Universal Wyner–Ziv coding of discrete memoryless sources with known side information statistics

Maxim Raginsky
Department of Electrical and Computer Engineering
Duke University, Durham NC, 27708, USA
E-mail: m.raginsky@duke.edu

***Abstract*— We consider a universal variant of the Wyner–Ziv problem of lossy source coding with decoder side information, where the marginal distribution of the side information is known, while the conditional distribution of the source sequence given the side information is only assumed to lie in a given parametric family. Our approach combines minimum-distance channel estimation with a recent scheme of Jalali, Verdú and Weissman based on discrete universal denoising with auxiliary information. Our scheme is universal under mild regularity conditions. We also give a concrete example of a family of sources satisfying the regularity conditions.**

I. Introduction

The Wyner–Ziv (WZ) problem [1] pertains to lossy compression with side information at the decoder. Since the original publication [1], this problem has been studied extensively, both theoretically and practically, but always under the assumption that the joint distribution of the source and the side information is known. Recently, however, there has been some progress on universal variants of the WZ problem, where the side information symbols are generated from the source symbols via a known memoryless channel, and the source symbols form either an unknown individual sequence [2] or a realization of an unknown stationary ergodic process [3].

In this paper, we make another contribution to the problem of universal WZ coding. We consider the situation when the marginal distribution of the decoder side information is known, but there is uncertainty regarding the conditional distribution of the source sequence given the side information. This set-up, shown in Fig. 1, is the reverse of the usual arrangement, where the side information sequence is viewed as a noisy observation of the source sequence. However, there are settings in which it is more natural to regard the source as a stochastic function of the side information rather than the other way around. As an example, consider a sensor network comprised by n sensors deployed at random, according to a known distribution, over some region of interest. The region is partitioned into L disjoint cells, and the exact physical location of the ith sensor, $1 \leq i \leq n$, is quantized to the index $y_i \in \{1, \ldots, L\}$ of the cell containing that sensor. Now, let x_i be the measurement collected by the ith sensor and suppose that it depends probabilistically both on y_i and on some unknown "state of nature" θ. Assuming that the sensors do not attempt to self-localize, the problem of rate-constrained transmission of the measurement vector x^n to a fusion center that keeps track of the location of each sensor can be formulated in terms of universal WZ coding with known side information statistics.

It was pointed out in [2] that, unless there is feedback, no WZ coding scheme can be universal w.r.t. the channel that generates the side information from the source sequence. The reason is that the encoder cannot learn the statistics of the side information. The setting of this paper, however, is different because here the encoder already knows the statistics of side information and can use this knowledge to learn the unknown "state of nature." Thus, under suitable regularity conditions, we can guarantee universality w.r.t. the channel that generates the source sequence from the side information. Apart from applications in the domain of sensor networks, universal WZ coding with known side information statistics may be useful in such settings as lossy functional compression [4] or learning from compressed observations [5]. We plan to explore these topics in future work.

II. Notation and preliminaries

We begin by fixing notation. The set of all probability distributions on a finite set $\mathcal{A}$ will be denoted by $\mathcal{M}(\mathcal{A})$; the set of all channels with input alphabet $\mathcal{A}$ and output alphabet $\mathcal{B}$ will be denoted by $\mathcal{C}(\mathcal{A} \to \mathcal{B})$. The variational distance between two $P, P' \in \mathcal{M}(\mathcal{A})$ is defined as

$$\|P - P'\|_1 \triangleq \sum_{a \in \mathcal{A}} |P(a) - P'(a)|.$$

The supremum distance between two $Q, Q' \in \mathcal{C}(\mathcal{A} \to \mathcal{B})$ is defined as

$$\|Q - Q'\|_\infty \triangleq \max_{a \in \mathcal{A}} \|Q(\cdot|a) - Q'(\cdot|a)\|_1,$$

where, for each $a \in \mathcal{A}$, $Q(\cdot|a)$ is the probability distribution $\{Q(b|a)\}_{b \in \mathcal{B}} \in \mathcal{M}(\mathcal{B})$. Finally, given a matrix A, we shall denote by $\|A\|_\infty$ the largest absolute value of its entries.

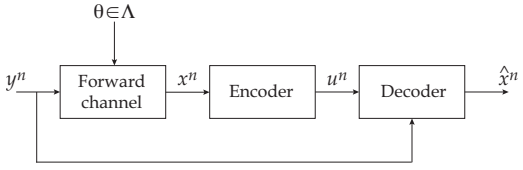

Fig. 1. Wyner–Ziv coding with known side information statistics and an unknown forward channel.

