Applications of Information Theory in Image Processing

Image Denoising

Gadiel Seroussi

March 21, 2023

1. DUDE: Discrete Universal DEnoising

Basic algorithm

References

Basic DUDE

T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M.J. Weinberger, “Universal discrete denoising: known channel,” *IEEE Transactions on Information Theory*, **51**, No. 1, pp. 5–28, January 2005.

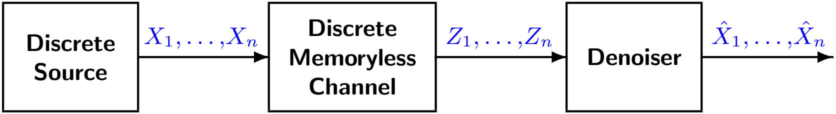
DUDE for B/W images

E. Ordentlich, G. Seroussi, S. Verdú, M. Weinberger, T. Weissman, “A discrete universal denoiser and its application to binary images,” *IEEE International Conference on Image Processing (ICIP 2005)*, Barcelona, Spain, September 2003.

DUDE for grayscale images

G. Motta, E. Ordentlich, I. Ramírez, G. Seroussi, and M. Weinberger, “The DUDE framework for continuous tone image denoising,” *IEEE Transactions on Image Processing*, **20**, No. 1, pp. 1–21, January 2011.

Discrete denoising



- X_i, Z_i, $\hat{X}_i$ take values from *finite alphabets*.
- *Goal:* Choose $\hat{X}_1, \ldots, \hat{X}_n$ on the basis of $Z_1, \ldots, Z_n$ to minimize a *fidelity criterion* (some notion of *distortion* of $\hat{X}_1, \ldots, \hat{X}_n$ relative to $X_1, \ldots, X_n$, *which we may not be able to measure!*).

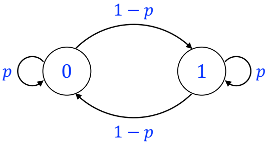
Applications

- Image Denoising
- Text Correction
- Reception of Uncoded Data
- Hidden Markov Model State Estimation
- . . . . . .

Example (non-universal)

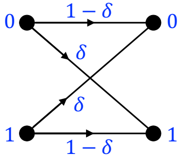
Source:

Binary Markov Chain



Channel:

BSC



$\ldots 0001111100001111100 \ldots$

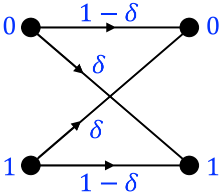
$\ldots 0001000001000001010 \ldots \Rightarrow \ldots 0000111101001110110 \ldots$

↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

- Objective: Minimize *bit error rate* given the observation of an n-block, knowing the parameters p, δ.
- Solution: Backward-Forward Dynamic Programming.
- Fundamental Limit: $\lim_{n\to\infty}(\text{Min Error Probability}) = f(\delta, p)$ still open.

Example (universal)

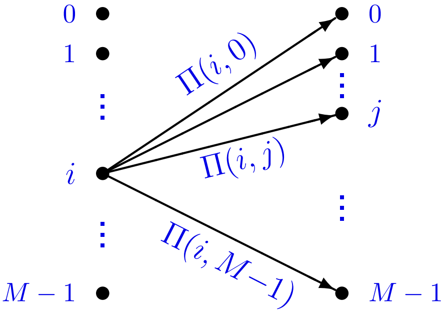
- Source: Binary, *nothing known about the distribution*
- Channel: BSC



- Objective: Minimize *bit error rate* given the observation of an n-block, knowing the parameter δ.
- Solution: ???
- Fundamental Limit: $\lim_{n\to\infty}$(Min Error Probability) = ???

Universal Setting: Basic Assumptions

- Unknown source distribution.
- Discrete Memoryless Channel (DMC) over alphabet $\mathcal{A} = \{0, 1, \ldots, M-1\}$, with a *known transition probability matrix*.



$\Pi(i,j) =$ Prob(j received | i sent)

Previous Approaches to Universal Discrete Denoising

- Occam filter [Natarajan 1993, 1995]
    - Lossy compression of the noisy signal, tuning the desired SNR to the expected noise level of the channel.
    - Experiments with specific lossy data compression algorithms
    - Shortcoming: No implementable universal optimal lossy compression is known.

Previous Approaches to Universal Discrete Denoising

- Kolmogorov Sampler [(Donoho 2002]
  - For all typical noise realizations, list the corresponding source realizations that explain the data. Then, do lossless compression of the source realizations and select the shortest one.
  - Shortcoming: Not implementable.
  - It does not attain the theoretically optimum distribution-dependent performance. *(It can be off by a factor of 2.)*
  - Example: Bernoulli(p) source corrupted by a BSC(δ). Trivial schemes "say what you see" (optimal for $p \geq \delta$) and "say all zeros" (optimal for $p \leq \delta$) each outperform the Kolmogorov sampler on more than half of the parameter space.
  - Does this mean it is suboptimal in the universal setting? Is the distribution-dependent performance attainable at all in this setting?

Notation and Additional Assumptions

- Alphabet: *Same finite alphabet* $\mathcal{A}$ for clean and noisy signals; w.l.o.g., $\mathcal{A} = \{0, 1, 2, \ldots, M-1\} \quad (|\mathcal{A}| = M)$.
- Channel: *Nonsingular* transition probability matrix:

$$\boldsymbol{\Pi} = \{\Pi(i,j)\}_{i,j\in\mathcal{A}} = [\underset{\uparrow}{\boldsymbol{\pi}_0} \mid \underset{\cdots}{\cdots} \mid \underset{\uparrow}{\boldsymbol{\pi}_{M-1}}]$$

columns

both assumptions above can be relaxed

- Sequence notation:
  - $x_i^j = x_i,\, x_{i+1},\, \ldots,\, x_j$
  - $x^n = x_1^n = x_1,\, x_2,\, \ldots,\, x_n$

  applies to symbols (e.g. x^n) or random variables (e.g. X^n)

Notation and Additional Assumptions (cont.)

