Coding Theorems for Reversible Embedding

Frans Willems and Ton Kalker
T.U. Eindhoven, Philips Research

DIMACS, March 16-19, 2003

Outline

1. Gelfand-Pinsker coding theorem

2. Noise-free embedding

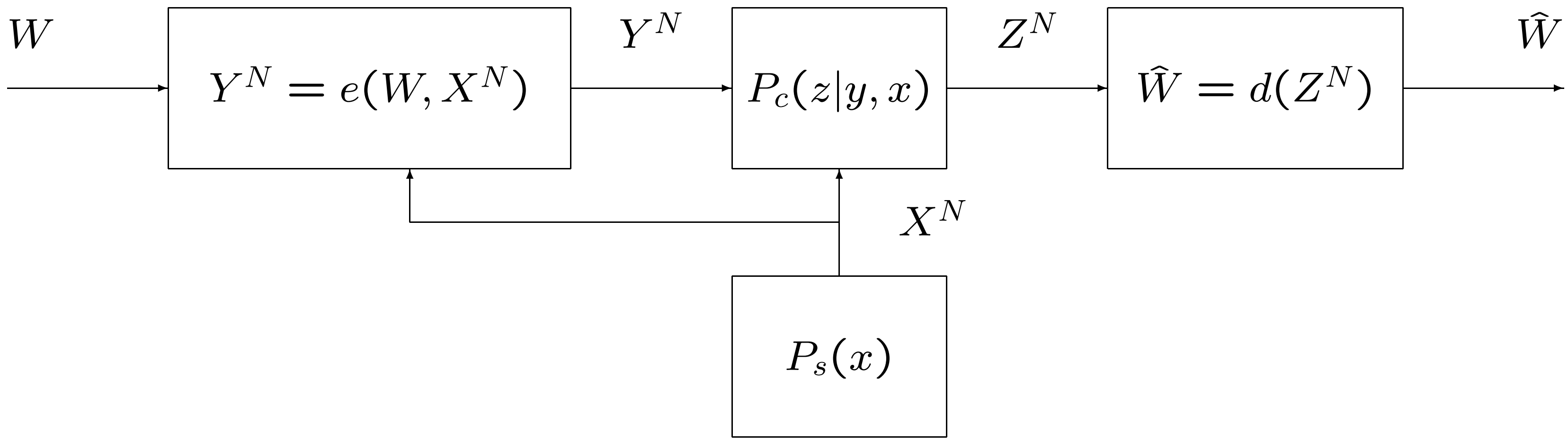
3. Reversible embedding

4. Robust and reversible embedding

5. Partially reversible embedding

6. Remarks

I. The Gelfand-Pinsker Coding Theorem



Messages: $\Pr\{W = w\} = 1/M$ for $w \in \{1, 2, \cdots, M\}$.

Side information: $\Pr\{X^N = x^N\} = \prod_{n=1}^{N} P_s(x_n)$ for $x^N \in \mathcal{X}^N$.

Channel: discrete memoryless $\{\mathcal{Y} \times \mathcal{X}, P_c(z|y, x), \mathcal{Z}\}$.

Error probability: $P_{\mathcal{E}} = \Pr\{\widehat{W} \neq W\}$.

Rate: $R = \frac{1}{N} \log_2(M)$.

Capacity

The **side-information capacity** $C_{\textsf{si}}$ is the largest ρ such that for all $\epsilon > 0$ there exist for all large enough N encoders and decoders with $R \geq \rho - \epsilon$ and $P_{\mathcal{E}} \leq \epsilon$.

THEOREM (Gelfand-Pinsker [1980]):

$$C_{\textsf{si}} = \max_{P_t(u,y|x)} I(U;Z) - I(U;X). \quad (1)$$

Achievability proof: Fix a **test-channel** $P_t(u,y|x)$. Consider sets $\mathcal{A}_\epsilon(\cdot)$ of **strongly typical sequences**, etc.

(a) For each message index $w \in \{1, \cdots, 2^{NR}\}$, generate 2^{NR_u} sequences u^N at random according to $P(u) = \sum_{x,y} P_s(x)P_t(u,y|x)$. Give these sequences the label w.

(b) When message index w has to be transmitted choose a sequence u^N having label w such that $(u^N, x^N) \in \mathcal{A}_\epsilon(U,X)$. Such a sequence exists almost always if $R_u > I(U;X)$ (roughly).

(c) The input sequence y^N results from applying the "channel" $P(y|u,x) = P_t(y,u|x)/\sum_y P_t(u,y|x)$ to u^N and x^N. Then y^N is transmitted.

(d) The decoder upon receiving z^N, looks for the unique sequence u^N such that $(u^N, z^N) \in \mathcal{A}_\epsilon(U,Z)$. If $R + R_u < I(U;Z)$ (roughly) such a unique sequence exists. The message index is the label of u^N.

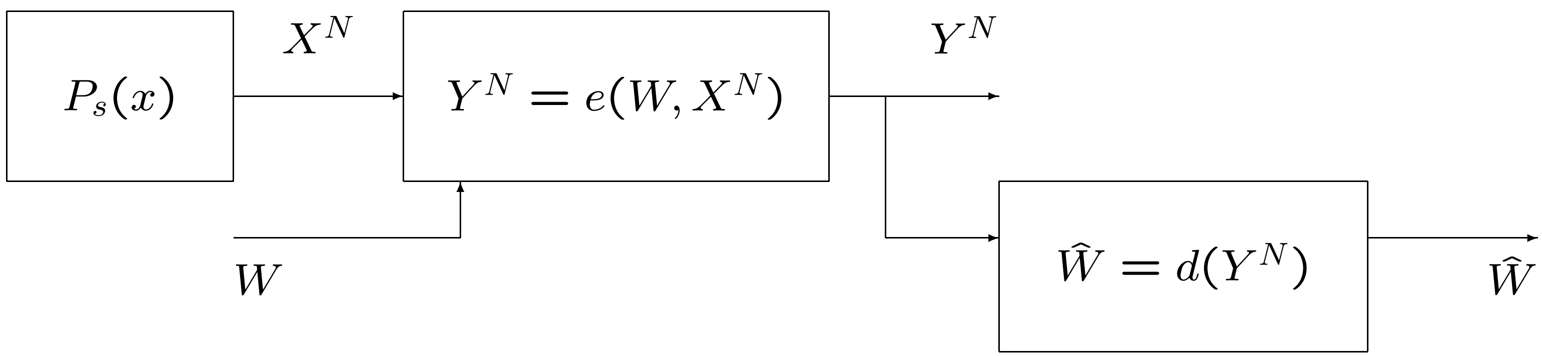
Conclusion is that $R < I(U;Z) - I(U;X)$ is achievable.

