Source Coding with Side Information

Stark Draper
scd@allegro.mit.edu

6.962, Week 2
Spring 2001
2/15/01

Thanks to: R. Barron for use of figures

Outline

I. **Problem Statement**

II. Practical Approach

III. Lossless Side-Information

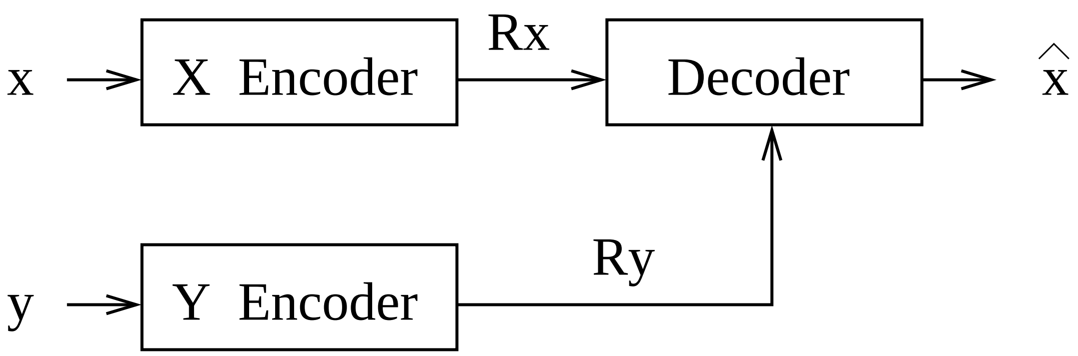
IV. Lossy Side-Information

V. Gaussian and Binary Sources

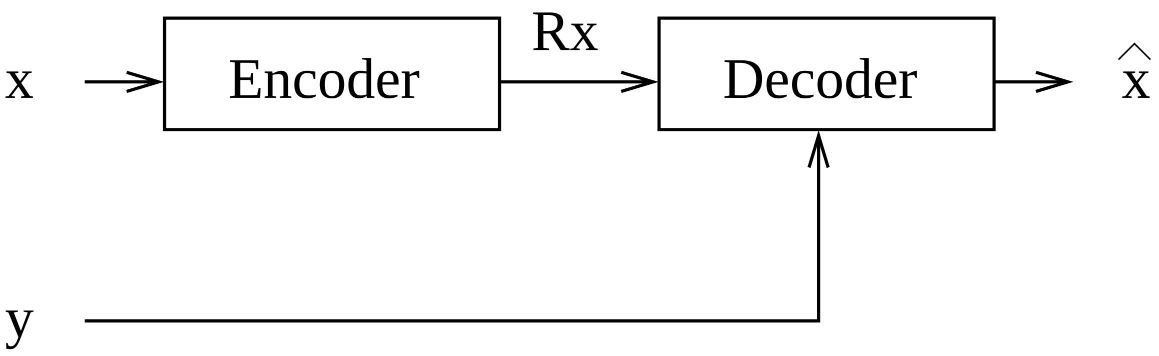
VI. Duality with Info Embedding

We use Side-Information to Help Estimate x

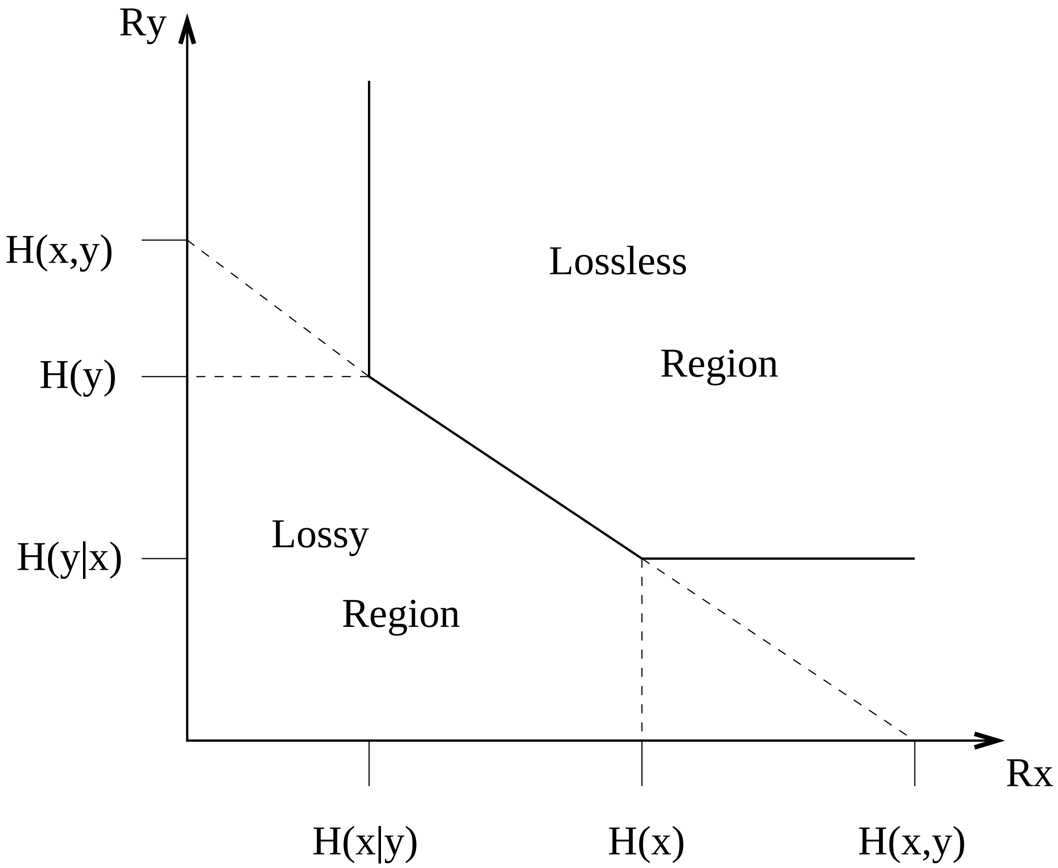
Lossless



Lossy



Joint Source Coding Splits into Lossy and Lossless Regions



Source Coding with Side-Info Has Many Applications

- Hybrid Analog/Digital Broadcasts (Radio, TV)

