Information Spectrum Approach to the Source Channel Separation Theorem

Nir Elkayam and Meir Feder
Department of Electrical Engineering - Systems
Tel-Aviv University, Israel
Email: nirelkayam@post.tau.ac.il, meir@eng.tau.ac.il

Abstract

A source-channel separation theorem for a general channel has recently been shown by Aggrawal et al. [1]. This theorem states that if there exists a coding scheme that achieves a maximum distortion level $d_{\text{max}}$ over a general channel $W$, then reliable communication can be accomplished over this channel at rates less than $R(d_{\text{max}})$, where $R(\cdot)$ is the rate distortion function of the source. The source, however, is essentially constrained to be discrete and memoryless (DMS). In this work we prove a stronger claim where the source is general, satisfying only a “sphere packing optimality” feature, and the channel is completely general. Furthermore, we show that if the channel satisfies the strong converse property as defined by Han & Verdu [2], then the same statement can be made with $d_{\text{avg}}$, the average distortion level, replacing $d_{\text{max}}$. Unlike the proofs in [1], we use information spectrum methods to prove the statements and the results can be quite easily extended to other situations.

I. INTRODUCTION

The source-channel separation theorem, first proved by Shannon [3] for the transmission of discrete memoryless source (DMS) over a discrete memoryless channel (DMC), states that the separation strategy is optimal. This means optimal performance can be attained by first compressing the source output to the desired distortion level and then reliably communicating the compressed bits over the channel. The separation theorem was later extended to indecomposable channels [4]. The almost lossless case (transmission codes) was handled at [5] and a general condition is given there for the separation theorem to hold.

Joint Source-Channel Coding (JSCC) refers to the case where such a separation is not used. In some cases, e.g., symmetric binary source over a binary symmetric channel (BSC) or a Gaussian source over an Additive White Gaussian Noise (AWGN) channel, separation can attain the optimal performance, yet a simpler JSCC strategy can be used, i.e., uncoded transmission [6].

In some cases separation is suboptimal. A simple example would be the symmetric binary source and a compound memoryless BSC where the flipping probability is drawn ahead of the block and stays fixed for the whole block. In this case uncoded transmission is optimal and it is strictly suboptimal to use separation.

Some caution is needed here, because there are two senses of optimality when a distortion measure is given - the maximal distortion level and average distortion level. The separation relative to the maximal distortion level is easier to accomplish, as this allows us to increase the distortion as long as we do not exceed the desired distortion level. So even if the average distortion in the original scheme is much less than the maximum distortion, the separation strategy yields the desired maximum distortion but the average distortion might be increased.

When average distortion level is used, we should follow the specific distortion, which can be large or small as the channel conditions are good or bad. This is much like the variable rate channel capacity [7], which tries to capture the whole spectrum of channel conditions - when the channel provides good conditions lower distortion level can be achieved or more bits can be reliably transmitted over the channel. A separation strategy in that case will use successive refinement of the source [8], [9] and then transmission of the bits over a channel with variable rate channel capacity so that the better the channel, the lower the distortion [10]. In many cases this separation strategy is suboptimal.

In transmitting a DMS over a DMC the cases of average distortion level and maximal distortion level coincide [11]. However, for other channels this is not always the case even for DMSs. We will see that for indecomposable channels, this is always true.

Information spectrum methods [2], [12] provide a very simple formalization and intuition into channel capacity in almost every communication situation including unicast, multiple access, broadcast, and other situations [13]. The problem is that the expressions given with these methods are usually not useful when we want to compute the channel capacity. Nevertheless, we use information spectrum methods in this work and attain the desired results.

In this paper we deal with sources for which the sphere packing bound is tight. This means that the probability of the intersection between the typical set and the sphere of radius $d$ around any reconstruction point is at most $2^{-nR(d)}$.

For these sources we first prove a generalization of the Han-Verdu converse lemma [2] that connects the rate distortion spectrum (that is, the probability that the random variable $D$, the instantaneous distortion level, exceeds some level $d$) to the
information spectrum of the channel. Then we use this to prove a very general separation theorem for the case where maximum distortion level criteria are given. For channels that satisfy the strong converse, which includes DMC and ergodic channels, we prove a separation theorem for the average distortion level as well. Actually, for these kinds of channels we prove that the average and maximal distortion levels coincide.

