

The Method of Types and Its Application to Information Hiding

Pierre Moulin

University of Illinois at Urbana-Champaign

www.ifp.uiuc.edu/~moulin/talks/eusipco05-slides.pdf

EUSIPCO

Antalya, September 7, 2005

Outline

- Part I: General Concepts
 - Introduction
 - Definitions
 - What is it useful for?
- Part II: Application to Information Hiding
 - Performance guarantees against omnipotent attacker?
 - Steganography, Watermarking, Fingerprinting

Part I: General Concepts

Reference Materials

- I. Csiszar, “The Method of Types”, *IEEE Trans. Information Theory*, Oct. 1998 (commemorative Shannon issue)
- A. Lapidoth and P. Narayan, “Reliable Communication under Channel Uncertainty”, same issue.
- Application areas:
 - capacity analyses
 - computation of error probabilities (exponential behavior)
 - universal coding/decoding
 - hypothesis testing

Basic Notation

- Discrete alphabets \mathcal{X} and \mathcal{Y}
- Random variables X, Y with joint pmf $p(x, y)$
- The entropy of X is $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$
(will sometimes be denoted by $H(p_X)$)
- Joint entropy $H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y)$
- The conditional entropy of Y given X is

$$\begin{aligned} H(Y|X) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= H(X, Y) - H(X) \end{aligned}$$

- The mutual information between X and Y is

$$\begin{aligned} I(X; Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= H(Y) - H(Y|X) \end{aligned}$$

- The Kullback-Leibler divergence between pmf's p and q is

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

Types

- Deterministic notion
- Given a length- n sequence $\mathbf{x} \in \mathcal{X}^n$, count the frequency of occurrence of each letter of the alphabet \mathcal{X}
- Example: $\mathcal{X} = \{0, 1\}$, $n = 12$,
 $\mathbf{x} = 110100101110$ contains 5 zeroes and 7 ones
 \Rightarrow the sequence \mathbf{x} has type $\hat{p}_{\mathbf{x}} = (\frac{5}{12}, \frac{7}{12})$
- $\hat{p}_{\mathbf{x}}$ is also called empirical pmf.
It may be viewed as a pmf over \mathcal{X}
- Each $\hat{p}_{\mathbf{x}}(x)$ is a multiple of $\frac{1}{n}$.

Joint Types

- Given two length- n sequences $\mathbf{x} \in \mathcal{X}^n$ and $\mathbf{y} \in \mathcal{Y}^n$, count the frequency of occurrence of each pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$
- Example: $\mathbf{x} = 110100101110$
 $\mathbf{y} = 111100101110$
- (\mathbf{x}, \mathbf{y}) have joint type $\hat{p}_{\mathbf{xy}} = \begin{pmatrix} 4/12 & 1/12 \\ 0 & 7/12 \end{pmatrix}$
- Empirical pmf over $\mathcal{X} \times \mathcal{Y}$

Conditional Types

- By analogy with Bayes rule, define the conditional type of \mathbf{y} given \mathbf{x} as

$$\hat{p}_{\mathbf{y}|\mathbf{x}}(y|x) = \frac{\hat{p}_{\mathbf{xy}}(x, y)}{\hat{p}_{\mathbf{x}}(x)}$$

which is an empirical conditional pmf

- Example: $\mathbf{x} = 110100101110$
 $\mathbf{y} = 11\mathbf{1}100101110$

$$\Rightarrow \hat{p}_{\mathbf{y}|\mathbf{x}} = \begin{pmatrix} 4/5 & 1/5 \\ 0 & 1 \end{pmatrix}$$

Type Classes

- The type class $T_{\mathbf{x}}$ is the set of all sequences that have the same type as \mathbf{x} .

