Noise Removal -
An Information Theoretic View

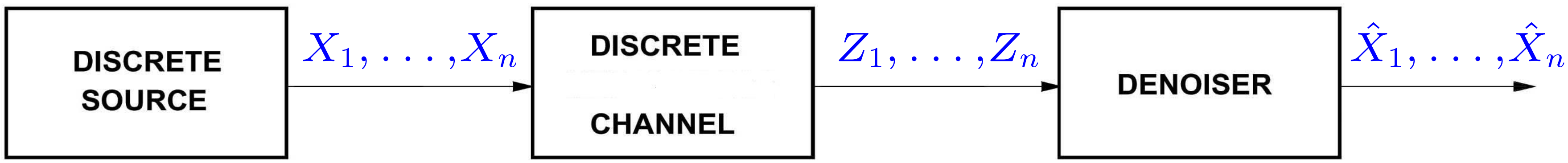
048703
Technion, EE Dept.

Spring 2008

Why Another Course on Noise Removal ?

Keep that thought. Answers later.

Discrete Denoising



- X_i, Z_i, $\hat{X}_i$ take values in *finite alphabets*
- Goal: Choose $\hat{X}_1, \ldots, \hat{X}_n$ on the basis of $Z_1, \ldots, Z_n$ which will be "close" to $X_1, \ldots, X_n$
- Closeness is under given "single-letter" loss function Λ

Why discrete ?

- Finite alphabets allow to focus on the essentials
- Discrete data becoming increasingly ubiquitous
- Insight from discrete case turns out fruitful also for the analogue world

Example I: Bit Streams

$$\ldots 0001111100001111100 \ldots \quad \overset{corruption}{\Longrightarrow} \quad \ldots 0001000001000001010 \ldots$$

Example II: Text

Original Text:

”What giants?” said Sancho Panza. ”Those thou seest there,” answered his master, ”with the long arms, and spne have them nearly two leagues long.” ”Look, your worship,” said Sancho; ”what we see there are not giants but windmills, and what seem to be their arms are the sails that turned by the wind make the millstone go.” ”It is easy to see,” replied Don Quixote, ”that thou art not used to this business of adventures; those are giants; and if thou are afraid, away with thee out of this and betake thyself to prayer while I engage them in fierce and unequal combat.”

Corrupted Text:

”Whar giants?” said Sancho Panza. ”Those thou seest theee,” snswered yis master, ”with the long arms, and spne have tgem ndarly two leagues long.” ”Look, ylur worship,” sair Sancho; ”what we see there zre not gianrs but windmills, and what seem to be their arms are the sails that turned by the wind make rhe millstpne go.” ”Kt is easy to see,” replied Don Quixote, ”that thou art not used to this business of adventures; fhose are giantz; and if thou arf wfraod, away with thee out of this and betake thysepf to prayer while I engage them in fierce and unequal combat.”

Example III: Biological Data

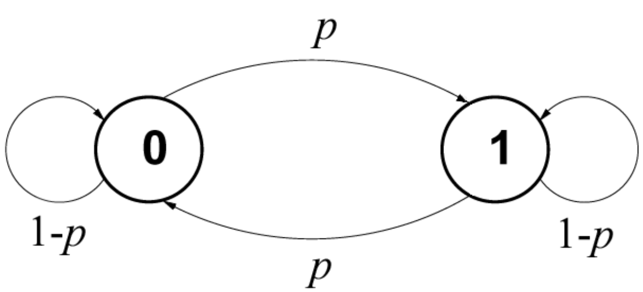
$$\ldots AGCATTCGATGCTTAAAGA \ldots \quad \stackrel{corruption}{\Longrightarrow} \quad \ldots AGC{\color{red}G}TTCGA{\color{red}A}GCTTA{\color{red}T}A{\color{red}C}A \ldots$$

The "easy" life: known $P_{\mathbf{X}}$ and channel

- Fundamental performance limits
- Optimal but non-universal schemes:
  - Bayes-optimal schemes (not necessarily so easy..)
  - But sometimes life is good: forward-backward recursions for noise-corrupted Markov processes

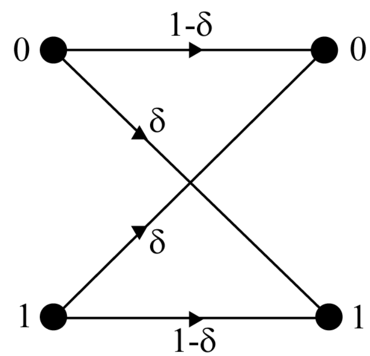
The Easy Life: Example I

Source: Binary Markov Chain



$\ldots 0001111100001111100 \ldots$

Channel: BSC

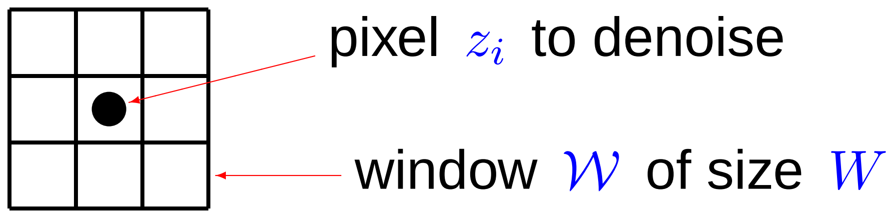


$\ldots 00010000001000001010 \ldots \quad \Rightarrow \ldots 0000111101001110110 \ldots$

- Objective: Minimize Bit Error Rate given the observation of n-block.
- Solution: Backward-Forward Dynamic Programming

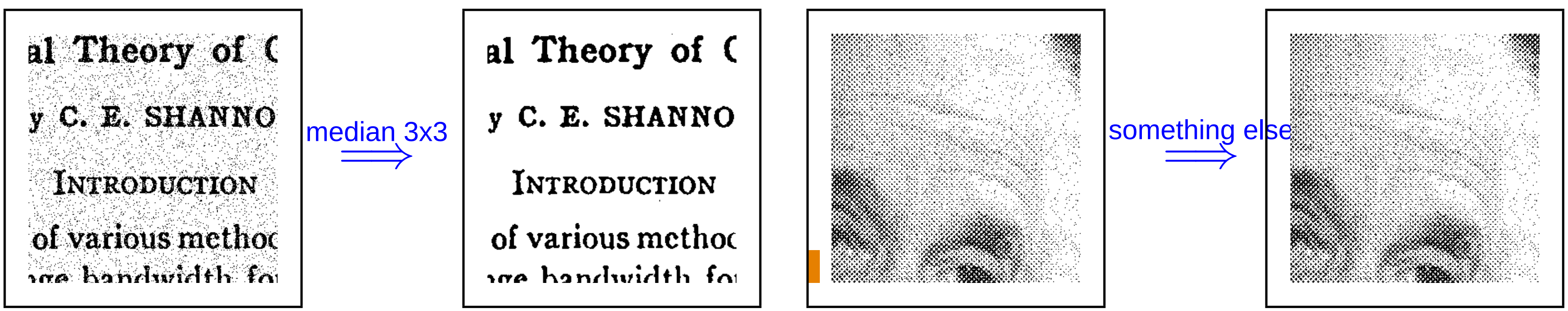
The Easy Life: Example II

- Many successful algorithms are *window-based*



$$\hat{x}_i = f(\mathcal{W})$$

- When type of data is known a priori, we may know which rule to use:

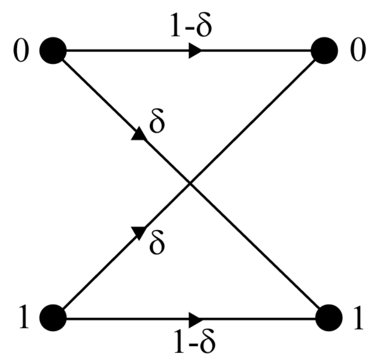


works well *would work well*

The Real Life: Example I

Source: ?

Channel: BSC



$\ldots 00011111100001111100 \ldots$ $\ldots 00010000001000001010 \ldots$ $\Rightarrow \ldots$?

- Objective: Minimize Bit Error Rate given the observation of n-block.

- Solution: ?

Initial Setting

- Unknown source of data
- Known corruption mechanism (memoryless channel)

$$\Pi(x, z) = \text{Prob}(z \text{ observed} \mid x \text{ clean})$$

- Given loss function

$$\mathbf{\Lambda}(x, \hat{x})$$

Approaches

- Numerous heuristics
- HMM-based plug-in techniques
- Compression-based approach
- DUDE

The DUDE Algorithm: General Idea

Fix *context length* k. For each letter x_i to be denoised, do:

- Find *left* k*-context* $(\ell_1, \ldots, \ell_k)$ and *right* k*-context* $(r_1, \ldots, r_k)$

ℓ_1	ℓ_2	$\cdots$	ℓ_k	$\bullet$	r_1	r_2	$\cdots$	r_k

- Count all occurrences of letters with left k-context $(\ell_1, \ldots, \ell_k)$ and right k-context $(r_1, \ldots, r_k)$.

- Decide on $\hat{x}_i$ according to

$$\hat{x}_i = \text{simple rule}(\mathbf{\Lambda}, \Pi, \text{count vector}, z_i)$$

Noiseless Text

We might place the restriction on allowable sequences that no spaces follow each other. $\cdots$ effect of statistical knowledge about the source in reducing the required capacity of the channel $\cdots$ the relative frequency of the digram $i\ j$. The letter frequencies $p(i)$, the transition probabilities $\cdots$ The resemblance to ordinary English text increases quite noticeably at each of the above steps. $\cdots$ This theorem, and the assumptions required for its proof, are in no way necessary for the present theory. $\cdots$ The real justification of these definitions, however, will reside in their implications. $\cdots$ H is then, for example, the H in Boltzmann's famous H theorem. We shall call $H = -\sum p_i \log p_i$ the entropy of the set of probabilities $p_1, \ldots, p_n$. $\cdots$ The theorem says that for large N this will be independent of q and equal to H. $\cdots$ The next two theorems show that H and H' can be determined by limiting operations directly from the statistics of the message sequences, without reference to the states and transition probabilities between states. $\cdots$ The Fundamental Theorem for a Noiseless Channel $\cdots$ The converse part of the theorem, that $\frac{C}{H}$ cannot be exceeded, may be proved by noting that the entropy $\cdots$ The first part of the theorem will be proved in two different ways. $\cdots$ Another method of performing this coding and thereby proving the theorem can be described as follows: $\cdots$ The content of Theorem 9 is that, although an exact match is $\cdots$ With a good code the logarithm of the reciprocal probability of a long message must be proportional to the duration of the corresponding signal $\cdots$

Noisy text

Wz right peace the rest iction on alksoable sequbole thgt wo spices fokiow eadh otxer. $\cdots$ egfbct of sraaistfcal keowleuge apolt tje souwce in recucilg the requihed clpagity ofythe clabbel $\cdots$ the relatrte pweqiency ofpthe digram $i\ j$. The setter freqbwncles $p(i)$, ghe rrahsibion probtbilities $\cdots$ The resemglahca to ordwnard Engdish tzxt ircreakes quitq noliceabcy at vach oftthe hbove steps. $\cdots$ Thus theorev, andlthe aszumptjona requiyed ffr its croof, arv il no wsy necqssrry forptfe prwwent theorz. $\cdots$ jhe reap juptifocation of dhese defikjtmons, doweyer, bill rehide inytheir imjlycajijes. $\cdots$ H is them, fol eskmqle, tle H in Bolgnmann's falous H themreg. We vhall cbll $H = -\sum p_i \log p_i$ the wntgopz rf thb set jf prwbabjlities $p_1, \ldots, p_n$. $\cdots$ The theorem sahs tyat fsr lawge N mhis gill we hndeypensdest of q aed vqunl tj H. $\cdots$ The neht txo theiremf scow tyat H and H' can be degereined jy likitkng operatiofs digectlt fgom the stgtissics of thk mfssagj siqufnves, bithout referenge ty the htates and trankituon krobabilitnes bejwekn ltates. $\cdots$ The Fundkmendal Theorem kor a Soiselesd Chjnnen $\cdots$ Lhe ronvegse jaht jf tketheorem, thlt $\frac{C}{H}$ calnot be excweded, may ke xroved ey hotijg tyat the enyropy $\cdots$ The first pajt if the theqrem will be ptoved in two kifferent wjys. $\cdots$ Another methjd of plrfolming shis goding ald thmreby proking toe oheorem can bexdescrined as folfows: $\cdots$ The contemt ov The rem 9 if thst, ajthorgh an ezacr mawwh is $\cdots$ Wotf a goul code therlogaretym of the rehitrocpl prossbilfly of a lylg mwgsage lust be priporyiopal to tha rurafirn of $\cdots$

