Applications of Random Coding and Algebraic Coding Theories to Universal Lossless Source Coding Performance Bounds

Gil I. Shamir

Department of Electrical & Computer Engineering

University of Utah

Salt Lake City, UT 84112

U.S.A.

DIMACS - 2003

Workshop on Algebraic Coding Theory and Information Theory

DIMACS Center, Rutgers University, Piscataway, NJ

December 15-18, 2003

Overview

Research Problem

- Average Case Universal Lossless Compression
- Performance Lower Bounds (on Redundancy - best possible performance of any scheme for a specific model)

Research Approach

- Use Redundancy-Capacity Theorems to obtain bounds
- Lower bound the relevant capacity for given source model

Models Discussed

- finite number of parameters parametric sources
- i.i.d. sources with large alphabets
- patterns induced by i.i.d. sources
- piecewise stationary sources
- piecewise stationary sources with slowly varying statistics
- switching sources

Universal Coding and Redundancy

Problem Layout

- A sequence x^n of length n, governed by P_θ,
- $\boldsymbol{\theta}$ unknown in a known class Λ,
- uniquely decipherable code $L(\cdot)$ may depend on Λ but **independent** of $\boldsymbol{\theta}$.
- Unknown parameters cost **redundancy**.

Average Redundancy

of code $L(\cdot)$ for n-sequences drawn by source $\boldsymbol{\theta}$

$$R_n(L, \boldsymbol{\theta}) \triangleq \frac{1}{n} E_\theta L(X^n) - H_\theta(X^n)$$

- E_θ - mean w.r.t. θ,
- H_θ - per symbol entropy.

Average Universality Measure of a Class Λ

- **Maximin** $R_n^-(\Lambda)$ and **Minimax** $R_n^+(\Lambda)$ average redundancies - best code for some worst average (over x^n) case. [Davisson, 1973]
- Average **redundancy for most sources** [Rissanen, 1984] (strongest sense).

Redundancy-Capacity Theorem

Weak Version [Implied from Davisson, 1973, Gallager, 1976]

Let $n \to \infty$. Let $\boldsymbol{\varphi}$ be a set of M points $\boldsymbol{\theta}$ in the class Λ_k, that are *distinguishable* by x^n. Then, the minimax and maximin redundancies satisfy

$$R_n^+ (\Lambda_k) = R_n^- (\Lambda_k) \geq (1 - \varepsilon) \frac{\log M}{n}$$

Strong Random Coding Version [Merhav & Feder, 1995, 1996]

Let $n \to \infty$. Define a distribution over Λ_k, and partition *most* of the class Λ_ε into disjoint countable sets $\boldsymbol{\varphi}$, where the marginal of each $\boldsymbol{\theta} \in \boldsymbol{\varphi}$ is equal, and there are $M_\phi \geq M$ sources in $\boldsymbol{\varphi}$, *distinguishable* by x^n. Then,

$$R_n (L, \boldsymbol{\theta}) \geq (1 - \varepsilon) \frac{\log M}{n},$$

for every code $L(\cdot)$, and almost every $\boldsymbol{\theta} \in \Lambda_k$.

Distinguishability

$\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ distinguishable if x^n generated by $\boldsymbol{\theta}$ appears to be generated by $\boldsymbol{\theta}'$ with probability that goes to 0 and vice versa.

Use of Redundancy-Capacity Theorem

Weak Version for $\boldsymbol{\Lambda_k}$

1. Demonstrate how to find $\boldsymbol{\varphi}$.
2. Lower bound M.
3. Prove that all $\boldsymbol{\theta} \in \boldsymbol{\varphi}$ are distinguishable by x^n.

