Optimized Couplings for Watermarking Large Language Models

Carol Xuan Long[*†], Dor Tsur[* ‡†], Claudio Mayrink Verdun [†],
Hsiang Hsu[a§], Haim Permuter[‡], Flavio P. Calmon[†]
[†]John A. Paulson School of Engineering and Applied Sciences, Harvard University
[‡]School of Electrical Engineering, Ben-Gurion University
[§]JPMorgan Chase Global Technology Applied Research

***Abstract*—THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD. 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.**

I. 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 largely indistinguishable from humans begets the creation of trustworthy text generation algorithms [1] that create safe [2], interpretable [3], and authentic [4] 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 [5]–[21].

[*]Equal contributions.

[a]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.

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, [5] – which ignited the recent interest in LLM watermarking – distorts Q_X by randomly choosing a set of tokens (as determined by S) to be on a "green list," 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 [7], [9], [10], [17], 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 instance, generated from the hash of previous tokens in a sequence [5] or sophisticated tournament-like sampling strategies [12]. 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 $X_i \sim \widetilde{Q}_{X|S_i}$.
2) *Detection:* Given a sequence $\{(X_i, S_i)\}_{i=1}^n$, compute the statistic $T_n = \frac{1}{n}\sum_{i=1}^n f(X_i, S_i)$ 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 <u>*does not*</u> 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 generated tokens Q_{X^n} was known, then a simple likelihood ratio test (LRT) would suffice for watermark detection. What makes LLM watermarking distinct from existing information-theoretic watermarking schemes (e.g.,

[22]–[27]) 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.

In this paper, we analyze the one-token watermarking process, i.e., when $n = 1$. Specifically, we study how to generate a coupling $\widetilde{Q}_{X,S}$ and the corresponding detection function f that maximizes the watermark detection probability, while controlling the textual quality. The latter is controlled through the distortion relative to Q_X – a quantity we call *perception*, following recent trends in the source coding literature [28]–[30]. We refer to this setting as *one-shot watermarking*. We jointly optimize $\widetilde{Q}_{X,S}$ and f given a perception constraint, with the case $\widetilde{Q}_X = Q_X$ 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 watermark detection vs. perception trade-off 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 [31, Section 7]. This formulation embeds TV-information with a new operational interpretation.

We then 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)). 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 partitions lead to low detection probabilities, we introduce randomness to f. We show in Theorem 3 that randomized partitions yield non-trivial detection probabilities across *all* distributions that satisfy the min-entropy constraint. We identify the optimal coupling and quantify the detection probability in closed form for any value of $|\mathcal{S}|$ for a given Q_X and distribution of randomness in f. We pair our analysis with a characterization of the optimal design of randomization of the partition. When $\lambda = 1/2$, S is binary, and $|\mathcal{X}|$ is large, we show that the loss in detection for not knowing Q_X is at most $\frac{1}{8}$. We complement the discussion with numerical results and the extensions of our scheme to sequential watermarking.

Related Work. Watermarking has been extensively studied in information theory [24]–[26], particularly through the Gelfand-Pinsker (GP) channel [22], [23], [27]. These approaches typically focus on watermarking sequences via joint typicality and assume perfect knowledge of the underlying source distribution. The work of [5] led to various developments in watermarking schemes [9]–[20], with several approaches focusing on distortion-free methods, e.g., [6], [7], [15], [16]. In particular, [17] proposes a watermark using error-correcting codes leading to correlated channels similar to the ones we find via optimizing couplings. In [32], 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 [10] 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?* [1]

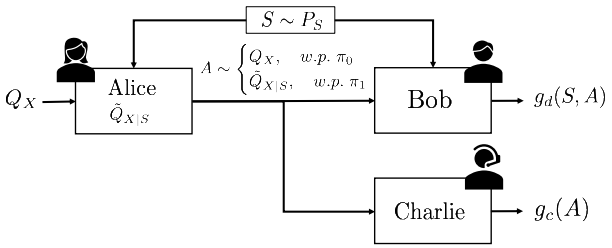

Fig. 1. Watermarking problem as a hypothesis test with side information.

II. 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 [5], [9], [10], [12]. 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 V.