- n-block denoiser: $\hat{X}^n : \mathcal{A}^n \to \mathcal{A}^n$
  on input z^n, returns output $\hat{x}^n$; *no sequentiality assumed*.
- Loss (or Cost) Function: $\Lambda : \mathcal{A}^2 \to [0, \infty)$, represented by the matrix

$$\boldsymbol{\Lambda} = \{\Lambda(i,j)\}_{i,j\in\mathcal{A}} = \left[\boldsymbol{\lambda}_0 \mid \cdots \mid \boldsymbol{\lambda}_{M-1}\right].$$

cost of guessing j when clean signal is i

Examples. Hamming loss: $\Lambda(i,j) = 1$ if $i \neq j,\ 0$ otherwise.
Quadratic (Euclidean) loss: $\Lambda(i,j) = (i-j)^2$.

- Normalized cumulative loss of the denoiser $\hat{X}^n$ when the observed sequence is $z^n \in \mathcal{A}^n$ and the channel input is $x^n \in \mathcal{A}^n$:

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

Denoiser output for i-th coordinate

Performance Benchmark

Optimum performance for a denoiser *when the input distribution is known*:

$$\lim_{n\to\infty} \min_{\hat{X}^n \in \mathcal{D}_n} EL_{\hat{X}^n}(X^n, Z^n)$$

where $\mathcal{D}_n$ is the class of all n-block denoisers, and expectation is with respect to the input distribution and the channel.

Notes

- Since $\mathcal{A}$ is finite, so is $\mathcal{D}_n$ for a given n.
- It would take a pretty powerful genie to compute this benchmark!
- Nevertheless, it is well defined (maybe with mild assumptions on the distributions).

The DUDE Algorithm: General Idea

- Fix *context length* k. For each symbol 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	●	r_1	r_2	$\cdots$	r_k

  - Count all occurrences of symbols with left k-context $(\ell_1, \ldots, \ell_k)$ and right k-context $(r_1, \ldots, r_k)$. This gives a *conditional empirical distribution* of the *noisy* symbol given the *noisy* contexts $(\ell_1, \ldots, \ell_k)$ and $(r_1, \ldots, r_k)$.
  - Use channel transition probability to estimate the conditional empirical distribution of the *clean* symbol given the *noisy* contexts $(\ell_1, \ldots, \ell_k)$ and $(r_1, \ldots, r_k)$.
  - Make decision on reconstructed symbol using
    - the loss function,
    - the channel transition probability,
    - the conditional empirical distribution
    - the observed symbol to be denoised.

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 $\cdots$

Noisy text

Wz right peace the restiction 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 ofthe 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 hndependest 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 tke theorem, 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 hndependest 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 tke theorem, 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 $k = 1$ e●r

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 hndependest 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 tke theorem, 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 $k = 1$ e●r counts

- e r : 8
- eor : 6
- eir : 2
- emr : 1
- eqr : 1

Context search $k = 2$ he•re

Wz right peace the rest iction on alksoable sequbole thgt wo spices fokiow eadh otxer. ··· egfbct of sraaistfcal keowleuge apolt tje souwce in recucilg the requihed clpagity ofythe clabbel ··· the relatrte pweqiency ofpthe digram i j. The setter freqbwncles $p(i)$, ghe rrahsibion probtbilities ··· The resemglahca to ordwnard Engdish tzxt ircreakes quitq noliceabcy at vach oftthe hbove steps. ··· Thus theorev, andlthe aszumptjona requiyed ffr its croof, arv il no wsy necqssrry forptfe prwwent theorz. ··· jhe reap juptifocation of dhese defikjtmons, doweyer, bill rehide inytheir imjlycajijes. ··· 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 $p_1, \ldots, p_n$. ··· The theorem sahs tyat fsr lawge N mhis gill we hndependest of q aed vqunl tj H. ··· 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. ··· The Fundkmendal Theorem kor a Soiselesd Chjnnen ··· Lhe ronvegse jaht jf tke theorem, thlt $\frac{C}{H}$ calnot be excweded, may ke xroved ey hotijg tyat the enyropy ··· The first pajt if the theqrem will be ptoved in two kifferent wjys. ··· Another methjd of plrfolming shis goding ald thmreby proking toe oheorem can bexdescrined as folfows: ··· The contemt ov The rem 9 if thst, ajthorgh an ezacr mawwh is ··· Wotf a goul code therlogaretym of the rehitrocpl prossbilfly of a lylg mwgsage lust be priporyiopal to tha rurafirn of ···

Context search $k = 2$ he●re counts

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

Notation for Context Counts

- $\mathbf{a} = a_1^n \in \mathcal{A}^n$, whole data sequence
- $\mathbf{b} \in \mathcal{A}^k$, left k-context string
- $\mathbf{c} \in \mathcal{A}^k$, right k-context string
- $\alpha \in \mathcal{A}$, arbitrary symbol, $0 \le \alpha \le M-1$
- $\mathbf{m}[\mathbf{a},\mathbf{b},\mathbf{c}] \triangleq M$-vector (column) with α-th component equal to the number of occurrences of the pattern $\boxed{\mathbf{b}}\boxed{\alpha}\boxed{\mathbf{c}}$ in $\mathbf{a}$:

$$\mathbf{m}[\mathbf{a},\mathbf{b},\mathbf{c}]_\alpha = \left|\left\{k+1 \le i \le n-k : a_{i-k}^{i-1} = \mathbf{b}, a_i = \alpha, a_{i+1}^{i+k} = \mathbf{c}\right\}\right|.$$

- Example:

$$\begin{array}{lcl} \mathbf{m}[\texttt{Shannon text}, he, re] & = & [0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,5\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,7]^T \\ & & \uparrow\uparrow\uparrow\uparrow\uparrow \\ & & imoq\texttt{sp} \end{array}$$

The Discrete Universal Denoiser (DUDE)

Fix k. Initialize $\mathbf{m}[z^n, \mathbf{b}, \mathbf{c}] = \mathbf{0}$ for all $\mathbf{b}, \mathbf{c} \in \mathcal{A}^k$.

- Pass 1 For every $k+1 \leq i \leq n-k$: increment the count of z_i in

$$\mathbf{m}[z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k}]$$

(build the count vectors $\mathbf{m}[z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k}]$).

