Information Theory

1 Foundations

Definitions

Information of an outcome x

$$h(x) = -\log(p(x))$$

Cross-Entropy between p (nature) and q (model)

$$H(p;q) = -\sum_x p(x)\log q(x)$$

Shannon Entropy

$$H(p) = H(p;p) = -\sum_x p(x)\log p(x)$$

Notation We identify outcomes x with integers $1, ..., m$ and associate probabilities $p(x) \geq 0$.
$H(\frac{1}{m})$ for $H(p)$ with $p(x) = \frac{1}{m}$ (uniform)
$H(X) = H(p) = \mathbb{E}(-\log(p(X)))$ where p is the pdf of X

Jensen's Inequality Let f be convex and $g : [m] \to \mathbb{R}$ be an arbitrary function that assigns a value to each outcome.

$$f\left(\sum_x p(x)g(x)\right) \leq \sum_x p(x)f(g(x)), \forall p(x) \geq 0, \sum_x p(x) = 1$$

alternatively

$$f(\mathbb{E}(g(X))) \leq \mathbb{E}(f(g(X)))$$

Applying this inequality to relate **Cross-Entropy** and **Entropy**, we get the following properties.

$$H(p;q) \geq H(p)$$

KL Divergence / Relative Entropy measures the difference of two distribution, i.e. the excess level of surprise when using model q when the actual distribution is p.

$$D(p||q) = \sum_x p(x)\log\frac{p(x)}{q(x)}$$

With this we get

$$H(p;q) = H(p) + D(p||q)$$

Further investigating KL divergence, we find

$$D(p||q) \geq 0$$

$$D(p||q) = 0 \iff p = q$$

A further consequence of (1) is that the uniform distribution **maximizes** entropy.

$$H\left(\frac{1}{m}\right) = \max_p H(p)$$

1.1 Conditional Entropy

Conditional Information $h(x|y) = -\log p(x|y)$
Conditional Entropy

$$H(X|Y=y) = -\sum_x p(x|y)\log p(x|y)$$

$$H(X|Y) = \sum_y p(y)H(X|Y=y)$$

Monotonicity $H(X|Y) \leq H(X)$
Joint Entropy $H(X,Y) = -\sum_{x,y} p(x,y)\log p(x,y)$
Chain Rule $H(X,Y) = H(X|Y) + H(Y)$
Subadditivity $H(X,Y) \leq H(X) + H(Y)$ with equality if $X \perp Y$
Function $H(\phi(X)) \leq H(X)$
Multiple Conditioning $H(X|Y,Y') \leq H(X|Y)$
Generalized to $X_1, ..., X_n$ we get

$$H(X_1, ..., X_n) = \sum_{i=1}^n H(X_i|X_1, ..., X_{i-1}) \leq \sum_{i=1}^n H(X_i)$$

1.2 Mutual Information

$$I(X;Y) := H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Symmetry $I(X;Y) = I(Y;X)$
Non-negative $0 \leq I(X;Y) \leq \min(H(X), H(Y))$
Independence $I(X;Y) = 0 \iff X \perp Y$
We further have

$$I(X;Y) = D(P(X,Y)||P(X)P(Y))$$

with $I(X;Y) = 0$ if $X \perp Y$.
Conditional Mutual Information

$$I(X;Y|Z) := H(X|Z) - H(X|Y,Z)$$

Conditional Independence If $X \perp Y|Z$

$$I(X;Y|Z) = 0 \text{ and } I(X;Y) \leq I(X;Z)$$

Data processing inequality

$$I(\phi(X);Y) \leq I(X;Y)$$

2 Compression

Definition - Code

A code C is a mapping from outcomes to codewords

$$C : \{1, ..., m\} \to \{0,1\}^*$$

Prefix Code If there is no codeword that is a prefix of another codeword, the code is a **prefix code**. Prefix codes retain injectivity when concatenating codewords.
Sets of codewords fulfilling the prefix property can be uniquely represented by the leaves of a binary tree. Since a leaf node has no children the prefix property is guaranteed.
Kraft's Inequality If $\{c_1, ..., c_m\}$ are codewords of a prefix code, then

$$\sum_x 2^{-l_x} \leq 1, \text{ where } l_x = |c_x|$$

Conversely, given $\{l_1, ..., l_m\} \subset \mathbb{N}$ satisfying the inequality, there exists a prefix code with those codeword lengths.
Succinct A prefix code is succinct, if Kraft's inequality holds with a equality. Else it can be optimized by pruning.
Succinct codes uniquely define a **dyadic probabilistic model**

$$q(x) = 2^{-l_x}$$

Expected codeword length of a prefix code C

$$L(C) = \sum_x p(x)l_x = \sum_x p(x)(-\log q(x)) = H(p;q)$$

Using $H(p;q) = H(p) + D(p||q)$ we can deduce that the **minimal** $L(C)$ for a binary prefix code C is

$$L^* = H(p) + \min_{q:\text{dyadic}} D(p||q)$$

Thus the closer q is to p, the more optimal the prefix code is. But since p doesn't have to be dyadic there can be an **inherent suboptimality** based on rounding.