II. Preliminaries

A. Notation

This paper uses lower case letters (e.g., $x$) to denote a particular value of the corresponding random variable denoted in capital letters (e.g., $X$). Calligraphic fonts (e.g., $\mathcal{X}$) represent a set.

The random variable $D$ represents the instantaneous distortion level. $\mathbb{E}(D) = d_{avg}$ is the average distortion level, and $d_{max}$ will be used to denote the maximum distortion level.

We will use the $o(\cdot)$ notation to denote terms that go to 0 w.r.t. the argument. Mainly, $o(n)$ will denote a sequence $\epsilon_n$ such that $\lim_{n \to \infty} \epsilon_n/n = 0$ and $o(1)$ denotes a sequence that converges to 0. Throughout this paper log will be defined to base 2 unless otherwise indicated. $Pr\{A\}$ will denote the probability of the event $A$.

For Information Spectrum notation we follow [2]. We will use $i(a;b)$ to denote the information density between two outcomes from correlated random variables $A$ and $B$. Specifically, $i(a;b) = \frac{1}{n} \log \left( \frac{p(a|b) p(b)}{p(a)} \right)$. We will omit the $n$ and the indication of which random variable produces the outcomes as it will be understood from the context.

B. Definition

Definition 1 ([12] Source and Reproduction alphabet, Distortion measure, Rate distortion function). A general source $S$ is defined as an infinite sequence of random variables on $\mathcal{S}_n$. The reproduction alphabet is defined over the set $\hat{\mathcal{S}}_n$. A distortion measure is a function $d_n : \mathcal{S}_n \times \hat{\mathcal{S}}_n \rightarrow \mathbb{R}_+$. There are several rate distortion functions that can be defined according to the different performance requirements, see [12] ch.5.3.

- $R_{fm}(d)$ - fixed length code, maximum distortion criterion.
- $R_{fa}(d)$ - fixed length code, average distortion criterion.
- $R_{vm}(d)$ - variable length code, maximum distortion criterion.
- $R_{va}(d)$ - variable length code, average distortion criterion.

Definition 2 (Sphere Packing Optimal Source (SPO)). A general source $S$ is said to be Sphere Packing Optimal if there exist subsets $A_n \subset \mathcal{S}_n$, and $k_n$, such that:

$$\lim_{n \to \infty} Pr\{s_n \notin A_n\} = 0$$

$$Pr\{s_n \in A_n, d_n(s_n, \hat{s}_n) \leq d\} \leq 2^{-n(R_{fm}(d)+k_n)}$$

$$\lim_{n \to \infty} k_n = 0$$

for each $d \geq 0$ and $\hat{s}_n \in \hat{\mathcal{S}}_n$.

Remark 1. In section V we demonstrate that memoryless sources (satisfying regularity conditions) are SPO.

Remark 2. For SPO sources that have a reference word (see [12] Theorem 5.3.1), the following different notations of rate functions are equal:

$$R_{fm}(d) = R_{fa}(d) = R_{vm}(d).$$

The common value will be denoted $R(d)$ without the subscript. This is also shown in the full paper [14].

Definition 3 (General Channel). A general channel is a sequence of transition matrices $W = \{ W_n : \mathcal{X}_n \rightarrow \mathcal{Y}_n \}$ where $W^n(y|x)$ denotes the conditional probability of $y$ given $x$. Throughout this paper we will assume that the channel has finite input and output space.

Remark 3. The source and channel input and output will be written as $\mathcal{S}_n, \mathcal{X}_n, \mathcal{Y}_n, \hat{\mathcal{S}}_n$ to indicate that it is usually not the $n^{th}$ order Cartesian product. As an example, let $\mathcal{S}_n = \mathcal{S}_n^{\mathcal{R}}, \hat{\mathcal{S}}_n = \hat{\mathcal{S}}_n^{\mathcal{R}}$, and $\mathcal{X}_n = \mathcal{X}_n, \mathcal{Y}_n = \mathcal{Y}_n$. This allows us to simplify the notation and to avoid unnecessary ”rate” indication. We will use $x$ and $y$ to denote the input and output. Occasionally, we will omit the subscript $n$ and assume that it is understood from the context.

Definition 4 (JSCC Scheme). A general JSCC Scheme of the source $S$ over the channel $W$, includes:

- Encoding function: $E_n : \mathcal{S}_n \rightarrow \mathcal{X}_n$

1 Extension of the results to abstract input and output spaces $(\mathcal{X}_n, \mathcal{Y}_n)$ requires subtle handling; see [12]. Yet, this extension is possible whenever information spectrum can be used.
Decoding function: $D_n : Y_n \to \hat{S}_n$.

