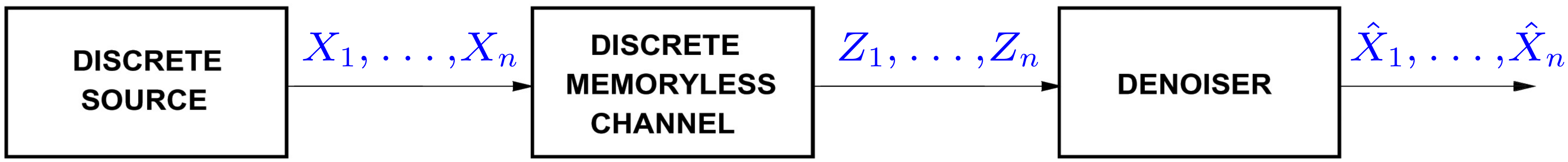
DUDE:

An Algorithm for Discrete Universal Denoising

Tsachy Weissman	HP Labs and Stanford
Erik Ordentlich	HP Labs
Gadiel Seroussi	HP Labs
Sergio Verdú	Princeton
Marcelo Weinberger	HP Labs

Discrete Denoising



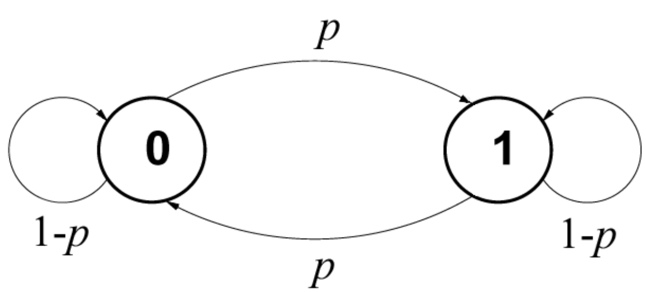
- X_i, Z_i, $\hat{X}_i$ take values from *finite alphabets*.
- Goal: Choose $\hat{X}_1, \ldots, \hat{X}_n$ on the basis of $Z_1, \ldots, Z_n$ to minimize a *fidelity criterion* (*which we might not be able to measure!*).

Applications

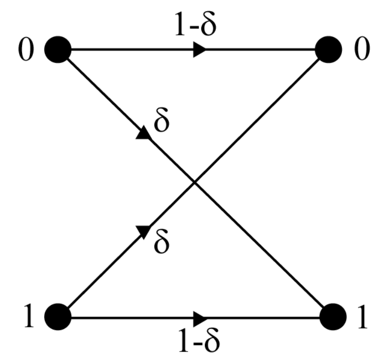
- Text Correction
- Image Denoising
- Reception of Uncoded Data
- DNA Sequence Analysis and Processing
- Hidden Markov Model State Estimation
- . . .

Example

Source: Binary Markov Chain



Channel: BSC



$\ldots 0001111100001111100 \ldots$

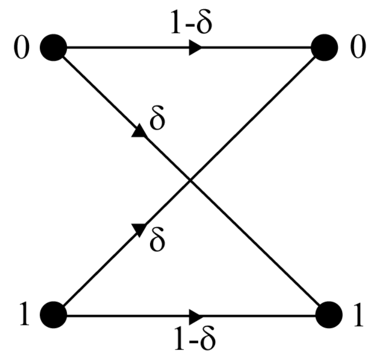
$\ldots 0001000001000001010 \ldots \quad \Rightarrow \ldots 0000111101001110110 \ldots$

- Objective: Minimize Bit Error Rate given the observation of n-block.
- Solution: Backward-Forward Dynamic Programming
- Fundamental Limit: $\lim_{n\to\infty}$ Min Error Probability $= f(\delta, p)$ still open.

