Lecture 11
Final Exam Review Session

- Basics of Probability
- Basics of Linear Algebra
- Basics of Statistical Inference
- Basics of Information Theory
- Basics of Stochastic Processes
- Inequalities in Information Theory

Contents

Common functions of random variables

proof: $\mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2 + (\mathbb{E}[X])^2 - 2X\mathbb{E}[X]] = \mathbb{E}[X^2] + \mathbb{E}[(\mathbb{E}[X])^2] - \mathbb{E}[2X\mathbb{E}[X]]$

$= \mathbb{E}[X^2] + (\mathbb{E}[X])^2 - 2\mathbb{E}[X]\mathbb{E}[X]$

$= \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

Random Variable X	Discrete	Continuous
Cumulative distribution function (cdf)	$F(a) = P\{X \leq a\}$	$F(a) = \int_{-\infty}^{a} f(x)\mathrm{d}x$
Probability mass function (pmf) or Probability density function (pdf)	$p(x) = P\{X = x\}$	$f(x) = \frac{\mathrm{d}}{\mathrm{d}x}F(x)$
Expected value $\mathbb{E}[X]$	$\sum_{x:\, p(x)>0} xp(x)$	$\int_{-\infty}^{\infty} xf(x)\mathrm{d}x$
Expected value of $g(x)$, $\mathbb{E}[g(x)]$	$\sum_{x:\, p(x)>0} g(x)p(x)$	$\int_{-\infty}^{\infty} g(x)f(x)\mathrm{d}x$
Variance of X, $\mathrm{Var}(X)$	$\mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$	
Standard deviation of X, $\mathrm{std}(X)$	$\sqrt{\mathrm{Var}(X)}$	

Common discrete random variables

Name	Probability mass function (pmf)	$E[X]$	$\text{var}(X)$
Uniform	$P(x)=\frac{1}{b-a+1},\quad x=a,a+1,\cdots,b$	$\frac{b+a}{2}$	$\frac{(b-a+1)^2-1}{12}$
Binomial	$P(x)=\binom{n}{x}p^x(1-p)^{n-x},\quad x=0,1,\cdots,n$	np	$np(1-p)$
Poisson	$P(x)=\frac{e^{-\lambda t}(\lambda t)^x}{x!},\quad x=0,1,\cdots^{20}$	λt	λt
Geometric	$P(x)=(1-p)^{x-1}p,\ x=1,2,\cdots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative Binomial	$P(x)=\binom{x-1}{r-1}p^r(1-p)^{x-r},\ x=r,r+1,\cdots$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$

rate λt, pmf: $\frac{e^{-\lambda t}(\lambda t)^x}{x!}$, $x=0,1,\cdots$

$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

Common continuous random variables

Name	Probability density function (pdf)	$E[X]$	$\text{var}(X)$
Uniform	$\frac{1}{b-a}$, $a \le x \le b$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, $x \in (-\infty, \infty)$	μ	σ^2
Exponential	$\lambda e^{-\lambda x}$, $x \ge 0$	$1/\lambda$	$1/\lambda^2$
Gamma	$\frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}$, $x \ge 0$, $\Gamma(a) = \int_0^\infty e^{-y} y^{a-1}$	α/λ	α/λ^2
Beta	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, $0 < x < 1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

$f(x) = \lambda e^{-\lambda x}$ is well-defined pdf:

$$\int_0^{+\infty} f(x)\,dx = \int_0^{+\infty} \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_0^{+\infty} = 1$$

Sample Question

1. (20 points) Roll a fair four-sided die twice. Let X be the outcome on the first roll, and Y be the sum of the two rolls. Calculate

   (i) $\mu_X = \mathbb{E}[X]$. (4 points)

   (ii) $\mu_Y = \mathbb{E}[Y]$. (4 points)

   (iii) $\sigma_X^2 = \mathrm{Var}(X)$. (4 points)

   (iv) $\sigma_Y^2 = \mathrm{Var}(Y)$. (4 points)

   (v) $\mathrm{Cov}(X, Y) := \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. (4 points)

Z: second outcome

$Y = X + Z$

(i) $\mu_X = \frac{1}{4}(1+2+3+4) = 2.5$

(ii) $E[Y] = E[X+Z] = E[X] + E[Z] = 5$

(iii) $Var(X) = E[X^2] - (E[X])^2$

$E[X^2] = \frac{1}{4}(1^2+2^2+3^2+4^2) = \frac{30}{4} = 7.5$

$Var(X) = 7.5 - 2.5^2 = 1.25$

(iv) $Var(Y) = Var(X+Z) = Var(X) + Var(Z) = 2.5$

(v) $E[XY] = E[X(X+Z)] = E[X^2] + E[XZ]$

$= 7.5 + E[X]E[Z]$

$= 7.5 + 2.5^2$

$E[X]E[Y] = 2.5 \times 5 = 12.5$

$\Rightarrow Cov(X,Y) = 7.5 + 6.25 - 12.5 = 1.25$

Contents

Definition

- Let A be an $n \times n$ matrix.
- A scalar λ is said to be an **eigenvalue** of A if there exists a nonzero vector $\mathbf{x}$ such that

$$A\mathbf{x} = \lambda\mathbf{x}.$$

eigenvalue

eigenvector associated with λ

- The vector $\mathbf{x}$ is said to be an **eigenvector belonging to** λ.

