9 July 2008

Source Coding and Simulation

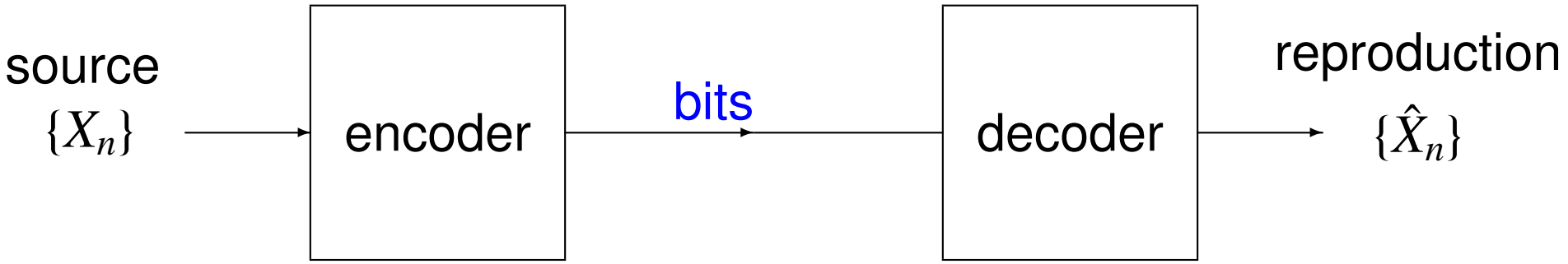
Robert M. Gray
Information Systems Laboratory
Department of Electrical Engineering
Stanford, CA 94305
rmgray@stanford.edu

http://ee.stanford.edu/~ gray

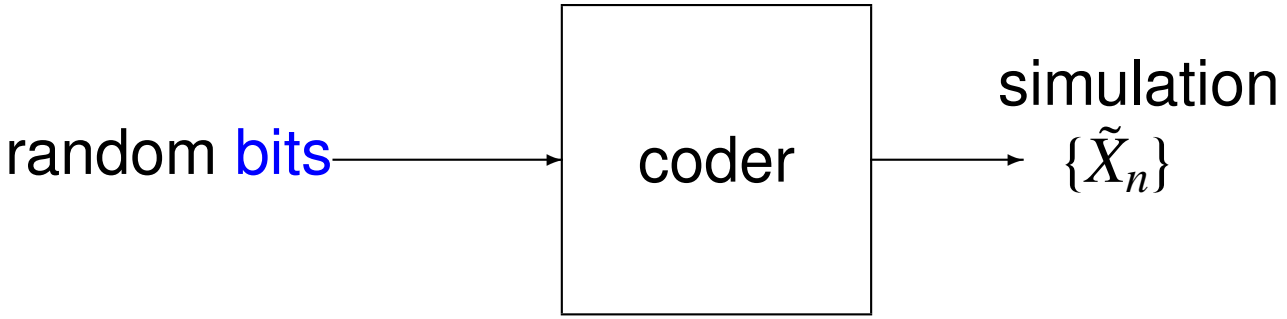
Historical and recent research described here was supported in part by

Source coding and simulation

Source coding/compression/quantization



Simulation/synthesis/fake process



Source

$X = \{X_n; n \in \mathcal{Z}\}$ stationary and ergodic *random process*, distribution μ

$X_n \in A_X =$ *alphabet:* discrete, continuous, or mixed

random vectors $X^N = (X_0, X_1, \cdots, X_{N-1})$, distribution μ^N

Shannon entropy

$$H(X^N) = H(\mu^N) = \begin{cases} -\sum_{x^N} \mu^N(x^N) \log \mu^N(x^N) & A_X \text{ discrete} \\ \infty & \text{otherwise} \end{cases}$$