Noisy text: Denoising m

Wz right peace the rest iction on alksoable sequbole thgt wo spices fokiow eadh otxer. $\cdots$ egfbct of sraaistfcal keowleuge apolt tje souwce in recucilg the requihed clpagity ofythe clabbel $\cdots$ the relatrte pweqiency ofpthe digram $i\ j$. The setter freqbwncles $p(i)$, ghe rrahsibion probtbilities $\cdots$ The resemglahca to ordwnard Engdish tzxt ircreakes quitq noliceabcy at vach oftthe hbove steps. $\cdots$ Thus theorev, andlthe aszumptjona requiyed ffr its croof, arv il no wsy necqssrry forptfe prwwent theorz. $\cdots$ jhe reap juptifocation of dhese defikjtmons, doweyer, bill rehide inytheir imjlycajijes. $\cdots$ H is them, fol eskmqle, tle H in Bolgnmann's falous H themreg. We vhall cbll $H = -\sum p_i \log p_i$ the wntgopz rf thb set jf prwbabjlities $p_1, \ldots, p_n$. $\cdots$ The theorem sahs tyat fsr lawge N mhis gill we hndeypensdest of q aed vqunl tj H. $\cdots$ The neht txo theiremf scow tyat H and H' can be degereined jy likitkng operatiofs digectlt fgom the stgtissics of thk mfssagj siqufnves, bithout referenge ty the htates and trankituon krobabilitnes bejwekn ltates. $\cdots$ The Fundkmendal Theorem kor a Soiselesd Chjnnen $\cdots$ Lhe ronvegse jaht jf tketheorem, thlt $\frac{C}{H}$ calnot be excweded, may ke xroved ey hotijg tyat the enyropy $\cdots$ The first pajt if the theqrem will be ptoved in two kifferent wjys. $\cdots$ Another methjd of plrfolming shis goding ald thmreby proking toe oheorem can bexdescrined as folfows: $\cdots$ The contemt ov The rem 9 if thst, ajthorgh an ezacr mawwh is $\cdots$ Wotf a goul code therlogaretym of the rehitrocpl prossbilfly of a lylg mwgsage lust be priporyiopal to tha rurafirn of $\cdots$

Context search $k = 2$ | h | e | • | r | e |

Wz right peace the rest iction on alksoable sequbole thgt wo spices fokiow eadh otxer. $\cdots$ egfbct of sraaistfcal keowleuge apolt tje souwce in recucilg the requihed clpagity ofythe clabbel $\cdots$ the relatrte pweqiency ofpthe digram $i\ j$. The setter freqbwncles $p(i)$, ghe rrahsibion probtbilities $\cdots$ The resemglahca to ordwnard Engdish tzxt ircreakes quitq noliceabcy at vach oftthe hbove steps. $\cdots$ Thus theorev, andlthe aszumptjona requiyed ffr its croof, arv il no wsy necqssrry forptfe prwwent theorz. $\cdots$ jhe reap juptifocation of dhese defikjtmons, doweyer, bill rehide inytheir imjlycajijes. $\cdots$ H is them, fol eskmqle, tle H in Bolgnmann's falous H themreg. We vhall cbll $H = -\sum p_i \log p_i$ the wntgopz rf thb set jf prwbabjlities $p_1, \ldots, p_n$. $\cdots$ The theorem sahs tyat fsr lawge N mhis gill we hndeypensdest of q aed vqunl tj H. $\cdots$ The neht txo theiremf scow tyat H and H' can be degereined jy likitkng operatiofs digectlt fgom the stgtissics of thk mfssagj siqufnves, bithout referenge ty the htates and trankituon krobabilitnes bejwekn ltates. $\cdots$ The Fundkmendal Theorem kor a Soiselesd Chjnnen $\cdots$ Lhe ronvegse jaht jf tketheorem, thlt $\frac{C}{H}$ calnot be excweded, may ke xroved ey hotijg tyat the enyropy $\cdots$ The first pajt if the theqrem will be ptoved in two kifferent wjys. $\cdots$ Another methjd of plrfolming shis goding ald thmreby proking toe oheorem can bexdescrined as folfows: $\cdots$ The contemt ov The rem 9 if thst, ajthorgh an ezacr mawwh is $\cdots$ Wotf a goul code therlogaretym of the rehitrocpl prossbilfly of a lylg mwgsage lust be priporyiopal to tha rurafirn of $\cdots$

Context search $k = 2$ $\boxed{\textbf{h}}\boxed{\textbf{e}}\boxed{\bullet}\boxed{\textbf{r}}\boxed{\textbf{e}}$ counts

- he re : 7, heore : 5, heire : 1, hemre : 1, heqre : 1

$$\mathbf{m}(\texttt{Shannon text}, he, re) = [0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,5\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,7]^T$$

$$\begin{array}{ccccc} \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ i & m & o & q & \texttt{sp} \end{array}$$

The reconstruction at the point i we looked at is:

$$\hat{x}_i = \text{simple rule}\,(\mathbf{\Lambda}, \Pi, \mathbf{m}(\texttt{Shannon text}, he, re), \texttt{m})$$

The DUDE Algorithm for Multi-D Data

Same algorithm.

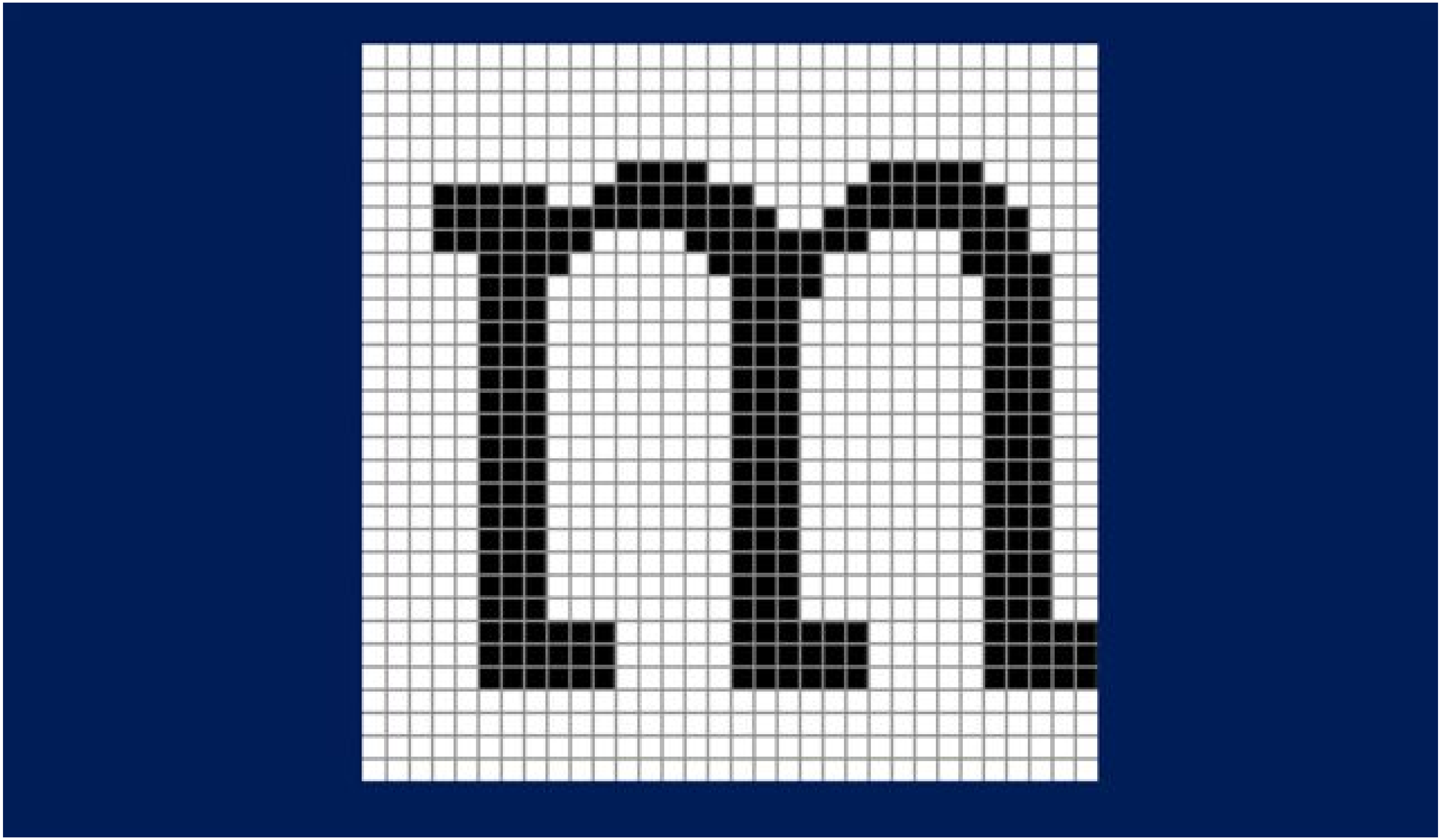
- Contexts are of form:

$$
\begin{array}{ccccc}
 & & c_9 & & \\
 & c_8 & c_1 & c_2 & \\
c_{12} & c_7 & \bullet & c_3 & c_{10} \\
 & c_6 & c_5 & c_4 & \\
 & & c_{11} & &
\end{array}
$$

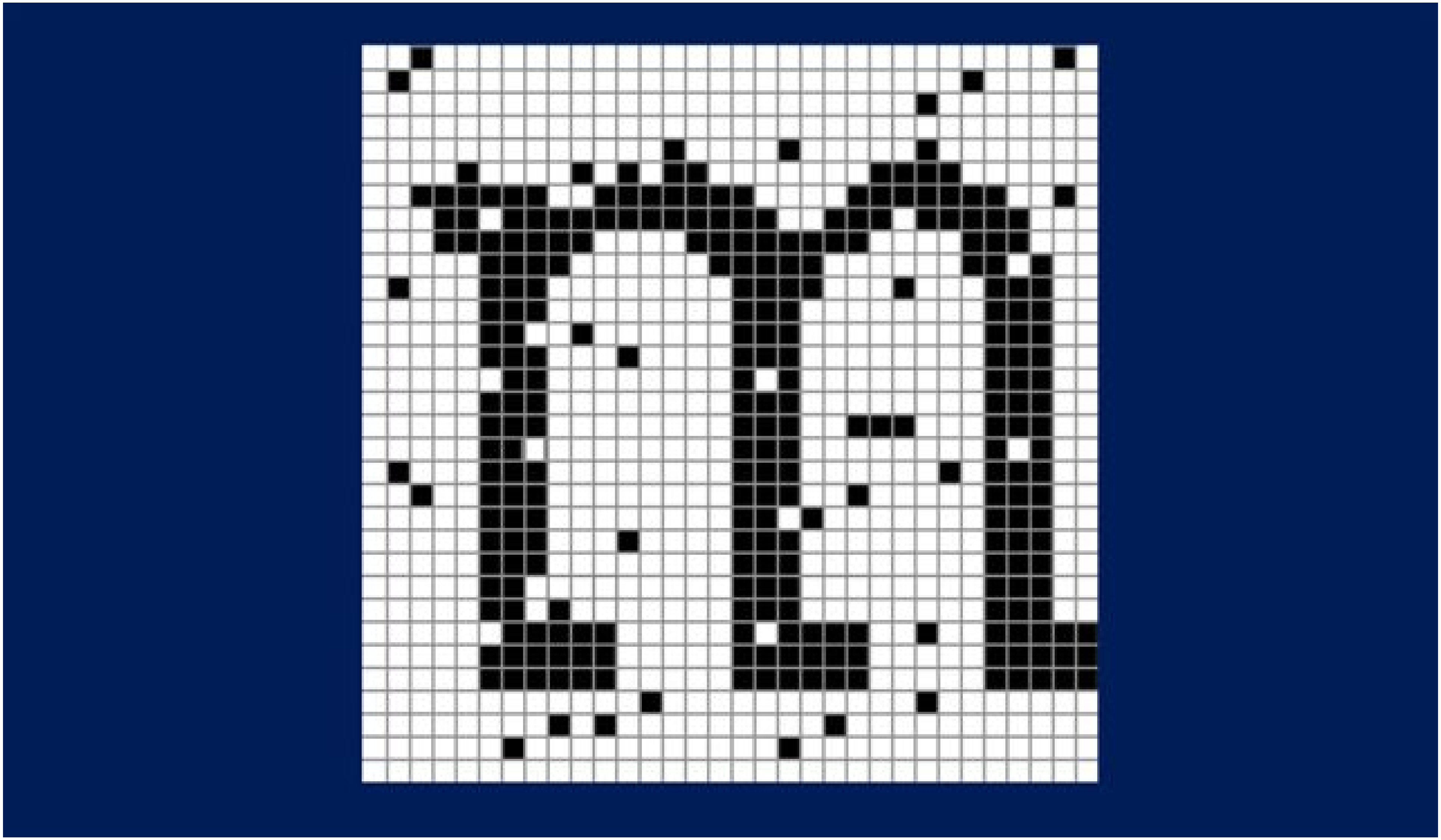
Example:

$$
\begin{aligned}
&K = 12 \\
&\{\,(0,\pm 1), \\
&\quad(\pm 1, 0), \\
&\quad(\pm 1, \pm 1), \\
&\quad(0, \pm 2), \\
&\quad(\pm 2, 0)\,\}
\end{aligned}
$$

