ISIT'2013 1569715129

Universal Wyner-Ziv Coding for Distortion Constrained General Side-Information

Shun Watanabe* and Shigeaki Kuzuoka†
*Department of Information Science and Intelligent Systems, University of Tokushima, Tokushima 770-8506, Japan,
and Institute for System Research, University of Maryland, College Park, MD 20742, USA,
Email: shun-wata@is.tokushima-u.ac.jp
†Department of Computer and Communication Sciences, Wakayama University, Wakayama 640-8510, Japan,
Email:kuzuoka@ieee.org.

***Abstract*—We investigate the Wyner-Ziv coding in which the statistics of the principal source is known but the statistics of the channel generating the side-information is unknown except that it is in a certain class. The class consists of channels such that the distortion between the principal source and the side-information is smaller than a threshold, but channels may be neither stationary nor ergodic. In this situation, we define a new rate-distortion function as the minimum rate such that there exists a Wyner-Ziv code that is universal for every channel in the class. Then, we show an upper bound and a lower bound on the rate-distortion function, and derive a matching condition such that the upper and lower bounds coincide.**

I. Introduction

In the seminal paper [1], Wyner and Ziv characterized the rate-distortion function of the lossy source coding with side-information at the decoder. In this paper, we consider a universal coding of this problem where the statistics of the principal source is known but the channel from the principal source to the side-information is unknown except that it is in a certain class.

To motivate the problem setting investigated in this paper, let us consider the following practical situation first. Suppose that the decoder already has a lossy compressed version of the principal source, and want to get a refined one. The encoder does not know how the previously transmitted lossy version is encoded, but knows that the quality of the lossy version is guaranteed to be above a certain level. What is the minimum additional rate that must be transmitted by the encoder so that the quality of the refined version is above a required level?

The above mentioned situation can be modeled as follows. The principal source X^n is a known i.i.d. source, and the side-information Y^n is generated from X^n through a channel W^n. The statistical property of the channel is unknown, but the distortion caused by the channel is smaller than a certain level E for a prescribed distortion measure. We assume that the distortion measure is additive, but the channel may be neither stationary nor ergodic. We consider the maximum distortion constraint and the average distortion constraint for the channel. Since we allow non-ergodic channel, the class of channels constrained by the maximum distortion and that constrained by the average distortion are different. In this problem formulation, we are interested in the minimum rate $R_m(D|E)$ and $R_a(D|E)$ such that the reproduction with distortion level D is possible at the decoder for any channel in the classes of channels satisfying the distortion level E with the maximum distortion constraint and the average distortion constrain respectively. In other word, we are interested in the minimum rate such that the universal coding is possible for each class.

For the maximum distortion constrained class, we show an upper bound and a lower bound on $R_m(D|E)$. We also derive a matching condition such that the upper and the lower bounds coincide. Especially, for the binary Hamming example, we show that the matching condition is satisfied, and thus $R_m(D|E)$ is completely characterized.

For the average distortion constrained class, we show an upper bound and a lower bound on $R_a(D|E)$. For the case with $D = 0$, i.e., the loss less reproduction case, we show that the upper and lower bounds coincide and thus $R_a(0|E)$ is completely characterized. Surprisingly, $R_a(0|E) = H(X)$, i.e., the side-information is completely useless, for any $E > 0$.

Some remarks on related literatures are in order.

For lossless source coding with side-information, i.e., the Slepian-Wolf network [2], the existence of universal code was first shown by Csiszár and Körner [3] (existence of linear universal code was also shown by Csiszár [4]). After that, the universal codings for the Slepian-Wolf network or other related lossless multi-terminal networks were studied by several researchers [5], [6].

For lossy source coding with side-information, i.e., the Wyner-Ziv network, the universal coding problem was investigated by Merhav and Ziv [7], Jalali *et. al.* [8], and Reani and Merhav [9]. It should be noted that the universal codes proposed in these literatures are universal for the statistics of the principal source but not for the channel generating the side-information, i.e., the statistics of the channel is known at the encoder. Under the same condition, i.e., known channel, it is also known that the universal code can be constructed for the network with several decoders [10].

