Optimal Couplings for Distortion-Free Watermarking

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. We present an information-theoretic framework for analyzing the fundamental limits of detection performance and perceptual quality of watermarking schemes for large language models (LLMs). We formulate the watermarking problem through the lens of hypothesis testing with side information, thereby characterizing the trade-off between watermark detection and perception in terms of E_γ divergence and total variance distance. We optimize randomized tests in a minimax setting — finding the distribution of randomness that maximizes mean detection for worst case token distribution. We propose a Correlated Channel (CC) watermarking scheme, which boils down to an optimal coupling between the randomized test and the side information through the Z-channel. We show that the optimum mean detection in the minimax setting is achieved by CC and the specified distribution of randomness. Our scheme operates under perfect perception, whereby for an average user of the LLM who lacks knowledge of the side information, the watermarked distribution equals to the original token distribution. Furthermore, we describe the sequential extension of our scheme and provide preliminary theoretical analysis. We validate our results numerically, demonstrating that CC matches the analytical optimal solution under a fixed distribution.**

I. Introduction

The emergence of powerful large-language models (LLMs) has presented the need to develop trustworthy text generation algorithms [1], with the purpose of creating safe [2], interpretable [3] and authentic [4] content. This work focuses on watermarking, which is the process of embedding a signal at token level in an LLM-generated text [5]–[21], carrying proof of its origin or authenticity. The strength of a given watermark can be evaluated across six domains [22], encompassing output quality, false positive and negative rates, efficiency, robustness to adversarial manipulations, and embedding capacity. Recent works have made analysis using various tools across subsets of those domains [10], [23], [24]. Herein, we focus on the trade-off between detection capability and generation quality.

[*]Equal contribution.

[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.

In the information theory literature, the problem of watermarking has been analyzed through the lenses of information hiding and embedding [25] and steganography [26]–[29], and connections were made to the Gelfand-Pinsker channel [30], [31]. Similar to the information hiding game, LLM watermarking can be abstracted into a game involving an encoder who transmits the side information and a decoder who aims to detect the side information. However, the assumptions of the constraints in the two problems are entirely different. While the information hiding problem is constrained by an attacker modeled as a noisy channel that aims to remove the hidden information, watermarking in our analysis is constrained by a regular user for whom the watermarked text should be indistinguishable and that generation quality is preserved.

We focus on the design of optimal couplings in the one-shot setting. While LLM watermarking is an auto-regressive and hence a sequential process, by characterizing the detection and perception limits in the one-shot setting, we delineates the framework, metrics, and components of the watermarking scheme that needs to be optimized in order to maximize detection and minimize the distinction between a watermarked and unwatermarked text. We characterize the fundamental limits of watermark detection versus perceptual quality through E_γ divergence, leading to precise bounds on achievable detection-perception pairs. With the assumption that the detector has no knowledge of the underlying distribution, given side information S and original distribution Q_X, we optimize randomized tests of the form $\mathbf{1}[f(X, B^m) = S]$ in a minimax setting — finding the distribution of randomness P_{B^m} that maximizes mean detection for worst case Q_X. In addition to providing the optimal coupling between $f(X)$ and S that maximizes detection, which has appeared in existing work [17], we proved the optimal distribution of randomness and specify adversarial Q_X for which mean detection in the minimax setting is optimized. Meanwhile, the design of our scheme is anchored on the perfect perception setting, where for an average user who lacks access to side information, the watermarked distribution equals to Q_X. We provide the sequential extension of our scheme and discusses preliminary analysis.

Related Work. 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 [32]–[34], particularly through the Gelfand-Pinsker (GP) channel [25], [30], [31], 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], [21]. 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 a minimax framework in the one-shot setting.

Theoretical Analysis of LLM Watermarking. Recent work has advanced our theoretical understanding of LLM watermarking limitations. Huang et al. [23] 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. [24] 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 error through the parameter of the E_γ objective. In the minimax formulation, we provide in closed form the optimal mean detection and characterized the optimal distribution of randomness under adversarial token distributions.

The development of the field is tracked through comprehensive benchmarks [35]–[39] and surveys [22], [40].