Universal Example

- Source: Binary; *unknown statistics*

- Channel: BSC



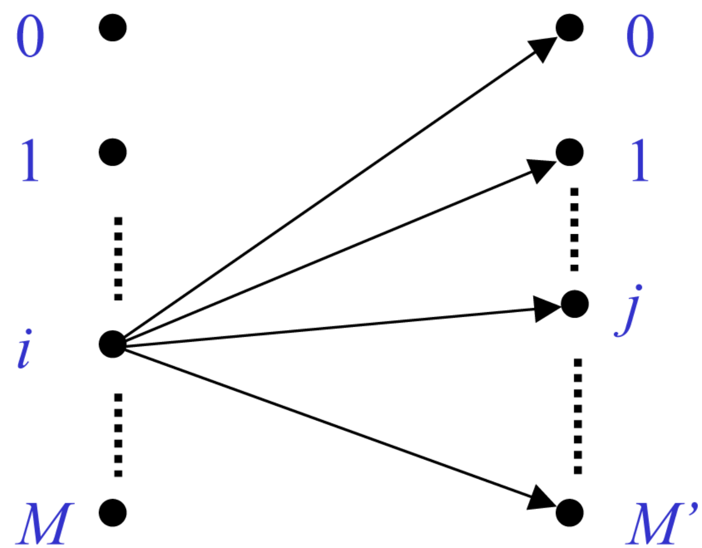
- Objective: Minimize Bit Error Rate given the observation of n-block.

- Solution: ???

- Fundamental Limit: $\lim_{n\to\infty}$ Min Error Probability = ???

Universal Setting: Basic Assumptions

- Unknown Source Statistics
- Discrete Memoryless Channel (DMC) with *known transition probability matrix*



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

Alphabet: Same finite alphabet for clean and noisy signals: $\mathcal{A},\ |\mathcal{A}| = M$.

Channel: Nonsingular transition probability matrix:

$$\mathbf{\Pi} = \{\Pi(i,j)\}_{i,j\in\mathcal{A}} = [\boldsymbol{\pi}_1 \mid \cdots \mid \boldsymbol{\pi}_M]$$

both assumptions above can be relaxed

n-block denoiser: $\hat{X}^n : \mathcal{A}^n \to \mathcal{A}^n$.

Loss Function: $\Lambda : \mathcal{A}^2 \to [0,\infty)$, represented by the matrix

$$\mathbf{\Lambda} = \{\Lambda(i,j)\}_{i,j\in\mathcal{A}} = [\boldsymbol{\lambda}_1 \mid \cdots \mid \boldsymbol{\lambda}_M] .$$

cost of guessing j when clean signal is i

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

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

Performance Benchmark

Optimum distribution-dependent performance for a denoiser that knows the input statistics:

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

where $\mathcal{D}_n$ is the class of all (including non-universal) n-block denoisers

The DUDE Algorithm: General Idea

- Fix *context length* k. For each letter 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 $(r_1, \ldots, r_k)$. This gives a *conditional empirical distribution* of the *noisy* symbol given the contexts $(\ell_1, \ldots, \ell_k)$ and $(r_1, \ldots, r_k)$.

- Use channel transition probability to estimate the conditional empirical distribution of the *noiseless* symbol given $(\ell_1, \ldots, \ell_k)$ and $(r_1, \ldots, r_k)$.

- Make decision using
  - the loss function,
  - the channel transition probability,
  - the conditional empirical distribution
  - the observed letter 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 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 $k = 1$ | 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 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 $k = 1$ $\boxed{\mathbf{e}}\boxed{\bullet}\boxed{\mathbf{r}}$ counts

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

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 $k = 2$ | h | e | • | r | e | counts

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

Notation for Context Counts

- $\mathbf{a} \in \mathcal{A}^n$, whole data sequence
- $\mathbf{b} \in \mathcal{A}^k$, left k-context string
- $\mathbf{c} \in \mathcal{A}^k$, right k-context string
- Let $\mathbf{m}[\mathbf{a}, \mathbf{b}, \mathbf{c}]$ denote the M-vector with α-th component equal to the number of occurrences of the pattern $\boxed{\mathbf{b}}\boxed{\alpha}\boxed{\mathbf{c}}$ in $\mathbf{a}$:

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

- Example:

$$\mathbf{m}[\texttt{Shannon text}, he, re] \quad = \quad \begin{array}{l} [0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,5\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,7]^T \end{array}$$