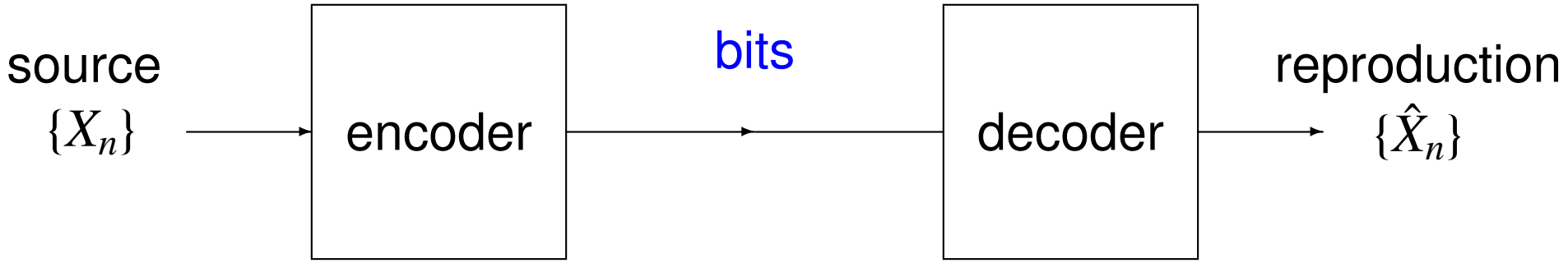
Shannon entropy (rate) $H(X) = H(\mu) = \inf_N H(X^N)/N = \lim_{N\to\infty} H(X^N)/N$

& other information measures

Source coding with a fidelity criterion

[Shannon (1959)]

Communicate a source $\{X_n\}$ to a user through a bit pipe

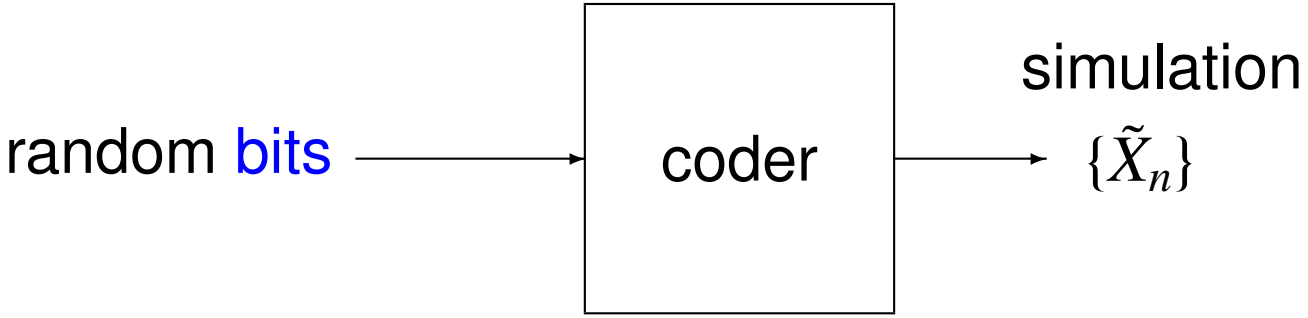


What is the *best tradeoff* between the *rate* in bits per source sample and the *quality* of the reproduction with respect to the input?

Shannon rate-distortion theory, source coding with a fidelity criterion, lossy data compression, quantization

The simulation problem (1977)

Simulate (synthesize, imitate, model, fake) a source $\{X_n\}$



What is the *best* simulation of the source given

- a simple random bit generation, e.g., coin flips (iid),
- a stationary (time-invariant) coder, and
- a constraint on # of bits (possibly infinite) per simulated symbol?

