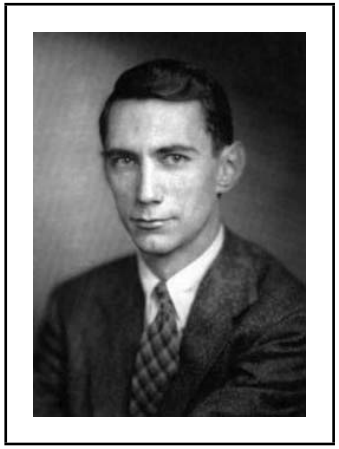
Stationary Codes



Shannon meets Ornstein

Robert M. Gray
Stanford University
rmgray@stanford.edu

Research partially supported by

Part I

Flipping coins, stationary codes, information sources, modeling, entropy, process distance, optimal fakes

Introduction: Flipping coins

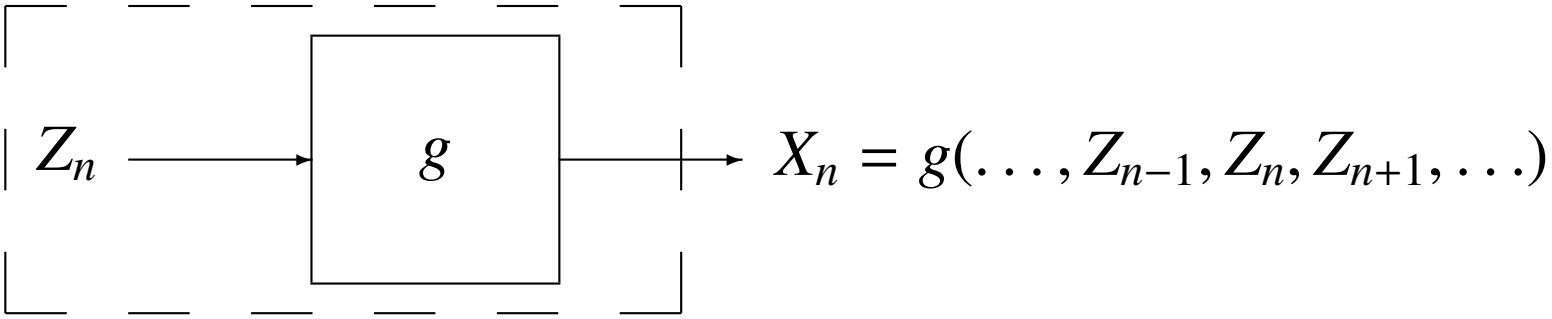
Arguably the simplest
nontrivial random process is a sequence
$Z = \{Z_n; n \in \mathcal{Z}\}$
of independent tosses of a fair coin

$$\cdots 01001100010100000101100111 \cdots$$

Process plays a basic role in the theory, practice, interpretation, and teaching of random processes and information theory

— moreover coin flips provide a building block for modeling more general processes and the process arises naturally inside optimal source codes

Modeling example: stationary coding of coin flips



stationary code = time-invariant (or shift-invariant) possibly nonlinear filter

$\Leftrightarrow$ Shift input sequence $\Rightarrow$ shift output sequence

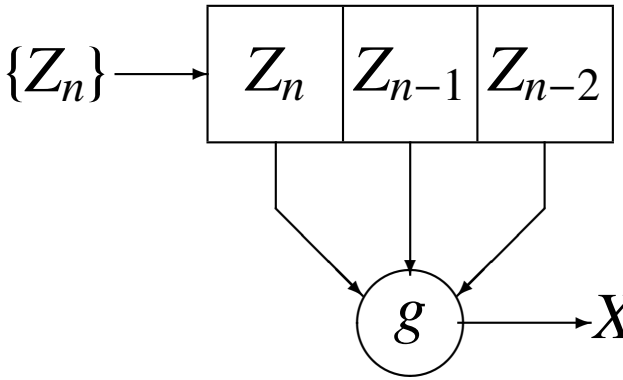
Nice property of stationary codes: preserve nice statistical properties of input: stationarity, ergodicity, mixing, K, B

(will define later)

How general a class of stationary processes has Z at its ♡?

Call this class $B(1)$: B = Bernoulli, 1=$\log_2$ (input alphabet size)

An example in $B(1)$

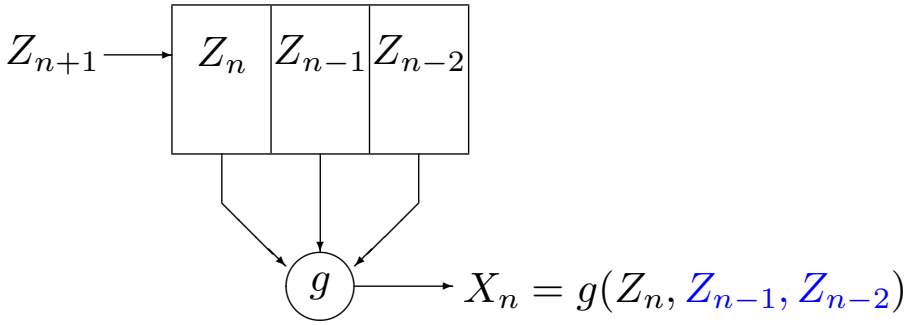


shift register

function, table

$Z_nZ_{n-1}Z_{n-2}$	X_n
000	0.7683
001	-0.4233
010	-0.1362
011	1.3286
100	0.4233
101	0.1362
110	-1.3286
111	-0.7683

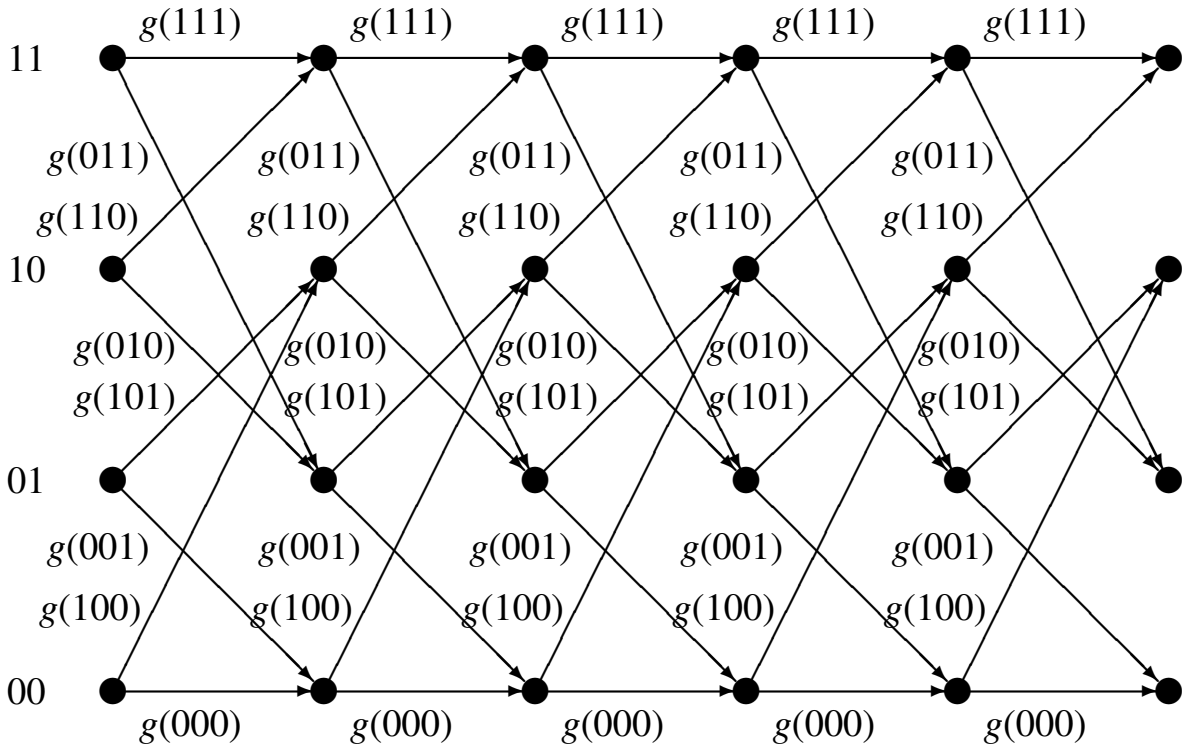
Output marginal distribution resembles $\mathcal{N}(0, 1)$


Z_{n+1}
Z_n
Z_{n-1}
Z_{n-2}
g
$X_n = g(Z_n, Z_{n-1}, Z_{n-2})$

Output process has constrained structure, sequences lie on a directed graph called a *trellis* (a *tree* if the shift register has infinite length)

Trellis of a stationary code

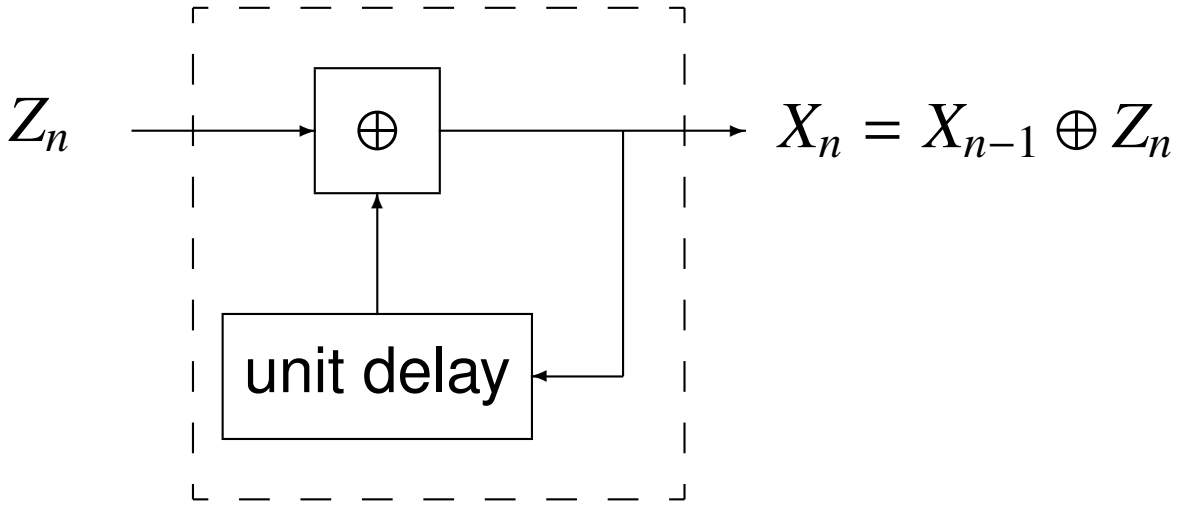
Nodes denote shift-register states, lines denote transitions or branches depending on state and input



If $g(z_0 z_1 z_2)$ are all distinct, can recover input sequence from output sequence. This stationary code is *invertible*

Another example: Binary autoregressive process

A linear (mod 2, GF(2)) time-invariant (LTI) filter:

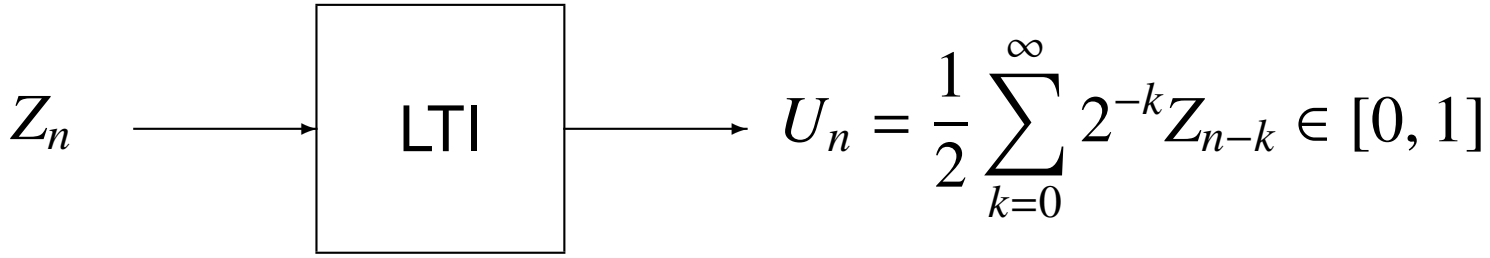


binary in, binary out — symmetric binary Markov/autoregressive process

Again invertible with stationary code: $Z_n = X_n \oplus X_{n-1}$

Another example: LTI with real arithmetic

More generally, *convolutional code with real arithmetic* $\Rightarrow$ linear time-invariant (LTI) filter, e.g.



— binary expansion of real number in unit interval

Discrete input alphabet, continuous output alphabet!

Fair coin flips in, output $U_n \sim \mathrm{U}([0,1])$,

uniform marginal distributions

Unlike block codes, infinite-length stationary codes make sense!

Like previous examples, this stationary code is invertible by another stationary code (again an LTI filter)

$$2U_n - U_{n-1} = \sum_{k=0}^{\infty} 2^{-k} Z_{n-k} - \frac{1}{2} \sum_{k=0}^{\infty} 2^{-k} Z_{n-1-k} = Z_n$$

Coin flips have binary alphabet $\{0, 1\}$, but output of stationary code might have same or larger alphabet A_X such as $\{1, 2, 3, 4, 5, 6\}$ (to resemble a fair die), *possibly even a continuous alphabet such as* $[0, 1]$ *or* $\mathbb{R}$ if the shift-register has infinite length!

Detour: block vs. stationary (sliding-block) codes

Quick discussion, aimed primarily at those with minimal information theory background

Code a process X with alphabet A_X into a process Y with alphabet A_Y:

Block coding Map each nonoverlapping block of source symbols into an index or block of encoded symbols (e.g., bits)

(standard for information theory)

Stationary coding Map overlapping blocks of source symbols into single encoded symbol (e.g., bit) (standard for ergodic theory)

Block Coding $\mathcal{E} : A_X^N \to A_Y^N$ (or other index set), N = block length

$$\begin{array}{ccccccc} \cdots, & \underbrace{X_{-N}, X_{-N+1}, \ldots, X_{-1},} & & \underbrace{X_0, X_1, \ldots, X_{N-1},} & & \underbrace{X_N, X_1, \ldots, X_{2N-1},} & \cdots \\ \cdots, & & \overbrace{\downarrow \mathcal{E}}^{} & & \overbrace{\downarrow \mathcal{E}}^{} & & \cdots \\ \cdots, & \overbrace{Y_{-N}, Y_{-N+1}, \ldots, Y_{-1},} & & \overbrace{Y_0, Y_1, \ldots, Y_{N-1},} & & \overbrace{Y_N, Y_1, \ldots, Y_{2N-1},} & \cdots \end{array}$$

Sliding-block Coding N=window length $= N_1 + N_2 + 1$, $f : A_X^N \to A_Y$

$$\cdots, \underbrace{X_{n-N_1}, X_{n-N_1+1}, \cdots, X_n, X_{n+1}, \cdots, X_{n+N_2}}, X_{n+N_2+1}, \cdots$$

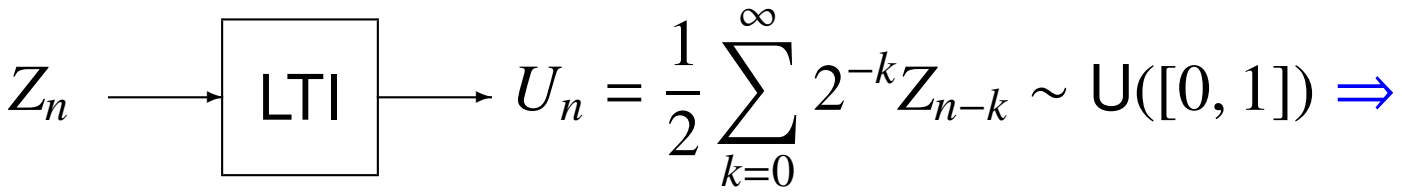
slide window $\longrightarrow$

$$Y_n = f(X_{n-N_1}, \ldots, X_n, \ldots, X_{n+N_2})$$

$$Y_{n+1} = f(X_{n-N_1+1}, \ldots, X_{n+1}, \ldots, X_{n+N_2+1})$$

Both structures induce mappings of sequences into sequences

Back to coding coin flips



can get *arbitrary* output marginal distribution via elementary probability trick:

Given cdf F on $\mathbb{R}$, e.g., cdf for $\mathcal{N}(0,1)$

Define F^{-1} (generalized) inverse cdf:

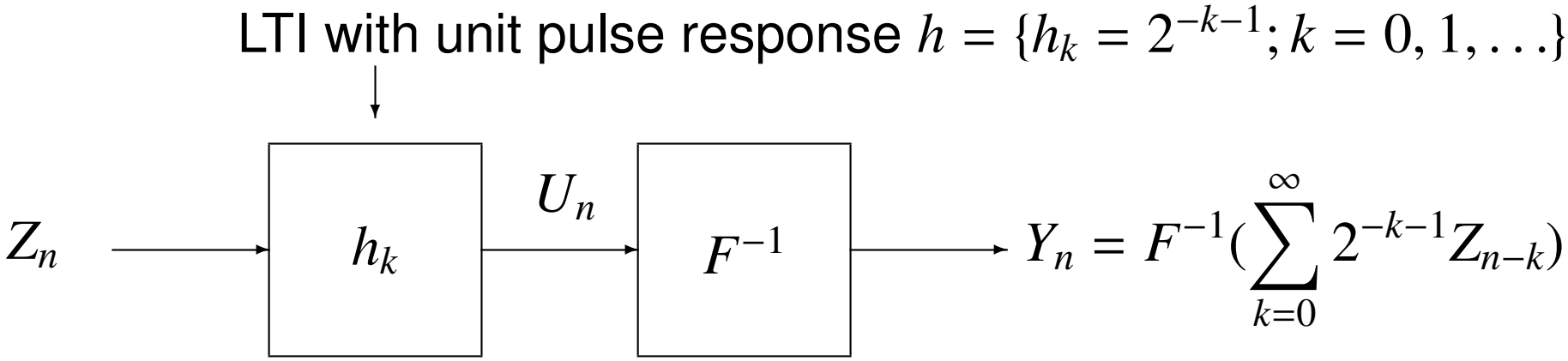
$F^{-1}(u) = \inf\{r\colon F(r) \geq u\} \quad \Rightarrow \quad Y_n = F^{-1}(U_n) \sim F$, e.g., Gaussian

Here stationary code = *LTI filter + memoryless nonlinearity*

Aside: Example of a Hammerstein nonlinear system =

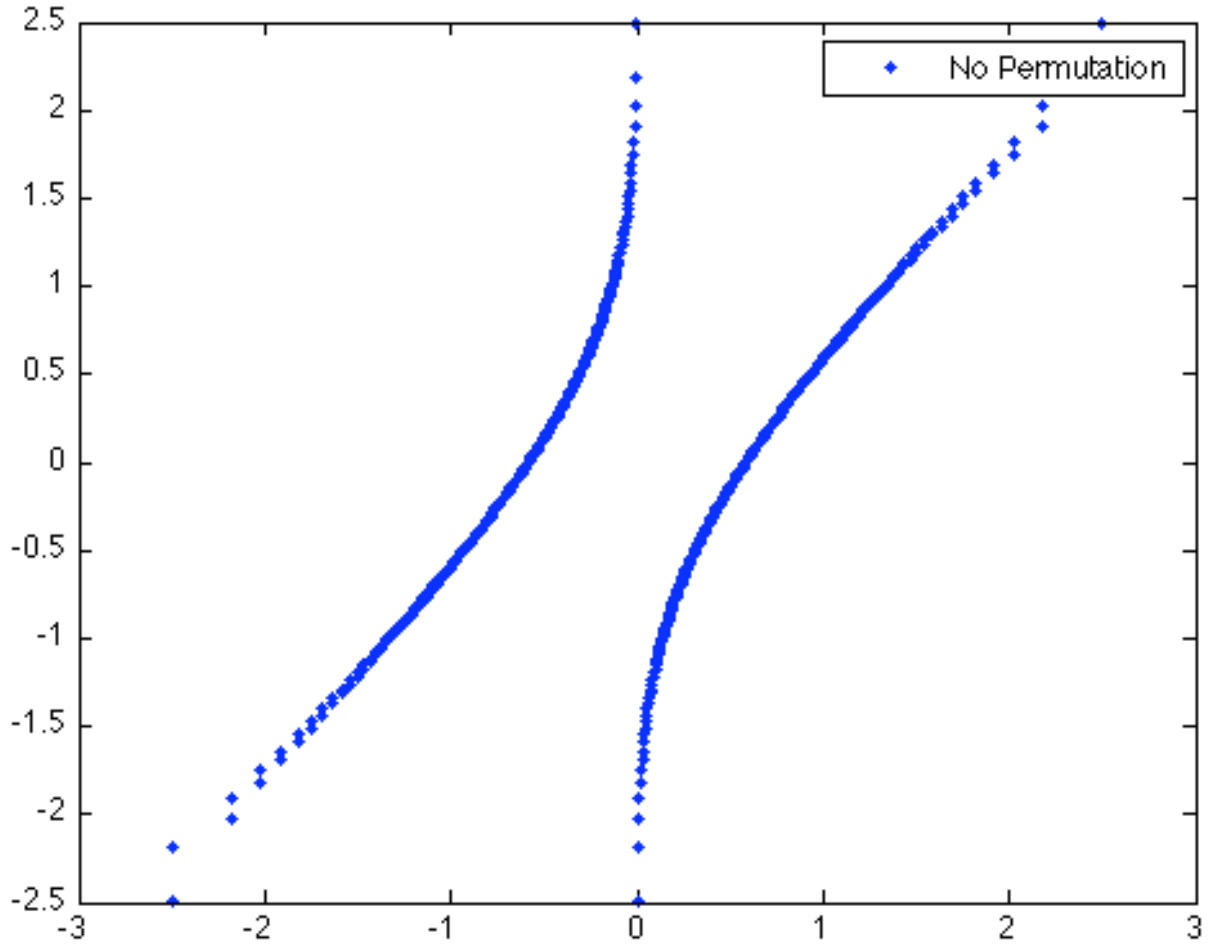
LTI filter + memoryless nonlinearity+ LTI filter

Have generated a process with Gaussian marginals from coin flips:



Is Y_n Gaussian?

No. The condtional probability distributions of Y_n given past values are discrete. Scatter plot of consecutive adjacent samples shows dependence:

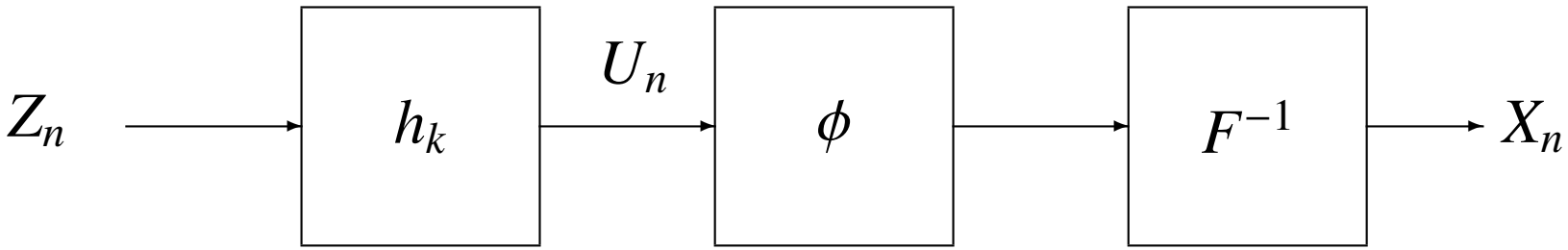


Fake white Gaussian process

Can tweak again and decorrelate:

$$\{Z_n\} \text{ coin flips, } U_n = \frac{1}{2}\sum_{k=0}^{\infty} 2^{-k} Z_{n-k}, \; F = \text{cdf of } \mathcal{N}(0,1)$$

Add: $\phi : [0,1) \to [0,1)$ satisfies $\phi(u) + \phi(u+1/2) = 1$, $u \in [0,1)$

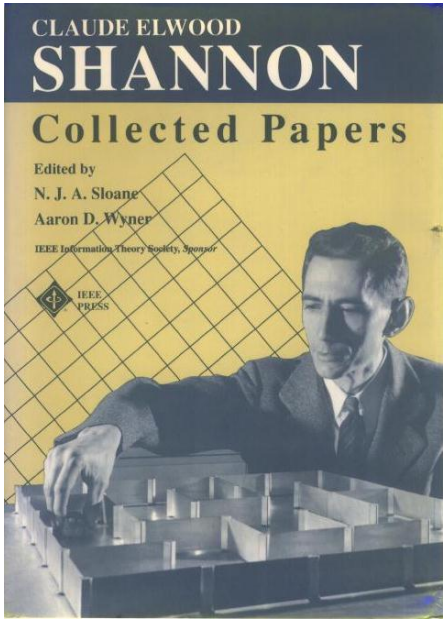


$$X_n = F^{-1}\left(\underbrace{\phi\left(\underbrace{\sum_{k=0}^{\infty} 2^{-k-1} Z_{n-k}}_{\sim \text{Unif}([0,1))}\right)}_{\sim \text{Unif}([0,1))}\right) \Rightarrow \text{Gaussian marginals \& uncorrelated!}$$

Is $\{X_n\}$ a Gaussian process?

No, can't be (Why??)

but how *close* to a Gaussian process *can* it be??



and what has all this
to do with information theory??

Questions raise issues in information theory *and ergodic theory* (especially Shannon and Ornstein):

- Taxonomy of information sources/random processes
- Entropy and entropy rate
- Stationary codes
- Distortion and distance between processes
- Modeling vs. compression (Simulation vs. source coding)

Sketch several familiar and perhaps less familiar relevant ideas at the border of information theory and ergodic theory, with a common thread of stationary codes.

Tools and intuition differ from ubiquitous block coding treatments.

Information sources

Discrete-time *information source* = discrete-time *random process* $X = \{X_n; n \in \mathcal{Z}\}$ described by a *process distribution* μ_X

i.e., Kolmogorov (directly-given) random process model = *distribution* μ_X *on sequence space* A_X^∞ + suitable sigma-field (*event space*)

$X_n \in A_X$ = *alphabet:* discrete or *maybe not*

The Shift

Ergodic theory focuses on the *shift transformation* on sequence space:

Shift $T : A_X^\infty \to A_X^\infty$: shift sequence left one time unit

$$Tx = T(\cdots, x_{n-1}, x_n, x_{n+1}, x_{n+2}, \cdots)$$

$$\swarrow$$

$$= (\cdots, x_n, x_{n+1}, x_{n+2}, x_{n+3}, \cdots)$$

A *dynamical system* in ergodic theory: $[A_X^\infty, \mu_X, T, X_0]$

$$\Rightarrow \text{process } X_n(x) = X_0(T^n x)$$

Generalization of random process (T might not be the shift)

Stationarity

An information source X is *stationary* (shift-invariant) if

$$\mu_X(T^{-1}F) = \mu_X(F) \text{ all events } F$$

where $T^{-1}F = \{x : Tx \in F\}$

shifting an event does not change its probability

Ergodic theory language: T is *measure preserving*

Ergodic theory = theory of measure preserving transformations
(and other related transformations)

i.e., of stationary random processes

and generalizations with similar behavior

Ergodicity

Information source is *ergodic* if invariant events $T^{-1}F = F$ must have probability 0 or 1

Emphasis in literature is on stationary/measure preserving and ergodic, but much remains true more generally

Random Vectors

Random process distribution $\mu_X \Rightarrow$ *random vectors*

$$X^N = (X_0, X_1, \cdots, X_{N-1}) \sim \mu_{X^N}$$

consistent family of distributions on A_X^N

Kolmogorov: $\mu_X \Leftrightarrow$ consistent family of distributions $\mu_{X_n, X_{n+1}, \ldots, X_{n+N-1}}$

IID Sources

X IID $\Leftrightarrow \mu_{X^N} = \mu_{X_0}^N$ = product distribution, $\mu_{X_n} = \mu_{X_0}$ all n

E.g., fair coin flips, biased coin flips, dice throws, IID uniform, IID Gaussian

IID process *most random possible process — no predictability, no sparse representation*

Bernoulli Processes and Shifts

Beware of the name *Bernoulli* —

Information theory: *Bernoulli process* = IID **binary** process with parameter p (coin bias), $p = 1/2$ for fair coin flips emphasized here

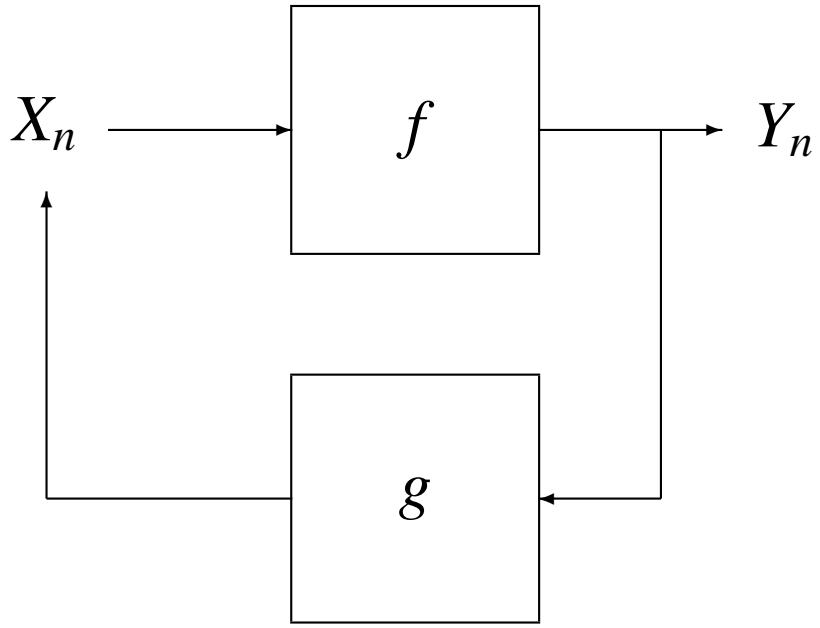
Ergodic theory: *Bernoulli shift* = IID process, discrete or non-discrete alphabet

Warning ☠**:** A minority of the ergodic theory literature uses "Bernoulli shift" differently:
(1) more narrowly — restricting name to finite alphabets (our definition becomes "generalized Bernoulli shift"),
(2) more generally — including any process *isomorphic* to an IID process

isomorphic??