There are several random variables defined:

- $X_n = E_n(S_n)$ - The random variable of the channel input.
- $Y_n$ - The output from the channel $W$ resulting from the input $X_n$.
- $D = d(s_n, D_n(Y_n))$ - The instantaneous distortion.

Sometimes we will write $x(s_n)$ for $E_n(s_n)$. The maximum distortion level $d_{\text{max}}$ is the infimum of the set of numbers $\alpha$, such that $Pr \{ D \geq \alpha \} \to 0$. The average distortion level is $E(D) = d_{\text{avg}}$.

Throughout the sequel we will need to use the following lemma, which says that if there exists a probability mass to the right of the mean of a non-negative random variable then there must be a probability mass to the left of it.

**Lemma 1.** Let $D$ be a non-negative random variable with $\mu = E(D) < \infty$. If there exist $d_1 > 0$ and $\epsilon_1 > 0$ such that: $Pr \{ D > \mu + d_1 \} > \epsilon_1$, then there exist $d_2 > 0$ and $\epsilon_2 > 0$ such that: $Pr \{ D < \mu - d_2 \} > \epsilon_2$.

### III. Joint Source Channel Lemma

In this section we state and prove a generalization of the Han-Verdú converse lemma [2], which relates the error rate probability to the information spectrum. Here we connect the instantaneous distortion level $D$ with the information spectrum of the channel.

A similar converse result was given in [15, Theorem 1] for a general source. We, however, provide the following lemma that holds for an SPO source, which is enough for our purposes. Let:

$$L_{R(d)}^\gamma = \{ (s_n, y) : i(x(s_n); y) \leq R(d) - \gamma \}$$

This is exactly the set that is used in the definition of channel capacity in terms of information spectrum.

**Lemma 2 (JSCH Converse lemma).** For any JSCH Scheme of the SPO source $S$ we have:

$$Pr \{ D > d \} \geq Pr \{ L_{R(d)}^\gamma \} - 2^{-n(\gamma + k_n)} - Pr \{ s_n \notin A_n \}$$

**Proof.** The term $Pr \{ L_{R(d)}^\gamma \}$ can be bounded by:

$$Pr \{ L_{R(d)}^\gamma \} \leq Pr \{ D > d \} + Pr \{ s_n \notin A_n \}$$

Continue with the last term in (4):

$$Pr \{ L_{R(d)}^\gamma \cap (D \leq d) \cap (s_n \in A_n) \} \leq \sum_{s_n \in A_n, (s_n, y) \in L_{R(d)}^\gamma \cap (D \leq d)} p(s_n) p(y) \cdot 2^{nR(d) - n\gamma}$$

where (a) follows because $(s_n, y) \in L_{R(d)}^\gamma \Rightarrow p(y|x(s_n)) \leq p(y) \cdot 2^{nR(d) - n\gamma}$ and (b) follows from the SPO assumption of the source $S$. Combining (4) and (5) completes the proof of the inequality. \qed
IV. SOURCE CHANNEL SEPARATION THEOREMS

A. Maximum distortion

Theorem 1 (Source channel separation, maximum distortion). If the maximum distortion in the joint source channel coding of an SPO source $S$ is $d_{\text{max}}$, then any rate smaller than $R(d_{\text{max}})$ is achievable, i.e., the capacity of the channel $W$ is at least $R(d_{\text{max}})$.

Proof. Fix a sequence $\gamma_n \to 0$ such that $n(\gamma_n + k_n) \to \infty$. Let $X_n = E_n(S_n)$ and $Y_n$ be the output of $X_n$ through the channel $W$. From lemma 2 we have:

$$Pr\{D > d_{\text{max}}\} \geq Pr\left\{L^n_{R(d_{\text{max}})}\right\} - 2^{-n(\gamma_n + k_n)} + o(1)$$

Since $\lim_{n \to \infty} Pr\{D > d_{\text{max}}\} = 0$ and $2^{-n(\gamma_n + k_n)} \to 0$, we get $\lim_{n \to \infty} Pr\left\{L^n_{R(d_{\text{max}})}\right\} = 0$. Since $\gamma_n \to 0$. This implies that:

$$\lim_{n \to \infty} Pr\{i(x(s_n); y) \leq R(d)\} = 0,$$

which proves that $I(X_n; Y_n) \geq R(d)$ so the channel capacity is at least $R(d)$ and separation is possible.