- Distributed Sensor Systems (Near/Far Sensors)

- Speech Enhancement

- Information Embedding

Outline

I. Problem Statement

II. **Practical Approach**

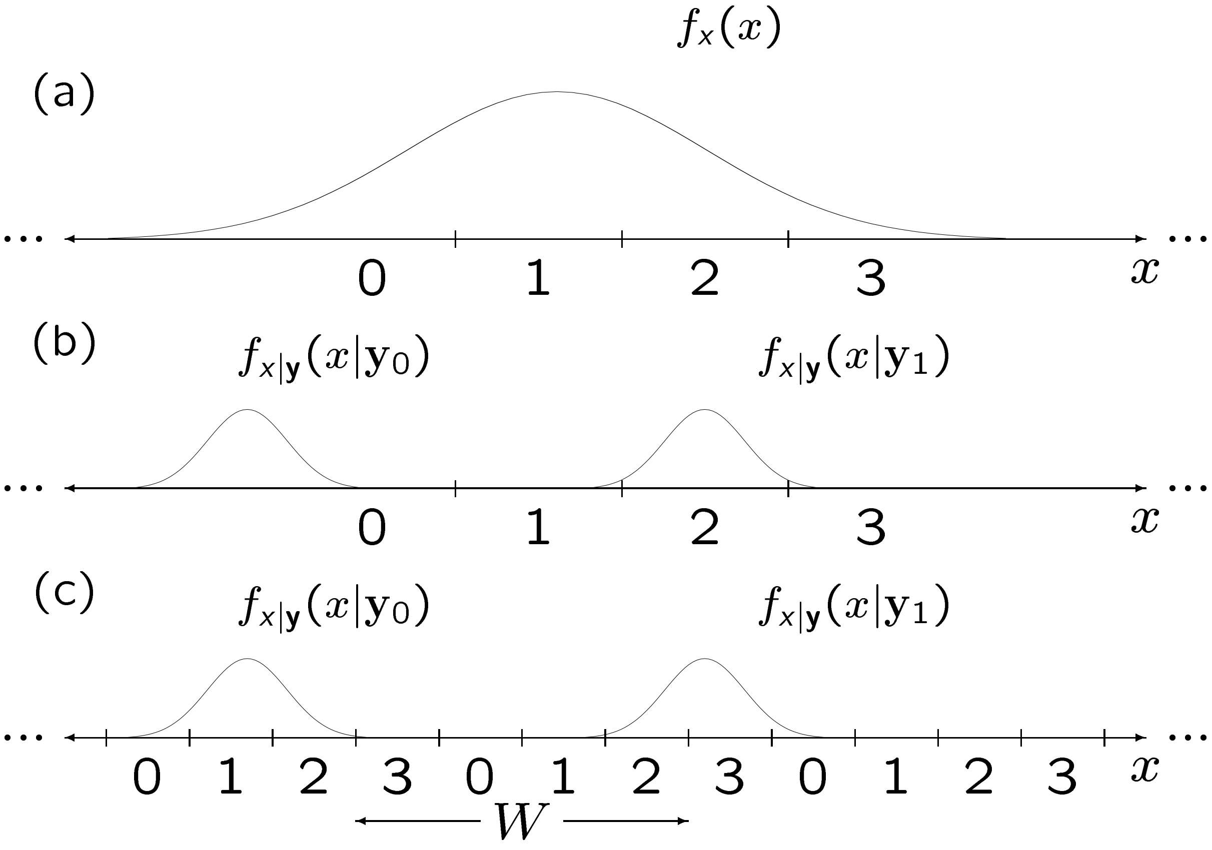
III. Lossless Side-Information

IV. Rate-Distortion Side-Information

V. Gaussian and Binary Sources

VI. Duality with Info Embedding

We can Build Quantizers to Exploit Side-Info*



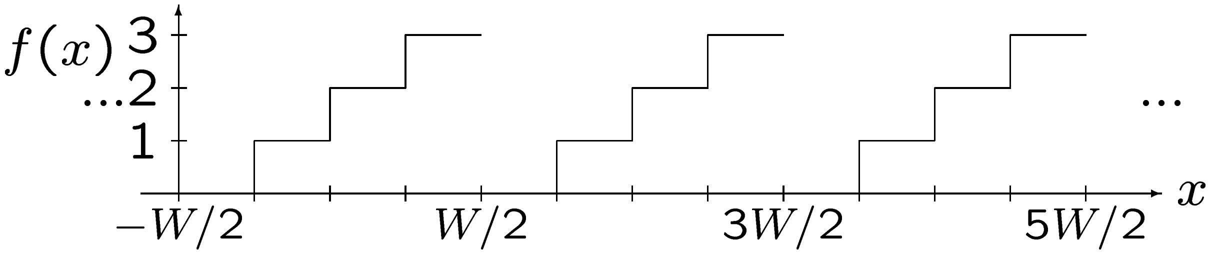
(a) Prior $f_x(x)$, standard quantizer.

(b) A posteriori $f_{x|\mathbf{y}}(x|\mathbf{y})$, standard quant.

(c) A posteriori $f_{x|\mathbf{y}}(x|\mathbf{y})$, side-info quant.

*Figure from R. Barron '00

Side Information Scalar Quantizers are Step Functions*



Encoder map, $f(x)$, exploiting side information. A 2 bit scalar example.

Another approach: Low-bit modulation.