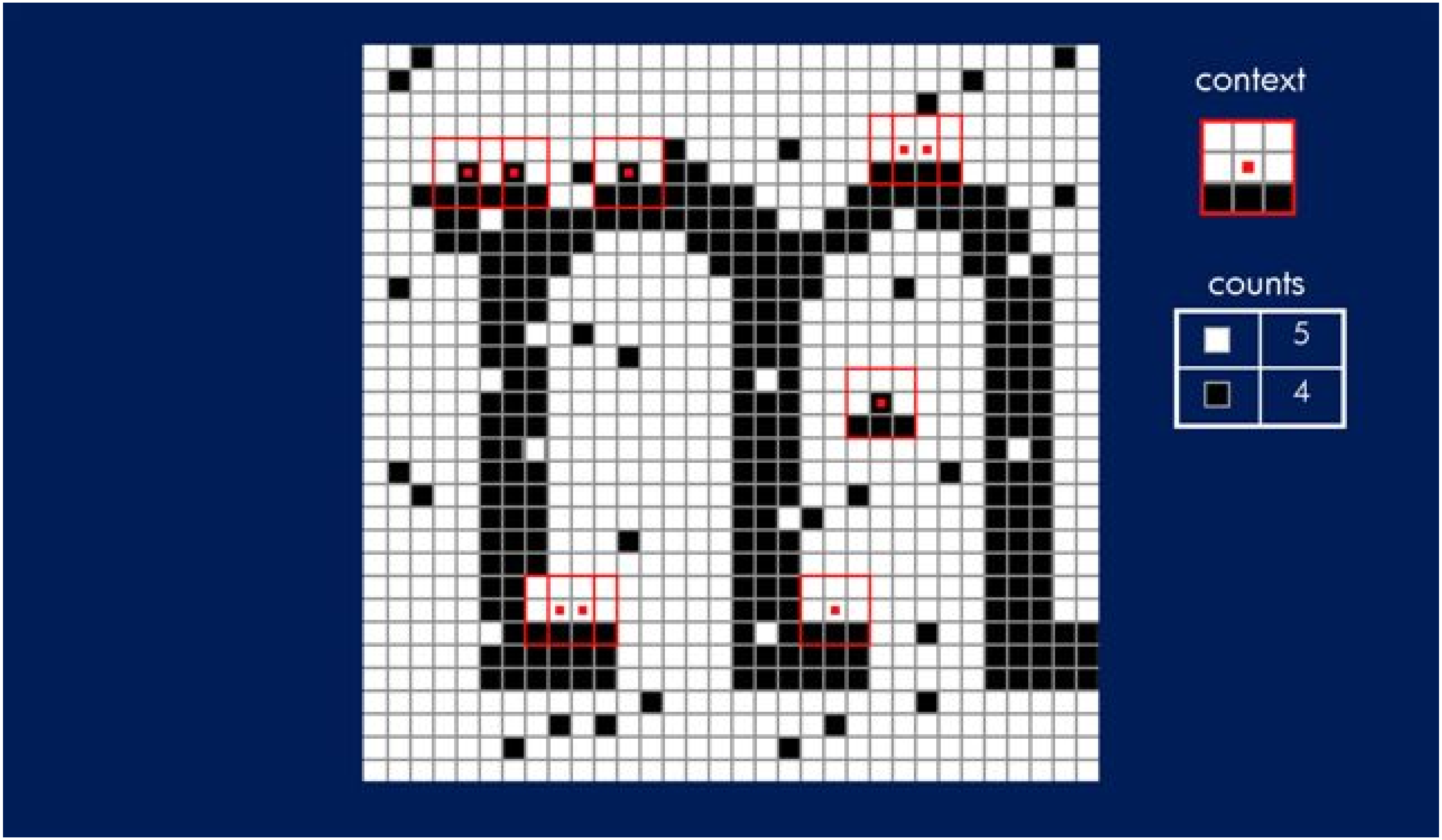
EXAMPLE: Binary Image



EXAMPLE: Noisy Binary Image



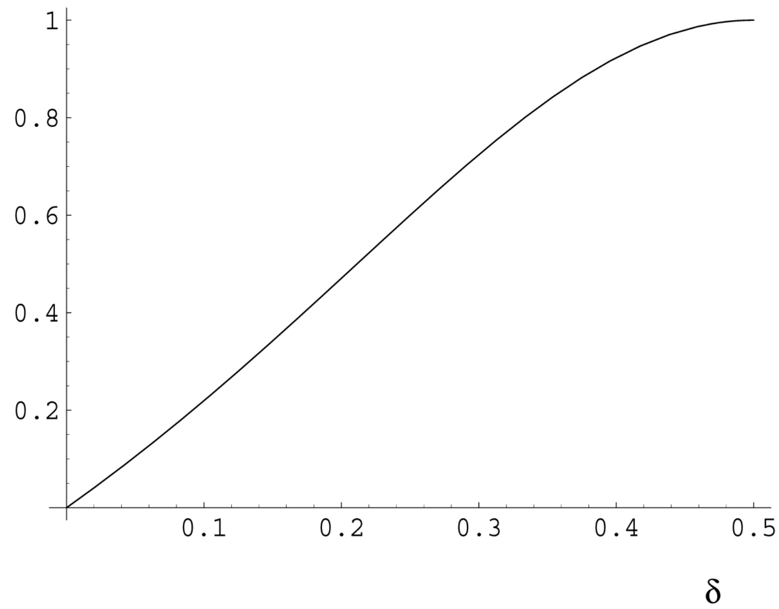
EXAMPLE: Context Symbol Counts



Example: BSC + BER

For each bit b, count how many bits that have the same left and right k-contexts are equal to b and how many are equal to $\bar{b}$. If the ratio of these counts is below

$$\frac{2\delta(1-\delta)}{(1-\delta)^2+\delta^2}$$



then b is deemed to be an error introduced by the BSC.

Example: M-ary erasure channel + Per-Symbol Error Rate

Correct every erasure with the most frequent symbol for its context

Choosing the Context Length k

- Tradeoff:
  - too short $\mapsto$ suboptimum performance
  - too long ($\Leftrightarrow$ too short n) $\mapsto$ counts are unreliable
- Our choice: $k = k_n = \left\lceil \frac{1}{2} \log_{|\mathcal{Z}|} n \right\rceil$

Computational Complexity

Linear

Experiment: Binary Markov Chain (p) $\rightarrow$ BSC (δ); $n = 10^5$

p	$\delta = 0.01$		$\delta = 0.10$		$\delta = 0.20$	
	DUDE	ForwBack	DUDE	ForwBack	DUDE	ForwBack
0.01	0.000723	0.000721	0.006648	0.005746	0.025301	0.016447
0.05	0.004223	0.004203	0.030084	0.029725	0.074936	0.071511
0.10	0.010213	0.010020	0.055976	0.055741	0.120420	0.118661
0.15	0.010169	0.010050	0.075474	0.075234	0.153182	0.152903
0.20	0.009994	0.009940	0.092304	0.092304	0.176354	0.176135

Table 1: Bit Error Rates

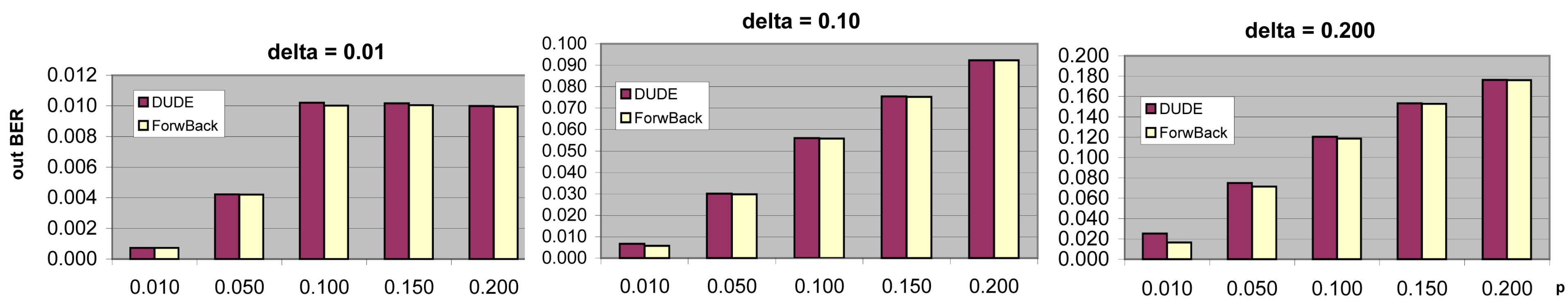


Image Denoising: δ=0.05

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

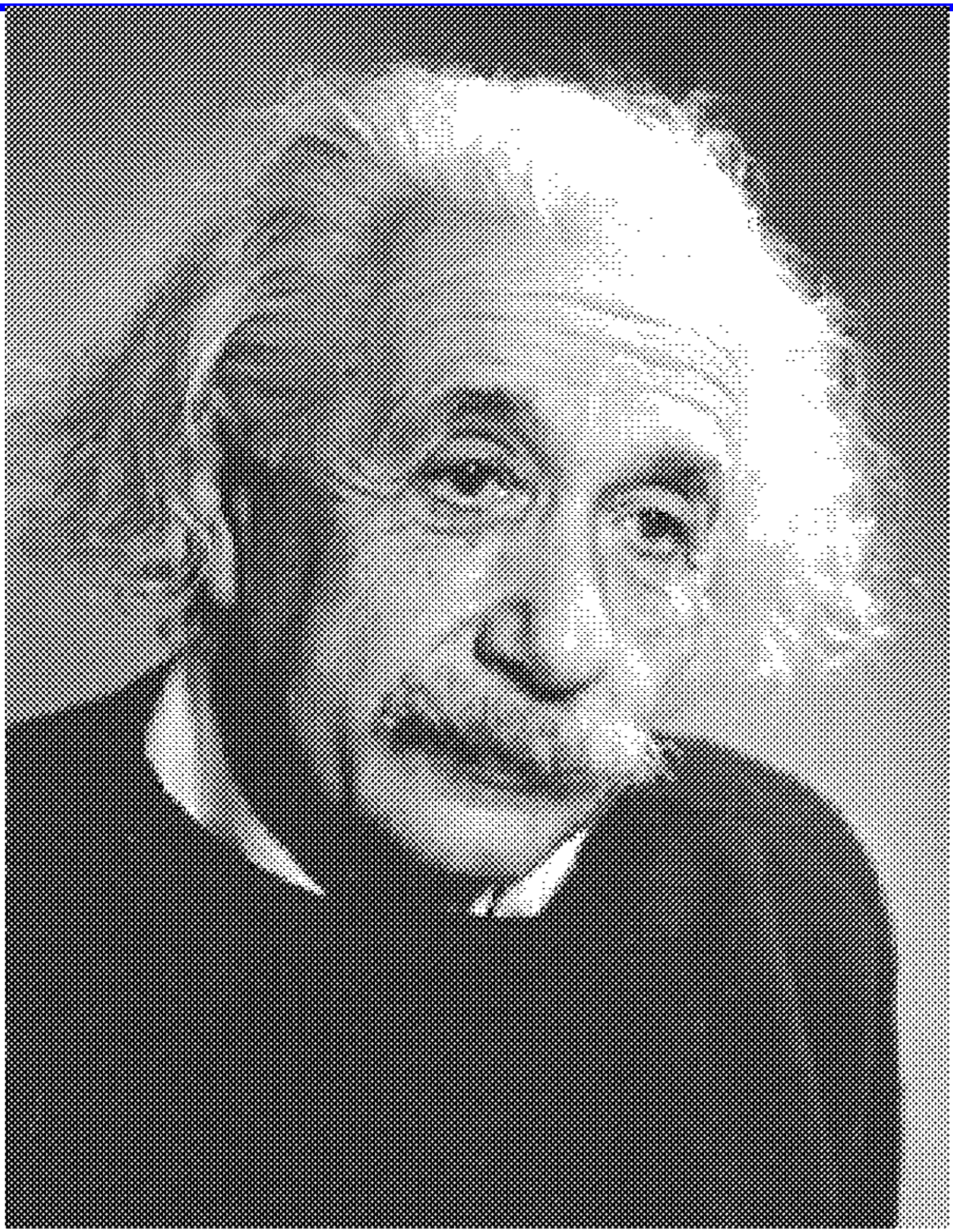
1. It is practically more useful. Parameters of engineering importance