Would like a simulated process to

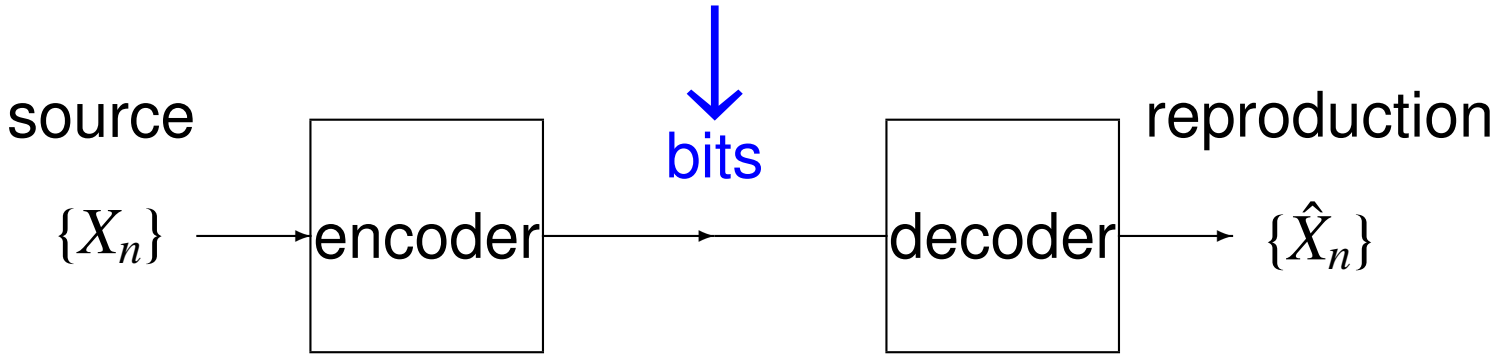
- have key properties of original process: stationarity, ergodicity, mixing, 0-1 law (purely nondeterministic, K)
- "resemble" original as closely as possible
- be *perfect* if bitrate sufficient. I.e., same distributions as original source. What stationary ergodic processes have exactly this form?

(Not all do!) (modeling, taxonomy of random processes)

An alternative notion of simulation introduced by Steinburg and Verdu (1996) and related to source coding. Does not require stationarity and ergodicity (or preservation of such properties).

An information theoretic "folk theorem"

If source code nearly optimal,
then bits ≈ iid fair coin flips



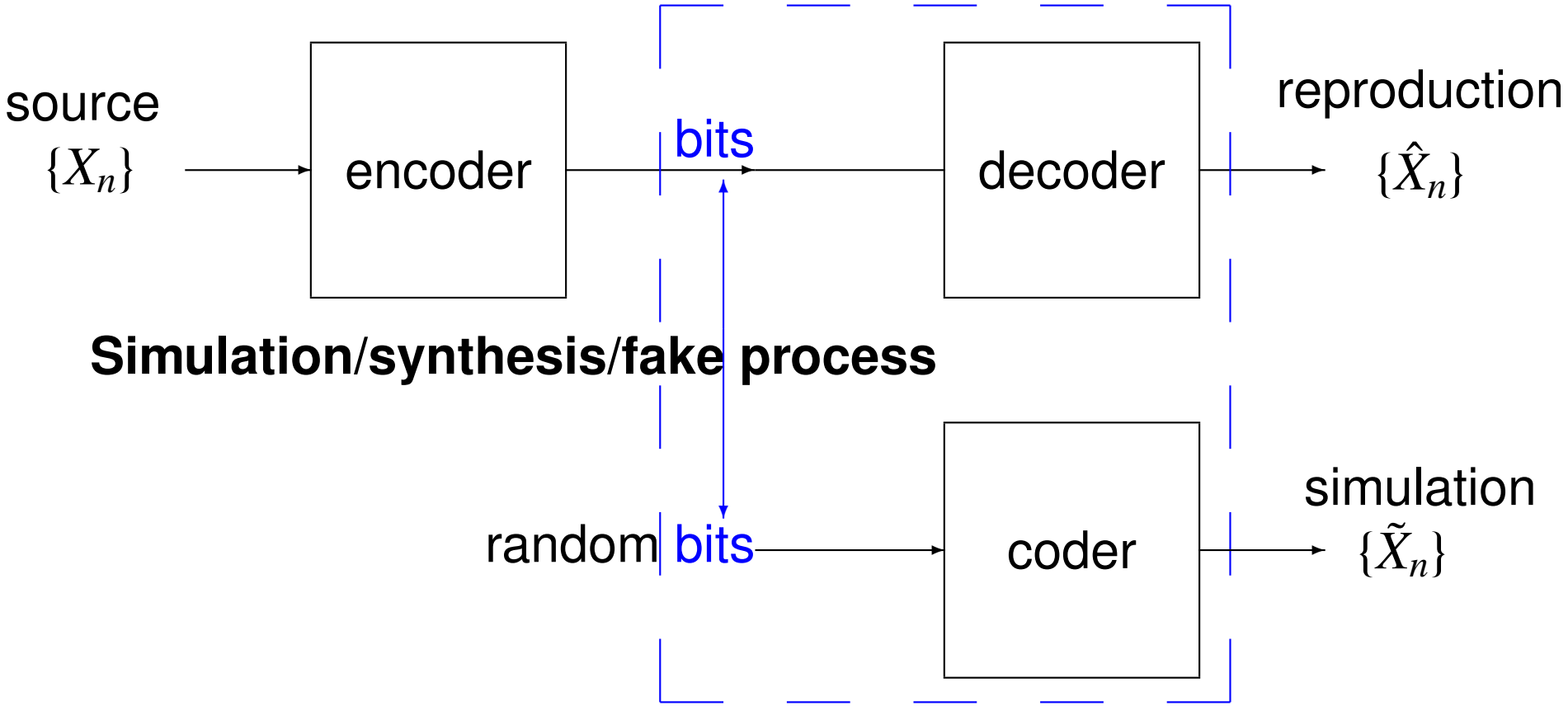
Bits are maximally informative, maximum entropy.