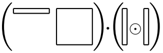
- Pass 2 Reconstruct according to:

Denoiser output for ith coordinate

$$\hat{X}_i^{n,k}(z^n) = g_{z^n}^k\left(z_{i-k}^{i-1}, z_i, z_{i+1}^{i+k}\right), \quad k+1 \leq i \leq n-k.$$

where

$$g_{\mathbf{a}}^k(\mathbf{b}, \alpha, \mathbf{c}) = \arg\min_{\hat{x} \in \mathcal{A}} \left(\mathbf{m}^T[\mathbf{a}, \mathbf{b}, \mathbf{c}]\, \mathbf{\Pi}^{-1}\right) \cdot (\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_\alpha)$$

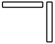


$$\begin{aligned} \mathbf{\Pi} &= \{\Pi(i,j)\}_{i,j \in \mathcal{A}} = [\boldsymbol{\pi}_0 \,|\, \cdots \,|\, \boldsymbol{\pi}_{M-1}] \\ \mathbf{\Lambda} &= \{\Lambda(i,j)\}_{i,j \in \mathcal{A}} = [\boldsymbol{\lambda}_0 \,|\, \cdots \,|\, \boldsymbol{\lambda}_{M-1}]\,. \\ (\mathbf{v} \odot \mathbf{w})_i &= v_i w_i \qquad \text{Schur product} \end{aligned}$$

DUDE Intuition

- Say you want to guess the value of a random variable $Y \sim P_Y$ over $\mathcal{A}$. Guessing $Y = \beta \in \mathcal{A}$ incurs an *expected* loss

$$\sum_{a \in \mathcal{A}} P_Y(a)\Lambda(a, \beta) = \mathbf{P}_Y^T \cdot \boldsymbol{\lambda}_\beta \,,$$



β-th column of cost matrix $\boldsymbol{\Lambda}$

where $\mathbf{P}_Y^T = [P_Y(0)\, P_Y(1) \,\ldots\, P_Y(M-1)]$.

- Expected loss is minimized with the estimate

$$\hat{Y} = \arg\min_{\beta \in \mathcal{A}} \ \mathbf{P}_Y^T \cdot \boldsymbol{\lambda}_\beta$$

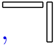
- We will apply this *MAP rule* to guess the value of a *clean* symbol x with an *estimate* of its distribution P_X (a guess based on another guess)

DUDE Intuition (cont.)

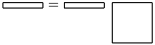
- Given an input r.v. $X \sim P_X$ going through the channel $\mathbf{\Pi}$, the output $Z \sim P_Z$ satisfies

a-th column of channel matrix $\mathbf{\Pi}$

$$P_Z(a) = \sum_{b=0}^{M-1} P_X(b)\Pi(b,a) = \mathbf{P}_X^T \cdot \boldsymbol{\pi}_a\,, \quad a = 0,1,2,\ldots,M-1$$



$$\Rightarrow \quad \mathbf{P}_Z^T = \mathbf{P}_X^T\,\mathbf{\Pi} \quad \Rightarrow \quad \mathbf{P}_X^T = \mathbf{P}_Z^T\,\mathbf{\Pi}^{-1}$$



- Let $\mathbf{b}_i{=}z_{i-k}^{i-1}$, $\mathbf{c}_i{=}z_{i+1}^{i+k}$. We take $\mathbf{m}[z^n,\mathbf{b}_i,\mathbf{c}_i]$ as an (unnormalized) estimate of $\mathbf{P}_Z(\,z_i \,|\, \mathbf{b}_i \bullet \mathbf{c}_i\,)$
- Then, we take $\mathbf{\Pi}^{-T}\mathbf{m}[z^n,\mathbf{b}_i,\mathbf{c}_i]$ as an estimate of

Notation: $\mathbf{\Pi}^{-T}{=}(\mathbf{\Pi}^{-1})^T$

$$\hat{\mathbf{P}}_X(\,x_i \,|\, \underbrace{\mathbf{b}_i \bullet \mathbf{c}_i}_{\text{noisy context}}\,)\,,$$

and $\left(\mathbf{\Pi}^{-T}\mathbf{m}[z^n,\mathbf{b}_i,\mathbf{c}_i]\right) \odot \boldsymbol{\pi}_\alpha$ as an estimate of

$$\hat{\mathbf{P}}_X(\,x_i \,|\, \underbrace{\mathbf{b}_i \bullet \mathbf{c}_i}_{\text{noisy context}}, \underbrace{z_i = \alpha}_{\text{noisy symbol}})\,,$$

incorporates ‘context’ and ‘what we see’ information about x_i; *relies on the channel being memoryless and independent of input*

DUDE Intuition - justification for last step of rule

Why if $\mathbf{\Pi}^{-T}\mathbf{m}[z^n, \mathbf{b}_i, \mathbf{c}_i]$ is an estimate of $\hat{P}_X(\cdot|\mathbf{b}_i \bullet \mathbf{c}_i)$, we can take $\left(\mathbf{\Pi}^{-T}\mathbf{m}[z^n, \mathbf{b}_i, \mathbf{c}_i]\right) \odot \boldsymbol{\pi}_\alpha$ as an estimate of $\hat{P}_X(\cdot|\mathbf{b}_i \bullet \mathbf{c}_i, z_i = \alpha)$.
All estimates unnormalized.
We have

$$\begin{aligned}
P(X = x_i|\mathbf{b}_i \bullet \mathbf{c}_i, z_i = \alpha) &= \frac{P(X = x_i, \mathbf{b}_i \bullet \mathbf{c}_i, z_i = \alpha)}{P(\mathbf{b}_i \bullet \mathbf{c}_i, z_i = \alpha)} \\
&= \frac{P(X = x_i, \mathbf{b}_i \bullet \mathbf{c}_i, z_i = \alpha)}{P(X = x_i, \mathbf{b}_i \bullet \mathbf{c}_i)} \frac{P(X = x_i, \mathbf{b}_i \bullet \mathbf{c}_i)}{P(\mathbf{b}_i \bullet \mathbf{c}_i)} \frac{P(\mathbf{b}_i \bullet \mathbf{c}_i)}{P(\mathbf{b}_i \bullet \mathbf{c}_i, z_i = \alpha)} \\
&= \frac{1}{P(z_i = \alpha|\mathbf{b}_i \bullet \mathbf{c}_i)} P(z_i = \alpha|X = x_i, \mathbf{b}_i \bullet \mathbf{c}_i) P(X = x_i|\mathbf{b}_i \bullet \mathbf{c}_i) \\
&= \frac{1}{P(z_i = \alpha|\mathbf{b}_i \bullet \mathbf{c}_i)} \Pi(x_i, \alpha) P(X = x_i|\mathbf{b}_i \bullet \mathbf{c}_i)
\end{aligned}$$

