1 Distribution

$\mathrm{Bin}(n,p): P_X(x) = \binom{n}{x} p^x q^{n-x}$ for $x \in [n]_0$.
$\mu = np, \sigma^2 = npq, H(X) = \frac{1}{2}\log(2\pi enpq) + O(\frac{1}{n})$.
$\mathrm{Pois}(\lambda): P_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x \in \mathbb{N}_0$.
$\mu = \sigma^2 = \lambda$.
$\mathrm{Geo}(p): P_X(x) = q^{x-1}p$ for $x \in \mathbb{N}$.
$\mu = \frac{1}{p}, \sigma^2 = \frac{q}{p^2}, H(X) = \frac{-q\log q - p\log p}{p}$.
$\mathrm{Exp}(\lambda): f_X(x) = \lambda e^{-\lambda x}$ for $x \in \mathbb{R}_0^+$.
$\mu = \frac{1}{\lambda}, \sigma^2 = \frac{1}{\lambda}$.
$\mathcal{N}(\mu,\sigma^2): f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.
$h(X) = \frac{1}{2}\log(2\pi e\sigma^2)$.
$\mathrm{Lap}(\mu,b): f_X(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$.
$\sigma^2 = 2b^2, h(X) = \log(2be)$.

2 Markov Chain

$X_1 - X_2 - \cdots - X_n := \forall n, x^n,\ P_{X_{n+1}|X^n}(x_{n+1}|x^n) = P_{X_{n+1}|X_n}(x_{n+1}|x_n)$.
Stationary: $P_{X_1,\ldots,X_n} = P_{X_{1+l},\ldots,X_{n+l}},\ \forall n,l \in \mathbb{N}$.

3 Central Limit Theorem

Khinchin WLLN: $X_1, X_2, \ldots,$ are i.i.d. with $E[|X_i|] < \infty$, then $\forall \epsilon > 0,\ \lim_{n\to\infty} \Pr\{|\bar{X}_n - \mu| \geq \epsilon\} = 0$.
Central limit theorem: $X_1, X_2, \ldots,$ are i.i.d. with $E[|X_i|] < \infty$, then $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathrm{d}} Z \sim N(0,1)$.
Berry-Esseen: $X_1, X_2, \ldots,$ are i.i.d. with $E[|X_i - \mu|^3] = \rho_3 < \infty$. Let $Z_n := \frac{S_n - n\mu}{\sqrt{n}\sigma}, Z \sim N(0,1)$. Then $|F_{Z_n}(z) - F_Z(z)| \leq c\frac{\rho_3}{\sigma^3}n^{-1/2},\ \forall z \in \mathbb{R}, n \in \mathbb{N}$ for constant $c \in (0.4, 0.5)$.

4 Representing An i.i.d. Sequence Almost Losslessly

DMS: discrete memoryless source. $\mathcal{B}(n,\epsilon)$ is an ϵ-high-probability set: $\Pr\{S^n \in \mathcal{B}(n,\epsilon)\} \geq 1-\epsilon$
s^n is δ-typical: $|\frac{1}{n}\sum_{i=1}^n \log P_S(s_i) + H(S)| \leq \delta$.
δ-typical set $\mathcal{A}_\delta^{(n)}(S) := \{s^n | s^n \text{ is } \delta\text{-typical}\}$.
Properties of typical sequences and typical sets:

- $\forall s^n \in \mathcal{A}_\delta^{(n)}(S), 2^{-n(H(S)+\delta)} \leq \Pr\{S^n = s^n\} \leq 2^{-n(H(S)-\delta)}$.
- $\Pr\{S^n \in \mathcal{A}_\delta^{(n)}(S)\} \geq 1-\epsilon$ for n large enough.
- $|\mathcal{A}_\delta^{(n)}(S)| \leq 2^{n(H(S)+\delta)}$.
- $|\mathcal{A}_\delta^{(n)}(S)| \geq (1-\epsilon)2^{n(H(S)-\delta)}$ for n large enough.

$s^n \to b^k \to \hat{s}^n$: (n,k) code.
(n,k,ϵ) code: (n,k) code with $P_e^{(n)} := \Pr\{S^n \neq \hat{S}^n\} \leq \epsilon$.
$k^*(n,\epsilon)$: the smallest k s.t. $\exists(n,k,\epsilon)$ code.
$R^*(\epsilon) := \lim_{n\to\infty} \frac{k^*(n,\epsilon)}{n}$.
A lossless source coding theorem for DMS: $R^*(\epsilon) = H(S),\ \forall \epsilon \in (0,1)$.
AEP (Asymptotic Equipartition Property): Entropy determines the asymptotic size of a typical set, and determines the probability of a typical sequence asymptotically.