978-1-4244-2571-6/08/$25.00 ©2008 IEEE

Let $\mathcal{X}$, $\mathcal{Y}$ and $\widehat{\mathcal{X}}$ denote the finite source, side information, and reproduction alphabets, respectively. Just as in [3], we assume for simplicity that these alphabets are all the same:

$$\mathcal{X} = \mathcal{Y} = \widehat{\mathcal{X}} = \{\alpha_1, \ldots, \alpha_M\};$$

extensions to more general cases are straightforward. Let $\{(X_i, Y_i)\}_{i=1}^{\infty}$ be a sequence of independent drawings of a pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ of dependent random variables. We assume that the marginal distribution P_Y is known, $P_Y \equiv P$, while the conditional distribution of X given Y belongs to a given parametric family $\{P^{\theta}_{X|Y} : \theta \in \Lambda\} \subset \mathcal{C}(\mathcal{Y} \to \mathcal{X})$, where the parameter set Λ is an open subset of a metric space. Let $\rho(\cdot, \cdot)$ denote the corresponding metric. We shall refer to each $P^{\theta}_{X|Y}$ as a *forward channel*; the corresponding *backward channel* $P^{\theta}_{Y|X} \in \mathcal{C}(\mathcal{X} \to \mathcal{Y})$ can be computed using Bayes' rule:

$$P^{\theta}_{Y|X}(y|x) = \frac{P^{\theta}_{X|Y}(x|y)P(y)}{\sum_{y' \in \mathcal{Y}} P^{\theta}_{X|Y}(x|y')P(y')}.$$

We assume that, for each $\theta \in \Lambda$, the transition matrix of $P^{\theta}_{Y|X}$ is invertible and denote its inverse by Q^{θ}.

Let $d : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, d_{\max}]$, $d_{\max} < \infty$, be the single-letter distortion function (fidelity criterion). For each $\theta \in \Lambda$, we can characterize the region of rate-distortion pairs achievable by block WZ codes as follows: Given the rate R, a WZ code of block length l is a pair (E_l, D_l), where $E_l : \mathcal{X}^l \to \{1, 2, \ldots, \lceil 2^{lR} \rceil\}$ is the encoder and $D_l : \{1, 2, \ldots, \lceil 2^{lR} \rceil\} \times \mathcal{Y}^l \to \widehat{\mathcal{X}}^l$ is the decoder. The performance is measured by the expected per-symbol distortion between the source l-sequence X^l and the reproduction l-sequence $\widehat{X}^l = D_l(E_l(X^l), Y^l)$,

$$\mathbb{E}_{\theta}\left[d_l(X^l, \widehat{X}^l)\right] \triangleq \frac{1}{l}\,\mathbb{E}_{\theta}\left[\sum_{i=1}^{l} d(X_i, \widehat{X}_i)\right],$$

where the expectation is w.r.t. the joint distribution P^{θ}_{XY}. A rate-distortion pair (R, D) is *achievable* for $\theta \in \Lambda$ if, for any $\varepsilon > 0$, there exist l and (E_l, D_l) operating at rate R, such that

$$\mathbb{E}_{\theta}\left[d_l(X^l, \widehat{X}^l)\right] \le D + \varepsilon.$$

The WZ distortion-rate function $D^{\theta}_{\mathrm{WZ}}(R)$ for θ is defined as the infimum of all achievable distortions at rate R:

$$D^{\theta}_{\mathrm{WZ}}(R) \triangleq \inf\Big\{D : (R, D) \text{ is achievable for } \theta\Big\}.$$

Given $\{P^{\theta}_{XY}\}$, we are interested in a sequence of WZ codes that asymptotically attains $D^{\theta}_{\mathrm{WZ}}(R)$ for every $\theta \in \Lambda$ without prior knowledge of which θ is in effect.

In a recent paper [3], Jalali, Verdú and Weissman showed that one can achieve points in the interior of the WZ rate-distortion region of a discrete stationary ergodic source using deterministic sliding block codes. Adapted to our notation, the theorem states:

Theorem 2.1 ([3]). Let (R, D) be an interior point of the WZ rate-distortion region of $\theta \in \Lambda$. Let $\{(X_i, Y_i)\}_{i=-\infty}^{\infty}$ be a two-sided i.i.d. sequence drawn from P^{θ}_{XY}. Then, for any $\varepsilon > 0$, there exist a sliding block encoder $f : \mathcal{X}^{2k+1} \to \mathcal{U}$, where the auxiliary alphabet $\mathcal{U}$ satisfies $\log|\mathcal{U}| \ge R$, and a sliding block decoder $g : \mathcal{Y}^{2l+1} \times \mathcal{U}^{2m+1} \to \widehat{\mathcal{X}}$, such that

$$\mathbb{E}_{\theta}\left[d(X_0, g(Y_{-l}^{l}, W_{-m}^{m}))\right] \le D + \varepsilon, \tag{2.1}$$

where $W_i = f(X_{i-k}^{i+k})$ for all i, and

$$\overline{H}(\boldsymbol{W}) = \lim_{n\to\infty} \frac{1}{n} H(W^n) \le R - \varepsilon' \tag{2.2}$$

for some $\varepsilon' > 0$.