[1] Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, April 1924, p. 324; "Certain Topics in Telegraph Transmission Theory," *A. I. E. E. Trans.*, v. 47, April 1928, p. 617.
[2] Hartley, R. V. L., "Transmission of Information," *Bell System Technical Journal*, July 1928, p. 535.

Published in THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 27, pp. 379-423, 623-656, July, October, 1948

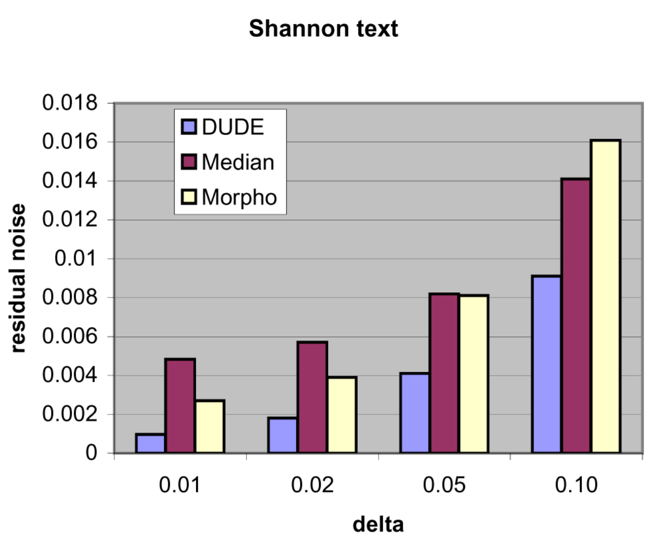
MONOGRAPH B-1598
Reissued December, 1957

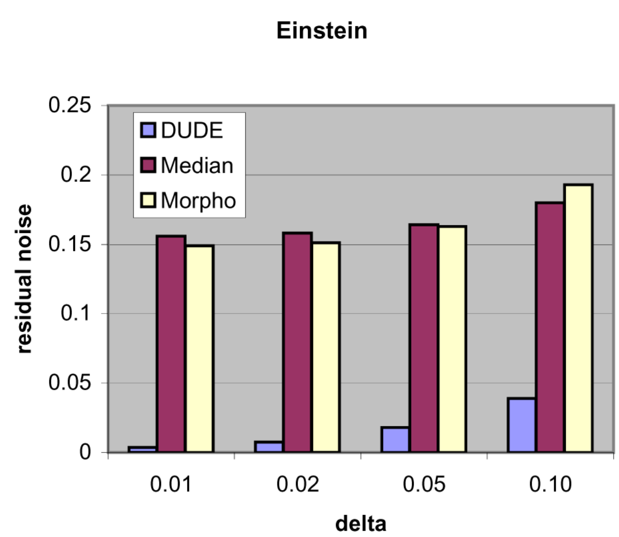
Image Denoising: $\delta = 0.02$



Comparison with known algorithms

		Channel parameter δ			
Image	Scheme	0.01	0.02	0.05	0.10
Shannon 1800×2160	DUDE	0.00096	0.0018	0.0041	0.0091
	median	0.00483	0.0057	0.0082	0.0141
	morpho.	0.00270	0.0039	0.0081	0.0161
Einstein 896×1160	DUDE	0.0035	0.0075	0.0181	0.0391
	median	0.156	0.158	0.164	0.180
	morpho.	0.149	0.151	0.163	0.193





Text Denoising: Don Quixote de La Mancha

Noisy Text (21 errors, 5% error rate):

”Whar giants?” said Sancho Panza. ”Those thou seest theee,” snswered yis master, ”with the long arms, and spne have tgem ndarly two leagues long.” ”Look, ylur worship,” sair Sancho; ”what we see there zre not gianrs but windmills, and what seem to be their arms are the sails that turned by the wind make rhe millstpne go.” ”Kt is easy to see,” replied Don Quixote, ”that thou art not used to this business of adventures; fhose are giantz; and if thou arf wfraod, away with thee out of this and betake thysepf to prayer while I engage them in fierce and unequal combat.”

DUDE output (4 errors):

”What giants?” said Sancho Panza. ”Those thou seest there,” answered his master, ”with the long arms, and spne have them nearly two leagues long.” ”Look, your worship,” said Sancho; ”what we see there are not giants but windmills, and what seem to be their arms are the sails that turned by the wind make the millstone go.” ”It is easy to see,” replied Don Quixote, ”that thou art not used to this business of adventures; fhose are giantz; and if thou are afraid, away with thee out of this and betake thyself to prayer while I engage them in fierce and unequal combat.”

Text Denoising: Don Quixote de La Mancha (cont.)

Noisy Text (4 errors):

... in the service of such a masger ws Dpn Qhixote ...

DUDE output, (0 errors):

... in the service of such a master as Don Quixote ...

Measure of Performance

[Normalized cumulative loss] of the denoiser $\hat{X}^n$ when the observed sequence is $z^n \in \mathcal{A}^n$ and the underlying clean sequence is $x^n \in \mathcal{A}^n$:

$$L_{\hat{X}^n}(x^n, z^n) = \frac{1}{n}\sum_{i=1}^{n} \Lambda(x_i, \hat{x}_i),$$

where

$$\hat{x}_i = \hat{X}^n(z^n)[i]$$

We denote the DUDE by

$$\hat{X}^n_{DUDE}$$

Optimality Result: Stochastic Setting

Theorem.

For every stationary noise-free signal $\mathbf{X}$,

$$\lim_{n\to\infty}\left[EL_{\hat{X}^n_{DUDE}}(X^n,Z^n)-\min_{\hat{X}^n\in\mathcal{D}_n}EL_{\hat{X}^n}(X^n,Z^n)\right]=0$$

where $\mathcal{D}_n$ is the class of all n-block denoisers.

Optimality Result: Semi-Stochastic Setting

Minimum k -*sliding-window loss* of (x^n, z^n) :

$$D_k(x^n, z^n) = \min_{f:\mathcal{A}^{2k+1}\to\mathcal{A}} \left[\frac{1}{n-2k} \sum_{i=k+1}^{n-k} \Lambda(x_i, f(z_{i-k}^{i+k})) \right]$$

Theorem. For all $\mathbf{x} \in \mathcal{A}^\infty$

$$\lim_{n\to\infty} \left[L_{\hat{X}^n_{DUDE}}(x^n, Z^n) - D_{k_n}(x^n, Z^n) \right] \;=\; 0 \quad a.s.$$

Some Further Directions we will Pursue

- Analogue Data
- Performance boosting tweaks for non-asymptotic regime
- Non-stationary data
- Channel Uncertainty
- Channels with Memory
- Sequentiality Constraint
- Applications to data compression and communications
- ⋮

Compression-based denoising

- Intuition and Philosophy
- Tools
  - Lossy compression preliminaries:
    - Rate distortion
    - Rate distortion theory for ergodic processes
    - Indirect rate distortion theory
    - Shannon lower bound
    - Empirical distribution of rate distortion codes
    - Universal lossy source coding:
      - Yang-Kieffer codes
      - Lossy compression via Markov chain Monte Carlo
- Universal denoising via lossy compression

Can DUDE accommodate large, even uncountable, alphabets?

- DUDE will perform poorly when alphabet is large
  - Repeated occurrence of contexts is rare
  - Is problem better viewed in the analogue world ?

- When alphabets are continuous
  Count statistic approach is inapplicable

Extension of “contextless” DUDE to continuous alphabets

Two-pass DUDE-like approach

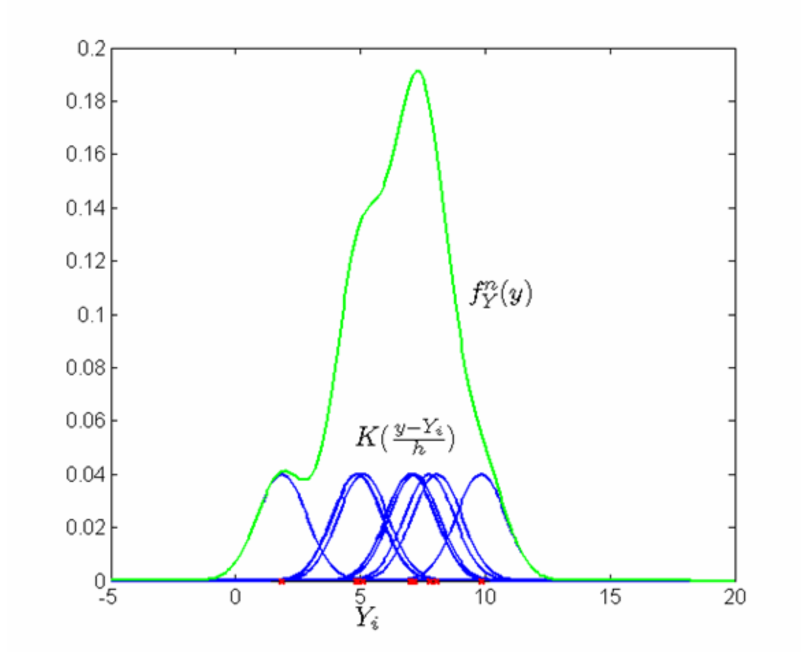
- Density estimation of the noisy symbol distribution
- Estimate empirical distribution of the underlying clean symbol
- Reconstruct to minimize the estimated conditional loss