*Figure from R. Barron '00

Outline

I. Problem Statement

II. Practical Approach

III. **Lossless Side-Information**

IV. Lossy Side-Information

V. Gaussian and Binary Sources

VI. Duality with Info Embedding

Coding Theorem – Lossless Case

Given (1) a pair of sources **x**, **y**

(2) $p(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} p(x_i, y_i)$

If (a) **x** is encoded at rate R_x,

(b) **y** is encoded at rate R_y,

Then we can recover **x** with arbitrarily small Pr(error)

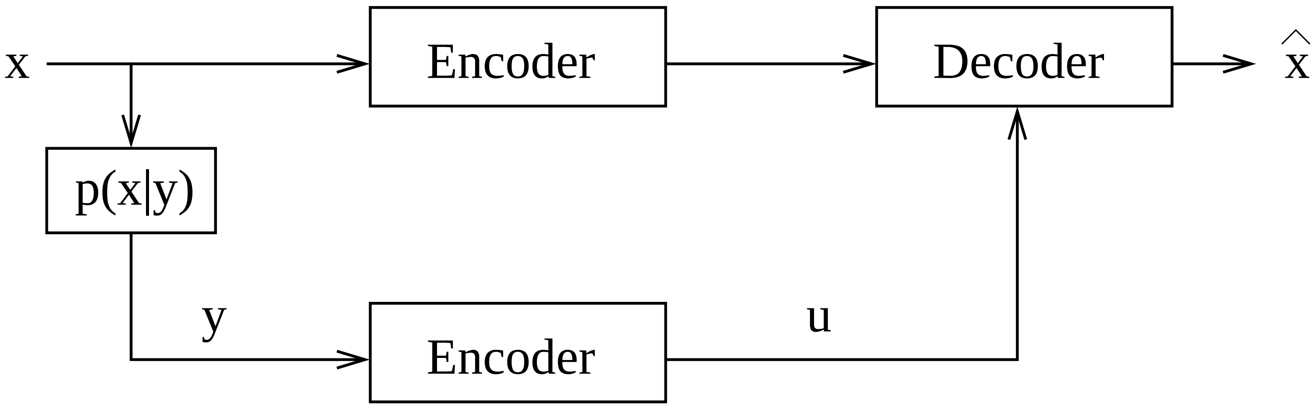
iff $\exists$ an auxiliary random variable u s.t.:

(i) $R_x \geq H(x|u)$,

(ii) $R_y \geq I(y; u)$,

(iii) $x \rightarrow y \rightarrow u$ (Markov).

Single-Letter View of Lossless Side-Information



Coding Theorem — Building Intuition

- Encode x conditioned upon knowledge *expected* about x from y.

- $I(y; u) = H(y) - H(y|u)$
  - Special Case (Slepian-Wolf): $u = y$

    $\rightarrow R_x = H(x|y)$.

    $\rightarrow R_y = H(y)$.
  - Generally $R_x \simeq H(x|u) \geq H(x|y)$.
  - Generally $R_y \simeq I(y; u) < H(y)$ since redundancy between x and y.

- Like Slepian-Wolf, but no estimate $\widehat{y}^n$.

Break Proof into 5 Steps

---

1. Random Binning — for $\mathbf{x}$

2. Random Side-Info Codebooks — for $\mathbf{y}$

3. Encode Separately

4. Decode Jointly

    - Typical Set Decoding

    - Markov Lemma (strong typicality)

5. Show Prob. of Error $\rightarrow 0$

Randomly Bin Source Words

- Uniformly assign each x^n sequence to one of $2^{nR_x} \simeq 2^{nH(x|u)}$ bins.

- $B(i) = \{x^n\}$ in the i^{th} bin, $i \in \{1, 2, \ldots, 2^{nR_x}\}$.

- Each bin contains roughly $\frac{2^{nH(x)}}{2^{nH(x|u)}} = 2^{nI(x;u)}$ typical **x**-sequences.

Randomly Generated Side-Info Codebook

- Generate $2^{nR_y} \simeq 2^{nI(y;u)}$ independent length-n codewords according to $\prod_{i=1}^{n} p(u)$.

- Index codewords $u^n(j)$, $j \in \{1, 2, \ldots, 2^{nR_y}\}$.

- $R_y \geq I(y;u) \geq I(x;u)$
  $\rightarrow$ By $x \rightarrow y \rightarrow u$ and Data Processing
  $\rightarrow$ Plausible you can do selection.

Encoders Act Separately

- Source Encoder:

  Transmit index i_0 of bin in which x^n falls.

- Side-Info Encoder:

  Look for indices $\{j\}$ of codewords $\{u^n(j)\}$ s.t. $(y^n, u^n(j)) \in T_\epsilon$, jointly typical.

  If only one such j transmit that j_0, else transmit any such j.

Decoder Acts Jointly

- Looks for unique $x^n \in B(i_0)$ s.t. $(x^n, u^n(j_0)) \in T_\epsilon$, jointly typical.

- If not unique, then error.

Break Error Analysis into 4 Sub-errors

Sub-errors:

$$
\begin{aligned}
E_0: &\quad (x^n, y^n) \notin T_\epsilon \\
E_1: &\quad (y^n, u^n(j)) \notin T_\epsilon \ \forall\ j \\
E_2: &\quad \exists\ \tilde{x}^n \in B(i_0) \text{ s.t.} \\
&\quad (\tilde{x}^n, u^n(j_0)) \in T_\epsilon \\
E_3: &\quad (y^n, u^n(j_0)) \in T_\epsilon, \text{ but} \\
&\quad (x^n, u^n(j_0)) \notin T_\epsilon
\end{aligned}
$$

We show $\Pr(E) \to 0$ where

$$
\begin{aligned}
E &= E_0 \cup E_1 \cup E_2 \cup E_3 \\
\Pr(E) &\leq \Pr(E_0) + \sum_{i=1}^{3} \Pr(E_i \cap E_0^c)
\end{aligned}
$$

$\Pr[E_0] < \epsilon$ **if** n **Big**

Error: $(x^n, y^n) \notin T_\epsilon$

For long block length, $\Pr(E_0) \to 0$.