Observations

A: As an intermediate result the **decoder recovers the sequence** u^N.

B: The **transmitted** u^N **is jointly typical with the side-info sequence** x^N, i.e. $(u^N, x^N) \in \mathcal{A}_\epsilon(U,X)$ thus **their joint composition is OK**. Note that $P(u,x) = \sum_y P_s(x)P_t(u,y|x)$.

II. Noise-free Embedding



Messages: $\Pr\{W = w\} = \frac{1}{M}$ for $w \in \{1, 2, \cdots, M\}$.

Source (host): $\Pr\{X^N = x^N\} = \Pi_{n=1,N} P_s(x_n)$ for $x^N \in \mathcal{X}^N$.

Error probability: $P_{\mathcal{E}} = \Pr\{\widehat{W} \neq W\}$.

Rate: $R = \frac{1}{N} \log_2(M)$.

Embedding distortion: $\overline{D_{xy}} = E[\frac{1}{N} \sum_{n=1,N} D_{xy}(X_n, e_n(W, X^N))]$ for some distortion matrix $\{D_{xy}(x, y), x \in \mathcal{X}, y \in \mathcal{Y}\}$.

Achievable region noise-free embedding

A rate-distortion pair (ρ, Δ_{xy}) is said to be achievable if for all $\epsilon > 0$ there exists for all large enough N encoders and decoders such that

$$\begin{aligned} R &\geq \rho - \epsilon, \\ \overline{D_{xy}} &\leq \Delta_{xy} + \epsilon, \\ P_{\mathcal{E}} &\leq \epsilon. \end{aligned}$$

THEOREM (Chen [2000], Barron [2000]): The set of achievable rate-distortion pairs is equal to $\mathcal{G}_{\text{nfe}}$ which is defined as

$$\begin{aligned} \mathcal{G}_{\text{nfe}} = \{(\rho, \Delta_{xy}) \quad : \quad & 0 \leq \rho \leq H(Y|X), \\ & \Delta_{xy} \geq \sum_{x,y} P(x,y) D_{xy}(x,y), \\ & \text{for } P(x,y) = P_s(x) P_t(y|x)\}. \end{aligned} \qquad (2)$$

Again $\{\mathcal{X}, P_t(y|x), \mathcal{Y}\}$ is called **test-channel**.

Proof:

Achievability: In the Gelfand-Pinsker achievability proof, note that $Z = Y$ (noiseless channel) and take the auxiliary random variable $U = Y$. Then $(x^N, y^N) \in \mathcal{A}_\epsilon(X, Y)$ hence $\overline{D_{xy}}$ is OK. For the embedding rate we obtain

$$R = I(U;Z) - I(U;X) = I(Y;Y) - I(Y;X) = H(Y|X).$$

Converse: Rate part:

$$\begin{aligned}
\log_2(M) &\leq H(W) - H(W|\widehat{W}) + \text{Fano term} \\
&\leq H(W|X^N) - H(W|X^N, Y^N) + \text{Fano term} \\
&= I(W; Y^N|X^N) + \text{Fano term} \\
&\leq H(Y^N|X^N) + \text{Fano term} \\
&\leq \sum_{n=1,N} H(Y_n|X_n) + \text{Fano term} \\
&\leq NH(Y|X) + \text{Fano term},
\end{aligned}$$

where X and Y are random variables with

$$\Pr\{(X,Y)=(x,y)\} = \frac{1}{N}\sum_{n=1,N}\Pr\{(X_n,Y_n)=(x,y)\},$$

for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Note that for $x \in \mathcal{X}$

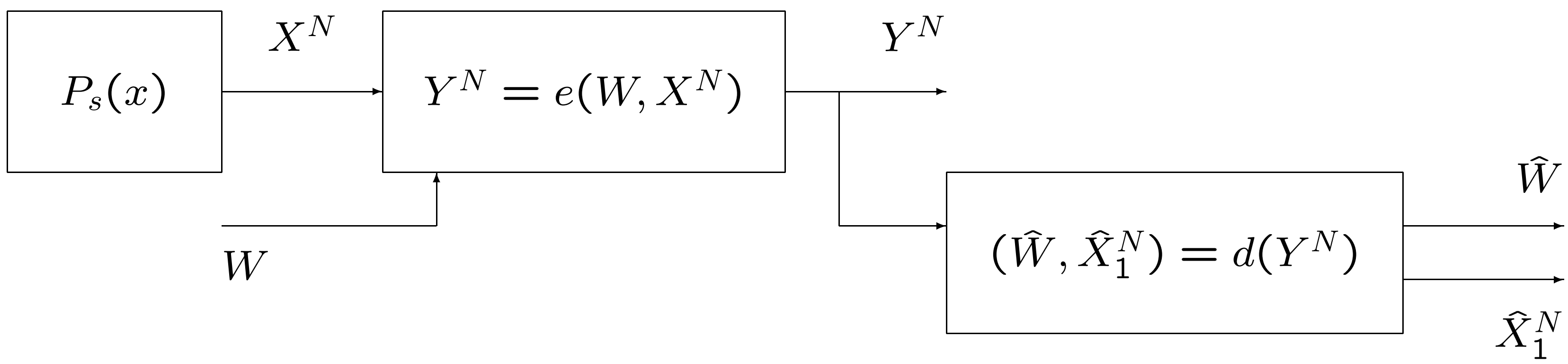
$$\Pr\{X=x\} = P_s(x).$$

Distortion part:

$$\begin{aligned}\overline{D_{xy}} &= \sum_{x^N,y^N}\Pr\{(X^N,Y^N)=(x^N,y^N)\}\frac{1}{N}\sum_n D_{xy}(x_n,y_n)\\ &= \sum_{x,y}\Pr\{(X,Y)=(x,y)\}D_{xy}(x,y).\end{aligned}$$

Let $P_{\mathcal{E}} \downarrow 0$, etc.