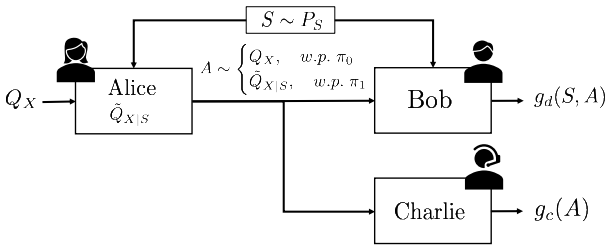

Fig. 1. Watermarking problem as a hypothesis test with side information.

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 tradeoff between detection probability and perceptual quality. While the optimal approach to watermarking would a sequence to sequence scheme, 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[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 $\widetilde{Q}_{X|S}$. Alice transmits a token according to a uniform prior:

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

where $C \perp\!\!\!\perp (X, \widetilde{X}, S)$. To detect the watermark, Bob performs a hypothesis test between $H_0 : Q_X$ and $H_1 : \widetilde{Q}_{X|S}$. On the other hand, Charlie is aware of the watermarking mechanism but is not aware of the specific sample of S. Therefore, Charlie performs an 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 [41]–[43], we adopt the

[1]Side information often corresponds to a secret shared key [7], [22]

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*

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

A perception test $g_p : \mathcal{X} \to \{0, 1\}$ is equivalently defined with 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.

C. Characterizing Optimal Trade-off

Following the Neyman-Pearson Lemma [44], 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., [45]–[47].

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

$$R_d = \pi_1 + \pi_0 \mathsf{E}_\gamma \left(\widetilde{Q}_{X|S}, Q_X | P_S\right) \tag{1}$$

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

where $\mathsf{E}_\gamma(P_{X|S}, Q_{X|S}|P_S) \triangleq \mathsf{E}_\gamma(P_{X,S}, Q_X P_S)$ is the E_γ-information.

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 $\mathcal{A}$,

$$E_\gamma(P||Q) \triangleq \max_{\mathcal{A}} [P(\mathcal{A}) - \gamma Q(\mathcal{A})],$$

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

$$\sup_{\widetilde{Q}_{X|S}} \mathsf{E}_\gamma \left(\widetilde{Q}_{X|S}, Q_X|P_S\right), \quad \text{s.t.} \quad \mathsf{TV}\left(\widetilde{Q}_X, Q_X\right) \leq \alpha_p \tag{3}$$

We are interested in characterizing the (R_d, R_p) trade-off region, which amounts to solving (3) as a function of α_p.

Note that (3) 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 $[1/2, 1]^2$.

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)**.** *Let P_S be uniform over $\mathcal{S}$, $|\mathcal{S}| \leq |\mathcal{X}|$ and let $\pi_1 = \frac{1}{2}$. Given watermark distribution Q_X and for $R_p = \frac{1}{2}$, we have*

$$\frac{1}{2} \leq \sup_{\widetilde{Q}_{X|S}} R_d \leq \mathsf{UB}(\gamma), \tag{4}$$

where,

$$\mathsf{UB}(\gamma) \triangleq \begin{cases} 1 - \frac{\gamma}{2|\mathcal{S}|}, & \gamma \leq \frac{|\mathcal{S}|}{2} \\ 1 - \frac{k\gamma}{2|\mathcal{S}|}, & \left(\frac{|\mathcal{S}|}{2(k)} \leq \gamma \leq \frac{|\mathcal{S}|}{2(k-1)}\right), k \in \left[2 : \left\lfloor \frac{|\mathcal{S}|}{2} \right\rfloor\right] \\ \frac{1}{2}, & \gamma > |\mathcal{S}|. \end{cases}$$

The upper bound emerges from jointly optimizing over both the watermarking strategy $\widetilde{Q}_{X|S}$ and the LLM distribution Q_X. This optimization reduces to a convex problem over the probability simplex, which ultimately becomes the counting problem of optimally assigning elements of $\mathcal{X}$. We complement this with a lower bound, achieved when Q_X concentrates on a single element.

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 (3) results in a nonconvex region, which generally lacks a closed form. This non-convexity is demonstrated in our experimental results (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, of course, a likelihood ratio test (LRT), and we have characterized the optimal detection-perception trade-off by total variation distance, LRT is infeasible since Bob lacks knowledge of Q_X. Under this core constraint, given side information S, we provide the optimal watermarking scheme of the form $\mathbf{1}[f(X) = S]$ for some $f : \mathcal{X} \to \mathcal{S}$.