Weak Law of Large Numbers

Let $Y_1, ..., Y_n$ be iid. random variables with mean μ. Then

$$\overline{Y}_n := \frac{1}{n}\sum_{i=1}^n Y_i \xrightarrow{\mathbb{P}} \mu \iff \lim_{n\to\infty} \mathbb{P}(|\overline{Y}_n - \mu| < \varepsilon) = 1, \forall\varepsilon > 0$$

Typicality - Asymptotic Equipartition
Let $X_1, ..., X_n \overset{\text{iid}}{\sim} p$. The ε-typical outcomes are

$$\mathcal{A}_\varepsilon^n = \left\{x \in \{1, ..., m\}^n : \left|H(p) + \frac{1}{n}\sum_{i=1}^n \log p(x_i)\right| < \varepsilon\right\}$$

By the law of large numbers for any $p, \varepsilon > 0$ and $\delta > 0$, there exists an n_0, s.t. $\forall n \geq n_0$

$$\mathbb{P}(A_\varepsilon^n) > 1 - \delta$$

in particular for $\delta = \varepsilon$.
For all $p, \varepsilon > 0$ and $n \in \mathbb{N}$, let $x \in \mathcal{A}_\varepsilon^n$, then

$$2^{-n(H(p)+\varepsilon)} \leq p(x) \leq 2^{-n(H(p)-\varepsilon)}$$

$$(1-\varepsilon)2^{n(H(p)-\varepsilon)} \leq |\mathcal{A}_\varepsilon^n| \leq 2^{n(H(p)+\varepsilon)}$$

$$\implies |\mathcal{A}_\varepsilon^n| \approx 2^{nH(p)} \text{ and for } x \in \mathcal{A}_\varepsilon^n : p(x) \approx 2^{-nH(p)}$$

We define the AEP Code to encode whole sequences

$$\text{AEP}_\varepsilon^n = \begin{cases} 0B^n(x) & \text{if } x \notin \mathcal{A}_\varepsilon^n \\ 1C^n(x) & \text{otherwise} \end{cases}$$

where we **enumerate** over the typical and atypical sequences.
Then the average codeword length **amortized** over the encoding of the sequence x of n outcomes is

$$\frac{1}{n}|C_\varepsilon^n(x)| \leq \frac{1}{n}(1 + \log|\mathcal{A}_\varepsilon^n|) \leq H(p) + \frac{1}{n} + \varepsilon$$

$$\frac{1}{n}|B^n(x)| \leq \frac{1}{n}(1 + \log m^n) \leq \log m + \frac{1}{n}$$

This result is theoretically optimal but practically not very useful.

Huffman Codes
Let X have outcomes $\{1, ..., m\}$ ordered (wlog) st. $p(1) \geq ... \geq p(m)$.
The Huffman contraction X' of X is defined as

$$X' = \min\{m-1, X\}$$

We define the Huffman Code C for X recursively from C' for the H. contraction X'

$$C(x) = \begin{cases} x-1 & \text{if } m = 2 \\ C'(x)0 & \text{if } x = m-1 \land m > 2 \\ C'(x-1)1 & \text{if } x = m \land m > 2 \\ C'(x) & \text{otherwise} \end{cases}$$

Let C be a length-optimal code, then

$$p(x) > p(x') \implies l_x \leq l_{x'}$$

$$\forall c \in \text{Img}(C) \text{ wt. } |c| \text{ maximal} : \exists c' \in \text{Img}(C).\ c' \text{ sibling of } c$$

Assume p_i ordered as above. Then a length-optimal prefix code C with $l_1 \leq ... \leq l_{m-1} = l_m$ and c_{m-1}, c_m only differing in last bit, is called **canonical**.
Huffman codes are length-optimal.

3 Prediction

A betting strategy b bets a fraction $b(x)$ on the x-th outcome.
The bookmaker provides odds 1-for-$q(x)$ for each outcome x.

$$S(X) = \frac{b(X)}{q(X)}$$

is a random variable which describes the wealth growth of a gamble.

