Problems and Solutions

Abhishek Sinha
LIDS, MIT

I. *Chain inequality* [1]

Let $X_1 \to X_2 \to X_3 \to X_4$ form a Markov Chain. Show that

$$I(X_1;X_3)+I(X_2;X_4) \leq I(X_1;X_4)+I(X_2;X_3)$$

Solution

We compute the following

$$\begin{aligned}
& I(X_1;X_4)+I(X_2;X_3)-(I(X_1;X_3)+I(X_2;X_4)) && (1)\\
= \;& \mathbb{E}\log\frac{p(X_1,X_4)p(X_2,X_3)p(X_1)p(X_3)p(X_2)p(X_4)}{p(X_1)p(X_4)p(X_2)p(X_3)p(X_1,X_3)p(X_2,X_4)}\\
= \;& \mathbb{E}\log\frac{p(X_1,X_4)p(X_2,X_3)}{p(X_1,X_3)p(X_2,X_4)}\\
= \;& \mathbb{E}\log\frac{p(X_4|X_1)p(X_3|X_2)}{p(X_3|X_1)p(X_4|X_2)}\\
= \;& \mathbb{E}\log\frac{p(X_3|X_1,X_2)}{p(X_3|X_1)}\frac{p(X_4|X_1)}{p(X_4|X_1,X_2)} && (2)\\
= \;& I(X_2;X_3|X_1)-I(X_2;X_4|X_1) && (3)\\
\geq \;& 0
\end{aligned}$$

Where Eqn 2 follows from the Markov property and Eqn 3 follows from the Information Inequality for the Markov Chain $X_2 \to X_3 \to X_4$, for each possible realization of X_1.

II. *Time-varying channel* [1]

A train pulls out of the station ar constant velocity. The received signal energy thus falls off with time as $\frac{1}{i^2}$. The total received signal at time i is

$$Y_i = \frac{1}{i}X_i + Z_i,$$

where $Z_1, Z_2, \ldots$ are i.i.d. $\sim N(0,\mathcal{N})$. The transmitter constraint for block length n is

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2(w) \leq P, \quad w \in \{1,2,\ldots,2^{nR}\}$$

Using Fano's inequality, show that the capacity C is zero for this channel.

Solution

We follow the same steps as in the proof of converse in section 9.2 upto Eqn. (9.52), to conclude that,

$$R_n \leq \frac{1}{2}\log\left(1+\frac{1}{N}\frac{1}{n}\sum_{i=1}^{n}\frac{P_i}{i^2}\right)+\epsilon_n \tag{4}$$

Where $\epsilon_n \to 0$. Now we simply use the inequality,

$$\sum_{i=1}^{n}\frac{P_i}{i^2} \leq \big(\sum_{i=1}^{n}P_i\big)\big(\sum_{i=1}^{n}\frac{1}{i^2}\big) \leq \frac{\pi^2}{6}P$$

Hence,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{P_i}{i^2} \leq \frac{\pi^2}{6n}P \searrow 0$$

From Eqn. (4), this implies,

$$\limsup R_n = 0$$

for any sequence of codebooks with $\epsilon_n \to 0$. Thus capacity of the channel is zero.

III. *Time-varying channels II.* [1]

Consider a time-varying discrete time memoryless channel. Let $Y_1, Y_2, \ldots, Y_n$ be conditionally independent given $X_1, X_2, \ldots, X_n$, with conditional distribution given by $p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} p_i(y_i|x_i)$. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n), \boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$. Find $\max_{p(\boldsymbol{x})} I(\boldsymbol{X};\boldsymbol{Y})$.

Solution

We simply compute,

$$\begin{aligned}
I(\boldsymbol{X};\boldsymbol{Y}) &= \mathbb{E}\log\frac{p(\boldsymbol{X},\boldsymbol{Y})}{p(\boldsymbol{X})p(\boldsymbol{Y})}\\
&= H(\boldsymbol{Y})-H(\boldsymbol{Y}|\boldsymbol{X})\\
&= H(\boldsymbol{Y})-\sum_{i=1}^{n}H(Y_i|X_i)\\
&\leq \sum_{i=1}^{n}\big(H(Y_i)-H(Y_i|X_i)\big)=\sum_{i=1}^{n}I_i(X_i;Y_i)
\end{aligned}$$

The inequality above becomes an equality iff the random variables $\{Y_i\}_{i=1}^{n}$ are independent, i.e. if $p(\boldsymbol{X})$ factorizes as,

$$p(\boldsymbol{X}) = \prod_{i=1}^{n} p_i(X_i)$$

For some distributtions $p_i(\cdot)$ on the input alphabet. Now, we can choose the distributions $p_i(\cdot)$, such that, it maximizes the one-slot capacity $I_i(X_i;Y_i)$. Call this value C_i. Hence, $\max_{p(\boldsymbol{x})} I(\boldsymbol{X};\boldsymbol{Y}) = \sum_{i=1}^{n} C_i$

IV. *Channels with two independent looks at Y.*[1]

Let Y_1 and Y_2 be conditionally independent and conditionally identically distributed given X.
Show that $I(X;Y_1,Y_2) = 2I(X;Y_1) - I(Y_1;Y_2)$

Solution

$$\begin{aligned}
I(X;Y_1,Y_2) &= H(Y_1,Y_2)-H(Y_1,Y_2|X)\\
&= H(Y_1,Y_2)-2H(Y_1)+2H(Y_1)-2H(Y_1|X)\\
&\quad \big(\text{Since, } p(Y_1,Y_2|X)=p(Y_1|X)p(Y_2|X)\big)\\
&= 2I(X;Y_1)-I(Y_1;Y_2)
\end{aligned}$$

V. *Tall, fat people* [1]

Suppose the average height of people in a room is 5 feet. Suppose that average weight is 100 lb.
(a) Argue that no more than one-third of the population is 15 feet tall.
(b) Find an upper bound on the fraction of 300-lb 10-footers in the room.

Solution

Assign a uniform distribution on the set of all people in the room. Hence, fraction of a population satisfying some attributes is simply probability of selecting a person having the same attributes.
(a)
Let us sample a person from the room and let the random variables H and W denote the height and weight of the person selected.
We have, using Markov's inequality :

$$\mathbb{P}(H = 15) \leq \mathbb{P}(H \geq 15) \leq \frac{\mathbb{E}(H)}{15} = \frac{5}{15} = \frac{1}{3}$$

(b)
Similarly, we have

$$\mathbb{P}(H \geq 10) \leq \frac{1}{2}$$
$$\mathbb{P}(W \geq 300) \leq \frac{1}{3}$$

Now,

$$\begin{aligned}\mathbb{P}(H = 10, W = 300) &\leq \mathbb{P}(H \geq 10, W \geq 300)\\ &\leq \mathbb{P}(W \geq 300) \leq \frac{1}{3}\end{aligned}$$

VI. *Bottleneck channel (Problem 7.25)* [1]

Solution

Assume $p(X)$ is chosen as the capacity achieving distribution for the channel $X \rightarrow Y$

$$\begin{aligned}C &= I(X;Y)\\ &\leq I(X;V)\\ &= H(V) - H(V|X)\\ &\leq H(V)\\ &\leq \log k\end{aligned}$$

Where the first inequality follows from data-processing inequality, the second inequality follows from non-negativity of conditional entropy and finally the third inequality follows from the fact that $|V| = k$.

VII. *Gaussian mutual information (Problem 9.11)* [1]

Solution

It is enough to derive the joint distribution of the random variable (X, Z). Since the random variables, (X, Y, Z) are jointly Gaussian, we know that (X, Z) is also jointly Gaussian. Hence, their distribution is completely specified in terms of the covariance matrix K_{XZ}, in particular the correlation coefficient ρ_{XZ}. We have,

$$\rho_{XZ} = \mathbb{E}(XZ) = \mathbb{E}_Y\mathbb{E}(XZ|Y) = \mathbb{E}_Y\mathbb{E}(X|Y)\mathbb{E}(Z|Y) \quad (5)$$

Where we have used the fact that $X \rightarrow Y \rightarrow Z$ forms a Markov chain. WOLOG, we may assume that $\mathbb{E}(X) = \mathbb{E}(Y) = \mathbb{E}(Z) = 0$. Thus the conditional expectations of bivariate normal distributions can be easily worked out to be

$$\mathbb{E}(X|Y) = \rho_1 \frac{\sigma_X}{\sigma_Y} Y$$
$$\mathbb{E}(Z|Y) = \rho_2 \frac{\sigma_Z}{\sigma_Y} Y$$

Please see John Tsitsiklis' probability text for reference on bivariate normal distribution.
Thus $\rho_{XZ} = \mathbb{E}_Y \rho_1 \rho_2 \frac{\sigma_X \sigma_Z}{\sigma_Y^2} Y^2 = \rho_1 \rho_2 \sigma_X \sigma_Z$. Hence,

$$\begin{aligned}I(X;Z) &= h(X) + h(Z) - h(X,Z)\\ &= \frac{1}{2} \ln \frac{\sigma_X^2 \sigma_Z^2}{\sigma_X^2 \sigma_Z^2 (1 - \rho_1^2 \rho_2^2)} = \frac{1}{2} \ln \frac{1}{1 - \rho_1^2 \rho_2^2} \blacksquare\end{aligned}$$

VIII. *An interesting elementary problem on functional equation*

Let $\mathcal{B}$ be the set of all functions f from the set of positive reals to the set of positive reals such that

$$f(3x) \geq f(f(2x)) + x, \qquad \forall x \in \mathbb{R}_+$$

Find the maximum real number τ such that for all such functions $f \in \mathcal{B}$, the following holds:

$$f(x) \geq \tau x$$

Solution

For any function $f \in \mathcal{B}$, we have the following series of inequalities valid for all $x \in \mathbb{R}_+$:

$$f(3x) \geq f(f(2x)) + x \geq \tau f(2x) + x \geq 2\tau^2 x + x = (1 + 2\tau^2)x,$$

i.e.,

$$f(x) \geq \frac{(1 + 2\tau^2)}{3} x, \forall x \in \mathbb{R}_+$$

Hence by definition of τ, we have $\tau \geq \frac{(1+2\tau^2)}{3}$, i.e. $\frac{1}{2} \leq \tau \leq 1$. Now to show that $\tau = \frac{1}{2}$, we simply note that the function $g(x) = \frac{1}{2}x, \forall x \in \mathbb{R}_+$ belongs to the set $\mathcal{B}$.

IX. *An interesting elementary problem on SNR*

Consider the positive quantities $\{a_1, a_2, \ldots, a_n\}$ and σ^2. Define the SNR for the user i as

$$\gamma_i = \frac{a_i}{\sum_{j \neq i} a_j + \sigma^2}, \qquad i = 1, 2, \ldots, n$$

Prove that for any $0 < \alpha < 1$, number of users with SNR at least α is at most $1 + \frac{1}{\alpha}$.

Solution

Assume that m is the number of users with $\gamma_i \geq \alpha$. Then for each such user i, we have by definition

$$\frac{a_i}{\sum_{j\neq i} a_j + \sigma^2} \geq \alpha$$

i.e.,

$$\frac{1}{\alpha} a_i \geq \sum_{j\neq i} a_j + \sigma^2 \geq \sum_{j\neq i} a_j$$

Adding a_i to both sides, we have

$$(1+\frac{1}{\alpha})a_i \geq \sum_j a_j = S$$

Where, $S \equiv \sum_j a_j$. Summing over all m users with $\gamma_i \geq \alpha$, we have

$$(1+\frac{1}{\alpha})S \geq mS$$

i.e.,

$$m \leq 1 + \frac{1}{\alpha}\blacksquare$$

X. *Exercise 14.8 from* RANDOMIZED ALGORITHMS [2]

From Lemma 14.17, we know that for any generator g of $\mathbb{Z}_p^*$, g^k is a quadratic residue if and only if k is even. Since $p-1$ is even, it is clear that exactly half of the elements of $\mathbb{Z}_p^*$ are quadratic residues.
To find a quadratic *non-residue*, just select an element a from $\mathbb{Z}_p^*$ u.a.r and compute the Legendre symbol $[\frac{1}{p}]$. If it is -1 then stop, otherwise repeat.
Clearly, the above algorithm has a probability of exactly $\frac{1}{2}$ of selecting a quadratic non-residue at each step. Hence on the average we need $\Omega(\log(p))$ iterations.

XI. *Exercise 9.1 from* RANDOMIZED ALGORITHMS[2]

Consider a set of points $S = \{(x_i, y_i), i = 1, 2, \ldots, n\}$, such that each of the points in the set forms a vertex of the conv(S). Then any algorithm which actually computes the convex hull of S would output the points in the set S in a sorted order. Hence we have at least $\Omega(n \log n)$ complexity of any algorithm for finding the convex hull of S.

XII. *Exercise 4.21 from* ALGORITHMS- VAZIRANI [3]

(a) Construct a graph $\mathcal{G}(V, E)$ as follows. The set of vertices are labelled by the currency set C. And the length of edge $l_{ij} = \log r_{ij}, \forall (i,j) \in C^2, i \neq j$. Then finding the shortest path in the graph $\mathcal{G}$ from the node s to the node t gives an optimum sequence of such exchanges.
(b) We just need to detect the presence of negative cycles in the graph $\mathcal{G}$.

XIII. *Exercise 4.20 from* ALGORITHMS- VAZIRANI [3]

Find the shortest path from $s \to t$ in all the graphs $\mathcal{G}(V, F_k')$, where $F_k' = E \cup \{e_k\}$, e_k being the k^{th} edge from the set E'. Find the minimum of all of the shortest path.

XIV. *Exercise 4.19 from* ALGORITHMS- VAZIRANI [3]

For every edge $e = (i, j)$, update its cost to $l_e' = l_e + c_i$. Find the shortest path from $s \to t, \forall t \in V$ in this updated edge-capacitated directed graph.

XV. *Problem 2 from* PUTNAM 1999

It is assumed implicitly that $p(x)$ is a polynomial with real coefficients. Then the roots of $p(x)$ are of the form $\alpha_i \pm j\beta_i$, i.e. we have, for some $C > 0$ and $m \in \mathbb{N}$

$$p(x) = C\prod_{i=1}^{m}\left((x-\alpha_i)^2 + \beta_i^2\right)$$

Expanding this product out, we get

$$p(x) = \sum_{i=1}^{k}(f_i(x))^2$$

For some polynomials f_i's and some $k \in \mathbb{N}$. $\blacksquare$

XVI. *Exercise 4.13 from* ALGORITHMS- VAZIRANI [3]

(a) Remove all the edges of the graph with length more than L. Run the BFS from s to determine whether t is reachable from s or not in this reduced graph.

XVII. *Exercise 4.5 from* ALGORITHMS- VAZIRANI [3]

Run BFS on the graph $\mathcal{G}$, noting the distances of each vertex v as they are discovered. If the vertex u is discovered at k $^{\text{th}}$ iteration (i.e. the shortest path length $s \to u$ is k), check how many vertices of in the neighbor of u was discovered at $k-1$ $^{\text{th}}$ iteration. Record the number of shortest path from $s \to u$ as $N(u) = \sum_v N(v)$, where the sum extends over all neighbors of v which are discovered at iteration $k-1$.