Formulating multiterminal source coding theorems in terms of deterministic sliding block codes [6] rather than the conventional stochastic block codes has several advantages. For one, the output of a sliding block coding of a stationary ergodic process is itself stationary ergodic. Also, with sliding block codes one can easily bound the mismatch due to coding one source with a code designed for another.

III. A UNIVERSAL WZ SCHEME

We now describe a WZ scheme which, under suitable regularity conditions, is universal w.r.t. the forward channel. Our approach is based on augmenting the recent scheme of [3] by a channel estimation step. In a nutshell, the encoder uses the source sequence to estimate the marginal distribution P^{θ}_X; provided that (a) the forward channel parameter θ is identifiable from P^{θ}_X, and (b) the backward channel is a sufficiently smooth function of θ, we can take the plug-in approach and use the scheme of [3] matched to the estimated backward channel. To keep the paper self-contained, we include a specification of the scheme of [3] in our description.

Let $\{k_n\}$, $\{l_n\}$, and $\{m_n\}$ be increasing sequences of positive integers, such that $k_n = o(n)$ and $t_n \triangleq \max\{l_n, m_n\}$ satisfies

$$t_n M^{2t_n} = o(\sqrt{n/\log n}); \tag{3.3}$$

let $\delta_1 = 1$ and $\delta_n = \sqrt{\log n / n}$ for $n > 1$. Given the rate R and the block length n, our scheme consists of the following two steps:

a) Channel estimation and encoding: Let s_n be a positive integer satisfying $1/s_n \le \delta_n/2M$, and let $\mathcal{T}_n$ be the set of all probability distributions on $\mathcal{X}$ that are of type s_n, i.e., the probability of each $x \in \mathcal{X}$ is an integer multiple of $1/s_n$. We then have that (a) $|\mathcal{T}_n| \le (s_n + 1)^M$ and (b) for any $\theta \in \Lambda$,

$$\min_{P \in \mathcal{T}_n} \left\|P^{\theta}_X - P\right\|_1 \le \frac{\delta_n}{2},$$

where (a) follows from elementary type-counting arguments [7], while (b) follows from the definition of $\|\cdot\|_1$ and from our definition of s_n. From this easily follows the existence of a set $\Lambda_n \subset \Lambda$ such that $|\Lambda_n| \le (s_n + 1)^M$ and

$$\min_{\theta' \in \Lambda_n} \left\|P^{\theta}_X - P^{\theta'}_X\right\|_1 \le \delta_n, \qquad \forall \theta \in \Lambda.$$

To each $\theta \in \Lambda_n$, we can associate a unique binary string $e_n(\theta)$ whose length is at most $M \log(s_n + 1)$ bits. Define the *minimum-distance estimator* (MDE) $\widehat{\theta}_n : \mathcal{X}^n \to \Lambda$ [8] by

$$\widehat{\theta}_n(x^n) \triangleq \arg\min_{\theta \in \Lambda_n} \left\|P^{\theta}_X - \mathsf{P}_{x^n}\right\|_1, \qquad x^n \in \mathcal{X}^n \tag{3.4}$$

with ties broken arbitrarily. Here, P_{x^n} denotes the type (empirical distribution) of $x^n \in \mathcal{X}^n$.

Given the source string x^n, the encoder first computes the MDE $\widehat{\theta}_n = \widehat{\theta}_n(x^n)$ of θ. Now, following [3], define the set

$$\mathcal{S}_n(x^n, R) \stackrel{\triangle}{=} \left\{ f : \mathcal{X}^{2k_n+1} \to \mathcal{U} \middle| \ell_{\mathrm{LZ}}(f(x^n)) \le nR \right\},$$

where $\ell_{\mathrm{LZ}}(\cdot)$ denotes the Lempel–Ziv encoding length [9, Ch. 13], and, for each f, $u^n = f(x^n)$ is defined as $u_i = f(x_{i-k_n}^{i+k_n})$, $k_n + 1 \le i \le n - k_n$, and $u_i = 0$ otherwise. For every $f \in \mathcal{S}_n(x^n, R)$, define next the quantity

$$V_n(f) \stackrel{\triangle}{=} \min \mathbb{E}_{\widehat{\theta}_n} \left[\sum_{i=k_n+1}^{n-k_n} d\left(x_i, g(Y_{i-l_n}^{i+l_n}, u_{i-m_n}^{i+m_n}) \right) \right],$$

where the minimization is over all mappings $g : \mathcal{Y}^{2l_n+1} \times \{1, 2, \ldots, \lceil 2^{nR} \rceil\}^{2m_n+1} \to \widehat{\mathcal{X}}$, and $u^n = f(x^n)$. The expectation is w.r.t. the estimated backward channel