(λ, x) is an eigen-pair of A

Sample Question

(w) $A^T = Q \Lambda Q^T$

2. (20 points) Suppose that A admits eigenvalue decomposition with eigenvalues $\lambda_1, \ldots, \lambda_n$. Show that

   (i) $\det(A) = \prod_{i=1}^n \lambda_i$. (5 points)

   (ii) $\text{Trace}(A) = \sum_{i=1}^n \lambda_i$. (5 points)

   (iii) The eigenvalues of A^k are $\lambda_1^k, \ldots, \lambda_n^k$. (5 points)

   (iv) The eigenvalues of $A^\top$ are $\lambda_1, \ldots, \lambda_n$ as well. (5 points)

$A = Q \Lambda Q^T$ s.t. $Q Q^T = I_d$, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$

(i) $\det(A) = \det(Q \Lambda Q^T) = \det(\Lambda Q^T Q) = \det(\Lambda)\det(Q^T Q) = \det(\Lambda)\det(Q Q^T)$

$= (\prod \lambda_i)\det(I_d) = \prod_{i=1}^n \lambda_i$

$\det(AB) = \det(BA)$

(ii) $\text{Tr}(A) = \text{Tr}(Q \Lambda Q^T) = \text{Tr}(\Lambda Q^T Q) = \text{Tr}(\Lambda) = \sum_{i=1}^n \lambda_i$

(iii) $A^2 = (Q \Lambda Q^T)(Q \Lambda Q^T) = Q \Lambda (Q^T Q) \Lambda Q^T = Q \Lambda^2 Q^T \Rightarrow A^k = Q \Lambda^k Q^T$, $\Lambda^k = \text{diag}(\lambda_1^k, \cdots, \lambda_n^k)$

Contents

Method of Maximum Likelihood

Suppose that X is a random variable with probability distribution $f(x; \theta)$, where θ is a single unknown parameter. Let $x_1, x_2, \ldots, x_n$ be the observed values in a random sample of size n. Then the **likelihood function** of the sample is

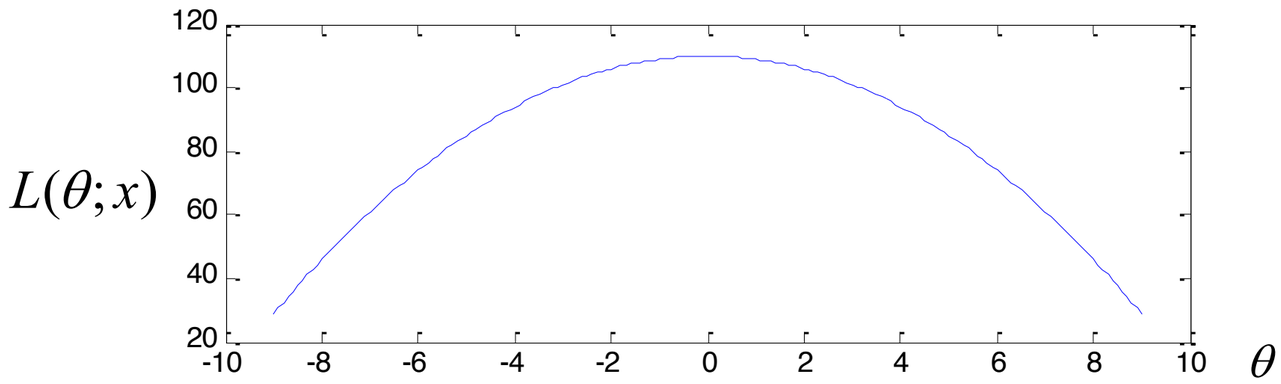
$$L(\theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \cdots \cdot f(x_n; \theta) \qquad (7\text{-}5)$$

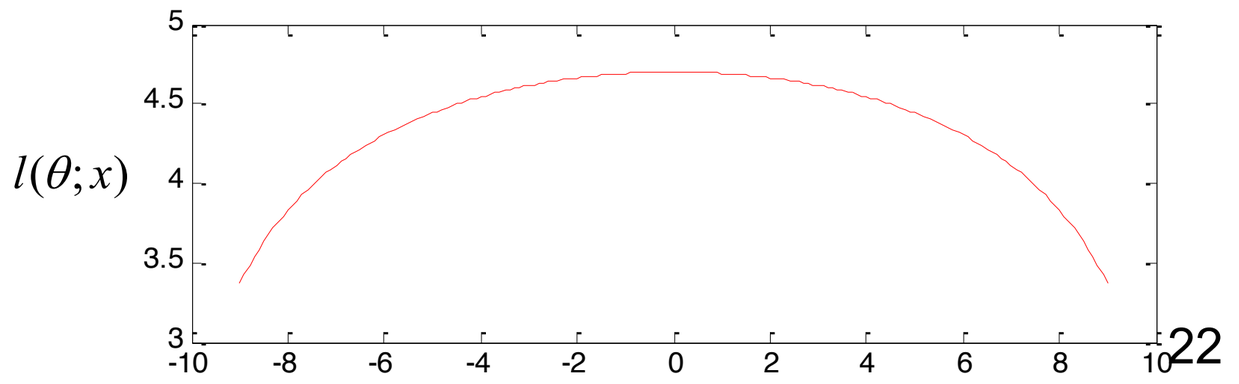
Note that the likelihood function is now a function of only the unknown parameter θ. The **maximum likelihood estimator** of θ is the value of θ that maximizes the likelihood function $L(\theta)$.

