark on a token-by-token basis. Moreover, small gains in single-token watermark detection compound to exponential gains in detection accuracy in threshold tests applied across multiple tokens.

We begin with an information-theoretic formulation for one-shot watermarking. We quantify the fundamental trade-off between watermark detection vs. perception when the underlying next-token distribution Q_X is known with the side information P_S uniformly distributed. This analysis yields a fundamental upper bound on one-shot watermark performance; see Theorems 1 and 2. Interestingly, when the watermark does not change the next-token probability (i.e., perfect perception), optimizing a one-shot watermark is equivalent to maximizing the TV-information $\mathsf{TV}\left(Q_{X,S} \| Q_X P_S\right)$ across the coupling $Q_{X|S}$ – a non-convex optimization problem (Polyanskiy & Wu, 2024, Section 7). This formulation embeds TV-information with a new operational interpretation.

We optimize one-shot watermarks when Q_X is unknown to the detector but satisfies a min-entropy constraint, i.e., $\|Q_X\|_\infty \leq \lambda$ (Eq. (6)), which corresponds to $H_\infty(Q_X) \geq -\log(\lambda)$. Operationally, lower values of λ correspond to higher entropy token distributions with greater uncertainty, while higher values of λ indicate more concentrated distributions where the next token is more predictable. Moreover, we optimize for detection tests of the form $\mathbf{1}[f(X) = S]$, where $f : \mathcal{X} \to \mathcal{S}$ forms a partition of $\mathcal{X}$.

Motivated by the fact that deterministic token partitions lead to low detection probabilities, we introduce randomness to f. In Theorem 3, we analyze the probability of detection of such detection tests under the worst-case token distribution. We pair our analysis with a characterization of

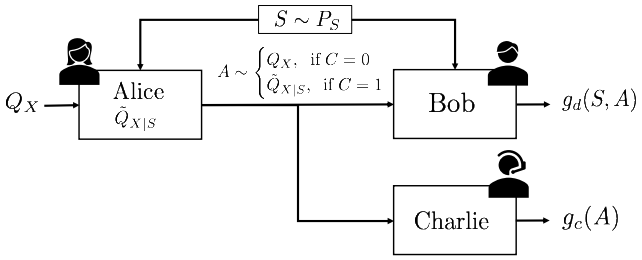


Figure 1: Watermarking problem as a hypothesis test with side information.

the optimal design of the partition randomization. In Theorem 4, we consider a simplified token partition strategy and show that it yields a near-optimal detection probability. Together, we provide a complete characterization of the minimax detection rate for a given vocabulary size, side information, and min-entropy constraint under the optimal and near-optimal partition randomization strategies. Lastly, we provide numerical results of the Correlated Channel (CC) benchmarked against Gurobi-based optimum couplingGurobi Optimization, LLC (2024) and the red/green watermarkKirchenbauer et al. (2023).

Related Work. Watermarking has been extensively studied in information theory Chen (2000); Moulin & O'Sullivan (2003); Martinian et al. (2005), particularly through the Gelfand-Pinsker (GP) channel Gel'Fand & Pinsker (1980); Villán et al. (2006); Willems (2000). These approaches typically focus on watermarking sequences via joint typicality and assume perfect knowledge of the underlying source distribution. The work of Kirchenbauer et al. (2023) led to various developments in watermarking schemes Aaronson (2023); He et al. (2024); Bahri et al. (2024); Dathathri et al. (2024); Yang et al. (2023); Ren et al. (2024); Hu et al. (2024); Zhao et al. (2024c); Chao et al. (2024); Qu et al. (2024); Xie et al. (2024); Liu & Bu (2024), with several approaches focusing on distortion-free methods, e.g., Kuditipudi et al. (2023); Hu et al. (2024); Zhao et al. (2024c); Christ et al. (2024). In particular, Chao et al. (2024) proposes a watermark using error-correcting codes leading to correlated channels similar to the ones we find via optimizing couplings. In Huang et al. (2023), the optimal Type-II error for bounded Type-I error is analyzed by comparing watermarking schemes to the uniformly most powerful watermark with knowledge of Q_X. The authors of He et al. (2024) characterize the universal Type II error while controlling the worst-case Type-I error by optimizing the watermarking scheme and detector. While these works operate on a token-level basis, they focus on the effect of a given strategy along a sequence. In contrast, we focus on a preliminary step and aim to answer the simple yet important question – *What is the optimal coupling when watermarking a single token?*

2 Optimal One-Shot Watermarking

In this section, we formulate the watermarking problem, derive the resulting optimization problem, and discuss the optimal solution structure. We focus on the fundamental trade-off between detection probability and perceptual quality. As mentioned above, while the optimal approach to watermarking considers sequence-to-sequence schemes, due to the autoregressive nature of token generation in LLMs most popular schemes focus on token level strategies Kirchenbauer et al. (2023); Aaronson (2023); He et al. (2024); Dathathri et al. (2024). As a first step towards token-level watermarking of sequences, we provide an extensive analysis of the one-shot setting. We discuss the extension to a token-level scheme in the sequential case in Section 4.

2.1 Problem Setting

We consider a hypothesis test using the private side information setting and textual quality of the model as the ability of an external observer to detect the watermark without access to the side information. Formally, let Q_X be the LLM distribution over some finite vocabulary of $|\mathcal{X}| = m$ tokens. We consider *Alice* (the watermarker), whose goal is to convey a single token to *Bob* (the detector), which, in turn, tries to detect whether the token is watermarked or not. Alice and Bob

share some random side information[1] $S \sim P_S$ with $|\mathcal{S}| = k$. Furthermore, we consider *Charlie* (average observer), which tries to detect the existence of the watermark but does not have access to the side information. The setting is depicted in Figure 1.

On Alice's end, the watermark design boils down to the construction of the conditional distribution $Q_{X|S}$. We consider a Bayesian setting, in which Alice transmits a token according to a uniform prior:

$$A = \begin{cases} X \sim Q_X & \text{if } C = 0, \\ \tilde{X} \sim Q_{X|S} & \text{if } C = 1. \end{cases}, \quad C \sim \mathsf{Ber}\left(\frac{1}{2}\right) \tag{1}$$

where $C \perp\!\!\!\perp (X, \tilde{X}, S)$. To detect the watermark, Bob performs the following hypothesis test

$$H_0 : A \sim Q_X$$
$$H_1 : A \sim Q_{X|S}.$$

We assume that Charlie is aware of the watermarking mechanism but is not aware of the specific sample of S. Therefore, Charlie performs an hypothesis test with a corresponding alternative hypothesis, i.e.

$$H_0 : A \sim Q_X$$
$$H_1 : A \sim \bar{Q}_X,$$

where $\bar{Q}_X \triangleq \mathbb{E}_S[Q_{X|S}]$ is the watermark distribution averaged w.r.t. the side information S.

2.2 A Detection-Perception Perspective

Given the hypothesis test formulation, we recast the problem of watermarking as a trade-off between two measures: Bob's *detection* and Charlie's *perception* probabilities. Motivated by recent advances in lossy source-coding Blau & Michaeli (2019); Theis & Wagner (2021); Chen et al. (2022), we adopt the notion of perceptual qualities of the data, which is quantified through a discrepancy measure between the two distributions, e.g. f-divergences, rather than a metric calculated directly on the random variables.

We define two fundamental metrics that capture the trade-off between detection capability for Bob and imperceptibility for Charlie. For Bob's detection capability, we weigh true negative (TN) detections with prior π_0 and true positive (TP) detections with prior $\pi_1 = 1 - \pi_0$. The tests are defined as follows:

Definition 1 (Watermark Tests and Error Probabilities) *A watermarking scheme comprises of a detection test $g_d : \mathcal{X} \times \mathcal{S} \to \{0, 1\}$, such that for $(A, S) \in \mathcal{X} \times \mathcal{S}$, we respectively define the detection probability with prior $\pi = (\pi_0, \pi_1)$ as*

$$R_d \triangleq \mathbb{E}_\pi \left[\Pr(g_d(S, A) = C)\right].$$

Perception probability R_p is similarly defined with a test $g_c : \mathcal{X} \to \{0, 1\}$ and a uniform prior $\pi_0 = 1/2$.

Optimally, we aim to optimize detection R_d while lowering R_p, which indicate Charlie's low perception of the watermark. The metrics detection and perception are formalized next.

2.3 Characterizing Optimal Trade-off

Following the Neyman-Pearson Lemma Lehmann et al. (1986), the likelihood ratio gives the optimal test statistic, and (R_d,R_p) have a simple form in terms of E_γ (or hockey-stick) divergence. The next proposition is a direct result of the well-known connection between E_γ and hypothesis testing; see, e.g., Polyanskiy (2010); Polyanskiy et al. (2010); Liu et al. (2016).

[1] Side information often corresponds to a secret shared key; see, e.g., Kuditipudi et al. (2023); Zhao et al. (2024b).

Proposition 1 *Fix $(P_S, Q_X, \tilde{Q}_{X|S})$ and error prior π. Let $\gamma = \frac{\pi_1}{\pi_0}$. Using the LRT, the optimal detection and perception probabilities are given by*

$$R_d = \pi_1 + \pi_0 \mathsf{E}_\gamma \left(\tilde{Q}_{X|S} P_S, Q_X P_S \right), \tag{2}$$

$$R_p = \frac{1}{2} + \frac{1}{2} \mathsf{TV} \left(\tilde{Q}_X, Q_X \right). \tag{3}$$

Remark 1 *The E_γ divergence characterizes the error of hypothesis tests with specified priors on TP and TN rates. It can be defined as*[2] *in Liu et al. (2016)*

$$\mathsf{E}_\gamma(P, Q) \triangleq \max_{\mathcal{A}} [P(\mathcal{A}) - \gamma Q(\mathcal{A})],$$

where $\mathcal{A}$ are rejection regions, $P(\mathcal{A})$ and $Q(\mathcal{A})$ are $1-$TN rate and TP rate, respectively. When. $\pi_0 = \pi_1$ and $\gamma = 1$, detection probability boils down to the total variation (TV) distance, in which case, we have $R_d = \frac{1}{2} + \frac{1}{2}\mathsf{TV}(\tilde{Q}_{X|S}, Q_X|P_S)$, where $\mathsf{TV}(\tilde{Q}_{X|S} P_S, Q_X P_S) = \mathsf{TV}(\tilde{Q}_{X|S}, Q_X|P_S)$.

Our hypothesis testing framework employs priors π_0 and π_1 to explicitly weight the importance of different error types in the detection process. Setting $\pi_0 = \pi_1 = \frac{1}{2}$ gives equal importance to both errors, whereas asymmetric values prioritize either minimizing false positives (incorrectly flagging human content as AI-generated) or false negatives (failing to detect AI-generated content). This Bayesian framework provides a principled approach to designing watermark schemes with detection rates optimized for specific operational requirements, where the relative costs of different error types may vary significantly across applications.

Due to Jensen's inequality, for any fixed $(P_S, \tilde{Q}_{X|S})$, we have $R_p \leq R_d$, i.e., Bob's access to the shared side information allows for a potentially higher detection probability. Generally, for any perception constraint $\alpha_p \in [1/2, 1]$, the optimal detection probability is given by the solution to the following optimization:

$$\sup_{\tilde{Q}_{X|S}} \mathsf{E}_\gamma \left(\tilde{Q}_{X|S}, Q_X|P_S \right), \quad \text{s.t.} \quad \mathsf{TV} \left(\tilde{Q}_X, Q_X \right) \leq \alpha_p. \tag{4}$$

We are interested in characterizing the (R_d, R_p) trade-off region, which amounts to solving (4) as a function of α_p.

Note that (4) is a non-convex optimization problem. However, in what follows, we characterize the several corner points of the optimal curve (i.e., $R_p = 0.5$), which, in turn, gives insight into the structure of the (R_d, R_p) region within the box $[\frac{1}{2}, 1]^2$.

We provide a complete characterization of the fundamental limits of detection probability under zero perception (where $\tilde{Q}_X = Q_X$). The following result establishes tight bounds on the optimal detection probability in this regime

Theorem 1 (Zero perception bounds) *Fix Q_X and let P_S be uniform over $\mathcal{S}$, $|\mathcal{S}| \leq |\mathcal{X}|$ and let $\pi_1 = \frac{1}{2}$. Then, for $R_p = \frac{1}{2}$, we have*

$$\frac{1}{2} \leq \sup_{\tilde{Q}_{X|S}} R_d \leq \max \left(\frac{1}{2}, 1 - \frac{\gamma}{2k} \right). \tag{5}$$

The upper bound emerges from jointly optimizing over both the coupling $\tilde{Q}_{X|S}$ and Q_X. This optimization reduces to a convex problem over the probability simplex, which we recast as counting the optimally assigning elements of $\mathcal{X}$. The lower bound is achieved when Q_X is a singleton.

Beyond characterizing the zero-distortion endpoints, we derive an upper bound on the detection probability that holds across all perception levels. The bound is given as follows:

Theorem 2 (Uniform Detection Upper Bound) *Let $Q_{\mathsf{min}} \triangleq \min_{x \in \mathcal{X}} Q_X(x)$. For any $R_p \geq 0$ we have $R_d \leq 1 - \frac{\gamma Q_{\mathsf{min}}}{2}$.*

[2]Some works include a residual term $(1-\gamma)_+$ Asoodeh et al. (2020), which we omit for convenience as it does not affect the optimization problem.