True??

Coin flips provide simple input mechanism for simulation.

Suggests connection between source coding and simulation:

Source coding/compression



Does nearly optimal performance $\Rightarrow$ "nearly" iid bits?

Are source decoders and source simulators equivalent?

Are source coding and simulation equivalent?

Coding

Two basic coding structures for coding a process X with alphabet A_X into Y with alphabet A_Y:

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

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

There are constructions in IT and ergodic theory to get BC from SBC & vice-versa.

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

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

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 $\longrightarrow$

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

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

Block coding

- Far more known about design: e.g., transform codes, vector quantization, clustering 

- Does not preserve key properties (stationarity, ergodicity, mixing, 0-1 law) 

  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 SBC of input.

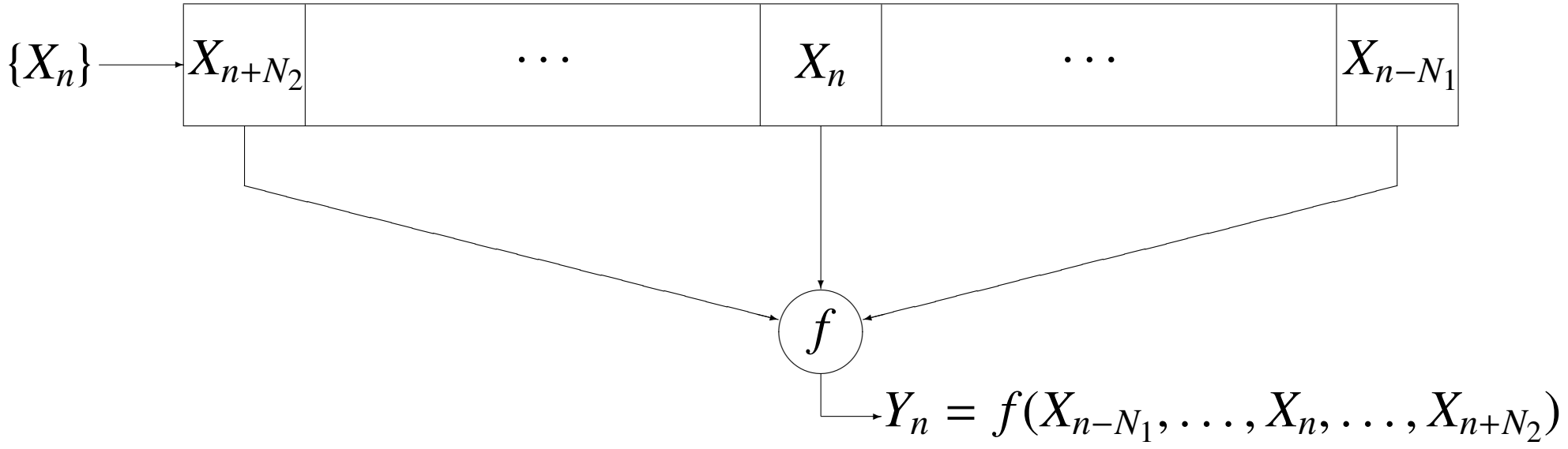
- Not defined for infinite block length, no limiting codes. 