$$P_{Y^n|x^n}^{\widehat{\theta}_n}(y^n|x^n) = \prod_{i=1}^{n} P_{Y|X}^{\widehat{\theta}_n}(y_i|x_i), \qquad \forall y^n \in \mathcal{Y}^n$$

where the conditioning is on the source sequence x^n. Let $f_n^* \in \mathcal{S}_n(x^n, R)$ be a minimizer of $V_n(f)$ over all $f \in \mathcal{S}_n(x^n, R)$. The encoder then communicates to the decoder a two-stage description $E_n^*(x^n)$ of x^n consisting of the binary encoding $e_n(\widehat{\theta}_n)$ of $\widehat{\theta}_n$, followed by the LZ encoding of $f_n^*(x^n)$.

b) Decoding: Again, following [3], decoding is based on an extension of the discrete universal denoiser (DUDE) [10] to the case when auxiliary information[1] is available. Roughly speaking, given an auxiliary information sequence u^n and assuming that the side information sequence y^n is generated from the source sequence x^n via the backward channel $P_{Y|X}^{\theta}$, the DUDE with auxiliary information amounts to simply applying the original DUDE [10] matched to the tensor-product channel $P_{Y|X}^{\theta} \otimes I \in \mathcal{C}(\mathcal{X} \times \mathcal{U} \to \mathcal{Y} \times \mathcal{U})$, where I is the noiseless channel on $\mathcal{U}$, to the combined sequence $((y_1, u_1), \ldots, (y_n, u_n))$. More precisely, for $t_n + 1 \le i \le n - t_n$, the ith denoised symbol $\widehat{x}_i$ is given by

$$\arg\min_{\widehat{x} \in \widehat{\mathcal{X}}} \boldsymbol{r}^T(y^n, u^n, y_{i-l_n}^{i-1}, y_{i+1}^{i+l_n}, u_{i-m_n}^{i+m_n}) Q^{\theta} \left[d_{\widehat{x}} \odot p_{y_i}^{\theta} \right].$$

In this expression, $\boldsymbol{r}(y^n, u^n, a^l, b^l, c_{-m}^m)$ is a column M-vector whose rth component is given by

$$\left| \left\{ t+1 \le j \le n-t : y_{j-l}^{j+l} = a^l \alpha_r b^l, u_{j-m}^{j+m} = c_{-m}^m \right\} \right|,$$

where $a^l \alpha_r b^l$ is a concatenation of the strings a^l, α_r, and b^l; $d_{\widehat{x}}$ is a column M-vector with components $d(\alpha_r, \widehat{x})$; $p_{y_i}^{\theta}$ is a column M-vector with components $P_{Y|X}^{\theta}(y_i|\alpha_r)$; and $\odot$ denotes componentwise multiplication of vectors. The remaining components of $\widehat{x}^n$ are arbitrary. We denote the resulting mapping from $\mathcal{Y}^n \times \mathcal{U}^n$ into $\widehat{\mathcal{X}}^n$ by $\widehat{X}_{\theta}^n$.

[1] In Ref. [3] this step is referred to as "denoising with side information." We opt for the term "auxiliary information" instead in order to conform to standard WZ terminology.

The decoder receives $E_n^*(x^n)$, decodes $\widehat{\theta}_n$ and $u^n = f_n^*(x^n)$, and then computes the reproduction

$$\widehat{x}^n \stackrel{\triangle}{=} \widehat{X}_{\widehat{\theta}_n}^n(y^n, u^n). \tag{3.5}$$

Theorem 3.1. Suppose that the family of discrete memoryless sources specified by the marginal distribution $P_Y = P$ and the forward channels $\{P_{X|Y}^{\theta} : \theta \in \Lambda\}$ satisfies the following three conditions:

(C1) $\inf_{\theta \in \Lambda} \left\| Q^{\theta} \right\|_{\infty} > 0$ and $\sup_{\theta \in \Lambda} \left\| Q^{\theta} \right\|_{\infty} < +\infty$.

(C2) For each $\theta \in \Lambda$, there are some $r_{\theta}, C_{\theta} > 0$, such that

$$\left\| P_X^{\theta'} - P_X^{\theta} \right\|_1 < r_{\theta} \quad \Rightarrow \quad \rho(\theta', \theta) < C_{\theta} \left\| P_X^{\theta'} - P_X^{\theta} \right\|_1.$$

(C3) For each $\theta \in \Lambda$, there are some $r'_{\theta}, C'_{\theta} > 0$, such that

$$\rho(\theta', \theta) < r'_{\theta} \quad \Rightarrow \quad \left\| P_{Y|X}^{\theta'} - P_{Y|X}^{\theta} \right\|_{\infty} < C'_{\theta} \rho(\theta', \theta).$$