Isomorphism and stationary codes

Two processes $X \sim \mu_X$ and $Y \sim \mu_Y$ are *isomorphic* if there is an *invertible* (with probability 1) stationary coding of μ_X with distribution equal to μ_Y



Can code from one source into the other in an *invertible way*, as in most earlier examples

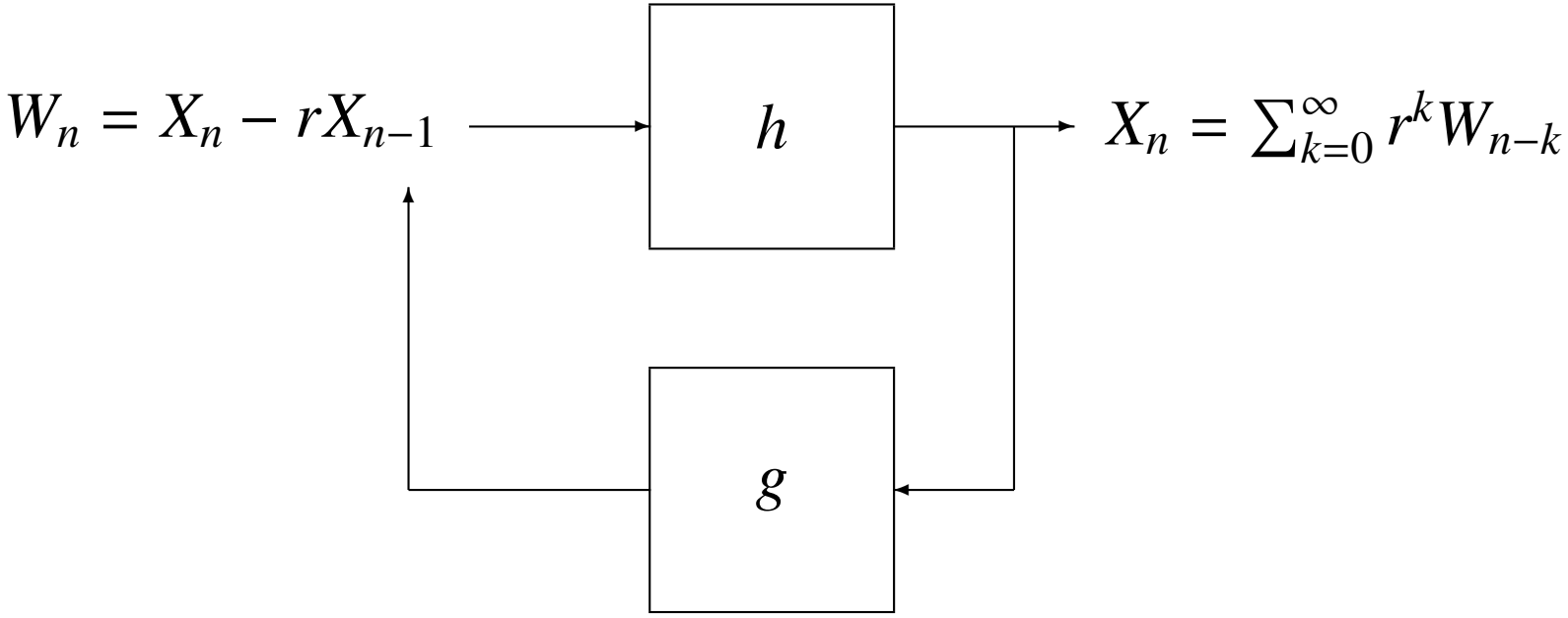
no "information" is lost!

Isomorphism = process/stationary coding analogue of Shannon lossless coding

Unlike Shannon, **well-defined for non-discrete alphabet sources**

E.g., $\{W_n\}$ Gaussian IID process and (stationary) Gauss autoregressive process $\{X_n\}$ are isomorphic, stationary code = invertible LTI filter!

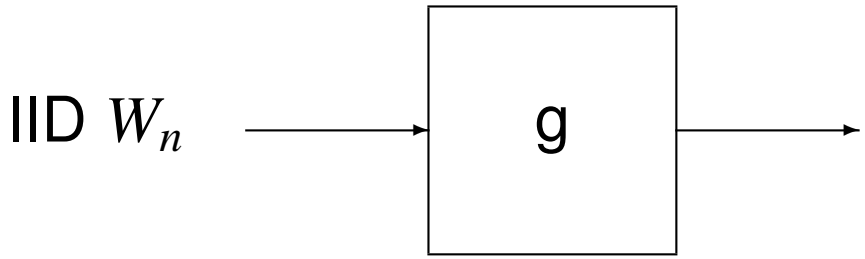
$h_k = r^k; k \geq 0; |r| < 1$ $\qquad$ $g_k = \delta_k - r\delta_{k-1}$



B-processes

Class of stationary codings of coin flips $\subset$ of the class of stationary codings of an IID source such as Z, dice, IID Gaussian

Ornstein's class of *B-processes*



(aka "Bernoulli processes")

Where do B-processes fit in taxonomy of random processes?

A taxonomy of random processes

IID $\subset B \subset K$ (Kolmogorov zero-one law) $\subset$ strongly mixing $\subset$ weakly mixing $\subset$ stationary and ergodic $\subset$ stationary $\subset$ block stationary $\subset$ asymptotically stationary $\subset$ asymptotically mean stationary $\Leftrightarrow$ sample averages converge

Mixing & ergodicity a form of asymptotic independence:

$$\lim_{n\to\infty} \mu(T^{-n}F \cap G) - \mu(F)\mu(G) = 0, \forall F, G: \text{ strong mixing}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \left|\mu(T^{-k}F \cap G) - \mu(F)\mu(G)\right| = 0, \forall F, G: \text{ weak mixing}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \left(\mu(T^{-k}F \cap G) - \mu(F)\mu(G)\right) = 0, \forall F, G: \text{ ergodic}$$

Reminder: Special case of B-processes: positive integer R, $B(R)$ = {all stationary codings of equiprobable IID processes with alphabet of size 2^R} $\subset B$ e.g., $B(1)$

B-processes arguably are the most fundamental for ergodic theory and there are many equivalent characterizations.

IMHO they are also basic to information theory

To sketch these results need two important tools used in both ergodic theory and information theory:

- Shannon entropy + Kolmogorov generalization (*Kolmogorov-Sinai invariant*, extension of Shannon entropy rate to general alphabets, dynamical systems, flows)
- d-bar distance between random processes (and, implicitly, Shannon fidelity criterion)

Entropy: Finite alphabet (Shannon)

Usual information theory treatment

Stationary source X, distribution $\mu_X \Rightarrow$ distributions μ_{X^N} for random vectors X^N, If alphabet A_X is finite, also denote pmf by μ_{X^N}

$$H(X^N) = H(\mu_{X^N}) = -\sum_{x^N} \mu_{X^N}(x^N) \log \mu_{X^N}(x^N)$$

$$H(X) = H(\mu_X) = \inf_N N^{-1} H(X^N) = \lim_{N\to\infty} N^{-1} H(X^N)$$

e.g., for coin flips $N^{-1}H(Z^N) = H(Z) = 1$ bit/symbol

Entropy: General alphabet (Kolmogorov)

Alphabet discrete, continuous, or mixed:

$$\text{vector entropy} \quad H(X^N) = \sup_q \underbrace{H(q(X^N))}_{\text{finite alphabet defn}}$$

supremum over all quantizers (finite output alphabet) q of A_X^N

$$\text{process entropy rate} \quad H(X) = \sup_g \underbrace{H(g(X))}_{\text{finite alphabet defn}}$$

supremum over all stationary codes g with finite output alphabet

Example: $\{X_n\}$ IID, $X_n \sim \mathcal{N}(0,1)$, $N^{-1}H(X^N) = H(X) = \infty$

well defined, but infinite!!

Warning ☠: Shannon differential entropy for continuous distributions is something different and lacks many of the important properties, intuition, and theorems of entropy

Entropy in Ergodic Theory

Entropy plays a fundamental role in ergodic theory (Shannon's idea adopted by Kolmogorov)

Two key results:

Sinai-Ornstein Theorem If μ_X and μ_Y are stationary and ergodic random processes and $H(\mu_X) \geq H(\mu_Y)$, then *there is a stationary coding of* X *with process distribution equal to* μ_Y

Ornstein Isomorphism Theorem A *necessary* condition for two stationary random processes μ_X and μ_Y to be isomorphic is that $H(\mu_X) = H(\mu_Y)$ (Kolmogorov, Sinai). *The condition is* sufficient *if both processes are* B*-processes*.

The class of B-processes is the most general class known for which equal entropy rate ensures isomorphism.

There exist K-processes having equal entropy which are not isomorphic (next most general class of stationary and ergodic processes)

Isomorphism theorem includes discrete and non-discrete alphabet processes and extends to continuous-time processes

Two B-processes can be coded into each other invertably iff they have equal entropy

---

If two stationary and ergodic processes have equal entropy rate, then each can be constructed as a stationary coding of the other, *but there will not be an invertible coding unless both processes are B*

Warning ☠

In general,

$$H(X) \leq \lim_{N\to\infty} N^{-1} H(X^N),$$

Not always equal! (equality if alphabets discrete)

E.g., $X_0 \sim \mathcal{N}(0,1)$, $X_n \equiv X_0$ all n

$$\Rightarrow H(X^N) = \infty \text{ all } N, \text{ but } H(X) = 0$$

Quantization and limit do not always interchange.

Short-term behavior might be misleading regarding long term behavior.

Might hope this is an extreme example, e.g., stationary, but *not ergodic* *— no such luck*

Fake white Gaussian revisited

A stationary coding of fair coin flips in earlier example yielded stationary, ergodic, uncorrelated process with Gaussian marginals

From definition of entropy rate, $H(X) \leq H(Z) = 1$ bit per symbol, so X can not be a stationary uncorrelated Gaussian process (= IID Gaussian process) since IID Gaussian has $H = \infty$

Note: X_n has continuous alphabet, stationary *and* ergodic, *but finite nonzero entropy rate!*

$\Rightarrow$ entropy (rate) distinguishes the fake (less than 1 bit) Gaussian with the correct spectrum and marginals from the real item

Process Distance

Such finite entropy rate processes masquerading as infinite entropy rate processes play a role in Shannon source coding (as will see)

How *good* a fake of μ_X (say IID $\mathcal{N}(0,1)$) is possible using coin flips?

Suppose have a notion of *"distance"* $\overline{d}(\mu_X, \mu_Y)$ between random processes (there are many)

Require at least

- $\overline{d}(\mu_X, \mu_X) = 0$, and
- $\overline{d}(\mu_X, \mu_Y) > 0$ if $\mu_X \neq \mu_Y$

Might also want triangle inequality (or something similar)

Given a class of random processes $\mathcal{G}$ (e.g., $B(R)$) & a stationary and ergodic target source μ_X to be faked by a $\mu_Y \in \mathcal{G}$, "best" fake is closest to target in $\overline{d}$ sense:

$$\overline{d}(\mu_X, \mu_Y) \geq \overline{d}(\mu_X, \mathcal{G}) \equiv \inf_{\mu_Y \in \mathcal{G}} \overline{d}(\mu_X, \mu_Y)$$

If $\mu_Y \in B(R)$, then $H(\mu_Y) \leq R$, so

$$\begin{aligned}
\overline{d}(\mu_X, B(R)) &= \inf_{g: Z_n \to \boxed{g} \to Y_n} \bar{d}(\mu_X, \mu_Y) \\
&\geq \overline{d}(\mu_X, \{B\text{-processes } \mu_Y : H(\mu_Y) \leq R\}) \\
&\geq \overline{d}(\mu_X, \underbrace{\{\text{stationary ergodic } \mu_Y : H(\mu_Y) \leq R\}}_{\mathcal{S}_e(R)})
\end{aligned}$$

are inequalities equalities?

Suppose that μ_Y approximately achieves $\overline{d}(\mu_X, \mathcal{S}_e(R))$:

$$H(\mu_Y) \le R \text{ and } \overline{d}(\mu_X, \mu_Y) \le \overline{d}(\mu_X, \mathcal{S}_e(R)) + \epsilon$$

Then by the Sinai-Ornstein theorem there is a stationary coding of an IID equiprobable source of entropy rate $H(\mu_Y) \le R$ with output distribution μ_Y, thus $\mu_Y \in B(R)$.

Thus for all $\epsilon > 0$, $\quad \overline{d}(\mu_X, \mathcal{S}_e(R)) + \epsilon \ge \overline{d}(\mu_X, \mu_Y) \ge \overline{d}(\mu_X, B(R))$

$$\begin{aligned} \overline{d}(\mu_X, B(R)) &= \overline{d}(\mu_X, \{B\text{-processes } \mu_Y : H(\mu_Y) \le R\}) \\ &= \overline{d}(\mu_X, \{\text{stationary ergodic } \mu_Y : H(\mu_Y) \le R\}) \qquad (\star) \end{aligned}$$

If $H(\mu_X) \le R$, then Sinai-Ornstein $\Rightarrow \overline{d}(\mu_X, B(R)) = 0$!

but what if $H(\mu_X) > R$?

Further questions on best fake

- What is a useful distance measure on random processes for information theory and ergodic theory?
- Can $\overline{d}(\mu_X, B(R))$ be evaluated for the case where $H(\mu_X) > R$?
- Can $\overline{d}(\mu_X, B(R))$ be achieved? I.e., *do optimal codes exist?*

  Is infimum a minimum?

  Already seen the answer is "yes" if $H(\mu_X) \leq R$.
- What are the properties of nearly optimal codes?
- Connections with Shannon rate-distortion/source coding theory? Lossy source code design?

Process Distance

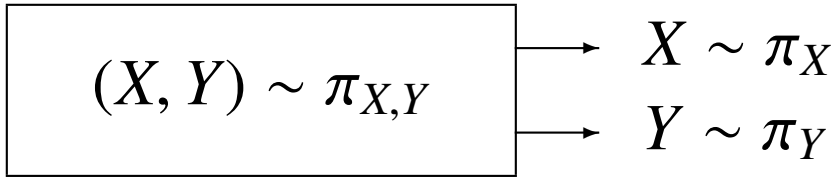
$\exists$ many of distances/metrics on probability distributions

One family particularly useful for ergodic theory and information theory: Monge/Kantorovich/transportation/Vassershtein/Ornstein etc.

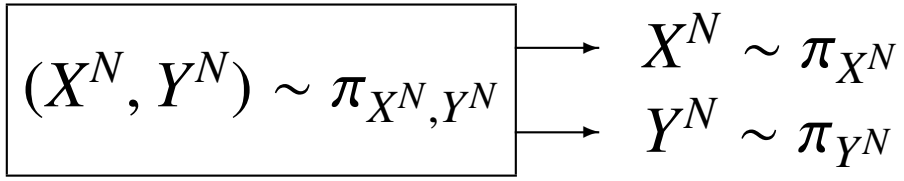
First need some notation

Detour: Pair Processes

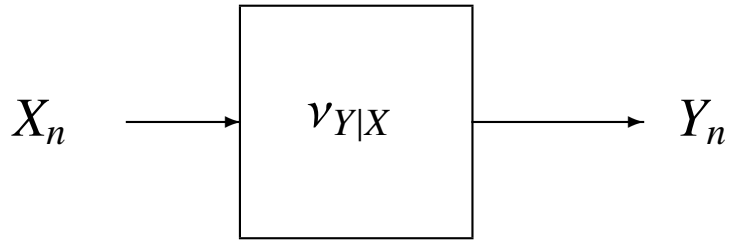
Pair random process $(X, Y) = \{X_n, Y_n\}$ described by joint process distribution $\pi_{XY} \Rightarrow \pi_X, \pi_Y$, marginal distributions



$\Rightarrow$ random vectors (X^N, Y^N) with distributions $\pi_{X^N,Y^N} \Rightarrow \pi_{X^N}, \pi_{Y^N}$, marginal distributions



Example of pair process = input/output of noisy channel, code, communication system



Pair process described by input distribution μ_X and conditional distribution $\nu_{Y|X}$ (deterministic if code)

Detour: Distortion measures and fidelity criteria

Suppose have two alphabets A_X, A_Y.