In our abstraction of the watermarking problem (Figure 1), Bob is assumed to access only the transmitted token A and side information S. Hence, Bob's decision function is

limited to the form $f : \mathcal{X} \mapsto \{0, 1\}$, where 1 rejects H_0 and declares the presence of a watermark and 0 its absence. This decision function effectively partitions $\mathcal{X}$ into two subsets: one where a watermark is declared and another where it is not [5]. To generalize this idea of binary partition, we consider $|S|$ partitions and assume Bob employs tests of the form $\mathbf{1}\{f(x) = s\}$, where $f : \mathcal{X} \to \mathcal{S}$ is a pre-determined function.

First, note that when the side information S is uniform over $\mathcal{S}$, a deterministic detection test f attains the same error of $1/k$ for any $Q \in \Delta_{|\mathcal{X}|}$. To optimize for detection rate, we consider randomized partitions next.

A. Optimal Randomized Partition – Correlated Channel

We now present the optimal watermarking scheme of the form $\mathbf{1}[f(X, B^m), S]$, where f is a random function of X and $B^m \triangleq (B_1, \ldots, B_m)$, a set of k-valued random variables. We formulate the optimal detection as a minimax optimization problem where we seek a distribution over $P^*_{B^m}$ that maximizes the worst-case value of R_d given an ℓ_∞ constraint on Q_X. This constraint excludes deterministic Q_X for which watermarking necessarily introduces perception. Specifically, we analyze:

$$R^*_d(\lambda) \triangleq \max_{P_{B^m}} \min_{\substack{Q_X \in \Delta_m \\ \|Q_X\|_\infty \le \lambda}} \mathbb{E}\left[R_d(Q_X, B^m)\right] \tag{5}$$

This formulation involves optimizing two interdependent components. Given a fixed P_{B^m} and Q_X, we maximize the probability of correct detection $\Pr[f(X, B^m) = S]$ by designing the optimal coupling between $f(X, B^m)$ and S. Then, we investigate the optimal P_{B^m} for worst-case adversarial Q_X in III-C.

Consider the mapping $f(x, b^m) = b_x$, which effectively randomly assigns a bin to each $x \in \mathcal{X}$. Consequently, the partition's probabilities are characterized by the distribution of the assignment $\widetilde{Y} \triangleq f(X, B^m)$. Our goal is to solve the following optimization:

$$\sup_{P_{S,\widetilde{Y}}} \Pr(S = \widetilde{Y}), \quad S \sim \mathsf{Unif}\left(\mathcal{S}\right), \widetilde{Y} \sim P_{\widetilde{Y}}. \tag{6}$$

This is a maximum coupling problem whose solution is given in the closed form, which is a direct consequence of the inf-representation of total variation distance [48].

Proposition 2. *Let $S \sim \mathsf{Unif}[k]$ and $P_{\widetilde{Y}} = \{p_1, \ldots, p_k\} \in \Delta_k$. Let Π be the set of all couplings of $(P_S, P_{\widetilde{Y}})$. Then, the solution to $\max_{\pi \in \Pi} \Pr(S = \widetilde{Y})$ is given by*

$$\pi(y, \widetilde{y}) = \begin{cases} \min(\frac{1}{k}, p_i), & s = \hat{y} = i \\ \frac{1}{k}(\frac{1}{k} - \min(\frac{1}{k}, p_j)), & \textit{otherwise} \end{cases}$$

When $k = 2$, the coupling of S and $\widetilde{Y}$ boils down to a Z-S channel, which is depicted in Figure 2.

The CC watermarking scheme consists of the following steps: Both Alice and Bob receive a sample (s, b^m). Alice generates a token from Q_X from the LLM model and $\widetilde{Q}_{X|S}$, which is given by the CC:

$$\widetilde{Q}_{X|s,b^m}(x) = Q_X(x) \frac{P_{S|\widetilde{Y}}(s|f(x, b^m))}{P_S(s)}. \tag{7}$$

Algorithm 1 Correlated Channel Watermark (CC)

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

1: **Alice:**
2: Sample one token from both Q_X and $\widetilde{Q}_{X|S,B^m}$ according to (7)
3: Flip a coin $C \sim \mathsf{Ber}(\frac{1}{2})$ and send composite token A according to the outcome of C
4: **Bob:**
5: **if** $S = f(A, B^m)$ – Declare **Watermarked**
6: **else** – Declare **Not watermarked**

Alice then communicates the mixture token a. Bob performs the detection test by checking whether $s = f(a, b^m)$. The complete list of steps is summarized in Algorithm 1.