The universal Wyner-Ziv coding is also related to the Heeger-Berger problem [11], in which there are several decoders that have their own side-information. The Heeger-Berger problem has not been solved in general, and it has only been solved under the condition that there is a degraded partial order between the channels generating the side-information [12], [13], [14] except some special cases [15], [16]. It should be noted that there is no degraded partial order among the channel class considered in this paper. Thus, the authors believe that the result in this paper also shed some light on the unsolved Heeger-Berger problem.

Our problem setting can be also viewed as a kind of

the successive refinement coding [17], [18]. The successive refinement coding consists of two layers of the encodings. If the method used by the first layer encoder is not known to the second layer encoder, this is exactly the situation of our problem setting.

Although the universal coding for distortion constrained class of channels is unfamiliar and new in the source coding scenario, this kind of channel is quite natural when the channel is cased by an adversary such as in the data hiding scenario. Indeed, this kind of channel class is commonly used in the information theoretical analysis of the data hiding [19], [20], [21].

There are some technical differences between the data hiding problem and our problem. First, in the data hiding problem, the channel output is only used for the decoding of the encoded message. On the other hand, in our problem, the side-information is not only used for the decoding of the encoded source, but also for the estimation at the decoder. This makes the problem difficult, and causes a gap between the upper bound and the lower bound derived in this paper. Second, in the data hiding problem for the average distortion constrained class of channels, it was shown that the achievable transmission rate is 0, i.e., the channel is completely useless [20]. On the other hand, in our problem for the average distortion constrained class of channels, the side-information is useless for bin coding, but it can be used for the estimation at the decoder. Thus, $R_a(D|E)$ can be strictly smaller than the rate-distortion function $R(D)$ without any side-information for $D > 0$, though $R_a(0|E) = H(X)$.

The rest of this paper is organized as follows. In Section II, we introduce notations and the formal definition of the problem. In Section III, we state our main theorems, and show a representative example, i.e., the binary Hamming example. Sketch of proofs will be presented in Section IV. The detail of the proofs can be found in the full version of this paper [22].

II. Preliminaries

A. Notations

Henceforth, we adopt the following notation conventions. Random variables will be denoted by capital letters such as X, while their realizations will be denoted by respective lower case letters such as x. A random vector of length n is denoted by $X^n = (X_1, \ldots, X_n)$, while its realization is denoted by $x^n = (x_1, \ldots, x_n)$. The alphabet of a random variable is denoted by a calligraphic letter such as $\mathcal{X}$, and its n-fold Cartesian product is denoted by $\mathcal{X}^n$. The probability distribution of random variable X is denoted by P_X, and its n-fold i.i.d. extension is denoted by P_X^n. For a given channel W, its n-fold i.i.d. extension is denoted by $W^{\times n}$, while W^n indicates a channel that is not necessarily i.i.d.. The set of all probability distribution on $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$. The set of all channel from $\mathcal{X}$ to $\mathcal{Y}$ is denoted by $\mathcal{P}(\mathcal{Y}|\mathcal{X})$. The indicator function is denoted by $\mathbf{1}[\cdot]$. The entropy and the mutual information is denoted in a standard notation such as $H(X)$ or $I(X;Y)$. For a input distribution P of a channel W, we sometimes use the notation $I(P, W)$ to designate the mutual information $I(X;Y)$, where the joint distribution of (X, Y) is $P(x)W(y|x)$.

B. Problem Formulation

Let $\boldsymbol{X} = \{X^n\}_{n=1}^{\infty}$ be an i.i.d. source. Let

$$e_n(x^n, y^n) := \frac{1}{n}\sum_{t=1}^{n} e(x_t, y_t)$$

be an additive distortion measure for side information. As a natural assumption, we assume that there exists y such that $e(x, y) = 0$ for each x. We also assume that the distortion is bounded, i.e., $e(x, y) \le e_{\max} < \infty$ for every (x, y). For a given distortion $E \ge 0$, we consider the following maximum distortion constraint on the side-information

$$\begin{aligned}\mathcal{W}_m(E) \;:=\; & \{\boldsymbol{W} = \{W^n\}_{n=1}^{\infty} : \forall \delta > 0\ \exists n_0(\delta) \text{ s.t.} \\ & \Pr\{e_n(X^n, Y^n) > E\} \le \delta\ \forall n \ge n_0(\delta)\},\end{aligned}$$