Sliding-block (stationary, sliding-window) coding

- preserves key properties of input process: stationarity, ergodicity, mixing, 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, nonlinear and linear time-invariant filtering, wavelet coefficient evaluation
- used to prove fundamental results in ergodic theory, e.g., the Kolmogorov-Sinai-Ornstein isomorphism theorem:

Sliding-block coding and isomorphism

A sliding-block code (SBC) has the form



Infinite N_1, N_2 are allowed. Two processes are *isomorphic* if there exists an *invertible* SBC from either process to the other.

A process is a *B-process* if it is an SBC of an iid process.

Ornstein proved (1970) that two B-processes are isomorphic iff their entropy rates are equal.

Source Coding: Block Coding

Distortion measure $d_N(x^N, y^N) = \frac{1}{N} \sum_{i=0}^{N-1} d_1(x_i, y_i)$

Codebook/Decoder $C_N = \{\mathcal{D}_N(i); i \in \mathcal{I}\}, |\mathcal{I}| = M$

Encoder $\mathcal{E}_N : A_X^N \to \mathcal{I}$

Distortion $D(\mathcal{E}_N, \mathcal{D}_N) = E\left(d_N(X^N, \mathcal{D}_N(\mathcal{E}_N(X^N)))\right)$

Rate $R(\mathcal{E}_N) = \begin{cases} \frac{1}{N} \log M & \text{fixed-rate} \\ N^{-1} H(\mathcal{E}_N(X^N)) & \text{variable-rate} \end{cases}$

Optimal performance? Operational distortion-rate function (DRF)

$$\delta_{\mathrm{BC}}^{(N)}(R) = \inf_{\mathcal{E}_N, \mathcal{D}_N : R(\mathcal{E}_N) \le R} D(\mathcal{E}_N, \mathcal{D}_N)$$

$$\delta_{\mathrm{BC}}(R) = \inf_N \delta_{\mathrm{BC}}^{(N)}(R) = \lim_{N \to \infty} \delta_{\mathrm{BC}}^{(N)}(R)$$

Not computable. Evaluate by Shannon DRF:

$$D_X(R) = \inf_N D_N(R) = \lim_{N \to \infty} D_N(R)$$

$$D_N(R) = \inf_{p^N : p^N \Rightarrow \mu^N, N^{-1} I(X^N, Y^N) \le R} E d_N(X^N, Y^N)$$

Block Source Coding Theorem: For a stationary and ergodic source*, $\delta_{\mathrm{BC}}(R) = D_X(R)$

*With the usual technical conditions.

Source Coding: Sliding-Block Coding

Encoder $f_N : A_X^N \to A_U,\ U_n = f_N(X_{n-N_1}, \ldots, X_{n+N_2})$

Decoder $g_K : A_U^K \to \hat{A}_X,\ \hat{X}_n = g_K(U_{n-K_1}, \ldots, U_{n+K_2})$

Distortion $D(f,g) = E\left(d_1(X_0, \hat{X}_0)\right)$, **Rate** $R(f) = \log|A_U|$

Optimal performance:

$$\begin{aligned} \delta_{\text{SBC}}^{(N,K)}(R) &= \inf_{f_N, g_K : R(f) \le R} D(f_N, g_K) \\ \delta_{\text{SBC}}(R) &= \inf_{N,K} \delta_{\text{SBC}}^{(N,K)}(R) = \inf_{f,g : R(f) \le R} D(f,g) \end{aligned}$$

Sliding-lock Source Coding Theorem: For a stationary and ergodic source*, $\delta_{\text{BC}}(R) = \delta_{\text{SBC}}(R) = D_X(R)$

*ditto

$$\underbrace{X_0, X_1, \ldots, X_{N-1}}_{\downarrow \mathcal{E}_N \ \underbrace{U_0, U_1, \ldots, U_{N-1}}_{\downarrow \mathcal{D}_N \ \hat{X}_0, \hat{X}_1, \ldots, \hat{X}_{N-1}}} \quad \underbrace{X_N, X_1, \ldots, X_{2N-1}}_{\downarrow \mathcal{E}_N \ \underbrace{U_N, U_{N+1}, \ldots, U_{2N-1}}_{\downarrow \mathcal{D}_N \ \hat{X}_N, \hat{X}_1, \ldots, \hat{X}_{2N-1}}} \quad \underbrace{X_{2N}, X_1, \ldots, X_{3N-1}}_{\downarrow \mathcal{E}_N \ \underbrace{U_{2N}, U_{2N+1}, \ldots, U_{3N-1}}_{\downarrow \mathcal{D}_N \ \hat{X}_{2N}, \hat{X}_1, \ldots, \hat{X}_{3N-1}}} \quad \cdots$$

VS.