XVIII. *Exercise 4.7 from* ALGORITHMS- VAZIRANI [3]

First traverse the tree and compute the distance of every node t from s along the tree. Call these distances $d(\cdot)$. Now check whether the Bellman-Ford equality holds for every vertex v, i.e.,

$$d(v) = \min_{u:(u,v)\in E}\left(d(u) + l_{uv}\right), \quad \forall v \in V$$

If it holds then the tree is a shortest-path tree, otherwise not. Note that the above steps can be completed in linear time.
The justification of the above algorithm is simple : the Bellman-Ford fixed point equation has an unique solution and that solution is optimal.

XIX. *Exercise 4.8 from* ALGORITHMS- VAZIRANI [3]

No it is not. Use the example of Figure 4.12 from the book with the added constant 3.

XX. *Exercise 5.3 from* ALGORITHMS- VAZIRANI [3]

To determine whether there exists such an edge or not, just check whether $|E| \geq |V|$. In that case there exists such an edge for the originally connected graph.
To determine such an edge, compute a spanning tree of the graph $\mathcal{G}$. Any edge that is not part of the spanning tree can be deleted without sacrificing the connectivity of $\mathcal{G}$.

XXI. *Exercise 5.4 from* Algorithms- Vazirani [3]

Take the i^{th} connected component of $\mathcal{G}$. If it has m_i edges and n_i vertices, we know that

$$m_i \geq n_i - 1$$

Summing over all k components we have $|E| \geq n - k$. ■

XXII. *Exercise 5.6 from* Algorithms- Vazirani [3]

Consider the edges of two trees T_1 and T_2 in the order of increasing weights. If the trees are not the same, there must exist an index where the edge-weights differ (and we have two different edges). Replacing the higher-weighted edge with the lower one and keeping everything else the same reduces the weight of one of the trees. Hence both the trees can't be MST. ■

XXIII. *Exercise 5.8 from* Algorithms- Vazirani [3]

No. Both contain the minimum-weight edge in it.

XXIV. *Exercise 5.20 from* Algorithms- Vazirani

Do a BFS from a node $s \in V$ and group the vertices in two sets depending on the parity of their distances from s. This results in a bipartite graph $\mathcal{B}$. Now apply Bipartite Matching algorithm on $\mathcal{B}$ to find the maximum matching and check whether it is perfect or not.
Second part: Find MST T with negative edge weights. Put all edges that are not in the tree T into the set E'.

XXV. *Exercise 5.21(c) from* Algorithms- Vazirani [3]

For a given edge $e = (a, b) \in E$, consider the graph $\mathcal{G}_e(V, E \setminus \{e\})$. Do a BFS on $\mathcal{G}_e$ to check whether a and b are connected in $\mathcal{G}_e$. If yes, then the path $\pi(a \rightarrow b)$ plus the edge e forms a cycle, else not.

XXVI. *Exercise 5.31 from* Algorithms- Vazirani [3]

Do the SJF (Shortest Job First), i.e. sort the customers according to their service times and serve the customers in the order of increasing service times.
To prove optimality, consider an exchange argument. Consider any two consecutive customers (in any scheduling order) and switch their order of service. Show that serving the customer with lower service time first always improves the overall cost.

XXVII. *Exercise 5.25 from* Algorithms- Vazirani [3]

Consider a graph $\mathcal{G}(V, E)$ with vertices denoting the variables and there is an edge between v_i and v_j if there is an equality constraint $x_i = x_j$. Now for every inequality constraint $x_j \neq x_l$, check whether there is a path from vertex v_j to v_l in the graph $\mathcal{G}$. If it is false for every inequality constraint then there exists a satisfying assignment.

XXVIII. *Exercise 5.26 from* Algorithms- Vazirani [3]

(a) Consider the degree-sequence $(3, 3, 1, 1)$. It satisfies the conditions mentioned in the problem. Now note that $d_1 = d_2 = 3$ implies that vertices 1 and 2 have edges to all other 3 vertices. This implies that any vertex in the graph can't have degree less than 2. Clearly vertices 3 and 4 violates this condition.

XXIX. *Exercise 5.23 from* Algorithms- Vazirani [3]

First remove the nodes U and all the edges that are incident on them. Since the nodes U are leaves in the desired spanning tree, the nodes $V \setminus U$ must form an MST. So find an MST T on the induced graph $V \setminus U$. Now for all $u \in U$, add the lightest edge connecting u to T.

XXX. *Exercise 6.2 from* Algorithms- Vazirani [3]

Construct a graph $\mathcal{G}(V, E)$ with node sets $V = [[n]]$ and edges $\{(i, j), i < j\}$ of length $(200 - (a_j - a_i))^2, \forall i < j$. CLearly $\mathcal{G}$ is a DAG. Find the shortest path on this DAG.

XXXI. *Exercise 6.8 from* Algorithms- Vazirani [3]

The sub-problems we consider for DP iteration are maximum length of longest commone substring ending precisely at the i^{th} and j^{th} position of the string-arrays x and y, which we call $P[i, j], i = 1, 2, \ldots, n, j = 1, 2, \ldots, m$. Then we have the following DP iterations:

```
for i=1:n
   for j=1:m
     if(i==1 || j==1)
       P(i,j)=(a[i]==b[j]);
     else
       if(a[i]==b[j])
         P(i,j)=P(i-1,j-1)+1;
      else
         P(i,j)=0;
     endif
   endfor
endfor
```

Then finding the $\max$ in the matrix $P[i, j]$ gives the maximum common substring length.

XXXII. *Exercise 6.9 from* Algorithms- Vazirani [3]

Assume that the cut-locations are provided in an array $a[\cdot]$. Consider the sub-problem of cutting the string from $a[i]$ to $a[j]$ and record the optimal cost in the $(i, j)^{\text{th}}$ element of the matrix $P[i, j]$. Then, we have

$$P[i, j] = \min_{k : i < k < j} (P[i, k] + P[k, j] + a[j] - a[i] + 1)$$

This can be readily coded following the same-style as in Matrix-Chain Multiplication.

XXXIII. *Exercise 6.10 from* ALGORITHMS- VAZIRANI [3]

Let $P[i,j]$ denote the probability of obtaining exactly j heads when the first i coins are tossed. We have the following DP recursion

$$P[i,j] = P[i-1,j-1]p_i + P[i-1,j](1-p_i)$$

XXXIV. *Exercise 4.3 from* RANDOMIZED ALGORITHMS [2]

Follows from the property of *bit-fixing* routing algorithm. Once two routes separate, the algorithm never make changes to the bit position (of each individual routes) where they separated from. Hence they can't merge in future.

XXXV. *Exercise 4.4 from* RANDOMIZED ALGORITHMS [2]

No, because before separation they can be in several queues waiting for transmission together.

XXXVI. *Exercise 4.9 from* RANDOMIZED ALGORITHMS [2]

An example of Doob's martingale.

XXXVII. *Exercise 6.4 from* RANDOMIZED ALGORITHMS [2]

Note that, from symmetry we have $h_{1n} = h_{n1} = \frac{1}{2}C_{1n}$. Applying Theorem 6.6 to the given line graph, we have

$$C_{1n} = 2(n-1)(n-1) = 2(n-1)^2$$

From which the result follows.

XXXVIII. *Corollary 6.20 from* RANDOMIZED ALGORITHMS [2]

Since the graph is d-regular, we have

$$e(W, V\setminus W) \leq |\Gamma(W)|d \tag{6}$$

Hence from Theorem 6.19, we have

$$\lambda_2 \geq d - \frac{n|\Gamma(W)|d}{|W||V\setminus W|} \tag{7}$$

Rearranging,

$$1 + \frac{n|\Gamma(W)|}{2|W||V\setminus W|} \geq 1 + (1-\lambda_2/d)/2 \tag{8}$$

Thus,

$$|W| + \frac{n}{2}\frac{|\Gamma(W)|}{|V\setminus W|} \geq (1 + (1-\lambda_2/d)/2)|W| \tag{9}$$

XXXIX. *Exercise 9.2 from* RANDOMIZED ALGORITHMS [2]

It is sufficient to check whether $x_{\min} \leq x \leq x_{\max}$ and $y_{\min} \leq y \leq y_{\max}$, which requires $\Theta(n)$ steps.

XL. *Exercise 11.2 from* RANDOMIZED ALGORITHMS [2]

For a fixed instance I, let us run the randomized algorithm N times and let X_i be the random variable denoting the output for the i^{th} run. Then we have for all $1 \leq i \leq N$

$$\mathbb{P}((1-\epsilon)\#(I) \leq X_i \leq (1+\epsilon)\#(I)) \geq \frac{3}{4} \tag{10}$$

This implies that,

$$\mathbb{P}(X_i \leq (1-\epsilon)\#(I)) \leq \frac{1}{4} \tag{11}$$

$$\mathbb{P}(X_i \geq (1+\epsilon)\#(I)) \leq \frac{1}{4} \tag{12}$$

Let the r.v. Y denote the median of $\{X_i\}, i = 1, 2, \ldots N$. Then we note that the event $Y \leq (1-\epsilon)\#(I)$ occurs only if at least half of the X_i's are less than or equal to $(1-\epsilon)\#(I)$. Define the bernoulli random variables Z_i as follows:

$$\begin{aligned} Z_i &= 1 \text{ if } \quad X_i \leq (1-\epsilon)\#(I) \qquad (13)\\ &= 0 \text{ o.w.} \qquad (14) \end{aligned}$$

Let $Z = \sum_{i=1}^{N} Z_i$. Then following the above discussion, we have

$$\begin{aligned} \mathbb{P}(Y &\leq (1-\epsilon)\#(I)) \leq \mathbb{P}(Z \geq N/2) \qquad (15)\\ &\leq \frac{\big(pe + (1-p)\big)^N}{e^{N/2}} \qquad (16)\\ &\leq \left(\frac{1 + \frac{1}{4}(e-1)}{\sqrt{e}}\right)^N \qquad (17)\\ &\leq \alpha^N \qquad (18) \end{aligned}$$

Where $\alpha = 0.87$. Similarly we have

$$\mathbb{P}\big(Y > (1+\epsilon)\#(I)\big) \leq \alpha^N \tag{19}$$

Hence, via union bound,

$$\mathbb{P}\big(\{Y > (1+\epsilon)\#(I)\} \cup \{Y < (1-\epsilon)\#(I)\big) \leq 2\alpha^N \tag{20}$$

Hence,

$$\mathbb{P}((1-\epsilon)\#(I) \leq Y \leq (1+\epsilon)\#(I)) \geq 1 - 2\alpha^N \tag{21}$$

Where the R.H.S. estimate is at least $1-\delta$ if $2\alpha^N \leq \delta$ if $N \geq \Theta(\ln(1/\delta))$.

XLI. *Exercise 4.2 from* RANDOMIZED ALGORITHMS [2]

Consider $2^{n/2-1} = \frac{1}{2}\sqrt{N}$ number of packets with the first digit of origin 1 and last $n/2$ digits of origin being zeros, i.e., we consider the packets on the vertices with address of the form $1xx\ldots xx||000\ldots 00$, where the separator is after $n/2$ bits. Because of the *bit-fixing* routing strategy, all these $\frac{1}{2}\sqrt{N}$ packets has to visit the vertex $000\ldots 00||000\ldots 00$ before taking the edge leading to $000\ldots 00||100\ldots 00$. Since a single packet can be sent over that edge per slot, this strategy requires at least $\Omega(\sqrt{N})$ steps.

XLII. *Microsoft Research Puzzle 1.*

Prove that the digit sum of any (positive integral) multiple of 11 is even.

Proof: Consider a positive integer n which reads $a_1a_2\ldots a_k$ when written in decimal ($a_i \in \{0,1,2,\ldots,9\}$). Now write $11n = (10+1)n = a_1a_2\ldots a_k0 + a_1a_2\ldots a_k$. Perfoming operations in modulo 10, we conclude that the digit sum (modulo 10) is $2(a_1+a_2+\ldots a_k) \mod 10$ which is even. ■

XLIII. *Problem 7.3 from* INFORMATION THEORY, COVER AND THOMAS [1]

Given that $\boldsymbol{Y} = \boldsymbol{X} \bigoplus \boldsymbol{Z}$ and $\boldsymbol{X} \perp \boldsymbol{Z}$ where $\big(Z_i \in \{0,1\}\big)$ are identically distributed but not necessarily independent. Assume a fixed distribution $p(\boldsymbol{X})$ of $\boldsymbol{X}$. We have,

$$\begin{aligned}
I(\boldsymbol{X};\boldsymbol{Y}) &= H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}) && (22)\\
&= H(\boldsymbol{Y}) - H(\boldsymbol{Z}|\boldsymbol{X}) && (23)\\
&= H(\boldsymbol{Y}) - H(\boldsymbol{Z}) && (24)\\
&\geq H(\boldsymbol{Y}|\boldsymbol{Z}) - H(\boldsymbol{Z}) && (25)\\
&= H(\boldsymbol{X}|\boldsymbol{Z}) - H(\boldsymbol{Z}) && (26)\\
&= H(\boldsymbol{X}) - H(\boldsymbol{Z}) && (27)\\
&\geq H(\boldsymbol{X}) - \sum_{i=1}^{n} H(Z_i) && (28)\\
&= H(\boldsymbol{X}) - nH(p, 1-p) && (29)
\end{aligned}$$

Where (23) follows from the fact $\boldsymbol{Y} = \boldsymbol{X} \bigoplus \boldsymbol{Z}$, (24) follows from $\boldsymbol{X} \perp \boldsymbol{Z}$, (25) follows from the fact that conditioning reduces entropy, (26) follows from $\boldsymbol{Y} = \boldsymbol{X} \bigoplus \boldsymbol{Z}$, (27) follows from $\boldsymbol{X} \perp \boldsymbol{Z}$, (28) follows from independence bound of entropy and (29) follows from the fact that $\{Z_i\}$ are identically distributed.
Now (29) holds for any fixed input distribution $p(\boldsymbol{X})$. Hence,

$$\begin{aligned}
\max_{p(\boldsymbol{X})} I(\boldsymbol{X};\boldsymbol{Y}) &\geq H(\boldsymbol{X})|_{\boldsymbol{p}\sim\boldsymbol{U}} - nH(p, 1-p) && (30)\\
&= n(1 - H(p, 1-p)) && (31)\\
&= nC && (32)
\end{aligned}$$

Where $\boldsymbol{U}$ denotes u.i.i.d. distribution on $\boldsymbol{X}$. ■

XLIV. *Problem 7.33 from* INFORMATION THEORY, C & T[1]

(a) Due to feedback, we have $X_{i+1} = Y_i, \forall i = 1, 2\ldots, n-1$. Hence we have,

$$\begin{aligned}
H(Y^n) &= H(Y_1) + \sum_{i=1}^{n-1} H(Y_{i+1}|Y^i)\\
&= H(Y_1) + \sum_{i=1}^{n-1} H(Y_{i+1}|X_2, X_3, \ldots, X_{i+1})\\
&= H(Y_1) + \sum_{i=1}^{n-1} H(Y_{i+1}|X_{i+1})\\
&= H(Y_1) + (n-1)H(p)
\end{aligned}$$

Similarly,

$$\begin{aligned}
H(Y^n|X^n) &= \sum_{i=1}^{n} H(Y_i|X^n, Y^{i-1})\\
&= 0 + 0 + \ldots + 0 + H(Y_n|X^n, Y^{n-1})\\
&= H(p)
\end{aligned}$$

Thus,

$$I(X^n;Y^n) = H(Y^n) - H(Y^n|X^n) = H(Y_1) + (n-2)H(p)$$

Since, $0 \leq H(Y_1) \leq 1$, we have

$$\lim_{n\to\infty} \frac{1}{n} I(X^n;Y^n) = H(p) \ ■$$

(b) We know that the capacity of a BSC without feedback is given by $C = 1 - H(p)$, where p is the parameter of the BSC. Clearly if $H(p) > 1 - H(p)$, i.e. $H(p) > \frac{1}{2}$, i.e. $0.11 < p < 0.88$ then the above limit is higher than the capacity of the BSC without feedback.