For simplicity assume $A_X, A_Y \subset \mathbb{R}$

A *fidelity criterion* is a family of *distortion measures* $d_N(x^N, y^N) \geq 0$ on (A_X^N, A_Y^N), $N = 1, 2, \ldots$

(as always, assume sets and functions are measurable wrt suitable sigma-fields)

Assume fidelity criterion *additive* or *single-letter* with per-letter distortion $d_0 = d$:

$$d_N(x^N, y^N) = \sum_{i=0}^{N-1} d(x_i, y_i)$$

By far most important examples are **Hamming distortion** $d(a,b) = 0$ if $a = b$ and 1 otherwise, and **squared error** distortion $d(a,b) = (a-b)^2$

(most everything generalizes to nonnegative powers r of a metric, with $r = 0$ indicating the Hamming distance)

Average Distortion

Given pair process $\pi_{X,Y}$ + fidelity criterion d_N, $N = 1, 2, \ldots$

Average distortion: $D_N(\pi_{X^N,Y^N}) = E_{\pi_{X^N,Y^N}}\left[d_N(X^N, Y^N)\right]$

Limiting average distortion: $D(\pi_{X,Y}) = \lim_{N\to\infty} \frac{1}{N} D_N(\pi_{X^N,Y^N})$

(if limit exists)

If $\pi_{X,Y}$ stationary, fidelity criterion additive

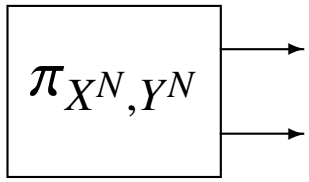
$$D(\pi_{X,Y}) = \frac{1}{N} D_N(\pi_{X^N,Y^N}) = D_1(\pi_{X_0,Y_0}) = E_{\pi_{X_0,Y_0}}\left[d(X_0, Y_0)\right]$$

single-letter characterization

Transportation (Kantorovich) Distance

For vectors:

μ_{X^N}, μ_{Y^N} fixed. *Coupling* π_{X^N,Y^N}



$\pi_{X^N} = \mu_{X^N}$

$\pi_{Y^N} = \mu_{Y^N}$

What is *best* coupling?

$$\mathcal{T}(\mu_{X^N}, \mu_{Y^N}) \equiv \inf_{\pi_{X^N,Y^N} \Rightarrow \mu_{X^N}, \mu_{Y^N}} D_N(\pi_{X^N,Y^N})$$

$\pi_{X^N,Y^N} \Rightarrow \mu_{X^N}, \mu_{Y^N}$ is shorthand for $\pi_{X^N} = \mu_{X^N}$ and $\pi_{Y^N} = \mu_{Y^N}$.

$\Rightarrow$ *transportation distance*

Is transportation "distance" really a distance (metric)?

Transportation distance for powers of a metric

Suppose have underlying metric m on $A_{X^N} \times A_{Y^N}$ and $d_N(x^N, y^N) = m(x^N, y^N)^r$, $r \geq 0$

(If $r = 0$, consider as Hamming distance)

$\mathcal{T}_r$ is transportation " distance"

- If $r \in [0, 1]$, then $\mathcal{T}_r(\mu_{X^N}, \mu_{Y^N})$ is a metric
- If $r \geq 1$, then $\mathcal{T}_r(\mu_{X^N}, \mu_{Y^N})^{1/r}$ is a metric

Summarize: For $r \geq 0$

$\mathcal{T}_r^{\min(1,1/r)}$ is a metric

Monge (1781)/Kantorovich (1942), Vasershtein/Wasserstein (1969), Mallows (1972), "earth mover's" (1998), Rachev and Rüschendorf (1998), Villani (2003, 2009). Villani has > 500 references!

Processes: $$\bar{d}(\mu_X, \mu_Y) \equiv \sup_N N^{-1} \mathcal{T}(\mu_{X^N}, \mu_{Y^N})$$

As with transportation distance on random vectors, if $d(a,b) = m(a,b)^r$, then $\overline{d}^{\min(1,1/r)}$ is a distance on random processes

Ornstein d-bar distance (1970) for average Hamming distance

$$\overline{d}_0 = \overline{d}\text{-distance based on } \mathcal{T}_0$$

Metric space alphabets and squared error (1975)

$$\overline{d}_2 = \overline{d}\text{-distance based on } \mathcal{T}_2$$

A few process distance properties

- If processes μ_X, μ_Y stationary, then

$$\bar{d}(\mu_X, \mu_Y) = \inf_{\pi_{X,Y} \Rightarrow \mu_X, \mu_Y} E_{\pi_{X,Y}} d(X_0, Y_0)$$

  where the infimum is over stationary pair processes.
  If μ_X, μ_Y are also ergodic, then infimum can be restricted to stationary and ergodic pair processes.

- $\bar{d}(\mu_X, \mu_Y)$ = amount by which a μ_X-frequency-typical sequence must be changed in the time-average d sense in order to confuse it with a μ_Y-frequency-typical sequence

- Class of finite-alphabet B-processes = class of all mixing finite-order Markov processes + $\overline{d}_0$-limits

- $H(\mu)$ is continuous in μ with respect to $\overline{d}_0$
- If μ_X and μ_Y are IID, then $\bar{d}(\mu_X, \mu_Y) = \mathcal{T}(\mu_{X_0}, \mu_{Y_0})$
- If squared-error distortion, IID processes

$$\begin{aligned} \overline{d}(\mu_X, \mu_Y) &= \mathcal{T}_2(\mu_{X_0}, \mu_{Y_0}) \\ &= \int_0^1 | F_{X_0}^{-1}(u) - F_{Y_0}^{-1}(u) |^2 \, du \end{aligned}$$

- If Hamming distance, IID discrete-alphabet processes,

$$\overline{d}(\mu_X, \mu_Y) = \mathcal{T}_0(\mu_{X_0}, \mu_{Y_0}) = \frac{1}{2} \sum_{x \in A} | \mu_{X_0}(x) - \mu_{Y_0}(x) |$$

- If μ_X, μ_Y 0 mean stationary with power spectral density

$$S_X(f) = \sum_{k=-\infty}^{\infty} R_X(k) e^{-j2\pi kf} \quad , \quad R_X(k) = E(X_n X_{n-k})$$

then

$$\overline{d}(\mu_X, \mu_Y) \geq \int_{-1/2}^{1/2} |\sqrt{S_X(f)} - \sqrt{S_Y(f)}|^2 df$$

with = if the processes are Gaussian

Part II

Mutual information, Shannon's distortion-rate function, source coding with a fidelity criterion, good fakes and source coding, optimality properties of stationary codes, trellis encoding IID sources

Re-enter Shannon — Mutual Information

Pair process $(X, Y) = \{X_n, Y_n\}$, process distribution $\pi_{X,Y}$

If alphabets discrete

$$I(X^N; Y^N) = I(\pi_{X^N,Y^N}) = H(X^N) + H(Y^N) - H(X^N, Y^N) \leq H(Y^N)$$

In general: $I(X^N; Y^N) = \sup_{\text{quantizers } q} I(q(X)^N; r(Y)^N) \leq H(Y^N)$

Information rate:

— If discrete alphabet $I(X; Y) = I(\pi_{X,Y}) = \lim_{n\to\infty} \frac{1}{N} I(X^N; Y^N) \leq H(Y)$

— In general (Kolmogorov, Dobrushin, Pinsker)

$I(X; Y) = \sup_{\text{quantizers } q,r} I(q(X); r(Y)) \leq H(Y) \Rightarrow$ $\boxed{I(X; Y) \leq H(Y)}$

Distortion-rate function lower bound

Apply to earlier inequality chain:

$$
\begin{aligned}
\overline{d}(\mu_X, B(R)) &= \overline{d}(\mu_X, \{B\text{-processes } \mu_Y : H(\mu_Y) \le R\}) \\
&= \inf_{\text{stationary ergodic } \mu_Y : H(\mu_Y) \le R} \overline{d}(\mu_X, \mu_Y) \\
&= \inf_{\mu_Y : H(\mu_Y) \le R} \left[\inf_{\pi_{X,Y} \Rightarrow \mu_X, \mu_Y} E_{\pi_{X,Y}} d(X_0, Y_0) \right] \\
&= \inf_{\pi_{X,Y} \Rightarrow \mu_X, H(\pi_Y) \le R} E_{\pi_{X,Y}} d(X_0, Y_0) \\
&\ge \inf_{\pi_{X,Y} \Rightarrow \mu_X, I(\pi_{X,Y}) \le R} E_{\pi_{X,Y}} d(X_0, Y_0) \equiv D_X(R) \qquad (\star\star)
\end{aligned}
$$

process definition of Shannon distortion-rate function!

More questions

- Not the traditional Shannon DRF definition. Equivalent? What about dual/inverse Shannon rate-distortion function?
- Shannon DRF/RDF familiar to information theorists as characterization of optimal source coding with a fidelity criterion = Shannon theory of data compression. How relate to current problem?
- *Is the inequality achievable?*

Process vs. traditional Shannon DRF, RDF

$$D_X(R) = \inf_{\pi_{X,Y} \Rightarrow \mu_X, I(\pi_{X,Y}) \le R} E_{\pi_{X,Y}} d(X_0, Y_0)$$

$$R_X(D) = \inf_{\pi_{X,Y} \Rightarrow \mu_X, E_{\pi_{X,Y}} d(X_0,Y_0) \le D} I(\pi_{X,Y})$$

$$D(R) = \inf_N N^{-1} D_N(R) = \lim_{N \to \infty} N^{-1} D_N(R)$$

$$\text{where } D_N(R) = \inf_{\pi^N : \pi^N \Rightarrow \mu_{X^N}, N^{-1} I(\pi^N) \le R} N^{-1} E d_N(X^N, Y^N)$$

$$R(D) = \inf_N N^{-1} R_N(D) = \lim_{N \to \infty} N^{-1} R_N(D)$$

$$\text{where } R_N(D) = \inf_{\pi^N : \pi^N \Rightarrow \mu_{X^N}, N^{-1} d(\pi^N) \le D} N^{-1} I(\pi^N)$$

Theorem Given a stationary and ergodic μ_X, under suitable technical assumptions $D(R) = D_X(R)$, $R(D) = R_X(D)$

Ideas behind proof

$D_X(R) \geq D(R)$: Suppose that π is a process distribution that (approximately) yields the process DRF, i.e., $I(\pi) \leq R$ and $D_X(R) \approx D(\pi)$

Then for large N, the induced vector distributions π^N yield $I(\pi^N) \leq NR$ (almost) and $D(\pi^N) = ND(\pi)$, which with a continuity argument $\Rightarrow$

$$D_X(R) \geq N^{-1} D_N(R) \geq D(R)$$

$\underline{D(R) \geq D_X(R)}$: Choose N large enough so that $N^{-1}D_N(R) \approx D(R)$ and suppose that π^N approximately yields $D_N(R)$, that is, $I(\pi^N) \leq R$ and $D(\pi^N) \approx D_N(R)$

A stationary (and ergodic) pair process π can be constructed using π^N in such a way that the information rate is $\leq R + \epsilon$ and the per-symbol distortion is close to $N^{-1}D_N(R)$, which again with continuity arguments $\Rightarrow N^{-1}D_N(R) \geq D_X(R) \Rightarrow D(R) \geq D_X(R)$

The construction of the process distribution from the block distributions involves using the conditional block distributions most of the time in a conditionally independent manner, but occasionally inserting some random spacing between blocks, which "stationarizes" the pair process

Advantage of traditional definitions

$R_N(D)$ is a convex optimization problem.

$\exists$ tools for analytical optimization of $R_N(D)$ (Gallager, Csiszár) and numerical optimization (Blahut, Rose)

Shannon and other lower bounds, sometimes hold with equality

Useful fact: Given a stationary and ergodic source μ_X, the infimumum in the process DRF

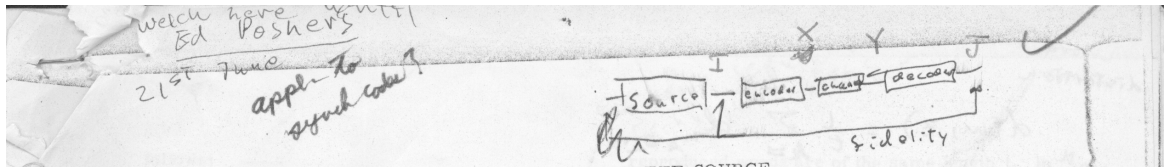
$$D_X(R) = \inf_{\pi_{X,Y} \Rightarrow \mu_X, I(\pi_{X,Y}) \leq R} E_{\pi_{X,Y}} d(X_0, Y_0)$$

is the same whether taken over all stationary and ergodic processes or over all stationary processes

This greatly simplifies proof of block coding theorem for sources with memory,

which brings us at last to **Shannon source coding**, the Shannon theory of data compression — the information theoretic approach to continuous or high entropy rate sources into relatively low entropy rate sources while minimizing distortion.

The classic work is Shannon's 59 paper



CODING THEOREMS FOR A DISCRETE SOURCE WITH A FIDELITY CRITERION*

Claude E. Shannon

Departments of Mathematics and Electrical Engineering
and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Summary

Consider a discrete source producing a sequence of message letters from a finite alphabet. A single letter distortion measure is given by a non-negative matrix (d_{ij}). The entry d_{ij} measures the "cost" or "distortion" if letter i is reproduced at the receiver as letter j. The average distortion of a communications system (source-coder-noisy channel-decoder) is taken to be $d = \sum_{i,j} P_{ij} d_{ij}$ where P_{ij} is the probability of i being reproduced as j. It is shown that there is a function R(d) that measures the "equivalent rate" of the source for a given level of distortion. For coding purposes where a level d of distortion can be tolerated, the source acts like one with information rate R(d). Methods are given for calculating R(d), and various properties discussed. Finally, generalizations to ergodic sources, to continuous sources, and to distortion measures involving blocks of letters are developed.

In this paper a study is made of the problem of coding a discrete source of information, given a fidelity criterion or a measure of the distortion of the final recovered message at the receiving point relative to the actual transmitted message. In a particular case there might be a certain tolerable level of distortion as determined by this measure. It is desired to so encode the information that the maximum possible signaling rate is obtained without exceeding the tolerable distortion level. This work is an expansion and detailed elaboration of ideas presented earlier[1], with particular reference to the discrete case.

We shall show that for a wide class of distortion measures and discrete sources of information there exists a function R(d) (depending on the particular distortion measure and source) which measures, in a sense, the equivalent rate R of the source (in bits per letter produced) when d is the allowed distortion level. Methods will be given for evaluating R(d) explicitly in certain simple cases and for evaluating R(d) by a limiting process in more complex cases. The basic results are roughly that it is impossible to signal at a rate faster than C/R(d) (source letters per second) over a memoryless channel of capacity C (bits per second) with a distortion measure less than or equal to d. On the other hand, by sufficiently long block codes it is possible to approach as closely as desired the rate C/R(d) with distortion level d.

Finally, some particular examples, using error probability per letter of message and other simple distortion measures, are worked out in detail.

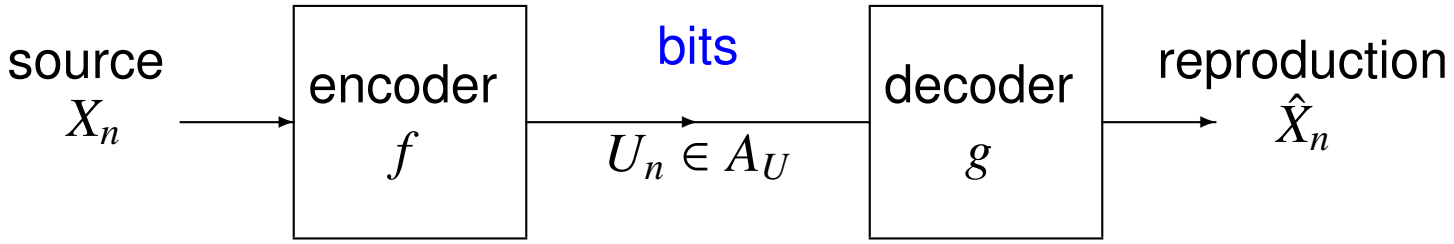
The Single-Letter Distortion Measure

Suppose that we have a discrete information source producing a sequence of letters or "word" $m = m_1, m_2, m_3, \ldots, m_t$, each chosen from a finite alphabet. These are to be transmitted over a channel and reproduced, at least approximately, at a receiving point. Let the reproduced word be $Z = z_1, z_2, \ldots, z_t$. The z_i letters may be from the same alphabet as the m_i letters or from an enlarged alphabet including, perhaps, special symbols for unknown or semi-unknown letters. In a noisy telegraph situation m and Z might be as

*This work was supported in part by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S. Navy (Office of Naval Research).