III. Reversible Embedding



Messages: $\Pr\{W = w\} = \frac{1}{M}$ for $w \in \{1, 2, \cdots, M\}$.

Source (host): $\Pr\{X^N = x^N\} = \Pi_{n=1,N} P_s(x_n)$ for $x^N \in \mathcal{X}^N$.

Error probability: $P_{\mathcal{E}} = \Pr\{\widehat{W} \neq W \vee \widehat{X}_1^N \neq X^N\}$.

Rate: $R = \frac{1}{N} \log_2(M)$.

Embedding distortion: $\overline{D_{xy}} = E[\frac{1}{N} \sum_{n=1,N} D_{xy}(X_n, e_n(W, X^N))]$ for some distortion matrix $\{D_{xy}(x, y), x \in \mathcal{X}, y \in \mathcal{Y}\}$.

Inspired by Fridrich, Goljan, and Du, "Lossless data embedding for all image formats," *Proc. SPIE, Security and Watermarking of Multimedia Contents*, San Jose, CA, 2002.

Achievable region for reversible embedding

A rate-distortion pair (ρ, Δ_{xy}) is said to be achievable if for all $\epsilon > 0$ there exists for all large enough N encoders and decoders such that

$$\begin{aligned} R &\geq \rho - \epsilon, \\ \overline{D_{xy}} &\leq \Delta_{xy} + \epsilon, \\ P_{\mathcal{E}} &\leq \epsilon. \end{aligned}$$

RESULT (Kalker-Willems [2002]): The set of achievable rate-distortion pairs is equal to $\mathcal{G}_{\text{re}}$ which is defined as

$$\begin{aligned} \mathcal{G}_{\text{re}} = \{(\rho, \Delta_{xy}) \;:\; & 0 \leq \rho \leq H(Y) - H(X), \\ & \Delta_{xy} \geq \sum_{x,y} P(x,y) D_{xy}(x,y), \\ & \text{for } P(x,y) = P_s(x) P_t(y|x)\}. \end{aligned} \quad (3)$$

Note that $\{\mathcal{X}, P_t(y|x), \mathcal{Y}\}$ is the test channel.

Proof:

Achievability: In the Gelfand-Pinsker achievability proof, note that $Z = Y$ (noiseless channel) and take the auxiliary random variable $U = [X, Y]$. Then x^N can be reconstructed by the decoder and $(x^N, y^N) \in \mathcal{A}_\epsilon(X, Y)$ hence $\overline{D_{xy}}$ is OK. For the embedding rate we obtain

$$R = I(U; Z) - I(U; X) = I([X, Y]; Y) - I([X, Y]; X) = H(Y) - H(X).$$

Converse: Rate part:

$$\begin{aligned}
\log_2(M) &\leq H(W) - H(W, X^N | \widehat{W}, \widehat{X}_1^N) + \text{Fano term} \\
&= H(W, X^N) - H(W, X^N | \widehat{W}, \widehat{X}_1^N) - H(X^N) + \text{Fano term} \\
&\leq H(W, X^N) - H(W, X^N | Y^N, \widehat{W}, \widehat{X}_1^N) - H(X^N) + \text{Fano term} \\
&= I(W, X^N; Y^N) - H(X^N) + \text{Fano term} \\
&= H(Y^N) - H(X^N) + \text{Fano term} \\
&\leq \sum_{n=1,N} [H(Y_n) - H(X_n)] + \text{Fano term} \\
&\leq N[H(Y) - H(X)] + \text{Fano term},
\end{aligned}$$

where X and Y are random variables with

$$\Pr\{(X,Y) = (x,y)\} = \frac{1}{N} \sum_{n=1,N} \Pr\{(X_n, Y_n) = (x,y)\},$$

for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Note that for $x \in \mathcal{X}$

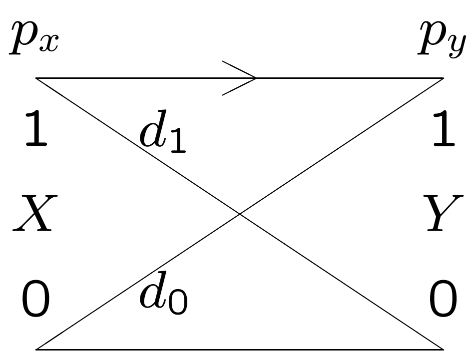
$$\Pr\{X = x\} = P_s(x).$$

Distortion part:

$$\begin{aligned} \overline{D_{xy}} &= \sum_{x^N, y^N} \Pr\{(X^N, Y^N) = (x^N, y^N)\} \frac{1}{N} \sum_n D_{xy}(x_n, y_n) \\ &= \sum_{x,y} \Pr\{(X,Y) = (x,y)\} D_{xy}(x,y). \end{aligned}$$

Let $P_{\mathcal{E}} \downarrow 0$, etc.

Example: Binary source, Hamming distortion



Since

$$\Delta_{xy} \geq p_x d_1 + (1 - p_x) d_0$$
$$p_y = p_x(1 - d_1) + (1 - p_x) d_0$$

we can write

$$p_y \leq \Delta_{xy} + p_x(1 - 2d_1).$$

Assume w.l.o.g. that $p_x \leq 1/2$. First let Δ_{xy} be such that $\Delta_{xy} + p_x \leq 1/2$ or $\Delta_{xy} \leq 1/2 - p_x$. Then we have

$$p_y \leq \Delta_{xy} + p_x \leq 1/2,$$

and hence

$$\rho \le h(p_y) - h(p_x) \le h(p_x + \Delta_{xy}) - h(p_x).$$

However $\rho = h(p_x + \Delta_{xy}) - h(p_x)$ is achievable with Δ_{xy} by taking