This bound emerges from analyzing a simple strategy of replacing each token with the least likely symbol in the LLM's vocabulary. The structure of the optimization (4) results in a nonconvex region, which generally lacks a closed form. This non-convexity is demonstrated in our experimental results, see Section 5, where exact solvers are used to compute the trade-off region. In light of this challenge, we will next derive a simple and tractable watermarking scheme.

3 A One-Shot Watermarking Scheme

While the optimal test that maximizes Bob's detection accuracy is the LRT, it is infeasible in practical scenarios where Bob is not assumed to have access to Q_X. To make use of the shared side information, Bob and Alice look for a mechanism that couples S with the token distribution. This can be done by applying a map $f : \mathcal{X} \to \mathcal{S}$. Alice uses $(f(X), S)$ to construct a watermarked distribution, and Bob uses $(f(A), S)$ to detect its presence. We note that a map f creates a partition of $\mathcal{X}$ into $\mathcal{S}$ bins. When $k = 2$, this can be interpreted as a partition of $\mathcal{X}$ into a rejection region and its complement. We note that considering deterministic mappings is insufficient, as for $S \sim \mathsf{Unif}([1 : k])$, the detection probability is $\frac{1}{k}$, independent of the choice of (f, Q_X). Therefore, we introduce randomness into our partitioning approach by making the function f stochastic rather than deterministic. Specifically, we define a randomized mapping that varies the way tokens are assigned to each partition based on additional random variables that both Alice and Bob can access.

3.1 Optimal Randomized Partition – Correlated Channel

We randomize f by introducing a set of m $\mathcal{S}$-valued random variables denoted B^m. We assume that B^m is publicly available to all parties and is therefore not considered a part of the private side information S. Our goal is therefore to couple the side information with the randomized mapping $f(X, B^m)$. This boils down to finding a coupling of Q_X and S through the design of partition randomness P_{B^m} and conditional distribution $Q_{X|S}$. We look for such $(P_{B^m}.Q_{X|S})$ that are optimal under the worst choice of token distribution Q_X within a given class. Our problem is therefore formally given by the following max-min expression

$$R_d^\star(\lambda) \triangleq \max_{P_{B^m}} \min_{\substack{Q_X \in \Delta_m \\ \|Q_X\|_\infty \leq \lambda}} \mathbb{E}\left[R_d(Q_X, B^m)\right], \tag{6}$$

where $\|Q_X\|_\infty = \max_{x\in\mathcal{X}} Q(x)$. As discussed in Section 1.1, we consider the constraint $\{Q_X \in \Delta_m, \|Q_X\|_\infty \leq \lambda\}$ which enables a more comprehensive analysis by allowing us to adjust the parameter λ. This flexibility provides insights across various scenarios: smaller λ values yield higher entropy token distributions with greater uncertainty, while larger λ values produce more deterministic distributions with reduced uncertainty about the next token.

According to (6), given a fixed pair (P_{B^m}, Q_X), we maximize $R_d(Q_X, B^m)$ by designing the coupling of $(f(X, B^m), S)$. We consider the mapping of the form[3] $f(x, b^m) = b_x$ under which, the partition's probabilities are characterized by the distribution of the random variable $Y \triangleq f(X, B^m)$. To this end, we first solve the following optimization problem:

$$\sup_{P_{S,Y}} \Pr(S = Y), \quad S \sim \mathsf{Unif}\left(\mathcal{S}\right), Y \sim P_Y. \tag{7}$$

This is a maximum coupling problem whose closed-form solution is given below. It is a direct consequence of the inf-representation of TV distance Polyanskiy & Wu (2022).

Proposition 2 *Let $S \sim \mathsf{Unif}[1 : k]$ and $P_Y = \{p_1, \ldots, p_k\} \in \Delta_k$, $t = \mathsf{TV}(P_S, Y)$ and let Π be the set of all couplings of (P_S, P_Y). Then, $\arg\max_{\pi\in\Pi} \Pr(S = Y)$ is given by*

$$\pi(Y = i, S = j) = \begin{cases} \min(\frac{1}{k}, p_i), & i = j, \\ \frac{1}{t}(\frac{1}{k} - p_i)(p_j - \frac{1}{k}), & (i \in A) \cap (j \in A^c), \\ 0, & \textit{otherwise}, \end{cases}$$

where $A = \{i : p_i \geq \frac{1}{k}\}$, and $A^c = [k] \setminus A$.

[3] We consider a vocabulary $\mathcal{X} = [1 : m]$, which can be thought of as the enumeration of the tokens.

Algorithm 1 Correlated Channel Watermark (CC)

Require: LLM distribution Q_X, Side information S, shared randomness B^m.
1: **Alice:**
2: Generate $\tilde{Q}_{X|S,B^m}$ according to (8)
3: Flip a coin $C \sim \mathsf{Ber}(\frac{1}{2})$ and sample A according to (1).
4: **Bob:**
5: **if** $S = f(A, B^m)$ – Declare **Watermarked**
6: **else** – Declare **Not watermarked**

Figure 2: Optimal coupling between side information S and random partition $Y = f(X, B^m)$ for $\tilde{p}_1 \leq 0.5$ (**left**), $\tilde{p}_0 \leq 0.5$ (**right**), with $\beta(p) = \frac{2p-1}{2p}$.

The resulting coupling can be thought of as a transition kernel that maps P_Y to P_S under maximum acceptance probability. When $k = 2$, the optimal coupling boils down to a binary asymmetric channel, known in information theory as the Z-channel Cover & Thomas (2006). That is, when $S = 0$, the mapping always outputs $Y = 0$, but when $S = 1$, the mapping may output either $Y = 1$ or $Y = 0$ with certain probabilities. This asymmetric structure is particularly effective for watermark detection because it creates a distinctive pattern that appears only in watermarked content. We therefore term this method as the correlated channel (CC) watermark. We note that CC was previously considered, for example, in Chao et al. (2024).

The CC scheme consists of the following steps: Both Alice and Bob observe (s, b^m). Alice samples $C \sim \mathsf{Ber}(\frac{1}{2})$. If $C = 0$, she samples $a \sim Q_X$ and sends it. Otherwise, she samples and sends $a \sim \tilde{Q}_{X|S=s}$, which is given by the CC:

$$\tilde{Q}_{X|s,b^m}(x) = Q_X(x)\frac{P_{S|Y}(s|f(x,b^m))}{P_S(s)}. \tag{8}$$

Bob performs the detection test by declaring that a is watermarked if $s = f(a, b^m)$. The complete list of steps is summarized in Algorithm 1. Note that by coupling (P_Y, P_S), we result with a coupling of (Q_X, P_S). Consequently, we have $Q_X = \mathbb{E}_S[Q_{X|S}] = \bar{Q}_X$, which implies that the CC watermark has zero perception.

3.2 Theoretical Analysis of the CC scheme

We provide a complete analysis of the CC scheme under $k = 2$. Given the optimal coupling, we give a closed-form expression for R_d in terms of the TV surrogate of mutual information in the resulting channel.

Proposition 3 *The CC watermark detection is given by*

$$R_d = \frac{1}{2}\left(1 + \mathsf{TV}\left(P_S, P_{S|Y}|P_Y\right)\right) = 1 - \frac{1}{2k} - \frac{1}{2}\mathsf{TV}\left(P_Y, P_S\right). \tag{9}$$

Proposition 3 provides a closed-form characterization of Bob's detection probability as a function. Specifically, for $k = 2$, we have $R_d = \frac{1}{2}(1 + \tilde{p})$, where $\tilde{p} \triangleq \min(\tilde{p}_0, \tilde{p}_1)$. This term is maximized when $Y \sim \mathsf{Ber}(\frac{1}{2})$, with maximum value of $\frac{3}{4}$. A consequence of Proposition 3 is that we are interested in designing a partition that is as close as possible to P_S as possible. As P_S is uniform over $\{1, \ldots, k\}$, our aim is to obtain a uniform distribution, i.e., a balanced partition of the token vocabulary $\mathcal{X}$, given the token distribution Q_X and the partition randomness P_{B^m}.

Remark 2 (Equivalence to the likelihood ratio test) *When we consider the indicator test* $\mathbf{1}\{f(x, b^m) = s\}$*, the decision region obtained by the CC watermark is equivalent to the one*

attained by the LRT with threshold value of $\tau = 1$. *This follows from the observation that* $\Pr[S|f(S,B^m)] \geq \frac{1}{2}$, *if and only if* $S = f(X,B^m)$.

Next, we discuss the design of randomness. Specifically, we analyze the dependence of the CC watermark detection probability on the distribution of B^m and propose an optimal design of P_{B^m}.

3.3 Optimizing the Partition

As seen in Equation (9), the distribution of the resulting partition governs the detection power of the CC watermark. The partition distribution is determined by the token distribution Q_X and the distribution of B^m. As Q_X cannot be controlled by the watermark designer, we aim to characterize the class of distributions P_{B^m} that maximizes R_d under the worst-case adversarial distribution Q_X. Due to the symmetry of the CC, we can restrict the optimization over permutation classes of P_{B^m}. First, we show that the optimal distribution P_{B^m} is permutation invariant.

Lemma 1 *Let* $F(P_{B^m}) \triangleq \min_{\substack{Q_X \in \Delta_m \\ \|Q_X\|_\infty \leq \lambda}} \mathbb{E}_{P_{B^m}}[R_d(Q_X, B^m)]$. *Let* $P^\star_{B^m}$ *be a distribution that maximizes* $F(P_{B^m})$. *Consider a permutation* $\phi : \mathcal{S}^m \to \mathcal{S}^m$ *and define* $\tilde{P}_\phi(B^m) = P^\star_{B^m}(\phi \circ B^m)$. *Then,* $F(P^*_{B^m}) = F(\tilde{P}_\phi)$.

Next, let $\mathcal{P}_m = \{\mathcal{B}_1, ..., \mathcal{B}_K\}$ be the partition of $\mathcal{S}^m$ into K sets of sequences that are identical up to a permutation. We refer to each $\mathcal{B}_i$ as a permutation class. We proceed to characterize the optimal mean detection probability $R^\star_d$ and the corresponding distribution $P^\star_{B^m}$.

Theorem 3 (Optimal max-min Detection) *Let* $|\mathcal{S}| = k$ *and* $\mathcal{X} = m$, *and assume that* m *is divisible by* k. *Given min-entropy constraint* $\lambda \in [0,1]$, *and let* $t = \lfloor \frac{1}{\lambda} \rfloor$. *The optimal minimax detection probability from Equation 6 is given by:*

$$R^*_d(\lambda) = 1 - \frac{1}{2k} - \frac{1}{4}\mathbb{E}[g(Q^*_\lambda, B^m)], \tag{10}$$

where

$$\mathbb{E}[g(Q^*_\lambda, B^m)] = k \sum_{c=0}^{t} \frac{\binom{m/k}{c}\binom{m-m/k}{t-c}}{\binom{m}{t}} \left(\left(\frac{(m/k)-c}{m-t} \right) \left| c\lambda + (1-\lambda t) - \frac{1}{k} \right| + \left(1 - \frac{(m/k)-c}{m-t} \right) \left| c\lambda - \frac{1}{k} \right| \right).$$

Furthermore, the optimal detection probability is achieved for $P^*_{B^m}$ *corresponding to uniform sampling over the permutation class of the sequence with an equal number of each element. For* $|\mathcal{S}| = 2$, $P_{B^m} = \mathsf{Unif}(B^\star)$, *where* $B^\star = \{b^m \in \{0,1\}^m | b^m$ *has equal number of 1's and 0's*$\}$.

Under additional assumptions, we can further simplify the optimal detection.

