SEMANTIC CODES

Frans M.J. Willems
(joint work w. Ton Kalker (HP Labs.) & Deran Maas (ABB))

Eindhoven University of Technology

IZS, March 3-5, 2010, Zurich, Switzerland

Abstract:

Shannon wrote in 1948: "The semantic aspects of communication are irrelevant to the engineering problem." He demonstrated indeed that the information generated by a source depends only on its statistics and not on the meaning of the source output. The authors derived the fundamental limits for semantic compaction, transmission and compression systems recently. These systems have the property that the codewords are semantic however, i.e. close to the source sequences. In the present article we determine the minimum distortion for semantic partial transmission systems. In these systems only a quantized version of each source symbol is transmitted to the receiver. It should be noted that our achievability proof is based on weak instead of strong typicality. This is unusual for Gelfand-Pinsker [1980] related setups as e.g. semantic coding and embedding.

Again the abstract:

`Shnnn wrte n 1948: "Th smntc spcts f cmmnctn r rrlvnt t th ngnrng prblm." H dmnstrtd ndd tht th nfrmtn gnrtd b src dpnds nl n ts sttstcs nd nt n th mnng f th src tpt. Th thrs drvd th fndmntl lmts fr smntc cmpctn, trnsmssn nd cmprssn sstms rcntl. Ths sstms hv th prprt tht th cdwrds r smntc hwvr, .. cls t th src sqncs. n th prsnt rtcl w dtrmn th mnmm dstrtn fr smntc prtl trnsmssn sstms. n ths sstms nl qntzd vrsn f ch src smbl s trnsmttd t th rcvr. t shld b ntd tht r chvblt prf s bsd n wk nstd f strng tpclt. Ths s nsl fr Glfnd-Pnskr [1980] rltd stps s .. smntic cdng nd mbddng.`

CONCLUSION: Codes (for compaction) need not be abstract, but can be meaningful (semantic) as well.

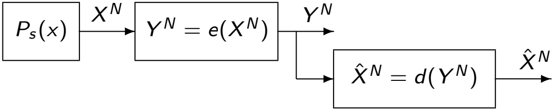
Semantic Coding: Three Settings

8bit / pixel
Original =

5bit / pixel
Codeword =

- **I. Semantic compaction:** Original is decoded from the semantic codeword. **Semantic codeword contains as many symbols as the original, but code alphabet size is smaller than source alphabet size.** Code-rate **above** entropy of the original.
- **II. Semantic compacted transmission:** Semantic codeword is transmitted over noisy channel. Original is decoded from channel output. Same number of symbols. Entropy **below** channel capacity.
- **III. Semantic compression:** From semantic codeword a (better) approximation of the original can be determined. Same numbers of symbols, different alphabet sizes. Code-rate **below** entropy of original.

I. Semantic Compaction



An i.i.d. source produces a sequence x^N with probability $\Pr\{X^N = x^N\} = \Pi_{n=1,N} P_s(x_n)$ for $x^N = (x_1, x_2, \cdots, x_N) \in \mathcal{X}^N$. Here $\mathcal{X}$ is the finite source alphabet.
Encoder $e(\cdot)$ transforms x^N into a **semantic code sequence** $y^N = (y_1, y_2, \cdots, y_N) \in \mathcal{Y}^N$ where $\mathcal{Y}$ is the finite code alphabet.
From y^N a decoder $d(\cdot)$ produces an estimate $\hat{x}^N$ of the original source sequence x^N.
The **error probability** $P_\mathcal{E}$ and **semantic distortion** $\overline{D_{xy}}$ are defined as

$$\begin{aligned} P_\mathcal{E} &\triangleq \Pr\{\hat{X}^N \neq X^N\}, \\ \overline{D_{xy}} &\triangleq E[\frac{1}{N} \sum_{n=1,N} D_{xy}(X_n, e_n(X^N))], \end{aligned}$$

where $e_n(\cdot)$ denotes the n'th component of the encoder output and $\{D_{xy}(x,y), x \in \mathcal{X}, y \in \mathcal{Y}\}$ is the **semantic distortion matrix** having only finite non-negative entries.
Semantic distortion level Δ_{xy} is said to be **achievable** if for all $\epsilon > 0$ there exists for all large enough N, encoders and decoders such that