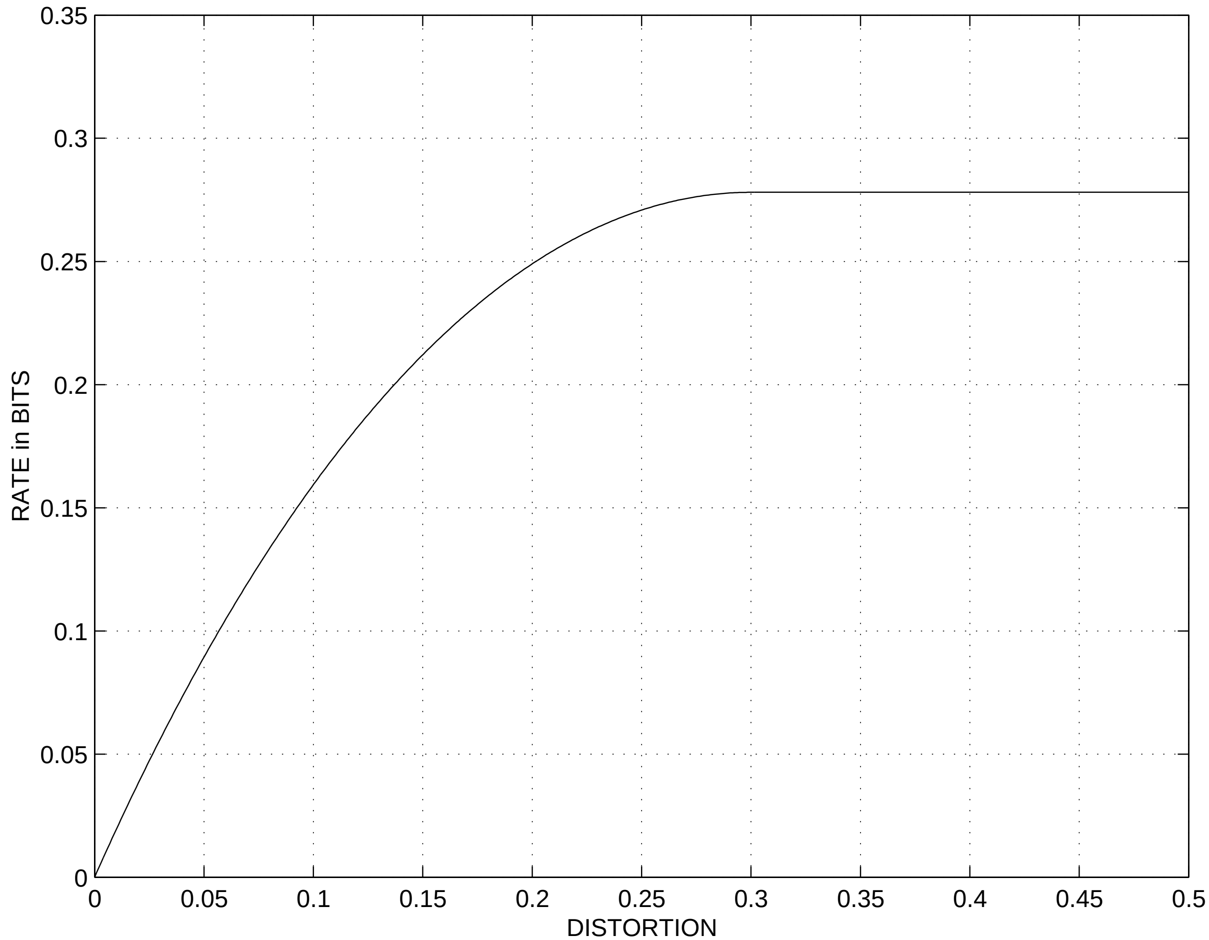
$$d_1 = 0 \quad \text{and} \quad d_0 = \frac{\Delta_{xy}}{1 - p_x}.$$

Note that the test channel is **not symmetric** and that

$$d_0 = \frac{\Delta_{xy}}{1 - p_x} \le \frac{1/2 - p_x}{1 - p_x} \le 1/2.$$

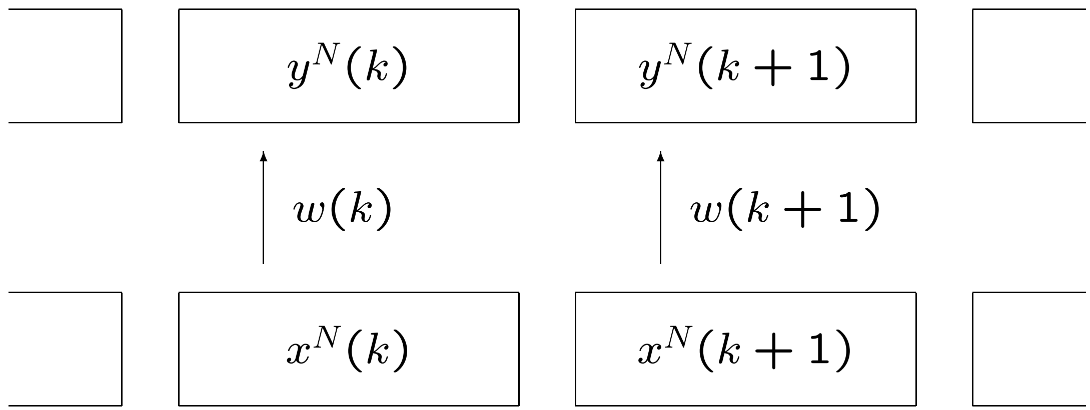
For $\Delta_{xy} + p_x \ge 1/2$ the rate is bounded as $\rho \le 1 - h(p_x)$ but also achievable.

Plot of rate-distortion region $\mathcal{G}_{\text{re}}$



Horizontal axis Δ_{xy}, vertical axis ρ, for $p_x = 0.2$. Maximum embedding rate $1 - h(0.2) \approx 0.278$.

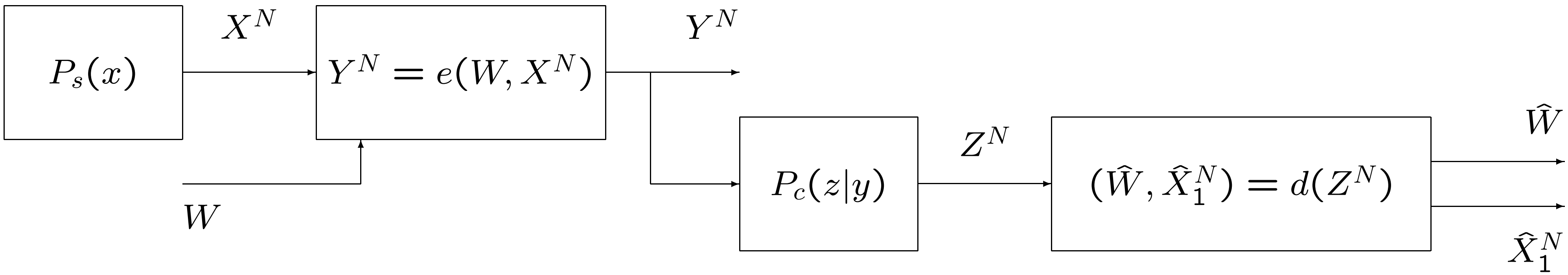
Another perspective



Consider a blocked system with blocks of length N.
In block k message bits can be (noise-free) embedded with rate $H(Y|X)$ and corresponding distortion.
Then in block $k+1$ message bits are embedded that allow for reconstruction of $x^N(k)$ given $y^N(k)$. This requires $NH(X|Y)$ bits. Therefore the resulting embedding rate is

$$\begin{aligned} R &= H(Y|X) - H(X|Y) \\ &= H(Y,X) - H(X) - H(X|Y) \\ &= H(Y) - H(X). \end{aligned}$$