Corollary 1 *Under the setting of Theorem 3, assume that* $\lambda = \frac{1}{k}$. *Then, we have*

$$R^*_d(\lambda) = 1 - \frac{1}{2k} - \frac{1}{2}\frac{\binom{(k-1)m/k}{k}}{\binom{m}{k}}. \tag{11}$$

Furthermore, if $k = 2$ *and* $\lambda \in [\frac{1}{3}, 1]$ *we have*

$$R^\star_d(\lambda) = \begin{cases} \frac{3}{4} - \frac{m\lambda - 1}{4(m-1)}, & \text{for} \quad \frac{1}{2} \leq \lambda \leq 1 \\ \frac{3}{4} - \frac{m-2}{8(m-1)}, & \text{for} \quad \frac{1}{3} \leq \lambda < \frac{1}{2}. \end{cases} \tag{12}$$

Here, we have characterized detection for the worst-case distributions $Q^\star_\lambda$, which lie at the extreme point of the feasible set — probabilities with bounded inf norm $\|Q_X\|_\infty \leq \lambda$). For example, for $\lambda \in [0.5, 1]$, the above minimax detection probability corresponds to token distributions with only two nonzero entries, i.e., Q_X takes the form $[\lambda, 1-\lambda, 0, ..., 0]$; for $\lambda \in [\frac{1}{3}, \frac{1}{2}]$, the worst-case token

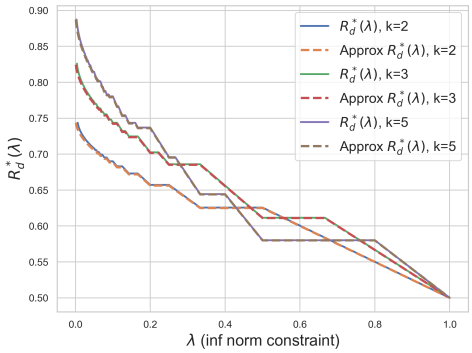


Figure 3: Optimal detection probability of CC in one-shot on the adversarial token distribution (Eq. 6) is plotted against the inf-norm constraint λ (or equivalently, an entropy constraint) on Q_X[3]. When $\lambda = 1$ (entropy $H(Q_X) = 0$) , Q_X is deterministic, and detection is random. As entropy of Q_X grows (moves to smaller λ values), single-token optimal detection probability reaches a maximum of around 0.75 for binary side information. If the side information one transmits contain a larger set of values, CC achieves a higher detection probability correspondingly. The actual detection rate (solid lines) and approximate solutions (dotted lines) overlap for large enough vocabulary size[4], and their exact forms are provided in Theorem 3 and 4.

distribution have 3 non-zero elements and has the form $[\lambda, \lambda, 1-2\lambda, 0, ..., 0]$. Furthermore, we note that due to Equation (9), when $k = 2$, R_d is upper bounded by $\frac{3}{4}$. Thus, the second term in (12) serves as a penalty when considering the max-min setting. Notably, for $\lambda \in [0.5, 1]$ and when m is large, this penalty equals $\frac{\lambda}{4}$, which implies that the cost of considering worst-case token distributions is lower bounded by $\frac{1}{8}$.

In addition to characterizing the minimax detection rate, Theorem 3 shows that the optimal sampling strategy for token partition B^m is to sample uniformly from a collection of sets with an equal number of each element in k. Next, we show that we can adopt a much simpler sampling strategy, sampling i.i.d. Bernoulli variables with probability $\frac{1}{k}$ and arrive at a near-optimal detection probability. In Figure 3, we plot the probability of detection of both sampling strategies and show that the Bernoulli sampling strategy results in negligible approximation error. To motivate i.i.d. Bernoulli sampling, we start with an alternative view of the optimal sampling strategy in Theorem 3. Sampling a b^m uniformly over $\mathcal{B}^*$ — containing sequences with equal numbers of each element in k — can be equivalently defined as the following process: given m elements with predefined proportions $[\frac{1}{k}, ..., \frac{1}{k}]$, sample m times with replacement. In the following theorem, we obtain an approximation of $R_d^\star$ for any λ by sampling without replacement. We also show that, by applying de Finetti's theorem on finite exchangeable sequencesDiaconis & Freedman (1980), the approximation error decays with $O(\frac{1}{m})$.

Theorem 4 (Approximation of Max-min Detection Rate) *Given $|S| = k$, $|\mathcal{X}| = m$, and the inf-norm constraint $\lambda \in [0, 1]$. Let $t = \lfloor \frac{1}{\lambda} \rfloor$, and $Y \sim Bin(t, \frac{1}{k})$ An approximation of the optimal minimax detection probability is given by:*

$$\tilde{R}_d^\star(\lambda) = 1 - \frac{1}{2k} - \frac{1}{4}\left[\sum_{c=0}^{t} \Pr[Y = c]\left(\left|(c-t)\lambda + (1-\frac{1}{k})\right| + (k-1)\left|c\lambda - \frac{1}{k}\right|\right)\right] \tag{13}$$

The approximation error decays as $O(\frac{1}{m})$. Specifically:

$$\left|\tilde{R}_d^\star(\lambda) - R_d^\star(\lambda)\right| \leq \frac{2k\lceil \frac{1}{\lambda} \rceil}{m} \tag{14}$$

[3]For discrete probability Q_X, inf-norm and entropy are connected via $H(Q_X) \geq -\log\|Q_X\|_\infty$, and we have $\lambda = \|Q_X\|_\infty$.

[4]We take $m = 100k$. Hence, existing LLMs with much larger vocabulary size would produce negligible approximation error.

We plot the results of Theorem 3 and 4 in Figure 3. For all λ and k values, the approximated maxmin detection coincides with the closed-form $R_d^\star(\lambda)$. We choose $m = 100 * k$. The overlap between the actual and approximated $R_d^\star(\lambda)$ in the plot testifies our result that the approximation error decays with m. In practice, since LLMs have a much large vocabulary, where $m \approx 100,000$Grattafiori et al. (2024), the approximation error will be negligible.

4 SEQUENTIAL WATERMARKING

While this paper focused on a single-shot analysis of token distribution watermarking, general text generation involves sequential prediction of long token sequences. A common approach involves applying a token-level watermarking of the next token distribution and designing token-level test statistics. This approach was shown to benefit from favorable performance Kirchenbauer et al. (2023); Aaronson (2023), albeit being theoretically suboptimalHe et al. (2024). We note that our one-shot method readily extends to a sequential token-level scheme as we can treat each step as a one-shot problem, and considering an average test $\frac{1}{n}\sum \mathbf{1}[f_i(A_i, B_i^m) = S_i]$ which we them compare with some threshold $\tau \in [0, 1]$. We leave the theoretical analysis of the token-level extension of our scheme to future work, while showing empirical results in Section 5. In the simplified case when X^n are i.i.d., we provide the following bounds on the detection probability (a related result was given in Chao et al. (2024) bounding mismatch proportion using entropy):

Proposition 4 *Let $Q^n = Q_X^{\otimes n}$ be the an i.i.d. token distribution, let $S^n \sim P_S^{\otimes n}$ and apply the one-shot CC on each step $i \in [1 : n]$, then*

$$1 - 2^{-\left(\frac{n}{2}+1\right)} \left(g(\tilde{p})\right)^n \leq R_d \leq \frac{1}{2}\left(1 + \sqrt{1 - \left(\frac{\left(g(\tilde{p})\right)^2}{2}\right)^n}\right),$$

where $\tilde{p} = \min(\tilde{p}_0, \tilde{p}_1)$ is similarly defined as in the on-shot case, and $g(p) \triangleq p + \sqrt{\frac{1-p}{2}}\left(1 + \sqrt{1-2p}\right), p \in [0, 0.5]$.

The proof utilizes bounds on TV in terms of the Hellinger distance, which benefits from a tensorization.

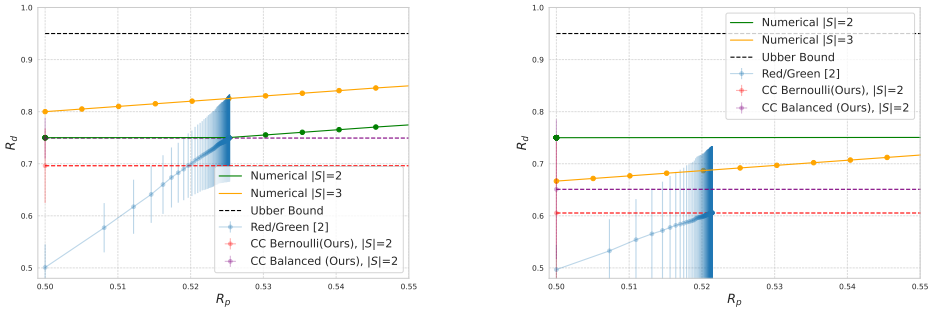
5 EXPERIMENTAL RESULTS

We numerically evaluate the CC watermark on synthetic distributions with various inf-norm constraints. We compareCC with the solution of an exact GUROBI-based numerical solution Gurobi Optimization, LLC (2024) of Eq. (4) and the red/green watermark Kirchenbauer et al. (2023). [5]

5.1 ONE-SHOT PERFORMANCE ANALYSIS

Detection-Perception Tradeoff: We present the (R_d, R_p) trade-off region for the one-shot watermarking setting. We consider the worst-case distribution within $\{Q_x, \|Q_X\|_\infty \leq \lambda\}$. When $\lambda = \frac{1}{m}$, the resulting distribution is simply the uniform distribution over $\mathcal{X}$ and when $\lambda \geq \frac{1}{2}$ it is given by a distribution with two nonzero entries valued $(\lambda, 1-\lambda)$. This distribution is representative of a next-token distribution in the low entropy regime (highly predictable next token). As seen in Figure 4a, for uniform Q_X, when we apply the CC scheme with P_{B^m} sampled over balanced partitions, we obtain a gain of ≈ 0.07 over sampling $B^m \overset{i.i.d.}{\sim} \mathsf{Ber}(\frac{1}{2})$, meeting the upper bound from (4). In contrast, the red-green detection coincides with ours in the limit of $\delta \to \infty$, intersecting with the suboptimal i.i.d. Bernoulli sampling method at $\delta \approx 7.6$. When $\delta = \frac{1}{2}$ we observe a decrease in the gain of sampling from the balanced partition sets. **Effect of k:** Next, we analyze the effect of the side information alphabet size on the CC scheme performance. We present a plot for $m = 10$ which serves as an extension of the performance we present in Figure 4a and a plot for $m = 60$, which allows us to further understand the effect. As seen in Figure 5, as k increases, the detection

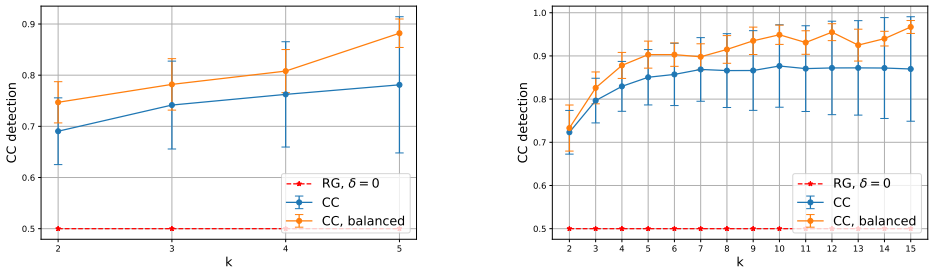
[5] Full implementation details and code are given in `https://github.com/Carol-Long/CC_Watermark`.



(a) $Q_X = \mathsf{Unif}[1:m]$ (b) $Q_X = [0.5, 0.5, 0, \dots]$

Figure 4: One-shot watermark detection results on $Q_X = \mathsf{Unif}(\mathcal{X})$. For $\alpha_p = 0$, CC achieves a detection probability of 0.75 and 0.7 with balanced and Bernoulli partitions, respectively. CC Balanced achieves the optimal detection (Eq. 4 with $\gamma = 1$ and $|\mathcal{S}| = 2$). Standard deviations plotted as two-sided bars.

rate of the CC watermark increases. However, the gain from increasing k decreases as k grows (or alternatively, as the ratio m/k decreases). Furthermore, we note that the performance depends on the divisibility of m by k; when m/k is not an integer, we experience a degradation of performance. This follows from the inability to construct equally sized partitions of $\mathcal{X}$, which, in turn, decreases the probability to result with a balanced partition.



(a) $m = 10$ (b) $m = 60$

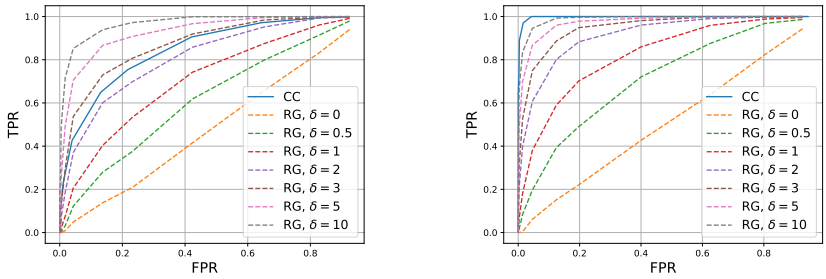
Figure 5: Detection probability vs. k for two values of m and a uniform token distribution Q_X.

5.2 Sequential Watermarking

We now present the performance of the CC watermark on a sequence level scheme. We present preliminary results on synthetically generated data, with the purpose of demonstrating the applicability of our method to a sequence-level test. To that end, we consider the generation of n tokens A^n, which are generated from a sequence of tokens $X^n \overset{i.i.d.}{\sim} Q_X$ using from n i.i.d. samples of side information s^n and randomness $(B^m(i))_{i=1}^n$. We apply the token-level watermarking scheme to each element X_i to generate A_i and apply the following sequence-level threshold test

$$r(A^n, S^n) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(f(A_i, B^m(i)) = S_i\right) \geq \tau \right\}$$

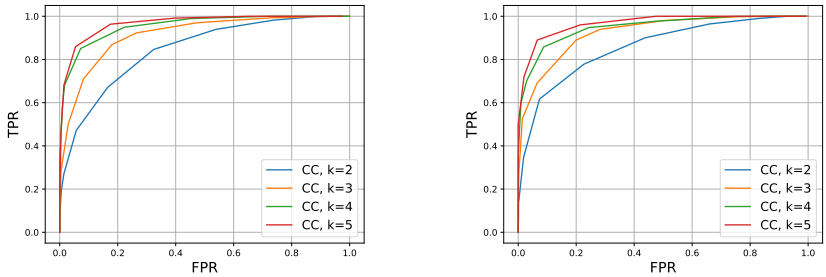
for some threshold $\tau \in [0, 1]$. To understand the performance of the proposed sequence-level generalization, we analyze the ROC of the results scheme. In out experiment, we consider $k = 2$, $m = 20$ and a sequence of $n = 50$ tokens. Figure 6 compares the ROC of the CC scheme (sampling from balanced sets) with the red-green scheme for a range of δ values. We note that, while the CC method is perceptionless, it results in a better ROC than the red-green method. Specifically, for $\lambda = 0.5$, the CC method demonstrated better detection than the red-green method for the considered range of δ values. However, when $\lambda = 0.8$, i.e., when the distribution is spikier, the red-green method with higher δ values result in a better ROC than the CC method, but at the cost of nonzero perception.