A. Problem Setting

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[2] $S \sim P_S$ 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 $\widetilde{Q}_{X|S}$. Alice transmits a token according to a uniform prior:

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

[1] A comprehensive overview of related work, detailed proofs, and code implementations is available at https://github.com/Carol-Long/CC_Watermark

[2] Side information often corresponds to a secret shared key [7], [33]

where $C \perp\!\!\!\perp (X, \widetilde{X}, S)$. To detect the watermark, Bob performs a hypothesis test between $H_0 : Q_X$ and $H_1 : \widetilde{Q}_{X|S}$. Charlie is aware of the watermarking mechanism but is not aware of the specific sample of S. Therefore, Charlie performs a hypothesis test between $H_0 : Q_X$ and $H_1 : \widetilde{Q}$, where $\widetilde{Q} \triangleq \mathbb{E}_S[\widetilde{Q}_{X|S}]$ is the watermark distribution averaged w.r.t. the side information S.

B. 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 [28]–[30], 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.

C. Characterizing Optimal Trade-off

Following the Neyman-Pearson Lemma [34], 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., [35]–[37].

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

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

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

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[3] [37]

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

[3]Some works include a residual term $(1 - \gamma)_+$ [38], which we omit as it does not affect the optimization problem.

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 $R_d = \frac{1}{2} + \frac{1}{2}\mathsf{TV}(\widetilde{Q}_{X|S}, Q_X|P_S)$, where $\mathsf{TV}(\widetilde{Q}_{X|S}P_S, Q_X P_S) = \mathsf{TV}(\widetilde{Q}_{X|S}, Q_X|P_S)$.

Due to Jensen's inequality, for any fixed $(P_S, \widetilde{Q}_{X|S})$, 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:

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

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 $\widetilde{Q}_X = Q_X$). The following result establishes tight bounds on the optimal detection probability in this regime

Theorem 1 (Optimal Corner points)**.** *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*

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

The upper bound emerges from jointly optimizing over both the coupling $\widetilde{Q}_{X|S}$ 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 (Detection Upper Bound)**.** *Let $Q_{\min} \triangleq \min_{x \in \mathcal{X}} Q_X(x)$. For any $R_p \geq 0$ we have $R_d \leq 1 - \frac{\gamma Q_{\min}}{2}$.*

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 IV, 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.

III. 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 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). To the end, in what follows, we consider randomized partition functions.

A. Optimal Randomized Partition – Correlated Channel

We model f as a random assignment by introducing a set of m $\mathcal{S}$-valued variables denoted B^m. We assume that B^m is publicly available and is therefore not considered as a part of the secret side information S. Our goal is to therefore couple the side information with the randomized mapping $f(X, B^m)$. This boils down to the finding of a coupling of Q_X and S through the design of partition randomness P_{B^m} and conditional distribution $\widetilde{Q}_{X|S}$. We look for such $(P_{B^m}.\widetilde{Q}_{X|S})$ 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

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

where $\|Q_X\|_\infty = \max_{x\in\mathcal{X}} Q(x)$.

According to (6), given a fixed (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 $f(x, b^m) = b_x$ under which, the partition's probabilities are characterized by the distribution of the random variable $\widetilde{Y} \triangleq f(X, B^m)$. To this end, we first solve the following optimization problem:

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

This is a maximum coupling problem whose solution is given in the closed form, which is a direct consequence of the inf-representation of TV distance [39].

Proposition 2. *Let $S \sim \mathsf{Unif}[1:k]$ and $P_{\widetilde{Y}} = \{p_1, \ldots, p_k\} \in \Delta_k$, $t = \mathsf{TV}(P_S, \widetilde{Y})$ and let Ξ be the set of all couplings of $(P_S, P_{\widetilde{Y}})$. Then, $\operatorname{argmax}_{\xi\in\Xi} \Pr(S = \widetilde{Y})$ is given by*

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

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

When $k = 2$, the optimal coupling boils down to a Z-channel, as depicted in Figure 2 used, e.g., in [17]. We, therefore, term this method as the correlated channel (CC) watermark.