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

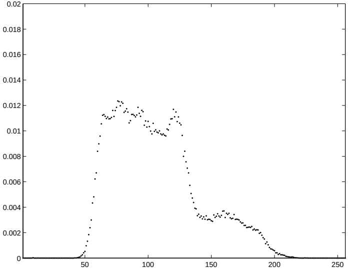
Theorem

For compaction the **set of achievable semantic distortions** *is equal to* $\mathcal{G}_{compact}$ *which is defined as*

$$\begin{aligned} \mathcal{G}_{compact} = \{\Delta_{xy} \quad : \quad & \Delta_{xy} \geq \sum_{x,y} P(x,y) D_{xy}(x,y), \\ & \textit{for } P(x,y) = P_s(x) P_t(y|x) \\ & \textit{such that } H(X) \leq H(Y)\}. \end{aligned}$$

NOTE that the code alphabet $\mathcal{Y}$ is assumed to be specified. Therefore the **compaction-rate** is actually $\log_2 |\mathcal{Y}|$ bits per symbol.

Back to Lena Again (blue part)



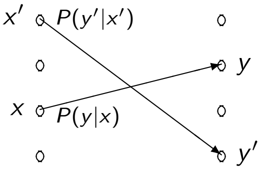
$\mathcal{X} = \{0, 1, 2, \cdots, 255\}$, empirical distribution (right) $\{P(x), x \in \mathcal{X}\}$, empirical entropy $H(X) = 6.968$ bit.
Take $\mathcal{Y} = \{0, 2, 4, \cdots, 254\}$ and $P(y) = 1/128$ (uniform) for all $y \in \mathcal{Y}$, then $H(Y) = \log_2 128 = 7$ bit.

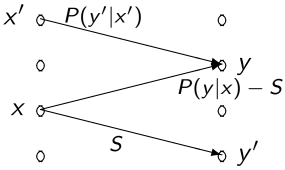
Question: If we consider squared-error distortion $D_{xy} = (y - x)^2$, what is the test-channel $P_t(y|x)$ that minimizes $\sum_{xy} P(x, y) D_{xy}(x, y)$?

Crossing Edges

Lemma

A test channel with both $P_t(y|x) > 0$ and $P_t(y'|x') > 0$ for $x' < x$ and $y' > y$ can be replaced by an alternative test channel that achieves the same output distribution but a smaller average distortion.





Proof: For $x' < x$ and $y' > y$ note that

$$\begin{aligned}
&(y - x')^2 + (y' - x)^2 - (y - x)^2 - (y' - x')^2 \\
&\quad = \quad -2yx' - 2y'x + 2yx + 2y'x' \\
&\quad = \quad 2(y - y')(x - x') < 0.
\end{aligned}$$

Consider the left "partial" channel in the figure. The contribution[1] to the average distortion of this partial channel is

$$p(x')P(y'|x')(y'-x')^2 + p(x)P(y|x)(y-x)^2.$$

If $p(x)P(y|x) > p(x')P(y'|x')$ then $P(y|x) > p(x')P(y'|x')/p(x)$. Now the right partial channel in the figure, if we choose $S = p(x')P(y'|x')/p(x)$, has the same contribution to the output distribution however its contribution to the average distortion is

$$\begin{aligned} & p(x')P(y'|x')(y-x')^2 + p(x)\left(P(y|x)-S\right)(y-x)^2 + p(x)S(y'-x)^2 \\ = \quad & p(x')P(y'|x')\left((y-x')^2+(y'-x)^2-(y-x)^2\right) + p(x)P(y|x)(y-x)^2 \\ < \quad & p(x')P(y'|x')(y'-x')^2 + p(x)P(y|x)(y-x)^2. \end{aligned}$$

For $p(x)P(y|x) < p(x')P(y'|x')$ we can give a similar argument.

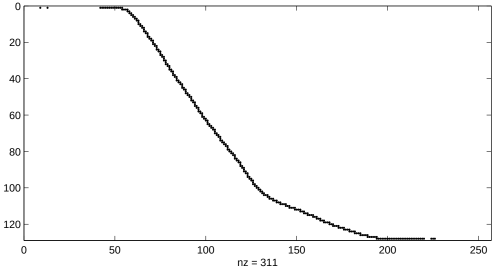
Lemma

for each output distribution $\{P(y) : y \in \mathcal{Y}\}$ there is a unique test channel without crossing edges. This test channel minimizes the average distortion.

---

[1] Denoting $P_s(\cdot)$ by $p(\cdot)$ and $P_t(\cdot|\cdot)$ by $P(\cdot|\cdot)$.