$$L(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta) = f(x_1; \theta) \ldots f(x_n; \theta)$$

$$l(\theta; x) = \sum_{i=1}^{n} \log[f(x_i; \theta)]$$

$$\hat{\Theta}(x) = \arg\max_{\theta} L(\theta; x) = \arg\max_{\theta} l(\theta; x)$$





7-61. A random variable x has probability density function

$$f(x;\theta) = \frac{1}{2\theta^3} x^2 e^{-x/\theta}, \qquad 0 < x < \infty, \quad 0 < \theta < \infty$$

Given samples $x_1, \ldots x_n$,
find the maximum likelihood estimator for θ

Example: Bernoulli

Let X be a Bernoulli random variable. The probability mass function is

$$f(x;p) = \begin{cases} p^x(1-p)^{1-x}, & x = 0, 1 \\ 0, & \text{otherwise} \end{cases}$$

where p is the parameter to be estimated. The likelihood function of a random sample of size n is

$$L(p) = p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\cdots p^{x_n}(1-p)^{1-x_n}$$
$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

$$\longrightarrow \quad \ln L(p) = \left(\sum_{i=1}^{n} x_i\right)\ln p + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p)$$

$$\longrightarrow \quad \frac{d\ln L(p)}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\left(n - \sum_{i=1}^{n} x_i\right)}{1-p} \quad \longrightarrow \quad \hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Example: normal

Let X be normally distributed with unknown μ and known variance σ^2. The likelihood function of a random sample of size n, say $X_1, X_2, \ldots, X_n$, is

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/(2\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)\sum_{i=1}^{n}(x_i-\mu)^2}$$

Now

$$\ln L(\mu) = -(n/2)\ln(2\pi\sigma^2) - (2\sigma^2)^{-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

and

$$\frac{d \ln L(\mu)}{d\mu} = (\sigma^2)^{-1}\sum_{i=1}^{n}(x_i - \mu)$$

$\longrightarrow$ What is the MLE for μ?

Example (Continued, unknown variance)

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

The solutions to the above equation yield the maximum likelihood estimators

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

MLE: Exponential

Let *X* be a exponential random variable with parameter λ. The likelihood function of a random sample of size n is:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

$$\ln L(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\hat{\lambda} = n \Big/ \sum_{i=1}^{n} x_i = 1/\bar{X} \quad \text{(same as moment estimator)}$$

Methods of Moments

Population and samples moments

Let $X_1, X_2, \ldots, X_n$ be a random sample from the probability distribution $f(x)$, where $f(x)$ can be a discrete probability mass function or a continuous probability density function. The kth **population moment** (or **distribution moment**) is $E(X^k)$, $k = 1, 2, \ldots$. The corresponding kth **sample moment** is $(1/n)\sum_{i=1}^{n} X_i^k$, $k = 1, 2, \ldots$.

Population moments

$$\mu'_k = \begin{cases} \int_x x^k f(x)dx & \textbf{If } \boldsymbol{x} \textbf{ is continuous} \\ \sum_x x^k f(x) & \textbf{If } \boldsymbol{x} \textbf{ is discrete} \end{cases}$$

Sample moments

$$m'_k = \frac{\sum_{i=1}^{n} X_i^k}{n}$$

Method of Moments

- **Equating empirical moments to theoretical moments**

Let $X_1, X_2, \ldots, X_n$ be a random sample from either a probability mass function or probability density function with m unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$. The **moment estimators** $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_m$ are found by equating the first m population moments to the first m sample moments and solving the resulting equations for the unknown parameters.

***m* equations for *m* parameters**

$$\begin{cases} m_1' = \mu_1' \\ m_2' = \mu_2' \\ \vdots \\ m_m' = \mu_m' \end{cases}$$

Sample Question

3. (20 points) A random variable has probability density function

$$f(x;\theta) = \frac{1}{\theta^2} x^{(1-\theta^2)/\theta^2}, \quad 0 < x < 1, 0 < \theta < \infty.$$

Now, given samples $X_1, \ldots, X_n$, derive the maximum likelihood estimator for the parameter θ.