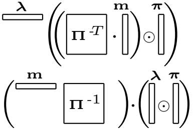
- Given $X = x_i$, z_i is independent of $\mathbf{b}_i \bullet \mathbf{c}_i$, so $P(z_i = \alpha|X = x_i, \mathbf{b}_i \bullet \mathbf{c}_i) = \Pi(x_i, \alpha)$.
- $P(z_i = \alpha|\mathbf{b}_i \bullet \mathbf{c}_i)$ is a constant in the $\arg\max$ iteration over x_i.

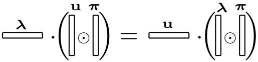
DUDE Intuition (cont.)

- Now, we use the MAP rule to guess the clean symbol x_i as

$$\arg\min_{\hat{x}\in\mathcal{A}} \boldsymbol{\lambda}_{\hat{x}}^T \cdot \Big(\big(\boldsymbol{\Pi}^{-T}\,\mathbf{m}[z^n,\mathbf{b}_i,\mathbf{c}_i]\big) \odot \boldsymbol{\pi}_\alpha \Big)$$

$$= \arg\min_{\hat{x}\in\mathcal{A}} \big(\mathbf{m}^T[z^n,\mathbf{b}_i,\mathbf{c}_i]\,\boldsymbol{\Pi}^{-1}\big) \cdot (\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_\alpha)$$



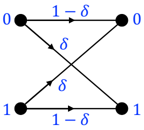


$$\boxed{\begin{aligned}(\boldsymbol{\pi}^T\odot\mathbf{u}^T)\cdot\boldsymbol{\lambda} &= \textstyle\sum_i(\pi_i u_i)\lambda_i\\ &= \textstyle\sum_i u_i(\lambda_i\pi_i)\\ &= \mathbf{u}^T\cdot(\boldsymbol{\lambda}\odot\boldsymbol{\pi})\end{aligned}}$$

- Issues:
    - lots of estimates and ‘estimates on estimates’
    - contexts are noisy, things that are in the same context in the clean data may be in different contexts in the noisy data (aside from being noisy themselves)
    - will all this work?

Example: Binary Symmetric Channel (BSC)

- BSC(δ), cost measured by Hamming distance (bit errors)



$$\mathbf{\Pi} = \begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix}, \quad \mathbf{\Lambda} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

$$\mathbf{\Pi}^{-1} = \begin{pmatrix} 1-\delta & -\delta \\ -\delta & 1-\delta \end{pmatrix} \frac{1}{1-2\delta}\,.$$

DUDE rule: $\arg\min_{\hat{x}\in\mathcal{A}} \left(\mathbf{m}^T[z^n,\mathbf{b}_i,\mathbf{c}_i]\,\mathbf{\Pi}^{-1}\right)\cdot(\boldsymbol{\lambda}_{\hat{x}}\odot\boldsymbol{\pi}_{\alpha})$

- Say we have $\mathbf{m}[z^n,\mathbf{b}_i,\mathbf{c}_i] = (n_0,\ n_1)^T$, and $z_i = 0$.

$$\mathbf{u} \stackrel{\Delta}{=} \mathbf{m}^T\mathbf{\Pi}^{-1} = \left[\, n_0(1-\delta) - n_1\delta,\ \ n_1(1-\delta) - n_0\delta \,\right]$$

$$\hat{x} = 0 \quad : \quad C_0 \stackrel{\Delta}{=} \mathbf{u}\cdot(\boldsymbol{\lambda}_0\odot\boldsymbol{\pi}_0) = \delta\,(n_1(1-\delta) - n_0\delta)$$

$$\hat{x} = 1 \quad : \quad C_1 \stackrel{\Delta}{=} \mathbf{u}\cdot(\boldsymbol{\lambda}_1\odot\boldsymbol{\pi}_0) = (1-\delta)\,(n_0(1-\delta) - n_1\delta)$$

With $z_i = 0$, we choose $\hat{x}_i = 0$ iff $C_0 \le C_1$, or

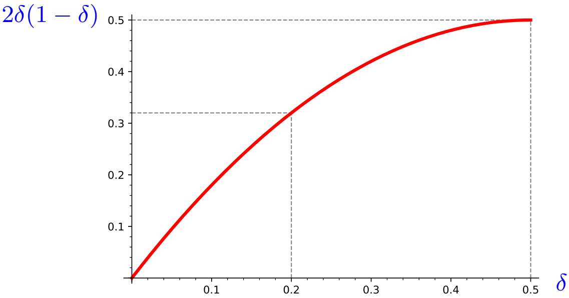
$$\frac{n_0}{n_1} \ge \frac{2\delta(1-\delta)}{(1-\delta)^2+\delta^2} \quad \overset{n=n_0+n_1}{\Longleftrightarrow} \quad \frac{n_0}{n} \ge 2\delta(1-\delta)$$

Example: BSC (cont.)

- In general, if $z_i = b$, we leave z_i alone if

$$\frac{n_b}{n} \geq 2\delta(1-\delta)\,.$$

Otherwise, we flip z_i.