XLV. *Problem 7.10 from* INFORMATION THEORY, C & T[1]

(a) Since the channel is symmetric, we have $H(Y|X) = H(1/2) = 1$. Also for the symmetric channel capacity achieving input distribution is uniform. Hence $C = \log 5 - 1 \approx 1.322$ bits/ch. use.
(b) Consider the following codebook of length 2

$$\mathcal{C} = \{11, 13, 31, 33, 45\} \tag{33}$$

$\mathcal{C}$ is a zero-error code of length 2, because given any two distinct codewords (a_1, b_1) and (a_2, b_2) from $\mathcal{C}$, it can be easily verified that either the first symbol a_1 and a_2 is *non-confusing* or the second symbol b_1 and b_2 is non-confusing. Hence the code $\mathcal{C}$ has rate $\log|\mathcal{C}|/2 = \log 5/2 \approx 1.16 > 1$■.

XLVI. AN UPPER-BOUND ON THE ZERO-ERROR-CAPACITY OF THE CHANNEL [1]

We can obtain an upper-bound on the ZEC of the channel by a simple *sphere-packing-argument*. Consider *any* zero-error-code $\mathcal{C}_n$ of length n for the given 5-symbol type-writer channel. Consider a particular codeword $\boldsymbol{x} \in \mathcal{C}$. Clearly, any word $\boldsymbol{y}$ in the box around $\boldsymbol{x}$ such that $y_i - x_i \mod (5) \leq 1, \forall i = 1, 2, \ldots n$ *can not* belong to $\mathcal{C}_n$, due to zero-error property (Otherwise $\boldsymbol{y}$ could be potentially confused with $\boldsymbol{x}$). Hence every such ball $\mathcal{B}_{\boldsymbol{x}}$ (in ℓ_1 norm) around each codeword $\boldsymbol{x} \in \mathcal{C}$ contains exactly one codeword. Also, it is easy to see that if $\boldsymbol{x} \neq \boldsymbol{y}$ then $\mathcal{B}_{\boldsymbol{x}} \cap \mathcal{B}_{\boldsymbol{y}} = \phi$. Otherwise, if $\boldsymbol{z} \in \mathcal{B}_{\boldsymbol{x}} \cap \mathcal{B}_{\boldsymbol{y}}$, then $\boldsymbol{z}$ could be potentially confused with both $\boldsymbol{x}$ and $\boldsymbol{y}$. Pick an index i such that $x_i - y_i \geq 2$. Since $z \in \mathcal{B}_{\boldsymbol{x}}$, we have $z_i - x_i = 0, 1$. Similarly, $z_i - y_i = 0, 1$, which is impossible. Now since total number of words of length n from an alphabet of size 5 is 5^n, the above argument yields

$$|\mathcal{C}|2^n \leq 5^n$$

i.e.,

$$\limsup_{n\to\infty} \frac{1}{n} \log|\mathcal{C}| \leq \log(2.5) \approx 1.322 \tag{34}$$

Which gives an upper-bound on the capacity of the channel.

XLVII. Differential Entropy Rate of an A-R Process

Problem : Let $\{Z_n\}$ be i.i.d. $\sim \mathcal{N}(0,\sigma^2)$. What is the differential entropy rate $h(X)$ of the stationary process

$$X_{n+1} = \frac{1}{2}X_n + \frac{1}{4}X_{n-1} + Z_{n+1}$$

Solution : We have, $h(X_{n+1}|X^n) = h(Z_{n+1}|X^n) = h(Z_{n+1}) = \frac{1}{2}\ln(2\pi e\sigma^2)$. Where we have used the independence of $\{Z_{n+1}\}$. Hence $h(X) = \lim_{n\to\infty} h(X_{n+1}|X^n) = \frac{1}{2}\ln(2\pi e\sigma^2)$ ■.

XLVIII. Matrix-Norm and Positive Semi-definiteness

Problem : Show that $||A||_2 \leq s$ iff $A^TA \preccurlyeq s^2 I$.
Solution (Only if direction): Consider the SVD of the matrix $A = U\Sigma V^T$ so that all the squared-diagonal entries $\sigma_i^2 \leq s^2, \forall i$. Now, for any $\boldsymbol{x} \in \mathbb{R}^n$ we compute the quadratic form $\boldsymbol{x}^T(s^2\boldsymbol{I} - A^TA)\boldsymbol{x} = \boldsymbol{x}^T(s^2\boldsymbol{I} - V\Sigma^2V^T)\boldsymbol{x} = s^2||\boldsymbol{x}||^2 - \sum_{i=1}^n \sigma_i^2(V^T\boldsymbol{x})_i^2 \geq s^2||\boldsymbol{x}||^2 - \sigma_{\max}^2||\boldsymbol{x}||^2 \geq 0$. This implies that $A^TA \preccurlyeq s^2 I$.
Solution (If direction) : Take $\boldsymbol{x}$ to correspond to the first column of $\boldsymbol{V}$. Then $(\boldsymbol{V}^T\boldsymbol{x})_i = \delta_{i,1}$. Hence, $\boldsymbol{x}^T(s^2\boldsymbol{I} - A^TA)\boldsymbol{x} = s^2 - \sigma_{\max}^2$, which is non-negative iff $||A||_2 \leq s$.

XLIX. Integral transform of a convex function

Problem : Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. Show that

$$\int_0^{2\pi} f(x)\cos(x)dx \geq 0$$

Solution First, for simplicity, assume that the function f is differentiable in the open interval $(0, 2\pi)$. Thus its derivative is non-decreasing. Integrating by parts, we have

$$\begin{aligned} I &= \int_0^{2\pi} f(x)\cos x dx \\ &= -\int_0^{2\pi} f'(x)\sin(x)dx \\ &= -\int_0^{\pi} f'(x)\sin(x)dx - \int_{\pi}^{2\pi} f'(x)\sin(x)dx \end{aligned}$$

Substituting $z = x - \pi$ in the second integral, we have

$$\begin{aligned} I &= -\int_0^{\pi} f'(x)\sin(x)dx + \int_0^{\pi} f'(z+\pi)\sin(z)dz \\ &= \int_0^{\pi} \big(f'(x+\pi) - f'(x)\big)\sin(x)dx \end{aligned}$$

Now since the derivative is non-decreasing, we have $f'(x+\pi) \geq f'(x), \forall x \in \mathbb{R}$. Also the sine function is non-negative in the interval of integration. Hence the integrand is non-negative. Thus $I \geq 0$.
However the above proof assumes the function f to be differentiable in the interval $(0, 2\pi)$. This assumption can be dropped by noting that a convex function possesses only a *countable* number points of non-differentiability in any open interval. See corollary 6.3 of this note.

L. Generalization of Fano's inequality [Han, Verdu]

Problem : Part (1) Define the binary divergence function $d(x||y)$ as the diveregence between the two-mass distributions $(x, 1-x), (y, 1-y)$. If X and Y take values on the same set, then

$$I(X;Y) \geq d(\mathbb{P}[X = Y]||\mathbb{P}[X = \bar{Y}])$$

where $\bar{Y}$ is independent of X and has the same distribution as Y.
Solution : Write the mutual information $I(X;Y)$ as

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_{\bar{Y}}(y)}$$

Now we use *log-sum inequality* to obtain,

$$\begin{aligned} I(X;Y) \geq \Big(\sum_{x=y} P_{XY}(x,y)\Big)\log\frac{\Big(\sum_{x=y} P_{XY}(x,y)\Big)}{\Big(\sum_{x=y} P_X(x)P_{\bar{Y}}(y)\Big)} + \\ \Big(\sum_{x\neq y} P_{XY}(x,y)\Big)\log\frac{\Big(\sum_{x\neq y} P_{XY}(x,y)\Big)}{\Big(\sum_{x\neq y} P_X(x)P_{\bar{Y}}(y)\Big)} \\ = d(\mathbb{P}[X = Y]||\mathbb{P}[X = \bar{Y}]) \end{aligned}$$

Problem : Part (2) If X and Y take values from the same set, then

$$I(X;Y) \geq \mathbb{P}[X = Y]\log\frac{1}{\max_{\omega\in\Omega} P_X(\omega)} - h(\mathbb{P}[X = Y])$$

Solution : Simply follows from the fact that $d(x||y) \geq x\log\frac{1}{y} - h(x)$. ■

LI. Mrs. Gerber's Lemma: Problem 2.5, El Gamal [4]

(a) It can be shown that the function $H(H^{-1}(v) * p)$ is convex in v for every $p \in [0,1]$.
(b) **Scalar MGL :** Given that X is a binary random variable and U is an arbitrary random variable. $Z \sim$ Bern(p) independent of (X, U) and $Y = X \oplus Z$. Fix $U = u$. Then the conditional distribution of the random-variable $Y|U = u \sim$ Bern$(H^{-1}(H(X|U = u)) * p)$. Hence $H(Y|U = u) = H(H^{-1}(H(X|U = u)) * p)$. Finally, taking expectations over the distribution of U, we have

$$\begin{aligned} H(Y|U) &= \mathbb{E}(H(Y|U = u)) \\ &= \mathbb{E}(H(H^{-1}(H(X|U = u)) * p)) \\ &\geq H(H^{-1}(\mathbb{E}H(X|U = u) * p)) \\ &= H(H^{-1}(H(X|U) * p)) \end{aligned}$$

Where we have used the Jensen's inequality on the convex function $H(H^{-1}(v) * p)$.

LII. Functional Analysis : Every Compact metric space (M,d) is Complete

Proof: Take a cauchy sequence $\{x_n\}$ in the space (M,d). Fix $\epsilon > 0$. Hence there exists $N_1 \in \mathbb{N}$ such that for all $m,n \geq N_1$ we have $d(x_m,x_n) \leq \epsilon/2$. Since the space is compact, we can choose a converging subsequence $\{n_k\}$ and a $\bar{x} \in M$ so that $x_{n_k} \to \bar{x}$. Hence there exists $N_2 \in \mathbb{N}$ so that $d(x_{n_k},\bar{x}) \leq \epsilon/2$, $\forall k \geq N_2$. Now consider $d(x_m,\bar{x}) \leq d(x_m,x_{n_k})+d(x_{n_k},\bar{x})$, where we choose $k \geq N_2$ large enough so that $n_k \geq N_1$. Now if $m \geq N_1$, we have $d(x_m,x_{n_k}) \leq \epsilon/2$ and $d(x_{n_k},\bar{x}) \leq \epsilon/2$. Thus $d(x_m,\bar{x}) \leq \epsilon$ for all $m \geq N_1$. Thus $x_m \to \bar{x}$. ■

LIII. Graph Connectivity

Problem : Given a graph $\mathcal{G}(V,E)$, if the degree of every vertex is at least $|V|/2$, show that the graph is connected.

Proof: We will proceed with induction. The proposition is trivial for $|V| = 2,3$. Assume that the proposition holds good for $|V| = n+1$. We will show that it holds good for $|V| = n+2$.

Suppose the contrary for $|V| = n+2$. Hence the graph can be partioned in two non-empty disjoint components $\mathcal{C}_1, \mathcal{C}_2 \subset V$. Let $u \in \mathcal{C}_1$ and $v \in \mathcal{C}_2$. Now remove the vertices u,v along with all edges incident on them. Thus we get a graph with n vertices such that all vertices has degree at least $\frac{n+2}{2}-1 = \frac{n}{2}$, this is because the components $\mathcal{C}_1$ and $\mathcal{C}_2$ were assumed to be *disjoint* and hence no edge exists between u and $\mathcal{C}_2$ and vice-versa.

Hence by the induction assumption, the components $\mathcal{C}_1 \setminus u$ and $\mathcal{C}_2 \setminus v$ are connected and hence the original graph $\mathcal{G}$ is connected. Hence the assumption was false and the induction step is now complete. ■

LIV. Independently generated codebooks: Problem 3.8, El Gamal [4]

Proof: Since the codebooks are generated unifromly at random, it suffices to consider, say, the first indices of both the codebooks $\mathcal{C}_1$ and $\mathcal{C}_2$.

We have $\mathbb{P}\Bigg((\boldsymbol{X}^n(1),\boldsymbol{Y}^n(1)) \in \mathcal{T}_\epsilon^{(n)}(X,Y)\Bigg) \doteq 2^{-nI(X;Y)}$ ■

Now for any two codebooks $\mathcal{C}_1, \mathcal{C}_2$

$$|\mathcal{C}| = \sum_{i=1}^{2^{nR_1}} \sum_{j=1}^{2^{nR_2}} \mathbf{1}_{(\boldsymbol{X}^n(i),\boldsymbol{Y}^n(j)) \in \mathcal{T}_\epsilon^{(n)}(X,Y)} \tag{35}$$

Taking expectation and using the linearity of expectation, we have

$$\begin{aligned}\mathbb{E}|\mathcal{C}| &= \sum_{i=1}^{2^{nR_1}} \sum_{j=1}^{2^{nR_2}} \mathbb{P}\Bigg((\boldsymbol{X}^n(i),\boldsymbol{Y}^n(j)) \in \mathcal{T}_\epsilon^{(n)}(X,Y)\Bigg) \\ &\doteq 2^{n(R_1+R_2-I(X;Y)}\end{aligned}$$

LV. Nonuniform message: Problem 3.7, El Gamal [4]

As suggested, consider a random ensemble of codes $\Phi_n = \phi_n \circ \sigma, \Psi'_n = \sigma^{-1} \circ \psi_n$, where σ is sampled independently and unformly from the set of all permutations and ϕ_n, ψ_n are the given coding-decoding pair sequences. Then for any index i

$$\begin{aligned}\mathbb{P}(\sigma(M') = i) &= \sum_{k=1}^{2^{nR}} \mathbb{P}(M' = k, \sigma(k) = i) \\ &\overset{(1)}{=} \sum_{k=1}^{2^{nR}} \mathbb{P}(M' = k)\mathbb{P}(\sigma(k) = i) \\ &\overset{(2)}{=} \frac{1}{2^{nR}} = \mathbb{P}(M = i)\end{aligned}$$

Where (1) follows from independence of σ and M' and (2) follows from uniformity of σ. Similarly we can show pairwise independence of different message sets. Thus the result follows.

LVI. List Codes: Problem 3.14, El Gamal [4]

(a) Consider the following coding strategy: Let $\mathcal{C}^n$ be a capacity achieving sequence of code of rate $R-L$. We group 2^{nR} messages into $2^{n(R-L)}$ groups of 2^{nL} codewords each. We associate a single codeword from $\mathcal{C}^n$ for each group and transmit this code over the DMC and decode via joint typicality. Finally, once we decode a group index, we reveal the associated 2^{nL} messages as list. Since $R-L < C$, via channel coding theorem, probability of error of the above coding strategy goes to zero.