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 and sends $a \sim Q_X$. Otherwise, she samples and sends $a \sim \widetilde{Q}_{X|S=s}$, which is given by the CC:

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

Bob performs the detection test by checking whether $s = f(a, b^m)$. The complete list of steps is summarized in Algorithm 1. As a result of coupling the side information and random partition, we result with a zero perception scheme, i.e. $Q_X = \mathbb{E}_S[\widetilde{Q}_{X|S}]$.

Algorithm 1 Correlated Channel Watermark (CC)

Input: LLM distribution Q_X, Side information S, shared randomness B^m.

1: **Alice:**
2: Generate $\widetilde{Q}_{X|S,B^m}$ 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**

B. 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 the mutual information in the resulting channel.

Proposition 3. *The CC watermark detection is given by*

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

Specifically, for $|\mathcal{S}| = 2$, we have $R_d = \frac{1}{2}(1 + \widetilde{p})$, where $\widetilde{p} \triangleq \min(\widetilde{p}_0, \widetilde{p}_1)$ with $\widetilde{p}_i = P_{\widetilde{Y}}(i)$.

Remark 2. The CC detection test is equivalent to the LRT with threshold $\tau = 1$, as both tests attain the same decision region. This follows from the observation that $\Pr(S|f(S, B^m) \ge \frac{1}{2}$, if and only if $S = f(X, B^m)$.

R_d for tests of the form $\mathbf{1}(f(X) = S)$ is a function of Q_X and B^m through (9), while its expectation depends on P_{B^m}.

C. Optimizing the Partition

As seen from (9), the distribution of the resulting partition governs the detection power of the CC watermark. As Q_X cannot be controlled by the 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 P_{B^m} is permutation invariant.

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

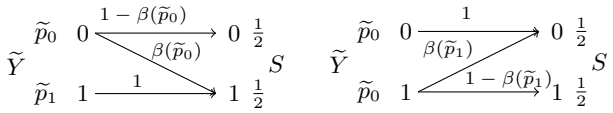

Fig. 2. Optimal coupling between side information S and random partition $\widetilde{Y} = f(X, B^m)$ for $\widetilde{p}_1 \leq 0.5$ (**left**), $\widetilde{p}_0 \leq 0.5$ (**right**), with $\beta(p) = \frac{2p-1}{2p}$.

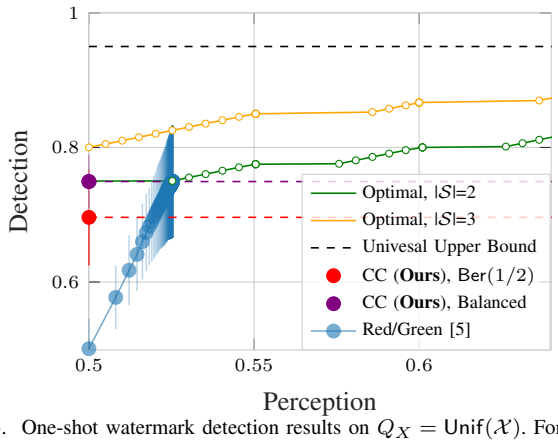

Fig. 3. One-shot watermark detection results on $Q_X = \mathsf{Unif}(\mathcal{X})$. For $\alpha_p = 0$, CC achieves a detection probability of 0.75 and 0.7 with balanced and Bernoulli partitions, respectively. CC Balanced achieves the optimal detection (Eq. 4 with $\gamma = 1$ and $|\mathcal{S}| = 2$). Standard deviations plotted as two-sided bars.

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

Theorem 3 (Optimal Max-min Detection)**.** *Let $\mathcal{S} = 2$ and $\mathcal{X} = m$. Given $\lambda \in [\frac{1}{2}, 1]$, let $\mathcal{Q}_{\infty,\lambda} \triangleq \{Q \in \Delta_m | \|Q\|_\infty \leq \lambda\}$. Then, for even m, the optimal detection probability is given by:*

$$R_d^*(\lambda) = \frac{3}{4} - \frac{m\lambda - 1}{4(m-1)}. \tag{10}$$

Furthermore, the optimal detection probability is achieved when $P_{B^m} = \mathsf{Unif}(B^*)$, where $B^* = \operatorname{argmax}_{\mathcal{P}_m} \frac{1}{|B_i|} \sum_{\mathbf{b} \in \mathcal{B}_i} (1 - 2\widetilde{p})$. The worst-case distribution Q_λ^* has two non-zero entries equal to λ and $1 - \lambda$.