Example: Salt and Pepper (S&P) Channel, Euclidean Cost

$\mathcal{A}=\{0,1,\ldots,M-1\}$. Sample goes to 0 or $M-1$ (uniformly) with probability δ, or stays intact with prob. $(1-\delta)$, $0\le\delta<1$.

$$\mathbf{\Pi}=\begin{bmatrix} 1-\frac{\delta}{2} & 0 & \cdots & 0 & \frac{\delta}{2} \\ \frac{\delta}{2} & 1-\delta & \cdots & 0 & \frac{\delta}{2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\delta}{2} & 0 & \cdots & 1-\delta & \frac{\delta}{2} \\ \frac{\delta}{2} & 0 & \cdots & 0 & 1-\frac{\delta}{2} \end{bmatrix} \quad \mathbf{\Pi}^{-1}=\frac{1}{1-\delta}\begin{bmatrix} 1-\frac{\delta}{2} & 0 & \cdots & 0 & -\frac{\delta}{2} \\ -\frac{\delta}{2} & 1 & \cdots & 0 & \vdots \\ \vdots & \vdots & \ddots & \vdots & -\frac{\delta}{2} \\ -\frac{\delta}{2} & 0 & \cdots & 1 & -\frac{\delta}{2} \\ -\frac{\delta}{2} & 0 & \cdots & 0 & 1-\frac{\delta}{2} \end{bmatrix}.$$

$$\boldsymbol{\lambda}_x=\left[x^2,\,(x-1)^2,\,(x-2)^2,\,\ldots,\,(x-M+1)^2\right]^T \quad (\boldsymbol{\lambda}_{xx}=0)$$

Say $\mathbf{m}=[n_0,n_1,\ldots,n_{M-1}]^T$. If $z_i=a\notin\{0,M-1\}$, we have

$$\left(\mathbf{\Pi}^{-T}\,\mathbf{m}\right)\odot\boldsymbol{\pi}_a=[0,\ldots,0,u_a,0,\ldots,0]^T,\quad u_a>0$$

$$\Rightarrow \boldsymbol{\lambda}_x^T\cdot\left((\mathbf{\Pi}^{-T}\,\mathbf{m})\odot\boldsymbol{\pi}_a\right)=0 \text{ if } x=a, \text{ otherwise positive}$$

$\Rightarrow$ rule is: choose $\hat{x}_i=z_i$ if $z_i\notin\{0,M-1\}$

Salt and Pepper (S&P) Channel, Euclidean Cost (cont.)

When $z_i = 0$

$$(\boldsymbol{\Pi}^{-T}\mathbf{m})\odot\boldsymbol{\pi}_0 = \tfrac{1}{1-\delta}\begin{bmatrix} n_0 - n\delta/2 \\ n_1 \\ \vdots \\ n_{M-2} \\ n_{M-1} - n\delta/2 \end{bmatrix} \odot \begin{bmatrix} 1-\delta/2 \\ \delta/2 \\ \vdots \\ \delta/2 \\ \delta/2 \end{bmatrix} \triangleq \gamma \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-2} \\ w_{M-1} \end{bmatrix}, \sum_i w_i = 1$$

with $\gamma = n_0$.

$$\boldsymbol{\lambda}_x \cdot \mathbf{w} = \sum_i (x-i)^2 w_i, \quad \text{minimized at } \hat{x} = \sum_i i w_i\,.$$

The rule for $z_i = 0$:

$$\hat{x}_i = \left[\frac{\delta}{2(1-\delta)n_0} \left(\sum_{i=0}^{M-1} i\, n_i - \frac{n\delta}{2}(M-1) \right) \right],$$

where $\left[\, x \,\right] =$ round x to nearest integer in $\mathcal{A}$.
Symmetric rule for $z_i = M{-}1$.

Salt and Pepper (S&P) Channel, Euclidean Cost – proofs

$$\mathbf{u} \triangleq (\mathbf{\Pi}^{-T}\mathbf{m}) \odot \boldsymbol{\pi}_0 = \frac{1}{1-\delta}\begin{bmatrix} n_0 - n\delta/2 \\ n_1 \\ \vdots \\ n_{M-2} \\ n_{M-1} - n\delta/2 \end{bmatrix} \odot \begin{bmatrix} 1-\delta/2 \\ \delta/2 \\ \vdots \\ \delta/2 \\ \delta/2 \end{bmatrix}$$

$$\begin{aligned}\sum_{i=0}^{M-1} u_i &= \frac{1}{1-\delta}\left[(n_0 - \frac{\delta}{2}n)(1-\frac{\delta}{2}) + \frac{\delta}{2}(n_1 + n_2 + \cdots + n_{M-1}) - \left(\frac{\delta}{2}\right)^2 n\right] \\ &= \frac{1}{1-\delta}\left[(n_0 - \frac{\delta}{2}n)(1-\frac{\delta}{2}) + \frac{\delta}{2}(n - n_0) - \left(\frac{\delta}{2}\right)^2 n\right] = n_0 \Rightarrow \mathbf{u} = n_0\mathbf{w}.\end{aligned}$$

$$\begin{aligned}\sum_{i=0}^{M-1} i w_i &= \frac{1}{(1-\delta)n_0}\left[\frac{\delta}{2}n_1 + 2\frac{\delta}{2}n_2 + \cdots + (M-1)\frac{\delta}{2}n_{M-1} - \left(\frac{\delta}{2}\right)^2 (M-1)n\right] \\ &= \frac{\delta}{2(1-\delta)n_0}\left[\sum_{i=0}^{M-1} i\, n_i - \frac{\delta}{2}(M-1)n\right].\end{aligned}$$

DUDE Optimality: Stochastic Setting

Define

$$\hat{X}^n_{\text{univ}} = \hat{X}^{n,k_n}, \qquad \text{with} \quad k_n \to \infty \quad \text{as} \quad k_n M^{2k_n} = o(n/\log n)$$

Theorem (universality in stochastic setting)

For every stationary ergodic input process,

$$\lim_{n\to\infty} EL_{\hat{X}^n_{\text{univ}}}(X^n, Z^n) = \lim_{n\to\infty} \min_{\hat{X}^n \in \mathcal{D}_n} EL_{\hat{X}^n}(X^n, Z^n)$$

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

DUDE Optimality: Semi-stochastic Setting

$$\hat{X}^n_{\text{univ}} = \hat{X}^{n,k_n}, \qquad \text{with} \quad k_n \to \infty \quad \text{as} \quad k_n M^{2k_n} = o(n/\log n)$$

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 (universality in semi-stochastic setting)

For any input sequence, a.s.

$$\limsup_{n\to\infty} \left[L_{\hat{X}^n_{\text{univ}}}(x^n, Z^n) - D_{k_n}(x^n, Z^n) \right] \le 0$$

Note: Among the competitors for each k, we have the function $f : \mathcal{A}^{2k+1} \to \mathcal{A}$ that minimizes the cost *given* the sequences x^n, z^n.