$\Pr[E_1 \cap E_0^c] < \epsilon$ **if** R_y **Big**

Error:	$(y^n, u^n(j)) \notin T_\epsilon \; \forall \; j$
Set-up:	Given y^n sequence. Generate 2^{nR_y} i.i.d. u^n sequences (side-info encoder codebook)
Question:	What size R_y guarantees $\exists \geq 1$ u^n sequence jointly typical with y^n?
Answer:	$\simeq 2^{nH(u)}$ typical u^n. $\simeq 2^{nH(u\|y)}$ typical u^n given y^n. Examine $\simeq \frac{2^{nH(u)}}{2^{nH(u\|y)}} = 2^{nI(u;y)}$. $Ry > I(y;u)$ $\to \Pr(E_1 \cap E_0^c) \to 0$.

$\Pr[E_2 \cap E_0^c] < \epsilon$ **if** R_x **Big**

Error: $\exists\, \tilde{x}^n \in B(i_0)$, s.t.
1) $\tilde{x}^n \neq x^n$
2) $\tilde{x}^n \in T_\epsilon$ and
3) $(\tilde{x}^n, u^n(j_0)) \in T_\epsilon$

Set-up: $\tilde{x}^n$ independent of $u^n(j_0)$

Question: How many bins?
(How few sequence in each bin?)

Answer: 1) $\Pr[(\tilde{x}^n, u^n) \in T_\epsilon] \leq 2^{-nI(x;u)}$.

2) $|B(i) \cap T_\epsilon(x^n)| \leq 2^{nH(x)} 2^{-nR_x}$

$$\begin{aligned}\Pr(E_3 \cap E_0^c) &\leq 2^{n(H(x)-R_x-I(x;u))} \\ &\leq 2^{n(H(x|u)-R_x)}\end{aligned}$$

$R_x \geq H(x|u)$
$\Rightarrow \Pr(E_3 \cap E_0^c) \to 0$

Pr$[E_3 \cap E_0^c] < \epsilon$ because Pairwise Typicality is Transitive for Markov Chains

---

Error:	$(x^n, y^n) \in T_\epsilon$, and $(y^n, u^n(j_0)) \in T_\epsilon$, but $(x^n, u^n(j_0)) \notin T_\epsilon$
Set-up:	Use $x \to y \to u$ (Markov)
Question:	Is pairwise typicality transitive?
Answer:	Yes, when relationship is Markov (Markov Typicality Lemma) (Uses Strong Typicality)
	$\to$ Pr$(E_2 \cap E_0^c) \to 0$.

Defining Typicality: Weak and Strong

Def: Weakly Typical Sequence
An i.i.d. seq. x^n is ϵ-weakly typical if

$$\left| -\frac{1}{n} \log p(x^n) - H(x) \right| \leq \epsilon.$$

Statement about empirical entropy.

Def: Strongly Typical Sequence
An i.i.d. seq. x^n is ϵ-strongly typical if

$$\begin{array}{ll} (1) & \left| \frac{1}{n} \log N(a|x^n) - p(a) \right| \leq \frac{\epsilon}{|\mathcal{X}|} \\ (2) & N(a|x^n) = 0 \text{ if } p(a) = 0. \end{array}$$

Statement about empirical distribution.

Weakly Typ. $\not\Rightarrow$ Strongly Typ.

Example: Fair Coin, $\Pr[H] = \Pr[T] = 0.5$.

$$\begin{aligned} H(x) &\simeq -\frac{1}{n}\log p(x^n) \\ &= -\frac{1}{n}\log \prod_{i=1}^{n} p(x_i) \\ &= -\frac{1}{n}\log \Pr[H]^{k_H}\Pr[T]^{k_T} \\ &= -\frac{1}{n}\log(0.5^{k_H})(0.5^{n-k_H}) \\ &= -\frac{1}{n}\log(0.5^n) \end{aligned}$$

	k_H	k_T
Weakly Typical	0	n
Strongly Typical	$n/2$	$n/2$

Weakly Typ. $\not\Rightarrow$ Strongly Typ.

Example: $|\mathcal{X}| = 3$:

$\Pr[a] = \Pr[b] = 0.25$, $\Pr[c] = 0.5$.

$$
\begin{aligned}
H(x) &\simeq -\frac{1}{n}\log p(x^n) \\
&= -\frac{1}{n}\log \prod_{i=1}^{n} p(x_i) \\
&= -\frac{1}{n}\log \Pr[a]^{k_a}\Pr[b]^{k_b}\Pr[c]^{n-k_a-k_b} \\
&= -\frac{1}{n}\log (0.25^{k_a})(0.25^{k_b})(0.5^{n-k_a-k_b}) \\
&= -\frac{1}{n}\log (0.25^{k_a+k_b})(0.5^{n-k_a-k_b}).
\end{aligned}
$$

	k_a	k_b	k_c
Weakly Typical	0	$n/2$	$n/2$
Strongly Typical	$n/4$	$n/4$	$n/2$

Markov Typicality Lemma

Given $x_k \to y_k \to u_k$ are i.i.d. Markov chains for $l = 1, \ldots n$

If (1) $(x^n, y^n) \in T_\epsilon$
(2) $(y^n, u^n) \in T_\epsilon$

Then $(x^n, u^n) \in T_\epsilon$

Example: Not true without Markov.
x^n, y^n independent and i.i.d. Bernoulli(0.5).
$u^n = x^n$.
Independently generate x^n, y^n, u^n.
Then $(x^n, y^n) \in T_\epsilon$, $(y^n, u^n) \in T_\epsilon$
But $(x^n, u^n) \notin T_\epsilon$

Conclusion of Proof

$\Pr(E) \leq \Pr(E_0) + \sum_{i=1}^{3} \Pr(E_i \cap E_0^c)$.

We showed: $\Pr(E_0) \to 0$,
$\Pr(E_0 \cap E_i^c) \to 0$ for $i \in \{1, 2, 3\}$.