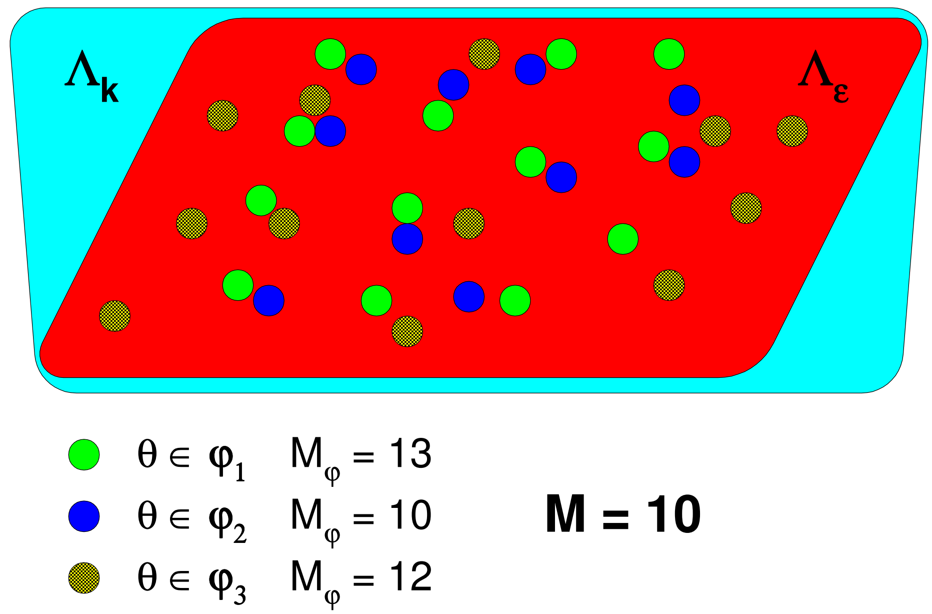
Strong Version for $\boldsymbol{\Lambda_k}$

1. Demonstrate how to define most of the class Λ_ε.
2. Show that Λ_ε is most of the class.
3. Show how to partition Λ_ε such that every source in Λ_ε is in **exactly** one $\boldsymbol{\varphi}$, and sources in $\boldsymbol{\varphi}$ are uniformly distributed with the uniform prior on Λ_k. Lower bound M.
4. Prove that for every valid $\boldsymbol{\varphi}$, all $\boldsymbol{\theta} \in \boldsymbol{\varphi}$ are distinguishable by x^n.

Compound Classes

If $\Lambda = \bigcup_k \Lambda_k$, redundancy for $\boldsymbol{\theta} \in \Lambda_k$ consists of Intra-class redundancy in Λ_k, and Inter-class redundancy distinguishing Λ_k from Λ.

Redundancy Capacity - Demo



- The volume of Λ_k outside Λ_ε assumed negligible.
- Any $\boldsymbol{\theta}$ is contained in a unique $\boldsymbol{\varphi}$ and has equal probability to other $\boldsymbol{\theta}' \in \boldsymbol{\varphi}$.
- In every $\boldsymbol{\varphi}$ all points distinguishable by x^n.

By theorem, for every code and almost every $\boldsymbol{\theta} \in \Lambda_k$,

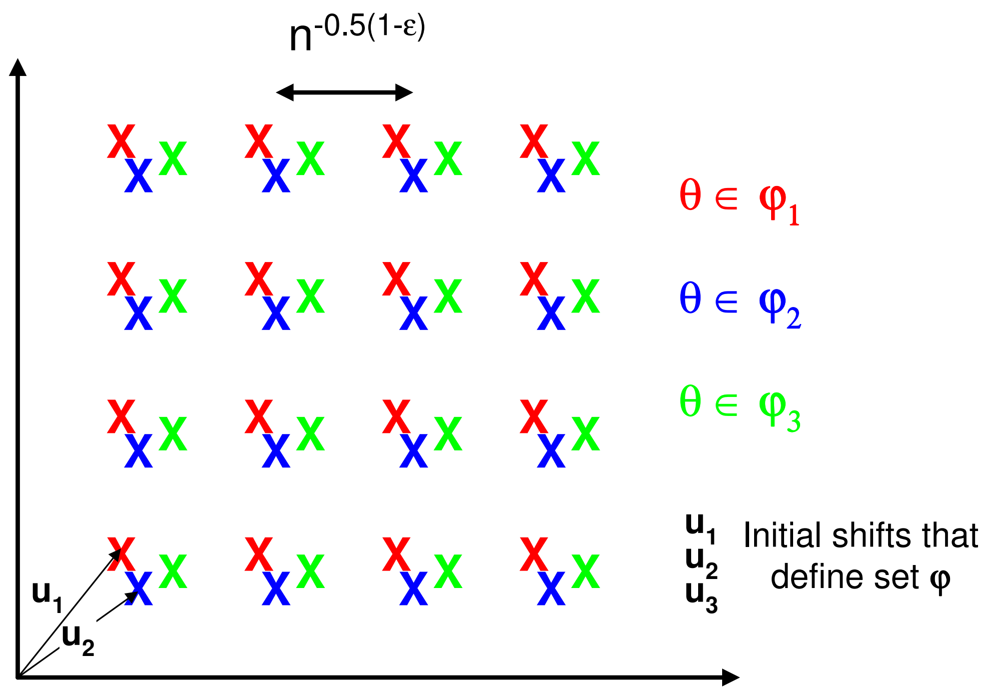
$$R_n(L, \boldsymbol{\theta}) \geq (1 - \varepsilon) \frac{\log 10}{n}$$