(a) $\lambda = 0.8$ (b) $\lambda = 0.5$

Figure 6: ROC of the sequence-level watermarking scheme. We compare the red-green method Kirchenbauer et al. (2023) with the CC scheme (Section 3). We consider a range of δ. An increase of δ increases detection, at the expense of higher perception (lower textual quality), while the CC method has fixed zero perception.

Finally, we analyze the effect of k on performance in the sequential setting by observing the ROC for a range of k values. Specifically, we consider $m = 20$ and apply the sequential generalization of the CC watermark for $k \in \{2, 3, 4, 5\}$. We consider two distributions within the bounded infinity norm set with $\lambda = 0.8$. As can be seen in Figure 7, as k increases, the ROC improves.



(a) Q_X has two nonzero values of $Q_1 = 0.8$ and $Q_2 = 0.2$.

(b) Q_X has a spike of $\lambda = 0.8$ and is uniform on the rest of its entries.

Figure 7: ROC of the sequence-level watermarking scheme under CC method for a range of k values.

6 Conclusion

This work presents a rigorous analysis of text watermarking in a one-shot setting through the lens of hypothesis testing with side information. We analyze the fundamental trade-off between watermark detection power and distortion in generated textual quality. A key insight of our approach is that effective watermark design hinges on generating a coupling between the side information shared with the watermark detector and a random partition of the LLM vocabulary. We develop a perfect perception watermarking scheme – the Correlated Channel Watermark (CC). Our analysis identifies the optimal coupling and randomization strategy under the worst-case LLM next-token distribution that satisfies a min-entropy constraint. Under the proposed scheme, we derive a closed-form expression of the resulting detection rate, quantifying the cost in a max-min sense. The CC scheme offers a framework that can potentially accommodate additional objectives of LLM watermarking, such as robustness against adversarial manipulations and embedding capacity. Additionally, we envision future work implementing the scheme for sequential watermarking and extending it to the positive-perception regime, where minor adjustments to token probabilities are permitted in exchange for superior detection.

Disclaimer. This paper was prepared by Hsiang Hsu prior to his employment at JPMorgan Chase & Co.. Therefore, this paper is not a product of the Research Department of JPMorgan Chase & Co. or its affiliates. Neither JPMorgan Chase & Co. nor any of its affiliates makes any explicit or implied representation or warranty and none of them accept any liability in connection with this paper, including, without limitation, with respect to the completeness, accuracy, or reliability of the information contained herein and the potential legal, compliance, tax, or accounting effects thereof. This document is not intended as investment research or investment advice, or as a recommendation, offer, or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction.

REFERENCES

Scott Aaronson. Watermarking of large language models. `https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17`, August 2023. Accessed: 2025-01-1-.

Shahab Asoodeh, Mario Diaz, and Flavio P Calmon. Contraction of e_γ-divergence and its applications to privacy. *arXiv preprint arXiv:2012.11035*, 2020.

Dara Bahri, John Wieting, Dana Alon, and Donald Metzler. A watermark for black-box language models. *arXiv preprint arXiv:2410.02099*, 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pp. 675–685. PMLR, 2019.

Patrick Chao, Edgar Dobriban, and Hamed Hassani. Watermarking language models with error correcting codes. *arXiv preprint arXiv:2406.10281*, 2024.

Brian Chen. *Design and analysis of digital watermarking, information embedding, and data hiding systems*. PhD thesis, Massachusetts Institute of Technology, 2000.

Jun Chen, Lei Yu, Jia Wang, Wuxian Shi, Yiqun Ge, and Wen Tong. On the rate-distortion-perception function. *IEEE Journal on Selected Areas in Information Theory*, 3(4):664–673, 2022.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1125–1139. PMLR, 2024.

Thomas M Cover and A Joy Thomas. *Elements of Information Theory*. Wiley, New-York, 2nd edition, 2006.

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.

Persi Diaconis and David Freedman. Finite exchangeable sequences. *The Annals of Probability*, pp. 745–764, 1980.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. Three bricks to consolidate watermarks for large language models. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. IEEE, 2023.

Israel Gel'Fand and Mark Pinsker. Coding for channels with random parameters. *Probl. Contr. Inform. Theory*, 9(1):19–31, 1980.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024. URL `https://www.gurobi.com`.

Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. Universally optimal watermarking schemes for llms: from theory to practice. *arXiv preprint arXiv:2410.02890*, 2024.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. TrustLLM: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

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

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

Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*, volume 3. Springer, 1986.

Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *arXiv preprint arXiv:2404.01245*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022.

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, 2024.

Jingbo Liu, Paul Cuff, and Sergio Verdú. e_γ-resolvability. *IEEE Transactions on Information Theory*, 63(5):2629–2658, 2016.

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

Emin Martinian, Gregory W Wornell, and Brian Chen. Authentication with distortion criteria. *IEEE Transactions on Information Theory*, 51(7):2523–2542, 2005.

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

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, et al. Markllm: An open-source toolkit for llm watermarking. *arXiv preprint arXiv:2405.10051*, 2024.

Julien Piet, Chawin Sitawarin, Vivian Fang, Norman Mu, and David Wagner. Mark my words: Analyzing and evaluating language model watermarks. *arXiv preprint arXiv:2312.00273*, 2023.

Yury Polyanskiy. *Channel coding: Non-asymptotic fundamental limits*. Princeton University, 2010.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2022.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2024.

Yury Polyanskiy, H Vincent Poor, and Sergio Verdú. Channel coding rate in the finite blocklength regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359, 2010.

Jielin Qiu, William Han, Xuandong Zhao, Shangbang Long, Christos Faloutsos, and Lei Li. Evaluating durability: Benchmark insights into image and text watermarking. *Journal of Data-centric Machine Learning Research*, 2024.

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

Yubing Ren, Ping Guo, Yanan Cao, and Wei Ma. Subtle signatures, strong shields: Advancing robust and imperceptible watermarking in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 5508–5519, 2024.

Lucas Theis and Aaron B Wagner. A coding theorem for the rate-distortion-perception function. *arXiv preprint arXiv:2104.13662*, 2021.

Shangqing Tu, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. Waterbench: Towards holistic evaluation of watermarks for large language models. *arXiv preprint arXiv:2311.07138*, 2023.

Renato Villán, Sviatoslav Voloshynovskiy, Oleksiy Koval, J Vila, Emre Topak, Frédéric Deguillaume, Yuri Rytsar, and Thierry Pun. Text data-hiding for digital and printed documents: Theoretical and practical considerations. In *Security, Steganography, and Watermarking of Multimedia Contents VIII*, volume 6072, pp. 406–416. SPIE, 2006.

Frans MJ Willems. An informationtheoretical approach to information embedding. In *2000 Symposium on Information Theory in the Benelux, SITB 2000*, pp. 255–260. Werkgemeenschap voor Informatie-en Communicatietheorie (WIC), 2000.

Yangxinyu Xie, Xiang Li, Tanwi Mallick, Weijie J Su, and Ruixun Zhang. Debiasing watermarks for large language models via maximal coupling. *arXiv preprint arXiv:2411.11203*, 2024.

Xi Yang, Kejiang Chen, Weiming Zhang, Chang Liu, Yuang Qi, Jie Zhang, Han Fang, and Nenghai Yu. Watermarking text generated by black-box language models. *arXiv preprint arXiv:2305.08883*, 2023.

Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for ai-generated text. In *The Twelfth International Conference on Learning Representations*, 2024a.

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

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

A PROOFS OF THEORETICAL RESULTS

In this appendix, we include comprehensive overview of related works, as well as detailed proofs of our theoretical results, which are presented in the main body of the paper.

A.1 RELATED WORKS

Given the extensive volume of work in LLM watermarking, we focus our discussion on works that inform and contrast with our main contribution: theoretical frameworks for analyzing the limits of LLM watermarking.

Classical Information-Theoretic Approaches. Post-process watermarking, where watermarks are embedded after content generation, has been extensively studied through information-theoretic lenses Chen (2000); Moulin & O'Sullivan (2003); Martinian et al. (2005), particularly through the Gelfand-Pinsker (GP) channel Gel'Fand & Pinsker (1980); Villán et al. (2006); Willems (2000), which treats the LLM token $X \sim Q_X$ as the channel state for constructing the watermarked token. The GP scheme constructs auxiliary random variables $U \sim P(U|X)$ and encodes the watermarked token as $A = f(U, X)$. These approaches differ from our approach in two key aspects: (1) they typically require long sequences for joint typicality to hold, which leads to schemes that are intractable in the online setting with a large token vocabulary, while we focus on optimizing the one-shot minimax setting motivated by auto-regressive generation; and (2) they generally assume perfect knowledge of the underlying distributions, whereas our scheme is designed to work with the assumption that the underlying distribution is unknown, only the sampled token and side information are available.

Modern LLM Watermarking. Kirchenbauer et al. Kirchenbauer et al. (2023) introduced the first watermarking scheme for LLMs, which divides the vocabulary into green and red lists and slightly enhances the probability of green tokens in the next token prediction (NTP) distribution. This seminal work sparked numerous developments Aaronson (2023); He et al. (2024); Bahri et al. (2024); Dathathri et al. (2024); Yang et al. (2023); Ren et al. (2024); Hu et al. (2024); Zhao et al. (2024c); Chao et al. (2024); Qu et al. (2024); Xie et al. (2024); Liu & Bu (2024), with several approaches focusing on distortion-free methods that maintain the original NTP distribution unchanged, e.g., Kuditipudi et al. (2023); Hu et al. (2024); Zhao et al. (2024c); Christ et al. (2024). Unlike these methods which primarily focus on implementation strategies, our work provides a theoretical framework that characterizes optimal detection-perception trade-offs. Most related to our approach, Chao et al. Chao et al. (2024) propose a watermark using optimal correlated channels, though our work differs by providing a complete characterization through joint optimization of the randomization distribution in the one-shot setting.

Theoretical Analysis of LLM Watermarking. Recent work has advanced our theoretical understanding of LLM watermarking limitations. Huang et al. Huang et al. (2023) designed an optimal watermarking scheme for a specific detector, but their approach requires knowledge of the original NTP distributions of the watermarked LLM, making it model-dependent. Li et al. Li et al. (2024) proposed detection rules using pivotal statistics, though their Type II error control relies on asymptotic techniques from large deviation theory and focuses on large-sample statistics, whereas our analysis addresses the fundamental one-shot case including explicit characterization of corner point cases and the development of an optimal correlated channel scheme. Most recently, He et al. He et al. (2024) characterizes the universal Type II error while controlling the worst-case Type-I error by optimizing the watermarking scheme and detector. In contrast to these approaches, we analyze optimal mean detection by formulating a minimax framework while balancing Type I and Type II errors through the use of an E_γ-information objective. In the minimax formulation, we provide the optimal mean detection in closed form and characterize the optimal distribution of randomness under adversarial token distributions.

The development of the field is tracked through comprehensive benchmarks Piet et al. (2023); Tu et al. (2023); Pan et al. (2024); Qiu et al. (2024) and surveys Zhao et al. (2024b); Liu et al. (2024).

A.2 PROOF FOR PROPOSITION 1

Fixed $(P_S, Q_X, \tilde{Q}_{X|S})$ and priors (π_0, π_1).

Eve's hypothesis testing problem can be formulated as distinguishing between $H_0 : A \sim Q_X$ and $H_1 : A \sim \tilde{Q}_X$. By the Neyman-Pearson Lemma, the optimal test statistic is given by the likelihood ratio $L(a) = Q_X(a)/\tilde{Q}_X(a)$. The optimal decision rule takes the form $\delta(a) = \mathbb{1}\{L(a) > \eta\}$ for some threshold η. The probability of correct detection for Eve can be expressed as:

$$\Pr(\hat{H}_E = C) = \frac{1}{2}\Pr(\delta(A) = 1|H_1) + \frac{1}{2}\Pr(\delta(A) = 0|H_0)$$

For the optimal threshold $\eta = 1$, this probability becomes:

$$\begin{aligned}\Pr(\hat{H}_E = C) &= \frac{1}{2} + \frac{1}{2}\sum_{a\in\mathcal{X}}|\tilde{Q}_X(a) - Q_X(a)| \\ &= \frac{1}{2} + \frac{1}{2}\text{TV}(\tilde{Q}_X, Q_X)\end{aligned}$$

Now, we turn to Bob's detection probability. Bob's hypothesis testing problem differs from Eve's due to his access to the side information S. His testing problem can be formulated as distinguishing between $H_0 : (A, S) \sim Q_{X|S} \times P_S$ and $H_1 : (A, S) \sim \tilde{Q}_{X|S} \times P_S$.

