# On Random Coding Error Exponents of Watermarking Codes

Neri Merhav*

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL
merhav@ee.technion.ac.il

**Abstract**

Watermarking codes are analyzed from an information-theoretic viewpoint as a game between an information hider and an active attacker. While the information hider embeds a secret message (watermark) in a covertext message (typically, text, image, sound, or video stream) within a certain distortion level, the attacker processes the resulting watermarked message, within limited additional distortion, in attempt to invalidate the watermark. For the case where the covertext source is memoryless (or, more generally, where there exists some transformation that makes it memoryless), we provide a single-letter characterization of the maximin game of the random coding error exponent associated with the average probability of erroneously decoding the watermark. This single-letter characterization is in effect because if the information hider utilizes a memoryless channel to generate random codewords for every covertext message, the (causal) attacker will maximize the damage by implementing a memoryless channel as well. Partial results for the dual minimax game and the conditions for the existence of a saddle point, are also presented.

**Index Terms:** Steganography, watermarking, information hiding, error exponent, random coding.

---

*Also with Hewlett-Packard Laboratories – Israel. Part of this work was done while the author was visiting Hewlett–Packard Laboratories, California, U.S.A.

# 1 Introduction

The fast development of digital information technology, combined with the simplicity of duplication and distribution of digital data across communication networks like the Internet (using, e.g., publicly accessible web sites), has exposed content providers to a real challenge of how to protect their electronic data (e.g., images, video, sound, text). This has stimulated many research efforts towards the design and study of sophisticated watermarking and information hiding methodologies for copyright protection (see, e.g., [1],[2],[4] and references therein). Watermarking techniques are about embedding secret information into, say, an image, that can neither be removed nor deciphered without access to some secret key. On the other hand, to maintain quality, the watermarked image should be perceptually identical or with nearly invisible difference from the original image. The main application is for proving ownership of the data and for protection against forgers. It is therefore crucial that no malicious attacker could remove the watermark without damaging the image to the degree that it becomes useless. A sophisticated attacker would then do his best to invalidate the watermark at the minimum possible cost of additional degradation in image quality. Watermarking techniques must, of course, be robust not only to hostile attacks, but also to standard operations performed on digital data, like printing, lossy compression (e.g., JPEG or MPEG), filtering, halftoning, and so on. The tradeoffs to be considered in the design of watermarking techniques are, therefore, among three principal parameters: (i) robustness to attacks and standard signal processing, (ii) high quality, and (iii) the relative amount of secret information to be embedded (in bits per pixel).

In contrast to the vast amount of research work reported in the signal and image processing literature, relatively little attention has been devoted to this problem from the information-theoretic perspective. An exception is a recent work by O'Sullivan, Moulin, and Ettinger [16], who characterized the highest achievable information rate, that is, the capacity of watermarking systems for the following game–theoretic model: A secret message, encoded at rate $R$, is hidden in a memoryless[1] covertext message (an image) within small degradation of quality, symbolized by distortion level $D_1$ with respect to (w.r.t.) some distortion measure. An active attacker may introduce additional distortion $D_2$ (relative to possibly another distortion measure) in attempt to disrupt the watermark. Finally, the resulting data set is decoded using information shared with the information hider (i.e., the original covertext message) to extract the watermark. In [16], the highest information rate $R$ has been found (as a function of $D_1$ and $D_2$) for which, even in the presence of optimum attack, there exist block encoders and decoders that guarantee arbitrarily small probability of erroneous watermark decoding. This critical rate is referred to as the maximin capacity. For example, when both distortion measures are the squared error, the maximin (and minimax) capacity is well-known to be the capacity of the Gaussian channel, $C = \frac{1}{2}\log(1 + D_1/D_2)$, attained by the saddle point where both parties implement Gaussian strategies with the maximum allowable distortion (see, also [17]).

In this paper, we go one step further and investigate error exponents. Our main result is a single-letter characterization of the maximin random coding error exponent under this model. It is heavily based on a fact, proved in Section 3, that for a memoryless covertext source and a given memoryless random coding distribution, an optimum causal[2] attack strategy subject to the above-mentioned distortion constraint, is to implement a *memoryless channel* on the watermarked data

[1] While the memoryless source model is rarely realistic when attributed to the raw data, it is fairly plausible when applied to certain domains of decorrelating transforms, e.g., the discrete Fourier transform, discrete cosine transform, or KLT, which are often used for preprocessing.

[2] The causality assumption, which corresponds to on–line attack, is used also in the jamming game [19]. In watermarking applications it might be reasonable in the presence of limitations on complexity and delay, or when applied to the superalphabet of blocks.

set. It is then clear that this memoryless channel is the one that minimizes the random coding error exponent subject to the single-letter distortion constraint. This maximin criterion corresponds to the common assumption (used also in cryptography) that the attacker knows the strategy used by the information hider, in our case, the random coding distribution. Generally speaking, a more plausible assumption would have been that the attacker knows moreover the specific watermark code in use. Nonetheless, knowledge of the random coding distribution but not the code itself, corresponds to a situation where the random choice of the code is controlled by a secret key shared by the encoder and decoder, but not the attacker.

We also show in Section 4, that under a certain condition (that holds at least for the Hamming distortion measure), the dual property holds as well: If the attack strategy is given by a memoryless channel, then the optimal random coding distribution is memoryless as well. It turns out then that, in this case, there is a minimax–maximin saddle point in the game between the information hider and the attacker, where in the presence of a memoryless covertext source, both parties implement memoryless channels.

Our results can also be viewed as a generalization of those of Stiglitz [18],[19] regarding error exponents for the jamming game. In Section 3, after presenting our main result, we will also explain in what way it is stronger and more general.

## 2 Notation, Problem Formulation, and Preliminaries

Throughout this paper, we adopt the following notation conventions.

Scalar random variables will be denoted by capital letters (e.g., $X$) and specific values they may take will be denoted by the corresponding lower case letters (e.g., $x$). All scalar random variables in this paper are assumed to take on values in the same finite alphabet $\mathcal{A}$. Similarly, random vectors of length $n$ ($n$ – positive integer) will be denoted by boldface capital letters (e.g., $\boldsymbol{X} = (X_1, \ldots, X_n)$), and specific values that they take will be denoted by the respective boldface lower case letters (e.g., $\boldsymbol{x} = (x_1, \ldots, x_n)$), where it is understood that the alphabet of each such vector is always $\mathcal{A}^n$, the $n$-fold Cartesian product of $\mathcal{A}$. When the dimension of a vector is different from $n$, a superscript will denote the actual dimension (e.g., $\boldsymbol{x}^{n-1} = (x_1, \ldots, x_{n-1})$). Thus $\boldsymbol{x}$ is shorthand notation for $\boldsymbol{x}^n$. Also, dimension zero ($\boldsymbol{x}^0$) corresponds to the null string.

Probability mass functions (PMF's) of single letters will also be denoted by capital letters with probabilities of specific letters denoted by the respective lower case letters (e.g., $Q = \{q(x),\ x \in \mathcal{A}\}$). Similarly, vector sources, or joint PMF's of $n$-vectors, will be denoted by boldface capital letters with probabilities of specific vector values denoted by the corresponding boldface lower case letters (e.g., $\boldsymbol{Q} = \{\boldsymbol{q}(\boldsymbol{x}),\ \boldsymbol{x} \in \mathcal{A}^n\}$). A source $\boldsymbol{Q}$ over $\mathcal{A}^n$ is said to assume a *product form* if $\boldsymbol{Q}(\boldsymbol{x}) = \prod_{i=1}^n q_i(x_i)$ for all $\boldsymbol{x} \in \mathcal{A}^n$. For shorthand notation, this fact will be denoted by $\boldsymbol{Q} = Q_1 \times \ldots \times Q_n$. If, in addition, $Q_i = Q$ for all $i$, then $Q \times \ldots \times Q$ will be denoted by $Q^n$ and $Q$ will be referred to as the *single-letter marginal* of $\boldsymbol{Q}$.

Similar conventions apply to conditional PMF's of channels: Single-letter conditional PMF's will be denoted by capital letters that symbolize matrices of conditional probabilities (e.g., $W = \{w(y|x),\ x, y \in \mathcal{A}\}$) and vector channels of dimension $n$ will be denoted by the respective boldface letters (e.g., $\boldsymbol{W} = \{\boldsymbol{w}(\boldsymbol{y}|\boldsymbol{x}),\ \boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}^n\}$. If the dimension is different from $n$, then the symbol that designates the channel will be superscripted by the actual dimension in parentheses (e.g., $\boldsymbol{W}^{(n-1)}$ is a vector channel of dimension $n-1$). A channel $\boldsymbol{W}$ is said to admit a product form if $\boldsymbol{w}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^n w_i(y_i|x_i)$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}^n$. For shorthand notation, this fact will be denoted by

$\boldsymbol{W} = W_1 \times \ldots \times W_n$. If, in addition, $W_i = W$ for all $i$, then $W \times \ldots \times W$ will be denoted by $W^n$ and $W$ will be referred to as the single-letter marginal of $\boldsymbol{W}$.

The probability of an event $\mathcal{E}$ will be denoted by $\Pr\{\mathcal{E}\}$ whenever the underlying probability measure is clear from the context, and the expectation operator will be denoted by $\mathbf{E}\{\cdot\}$. Summations and products of indexed terms over the entire index set will sometimes be subscripted only by the symbol of that index. For example, $\sum_x f(x)$ will mean that the summation is over the entire alphabet $\mathcal{A}$, $\prod_i q(x_i)$ means that the product is taken from $i = 1$ to $n$, etc.

Let $\boldsymbol{P} = P^n$ designate a stationary, memoryless vector source generating random vectors $\boldsymbol{U} = (U_1, \ldots, U_n)$ over $\mathcal{A}^n$. The data vector $\boldsymbol{U}$ will designate the *covertext* message within which the watermark will be hidden. Let $d_1 : \mathcal{A} \times \mathcal{A} \to \mathbb{R}_+$ denote a single-letter distortion measure.