- Maximizing $\mathbb{E}(S(X))$ over all possible b results in betting all on the highest probable outcome.
- Let $X_1, ..., X_n \sim p$ iid.
$$S_n := S(X_1, ..., X_n) = \prod_{i=1}^{n} S(X_i)$$
- Any strategy with $b(x) = 0, p(x) > 0$ for some x will almost surely fail for increasing n.
- **Doubling Rate**
$$W(b) = \mathbb{E}(\log S(X)) = \sum_x p(x) \log \frac{b(x)}{q(x)}$$
- Odds 1-for-q are **fair**, if $\sum_x q(x) = 1$. super-fair: $\sum < 1$
- In general, for fair odds, we have
$$W(b) = D(p||q) - D(p||b)$$
which is **optimal** for $b = p$, since then $D(p||b) = 0$.
- **Conservation Theorem.** For q uniform, fair and $b = p$.
$$W(b) + H(p) = \log m$$
- With fair odds, withholding part of the budget doesn't gain anything.
- If we have $Q = \sum_x q(x) < 1$, Kelly-betting ($b = p$) remains optimal in expectation. But the **Dutch book**
$$b(x) := \frac{q(x)}{Q} > q(x) \implies S(X) = \frac{b(X)}{q(X)} = \frac{1}{Q} > 1$$
has a **guaranteed** doubling rate $W(b) = -\log Q > 0$.

Kelly Criterion

Consider an offered bet, where we can bet $b \in [0, 1]$. We receive αb on a win and pay βb on a loss. Then

$$W(b) = p \log(1 + \alpha b) + (1 - p) \log(1 - \beta b)$$

$$b^* = \min\{1, \max\{0, b\}\}, \qquad b = \frac{p}{\beta} - \frac{1-p}{\alpha}$$

4 Processes

A semi-infinite sequence of random variables $X_1, X_2, ...$

- A process is **stationary** if, for any $n \in \mathbb{N}$ and any $\Delta \geq 0$
$$\mathbb{P}(X_1, ..., X_n) = \mathbb{P}(X_{1+\Delta}, ..., X_{n+\Delta})$$
- **Conditional Entropy Rate.**
$$H(X) = \lim_{t\to\infty} H(X_{t+1}|X_t, ..., X_1)$$
- X stationary $\implies$ $H(X)$ well-defined.
- **Entropy Rate.**
$$H'(X) = \lim_{t\to\infty} \frac{1}{t} H(X_1, ..., X_t)$$

4.1 Markov Chains

A **Markov Chain** is a stochastic process for which
$$X_{t+1} \perp X_{t-1}, ..., X_1 | X_t$$
Let X be a Markov Chain and $\pi := P(X_1)$.

- then by the independence from past and future
$$\mathbb{P}(X_1, ..., X_t) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1)\mathbb{P}(X_3|X_2) \cdots \mathbb{P}(X_t|X_{t-1})$$
- X is **time-homogeneous**, if
$$\mathbb{P}(X_{t+1}|X_t) = \mathbb{P}(X_2|X_1), \qquad \forall t \geq 1$$
- A time-homogeneous Markov Chain is fully characterized by its initial distribution and the **transition matrix** P with
$$P_{ij} := \mathbb{P}(X_2 = i | X_1 = j)$$
then
$$\mathbb{P}(X_{i+r} = b | X_i = a) = (P^r)_{ba}$$
- M.C. X **stationary** $\iff$ π stationary and X time-homogeneous $\iff$ $P\pi = \pi$
- **Entropy Rate** of a stationary time-homogeneous M.C.
$$H'(X) = H(X) = \sum_a \pi_a \left(-\sum_b P_{ba} \log P_{ba} \right)$$
Note that $H' = H$ if X is **stationary**.
Also, we have that $H(X) = \lim_{t\to\infty} H(X_{t+1}|X_t) = H(X_2|X_1)$
- A M.C. is **ergodic**, iff. $\exists t \geq 1$ s.t. $(P^t)_{ij} > 0, \forall i, j$.
- An M.C. **ergodic** $\iff$ has a unique stationary distribution
- For stationary M.C. $H(X_t|X_1) \leq H(X_{t+1}|X_1)$.

Reversible Chains.
For any finite Markove Chain X
$$\mathbb{P}(X_1, ..., X_t) = \mathbb{P}(X_t)\mathbb{P}(X_{t-1}|X_t) \cdots \mathbb{P}(X_1|X_2)$$
For a t.-h. M.C. X with stationary distribution $\pi > 0$ and transition matrix P, then the backwards transitions are characterized by
$$U_{ab} = P_{ba} \frac{\pi_a}{\pi_b}$$
and X is **reversible** $\iff$ $P = U$ $\iff$ $P_{ba}\pi_a = P_{ab}\pi_b$.

4.2 Random Walks on Graphs

Consider an undirected graph with nodes $\{1, ..., m\}$ and edge weights $w_{ab} = w_{ba} \geq 0$. We define a random walk as a Markov Chain
$$P_{ba} = \frac{w_{ab}}{W_a}, \qquad W_a = \sum_b w_{ab}$$

- It has a stationary distribution $\pi_a = \frac{W_a}{W}$ with $W = \sum_a W_a$.
- Graph connected $\implies$ this stationary distribution is unique.
- A random walk on an undirected graph is reversible.
- Every t.-r. M.C. is equivalent to a random walk on a graph.