LVII. Energy-aware Krafts inequality

Problem : Suppose we have a binary channel where a 0 takes 1 unit of energy to transmit and 1 takes 2 units of energy transmit. Suppose there exists a prefix-free code for a universe $U = \{a_1,\ldots,a_n\}$ such that the codeword for a_i takes e_i units of energy to transmit. Show that

$$\sum_{i=1}^{n} \phi^{e_i} \leq 1$$

where $\phi = \frac{\sqrt{5}-1}{2}$.

Proof The key to the problem is the obsservation $\phi+\phi^2 = 1$. Now we follow exactly the same line of inductive argument used to prove the Krafts inequality. Since the code is Prefix-free, the codewords can be arranged on the leaves of a binary tree. Hence we will be done if we prove the inequality for all binary trees. For this we do induction on maximal codeword lengths.

Let (n_{0i}, n_{1i}) denote the number of zeros and ones for the i^{th} codeword. The base case for $n_{11}+n_{01} = 1$ is trivial. Assume that the induction hypothesis hold for a subtree with codeword energies $e_i, i = 1,2,\ldots,n-1$. To complete the induction step, we split a leaf and put two leafs as its children. If the leaf to be split has n_1 1 s and n_0 zeros, then the contribution of the term in the given sum is $\phi^{2n_1+n_0}$. On the other hand, in the new code book, we get (atmost) two different codewords with contributions $\phi^{2n_1+2+n_0}$ and $\phi^{2n_1+n_0+1}$. Hence their combined contribution to the sum is $\phi^{2n_1+n_0}(\phi^2+\phi) = \phi^{2n_1+n_0}$, which is the same as previous split leaf. Hence the proof follows by the induction assumption.

LVIII. Exercise 8.16 from Algorithms-Vazirani [5]

Since 3SAT is poly-time reducible to Independent Set (see text), a poly-time reduction from Independent Set to Experimental Cuisine exhibits the desired reduction.
Let $\mathcal{G}(V, E)$ be an instance of Independent Set problem. Let A be the adjacency matrix of the graph $\mathcal{G}$. We now invoke Experimental Cuisine with the discord matrix $D = A$ and $p = 0$. A little thought shows that the output is the maximum size of independent set of the graph $\mathcal{G}$. This completes the reduction. ■

LIX. Exercise 8.7 from Algorithms-Vazirani [5]

Consider a bipartite graph $\mathcal{G}(V_1, V_2, E)$, where the each of the vertices in V_1 corresponds to a clause and each of the vertices in V_2 corresponds to a variable. If the variable appears in a clause (in either complemented or uncomplemented form), there is an edge $e \in E$ between them. Now first we show that the graph $\mathcal{G}$ has a matching saturating V_1.
Take any subset $S \subset V_1$. Let $E_S \subset E$ be the set of edges incident on any vertex in S and let $\Gamma(S) \subset V_2$ be the set of neighbours of S. Then we have,

$$3|S| = |E_S| \leq 3|\Gamma(S)|$$

Hence $|\Gamma(S)| \geq |S|$ for any $S \subset V_1$ and by **Hall's theorem**, we conclude that $\mathcal{G}$ has a matching saturating V_1.
Let $\mathcal{M}$ be such a matching. If $v_1v_2 \in \mathcal{M}$, where $v_1 \in V_1$ and $v_2 \in V_2$, we set $x_{v_2} = 1$ if x_{v_2} appears in the clause corresponding v_1 in uncomplemented form or set it to 0 otherwise. The above construction shows that there always exists a satisfying assignment and gives an efficient method of finding it. ■

LX. Exercise 8.22 from Algorithms-Vazirani [5]

(b) Suppose the given undirected graph has a vertex cover given by the set $S \subset V$, with $|S| = b$. We will show that there exists a Feedback arc set of size b in the constructed graph $\mathcal{G}'$.
For this, we remove the edges from the graph $\mathcal{G}'$ as follows. For all $v \in S$, remove the edge vv' from the graph $\mathcal{G}'$. Clearly there are b of them. We will show that the resulting graph is acyclic.
First note that, by construction, any directed cycle in the graph $\mathcal{G}'$ must be of the form $aa'bb' \ldots zz'a$. Also, since the set S forms a vertex cover of the graph $\mathcal{G}$, every edge $e \in E$ must be incident on some node in S. Translating this condition on the graph $\mathcal{G}'$, we conclude that, for every edge of the form $m'n$ in the graph $\mathcal{G}'$, either the corresponding node $m \in S$ or the node $n \in S$. Hence either the edge mm' or the edge nn' was removed from the graph $\mathcal{G}'$ by the above process. Hence the directed cycle is not possible and this proves the statement. ■
(c) Now assume that $\mathcal{G}'$ contains a feedback arc set of size b. Hence there exists edge set E_F with $|E_F| = b$ such that $\mathcal{G}' \setminus E_F$ is acyclic. The edges in the set E_F are of the form aa' or $a'c$, for some nodes a, c in the graph $\mathcal{G}$. Note that, since all the cycles involving edges of the form $a'c$ include the edge cc', we can instead retain the edge $a'c$ and remove the edge cc' (if it is not already removed) to obtain another feedback arc set of size at most b, where all the removed edges E_F is of the form aa', for some $a \in V$ and $|E_F| \leq b$.
Now in the graph $\mathcal{G}$, we put the vertices v corresponding to the removed edges vv', into the the set S. We claim that S is a vertex cover of $\mathcal{G}$.
To prove this, assume the contrary. Hence there exists an edge $ac \in E$ such that neither a nor c is in S. Translating this condition in the graph $\mathcal{G}' \setminus E_F$, we conclude that there exists a cycle $aa'cc'a$ which is a contradiction to the fact that $\mathcal{G}' \setminus E_F$ is acyclic. ■

LXI. Exercise 9.6 from Algorithms-Vazirani[5]

Double the edges of the optimum Steiner Tree and obtain S. Consider an Eulerian tour over S. Let T be a spanning tree over the terminal vertices only. Then due to the triangle inequality, cost of T may be upper bounded by the cost of the tour of S which is double of OPT. ■

LXII. Exercise 6.4 from Algorithms-Vazirani [5]

(a) Initialize a binary array $V[\cdot]$ of length n. The value $V[i]$ will be 1 if dict$(w[i], n)$ is TRUE and zero otherwise. Clearly, the entered text will be a sequence of valid words iff $V[1]$ is 1. We have the following dynamic programming backward recursion equation:

$$V[i-1] = \max_{k=i-1}^{n} \text{dict}(w[i-1], w(k))V(k+1)$$

Which has a running time of at most $\mathcal{O}(n^2)$.

LXIII. Computational and Statistical Learning Theory : HW2

http://ttic.uchicago.edu/~nati/Teaching/TTIC31120/hw2.pdf
1. **Shatter Lemma :** Setup: $S = \{x_1, x_2, \ldots, x_m\}$, $\mathcal{H}_{x_1,x_2,\ldots,x_m} = \{\big(h(x_1), \ldots, h(x_m)\big) \in \{\pm 1\}^m : h \in \mathcal{H}\}$. and the *growth function* of the hypothesis class $\mathcal{H}$ is given by

$$\Pi_{\mathcal{H}}(m) = \sup_{x_1,x_2,\ldots,x_m} |\mathcal{H}_{x_1,x_2,\ldots,x_m}|$$

We say that $\mathcal{H}$ is *shattered* by $\mathcal{H}$ if $|\mathcal{H}_{x_1\ldots,x_m}| = 2^m$. The VC dimension is defined as the largest m such that $\Pi_{\mathcal{H}}(m) = 2^m$. We want to show that if $\mathcal{H}$ has VC dimension of d then

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

Proof: In oreder tp prove that, we actually prove the following statement: for any $S = \{x_1, x_2, \ldots, x_m\}$:

$$|\mathcal{H}_S| \leq |\{B \subset S : B \text{ is shattered by } \mathcal{H}\}| \qquad (36)$$

(a) When S is empty, both sides of Eqn. (36) is zero and hence the base case is clear.
(b) Define $S' = S \cup \{x'\}$ and write $\mathcal{H} = \mathcal{H}^- \cup \mathcal{H}^+$, where:

$$\mathcal{H}^+ = \{h \in \mathcal{H} : h(x') = +1\}$$
$$\mathcal{H}^- = \{h \in \mathcal{H} : h(x') = -1\}$$

Then clearly, we have

$$|\mathcal{H}_{S'}| = |\mathcal{H}_S^+| + |\mathcal{H}_S^-|$$

(d) We have,

$$\begin{aligned}\binom{m}{i} &= \frac{m(m-1)(m-2)\dots(m-(i-1))}{i!} \\ &\le \frac{m^i}{i!} \\ &= (\frac{m}{d})^i\frac{d^i}{i!} \\ &\le (\frac{m}{d})^d\frac{d^i}{i!}\end{aligned}$$

Summing over $i = 0$ to d , we have

$$\sum_{i=0}^{d}\binom{m}{i} \le (\frac{m}{d})^d\sum_{i=0}^{d}\frac{d^i}{i!} \le \left(\frac{em}{d}\right)^d$$

$\blacksquare$

LXIV. EXERCISE 6.29 FROM ALGORITHMS-VAZIRANI

Let $W[i,j]$ be the function that returns the local weight of the segment $x[i,j]$. We assume that the function $W[\cdot]$ is efficiently computable.
Let $S[\cdot]$ denote the $1 \times n$ array, whose i^{th} component stores the optimum weight of the segment $x[1,i]$. Our objective is to find $S[n]$.
For this we have the following forward DP recursions:

$$\begin{aligned}S[0] &= 0 \\ S[i+1] &= \max\{S(i), \max_{1\le j\le i}\{S(j-1)+W[j,i+1]\}\}\end{aligned}$$

If the maximum is attained by the later term in the above equation, we simply store the index of *an* $\arg\max$ term corresponding to the term i in an array $A[\cdot]$, otherwise we store a null character corresponding to it (meaning that the i^{th} term was not used). The optimal exon chaning can be recosntructed with the array $A[\cdot]$.

LXV. EXERCISE 6.30 FROM ALGORITHMS-VAZIRANI [5]

(b) It is clear that we can optimize over one position at a time. Let N denote the array of all internal nodes and L denotes the set of all leaf nodes of the given tree, sorted in a way such that the children of a node appears before a node in the array. A simple post-order traversal over the tree constructs this array.
Let $\mathcal{C}[n]$ denote the array of children of the internal node $n \in N$. For each position l of the given strings, $1 \le l \le k$, and for each internal node $n \in N$, we compute four variables $D^l[n,A], D^l[n,G], D^l[n,T], D^l[n,C]$, as follows:

$$D^l[n,\cdot] = \min_{\alpha_1,\alpha_2,\dots,\alpha_{|\mathcal{C}(n)|}}\sum_{i=1}^{|\mathcal{C}(n)|}\left(D^l[\mathcal{C}(n,i),\alpha_i]+\delta(\cdot,\alpha_i)\right)$$

And for each leaf node $n \in L$, we initialize by

$$\begin{aligned}D^l[n,\alpha] &= 0, \text{ if } \quad n(l)=\alpha \\ &= \infty, \text{ o.w.}\end{aligned}$$

Where in the above minimization each α_i takes all four values $\{A,G,T,C\}$. $\delta(\alpha,\beta) = 0$ if $\alpha = \beta$ and is 1 otherwise. If the internal degree of the tree is bounded (e.g., a binary tree), the above procedure is efficient.

LXVI. *Variations on the joint typicality lemma*: PROBLEM 2.14, EL GAMAL [4]

(a) Given $(X^n,Y^n) \sim \prod_{i=1}^n p_{X,Y}(x_i,y_i)$ and $\tilde{Z}^n|\{X^n = x^n, Y^n = y^n\} \sim \prod_{i=1}^n p_{Z|X}(\tilde{z}_i|x_i)$. Hence

$$\begin{aligned}&\mathbb{P}\{(X^n,Y^n,\tilde{Z}^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y,Z)\} \\ =& \sum_{(x^n,y^n,z^n)\in\mathcal{T}_\epsilon^{(n)}(X,Y,Z)}\prod_{i=1}^n p_{X,Y}(x_i,y_i)\prod_{i=1}^n p_{Z|X}(z_i|x_i) \\ \doteq& 2^{nH(X,Y,Z)}2^{-nH(X,Y)}2^{-nH(Z|X)} \\ =& 2^{-n(H(Z|X)-H(Z|X,Y)} = 2^{-nI(Y;Z|X)}\blacksquare\end{aligned}$$

(b) Given $(x^n,y^n) \in \mathcal{T}_{\epsilon'}^{(n)}$ and $\tilde{Z}^n \sim$ Unif$(\mathcal{T}_\epsilon^{(n)}(Z|x^n))$. Hence,

$$\begin{aligned}&\mathbb{P}\{(x^n,y^n,\tilde{Z}^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y,Z)\} \\ =& \sum_{z^n\in\mathcal{T}_\epsilon^{(n)}(Z|x^n,y^n)}\mathbb{P}(\tilde{Z}^n = z^n)\end{aligned}$$

Now note that $\mathcal{T}_\epsilon^{(n)}(Z|x^n,y^n) \subset \mathcal{T}_\epsilon^{(n)}(Z|x^n)$. Hence we can write the above probability as

$$\begin{aligned}=& \frac{|\mathcal{T}_\epsilon^{(n)}(Z|x^n,y^n)|}{|\mathcal{T}_\epsilon^{(n)}(Z|x^n)|} \\ \doteq& 2^{-n(H(Z|X)-H(Z|X,Y))} = 2^{-nI(Y;Z|X)}\blacksquare\end{aligned}$$

(c) Let the random variable $\tilde{Z}^n$ have the distribution as given. Then,

$$\begin{aligned}&\mathbb{P}\{(x^n,\tilde{y}^n,\tilde{Z}^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y,Z)) \\ =& \sum_{z^n\in\mathcal{T}_\epsilon^{(n)}(Z|x^n,\tilde{y}^n)}p(z^n|x^n) \\ \le& 2^{-n(H(Z|X)-H(Z|X,Y)-\delta(\epsilon))} \le 2^{-n(I(Y;Z|X)-\delta(\epsilon))}\blacksquare\end{aligned}$$

(d) We have,

$$\begin{aligned}&\mathbb{P}\{(\tilde{X}^n,\tilde{Y}^n,\tilde{Z}^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y,Z)\} \\ =& \sum_{(x^n,y^n,z^n)\in\mathcal{T}_\epsilon^{(n)}(X,Y,Z)}\prod_{i=1}^n p_X(x_i)p_Y(y_i)p_{Z|X,Y}(z_i|x_i,y_i) \\ \doteq& 2^{nH(X,Y,Z)}2^{-nH(X)}2^{-nH(Y)}2^{-nH(Z|X,Y)} \\ =& 2^{-nI(X;Y)}\blacksquare\end{aligned}$$

LXVII. *Need for both ϵ and ϵ'*: PROBLEM 2.17, EL GAMAL [4]

(a) We have $\pi(1|x^i) = k/n$, where $\frac{n}{2}(1+\epsilon)-1 < k \le \frac{n}{2}(1+\epsilon)$. Thus,

$$\frac{\epsilon}{2}-\frac{1}{n} \le \pi(1|x^i)-\frac{1}{2} \le \frac{\epsilon}{2}$$