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

$$\downarrow f$$

$$\cdots, U_{n-K_1-1}, \underbrace{U_{n-K_1}, \cdots, U_n, \cdots, U_{n+K_2}}, U_{n+K_2+1}, \cdots$$

$$\downarrow g$$

$$\cdots, \hat{X}_{n-1}, \hat{X}_n, \hat{X}_{n+1}, \cdots$$

If coding nearly optimal, is U_n nearly iid?

Process Distance Measures

How quantify “nearly iid”?

Related: How quantify “best” simulation?

One approach: process distortion measures

Useful example in information theory and ergodic theory:

$\bar{d}$-distance: Kantorovich/Vasershtein/Ornstein distance

Basic ideas:

Two stationary random processes, X with distribution μ, Y with distribution ν. Vector distortion d_N.

$$\begin{aligned} \bar{d}_N(\mu^N, \nu^N) &= \inf_{p^N \Rightarrow \mu^N, \nu^N} E_{p^N} d_N(X^N, Y^N) \\ \bar{d}(\mu, \nu) &= \sup_N \bar{d}_N(\mu^N, \nu^N) = \inf_{p \Rightarrow \mu, \nu} E_p d_1(X_0, Y_0) \end{aligned}$$

Smallest achievable distortion between two processes with given marginals over all joint distributions consistent with marginals.

Many equivalent definitions. E.g., how much have to change one typical sequence of one source to get a typical sequence of another.

Historical aside

$\bar{d}_N$ rediscovered and renamed numerous times.

Kantorovich (1942): metrics on compact metric spaces. Often called the Kantorovich or transportation metric. Inseparable from development of linear programming.

Early focus on scalar case and ℓ_r norms: Dall'Aglio (1956), Frechet (1956), Vasershtein/Wasserstein (1969), Mallows (1972), Vallender (1973).

Ornstein (1970-73) used the idea with the Hamming distance on vectors and processes. Called the $\bar{d}$ distance. First appearance as distance measure on *processes*.

Gray, Neuhoff, and Shields (1975) considered vector and process case using additive distortion measures, including $d_1(x,y) = |x-y|^2$, calling the distortion $\bar{\rho}$ after Ornstein. Vector case equivalent to subsequent development of Kantorovich for ℓ_r norms to vectors (L_r-minimal metric):

$$\begin{aligned}\bar{\rho}^{1/r}(\mu^N,\nu^N) &= \ell_r(\mu^N,\nu^N) \stackrel{\Delta}{=} \left[N\bar{d}_N(\mu^N,\nu^N)\right]^{\frac{1}{r}} \\ &= \inf_{p^N \Rightarrow \mu^N,\nu^N} \left[E(\|X^N - Y^N\|_r^r)\right]^{\frac{1}{r}}\end{aligned}$$

Usually reserve notation $\bar{d}$ for Ornstein (Hamming), use $\bar{\rho}$ for ℓ_r^r

Rediscovered as "earth mover's distance" in CS literature, used in clustering algorithms for pattern recognition. Later renamed (1981) "Mallows distance" after 1972 rediscovery of scalar Kantorovich.

Properties

- Ornstein $\bar{d}$ distance and L_r-minimal distance/$\bar{\rho}^{1/r}$ are metrics.
- Infimum is actually a minimum.
- The class of all B-processes of a given alphabet is the closure under Ornstein's $\bar{d}$ of all k-step mixing Markov processes of that alphabet.
- Entropy rate is continuous in $\bar{d}$, Shannon DRF in $\bar{\rho}$
- Can evaluate $\bar{\rho}$ for iid, purely nondeterministic Gaussian processes, filtered uniform iid, $\bar{d}$ for discrete iid. In general a *linear programming problem*.