where Y^n is the output of channel W^n with input X^n. It should be noted that $n_0(\delta)$ depends on δ but not on $\boldsymbol{W}$. We also consider the average distortion constraint

$$\begin{aligned}&\mathcal{W}_a(E) \\ &:= \{\boldsymbol{W} = \{W^n\}_{n=1}^{\infty} : e_n(P_{X^n}, W^n) \le E\ \forall n \ge 1\} \quad (1)\end{aligned}$$

where

$$\begin{aligned}e_n(P_{X^n}, W^n) \;:=\; & \mathbb{E}[e_n(X^n, Y^n)] \\ =\; & \sum_{x^n, y^n} P_X^n(x^n) W^n(y^n|x^n) e_n(x^n, y^n).\end{aligned}$$

As it will be clarified later, the maximum distortion constraint and the average distortion constraint are completely different.

Let $\hat{\mathcal{X}}$ be the reproduction alphabet. Then, let

$$d_n(x^n, \hat{x}^n) := \frac{1}{n}\sum_{t=1}^{n} d(x_t, \hat{x}_t) \quad (2)$$

be an additive distortion measure for reproduction. We assume $d(x, \hat{x}) \le d_{\max} < \infty$ for every $(x, \hat{x})$.

We consider (possibly stochastic) encoder $\varphi_n : \mathcal{X}^n \to \mathcal{M}_n$ and decoder $\psi_n : \mathcal{M}_n \times \mathcal{Y}^n \to \hat{\mathcal{X}}^n$.

Definition 1: For any $\varepsilon > 0$, if there exists $n_0(\varepsilon)$ and a sequence of codes $\{(\varphi_n, \psi_n)\}_{n=1}^{\infty}$ such that $\frac{1}{n}\log|\mathcal{M}_n| \le R + \varepsilon$ and $\mathbb{E}[d_n(X^n, \psi_n(\varphi_n(X^n), Y^n))] \le D + \varepsilon$ for every $\boldsymbol{W} \in \mathcal{W}_m(E)$ and $n \ge n_0(\varepsilon)$, then we define the rate R to be *achievable*. We also define the rate distortion function

$$R_m(D|E) := \inf\{R : R \text{ is achievable}\}.$$

We also define $R_a(D|E)$ by replacing $\mathcal{W}_m(E)$ with $\mathcal{W}_a(E)$.

Let $R_{WZ}(D|W)$ be the rate distortion function of the ordinary Wyner-Ziv problem in which the principal source is X and the side-information Y is the output of the channel $W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$.

The rate distortion function $R_m(D|E)$ (or $R_a(D|E)$) means that if $R > R_m(D|E)$ there exists a *universal* code that works well for every $\boldsymbol{W} \in \mathcal{W}_m(E)$ (or $\boldsymbol{W} \in \mathcal{W}_a(E)$). It should be noted that this definition of universality is different from the ordinary definition of the universality. Let

$$\mathcal{W}_{WZ}(R, D) := \{W \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : R_{WZ}(D|W) \le R\}. \quad (3)$$

In the ordinary definition of the universality, we require that there exists a code that works well for every $W \in$

$\mathcal{W}_{WZ}(R,D)$. This requirement seems much severe than the requirement of $R_m(D|E)$ (or $R_a(D|E)$).

In this paper, we also use some terminologies from the Heegard-Berger problem [11] (see also [23]), which is a rate-distortion problem with one encoder and two decoders. Fix an i.i.d. source P_X. Then two side-information channel $W_1 : \mathcal{X} \to \mathcal{Y}$ and $W_2 : \mathcal{X} \to \mathcal{Y}$ define an i.i.d. joint source (X, Y_1, Y_2) whose joint distribution $P_{XY_1Y_2}$ is given by $P_{XY_1Y_2}(x, y_1, y_2) = P_X(x)W_1(y_1|x)W_2(y_2|x)$, where $x \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$. In the following, we denote by $R_{HB}(D_1, D_2|W_1, W_2)$ the HB rate-distortion function $R_{HB}(D_1, D_2|X, Y_1, Y_2)$ for (X, Y_1, Y_2) defined by W_1 and W_2.