Hence $\Pr(E) \to 0$.

Elements of Converse

1. Information Inequalities
2. "Subtle" Markov Relationship
3. Single Letterization via Time-Sharing

   (See C&T)

Explaining the "Subtle" Markov Relationship

$$H(\mathsf{x}_i|\mathsf{T}, \mathsf{x}^{i-1}, \mathsf{y}^{i-1}) \;=\; H(\mathsf{x}_i|\mathsf{T}, \mathsf{y}^{i-1})$$

See using Bayes' rule:

$p(\mathsf{x}_i|\mathsf{T}, \mathsf{x}^{i-1}, \mathsf{y}^{i-1}) =$

$$= \frac{p(f(\mathsf{y}^n), \mathsf{x}_i, \mathsf{x}^{i-1}, \mathsf{y}^{i-1})}{p(f(\mathsf{y}^n), \mathsf{x}^{i-1}, \mathsf{y}^{i-1})}$$

$$= \frac{p(f(\mathsf{y}^n)|\mathsf{x}_i, \mathsf{x}^{i-1}, \mathsf{y}^{i-1})}{p(f(\mathsf{y}^n)|\mathsf{x}^{i-1}, \mathsf{y}^{i-1})} p(\mathsf{x}_i|\mathsf{x}^{i-1}, \mathsf{y}^{i-1})$$

$$= \frac{p(f(\mathsf{y}^n)|\mathsf{x}_i, \mathsf{y}^{i-1})}{p(f(\mathsf{y}^n)|\mathsf{y}^{i-1})} p(\mathsf{x}_i|\mathsf{y}^{i-1})$$

$$= p(\mathsf{x}_i|f(\mathsf{y}^n), \mathsf{y}^{i-1})$$

$$= p(\mathsf{x}_i|\mathsf{T}, \mathsf{y}^{i-1}).$$

Outline

I. Problem Statement

II. Practical Approach

III. Lossless Side-Information

IV. **Lossy Side-Information**

V. Gaussian and Binary Sources

VI. Duality with Info Embedding

Coding Theorem – Lossy Case

Given (1) a pair of sources **x**, **y**,
(2) $p(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} p(x_i, y_i)$,
(3) A decoder that observes **y** directly,
(4) $d(x^n, \widehat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, \widehat{x}_i)$.

If **x** is encoded at rate R_{x},

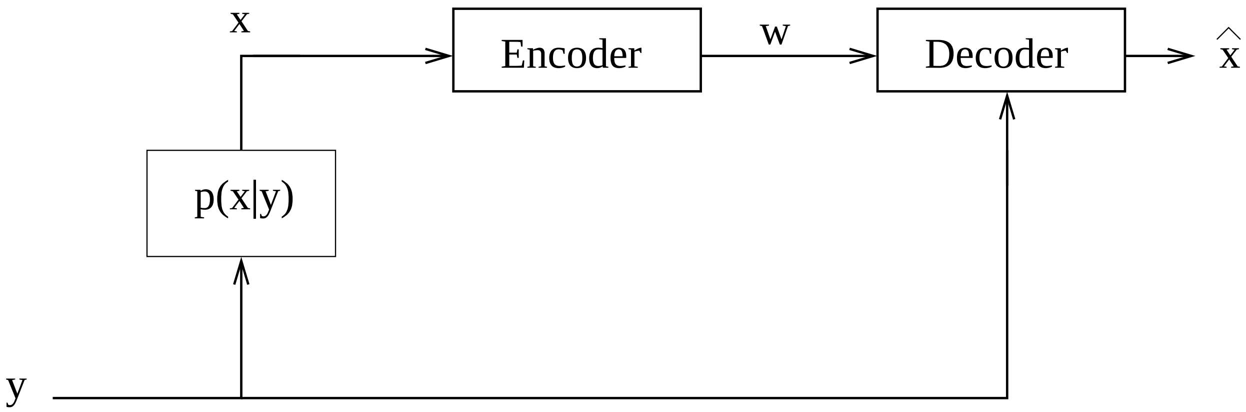
Then we can recover **x** to within distortion D_{x} with arbitrarily small Pr(failure)

iff (i) $R_{\mathsf{x}} \geq R^{WZ}_{\mathsf{x}|\mathsf{y}}(D_{\mathsf{x}}) =$
$= \min_{p(w|x)} \min_f (I(x; w) - I(y; w))$,

(ii) $y \to x \to w$ (Markov),

(iii) $E\left[d(x, f(y, w))\right] \leq D_{\mathsf{x}}$.

Single-Letter View of Lossy Side-Information



Break Proof into Steps

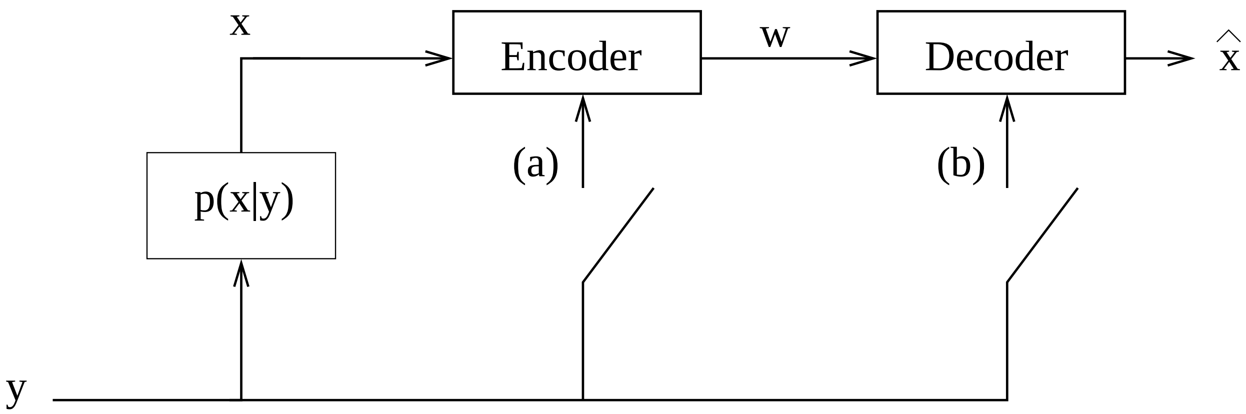
1. $2^{nI(x;w)}$ **w**-seq well describe x^n.