Hence,

$$|\pi(1|x^i)-p(1)| \le \epsilon p(1)$$

Similarly, it can be shown that

$$|\pi(0|x^i)-p(0)| \le \epsilon p(0)$$

Thus, $x^n \in \mathcal{T}_\epsilon^{(n)}(X)$. $\blacksquare$
(b) Now Y^n is given to be an i.i.d. Bern(1/2) sequence,

independent of x^n. By definition, a sequence $(x^n, y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)$ only if,

$$\frac{1}{n}\pi(1,1|x^n,y^n) - \frac{1}{4} \le \frac{\epsilon}{4}$$

i.e.,

$$\pi(1,1|x^n,y^n) \le \frac{n}{4}(1+\epsilon) \qquad (37)$$

Note that since the first k symbols of x^n is 1 and the rest are zero, we can write $\pi(1,1|x^n,y^n) = \sum_{i=1}^{k} Y_i$. Now if $\sum_{i=1}^{k} Y_i \ge \frac{(k+1)}{2}$, then we clearly have $\sum_{i=1}^{k} Y_i > \frac{n}{4}(1+\epsilon)$ and the necessary condition (37) is violated. Thus, a sequence $(x^n,y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)$ only if $\sum_{i=1}^{k} Y_i < \frac{(k+1)}{2}$. Hence,

$$\mathbb{P}\{(x^n,y^n) \in \mathcal{T}_\epsilon^{(n)}(X,Y)\} \le \mathbb{P}\left\{\sum_{i=1}^{k} Y_i < \frac{(k+1)}{2}\right\} \blacksquare$$

LXVIII. Entropy Inequality

Problem : If X, Y, Z are three random variables, show that

$$2H(X,Y,Z) \le H(X,Y) + H(Y,Z) + H(Z,X)$$

Proof: We have,

$$\begin{aligned} H(X,Y,Z) &= H(X,Y) + H(Z|X,Y) \\ &= H(Y,Z) + H(X|Y,Z) \\ &= H(Z,X) + H(Y|Z,X) \end{aligned}$$

Adding the above three equalities, we obtain

$$\begin{aligned} 3H(X,Y,Z) &= H(X,Y) + H(Y,Z) + H(Z,X) + \\ &\quad (H(Z|X,Y) + H(X|Y,Z) + H(Y|Z,X)) \end{aligned} \qquad (38)$$

Again, we have

$$\begin{aligned} H(X,Y,Z) &= H(X) + H(Y|X) + H(Z|X,Y) \\ &\ge H(X|Y,Z) + H(Y|X,Z) + H(Z|X,Y) \end{aligned}$$

Where we have used the fact that conditioning reduces entropy. Using the above two equations, we obtain the desired result.

■

LXIX. Entropy Inequality II: Gallager 2.17(b) [6]

Problem Let $a_1, a_2, \ldots, a_k$ be a set of disjoint events and let $\big(P(a_1), P(a_2)\ldots, P(a_k)\big)$ and $\big(Q(a_1), Q(a_2), \ldots, Q(a_k)\big)$ be two probability assignments on the events. Show that,

$$\sum_{k=1}^{K} \frac{[P(a_k)]^2}{Q(a_k)} \ge 1$$

Proof: Note that the function $f(x) = \frac{1}{x}$ is convex. Using Jensen's inequality for any probability distribution, we have for any r.v. X

$$\mathbb{E}f(X) \ge f(\mathbb{E}X) \qquad (39)$$

Now, the given quantity is

$$\sum_{k=1}^{K} \frac{[P(a_k)]^2}{Q(a_k)} = \sum_{k=1}^{K} P(a_k)\left(\frac{Q(a_k)}{P(a_k)}\right)^{-1} = \mathbb{E}f(Z)$$

where Z is a random variable which takes the value $\frac{Q(a_k)}{P(a_k)}$ w.p. $P(a_k), k = 1,2,\ldots,K$. Thus $\mathbb{E}Z = \sum_{k=1}^{K} P(a_k)\frac{Q(a_k)}{P(a_k)} = \sum_{k=1}^{K} Q(a_k) = 1$.
Using the inequality (39), we have

$$\sum_{k=1}^{K} \frac{[P(a_k)]^2}{Q(a_k)} \ge f(\mathbb{E}Z) = f(1) = 1 \quad \blacksquare$$

■

LXX. Parallel Channels: Gallager 4.18 (a) [6]

Problem : Consider n (in general different) DMC's with capacities $C_1, C_2, \ldots, C_n$. The "sum" channel associated therewith is that channel whose input and output alphabets are the unions of those of the original channels: i.e., the sum channel has all n channels available for use but only one channel may be used at any given time. Show that the capacity of the sum channel is given by

$$C = \log_2 \sum_{i=1}^{n} 2^{C_i}$$

and find $q(i)$, the probability of using the i^{th} channel.

Proof: Here we are coding over the channel C and the symbol of that particular channel X. The joint distribution of the variables (C,X,Y) is given by

$$p(C,X,Y) = q(C)\tilde{p}(X|C)P(Y|X,C)$$

Hence we would like to maximize the mutual information $I(C,X;Y)$ over all input distributions $q(C)\tilde{p}(X|C)$. Now,

$$\begin{aligned} I(C,X;Y) &= I(C;Y) + I(X;Y|C) \\ &= H(C) - H(C|Y) + \sum_{i=1}^{n} q(i)I(X_i;Y_i|C=i) \\ &\overset{(a)}{=} H(\boldsymbol{q}) + \sum_{i=1}^{n} q(i)I(X_i;Y_i|C=i) \end{aligned}$$

Where we have used the fact that the ouput alphabets of different channels are non-overlapping. Hence

$$\sup_{\tilde{p}(X_i|C=i)} I(C,X;Y) = H(\boldsymbol{q}) + \sum_{i=1}^{n} q(i)C_i$$

To maximize the above expression w.r.t. $\boldsymbol{q}$, we use the non-negativity property of the differential entropy. Let $\boldsymbol{r}$ denote the following distribution

$$r(i) = \frac{2^{C_i}}{\sum_k 2^{C_k}}, \quad \forall i = 1,2,\ldots,n$$

We have,

$$D(\boldsymbol{q}||\boldsymbol{r}) \ge 0$$

i.e.,

$$\begin{aligned} \sum_{i=1}^{n} q_i \log(q_i/r_i) &\ge 0 \\ -H(\boldsymbol{q}) - \sum_{i=1}^{n} q_i\big(C_i - \log\sum_{k=1}^{n} 2^{C_k}\big) &\ge 0 \end{aligned}$$

i.e.,

$$H(\boldsymbol{q}) + \sum_{i=1}^{n} q(i)C_i \leq \log\big(\sum_{k=1}^{n} 2^{C_k}\big)$$

and the equality is achievable when $\boldsymbol{q} = \boldsymbol{r}$. ■

LXXI. An Upper bound on probability of error: Gallager 4.7 [6]

Problem : Letting U and V have the same sample space, say $a_1, a_2, \ldots, a_k$, minimum error probability decoding has the property that $P_{U|V}(a_i|a_i) \geq P_{U|V}(a_k|a_i), \forall k \neq i$. Using this property, show that

$$P_e \leq H(U|V) \text{ nats}$$

Proof: Suppose we observe $V = v$. In this even, the probability of decoding error is given by $P_{e|v} = 1 - \max_u P_{U|V}(u|v)$. Now,

$$\begin{aligned}
& H(U|V=v) \\
= \; & -\sum_u P_{U|v}(u|V=v)\log\big(P_{U|V=v}(u|V=v)\big) \\
\overset{(1)}{\geq} \; & \sum_u P_{U|v}(u|V=v)\big(1 - P_{U|V=v}(u|V=v)\big) \\
\geq \; & \sum_u P_{U|v}(u|V=v)\big(1 - \max_u P_{U|V=v}(u|V=v)\big) \\
= \; & P_{e|v}
\end{aligned}$$

where in (1) we have used the inequality $\exp(x) \geq 1 + x$. Taking expecation of the above inequality w.r.t. the distribution of the r.v. V, the result follows. ■

LXXII. Properties of $R(D)$: Cover and Thomas 10.4

Consider a test-channel $p(\cdot|\cdot)$ for the distortion measure $d(\cdot,\cdot)$ with expected distortion $D + \bar{w}$. The expected distortion for the distortion measure $d'(\cdot|\cdot)$ with the channel $p(\cdot|\cdot)$ is given by

$$\begin{aligned}
\mathbb{E}d'(X,\hat{X}) &= \sum_{i,j} p(i)p(j|i)d'(i,j) \\
&= \sum_{i,j} p(i)p(j|i)\big(d(i,j) - w(i)\big) \\
&= D + \bar{w} - \bar{w} \\
&= D
\end{aligned}$$

Thus, we conlcude that

$$R'(D) \leq R(D + \bar{w}) \tag{40}$$

Since, $d(i,j) = d'(i,j) + w_i$, by a similar argument, we have

$$R(D) \leq R'(D - \bar{w}) \tag{41}$$

Combining Eqns (40) and (41), we obtain the result.

LXXIII. Shannon Lower Bound for the Rate Distortion Function: Cover and Thomas 10.4

Given

$$\phi(D) = \max_{\boldsymbol{p}:\sum_{i=1}^{m} p_i d_i \leq D} H(\boldsymbol{p}) \tag{42}$$

(a) Let $0 \leq \lambda \leq 1$. Let $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ be two distributions achieving the optimality in Eqn. (42) for distortions D_1 and D_2 resp. Consider another PMF $\boldsymbol{p}^\lambda = \lambda\boldsymbol{p}_1 + (1-\lambda)\boldsymbol{p}_2$. Clearly, $\sum_{i=1}^{n} p_i^\lambda d_i \leq \lambda D_1 + (1-\lambda)D_2$. Hence,

$$\begin{aligned}
\phi(\lambda D_1 + (1-\lambda)D_2) &\geq H(\boldsymbol{p}^\lambda) \\
&= H(\lambda\boldsymbol{p}_1 + (1-\lambda)\boldsymbol{p}_2) \\
&\overset{(1)}{\geq} \lambda H(\boldsymbol{p}_1) + (1-\lambda)H(\boldsymbol{p}_2) \\
&= \lambda\phi(D_1) + (1-\lambda)\phi(D_2)
\end{aligned}$$

where the inequality (1) follows from concavity of $H(\cdot)$. The above inequality shows that $\phi(D)$ is a concave function of D. ■

LXXIV. Submodularity of cuts

Problem : Given a directed graph $\mathcal{G}(V,E)$, denote the capacity of the cut associated with a vertex set $S \subset V$ by $\delta(S)$. Then for any sets $S, A \subset V$,

$$\delta(S \cup A) + \delta(S \cap A) \leq \delta(S) + \delta(A)$$

Proof: With a slight abuse of notation, let us denote number of directed edges going from set A to B by $\delta(A \to B)$. Then,

$$\begin{aligned}
&= \delta(S \cup A) + \delta(S \cap A) \\
&= \delta(S \cup A \to S^c \cap A^c) + \delta(S \cap A \to S^c \cup A^c) \\
&= \delta(S \to S^c \cap A^c) + \delta(A \to S^c \cap A^c) \\
&\quad -\delta(S \cap A \to S^c \cap A^c) + \delta(S \cap A \to S^c \cup A^c) \\
&= \delta(S \to S^c \cap A^c) + \delta(A \to S^c \cap A^c) + \\
&\quad \delta(S \cap A \to S^c \cap A) + \delta(S \cap A \to S \cap A^c) \\
&\leq \big(\delta(S \to S^c \cap A^c) + \delta(S \to S^c \cap A)\big) \\
&\quad +\big(\delta(A \to S^c \cap A^c) + \delta(A \to S \cap A^c)\big) \\
&= \delta(S \to S^c) + \delta(A \to A^c) \\
&= \delta(S) + \delta(A)
\end{aligned}$$

■

LXXV. Rate Distortion : Cover and Thomas 10.20 [1]

Given two distortion measures $d_1(x,\hat{x})$ and $d_2(x,\hat{x})$ such that $d_1(x,\hat{x}) \leq d_2(x,\hat{x}), \forall x \in X, \hat{x} \in \hat{X}$. Let $R_1(D)$ and $R_2(D)$ be the corresponding distortion measures.
(a) Since $d_1(x,\hat{x}) \leq d_2(x,\hat{x}), \forall x \in X, \hat{x} \in \hat{X}$, the optimal test channel for the second distortion measure $\boldsymbol{p}_2(\hat{x}|x)$ is also a feasible test channel for the first distortion measure. Hence,

$$R_1(D) \leq R_2(D)$$

(b) Since $d_2(X^n, \hat{X}_2^n) \leq D$ implies $d_1(X^n, \hat{X}_1^n) \leq D$, it is sufficient to describe the source at rate R_2.

LXXVI. Lower Bound : Cover and Thomas 10.11 [1]

Given,

$$X \sim \boldsymbol{p} \equiv \frac{e^{-x^4}}{\int_{-\infty}^{\infty} e^{-x^4} dx}$$

with $\mathbb{E}X^4 \leq c$. First we show that $\boldsymbol{p}$ has maximum entropy among all distributions $\boldsymbol{q}$ with $\mathbb{E}_q X^4 = c$. To prove this fact, consider the differential entropy

$$\begin{aligned} D(\boldsymbol{q}||\boldsymbol{p}) &= \int_{-\infty}^{\infty} q(x) \log \frac{q(x)}{p(x)} dx \\ &= -h(\boldsymbol{q}) + \int_{-\infty}^{\infty} x^4 q(x) dx + C_1 \end{aligned}$$

where, $C_1 = \log(\int_{\infty}^{\infty} e^{-x^4} dx)$, an absolute constant. Since $\mathbb{E}_q X^4 \leq c$, $D(\boldsymbol{q}||\boldsymbol{p})$ is upper bounded by $-h(\boldsymbol{q}) + c + C_1$ and lower bounded by zero. Hence, we obtain that

$$h(\boldsymbol{q}) \leq c + C_1$$

where the equality is achieved, i.e., the differential entropy is maximized when $\boldsymbol{q} = \boldsymbol{p}$ a.e.

The proof is now completed by invoking the Shannon lower bound of the rate distortion, i.e.,

$$R(D) \geq H(X) - g(D) \overset{(a)}{=} g(c) - g(D)$$

where the equality (a) follows from the above argument.

LXXVII. Adding a column to the distortion matrix: Cover and Thomas 10.12

We always have the option of not using the additional reproduction symbol. Hence this can only reduce $R(D)$.

LXXVIII. Rate Distortion for two independent sources: Cover and Thomas 10.14 [1]

(a) First we note that if $X \perp Y$, then

$$\begin{aligned} I(X,Y;\hat{X},\hat{Y}) &= H(X,Y) - H(X,Y|\hat{X},\hat{Y}) \\ &\overset{(a)}{=} H(X) + H(Y) - H(X,Y|\hat{X},\hat{Y}) \\ &\overset{(b)}{\geq} H(X) + H(Y) - H(X|\hat{X}) - H(Y|\hat{Y}) \\ &= I(X;\hat{X}) + I(Y;\hat{Y}) \end{aligned} \tag{43}$$

where (a) follows from the independence of X and Y and (b) follows from the fact that conditioning reduces entropy.