5 Entropy

$H(X|Y) = \sum_y P_Y(y)H(X|Y=y) = \sum_{x,y} P_{X,Y}(x,y)\log\frac{1}{P_{X|Y}(x,y)}$.
$0 \leq H(X) \leq \log|\mathcal{X}|$, where $H(X) = \log|\mathcal{X}| \iff$

X is uniform distributed over $\mathcal{X}$.

$H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$.

$H(X|Y) \le H(X)$, but $H(X|Y=y)$ may $> H(X)$.

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

$H(X|Y,Z) \le H(X|Y)$.

The above still holds for h.

Exercise 4: $H(X,Y,Z) \le H(X,Y) + H(X,Z) - H(X)$.

Concavity of Entropy: $H(\mathbf{p}) := -\sum_{i=1}^{d} p_i \log p_i$ is concave in $\mathbf{p}$.

That is, $H(\lambda\mathbf{p_1} + (1-\lambda)\mathbf{p_2}) \ge \lambda H(\mathbf{p_1}) + (1-\lambda)H(\mathbf{p_2})$.

Fano's inequality: $H(U|V) \le H_b(P_e) + P_e \log|\mathcal{U}|$, where $P_e := \Pr\{U \ne V\}$.

$\Rightarrow \Pr\{U \ne V\} \ge \dfrac{H(U|V) - 1}{\log|\mathcal{U}|}$.

Exercise 5: if U, V both take values in $\mathcal{U}$, then $H(U|V) \le H_b(P_e) + P_e \log(|\mathcal{U}| - 1)$.

6 Representing A Sequence with Memory Almost Losslessly

Entropy rate:

- $\mathrm{H}(\{X_i\}) := \lim_{n\to\infty} \frac{1}{n} H(X_1,\dots,X_n)$ if exists.
- $\tilde{\mathrm{H}}(\{X_i\}) := \lim_{n\to\infty} H(X_n|X^{n-1})$ if exists.

H and $\tilde{\mathrm{H}}$ may be different: consider $X_1, X_3, \dots$ are i.i.d. and $X_{2k} = X_{2k-1}$.

If $\{X_i\}$ is stationary, then $H(X_n|X^{n-1})$ is decreasing in n.

If $\{X_i\}$ is stationary, then $\mathrm{H}(\{X_i\}) = \tilde{\mathrm{H}}(\{X_i\})$.

Stationary ergodic processes: $\frac{1}{n}\sum_{l=0}^{n-1} f(X_{k_1+l},\dots,X_{k_m+l}) \overset{\text{a.s.},L^1}{\to} \mathrm{E}[f(X_{k_1},\dots,X_{k_m})]$ as $n \to \infty$.

Shannon-McMillan-Breiman theorem: if $\{S_i\}$ is stationary ergodic, then $\frac{1}{n}\log\frac{1}{P(S^n)} \overset{\text{a.s.},L^1}{\to} \mathrm{H}(\{S_i\})$ as $n \to \infty$.

A Lossless Source Coding Theorem for Ergodic DSS: For a discrete stationary ergodic source $\{S_i\}$, $R^*(\epsilon) = H(\{S_i\}) \forall \epsilon \in (0,1)$.

Let $\mathcal{X}$ be the state space of a Markov process.

1. A Markov process is irreducible if $\forall x, y \in \mathcal{X}$, it is possible to reach to start at x and reach y in a finite number of steps.

2. The period of a state is the g.c.d. of the # of times that a state can return to itself. A Markov process is aperiodic if all states have period $= 1$.

3. A Markov process is homogeneous (or time-invariant) if $\forall n > 1$, $P_{X_n|X_{n-1}} = P_{X_2|X_1}$. Hence, a homogeneous Markov process is completely defined by its initial state distribution P_{X_1} and transition probability $P_{X_2|X_1}$.

4. A steady-state distribution $\pi : \mathcal{X} \to [0,1]$ is one such that the distribution does not change after one transition: $\pi(x) = \sum_{y\in X} \pi(y) P_{X_{n+1}|X_n}(x|y)$, $\forall x \in \mathcal{X}$, $n \in \mathbb{N}$. For a finite-alphabet homogeneous Markov process, steady-state distribution always exists, and it is unique if the process is irreducible.

5. For a finite-alphabet homogeneous Markov process that is both irreducible and aperiodic, $\lim_{n\to\infty} \Pr\{X_{n+1} = y | X_1 = x\} = \pi(y)$, $\forall x, y \in \mathcal{X}$, where $\pi(\cdot)$ is the unique steady-state distribution. If $P_{X_1} = \pi$, the Markov process becomes a stationary process.