A rate $R$ *watermarking code* of size $n$, subject to distortion level $D_1$, is a triple $(\mathcal{M}_n, f_n, g_n)$ with the following ingredients. The set $\mathcal{M}_n$ is the watermarking message set whose size is $\lfloor 2^{nR} \rfloor$. The function $f_n : \mathcal{A}^n \times \mathcal{M}_n \to \mathcal{A}^n$ is the encoder that maps every combination of a covertext data set $\boldsymbol{u} \in \mathcal{A}^n$ and a *watermark message* $m \in \mathcal{M}_n$ into a *stegotext* message (or, watermarked message) $\boldsymbol{x} \in \mathcal{A}^n$ such that $\mathbf{E} d_1(u_i, X_i) \leq D_1$ for all $\boldsymbol{u} \in \mathcal{A}^n$, $1 \leq i \leq n$, that is, the conditional expected distortion in each coordinate given $\boldsymbol{U} = \boldsymbol{u}$ never exceeds $D_1$ for all $\boldsymbol{u}$.[3] The function $g_n : \mathcal{A}^n \times \mathcal{A}^n \to \mathcal{M}_n$ is the decoder, which maps the original covertext message (shared with the encoder) together with a stegotext message (possibly modified or corrupted by an attacker) back to a watermark message.

Let $d_2 : \mathcal{A} \times \mathcal{A} \to \mathbb{R}_+$ denote a single-letter distortion measure. A causal *attack* subject to distortion $D_2$ is a (possibly randomized) map $h_n : \mathcal{A}^n \to \mathcal{A}^n$ where

[3]Note that if the stochastic map from $\boldsymbol{u}$ to $\boldsymbol{x}$ is thought of as a stationary channel and $d_1$ is a difference distortion measure, i.e., $d_1(u, x) = d(x - u)$, all these constraints boil down to a single constraint $\mathbf{E} d(V) \leq D_1$, where $V = X - U$ designates the additive noise.

(i) Each component $y_i$ of the output depends on $(x_1, \ldots, x_i)$ and $(y_1, \ldots, y_{i-1})$, but not on $(x_{i+1}, \ldots, x_n)$, $i = 1, \ldots, n$. (ii) $\frac{1}{n}\sum_{i=1}^n \delta_i \le D_2$, where $\delta_i \triangleq \max\{\mathbf{E}\{d_2(x_i, Y_i)|\boldsymbol{x}^i, \boldsymbol{y}^{i-1}\}|\ \boldsymbol{x}^i \in \mathcal{A}^i,\ \boldsymbol{y}^{i-1} \in \mathcal{A}^{i-1}\}$.

Given distortion levels $D_1$ and $D_2$ w.r.t. distortion measures $d_1$ and $d_2$, respectively, the error probability is defined as

$$P_e(f_n, g_n, h_n) = \Pr\{g_n(\boldsymbol{U}, h_n(f_n(\boldsymbol{U}, M))) \neq M\}, \tag{1}$$

where $M$ designates a random variable in $\mathcal{M}_n$ with a uniform PMF independently of $\boldsymbol{U}$.

An *achievable rate* $R$ is one for which there exists a sequence of rate $R$ watermarking codes $\{(\mathcal{M}_n, f_n, g_n)\}_{n\ge 1}$ subject to distortion level $D_1$ such that

$$\lim_{n\to\infty} \sup_{h_n} P_e(f_n, g_n, h_n) = 0, \tag{2}$$

where the supremum is over all causal attack strategies subject to distortion $D_2$. The capacity of the system, $C = C(D_1, D_2)$, is the supremum of all achievable rates.

Similarly as in [16],[4] it can be shown that for a memoryless covertext source $\boldsymbol{P} = P^n$, the capacity is given by

$$C(D_1, D_2) = \max\min I(X; Y|U) \tag{3}$$

where (i) $U \ominus X \ominus Y$ is assumed a Markov chain, (ii) the marginal PMF of $U$ is given by $P$, (iii) the minimum is over all $W = \{w(y|x)\}$ such that $\max_x \mathbf{E} d_2(x, Y) \le D_2$, and (iv) the maximum is over all $Q = \{q(x|u)\}$ such that $\max_u \mathbf{E} d_1(u, X) \le D_1$.

In this paper, we refine the analysis of the probability of error at a given information rate $R < C(D_1, D_2)$. To this end, we assume a random coding regime and we focus on the average probability of error w.r.t. the ensemble of randomly chosen codes given $\boldsymbol{U}$. To meet the $D_1$ distortion constraint,

[4] The above-described model assumptions are somewhat different from those of [16].

we adopt the following random coding mechanism: For each $\boldsymbol{u} \in \mathcal{A}^n$, we generate independently, at random $\lfloor 2^{nR} \rfloor$ codewords by using a certain channel $\boldsymbol{Q} = \{\boldsymbol{q}(\boldsymbol{x}|\boldsymbol{u}),\ \boldsymbol{u}, \boldsymbol{x} \in \mathcal{A}^n\} \in \mathcal{Q}_n(D_1)$ fed by $\boldsymbol{u}$, where

$$\mathcal{Q}_n(D_1) \triangleq \{\boldsymbol{Q} :\ \sum_{x_i} q(x_i|\boldsymbol{x}^{i-1}, \boldsymbol{u}) d_1(u_i, x_i) \leq D_1, \quad \forall\, i, \boldsymbol{x}^{i-1}, \boldsymbol{u}\}.$$

The channel $\boldsymbol{Q}$ will be referred to as the *watermarking channel* [16]. For $n = 1$, $\mathcal{Q}_n(D_1) = \mathcal{Q}_1(D_1)$, which is a set of single-letter channels $Q$, will be abbreviated by $\mathcal{Q}(D_1)$.

Since the attacker is allowed to use a randomized map, we will symbolize the attack strategy by a another channel $\boldsymbol{W} = \{\boldsymbol{w}(\boldsymbol{y}|\boldsymbol{x}),\ \boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}^n\}$, henceforth referred to as the *attack channel* [16]. The fact that the attack channel is causal means that

$$\boldsymbol{W}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} w(y_i|\boldsymbol{y}^{i-1}\boldsymbol{x}^i), \qquad \forall\, \boldsymbol{x}, \boldsymbol{y}. \tag{4}$$

The fact that the attack is subjected to distortion constraint $D_2$ means that $\boldsymbol{W}$ lies in the set

$$\mathcal{W}_n(D_2) = \{\boldsymbol{W} :\ \frac{1}{n}\sum_{i=1}^{n} \delta_i \leq D_2\},$$

where as defined above,

$$\delta_i = \max_{\boldsymbol{x}^i, \boldsymbol{y}^{i-1}} \sum_{y_i} w(y_i|\boldsymbol{y}^{i-1}\boldsymbol{x}^i) d_2(x_i, y_i). \tag{5}$$

For $n = 1$, $\mathcal{W}_n(D_2) = \mathcal{W}_1(D_2)$, which is a set of single-letter channels $W$, will be abbreviated by $\mathcal{W}(D_2)$.

The average error probability $\bar{P}_e(\boldsymbol{Q}, \boldsymbol{W})$ is defined as the expectation of $P_e$ w.r.t. the joint ensemble of random codes drawn by $\boldsymbol{Q}$ and the randomized attacks governed by $\boldsymbol{W}$.

We will be interested primarily in the quantity

$$\min_{\boldsymbol{Q} \in \mathcal{Q}_n(D_1)} \max_{\boldsymbol{W} \in \mathcal{W}_n(D_2)} \bar{P}_e(\boldsymbol{Q}, \boldsymbol{W})$$

and in characterizing the pairs $(\boldsymbol{Q}, \boldsymbol{W})$ that achieve this minimax.[5] For completeness, we will also consider the dual problem

$$\max_{\boldsymbol{W} \in \mathcal{W}_n(D_2)} \min_{\boldsymbol{Q} \in \mathcal{Q}_n(D_1)} \bar{P}_e(\boldsymbol{Q}, \boldsymbol{W})$$

as well as on conditions for equivalence with the minimax problem and existence of a saddle point. Since an exact expression of $\bar{P}_e(\boldsymbol{Q}, \boldsymbol{W})$ is unavailable, we will focus on a Gallager–type upper bound [9] to $\bar{P}_e(\boldsymbol{Q}, \boldsymbol{W})$. This is also motivated by the fact that the random–coding error exponent is not merely an upper bound, but it is also tight in that it provides the exact exponential behavior of average error probability in (block) memoryless systems [10].

# 3 The Worst Causal Attack Channel is Memoryless

In this section, we confine attention to the minimax criterion, corresponding to the assumption that the attacker has perfect knowledge of the watermarking channel, which is in turn assumed here to admit a product form $\boldsymbol{Q} = Q^n$.

The first observation is that the problem at hand is no different from an ordinary channel coding problem with side information at both encoder and decoder: A message $m \in \mathcal{M}_n$ of length $nR$ bits is mapped to a codeword $\boldsymbol{x} \in \mathcal{A}^n$ depending on the side information $\boldsymbol{u} \in \mathcal{A}^n$. (In other words, there is a different codebook for every $\boldsymbol{u}$.) Then, $\boldsymbol{x}$ is transmitted across a channel $\boldsymbol{W}$, whose output $\boldsymbol{y}$ is processed in the presence of $\boldsymbol{u}$ to decode the message. Note that this is not as general as the ordinary definition of the problem of communication with side information, because in the more general setting, the channel is also allowed to depend on the side information.