Finite k-dimensional Parametric Sources

- φ determined by initial shift $\mathbf{u}$ in a grid (one φ sufficient for maximin)
- $\boldsymbol{\theta} \in \varphi$ distinguishable if φ is a grid with spacing $n^{-0.5(1-\varepsilon)}$

$$R_n(L, \boldsymbol{\theta}) \geq (1-\varepsilon)\frac{k}{2}\frac{\log n}{n}$$

for every code $L(\cdot)$ and almost every $\boldsymbol{\theta} \in \Lambda$ [Rissanen, 1984]



Distinguishability

Setting and Proof in most sources sense

- Choose a random grid $\boldsymbol{\varphi}$ (as in random coding).
- Generate x^n by a given $\boldsymbol{\theta} \in \boldsymbol{\varphi}$.
- Let $\hat{\boldsymbol{\theta}}$ be the Maximum Likelihood estimator of $\boldsymbol{\theta}$ from x^n.
- Let $\hat{\boldsymbol{\theta}}_g$ be the grid point whose components are nearest $\hat{\boldsymbol{\theta}}$.
- Prove that $P_e = \Pr\left(\hat{\boldsymbol{\theta}}_g \neq \boldsymbol{\theta} \mid \boldsymbol{\theta}\right) \rightarrow 0$ as $n \rightarrow \infty$.

Use union bound on components of $\boldsymbol{\theta}$:

$$\begin{aligned} P_e &\leq \sum_{j=1}^{k} \Pr\left(\hat{\theta}_{gi} \neq \theta_i\right) \\ &\leq \sum_{j=1}^{k} n \cdot 2^{-n \cdot \min_{x^n \in A_i} D\left(P_{\hat{\theta}_i} \,||\, P_{\theta_i}\right)} \\ &\leq 2^{(\log k) + (\log n) - cn^{\varepsilon/2}} \longrightarrow 0. \end{aligned}$$

- A_i - the event that $\hat{\theta}_{gi} \neq \theta_i$.
- $D\left(P_{\hat{\theta}_i} \,||\, P_{\theta_i}\right) \geq \frac{c}{n^{1-\varepsilon}}$ for $\hat{\boldsymbol{\theta}} \in A_i$, c is constant.

I.I.D. Sources - Large Alphabet k - Minimax

[Shamir, 2003]

Problems with Large k

- Volume of Λ_k is $1/(k-1)!$ (decreases with n), because

$$\sum_{i=1}^{k-1} \theta_i \leq 1.$$

- Too large spacing in grid $n^{-0.5(1-\varepsilon)}$ results in loose bound.
- Too small spacing $(nk)^{-0.5(1-\varepsilon)}$ results in lack of distinguishability in grids.

Solution

- Build non-uniform grids.
- Spacing near $\frac{a}{n}$ proportional to $\frac{\sqrt{a}}{n^{1-\varepsilon/2}}$.
- Number of grid points preceding $\frac{a}{n}$ proportional to $\frac{\sqrt{a}}{n^{\varepsilon/2}}$.