Unfortunately, finding a single-letter expression for $R_{HB}(D_1, D_2|W_1, W_2)$ has been a long-standing open problem. So, we consider a special case. Let

$$E_* := \min_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} P_X(x)e(x, y)$$

and $y_* \in \mathcal{Y}$ be a symbol which attains the minimum. Further, let $W_* : \mathcal{X} \to \mathcal{Y}$ be a side-information channel such that $W_*(y_*|x) = 1$ irrespective $x \in \mathcal{X}$. Then, let us consider a special case where $W_1 = W_*$. This case is equivalent to the problem of "lossy coding when side-information may be absent", and the single letter characterization for this special case has been solved in [11].

III. Main Result

A. Convex Form of WZ Rate-Distortion Function

We need convex form of the Wyner-Ziv rate-distortion function introduced in [24]. Let $\mathcal{U}$ be the set of all functions from $\mathcal{Y}$ to $\hat{\mathcal{X}}$. The set $\mathcal{U}$ includes a constant function, i.e., $u(y) = \hat{x}\ \forall y \in \mathcal{Y}$ for each $\hat{x} \in \hat{\mathcal{X}}$. We denote the set of constant functions by $\bar{\mathcal{U}} \subset \mathcal{U}$. For fixed channel $W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ and fixed test channel $V \in \mathcal{P}(\mathcal{U}|\mathcal{X})$, we denote

$$d(V, W) := \sum_{u,x,y} P_X(x)V(u|x)W(y|x)d(x, u(y)).$$

For a fixed channel $W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$, let

$$\mathcal{V}(W, D) := \{V \in \mathcal{P}(\mathcal{U}|\mathcal{X}) : d(V, W) \le D\}.$$

Let

$$\mathcal{W}_1(E) := \{W \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : e(P_X, W) \le E\}$$

and

$$\mathcal{V}(E, D) := \{V \in \mathcal{P}(\mathcal{U}|\mathcal{X}) : d(V, W) \le D\ \forall W \in \mathcal{W}_1(E)\}.$$

For $(V, W) \in \mathcal{P}(\mathcal{U}|\mathcal{X}) \times \mathcal{P}(\mathcal{Y}|\mathcal{X})$, let

$$\begin{aligned} \phi(V, W) &:= I(U; X) - I(U; Y) && (4)\\ &= I(U; X|Y). && (5)\end{aligned}$$

Note that $\phi(\cdot, W)$ is a convex function for fixed W, which can be confirmed from (5), and $\phi(V, \cdot)$ is a concave function for fixed V, which can be confirmed from (4).

By the above notations, the Wyner-Ziv rate-distortion function is given by

$$R_{WZ}(D|W) = \min_{V\in\mathcal{V}(W,D)} \phi(V, W).$$

Let

$$\tilde{R}_{WZ}(D|W, E) = \min_{V\in\mathcal{V}(E,D)} \phi(V, W)$$

be the pseudo rate-distortion function.

Lemma 1: The pseudo rate-distortion function $\tilde{R}_{WZ}(D|W, E)$ is a concave function of the channel, i.e.,

$$\begin{aligned}&\tilde{R}_{WZ}(D|\lambda W_1 + (1-\lambda)W_2, E)\\ &\quad\ge\ \lambda\tilde{R}_{WZ}(D|W_1, E) + (1-\lambda)\tilde{R}_{WZ}(D|W_2, E)\end{aligned}$$

holds for $W_1, W_2 \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ and $0 \le \lambda \le 1$.

B. Statements of General Results

For the maximum distortion class, we have the following.

Theorem 1: We have

$$\begin{aligned} R_m(D|E) &\ge \max_{W\in\mathcal{W}_1(E)} R_{WZ}(D|W) && (6)\\ &= \max_{W\in\mathcal{W}_1(E)} \min_{V\in\mathcal{V}(W,D)} \phi(V, W) && (7)\end{aligned}$$

and

$$\begin{aligned} R_m(D|E) &\le \min_{V\in\mathcal{V}(E,D)} \max_{W\in\mathcal{W}_1(E)} \phi(V, W) && (8)\\ &= \max_{W\in\mathcal{W}_1(E)} \min_{V\in\mathcal{V}(E,D)} \phi(V, W). && (9)\\ &= \max_{W\in\mathcal{W}_1(E)} \tilde{R}_{WZ}(D|W, E) && (10)\end{aligned}$$

The difference between (7) and (9) are $\mathcal{V}(W, D)$ and $\mathcal{V}(E, D)$. Thus, we have the following matching conditions.