Assuming the random coding regime described in Section 2 and maximum likelihood decoding,

[5] Since the probability of error is a decreasing function of the error probability, the minimax error probability is equivalent to the maximin error exponent and vice versa.

the following upper bound on the average error probability is easily derived as a straightforward extension of [9, Theorem 5.6.1, p. 135] to the case where side information is available at both encoder and decoder:

$$\bar{P}_e(\boldsymbol{Q}, \boldsymbol{W}) \le B(\rho, \boldsymbol{Q}, \boldsymbol{W}) \stackrel{\triangle}{=} 2^{\rho n R} F(\rho, \boldsymbol{Q}, \boldsymbol{W}), \qquad 0 \le \rho \le 1 \tag{6}$$

where

$$F(\rho, \boldsymbol{Q}, \boldsymbol{W}) \stackrel{\triangle}{=} \sum_{\boldsymbol{u}} \boldsymbol{p}(\boldsymbol{u}) \sum_{\boldsymbol{y}} \left[ \sum_{\boldsymbol{x}} \boldsymbol{q}(\boldsymbol{x}|\boldsymbol{u}) \boldsymbol{w}(\boldsymbol{y}|\boldsymbol{x})^{1/(1+\rho)} \right]^{1+\rho} \tag{7}$$

with $\boldsymbol{P} = \{\boldsymbol{p}(\boldsymbol{u}),\ \boldsymbol{u} \in \mathcal{A}^n\}$ being a given memoryless source $P^n$, $\boldsymbol{Q} = Q^n \in \mathcal{Q}_n(D_2)$, and $\boldsymbol{W} \in \mathcal{W}_n(D_2)$.

The objective function hereafter will be the upper bound on the average error probability $B(\rho, \boldsymbol{Q}, \boldsymbol{W})$. This function involves the auxiliary parameter $\rho$, which can be chosen so as to obtain the tightest upper bound, namely, $\min_{0 \le \rho \le 1} B(\rho, \boldsymbol{Q}, \boldsymbol{W})$. Thus, for a memoryless watermarking channel $\boldsymbol{Q} = Q^n$, we will be interested in characterizing the minimax game

$$\min_{Q \in \mathcal{Q}(D_1)} \max_{\boldsymbol{W} \in \mathcal{W}_n(D_2)} \min_{0 \le \rho \le 1} B(\rho, Q^n, \boldsymbol{W}). \tag{8}$$

For later use, we also define $F(\rho, Q, W)$ and $B(\rho, Q, W)$ as the same functionals as above, applied to the case $n = 1$, that is:

$$B(\rho, Q, W) = 2^{\rho R} F(\rho, Q, W) = 2^{\rho R} \sum_{u} p(u) \sum_{y} \left[ \sum_{x} q(x|u) w(y|x)^{1/(1+\rho)} \right]^{1+\rho}. \tag{9}$$

Note that for $\boldsymbol{P} = P^n$, $B(\rho, Q^n, W^n) = B(\rho, Q, W)^n$ and $F(\rho, Q^n, W^n) = F(\rho, Q, W)^n$. We also define the random coding error exponent as

$$E_r(\rho, Q, W) \stackrel{\triangle}{=} -\log B(\rho, Q, W) = -\log F(\rho, Q, W) - \rho R. \tag{10}$$

Our first result states that for a memoryless covertext source and a memoryless random coding distribution, the worst channel $\boldsymbol{W}$, in the sense of maximizing $B$ subject a distortion constraint, is memoryless as well. Since $B$ and $F$ are the same within a factor that does not depend on $\boldsymbol{Q}$ and $\boldsymbol{W}$, we can state this result in terms of $F$.

**Theorem 1** *Given $\rho$, $\boldsymbol{P} = P^n$, $Q$, and distortion level $D_2$ w.r.t. distortion measure $d_2$, it is a stationary memoryless channel $\boldsymbol{W}_* = W_*^n$ that attains the maximum of $F(\rho, Q^n, \boldsymbol{W})$ w.r.t. $\boldsymbol{W}$ subject to the constraints*

$$
\begin{aligned}
\boldsymbol{w}(\boldsymbol{y}|\boldsymbol{x}) &\geq 0 \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}^n && (11)\\
\sum_{\boldsymbol{y}} \boldsymbol{w}(\boldsymbol{y}|\boldsymbol{x}) &= 1 \qquad \forall \boldsymbol{x} \in \mathcal{A}^n && (12)\\
\frac{1}{n}\sum_{i=1}^{n} \delta_i &\leq D_2 && (13)
\end{aligned}
$$

*where $\delta_i$ is defined as in eq. (5) The single-letter marginal $W_*$ of $\boldsymbol{W}_*$ is given by the maximizer of $F(\rho, Q, W)$ subject to the same constraints applied to the case $n = 1$, i.e.,*

$$
\max_{\boldsymbol{W} \in \mathcal{W}_n(D_2)} F(\rho, Q^n, \boldsymbol{W}) = \left[\max_{W \in \mathcal{W}(D_2)} F(\rho, Q, W)\right]^n . \tag{14}
$$

The proof appears in Section 6. Theorem 1 leads to our main result, which is the following.

**Theorem 2** *For a memoryless source $\boldsymbol{P} = P^n$,*

$$
\min_{Q \in \mathcal{Q}(D_1)} \max_{\boldsymbol{W} \in \mathcal{W}_n(D_2)} \min_{0 \leq \rho \leq 1} B(\rho, Q^n, \boldsymbol{W}) = 2^{-nE(R, D_1, D_2)}, \tag{15}
$$

*where*

$$
\begin{aligned}
E(R, D_1, D_2) &= \max_{Q \in \mathcal{Q}(D_1)} \min_{W \in \mathcal{W}(D_2)} \max_{0 \leq \rho \leq 1} E_r(\rho, Q, W)\\
&= \max_{0 \leq \rho \leq 1} \max_{Q \in \mathcal{Q}(D_1)} \min_{W \in \mathcal{W}(D_2)} E_r(\rho, Q, W). && (16)
\end{aligned}
$$

**Discussion:** Before proving Theorem 2, several comments are in order.

First, observe that the two different expressions of $E(R, D_1, D_2)$ in eq. (16) indicate that the order of optimizations of $E_r(\rho, Q, W)$ is interchangeable. The second expression in (16) has the convenience that for every fixed $\rho$, the functional $B(\rho, Q, W) = 2^{-E_r(\rho,Q,W)}$ as we shall see, is concave in $W$ and convex in $Q$.

As mentioned in the Introduction, Theorem 2 can also be viewed as an extension of the results of Stiglitz [18], [19]. In [18], Stiglitz considered the maximin error exponent for a game where the communicator picks the best (memoryless, time–invariant) random coding distribution and the best additive decoding metric for the worst arbitrarily–varying memoryless channel $\boldsymbol{W}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^n W_i(y_i|x_i)$, with each $W_i$ being chosen from some fixed convex set of channels. The fact that the channel is assumed memoryless means that the choice of $W_i$ does not depend on past events. The main result in [18] is that the worst such channel is time-invariant and so, the maximin exponent is given by a simple single–letter expression given by the maximin value of Gallager's random coding error exponent (see also Ericson [6], Geraniotis [11], Hughes and Thomas [13]). In [19], the same result is shown to continue to hold even if the jammer is allowed, more generally, to pick, at each time instant $i$, a causal channel $w(\cdot|\boldsymbol{y}^{i-1}, \boldsymbol{x}^i)$ within the same convex set, based on past inputs, past outputs, and past choices of the channel. Theorem 2 extends and improves these results in several aspects:

- It incorporates side information. It is easy to show that our results extend to the aforementioned, more general setting of communication in the presence of side information, where the channel is allowed to depend non–causally on the side information. Specifically, among all channels of the form $\boldsymbol{W}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{u}) = \prod_i w(y_i|\boldsymbol{y}^{i-1}, \boldsymbol{x}^i, \boldsymbol{u})$, the one that maximizes $F$ subject to the distortion constraint (which now must hold also for every $\boldsymbol{u}$), is memoryless, i.e., of the form $\prod_i w(y_i|x_i, u_i)$. In this application, the distortion constraints can be thought of as power constraints.

- In [19] the optimization of the exponent w.r.t. $\rho$ is the outer–most, whereas Theorem 2 refers also to the expression where this optimization is the inner–most. The latter is, of course, stronger and more relevant since $\rho$ is only a parameter of the bound.

- While in [19] only the maximin exponent was considered, here we give some results on the minimax exponent as well (see next section).

- As indicated above, there are two basic assumptions in [19]: (i) The decoding metric is additive, i.e., it is of the form $m(\boldsymbol{x}, \boldsymbol{y}) = \sum_i m(x_i, y_i)$, and (ii) This metric is chosen so as to minimize a (Gallager–type) upper bound to the maximum average probability of error over all allowable channels. Actually, both assumptions are unrealistically pessimistic. The first assumption makes sense only if one knows a-priori that the jammer (or the attacker) would apply a memoryless channel. In a way, this implicitly assumes already that the worst allowable channel is memoryless. But even if the first assumption is accepted, the second assumption is still pessimistic because in many cases the channel could be estimated from the observed output and/or from training data if available. Moreover, it is well known to date that for certain classes of channels (see, e.g., [5],[7], [12],[14],[15], [21]) there exist *universal decoders* that are independent of the unknown channel and yet attain the random coding exponent as the maximum likelihood decoder. This suggests that the random coding error exponent considered here, which corresponds to (asymptotically) optimum decoding for the actual channel, is a more relevant objective function than the one corresponding to the upper bound of [19]. On the other hand, it should be emphasized that the elegant proof in [19] (for the fact that the worst channel is memoryless) does not seem to have a natural parallel when it comes to the random coding error exponent, and so, more powerful techniques of convex analysis are required.