This leads to the **optimal test channel** for (the blue part of) Lena:



The optimal test channel has

$$\begin{aligned} H(Y|X) &= 0.8366 \text{ bit}, \\ H(X|Y) &= 0.8051 \text{ bit}, \\ \text{and } \sum_{xy} P(x,y)(y-x)^2 &= 2249.6. \end{aligned}$$

To "see" the distortion we have generated for every element x in the image an element y at random with probability $\{P_t(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$. This resulted in the images below (left=original, right=code):

Postprocessing

Note that now we have taken the "reconstruction" alphabet $\mathcal{Y} = \{0, 2, \cdots, 254\}$. Doing some postprocessing can decrease the squared-error distortion enormously however. Taking

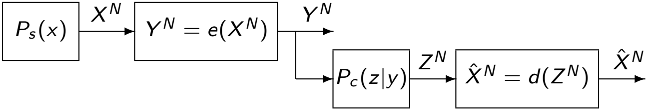
$$\hat{x}(y) = \sum_x x \frac{P(x,y)}{P(y)},$$

(MMSE estimate) will lead to the best results if the value y is replaced by $\hat{x}(y)$.
For our example (blue part of Lena) we now get

$$\sum_{xy} P(x,y)(\hat{x}(y) - x)^2 = 0.554.$$

This is more or less the inverse of the "mapping" that $P_t(y|x)$ represents.

II. Semantic Compacted Transmission



The source is again i.i.d..
An encoder $e(\cdot)$ transforms the source sequence x^N into a **semantic channel input sequence** $y^N \in \mathcal{Y}^N$.
The discrete channel $\{P_c(z|y), y \in \mathcal{Y}, z \in \mathcal{Z}\}$ with finite input alphabet $\mathcal{Y}$ and finite output alphabet $\mathcal{Z}$ is memoryless. Therefore $\Pr\{Z^N = z^N | Y^N = y^N\} = \Pi_{n=1,N} P_c(z_n|y_n)$ where $z^N = (z_1, z_2, \cdots, z_N) \in \mathcal{Z}^N$ is a channel output sequence.
From this channel output z^N a decoder $d(\cdot)$ constructs an estimate $\hat{x}^N$ of the original source sequence x^N.

The **error probability** $P_{\mathcal{E}}$ and **semantic distortion** $\overline{D_{xy}}$ are again defined as

$$\begin{aligned} P_{\mathcal{E}} &\triangleq \Pr\{\hat{X}^N \neq X^N\}, \\ \overline{D_{xy}} &\triangleq E[\frac{1}{N}\sum_{n=1,N} D_{xy}(X_n, e_n(X^N))]. \end{aligned}$$

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

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

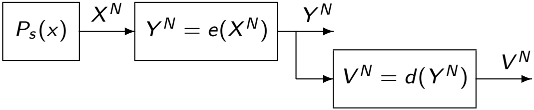
Theorem

For transmission the **set of achievable semantic distortions** *is equal to* $\mathcal{G}_{transm}$ *which is defined as*

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

Semantic transmission **impossible** if $H(X) > \max_{P_t(y)} I(Y;Z)$.

III. Semantic Compression



The source is i.i.d..
Encoder transforms the source sequence x^N into a **semantic code sequence** $y^N \in \mathcal{Y}^N$. The **semantic distortion** between x^N and y^N is now defined as

$$\overline{D_{xy}} \triangleq E[\frac{1}{N} \sum_{n=1,N} D_{xy}(X_n, e_n(X^N))].$$

The decoder $d(\cdot)$ determines from the code sequence y^N the restoration vector $v^N = (v_1.v_2, \cdots, v_N)$ with components from the finite alphabet $\mathcal{V}$. The **vector distortion** between x^N and v^N is defined as

$$\overline{D_{xv}} \triangleq E[\frac{1}{N} \sum_{n=1,N} D_{xv}(X_n, d_n(e(X^N)))]$$

for a second **vector distortion matrix** $\{D_{xv}(x, v), x \in \mathcal{X}, v \in \mathcal{V}\}$. Also this matrix has only finite non-negative entries.
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

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

Theorem

For semantic compression the **set of achievable distortion pairs** *is equal to* $\mathcal{G}_{compress}$ *which is defined as*

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

NOTE that the **compression-rate** is $\log_2 |\mathcal{Y}|$ nats per symbol.