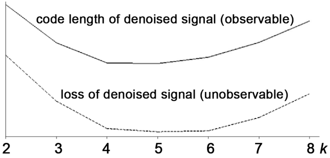
Choosing the Context Length *k*

- Tradeoff:
  - too short ↦ suboptimum performance;
  - too long (⇔ too short n) ↦ counts are unreliable
- $k = k_n$ s.t. $k_n M^{2k_n} = o(n/\log n)$ guarantees asymptotic optimality (e.g., $k_n = \lceil c \log_M n \rceil$, $c < \frac{1}{2}$): *not very meaningful in practice*
- DUDE optimality result has a "redundancy-like" term (convergence to optimal performance): *model cost*

$$\sqrt{\tfrac{2}{\pi}} C_{\mathbf{\Lambda},\mathbf{\Pi}} V_{\mathbf{\Pi}} M^k \sqrt{\tfrac{k+1}{n-2k}} + C_{\mathbf{\Lambda},\mathbf{\Pi}} M^{2k+2} \tfrac{k+1}{n-2k}$$

  where $C_{\mathbf{\Lambda},\mathbf{\Pi}}, V_{\mathbf{\Pi}}$ depend on the channel and cost function

- "Best k" for given sequence: open problem
  - Output compressibility heuristic
  - Dynamic, asymmetric context lengths (tree-like)
  - Using an estimator of the DUDE loss derived from *observables*



Computational Complexity

Time: $O(n)$ *register level* operations

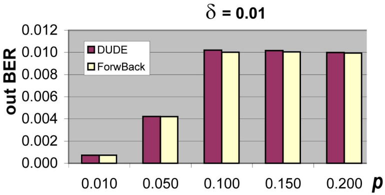
Space: $o(n)$ *working storage* (linear if storage for buffering sequence is counted)

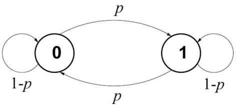
- **Preprocessing:** $O(M^3)$ operations
- **Computation of counts:** $O(n)$ operations (finite state automaton with M^{2k} states)
- **Pre-computations for the second pass:** $O(M^{2k})$ operations
- **Denoising:** O(n) operations

With $k < \frac{1}{2}\log_M n$, we have $M^{2k} = o(n)$.

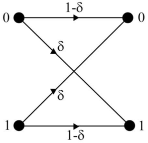
Denoising: Binary Markov Chain over BSC

Sequence length: $n = 10^6$ bits

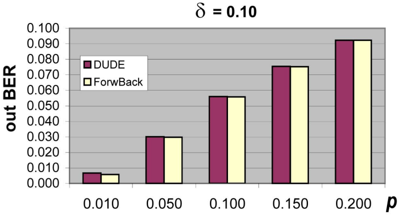


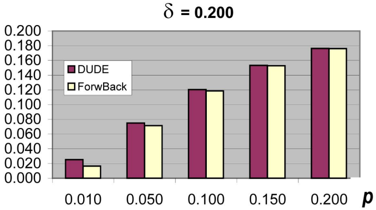


Source: Binary Markov



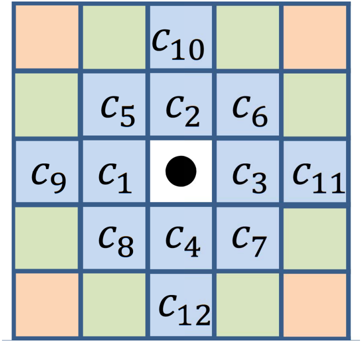
Channel: BSC(δ)





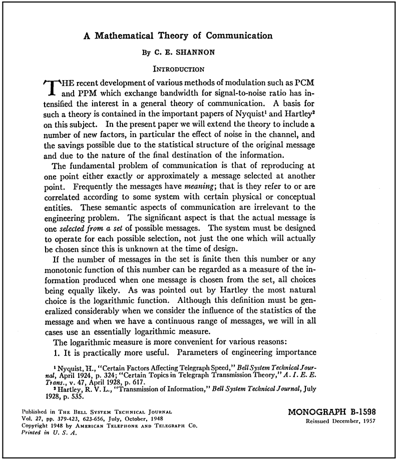
Non-causal 2D Contexts

Example: $K = 12$ (taking closest samples in Euclidean distance)



		c_{10}		
	c_5	c_2	c_6	
c_9	c_1	●	c_3	c_{11}
	c_8	c_4	c_7	
		c_{12}		

Examples: Binary images + BSC + Hamming distance



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
Copyright 1948 by AMERICAN TELEPHONE AND TELEGRAPH Co.
Printed in U. S. A.

MONOGRAPH B-1598
Reissued December, 1957

image: 1800x2104 scan

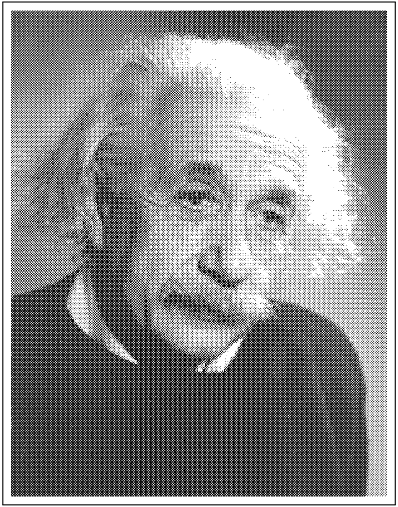
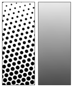
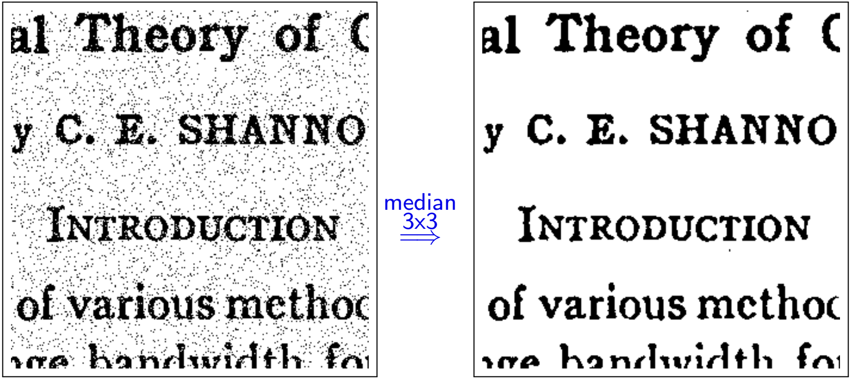


image: 896x1160 half-tone



The Importance of Universality

Scanned text through Binary Symmetric Channel (BSC)