*Proof of Theorem 2.* For a given $Q \in \mathcal{Q}(D_1)$, consider the expression

$$\max_{\boldsymbol{W} \in \mathcal{W}_n(D_2)} \min_{0 \le \rho \le 1} B(\rho, Q^n, \boldsymbol{W}) = \exp_2 \left\{ \max_{\boldsymbol{W} \in \mathcal{W}_n(D_2)} \min_{0 \le \rho \le 1} [\log F(\rho, Q^n, \boldsymbol{W}) + \rho n R] \right\}. \tag{17}$$

Since $F(\rho, Q^n, \boldsymbol{W})$ is concave in $\boldsymbol{W}$ (see Section 6, Lemma 1) and the logarithmic function is monotonic and concave, then $\log F(\rho, Q^n, \boldsymbol{W})$ is concave as well. Also, a straightforward extension of [8, p. 17, Proof of Theorem 2] to the case of side information, shows that $\log F(\rho, Q^n, \boldsymbol{W})$, and hence also $[\log F(\rho, Q^n, \boldsymbol{W}) + \rho n R]$ is convex in $\rho$. Also, $\mathcal{W}_n(D_2)$ and the interval $[0, 1]$ are both compact convex sets. Therefore, by the minimax theorem [20, p. 232, Theorem (6.3.7)], the maximization and the minimization on the r.h.s. of eq. (17) are interchangeable, and so, it can be written also as

$$\begin{aligned} \exp_2 \left\{ \min_{0 \le \rho \le 1} \max_{\boldsymbol{W} \in \mathcal{W}_n(D_2)} [\log F(\rho, Q^n, \boldsymbol{W}) + \rho n R] \right\} &= \min_{0 \le \rho \le 1} \left[ 2^{\rho n R} \max_{\boldsymbol{W} \in \mathcal{W}_n(D_2)} F(\rho, Q^n, \boldsymbol{W}) \right] \\ &= \left[ \min_{0 \le \rho \le 1} 2^{\rho R} \max_{W \in \mathcal{W}(D_2)} F(\rho, Q, W) \right]^n \end{aligned} \tag{18}$$

where the second equality follows from Theorem 1. The assertion of the theorem now follows by taking the minimum over $Q \in \mathcal{Q}(D_1)$. □

# 4 When Is the Best Watermarking Channel Memoryless?

In duality to the minimax problem of $\min_\rho B(\rho, \boldsymbol{Q}, \boldsymbol{W})$ that was studied in Section 3, consider now the maximin problem. In other words, we now ask what is the best watermarking channel for the worst memoryless attack channel. Seemingly, this problem is less relevant from the engineering point of view, because as mentioned earlier, the usual working assumption is that the attacker knows the strategy of the information hider and not vice versa. Nonetheless, this dual maximin problem is interesting because it provides us with the viewpoint of the attacker as well, and thereby completes the whole picture regarding the game between the information hider and the attacker. Moreover, if there is a saddle point between the two parties of this game, and hence the minimax and the maximin coincide, this means that there is a point of equilibrium in the sense that no party would have incentive to change his strategy unilaterally.

Unlike in the minimax problem, however, here we have only partial results. In duality to Theorem 1, here too we can prove that for a given $\rho$ and a given memoryless attack channel, the best watermarking channel is memoryless as well. Thus, we can provide a single–letter characterization of $\inf_{0\leq\rho\leq 1}\max_{W\in\mathcal{W}(D_2)}\min_{\boldsymbol{Q}\in\mathcal{Q}_n(D_1)} B(\rho,\boldsymbol{Q},W^n)$. Unfortunately, it is not clear that this, in general, is equivalent to $\max_{W\in\mathcal{W}(D_2)}\min_{\boldsymbol{Q}\in\mathcal{Q}_n(D_1)}\min_{0\leq\rho\leq 1} B(\rho,\boldsymbol{Q},W^n)$, which is, in fact, the relevant quantity of interest for the maximin problem.[6]

We first present the aforementioned dual to Theorem 1 and then give a simple sufficient condition for the existence of a saddle-point of $\min_{0\leq\rho\leq 1} B(\rho,\boldsymbol{Q},\boldsymbol{W})$, which leads to equivalence between the minimax and the maximin problems. In this saddle point, both $\boldsymbol{Q}$ and $\boldsymbol{W}$ are memoryless channels. As will be shown in the next section, the above-mentioned sufficient condition holds at least for the important special case where $\boldsymbol{P}$ is uniform and both $d_1$ and $d_2$ are the Hamming distortion measure. In this case, there is also a closed–form expression for the error exponent up to the step of maximizing over $\rho$. We leave open the question whether it is possible to relax this sufficient condition and still prove that $\min_\rho B(\rho,\boldsymbol{Q},\boldsymbol{W})$ has a saddle point given by a pair of memoryless

[6] The proof of this conjecture may require either proving first that $\min_{0\leq\rho\leq 1} B(\rho,Q,W)$ is convex in $Q$, or alternatively, that $B(\rho,Q,W)$ is convex in the pair $(\rho,Q)$. (As mentioned earlier, it is known that $B$ is convex in $\rho$ and in $Q$ separately.)

channels.

**Theorem 3** *Given $\rho$, $\boldsymbol{P} = P^n$, $W$, and a distortion level $D_1$ w.r.t. distortion measure $d_1$, it is a stationary memoryless channel $\boldsymbol{Q}_* = Q_*^n$ that attains the minimum of $F(\rho, \boldsymbol{Q}, W^n)$ w.r.t. $\boldsymbol{Q}$ subject to the constraints*

$$\boldsymbol{q}(\boldsymbol{x}|\boldsymbol{u}) \geq 0 \qquad \forall \boldsymbol{u}, \boldsymbol{x} \in \mathcal{A}^n \tag{19}$$

$$\sum_{\boldsymbol{x}} \boldsymbol{q}(\boldsymbol{x}|\boldsymbol{u}) = 1 \qquad \forall \boldsymbol{u} \in \mathcal{A}^n \tag{20}$$

$$\sum_{x_i} \boldsymbol{q}(x_i|\boldsymbol{x}^{i-1}, \boldsymbol{u}) d_1(u_i, x_i) \leq D_1, \qquad \forall \boldsymbol{u} \in \mathcal{A}^n,\ \boldsymbol{x}^{i-1},\ 1 \leq i \leq n \tag{21}$$

*The single-letter marginal $Q_*$ of $\boldsymbol{Q}_*$ is given by the minimizer of $F(\rho, Q, W)$ subject to the same constraints applied to the case $n = 1$, i.e.,*

$$\min_{\boldsymbol{Q} \in \mathcal{Q}_n(D_1)} F(\rho, \boldsymbol{Q}, W^n) = \left[\min_{Q \in \mathcal{Q}(D_1)} F(\rho, Q, W)\right]^n. \tag{22}$$

The proof of Theorem 3 is sketched in Section 7.

*Comment:* The fact that a memoryless input maximizes the random coding error exponent of a memoryless channel, in the absence of distortion constraints, is well-known as the *parallel channels theorem* due to Gallager [8, Theorem 5],[9, p. 149, Example 4]. Theorem 3 tells us that this continues to be the case in the presence of distortion constraints.

Observe that for every fixed $\rho$, $F(\rho, Q, W)$, and hence also $B(\rho, Q, W)$, is convex in $Q$ [8, Theorem 4] and concave in $W$ (Section 6, Lemma 1). Also, $\mathcal{Q}(D_1)$ and $\mathcal{W}(D_2)$ are both compact convex sets. Therefore,

$$\max_{W \in \mathcal{W}(D_2)} \min_{Q \in \mathcal{Q}(D_1)} B(\rho, Q, W) = \min_{Q \in \mathcal{Q}(D_1)} \max_{W \in \mathcal{W}(D_2)} B(\rho, Q, W) \tag{23}$$

and hence there exists a saddle point [20, p. 229, Theorem 6.2.9(ii)] $(Q_\rho, W_\rho)$, which in general, depends on $\rho$. (In particular, one may minimize both sides of eq. (23) w.r.t. $\rho$.) The following theorem tells us what happens if the saddle point $(Q_\rho, W_\rho)$ happens to be independent of $\rho$.

**Theorem 4** *For a memoryless source $\boldsymbol{P} = P^n$, suppose that $(Q_*, W_*)$ is a saddle point of $B(\rho, Q, W)$ for every $0 \le \rho \le 1$. Then, $(Q_*^n, W_*^n)$ is a saddle point of $\min_{0\le\rho\le 1} B(\rho, \boldsymbol{Q}, \boldsymbol{W})$, that is,*

$$\begin{aligned}
\max_{\boldsymbol{W}\in\mathcal{W}_n(D_2)} \min_{\boldsymbol{Q}\in\mathcal{Q}_n(D_1)} \min_{0\le\rho\le 1} B(\rho, \boldsymbol{Q}, \boldsymbol{W}) &= \min_{\boldsymbol{Q}\in\mathcal{Q}_n(D_1)} \max_{\boldsymbol{W}\in\mathcal{W}_n(D_2)} \min_{0\le\rho\le 1} B(\rho, \boldsymbol{Q}, \boldsymbol{W}) \\
&= 2^{-nE(R,D_1,D_2)}, \qquad (24)
\end{aligned}$$

*where*

$$\begin{aligned}
E(R, D_1, D_2) &= \max_{0\le\rho\le 1} \max_{Q\in\mathcal{Q}(D_1)} \min_{W\in\mathcal{W}(D_2)} E_r(\rho, Q, W) \\
&= \max_{0\le\rho\le 1} \min_{W\in\mathcal{W}(D_2)} \max_{Q\in\mathcal{Q}(D_1)} E_r(\rho, Q, W) \\
&= \min_{W\in\mathcal{W}(D_2)} \max_{Q\in\mathcal{Q}(D_1)} \max_{0\le\rho\le 1} E_r(\rho, Q, W) \\
&= \max_{Q\in\mathcal{Q}(D_1)} \min_{W\in\mathcal{W}(D_2)} \max_{0\le\rho\le 1} E_r(\rho, Q, W) \\
&= \max_{0\le\rho\le 1} E_r(\rho, Q_*, W_*). \qquad (25)
\end{aligned}$$

*Proof.* Let $(Q_*, W_*)$ be the saddle point of $B(\rho, Q, W)$. By Theorem 1, for every $\boldsymbol{W} \in \mathcal{W}_n(D_2)$,

$$\begin{aligned}
B(\rho, Q_*^n, \boldsymbol{W}) &\le \max_{\boldsymbol{W}\in\mathcal{W}_n(D_2)} B(\rho, Q_*^n, \boldsymbol{W}) \\
&= \left[\max_{W\in\mathcal{W}(D_2)} B(\rho, Q_*, W)\right]^n \\
&= B(\rho, Q_*, W_*)^n \\
&= B(\rho, Q_*^n, W_*^n). \qquad (26)
\end{aligned}$$