IV. Robust and Reversible Embedding



Messages: $\Pr\{W = w\} = \frac{1}{M}$ for $w \in \{1, 2, \cdots, M\}$.

Source (host): $\Pr\{X^N = x^N\} = \Pi_{n=1,N} P_s(x_n)$ for $x^N \in \mathcal{X}^N$.

Channel: discrete memoryless $\{\mathcal{Y}, P_c(z|y), \mathcal{Z}\}$.

Error probability: $P_{\mathcal{E}} = \Pr\{\hat{W} \neq W \vee \hat{X}_1^N \neq X^N\}$.

Rate: $R = \frac{1}{N} \log_2(M)$.

Embedding distortion: $\overline{D_{xy}} = E[\frac{1}{N} \sum_{n=1,N} D_{xy}(X_n, e_n(W, X^N))]$ for some distortion matrix $\{D_{xy}(x,y), x \in \mathcal{X}, y \in \mathcal{Y}\}$.

Achievable region for robust and reversible embedding

A rate-distortion pair (ρ, Δ_{xy}) is said to be achievable if for all $\epsilon > 0$ there exists for all large enough N encoders and decoders such that

$$\begin{aligned} R &\geq \rho - \epsilon, \\ \overline{D_{xy}} &\leq \Delta_{xy} + \epsilon, \\ P_{\mathcal{E}} &\leq \epsilon. \end{aligned}$$

RESULT (Willems-Kalker [2003]): The set of achievable rate-distortion pairs is equal to $\mathcal{G}_{\text{rre}}$ which is defined as

$$\begin{aligned} \mathcal{G}_{\text{rre}} = \{(\rho, \Delta_{xy}) \ : \ & 0 \leq \rho \leq I(Y;Z) - H(X), \\ & \Delta_{xy} \geq \sum_{x,y} P(x,y) D_{xy}(x,y), \\ & \text{for } P(x,y,z) = P_s(x) P_t(y|x) P_c(z|y)\}. \end{aligned} \qquad (4)$$

Proof:

Achievability: In the Gelfand-Pinsker achievability proof again take the auxiliary random variable $U = [X, Y]$. Then x^N can be reconstructed by the decoder and since $(x^N, y^N) \in \mathcal{A}_\epsilon(X, Y)$ the embedding distortion $\overline{D_{xy}}$ is OK. For the embedding rate we obtain

$$R = I(U; Z) - I(U; X) = I([X, Y]; Z) - I([X, Y]; X) = I(Y; Z) - H(X).$$

Converse: Rate part:

$$\begin{aligned}
\log_2(M) &\leq H(W) - H(W, X^N | \widehat{W}, \widehat{X}_1^N) + \text{Fano term} \\
&= H(W, X^N) - H(W, X^N | \widehat{W}, \widehat{X}_1^N) - H(X^N) + \text{Fano term} \\
&\leq H(W, X^N) - H(W, X^N | Z^N, \widehat{W}, \widehat{X}_1^N) - H(X^N) + \text{Fano term} \\
&= I(W, X^N; Z^N) - H(X^N) + \text{Fano term} \\
&= I(Y^N; Z^N) - H(X^N) + \text{Fano term} \\
&\leq \sum_{n=1,N} [I(Y_n; Z_n) - H(X_n)] + \text{Fano term} \\
&\leq N[I(Y; Z) - H(X)] + \text{Fano term},
\end{aligned}$$

where X, Y and Z are random variables with

$$\text{Pr}\{(X,Y,Z) = (x,y,z)\} = \frac{1}{N} \sum_{n=1,N} \text{Pr}\{(X_n, Y_n, Z_n) = (x,y,z)\},$$

for $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathcal{Z}$. Note that for $x \in \mathcal{X}$

$$\text{Pr}\{X = x\} = P_s(x).$$

and for $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$

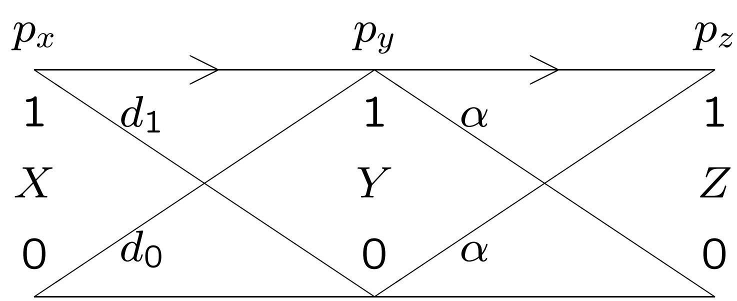
$$\text{Pr}\{Z = z|Y = y\} = P_c(z|y).$$

Distortion part:

$$\begin{aligned} \overline{D_{xy}} &= \sum_{x^N, y^N} \text{Pr}\{(X^N, Y^N) = (x^N, y^N)\} \frac{1}{N} \sum_n D_{xy}(x_n, y_n) \\ &= \sum_{x,y} \text{Pr}\{(X,Y) = (x,y)\} D_{xy}(x,y). \end{aligned}$$