Then the WZ scheme described above is *weakly minimax universal* for P and $\{P_{X|Y}^{\theta}\}$, i.e.,

$$\limsup_{n \to \infty} \mathbb{E}_{\theta} \left[\frac{1}{n} \sum_{i=1}^{n} d(X_i, \widehat{X}_i) \right] \le D_{\mathrm{WZ}}^{\theta}(R). \tag{3.6}$$

Remark 3.1. Condition (C1) is needed to ensure that the plug-in universal denoiser (3.5) eventually "zeroes in" on the correct backward channel. Condition (C2) states that the forward channel parameter θ should be identifiable from the marginal distribution P_X^{θ}. Finally, Condition (C3) ensures that the backward channel can be pinned down with sufficient accuracy given the knowledge of the forward channel parameter.

Proof: See Appendix. ∎

IV. An example

As an example, consider the binary alphabet case where the side information sequence is Bernoulli(p), i.e., for each i, $P(Y_i = 0) = p, P(Y_i = 1) = 1 - p$, and the source sequence is generated from the side information via a binary symmetric channel (BSC) with crossover probability θ. The backward channel corresponding to a given θ has the transition matrix

$$P_{Y|X}^{\theta} = \begin{pmatrix} \frac{(1-\theta)p}{(1-\theta)p + \theta(1-p)} & \frac{\theta(1-p)}{(1-\theta)p + \theta(1-p)} \\ \frac{\theta p}{\theta p + (1-\theta)(1-p)} & \frac{(1-\theta)(1-p)}{\theta p + (1-\theta)(1-p)} \end{pmatrix}. \tag{4.7}$$

This matrix is singular when $\theta = 1/2$ and when $p = 0, 1$. Hence, admissible pairs (θ, p) are contained in the set $\Omega = \Gamma \times (0, 1)$, where Γ is $[0, 1]$ with the point $\theta = 1/2$ deleted.

Lemma 4.1. For any $(\theta, p) \in \Omega$, the inverse $Q_{Y|X}^{\theta}$ of $P_{Y|X}^{\theta}$ can be written in the form

$$Q_{Y|X}^{\theta} = \begin{pmatrix} f_1(\theta) + \frac{1-p}{p} g(\theta) & -f_2(\theta) - \frac{1-p}{p} g(\theta) \\ -f_2(\theta) - \frac{p}{1-p} g(\theta) & f_1(\theta) + \frac{p}{1-p} g(\theta) \end{pmatrix},$$

where the functions f_1, f_2 and g satisfy

$$f_1(0) = 1, f_1(1) = 0, \lim_{\theta \to 1/2^-} f_1(\theta) = +\infty, \lim_{\theta \to 1/2^+} = -\infty$$

$$f_2(0) = 0, f_2(1) = -1, \lim_{\theta \to 1/2^-} f_2(\theta) = +\infty, \lim_{\theta \to 1/2^+} = -\infty$$

$$g(0) = g(1) = 0, \lim_{\theta \to 1/2^-} g(\theta) = +\infty, \lim_{\theta \to 1/2^+} = -\infty,$$

and are continuous and increasing on $[0, 1/2)$ and on $(1/2, 1]$.

From this lemma it follows that Condition (C1) will be met if $p \in (0,1)$ and if there exist some $0 < \zeta^- < \zeta^+ < 1/2 < \xi^- < \xi^+ < 1$, such that $\Lambda \subseteq [\zeta^-, \zeta^+] \cup [\xi^-, \xi^+]$. Turning now to Condition (C2), we observe that $\left\| P_X^\theta - P_X^{\theta'} \right\|_1 = 2|1-2p| \cdot |\theta - \theta'|$. Hence, Condition (C2) will hold with $\rho(\theta, \theta') = |\theta - \theta'|$ provided $p \neq 1/2$. Finally, we establish Condition (C3):

Lemma 4.2. For $\theta \in [0,1]$ and $p \in \operatorname{Int} \Gamma$, the entries of the transition matrix (4.7) of the backward channel $P_{Y|X}^\theta$ have continuous derivatives w.r.t. θ with absolute values bounded from above by $\max\{1/p, p/(1-p)\}$.

Using this lemma and the mean value theorem, we have, for $\theta, \theta' \in [0,1]$ and $p \in \operatorname{Int} \Gamma$,

$$\begin{aligned} \left\| P_{Y|X}^\theta - P_{Y|X}^{\theta'} \right\|_\infty &= \max_{x \in \{0,1\}} \left\| P_{Y|X}^\theta(\cdot|x) - P_{Y|X}^{\theta'}(\cdot|x) \right\|_1 \\ &\leq 2 \max\left\{ \frac{1}{p}, \frac{p}{1-p} \right\} |\theta - \theta'|. \end{aligned}$$

Summarizing all these observations, we obtain the following:

Proposition 4.1. Suppose the side information sequence is Bernoulli(p), and the source sequence is generated from the side information via a BSC with crossover probability θ. Then Conditions (C1)-(C3) are satisfied provided $p \notin \{0, 1/2, 1\}$ and there exist some $0 < \zeta^- < \zeta^+ < 1/2 < \xi^- < \xi^+ < 1$ such that $\theta \in \Lambda \subseteq [\zeta^-, \zeta^+] \cup [\xi^-, \xi^+]$.