2. Bin in $2^{nR_x} \simeq 2^{n(I(x;w)-I(y;w))}$ bins.

3. Roughly $2^{nI(y;w)}$ **w**-seq per bin.

4. Side info can select between $\simeq 2^{nI(y;w)}$ **w**-seq by joint typicality.

Problems Related to Lossy Side-Info



	(a)	(b)	Rate R_x	Conditions
Rate-Dist	○	○	$I(x;\hat{x})$	$E[d(x,\hat{x})] \leq D_x$
Cond. R-D	●	●	$I(x;w\|y)$	$E[d(x,\hat{x})] \leq D_x$
Lossy Side-Info	○	●	$I(x;w\|y)$	$E[d(x,\hat{x})] \leq D_x$ $y \rightarrow x \rightarrow w$

Outline

I. Problem Statement

II. Practical Approach

III. Lossless Side-Information

IV. Lossy Side-Information

V. **Gaussian and Binary Sources**

VI. Duality with Info Embedding

Example: Gaussian Source, Squared Distortion

Setup: x, y jointly Gaussian, 0-mean, r.v.'s

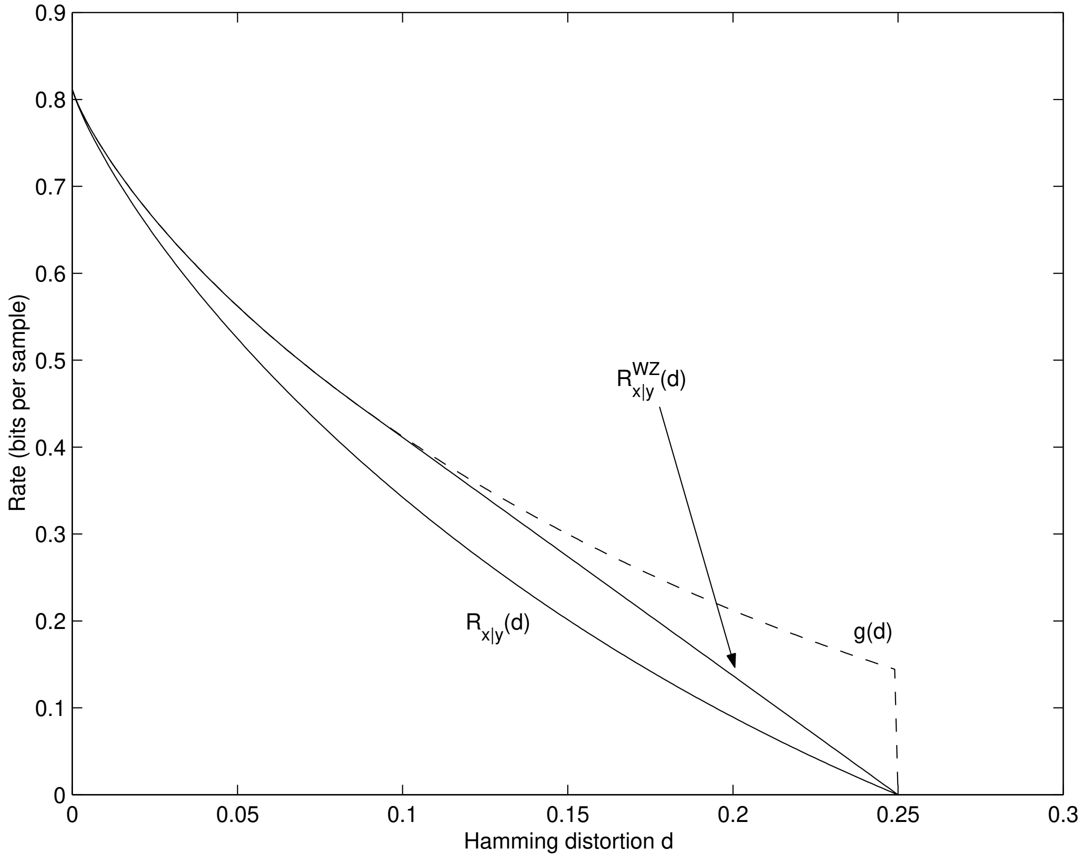
Rate-Distortion Bounds:

$$R^{WZ}_{x|y}(D_x) = \begin{cases} \frac{1}{2}\log\frac{\lambda_{x|y}}{D_x}, & 0 \le D_x \le \lambda_{x|y} \\ 0, & \lambda_{x|y} < D_x \end{cases},$$

where $\lambda_{x|y}$ is the Bayesian Estimation Error:

$$\begin{aligned} \lambda_{x|y} &= E\left[(\hat{x}(y) - x)^2\right] \\ &= \lambda_x - \lambda_{x,y}^2 \lambda_y^{-1}. \end{aligned}$$

Example: Binary Symmetric Source, Hamming Distortion*



Binary symmetric source where Pr$(x_i \neq y_i) = p =$ 0.25, $H(0.25) \simeq 0.8$.

The dashed line is $g(d) = H(p * d) - H(d)$.