By the Neyman-Pearson Lemma, the optimal test statistic in this case is $L(a, s) = Q_{X|S}(a|s)/\tilde{Q}_{X|S}(a|s)$. Given priors (π_0, π_1) and let $\gamma = \frac{\pi_1}{\pi_0}$, the conditional probability of correct detection given $S = s$ is:

$$\Pr(\hat{H}_B = C|S = s) = \pi_0 \Pr(\delta(A) = 0|H_0) + \pi_1 \Pr(\delta(A) = 1|H_1) \tag{15}$$

$$= \pi_0 Q_{X|S}[L(a, s) \geq \gamma] + \pi_1 \tilde{Q}_{X|S}[L(a, s) \leq \gamma] \tag{16}$$

$$= \pi_1 + \pi_0 Q_{X|S}[L(a, s) \geq \gamma] - \pi_1 \tilde{Q}_{X|S}[L(a, s) \geq \gamma] \tag{17}$$

$$= \pi_1 + \pi_0 \left[Q_{X|S}[L(a, s) \geq \gamma] - \frac{\pi_1}{\pi_0}\tilde{Q}_{X|S}[L(a, s) \geq \gamma\right] \tag{18}$$

$$= \pi_1 + \pi_0 E_\gamma(Q_{X|S}||\tilde{Q}_{X|S}). \tag{19}$$

The last equality comes from the alternative formula for E_γ where $E_\gamma(P||Q) = \max_{\mathcal{A}}[P(\mathcal{A}) - \gamma Q(\mathcal{A})]$, and supremum is attained with $A = \{a|L(a, s) \geq \gamma\}$.

A.3 Proof of Theorem 1

By the assumption of a uniform prior, we are looking for bounds on the quantity $\frac{1}{2}(1 + E_\gamma(\tilde{Q}_{X|S}\|Q_X|P_S))$, which boils down to bounding $E_\gamma(\tilde{Q}_{X|S}\|Q_X|P_S) = \mathbb{E}_S\left[E_\gamma(\tilde{Q}_{X|S}\|Q_X)\right]$. First, note that under a uniform prior, this quantity is lower bounded by the performance of a random guess, i.e., $\frac{1}{2} \leq R_d$. In what follows, we develop an upper for $E_\gamma(\tilde{Q}_{X|S}\|Q_X|P_S)$. For simplicity, denote $|\mathcal{X}| = d$ and $|\mathcal{S}| = m$. Let $Q_{X|S=s_i} = p_i$ such that $p_1, ..., p_m \in \Delta_d$, where Δ_d denotes the d-dimensional simplex. Assume that $S \sim \mathsf{Unif}[m]$. Following the zero perception assumption, we have $\tilde{Q}_X = Q_X$, i.e., $\frac{1}{m}\sum_{i=1}^m p_i = Q_X$. Consequently, our TV-optimization, when jointly optimized also over the marginal distribution Q_X is of the form:

$$\max_{p_1,...,p_m\in\Delta_d} \frac{1}{m}\sum_{i=1}^m \left\| p_i - \frac{\gamma}{m}\sum_{i=1}^m p_i \right\|_+, \tag{20}$$

where $\|x\|_+ \triangleq \sum_i (x_i)_+$ for $d \geq m$. We are maximizing a convex function over a polytope, so the optimal solution lies on the extreme points. Thus $p_i = e_j$ for some $j \leq d$, where e_j is the indicator vector with j-th entry equal to one. The problem boils down to determining how many times each vector e_j shows up.

Denote with q the probability vector corresponding to the distribution Q_X. We note that q can be rewritten as

$$q \triangleq \frac{1}{m}\sum_{i=1}^m p_i = \frac{1}{m}\sum_{j=1}^d n_j e_j, \tag{21}$$

where $\sum_j n_j = m$ and $n_j \in \mathbb{N}$. Denote the j-th entry of q by q_j. We have $\|e_j - q\|_+ = (1-q_j)_+ = 1 - q_j$. Therefore:

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\|p_i - \gamma q\|_+ &\overset{a}{=} \frac{1}{m}\sum_{j=1}^{d} n_j \|e_j - \gamma q\|_+ \\ &= \frac{1}{m}\sum_{j=1}^{d} n_j (1-\gamma q_j)_+ \\ &\overset{b}{=} \sum_{j=1}^{d} q_j (1-\gamma q_j)_+\end{aligned}$$

where (a) follows from from rewriting the sum in terms of e_j and (b) follows from the relation $q_j = \frac{n_j}{m}$, as can be seen from (21) and by the definition of the indicator. Out optimization problem had therefore boiled down to maximizing on the quantity

$$\sum_{j=1}^{d} q_j (1-\gamma q_j)_+ \text{ such that } q_j = k/m, k \in \mathbb{Z}, \sum_{j=1}^{d} q_j = 1. \tag{22}$$

To solve (22), we will examine various settings of the value of γ.

$\gamma \leq 1$ First, note that when $\gamma = 0$ the objective sums up to 1 by the constraints. Otherwise, note that whenever $\gamma \leq 1$, we have $(1-\gamma q_j)_+ = 1 - \gamma q_j$. Thus, we have

$$\sum_{j=1}^{d} q_j (1-\gamma q_j)_+ = 1 - \gamma \sum_{j=1}^{n} q_j^2.$$

Thus, maximization of the objective, boils down to the minimization of the sum of squares. We note that as q is a probability vectors, the sum of square minimizes under the uniform distribution, with the minimum being $\frac{1}{m}$. Thus, we have the upper bound

$$\frac{1}{2}(1 + E_\gamma(\tilde{Q}_{X|S}\|Q_X|P_S)) \leq \frac{1}{2}\left(1 + 1 - \frac{\gamma}{m}\right) = 1 - \frac{\gamma}{2m}.$$

$\gamma > 1$ In this case, we are not guaranteed with the positivity of $(1-\gamma q_j)$. We will look for a strategy to choose the values of $(q_j)_j$ such that the considered sum is maximized, while not passing the threshold that nullifies the terms $(1-\gamma q_j)$. For each j, denote each summand as $f(q_j)$, whose value is

$$f(q_j) = \begin{cases} q_j - \gamma q_j^2, & q_j \leq \frac{1}{\gamma} \\ 0, & \text{else.} \end{cases}$$

Consequently, as q_j is constrained to the set $(\frac{k}{m})_{k=0}^{m}$, whenever $\gamma \geq m$, no positive value of q_j will result in a positive value of $f(q_j)$. Thus, the resulting sum is 0, which implies that $R_d = \frac{1}{2}$. Thus we will focus on $\gamma \in (1, m)$. In this case, there is at least one possible value for each q_j that results in a nonnegative value of $f(q_j)$. First, we note that the mapping $x \mapsto x - \gamma x^2$ is a concave function of x for $\gamma > 0$, whose maximum is obtained in $x^\star = \frac{1}{2\gamma}$. Therefore, we would like to set $q_j = \frac{1}{2\gamma}$ as this will maximize a single summand. However, in most cases $\frac{1}{2\gamma} \notin (\frac{k}{m})_{k=1}^{m}$. To that end, we will set the closes possible value to $\frac{1}{2\gamma}$ within the allowed set. Second, we we would like to set as many q_j's to the value $\frac{1}{2\gamma}$ while following the constraint $\sum_{j=1}^{d} q_j = 1$, we will choose the lower value. To summarize, for each interval $\frac{k}{m} \leq \frac{1}{2\gamma} \leq \frac{k+1}{m}$, we will set $q_j = \frac{k}{m}$. The maximal amount of such q_j we can set while following the sum constraint is $\lfloor \frac{m}{k} \rfloor$. Thus, we have the following

$$\begin{aligned}E_\gamma(\tilde{Q}_{X|S}\|Q_X|P_S) &= \left\lfloor \frac{m}{k} \right\rfloor \left(\frac{k}{m} - \gamma\left(\frac{k}{m}\right)^2\right) \\ &\leq 1 - \frac{\gamma k}{m}.\end{aligned}$$

The corresponding bound on R_d is $1 - \frac{\gamma k}{2m}$. The bound is achievable whenever m is divisible by k within the resulting interval. Note that the interval $\frac{k}{m} \leq \frac{1}{2\gamma} \leq \frac{k+1}{m}$ corresponds to the interval $\frac{m}{2(k+1)} \leq \gamma \leq \frac{m}{2k}$. However, we already know the resulting bounds for $\gamma \geq m$ and $\gamma \leq 1$. Thus, the relevant values of k that correspond to this case are $k \in [1 : \frac{m}{2}]$. Finally, when $\frac{1}{2m} < \frac{1}{2\gamma} < \frac{1}{m}$ we cannot take the lower value $(k = 0)$, and will therefore take higher value $k = 1$. However, note that $\frac{1}{2m} < \frac{1}{2\gamma}$ corresponds to $\gamma > m$. Thus, this sub-case ($\frac{1}{2m} < \frac{1}{2\gamma} \leq \frac{1}{m}$) boils down to $\gamma < \frac{m}{2}$ with corresponding upper bound of $1 - \frac{\gamma}{m}$, which will merge with the interval $\gamma \leq 1$. This concludes the proof □

A.4 Proof of Theorem 2

Let $Q_i \triangleq Q_{X|S=s_i}$ The proof follows from analyzing the following steps:

$$\begin{aligned}
\sup_{\tilde{Q}_{X|S}} \sum_{s\in\mathcal{S}} P_S(s) E_\gamma(\tilde{Q}_{X|S=s}, Q_X) &= \sup_{\tilde{Q}_{X|S}} \frac{1}{2|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|Q_i - \gamma Q_x\|_1 \\
&= \frac{1}{2|\mathcal{S}|} \sup_{f:\mathcal{S}\to\mathcal{X}} \sum_{i=1}^{|\mathcal{S}|} \|Q_{f(i)} - \gamma Q_x\|_1 \\
&\leq \frac{1}{2} \sup_{i\in\mathcal{X}} \|Q_i - \gamma Q_x\|_1 \\
&= \sup_{i\in\mathcal{X}} |1 - \gamma Q_x(i)| \\
&= 1 - \gamma Q_{\mathsf{min}}
\end{aligned}$$

Therefore,

$$R_d \leq \frac{1}{2}\left(1 + 1 - \gamma Q_{\mathsf{min}}\right) = 1 - \frac{\gamma Q_{\mathsf{min}}}{2}$$

For the second equality, note that argmax of a convex function lies in the corner of the probability simplex. □

A.5 Proof of Correlated Channel (CC) with Perfect Perception

We prove that CC is a perfect perception scheme, i.e. $\mathbb{E}_S\left[\tilde{Q}_{X|S}\right](x) = Q_X(x)$. Recall that $S = (Y, B^m)$.We have the following

$$\begin{aligned}
\mathbb{E}_S\left[\tilde{Q}_{X|S}\right](x) &= \sum_{y,b^m} \mu_{B^m}(b^m) P_Y(y) Q_X(x) \frac{P_{Y|\tilde{Y}}(y|f(x,b^m))}{P_Y(y)} \\
&= Q_X(x) \sum_{y,b^m} \mu_{B^m}(b^m) P_{Y|\tilde{Y}}(y|f(x,b^m)).
\end{aligned}$$

Denote by $\mathcal{B}_1(x) \triangleq \{b^m : f(x,b^m) = 1\}$ and denote $\mathcal{B}_0(x)$ by the same token. We have

$$\begin{aligned}
&\mathbb{E}_S\left[\tilde{Q}_{X|S}\right](x) \\
&= Q_X(x) \left(\sum_{b^m\in\mathcal{B}_1(x)} \mu_{B^m}(b^m) \underbrace{\sum_{y=0,1} (b^m) P_{Y|\tilde{Y}}(y|1)}_{=1} + \sum_{b^m\in\mathcal{B}_0(x)} \mu_{B^m} \underbrace{\sum_{y=0,1} \mu_{B^m}(b^m) P_{Y|\tilde{Y}}(y|0)}_{=1} \right) \\
&= Q_X(x).
\end{aligned}$$

This concludes the proof. □

A.6 PROOF OF PROPOSITION 2

By the dual representation of the total variation

$$\mathsf{TV}(P,Q) = \min_{P_{XY}} \{\mathbb{P}[X \neq Y] : P_X = P, P_Y = Q\}, \tag{23}$$

Given $S \sim \mathsf{Unif}[k]$ and $P_{\tilde{Y}} = \{p_1, ..., p_k\} \in \Delta_k$. We have $\mathsf{TV}(P_S, P_{\tilde{Y}}) = 1 - \sum_{i=1}^{k} \min(\frac{1}{k}, p_i)$.

We propose a coupling and shows that it achieves $\mathsf{TV}(P_S, P_{\tilde{Y}})$.

To simplify notation, let the distribution of S and $\tilde{Y}$ be P and Q. Let $t = \mathsf{TV}(P,Q)$. Assume that $0 < t < 1$. Define three probability distributions $R = \frac{P \wedge Q}{1-t}$, $P' = \frac{P - P \wedge Q}{t}$ and $Q' = \frac{Q - P \wedge Q}{t}$. Construct P_{XY} as follows:

1. Generate $B \sim \text{Bernoulli}(t)$.
2. If $B = 0$, draw $Z \sim R$ and set $S = \tilde{Y} = Z$.
3. If $B = 1$, draw $S \sim P'$ and $\tilde{Y} \sim Q'$ independently.