Corollary 1: Let (V^*, W^*) be a saddle point satisfying

$$\phi(V^*, W^*) = \max_{W\in\mathcal{W}_1(E)} \min_{V\in\mathcal{V}(E,D)} \phi(V, W).$$

Suppose that

$$\hat{V} := \operatorname*{argmin}_{V\in\mathcal{V}(W^*,D)} \phi(V, W^*) \in \mathcal{V}(E, D).$$

Then, we have

$$R_m(D|E) = \phi(V^*, W^*) = \max_{W\in\mathcal{W}_1(E)} \min_{V\in\mathcal{V}(E,D)} \phi(V, W).$$

Corollary 2: Under the same notations as Corollary 1, suppose that $\mathrm{supp}(\hat{V}) \subset \bar{\mathcal{U}}$. Then, we have

$$R_m(D|E) = \phi(V^*, W^*) = \max_{W\in\mathcal{W}_1(E)} \min_{V\in\mathcal{V}(E,D)} \phi(V, W).$$

For the average distortion class, we have the following.

Theorem 2: We have

$$\begin{aligned}&R_a(D|E)\\ &\ge \max_{\substack{\lambda,E_1,E_2,W_1,W_2:\\ \lambda E_1+(1-\lambda)E_2\le E\\ W_1\in\mathcal{W}_1(E_1),W_2\in\mathcal{W}_1(E_2)}} \ \min_{\substack{D_1,D_2:\\ \lambda D_1+(1-\lambda)D_2\le D}} R_{HB}(D_1, D_2|W_1, W_2),\end{aligned} \tag{11}$$

where (i) max is taken over all $0 \le \lambda \le 1$, $E_j \ge 0$, and side information channels W_1, W_2 such that $\lambda E_1 + (1-\lambda)E_2 \le E$ and $W_j \in \mathcal{W}_1(E_j)$ $(j = 1, 2)$ and (ii) min is taken over

all $D_1, D_2 \in [0, d_{\max}]$ such that $\lambda D_1 + (1-\lambda)D_2 \le D$. Especially,

$$R_a(D|E) \ge \max_{\substack{\lambda, E_2, W_2 \in \mathcal{W}_1(E_2) \\ \lambda E_* + (1-\lambda)E_2 \le E}} \min_{\substack{D_1, D_2: \\ \lambda D_1 + (1-\lambda)D_2 \le D}} R_{HB}(D_1, D_2 | W_*, W_2) \tag{12}$$

holds. We also have

$$R_a(D|E) \le \min_{V \in \mathcal{V}(E,D)} I(P_X, V). \tag{13}$$

Remark 1: Note that (12) is obtained from (11) by letting $E_1 = E_*$ and $W_1 = W_*$. Thus, (11) is tighter than (12). However, we cannot give a single letter expression for the right hand side of (11), while we can for (12).

By setting $\lambda = 0$ and $E_2 = E$ in (12), we have the following corollary.

Corollary 3: We have

$$R_a(D|E) \ge \max_{W \in \mathcal{W}_1(E)} R_{WZ}(D|W).$$

For the lossless case, we also have the following corollary.

Corollary 4: For $D = 0$ and $E > 0$, we have[1]

$$R_a(0|E) = H(X). \tag{14}$$

This corollary indicates that the side information is completely useless when $D = 0$ and $E > 0$. It should be emphasized that Corollary 3 does not give Corollary 4 in general. This means that our result (12) is tighter than Corollary 3.

C. Binary Hamming Example

To provide some insight on our results, we consider the binary Hamming example, i.e., we assume that $\mathcal{X} = \mathcal{Y} = \hat{\mathcal{X}} = \{0,1\}$, $P_X(0) = P_X(1) = \frac{1}{2}$, and

$$e(x,y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{else} \end{cases}, \quad d(x,\hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{else} \end{cases}.$$

In this section, we assume that $E \le \frac{1}{2}$.

We first consider the maximum distortion class. In this case, the set $\mathcal{W}_1(E)$ can be parametrized by two parameters (α, β) satisfying $(\alpha + \beta)/2 \le E$.