Now fix a test channel $p(\hat{x},\hat{y}|x,y)$ achieving $R(D_1,D_2)$. The corresponding marginals $p(\hat{x}|x)$ and $p(\hat{y}|y)$ can be computed as follows. We have,

$$p(\hat{x},\hat{y}|x,y) = \frac{p(\hat{x},\hat{y},y|x)}{p(y|x)} = \frac{p(\hat{x},\hat{y},y|x)}{p(y)}$$

where we have used independence of X and Y in the last equality. Thus,

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

Hence,

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

Similarly,

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

These two conditional probabilities satisfy the constraints $\mathbb{E}d(X,\hat{X}) \leq D_1, \mathbb{E}d(Y,\hat{Y}) \leq D_2$. Hence from Eqn. (43), we conclude that

$$R_{X,Y}(D_1,D_2) \geq R_X(D_1) + R_Y(D_2) \quad \blacksquare$$

(b) From the fundamental definition of rate-distortion, it follows that if we simply use the optimal rate distortion code for X and Y separately, and transmit them as a cartesian product, we incur a rate of $R_X(D_1) + R_Y(D_2)$ with the prescribed distortion for both X and Y. From the operational definition, it means to use $p(\hat{x},\hat{y}|x,y) = p(\hat{x}|x)p(\hat{y}|y)$ where $p(\hat{x}|x)$ and $p(\hat{y}|y)$ are the optimal test channels for X and Y with distortion measure D_1 and D_2 respectively. Hence the equality can be definitely achieved.

From the above argument it follows that one **can not** compress two independent sources simultaneously better than by compressing the sources individually.

LXXIX. Concavity of determinants: Cover and Thomas 8.2 [1]

As given in the hint, let $\boldsymbol{Z} = \boldsymbol{X}_\theta$, where $\boldsymbol{X}_1 \sim N(0,K_1), \boldsymbol{X}_2 \sim N(0,K_2)$ and $\theta =$ Bernoulli(λ). Then we use $h(\boldsymbol{Z}|\theta) \leq h(\boldsymbol{Z})$. We have

$$h(\boldsymbol{Z}|\theta) = \frac{1}{2}\log(2\pi e)^n + \big(\lambda \log|K_1| + \bar{\lambda}\log|K_2|\big) \tag{44}$$

On the other hand, the covariance for $\boldsymbol{Z}$ is given by

$$\begin{aligned} \mathbb{E}\boldsymbol{Z}\boldsymbol{Z}^T &= \mathbb{E}_\theta \mathbb{E}_{\boldsymbol{Z}|\theta} \boldsymbol{Z}\boldsymbol{Z}^T = \lambda \mathbb{E}\boldsymbol{X_1}\boldsymbol{X_1}^T + \bar{\lambda}\mathbb{E}\boldsymbol{X_2}\boldsymbol{X_2}^T \\ &= \lambda K_1 + \bar{\lambda} K_2 \end{aligned}$$

Since with a specified covariance matrix, jointly normal distribution maximizes differential entropy, we have

$$h(\boldsymbol{Z}) \leq \frac{1}{2}\log(2\pi e)^n + \log|\lambda K_1 + \bar{\lambda} K_2| \tag{45}$$

Hence combining equations (44) and (45) with the given inequality, the result follows. $\blacksquare$

LXXX. Martingales: Stochastic Processes, Ross, Problem 6.2 [7]

We have the Martinagle sequence $\{Z_n, n \geq 1\}, Z_0 \equiv 0$ and the difference sequence $\{X_i\}$, where $X_i = Z_i - Z_{i-1}, i \geq 1$. Clearly, $\mathbb{E}Z_i = 0, \mathbb{E}X_i = 0, \forall i \geq 1$. Hence Var$(Z_i) = \mathbb{E}Z_i^2$, Var$(X_i) = \mathbb{E}X_i^2, \forall i \geq 1$. Now,

$$\begin{aligned} \text{Var}(Z_n) &= \mathbb{E}Z_n^2 \\ &= \mathbb{E}(X_n + Z_{n-1})^2 \\ &= \mathbb{E}(X_n^2 + Z_{n-1}^2 + 2X_n Z_{n-1}) \\ &= \mathbb{E}Z_{n-1}^2 + \mathbb{E}X_n^2 + 2\mathbb{E}Z_{n-1}(Z_n - Z_{n-1}) \\ &\overset{(a)}{=} \text{Var}(Z_{n-1}) + \mathbb{E}X_n^2 + 2\mathbb{E}(\mathbb{E}Z_{n-1}(Z_n - Z_{n-1})|\mathcal{F}_{n-1})) \\ &= \text{Var}(Z_{n-1}) + \mathbb{E}X_n^2 \end{aligned}$$

The equation (a) follows due to the property of Martinagle differences. The final result follows now by induction.

LXXXI. Martingales : Stochastic Processes, Ross, Problem 6.11 [7]

Let us denote the gain of the gambler on the i^{th} game by the random variable $X_i, i \geq 1$, where X_i's are i.i.d. Each of the X_i s take values from the set $\{1, -1\}$ with equal probability, thus $\mathbb{E}X_i = 0, \mathbb{E}X_i^2 = 1, \forall i \geq 1$. Total winnings of the gambler upto n th game is given by the r.v. S_n where $S_n = \sum_{i=1}^n X_i$. The gambler quits when S_n reaches either a or $-b$. Let T denote the random time when the gambler quits. Note that T is a stopping time w.r.t. the σ field generated by $\{S_n\}_{n \geq 1}$. It is also clear that the sequence $|S_{n \wedge T}|$ is uniformly bounded. Hence using the optional stopping theorem on the zero-mean martingale Sequence $\{S_n\}$ and the stopping time T we conclude that,

$$\mathbb{E}S_T = 0$$

i.e.,

$$a\mathbb{P}(S_T = a) - b\mathbb{P}(S_T = -b) = 0$$

Hence, $\mathbb{P}(S_T = a) = \frac{b}{a+b}, \mathbb{P}(S_T = -b) = \frac{a}{a+b}$ Finally, we check that the sequence of random variables $\{Z_n\}, n \geq 1$, where $Z_n = S_n^2 - n$ is also a zero-mean martingale. Hence using the Optional Stopping Theorem on $\{Z_n\}$ with the stopping time T, we conclude that

$$\mathbb{E}Z_T = 0, \text{i.e. } \mathbb{E}\big(S_T^2 - T\big) = 0$$

Hence,

$$\mathbb{E}(T) = \mathbb{E}(S_T^2) = a^2\frac{b}{a+b} + (-b)^2\frac{a}{a+b} = ab \quad \blacksquare$$

LXXXII. Martingales : Stochastic Processes, Ross, Problem 6.19 [7]

Define $D(\boldsymbol{x}) = \min_{y \in A} \rho(\boldsymbol{x}, \boldsymbol{y})$. Note that if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ differs in only one place then,

$$|D(\boldsymbol{x}_1) - D(\boldsymbol{x}_2)| \leq 1$$

Hence, using the corollary 6.3.4 we have for $a > 0$:

$$\mathbb{P}(D(\boldsymbol{X}) - \mathbb{E}D \geq a) \leq \exp\big(-\frac{a^2}{2n}\big)$$

Hence for $b > \mathbb{E}(D) = \mu$, we have

$$\begin{aligned}\mathbb{P}(D(\boldsymbol{X}) \geq b) &= \mathbb{P}(D(\boldsymbol{X}) - \mathbb{E}D \geq b - \mu) \\ &\leq \exp\big(-\frac{(b-\mu)^2}{2n}\big) \quad \blacksquare\end{aligned}$$

LXXXIII. Martingales : Stochastic Processes, Ross, Problem 6.22

It is clear that the random variables $\{X_n\}_{n \geq 1}$ are non-negative. Now we compute

$$\begin{aligned}&\mathbb{E}(X_{n+1}|X_1, X_2, \ldots, X_n) \\ =\; & X_n(\alpha X_n + 1 - \alpha) + (1 - X_n)\alpha X_n \\ =\; & X_n\end{aligned}$$

Also, since $\{X_n\}$ s are non-negative, $\mathbb{E}|X_n| = \mathbb{E}X_n = \mathbb{E}X_0 = \frac{1}{2} < \infty$. Hence $\{X_n\}_{n \geq 0}$ is a non-negative Martingale Sequence. Hence using the corollary 6.4.7 of Martingale Convergence Theorem, we conclude that $\lim_{n \to \infty} X_n$ exists and is finite w.p. 1. $\blacksquare$

LXXXIV. Martingales : Stochastic Processes, Ross, Problem 6.26 [7]

Consider the sequence of random variables $\{Z_n\}_{n \geq 1}$, where $Z_n = \sum_{i=1}^n X_i$. Since $\mathbb{E}X_i = 0, \mathbb{E}|X_i| < \infty, \forall i$, we have $\mathbb{E}|S_n| \leq \sum_{i=1}^n \mathbb{E}|X_i| < \infty, \forall n \geq 1$ and $\mathbb{E}(S_n|S_1, S_2, \ldots, S_{n-1}) = S_{n-1}, \forall n \geq 1$, since $\{X_i\}$'s are independent. This proves that $\{S_n\}_{n \geq 1}$ constitutes a Martingale sequence.

Now consider the event $E = \{\limsup Z_n > -\infty\}$. Clearly, $E \subset \{X_n = 2^n - 1$ *infinitely often*$\}$. Hence $\mathbb{P}(E) \leq \mathbb{P}(\{X_n = 2^n - 1$ *infinitely often*$\}$. Now,

$$\sum_{i=1}^{\infty} \mathbb{P}(X_n = 2^n - 1) = \sum_{i=1}^{n} \frac{1}{2^n} < \infty$$

Hence, using the Borel-Cantelli Lemma, we conclude that

$$\mathbb{P}(E) \leq 0$$

i.e.,

$$\mathbb{P}(\lim_{n \to \infty} Z_n = -\infty) = 1 \quad \blacksquare$$

LXXXV. Analysis : Lower semicontinuous functions and their level sets

Lemma : A function $f : \mathbb{R}^n \to \mathbb{R}$ is lower semicontinuous iff its level sets are closed.

Proof: **If part** : Assume that the level sets of the functions are closed, i.e. for the set C_α defined by

$$C_\alpha = \{\boldsymbol{x} \in \mathbb{R}^n : f(\boldsymbol{x}) \leq \alpha\}$$

we have,

$$C_\alpha = \bar{C_\alpha}, \quad \forall \alpha \in \mathbb{R}$$

Now consider a sequence $\boldsymbol{x}_n \to \boldsymbol{x}$ and the associated sequence $\{f(\boldsymbol{x}_n)\}$. Let $\liminf f(\boldsymbol{x}_n) = \alpha$. Fix $\epsilon > 0$. Hence there exists a subsequence $\{n_k\}$ such that for k large enough

$$f(\boldsymbol{x}_{n_k}) \leq \liminf f(\boldsymbol{x}_n) + \epsilon$$

Hence by the closure property of the level sets, we conclude

$$f(\boldsymbol{x}) \leq \liminf f(\boldsymbol{x}_n) + \epsilon \tag{46}$$

Since (46) holds for any $\epsilon > 0$, by taking $\epsilon \searrow 0$, we conclude that

$$f(\boldsymbol{x}) \leq \liminf f(\boldsymbol{x}_n)$$

Only if part: Assume that for any $\boldsymbol{x}_n \to \boldsymbol{x}$, we have

$$f(\boldsymbol{x}) \leq \liminf f(\boldsymbol{x}_n)$$

Fix $\alpha \in \mathbb{R}$. Consider the level set

$$C_\alpha = \{\boldsymbol{x} \in \mathbb{R}^n : f(\boldsymbol{x}) \leq \alpha\}$$

Consider any sequence $\{\boldsymbol{x}_n\} \in C_\alpha$. Thus,

$$f(\boldsymbol{x}_n) \leq \alpha$$

Hence,

$$f(\boldsymbol{x}) \leq \liminf f(\boldsymbol{x}_n) \leq \alpha$$

Thus $\boldsymbol{x} \in C_\alpha$.

$\blacksquare$

LXXXVI. Random 20 questions : Cover and Thomas, Problem 5.45 [1]

(a) Object 2 will yield the same answer as object 1, if we query those sets in which both $1, 2$ are present or neither of them are present. Number of such sets is 2^{n-1}. Hence probability that they yield the same answer on a single query is $\frac{1}{2}$. Since the successive queries are independent, probability that object 2 yields the same answer as in object 1 for k queries is simply $\frac{1}{2^k}$.
(b) Let Z denote the random variable denoting the number of objects in $\{2, 3, \ldots, m\}$ having same answer to questions as does the correct object. Also, let $X_i, i = 2, \ldots, m$ denote the indicator variable denoting that object i has the same answer to all questions as object 1. Thus we have,

$$Z = \sum_{i=2}^{m} X_i$$

Hence,

$$\mathbb{E}Z = \sum_{i=2}^{m} \mathbb{E}X_i = \frac{m-1}{2^k} = \frac{2^n - 1}{2^k}$$

(c) If $k = n + \sqrt{n}$, the above expectation becomes

$$\mathbb{E}Z \leq \frac{1}{2^{\sqrt{n}}}$$

(d) Using Markov inequality, the probability of error is upper-bounded as

$$\mathbb{P}(Z \geq 1) \leq \frac{1}{2^{\sqrt{n}}} \searrow 0 \quad \blacksquare$$

LXXXVII. Conditional mutual information : Cover and Thomas, Problem 2.23

Given that

$$\begin{aligned} p(X_1, X_2, \ldots, X_n) &= \frac{1}{2^{n-1}}, \quad \text{if} \sum_{i=1}^{n} X_i \equiv 0 \mod (2) \\ &= 0 \quad \text{o.w.} \end{aligned}$$

Hence, it follows that

$$p(X_1, X_2, \ldots, X_{n-1}) = \frac{1}{2^{n-1}} = \prod_{i=1}^{n-1} p(X_i) \tag{47}$$

where $p(X_i = 0) = p(X_i = 1) = \frac{1}{2}$. Hence the random variables $(X_1, X_2, \ldots, X_{n-1})$ are i.i.d.
Thus, $I(X_1; X_2) = 0$, $I(X_2; X_3|X_1) = 0$,, $\ldots, I(X_{n-2}; X_{n-1}|X_1, X_2, \ldots, X_{n-3}) = 0$. We compute the last conditional mutual information as follows :

$$\begin{aligned} & I(X_{n-1}; X_n|X_1, X_2, \ldots, X_{n-2}) \\ =\ & H(X_{n-1}|X_1, X_2, \ldots, X_{n-2}) - H(X_n|X_1, X_2, \ldots, X_{n-1}) \\ =\ & H(X_{n-1}) - 0 \\ =\ & 1 \quad \blacksquare \end{aligned}$$

LXXXVIII. Drawing with and without replacement : Cover and Thomas, Problem 2.8 [1]

Let the random variables X_i and Y_i denote the outcome (color of the balls) of the i^{th} draw when the balls are drawn with and without replacements respectively.
Fact: X_i's are iid and $Y_i \sim X_1$.
Now we have

$$\begin{aligned} H(Y_1, Y_2, \ldots, Y_k) &\leq \sum_{i=1}^{k} H(Y_i) \qquad (48) \\ &= \sum_{i=1}^{k} H(X_i) \qquad (49) \\ &= H(X_1, X_2, \ldots, X_k) \qquad (50) \end{aligned}$$

here Eqn. (49) follows from the fact that $Y_i \sim X_1 \sim X_i$ and Eqn. (50) follows from the fact that X_i's are iid. Hence drawing with replacement has higher entropy. $\blacksquare$

LXXXIX. Random questions : Cover and Thomas, Problem 2.41

(a) Given that $X \perp Q$ and $A(x, q)$ is deterministic. Hence

$$\begin{aligned} I(X; Q, A) &= I(X; Q) + I(X; A|Q) \\ &= I(X; A|Q) \qquad (51) \\ &= I(A; X|Q) \\ &= H(A|Q) - H(A|X, Q) \\ &= H(A|Q) \qquad (52) \end{aligned}$$

where in Eqn. (51), we have used the fact $X \perp Q$ and in Eqn. (52), we have used the fact that A is a deterministic function of X, Q $\blacksquare$.
(b) We have,

$$\begin{aligned} & I(X; Q_1, A_1, Q_2, A_2) \qquad (53) \\ =\ & I(X; Q_1, Q_2) + I(X; A_1, A_2|Q_1, Q_2) \\ =\ & I(A_1, A_2; X|Q_1, Q_2) \qquad (54) \\ =\ & H(A_1, A_2|Q_1, Q_2) - H(A_1, A_2|X, Q_1, Q_2) \\ =\ & H(A_1, A_2|Q_1, Q_2) \qquad (55) \\ \leq\ & H(A_1|Q_1, Q_2) + H(A_2|Q_1, Q_2) \qquad (56) \\ \leq\ & H(A_1|Q_1) + H(A_2|Q_2) \qquad (57) \\ =\ & 2H(A_1|Q_1) = 2I(X; Q_1, A_1) \qquad (58) \end{aligned}$$

where Eqn. (54) follows from $X \perp (Q_1, Q_2)$, Eqn. (55) follows from deterministic properties of the answers, Eqn. (57) follows from the fact conditioning reduces entropy and Eqn. (58) follows from the fact that $Q_1 \sim Q_2$ $\blacksquare$.