Source Coding

Given source X, fidelity criterion:



$\pi_{f,g}$ induced distribution of $(X, \hat{X})$

Average distortion $D_{\mu_X}(f,g) \equiv D(\pi_{f,g}) = E_{\pi_{f,g}} d(X_0, Y_0)$

Optimize: For a given class of codes C what is *best* possible code performance? $\delta_X(R) \equiv \inf_{f,g \in C: \log |A_U| \le R} D_{\mu_X}(f,g)$

Operational DRF

Shannon and most everybody since considers *block codes*

Emphasis here is stationary codes — compare the 2 structures

Block vs. Sliding-block

Block coding

- Far more known about design: e.g., transform codes, vector quantization, optimality properties, clustering 👍
- Does not preserve key properties (stationarity, ergodicity, mixing, 0-1 law, B) 👎

  In general output neither stationary nor ergodic (it is N-stationary and can have a periodic structure, not necessarily N-ergodic).

  Can "stationarize" with uniform random start, but retains possible periodicities. Not equivalent to stationary coding of input.
- Not defined for infinite block length, no limiting codes as blocklength grows. 👎

Stationary coding

- preserves key properties of input process: stationarity, ergodicity, mixing, B, 0-1 law 👍
- well-defined for $N = \infty$. Infinite codes can be approximated by finite codes. Sequence of finite codes can converge 👍
- models many communication and signal processing techinques: time-invariant convolutional codes, predictive quantization, Δ-modulation, $\Sigma\Delta$-modulation, nonlinear and linear time-invariant filtering, wavelet coefficient evaluation by LTI filters
- used to prove key results in ergodic theory (Ornstein isomorphism theorem, Sinai-Ornstein theorem)

There are constructions in ergodic theory and information theory to get stationary codes from block codes & vice-versa

Source Coding Theorem

X stationary and ergodic, additive (or subadditive) fidelity criterion with reverence letter, i.e., $\exists\, a^* \in A_X$ for which $E[d(X_0, a^*)] < \infty$, then for block codes and for stationary codes

$$\boxed{\delta_X(R) = D_X(R)}$$

Positive coding theorem is hard — Block coding theorem uses traditional Shannon random coding argument

Stationary coding uses positive block coding theorem to get good blocks, embed in sliding-block code structure using "stationarization" — code long sequences of blocks with occasional spacing based on past source information

No shortcuts using stationary codes

Converse coding theorem is simple for stationary codes

Similar to Shannon DRF lower bound for $\overline{d}(\mu_X, B(R))$

- Cascade of stationary encoder and decoder is a stationary code
- Channel process between encoder and decoder is a 2^R-ary alphabet process with entropy $\leq R \Rightarrow$

$$\begin{aligned}
\delta_X(R) &= \inf_{f,g \in C: \log|A_U| \leq R} D_{\mu_X}(f,g) \\
&\geq \inf_{\pi_{X,Y} \Rightarrow \mu_X, H(\mu_Y) \leq R} D(\pi_{X,Y}) = \underbrace{\inf_{\mu_Y: H(\mu_Y) \leq R} \overline{d}(\mu_X, \mu_Y)}_{= \overline{d}(\mu_X . B(R)) \text{ from } (\star\star)} \\
&\geq \inf_{\pi_{X,Y} \Rightarrow \mu_X, I(\pi_{X,Y}) \leq R} D(\pi_{X,Y}) = D_X(R)
\end{aligned}$$

In particular,

$$\delta_X(R) \geq \overline{d}(\mu_X, B(R)) \geq D_X(R)$$

So positive coding theorem $\Rightarrow$

$$\delta_X(R) = \overline{d}(\mu_X, B(R)) = \inf_{\mu_Y \in B(R)} \overline{d}(\mu_X, \mu_Y) = D_X(R)$$

$\overline{d}(\mu_X, B(R))$ "sandwiched" between equal quantities

$\Rightarrow$ Shannon DRF solves $\overline{d}(\mu_X, B(R))$ evaluation problem and provides geometric interpretation of source coding

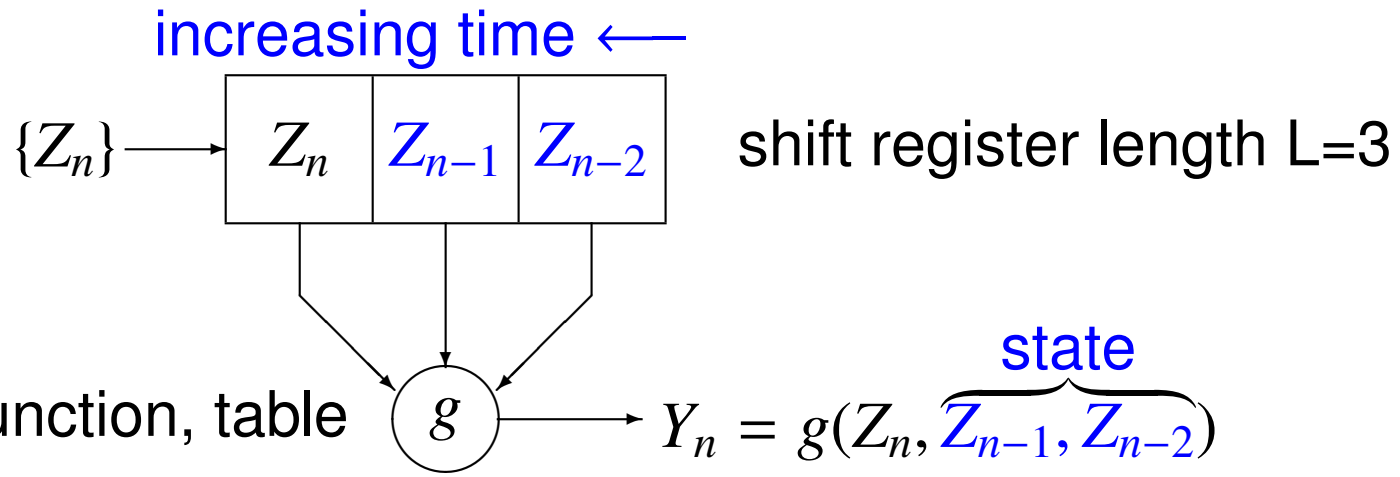
- *Is there an intuition behind this?*

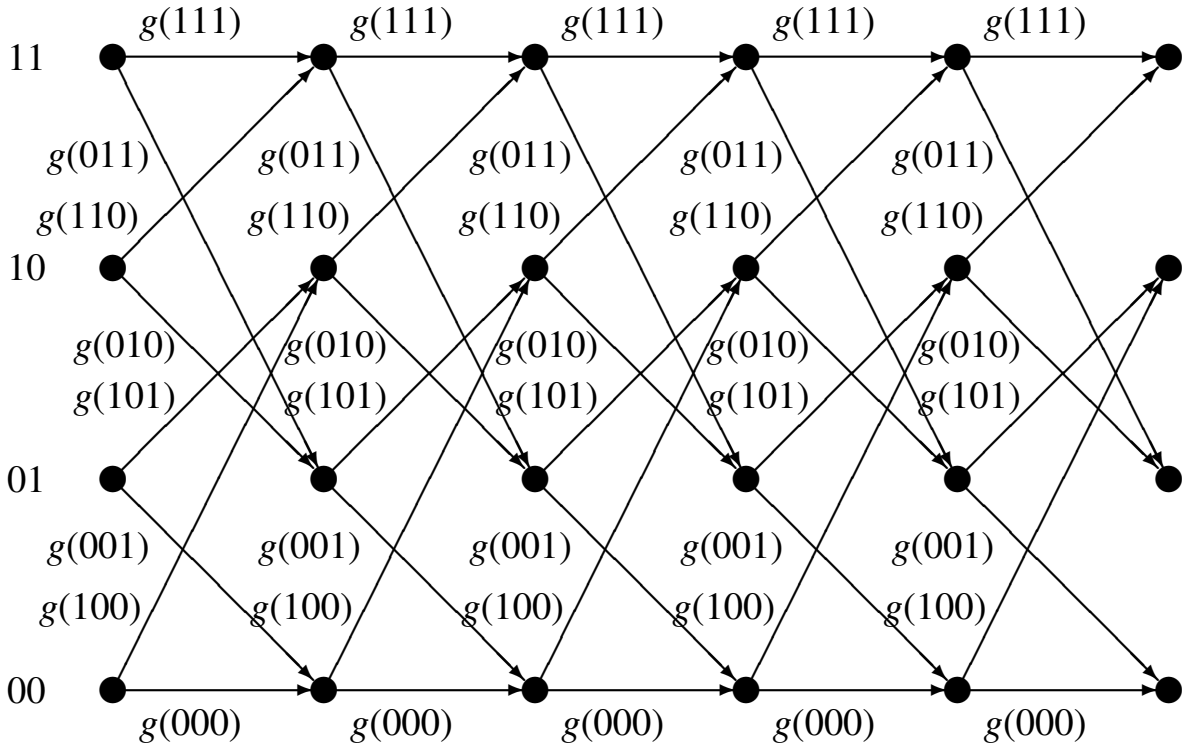
Yes – can use a good stationary fake of a process X to design a good source code for X

Fakes and good source codes

Suppose have good finite length stationary code g of coinflips, i.e., yields B-process Y_n with $\overline{d}(\mu_X, \mu_Y) \approx D_X(R)$

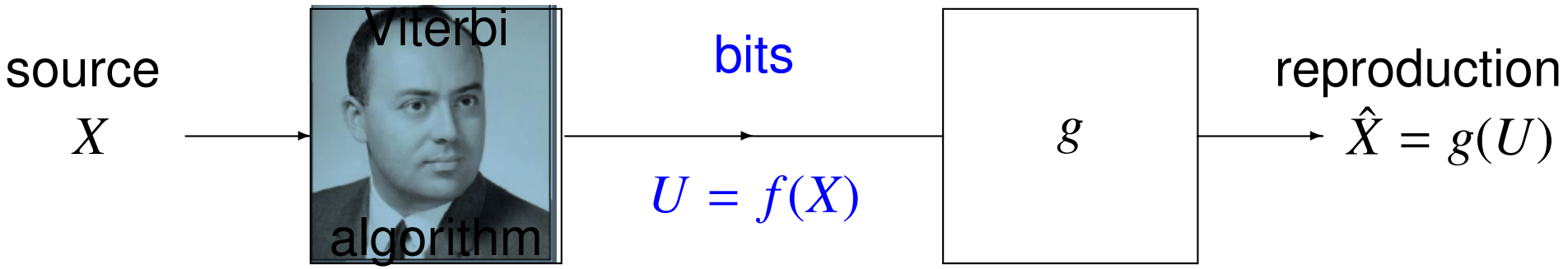
$\Rightarrow$ possible reproduction sequences can be depicted on a trellis diagram:

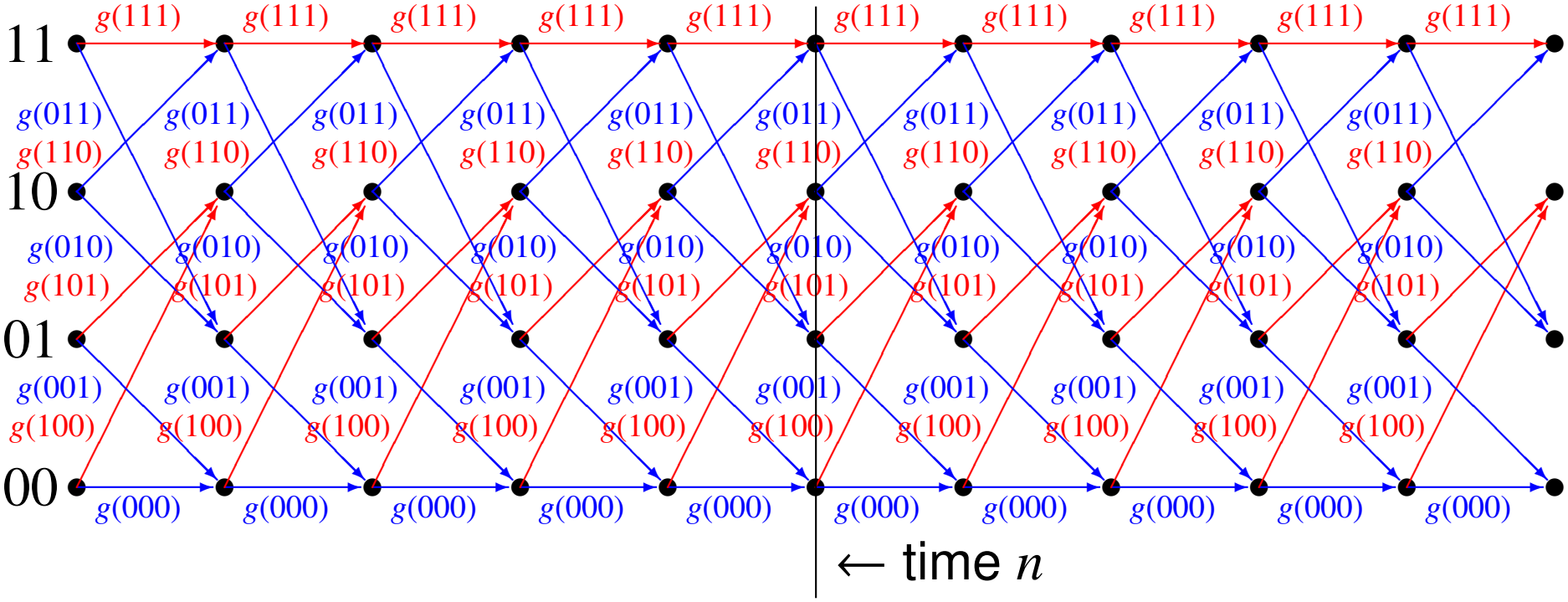




Given a long input sequence, an encoder can find *minimum distortion path through trellis* using the Viterbi algorithm (**dynamic programming** search of minimum distortion path through a directed graph)

Trellis encoder





If f=Viterbi algorithm encoder (or sliding-block approximation), then

$$D_{\mu_X}(f, g) \approx \overline{d}(\mu_X, B(R)) = D_X(R)$$

Most natural use is as a hybrid code — block Viterbi algorithm matched to stationary source decoder = nearly optimal simulator, but can stationarize encoder to match theory

Underlying theory assumes $R < H(\mu_X) \Rightarrow$ not enough bits to get $D_{\mu_X}(f, g) = 0$ and isomorphism.

Source coding can be viewed as a "sloppy" isomorphism — do not have invertible code, doomed to distortion $\geq D_X(R)$ if encoding into R bits per symbol

Good news is can do nearly this well!

but will see can not actually achieve $D_X(R)$

How find a good source decoder/rate-constrained simulator g?