By the concavity of $\tilde{R}_{WZ}(D|W,E)$ with respect to W (Lemma 1) and by the symmetry with respect to α and β, we have

$$\operatorname*{argmax}_{W \in \mathcal{W}_1(E)} \tilde{R}_{WZ}(D|W,E) = \text{BSC}(E).$$

Let $0, 1 \in \mathcal{U}$ be constant functions that output 0 or 1 irrespective of y and let y be the function that output y itself. Similarly, let $\bar{y}$ be the function that outputs $y \oplus 1$. In the binary Hamming case, $\mathcal{U} = \{0, 1, y, \bar{y}\}$. For $W^* = \text{BSC}(E)$, it is known that

$$R_{WZ}(D|W^*) = \min_{V \in \mathcal{V}(W^*,D)} \phi(V, W^*)$$

[1]We need the condition $E > 0$ because we need to take $\lambda > 0$ in (12).

is achieved by the test channel of the form

$$\hat{V}(u|x) = \begin{cases} \lambda(1-q) & \text{if } u = x \\ \lambda q & \text{if } u = x \oplus 1 \\ (1-\lambda) & \text{if } u = y \end{cases}$$

for some $0 \le \lambda \le 1$ and $0 \le q \le \frac{1}{2}$ (λ represents the time sharing). In this case, the distortion is given by

$$\begin{aligned} & \lambda \sum_{\hat{x},x} P_X(x) V_q(\hat{x}|x) d(x,\hat{x}) \\ & \qquad + (1-\lambda) \sum_{x,y} P_X(x) W^*(y|x) d(x,y) \\ = \;& \lambda \sum_{\hat{x},x} P_X(x) V_q(\hat{x}|x) d(x,\hat{x}) + (1-\lambda) E \\ \le \;& D, \end{aligned}$$

where $V_q = \text{BSC}(q)$. Since every channel $W \in \mathcal{W}_1(E)$ satisfies

$$\sum_{x,y} P_X(x) W(y|x) d(x,y) \le E,$$

we find that $\hat{V} \in \mathcal{V}(E,D)$. Thus, the matching condition of Corollary 1 is satisfied for this binary Hamming example.

Next, we consider the average distortion class. We evaluate the upper bound (13). We first fix W^* to be BSC(E). Note that

$$\min_{V \in \mathcal{V}(E,D)} I(P_X, V) \ge \min_{V \in \mathcal{V}(W^*,D)} I(P_X, V). \tag{15}$$

For a test channel $V \in \mathcal{V}(W^*, D)$, let $\bar{V}$ be a test channel such that $\bar{V}(u|x) = V(u \oplus 1 | x \oplus 1)$ for $u \in \bar{\mathcal{U}}$ and $\bar{V}(u|x) = V(u|x \oplus 1)$ for $u \in \{y, \bar{y}\}$. Then, by the symmetry of the BSC and the source P_X, we have $\bar{V} \in \mathcal{V}(W^*, D)$ and $I(P_X, V) = I(P_X, \bar{V})$. By the convexity of the mutual information for channel, we have

$$I(P_X, \tilde{V}) \le \frac{1}{2} I(P_X, V) + \frac{1}{2} I(P_X, \bar{V}),$$

where $\tilde{V} = \frac{1}{2}V + \frac{1}{2}\bar{V}$. This means that the minimum in the right hand side of (15) is achieved by a symmetric test channel, i.e., $V(u|x) = V(u \oplus 1 | x \oplus 1)$ for $u \in \bar{\mathcal{U}}$ and $V(u|x) = V(u|x \oplus 1)$ for $u \in \{y, \bar{y}\}$. Furthermore, for $E \le \frac{1}{2}$, we can assume that $V(\bar{y}|x) = 0$ because using $\bar{y}$ only makes the distortion larger. We also note that such a symmetric test channel satisfies $V \in \mathcal{V}(E,D)$. Thus, the equality in (15) actually holds. Consequently, the upper bound on $R_a(D|E)$ in this example is the time sharing between the ordinary rate-distortion function and the distortion that can be achieved only by the estimation, i.e., the point $(E, 0)$.