Entropy Rate In an undirected graph without edge weights. E is the number of edges.
$$H(X) = \log(2E) - H(\frac{E_1}{2E}, \ldots, \frac{E_m}{2E})$$

4.3 Thermodynamics

Let (X, Y) and X', Y' be R.V. pairs over the same probability space, then
$$D(\mathbb{P}(X, Y)||\mathbb{P}(X', Y')) = D(\mathbb{P}(X)||\mathbb{P}(X')) + D(\mathbb{P}(Y|X)||\mathbb{P}(Y'|X'))$$
Let X be a t.-h. M.C. with μ, v different PMF over states, then
$$D(P\mu||Pv) \leq D(\mu||v) \quad \forall \mu, v$$
with π stationary
$$D(\mu||\pi) \geq D(P\mu||\pi)$$
if addionally X is reversible
$$D(\mu||\pi) > D(P\mu||\pi), \quad \forall \mu \neq \pi.$$

5 Universal Coding

A CDF F_X induces a partition of $[0; 1)$ into
$$\{I_x : x \in X(\Omega)\}, \qquad I_x := [F_X(x) - p_X(x); F_X(x))$$
For $I = [a, b) \subseteq I$ there exists $z \in I, (z)_2 = 0.z_1 z_2 ... z_l$ with $l = \lceil -\log(b - a) \rceil$ (*there's short representation for every Interval I*).

Shannon-Fano-Elias Codes

- pick midpoint $z_x = \sum_{x' < x} p_X(x') + \frac{1}{2} p_X(x)$
- truncate to $\ell_i = \lceil -\log p_X(x_i) \rceil + 1$ bits

Prefix-free, but not wasteful.

5.1 Arithmetic Coding

Note that Huffman Codes require knowledge of the distribution and cannot easiy be adapted to changing distributions.
Consider a stochastic process made of $X_t : \Omega \to A, A := \{0, ..., m-1\}$.
We define
$$Z = (0.X_1 X_2 ...)_m \in [0, 1] \subset \mathbb{R}$$
For any Z. If F_Z is a **bijection**, then $U := F_Z(Z) \sim \mathcal{U}([0, 1])$.
Then $F_U(u) = u$ and $U = (0.U_1 U_2 ...) \implies U_i \overset{iid}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right)$.

- encoder $x_1, x_2, ... \mapsto z = (0.x_1 x_2 ...)_m \overset{F_Z}{\mapsto} u = (0.u_1 u_2 ...)_2$

- decoder $u \overset{F_Z^{-1}}{\mapsto} z = (0.x_1x_2...)_m \mapsto x_1x_2$

Consider $X = (X_1, ..., X_n)$ with $X : \Omega \to A^n$ and thus m^n possible outcomes.
F_X induces a partitioning $\{I_x : x = (x_1, ..., x_n)\}$ as defined above.
The **arithmetic code** is given by the S.-F.-E. code for x.
For the arithmetic code C, we have

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}(|C(X_1, ..., X_n)|) = H(\{X_t\})$$

Incremental refinement of I_x for $x = (x_1, ..., x_n)$ into m subinterval using

$$\mathbb{P}(X_{n+1}|X_1 = x_1, ..., X_n = x_n)$$

Lempel-Ziv Code
We denote a string to be compressed as

$$x = x_{-s}x_{-s+1}...x_{-1}x_0x_1...x_t$$

where $x_0, ..., x_t$ still needs to be encoded.
Matching.

$$\text{match}(x) := \{(j, l) \mid x_{-j}...x_{-j+l-1} = x_0...x_{l-1} \wedge j, l \geq 1\}$$

The **Maximal Matching** is $(j^*, l^*) \in \text{match}(x)$ is maximal in l and as tiebreaker minimal in j.
LZ77.

$$c(x) = \begin{cases} (0, x_0)c(x_1...x_t) & \text{if match}(x) = \emptyset \\ (1, j^*, l^*)c(x_{l^*}...x_t) & \text{otherwise} \end{cases}$$

An integer x can be encoded with $\leq \log x + 2\log\log x + 4$ bits.

$$C'(x) = 00...01(x)_2, \quad |C'(x)| = 2\lceil \log x\rceil + 1$$

and we construct

$$C(x) = C'(\lceil \log x\rceil)1(x)_2, \quad |C(x)| \leq \log x + 2\log\log x + 4$$

LZ77 is optimal (i.e. reaches the shannon limit).