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

The Discrete Universal Denoiser

Fix k

Pass 1 For every $k+1 \le i \le n-k$:
compute the count vectors $\mathbf{m}(z_1^n, z_{i-k}^{i-1}, z_{i+1}^{i+k})$

Pass 2 Correct according to:

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

where

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

$$\begin{aligned}
\mathbf{\Pi} &= \{\Pi(i,j)\}_{i,j\in\mathcal{A}} = [\boldsymbol{\pi}_1 \mid \cdots \mid \boldsymbol{\pi}_M] \\
\mathbf{\Lambda} &= \{\Lambda(i,j)\}_{i,j\in\mathcal{A}} = [\boldsymbol{\lambda}_1 \mid \cdots \mid \boldsymbol{\lambda}_M]\,. \\
(\mathbf{v} \odot \mathbf{w})_i &= v_i w_i
\end{aligned}$$

Motivation

- Given a distribution on an unknown random variable Y, the minimum expected distortion guess is

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

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

- Key:

$$\left\{ \boldsymbol{\pi}_{z_i} \odot \boldsymbol{\Pi}^{-T} \mathbf{m}[z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k}] \right\}$$

= unnormalized empirical estimate of $P_{X_i | Z^n = z^n}$

Motivation

$$\left\{\boldsymbol{\pi}_{z_i} \odot \boldsymbol{\Pi}^{-T}\mathbf{m}[z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k}]\right\}$$

unnormalized empirical estimate of $P_{X_i|Z^n=z^n}$

If $n = 1$, no context:

$$\begin{aligned}
\mathbf{P}_Z &= \boldsymbol{\Pi}^T\mathbf{P}_X \\
\mathbf{P}_{X|z}(a) &= \frac{P(Z=z|X=a)\mathbf{P}_X(a)}{\mathbf{P}_Z(z)} = \frac{\Pi(a,z)[\boldsymbol{\Pi}^{-T}\mathbf{P}_Z](a)}{\mathbf{P}_Z(z)}, \\
&\Leftrightarrow \\
\mathbf{P}_{X|z} &= \frac{1}{\mathbf{P}_Z(z)}\boldsymbol{\pi}_z \odot [\boldsymbol{\Pi}^{-T}\mathbf{P}_Z].
\end{aligned}$$

Motivation

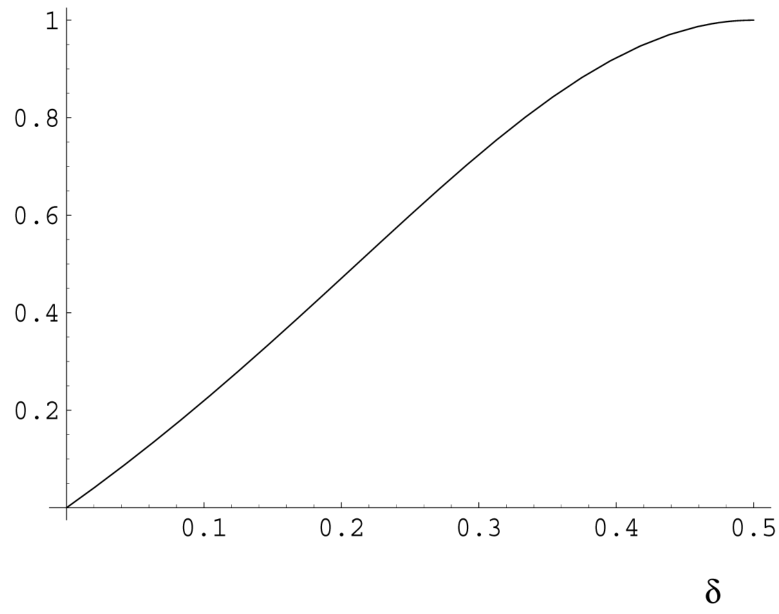
Because the channel is memoryless:

$$\mathbf{P}_{X_t|Z(T)=z(T)} \quad = \quad \frac{1}{\mathbf{P}_{Z_t|z(T\setminus t)}(z_t)} \boldsymbol{\pi}_{z_t} \odot \left[\mathbf{\Pi}^{-T} \mathbf{P}_{Z_t|z(T\setminus t)} \right].$$

Example: BSC + Error Probability

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.