To show that this is a valid coupling, we verify the marginal distribution is kept the same. We have:

$$\begin{aligned} P_S(a) &= \mathbb{P}(B=0)R(a) + \mathbb{P}(B=1)P'(a) \\ &= (1-t)\left(\frac{P \wedge Q}{1-t}\right)(a) + t\left(\frac{P - P \wedge Q}{t}\right)(a) \\ &= P(a) \end{aligned}$$

Similarly,

$$\begin{aligned} P_{\tilde{Y}}(a) &= \mathbb{P}(B=0)R(a) + \mathbb{P}(B=1)Q'(a) \\ &= (1-t)\left(\frac{P \wedge Q}{1-t}\right)(a) + t\left(\frac{Q - P \wedge Q}{t}\right)(a) \\ &= Q(a) \end{aligned}$$

Therefore $P_{S\tilde{Y}}$ is a valid coupling.

Lastly, we show that for the specific coupling, $\mathbf{P}(\tilde{Y} \neq S) = \mathsf{TV}(P_S, P_{\tilde{Y}})$

$$\begin{aligned} \mathbf{P}(\tilde{Y} \neq S) &= 1 - \mathbf{P}(\tilde{Y} = S) \\ &= 1 - (1-t) \\ &= t \\ &= \mathsf{TV}(P_S, P_{\tilde{Y}}) \end{aligned}$$

Thus, we have constructed a coupling $P_{S\tilde{Y}}$ that minimizes $\mathbf{P}(\tilde{Y} \neq S)$, which means that it maximizes $\mathbf{P}(\tilde{Y} = S)$. □

A.7 PROOF OF REMARK 2

The hypothesis test is the following: $H_0 : X \sim Q_X$ and $H_1 : X \sim \tilde{Q}_{X|S,B^m}$, where $\tilde{Q}_{X|S,B^m}$ is the CC-watermark distribution shown in equation (8), and side information $S \sim \mathsf{Ber}(0.5)$. We show H_0 is rejected by the CC detection test $S = f(X, B^m)$ if and only if it is also rejected by the likelihood ratio test (LRT).

If H_0 is rejected by CC detection test, then $S = f(X, B^m)$. Then, consider the likelihood ratio:

$$\frac{Q_X(X)}{\tilde{Q}_{X|B^m,S}(X)} = \frac{Q(X)}{Q_X(X)\frac{1}{P_S(S)}P_{S|\tilde{Y}}(S|f(X,B^m)} \tag{24}$$

$$= \frac{2}{P_{S|\tilde{Y}}(S|f(X,B^m)} \tag{25}$$

$$< 1, \tag{26}$$

The density of $\tilde{Q}_{X|B^m,S}(X)$ follows from the CC-watermark, side information $P_S(S) = 0.5$. The last inequality come from the Z-S channel construction: $\Pr_{S|\tilde{Y}}(S|f(S, B^m) \geq \frac{1}{2}$, if and only if $S = f(X, B^m)$. Since the likelihood ratio is less than 1, H_0 is rejected by the LRT.

If H_0 is rejected by the LRT with threshold 1, then we have

$$\frac{Q_X(X)}{\tilde{Q}_{X|B^m,S}(X)} < 1.$$

Expanding the likelihood ratio as above, this implies: $P_{S|\tilde{Y}}(S|f(X, B^m) < \frac{1}{2}$. By construction of the Z-S channel, $S = f(X, B^m)$. Hence, H_0 is rejected by CC detection test.

A.8 Proof of Proposition 3

We start by proving the following identity:

$$\mathsf{TV}\left(Q_X, \tilde{Q}_{X|(S,B^m)}|P_{S,B^m}\right) = \mathsf{TV}\left(P_S, P_{S|\tilde{Y}}|P_{\tilde{Y}}\right)$$

Proof: Recall that in the correlated channel watermark we have side information S and partition bits B^m. By definition, we have

$$\mathsf{TV}(Q_X, \tilde{Q}_{X|S,B^m}|P_{S,B^m}) = \sum_{b^m} \sum_{s=0,1} \mu(b^m) P_S(s) \mathsf{TV}(Q_X, \tilde{Q}_{X|b^m,s}). \tag{27}$$

Next, we simplify the TV expression within the sum. For any (b^m, s) we have

$$\begin{aligned}
\mathsf{TV}(Q_X, \tilde{Q}_{X|(b^m,s)}) &= \sum_x \left| Q_X(x) - Q_X(x) \frac{P_{S|\tilde{Y}}(s|f(x, b^m))}{P_S(s)} \right| \\
&= 2 \sum_x Q_X(x) \left| \frac{1}{2} - p_{S|\tilde{Y}}(s|\tilde{y}) \right|,
\end{aligned}$$

where recall that $\tilde{Y} = f(X, B^m)$, $p_{S|\tilde{Y}}(s|\tilde{y})$ is the corresponding coupling channel parameter, and $S \sim \mathsf{Ber}(\frac{1}{2})$. We define the pre-image of f for a fixed b^m as $f^{-1}(\cdot, b^m) : \{0, 1\} \to 2^{\mathcal{X}}$, with $f^{-1}(0), f^{-1}(1) \subseteq \mathcal{X}$. Plugging the simplified TV expression back into (27), we have

$$\begin{aligned}
&\mathsf{TV}(Q_X, \tilde{Q}_{X|(b^m,s)}) \\
&= \sum_{b^m} \mu(b^m) \sum_{s=0,1} \sum_x Q_X(x) \left| \frac{1}{2} - p_{S|\tilde{Y}}(s|\tilde{y}) \right| \\
&= \sum_{b^m} \mu(b^m) \sum_{s=0,1} \left(\sum_{x \in f^{-1}(0,b^m)} Q_X(x) \left| \frac{1}{2} - p_{S|\tilde{Y}}(s|0) \right| + \sum_{x \in f^{-1}(1,b^m)} Q_X(x) \left| \frac{1}{2} - p_{S|\tilde{Y}}(s|1) \right| \right) \\
&= \sum_{b^m} \mu(b^m) \left(P_{\tilde{Y}}(0) \sum_{s=0,1} \left| \frac{1}{2} - p_{S|\tilde{Y}}(y|0) \right| + P_{\tilde{Y}}(1) \sum_{s=0,1} \left| \frac{1}{2} - p_{S|\tilde{Y}}(s|1) \right| \right) \\
&= \mathsf{TV}\left(P_S, P_{S|\tilde{Y}}|P_{\tilde{Y}}\right),
\end{aligned}$$

where the randomness of $\tilde{Y}$ is determined by the pair (Q_X, μ). This concludes the proof. □

With this, we proceed to showing CC's detection rate. By Theorem 2, CC's detection rate is equal to that of likelihood ratio test. By Proposition 1 and under equal priors on TPR and TNR, we have

$$\begin{aligned}
R_d &= \frac{1}{2}(1 + \mathsf{TV}(Q_X, \tilde{Q}_{X|S,B^m}|P_{S,B^m})) &(28)\\
&= \frac{1}{2}\left(1 + \mathsf{TV}(P_S, P_{S|\tilde{Y}}|P_{\tilde{Y}})\right), &(29)
\end{aligned}$$

where the last equality is due to the identity above.

Next, we obtain a closed form for $\mathsf{TV}(P_S, P_{S|\tilde{Y}}|P_{\tilde{Y}})$. By definition, we have

$$\mathsf{TV}\left(P_S, P_{S|\tilde{Y}}|P_{\tilde{Y}}\right) = \tilde{p}_0 \mathsf{TV}\left(P_S, P_{S|\tilde{Y}=0}\right) + \tilde{p}_1 \mathsf{TV}\left(P_S, P_{S|\tilde{Y}=1}\right).$$

Following Proposition 2, the nature of the TV terms depends on wether $\tilde{p}_1 \leq \frac{1}{2}$ or $\tilde{p}_0 \leq \frac{1}{2}$. For $\tilde{p}_0 \leq \frac{1}{2}$, the optimal coupling is given by a Z-channel, whose parameter is $\frac{2\tilde{p}_1 - 1}{2\tilde{p}_1}$. The TV terms are therefore given by

$$\mathsf{TV}\left(P_S, P_{S|\tilde{Y}=0}\right) = \frac{1}{2}\left|\frac{1}{2} - 1\right| + \frac{1}{2}\left|\frac{1}{2}\right| = \frac{1}{2}$$

$$\begin{aligned}
\mathsf{TV}\left(P_S, P_{S|\tilde{Y}=1}\right) &= \frac{1}{2}\left(\left|\frac{1}{2} - \frac{2\tilde{p}_1 - 1}{2\tilde{p}_1}\right| + \left|\frac{1}{2} - \frac{1}{2\tilde{p}_1}\right|\right) \\
&= \frac{1}{2}\left(\left|\frac{1 - \tilde{p}_1}{2\tilde{p}_1}\right| + \left|\frac{\tilde{p}_1 - 1}{2\tilde{p}_1}\right|\right) \\
&= \frac{\tilde{p}_0}{2\tilde{p}_1}.
\end{aligned}$$

Thus, we have

$$\mathsf{TV}\left(P_S, P_{S|\tilde{Y}}|P_{\tilde{Y}}\right) = \tilde{p}_0.$$

By the symmetry of the optimal coupling, for $\tilde{p}_1 \leq \frac{1}{2}$ we have

$$\mathsf{TV}\left(P_S, P_{S|\tilde{Y}}|P_{\tilde{Y}}\right) = \tilde{p}_1.$$

Hence, CC's detection rate is given by $R_d = \frac{1}{2}(1 + \min(\tilde{p_0}, \tilde{p_1}))$. □

A.9 Proof of Theorem 3

We begin by proving Lemma 1.

A.9.1 Proof of Lemma 1

Let $\mathcal{S} = [k]$ and $\mathcal{X} = [m]$. For a given $Q_X = \mathbf{q} = (q_1, \ldots, q_m) \in \Delta_m$ and an m-length sequence $\mathbf{b} = (b_1, \ldots, b_m) \in \mathcal{S}^m$, we define the function $f : \mathcal{X} \times \mathcal{S}^m \to \mathcal{S}$ as

$$f(i, \mathbf{b}) = b_i. \tag{30}$$

A sequence $\mathbf{b}$ induces a probability distribution $\hat{P}(\mathbf{q}, \mathbf{b})$ over $\mathcal{S}$ denoted as (with a slight abuse of notation)

$$\hat{P}(s, \mathbf{q}, \mathbf{b}) = \sum_{i=1}^{m} q_i \mathbf{1}\left[b_i = s\right] \ \forall s \in [k]. \tag{31}$$

For a fixed $\mathbf{b}$ and $\mathbf{q}$ and assuming that Alice uses the optimal coupling, Bob's probability of detection is given by the quantity

$$R_d(\mathbf{q}, \mathbf{b}) \triangleq 1 - \frac{1}{2}\mathsf{TV}\left(Q_S \| \hat{P}(\mathbf{q}, \mathbf{b})\right) - \frac{1}{2k}\sum_{s=1}^{k} \hat{P}(s, \mathbf{q}, \mathbf{b}) \tag{32}$$

$$= 1 - \frac{1}{2k} - \frac{1}{4} g(\mathbf{q}, \mathbf{b}), \tag{33}$$

where

$$g(\mathbf{q}, \mathbf{b}) \triangleq \sum_{s=1}^{k} \left|\hat{P}(s, \mathbf{q}, \mathbf{b}) - \frac{1}{k}\right| \tag{34}$$

where Q_S is the uniform distribution. Our goal is to find a distribution over $P^*_{B^m}$ that maximizes the worst-case value of R_d given a set of constraints on $\mathbf{q}$. Specifically, we analyze:

$$R^*_d(\lambda) \triangleq \max_{P_{B^m}} \min_{\substack{\mathbf{q}\in\Delta_m \\ \|\mathbf{q}\|_\infty \le \lambda}} \mathbb{E}\left[R_d(\mathbf{q}, B^m)\right] \tag{35}$$

$$= 1 - \frac{1}{2k} - \frac{1}{4} \min_{P_{B^m}} \max_{\substack{\mathbf{q}\in\Delta_m \\ \|\mathbf{q}\|_\infty \le \lambda}} \sum_{\mathbf{b}\in\mathcal{S}^m} P_{B^m}(\mathbf{b}) g(\mathbf{q}, \mathbf{b}). \tag{36}$$

The function

$$H(P_{B^m}) \triangleq \max_{\substack{\mathbf{q}\in\Delta_m \\ \|\mathbf{q}\|_\infty \le \lambda}} \mathbb{E}\left[g(\mathbf{q}, B^m)\right] \tag{37}$$

is convex in the distribution P_{B^m}, since it is the maximum of linear functions. Let $P^*_{B^m}$ be a distribution that minimized H and consider the permutation $\pi : \mathcal{S}^m \to \mathcal{S}^m$, define $\tilde{P}_\pi(\mathbf{b}) = P^*_{B^m}(\pi \circ \mathbf{b})$.