IV. Sketch of Proofs

A. Proof of Theorem 1

A proof of the converse part is very simple. For arbitrary $\delta > 0$, the definition of $\mathcal{W}_m(E)$ implies $\{W^{\times n}\}_{n=1}^{\infty} \in \mathcal{W}_m(E)$ for some $W \in \mathcal{W}_1(E - \delta)$, which implies $R_m(D|E) \ge \max_{W \in \mathcal{W}_1(E-\delta)} R_{WZ}(D|W)$.

In a proof of the direct part, first note that the function $\phi(\cdot, W)$ is a convex function for fixed W, $\phi(V, \cdot)$ is a concave

function for fixed V, and $\mathcal{W}_1(E)$ and $\mathcal{V}(E,D)$ are convex sets. Thus, (9) is derived from (8) by applying the saddle point theorem [25]. We prove (8) by three steps. First, we prove that there exists a universal code for i.i.d. channels. Then, we show that there exists a randomized universal code for permutation invariant channels. Finally, we de-randomize the randomized universal code by using the technique of [26].

In the first step, we construct a universal Wyner-Ziv code for a fixed test channel such that it works well for every $W \in \mathcal{W}_1(E) \cap \mathcal{P}_n(\mathcal{Y}|\mathcal{X})$. We construct a universal Wyner-Ziv code by using the output statistics of random binning argument recently introduced by [27]. We note that a universal Wyner-Ziv code can be also constructed from the coding method in [28]. The goal of the first step is to prove the following lemma.

Lemma 2: For any $V \in \mathcal{V}(E,D)$ and any $\delta > 0$, there exists a universal code (φ_n, ψ_n) and a constant $\mu > 0$ such that

$$\frac{1}{n}\log|\mathcal{M}_n| \le \max_{W \in \mathcal{W}_1(E)} \phi(V,W) + 2\delta$$

and

$$\Pr\{d_n(X^n, \psi_n(\varphi_n(X^n), Y^n)) > D + \delta\} \le 2^{-\mu n}$$

for every $W \in \mathcal{W}_1(E) \cap \mathcal{P}_n(\mathcal{Y}|\mathcal{X})$ provided that n is sufficiently large.

In the second step, we first apply the random permutation π_n on $\{1,\ldots,n\}$ to the sequence (X^n, Y^n), and then use the code given by Lemma 2 for channels in $\mathcal{W}_m(E)$. More precisely, the goal of the second step is to prove the following lemma.

Lemma 3: For any $V \in \mathcal{V}(E + \delta e_{\max}, D)$, any $\delta > 0$, and any $\varepsilon > 0$, there exists a universal code (φ_n, ψ_n) such that

$$\frac{1}{n}\log|\mathcal{M}_n| \le \max_{W \in \mathcal{W}_1(E+\delta e_{\max})} \phi(V,W) + 2\delta$$

and

$$\begin{aligned}\mathbb{E}_{\pi_n}\left[\Pr\{d_n(\pi_n(X^n), \psi_n(\varphi_n(\pi_n(X^n)), \pi_n(Y^n))) > D + \delta\}\right] \\ \le \quad \varepsilon \end{aligned} \tag{16}$$

for every $\mathbf{W} \in \mathcal{W}_m(E)$ provided that n is sufficiently large.

In the third step, we reduce the size of the random permutation π_n by using the technique of [26].

B. Proof of Theorem 2

In contrast to the converse part of Theorem 1, the converse part of Theorem 2 is quite complicated. We only prove (11) because (12) is obtained from (11) by letting $E_1 = E_*$ and $W_1 = W_*$. Assume that R is achievable and fix λ, E_1, E_2, W_1, and W_2 such that $\lambda E_1 + (1-\lambda)E_2 \le E$ and $W_j \in \mathcal{W}_j(E_j)$ for $j = 1, 2$. To prove (11), it is sufficient to show that there exists a pair (D_1, D_2) such that $\lambda D_1 + (1-\lambda)D_2 \le D$ and $R \ge R_{HB}(D_1, D_2|W_1, W_2)$. An idea to prove this is to introduce the compound channel $W^n = \lambda W_1^{\times n} + (1-\lambda)W_2^{\times n}$, and relate the Wyner-Ziv problem for W^n to the Heegerd-Berger problem for $(W_1^{\times n}, W_2^{\times n})$.

In a similar manner to the direct part of Theorem 1, the direct part of is Theorem 2 proved by three steps.