Intuitively, Conditions (C1)-(C3) exclude the cases when either the side information sequence is uninformative about the source sequence ($p = 0$ or 1) or the source sequence is nearly uninformative about the side information sequence ($\theta \sim 1/2$).

V. Summary

We have presented a WZ coding scheme which is weakly minimax universal in the setting where the statistics of the side information are known, and there is uncertainty regarding the conditional distribution of the source sequence given the side information. This setting may be useful for certain problems related to lossy functional compression (and distributed source coding in general) or learning from compressed observations.

Appendix

The proof follows more or less the reasoning of [3], except that we also need to control the mismatch due to the channel estimation step, both at the encoder and at the decoder. Let $\theta \in \Lambda$ be the (unknown) active forward channel. Given the rate R and some $\varepsilon > 0$, $(R, D_{\mathrm{WZ}}^\theta(R) + \varepsilon)$ is an interior point of the WZ rate-distortion region for θ. Hence, Theorem 2.1 guarantees the existence of a sliding-block WZ encoder $f_\theta : \mathcal{X}^{2k+1} \to \mathcal{U}$, a sliding-block WZ decoder $g_\theta : \mathcal{Y}^{2l+1} \times \mathcal{U}^{2m+1} \to \widehat{\mathcal{X}}$, and some $\varepsilon' > 0$, such that (2.1) and (2.2) hold. The process $\{W_i\}$ obtained by a sliding block encoding of a stationary memoryless source $\{X_i\}$ is stationary and ergodic. Because LZ coding is universal for discrete stationary ergodic sources [9, Ch. 13], given any ω, there exists a sufficiently large N_ω, such that

$$\frac{1}{n} \ell_{\mathrm{LZ}}(W^n) \leq \overline{H}(\boldsymbol{W}) + \omega$$

for all $n > N_\omega$. Letting $\omega = \varepsilon'/2$, we then have

$$\frac{1}{n} \ell_{\mathrm{LZ}}(W^n) < R \tag{A.1}$$

for $n > N_\omega$. Hence, if the block length n is so large that (A.1) holds and $k_n > k, t_n > t \triangleq \max\{l, m\}$, then f_θ will belong to $\mathcal{S}_n(x^n, R)$ for any $x^n \in \mathcal{X}^n$. This implies, in turn, that

$$V_n(f_n^*) \leq V_n(f_\theta). \tag{A.2}$$

Now, let $\widehat{x}^n = \widehat{X}_{\widehat{\theta}_n}^n(y^n, u^n)$ be the reproduction sequence computed by the decoder. Then

$$\begin{aligned} \mathbb{E}_\theta \left[\sum_{i=k+1}^{n-k} d(x_i, \widehat{x}_i) \right] \leq \mathbb{E}_{\widehat{\theta}_n} \left[\sum_{i=k+1}^{n-k} d(x_i, \widehat{x}_i) \right] \\ + d_{\max}(n-2k)(2t_n+1) \left\| P_{Y|X}^\theta - P_{Y|X}^{\widehat{\theta}_n} \right\|_\infty, \end{aligned}$$

where expectations are taken w.r.t. the backward channel, while the source sequence is held fixed. Given any $\eta \in \Lambda$, the DUDE with auxiliary information, provided that $t_n M^{2t_n} = o(n/\log n)$ [which holds *a fortiori* for our choice of $\{l_n\}, \{m_n\}$ because of (3.3)], satisfies [3]

$$\begin{aligned} \lim_{n \to \infty} \Big[\frac{1}{n-2t_n} \sum_{i=t_n+1}^{n-t_n} d\left(x_i, \widehat{X}_\eta^n(Y^n, u^n)[i]\right) \\ - \frac{1}{n-2t_n} D_\eta^n(x^n, u^n) \Big] = 0, \ \ P^\eta - \text{a.s.} \end{aligned} \tag{A.3}$$

where

$$D_\eta^n(x^n, u^n) \triangleq \min_g \mathbb{E}_\eta \left[\sum_{i=t_n+1}^{n-t_n} d\left(x_i, g(Y_{i-l_n}^{i+l_n}, u_{i-m_n}^{i+m_n})\right) \right],$$

and the minimum is taken over all mappings $g : \mathcal{Y}^{2l_n+1} \times \mathcal{U}^{2m_n+1} \to \widehat{\mathcal{X}}$. Moreover, using the Borel–Cantelli lemma together with a large-deviation bound for the universal denoiser from Theorem 2 in [11], namely