- $\ell(\theta) = \sum_{i=1}^{n} \log f(X_i;\theta)$

$= \sum_{i=1}^{n} \log\left(\frac{1}{\theta^2} X_i^{(1-\theta^2)/\theta^2}\right)$

$= \sum_{i=1}^{n} \log\left(\frac{1}{\theta^2}\right) + \log\left(X_i^{(1-\theta^2)/\theta^2}\right)$

$= \sum_{i=1}^{n} -2\log\theta + \frac{1-\theta^2}{\theta^2}\log X_i$

$= -2n\log\theta + \left(\frac{1}{\theta^2} - 1\right)\sum_{i=1}^{n}\log X_i$

- $\nabla\ell(\theta) = \frac{-2n}{\theta} - \frac{2}{\theta^3}\sum_{i=1}^{n}\log X_i = 0$

$\Rightarrow -\frac{2}{\theta^3}\sum_{i=1}^{n}\log X_i = \frac{2n}{\theta}$

$\Rightarrow -\frac{1}{n}\sum_{i=1}^{n}\log X_i = \log\theta^2$

$\Rightarrow \theta^2 = \frac{1}{n}\sum_{i=1}^{n}\log X_i$

$\Rightarrow \hat{\theta}_{MLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\log X_i}$

Contents

- Basics of Probability
- Basics of Linear Algebra
- Basics of Statistical Inference
- Basics of Information Theory
- Basics of Stochastic Processes
- Inequalities in Information Theory

Understanding Entropy

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

$$= \mathbb{E}_x[-\log_2 p(x)]$$

$$= \mathbb{E}_x\left[\log \frac{1}{p(x)}\right]$$

- Uncertainty in a single random variable
- Can also be written as:

$$H(X) = \mathbb{E}\left\{\log \frac{1}{p(X)}\right\}$$

- Intuition: $H = \log(\texttt{\#of outcomes/states})$
- Entropy is a functional of $p(x)$
- Entropy is a lower bound on the number of bits need to represent a RV. E.g.: a RV that has uniform distribution over 32 outcomes

$\mathcal{X}$ $\{1, 2, 3, \ldots, 8\}$

1 → 001 4 → 100
2 → 010 5 → 101
3 → 011 6 → 110
7 → 111
8 → 000

Properties of entropy

- $H(X) \geq 0$

$H(X) = E_X\left[\log \frac{1}{P(x)}\right]$ $\log \frac{1}{P(x)} \geq 0$

- Definition, for Bernoulli random variable, $X = 1$ w.p. p,

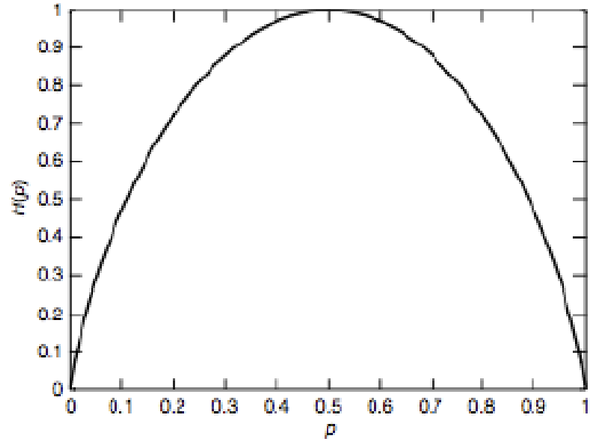
  $X = 0$ w.p. $1 - p$

$$H(p) = -p \log p - (1-p) \log(1-p)$$

$H'(p) = -(1 + \log p) - \left[(1-p) \cdot \frac{1}{1-p} \cdot -1 \right.$

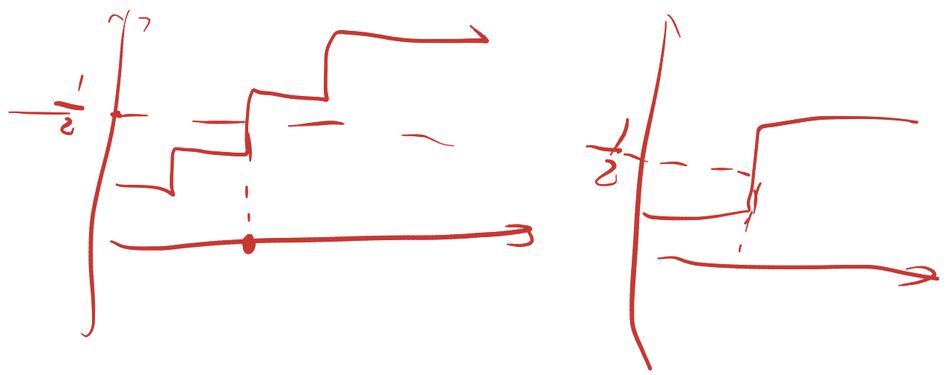
$\left. -\log(1-p)\right]$

$= -1 - \log p + 1 + \log(1-p)$



$H''(p) = -\frac{1}{p} - \frac{1}{1-p} \leq 0$

- **Concave**
- Maximizes at $p = 1/2$
- Example: how to ask questions?