Example: all sequences with 5 zeroes and 7 ones

- The joint type class $T_{\mathbf{xy}}$ is the set of all sequences that have the same joint type as (\mathbf{x}, \mathbf{y})
- The conditional type class $T_{\mathbf{y}|\mathbf{x}}$ is the set of all sequences \mathbf{y}' that have the same type as \mathbf{y} , conditioned on \mathbf{x}

Information Measures

- Any type may be represented by a dummy sequence
- Can define empirical information measures:

$$H(\mathbf{x}) \triangleq H(\hat{p}_{\mathbf{x}})$$

$$H(\mathbf{y}|\mathbf{x}) \triangleq H(\hat{p}_{\mathbf{y}|\mathbf{x}})$$

$$I(\mathbf{x}; \mathbf{y}) \triangleq I(X; Y) \quad \text{for } (X, Y) \sim \hat{p}_{\mathbf{xy}}$$

- Will be useful to design universal decoders

Typicality

- Consider pmf p over \mathcal{X}
 - Length- n sequence $\mathbf{x} \sim$ i.i.d. p . Notation: $\mathbf{x} \sim p^n$
 - Example: $\mathcal{X} = \{0, 1\}$, $n = 12$, $\mathbf{x} = 110100101110$
 - For large n , all *typical* sequences have approximately composition p
 - This can be measured in various ways:
 - Entropy ϵ -typicality: $|\frac{1}{n} \log p^n(\mathbf{x}) - H(X)| < \epsilon$
 - Strong ϵ -typicality: $\max_{x \in \mathcal{X}} |\hat{p}_{\mathbf{x}}(x) - p(x)| < \epsilon$
- both define sets of typical sequences

Application to Channel Coding

- Channel input $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$,
output $\mathbf{y} = (y_1, \dots, y_n) \in \mathcal{Y}^n$
- Discrete Memoryless Channel (DMC): $p^n(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^n p(y_i|x_i)$
- Many fundamental coding theorems can be proven using the concept of entropy typicality. Examples:
 - Shannon's coding theorem (capacity of DMC)
 - Rate-distortion bound for memoryless sources

- Many fundamental coding theorems **cannot** be proved using the concept of entropy typicality. Examples:
 - precise calculations of error log-probability
 - various kinds of unknown channels
- So let's derive some useful facts about types
- Number of types $\leq (n + 1)^{|\mathcal{X}|}$ (polynomial in n)
- Size of type class $T_{\mathbf{x}}$:

$$(n + 1)^{-|\mathcal{X}|} e^{nH(\hat{p}_{\mathbf{x}})} \leq |T_{\mathbf{x}}| \leq e^{nH(\hat{p}_{\mathbf{x}})}$$

Ignoring polynomial terms, we write

$$|T_{\mathbf{x}}| \doteq e^{nH(\hat{p}_{\mathbf{x}})}$$

- Probability of \mathbf{x} under distribution p^n :

$$\begin{aligned}
 p^n(\mathbf{x}) &= \prod_{x \in \mathcal{X}} p(x)^{n\hat{p}_{\mathbf{x}}(x)} \\
 &= e^{-n \sum_{x \in \mathcal{X}} \hat{p}_{\mathbf{x}}(x) \log p(x)} \\
 &= e^{-n[H(\hat{p}_{\mathbf{x}}) + D(\hat{p}_{\mathbf{x}}||p)]}
 \end{aligned}$$

same for all \mathbf{x} in the same type class

- Probability of type class $T_{\mathbf{x}}$ under distribution p^n :

$$P^n(T_{\mathbf{x}}) = |T_{\mathbf{x}}| p^n(\mathbf{x}) \doteq e^{-nD(\hat{p}_{\mathbf{x}}||p)}$$

- Similarly:

$$\begin{aligned}
 |T_{\mathbf{y}|\mathbf{x}}| &\doteq e^{nH(\hat{p}_{\mathbf{y}|\mathbf{x}})} \\
 P_{Y|X}^n(T_{\mathbf{y}|\mathbf{x}}|\mathbf{x}) &\doteq e^{-nD(\hat{p}_{\mathbf{xy}}||p_{Y|X}\hat{p}_{\mathbf{x}})}
 \end{aligned}$$

Constant-Composition Codes

- All codewords have the same type $\hat{p}_{\mathbf{x}}$
- **Random coding:** generate codewords \mathbf{x}_m , $m \in \mathcal{M}$ randomly and independently from uniform pmf on type class $T_{\mathbf{x}}$
- Note that channel outputs have different types in general