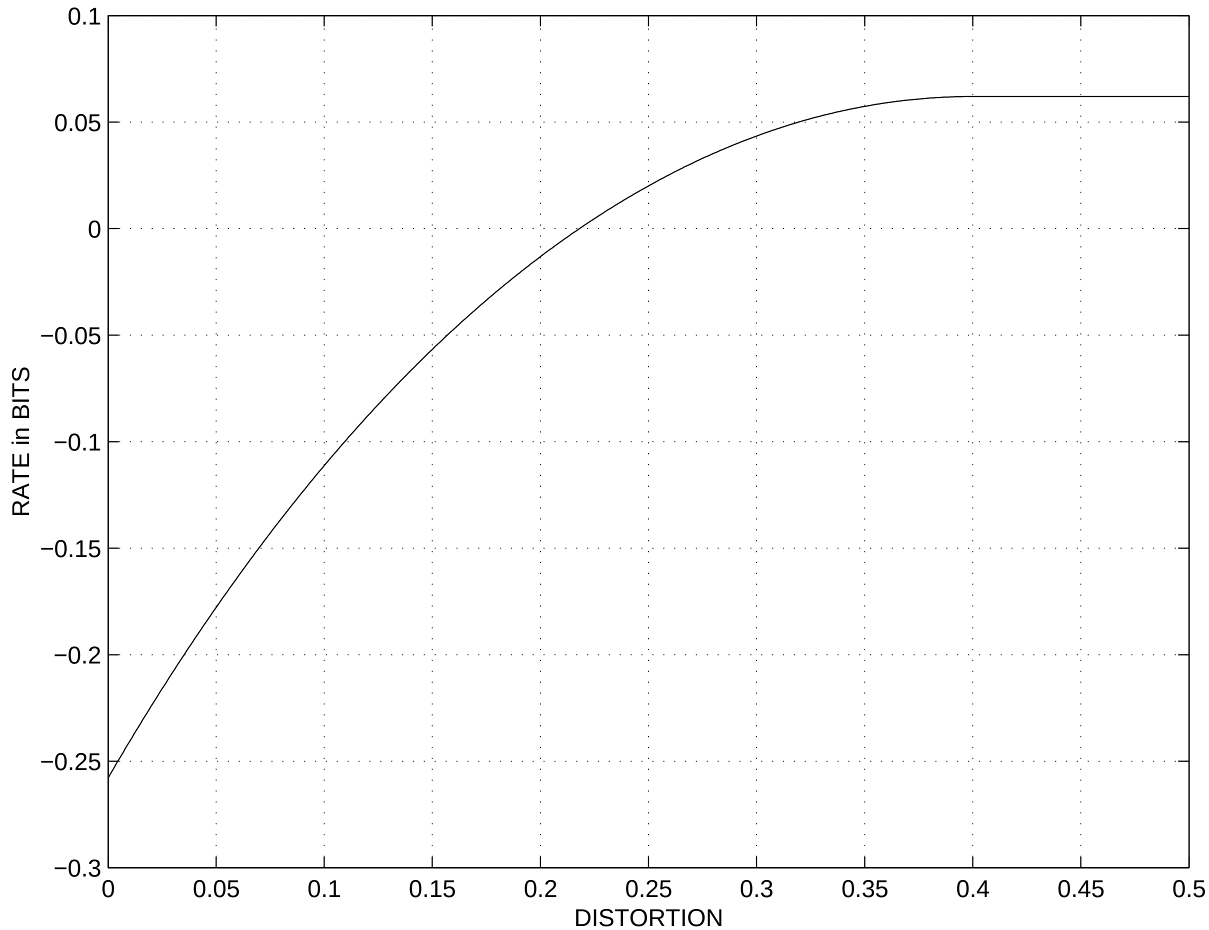
Let $P_{\mathcal{E}} \downarrow 0$, etc.

Example: Binary source, Hamming distortion, binary symmetric channel



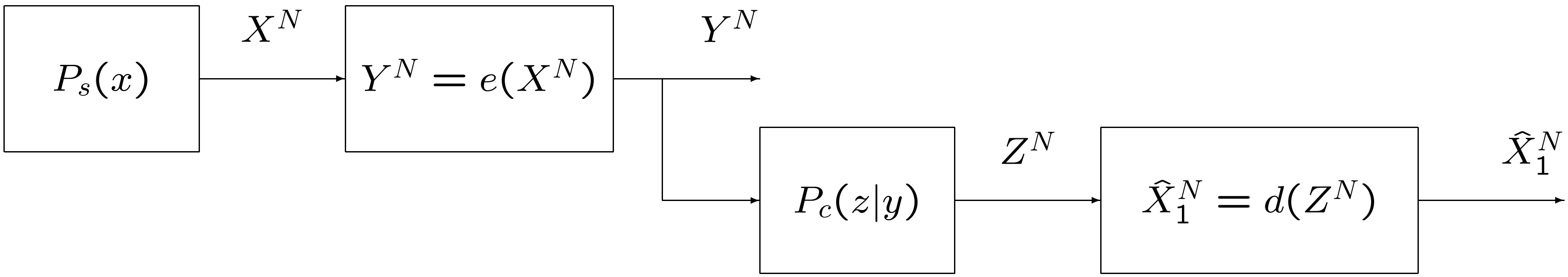
Similar analysis as before.

Plot of achievable region $\mathcal{G}_{\text{rre}}$



Horizontal axis Δ_{xy}, vertical axis ρ for $p_x = \alpha = 0.1$. Minimal distortion ≈ 0.218, maximum embedding rate $1 - h(0.1) - h(0.1) \approx 0.062$.

The zero-rate case: Robustification



Source (host): $\Pr\{X^N = x^N\} = \Pi_{n=1,N} P_s(x_n)$ for $x^N \in \mathcal{X}^N$.

Channel: discrete memoryless $\{\mathcal{Y}, P_c(z|y), \mathcal{Z}\}$.

Error probability: $P_{\mathcal{E}} = \Pr\{\hat{X}_1^N \neq X^N\}$.

Robustification distortion: $\overline{D_{xy}} = E[\frac{1}{N}\sum_{n=1,N} D_{xy}(X_n, e_n(W, X^N))]$ for some distortion matrix $\{D_{xy}(x,y), x \in \mathcal{X}, y \in \mathcal{Y}\}$.