B. Average distortion

Theorem 2 (Joint Source channel Separation, average distortion). For an optimal JSIS scheme of an SPO source over a channel that satisfies the strong converse property, the notion of average distortion and maximum distortion coincide.

Proof. Suppose not. Then there exist $\tau_1$ and $\epsilon_1$ such that:

$$\limsup_{n \to \infty} Pr\{D > E(D) + \tau_1\} \geq \epsilon_1.$$

By lemma 1 we have: $\limsup_{n \to \infty} Pr\{D \leq E(D) + \epsilon\} \geq \epsilon$ for some $\tau > 0$ and $\epsilon > 0$, which implies that $\liminf_{n \to \infty} Pr\{D > E(D) - \tau\} \geq 1 - \epsilon$. Fix a sequence $\gamma_n \to 0$ such that $n(\gamma_n + k_n) \to \infty$. Now:

$$1 - \epsilon \geq \liminf_{n \to \infty} Pr\{D > E(D) - \tau\} \geq \liminf_{n \to \infty} \left(Pr\left\{L^n_{R(E(D)-\tau)}\right\} - 2^{-n(\gamma_n + k_n)} + o(1)\right)$$

(6)

For $n$ large enough we have that $\liminf_{n \to \infty} Pr\left\{L^n_{R(E(D)-\tau)}\right\}$ bound away from 1, so we can transmit at rate $R(E(D) - \tau) - \gamma_n$ with error probability less than 1. Since the channel satisfies the strong converse, this means that we can transmit at rate $R(E(D) - \tau)$ with error approaching 0, so the JSIC is not optimal because with separation we can transmit at maximum distortion level $E(D) - \tau$, which is better the average $E(D)$. (We used here the fact that $R_{\text{fm}}(d) = R_{\text{fa}}(d)$.)

C. The operational point of view

The operational approach used in 1 can also be used here to prove the separation theorem for SPO sources when the maximum distortion criteria are used. Suppose that a JSIC system operates over a channel $W$ and achieves maximum distortion $d$, then a reliable communication system can be built over the JSIC system and provide communication at rate that approaches $R(d)$. Let the block-length be $n$ and fix a sequence $\gamma_n \to 0$ such that $n(\gamma_n + k_n) \to \infty$.

- **Codebook generation**: Generate $2^{n(R(d) - \gamma_n)}$ codewords independently from the source $S$.
- **Joint typicality**: $(s_n, s_n)$ are said to be jointly typical if $s_n \in A_n$ and $d(s_n, s_n) \leq d$.
- **Decoding**: let $s_n$ be received at the output of the channel. If there exists a unique codeword $s_n$ that is jointly typical with $\hat{s}_n$, then declare that $s_n$ was transmitted, else declare error.

The true word obviously would be jointly typical with the transmitted word with probability that approaches 1 and it remains to see that the probability of all the other words to be jointly typical approaches 0. Using the union bound and the definition of the SPO source this follows easily.

As mentioned in [1], this argument can also be applied over general networks.

D. The Multiple Access Case

The results can be extended to the multiple access channel. As mentioned in [1], for the maximum distortion notion the separation follows from the unicast case. For average distortion we should follow the information spectrum approach to use the strong converse property of the channel. See [14] for further discussion.
V. THE D-TILTED INFORMATION AND RELATION TO SPO SOURCES

We follow the notation of Verdú and Kostina [15]. Denote for a given source $S$ with distribution $P(S)$ on $S$ and distortion measure $d : S \times \hat{S} \mapsto [0, +\infty)$,

$$R_S(d) = \inf_{P(\hat{S}|S)}: \mathbb{E}(d(S, \hat{S})) \leq d I(S; \hat{S}).$$

(7)

We impose the following basic restrictions on $S$ and the distortion measure; see also [16]:

1) $R_S(d)$ is finite for some $d$, i.e., $d_{\text{min}} < \infty$, where

$$d_{\text{min}} = \inf \{d : R_S(d) < \infty\}.$$  

(8)

2) The infimum in (7) is achieved by a unique $P(\hat{S}|S)$.