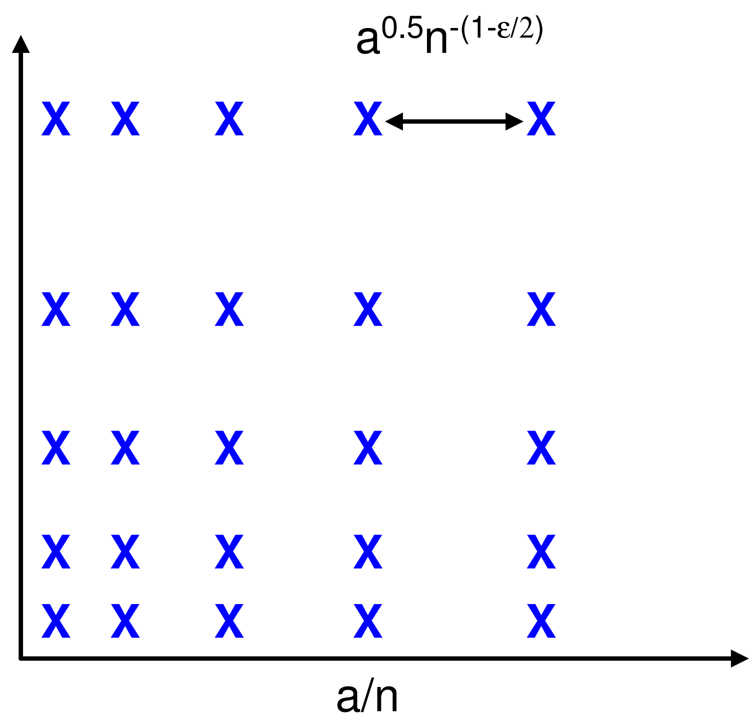
Drawback

- This structure violates the requirements of the strong version, and thus is only good for minimax/maximin redundancies.

Minimax/Maximin Redundancy - I.I.D. Large k

- $\boldsymbol{\varphi}$ is grid below,
- $\boldsymbol{\theta} \in \boldsymbol{\varphi}$ distinguishable by above definition (proved as in finite parametric case),
- bounding number of points in grid results in

$$R_n^+(\Lambda_k) = R_n^-(\Lambda_k) \geq (1-\varepsilon)\frac{(k-1)}{2n}\log\frac{n}{k}$$



Most Sources - I.I.D. Large k

Key Realizations

- Non-uniform grid above is not useful here.
- All sources outside a $k-1$ dimensional sphere with radius $r = n^{-0.5(1-\varepsilon)}$ around $\boldsymbol{\theta}$ are distinguishable from $\boldsymbol{\theta}$ by x^n.

Method

- Pack as many as possible spheres with radius r and volume $V_{k-1}(r)$ in the $k-1$ dimensional space Λ_k of volume $1/(k-1)!$.
- Place $\boldsymbol{\theta} \in \boldsymbol{\varphi}$ at centers of the spheres (whole grid shifted for random selection).
- Factor in packing density $2^{-(k-1)}$ to reduce number of points.

$$M \geq \frac{1}{(k-1)! V_{k-1}(r) 2^{(k-1)}}.$$

Result

$$R_n(L, \boldsymbol{\theta}) \geq (1-\varepsilon) \frac{(k-1)}{2n} \log \frac{n}{k}$$

for every code $L(\cdot)$ and almost every $\boldsymbol{\theta} \in \Lambda_k$. [Shamir, 2003]

Note: Second order term is lower than that of minimax/maximin bound.

Patterns Induced by I.I.D. Sources

Motivation

- Classical compression considers known small alphabets.
- Sometimes alphabet is unknown and possibly large.
- Coding cost of unknown alphabet is inevitable.

Approach

- Use the inevitable cost to improve compression.
- Code sequence *patterns* in a second stage.

Patterns

- Indices assigned to original sequence letters in order of first occurrence.
- **Example:** The strings: $x^n =$ 'lossless', 'sellsoll', '12331433', '76887288' all have the same pattern $\Psi(x^n) =$ '12331433'.
- Individual sequence redundancy studied in [Aberg, *et al.*, 1997, Orlitsky *et al.*, 2002-].

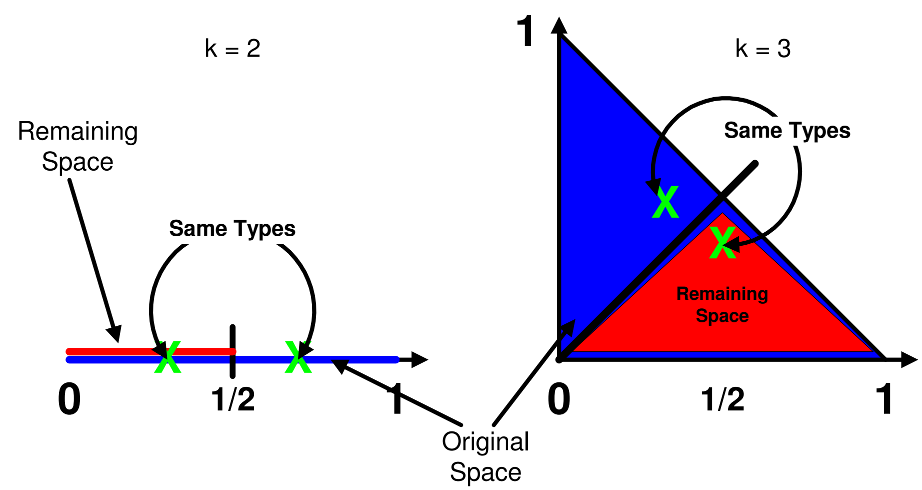
I.I.D. Induced Patterns - Derivation

- Any $\boldsymbol{\theta}'$ which is a permutation of $\boldsymbol{\theta}$ appears to be the same source.