Since $\mathbb{E}_{P^*_{B^m}}\left[g(\mathbf{q}, B^m)\right] = \mathbb{E}_{\tilde{P}_\pi}\left[g(\pi \circ \mathbf{q}, B^m)\right]$ for all $\mathbf{q}$, $H(\tilde{P}_\pi) = H(P_{B^m})$ from the symmetry of the maximum. Hence, from the equality in (36) $F(\tilde{P}_\pi) = F(P_{B^m})$ for $F(P_{B^m}) \triangleq \min_{\substack{\mathbf{q}\in\Delta_m \\ \|\mathbf{q}\|_\infty \le \lambda}} \mathbb{E}_{P_{B^m}}\left[R_d(Q_X, B^m)\right]$. □

Next, we proceed with the proof of Theorem 3.

Let $C = m!$ be the number of permutations of an m-length sequence, we have

$$F\left(\frac{1}{C}\sum_{\pi} \tilde{P}_\pi\right) \le F(P^*_{B^m}). \tag{38}$$

Consequently, it is sufficient to restrict the minimization in P_{B^m} to distributions that assign equal probability mass to sequences that are identical up to a permutation.

Denote by $\mathcal{P}_m$ the partition of $\mathcal{S}^m$ into sets of sequences that are equal up to a permutation, with $|\mathcal{P}_m| = K$. For simplicity, we denote $\mathcal{P}_m = (\mathcal{B}_1, \ldots, \mathcal{B}_K)$ and refer to $\mathcal{B}_i$ as a *permutation class* (alternatively, we could have named it orbits or type classes). Then

$$\min_{P_{B^m}} F(P_{B^m}) = \min_{\mathbf{w}\in\Delta_K} \max_{\substack{\mathbf{q}\in\Delta_m \\ \|\mathbf{q}\|_\infty \le \lambda}} \sum_{i=1}^{K} \frac{w_i}{|\mathcal{B}_i|} \sum_{\mathbf{b}\in\mathcal{B}_i} g(\mathbf{q}, \mathbf{b}). \tag{39}$$

Observe that $g(\mathbf{q}, \mathbf{b})$ is convex in $\mathbf{q}$ (since it is the absolute value of a linear function in $\mathbf{q}$), and thus the inner maximum is achieved at a vertex of the feasible set. The vertices of the polytope $\{\mathbf{q} \in \Delta_m \mid \|\mathbf{q}\|_\infty \le \lambda\}$ are permutations of the vector

$$\mathbf{q}^*_\lambda = (\lambda, \ldots, \lambda, 1 - t\lambda, 0, \ldots, 0),$$

where $\mathbf{q}^*$ has (i) exactly t entries equal to λ and t is the largest integer such that $t\lambda \le 1$ (assuming $\lambda \le 1$), (ii) one entry equal to $1 - t\lambda$, and (iii) the remaining entries equal to 0.

Since the vertices are identical up to a permutation, and for any permutation π

$$\sum_{\mathbf{b}\in\mathcal{B}_i} g(\mathbf{q}, \mathbf{b}) = \sum_{\mathbf{b}\in\mathcal{B}_i} g(\pi \circ \mathbf{q}, \mathbf{b}), \tag{40}$$

it is sufficient to select a vertex of the form $\mathbf{q}^*_\lambda$. Thus,

$$\min_{P_{B^m}} F(P_{B^m}) = \min_{\mathbf{w}\in\Delta_K} \sum_{i=1}^{K} \frac{w_i}{|\mathcal{B}_i|} \sum_{\mathbf{b}\in\mathcal{B}_i} g(\mathbf{q}^*_\lambda, \mathbf{b}), \tag{41}$$

and it sufficient to consider the optimal distribution $P^*_{B^m}$ as a distribution that selects a $\mathbf{b}$ uniformly over a *single* permutation class in $\mathcal{P}_m$; namely the one that maximizes $\frac{1}{|\mathcal{B}_i|}\sum_{\mathbf{b}\in\mathcal{B}_i} g(\mathbf{q}^*_\lambda, \mathbf{b})$.

Next, we aim to characterize $R^*_d(\lambda)$ for different values of λ. We denote by $P_{\mathcal{B}}$ the distribution resulting from drawing a sequence at random from the permutation class $\mathcal{B} \in \mathcal{P}_m$.

Our goal is to compute

$$\mathbb{E}\left[g(\mathbf{q}_\lambda^*, B^m)\right] = \sum_{s=1}^{k} \mathbb{E}\left[\left|\hat{P}(s, \mathbf{q}_\lambda^*, B^m) - \frac{1}{k}\right|\right] \tag{42}$$

Recall that the optimal choice of P_{B^m} is to draw sequences uniformly from a single permutation class. Assuming w.l.o.g. that $\mathcal{S} = [k]$, fix a sequence $\mathbf{b} \in \mathcal{S}^m$ with n_i entries equal to i, $i \in [k]$. For example, if $k = 2$, n_1 is the number of entries equal to 1 and n_2 is the number of entries equal to 2. Naturally, $\sum_{i=1}^{k} n_i = m$.

Now, for a fixed $s \in \mathcal{S}$, we can write

$$P(s, \mathbf{q}_\lambda^*, B^m) = \lambda \sum_{i=1}^{t} X_i + (1 - t\lambda) X_{t+1}, \tag{43}$$

where $t = \lfloor 1/\lambda \rfloor$ and $X_i \triangleq \mathbf{1}\left(B_i = s\right)$. We can expand the expectation in the lhs of (42) as

$$\mathbb{E}\left[\left|\hat{P}(s, \mathbf{q}_\lambda^*, B^m) - \frac{1}{k}\right|\right] = \mathbb{E}\left[\mathbb{E}\left[\left|\hat{P}(s, \mathbf{q}_\lambda^*, B^m) - \frac{1}{k}\right| \middle| \sum_{i=1}^{t} X_i\right]\right] \tag{44}$$

$$= \sum_{c=0}^{t} \Pr\left(\sum_{i=1}^{t} X_i = c\right) \left(\Pr\left(X_{t+1} = 1 \middle| \sum_{i=1}^{t} X_i = c\right) \left|c\lambda + (1 - \lambda t) - \frac{1}{k}\right| \right. \tag{45}$$

$$\left. + \Pr\left(X_{t+1} = 0 \middle| \sum_{i=1}^{t} X_i = c\right) \left|c\lambda - \frac{1}{k}\right|\right). \tag{46}$$

For our sampling without replacement strategy, we have

$$\Pr\left(\sum_{i=1}^{t} X_i = c\right) = \frac{\binom{n_s}{c}\binom{m-n_s}{t-c}}{\binom{m}{t}},$$

$$\Pr\left(X_{t+1} = 1 \middle| \sum_{i=1}^{t} X_i = c\right) = \frac{n_s - c}{m - t}.$$

Plugging these expressions in, we have:

$$\mathbb{E}\left[g(\mathbf{q}_\lambda^*, B^m)\right] = \sum_{s=1}^{k} \sum_{c=0}^{t} \frac{\binom{n_s}{c}\binom{m-n_s}{t-c}}{\binom{m}{t}} \left(\left(\frac{n_s - c}{m - t}\right) \left|c\lambda + (1 - \lambda t) - \frac{1}{k}\right| + \left(1 - \frac{n_s - c}{m - t}\right) \left|c\lambda - \frac{1}{k}\right|\right) \tag{47}$$

When we have an equal number of elements of each kind in the permutation class and m is divisible by k, i.e., $n_1 = \cdots = n_k = m/k$, the expression simplifies to:

$$\mathbb{E}\left[g(\mathbf{q}_\lambda^*, B^m)\right] = k \sum_{c=0}^{t} \frac{\binom{m/k}{c}\binom{m-m/k}{t-c}}{\binom{m}{t}} \left(\left(\frac{(m/k) - c}{m - t}\right) \left|c\lambda + (1 - \lambda t) - \frac{1}{k}\right| + \left(1 - \frac{(m/k) - c}{m - t}\right) \left|c\lambda - \frac{1}{k}\right|\right) \tag{48}$$

We can simplify this even further in the special case that $\lambda = 1/k$. In this case, $t = k$, and we don't have to consider the special case of X_{t+1} – $\mathbf{q}_\lambda^*$ has k entries equal to λ. In this case, denoting

$Z_k = \sum_{i=1}^{k} X_i$ (46), simplifies to:

$$\mathbb{E}\left[\left|\hat{P}(s, \mathbf{q}_\lambda^*, B^m) - \frac{1}{k}\right|\right] = \frac{1}{k}\sum_{c=0}^{k} \Pr\left(Z_k = c\right)|c-1| \tag{49}$$

$$= \frac{1}{k}\left(\Pr\left(Z_k = 0\right) + \sum_{c=1}^{k} \Pr\left(Z_k = c\right)(c-1)\right) \tag{50}$$

$$= \frac{1}{k}\left(2\Pr\left(Z_k = 0\right) - 1 + \mathbb{E}[Z_k]\right) \tag{51}$$

$$= \frac{2}{k}\Pr\left(Z_k = 0\right) \tag{52}$$

$$= \frac{2}{k} \times \frac{\binom{(k-1)m/k}{k}}{\binom{m}{k}} \tag{53}$$

and, consequently, we arrive at the elegant expression

$$\mathbb{E}\left[g(\mathbf{q}_\lambda^*, B^m)\right] = 2 \times \frac{\binom{(k-1)m/k}{k}}{\binom{m}{k}}. \tag{54}$$

Hence, for any given m, k, λ, that satisfies $\lambda = \frac{1}{k}$ and m divisible by k, we have (following Eq. (36)):

$$R_d^*(\lambda) = 1 - \frac{1}{2k} - \frac{1}{4}\mathbb{E}\left[g(\mathbf{q}_\lambda^*, B^m)\right] \tag{55}$$

$$= 1 - \frac{1}{2k} - \frac{1}{2}\frac{\binom{(k-1)m/k}{k}}{\binom{m}{k}} \tag{56}$$

For $1/2 \leq \lambda < 1$, $\mathbf{q}_\lambda^*$ has two non-zero entries equal to λ and $1-\lambda$. Consequently, $\hat{P}(\mathbf{q}_\lambda^*, \mathbf{b})$ assigns probability 1 to one value of S if $b_1 = b_2$, otherwise assigns mass $1-\lambda$ and λ to two separate values of s. Thus for a fixed distribution $P_{\mathcal{B}}$

$$\mathbb{E}_{P_{\mathcal{B}}}\left[R_d(\mathbf{q}_\lambda^*, B^m)\right] = 1 - \frac{1}{2k} - \Pr(B_1 = B_2) \times \frac{k-1}{2k} - \frac{1}{4}\Pr(B_1 \neq B_2) \times \left(1 - \frac{2}{k} + \left|\lambda - \frac{1}{k}\right| + \left|1 - \lambda - \frac{1}{k}\right|\right). \tag{57}$$

We need to select the set $\mathcal{B}$ that maximizes $\Pr(B_1 \neq B_2)$. For m even and $k = 2$ (i.e., $\mathcal{S}$ binary), $\mathcal{B}$ is the permutation class of the sequence of equal number of each element, we have $\Pr(B_1 = B_2) = \frac{m-2}{2(m-1)}$, $\Pr(B_1 \neq B_2) = \frac{m}{2(m-1)}$, which simplifies $R_d(\lambda)^*$ to

$$R_d^*(\lambda) = \frac{3}{4} - \frac{m\lambda - 1}{4(m-1)} \text{ for } k = 2,\ \frac{1}{2} \leq \lambda \leq 1. \tag{58}$$

As $m \to \infty$, $R_d^\star(\lambda) \to \frac{3}{4} - \frac{m}{4}$.

Remark 3 *We make precise why in the case for $\frac{1}{2} \leq \lambda < 1$, $k = 2$ and m even, $\mathcal{B}^* = \{b^m :$ equal number of 0's and 1's$\}$. For $S = \{0, 1\}$, i.e. $k = 2$, permutation classes are characterized by the number of $1's$. Let α be the number of $1's$ in $\mathcal{B}$ and $m - \alpha$ be the number of 0's. From Eq* (57)*, we need to select the set $\mathcal{B}$ that maximizes $\Pr(B_1 \neq B_2)$:*

$$\alpha^* = \arg\max_{\alpha \in [m]} \Pr[B_1 \neq B_2] = \arg\max_{\alpha \in [m]} 2\frac{\alpha(m-\alpha)}{m(m-1)} = \frac{m}{2}. \tag{59}$$

Next, we consider the case for $\frac{1}{3} \leq \lambda < \frac{1}{2}$. $\mathbf{q}^*_\lambda$ has three non-zero entries: $\mathbf{q}^*_\lambda = (\lambda, \lambda, 1 - 2\lambda, 0, ..., 0)$. Consequently, there are 4 cases with the corresponding $\hat{P}(\mathbf{q}^*_\lambda, \mathbf{b})$ and $g(\mathbf{q}^*_\lambda, \mathbf{b})$:

$$a. B_1 = B_2 = B_3: \quad \hat{P} = [1, 0, ..., 0] \quad g(\mathbf{q}^*_\lambda, \mathbf{b}) = 2(1 - \frac{1}{k})$$

$$b. B_1 = B_2, B_3 :\neq B_1 \quad \hat{P} = [2\lambda, 1 - 2\lambda, 0..., 0] \quad g(\mathbf{q}^*_\lambda, \mathbf{b}) = (2\lambda - \frac{1}{k}) + |1 - 2\lambda - \frac{1}{k}| + \frac{1}{k}(k - 2)$$

$$c. B_1 \neq B_2, B_3 = (B_1 \vee B_2): \quad \hat{P} = [1 - \lambda, \lambda, 0..., 0] \quad g(\mathbf{q}^*_\lambda, \mathbf{b}) = |\lambda - \frac{1}{k}| + |1 - \lambda - \frac{1}{k}| + \frac{1}{k}(k - 2)$$

$$d. B_1 \neq B_2 \neq B_3: \quad \hat{P} = [\lambda, \lambda, 1 - 2\lambda, 0..., 0] \quad g(\mathbf{q}^*_\lambda, \mathbf{b}) = 2|\lambda - \frac{1}{k}| + |1 - 2\lambda - \frac{1}{k}| + \frac{1}{k}(k - 3)$$

Recall that to maximize $\mathbb{E}_{P_\mathcal{B}}[R_d(\mathbf{q}^*_\lambda, B^m)]$, we need to minimize $\mathbb{E}_{P_\mathcal{B}}[g(\mathbf{q}^*_\lambda, B^m)]$.