On the other hand, by Theorem 3, for every $\boldsymbol{Q} \in \mathcal{Q}_n(D_1)$,

$$B(\rho, \boldsymbol{Q}, W_*^n) \ge \min_{\boldsymbol{Q}\in\mathcal{Q}_n(D_1)} B(\rho, \boldsymbol{Q}, W_*^n)$$

$$
\begin{aligned}
&= \left[\min_{Q\in\mathcal{Q}(D_1)} B(\rho,Q_*,W)\right]^n \\
&= B(\rho,Q_*,W_*)^n \\
&= B(\rho,Q_*^n,W_*^n). \qquad (27)
\end{aligned}
$$

Summarizing the last two inequalities, we have

$$
B(\rho,Q_*^n,\boldsymbol{W}) \le B(\rho,Q_*^n,W_*^n) \le B(\rho,\boldsymbol{Q},W_*^n), \qquad (28)
$$

that is, $(Q_*^n,W_*^n)$ is a saddle point of $B(\rho,\boldsymbol{Q},\boldsymbol{W})$ for every $\rho$. Minimizing w.r.t. $\rho$, we get

$$
\min_{0\le\rho\le 1} B(\rho,Q_*^n,\boldsymbol{W}) \le \min_{0\le\rho\le 1} B(\rho,Q_*^n,W_*^n) \le \min_{0\le\rho\le 1} B(\rho,\boldsymbol{Q},W_*^n). \qquad (29)
$$

Since this is true for every $(\boldsymbol{Q},\boldsymbol{W}) \in \mathcal{Q}_n(D_1)\times\mathcal{W}_n(D_2)$, we may maximize the l.h.s. over $\boldsymbol{W}\in\mathcal{W}_n(D_2)$ and minimize the r.h.s. over $\boldsymbol{Q}\in\mathcal{Q}_n(D_1)$, which leads to

$$
\max_{\boldsymbol{W}\in\mathcal{W}_n(D_2)} \min_{0\le\rho\le 1} B(\rho,Q_*^n,\boldsymbol{W}) \le \min_{0\le\rho\le 1} B(\rho,Q_*^n,W_*^n) \le \min_{\boldsymbol{Q}\in\mathcal{Q}_n(D_1)} \min_{0\le\rho\le 1} B(\rho,\boldsymbol{Q},W_*^n). \qquad (30)
$$

The left-most side of eq. (30) is further lower bounded by

$$
\min_{\boldsymbol{Q}\in\mathcal{Q}_n(D_1)} \max_{\boldsymbol{W}\in\mathcal{W}_n(D_2)} \min_{0\le\rho\le 1} B(\rho,\boldsymbol{Q},\boldsymbol{W})
$$

whereas the right-most side of eq. (30) is upper bounded by

$$
\max_{\boldsymbol{W}\in\mathcal{W}_n(D_2)} \min_{\boldsymbol{Q}\in\mathcal{Q}_n(D_1)} \min_{0\le\rho\le 1} B(\rho,\boldsymbol{Q},\boldsymbol{W}).
$$

On the other hand, since min–max is never smaller than max–min, both inequalities of eq. (30) must be equalities, i.e.,

$$
\begin{aligned}
\min_{\boldsymbol{Q}\in\mathcal{Q}_n(D_1)} \max_{\boldsymbol{W}\in\mathcal{W}_n(D_2)} \min_{0\le\rho\le 1} B(\rho,\boldsymbol{Q},\boldsymbol{W}) &= \min_{0\le\rho\le 1} B(\rho,Q_*^n,W_*^n) \\
&= \max_{\boldsymbol{W}\in\mathcal{W}_n(D_2)} \min_{\boldsymbol{Q}\in\mathcal{Q}_n(D_1)} \min_{0\le\rho\le 1} B(\rho,\boldsymbol{Q},\boldsymbol{W}), \qquad (31)
\end{aligned}
$$

establishing the first equality in eq. (24). Moreover, eq. (31) tells us that $(Q_*^n, W_*^n)$ is also a saddle point of $\min_{0\le\rho\le 1} B(\rho, \boldsymbol{Q}, \boldsymbol{W})$, and so,

$$\begin{aligned}
-\frac{1}{n}\log \min_{0\le\rho\le 1} B(\rho, Q_*^n, W_*^n) &= \max_{0\le\rho\le 1} E_r(\rho, Q_*, W_*)\\
&= \max_{0\le\rho\le 1}\max_{Q\in\mathcal{Q}(D_1)}\min_{W\in\mathcal{W}(D_2)} E_r(\rho, Q, W)\\
&= \max_{0\le\rho\le 1}\min_{W\in\mathcal{W}(D_2)}\max_{Q\in\mathcal{Q}(D_1)} E_r(\rho, Q, W)\\
&= \max_{Q\in\mathcal{Q}(D_1)}\min_{W\in\mathcal{W}(D_2)}\max_{0\le\rho\le 1} E_r(\rho, Q, W)\\
&= \min_{W\in\mathcal{W}(D_2)}\max_{Q\in\mathcal{Q}(D_1)}\max_{0\le\rho\le 1} E_r(\rho, Q, W)\\
&= E(R, D_1, D_2). \qquad \square \qquad (32)
\end{aligned}$$

## 5 An Example

In this section, we characterize the saddle point and the error exponent for the special case where $\boldsymbol{P}$ is the uniform distribution, and $d_1 = d_2$ is the Hamming distortion measure. We demonstrate that, in this case, the saddle point $(Q_*, W_*)$ is independent of $\rho$, and thus the condition of Theorem 4 is satisfied.

Let $p(u) = 1/A$, $A \stackrel{\Delta}{=} |\mathcal{A}|$, for all $u \in \mathcal{A}$, and let both $d_1$ and $d_2$ be the Hamming distortion measure. We claim that, in this case, as intuition suggests, the saddle point channels $Q_* = \{q_*(x|u),\ u, x \in \mathcal{A}\}$, $W_* = \{w_*(y|x),\ x, y \in \mathcal{A}\}$ are:

$$q_*(x|u) = \begin{cases} 1 - D_1 & \text{if } x = u \\ \frac{D_1}{A-1} & \text{if } x \ne u \end{cases} \qquad (33)$$

$$w_*(y|x) = \begin{cases} 1 - D_2 & \text{if } y = x \\ \frac{D_2}{A-1} & \text{if } y \ne x \end{cases} \qquad (34)$$

independently of $\rho$, where $D_1, D_2 \in [0, 1 - 1/A]$.

To see why this is true, fix $\rho \in [0,1]$ and consider the minimization of $B(\rho, Q, W_*)$ over $Q \in \mathcal{Q}(D_1)$, where $W_*$ is defined as above. For this choice of $W_*$, we have

$$\begin{aligned}
B(\rho, Q, W_*) &= 2^{\rho R} \sum_u p(u) \sum_y \left[ \sum_x q(x|u) w(y|x)^{1/(1+\rho)} \right]^{1+\rho} \\
&= 2^{\rho R} \sum_u p(u) \sum_y \left[ q(y|u)(1-D_2)^{1/(1+\rho)} + (1-q(y|u)) \left( \frac{D_2}{A-1} \right)^{1/(1+\rho)} \right]^{1+\rho} \quad (35)
\end{aligned}$$

Let $Q_1 = \{q_1(x|u),\ u, x \in \mathcal{A}\}$ be an arbitrary channel in $\mathcal{Q}(D_1)$. In the case of the Hamming distortion measure, for $Q_1$ to be a member of $\mathcal{Q}(D_1)$, it has to satisfy $\delta(u) \stackrel{\Delta}{=} 1 - q_1(u|u) \le D_1$ for all $u \in \mathcal{A}$. Now, for each $u$, consider $(A-1)! - 1$ additional channels generated by taking all permutations of the probability assignment $q(\cdot|u)$ on $\mathcal{A} - \{u\}$, while $q_1(u|u)$ and $q_1(\cdot|u')$, $u' \neq u$, all remain unchanged. Repeating this step for every $u$ at a time, we end up with a total of $m = A! - A + 1$ channels, denoted $Q_1, Q_2, ..., Q_m$, which are all in $Q(D_1)$. Obviously,

$$B(\rho, Q_1, W_*) = B(\rho, Q_2, W_*) = ... = B(\rho, Q_m, W_*)$$

since the permutations only cause a change in the order of summation over $y$ in the r.h.s. of eq. (35). Consider now the uniform mixture $\bar{Q} = \frac{1}{m}\sum_{i=1}^m Q_i$, which has the form

$$\bar{q}(x|u) = \begin{cases} 1 - \delta(u) & \text{if } x = u \\ \delta(u)/(A-1) & \text{if } x \neq u \end{cases}. \quad (36)$$

Then, by the convexity of $B$ w.r.t. $Q$, we have

$$B(\rho, \bar{Q}, W_*) \le \frac{1}{m} \sum_{i=1}^m B(\rho, Q_i, W_*) = B(\rho, Q_1, W_*), \quad (37)$$

which means that for a given $\delta(u)$, making $q(\cdot|u)$ uniform over $\mathcal{A} - \{u\}$ can only reduce $B(\rho, Q, W_*)$. It remains to show then that $B$ would further be reduced by setting $\delta(u) = D_1$ for all $u$. Indeed, by substituting eq. (36) into eq. (35), we get

$$B(\rho, \bar{Q}, W_*) = 2^{\rho R} \sum_u p(u) \left\{ \left[ (1-\delta(u))(1-D_2)^{1/(1+\rho)} + \delta(u) \left( \frac{D_2}{A-1} \right)^{1/(1+\rho)} \right]^{1+\rho} \right.$$

$$+(A-1)\left[\frac{\delta(u)}{A-1}(1-D_2)^{1/(1+\rho)}+\left(1-\frac{\delta(u)}{A-1}\right)\left(\frac{D_2}{A-1}\right)^{1/(1+\rho)}\right]^{1+\rho}\Bigg\}. \quad (38)$$

Now, each term of this summation over $u$ is of the form $s(\delta) = f(1-\delta)^{1+\rho} + (A-1)f(\delta/(A-1))^{1+\rho}$ where $f(z) = z(1-D_2)^{1/(1+\rho)} + (1-z)[D_2/(A-1)]^{1/(1+\rho)}$. A straightforward calculation shows that the derivative of $s$ vanishes at $\delta = 1 - 1/A$, and hence (by convexity) is negative for $\delta < 1 - 1/A$. Therefore, by monotonicity, $\min_{\delta \le D_1} s(\delta)$ is achieved at $\delta = D_1$ whenever $D_1 \le 1 - 1/A$. Note that thus far, we have not used the assumption that $P$ is the uniform distribution on $\mathcal{A}$.