6 Channel Coding

A noisy channel is a conditional probability distribution $\mathbb{P}_{Y|X}$. The noise reduces the information from $H(X)$ to $I(X; Y)$.
Channel Capacity Given a channel with $\mathbb{P}_{Y|X}$ its capacity is

$$R^* = \max_{\mathbb{P}_X} I(X; Y)$$

BSC	$C = 1 - H(p)$
Binary erasure channel	$C = 1 - \alpha$
Symm. Ternary Channel	$C = \log_2(3) - H(p)$
Each col has 1 element $\neq 0$	$C = \log_2(X)$
General	$C = \log_2 \lvert Y\rvert - H(r)$

Channel Rate

$$\frac{\text{\# of message bits}}{\text{\# of code bits}}$$

Symmetric Channels A channel is said to be symmetric if the rows of the channel transition matrix $P_{Y|X}$ are permutations of each other and the columns are permutations of each other. A channel is said to be weakly symmetric if every row of the transition matrix is a permutation of every other row and all the column sums $\sum_x P_{Y|X}(y|x)$ are equal.
Parallel Channels

$$2^C = 2^{C_1} + 2^{C_2}$$

Binary Symmetric Channel.
A BSC(n, η) is a channel with $\mathcal{X} = \mathcal{Y} = \{0, 1\}^n$, with

$$\mathbb{P}(Y_{1:n}|X_{1:n}) = \prod_{i=1}^{n} \mathbb{P}(Y_i|X_i), \text{ where } Y_i = X_i \oplus N_i,\ N_i \overset{iid}{\sim} \text{Ber}(\eta)$$

The probability of communication error is $P_e = 1 - (1 - \eta)^n$.
A **codebook** $\mathcal{C} = \{x(1), ..., x(m)\} \subseteq \{0, 1\}^n$ maps each message $w \in [1 : m]$ to a unique binary codeword $x(w)$.
Bayes-optimal Decoder. For a codebook $\mathcal{C}$ the min. error probability decoder is given by

$$\hat{w} = g(y) = f^{-1}(\hat{x}), \quad \hat{x} = \arg\max_{x\in\mathcal{C}} \mathbb{P}(x|y)$$

For X uniform, the Bayes-optimal decoder is the **maximum likelihood** decoder, i.e. $\hat{x} = \arg\max_{x\in\mathcal{C}} \mathbb{P}(y|x)$.
Joint Typicality. Consider n pairs of R.V. $(X_i, Y_i) \overset{iid}{\sim} \mathbb{P}$. Then for any $\varepsilon > 0$ define the jointly typical sets via

$$\begin{aligned}\mathcal{B}_\varepsilon^n =& \{(x_{1:n}, y_{1:n}) : |H(X, Y) + \frac{1}{n}\log p(x_{1:n}, y_{1:n})| < \varepsilon \wedge \\ & |H(X) + \frac{1}{n}\log p(x_{1:n})| < \varepsilon \wedge |H(Y) + \frac{1}{n}\log p(y_{1:n})| < \varepsilon\}\end{aligned}$$

Joint AEP. In the same setting as above

(i) $\mathbb{P}(\mathcal{B}_\varepsilon^n) \to 1$, as $n \to \infty$

(ii) $\log |\mathcal{B}_\varepsilon^n| \leq n(H(X, Y) + \varepsilon)$, for all n large enough.

Decoding by Joint Typicality. Assume codeb. $\mathcal{C}$ and y received.

1. $\mathcal{C}_\varepsilon(y) := \{x \in \mathcal{C} : (x, y) \in \mathcal{B}_\varepsilon^n\}$
2. If $\mathcal{C}_\varepsilon(y) = \{x\}$, then decode $g(y) = f^{-1}(x)$.
3. otherwise declare an inability to decode by setting, $g(y) = 0$.

Random Codebooks. Generate a random codebook from $m \cdot n$ fair coin tosses

$$X_i(w) \overset{iid}{\sim} \text{Ber}\left(\frac{1}{2}\right), \qquad w \in [1 : m], i \in [1 : n]$$

Define events

$$E_w = (X(w), Y(1)) \in \mathcal{B}_\varepsilon^n$$

Expected probability error of typicality Decoding

$$P_\varepsilon = \mathbb{P}(E_1^c \cup E_2 \cup ... \cup E_m)$$

With Union Bound and typicality ($\mathbb{P}(E_1^c) \leq \varepsilon$ for all $n \geq n_0(\varepsilon)$)

$$P_\varepsilon \leq \mathbb{P}(E_1^c) + \sum_{i=2}^{m} \mathbb{P}(E_i) \leq \varepsilon + m\mathbb{P}(E_2)$$

To further bound the error we consider the following **Lemma:**
Let $(X_i, Y_i) \overset{iid}{\sim} \mathbb{P}(X, Y)$ and $(\overline{X}_i, \overline{Y}_i) \overset{iid}{\sim} \mathbb{P}(X)\mathbb{P}(Y), i \in [1 : n]$. For large enough n

$$\mathbb{P}((\overline{X}_{1:n}, \overline{Y}_{1:n}) \in \mathcal{B}_\varepsilon^n) \leq 2^{-n(I(X;Y)-3\varepsilon)}$$

With the lemma and noting $m = 2^{nR}$, we get for any $\varepsilon > 0$ and large enough n

$$P_\varepsilon \leq \varepsilon + 2^{nR}2^{-n(I(X;Y)-3\varepsilon)} \leq \varepsilon + 2^{-n\kappa}$$

with $\kappa = I(X; Y) - R - 3\varepsilon$. Thus as long as $R < I(X; Y)$ we have $P_\varepsilon \to 0$ for $n \to \infty$. Note then $\kappa > 0$ since $\varepsilon > 0$ can be arbitrary.