Application 1: Geometric view of source coding

$$\delta_{\mathrm{BC}}(R) = \delta_{\mathrm{SBC}}(R) = D_X(R) = \inf_{\nu: H(\nu) \le R} \bar{\rho}(\mu, \nu)$$

[Gray, Neuhoff, and Omura (1974)]

A form of simulation, but cannot say ν generated from iid.

Distance to "closest" process in $\bar{\rho}$ with entropy rate $\le R$

Compare with process version of Shannon DRF [Marton (1972)]:

$$D_X(R) = \inf_{p: p \Rightarrow \mu, I(X,Y) \le R} E[d_1(X_0, Y_0)]$$

Application 2: Quantization as distribution approximation

[Pollard (1982), Graf and Luschgy (2000)]

(Vector) Quantizer $\Leftrightarrow$ probability distribution on codebook

Block coding/quantization: fixed rate

$$\delta_{\mathrm{BC}}^{(N)}(R) = \inf_{\nu^N} \bar{\rho}_N(\mu^N, \nu^N)$$

Minimum is over all discrete distributions ν^N with 2^{NR} atoms.

Suppose discrete distribution $(\pi, C_N) = \{\pi_i, y_i; i = 1, \ldots, 2^{NR}\}$, $\sum_{i=1}^{2^{NR}} \pi_i = 1$, $y_i \in A_X^N$, solves minimization $\Rightarrow$
a discrete simulation of X^N $\Rightarrow$ block independent, N-stationary process simulation

Application 3: Optimal simulation and source coding

A definition of optimal simulation of process $X \sim \mu$ using an SBC of an iid process Z [Gray (1977)]:

$$\Delta(X|Z) = \inf_{\tilde{\mu}: Z_\pi \to [g] \to \tilde{X}_n \sim \tilde{\mu}} \bar{\rho}(\mu, \tilde{\mu})$$

Sliding-block coding reduces entropy $\Rightarrow H(Z) \geq H(\tilde{\mu}) \Rightarrow$

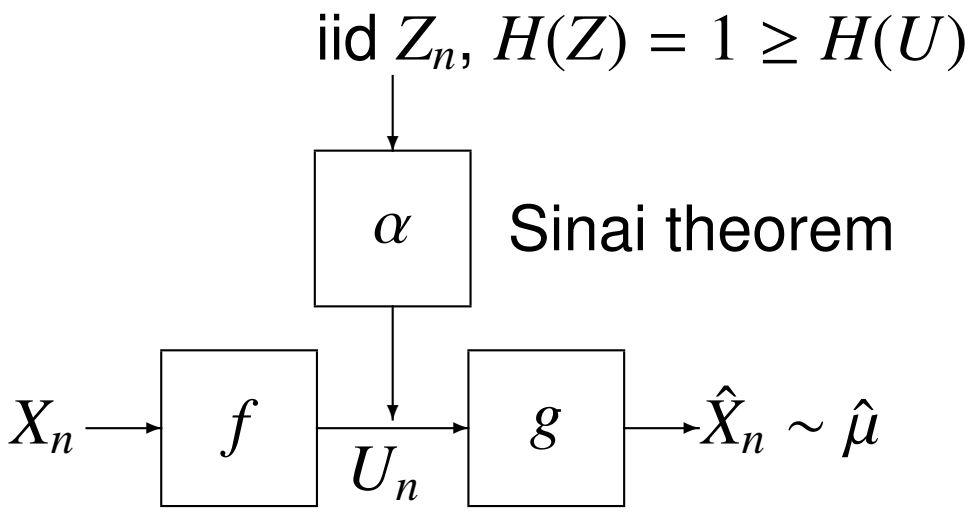
$$\Delta(X|Z) \geq \inf_{\text{stationary ergodic } \hat{\mu}: H(\hat{\mu}) \leq H(Z)} \bar{\rho}(\mu, \hat{\mu}) = D_X(H(Z))$$

If X is a B-process, converse is true and

$$\Delta(X|Z) = D_X(H(Z)) = \delta_{\text{SBC}}(H(Z))$$

$\Rightarrow$ the source coding and simulation problems are equivalent if the source is a B-process

Proof: Choose f, g: $Ed_1(X_0, \hat{X}_0) \approx D_X(1)$,



Cascade $\beta = g\alpha$ is SBC producing $\hat{X}$ from Z, $Ed_1(X_0, \hat{X}_0) \geq \Delta(X|Z)$. □

Bit behavior for near optimal codes

Suppose use block code C_N to code source. Let π denote induced index pmf.