For a homogeneous, irreducible, and aperiodic Markov process $\{X_i\}$, $\mathrm{H}(\{X_i\}) = \tilde{\mathrm{H}}(\{X_i\}) = H(X_2|X_1)|_{P_{X_1}=\pi} = \sum_{x\in\mathcal{X}} \pi(x)H(X_2|X_1 = x)$, where π is the unique steady-state distribution.

7 Information for Continuous Distributions

The covariance of n-dimensional X is k, then $h(X) \le h(X^G) = \frac{1}{2}\log((2\pi e)^n \det(k))$.

8 Learning a Bit of Information

$\pi_{1|0}(\phi)$: false alarm, false positive, false rejection, type I error.

$\pi_{0|1}(\phi)$: miss detection, false negative, false acceptance, type II error.

$\mathcal{A}_\theta(\phi)$: acceptance region of H_θ.

Likelihood ratio $LR(x) := \frac{P_1(x)}{P_0(x)}$, Log likelihood ratio $LLR(x) := \log LR(x)$.

Likelihood ratio test (LRT) with parameter $\tau \in \mathbb{R}_0^+$ is $\phi_\tau^{LRT}(x) := \mathbb{I}\{LR(x) > \tau\}$.

(Randomized) LRT $\phi_{\gamma,\tau}(x) = \begin{cases} 1, \text{ if } LR(x) > \tau \\ \gamma, \text{ if } LR(x) = \tau \\ 0, \text{ if } LR(x) < \tau \end{cases}$.

Neyman-Pearson problem: minimize $\pi_{0|1}(\phi)$ subject to $\pi_{1|0}(\phi) \le \epsilon$.

Neyman-Pearson: LRT is optimal.

Generalized to n i.i.d.: $\phi^n_{\eta_n,\gamma_n}(x^n) = \begin{cases} 1/0, \text{ if } \sum_{i=1}^n LLR(x_i) \gtrless \eta_n \\ \gamma_n, \text{ if } \sum_{i=1}^n LLR(x_i) = \eta_n \end{cases}$.

Chernoff-Stein lemma: $\lim_{n\to\infty} -\frac{1}{n}\log\omega^*_{0|1}(n,\epsilon) = D(P_0\|P_1)$.

Typical set:

9 Information Divergence

$D(P\|Q) := \sum_a P(a)\log\frac{P(a)}{Q(a)}$.

$D(P\|Q) \ge 0$, with equality $\iff P(x) = Q(x),\ \forall x$.

$D(P_{Y|X}\|Q_{Y|X}|P_X) := \mathrm{E}_{X\sim P_X}[D(P_{Y|X}(\cdot|X)\|Q_{Y|X}(\cdot|X))]$.

Chain rule for information divergence: $D(P_{X,Y}\|Q_{X,Y}) = D(P_{Y|X}\|Q_{Y|X}|P_X) + D(P_X\|Q_X)$.

$D(P_Y\|Q_Y) \le D(P_{Y|X}\|Q_{Y|X}|P_X)$, with equality iff $D(P_{X|Y}\|Q_{X|Y}|P_Y) = 0$.

Donsker-Varadhan theorem: $D(P\|Q) = \max_{f:\mathcal{X}\to\mathbb{R}} E_{X\sim P}[f(X)] - \log \mathrm{E}_{X\sim Q}[2^{f(X)}]$ s.t. $E_{X\sim Q}[2^{f(X)}] < \infty$.

10 Error Exponents and Chernoff Information

P_0, P_1 are given.

$P_\lambda(a) := \frac{P_0(a)^{1-\lambda}P_1(a)^\lambda}{\sum_b P_0(b)^{1-\lambda}P_1(b)^\lambda}$.

Exercise 6: $D(P_\lambda\|P_0)$ is a continuous and strictly increasing function of λ for $\lambda \in [0,1)$.

$P_e^*(\pi(=(\pi_0,\pi_1)),n) := \min_\phi\{\pi_0\pi^{(n)}_{1|0}(\phi) + \pi_1\pi^{(n)}_{0|1}(\phi)\}$.

$\bar{P}_e^*(n) := \min_\phi\{\max\{\pi^{(n)}_{1|0}, \pi^{(n)}_{0|1}\}\}$.

Chernoff Information: $CI(P_0,P_1) :=$

$$\max_{\lambda\in(0,1)} \underbrace{-\log\sum_{a\in\mathcal{X}} P_0(a)^{1-\lambda}P_1(a)^{\lambda}}_{f(\lambda)}.$$

Theorem 11: $\lim_{n\to\infty}\{-\frac{1}{n}\log P_e^*(\pi,n)\} = \lim_{n\to\infty}\{-\frac{1}{n}\log \bar{P}_e^*(n)\} = CI(P_0,P_1)$.