We note that due to (9), when $k = 2$, R_d is upper bounded by $\frac{3}{4}$. Thus, the second term in (10) serves as a penalty when considering the max-min setting. Notably, 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 $\frac{1}{8}$.

IV. Experimental Results

We numerically evaluate the CC watermark on a uniform source over $m = 10$ tokens and binary side information. We compare our results with the solution of an exact GUROBI-based numerical solution [40] of (4) and a popular watermarking scheme termed red/green list [5]. The red/green watermark tilts Q_X according to the value of some $\delta \in \mathbb{R}_{\geq 0}$ which implicitly controls the downstream R_p. As seen from Figure 3, when we apply the CC scheme with P_{B^m} sampled over balanced partitions, we obtain a gain of ≈ 0.05 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 $\delta \approx 7.6$.

V. 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 [5], [9], [10], albeit being theoretically suboptimal. We note that our one-shot method readily extends to a sequential token-level scheme as we can treat each step as a single-shot problem, and considering an average test $\frac{1}{n}\sum \mathbf{1}[f_i(A_i, B_i^m) = S_i]$. We leave the analysis of the token-level extension of our scheme to future work. In the simplified case when X^n are i.i.d., we provide bounds on the detection probability (a related result was given in [17] 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*

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

where $\widetilde{p} = \min(\widetilde{p}_0, \widetilde{p}_1)$ 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 tensortization.

VI. 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.

REFERENCES

[1] 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.

[2] 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.

[3] 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*, pages 5484–5495, 2021.

[4] 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)*, pages 3214–3252, 2022.

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

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

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

[8] 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*, 2024.

[9] 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-.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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*, pages 5508–5519, 2024.

[15] 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.

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

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

[18] 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.

[19] 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.

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

[21] 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)*, pages 1–6. IEEE, 2023.

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

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

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

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

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

[27] 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, pages 406–416. SPIE, 2006.

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

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

[30] 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.

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

[32] 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.

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

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

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

[36] 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.

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

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

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

[40] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

APPENDIX

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. 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 [24]–[26], particularly through the Gelfand-Pinsker (GP) channel [22], [23], [27], 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. [5] 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 [9]–[20], with several approaches focusing on distortion-free methods that maintain the original NTP distribution unchanged, e.g., [6], [7], [15], [16]. 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. [17] 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. [32] 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. [41] 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. [10] 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 [42]–[45] and surveys [33], [46].

B. Proof for Proposition 1

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

Eve's hypothesis testing problem can be formulated as distinguishing between $H_0 : A \sim Q_X$ and $H_1 : A \sim \widetilde{Q}_X$. By the Neyman-Pearson Lemma, the optimal test statistic is given by the likelihood ratio $L(a) = Q_X(a)/\widetilde{Q}_X(a)$. 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:

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

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 $H_1 : (A, S) \sim \widetilde{Q}_{X|S} \times P_S$.

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

$$\begin{aligned}
\Pr(\hat{H}_B = C|S = s) &= \pi_0 \Pr(\delta(A) = 0|H_0) + \pi_1 \Pr(\delta(A) = 1|H_1) && (11)\\
&= \pi_0 Q_{X|S}[L(a,s) \geq \gamma] + \pi_1 \widetilde{Q}_{X|S}[L(a,s) \leq \gamma] && (12)\\
&= \pi_1 + \pi_0 Q_{X|S}[L(a,s) \geq \gamma] - \pi_1 \widetilde{Q}_{X|S}[L(a,s) \geq \gamma] && (13)\\
&= \pi_1 + \pi_0 \left[Q_{X|S}[L(a,s) \geq \gamma] - \frac{\pi_1}{\pi_0} \widetilde{Q}_{X|S}[L(a,s) \geq \gamma \right] && (14)\\
&= \pi_1 + \pi_0 E_\gamma(Q_{X|S} || \widetilde{Q}_{X|S}). && (15)
\end{aligned}$$

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 $A = \{a|L(a,s) \geq \gamma\}$. □