Note that P_ε is the **expected error** over all possible codebooks (uniformly chosen). To find a codebook with an error $\leq P_\varepsilon$ one could do an exhaustive search over all codebooks (**not practical**).
Channel Coding Theorem. Consider a BSC(n, η). For any rate $R < R^* = 1 - H(\eta)$ we can communicate $m = 2^{nR}$ distinct messages with an average error $P_\varepsilon \to 0$ for $n \to \infty$.
The **Hamming Distance** between $x, x' \in \{0, 1\}^n$ is the number of bits they differ.

$$d_H(x, x') = \sum_{i=1}^{n} x_i(1 - x'_i) + (1 - x_i)x'_i = \sum_{i=1}^{n}(x_i + x'_i) - 2x_ix'_i$$

In BSC(n, η) with $\mathbb{P}(X)$ **uniform (!)** and $\eta < \frac{1}{2}$, optimal decoding wrt. to $\mathcal{C}$ is characterized by

$$x^* = \arg\max_{x\in\mathcal{C}} \mathbb{P}(y|x) = \min_{x\in\mathcal{C}} d_H(x, y)$$

Detecting single-bit corruptions.
Let $\mathcal{C} = \{0, 1\}^n$ be a codebook and y be received with atmost one bit corruption from input x. We construct $\mathcal{C}' \subseteq \{0, 1\}^{n+1}$ with an appended parity bit for every $x \in \mathcal{C}$.

$$x \mapsto x' = \left(x_1, ..., x_n, \sum_{i=1}^{n} x_i \mod 2\right)$$

Then $\min_{x'\neq z'\in\mathcal{C}'} d_H(x', z') \geq 2$
Hamming Code.
Blocklength $n = 2^k - 1$ with k parity bits and $2^k - k - 1$ data bits.
The i-th parity bit is at position 2^{i-1} with $i \in [1 : k]$ and covers all positions where the i-th least significant bit is set (including itself).
The sum of positions of the mismatched parity bits gives the position of the corrupted bit.
The rate achieved is

$$R = \frac{2^k - k - 1}{2^k - 1}$$

7 Lossy Coding

Rate-Distortion Theory
Distortion measure $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, s.t. $d(x, x) = 0 (\forall x)$.
Examples include Hamming Distance and MSE. We consider $\mathbb{E}(d(X, \hat{X}))$, where X is original data and $\hat{X}$ its reconstruction.
Rate Distortion Theorem. The maximal rate $R(D)$ at which X ($\mathbb{P}(X)$ given) can be encoded with $\mathbb{E}(d(X, \hat{X})) \leq D$ is given by

$$R(D) = \underset{\mathbb{P}(\hat{X}|X):\mathbb{E}(d(X,\hat{X}))\leq D}{I(X, \hat{X})}$$

Consider the specific Case $X_t \overset{iid}{\sim} \text{Ber}(p)$ and $d(x^n, \hat{x}^n) = \frac{1}{n}d_H(x^n, \hat{x}^n)$.
Then requiring $\mathbb{P}(X_t \neq \hat{X}_t) \leq \eta$ (wlog. $\eta \leq p \leq 1/2$), and minimizing the mutual information $I(X; \hat{X})$ gives us a symmetric backwards channel

$$\mathbb{P}(X|\hat{X}) = \begin{bmatrix} 1-\eta & \eta \\ \eta & 1-\eta \end{bmatrix}, \text{ and thus } \hat{X}_t \overset{iid}{\sim} \text{Ber}(q) \text{ with } q = \frac{p-\eta}{1-2\eta}$$

which gives us a optimal forward channel (it's asymmetric).
This gives a rate of $R(\eta) = H(p) - H(\eta)$ as a specific case of the Rate-Distortion Theorem.
Distortion-Typicality. A pair $(x, \hat{x})$ is (ε, δ)-d-typical, if it is jointly ε-typical and $|\eta - d(x^n, \hat{x}^n)| < \delta$ (η is the expected bit error).
Uniform Quantization Let $U : \Omega \to R \subseteq \mathbb{R}$ (Scalar Quantization).