*Figure from R. Barron '00

Outline

I. Problem Statement

II. Practical Approach

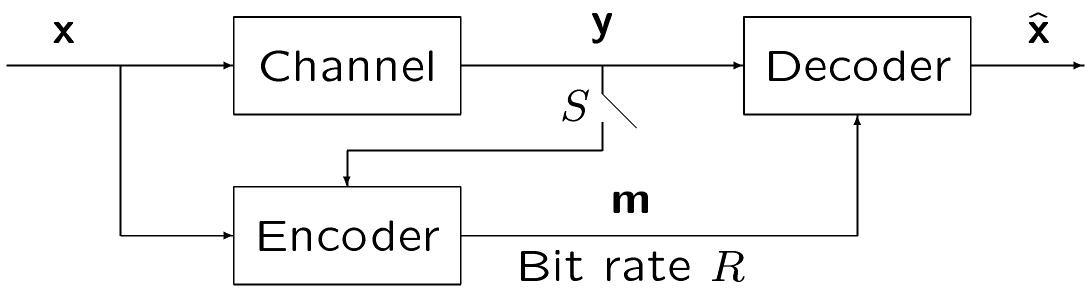
III. Lossless Side-Information

IV. Lossy Side-Information

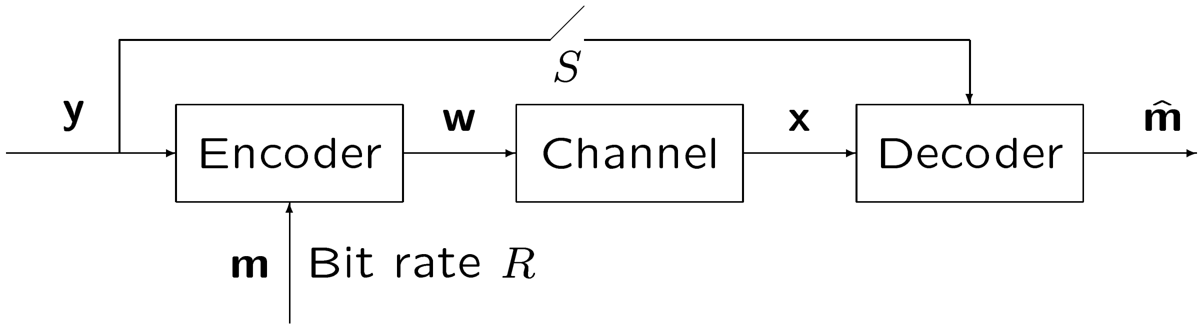
V. Gaussian and Binary Sources

VI. **Duality with Info Embedding**

Source Coding with Side-Info & Info Embedding are Dual Problems*



Source coding with side information

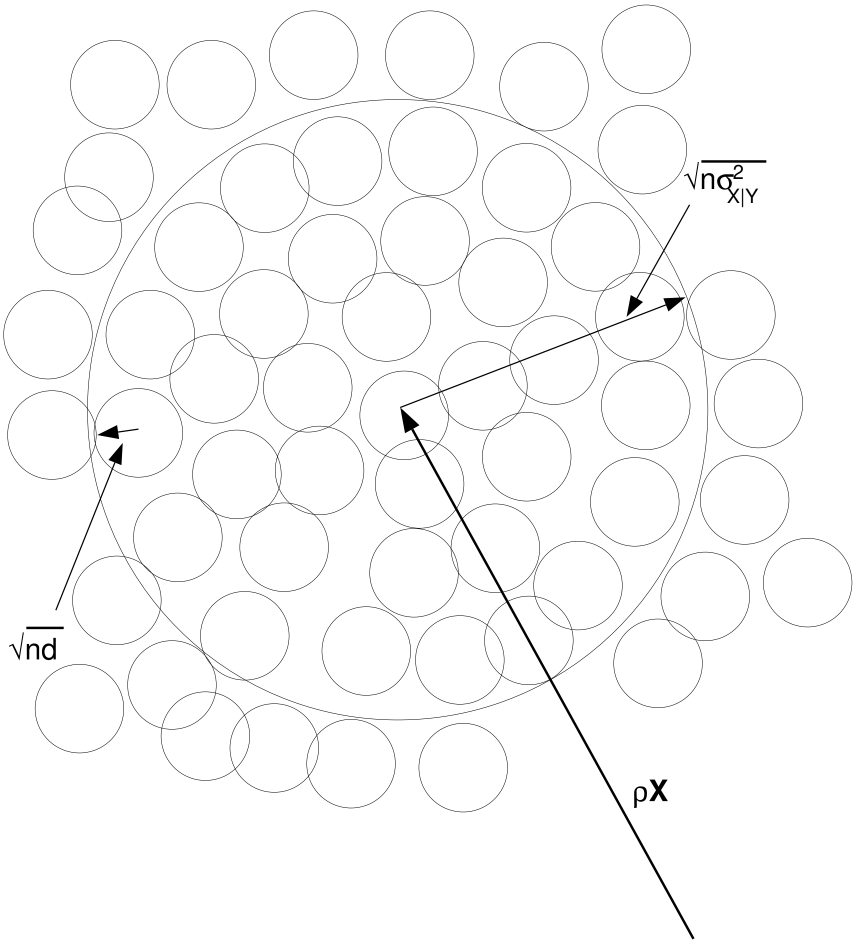


Information embedding

- A good side information encoder (decoder) is a good info embedding decoder (encoder)
- Low-bit modulation