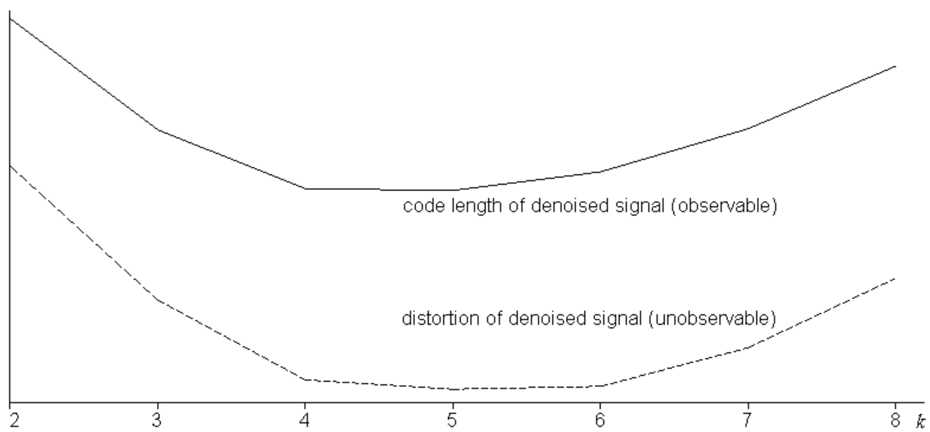
Example: M-ary erasure channel + Error Probability

$$
\begin{aligned}
g_{z^n}^k(u_{-k}^{-1} e u_1^k) &= \arg\min_{\hat{x}\in\{1,\ldots,M\}} \frac{\delta}{1-\delta} \sum_{a\in\{1,\ldots,M\}:a\neq\hat{x}} \mathbf{m}[z^n, u_{-k}^{-1}, u_1^k](a) \\
&= \arg\min_{\hat{x}\in\{1,\ldots,M\}} \left[\left(\sum_{a\in\{1,\ldots,M\}} \mathbf{m}[z^n, u_{-k}^{-1}, u_1^k](a) \right) - \mathbf{m}[z^n, u_{-k}^{-1}, u_1^k](\hat{x}) \right] \\
&= \arg\max_{\hat{x}\in\{1,\ldots,M\}} \mathbf{m}[z^n, u_{-k}^{-1}, u_1^k](\hat{x}).
\end{aligned}
$$

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
- $k = k_n$ s.t. $k_n M^{2k_n} = o(n/\log n)$ guarantees asymptotic optimality (e.g., $k_n = \lceil c \log_M n \rceil$, $c < \frac{1}{2}$).
- "Best k" for given sequence: open problem
  - Output compressibility heuristic
  - Dynamic, asymmetric context lengths



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 $M^{2k} = o(n)$.

Optimality Result: Stochastic Setting

Define

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

Theorem. (No asymptotic penalty for universality)

For every stationary ergodic input process,

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

where $\mathcal{D}_n$ is the class of all (including non-universal) n-block denoisers

Optimality Result: Semi-Stochastic Setting

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

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. For any input sequence, a.s.

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

Proof Sketch

- $P_{X_i|Z_{i-k}^{i+k}=z_{i-k}^{i+k}}$ is not available because the source statistics are unknown.