We partition R into intervals R_j of length Δ, assuming $|R| < \infty$. We can then approximate the pdf $p(u)$ by a step function $\hat{p}(u)$, which is constant over R_j.

$$\hat{p}(u) = \sum_j \mathbb{I}\{u \in R_j\} \frac{\mathbb{P}(u \in R_j)}{\Delta} = \sum_j \mathbb{I}\{u \in R_j\} \frac{\int_{R_j} p(u)du}{\Delta}$$

Let V be an RV characterized by $\hat{p}$. Then $\mathbb{E}((U-V)^2) \approx \frac{\Delta^2}{12}$.
If $U \sim \mathcal{U}([a,b])$ and we want the constant x minimizing MSE. Then the optimal point is $x = \frac{a+b}{2}$ with an MSE of $\frac{(a+b)^2}{12}$.

Differential Entropy

Let $U : \Omega \to R \subseteq \mathbb{R}$. The differential entropy is defined as

$$h(U) = -\int_R p(u) \log p(u) du$$

We have $H(V) \approx h(U) - \log \Delta$.
To be more precise we have

$$H(V) + \log \Delta \to h(U), \quad \text{as } \Delta \to 0$$

Let $U \sim \mathcal{N}(0, \sigma^2)$. Then if we accept a MSE of at most η, U can be quantized at a rate

$$R(\eta) = \begin{cases} \frac{1}{2}(\log \sigma^2 - \log \eta) & \text{if } \eta \leq \sigma^2 \\ 0 & \text{otherwise} \end{cases}$$

Shannon's Lower Bound. Let $U : \Omega \to R \subseteq \mathbb{R}$ with $h(U) < \infty$ and h^* the differential entropy $h^*(\sigma^2)$ of a gaussian. Then U can be encoded with MSE at most η at a rate $R > R(\eta)$, where

$$R(\eta) \geq h(U) - h*(\eta)$$

For $U : \Omega \to R \subseteq \mathbb{R}^d, d > 1$ we speak of *vector quantization*. Generally if we are given $\{y_1, ..., y_m\} \subseteq R$, the optimal quantizer V is characterized by

$$u \mapsto \hat{u} = y_k, \quad k \in \arg\min_j ||u - y_j||$$

This induces Voronoi cells around the y_j, which individually are convex regions (for $d = 1$ these are intervals).
Centroid Condition. Given a partition $\{R_j\}$ of $R \subseteq \mathbb{R}^n$, the optimal code points are

$$y_j = \mathbb{E}(U | U \in R_j), \text{ empirically w/ dataset } \mathcal{X} : \frac{\sum_{x \in \mathcal{X} \cap R_j} x}{|\mathcal{X} \cap R_j|}$$

Loyld's Algorithm

1. Generate random code points y_j at random and the induced Voronoi cells R_j.
2. For each cell, recompute the centroid. Then update the newly induced Voronoi cells. Repeat until convergence.

8 Estimation

Laplace Distribution is characterized by the pdf

$$p(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$

with parameters $\theta = (\mu, b) \in \Theta = \mathbb{R} \times \mathbb{R}_+$.
Any function ϕ of a RV X is called a **statistic** of X.
A statistic ϕ of X is **sufficient** for Y, if $X \perp Y | \phi(X)$.
It is **minimally sufficient**, if for all other statistics ϕ', $X \perp \phi'(X) | \phi(X)$. $p_\theta = h(x) g_\theta(x)$.

Exponential Family

$\phi(x)$ sufficient statistic (measures everything of x that influences the probability)

$h(x)$ positive function only depending on x

$A(\theta)$ $= \ln \int h(x) \exp(\theta \cdot \phi(x)) dx$

$$p(x; \theta) = h(x) \exp(\theta \cdot \phi(x) - A(\theta)),$$

	$h(x)$	$\phi(x)$	θ	$A(\theta)$
Ber	1	x	$\ln \frac{p}{1-p}$	$\ln(1 + e^\theta)$
Bin	$\binom{n}{x}$	x	$\ln \frac{p}{1-p}$	$n \ln(1 + e^\theta)$
Norm	$\frac{1}{\sqrt{2}}$	$(x, x^2)^\top$	$\left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$	$-\frac{\theta_1^2}{4\theta_2} + \ln\left(\frac{1}{\sqrt{-2\theta_2}}\right)$
Exp	1	$-x$	λ	$-\ln \lambda = -\ln \theta$

Let $X_1, ..., X_n$ be iid. with pdf $p(x; \theta)$ in an exponential family. Then the joint distribution is also an expenonetial family with sufficient statistic $\phi(X_1, ..., X_n) = \sum_{i=1}^n \phi(X_i)$.
iff. the distributions can be represented as an exponential family.