To show that $W_*$ maximizes $B(\rho, Q_*, W)$ over $W \in \mathcal{W}(D_2)$, it is enough to show that it satisfies the Kuhn–Tucker conditions [3] (see also Section 6, proof of Lemma 2). In this case, since $w_*(y|x)$ are all positive, these conditions are as follows: For every $x \in \mathcal{A}$, there exists constants $\mu(x)$ and $\lambda(x)$ such that

$$\begin{aligned} V(Q_*, W_*, x, y) &\stackrel{\Delta}{=} \left[\frac{\partial F(\rho, Q_*, W)}{\partial w(y|x)}\right]_{W=W_*} \\ &\equiv w_*(y|x)^{-\rho/(1+\rho)} \sum_u p(u) q_*(x|u) \left[\sum_{x'} q_*(x'|u) w_*(y|x')^{1/(1+\rho)}\right]^\rho \\ &= \mu(x) + \lambda(x) d_2(x, y). \end{aligned} \quad (39)$$

Carrying out the calculation of $V(Q_*, W_*, x, y)$ for the indicated channels and the uniform distribution $P$, we obtain

$$V(Q_*, W_*, x, y) = \begin{cases} \xi & \text{if } y = x \\ \eta & \text{if } y \neq x \end{cases} \quad (40)$$

where

$$\xi = \frac{1}{A}(1-D_2)^{-\rho/(1+\rho)}[(1-D_1)\psi_1^\rho + D_1 \psi_2^\rho] \quad (41)$$

and

$$\eta = \frac{1}{A}\left(\frac{D_2}{A-1}\right)^{-\rho/(1+\rho)}\left[\frac{D_1}{A-1}\psi_1^\rho + \left(1 - \frac{D_1}{A-1}\right)\psi_2^\rho\right] \quad (42)$$

with

$$\psi_1 = D_1 \left(\frac{D_2}{A-1}\right)^{1/(1+\rho)} + (1-D_1)(1-D_2)^{1/(1+\rho)} \tag{43}$$

and

$$\psi_2 = \left(1 - \frac{D_1}{A-1}\right)\left(\frac{D_2}{A-1}\right)^{1/(1+\rho)} + \frac{D_1}{A-1}(1-D_2)^{1/(1+\rho)}. \tag{44}$$

It is not difficult to verify that $\eta \geq \xi$ for $D_1$ and $D_2$ in the indicated range. Thus, $W_*$ satisfies the Kuhn–Tucker conditions with $\mu(x) = \xi \geq 0$ and $\lambda(x) = \eta - \xi \geq 0$ for all $x \in \mathcal{A}$.

# 6 Proof of Theorem 1

The first important property of $F$ is concavity w.r.t. $\boldsymbol{W}$. This is stated in the following lemma.

**Lemma 1** *The functional $F(\rho, \boldsymbol{Q}, \boldsymbol{W})$ is concave in $\boldsymbol{W}$ for fixed $\boldsymbol{Q}$ and $\rho$.*

*Proof.* It is sufficient to show that for every $(\boldsymbol{u}, \boldsymbol{y}) \in \mathcal{A}^n \times \mathcal{A}^n$, each summand of the r.h.s. of eq. (7) is concave in $\boldsymbol{W}$. Specifically, let $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ be two arbitrary channels and let $\alpha$ be an arbitrary number in $(0,1)$. We wish to show then that

$$\begin{aligned}
& \left\{\sum_{\boldsymbol{x}} q(\boldsymbol{x}|\boldsymbol{u})\left[\alpha \boldsymbol{w}_1(\boldsymbol{y}|\boldsymbol{x}) + (1-\alpha)\boldsymbol{w}_2(\boldsymbol{y}|\boldsymbol{x})\right]^{1/(1+\rho)}\right\}^{1+\rho} \\
\geq\ & \alpha \left[\sum_{\boldsymbol{x}} \boldsymbol{q}(\boldsymbol{x}|\boldsymbol{u})\boldsymbol{w}_1(\boldsymbol{y}|\boldsymbol{x})^{1/(1+\rho)}\right]^{1+\rho} + (1-\alpha)\left[\sum_{\boldsymbol{x}} \boldsymbol{q}(\boldsymbol{x}|\boldsymbol{u})\boldsymbol{w}_2(\boldsymbol{y}|\boldsymbol{x})^{1/(1+\rho)}\right]^{1+\rho} \\
=\ & \left\{\sum_{\boldsymbol{x}} [\alpha \boldsymbol{q}(\boldsymbol{x}|\boldsymbol{u})^{1+\rho}\boldsymbol{w}_1(\boldsymbol{y}|\boldsymbol{x})]^{1/(1+\rho)}\right\}^{1+\rho} + \left\{\sum_{\boldsymbol{x}} [(1-\alpha)\boldsymbol{q}(\boldsymbol{x}|\boldsymbol{u})^{1+\rho}\boldsymbol{w}_2(\boldsymbol{y}|\boldsymbol{x})]^{1/(1+\rho)}\right\}^{1+\rho}.
\end{aligned} \tag{45}$$

Now, Minkowski's inequality (see, e.g., [9, p. 523, Problem 4.15(g)]) states that for a set of non-negative numbers $\{a_{jk}\}$, $1 \leq j \leq J$, $1 \leq k \leq K$, and $r > 1$,

$$\left[\sum_j \left(\sum_k a_{jk}\right)^{1/r}\right]^r \geq \sum_k \left(\sum_j a_{jk}^{1/r}\right)^r. \tag{46}$$

It is easy then to see that eq. (45) is obtained as a special case of Minkowski's inequality with $r = 1+\rho$, $K = 2$, $j = \boldsymbol{x}$, $a_{j1} = \alpha \boldsymbol{q}(\boldsymbol{x}|\boldsymbol{u})^{1+\rho}\boldsymbol{w}_1(\boldsymbol{y}|\boldsymbol{x})$, and $a_{j2} = (1-\alpha)\boldsymbol{q}(\boldsymbol{x}|\boldsymbol{u})^{1+\rho}\boldsymbol{w}_2(\boldsymbol{y}|\boldsymbol{x})$. This completes the proof of Lemma 1. $\square$

For a given average distortion level $D_2$, let us assign instantaneous distortion levels $\Delta_1, \Delta_2, \ldots, \Delta_n$ at times $t = 1, 2, \ldots, n$, respectively, where $\frac{1}{n}\sum_{i=1}^{n}\Delta_i \leq D_2$. Now, let us select a certain pair $(\boldsymbol{x}^{n-1}, \boldsymbol{y}^{n-1})$ and consider the constrained maximization of $F$ w.r.t. $\{w(\cdot|\boldsymbol{y}^{n-1}, \boldsymbol{x}^n),\ x_n \in \mathcal{A}\}$ where all other conditional PMF's derived from $\boldsymbol{W}$, $\{w(\cdot|\boldsymbol{y}^{i-1}, \boldsymbol{x}^i),\ \boldsymbol{x}^i \in \mathcal{A}^i,\ \boldsymbol{y}^{i-1} \in \mathcal{A}^{i-1}\}_{i=1}^{n-1}$ and $\{w(\cdot|\tilde{\boldsymbol{y}}^{n-1}, \tilde{\boldsymbol{x}}^n), (\tilde{\boldsymbol{x}}^{n-1}, \tilde{\boldsymbol{y}}^{n-1}) \neq (\boldsymbol{x}^{n-1}, \boldsymbol{y}^{n-1})\}$, are held fixed. The constraints on $\{w(\cdot|\boldsymbol{y}^{n-1}, \boldsymbol{x}^n),\ x_n \in \mathcal{A}\}$ are:

$$
\begin{aligned}
w(y_n|\boldsymbol{y}^{n-1}, \boldsymbol{x}^n) &\geq 0 \quad \forall\, x_n, y_n \in \mathcal{A} &\quad (47)\\
\sum_{y_n} w(y_n|\boldsymbol{y}^{n-1}, \boldsymbol{x}^n) &= 1,\ \forall\, x_n \in \mathcal{A} &\quad (48)\\
\sum_{y_n} w(y_n|\boldsymbol{y}^{n-1}, \boldsymbol{x}^n) d_2(x_n, y_n) &\leq \Delta_n,\ \forall\, x_n \in \mathcal{A}. &\quad (49)
\end{aligned}
$$

From Lemma 1 it is obvious that $F$ is concave in $\{w(\cdot|\boldsymbol{y}^{n-1}, \boldsymbol{x}^n),\ x_n \in \mathcal{A}\}$ given all other conditional PMF's, and so, this constrained maximization problem is a convex program. Therefore, the Kuhn–Tucker conditions are both necessary and sufficient for optimality [3, Theorems 4.38, 4.39]. Specifically, let

$$
\begin{aligned}
V(\boldsymbol{Q}, \boldsymbol{W}, \boldsymbol{x}, \boldsymbol{y}) &\stackrel{\Delta}{=} \frac{\partial F(\rho, \boldsymbol{Q}, \boldsymbol{W})}{\partial w(y_n|\boldsymbol{y}^{n-1}, \boldsymbol{x}^n)}\\
&= \boldsymbol{w}(\boldsymbol{y}^{n-1}|\boldsymbol{x}^{n-1})\boldsymbol{w}(\boldsymbol{y}|\boldsymbol{x})^{-\rho/(1+\rho)} \times\\
&\quad \sum_{\boldsymbol{u}} \boldsymbol{p}(\boldsymbol{u})\boldsymbol{q}(\boldsymbol{x}|\boldsymbol{u})\left[\sum_{\boldsymbol{x}'} \boldsymbol{q}(\boldsymbol{x}'|\boldsymbol{u})\boldsymbol{w}(\boldsymbol{y}|\boldsymbol{x}')^{1/(1+\rho)}\right]^{\rho}. \qquad (50)
\end{aligned}
$$