Estimation of Output Statistics

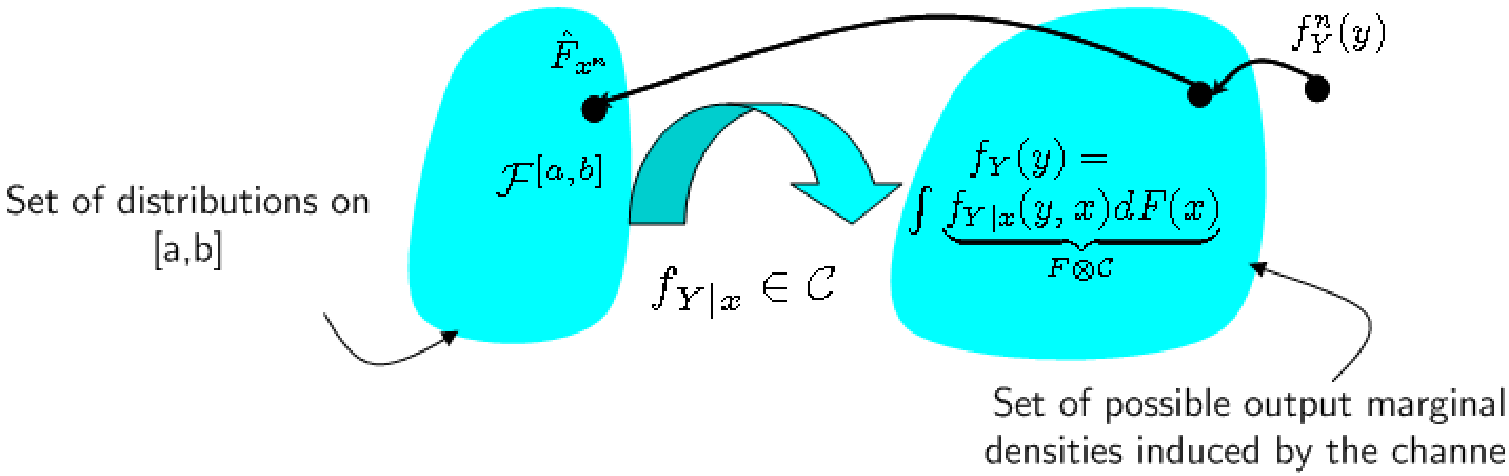
$Y^n = \{Y_1, Y_2, \cdots, Y_n\}$ is the sequence of noisy observations in $\mathbb{R}$

Kernel Density Estimate

$$f_Y^n = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_n}\right) \quad (1)$$



Projection of Channel Output to Input Statistics



$$\hat{F}_{x^n} = \arg\min_{F\in\mathcal{F}[a,b]} d\left(f_Y^n(y), \underbrace{\int f_{Y|x}(y,x)dF(x)}_{[F\otimes\mathcal{C}]_Y}\right)$$

$$d(f,g) = \int |f(y) - g(y)|\, dy \tag{2}$$

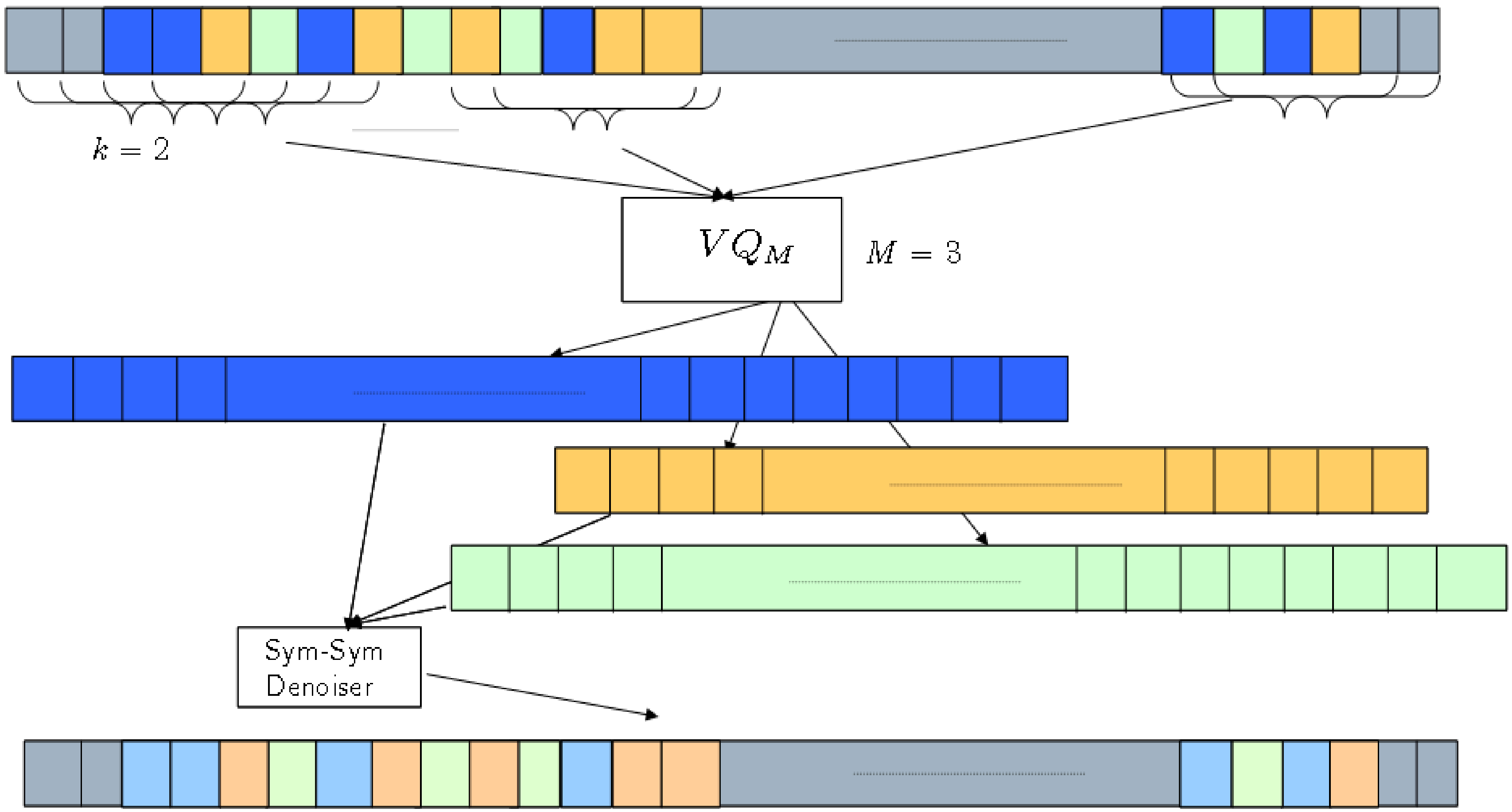
Goodness of Estimation of Clean Signal Statistics

With $\lambda(F, G)$ denoting Levy distance between F and G

Theorem 1. $\lambda\left(F_{x^n}, \hat{F}_{x^n}\right) \longrightarrow 0$ *a.s.* $\forall \mathbf{x}$

DUDE-inspired Denoiser: Quantized Contexts

For context length, k, number of quantization levels, M, and quantizer Q_M



Computational Complexity

- linear in n
- logarithmic in M
- independent of k

Performance Guarantees

Under benign conditions on the channel:

- Can identify the right rate for increase of:
  - Quantization resolution (with an asymptotically fine partition)
  - Context lengths
- Performance guarantees analogous to those of DUDE in both
  - semi-stochastic setting
  - stochastic setting

Another Approach for Analogue World

Via kernel techniques for vector density estimation

Experimental Results

Original Image

Image corrupted by AWGN, $\sigma = 20$

Denoised Images

Ours RMSE= 7.842

Wavelet-based thresholding, RMSE= 11.1782

Multiplicative Noise Example

Corrupted by Multiplicative Noise, $\mathcal{N}(1, 0.2)$

Denoised Images

Denoised Using BLS-GSM

Denoised Using the Proposed Scheme

Back to Discrete World: Performance Boosts

- Dynamic contexts
- Context aggregation (inspired by scheme from analogue world)
- Iterated DUDE

Performance Boost Example I: DUDE with Context Aggregation

Given

- Distance Function: $d(c, \tilde{c})$
- Weight Function: $w(c, \tilde{c})$

Outline of CA DUDE Algorithm:

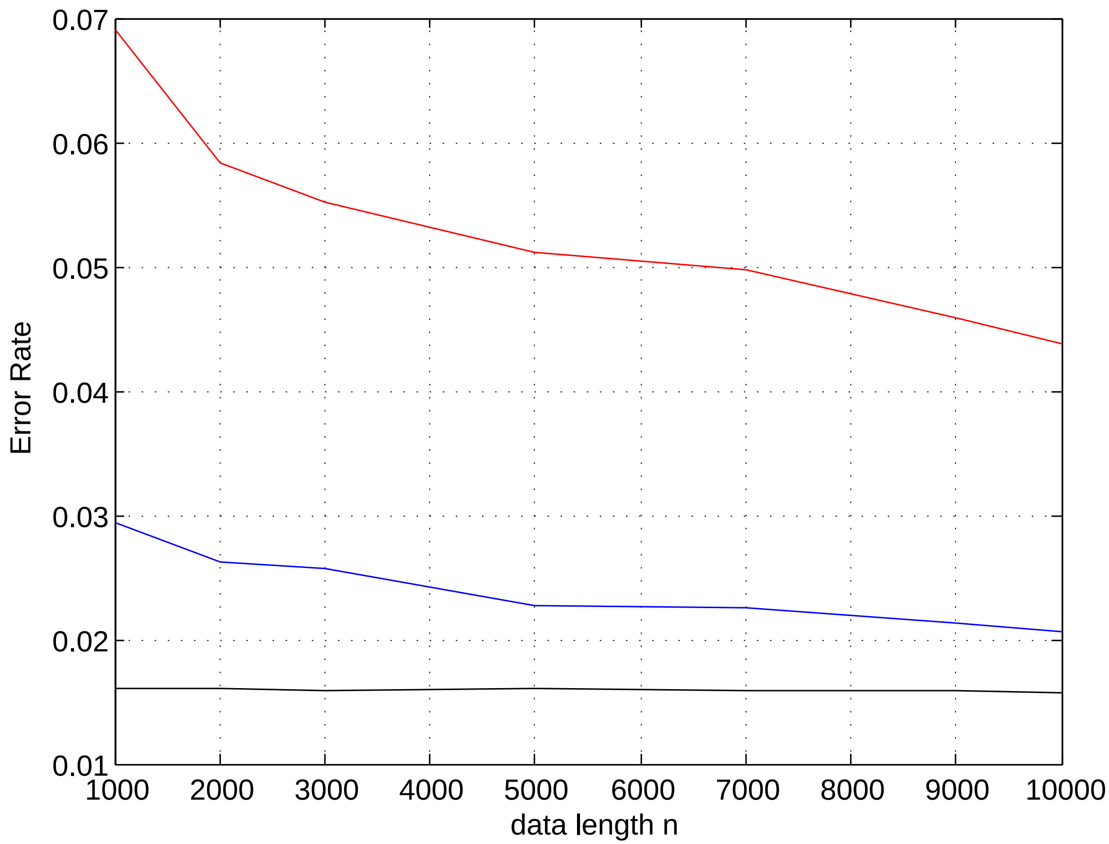
1. Compute count vectors (same as DUDE)
2. Aggregate the counts for similar contexts: for each context c,
   - Step 1: Find $\mathcal{A} = \{\tilde{c} \,|\, d(c, \tilde{c}) \leq D\}$
   - Step 2: Compute new context count, $m_c = \sum_{\tilde{c} \in \mathcal{A}} w(c, \tilde{c}) m(\tilde{c})$
3. Denoising decision made based on new context count, m_c

DUDE with Context Aggregation

- Possible distance and weight functions include:
  - $d(c, \tilde{c}) = P_\pi(\tilde{c}|c)$: Distance based on channel crossover probabilities
  - $w(c, \tilde{c}) = \alpha e^{-\gamma d(c,\tilde{c})}$: Closer contexts contribute higher weights