Joint entropy

- Extend the notion to a pair of discrete RVs (X, Y)
- Nothing new: can be considered as a single vector-valued RV
- Useful to measure dependence of two random variables

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

$$H(X,Y) = -\mathbb{E}\log p(X,Y)$$

Conditional Entropy

$(X, Y) \sim p(x, y)$

$Y|X \sim p(Y|X)$

$Y|X=x \sim p(Y|X=x) = \frac{p(x,y)}{p(x)}$, $\forall\, y \in \text{support}(Y)$

- Conditional entropy: entropy of a RV given another RV. If

  $(X, Y) \sim p(x, y)$

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

- Various ways of writing this

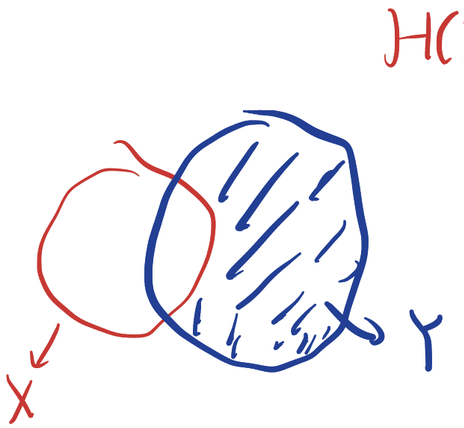
$$H(Y|X=x) = -\sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x)$$

$$H(Y|X) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x)\, p(y|x) \log p(y|x)$$

$$= -\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log p(y|x)$$

$$= -\mathbb{E} \log p(Y|X)$$

Chain rule for entropy



$H(Y|X) = H(X,Y) - H(X)$

- Entropy of a pair of RVs = entropy of one + conditional entropy of the other:

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

- Proof:
- $H(Y|X) \neq H(X|Y)$
- $H(X) - H(X|Y) = H(Y) - H(Y|X)$

$$H(X,Y) = -\sum_{(x,y)} p(x,y) \log p(x,y)$$

$$= -\sum_{(x,y)} p(x,y) \log p(x)p(y|x)$$

$$= -\sum_{(x,y)} p(x,y) \log p(x) - \sum_{(x,y)} p(x,y) \log p(y|x)$$

$H(Y|X)$

$$-\sum_x \left(\sum_y p(x,y) \right) \log p(x) = -\sum_x p(x) \log p(x) = H(X)$$

Sample Question

4. (20 points) Two random variables X and Y have the following joint distribution:

$$\Pr(X=0,Y=0)=0.2,$$
$$\Pr(X=0,Y=1)=0.3,$$
$$\Pr(X=1,Y=0)=0.1,$$
$$\Pr(X=1,Y=1)=0.4.$$

Calculate $H(X), H(Y), H(X \mid Y), H(Y \mid X)$, and $H(X,Y)$ ≈ 1.847

- $\Pr(X=0) = 0.5 = \Pr(X=1)$

$\Rightarrow H(X) = -0.5\log_2 0.5 - 0.5\log_2 0.5 = 1$

- $\Pr(Y=0) = 0.3 \quad \Pr(Y=1) = 0.7$

$\Rightarrow H(Y) = -0.3\log_2 0.3 - 0.7\log_2 0.7$

$H(Y \mid X) = 0.847$

- $H(X \mid Y) = \Pr(Y=0)\cdot H(X \mid Y=0) + \Pr(Y=1)\cdot H(X \mid Y=1) = 0.965$

$\Pr(X=0 \mid Y=0) = \frac{2}{3} \quad \Pr(X=1 \mid Y=0) = \frac{1}{3}$

$\Downarrow$

$-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$

$\Pr(X=0 \mid Y=1) = \frac{3}{7} \quad \Pr(X=1 \mid Y=1) = \frac{4}{7}$

$\Downarrow$

$-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$

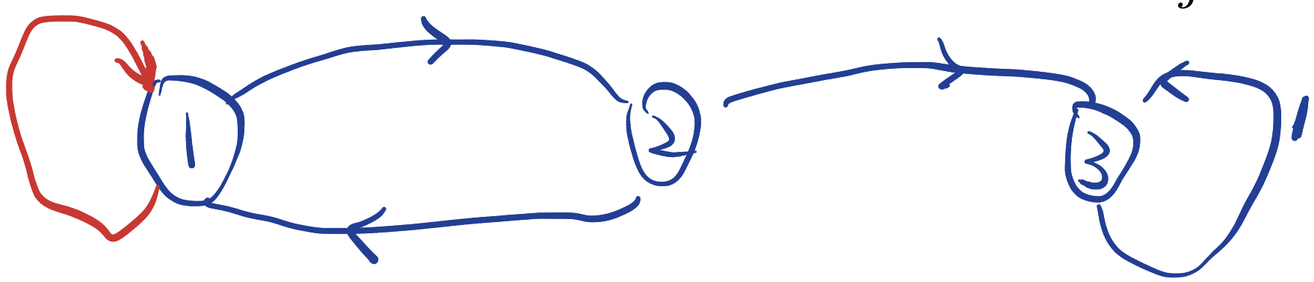
Contents

- Basics of Probability
- Basics of Linear Algebra
- Basics of Statistical Inference
- Basics of Information Theory
- Basics of Stochastic Processes
- Inequalities in Information Theory