  Example: typical sequences - similar patterns

$$\boldsymbol{\theta} = \{0.1, 0.2\} \qquad \boldsymbol{\theta}' = \{0.7, 0.2\}$$
$$x^n = 1223333333 \qquad x^n = 3221111111$$
$$\Psi(x^n) = 1223333333 \qquad \Psi(x^n) = 1223333333$$

- There are at most $k!$ such permutations.



Note: for $k = 3$ this is true for any combination of 2 out of 3 letters.

Pattern Redundancy Bounds

- The grid (in both maximin and most source cases) reduces

$$M_\Psi \geq \frac{M_{\text{i.i.d.}}}{k!}$$

- For $k \geq n^{1/3}$ too many permutations eliminated more than once, but worst smaller k can be assumed.
- More sequences contribute to correct decision in the grid to allow distinguishability.

Bounds [Shamir, 2003]

- Average minimax lower bound

$$R_n^+ \left[\Psi\left(\Lambda_k\right)\right] \geq \begin{cases} \frac{k-1}{2n} \log \frac{n^{1-\varepsilon}}{k^3} + \frac{k-1}{2n} \log \frac{\pi e^3}{2} - O\left(\frac{\log k}{n}\right), & \text{for } k \leq \left(\frac{\pi n^{1-\varepsilon}}{2}\right)^{1/3} \\ \left(\frac{\pi}{2}\right)^{1/3} \cdot (1.5 \log e) \cdot n^{-(2+\varepsilon)/3} - O\left(\frac{\log n}{n}\right), & \text{for } k > \left(\frac{\pi n^{1-\varepsilon}}{2}\right)^{1/3} \end{cases}$$

- Average most-sources lower bound

$$R_n \left[L, \boldsymbol{\psi}\left(\boldsymbol{\theta}\right)\right] \geq \begin{cases} \frac{k-1}{2n} \log \frac{n^{1-\varepsilon}}{k^3} - \frac{k-1}{2n} \log \frac{8\pi}{e^3} - O\left(\frac{\log k}{n}\right), & \text{for } k \leq \frac{1}{2} \cdot \left(\frac{n^{1-\varepsilon}}{\pi}\right)^{1/3} \\ \frac{1.5 \log e}{2\pi^{1/3}} \cdot n^{-(2+\varepsilon)/3} - O\left(\frac{\log n}{n}\right), & \text{for } k > \frac{1}{2} \cdot \left(\frac{n^{1-\varepsilon}}{\pi}\right)^{1/3} \end{cases}$$

Piecewise Stationary Sources - PSS's

Definition of PSS $\boldsymbol{\psi} \triangleq (\boldsymbol{\theta}, \mathbf{t}) \in \boldsymbol{\Lambda_q} \subset \boldsymbol{\Lambda}$

- PSS - emits data divided into independent stationary segments separated by abrupt changes in statistics
- Λ - nth order class of PSS's (contains all possible combinations of the k-dimensional parameters for n-sequences)
- Λ_q - All PSS's in Λ with q segments
- $\boldsymbol{\theta} \triangleq \{\theta_1, \theta_2, \ldots, \theta_q\}$ - *segmental parameters*
- $\mathbf{t} \triangleq \{t_1, t_2, \ldots, t_{q-1}\}$ - *transition path* (TP)

Redundancy bound [Shamir, 2000]

$$R_n\left(L, \boldsymbol{\psi}\right) \geq (1-\varepsilon)\left(\frac{1}{2}kq + q - 1\right)\frac{\log(n/q)}{n}$$

for every $L(\cdot)$, for almost every $\boldsymbol{\psi} \in \Lambda_q$, for every q.
in the minimax/maximin senses.