DUDE with Context Aggregation

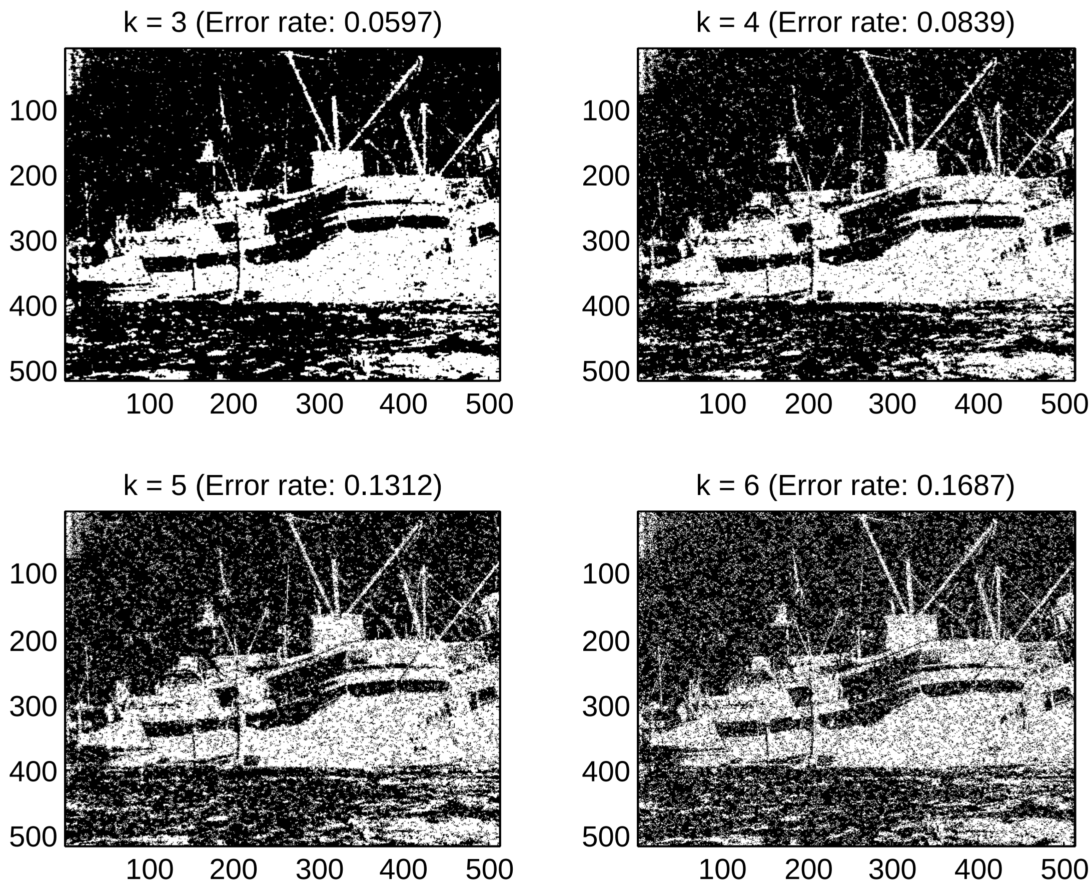
Test Results: Binary Markov Source $(p = 0.01, \delta = 0.2)$.



Blue: CA DUDE, Red: DUDE, Black: Forward-Backward Recursions

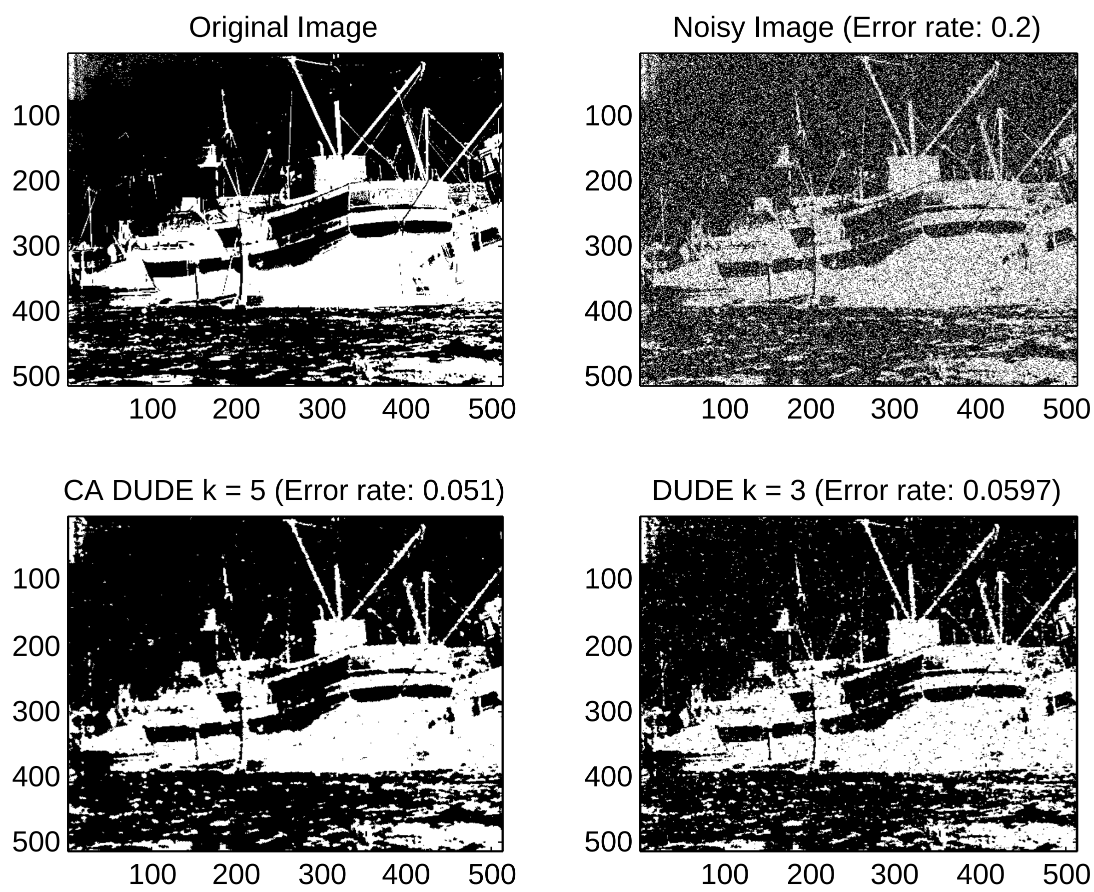
DUDE with Context Aggregation

DUDE: Performance degrades when k is too large



DUDE with Context Aggregation

Test Results: Bi-Level image corrupted with BSC $\delta = 0.2$



Performance Boost Example II: Iterated DUDE

Possible approaches (in increasing order of sophistication):

- Empirically find the transition matrix H from z^n to $\hat{x}^n$ (previous reconstruction), and employ DUDE with $\Pi \cdot H$
  Simplistic but surprisingly effective:

Table 2: Trial 1 for sequence length of 10^3, $\delta = 0.2$, ($k = 5$)

iteration	0	1	2	3	error rate
# of errors left	198	34	26	25	0.025
Forward-Backward					0.019

Table 3: Trial 1 for sequence length of 10^4, $\delta = 0.2$, ($k = 5$)

iteration	0	1	2	3	error rate
# of errors left	2003	213	141	136	0.0136
Forward-Backward					0.0125

Performance Boost Example II: Iterated DUDE (cont.)

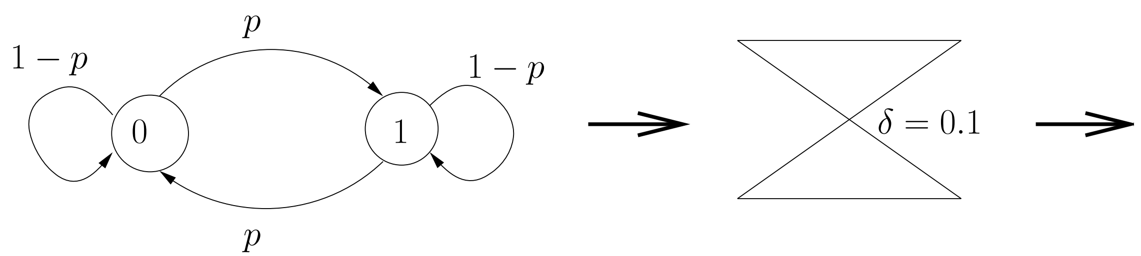
- Compute new effective channel at each iteration, and employ DUDE

- Same as previous approach, taking channel memory into account [see “Channels with Memory” below]

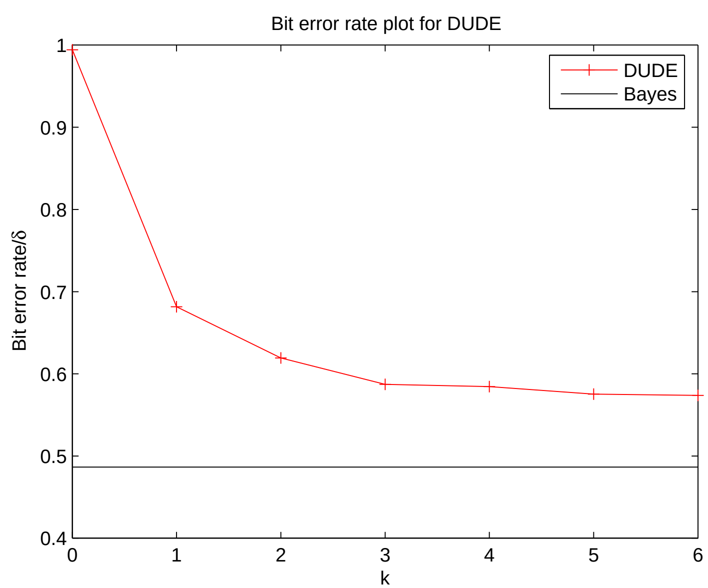
Perf. boost Ex. III: Accommodating Non-Stationarity

Consider following simplistic motivating example:

- “switching” binary symmetric Markov chain corrupted by BSC ($n = 10^6$)



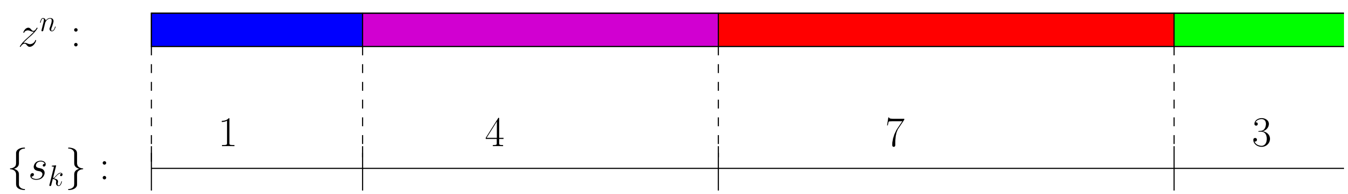
  - suppose $p = p_1 = 0.01 \rightarrow p = p_2 = 0.2$ at $t^* = 5 \times 10^5$ (midpoint)



ShifTing Discrete Universal Denoiser (STUD) - 1D data

- can we learn the switch of the source based only on the noisy observation?
  - if so, can we do it efficiently?

- reference class: class of k-th order denoisers that allow at most m shifts



- $D_{k,m}(x^n, z^n)$: best performance among $\mathcal{S}^n_{k,m}$ ($\leq D_k(x^n, z^n)$)

STUD - Performance Guarantees

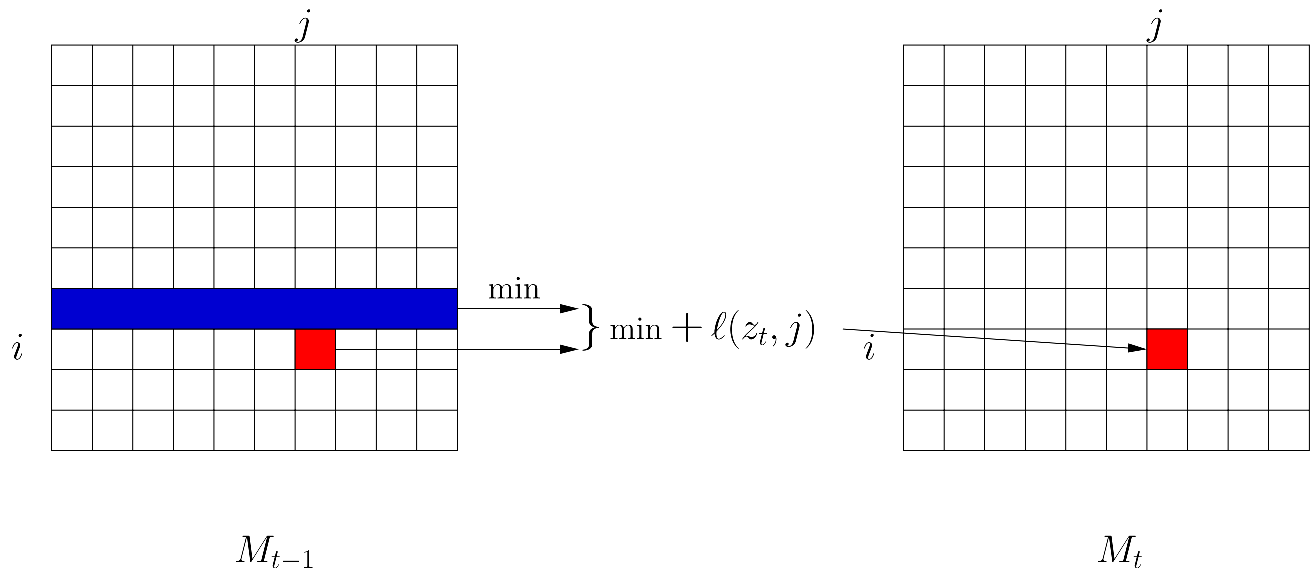
- **direct** (semi-stochastic setting):
  when $m = o(n)$, for all $\mathbf{x}$,
  $$\lim_{n\to\infty} \left[L_{\hat{\mathbf{X}}^{n,k,m}_{\mathrm{STUD}}}(x^n, Z^n) - D_{k,m}(x^n, Z^n) \right] = 0 \quad \textit{a.s.}$$