Look at properties of good codes

Shannon Optimal Reproduction Distribution

- For any random vector X^N, Shannon RDF is an information theoretic (convex) optimization
- For IID source, $R_X(D) = R_{X_0}(D)$ = infimum of $I(\pi)$ over joint distributions π on $\mathbb{R}^2$ with input marginal μ_X and $d(\pi) \leq R$. If optimizing π exists, resulting μ_{Y_0} is a *Shannon optimal reproduction distribution*

  For IID source, $\Rightarrow N$-dimensional Shannon optimal reproduction distribution = $\mu_{Y^N} = \mu_{Y_0}^N$, product distribution

  For IID source, optimal process distribution is IID with marginal distribution μ_{Y_0}, the distribution yielding the Shannon RDF

- Csiszár (1974): For random vectors X (abbreviate X^N)
  - There exists a distribution π achieving finite-order $R(D)$ (under more general conditions than those considered here)
  - If a sequence of joint distributions $\pi^{(n)}$, $n = 1, 2, \ldots$ with marginals μ_X and $\mu_Y^{(n)}$ satisfy

$$\begin{aligned} I(\pi^{(n)}) &= I(X, Y^{(n)}) \leq R, n = 1, 2, \ldots \\ \lim_{n\to\infty} E_{\pi^{(n)}}[d(X, Y^{(n)})] &= D_X(R) \end{aligned}$$

    then $\mu_Y^{(n)}$ has a subsequence that converges to a Shannon optimal reproduction distribution weakly and in squared error transportation distance.
  - If the Shannon optimal distribution is unique (e.g., it is Gaussian) , then $\mu_Y^{(n)}$ converges to it.

Finding a Shannon optimal reproduction distribution

- For squared-error: Shannon optimal reproduction alphabet is *finite* unless Shannon lower bound holds with equality, e.g., Gaussian [Fix (1977), Rose (1994)]
- Shannon optimal reproduction distribution for Gaussian $\mathcal{N}(0,1)$, squared-error $\Rightarrow$ another Gaussian $\mathcal{N}(0,0.75)$ ($R=1$, SLB holds)
- Rose's algorithm: mapping approach $\Rightarrow$ alternative to Blahut algorithm, gives better results for continuous alphabets when SLB does not hold
- Shannon optimal reproduction distribution for Uniform $(0,1)$, $R=1$, is a discrete distribution with alphabet size 3, pmf:

y	0.2	0.5	0.8
$p_Y(y)$	0.368	0.264	0.368

Asymptotically Optimal Codes

- Codes f_n, g_n, $n = 1, 2, \ldots$ are *asymptotically optimal (a.o.)* if

  Source coding: $\lim_{n\to\infty} D_{\mu_X}(f_n, g_n) = \delta_X(R) = D_X(R)$

  Simulation/fake process: $\lim_{n\to\infty} \overline{d}(\mu_X, \mu_{\overline{g}_n(Z)}) = \overline{d}(\mu_X, B(R))$

- An optimal code (f, g) (if it exists) is trivially asymptotically optimal — set $(f_n, g_n) = (f, g)$ all n. If (f, g) optimal, then necessarily
Source coding: $D_{\mu_X}(f, g) = \delta_X(R) = D_X(R)$ Simulation:
$\overline{d}(\mu_X, \mu_{\overline{g}(Z)}) = \overline{d}(\mu_X, B(R))$

Focus on source coding, similar properties for fake process problem

A design approach

• Code design idea: Find necessary conditions for a.o. codes, use as guidelines for code design. Has much in common with historical methods for block codes such as **Lloyd clustering**

Detour: Review Lloyd optimality conditions for block source codes (vector quantizers). Here fix blocklength N. "Optimum" here means equals operational DRF for blocklength N.

Note: This block code optimality stuff will NOT be covered if I am running behind.

Lloyd Optimality Conditions for Block Codes

Steinhaus (1956) for squared error, vectors, Lloyd (1957) for random variables, general distortion (easily generalized to vectors)

Rediscovered many times, e.g., k-means (1967), principal points (1990), alternating optimization (2002)

Abbreviate notation by dropping the superscript N: X is the N-D random vector and $\mathcal{E}$ and $\mathcal{D}$ denote a blocklength N encoder and decoder, respectively.

Lloyd Quantizer Optimality Properties An optimal quantizer must satisfy the following two conditions

Optimum encoder for a given decoder Given a decoder with reproduction codebook $C = \{\hat{x}_i; i = 1, 2, \ldots, 2^{NR}\}$, the optimal encoder satisfies

$$\mathcal{E}(x) = \operatorname*{argmin}_{i \in \mathbb{I}} \rho(x, \hat{x}_i).$$

minimum distortion encoder is optimal

Optimum decoder for a given encoder Given an encoder $\mathcal{E}$, the optimal decoder is the generalized centroid

$$\mathcal{D}(i) = \operatorname*{argmin}_{y \in \hat{A}} E[d(X, y) \mid X \in \{x : \mathcal{E}(x) = i\}]$$

centroid decoder is optimal

Application to an empirical distribution (a training or learning set) yields an iterative codebook improvement algorithm, an early clustering/learning algorithm! **Lloyd algorithm**

Back to stationary codes, where “optimal” means close to the Shannon DRF —

Necessary condition 1: Process Approximation

- $f_n, g_n, n = 1, 2, \ldots$: asymptotically optimal sequence of stationary source codes of a stationary ergodic source $\{X_n\}$
- $U^{(n)}$: encoder output/decoder input process with alphabet of size 2^R for integer rate R,
- $\hat{X}^{(n)}$: the resulting reproduction processes.

Then

$$\begin{aligned} \lim_{n\to\infty} \overline{d}(\mu_X, \mu_{\hat{X}^{(n)}}) &= D_X(R) \\ \lim_{n\to\infty} H(\hat{X}^{(n)}) &= \lim_{n\to\infty} H(U^{(n)}) = R \\ \lim_{n\to\infty} \bar{d}_0(U^{(n)}, Z) &= 0 \end{aligned}$$

If $R = 1$, encoded process $U^{(n)} \to$ fair coin flips!

Necessary condition 2: Moment Conditions

Resemble block code moment conditions, $\epsilon_0^{(n)} = \hat{X}_0^{(n)} - X_0$

$$\lim_{n\to\infty} E(\hat{X}_0^{(n)}) = E(X_0)$$

$$\lim_{n\to\infty} \frac{\mathrm{COV}(X_0, \hat{X}_0^{(n)})}{\sigma^2_{\hat{X}_0^{(n)}}} = 1$$

$$\lim_{n\to\infty} \sigma^2_{\hat{X}_0^{(n)}} = \sigma^2_{X_0} - D_X(R), \text{ or}$$

$$\lim_{n\to\infty} E(\epsilon_0^{(n)}) = 0$$

$$\lim_{n\to\infty} E(\epsilon_0^{(n)} \hat{X}_0^{(n)}) = 0$$

$$\lim_{n\to\infty} \sigma^2_{\epsilon_0^{(n)}} = D_X(R)$$

Necessary Condition 3: Marginal distribution Shannon condition for IID processes

If X is IID, then

- A subsequence of the marginal distribution of the reproduction process, $\mu_{\hat{X}_0^{(n)}}$ converges weakly and in squared error transportation distance to a Shannon optimal reproduction distribution.
- If the Shannon optimal reproduction distribution is unique, then $\mu_{\hat{X}_0^{(n)}}$ converges to it.
- *If a code is optimal, then* $\mu_{\hat{X}_0} =$ *Shannon optimal distribution*

Necessary Condition 4: Finite-dimensional distributions Shannon condition

If X is IID,

- then a subsequence of the N-dimensional reproduction distribution $\mu_{(\hat{X}^{(n)})^N}$ converges weakly and in $\mathcal{T}_2$ to the N-fold product of a Shannon optimal marginal distribution
- If the one-dimensional Shannon optimal distribution is unique, then $\mu_{(\hat{X}^{(n)})^N}$ converges weakly and in $\mathcal{T}_2$ to its N-fold Shannon optimal product distribution
- If a code (f, g) is optimal, then $\mu_{\hat{X}^N}$ = the N-fold product of a Shannon optimal marginal distribution

If code optimal, then Condition 4 $\Rightarrow \hat{X}$ is also IID with the Shannon optimal marginal reproduction marginal. This yields a contradiction since $H(\hat{X}) \leq R < H(Y) = \infty$

optimal codes do not exist for the IID Gaussian source with squared-error !!

Asymptotically Uncorrelated Condition

Covariance function of $\hat{X}^{(n)}$: $K_{\hat{X}^{(n)}}(k) = \text{COV}(\hat{X}_i^{(n)}, \hat{X}_{i-k}^{(n)})$ $\forall$ integer k.

Given: IID process X with distribution μ_X, f_n, g_n an a.o. sequence of stationary source encoder/decoder pairs with common alphabet of size 2^R For all $k \neq 0$,

$$\lim_{n\to\infty} K_{\hat{X}^{(n)}}(k) = 0$$

and hence the reproduction processes are asymptotically uncorrelated.

Do optimal codes exist?

Already seen answer is *no* for Gaussian IID, $R = 1$

On the other hand, suppose (f, g) is an optimal source code for a source μ_X where μ_X is a B-process with $H(X) \leq R = 1$. Then $\exists$ invertible stationary mapping into an IID binary process $\Rightarrow$ code has zero distortion and entropy rate of binary channel process and reproduction process of $H(X) \leq 1$.

If μ_X is an IID process and $\infty > H(X) = H(X_0) > R = 1$, then from the optimality properties an optimal code must have $\mu_{\hat{X}^N} = \pi_{Y^N} = \pi_{Y_0}^N$ for all N.

Linder has shown using Csiszár's results that still $H(Y_0) > 1 = H(X_0)$ # and hence no optimal code exists, as in the Gaussian IID case

There is a discontinuity between the 0 distortion result (Ornstein isomorphism theorem) and the nonzero distortion result (Shannon rate-distortion theorem) — *optimal stationary codes exist for the former if the source is B, but not for the latter if the source is IID*

You can get as close as you like to $D_X(R)$, but you can never achieve it for source coding or faking

A Code Design Algorithm

Recall question: *How find a good simulator/decoder g?*

One approach: Find g which at least satisfies necessary conditions for approximate optimality. (provably or numerically)

Consider a sliding-block code g_L of length L of an equiprobable binary IID process Z which produces an output process $\tilde{X}$ defined by

$$\tilde{X}_n = g_L(\underbrace{Z_n, Z_{n-1}, \cdots, Z_{n-L+1}}_{\text{binary } L\text{-tuple}})$$

In the trellis setting, $(Z_{n-1}, \cdots, Z_{n-L+1})$ is the state, Z_n is the next input bit.

Suppose that the ideal distribution for $\tilde{X}_n$ is given by a CDF F_{Y_0} of the Shannon optimal marginal reproduction distribution.

Given binary L-tuple $u^L = (u_0, u_1, \ldots, u_{L-1})$,

define

$$b(u^L) = \sum_{i=0}^{L-1} u_i 2^{-i-1} + 2^{-L-1} \in (0,1). \tag{1}$$

Let

$$\tilde{X}_n = g(Z_n, Z_{n-1}, \cdots, Z_{n-L+1}) = F_{Y_0}^{-1}(b(Z_n, Z_{n-1}, \cdots, Z_{n-L+1})).$$

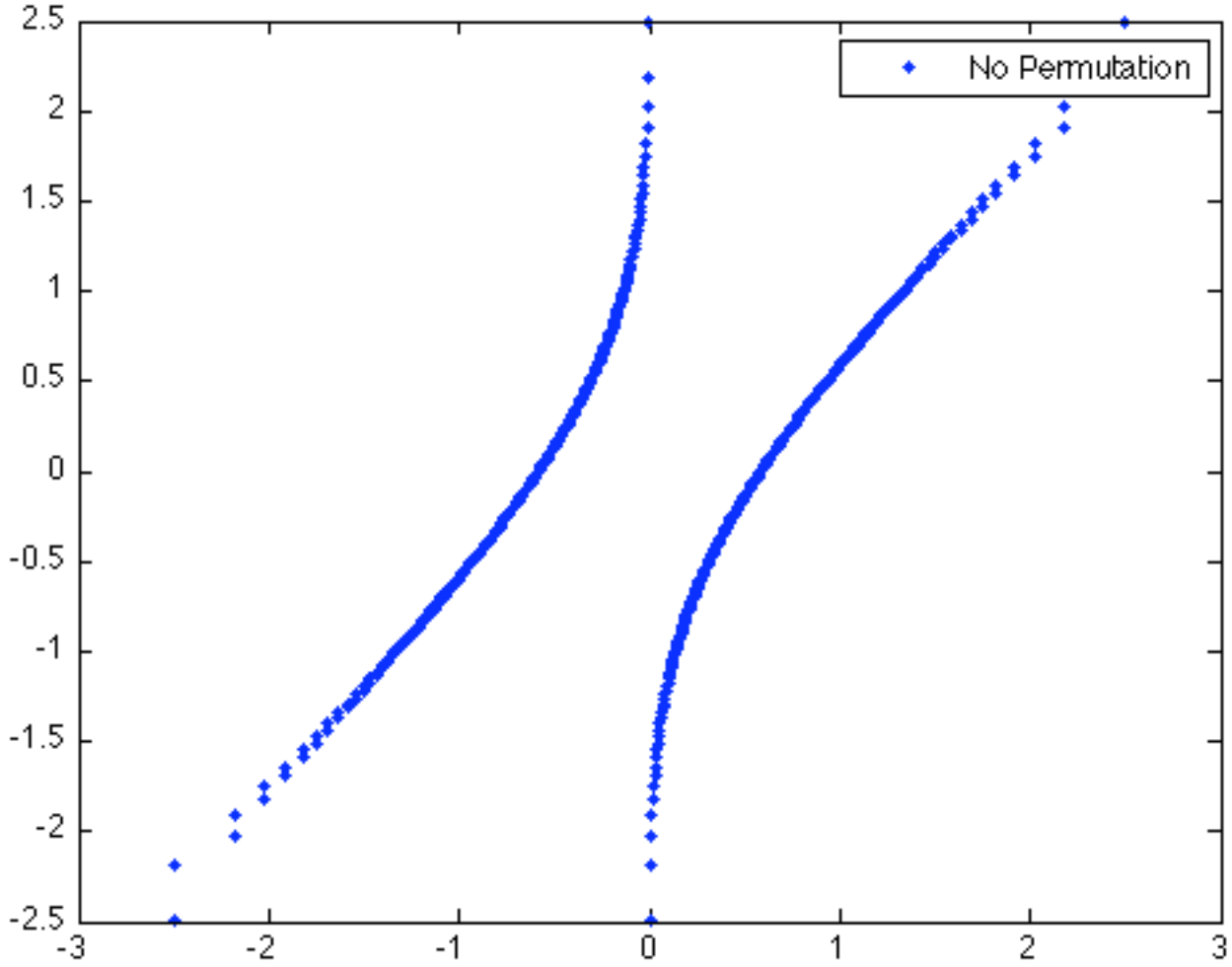
As $L \to \infty$, $b(Z_n, Z_{n-1}, \cdots, Z_{n-L+1})$ converges weakly to uniform on $(0,1)$, and

$g_L(Z_n, Z_{n-1}, \cdots, Z_{n-L+1})$ converges weakly to the (1D) Shannon optimal marginal distribution.

Works for both continuous and discrete Shannon optimal marginal distribution.

g satisfies the 1D moment conditions and the 1D Shannon marginal distribution condition.

Problem: Only matches Shannon marginal distribution, successive outputs on the trellis are highly correlated (reminder plot on next page), poor fake of IID process



Scatter plot of successive samples.