11 Deviverling Information Reliably

$BSC(p)$: flip the bit bit i.i.d. with probability $p\in(0,\frac{1}{2})$.

12 Mutual Information

$I(X;Y) = D(P_{X,Y}\|P_X\times P_Y)$.

Exercise 1: $I(X;Y) = \min_{Q_Y:D(P_Y\|Q_Y)<\infty} D(P_{Y|X}\|Q_Y|P_X)$.

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

Chain rule: $I(X;Y^n) = \sum_{i=1}^{n} I(X;Y_i|Y^{i-1})$.

$X-Y-Z$, then $I(X;Y)\geq I(X;Z)$.

$X-Y-Z$, then $I(X;Y)\geq I(X;Y|Z)$.

13 Noisy Channel Coding Theorem

An (n,k) code with $P_e^{(n)} := \Pr\{W\neq\hat{W}\}\leq\epsilon$ is called an (n,k,ϵ) code.

$k^*(n,k)$ is the largest k s.t. $\exists(n,k,\epsilon)$ code.

$C(\epsilon) := \lim_{n\to\infty}\frac{1}{n}k^*(n,\epsilon)$.

Channel coding theorem for DMC without feedback: $C(\epsilon) = C^I := \max_{P_X} I(X;Y),\ \forall\epsilon\in(0,1)$.

x^n is robust typical sequence: $|\hat{P}_{x^n}(a) - P_X(a)|\leq\epsilon P_X(a)$, where $\hat{P}_{x^n}(a) := \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{x_i=a\}$.

The set of ϵ-robust typical sequence with respect to X: $\mathcal{T}_\epsilon^{(n)}(X)$.

14 Channel Coding with a Cost Constraint

Constraint: $\frac{1}{n}\sum_{i=1}^{n} b(x_i)\leq B$.

$(n,\lceil nR\rceil,B)$ code.

$C(B) := \sup\{R|R:\text{achievable }\}$.

Channel coding for DMC with average input cost constraint: $C(B) = C^I(B) := \max_{P_X:\mathrm{E}_{P_X}[b(X)]\leq B} I(X;Y)$.

The above also holds for CMC.

$C^I(B)$ is non-decreasing, concave, continuous in B.

AWGN (additive with Gaussian noise) channel: noise is Gaussian and independent of others, and constraint: $\frac{1}{n}\sum_{i=1}^{n}|x_i|^2\leq B$.

The capacity of the AWGN channel with input power constraint B and noise variance σ^2 is given by $C(B) = \sup_{X:E[|X^2|]\leq B} I(X;Y) = \frac{1}{2}\log(1+\frac{B}{\sigma^2})$, which is achieved by $X\sim N(0,B)$.

Proposition 2: $X^G\sim N(0,B)$, $Y = X^G+Z$ where $\mathrm{Var}[Z]=\sigma^2$, $Z\perp X^G$, then $I(X^G;Y)\geq\frac{1}{2}\log(1+\frac{B}{\sigma^2})$.

15 Lossy Source Coding

$d(s^n,\hat{s}^n) := \frac{1}{n}\sum_{i=1}^{n} d(s_i,\hat{s}_i)$, where $d(s,\hat{s}) := (s-\hat{s})^2$.

(R,D) achievable: $\exists$ sequence of $(n,\lfloor nR\rfloor)$ codes s.t. $\limsup_{n\to\infty} D^{(n)}\leq D$.

Rate distortion function $R(D) := \inf\{R|(R,D):\text{achievable }\}$.

$D_{\min} := \min_{\hat{s}(s)} \mathrm{E}[d(S, \hat{s}(S))]$.

$D_{\max} := \min_{\hat{s}} \mathrm{E}[d(S, \hat{s})]$.

$R(D) = R^I(D) := \min_{P_{\hat{S}|S}: \mathrm{E}[d(S,\hat{S})] \le D} I(S; \hat{S})$.

$R^I(D_{\min}) \le H(S), R^I(D) = 0$ if $D \ge D_{\max}$.

Ber(p) source: $R(D) = \begin{cases} H_b(p) - H_b(D), \text{ if } 0 \le D \le \min\{p, 1-p\} \\ 0, \text{ if } D > \min\{p, 1-p\} \end{cases}$.

Gaussian source: $R(D) = \begin{cases} \frac{1}{2}\log(\frac{\sigma^2}{D}), \text{ if } 0 \le D \le \sigma^2 \\ 0, \text{ if } D > \sigma^2 \end{cases}$.

$R(D) \le R^G(D)$.