C. Proof of Theorem 1

By the assumption of a uniform prior, we are looking for bounds on the quantity $\frac{1}{2}(1 + E_\gamma(\widetilde{Q}_{X|S} \| Q_X|P_S))$, which boils down to bounding $E_\gamma(\widetilde{Q}_{X|S} \| Q_X|P_S) = \mathbb{E}_S \left[E_\gamma(\widetilde{Q}_{X|S} \| Q_X) \right]$. 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 $E_\gamma(\widetilde{Q}_{X|S} \| Q_X|P_S)$. Recall that $|\mathcal{X}| = m$ and $|\mathcal{S}| = k$. Let $Q_{X|S=s_i} = p_i$ such that $p_1, ..., p_m \in \Delta_m$, where Δ_m denotes the m-dimensional simplex. Assume that $S \sim \mathsf{Unif}[k]$. Following the zero perception assumption, we have $\widetilde{Q}_X = Q_X$, i.e., $\frac{1}{k} \sum_{i=1}^k p_i = Q_X$. Consequently, our TV-optimization, when jointly optimized also over the marginal distribution Q_X is of the form:

$$\max_{p_1,\ldots,p_k \in \Delta_m} \frac{1}{k} \sum_{i=1}^k \left\| p_i - \frac{\gamma}{k} \sum_{i=1}^k p_i \right\|_+, \tag{16}$$

where $\|x\|_+ \triangleq \sum_i (x_i)_+$ for $m \geq k$. 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 $j \leq m$, 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

$$q \triangleq \frac{1}{k} \sum_{i=1}^k p_i = \frac{1}{m} \sum_{j=1}^m n_j e_j, \tag{17}$$

where $\sum_j n_j = k$ 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:

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

where (a) follows from from rewriting the sum in terms of e_j and (b) follows from the relation $q_j = \frac{n_j}{k}$, as can be seen from (17) and by the definition of the indicator. Out optimization problem had therefore boiled down to maximizing on the quantity

$$\sum_{j=1}^m q_j (1 - \gamma q_j)_+ \text{ such that } q_j = \ell/k, \ell = 0, \ldots, k, \sum_{j=1}^m q_j = 1. \tag{18}$$

To solve (18), we will examine various settings of the value of γ.

1) $\gamma \leq 1$: 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

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

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 $\frac{1}{k}$. Thus, we have the upper bound

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

2) $\gamma > 1$: 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 $\left\{\frac{k}{m}, k=0,\ldots,m\right\}$, 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 $f(q_j)$. A simple strategy would choose to set as many q_j values to $\frac{1}{m}$. Within the considered range, we could set m such q_j values to $\frac{1}{m}$, resulting in

$$E_\gamma(\widetilde{Q}_{X|S}\|Q_X|P_S) = 1-\frac{\gamma}{m}.$$

Consequently, the detection rate is given by $R_d = 1-\gamma/(2m)$, which coincides with the case of $\gamma \leq 1$. Thus, a unifying term for the detection upper bound is given by

$$R_d \leq \max\left(\frac{1}{2}, 1-\frac{\gamma}{2m}\right),$$

which concludes the proof.

We complement the proof by giving an additional upper bound when allowing for a finer granularity over the possible values of γ, again focusing on the range $\gamma \in (1,k)$. 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 $\frac{1}{2\gamma} \notin (\frac{\ell}{k})_{\ell=1}^k$. 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 $\sum_{j=1}^m q_j = 1$, we will choose the lower value. To summarize, for each interval $\frac{\ell}{k} \leq \frac{1}{2\gamma} \leq \frac{\ell+1}{k}$, we will set $q_j = \frac{\ell}{k}$. The maximal amount of such q_j we can set while following the sum constraint is $\lfloor \frac{\ell}{k} \rfloor$. Thus, we have the following

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

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

D. Proof of Theorem 2

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

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

Therefore,

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

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

E. Proof of Correlated Channel (CC) with Perfect Perception

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

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

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

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

This concludes the proof. □

F. Proof of Proposition 2

By the dual representation of the total variation

$$\mathsf{TV}(P,Q) = \min_{P_{XY}}\{\mathbb{P}[X\neq Y] : P_X = P, P_Y = Q\}, \tag{19}$$