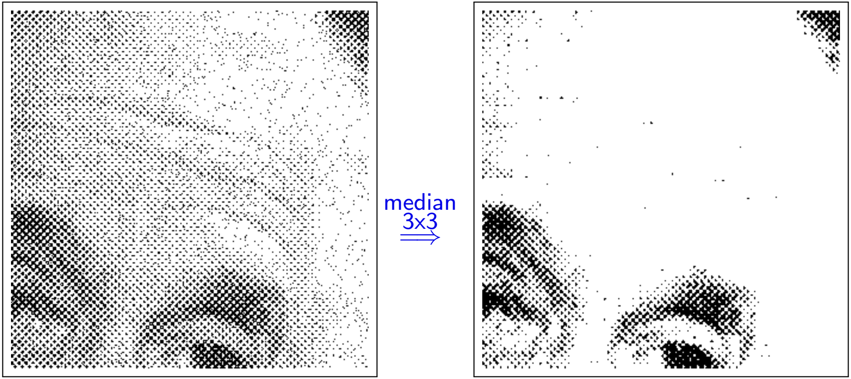
al Theory of C
y C. E. SHANNO
INTRODUCTION
of various methoc
age bandwidth for

al Theory of C
y C. E. SHANNO
INTRODUCTION
of various methoc
age bandwidth for

some simple denoising filters work well for some types of input data

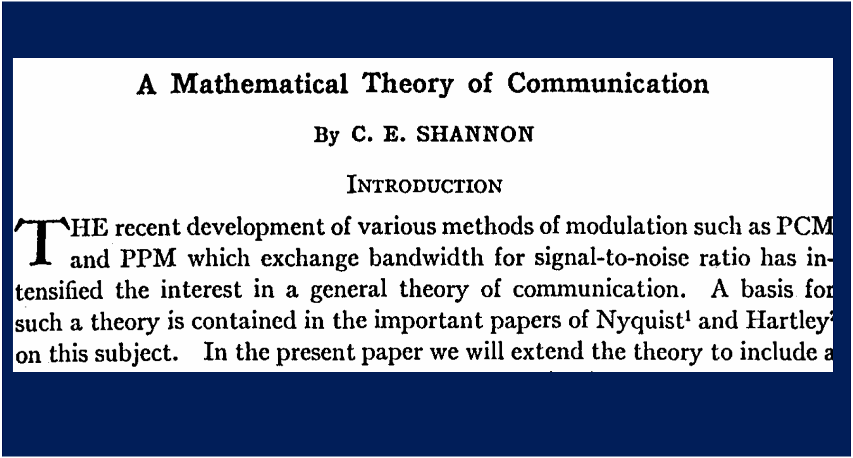
The Importance of Universality

Halftone image through Binary Symmetric Channel (BSC)



but are catastrophic for others!

Binary image: original (i)



A Mathematical Theory of Communication

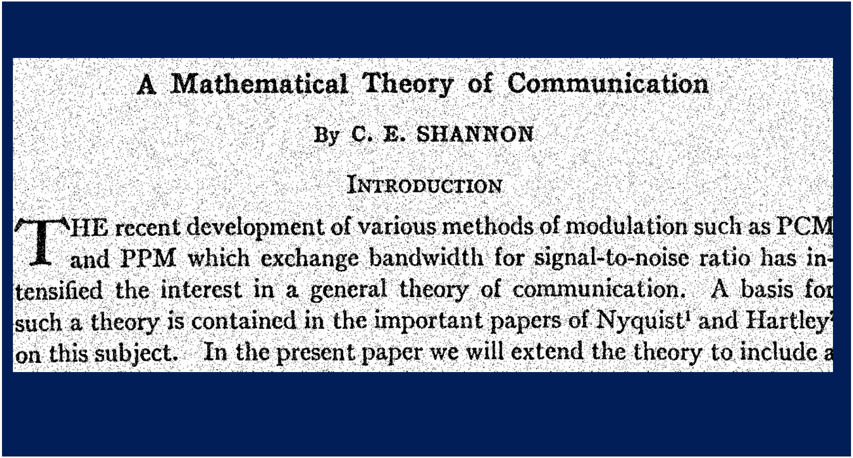
By C. E. SHANNON

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

image: 1800x2104 scan (1296x496 segment shown)

Binary image: corrupted by BSC (i)



A Mathematical Theory of Communication

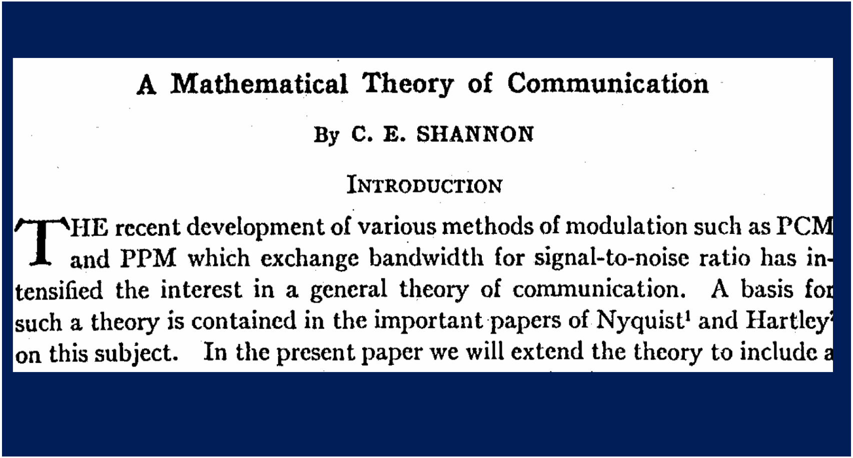
By C. E. SHANNON

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

random bit error rate: 5.0%

Binary image: denoised by DUDE (i)



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

denoised bit error rate: 0.4%

Binary image: original (ii)