Probability and expected time to absorption

- Suppose the state space $\mathcal{S} = \{1, 2, \ldots, M\}$
- From state $i \in \mathcal{S}$, denote the probability to reach a specific absorbing state s as a_i.
- It holds that $a_s = 1$ and for all absorbing states $i \neq s$, $a_i = 0$.
- For all transient states i,

$$a_i = \sum_{j=1}^{M} a_j p_{ij}.$$



$a_3 = 1$

$a_1 = a_2 P_{12} + a_1 P_{11}$

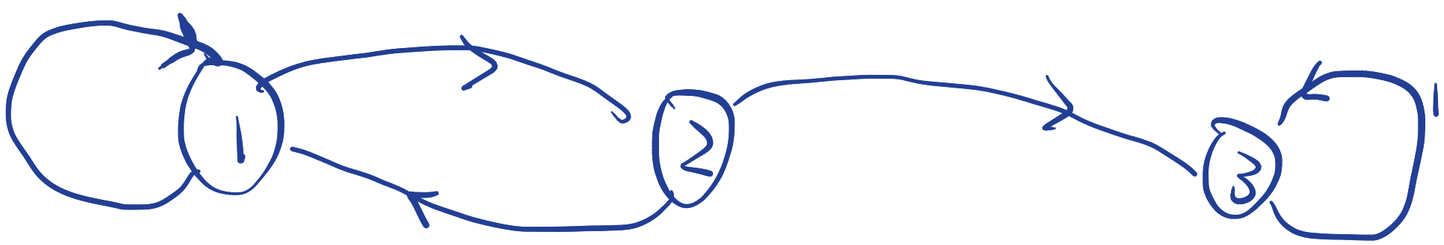
$a_2 = a_1 P_{21} + a_3 P_{23}$

$= P_{23} + a_1 P_{21}$

Probability and expected time to absorption

- Suppose the state space $\mathcal{S} = \{1, 2, \ldots, M\}$
- From state $i \in \mathcal{S}$, denote the expected times to absorption as $\mu_1, \ldots, \mu_M$.
- $\{\mu_i\}_{i \in \mathcal{S}}$ is the unique solution to the system of equations
  - $\mu_i = 0$ for all absorbing state(s) i
  - For all transient states i, $\mu_i = 1 + \sum_{j=1}^{M} p_{ij} \mu_i$.

$$\mu_i = 1 + \sum_{j=1}^{M} p_{ij} \mu_j$$



$\mu_3 = 0$

$\mu_1 = 1 + \mu_1 p_{11} + \mu_2 p_{12}$

$\mu_2 = 1 + \mu_1 p_{21} + \mu_3 p_{23}$

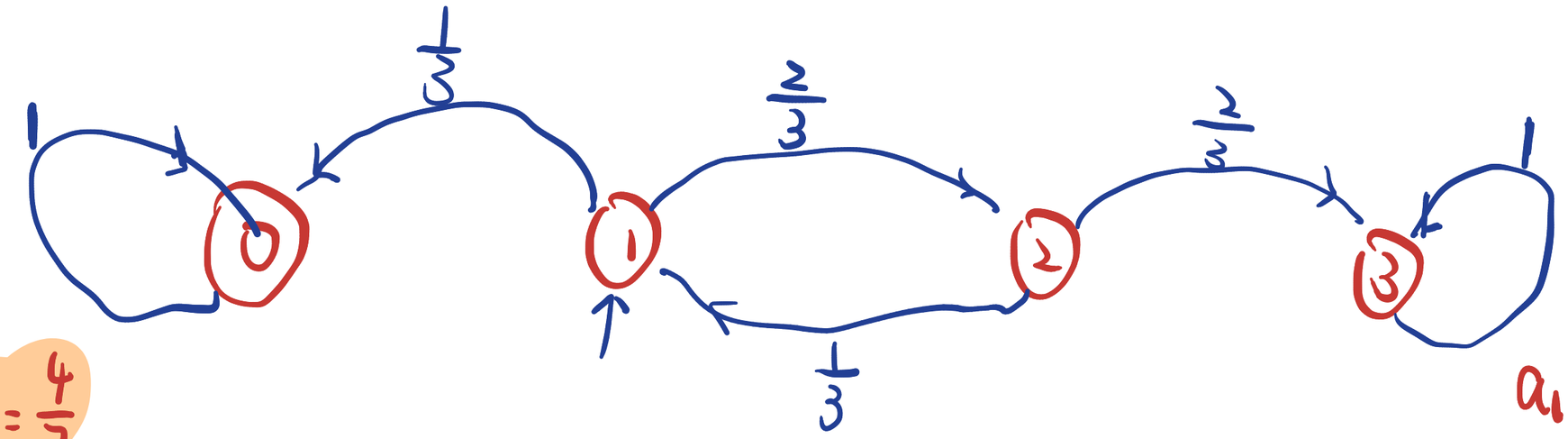
Example

$S = \{(3,0), (2,1), (1,2), (0,3)\}$

$S = \{$ # of money M have $\} = \{0, 1, 2, 3\}$.

a_i: Prob that M reaches state 3 starting from state i.