$$\begin{aligned} P^\eta \Big\{ \frac{1}{n-2t_n} \sum_{i=t_n+1}^{n-t_n} d\left(x_i, \widehat{X}_\eta^n(Y^n, u^n)[i]\right) > \\ \frac{1}{n-2t_n} D_\eta^n(x^n, u^n) + \varepsilon \Big\} \\ \leq F(M, t_n, d_{\max}, \|Q^\eta\|_\infty) e^{-nG(\varepsilon, t_n, d_{\max}, \|Q^\eta\|_\infty)}, \end{aligned}$$

where $F(\cdot)$ and $G(\cdot)$ are both increasing functions of $\|Q^\eta\|_\infty^{-1}$, we see that, provided Condition (C1) holds, the convergence in (A.3) is *uniform* in $\eta \in \Lambda$. Thus, for n sufficiently large,

$$\begin{aligned} \frac{1}{n-2k} \mathbb{E}_{\widehat{\theta}_n} \left[\sum_{i=k+1}^{n-k} d(x_i, \widehat{x}_i) \right] \leq \frac{2d_{\max}(k_n - k)}{n-2k} + \\ \min_g \frac{1}{n-2k} \mathbb{E}_{\widehat{\theta}_n} \left[\sum_{i=k_n+1}^{n-k_n} d\left(x_i, g(Y_{i-l}^{i+l}, u_{i-m}^{i+m})\right) \right] + \varepsilon, \end{aligned}$$

where the minimization is over all mappings $g : \mathcal{Y}^{2l+1} \times \mathcal{U}^{2m+1} \to \widehat{\mathcal{X}}$. Next, recalling that $u^n = f_n^*(x^n)$ and using (A.2), we can write

$$\begin{aligned}\min_g \mathbb{E}_{\widehat{\theta}_n}\left[\sum_{i=k_n+1}^{n-k_n} d\left(x_i, g(Y_{i-l}^{i+l}, u_{i-m}^{i+m})\right)\right] &= V_n(f_n^*)\\ &\le V_n(f_\theta)\\ &\le \mathbb{E}_{\widehat{\theta}_n}\left[\sum_{i=k+1}^{n-k} d\left(x_i, g_\theta(Y_{i-l}^{i+l}, w_{i-m}^{i+m})\right)\right],\end{aligned}$$

where $w^n = f_\theta(x^n)$. Now, we also have

$$\begin{aligned}&\mathbb{E}_{\widehat{\theta}_n}\left[\sum_{i=k+1}^{n-k} d\left(x_i, g_\theta(Y_{i-l}^{i+l}, w_{i-m}^{i+m})\right)\right] \le\\ &\quad \mathbb{E}_\theta\left[\sum_{i=k+1}^{n-k} d\left(x_i, g_\theta(Y_{i-l}^{i+l}, w_{i-m}^{i+m})\right)\right]\\ &\quad +d_{\max}(n-2k)(2t+1)\left\|P_{Y|X}^\theta - P_{Y|X}^{\widehat{\theta}_n}\right\|_\infty.\end{aligned}$$

Putting all of this together, we have, for n sufficiently large,

$$\begin{aligned}\frac{1}{n-2k}\mathbb{E}_\theta\left[\sum_{i=k+1}^{n-k} d(x_i, \widehat{x}_i)\right] &\le \frac{2d_{\max}(k_n-k)}{n-2k}\\ &+\frac{1}{n-2k}\mathbb{E}_\theta\left[\sum_{i=k+1}^{n-k} d(x_i, g_\theta(Y_{i-l}^{i+l}, w_{i-m}^{i+m}))\right]\\ &+2d_{\max}(2t_n+1)\left\|P_{Y|X}^\theta - P_{Y|X}^{\widehat{\theta}_n}\right\|_\infty + \varepsilon.\end{aligned}$$

By the ergodic theorem, with probability one there exists some $N > 0$, such that for all $n > N$,

$$\begin{aligned}&\frac{1}{n-2k}\mathbb{E}_\theta\left[\sum_{i=k+1}^{n-k} d\left(x_i, g_\theta(Y_{i-l}^{i+l}, w_{i-m}^{i+m})\right)\right]\\ &\quad\le \mathbb{E}_\theta\left[d(X_0, g_\theta(Y_{-l}^l, W_{-m}^m))\right] + \varepsilon\\ &\quad\le D_{\mathrm{WZ}}^\theta(R) + 3\varepsilon.\end{aligned}$$