Unknown DMC's – Universal Codes

- Channel $p_{Y|X}$ is revealed neither to encoder nor to decoder
⇒ neither encoding rule nor decoding rule may depend on $p_{Y|X}$

$$C = \max_{p_X} \min_{p_{Y|X}} I(X; Y)$$

- Universal codes: same error exponent as in known- $p_{Y|X}$ case (existence?)
- **Encoder:** select $T_{\mathbf{x}}$, use constant-composition codes
- **Decoder:** uses *Maximum Mutual Information* rule

$$\begin{aligned} \hat{m} &= \operatorname{argmax}_{m \in \mathcal{M}} I(\mathbf{x}_m; \mathbf{y}) \\ &= \operatorname{argmin}_{m \in \mathcal{M}} H(\mathbf{y} | \mathbf{x}_m) \end{aligned}$$

- Note: the GLRT decoder is in general not universal
(GLRT: first estimate $p_{Y|X}$, then plug in ML decoding rule)

Key idea in proof

- Denote by $\mathcal{D}_m \subset \mathcal{Y}^n$ the decoding region for message m
- Polynomial number of type classes, forming a partition of \mathcal{Y}^n
- Given that m was transmitted, partition error event

$$\mathbf{y} \in \mathcal{Y}^n \setminus \mathcal{D}_m$$

into a union over type classes:

$$\mathbf{y} \in \bigcup_{T_{\mathbf{y}|\mathbf{x}_m}} T_{\mathbf{y}|\mathbf{x}_m} \setminus \mathcal{D}_m$$

- The probability of the error event is therefore given by

$$\begin{aligned}
Pr[\text{error}|m] &= Pr \left[\bigcup_{T_{\mathbf{y}|\mathbf{x}_m}} T_{\mathbf{y}|\mathbf{x}_m} \setminus \mathcal{D}_m \right] \\
&\leq \sum_{T_{\mathbf{y}|\mathbf{x}_m}} Pr [T_{\mathbf{y}|\mathbf{x}_m} \setminus \mathcal{D}_m] \\
&\doteq \max_{T_{\mathbf{y}|\mathbf{x}_m}} Pr [T_{\mathbf{y}|\mathbf{x}_m} \setminus \mathcal{D}_m] \\
&= \max_{T_{\mathbf{y}|\mathbf{x}_m}} Pr[T_{\mathbf{y}|\mathbf{x}_m}] \frac{|T_{\mathbf{y}|\mathbf{x}_m} \setminus \mathcal{D}_m|}{|T_{\mathbf{y}|\mathbf{x}_m}|} \\
&\doteq \max_{T_{\mathbf{y}|\mathbf{x}_m}} e^{-nD(\hat{p}_{\mathbf{x}_m\mathbf{y}}||p_{Y|X}\hat{p}_{\mathbf{x}_m})} \frac{|T_{\mathbf{y}|\mathbf{x}_m} \setminus \mathcal{D}_m|}{|T_{\mathbf{y}|\mathbf{x}_m}|}
\end{aligned}$$

\Rightarrow the worst conditional type class dominates error probability

- Calculation mostly involves combinatorics: finding out

$$|T_{\mathbf{y}|\mathbf{x}_m} \setminus \mathcal{D}_m|$$

Extensions

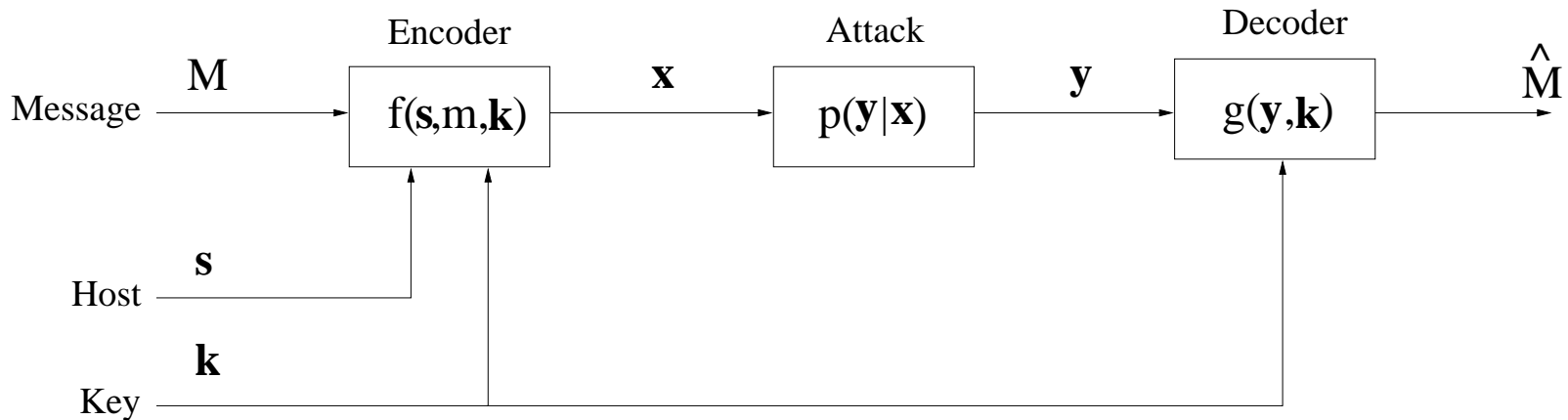
- Channels with memory
- “Arbitrary Varying” Channels \Rightarrow randomized codes
- Continuous alphabets (difficult!)