- **direct** (stochastic setting):
  when $m = o(n)$, achieves optimum performance for any *piecewise stationary* $\mathbf{X}$

- **converse:**
  if $m = \Theta(n)$, no denoiser can achieve above

Two-pass algorithm

- **first pass**: forward recursion - update M_t (dynamic programming)

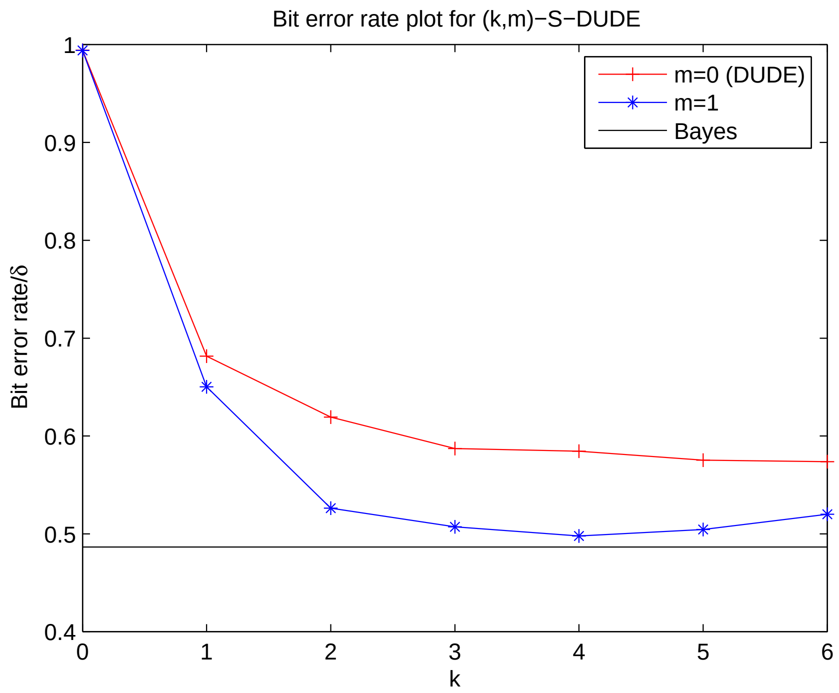


$$M_t(i,j) = \ell(z_t, j) + \min\left\{M_{t-1}(i,j), \min_{1\le k\le|\mathcal{S}|} M_{t-1}(i-1,k)\right\}$$

- **second pass**: backward recursion - extract $\hat{\mathbf{S}}$ and denoise
  - linear complexity in both n and m

Example - 1D data (revisited)

- can STUD achieve the optimal BER ?



- m is another "design parameter" for devising a discrete denoiser

Extension to 2D data

- what about 2D data?
  - we need to learn the best segmentation of data
  - 1D : disjoint intervals $\Leftrightarrow$ 2D : ?

STUD - 2D data

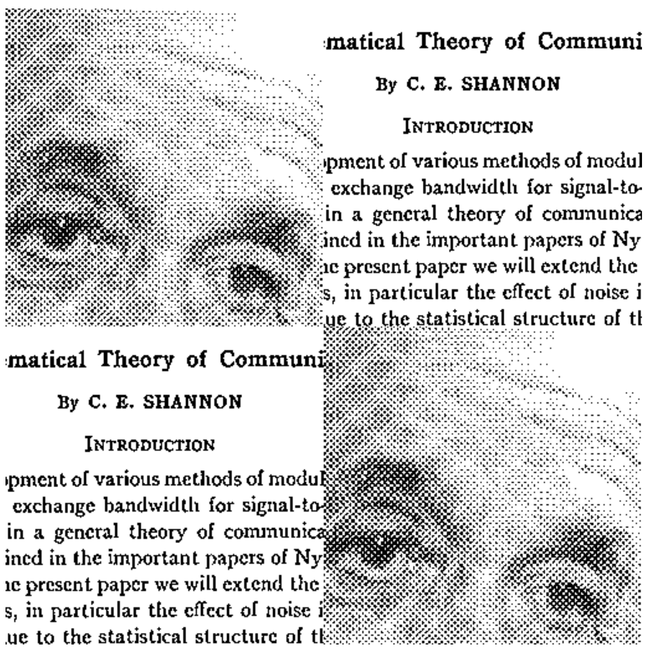
- reference class: class of 2D k-th order denoisers that allow at most m shifts along the "quadtree decomposed" regions
- $D_{k,m}(x^n, z^n)$: best performance among $\mathcal{S}^n_{k,m}$
- $\hat{\mathbf{X}}^{n,k,m}_{\text{2D STUD}}$ defined in similar way as in 1D case
- **guarantee:** when $m \ln m = o(n)$, for all $\mathbf{x} \in \mathcal{X}^\infty$,

$$\lim_{n\to\infty} \left[L_{\hat{\mathbf{X}}^{n,k,m}_{\text{2D STUD}}}(x^n, Z^n) - D_{k,m}(x^n, Z^n) \right] = 0 \quad a.s.$$

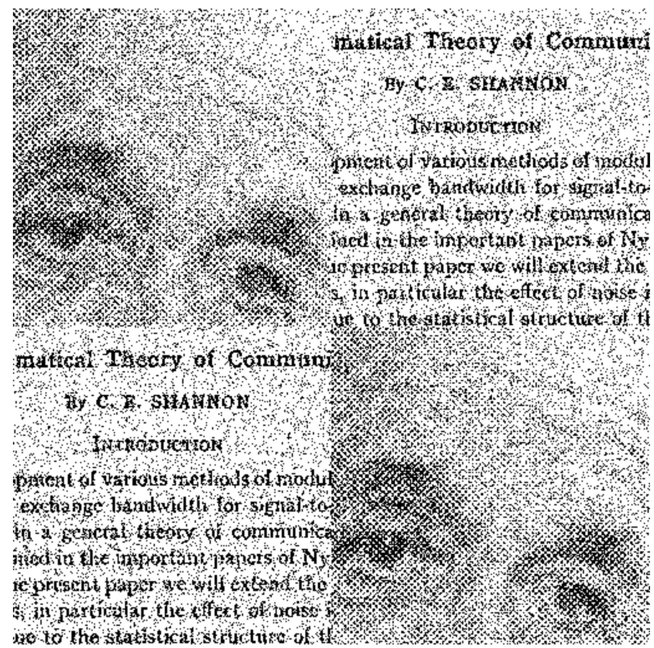
- we have a practical scheme with linear complexity in both n and m

Example - 2D data

- experimental results ($\delta = 0.1$)



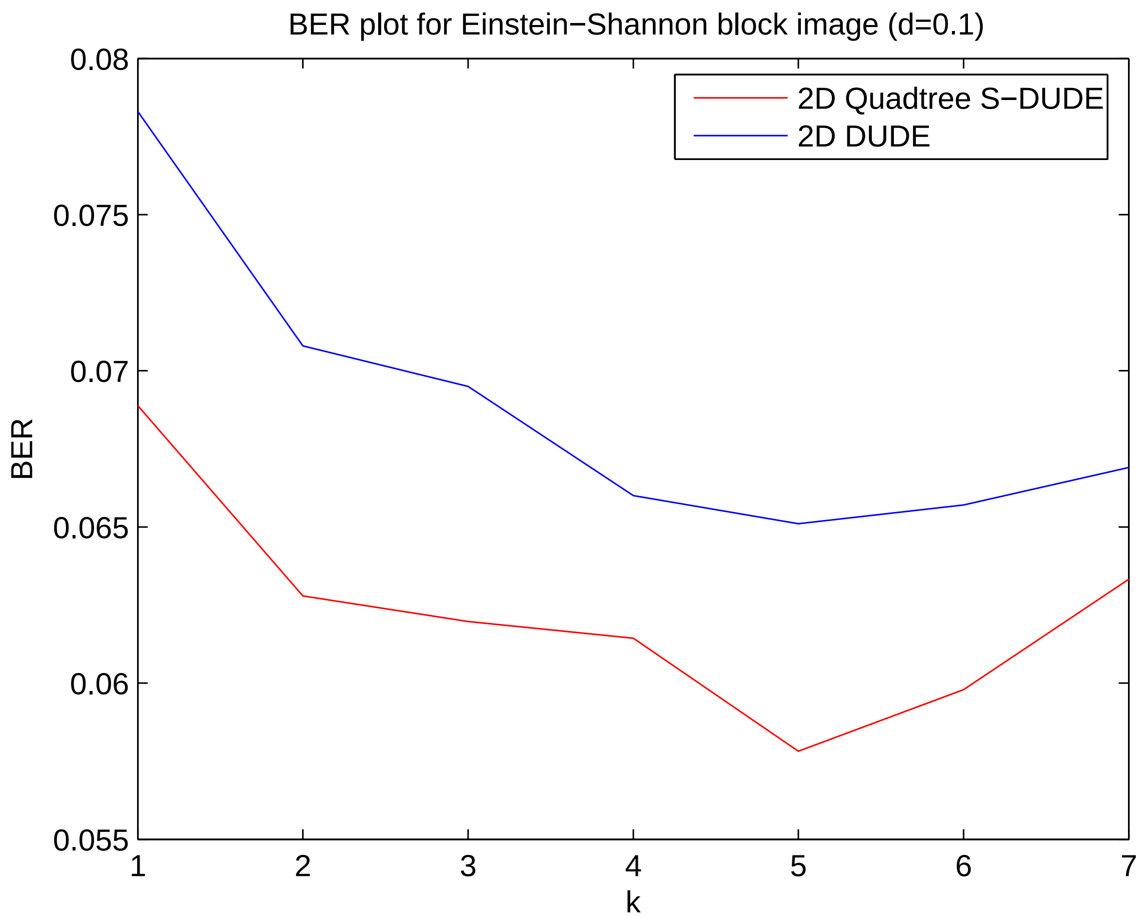
(a) clean image



(b) noisy image

Example - 2D data (cont'd)

- experimental results ($\delta = 0.1$)



Channel Uncertainty

Question: In the case of channel uncertainty is there still hope to find a denoiser with the theoretical performance guarantees of the DUDE?

Channel Uncertainty

Question: In the case of channel uncertainty is there still hope to find a denoiser with the theoretical performance guarantees of the DUDE?