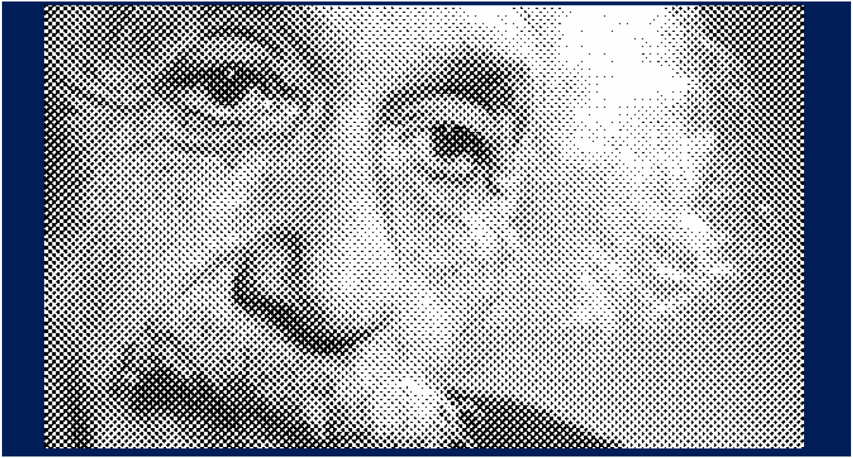
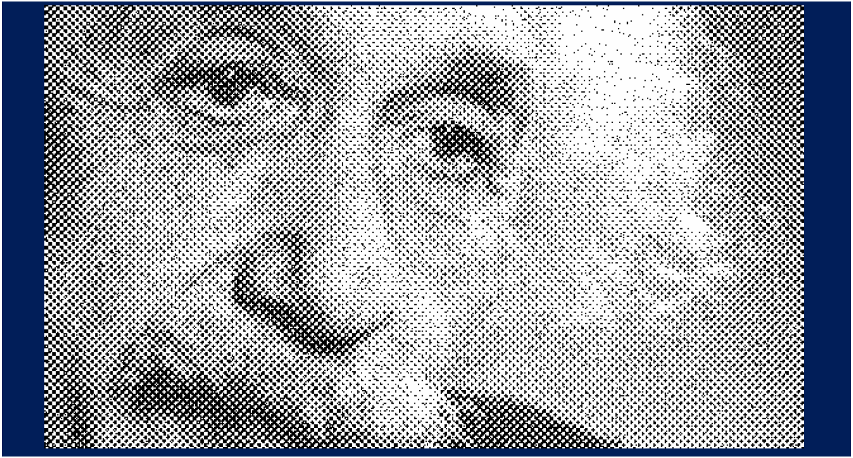


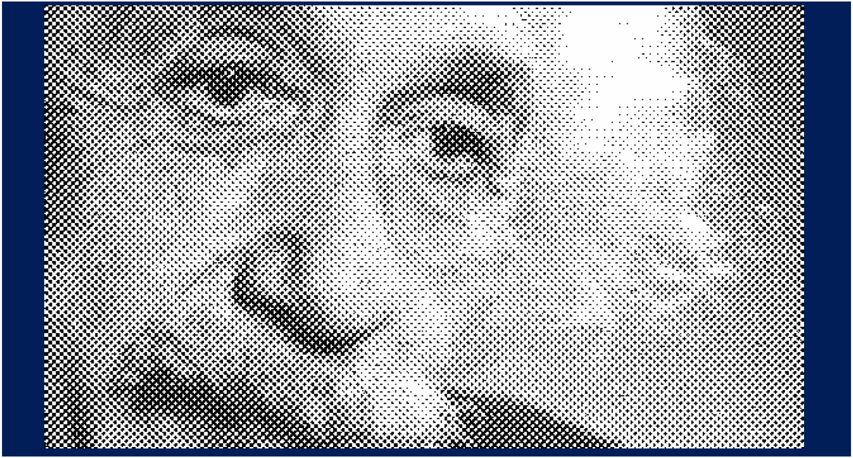
image: 896x1160 half-tone (600x350 segment shown)

Binary image: corrupted by BSC (ii)



random bit error rate: 2.0%

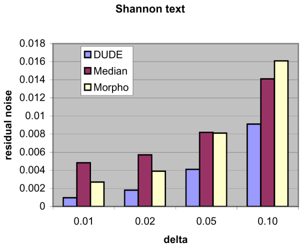
Binary image: denoised by DUDE (i)

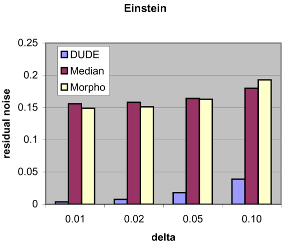


denoised bit error rate: 0.7%

Comparison with some known algorithms

		Channel parameter δ			
Image	Scheme	0.01	0.02	0.05	0.10
Shannon 1800×2160	DUDE	0.00096 $K=11$	0.0018 $K=12$	0.0041 $K=12$	0.0091 $K=12$
	median	0.00483	0.0057	0.0082	0.0141
	morpho.	0.00270	0.0039	0.0081	0.0161
Einstein 896×1160	DUDE	0.0035 $K=18$	0.0075 $K=14$	0.0181 $K=12$	0.0391 $K=12$
	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):

"Wha*r* giants?" said Sancho Panza. "Those thou seest the*e*e," *s*nswered *y*is master, "with the long arms, and s*pn*e have t*g*em n*d*arly two leagues long." "Look, y*l*ur worship," sai*r* Sancho; "what we see there *z*re not gian*r*s but windmills, and what seem to be their arms are the sails that turned by the wind make *r*he millst*p*ne go." "*K*t is easy to see," replied Don Quixote, "that thou art not used to this business of adventures; *f*hose are giant*z*; and if thou ar*f* *w*fra*o*d, away with thee out of this and betake thyse*p*f to prayer while I engage them in fierce and unequal combat."

DUDE output, $k = 2$ (7 errors):

"What giants?" said Sancho Panza. "Those thou seest there," answered his master, "with the long arms, and s*pn*e 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; *f*hose are giant*z*; and if thou ar*f* *w*fra*o*d, away with thee out of this and betake thyself to prayer while I engage them in fierce and unequal combat."