B. Theoretical Analysis of Correlated Channel

We provide a complete analysis of the CC scheme under $k = 2$. First, we show that the CC is perceptionless by definition

Lemma 1. *Under the CC scheme $Q_X = \mathbb{E}_S\left[\widetilde{Q}_{X|S}\right]$.*

Next, we show Bob's detection probability in closed form. Given the optimal coupling, detection probability can be characterized by total variation distance of $\widetilde{Y}$ and S.

Proposition 3. *Let $\widetilde{p} \triangleq \min(\widetilde{p}_0, \widetilde{p}_1)$ with $\widetilde{p}_i = P_{\widetilde{Y}}(i)$. The CC watermark detection is given by*

$$R_d = \frac{1}{2}\left(1 + \mathsf{TV}\left(P_S, P_{S|\widetilde{Y}} | P_{\widetilde{Y}}\right)\right) = \frac{1}{2}(1 + \widetilde{p}) \tag{8}$$

Remark 2. The CC detection test is equivalent to the likelihood ratio test with threshold $\tau = 1$, because both test attain the same decision region. This can be attained from the observation that $\Pr_{S|\widetilde{Y}}(S|f(S, B^m) \ge \frac{1}{2}$, if and only if $S = f(X, B^m)$.

Thus, R_d for tests of the form $\mathbf{1}(f(X) = S)$ is a function of Q_X and B^m through (8), where

$$\widetilde{p} = \min_{i=\{0,1\}} [\mathbb{E}_{Q_X}(\mathbf{1}[f(X, B^m) = i]))] \tag{9}$$

C. Optimizing the Partition

As seen from (8), the distribution of the resulting partition governs the detection power of the CC watermark. The distribution Q_X cannot be controlled by the designer. Therefore, we aim to investigate the optimal choice of $P(B^m)$ on the detection probability under the worst-case adversarial distribution Q_X.

Due to the symmetry in the problem, we can restrict the optimization over P_{B^m} to distributions that assign equal probabilities to sequences that are identical up to a permutation.

Lemma 2. *Let $F(P_{B^m}) \triangleq \min_{\substack{\mathbf{q} \in \Delta_m \\ \|\mathbf{q}\|_\infty \le \lambda}} \mathbb{E}_{P_{B^m}}\left[R_d(Q_X, B^m)\right]$. Let $P^*_{B^m}$ be a distribution that maximizes $F(P_{B^m})$. Consider a permutation: $\pi : \mathcal{S}^m \mapsto \mathcal{S}^m$. Define $\widetilde{P}_\pi(B^m) = P^*_{B^m}(\pi \circ B^m)$ Then, $F(P_{B^m}) = F(\widetilde{P}_\pi)$.*

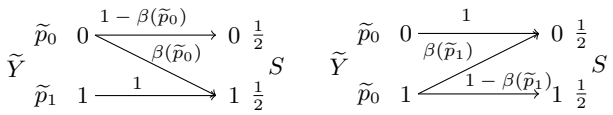

Fig. 2. Optimal coupling between side information S and random partition $\widetilde{Y} = f(X, B^m)$ for $\widetilde{p}_1 \leq 0.5$ (**left**), $\widetilde{p}_0 \leq 0.5$ (**right**), with $\beta(p) = \frac{2p-1}{2p}$.