Achievable distortions for robustification

A distortion Δ_{xy} is said to be achievable if for all $\epsilon > 0$ there exists for all large enough N encoders and decoders such that

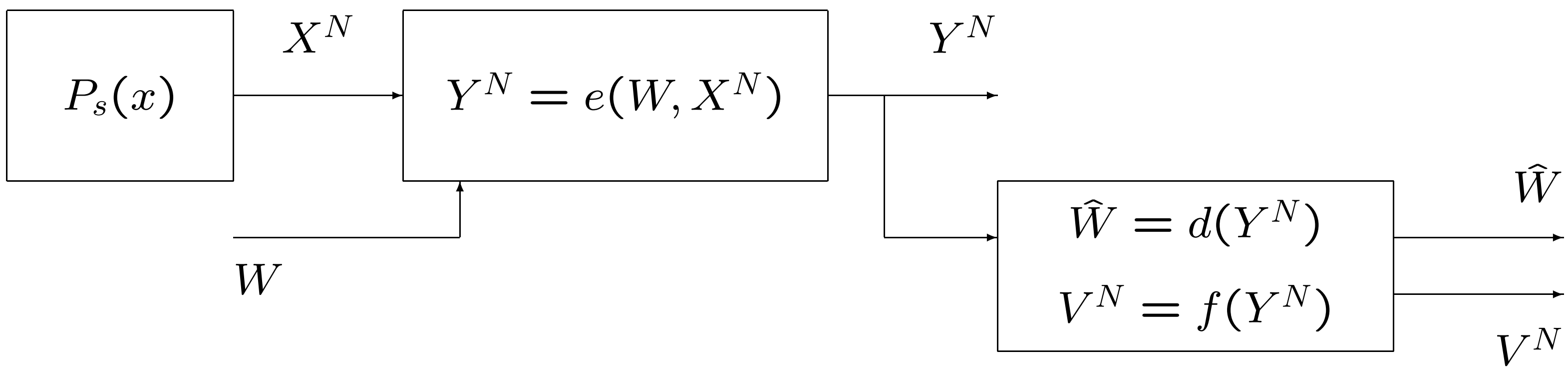
$$\begin{aligned} \overline{D_{xy}} &\leq \Delta_{xy} + \epsilon, \\ P_{\mathcal{E}} &\leq \epsilon. \end{aligned}$$

RESULT: The set of achievable distortions is equal to $\mathcal{G}_{\text{rob}}$ which is defined as

$$\begin{aligned} \mathcal{G}_{\text{rob}} = \{\Delta_{xy} \ : \ & \Delta_{xy} \geq \sum_{x,y} P(x,y) D_{xy}(x,y), \\ & \text{for } P(x,y,z) = P_s(x) P_t(y|x) P_c(z|y) \\ & \text{such that } H(X) \leq I(Y;Z)\}. \end{aligned} \quad (5)$$

Related to Shannon's separation principle ! Robustification is not possible if $H(X) > \max_{P_t(y)} I(Y;Z)$.

V. Partially Reversible Embedding



Messages: $\Pr\{W = w\} = \frac{1}{M}$ for $w \in \{1, 2, \cdots, M\}$.

Source (host): $\Pr\{X^N = x^N\} = \prod_{n=1,N} P_s(x_n)$ for $x^N \in \mathcal{X}^N$.

Error probability: $P_{\mathcal{E}} = \Pr\{\widehat{W} \neq W\}$.

Rate: $R = \frac{1}{N} \log_2(M)$.

Embedding distortion: $\overline{D_{xy}} = E[\frac{1}{N}\sum_{n=1,N} D_{xy}(X_n, e_n(W, X^N))]$ for some distortion matrix $\{D_{xy}(x,y), x \in \mathcal{X}, y \in \mathcal{Y}\}$.

Restoration distortion: $\overline{D_{xv}} = E[\frac{1}{N}\sum_{n=1,N} D_{xv}(X_n, f_n(Y^N))]$ for a distortion matrix $\{D_{xv}(x,v), x \in \mathcal{X}, z \in \mathcal{V}\}$.

Achievable region for partially reversible embedding

A rate-distortion triple $(\rho, \Delta_{xy}, \Delta_{xv})$ is said to be achievable if for all $\epsilon > 0$ there exists for all large enough N encoders and decoders such that

$$\begin{aligned} R &\geq \rho - \epsilon, \\ \overline{D_{xy}} &\leq \Delta_{xy} + \epsilon, \\ \overline{D_{xv}} &\leq \Delta_{xv} + \epsilon, \\ P_{\mathcal{E}} &\leq \epsilon. \end{aligned}$$

RESULT (Willems-Kalker [2002]): The set of achievable rate-distortion triples is given by $\mathcal{G}_{\text{pre}}$ which is defined as

$$\begin{aligned} \mathcal{G}_{\text{pre}} = \{(\rho, \Delta_{xy}, \Delta_{xv}) \ : \ & 0 \leq \rho \leq H(Y) - I(X;Y,V), \\ & \Delta_{xy} \geq \sum_{x,y,v} P(x,y,v) D_{xy}(x,y), \\ & \Delta_{xv} \geq \sum_{x,y,v} P(x,y,v) D_{xv}(x,v), \\ & \text{for } P(x,y,v) = P_s(x) P_t(y,v|x)\}. \end{aligned} \tag{6}$$

Proof:

Achievability: In the Gelfand-Pinsker achievability proof, note again that $Z = Y$ (noiseless channel) and take the auxiliary random variable $U = [Y, V]$. Then v^N can be reconstructed by the decoder and since $(x^N, y^N, v^N) \in \mathcal{A}_\epsilon(X, Y, V)$ both $\overline{D_{xy}}$ and $\overline{D_{xv}}$ are OK. For the embedding rate we obtain

$$R = I(U;Z) - I(U;X) = I([Y,V];Y) - I([Y,V];X) = H(Y) - I(X;Y,V).$$

Converse: Rate part:

$$\begin{aligned}
\log_2(M) &= H(Y^N, V^N, W) - H(Y^N, V^N|W) \\
&\leq H(Y^N) + H(W|\widehat{W}) - I(Y^N, V^N; X^N|W) \\
&\leq H(Y^N) - H(X^N|W) + H(X^N|W, Y^N, V^N) + \text{Fano term} \\
&\leq H(Y^N) - H(X^N) + H(X^N|Y^N, V^N) + \text{Fano term} \\
&\leq \sum_{n=1,N} H(Y_n) - \sum_{n=1,N} H(X_n) + \sum_{n=1,N} H(X_n|Y_n, V_n) + \text{Fano term} \\
&\leq N[H(Y) - H(X) + H(X|Y,V)] + \text{Fano term} \\
&= N[H(Y) - I(X;Y,V)] + \text{Fano term},
\end{aligned}$$