The Kuhn–Tucker conditions for the optimality of a conditional PMF $\{w(\cdot|\boldsymbol{y}^{n-1},\boldsymbol{x}^n)$ are that there exist constants $\{\boldsymbol{\mu}(\boldsymbol{x}^n,\boldsymbol{y}^{n-1})\}$ and $\{\boldsymbol{\lambda}(\boldsymbol{x}^n,\boldsymbol{y}^{n-1})\}$, such that for every $y_n \in \mathcal{A}$,

$$V(\boldsymbol{Q},\boldsymbol{W},\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{\mu}(\boldsymbol{x}^n,\boldsymbol{y}^{n-1}) + \boldsymbol{\lambda}(\boldsymbol{x}^n,\boldsymbol{y}^{n-1})d_2(x_n,y_n), \quad w(y_n|\boldsymbol{y}^{n-1},\boldsymbol{x}^n) > 0 \tag{51}$$

$$V(\boldsymbol{Q},\boldsymbol{W},\boldsymbol{x},\boldsymbol{y}) \le \boldsymbol{\mu}(\boldsymbol{x}^n,\boldsymbol{y}^{n-1}) + \boldsymbol{\lambda}(\boldsymbol{x}^n,\boldsymbol{y}^{n-1})d_2(x_n,y_n), \quad w(y_n|\boldsymbol{y}^{n-1},\boldsymbol{x}^n) = 0 \tag{52}$$

$$\boldsymbol{\lambda}(\boldsymbol{x}^n,\boldsymbol{y}^{n-1}) \ge 0 \tag{53}$$

$$\boldsymbol{\lambda}(\boldsymbol{x}^n,\boldsymbol{y}^{n-1})[\sum_{y_n} w(y_n|\boldsymbol{y}^{n-1},\boldsymbol{x}^n)d_2(x_n,y_n) - \Delta_n] = 0. \tag{54}$$

Consider now the one-dimensional problem $(n=1)$ of maximizing $F(\rho,Q,W)$ w.r.t. $W$ subject to the constraints

$$\begin{aligned} w(y|x) &\ge 0 \quad \forall\, x,y \in \mathcal{A} & (55)\\ \sum_y w(y|x) &= 1 \quad \forall\, x \in \mathcal{A} & (56)\\ \sum_y w(y|x)d_2(x,y) &\le \Delta_n, \quad \forall\, x \in \mathcal{A}. & (57) \end{aligned}$$

Now, by the necessity of the Kuhn–Tucker conditions there must exist constants $\{\lambda(x),\ x \in \mathcal{A}\}$, and $\{\mu(x),\ x \in \mathcal{A}\}$ such that for the optimum conditional PMF $W_n^* = \{w_n^*(y|x)\}$, we have for every $x,y \in \mathcal{A}$:

$$V(Q,W_n^*,x,y) = \mu(x) + \lambda(x)d_2(x,y), \quad w_n^*(y|x) > 0 \tag{58}$$

$$V(Q,W_n^*,x,y) \le \mu(x) + \lambda(x)d_2(x,y), \quad w_n^*(y|x) = 0 \tag{59}$$

$$\lambda(x) \ge 0 \tag{60}$$

$$\lambda(x)[\sum_y w_n^*(y|x)d_2(x,y) - \Delta_n] = 0 \tag{61}$$

where for the one–dimensional case,

$$\begin{aligned} V(Q,W,x,y) &= \frac{\partial F(\rho,Q,W)}{\partial w(y|x)} \\ &= w(y|x)^{-\rho/(1+\rho)} \sum_u p(u)q(x|u) \left[ \sum_{x'} q(x'|u) w(y|x)^{1/(1+\rho)} \right]^\rho . \end{aligned} \tag{62}$$

For every $(x,y) = (x_n, y_n) \in \mathcal{A}$, we now multiply eqs. (58)–(61) by

$$\begin{aligned} \phi(\boldsymbol{x}^{n-1}, \boldsymbol{y}^{n-1}) \stackrel{\Delta}{=} & \boldsymbol{w}(\boldsymbol{y}^{n-1}|\boldsymbol{x}^{n-1})^{1/(1+\rho)} \sum_{\boldsymbol{u}^{n-1}} \boldsymbol{p}(\boldsymbol{u}^{n-1}) \boldsymbol{q}(\boldsymbol{x}^{n-1}|\boldsymbol{u}^{n-1}) \times \\ & \left[ \sum_{\tilde{\boldsymbol{x}}^{n-1}} q(\tilde{\boldsymbol{x}}^{n-1}|\boldsymbol{u}^{n-1}) \boldsymbol{w}(\boldsymbol{y}^{n-1}|\tilde{\boldsymbol{x}}^{n-1})^{1/(1+\rho)} \right]^\rho . \end{aligned} \tag{63}$$

Based on the assumption that $\boldsymbol{W}$ is a causal channel, it is easy to verify that

$$\phi(\boldsymbol{x}^{n-1}, \boldsymbol{y}^{n-1}) V(Q, W_n^*, x_n, y_n) = [V(Q^n, \boldsymbol{W}, \boldsymbol{x}, \boldsymbol{y})]_{w(y_n|\boldsymbol{y}^{n-1}, \boldsymbol{x}^n) = w_n^*(y_n|x_n)} . \tag{64}$$

Thus, since $\phi(\boldsymbol{x}^{n-1}, \boldsymbol{y}^{n-1}) \geq 0$, we have that $w(\cdot|y^{n-1}, \boldsymbol{x}^n) = w_n^*(\cdot|x_n)$, satisfies the Kuhn–Tucker conditions with Lagrange multipliers given by

$$\boldsymbol{\lambda}(\boldsymbol{x}^n, \boldsymbol{y}^{n-1}) = \lambda(x_n) \phi(\boldsymbol{x}^{n-1}, \boldsymbol{y}^{n-1})$$

and

$$\boldsymbol{\mu}(\boldsymbol{x}^n, \boldsymbol{y}^{n-1}) = \mu(x_n) \phi(\boldsymbol{x}^{n-1}, \boldsymbol{y}^{n-1}).$$

Therefore, by the sufficiency of the Kuhn–Tucker conditions, $\{w_n(\cdot|y^{n-1}, \boldsymbol{x}^n) = w_n^*(\cdot|x_n),\ x_n \in \mathcal{A}\}$ maximizes $F$ subject to the indicated constraints. By repeating the same argument for all $\boldsymbol{x}^{n-1}$ and $\boldsymbol{y}^{n-1}$, it becomes apparent that for the optimum channel $\boldsymbol{W}$, $y_n$ depends only on $x_n$ and not on past channel inputs and outputs. For a channel with this property,

$$F(\rho, Q^n, \boldsymbol{W}^{(n)}) = F(\rho, Q^{n-1}, \boldsymbol{W}^{(n-1)}) F(\rho, Q, W_n^*),$$

Note that since $W_n^*$ is independent of $\boldsymbol{W}^{(n-1)}$, the problem is reduced from dimension $n$ to $n-1$, and the next step is to maximize $F(\rho, Q^{n-1}, \boldsymbol{W}^{(n-1)})$ w.r.t. $\{w(\cdot|y^{n-2}, \boldsymbol{x}^{n-1}),\ x_{n-1} \in \mathcal{A}\}$ subject to distortion constraint $\Delta_{n-1}$. By iterative repetition of the above proof, we finally conclude that the optimum $n$–dimensional channel for individual distortion levels $\Delta_1, ..., \Delta_n$ is the product form channel $W_1^* \times \ldots \times W_n^*$, with each $W_i^*$ being the optimal solution for the one–dimensional problem with distortion level $\Delta_i$, $i = 1, 2, \ldots, n$.

Finally, we show that the best distortion assignment $(\Delta_1, \ldots, \Delta_n)$ is uniform. Let $F_{\max}(\Delta)$ denote the maximum of $F(\rho, Q, W)$ subject to the constraints

$$w(y|x) \geq 0 \quad \forall x, y \tag{65}$$

$$\sum_y w(y|x) = 1 \quad \forall x \tag{66}$$

$$\sum_y w(y|x) d_2(x, y) \leq \Delta \quad \forall x. \tag{67}$$

Since $F(\rho, Q, W)$ is concave in $W$ and since the constraint set is a convex set, then it is clear that $F_{\max}(\Delta)$ is concave as well. Since the logarithmic function is monotonic and concave, then so is $\log F_{\max}(\Delta)$. Given $\Delta_1, \ldots, \Delta_n$, the channel $W_1^* \times \ldots \times W_n^*$ achieves

$$\log F(\rho, Q^n, W_1^* \times \ldots \times W_n^*) = \sum_{i=1}^n \log F_{\max}(\Delta_i).$$

Now, since $\frac{1}{n} \sum_{i=1}^n \Delta_i \leq D_2$, we have

$$\begin{aligned}
\sum_{i=1}^n \log F_{\max}(\Delta_i) &= n \cdot \frac{1}{n} \sum_{i=1}^n \log F_{\max}(\Delta_i) \\
&\leq n \log F_{\max}\left(\frac{1}{n} \sum_{i=1}^n \Delta_i\right) \\
&\leq n \log F_{\max}(D_2),
\end{aligned} \tag{68}$$

where the first inequality follows from the concavity and the second inequality follows from the monotonicity of $\log F_{\max}(\cdot)$. Both inequalities become equalities if $\Delta_i = D_2$ for all $i$. This means that the optimum channel subject to the time–average distortion constraint with distortion level $D_2$ is a DMC $\boldsymbol{W}_* = W_*^n$ whose single–letter marginal $W_*$ is optimum for the one–dimensional version of this problem. □