Part II: Applications to WM

Reference Materials

- [SM'03] A. Somekh-Baruch and N. Merhav, “On the Error Exponent and Capacity Games of Private Watermarking Systems,” *IEEE Trans. Information Theory*, March 2003
- [SM'04] A. Somekh-Baruch and N. Merhav, “On the Capacity Game of Public Watermarking Systems,” *IEEE Trans. Information Theory*, March 2004
- [MO'03] P. Moulin and J. O’Sullivan, “Information-Theoretic Analysis of Information Hiding,” *IEEE Trans. Information Theory*, March 2003
- [MW'04] P. Moulin and Y. Wang, “Error Exponents for Channel Coding with Side Information,” *preprint*, Sep. 2004

Communication Model for Data Hiding



- Memoryless host sequence \mathbf{s}
- Message M uniformly distributed over $\{1, 2, \dots, 2^{nR}\}$
- **Unknown** attack channel $p(\mathbf{y}|\mathbf{x})$
- Randomization via secret key sequence \mathbf{k} , arbitrary alphabet \mathcal{K}

Attack Channel Model

- First IT formulations of this problem assumed a fixed attack channel (e.g., AWGN) or a family of memoryless channels (1998-1999)
- Memoryless assumption was later relaxed (2001)
- We'll just require the following distortion constraint:

$$d^n(\mathbf{x}, \mathbf{y}) \triangleq \sum_{i=1}^n d(x_i, y_i) \leq D_2 \quad \forall \mathbf{x}, \mathbf{y} \quad (\text{wp1})$$

\Rightarrow unknown channel with arbitrary memory

- Similarly the following embedding constraint will be assumed:

$$d^n(\mathbf{s}, \mathbf{x}) \leq D_1 \quad \forall \mathbf{s}, \mathbf{k}, m, \mathbf{x} \quad (\text{wp1})$$

Data-Hiding Capacity [SM'04]

- Single-letter formula:

$$C(D_1, D_2) = \sup_{p(x,u|s) \in \mathcal{Q}(D_1)} \min_{p(y|x) \in \mathcal{A}(D_2)} [I(U; Y) - I(U; S)]$$

where U is an auxiliary random variable

$$\mathcal{Q}(D_1) = \{p_{XU|S} : \sum_{x,u,s} p(x,u|s)p(s)d(s,x) \leq D_1\}$$
$$\mathcal{A}(D_2) = \{p_{Y|X} : \sum_{x,y} p(y|x)p(x)d(x,y) \leq D_2\}$$

- Same capacity formula as in [MO'03], where $p(\mathbf{y}|\mathbf{x})$ was constrained to belong to the family $\mathcal{A}^n(D_2)$ of *memoryless* channels
- Why?

Achievability – sketch of the proof

- Random binning construction
- Randomly-permuted, constant-composition code
- Given n , solve minmax problem over types $\hat{p}(x, u|s)$ and $\hat{p}(y|x)$.

$$\Rightarrow \text{solution } (\hat{p}_{\mathbf{x}^* \mathbf{u}^* | \mathbf{s}^*}, \hat{p}_{\mathbf{y}^* | \mathbf{x}^*})$$

- All codewords $\mathbf{u}(l, m)$ are drawn uniformly from type class $T_{\mathbf{u}^*}$
- Given m, \mathbf{s} , select codeword $\mathbf{u}(l, m)$ such that

$$(\mathbf{u}(l, m), \mathbf{s}) \in T_{\mathbf{u}^* \mathbf{s}^*}$$

Then generate \mathbf{x} uniformly from cond'l type class $T_{\mathbf{x}^* | \mathbf{u}(l, m), \mathbf{s}}$

- Maximum mutual information decoder:

$$\hat{m} = \operatorname{argmax}_{l, m} I(\mathbf{u}(l, m); \mathbf{y})$$

Achievability – What is the worst attack?