For k=2, case d is invalid and case c produces the minimum $g(\mathbf{q}^*_\lambda, \mathbf{b})$. Hence, we select the set $\mathcal{B}$ that maximizes $\Pr[B_1 \neq B_2, B_3 = (B_1 \vee B_2)]$, which is equivalent to maximizing $\Pr[B_1 \neq B_2]$. Following (59), $\mathcal{B}^* = \{b^m : \text{equal number of 0's and 1's}\}$. We have $\Pr[B_1 = B_2 = B_3] = \frac{m-4}{4(m-1)}$, $\Pr[B_1 = B_2, B_3 \neq B_1] = \frac{m}{4(m-1)}$ and $\Pr[B_1 \neq B_2, B_3 = (B_1 \vee B_2)] = \frac{m}{2(m-1)}$.

The resulting $R_d(\lambda)^*$ is:

$$R^*_d(\lambda) = \frac{3}{4} - \frac{m-2}{8(m-1)} \text{ for } k = 2, \ \frac{1}{3} \leq \lambda < \frac{1}{2}. \tag{60}$$

As $m \to \infty$, $R^*_d(\lambda) \to \frac{5}{8}$.

A.10 Proof of Theorem 4

Our results so far have been based on the discussion that it is sufficient to consider the optimal distribution $P^*_{B^m}$ as one that selects $\mathbf{b}$ uniformly over a single permutation class $\mathcal{B}^* \in \mathcal{P}_m$. Recall that $\mathbf{b}$ is a sequence of m elements each take a value in $\mathcal{S}$: $|\mathbf{b}| = m$ and $\mathcal{S} = k$. Recall as well that $\mathcal{B}$ can be characterized by the proportion of each element of S: for $i \in [k]$, denote the proportions as $[p_1, ..., p_k]$, where

$$p_s = \frac{\sum_{i=1}^m \mathbf{1}[\mathbf{b}_i == s]}{m} \quad \forall \mathbf{b} \in \mathcal{B}.$$

Hence, sampling an $\mathbf{b}$ uniformly over $\mathcal{B}^*$ can be equivalently defined as the following process: given m elements with predefined proportions $[p_1, ..., p_k]$, sample m times with replacement.

To generalize the analysis for other ranges of λ, k, and m, we consider an alternative process in which rather than fixing the proportions over m elements, we take $[p_1, ..., p_k]$ as probabilities. $\mathbf{b}$ amounts to m i.i.d samples from a categorical distribution: $\mathbf{b}_i \overset{i.i.d}{\sim} \text{CATEGORICAL}(p_1, ..., p_k)$. Recall that optimal B^* amounts to having an equal number for each element in $\mathcal{S}$. Hence, for all $i \in [k]$, $p^*_i = \frac{1}{k}$.

Furthermore, recall that the adversarial distribution for a given min-entropy constraint λ is: $\mathbf{q}^* = [\lambda, \lambda, ..., 1 - t\lambda, 0, ..., 0]$, where $t = \lfloor \frac{1}{\lambda} \rfloor$. For the purpose of characterizing $\mathbb{E}_{P_\mathcal{B}} g(\mathbf{q}^*, \mathbf{b})$, only the color of the first $t + 1$ draws matter, because the rest have 0 probabilities.

Let $X_i \triangleq \mathbf{1}(B_i = s)$, for a fixed $s \in \mathcal{S}$. $X_i \overset{i.i.d}{\sim} \text{BER}(\frac{1}{k})$. We can compute $\mathbb{E}_{P_\mathcal{B}} g(\mathbf{q}^*, \mathbf{b})$ in closed form. Following (42) and (46), for sampling with replacement, we have:

$$\mathbb{E}\left[g(\mathbf{q}_\lambda^*, B^m)\right] = \sum_{s=1}^{k} \mathbb{E}\left[\left|\hat{P}(s, \mathbf{q}_\lambda^*, B^m) - \frac{1}{k}\right|\right] \tag{61}$$

$$= \sum_{s=1}^{k}\sum_{c=0}^{t} \Pr\left(\sum_{i=1}^{t} X_i = c\right)\left(\Pr\left(X_{t+1}=1 \,\middle|\, \sum_{i=1}^{t} X_i = c\right)\left|c\lambda + (1-\lambda t) - \frac{1}{k}\right|\right. \tag{62}$$

$$\left. + \Pr\left(X_{t+1}=0 \,\middle|\, \sum_{i=1}^{t} X_i = c\right)\left|c\lambda - \frac{1}{k}\right|\right) \tag{63}$$

$$= k\sum_{c=0}^{t} \Pr[Y=c]\left(\frac{1}{k}\left|c\lambda + (1-\lambda t) - \frac{1}{k}\right| + (1-\frac{1}{k})\left|c\lambda - \frac{1}{k}\right|\right) \tag{64}$$

$$= \sum_{c=0}^{t} \Pr[Y=c]\left(\left|(c-t)\lambda + (1-\frac{1}{k})\right| + (k-1)\left|c\lambda - \frac{1}{k}\right|\right) \tag{65}$$

where $Y \sim Bin(t, \frac{1}{k})$, and hence $\Pr[Y=c] = \binom{t}{c}(\frac{1}{k})^c(1-\frac{1}{k})^{t-c}$

By Eq. 36, the approximated minimax detection is given by:

$$\tilde{R}_d^\star(\lambda) = 1 - \frac{1}{2k} - \frac{1}{4}\left[\sum_{c=0}^{t} \Pr[Y=c]\left(\left|(c-t)\lambda + (1-\frac{1}{k})\right| + (k-1)\left|c\lambda - \frac{1}{k}\right|\right)\right] \tag{66}$$

Finally, we analyze the approximation error of $\tilde{R}_d^\star(\lambda)$. Define $H_{\mathbf{b}}$ and $M_{\mathbf{b}}$ as the distribution of $\mathbf{b}$ when we sample without (which yields $R_d^\star(\lambda)$) and with replacement (which yields $\tilde{R}_d^\star(\lambda)$). First, notice that $g(\mathbf{q}^*, \mathbf{b}) \leq \frac{2(k-1)}{k} \leq 2$ by considering the TV between singleton distribution and uniform. Then, by triangular inequality, we have:

$$\left|\tilde{R}_d^\star(\lambda) - R_d^\star(\lambda)\right| = \frac{1}{4}\left|\left(\mathbb{E}_{\mathbf{b}\sim H_{\mathbf{b}}} g(\mathbf{q}^*, \mathbf{b}) - \mathbb{E}_{\mathbf{b}\sim M_{\mathbf{b}}} g(\mathbf{q}^*, \mathbf{b})\right)\right| \tag{67}$$

$$= \frac{1}{4}\left|\sum_{\mathbf{b}} g(\mathbf{q}^*, \mathbf{b})(H_{\mathbf{b}}(\mathbf{b}) - M_{\mathbf{b}}(\mathbf{b}))\right| \tag{68}$$

$$\leq \frac{1}{4} * 2\left|\sum_{\mathbf{b}} (H_{\mathbf{b}}(\mathbf{b}) - M_{\mathbf{b}}(\mathbf{b}))\right| \tag{69}$$

$$\leq \frac{1}{2}\sum_{\mathbf{b}} \left|(H_{\mathbf{b}}(\mathbf{b}) - M_{\mathbf{b}}(\mathbf{b}))\right| \tag{70}$$

$$= \mathsf{TV}(M_{\mathbf{b}}, H_{\mathbf{b}}) \tag{71}$$

$$\leq \frac{2k\lceil\frac{1}{\lambda}\rceil}{m} \tag{72}$$

The last inequality follows from de Finetti's Finite Exchangeable SequencesDiaconis & Freedman (1980).

A.11 Proof of Proposition 4

Let $n < \infty$ and assume that $X^n \sim Q^{\otimes n}$, $S^n \sim P^{\otimes n}$ and $(B_i^m)_{i=1}^n \sim P_{B^m}^{\otimes n}$. Consequently, the CC watermarked distribution is also i.i.d. distributed according $\tilde{Q} = Q_{X|S}$. On Bob's end, the detection probability is given by the expression

$$R_d = \frac{1}{2}\left(1 + \mathsf{TV}\left((PQ)^{\otimes n}, (P\tilde{Q})^{\otimes n}\right)\right),$$

where $P\tilde{Q}(S,X) = P(S)\tilde{Q}(X|S)$ To that end, we focus on obtaining bounds on the aforementioned TV term. For a pair of distributions P, Q, we have the following Hellinger bounds on the TV distance Polyanskiy & Wu (2024):

$$\frac{1}{2}H^2(P,Q) \leq \mathsf{TV}(P,Q) \leq H(P,Q)\sqrt{1-\frac{1}{4}H^2(P,Q)}, \tag{73}$$

where, for two measures P, Q on a finite alphabet $\mathcal{X}$, the squarred Hellinger divergence is given by the following equivalent forms

$$H^2(P,Q) \triangleq \mathbb{E}_Q\left[\left(1-\sqrt{\frac{P}{Q}}\right)^2\right] = \sum_{x\in\mathcal{X}}\left(\sqrt{P(x)}-\sqrt{Q(x)}\right)^2 = 2-2\sum_{x\in\mathcal{X}}\sqrt{P(x)Q(x)}.$$

For a pair of product distributions $(P^{\otimes n}, Q^{\otimes n})$, the squarred Hellinger divergence benefits from the relation Polyanskiy & Wu (2024)

$$H^2\left(P^{\otimes n}, Q^{\otimes n}\right) = 2-\left(1-\frac{1}{2}H^2(P,Q)\right)^n.$$

Our problem therefore boils down to characterize $H^2\left(PQ, P\tilde{Q}\right)$. We have

$$\begin{aligned} H^2\left(PQ, P\tilde{Q}\right) &= \sum_{x,s}P(s)\left(\sqrt{Q(x)}-\sqrt{\tilde{Q}(x|s)}\right)^2 \\ &= \mathbb{E}_S\left[H^2(Q(X),Q(X|S))\right]. \end{aligned}$$

For a given s, b^m),we have

$$\begin{aligned} H^2(Q(X),Q(X|S=s) &= 2-2\sum_x\sqrt{Q(x)\tilde{Q}(x|s)} \\ &= 2-2\sum_x Q(x)\sqrt{\frac{P_{S|Y}(s|y(x,b^m))}{P(s)}} \\ &= 2\mathbb{E}_X\left[1-\sqrt{\frac{P_{S|Y}(s|Y(X,b^m))}{P(s)}}\right], \end{aligned}$$

where $P(S|Y)$ is the correlated channel. Assuming $S \sim \mathsf{Ber}\left(\frac{1}{2}\right)$, we have

$$\begin{aligned} H^2\left(PQ, P\tilde{Q}\right) &= 2\mathbb{E}_{S,X}\left[1-\sqrt{\frac{P_{S|Y}(S|Y(X,b^m))}{P(S)}}\right] \\ &= \mathbb{E}_Y\left[1-\sqrt{2P(0|Y)}\right]+\mathbb{E}_Y\left[1-\sqrt{2P(1|Y)}\right] \\ &= 2-\sqrt{2}\mathbb{E}_Y\left[P(0|Y)+P(1|Y)\right] \\ &= 2-\sqrt{2}\left(\tilde{p}_0\left(\sqrt{p(0|0)}+\sqrt{p(1|0)}\right)+\tilde{p}_1\left(\sqrt{p(0|1)}+\sqrt{p(1|1)}\right)\right), \end{aligned}$$

where $Y \sim \mathsf{Ber}(\tilde{p}_0, \tilde{p}_1)$. Due to the symmetry of the correlated channel, we have for $\tilde{p} \triangleq \min(\tilde{p}_0, \tilde{p}_1)$

$$H^2\left(PQ, P\tilde{Q}\right) = 2-\sqrt{2}f(\tilde{p})$$

where

$$f(\tilde{p}) \triangleq \tilde{p}+\sqrt{\frac{1-\tilde{p}}{2}}\left(1+\sqrt{1-2\tilde{p}}\right),$$

which implies that

$$H^2\left(P^{\otimes n}, Q^{\otimes n}\right) = 2-2^{1-\frac{n}{2}}\left(f(\tilde{p})\right)^n.$$

The bounds on the detection probability then follow by plugging the squarred Hellinger distance into (73). □