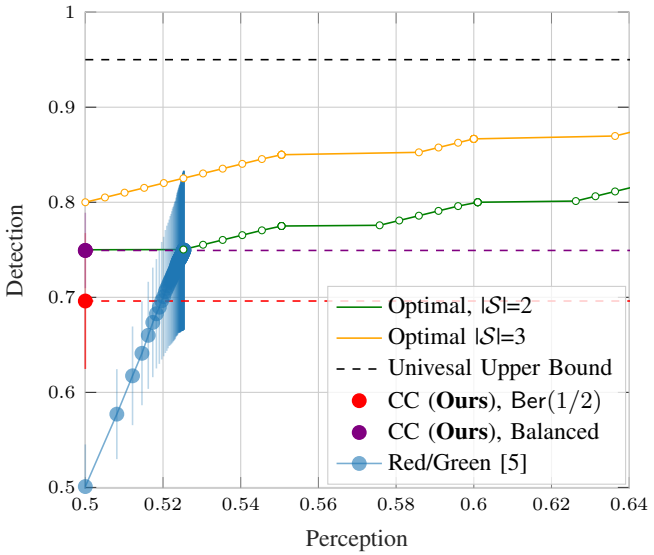

Fig. 3. Experimental results of one-shot watermark detection on the uniform distribution with $|\mathcal{X}| = 10$. Standard deviations are plotted as one-sided error bars. At perfect perception, CC (our method) achieves a detection probability of 0.75 and 0.7 with balanced and Bernoulli partition respectively. CC Balanced matches the TV optimal solution (Eq. 3 with $\gamma = 1$ and $|S| = 2$).

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. *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 m even, the optimal detection probability is given by:*

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

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

IV. Experimental Results

We numerically evaluate the CC watermarking scheme on a uniform token distribution over 10 elements. By Theorem 1 and Proposition 3, CC under a uniform distribution with balanced partition achieves the upper bound of optimal detection rate for $|S| = 2$ and $\gamma = 1$, which we aim to validate through experimental results. Specifically, we compare our scheme with $k = 2$ with the optimal solution from (3) for two values of k. These solutions are obtained by solving the optimization via GUROBI [49]. Furthermore, we compare our method with a one-shot version of the popular red-green watermark [5], which applies a tilting operation of Q_X according to a binary partition and some $\delta \in \mathbb{R}_{\geq 0}$. The results are presented in Figure 3. For our scheme, we consider sampling B^m from the product of Bernoulli distributions and from the subset of balanced partitions (i.e., an equal number of 0's and 1's). For this method, we consider $\delta \in [0, 100]$. The detection results are averaged over $n = 100$ experiments, and error bars are calculated from standard deviation. As expected, our scheme achieves the optimal zero-perception detection probability when sampling balanced partitions. Sampling from the balanced set is attainable in practice, as it can be viewed as a random permutation of any balanced starter partition. Even if we consider the sub-optimal scheme of sampling over all B^m, the resulting detection probability is significantly higher than the one attained by the red-green list, and the intersection occurs on $\delta \approx 7.6$.

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 sequence. The common paradigm of LLM watermarking follows token-level methods [5], [9], [10], which apply the watermark to each distribution $Q_X(\cdot|x^{i-1})$ separately and focuses on detection tests of the form $f(x_i, s_i, b_i^m)$. We note that our scheme naturally extends to the sequential nature by applying the CC scheme (Algorithm 1) on each time step separately. The resulting test statistic is given by $\frac{1}{n} \sum \mathbf{1}[f_i(A_i, B_i^m) = S_i]$.

Although applying a series of token-level strategies is potentially suboptimal, such schemes have proven successful in the task of LLM watermarking. We leave the full characterization of the gap for future work but provide detection bounds in the i.i.d. setting:

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*

$$1 - 2^{-\left(\frac{n}{2}+1\right)} \left(g(\widetilde{p})\right)^n \leq R_d \leq \frac{1}{2}\left(1 + \sqrt{1 - \left(\frac{(g(\widetilde{p}))^2}{2}\right)^n}\right), \tag{10}$$

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

Using the setting of hypothesis testing with side information, this work studied the trade-off between detection and perception in one-shot token watermarking, which is a fundamental

problem in trustworthy LLMs. Furthermore, we developed a simple and scalable zero-perception watermarking scheme, which boils down to the design of an auxiliary channel between a random partition and shared side information. We studied the proposed scheme, showing optimality and characterizing the role of partition randomization. We then compared the CC scheme to existing schemes and the fundamental trade-off curve. Due to the simplicity and mathematical grounding of the CC scheme, we believe it can be further extended to accommodate additional facets of token watermarking, such as robustness. For future work, we plan to implement the scheme to text-sequence watermarks and extend the CC method to the positive perception regime.

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] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Niels Provos and Peter Honeyman. Hide and seek: An introduction to steganography. *IEEE security & privacy*, 1(3):32–44, 2003.

[27] Mohammed Abdul Majeed, Rossilawati Sulaiman, Zarina Shukur, and Mohammad Kamrul Hasan. A review on text steganography techniques. *Mathematics*, 9(21):2829, 2021.

[28] Jiaming Shen, Heng Ji, and Jiawei Han. Near-imperceptible neural linguistic steganography via self-adjusting arithmetic coding. *arXiv preprint arXiv:2010.00677*, 2020.