REFERENCES

[1] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. 22, no. 1, pp. 1–10, January 1976.

[2] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, no. 4, pp. 471–480, July 1973.

[3] I. Csiszár and Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inform. Theory*, vol. 27, no. 1, pp. 5–12, January 1981.

[4] I. Csiszár, "Linear codes for sources and source networks: Error exponents, universal coding," *IEEE Trans. Inform. Theory*, vol. 28, no. 4, pp. 585–592, July 1982.

[5] Y. Oohama and T. S. Han, "Universal coding for the Slepian-Wolf data compression system and the strong converse theorem," *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1908–1919, November 1994.

[6] A. Kimura, T. Uyematsu, S. Kuzuoka, and S. Watanabe, "Universal source coding over generalized complementary delivery networks," *IEEE Trans. Inform. Theory*, vol. 55, no. 3, pp. 1360–1373, March 2009.

[7] N. Merhav and J. Ziv, "On the Wyner-Ziv problem for individual sequences," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 867–873, March 2006.

[8] S. Jalali, S. Verdu, and T. Weissman, "A universal scheme for Wyner-Ziv coding of discrete sources," *IEEE Trans. Inform. Theory*, vol. 56, no. 4, pp. 1737–1750, April 2010.

[9] A. Reani and N. Merhav, "Efficient on-line scheme for encoding individual sequences with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. 57, no. 10, pp. 6860–6876, October 2011.

[10] S. Kuzuoka, A. Kimura, and T. Uyematsu, "Universal source coding for multiple decoders with side information," in *IEEE International Symposium on Information Theory*, 2010, pp. 1–5.

[11] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *IEEE Trans. Inform. Theory*, vol. 31, no. 6, pp. 727–734, November 1985.

[12] Y. Steinberg and N. Merhav, "On successive refinement for the Wyner-Ziv problem," *IEEE Trans. Inform. Theory*, vol. 50, no. 8, pp. 1636–1654, August 2004.

[13] C. Tian and S. Diggavi, "On multistage successive refinement for Wyner-Ziv source coding with degraded side informations," *IEEE Trans. Inform. Theory*, vol. 53, no. 8, pp. 2946–2960, August 2007.

[14] ——, "Side-information scalable source coding," *IEEE Trans. Inform. Theory*, vol. 54, no. 12, pp. 5591–508, December 2008.

[15] R. Timo, T. Chan, and A. Grant, "Rate distortion with side-information at many decoders," *IEEE Trans. Inform. Theory*, vol. 57, no. 8, pp. 5240–5257, August 2011.

[16] S. Watanabe, "The rate-distortion function for product of two sources with side-information at decoders," in *Proc. IEEE Int. Symp. Inf. Theory 2011*, Saint Petersburg, Russia, 2011, pp. 2862–2866, arXiv:1105.2864.

[17] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 269–275, March 1991.

[18] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Trans. Inform. Theory*, vol. 40, no. 1, pp. 253–259, January 1994.

[19] P. Mouline and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 563–593, March 2003.

[20] A. S. Cohen and A. Lapidoth, "The gaussian watermarking game," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1639–1667, June 2002.

[21] A. S.-Baruch and N. Merhav, "On the error exponent and capacity games of private watermarking systems," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 537–562, March 2003.

[22] S. Watanabe and S. Kuzuoka, "Universal wyner-ziv coding for distortion constrained general side-information," 2013, arXiv:1302.0050.

[23] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, 2011.

[24] F. M. J. Willems, "Computation of the Wyner-Ziv rate distortion function," in *Eindhoven Univ. Tech. Rep. 83-E-140*, 1983.

[25] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Athena Scientific, 2003.

[26] R. Ahlswede, "A method of coding and an application to arbitrary varying channel," *J. Comb. Inform. Syst.*, vol. 5, no. 1, pp. 10–35, 1980.

[27] M. H. Yassaee, M. R. Aref, and A. Gohari, "Achievability proof via output statistics of random binning," 2012, arXiv:1203.0730.

[28] B. G. Kelly and A. B. Wagner, "Reliability in source coding with side information," *IEEE Trans. Inform. Theory*, vol. 58, no. 8, pp. 5086–5111, August 2012, arXiv:1109.0923.