Random permutation

A possible solution: Randomly generate a permutation on binary L-tuples: $\mathcal{P} : \{0,1\}^L \to \{0,1\}^L$ then set

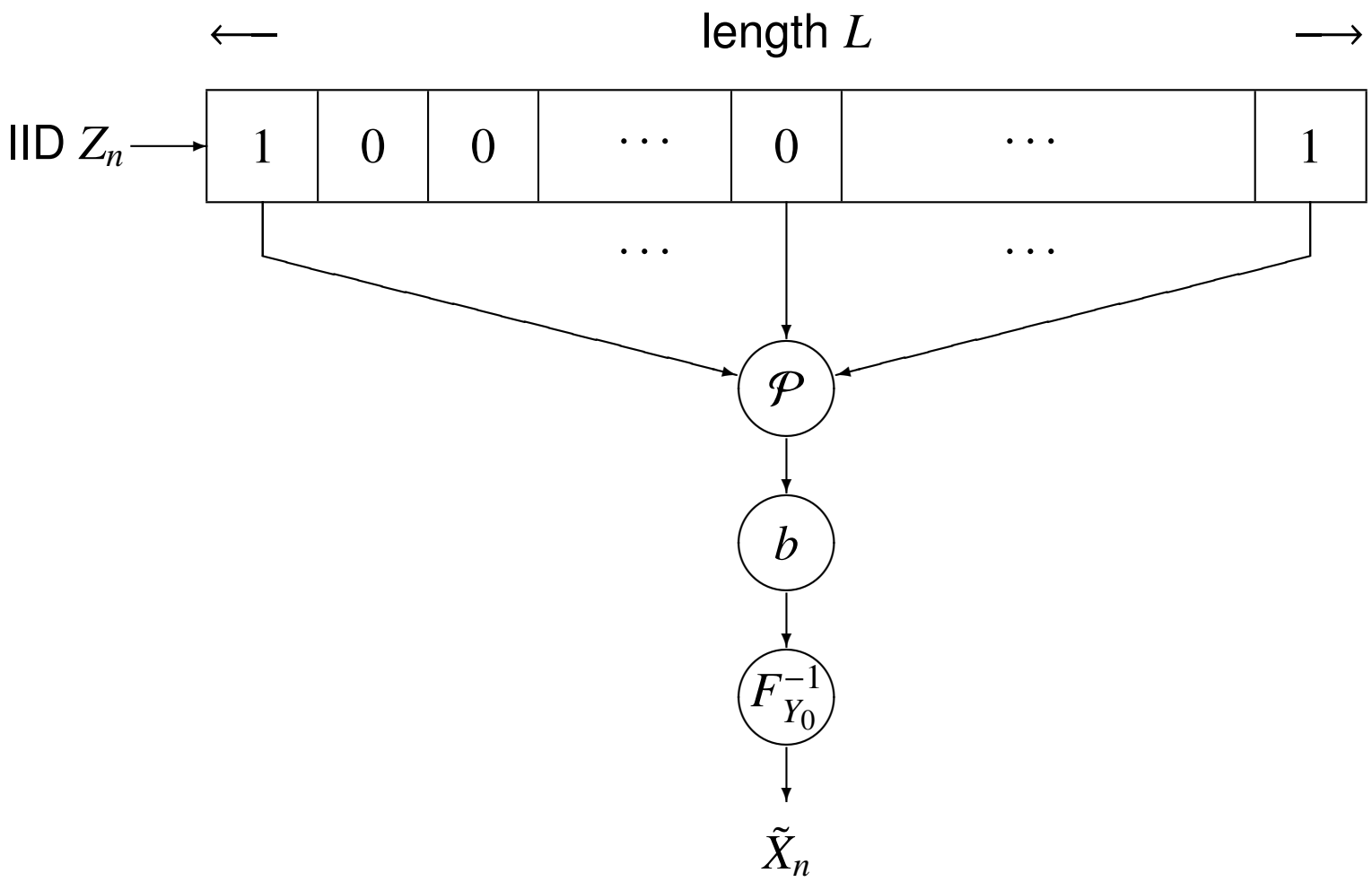
$$g(u^L) = F_{Y_0}^{-1}(b(\mathcal{P}(u^L))$$

Permutation is then *fixed for all time* and used repeatedly.

Now

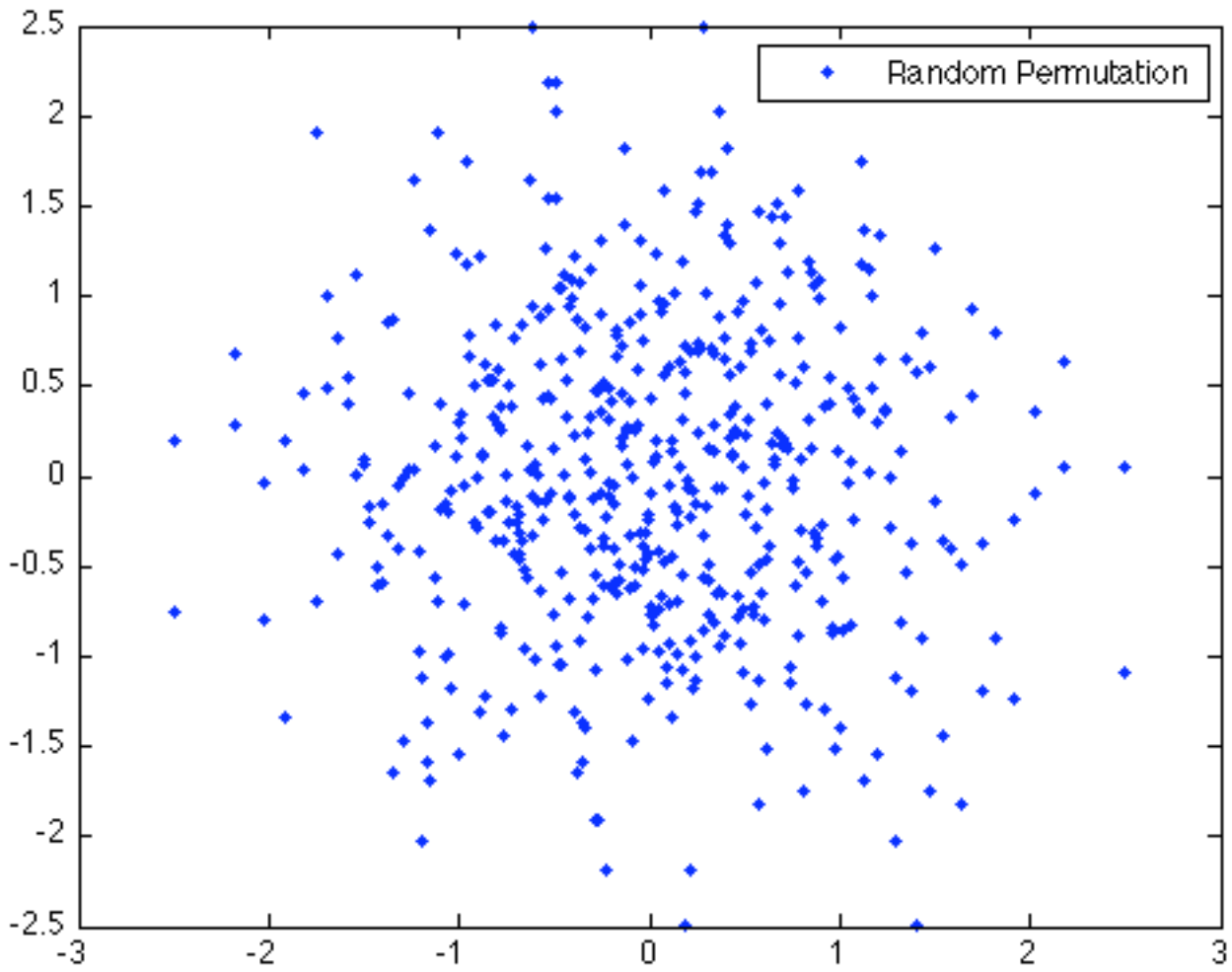
$$g(Z_n, Z_{n-1}, \cdots, Z_{n-L+1}) = F_{Y_0}^{-1}(b(\mathcal{P}(Z_n, Z_{n-1}, \cdots, Z_{n-L+1})))$$

Simulation coder/source decoder becomes

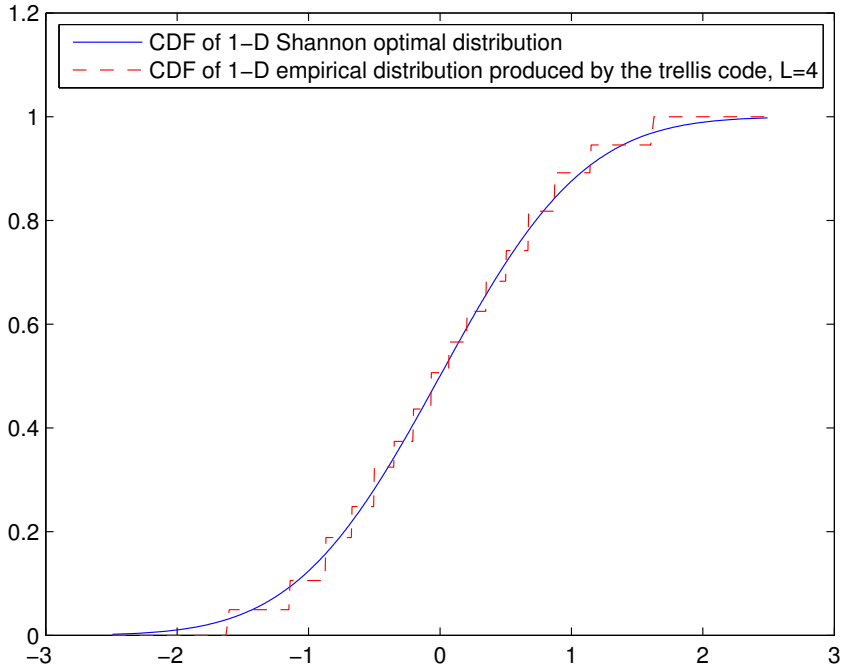


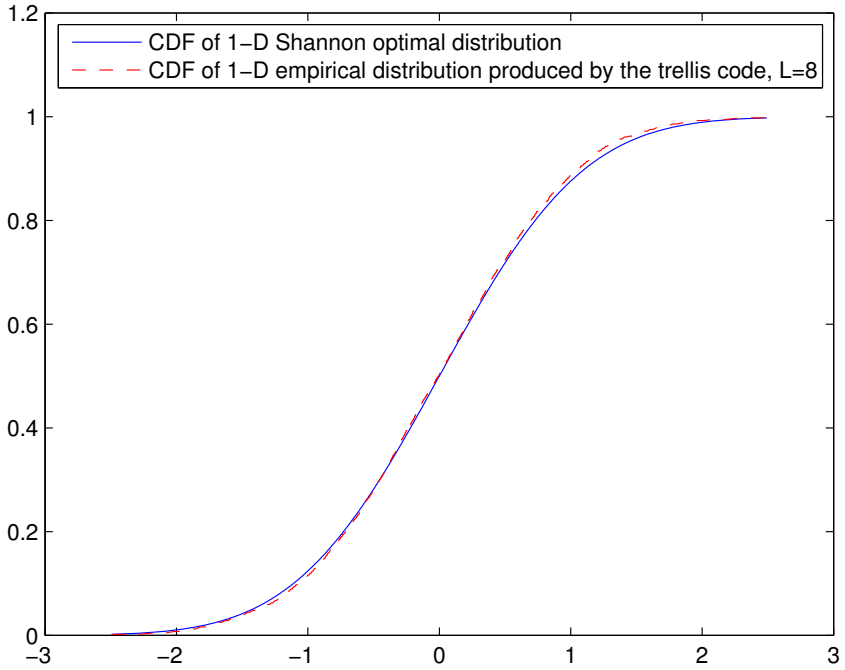
Coupled with Viterbi algorithm yields a source code.

Driven by fair coin flips yields white Gaussian fake



Scatter plot with permutation





CDF of the 1-D empirical reproduction distributions

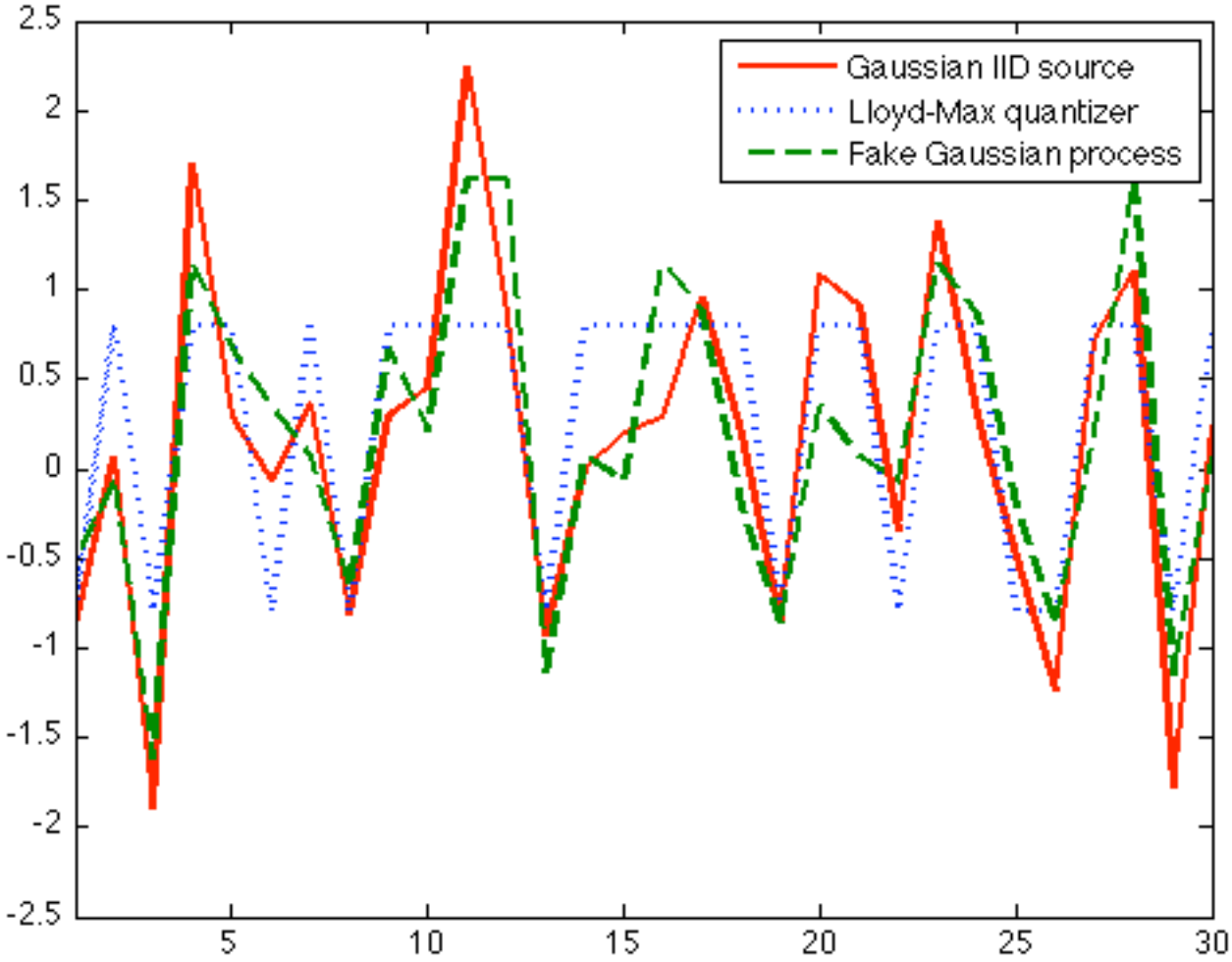
Empirically: Spectrum $\approx$ flat, H(reproduction process) $\approx$ H(binary sequence) $\approx 1 \Rightarrow$ close to fair coin flips in $\overline{d}_0$ (using Marton's inequality (1996) relating $\overline{d}_0$ to limiting relative entropy/Kullback-Leibler rate)

Performance in source coding/compression:

Numeric Results: IID Gaussian

	Rate(bits)	MSE	SNR(dB)
RP_8	1	0.2989	5.24
RP_9	1	0.2913	5.36
RP_10	1	0.2835	5.47
RP_12	1	0.2740	5.62
RP_16	1	0.2638	5.79
RP_20	1	0.2582	5.88
RP_24	1	0.2557	5.92
RP_28	1	0.2542	5.95
$D_X(R)$	1	0.25	6.02
TCQ9	1	0.3105	5.08
TCQ(opt)_9	1	0.2780	5.56
Pearlman_10	1	0.292	5.35
Stewart_10	1	0.293	5.33
Linde/Gray_9	1	0.31	5.09
LC_10	1	0.2698	5.69
LC(opt)_10	1	0.2673	5.73

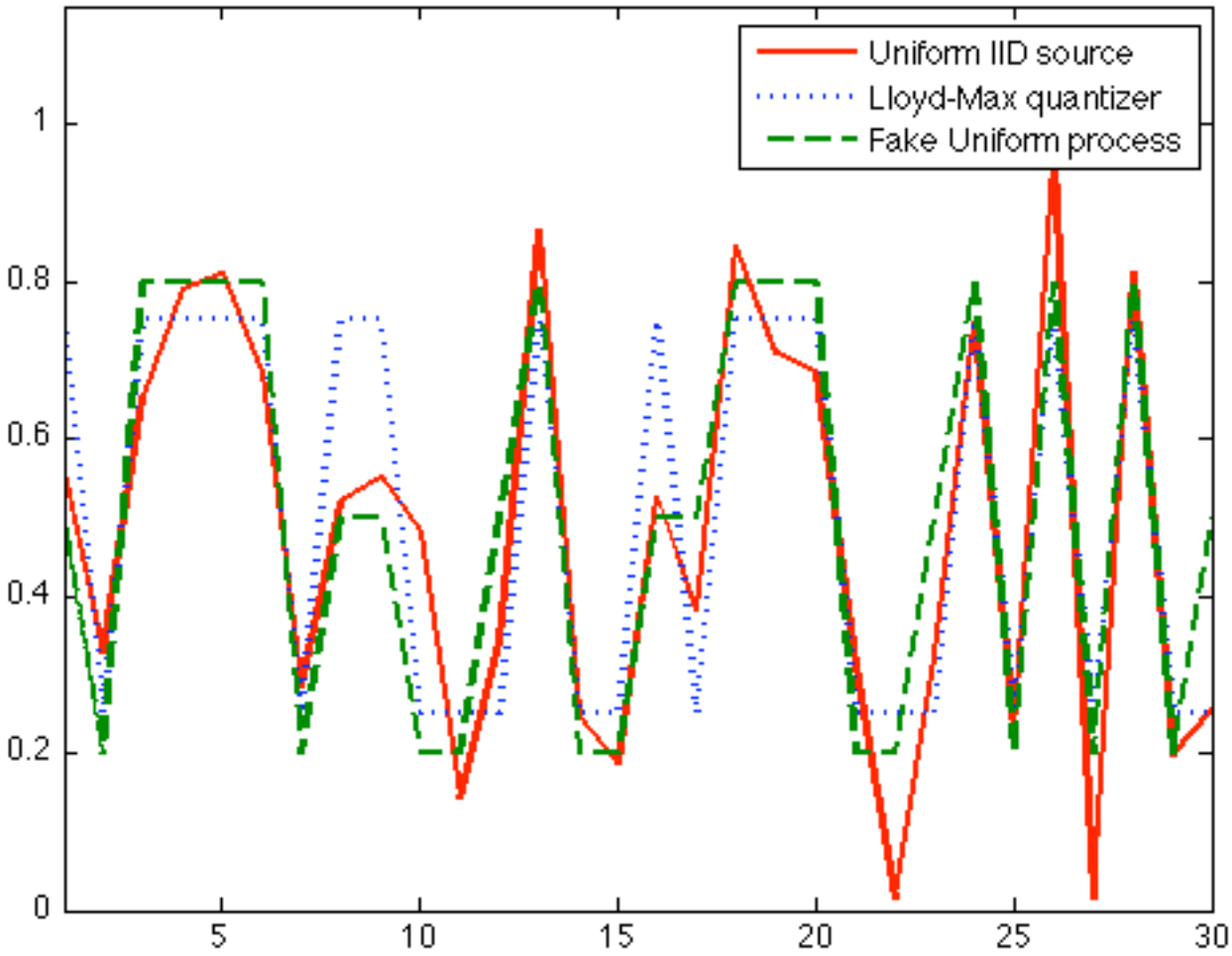
1 bit fake Gaussian



Numeric Results: Uniform $[0, 1)$

	Rate(bits)	MSE	SNR(dB)
RP_8	1	0.0203	6.13
RP_9	1	0.0195	6.30
RP_10	1	0.0190	6.42
RP_12	1	0.0184	6.55
RP_16	1	0.0179	6.69
RP_20	1	0.0176	6.75
RP_24	1	0.0175	6.78
RP_28	1	0.0174	6.79
$D_X(R)$	1	0.0173	6.84
TCQ_9	1	0.0194	6.33
TCQ(opt)_9	1	0.0183	6.58
LC_10	1	0.0191	6.40
LC(opt)_10	1	0.0179	6.67

1 bit fake Uniform

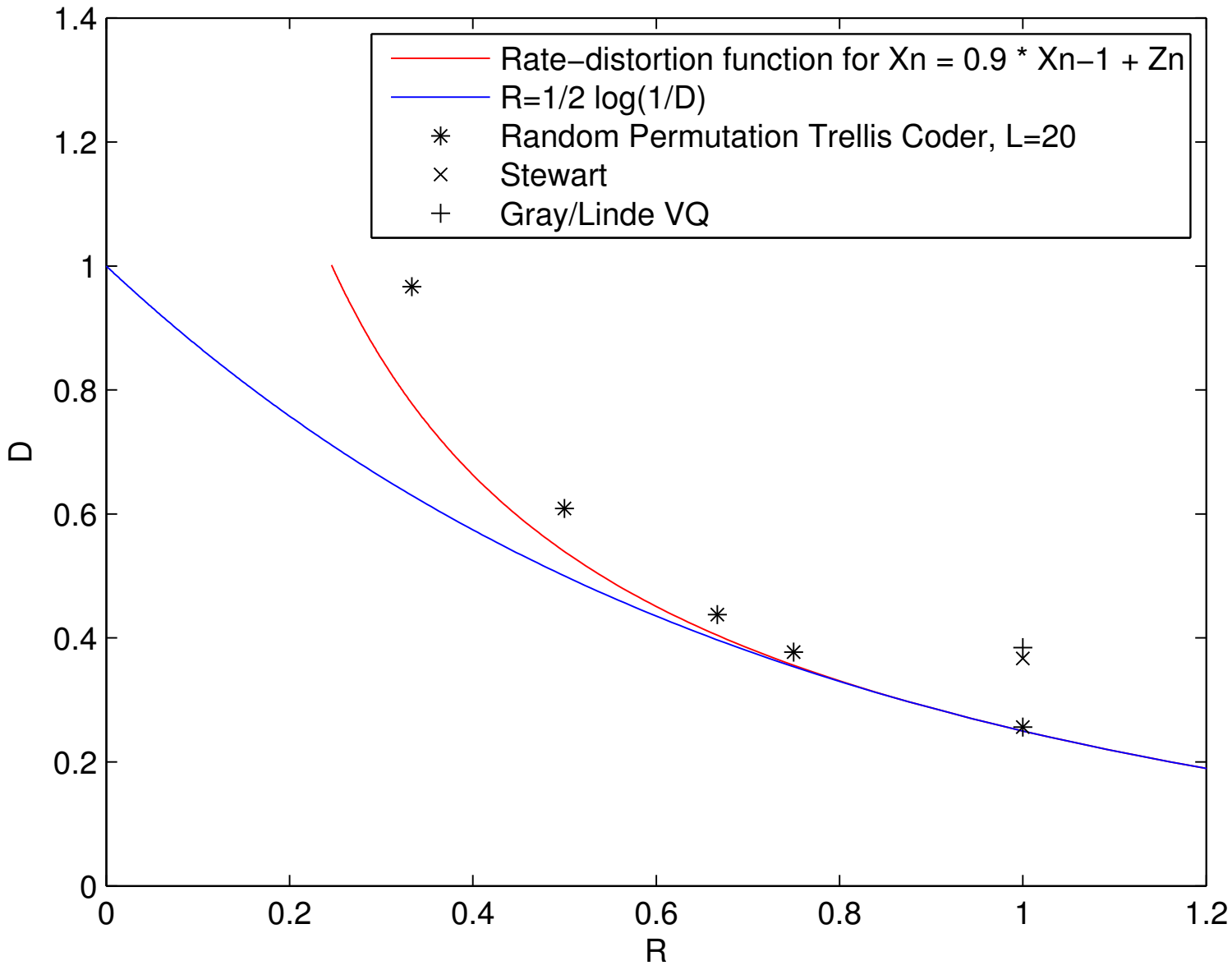


More general sources?

Finite-order distribution convergence proofs exist only for IID sources. Conjecture more generally the convergence is to the corresponding finite-order distribution of the Shannon optimal reproduction *process* distribution from process definition of the DRF.

Mao (2011) used this idea to design a trellis encoding system for a Gauss Markov source $X_n = 0.9X_{n-1} + W_n$, where W_n is IID $\mathcal{N}(0, 1)$.

Fake Gauss Markov designed by using a fake Gauss IID to drive an autoregressive filter chosen so the output reproduction process had a covariance close to that of the Shannon optimal reproduction process.


Rate−distortion function for Xn = 0.9 * Xn−1 + Zn
R=1/2 log(1/D)
Random Permutation Trellis Coder, L=20
Stewart
Gray/Linde VQ
D
R
0
0.2
0.4
0.6
0.8
1
1.2
1.4

Random Closing Observations

- The 1 bit per symbol fake Gaussian passes the Kolmogorov-Smirnov Goodness-of-Fit (with significance level $\alpha = 0.05$) for marginals *and the conditional distributions for past of length* ≤ 4 as being Gauss IID.
- Weissman and Ordentlich (2005) developed results in the spirit of the asymptotically optimal reproduction distributions for block codes by considering empirical reproduction distributions for IID sources and sources satisfying the Shannon lower bound.
- Can use Lloyd centroid property to fine-tune a trellis encoder to a training set. Helps a little when the shift-register length is short, but improvement is negligible otherwise.

- Many of the fundamental underlying results remain very hard to prove, e.g., Ornstein wrote a book proving his general isomorphism theorem. Ornstein theory is harder than Shannon theory because one needs to construct a sequence of codes that get better *and converges*.
- **Mismatch** The $\overline{d}$-distance yields several mismatch results of the following form: Suppose that you design a nearly optimal source code for a source X, but you then you apply the code to another source Y. Then the resulting difference in performance is bound above by $\overline{d}(\mu_X, \mu_Y)$, Similarly, both operational and Shannon DRFs are continuous with respect to $\overline{d}$.

Acknowledgements

These slides include material from over four decades of work with students and colleagues. Particularly influential collaborators on these specific topics include Dave Neuhoff, Paul Shields, Don Ornstein, Tamás Linder, and Mark Mao, none of whom bare any blame for any errors in these slides.

Suggested Reading

I. Csiszár. On an extremum problem of information theory. *Studia Scientiarum Mathematicarum Hungarica*, pages 57–70, 1974.

S.L. Fix. *Rate distortion functions for continuous alphabet memoryless sources*. PhD thesis, University of Michigan, Ann Arbor, Michigan, 1977.

A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Boston, 1992.

R. M. Gray. Sliding-block source coding. *IEEE Trans. Inform. Theory*, IT-21(4):357–368, July 1975.

R. M. Gray. Time–invariant trellis encoding of ergodic discrete–time sources with a fidelity criterion, *IEEE Trans. on Info. Theory*, Vol. 23, pp. 71–83, Jan. 1977.

R. M. Gray. *Probability, Random Processes, and Ergodic Properties*. Springer-Verlag, New York, 1988. Second Edition, Springer, 2009.

R. M. Gray. *Entropy and Information Theory*, Springer–Verlag, 1990. Second edition, Springer, 2011.

R. M. Gray, D. L. Neuhoff, and P. C. Shields. A generalization of ornstein's d-bar distance with applications to information theory. *Ann. Probab.*, 3:315–328, April 1975.

A. J. Khinchine. The entropy concept in probability theory. *Uspekhi Matematicheskikh Nauk.*, 8:3–20, 1953. Translated in *Mathematical Foundations of Information Theory*,Dover New York (1957).

A. N. Kolmogorov. On the Shannon theory of information in the case of continuous signals. *IRE Transactions Inform. Theory*, IT-2:102–108, 1956.

A. N. Kolmogorov. A new metric invariant of transitive dynamic systems and automorphisms in Lebesgue spaces. *Dokl. Akad. Nauk SSR*, 119:861–864, 1958. (In Russian.).

S. P. Lloyd. Least squares quantization in PCM. Unpublished Bell Laboratories Technical Note. Portions presented at the Institute of Mathematical Statistics Meeting Atlantic City New Jersey September 1957. Published in the March 1982 special issue on quantization of the *IEEE Transactions on Information Theory*, 1957.

Mark Z. Mao, Robert M. Gray, and Tamás Linder. Rate-constrained simulation and source coding iid sources. *IEEE Transactions on Information Theory*, to appear.

Mark Z. Mao, *On Asymptotically Optimal Source Coding and Simulation of Stationary Sources*, PhD Dissertation, Department of Electrical Engineering, Stanford University, June, 2011.

K. Marton. On the rate distortion function of stationary sources. *Problems of Control and Information Theory*, 4:289–297, 1975.

K. Marton. Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration. *Annals of Probability*, 24(2):857–866, 1996.

D. Ornstein. Bernoulli shifts with the same entropy are isomorphic. *Advances in Math.*, 4:337–352, 1970.

D. Ornstein. An application of ergodic theory to probability theory. *Ann. Probab.*, 1:43–58, 1973.

D. Ornstein. *Ergodic Theory,Randomness,and Dynamical Systems*. Yale University Press, New Haven, 1975.

S.T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. John Wiley & Sons Ltd, Chichester, 1991.

S.T. Rachev and L. Rüschendorf. *Mass Transportation Problems Vol. I: Theory, Vol. II: Applications*. Probability and its applications. Springer- Verlag, New York, 1998.

K. Rose. A mapping approach to rate-distortion computation and analysis. *EEE Trans. Inform. Theory*, 40(6):1939–1952, Nov. 1994.

P. C. Shields. *The Theory of Bernoulli Shifts*. The University of Chicago Press, Chicago,Ill., 1973.

P. C. Shields. The interactions between ergodic theory and information theory. *IEEE Trans. Inform. Theory*, 40:2079–2093, 1998.

P. C. Shields and D. L. Neuhoff. Block and sliding-block source coding. *IEEE Trans. Inform. Theory*, IT-23:211–215, 1977.

M. Smorodinsky. A partition on a bernoulli shift which is not weakly bernoulli. *Theory of Computing Systems*, 5(3):201–203, 1971.

H. Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci.*, IV(C1. III):801–804, 1956.

L. C. Stewart, R. M. Gray, and Y. Linde. The design of trellis waveform coders. *IEEE Trans. Comm.*, COM-30:702–710, April 1982.

S. S. Vallender. Computing the wasserstein distance between probability distributions on the line. *Theory Probab. Appl.*, 18:824–827, 1973.

L. N. Vasershtein. Markov processes on countable product space describing large systems of automata. *Problemy Peredachi Informatsii*, 5:64–73, 1969.

C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

C. Villani. *Optimal Transport, Old and New*, volume 338 of *Grundlehren der Mathemtischen Wissenschaften*. Springer, 2009.

T. Weissman and E. Ordentlich. The Empirical Distribution of Rate-Constrained Source Codes, *IEEE Transactions on Information Theory*, Vo. 51, No. 11, Nov 2005, pp. 3718-3733.