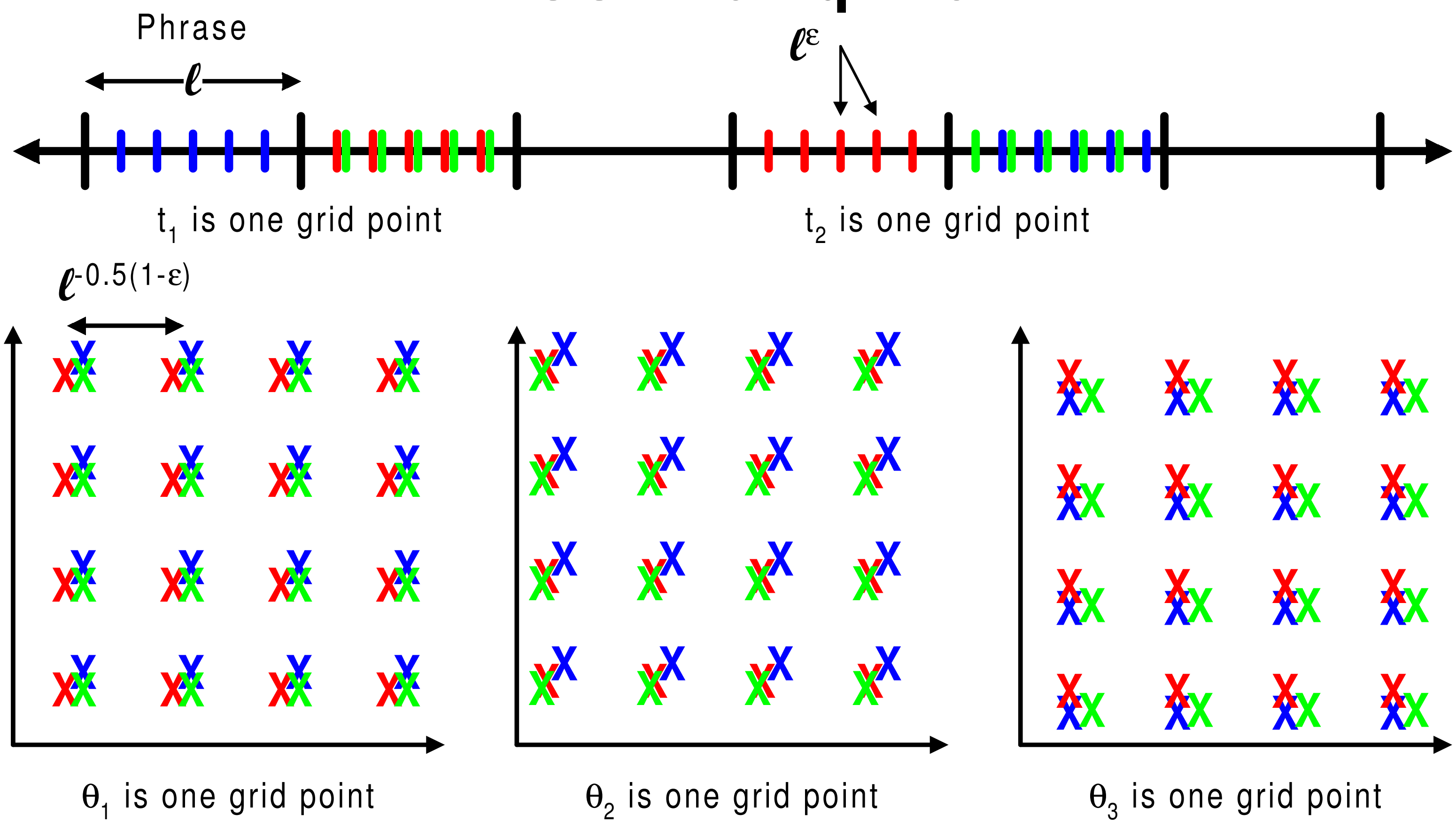
Bound Derivation - PSS's

Finite Number of Segments q

1. Λ_ε contains all $\boldsymbol{\psi}$ for which **all** segments long (longer than $n^{1-\varepsilon/2}$), and **all** transitions are large.
2. Λ_ε is most class for fixed q.
3. Partition Λ_ε into sets as follows:
   - Parse n-tuple to **phrases** of length $l = n^{1-\varepsilon}$.
   - For all $\boldsymbol{\psi} \in \boldsymbol{\varphi}$, $\forall i$, t_i is a point in the same phrase in a grid with spacing l^ε.
   - θ_i is a point in a grid as defined for stationary sources.
   - $\forall \boldsymbol{\psi} \in \boldsymbol{\varphi}$, t_i and θ_i must be from grids with identical initial shifts.
4. Distinguish among $\boldsymbol{\psi} \in \boldsymbol{\varphi}$ as follows:
   - Use phrases entirely inside segments to estimate θ_i.
   - Given $\hat{\boldsymbol{\theta}}$, estimate transitions from respective grids.