8.1 MLE

For a parametric family given iid $x_1, ..., x_n$:

$$\hat{\theta} = \arg\max \sum_{i=1}^n \ln p_\theta(x_i)$$

Binomial	$\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$
Normal	$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i, \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$
Laplace	$\hat{\mu} = \text{median}(x_1, ..., x_n), \hat{b} = \frac{1}{n} \sum_{i=1}^n \lvert x_i - \hat{\mu} \rvert$

For an exponential family we have

$$\nabla_\theta \sum_{i=1}^n \log p_\theta(x_i) = 0 \iff \mathbb{E}_\theta(\phi(X)) = \frac{1}{n} \sum_{i=1}^n \phi(x_i)$$

$$\frac{d}{d\theta} A(\hat{\theta}) = \frac{1}{s} \sum_{t=1}^s \phi(x_t)$$

8.2 Maximum Entropy Inference

Given constraint functions $\phi_i : R \subseteq \mathbb{R} \to \mathbb{R}$ and empirical values $\bar{\phi}_i$. Consider

$$\mathcal{P} = \{p : R \to [0,1] \mid \mathbb{E}_p(\phi_i(X)) = \bar{\phi}_i(\forall i)\} \neq \emptyset$$

Define the exponential family $p_\theta(x) = \exp(\phi(x) \cdot \theta - A(\theta))$.
Let $p_{\hat{\theta}}$ be the pdf obtained for the MLE $\hat{\theta}$ based on data summaries $\bar{\phi}_i$, then

$$p_{\hat{\theta}} = \arg\max_{p \in \mathcal{P}} h(p), \text{ with } h(p) = -\int p(x) \ln p(x) d$$

8.3 Fisher Information

Let p_θ be a parametric family of pdfs.
For an estimator $\hat{\theta}$ and the true parameter θ^*, we define $\text{MSE}(\hat{\theta}) = \mathbb{E}((\hat{\theta}(X_1, ..., X_n) - \theta^*)^2)$.
The **score** is an RV $S(X) = \nabla_\theta \log p_\theta(X)$.
We have $\mathbb{E}(S(X)|\theta) = 0$, and the **Fisher Information** is symmetric positive semi-definite matrix $\mathcal{I}(\theta) = \mathbb{E}(S(X)S(X)^\top | \theta)$.
For $X_1, ..., X_n \overset{iid}{\sim} p_\theta$ we have $S(X_1, ..., X_n) = nS(X_1)$.
For $X, Y \sim p_\theta$ we have $S(X, Y) = S(X|Y) + S(Y)$ and by that $\mathcal{I}_{X,Y}(\theta) = \mathcal{I}_{Y|X}(\theta) + \mathcal{I}_X(\theta)$.
If p_θ twice-differentiable, then $\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \ln(p_\theta(X))\right)^2\right]$.

8.4 Cramér-Rao Bound

Let $p_\theta(x)$ be a parameterized family of pdfs for a real-valued RV X. Let $\hat{\theta} = T(X)$ be an unbiased estimator, i.e. $\mathbb{E}_\theta(T(X) - \theta) = 0$, then

$$\text{MSE}(\hat{\theta}) \geq \mathcal{I}(\theta)^{-1}$$

Unbiased estimator $\hat{\theta}$ is **efficient**, if attains Cramér-Rao w/ equality.
Rao-Blackwell Theorem Given an estimator $\hat{\theta}$ for the parameter of a family with sufficient statistics T define an estimator $\bar{\theta} = \mathbb{E}(\hat{\theta}|T)$, then

$$\text{MSE}(\bar{\theta}) \leq \text{MSE}(\hat{\theta})$$

9 Language

Perplexity

When p is the distribution of the test data, the perplexity $PP(q)$ of a model q is

$$PP(q) = 2^{H(p;q)}$$

10 Examples

Exponential distribution $f(x; \lambda) = \lambda e^{-\lambda x}$

Normal distribution $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
Bottleneck channel Two channels in series, $X \to V \to Y$ with $v = \{1, 2, ..., k\}$.
Then $C = max_{p(x)} I(X;Y)$, and $I(X;Y) \leq I(V;Y) = H(V) - H(V|Y) \leq H(V) \leq \log k$ by data processing inequality.
Calculate Entropy

$$H(\frac{1}{3}, \frac{2}{3}) = -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} = \frac{1}{3} \log 3 + \frac{2}{3} \log 3 - \frac{2}{3} \log 2$$

Arithmetic Codes for Markov Chains

$$F(x^n) = \sum_{k=1}^n p(x^{k-1}0) x_k$$

Find R(D) for $x \sim \text{Ber}(\frac{1}{2})$, $d(x, \hat{x}) = \begin{bmatrix} 0 & \infty \\ 1 & 0 \end{bmatrix}$

1. $p(x, \hat{x}) = \begin{bmatrix} \frac{1}{2} & 0 \\ D & \frac{1}{2} - D \end{bmatrix}$
2. $R(D) = 1 + \frac{1}{2} \log \frac{1/2}{1/2+D} + D \cdot \log \frac{D}{1/2+D}$ if $0 \leq D \leq \frac{1}{2}$