Gaussian Source, MSE Distortion

Consider a Gaussian source with mean zero and variance $\sigma_X^2 = 1$, i.e.

$$P_s(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}).$$

If $P_t(v|x)$ achieves a point on the rate-distortion function with rate R for the mean-squared error case, then $X = V + N$ where V and N are independent zero-mean Gaussians with $\sigma_N^2 = 2^{-2R}$ and $\sigma_V^2 = 1 - 2^{-2R}$ respectively since

$$I(X;V) = h(X) - h(N) = \frac{1}{2} \log_2 \frac{\sigma_X^2}{\sigma_N^2} = R.$$

The vector-distortion is $E[D_{xv}] = E[(X-V)^2] = \sigma_N^2 = 2^{-2R}$. For the semantic distortion we can write

$$\begin{aligned} E[D_{xy}] = E[(X-Y)^2] &= E[(V+N-Y)^2] \\ &= E[(V-Y)^2] + 2E[(V-Y)N] + E[N^2] \\ &= E[(V-Y)^2] + 2^{-2R}. \end{aligned}$$

Note that the cross-term $E[(V-Y)N] = 0$ since N is independent of $Y - V$ since $I(X;Y|V) = 0$ and therefore $E[(V-Y)N] = E[V-Y]E[N] = 0$.

In order to minimize $E[(V-Y)^2]$ the random variable Y should be the Lloyd quantizer for V with 2^R quantization intervals. If a Lloyd quantizer with L intervals for a Gaussian unit-variance source yields distortion $\Delta_{\text{Lloyd}}(L)$ we obtain

$$E[(V-Y)^2] = (1-2^{-2R})\Delta_{\text{Lloyd}}(2^R)$$

since $\sigma_V^2 = 1-2^{-2R}$. All this leads to the following table of mean-squared error distortions. Note that $E[(X-V)^2]$ is the distortion of the best vector quantizer and Δ_{Lloyd} the distortion of the best scalar quantizer.

R	$E[(X-V)^2]$	$\Delta_{\text{Lloyd}}(2^R)$	$E[(X-Y)^2] = (1-2^{-2R})\Delta_{\text{Lloyd}} + 2^{-2R}$
1	0.2500	0.3634	0.5225
2	0.0625	0.1175	0.1722
3	0.0156	0.0345	0.0496
4	0.0039	0.0095	0.0134
5	0.0010	0.0025	0.0035

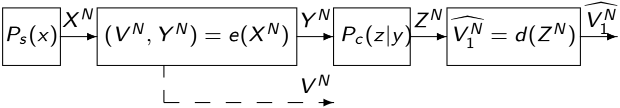
For large values of the rate R we obtain for the semantic distortion that

$$E[D_{xy}] = E[(X-Y)^2] \approx \Delta_{\text{Lloyd}}(2^R) + 2^{-2R}$$

i.e. the sum of the smallest possible distortion achievable with a scalar quantizer plus the smallest possible distortion achievable with a vector quantizer.

Achievability Proofs: Outline

Consider the joint distribution $P(x, v, y, z) = P_s(x)P_t(v, y|x)P_c(z|y)$ where $\mathcal{V}$ is a finite auxiliary alphabet, $P_t(\cdot|\cdot, \cdot)$ some test-channel between $\mathcal{X}$ and $\mathcal{V} \times \mathcal{Y}$, and $P_c(\cdot|\cdot)$ some discrete memoryless channel between $\mathcal{Y}$ and $\mathcal{Z}$.



We claim that for

$$I(V; X) < I(V; Z)$$

there exists a set $\mathcal{S}$ of sequences $v^N \in \mathcal{V}^N$ such that:

- With probability ≈ 1 there is at least one $v^N \in \mathcal{S}$ that is jointly typical with x^N.
- With probability ≈ 1 this v^N is the unique $v^N \in \mathcal{S}$ that is jointly typical with z^N.

- **I. Compaction:** Setting $V \equiv (X, Y)$ and $Z \equiv Y$ (no channel) results in
$$H(X) < H(Y).$$
- **II. Transmission:** Setting $V \equiv (X, Y)$ results in
$$H(X) < I(Y; Z).$$
- **III. Compression:** Setting $V \equiv (Y, V')$ and $Z \equiv Y$ (no channel) results in
$$I(X; Y, V') < H(Y).$$

Typicality takes care of the **distortions**.

NOTE that our coding method is a simplified case of Gelfand-Pinsker [1980] coding.

Concluding Remarks and Problems

- **Converses** are standard.
- Semantic coding comes for **FREE**, is always possible to some extent.
- Semantic compaction, transmission, and compression are special cases (no embedding) of **Reversible Embedding**.
- **CODE CONSTRUCTIONS ???**
- **Sources with MEMORY**. Blocking ???
- **APPLICATIONS** ???