**Definition 5** ($d$–tilted information, [15]). For $d > d_{\text{min}}$, the $d$–tilted information in $s$ is defined as

$$j_S(s, d) = \log \frac{1}{\mathbb{E}(\exp \{\lambda^* d - \lambda^* d(s, \hat{s})\})},$$

where the expectation is with respect to the unconditional distribution of $\hat{S}$, and

$$\lambda^* = -R_S^*(d).$$

(10)

The following properties of $d$–tilted information, proven in [17], are used in the sequel.

$$j_S(s, d) = i(s; \hat{s}) + \lambda^* d(s, \hat{s}) - \lambda^* d$$

(11)

$$\mathbb{E}(j_S(s, d)) = R_S(d)$$

(12)

$$\mathbb{E}(\exp \{\lambda^* d - \lambda^* d(S, \hat{s}) + j_S(s, d)\}) \leq 1,$$

(13)

where (11) holds for $P(\hat{S})$–almost every $\hat{s}$, while (15) holds for all $s \in \hat{S}$, and

$$i(s; \hat{s}) = \log \left(\frac{P(\hat{s}|s)}{P(\hat{s})}\right)$$

(14)

denotes the information density of the joint distribution $P(\hat{s}|s)$ at $(s, \hat{s})$.

The next calculation can be used to bound the $d$-ball over any reproduction word:

**Lemma 3.**

$$\Pr \{d(S, \hat{s}) < d, j_S(s, d) > \alpha\} \leq e^{-\alpha}.$$  

(15)

**Proof.**

$$\Pr \{d(S, \hat{s}) < d, j_S(s, d) > \alpha\} =$$

$$= \sum_{s : d(s, \hat{s}) < d} \Pr \{s\} \leq \sum_{s : d(s, \hat{s}) < d, j_S(s, d) > \alpha} \Pr \{s\} \leq$$

$$\leq \sum_{s : d(s, \hat{s}) < d, j_S(s, d) > \alpha} \Pr \{s\} \cdot e^{j_S(s, d)} \cdot e^{\lambda^* d(S, \hat{s})}$$

$$\leq e^{-\alpha} \cdot \mathbb{E}(\exp \{\lambda^* d - \lambda^* d(S, \hat{s}) + j_S(s, d)\})$$

$$\leq e^{-\alpha} ,$$

where (a) follows because we multiply by terms bigger than 1 and (b) follows from (13).

For a sequence of sources $S_n$, we apply the above definition and properties of the $d$–tilted information to every $n$, and assume the regularity condition holds for every $n$.

**Proposition 3.** If there exists a sequence $\gamma_n$ such that $A_n = \{s_n : j_S(s_n, d) > n \cdot R_S(d) - n k_n\}$ with $k_n \rightarrow_{n \rightarrow \infty} 0$ and $\Pr \{A_n\} \rightarrow_{n \rightarrow \infty} 1$, then the source $S_n$ is SPO.

**Proof.** The conditions (1), (2) are by definition. (2) follows from lemma 3.

\[ \]
For memoryless sources with $S_n = S^n$, obviously $\hat{S}_n = \hat{S}^n$, i.e., the infimum in (7) is achieved by memoryless reproduction sequence, and it follows from the law of large numbers $\mathbb{E}(\mathbb{J}^n(s^n, d)) \approx \mathbb{E}(\mathbb{J}^n(s^n, d)) = R_S(d) = n \cdot R_S(d)$ with probability that approaches 1.

**Theorem 4.** Every memoryless source satisfying the regularity conditions is SPO.

**VI. POSSIBLE EXTENSIONS**

**A. Stationary ergodic sources**

Suppose that we also have $I(S; Z) = I(S; Z) = I(S; Z)$, and $\mathbb{E}(D) = D$. These assumptions mean that $I(s; z) \approx I(S; Z) = nR(d)$ and $d(s, z) \approx d$ with probability that approaches 1 over the joint distribution of $s, z$. Now by (11), this means that $\mathbb{J}^n(s, d) \approx nR(d)$ with probability 1 and the source is SPO. Here we conjecture that ergodic stationary sources with a subadditive distortion metric are SPO according to [12, 5.9]. A full rigorous proof obviously still needs more delicate arguments.

**B. Network and correlated sources**

The extension of the results to a general network is straightforward for independent sources. It is not clear what is the "right" definition of SPO when there exists correlation between the sources (see [18]) and whether the information spectrum approach applies. Another interesting extension would be to prove when the separation is optimal when the channel does not satisfy the strong converse property and the average distortion criterion is used.

**REFERENCES**