By definition of the grids, the bound for finite q [Merhav, 1993] results.

PSS with q = 3



$\psi \in \varphi_1 : \theta_1, \theta_2, \theta_3, t_1, t_2$ **are only red points**

$\psi \in \varphi_2 : \theta_1, \theta_2, \theta_3, t_1, t_2$ **are only blue points**

$\psi \in \varphi_3 : \theta_1, \theta_2, \theta_3, t_1, t_2$ **are only green points**

Set φ_i contains all combinations with one point from each of the five grids

General Bound Derivation - PSS's

Large q

1. Λ_ε defined is **not** most class.
2. For very large q, probability of error in **at least one** of the source parameters significantly increases the overall error probability.

Solutions to Asymptotic Problems

1. Λ_ε contains sources for which **most** segments are long and **most** transitions are large.
2. Reduce sets $\boldsymbol{\varphi}$ to improve distinguishability for very large q.

Two different Cases

- $q \not\gg n/q$ - almost similar to fixed q (modified according to modification 1 above).
- $q \gg n/q$ - requires additional **algebraic coding** techniques for distinguishability.

General Bound Derivation - PSS's, Cont.

Second Case: $q \gg n/q$

- Too many parameters.
- Error in estimating one results in error in estimating $\boldsymbol{\psi}$.

Solution - Reduce φ by Linear Block Codes:

- Let $\eta > 0$ be arbitrarily small,
- q' - number of 'free' segmental parameters,
- c' - number of 'free' transition times.

- $(1-\eta)\,q'$ segmental parameters and $(1-\eta)\,c'$ transitions chosen from grids.
- Remaining parameters are parity checks.
- Grids' resolutions chosen to yield Galois Fields.
- Each grid point is assigned an element in the proper Galois Field.
- Codes designed to correct up to $\alpha\eta q'$ errors (exist: Gilbert-Varshamov).

Guarantees distinguishability even for $q \gg n/q$, resulting in the same asymptotic bound (ε is now larger).

Additional Source Classes

PSS's with Slowly Linearly Varying Statistics [Shamir, 2001]

- q segments, transition duration of $(n/q)^{\alpha}$:

$$R_n(L,\boldsymbol{\psi}) \geq (1-\varepsilon)\left[\frac{kq}{2} + (q-1)\left(1-\frac{\alpha}{2}\right)\right]\frac{\log(n/q)}{n}$$

- If durations unknown,

$$R_n(L,\boldsymbol{\psi}) \geq (1-\varepsilon)\left(\frac{1}{2}kq + q - 1\right)\frac{\log(n/q)}{n}$$

Hierarchical version of redundancy-capacity for compound class must be used. Insignificant cost above PSS's.

Switching Sources - *s* states [Shamir, 2001]

If $s \leq (n/q)^{0.5k(1-\varepsilon)}$. Then, for every code $L(\cdot)$ and almost all sources

$$R_n(L,\boldsymbol{\psi}) \geq \frac{(1-\varepsilon)}{n}\left[\frac{ks}{2}\log(n/s) + (q-1)\log(n/q) + (q-s)\log s\right]$$

Otherwise,

$$R_n(L,\boldsymbol{\psi}) \geq \frac{(1-\varepsilon)}{n}\left(\frac{1}{2}kq + q - 1\right)\log(n/q)$$

Summary and Conclusions

1. The redundancy-capacity theorem is very useful to derive lower bounds on
    - minimax/maximin redundancy in universal coding,
    - redundancy for most sources in universal coding.
2. Lower bounds on redundancy in both cases were obtained for
    - finite number of parameters parametric sources,
    - i.i.d. sources with large alphabets,
    - patterns induced by i.i.d. sources,
    - piecewise stationary sources,
    - piecewise stationary sources with slowly varying statistics,
    - switching sources.
3. Different techniques from coding theory were used:
    - random coding,
    - sphere packing,
    - algebraic code distance bounds.