- A genie that sees both x^n and z^n can give an empirical (unnormalized) estimate $\mathbf{q}[Z^n, x^n, u_{-k}^k] = \tilde{P}_{X_i|Z_{i-k}^{i+k}=u_{-k}^k}$

- Even without the genie, DUDE can get very close to that estimate:

$$\Bigg\| \underbrace{\mathbf{q}[Z^n, x^n, u_{-k}^k]}_{\text{genie}} - \underbrace{\boldsymbol{\pi}_{u_0} \odot \left[\boldsymbol{\Pi}^{-T}\mathbf{m}[Z^n, u_{-k}^{-1}, u_1^k]\right]}_{\text{DUDE}} \Bigg\|_1$$

  can be bounded tightly enough that:

- $\Pr\left(L_{\hat{X}^{n,k}}(x_{k+1}^{n-k}, Z^n) - D_k(x^n, Z^n) > \epsilon\right) \leq K_1(k+1)M^{2k+1}\exp\left(-\frac{(n-2k)\epsilon^2}{4(k+1)M^{2k}F_{\boldsymbol{\Pi}}C_{\boldsymbol{\Lambda},\boldsymbol{\Pi}}^2}\right)$

- RHS summable for $k_n M^{2k_n} = o(n/\log n) \Rightarrow$ a.s. result by BC Lemma

Nonsquare/Nonsingular $\mathbf{\Pi}$

- If $\mathbf{\Pi}$ is full-row rank, then replace $\mathbf{\Pi}^{-1}$ by a pseudo-inverse, e.g. the Moore-Penrose pseudo-inverse $\mathbf{\Pi}^T\left(\mathbf{\Pi}\mathbf{\Pi}^T\right)^{-1}$. Same optimality results.

- If $\mathbf{\Pi}$ is not full-row rank, then, in general, optimum distribution-dependent performance cannot be obtained by a universal algorithm.

Experiment: Binary Markov Chain (p) $\rightarrow$ BSC (δ); $n = 10^6$

p	$\delta = 0.01$		$\delta = 0.10$		$\delta = 0.20$	
	DUDE	ForwBack	DUDE	ForwBack	DUDE	ForwBack
0.01	0.000723 [3]	0.000721	0.006648 [5]	0.005746	0.025301 [6]	0.016447
0.05	0.004223 [3]	0.004203	0.030084 [5]	0.029725	0.074936 [5]	0.071511
0.10	0.010213 [8]	0.010020	0.055976 [3]	0.055741	0.120420 [4]	0.118661
0.15	0.010169 [8]	0.010050	0.075474 [5]	0.075234	0.153182 [4]	0.152903
0.20	0.009994 [8]	0.009940	0.092304 [3]	0.092304	0.176354 [4]	0.176135

Table 1: Bit Error Rates; context length $[k]$ chosen according to LZ heuristic

Image Denoising: δ=0.05, k=12 (2D)

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

Printed in U. S. A.

MONOGRAPH B-1598
Reissued December, 1957

Image Denoising: $\delta=0.02$, $k=14$ (1D)