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

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

To simplify notation, let the distribution of S and $\widetilde{Y}$ 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 $S = \widetilde{Y} = Z$.
3) If $B = 1$, draw $S \sim P'$ and $\widetilde{Y} \sim Q'$ 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,

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

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

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

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

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

G. Proof of Remark 2

The hypothesis test is the following: $H_0 : X \sim Q_X$ and $H_1 : X \sim \widetilde{Q}_{X|S,B^m}$, where $\widetilde{Q}_{X|S,B^m}$ 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:

$$\begin{aligned}\frac{Q_X(X)}{\widetilde{Q}_{X|B^m,S}(X)} &= \frac{Q(X)}{Q_X(X)\frac{1}{P_S(S)}P_{S|\widetilde{Y}}(S|f(X,B^m)} && (20)\\ &= \frac{2}{P_{S|\widetilde{Y}}(S|f(X,B^m)} && (21)\\ &< 1, && (22)\end{aligned}$$

The density of $\widetilde{Q}_{X|B^m,S}(X)$ follows from the CC-watermark, side information $P_S(S) = 0.5$. The last inequality come from the Z-S channel construction: $\Pr_{S|\widetilde{Y}}(S|f(S,B^m) \geq \frac{1}{2}$, 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

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

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

H. Proof of Proposition 3

We start by proving the following identity:

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

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

$$\mathsf{TV}(Q_X, \widetilde{Q}_{X|S,B^m}|P_{S,B^m}) = \sum_{b^m}\sum_{s=0,1}\mu(b^m)P_S(s)\mathsf{TV}(Q_X, \widetilde{Q}_{X|b^m,s}). \quad (23)$$

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

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

where recall that $\widetilde{Y} = f(X, B^m)$, $p_{S|\widetilde{Y}}(s|\widetilde{y})$ 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 (23), we have

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

where the randomness of $\widetilde{Y}$ 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

$$R_d = \frac{1}{2}(1 + \mathsf{TV}(Q_X, \widetilde{Q}_{X|S,B^m} | P_{S,B^m})) \tag{24}$$

$$= \frac{1}{2}\left(1 + \mathsf{TV}(P_S, P_{S|\widetilde{Y}} | P_{\widetilde{Y}})\right), \tag{25}$$

where the last equality is due to the identity above.

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

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

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

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

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

Thus, we have

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

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

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

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

I. Proof of Theorem 3

We begin by proving Lemma 1.

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

$$f(i, \mathbf{b}) = b_i. \tag{26}$$

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)

$$\hat{P}(s, \mathbf{q}, \mathbf{b}) = \sum_{i=1}^{m} q_i \mathbf{1}\left[b_i = s\right] \ \ \forall s \in [k]. \tag{27}$$

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

$$R_d(\mathbf{q}, \mathbf{b}) \triangleq 1 - \frac{1}{2}\mathsf{TV}\left(Q_S \| \hat{P}(\mathbf{q}, \mathbf{b})\right) - \frac{1}{2k}\sum_{s=1}^{k} \hat{P}(s, \mathbf{q}, \mathbf{b}) \tag{28}$$

$$= 1 - \frac{1}{2k} - \frac{1}{4} g(\mathbf{q}, \mathbf{b}), \tag{29}$$

where

$$g(\mathbf{q}, \mathbf{b}) \triangleq \sum_{s=1}^{k} \left| \hat{P}(s, \mathbf{q}, \mathbf{b}) - \frac{1}{k} \right| \tag{30}$$

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:

$$R^*_d(\lambda) \triangleq \max_{P_{B^m}} \min_{\substack{\mathbf{q} \in \Delta_m \\ \|\mathbf{q}\|_\infty \leq \lambda}} \mathbb{E}\left[R_d(\mathbf{q}, B^m)\right] \tag{31}$$

$$= 1 - \frac{1}{2k} - \frac{1}{4} \min_{P_{B^m}} \max_{\substack{\mathbf{q} \in \Delta_m \\ \|\mathbf{q}\|_\infty \leq \lambda}} \sum_{\mathbf{b} \in \mathcal{S}^m} P_{B^m}(\mathbf{b}) g(\mathbf{q}, \mathbf{b}). \tag{32}$$

The function

$$H(P_{B^m}) \triangleq \max_{\substack{\mathbf{q} \in \Delta_m \\ \|\mathbf{q}\|_\infty \leq \lambda}} \mathbb{E}\left[g(\mathbf{q}, B^m)\right] \tag{33}$$

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 $\widetilde{P}_\pi(\mathbf{b}) = P^*_{B^m}(\pi \circ \mathbf{b})$.

