This week's Topics

Shannon's Work.

- Mathematical/Probabilistic Model of Communication.
- Definitions of Information, Entropy, Randomness.
- Noiseless Channel & Coding Theorem.
- Noisy Channel & Coding Theorem.
- Converses.
- Algorithmic challenges.

Detour from Error-correcting codes?

Shannon's Framework (1948)

Three entities: Source, Channel, and Receiver.

Source: Generates "message" - a sequence of bits/ symbols - according to some "stochastic" process S.

Communication Channel: Means of passing information from source to receiver. May introduce errors, where the error sequence is another stochastic process E.

Receiver: Knows the processes S and E, but would like to know what sequence was generated by the source.

Goals/Options

- Noiseless case: Channel precious commodity. Would like to optimize usage.
- Noisy case: Would like to recover message despite errors.
- Source can "Encode" information.
- Receiver can "Decode" information.

Theories are very general: We will describe very specific cases only!

Noiseless case: Example

- Channel transmits bits: $0 \to 0, 1 \to 1$. 1 bit per unit of time.
- Source produces a sequence of independent bits: 0 with probability $1-p$ and 1 with probability p.
- Question: Expected time to transmit n bits, generated by this source?

Noiseless Coding Theorem (for Example)

Let $H_2(p) = -(p \log_2 p + (1-p) \log_2 (1-p))$.

Noiseless Coding Theorem: Informally, expected time $\to H(p) \cdot n$ as $n \to \infty$.

Formally, for every $\epsilon > 0$, there exists n_0 s.t. for every $n \geq n_0$,

$\exists E : \{0,1\}^n \to \{0,1\}^*$ and $D : \{0,1\}^* \to \{0,1\}^n$ s.t.

- For all $x \in \{0,1\}^n, D(E(x)) = x$.
- $\mathbf{E}_x[|E(x)|] \leq (H(p) + \epsilon)n$.

Proof: Exercise.

Entropy of a source

- Distribution $\mathcal{D}$ on finite set S is $\mathcal{D} : S \to [0,1]$ with $\sum_{x \in S} \mathcal{D}(x) = 1$.
- Entropy: $H(\mathcal{D}) = \sum_{x \in S} -\mathcal{D}(x) \log_2 \mathcal{D}(x)$.
- Entropy of p-biased bit $H_2(p)$.
- Entropy quantifies randomness in a distribution.
- Coding theorem: Suffices to specify entropy # of bits (amortized, in expectation) to specify the point of the probability space.
- Fundamental notion in probability/information theory.

Binary Entropy Function $H_2(p)$

- Plot $H(p)$.
- Main significance?
  - Let $B_2(y,r) = \{x \in \{0,1\}^n | \Delta(x,y) \leq r\}$ (n implied).
  - Let $\mathrm{Vol}_2(r,n) = |B_2(0,r)|$.
  - Then $\mathrm{Vol}_2(pn,n) = 2^{(H(p)+o(1))n}$

Noisy Case: Example

- Source produces 0/1 w.p. 1/2.
- Error channel: Binary Symmetric Channel with probability p (BSC_p), transmits 1 bit per unit of time faithfully with probability $1-p$ and flips it with probability p.
- Goal: How many source bits can be transmitted in n time units?
  - Can permit some error in recovery.
  - Error probability during recovery should be close to zero.
- Prevailing belief: Can only transmit $o(n)$ bits.

Noisy Coding Theorem (for Example)

Theorem: (Informally) Can transmit $(1 - H(p)) \cdot n$ bits, with error probability going to zero exponentially fast.

(Formally) $\forall \epsilon > 0, \exists \delta > 0$ s.t. for all n:

Let $k = (1 - H(p + \epsilon))n$. Then $\exists E : \{0,1\}^k \to \{0,1\}^n$ and $\exists D : \{0,1\}^n \to \{0,1\}^k$ s.t.

$$\Pr_{\eta,x}[D(E(x) + \eta) \neq x] \leq \exp(-\delta n),$$

where x is chosen according to the source and η independently according to BSC_p.

The Encoding and Decoding Functions

- E chosen at random from all functions mapping $\{0,1\}^k \to \{0,1\}^n$.
- D chosen to be the brute force algorithm - for every y, $D(y)$ is the vector x that minimizes $\Delta(E(x), y)$.
- Far from constructive!!!
- But its a proof of concept!
- Main lemma: For E, D as above, the probability of decoding failure is exponentially small, for any fixed message x.
- Power of the probabilistic method!