Player M has \$1 and player N has \$2. Each game gives the winner \$1 from the other. As a better player, M wins $2/3$ of the games. They play until one of them is bankrupt. What is the probability that M wins?



$a_0 = 0$

$a_3 = 1$

$a_1 = \frac{1}{3} a_0 + \frac{2}{3} a_2 = \frac{2}{3} a_2$

$a_2 = \frac{1}{3} a_1 + \frac{2}{3} a_3$

$= \frac{2}{3} + \frac{1}{3} a_1$

$a_2 = \frac{1}{3} + \frac{1}{3}\frac{2}{3} a_2 \Leftrightarrow a_2 - \frac{2}{9} a_2 = \frac{1}{3} \Leftrightarrow \frac{7}{9} a_2 = \frac{2}{3}$

$\Leftrightarrow a_2 = \frac{2}{3} \times \frac{9}{7} = \frac{6}{7}$

$a_1 = \frac{2}{3} a_2 = \frac{4}{7}$

Sample Question

5. (10 points) Two players, A, and B, start with 2 and 3 dollars respectively. Player A wins each round with probability $p = 0.4$. These two players play such a game until one is ruined.

   (i) Find the probability that A wins all money (ruins B). (5 points)

   (ii) Compute the expected number of rounds until the game ends. (5 points)

Contents

$E[f(z)] \geq f(E[z])$, if f is convex

Log sum inequality

$b_i' = \frac{b_i}{\sum_{i=1}^n b_i}$ $\quad t_i = \frac{a_i}{b_i}$ $\quad f(x) = x \log x$

$$\Rightarrow \text{LHS} = \left(\sum_{i=1}^n b_i' t_i \log t_i\right) \sum_{i=1}^n b_i \quad \Rightarrow \quad \sum_{i=1}^n b_i' f(t_i) \geq f\left(\sum_{i=1}^n b_i' t_i\right)$$

$$= \sum_{i=1}^n b_i' t_i \log\left(\sum_{i=1}^n b_i' t_i\right)$$

- Consequence of concavity of log
- **Theorem.** For nonnegative $a_1, \cdots, a_n$ and $b_1, \cdots, b_n$

$$= \sum_{i=1}^n \frac{a_i}{\sum_{i=1}^n b_i} \log\left(\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}\right)$$

$(a_1+a_2)\log\frac{a_1+a_2}{b_1+b_2} \leq a_1 \log\frac{a_1}{b_1} + a_2 \log\frac{a_2}{b_2}$

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i\right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}$$

$f\left(\frac{a_i}{b_i}\right) = \frac{a_i}{b_i} \log \frac{a_i}{b_i}$

$$= \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \log\left(\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}\right)$$

Equality iff $a_i/b_i =$ constant.

- Proof by Jensen's inequality using convexity of $f(x) = x \log x$.

$$\text{LHS} \geq \times \sum_{i=1}^n b_i = \sum_{i=1}^n a_i \log\left(\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}\right)$$

Application of log-sum inequality

- Very handy in proof: e.g., prove $D(p\|q) \geq 0$:

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

$$\geq \left(\sum_x p(x) \right) \log \frac{\sum_x p(x)}{\sum_x q(x)} = 1 \log 1 = 0.$$

Convexity of relative entropy

Theorem. $D(p\|q)$ is convex in the pair (p, q): given two pairs of pdf,

$$D(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \le \lambda D(p_1\|q_1) + (1-\lambda)D(p_2\|q_2)$$

for all $0 \le \lambda \le 1$.

Proof: By definition and log-sum inequality

$$\begin{aligned}
&D(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \\
&= \sum_{x\in\mathcal{X}} (\lambda p_1 + (1-\lambda)p_2) \log \frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2} \\
&\le \sum_{x\in\mathcal{X}} \lambda p_1 \log \frac{\lambda p_1}{\lambda q_1} + (1-\lambda) \log \frac{(1-\lambda)p_2}{(1-\lambda)q_2} \\
&= \lambda D(p_1\|q_1) + (1-\lambda)D(p_2\|q_2)
\end{aligned}$$

log-sum inequality

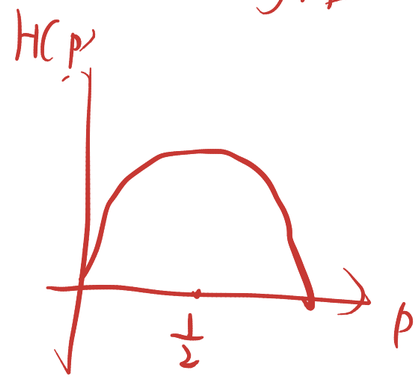
Concavity of entropy

Entropy

$$H(p) = -\sum_i p_i \log p_i$$

$\mathcal{X} = \{0,1\}$

p $H(p) = -p\log p$ $-(1-p)\log(1-p)$



is concave in p

Proof 1:

$$H(p) = -\sum_{i\in\mathcal{X}} p_i \log p_i = -\sum_{i\in\mathcal{X}} p_i \log \frac{p_i}{u_i} u_i$$

$$= -\sum_{i\in\mathcal{X}} p_i \log \frac{p_i}{u_i} - \sum_{i\in\mathcal{X}} p_i \log u_i$$

$$= -D(p\|u) - \log \frac{1}{|\mathcal{X}|} \sum_{i\in\mathcal{X}} p_i$$

$$= \log |\mathcal{X}| - D(p\|u)$$

Concavity of entropy (Proof 2)