[29] Yu-Shin Huang, Peter Just, Krishna Narayanan, and Chao Tian. Od-stega: Llm-based near-imperceptible steganography via optimized distributions. *arXiv preprint arXiv:2410.04328*, 2024.

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

[31] 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.

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

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

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

[35] 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.

[36] 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.

[37] 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.

[38] Jiacheng Liang, Zian Wang, Lauren Hong, Shouling Ji, and Ting Wang. Waterpark: A robustness assessment of language model watermarking. *arXiv preprint arXiv:2411.13425*, 2024.

[39] 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.

[40] 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.

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

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

[43] 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.

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

[45] Yury Polyanskiy. *Channel coding: Non-asymptotic fundamental limits*. Princeton University, 2010.
[46] 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.
[47] Jingbo Liu, Paul Cuff, and Sergio Verdú. e_γ-resolvability. *IEEE Transactions on Information Theory*, 63(5):2629–2658, 2016.
[48] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2022.
[49] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024.
[50] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

APPENDIX

In this appendix, we include detailed proofs of our theoretical results, which are presented in the main body of the paper.

A. 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:

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

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 $A = \{a|L(a, s) \geq \gamma\}$. □

B. 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 $E_\gamma(\widetilde{Q}_{X|S}\|Q_X|P_S)$. 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 $\widetilde{Q}_X = Q_X$, 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:

$$\max_{p_1,...,p_m\in\Delta_d}\frac{1}{m}\sum_{i=1}^m\left\|p_i - \frac{\gamma}{m}\sum_{i=1}^m p_i\right\|_+, \tag{16}$$

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

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

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 (17) and by the definition of the indicator. Out optimization problem had therefore boiled down to maximizing on the quantity

$$\sum_{j=1}^{d} q_j (1 - \gamma q_j)_+ \text{ such that } q_j = k/m, k \in \mathbb{Z}, \sum_{j=1}^{d} q_j = 1. \tag{18}$$

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

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

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

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

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 $(\frac{k}{m})_{k=0}^{m}$, 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

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

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 $\square$

C. Proof of Theorem 2

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

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

Therefore,

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

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

D. Proof of Lemma 1

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. □

E. Proof of Proposition 2

By the dual representation of the total variation

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

Given $P_S \sim \mathsf{Unif}[k]$ and $P_{\widetilde{Y}} = \{p_1, ..., p_k\} \in \Delta_k$. For the proposed coupling, we show that the coupling achieves $\mathsf{TV}(P_S, P_{\widetilde{Y}})$.

$$\begin{aligned}
\mathbf{P}(\widetilde{Y} \ne S) &= 1 - \mathbf{P}(\widetilde{Y} = S) \\
&= 1 - \sum_{i=1}^{m} \min(\frac{1}{m}, p_i) \\
&= \mathsf{TV}(P_S, P_{\widetilde{Y}})
\end{aligned}$$

F. 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 (7), 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:

$$\frac{Q_X(X)}{\widetilde{Q}_{X|B^m,S}(X)} = \frac{Q(X)}{Q_X(X)\frac{1}{P_S(S)}P_{S|\widetilde{Y}}(S|f(X,B^m)} \tag{20}$$

$$= \frac{2}{P_{S|\widetilde{Y}}(S|f(X,B^m)} \tag{21}$$

$$< 1, \tag{22}$$

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.

G. 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

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

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 $R_d = \frac{1}{2}(1 + \min(\widetilde{p}_0, \widetilde{p}_1))$. □

H. Proof of Theorem 3

We begin by proving Lemma 2.

1) Proof of Lemma 2: 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:

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

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

The function

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

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

Next, we proceed with the proof of Theorem 3.

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

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

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

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

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 $\{\mathbf{q} \in \Delta_m \mid \|\mathbf{q}\|_\infty \le \lambda\}$ 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

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

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 $P_\mathcal{B}$

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

We need to select the set $\mathcal{B}$ that maximizes $\Pr(B_1 \ne B_2)$. 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 $\Pr(B_1 = B_2) = \frac{m-2}{2(m-1)}$, $\Pr(B_1 \ne B_2) = \frac{m}{2(m-1)}$, which simplifies $R_d(\lambda)^*$ to

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

I. 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 [50]:

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

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 [50]

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

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). □