Proof of Lemma

- Will fix $x \in \{0,1\}^k$ and $E(x)$ first and pick error η next, and then the rest of E last!
- η is *Bad* if it has weight more than $(p+\epsilon)n$.

  $$\Pr_{\eta}[\eta \mathrm{Bad}] \leq 2^{-\delta n}$$

  (Chernoff bounds).
- x' Bad for x, η if $E(x') \in B_2(E(x)+\eta, (p+\epsilon)n)$.

  $$\Pr_{E(x')}[x' \mathrm{Bad\ for\ } x, \eta] \leq 2^{H(p+\epsilon)n}/2^n$$

- $\Pr_E[\exists x' \text{ Bad for } x, \eta] \leq 2^{k+H(p)\cdot n - n}$

- If η is not Bad, and no $x' \neq x$ is Bad for x, then $D(E(x) + \eta) = x$.
- Conclude that decoding fails with probability at most $e^{-\Omega(n)}$, over random choice of E, η (for every x, and so also if x is chosen at random).
- Conclude there exists E such that encoding and decoding lead to exponentially small error probability, provided $k + H(p) \cdot n \ll n$.

Converse to Coding Theorems

- Shannon also showed his results to be tight.
- For noisy case, $1-H(p)$ is the best possible rate ...
- ... no matter what E, D are!
- How to prove this?
- Intuition: Say we transmit $E(x)$. W.h.p. # erroneous bits is $\approx pn$. In such case, symmetry implies no one received vector is likely w.p. more that $\binom{n}{pn} \approx 2^{-H(p)n}$. To have error probability close to zero, at least $2^{H(p)n}$ received vectors must decode to x. But then need $2^k \leq 2^n/2^{H(p)n}$.

Formal proof of the Converse

- η Easy if weight $\leq (p-\epsilon)n$. $\Pr_\eta[\eta$ Easy $] \leq \exp(-n)$. For any y of weight $\geq (p-\epsilon)n$, $\Pr[\eta = y] \leq 2^{-H(p-\epsilon)n}$.
- For $x \in \{0,1\}^k$ let $S_x \subseteq \{0,1\}^n = \{y | D(y) = x\}$. Have $\sum_x |S_x| = 2^n$.
- $\Pr[$ Decoding correctly$]$

$$= 2^{-k} \sum_{x \in \{0,1\}^k} \sum_{y \in S_x} \Pr_\eta[\eta = y - E(x)]$$

$$= \Pr_\eta[\eta \text{ Easy}] + 2^{-k} \sum_x \sum_{y \in S_x} \Pr_\eta[\eta = y - E(x) | \eta \text{H}$$

$$= \exp(-n) + 2^{-k} \cdot 2^{-H(p)n} \cdot 2^n$$
$$= \exp(-n)$$

Importance of Shannon's Framework

- Examples considered so far are the baby examples!
- Theory is wide and general.
- But, essentially probabilistic + "information-theoretic" not computational.
- For example, give explicit E! Give efficient D! Shannon's work does not.

More general source

- Allows for Markovian sources.
- Source described by a finite collection of states with a probability transition matrix.
- Each state corresponds to a fixed symbol of the output.
- Interesting example in the original paper: Markovian model of English. Computes the rate of English!

More general models of error

- i.i.d. case generally is a transition matrix from Σ to Γ. (Σ, Γ need not be finite! (Additive White Gaussian Channel). Yet capacity might be finite.)
- Also allows for Markovian error models. May be captured by a state diagram, with each state having its own transition matrix from Σ to Γ.

General theorem

- Every source has a Rate (based on entropy of the distribution it generates).
- Every channel has a Capacity.

Theorem: If Rate $<$ Capacity, information transmission is feasible with error decreasing exponentially with length of transmission. If Rate $>$ Capacity, information transmission is not feasible.

Contrast with Hamming

- Main goal of Shannon Theory:
  - Constructive (polytime/linear-time/etc.) E, D.
  - Maximize rate $= k/n$ where $E : \{0,1\}^k \to \{0,1\}^n$.
  - While minimizing $P_{\rm err} = \Pr_{x,\eta}[D(E(x)+\eta) \neq x]$
- Hamming theory:
  - Explicit description of $\{E(x)\}_x$.
  - No focus on E, D itself.
  - Maximize k/n and d/n, where $d = \min_{x_1,x_2}\{\Delta(E(x_1), E(x_2))\}$.
- Interpretations: Shannon theory deals with probabilistic error. Hamming with

adversarial error. Engineering need: Closer to Shannon theory. However Hamming theory provided solutions, since min. distance seemed easier to analyze than $P_{\rm err}$.