*Figures from R. Barron '00

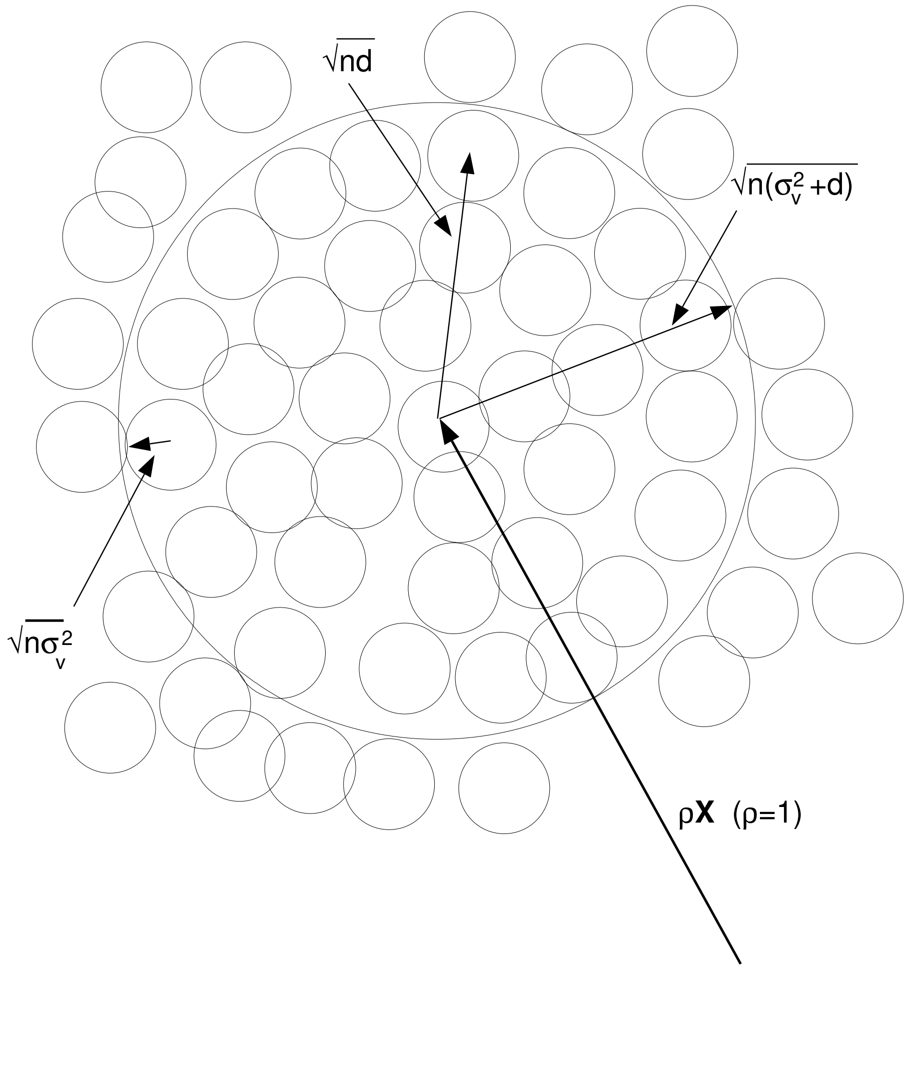
Sphere-Packing View of Side-Info*



The vector $\rho\mathbf{y}$ is source estimate from $\mathbf{y}$.

*Figure from R. Barron '00

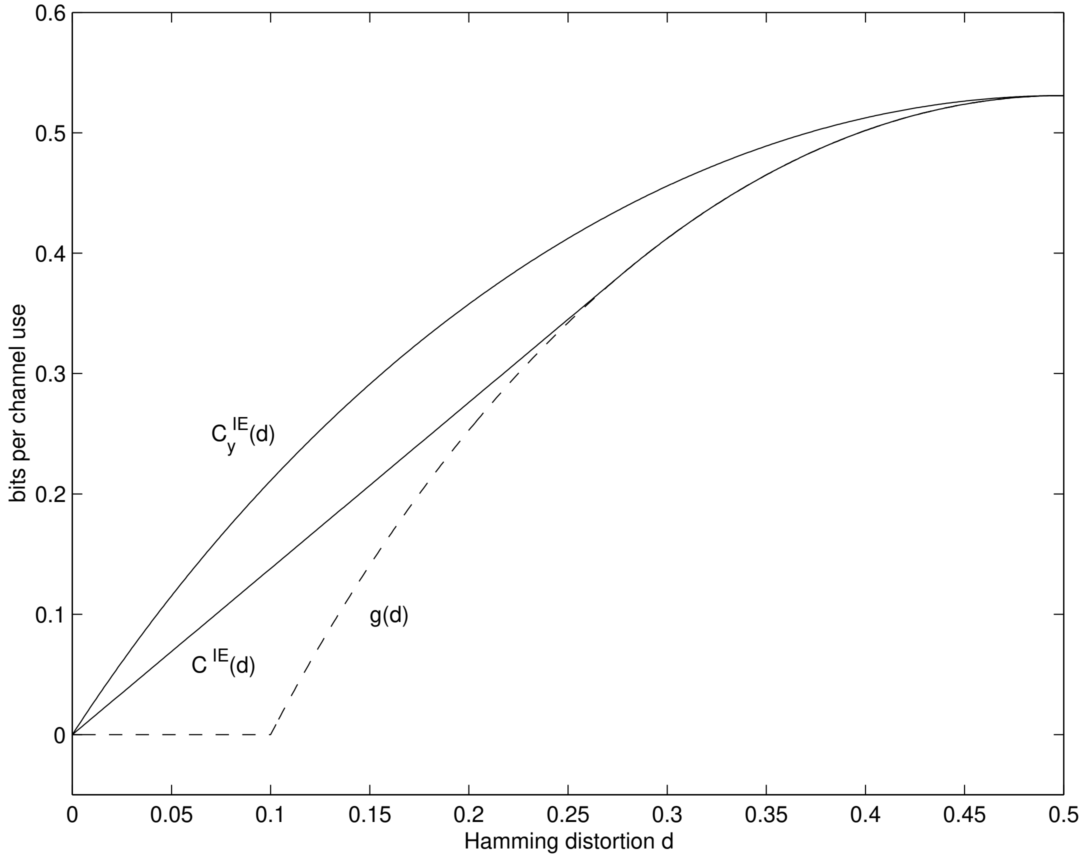
Sphere-Filling View of Info Embedding*



The vector $\rho\mathbf{y}$ is the host signal.

*Figure from R. Barron '00

Example: Binary Symmetric Source and Channel*



Binary symmetric case where the channel transition probability is $p = 0.1$.

The dashed line is $g(d) = H(d) - H(p)$.

*Figure from R. Barron '00