XC. AEP : Cover and Thomas, Problem 3.4 [1]

(a) By WLLN, we have $\mathbb{P}(X^n \in A^n) \to 1$.
(b) Fix any $\delta > 0$. For n large enough, we have

$$\begin{aligned} \mathbb{P}(X^n \in A^n) &\geq 1 - \delta/2 \qquad (59) \\ \mathbb{P}(X^n \in B^n) &\geq 1 - \delta/2 \quad \text{(WLLN)} \qquad (60) \end{aligned}$$

By union bound, we have

$$\mathbb{P}(X^n \in A^n \cap B^n) \geq 1 - \delta \tag{61}$$

which affirmatively shows the result.
(c) We have,

$$|A^n \cup B^n| \leq |A^n| \leq 2^{n(H+\epsilon)} \tag{62}$$

(d) From part (b), for n sufficiently large, we have

$$\mathbb{P}(X^n \in A^n \cap B^n) \geq \frac{1}{2} \tag{63}$$

Since all elements $x^n \in A_n$ has $p(x^n) \leq 2^{-n(H-\epsilon)}$, we have

$$|A^n \cap B^n| 2^{-n(H-\epsilon)} \geq \frac{1}{2} \tag{64}$$

From which the result follows.

XCI. AEP and source coding: Cover and Thomas, Problem 3.7 [1]

(a) Number of sequence N of 100 bits with three or fewer ones is simply

$$N = \binom{100}{0} + \binom{100}{1} + \binom{100}{2} + \binom{100}{3} = 166751$$

Since all codewords are of the same length, minimum length binary codeword required for providing a distinct codeword for each of the N sequence is simply

$$L^* = \lceil \log N \rceil = 18 \tag{65}$$

(b) Probability of observing a source sequence for which no codeword has been assigned is simply

$$P_{\text{bad}} = 1 - \sum_{i=0}^{3} \binom{100}{i} p(1)^i p(0)^{100-i} \approx 0.001673 \tag{66}$$

(c) Let the r.v. L denote number of 1's in a sequence, we have $\mu = \mathbb{E}L = 100 \times 0.005 = 0.5$ and $\text{Var}(L) = 100 \times 0.005 \times 0.995 = 0.4975$. The Using Chebysev's inequality, we have

$$\begin{aligned} P_{\text{bad}} = \mathbb{P}(L \geq 4) = \mathbb{P}(L - \mu \geq 3.5) &= \mathbb{P}(|L - \mu| \geq 3.5) \\ &\leq \frac{0.4975}{3.5^2} \approx 0.04 \end{aligned}$$

XCII. Entropy of a random tree: Cover and Thomas, Problem 4.5

Let T_n denote a random tree with n terminal node and $N_1(T_n)$ denote the number of terminal on the right half of the tree. Then we have,

$$\begin{aligned} H_n &= H(T_n) && (67) \\ &= H(N_1, T_n) \quad (\text{since } H(N_1|T_n) = 0) && (68) \\ &= H(N_1) + H(T_n|N_1) && (69) \\ &= \log(n-1) + \frac{1}{n-1} \sum_{k=1}^{n-1} (H(T_k) + H(T_{n-k})) && (70) \\ &= \log(n-1) + \frac{2}{n-1} \sum_{k=1}^{n-1} H_k && (71) \end{aligned}$$

where Eqn. (70) follows from the definition of conditional entropy and the fact that given number of terminal points, the left tree is independent of the right tree.

From Eqn. (71) we conclude that
(f)

$$(n-1)H_n = (n-1)\log(n-1) + 2\sum_{k=1}^{n-1} H_k \tag{72}$$

Changing n to $n-1$, we have

$$(n-2)H_{n-1} = (n-2)\log(n-2) + 2\sum_{k=1}^{n-2} H_k \tag{73}$$

subtracting Eqn. (73) from (72), we have

$$\begin{aligned} &(n-1)H_n - (n-2)H_{n-1} = \\ &(n-1)\log(n-1) - (n-2)\log(n-2) + 2H_{n-1} \end{aligned}$$

i.e.,

$$(n-1)H_n = nH_{n-1} + (n-1)\log(n-1) - (n-2)\log(n-2) \tag{74}$$

Hence,

$$\frac{H_n}{n} = \frac{H_{n-1}}{n-1} + c_n \tag{75}$$

where $c_n \equiv \frac{\log(n-1)}{n} - \frac{(n-2)\log(n-2)}{n(n-1)}$. It can be simplified further as follows :

$$c_n = \left(\frac{\log(n-1)}{n} - \frac{\log(n-2)}{n-1} \right) + 2\frac{\log(n-2)}{n(n-1)} \tag{76}$$

In the sum $\sum_{n\geq 3} c_n$, the sum of the part within the brace telescopes to 0. The second part of the sum may be upper bounded as follows. We have $2\frac{\log(n-2)}{n(n-1)} \leq 2\frac{\log(n-2)}{(n-1)^2}$.
Thus,

$$\begin{aligned} 2\sum_{n\geq 3} \frac{\log(n-2)}{n(n-1)} &\leq 2\sum_{n\geq 3} \frac{\log(n-2)}{(n-1)^2} \leq 2\sum_{n\geq 2} \frac{\log(n)}{n^2} \\ &\leq 2\int_1^{\infty} \frac{\log(x)}{x^2} dx = 2 \end{aligned}$$

Hence $\sum c_n \leq 2$ and $\frac{H_n}{n}$ converges to constant value less than or equal to 2.

XCIII. Initial conditions: Cover and Thomas, Problem 4.9 [1]

Proof: First, we show that for a Markov Chain $\{X_n\}$, we have

$$p(X_0|X_{n-1}, X_n) = p(X_0|X_{n-1}) \tag{77}$$

To show this, we simply compute

$$\begin{aligned} p(X_0|X_{n-1}, X_n) &= \frac{p(X_0, X_{n-1}, X_n)}{p(X_{n-1}, X_n)} \\ &= \frac{p(X_0, X_{n-1})p(X_n|X_0, X_{n-1})}{p(X_{n-1})p(X_n|X_{n-1})} \\ &= p(X_0|X_{n-1}) \end{aligned}$$

where the last step follows from the Markovian assumption. The above implies :

$$\begin{aligned} H(X_0|X_{n-1}) &= H(X_0|X_n, X_{n-1}) && (78) \\ &\leq H(X_0|X_n) && (79) \end{aligned}$$

where the last step follows from the fact that conditioning reduces entropy. ■

XCIV. Chains and Antichains: Approximation Algorithms, Vazirani Problem 1.7 [5]

Proof: Let B be the size of a minimum antichain cover and m be the size of a maximum chain. First note that size of any antichain cover is *atleast* the size of any chain. This is simply because all elements in a chain must belong to different elements of any antichain cover. Thus, we have

$$m \leq B \tag{80}$$

Now for all $a \in A$, define $\phi(a)$ to be the size of the largest chain in which a is the smallest element. Consider the partition:

$$A_i = \{a \in A | \phi(a) = i\}, \;\; i = 1, 2, \ldots, m \tag{81}$$

Clealy, $\bigcup A_i = A$ and hence A_i covers A. Now we show that A_i's are all antichain. If not, let A_j is not an antichain and there exists $a, b \in A_j$ such that $a \leq b$. Hence concatanating a with the largest chain with b as the smallest element, we obtain a chain of length atleast $j+1$ with a as the smallest element. This violates our definition that $\phi(a) = j$. Hence,

$$B \leq m \tag{82}$$

Combining Eqns. (80) and (82), we get the desired result. ■

XCV. Gomory-Hu Cuts : Approximation Algorithms, Vazirani Problem 4.3 [5]

Proof: Consider a min-cut $\mathcal{C}$ between nodes u and v. This min-cut partitions the node in V into two sets : S_u and S_v. The node w must belong to any one of these sets. If $w \in S_v$, C is a valid $u - w$ cut. Thus we have

$$f(u, w) \leq f(u, v) \tag{83}$$

On the other hand, if $w \in S_u$, $\mathcal{C}$ is a valid $v - w$ cut, thus

$$f(v, w) \leq f(u, v) \tag{84}$$

This proves (3):

$$f(u, v) \geq \min\{f(u, w), f(w, v)\} \tag{85}$$

■

Proof: (4) follows easily by induction : Base case follows from (3). Let us assume that the result holds for $r = k - 1$. Thus we have

$$f(u, v) \geq \min\{f(u, w_1), f(w_1, w_2), \ldots, f(w_{k-1}, v)\} \tag{86}$$

Now, by (3)

$$f(w_{k-1}, v) \geq \min\{f(w_{k-1}, w_k), f(w_k, v)\} \tag{87}$$

The result follows by combining the above two equations. ■

Proof: 1. From 3 above, we have

$$f(u, w) \leq f(u, v) \leq f(u, w) \tag{88}$$

Hence,

$$f(u, v) = f(u, w) \tag{89}$$

■

XCVI. MAX $k-$CUT : Approximation Algorithms, Vazirani Problem 2.14 [5]

The greedy algorithm grows k sets $S_1, S_2, \ldots, S_k$ from scratch as follows : At any iteration we peak a vertex v which is not present in any of the sets $S_1, S_2, \ldots, S_k$. Then greedily put v in that set S_i which contains least number of neighbors of v (among the sets S_i).
Note that each edge is examined exactly once during the execution of the algorithm. If during the iteration when the node v is considered, there are m_v edges incident on the sets $\cup_i S_i$, then there must exist a set with at most $\frac{m_v}{k}$ neighbors. Thus inclusion of v contributes a cut of at least $m_v(1 - \frac{1}{k})$. Summing over all the vertices, we obtain

$$\text{CUT} \geq \sum_v m_v(1 - \frac{1}{k}) \geq |E|(1 - \frac{1}{k}) \geq \text{OPT}(1 - \frac{1}{k}) \tag{90}$$

XCVII. MAX-SAT : Approximation Algorithms, Vazirani Problem 16.2

Proof: Note that, if a clause is not satisfied in the truth assignment τ, it is definitely satisfied in the truth assignment τ'. Hence total number of satisfied clauses in τ and τ' is at least $\mathcal{C}$. Since we take the better of the assignments τ and τ', at least half of the clauses are satisfied. ■

XCVIII. Exercise 5.2 : Large Deviations For Performance Analysis [8]

Proof: We start with the following result:

$$\lim_{n \to \infty} \frac{\exp(-n) n^n}{n!} = 0 \tag{91}$$

The result follows simply from using the Starling's approximation for $n!$. Hence the above result implies that for any given $\delta > 0$, there exists n_0 such that for all $n \geq n_0$, we have

$$\frac{\exp(-n) n^n}{n!} < \delta \tag{92}$$

Now WOLOG, assume that $x_i(t)$'s are Poisson random-variables of mean $\lambda = 1$. Thus

$$\frac{1}{n} \sum_{i=1}^{n} x_i(t) \overset{(d)}{=} N(t) \tag{93}$$

Where $N(t)$ is a Poisson random variable of mean 1.
Fix any $\eta > 0$. Take $\delta = \frac{\eta}{2\epsilon}$ and any $n \geq n_0(\delta)$. Now the required probability is simply

$$\begin{aligned}
\mathbb{P}\bigg(\sup_{t \geq 0} |N(t) - t| &\geq \epsilon \bigg) \geq \mathbb{P}\bigg(|N(n) - n| \geq \epsilon \bigg) \\
&= 1 - \mathbb{P}\bigg(|N(n) - n| \leq \epsilon \bigg) \\
&\geq 1 - 2\epsilon \frac{\exp(-n) n^n}{n!} \\
&\geq 1 - 2\epsilon\delta \\
&= 1 - \eta
\end{aligned}$$

Hence, for any $\eta > 0$, we have

$$1 \geq \mathbb{P}\bigg(\sup_{t \geq 0} |N(t) - t| \;\; \geq \;\; \epsilon \bigg) \geq 1 - \eta$$

Hence for any $\epsilon > 0$:

$$\mathbb{P}\left(\sup_{t\geq 0} |\frac{1}{n}\sum_{i=1}^{n} x_i(t) - t| \quad \geq \quad \epsilon \right) = 1$$

■

XCIX. Bounds on Jensen-Shannon Divergence

Problem: For two discrete probability measures $(\boldsymbol{P}, \boldsymbol{Q})$, the Jensen-Shannon divergence JSD$(\boldsymbol{P}||\boldsymbol{Q})$ is defined as follows

$$\text{JSD}(\boldsymbol{P}||\boldsymbol{Q}) = D(\boldsymbol{P}||\boldsymbol{M}) + D(\boldsymbol{Q}||\boldsymbol{M})$$

where $\boldsymbol{M} = \frac{1}{2}\big(\boldsymbol{P} + \boldsymbol{Q}\big)$. Prove that

$$0 \leq \text{JSD}(\boldsymbol{P}||\boldsymbol{Q}) \leq 2$$

Proof: The lower-bound follows directly from the non-negativity of the KL-divergence $D(\cdot||\cdot)$. To prove the upper bound, we use the inequality $\ln(x) \leq x - 1$ to obtain

$$\begin{aligned}
& \text{JSD}(\boldsymbol{P}||\boldsymbol{Q}) \\
= \; & \sum_i p_i \ln \frac{p_i}{\frac{1}{2}(p_i+q_i)} + \sum_i q_i \ln \frac{q_i}{\frac{1}{2}(p_i+q_i)} \\
\leq \; & \sum_i p_i \big(\frac{p_i}{\frac{1}{2}(p_i+q_i)} - 1\big) + \sum_i q_i \big(\frac{q_i}{\frac{1}{2}(p_i+q_i)} - 1\big) \\
= \; & \sum_i p_i \frac{p_i - q_i}{p_i + q_i} - \sum_i q_i \frac{p_i - q_i}{p_i + q_i} \\
= \; & \sum_i \frac{(p_i - q_i)^2}{p_i + q_i} \\
= \; & \sum_i \frac{(p_i + q_i)^2 - 4p_iq_i}{p_i + q_i} \\
= \; & \sum_i \big(p_i + q_i\big) - 4\sum_i \frac{p_iq_i}{p_i + q_i} \\
\leq \; & 2
\end{aligned}$$