- The worst $p(\mathbf{y}|\mathbf{x})$ is uniform over a single conditional type
- Example: $\mathbf{x}^* = 110100101110$

$$\mathbf{y}^* = 11\mathbf{1}100101110$$

Then all sequences \mathbf{y} that differ from \mathbf{x}^* by exactly one bit are equally likely

- For a memoryless attack, $p(\mathbf{y}|\mathbf{x})$ is uniform over multiple conditional types
- Memory does not help the attacker!
- Capacity is the same as in the memoryless case

Converse

- For any code with rate $R > C(D_1, D_2)$, there exists an attack $p(\mathbf{y}|\mathbf{x})$ such that reliable decoding is impossible
- Claim is proven by restricting search over attack channels that are uniform over a single conditional type
- Proof is similar to memoryless case, using Fano's inequality and Marton's telescoping technique

Error Exponents [MW'04]

- Obtain $E_r(R) \leq E(R) \leq E_{sp}(R)$ for all $R < C$ where

$$E(R) \triangleq \limsup_{n \rightarrow \infty} \left[-\frac{1}{n} \log P_{e,n} \right]$$

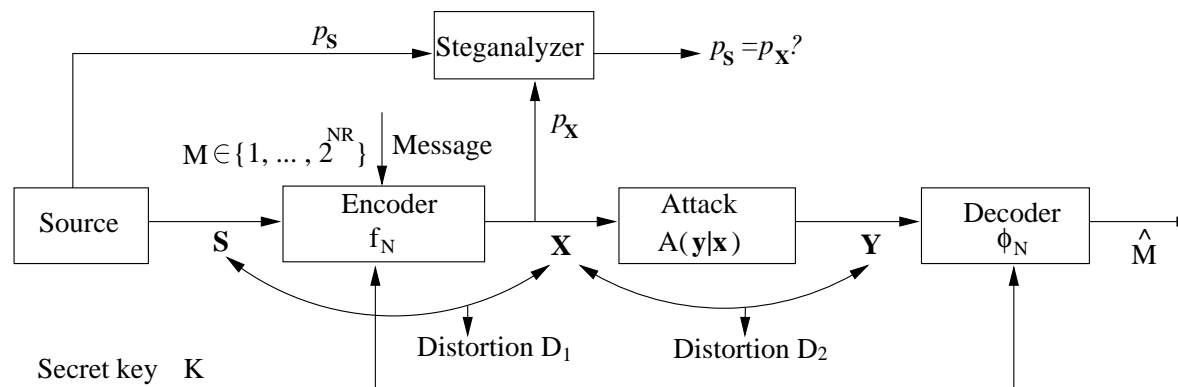
and

$$P_{e,n} = \min_{p_{F_n, G_n}} \max_{p_{\mathbf{Y}|\mathbf{X}}} P_e(F_N, G_N, p_{\mathbf{Y}|\mathbf{X}})$$

is the minmax probability of error

- Random-coding exponent $E_r(R)$ is obtained using a modification of the binning method above. Still randomly-permuted, constant-composition codes.
- Sphere-packing exponent $E_{sp}(R)$ is obtained by restricting search over attack channels that are uniform over a single conditional type

Communication Model for Steganography



- Would like $p_S = p_X$ for *perfect security*
- This is hard – matching n -dim pmf's
- Can be done using the randomization techniques described earlier \Rightarrow make \mathbf{x} is uniform over type classes
- Capacity formula is still of the form

$$C(D_1, D_2) = \sup_{p(x, u|s)} \min_{p(y|x)} [I(U; Y) - I(U; S)]$$

Conclusion

- Method of types is based on combinatorics
- Polynomial number of types
- Useful to determine capacity and error exponents
- Randomized codes, universal decoders
- Natural concepts in presence of an adversary