Answer: Unfortunately not

Channel Uncertainty

Approaches that are fruitful in practice:

- DUDE with a “knob”
- DUDE with a channel estimate
- DUDE-like scheme with a channel-independent rule

Channels with Memory

- “Single-letter” nature of the DUDE is lost
- Can devise denoisers with performance guarantees analogous to those of DUDE
- Case of “additive” noise yields a graceful solution

The Sequential LZ-DUDE

- LZ78 incremental parsing: Defined recursively to include shortest phrase not previously parsed: $00000010001110z_t \to 0, 00, 000, 1, 0001, 11, 0z_t$

- At any time t let k_t be the position of z_t in current phrase. Consider subsequence of past data symbols which are the k_t-th symbol in phrases that are identical to the current phrase up through time $t-1$: $0, 00, 000, 1, 0001, 11, 0z_t$

- Reconstruct at time t, $\hat{x}_t$, as the DUDE would, using as counts those of the node (in the LZ tree) corresponding to z_t

The Sequential LZ-DUDE: Performance Guarantees

- Performance guarantees analogous to those of DUDE in:
  - semi-stochastic setting
    - reference class not only of Markov but of *Finite-State* filters
  - stochastic setting
- Fundamental limit different (worse) than for non-sequential case
- Unlike LZ-based predictor, LZ-DUDE does *not* need to randomize

Filtering (causal estimation) $\Leftrightarrow$ Prediction $\Leftrightarrow$ Compression

We will derive, make mathematically precise, and exploit the following relationships:

- Filtering $\Leftrightarrow$ Prediction $\Leftrightarrow$ Lossless compression

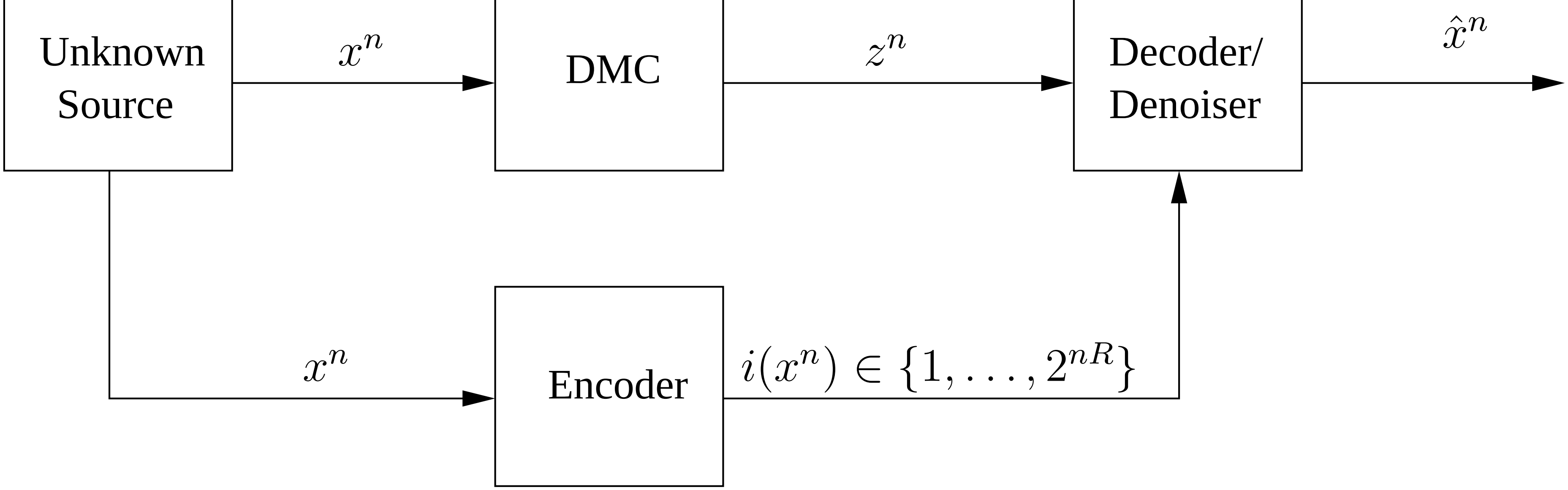
  $\Downarrow$

- Universal compression $\Leftrightarrow$ universal predictor $\Rightarrow$ universal filter

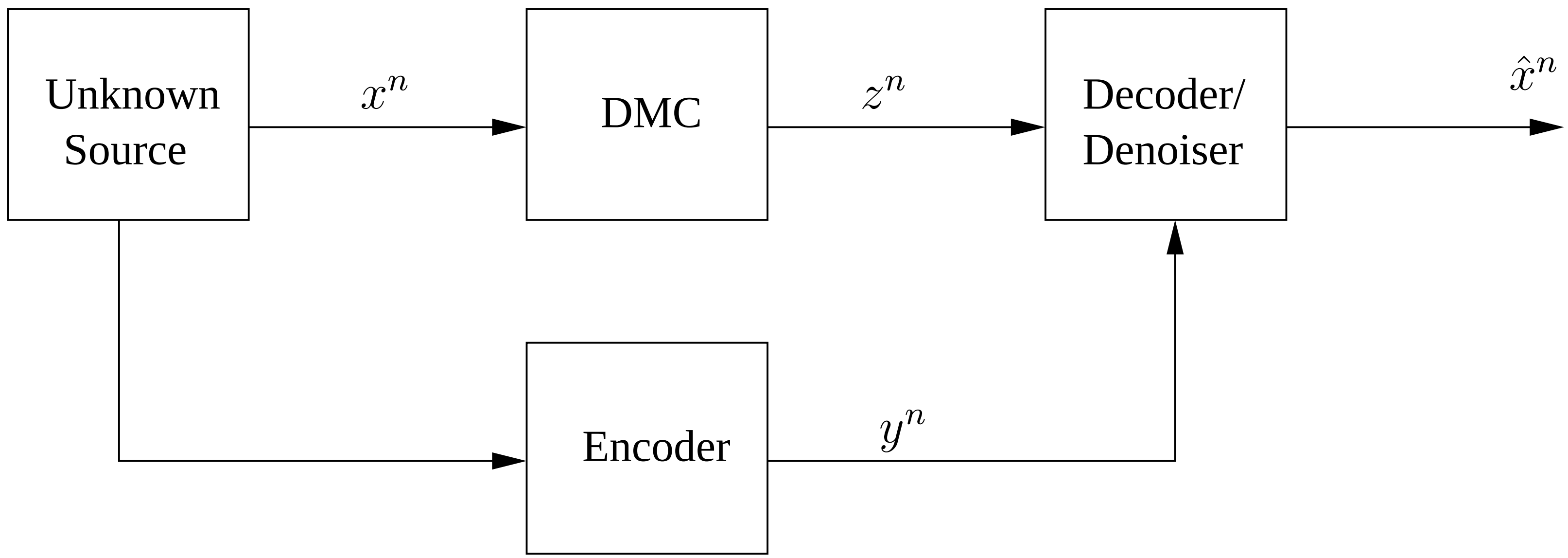
  $\Downarrow$

- LZ compression $\Rightarrow$ LZ predictor $\Rightarrow$ LZ-DUDE

Application Example: Wyner-Ziv Problem



Wyner-Ziv DUDE



- Encoding: among y^n s.t. $LZ(y^n) \leq nR$, describe y^n most conducive to "DUDE with S.I." decoder

- Decoding: "DUDE with S.I.", with y^n as a side information sequence

Wyner-Ziv DUDE: Main Theoretical Result

- For a source $\mathbf{X}$ define:

$$D_{\mathbf{X}}(R) = \inf\{D : (R, D) \text{ is achievable}\}$$

Theorem: For any $R \geq 0$, and any stationary ergodic source $\mathbf{X}$,

$$\lim_{n\to\infty} E[\text{distortion}\,(X^n, \text{Reconstruction using Wyner-Ziv DUDE})] = D_{\mathbf{X}}(R)$$

Example: Binary Image + WZ-DUDE

Original

BSC(0.15)-corrupted version

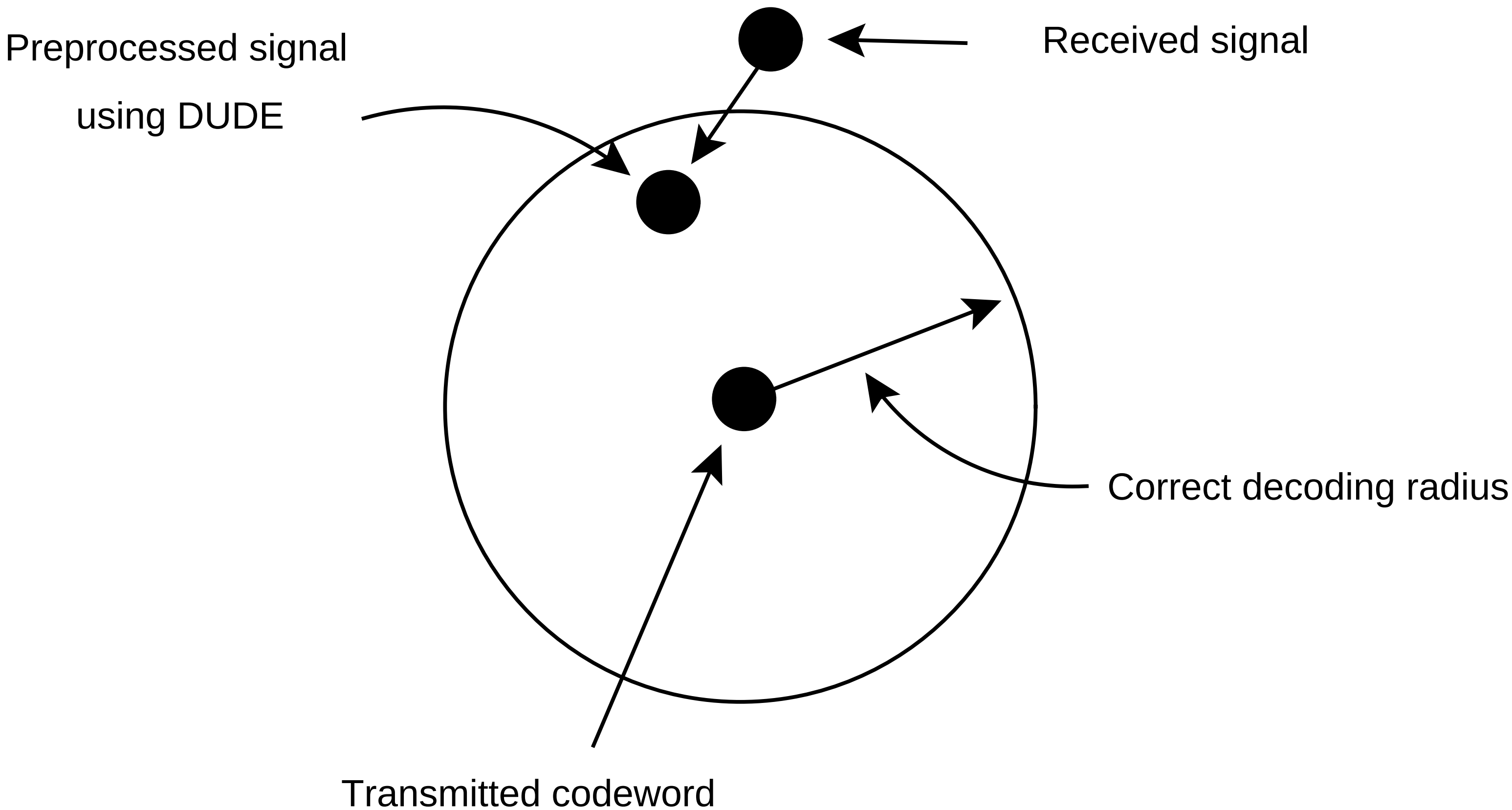
Example: Binary Image + WZ-DUDE (cont.)

`Left`: Lossy JPEG coding of original image: $R = 0.22$ b.p.p., $BER = 0.0556$

`Center`: DUDE output: $BER = 0.0635$

`Right`: WZ-DUDE output: $R = 0.22$ b.p.p., $BER = 0.0407$

DUDE for Error Correction



So why take this course ?

- Intellectual + practical value of the specific problems considered
- An excuse to learn other topics in information theory
- Opportunity to acquire some tools and see how they are applied
- Learn IT approach to universality

Excuse to learn other topics in IT

Beyond our “target” topics, we will pick up:

- State estimation in HMPs and the Forward-Backward scheme
- R-D theory for ergodic sources
- Shannon Lower Bound
- Empirical distribution of good codes
- Indirect R-D
- Ziv-Lempel compression
- Universal prediction
- Compound sequential decision problem
- R-D with decoder side information (Wyner-Ziv problem)
- Systematic channel coding

Opportunity to learn some tools and how they are applied

- Martingales
- Concentration Inequalities
- Dynamic Programming
- Markov Chain Monte Carlo
- Density Estimation Techniques

Learn IT approach to Universality

Typical IT way of viewing problems:

- Characterization of fundamental limits
- Existence of universal schemes ?
- Universality
  - Stochastic setting
  - Individual sequence setting
- Low complexity, practicality, cuteness and grace of schemes

We'll see this structure for denoising, lossy compression, lossless compression, prediction, filtering, Wyner-Ziv coding, . . .

Can then apply to your own problems