Where we use the fact that

$$\sum_i p_i = \sum_i q_i = 1 \tag{94}$$

■

C. Exercise 8.1 : Approximation Algorithms [5]

Proof: Consider the following (parameterized) problem where we have a knapsack of size 1 and have two items a, b such that size$(a) = 1$, profit$(a) = 1$, and size$(b) = \epsilon$, profit$(b) = 2\epsilon$, where $0 < \epsilon < 1$. Then the optimal allocation is to pick item a and have a total profit of 1. On the other hand, the greedy algorithm picks the item b (since it has a higher profit to size ratio) and end up having a profit of only 2ϵ, which can be made arbitrarily small by taking small enough ϵ. ■

CI. Exercise 8.2 : Approximation Algorithms [5]

Proof: Let the objects $\{a_1, a_2, \ldots, a_n\}$ be sorted according to the decreasing order of profit to size ratios. Also let p_i and s_i denote the profit and the size associated with the i^{th} object respectively. Then it is clear that for a knapsack of size $\sum_{i=1}^{l} s_l$, the optimal profit is exactly $\sum_{i=1}^{l} p_l$. Let

$$\sum_{i=1}^{k-1} s_i \leq B < \sum_{i=1}^{k} s_i \tag{95}$$

Then it is easy to see that the "extra" profit that we might get by squeezing in some objects in the left-over space (if there is any) $B - \sum_{i=1}^{k-1} s_i$ is at most $\frac{p_k}{s_k}(B - \sum_{i=1}^{k-1} s_i)$. Hence,

$$\begin{aligned}
\text{OPT} \quad & \leq \quad \sum_{i=1}^{k-1} p_i + \frac{p_k}{s_k}(B - \sum_{i=1}^{k-1} s_i) \leq \sum_{i=1}^{k-1} p_i + p_k \quad (96) \\
& \leq \quad 2\max\{\sum_{i=1}^{k-1} p_i, p_k\} \quad (97)
\end{aligned}$$

This shows that the given algorithm achieves an approximation ratio of $\frac{1}{2}$. ■

CII. Geometric Properties of the Convex Hull of Independent Sets

Problem: Consider an undirected graph $\mathcal{G}(V, E)$ on $|V| = n$ nodes and let the set of $\{0, 1\}$ vectors $\{v_i \in \{0, 1\}^n\}$ denote the incidence vectors of corresponding to the independent sets of the graph $\mathcal{G}$. Let the set $\Lambda \subset \mathbb{R}_+^n$ denote the convex-hull of the vectors $\{v_i\}$. For any given non-negative vector $\lambda \in \mathbb{R}_+^n$, define its load-factor $\rho(\lambda)$ as follows

$$\rho(\lambda) = \min\{\alpha > 0 : \lambda \in \alpha\Lambda\}$$

Prove the following properties of $\rho(\lambda)$:

$$\begin{array}{ll}
(a) & \lambda \leq \lambda' \Longrightarrow \rho(\lambda) \leq \rho(\lambda') \\
(b) & \rho(\lambda + \lambda') \leq \rho(\lambda) + \rho(\lambda') \\
(c) & \rho(c\lambda) = c\rho(\lambda) \\
(d) & <\lambda, \mathbf{1}> = \nu \Longrightarrow \nu/n \leq \rho(\lambda) \leq \nu \\
(e) & 1 \leq \rho(\mathbf{1}) \leq n
\end{array}$$

Proof: We will use two properties of the set Λ, that it is down-ward closed and that it is convex.

(a) By the given assumption, there exists $\mu \in \Lambda$ such that, $\lambda' \leq \rho(\lambda')\mu$. Hence, $\lambda \leq \lambda' \leq \rho(\lambda')\mu$. By down-ward closeness, we conclude that $\lambda \in \rho(\lambda')\Lambda$. Hence by the definition of load-factor, we have $\rho(\lambda) \leq \rho(\lambda')$.

(b) By the given assumption, there exists two vectors $\mu_1, \mu_2 \in \Lambda$, such that

$$\mu_1, \mu_2 \in \Lambda, \lambda = \rho(\lambda)\mu_1, \lambda' = \rho(\lambda')\mu_2$$

Because of convexity,

$$\frac{\lambda + \lambda'}{\rho(\lambda) + \rho(\lambda')} = \frac{\rho(\lambda)\mu_1 + \rho(\lambda')\mu_2}{\rho(\lambda) + \rho(\lambda')} \in \Lambda$$

Hence,

$$\lambda + \lambda' \in (\rho(\lambda) + \rho(\lambda'))\Lambda$$

Thus, by definition

$$\rho(\lambda+\lambda') \leq \rho(\lambda)+\rho(\lambda')$$

(c) Trivial from the definition.
(d) Let $< \lambda, 1 >= \nu$. Then from the definition, it follows that, there exists a $\mu \in \Lambda$ such that

$$\lambda \leq \rho(\lambda)\mu$$

Hence,

$$\nu =< \lambda, \mathbf{1} >\leq \rho(\lambda) < \mu, \mathbf{1} >\leq \rho(\lambda)n$$

Where the last inequality follows from the fact that each component of μ is bounded by unity. Hence, we have

$$\rho(\lambda) \geq \frac{\nu}{n}$$

To prove the other inequality, note that the unit vectors $e_i, i = 1, 2, \ldots, n$ belong to Λ. Hence, by the convexity Λ, we have

$$\frac{\lambda}{\nu} \in \Lambda$$

i.e., $\lambda \in \nu\Lambda$, which implies that $\rho(\lambda) \leq \nu$.
(e) Follows from part (d) with $\lambda = 1$. ■

CIII. Properties of Euclidean-Distance Matrices

Problem (a) Let A and B be two matrices of dimension $m \times n$. Show that

$$\text{rank}(A+B) \leq \text{rank}(A) + \text{rank}(B)$$

Proof: Let the columns of the matrices A and B be given by $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$ respectively. Then the columns of the matrix $(A+B)$ is given by $\{(a_i+b_i)\}_{i=1}^n$. By the rank assumption, we conclude that there are rank(A) linearly independent vectors $e_j^A, j = 1, 2, \ldots, \text{rank}(A)$ such that each column of A can be expressed as a linear combination of e_j^A's, i.e., for some scalars λ_{ij}, we have

$$a_i = \sum_{j=1}^{\text{rank}(A)} \lambda_{ij} e_j^A$$

Similarly, we have

$$b_i = \sum_{j=1}^{\text{rank}(B)} \mu_{ij} e_j^B$$

where the vectors e_j^B and the scalars μ_{ij} have been defined similarly.
From the above two expressions, it is clear that any linear combination of the columns of the matrix $(A+B)$ can be expressed as a linear combination of rank(A) + rank(B) number of vectors: $\{e_j^A\}_{j=1}^{\text{rank}(A)}$'s and $\{e_j^B\}_{j=1}^{\text{rank}(B)}$'s. Since rank of a matrix is also dimension of its column space, we have:

$$\text{rank}(A+B) \leq \text{rank}(A) + \text{rank}(B) \tag{98}$$

■

Problem (b) Let $\{\boldsymbol{x}_i\}_{i=1}^N$ be N points all belonging to $\mathbb{R}^d$, where $N \geq d$. Define the $N \times N$ **Euclidean-distance matrix** $\boldsymbol{E}(\boldsymbol{x})$, corresponding to the points $\boldsymbol{x}_i$'s as follows

$$\boldsymbol{E}(\boldsymbol{x})_{i,j} = ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2$$

Show that,

$$\text{rank}(\boldsymbol{E}(\boldsymbol{x})) \leq d+2 \tag{99}$$

Proof: We have,

$$||\boldsymbol{x}_i - \boldsymbol{x}_j||^2 = ||\boldsymbol{x}_i||^2 + ||\boldsymbol{x}_j||^2 - 2\boldsymbol{x}_i \cdot \boldsymbol{x}_j \tag{100}$$

Let $\boldsymbol{\alpha} = \bigg(||\boldsymbol{x}_1||^2, ||\boldsymbol{x}_2||^2, \ldots, ||\boldsymbol{x}_N||^2\bigg)'$. Hence,

$$E(\boldsymbol{x}) = \boldsymbol{e}_N \boldsymbol{\alpha}' + \boldsymbol{\alpha}\boldsymbol{e}_N' - 2\text{Gram}(\boldsymbol{x}) \tag{101}$$

Where $\boldsymbol{e}_N$ is the N dimensional vector with all entries unity and the Gram matrix is defined as follows

$$\text{Gram}(\boldsymbol{x})_{i,j} = \boldsymbol{x}_i \cdot \boldsymbol{x}_j \tag{102}$$

Claim: rank(Gram$(\boldsymbol{x})$) $\leq d$

Proof: To prove this claim, take any $(d+1)$ rows of the matrix Gram(x). Let the indices of the rows be $i_k, k = 1, 2, \ldots, d+1$. Since $\boldsymbol{x}_i \in \mathbb{R}^d$, there exists scalars $\lambda_{i_k}, k = 1, 2 \ldots, d+1$, not all zero, such that

$$\sum_{k=1}^{d+1} \lambda_{i_k} \boldsymbol{x}_{i_k} = \mathbf{0}$$

Hence, for any j, we have

$$\sum_{k=1}^{d+1} \lambda_{i_k} \boldsymbol{x}_{i_k} \cdot \boldsymbol{x}_j = 0$$

This shows that any $(d+1)$ rows of the Gram matrix is linearly dependent. Hence

$$\text{rank}(\text{Gram}(\boldsymbol{x})) \leq d$$

■

The final result follows by combining the result of part(a) with the above claim in Eqn. (101) and noting that the other two matrices are of rank at most 1. ■

CIV. Number of Edge-Disjoint Spanning Trees

Problem Consider a graph $\mathcal{G}(V, E)$ which is $2k$ edge-connected. Prove that $\mathcal{G}$ has at least k edge-disjoint spanning trees.

Proof: We will be using the theorem of Nash-Williams-Tutte, given as follows

Theorem 1 (Nash-Williams-Tutte): A graph $\mathcal{G}(V, E)$ contains k edge-disjoint spanning trees if and only if

$$E_{\mathcal{G}}(\mathcal{P}) \geq k(t-1)$$

for every partition $\mathcal{P} = \{V_1, V_2, \ldots, V_t\}$ of V into non-empty subsets, where $E_{\mathcal{G}}(\mathcal{P})$ denotes the number of edges between distinct classes of $\mathcal{P}$.
Now as in the previous theorem, consider any partition $\mathcal{P} = \{V_1, V_2, \ldots, V_t\}$ of V into non-empty subsets. It is clear that number of outgoing edges from any of the above partition is at least $2k$, otherwise , we could remove less than $2k$ edges to disconnect the graph $\mathcal{G}$. Hence, by hand-shake lemma, we have

$$\sum_{P \in \mathcal{P}} \deg(P) = 2E_{\mathcal{G}}(\mathcal{P}) \geq 2k(t-1)$$

The result now follows from the above theorem. ■

CV. Problem 11.7 Fisher Information and Relative Entropy [1]

Proof: We need to show that

$$\lim_{\theta'\to\theta}\frac{1}{(\theta-\theta')^2}D(p_\theta||p_{\theta'}) = \frac{1}{\ln 4}J(\theta) \tag{103}$$

To prove this result, we expand the LHS in θ' around the point θ in Taylor series (upto first order), by substituting $\theta' = \theta + h$, where $h \to 0$ to obtain:

$$\begin{aligned}
& \frac{1}{h^2}D(p_\theta||p_{\theta+h}) \\
= \quad & -\frac{1}{h^2}\mathbb{E}\log\frac{p_{\theta+h}(X)}{p_\theta(X)} \\
\approx \quad & -\frac{1}{h^2}\mathbb{E}\log\frac{p_\theta(X) + h\partial p_\theta(X)/\partial\theta}{p_\theta(X)} \\
= \quad & -\frac{1}{h^2\ln 2}\mathbb{E}\ln\left(1 + h\frac{\partial\log(p_\theta(X))}{\partial\theta}\right) \\
\approx \quad & -\frac{1}{h^2\ln 2}\mathbb{E}\Bigg(h\frac{\partial\log(p_\theta(X))}{\partial\theta} - \frac{h^2}{2}\big(\frac{\partial\log(p_\theta(X))}{\partial\theta}\big)^2\Bigg) \\
= \quad & \frac{1}{\ln 4}\mathbb{E}\big(\frac{\partial\log(p_\theta(X))}{\partial\theta}\big)^2 \\
= \quad & \frac{1}{\ln 4}J(\theta)
\end{aligned}$$

In the above, we have used the fact that $\mathbb{E}\frac{\partial\log p_\theta(X)}{\partial\theta} = 0$, which follows from interchanging the order of differentiation and integration and dropped the higher order terms in h appropriately. ■

CVI. Problem 4.35 [1]: Concavity of Second Law

Proof: We start with the RHS,

$$\begin{aligned}
& -I(X_1;X_{n-1}|X_0,X_n) \\
\overset{(a)}{=} \quad & H(X_1|X_0,X_{n-1},X_n) - H(X_1|X_0,X_n) \\
\overset{(b)}{=} \quad & H(X_0,X_1,X_{n-1},X_n) - H(X_0,X_{n-1},X_n) \\
& -H(X_0,X_1,X_n) + H(X_0,X_n) \\
\overset{(c)}{=} \quad & H(X_0) + H(X_1|X_0) + H(X_{n-1}|X_1) + H(X_n|X_{n-1}) \\
& -H(X_0) - H(X_{n-1}|X_0) - H(X_n|X_{n-1}) - H(X_0) \\
& -H(X_1|X_0) - H(X_n|X_1) + H(X_0) + H(X_n|X_0) \\
\overset{(d)}{=} \quad & H(X_n|X_0) - H(X_{n-1}|X_0) \\
& -(H(X_{n-1}|X_0) - H(X_{n-2}|X_0))
\end{aligned}$$

which equals the LHS. Here in (a) we have used the definition of conditional mutual information, in (b) we have used the chain rule, in (c) we have used the Markovian property of $\{X_k\}_{-\infty}^{\infty}$ and finally in (d) we have made use of stationarity of $\{X_k\}_{-\infty}^{\infty}$.

■

References

[1] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[2] R. Motwani and P. Raghavan, *Randomized algorithms*. Chapman & Hall/CRC, 2010.

[3] S. Dasgupta, C. H. Papadimitriou, and U. Vazirani, *Algorithms*. McGraw-Hill, Inc., 2006.

[4] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge university press, 2011.

[5] V. V. Vazirani, *Approximation algorithms*. Springer Science & Business Media, 2013.

[6] R. G. Gallager, *Information theory and reliable communication*. Springer, 1968, vol. 2.

[7] S. M. Ross *et al.*, *Stochastic processes*. John Wiley & Sons New York, 1996, vol. 2.

[8] A. Shwartz and A. Weiss, *Large deviations for performance analysis: queues, communication and computing*. CRC Press, 1995, vol. 5.