where X, Y and V are random variables with

$$\Pr\{(X,Y,V) = (x,y,v)\} = \frac{1}{N} \sum_{n=1,N} \Pr\{(X_n, Y_n, V_n) = (x,y,v)\},$$

for $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $v \in \mathcal{V}$. Note that for $x \in \mathcal{X}$

$$\Pr\{X = x\} = P_s(x).$$

Distortion parts:

$$\begin{aligned} \overline{D_{xy}} &= \sum_{x^N, y^N} \Pr\{(X^N, Y^N) = (x^N, y^N)\} \frac{1}{N} \sum_n D_{xy}(x_n, y_n) \\ &= \sum_{x,y} \Pr\{(X,Y) = (x,y)\} D_{xy}(x,y). \end{aligned}$$

$$\begin{aligned} \overline{D_{xv}} &= \sum_{x^N, v^N} \Pr\{(X^N, V^N) = (x^N, v^N)\} \frac{1}{N} \sum_n D_{xv}(x_n, v_n) \\ &= \sum_{x,v} \Pr\{(X,V) = (x,v)\} D_{xv}(x,v). \end{aligned}$$

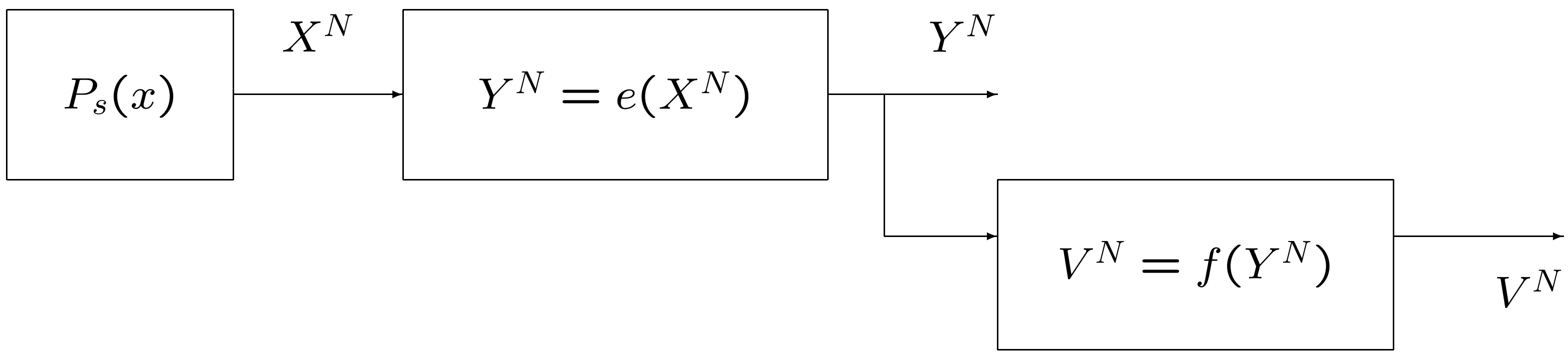
Let $P_{\mathcal{E}} \downarrow 0$, etc.

The other perspective again

Consider a blocked system with blocks of length N.
In block k a message can be (noise-free) embedded with rate $H(Y|X)$ and the corresponding distortion.
Then in block $k+1$ data is embedded that specifies a restoration sequence $v^N(k)$ given $y^N(k)$. This requires $NI(X;V|Y)$ bits. Therefore the remaining embedding rate is

$$\begin{aligned} R &= H(Y|X) - I(X;V|Y) \\ &= H(Y|X) - H(X|Y) + H(X|Y,V) \\ &= H(Y) - H(X) + H(X|Y,V) = H(Y) - I(X;Y,V). \end{aligned}$$

The zero-rate case: Self-Embedding



Source (host): $\Pr\{X^N = x^N\} = \prod_{n=1,N} P_s(x_n)$ for $x^N \in \mathcal{X}^N$.

Embedding distortion: $\overline{D_{xy}} = E[\frac{1}{N}\sum_{n=1,N} D_{xy}(X_n, e_n(W, X^N))]$ for some distortion matrix $\{D_{xy}(x,y), x \in \mathcal{X}, y \in \mathcal{Y}\}$.

Restoration distortion: $\overline{D_{xv}} = E[\frac{1}{N}\sum_{n=1,N} D_{xv}(X_n, f_n(Y^N))]$ for a distortion matrix $\{D_{xv}(x,v), x \in \mathcal{X}, z \in \mathcal{V}\}$.

Achievable distortions for self-embedding

A distortion pair $(\Delta_{xy}, \Delta_{xv})$ is said to be achievable if for all $\epsilon > 0$ there exists for all large enough N encoders and decoders such that

$$\begin{aligned} \overline{D_{xy}} &\leq \Delta_{xy} + \epsilon, \\ \overline{D_{xv}} &\leq \Delta_{xv} + \epsilon. \end{aligned}$$

RESULT (Willems-Kalker [2002]): The set of achievable distortion pairs is equal to $\mathcal{G}_{\text{se}}$ which is defined as

$$\begin{aligned} \mathcal{G}_{\text{se}} = \{(\Delta_{xy}, \Delta_{xv}) \ : \ & \Delta_{xy} \geq \sum_{x,y,v} P(x,y,v) D_{xy}(x,y), \\ & \Delta_{xv} \geq \sum_{x,y,v} P(x,y,v) D_{xv}(x,v), \\ & \text{for } P(x,y,v) = P_s(x) P_t(y,v|x) \\ & \text{such that } H(Y) \geq I(X;Y,V)\}. \end{aligned} \tag{7}$$

Self-embedding is putting a vector quantizer into a scalar quantizer. Or making an abstract index to a restoration vector v^N meaningful.

VI. Remarks

1. Our results are related to results of Sutivong, Cover, et al. Slightly different setups however. Embedding distortion.

2. We cannot do the partially reversible AND robust case. An achievable region would be similar to the Sutivong, Cover, Chiang, Kim [2002] region. No converse.

3. Coding techniques for the reversible case have been studied (with Deran Maas [2002]).

4. Open problems: (A) Arimoto-Blahut methods to compute the rate-distortion functions, (B) Coding techniques, especially for the zero-rate cases.