Thus, eventually almost surely,

$$\begin{aligned}\frac{1}{n}\mathbb{E}_\theta\left[\sum_{i=1}^{n} d(x_i, \widehat{x}_i)\right] &\le D_{\mathrm{WZ}}^\theta(R)\\ &+\frac{4d_{\max}k_n}{n} + 2(2t_n+1)\left\|P_{Y|X}^\theta - P_{Y|X}^{\widehat{\theta}_n}\right\|_\infty + 4\varepsilon.\end{aligned}$$

Moreover, if n is so large that $\delta_n < r_\theta$ and $C_\theta\delta_n < r'_\theta$, then, assuming Conditions (C2) and (C3) hold, then, on the event that $\left\|P_X^\theta - P_X^{\widehat{\theta}_n}\right\|_1 \le \delta_n$, we shall have

$$(2t_n+1)\left\|P_{Y|X}^\theta - P_{Y|X}^{\widehat{\theta}_n}\right\|_\infty < C'_\theta C_\theta(2t_n+1)\delta_n.$$

To bound the probability of this event, we use the following:

Lemma A.1. Let $X^n = (X_1, \ldots, X_n)$ be an i.i.d. sequence drawn from P_X^θ. Then the minimum-distance estimator (3.4) satisfies

$$P^\theta\left(\left\|P_X^{\widehat{\theta}_n(X^n)} - P_X^\theta\right\|_1 > \delta_n\right) \le \frac{8M}{n\delta_n^2} + \delta_n.$$

Proof: Given θ, there exists $\theta^* \in \Lambda_n$ such that $\left\|P_X^\theta - P_X^{\theta^*}\right\|_1 \le \delta_n$. We begin by observing that

$$\begin{aligned}\left\|P_X^\theta - P_X^{\widehat{\theta}_n(x^n)}\right\|_1 &\overset{(a)}{\le} \left\|P_X^\theta - \mathsf{P}_{x^n}\right\|_1 + \left\|\mathsf{P}_{x^n} - P_X^{\widehat{\theta}_n(x^n)}\right\|_1\\ &\overset{(b)}{\le} \left\|P_X^\theta - \mathsf{P}_{x^n}\right\|_1 + \left\|\mathsf{P}_{x^n} - P_X^{\theta^*}\right\|_1,\end{aligned}$$

where (a) follows from the triangle inequality and (b) from the definition of $\widehat{\theta}_n$. Hence,

$$\begin{aligned}P^\theta\left(\left\|P_X^{\widehat{\theta}_n} - P_X^\theta\right\|_1 > \delta_n\right) &\le P^\theta\left(\left\|\mathsf{P}_{X^n} - P_X^\theta\right\|_1 > \frac{\delta_n}{2}\right)\\ &\quad +P^\theta\left(\left\|\mathsf{P}_{X^n} - P_X^{\theta^*}\right\|_1 > \frac{\delta_n}{2}\right)\\ &\overset{(a)}{\le} 2\sup_{\theta\in\Lambda} P^\theta\left(\left\|\mathsf{P}_{X^n} - P_X^\theta\right\|_1 > \frac{\delta_n}{2}\right) + \left\|P_X^\theta - P_X^{\theta^*}\right\|_1\\ &\overset{(b)}{\le} \frac{8M}{n\delta_n^2} + \delta_n,\end{aligned}$$

where (a) follows from the definition of $\|\cdot\|_1$ and (b) from the Chebyshev inequality and the definition of θ^*. ■

Using the lemma with our choice of $\{\delta_n\}$, $\{k_n\}$ and $\{t_n\}$ and taking expectations w.r.t. P^θ, we get (3.6), and the theorem is proved.

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. IT-22, no. 1, pp. 1–10, January 1976.

[2] 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.

[3] S. Jalali, S. Verdú, and T. Weissman, "A universal Wyner–Ziv scheme for discrete sources," in *Proc. IEEE Int. Symp. Inform. Theory*, Nice, France, June 2007, pp. 1951–1955.

[4] V. Doshi, D. Shah, and M. Médard, "Source coding with distortion through graph coloring," in *Proc. Int. IEEE Symp. on Inform. Theory*, Nice, France, June 2007, pp. 1501–1505.

[5] M. Raginsky, "Learning from compressed observations," in *Proc. IEEE Inform. Theory Workshop*, Lake Tahoe, CA, September 2007, pp. 420–425.

[6] J. C. Kieffer, "A method for proving multiterminal source coding theorems," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 5, pp. 565–570, September 1981.

[7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Sources*. Budapest: Akadémiai Kiadó, 1981.

[8] J. Wolfowitz, "The minimum distance method," *Ann. Math. Statist.*, vol. 28, no. 1, pp. 75–88, 1957.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[10] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. J. Weinberger, "Discrete universal denoising: known channel," *IEEE Trans. Inform. Theory*, vol. 51, no. 1, pp. 5–28, January 2005.

[11] A. Dembo and T. Weissman, "Universal denoising for the finite-input general-output channel," *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1507–1517, April 2005.