What can be said about π if code performance near Shannon optimal?

Approximately uniform, like 2^N coin flips?

Sort of . . .

Shannon $\Rightarrow$ there is an asymptotically optimal sequence of block codes $C^{(N)}$ for which $D_N = Ed_N(X^N, \hat{X}^N) \downarrow D_X(1)$

$R_X(D)$ is a continuous function, hence

$$\begin{aligned} 1 &= N^{-1}\log_2 2^N \geq N^{-1}H(\mathcal{E}(X^N)) \geq N^{-1}H(\hat{X}^N) \\ &\geq N^{-1}I(X^N;\hat{X}^N) \geq R_N(D_N) \geq R_X(D_N) \underset{N\to\infty}{\to} 1 \end{aligned}$$

As blocklength grows, indexes have maximal per symbol entropy and hence can be thought of as approximately uniformly distributed, but not stationary or ergodic and can not get process theorem — does not determine *entropy rate* or show that overall process behavior is like coin flips, even if stationarize.

If use SBCs, can get rigorous process version:

Choose $f^{(N)}, g^{(N)}$ so that $D_N = D(f^{(N)}, g^{(N)}) \downarrow D_X(1)$

Let $U^{(N)}$, $\hat{X}^{(N)}$ denote encoded and reproduction processes (necessarily stationary and ergodic)

$$\begin{aligned} 1 &\geq H(U^{(N)}) \geq H(\hat{X}^{(N)}) \geq I(X, \hat{X}^{(N)}) \\ &\geq R(D_N) \underset{N\to\infty}{\to} 1 \end{aligned}$$

$$\lim_{N\to\infty} H(U^{(N)}) = 1 \Rightarrow \lim_{N\to\infty} \bar{d}(U^{(N)}, Z) = 0$$

Proof: Marton's inequality for relative entropy and $\bar{d}$ (T. Linder)

As average distortion nears Shannon limit for stationary ergodic source, binary channel process approaches coin flips in $\bar{d}$

Recap

Old: If source is a stationary filtering of an iid process (a B-process, discrete or continuous alphabet), then the source coding problem and the simulation problem have the same solution (and the optimal simulator and decoder are equivalent).

New: If stationary source coding performs close to Shannon optimum, encoded process is close to iid in $\bar{d}$

Frosting: An excuse to present ideas of modeling, coding, and process distance measures common to ergodic theory and information theory.

A few final thoughts and questions

• The $\bar{d}$ close to iid property is nice for intuition, but does it actually help?

E.g., B-processes (SBC of iid process) have many special properties. Are there weak versions of those properties for processes that = SBC of a process $\bar{d}$-close to iid?

• Does equivalence of source coding and simulation hold for the more general case of stationary and ergodic sources? — Steinberg/Verdu results hold more generally, but in ergodic theory it is known that there are stationary, ergodic, mixing, purely nondeterministic processes which are *not $\bar{d}$ close to a B-process*.

• Source coding as "almost isomorphism," avoids hard part (invertibility).

• How does fitting a model using $\bar{\rho}$ compare to the Itakura-Saito distortion used in speech processing to fit autoregressive speech models to real speech? Can Marton/Talagrand inequalities be extended? (Steinberg/Verdu considered relative entropy rates in their simulation problem formulation.)

• Shortcoming of B-processes: In speech, only model unvoiced sounds well. Voiced sounds better modeled by periodic input to same filter type: 0-entropy. Composite models? Connections to Pinsker's (☠ disproved) conjecture regarding products of K processes and 0-entropy processes?

• Simulator design, e.g., best fake Gaussian from bits?

♫