Since $\mathbb{E}_{P^*_{B^m}}\left[g(\mathbf{q}, B^m)\right] = \mathbb{E}_{\widetilde{P}_\pi}\left[g(\pi \circ \mathbf{q}, B^m)\right]$ for all $\mathbf{q}$, $H(\widetilde{P}_\pi) = H(P_{B^m})$ from the symmetry of the maximum. Hence, from the equality in (32) $F(\widetilde{P}_\pi) = F(P_{B^m})$ for $F(P_{B^m}) \triangleq \min_{\substack{\mathbf{q} \in \Delta_m \\ \|\mathbf{q}\|_\infty \leq \lambda}} \mathbb{E}_{P_{B^m}}\left[R_d(Q_X, B^m)\right]$. □

Next, we proceed with the proof of Theorem 3.

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

$$F\left(\frac{1}{C}\sum_{\pi} \widetilde{P}_\pi\right) \leq F(P^*_{B^m}). \tag{34}$$

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

$$\min_{P_{B^m}} F(P_{B^m}) = \min_{\mathbf{w} \in \Delta_K} \max_{\substack{\mathbf{q} \in \Delta_m \\ \|\mathbf{q}\|_\infty \leq \lambda}} \sum_{i=1}^{K} \frac{w_i}{|\mathcal{B}_i|} \sum_{\mathbf{b} \in \mathcal{B}_i} g(\mathbf{q}, \mathbf{b}). \tag{35}$$

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 $\{\mathbf{q} \in \Delta_m \mid \|\mathbf{q}\|_\infty \leq \lambda\}$ 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 $t\lambda \leq 1$ (assuming $\lambda \leq 1$), (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

$$\sum_{\mathbf{b}\in\mathcal{B}_i} g(\mathbf{q},\mathbf{b}) = \sum_{\mathbf{b}\in\mathcal{B}_i} g(\pi\circ\mathbf{q},\mathbf{b}), \tag{36}$$

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

$$\min_{P_{B^m}} F(P_{B^m}) = \min_{\mathbf{w}\in\Delta_K} \sum_{i=1}^{K} \frac{w_i}{|\mathcal{B}_i|} \sum_{\mathbf{b}\in\mathcal{B}_i} g(\mathbf{q}_\lambda^*,\mathbf{b}), \tag{37}$$

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 $\mathcal{B}\in\mathcal{P}_m$.

For $1/2\le\lambda<1$, $\mathbf{q}_\lambda^*$ 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 $P_{\mathcal{B}}$

$$\mathbb{E}_{P_{\mathcal{B}}}\left[R_d(\mathbf{q}_\lambda^*,B^m)\right] = 1-\frac{1}{2k}-\Pr(B_1=B_2)\times\frac{k-1}{2k}-\frac{1}{4}\Pr(B_1\neq B_2)\times\left(1-\frac{2}{k}+\left|\lambda-\frac{1}{k}\right|+\left|1-\lambda-\frac{1}{k}\right|\right). \tag{38}$$

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

$$R_d^*(\lambda) = \frac{3}{4}-\frac{m\lambda-1}{4(m-1)} \text{ for } k=2,\ \frac{1}{2}\le\lambda\le 1. \tag{39}$$

J. 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 $\widetilde{Q}=\widetilde{Q}_{X|S}$. On Bob's end, the detection probability is given by the expression

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

where $P\widetilde{Q}(S,X)=P(S)\widetilde{Q}(X|S)$ 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 [31]:

$$\frac{1}{2}H^2(P,Q)\le\mathsf{TV}(P,Q)\le H(P,Q)\sqrt{1-\frac{1}{4}H^2(P,Q)}, \tag{40}$$

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 [31]

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

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

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

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

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

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

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

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

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

where

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

which implies that

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

The bounds on the detection probability then follow by plugging the squarred Hellinger distance into (40). □