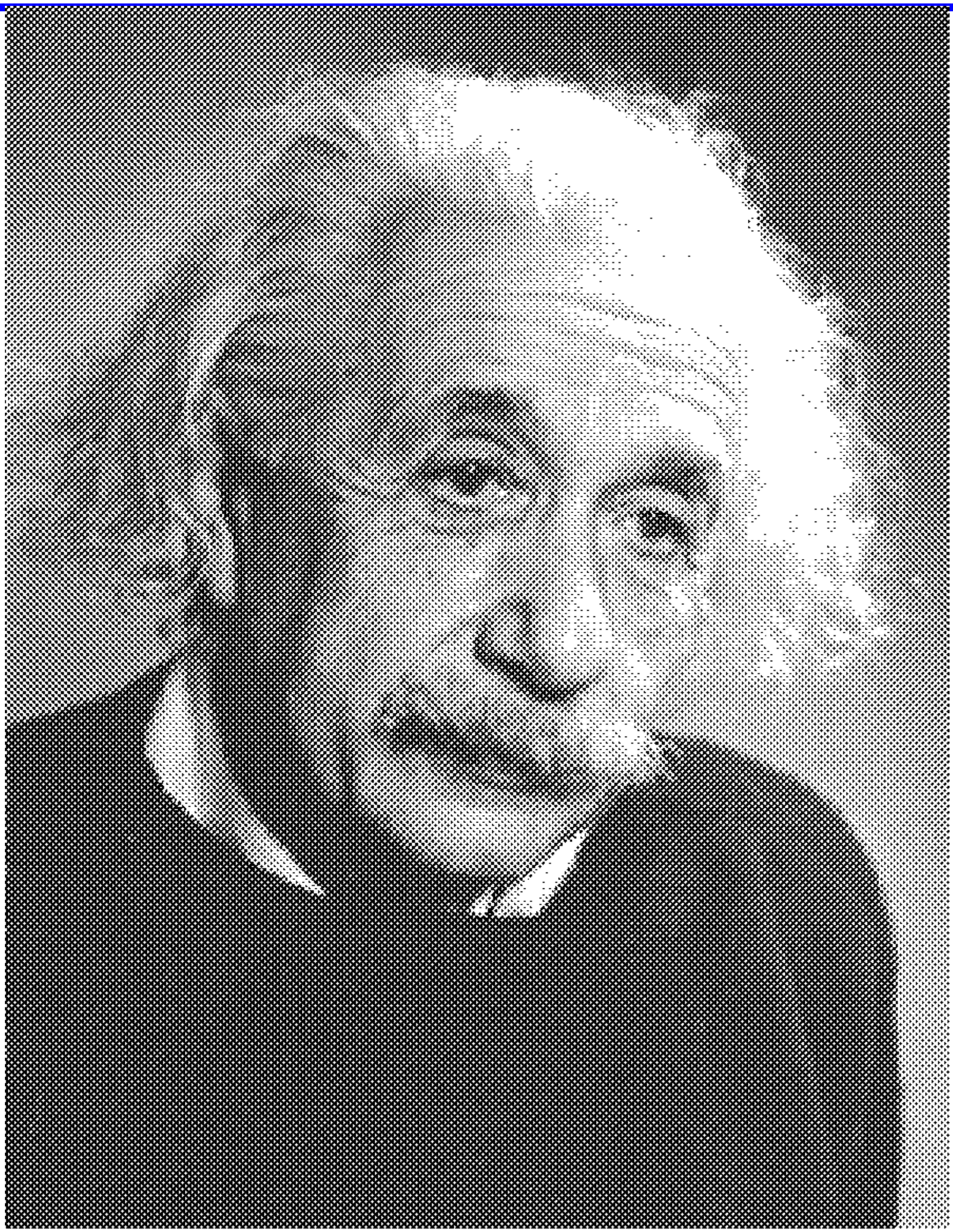


Image Denoising: Comparison with existing algorithms

		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 $K=14^\dagger$	0.0181 $K=12^\dagger$	0.0391 $K=12^\dagger$
	median	0.156	0.158	0.164	0.180
	morpho.	0.149	0.151	0.163	0.193

Table 2: Denoising results for binary images

Text Denoising: Don Quixote de La Mancha

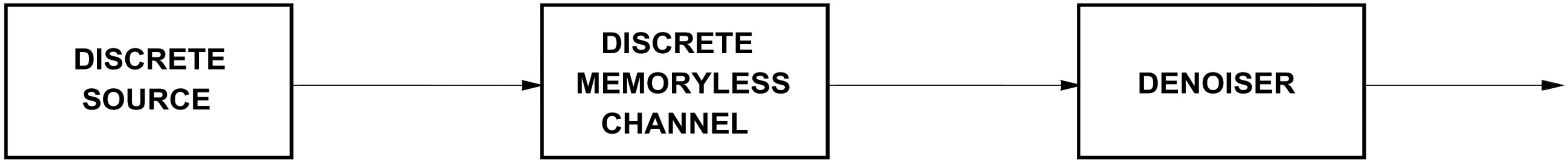
Noisy Text (21 errors, 5% error rate):

”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, $k = 2$ (7 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 arf wfraod, away with thee out of this and betake thyself to prayer while I engage them in fierce and unequal combat.”

Discrete Denoising



“If the source already has a certain redundancy and no attempt is made to eliminate it... a sizable fraction of the letters can be received incorrectly and still reconstructed by the context.” (Claude Shannon, 1948)