Proof 2: We want to prove

$$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2).$$

A neat idea: introduce auxiliary RV:

$$\theta = \begin{cases} 1, & \text{w.p. } \lambda \\ 2, & \text{w.p. } 1-\lambda. \end{cases}$$

Let $Z = X_\theta$, distribution of Z is $\lambda p_1 + (1-\lambda)p_2$. Conditioning reduces entropy:

$$H(Z) \geq H(Z|\theta)$$

By their definitions

$$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2).$$

Concavity and convexity of mutual information

Mutual information $I(X;Y)$ is:

- a concave function of $p(x)$ for fixed $p(y|x)$
- convex function of $p(y|x)$ for fixed $p(x)$

Mixing two gases of equal entropy results in a gas with higher entropy.

Proof of mutual information properties

Proof: write $I(X;Y)$ as a function of $p(x)$ and $p(y|x)$:

$$I(X;Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(y|x)}{p(y)}$$

$$= \sum_{x,y} p(x)p(y|x) \log p(y|x)$$

$$- \sum_{y} \left\{ \sum_{x} p(x)p(y|x) \right\} \log \left\{ \sum_{x} p(y|x)p(x) \right\}$$

- (a): Fixing $p(y|x)$, first linear in $p(x)$, second term concave in $p(x)$
- (b): Fixing $p(x)$, complicated in $p(y|x)$. Instead of verify it directly, we will relate it to something we know.

Strategy for convexity proof

Our strategy is to introduce auxiliary RV $\tilde{Y}$ with a mixing distribution

$$p(\tilde{y}|x) = \lambda p_1(y|x) + (1-\lambda)p_2(y|x).$$

To prove convexity, we need to prove:

$$I(X;\tilde{Y}) \leq \lambda I_{p_1}(X;Y) + (1-\lambda)I_{p_2}(X;Y)$$

Since

$$I(X;\tilde{Y}) = D(p(x,\tilde{y})\|p(x)p(\tilde{y}))$$

We want to use the fact that $D(p\|q)$ is convex in the pair (p,q).

Completing the convexity proof

What we need is to find out the pdfs:

$$p(\tilde{y}) = \sum_x [\lambda p_1(y|x)p(x) + (1-\lambda)p_2(y|x)p(x)] = \lambda p_1(y) + (1-\lambda)p_2(y)$$

$p(\tilde{y}) = \sum_x p(x)\, p(\tilde{y}|x)$

We also need

$$p(x, \tilde{y}) = p(\tilde{y}|x)p(x) = \lambda p_1(x,y) + (1-\lambda)p_2(x,y)$$

Finally, we get, from convexity of $D(p\|q)$:

$$D(p(x,\tilde{y})\|p(x)p(\tilde{y}))$$

$$= D(\lambda p_1(y|x)p(x) + (1-\lambda)p_2(y|x)p(x)\|\lambda p(x)p_1(y) + (1-\lambda)p(x)p_2(y))$$

$$\leq \lambda D(p_1(x,y)\|p(x)p_1(y)]) + (1-\lambda)D(p_2(x,y)\|p(x)p_2(y))$$

$$= \lambda I_{p_1}(X;Y) + (1-\lambda)I_{p_2}(X;Y)$$

Sample Question

6. (10 points) For a discrete probability distribution $P = (p_1, \ldots, p_n)$ and parameter $\alpha \in (0, 1)$, the the Rényi entropy is defined as

$$H_\alpha(P) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n} p_i^\alpha \right)$$

(i) Show that as $\alpha \to 1$, $H_\alpha(P)$ converges to the Shannon entropy $H(P) = -\sum_{i=1}^{n} p_i \log p_i$. (5 points)

(ii) Show that $H_\alpha(P)$ is concave in p if $\alpha \in (0, 1)$. (5 points)

(i) $\lim_{\alpha \to 1} \frac{\log\left(\sum_{i=1}^n p_i^\alpha\right)}{1-\alpha} = \lim_{\alpha \to 1} \frac{\frac{1}{\sum_{i=1}^n p_i^\alpha} \sum_{i=1}^n p_i^\alpha \log p_i}{-1} = -\lim_{\alpha \to 1} \frac{\sum_{i=1}^n p_i^\alpha \log p_i}{\sum_{i=1}^n p_i^\alpha} = -\sum_{i=1}^n p_i \log p_i$

(ii) $g(\lambda x_1 + (1-\lambda) x_2) \geq \lambda g(x_1) + (1-\lambda) g(x_2)$

$f[g(\lambda x_1 + (1-\lambda) x_2)] \geq f[\lambda g(x_1) + (1-\lambda) g(x_2)]$

$\forall\ x_1 \geq x_2$, $f(x_1) \geq f(x_2)$