# 7 Outline of the Proof of Theorem 3

The proof is essentially analog to that of Theorem 1 with a few slight differences. Therefore, we give here a brief sketch only. The fact that $F(\rho, \boldsymbol{Q}, \boldsymbol{W})$ is convex w.r.t. $\boldsymbol{Q}$ is well-known [8, Theorem 4]. We now wish to minimize $F$ w.r.t. $q(\cdot|\boldsymbol{x}^{n-1}, \boldsymbol{u})$ subject to the constraints that $q(\cdot|\boldsymbol{x}^{n-1}, \boldsymbol{u})$ is a PMF and the distortion constraint $\sum_{x_n} q(x_n|\boldsymbol{x}^{n-1}, \boldsymbol{u}) d_1(u_n, x_n) \le D_1$. Let

$$\begin{aligned} U(\boldsymbol{Q}, \boldsymbol{W}, \boldsymbol{u}, \boldsymbol{x}) &\triangleq \frac{1}{1+\rho} \cdot \frac{\partial F(\rho, \boldsymbol{Q}, W^n)}{\partial q(x_n|\boldsymbol{x}^{n-1}, \boldsymbol{u})} \\ &= \boldsymbol{q}(\boldsymbol{x}^{n-1}|\boldsymbol{u}) \boldsymbol{p}(\boldsymbol{u}) \sum_{\boldsymbol{y}} \boldsymbol{w}(\boldsymbol{y}|\boldsymbol{x})^{1/(1+\rho)} \left[ \sum_{\boldsymbol{x}'} \boldsymbol{q}(\boldsymbol{x}'|\boldsymbol{u}) \boldsymbol{w}(\boldsymbol{y}|\boldsymbol{x}')^{1/(1+\rho)} \right]^{\rho}. \end{aligned} \quad (69)$$

The Kuhn–Tucker conditions for optimality of a conditional PMF $q(\cdot|\boldsymbol{x}^{n-1}, \boldsymbol{u})$ are the existence of non-negative constants $\boldsymbol{\mu}(\boldsymbol{u}, \boldsymbol{x}^{n-1})$ and $\boldsymbol{\lambda}(\boldsymbol{u}, \boldsymbol{x}^{n-1})$ such that

$$U(\boldsymbol{Q}, \boldsymbol{W}, \boldsymbol{u}, \boldsymbol{x}) = \boldsymbol{\mu}(\boldsymbol{u}, \boldsymbol{x}^{n-1}) - \boldsymbol{\lambda}(\boldsymbol{u}, \boldsymbol{x}^{n-1}) d_1(u_n, x_n) \quad q(x_n|\boldsymbol{x}^{n-1}, \boldsymbol{u}) > 0 \quad (70)$$

$$U(\boldsymbol{Q}, \boldsymbol{W}, \boldsymbol{u}, \boldsymbol{x}) \ge \boldsymbol{\mu}(\boldsymbol{u}, \boldsymbol{x}^{n-1}) - \boldsymbol{\lambda}(\boldsymbol{u}, \boldsymbol{x}^{n-1}) d_1(u_n, x_n) \quad q(x_n|\boldsymbol{x}^{n-1}, \boldsymbol{u}) = 0 \quad (71)$$

$$\boldsymbol{\lambda}(\boldsymbol{u}, \boldsymbol{x}^{n-1}) \ge 0 \quad (72)$$

$$\boldsymbol{\lambda}(\boldsymbol{u}, \boldsymbol{x}^{n-1}) [\sum_{x_n} q(x_n|\boldsymbol{x}^{n-1}, \boldsymbol{u}) d_1(u_n, x_n) - D_1] = 0. \quad (73)$$

Let $Q^* = \{q^*(x|u)\}$ minimize $F(\rho, Q, W)$ with distortion constraint $D_1$ and let $\{\mu(u)\}$ and $\{\lambda(u)\}$ be the Lagrange multipliers that satisfy the corresponding Kuhn–Tucker conditions regarding the value of

$$U(Q, W, u, x) = \frac{1}{1+\rho} \cdot \frac{\partial F(\rho, Q, W)}{\partial q(x_n|u_n)}$$

$$= \quad p(u) \sum_y w(y|x)^{1/(1+\rho)} \left[ \sum_{x'} q(x'|u) w(y|x')^{1/(1+\rho)} \right]^{\rho}. \tag{74}$$

Let

$$\begin{aligned}\psi(\boldsymbol{u}, \boldsymbol{x}^{n-1}) \quad &\stackrel{\Delta}{=} \quad \boldsymbol{q}(\boldsymbol{x}^{n-1}|\boldsymbol{u}) \boldsymbol{p}(\boldsymbol{u}^{n-1}) \sum_{\boldsymbol{y}^{n-1}} \boldsymbol{w}(\boldsymbol{y}^{n-1}|\boldsymbol{x}^{n-1})^{1/(1+\rho)} \times \\ & \left[ \sum_{\tilde{\boldsymbol{x}}^{n-1}} \boldsymbol{q}(\tilde{\boldsymbol{x}}^{n-1}|\boldsymbol{u}) \boldsymbol{w}(\boldsymbol{y}^{n-1}|\tilde{\boldsymbol{x}}^{n-1})^{1/(1+\rho)} \right]^{\rho}. \end{aligned} \tag{75}$$

It is now easy to verify that for $\boldsymbol{W} = W^n$ (even when $\boldsymbol{Q}$ is non-causal),

$$U(Q, W, u_n, x_n) \psi(\boldsymbol{u}, \boldsymbol{x}^{n-1}) = [U(\boldsymbol{Q}, W^n, \boldsymbol{u}, \boldsymbol{x})]_{q(\cdot|\boldsymbol{x}^{n-1}, \boldsymbol{u}) = q^*(\cdot|u_n)} \tag{76}$$

and so, by multiplying the Kuhn–Tucker conditions of the one–dimensional case by $\psi(\boldsymbol{u}, \boldsymbol{x}^{n-1})$ we obtain the Kuhn–Tucker conditions for $q(\cdot|\boldsymbol{x}^{n-1}, \boldsymbol{u}) = q^*(\cdot|u_n)$. Similarly as in the proof of Theorem 1, induction leads to the optimality of a time–invariant DMC as the watermarking channel.

*Comment:* Note that here, unlike in the last step of the proof of Theorem 1, it is not obvious that we can relax the instantaneous distortion constraints and impose instead only a constraint on the time–average distortion. The reason is that the logarithm of the convex function $F_{\min}(\Delta) = \min\{F(\rho, Q, W) : \sum_x q(x|u) d_1(u, x) \leq \Delta, \ Q \text{ is a PMF}\}$ might not be convex. However, given an average distortion level $D_1$, one might be able to minimize $\prod_i F_{\min}(\Delta_i)$ s.t. $\sum_i \Delta_i \leq nD_1$ and thereby find the optimum product form $\boldsymbol{Q}$.

## Acknowledgments

I would like to thank Ronny Roth for useful discussions. I am also grateful to the anonymous reviewers for their helpful comments. In particular, special thanks to one of the reviewers, who pointed out a gap in a proof that appeared in the original version of this paper.

## References


[1] R. Anderson, "Stretching the limits of steganography," in *Information Hiding*, Springer Lecture Notes in Computer Science vol. 1174, pp. 39–48, 1996. Also, available at `http://www.bluespike.com/research/research.html`.

[2] R. J. Anderson and A. Petitcolas, "On the limits of steganography," *IEEE J. Select. Areas in Communications*, vol. 16, no. 4, pp. 463-473, May 1998.

[3] M. Avriel, *Nonlinear Programming Analysis and Methods.* Prentice-Hall, 1976.

[4] S. Craver, N. Memon, B.-L. Yeo, and M. M. Yeung, "Resolving rightful ownerships with invisible watermarking techniques: limitations, attacks, and implications," *IEEE J. Select. Areas in Commun.,* vol. 16, no. 4, pp. 573–586, May 1998.

[5] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press 1981.

[6] T. Ericson, "Exponential error bounds for random codes in arbitrarily varying channels," *IEEE Trans. Inform. Theory*, vol. IT–31, no. 1, pp. 42–48, January 1985.

[7] M. Feder and A. Lapidoth, "Universal decoders for channels with memory," *IEEE Trans. Inform. Theory*, vol. IT–44, no. 5, pp. 1726–1745, September 1998.

[8] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inform. Theory*, vol. IT–11, no. 1, pp. 3–18, 1965.

[9] R. G. Gallager, *Information Theory and Reliable Communication*, J. Wiley & Sons, 1968.

[10] R. G. Gallager, "The random coding bound is tight for the average code," *IEEE Trans. Inform. Theory*, pp. 244–246, March 1973.

[11] E. Geraniotis, "Minimax robust coding for channels with uncertain statistics," *IEEE Trans. Inform. Theory*, vol. IT–31, no. 6, pp. 802–811, November 1985.

[12] V. D. Goppa, "Nonprobabilistic mutual information without memory," *Probl. Contr. Inform. Theory*, vol. 4, pp. 97–102, 1975.

[13] B. L. Hughes and T. G. Thomas, "On error exponents for arbitrarily varying channels," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 87–98, January 1996.

[14] A. Lapidoth and J. Ziv, "On the universality of the LZ-based decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT–44, no. 5, pp. 1746–1755, September 1998.

[15] N. Merhav, "Universal decoding for memoryless Gaussian channels with a deterministic interference," *IEEE Trans. Inform. Theory*, vol. IT-39, no. 4, pp. 1261–1269, July 1993.

[16] J. A. O'Sullivan, P. Moulin, and J. M. Ettinger, "Information theoretic analysis of steganography," *Proc. ISIT '98*, p. 297, 1998.

[17] S. D. Servetto, C. I. Podilchuk, and K. Ramchandran, "Capacity issues in digital image watermarking," *Proc. ICIP '98*, 1998.

[18] I. G. Stiglitz, "Coding for a class of unknown channels," *IEEE Trans. Inform. Theory*, vol. IT–12, no. 2, pp. 189–195, April 1966. (correction: vol. IT–13, no. 2, p. 195).

[19] I. G. Stiglitz, "A coding theorem for a class of unknown channels," *IEEE Trans. Inform. Theory*, vol. IT–13, no. 2, pp. 217–220, April 1967.

[20] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions I*, Springer–Verlag, New York, 1970.

[21] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. IT–31, no. 4, pp. 453–460, July 1985.