ASSIGNMENT 2

An assignment report
submitted to **Prof. Ganesh Ramakrishnan**
in the subject of Optimization in Machine Learning

by

Sandarbh Yadav
(22D0374)

INDIAN INSTITUTE OF TECHNOLOGY BOMBAY

2023

Q 1. Let there be a graph G with V vertices and E edges. Comment on the modularity of the following functions and give reasons.

(1) Let $I(V_1) =$ set of edges with at least one end point in $V_1, V_1 \in V(G)$. Then $|I|$ is?

Submodular. Suppose $A \subseteq B \subseteq V(\mathcal{G})$ and let $a \in V(\mathcal{G}) - B$. Then, we have

$$I(A \cup a) = I(A) \biguplus [I(a) - I(A)]$$

$$I(B \cup a) = I(B) \biguplus [I(a) - I(B)]$$

where $\biguplus$ denotes disjoint union operator. It is easy to see that

$$I(a) - I(A) \supseteq I(a) - I(B)$$

Thus, the gains are diminishing. Hence, $|I|$ is submodular.

(2) Let cut $(V_1) =$ set of all branches with only one endpoint in $V_1, V_1 \in V(G)$. Then $|cut|$ is?

Submodular. Suppose $A \subseteq B \subseteq V(\mathcal{G})$ and let $a \in V(\mathcal{G}) - B$. Then, we have

$$\text{cut}(A \cup a) = (\text{cut}(A) - \text{cut}(A) \cap \text{cut}(a)) \biguplus (\text{cut}(a) - \text{cut}(A))$$
$$\text{cut}(B \cup a) = (\text{cut}(B) - \text{cut}(B) \cap \text{cut}(a)) \biguplus (\text{cut}(a) - \text{cut}(B))$$

where $\biguplus$ denotes disjoint union operator. It is easy to see that

$$\text{cut}(a) - \text{cut}(A) \supseteq \text{cut}(a) - \text{cut}(B) \text{ and } \text{cut}(A) \cap \text{cut}(a) \subseteq \text{cut}(B) \cap \text{cut}(a)$$

Thus, the gains are diminishing. Hence, $|cut|$ is submodular.

(3) $|I(V_1)| + |\text{cut}(V_1)|$ is?

Submodular. We know that the non-negative linear combination of submodular functions is also submodular. Here, $|I(V_1)|$ and $|\text{cut}(V_1)|$ are submodular. Hence, $|I(V_1)| + |\text{cut}(V_1)|$ is also submodular.

(4) Let $B = (V_L, V_R, E)$. Let $E_L(X) =$ set of all vertices in V_R adjacent only to vertices in $X, X \in V_L$. E_R is defined similarly on subsets of V_R. Then $|E_L|, |E_R|$ are?

Supermodular. Suppose $A \subseteq B \subseteq V(\mathcal{G})$ and let $a \in V(\mathcal{G}) - B$. Now,

$$E_L(C \cup a) = E_L(C) \biguplus E_L^{Ca}, \ C \subseteq V(\mathcal{G}), \ a \in V(\mathcal{G}) - C$$

where E_L^{Ca} is the set of all vertices in V_R adjacent only to vertices in $C \cup a$ and adjacent to a. It is easy to see that

$$E_L^{Aa} \subseteq E_L^{Ba}$$

Thus, the gains are increasing. Hence, $|E_L|$ is supermodular. Similarly, $|E_R|$ is also supermodular.

Q 2. Let Ω be a universal set and $A, B \subseteq \Omega$. For a monotone submodular f, submodular mutual information between the sets A, B denoted by $I_f(A; B)$ is defined as $I_f(A; B) \triangleq f(A) + f(B) - f(A \cup B)$.

(1) What would the expression for of $I_f\left(A_1; A_2; \ldots; A_k\right)$?

Solution: The expression for $I_f\left(A_1; A_2; \ldots; A_k\right)$ is given below:

$$I_f\left(A_1; A_2; \ldots; A_k\right) = -\sum_{T \subseteq [k]} (-1)^{|T|} f\left(\cup_{i \in T} A_i\right)$$

It has been defined using the inclusion-exclusion principle.

(2) Show that $I_f(A; B) \geq 0$ and $I_f(A; B \mid C) \geq 0$.

Proof: First, we prove that $I_f(A; B) \geq 0$. Submodular mutual information is defined as

$$I_f(A; B) = f(A) + f(B) - f(A \cup B)$$

It is given that f is monotone submodular. So, we have

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$$

$$\Longrightarrow f(A) + f(B) - f(A \cup B) \geq f(A \cap B) \geq 0$$

Hence, $I_f(A; B) \geq 0$.

Now, we prove that $I_f(A; B \mid C) \geq 0$. Conditional submodular mutual information is defined as

$$\begin{aligned} I_f(A; B \mid C) &= f(A \mid C) + f(B \mid C) - f(A \cup B \mid C) \\ &= f(A \cup C) - f(C) + f(B \cup C) - f(C) - f(A \cup B \cup C) + f(C) \\ &= f(A \cup C) + f(B \cup C) - f(A \cup B \cup C) - f(C) \end{aligned}$$

It is given that f is monotone submodular. So, we have

$$\begin{aligned} f(A \cup C) + f(B \cup C) &\geq f(A \cup B \cup C) + f([A \cup C] \cap [B \cup C]) \\ &= f(A \cup B \cup C) + f([A \cap B] \cup C) \\ &\geq f(A \cup B \cup C) + f(C) \end{aligned}$$

The last inequality holds because f is monotone.

$$\Longrightarrow f(A \cup C) + f(B \cup C) - f(A \cup B \cup C) - f(C) \geq 0$$

Hence, $I_f(A; B \mid C) \geq 0$.

(3) Show that $\min(f(A), f(B)) \geq I_f(A; B) \geq f(A \cap B)$.

Proof: First, we prove the lower bound. Submodular mutual information is defined as

$$I_f(A; B) = f(A) + f(B) - f(A \cup B)$$

It is given that f is submodular. So, we have

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$$

$$\implies f(A) + f(B) - f(A \cup B) \geq f(A \cap B)$$

Hence, $I_f(A; B) \geq f(A \cap B)$.

Now, we prove the upper bound. Submodular mutual information can be written as

$$I_f(A; B) = f(A) - f(A \mid B) = f(B) - f(B \mid A)$$

It is given that f is monotone submodular. So, we have

$$f(A \mid B) \geq 0 \text{ and } f(B \mid A) \geq 0 \ \forall A, B \subseteq \Omega$$

$$\implies I_f(A; B) \leq f(A) \text{ and } I_f(A; B) \leq f(B)$$

$$\implies I_f(A; B) \leq \min(f(A), f(B))$$

Hence, $\min(f(A), f(B)) \geq I_f(A; B)$.

(4) Show that $\min(f(A \mid C), f(B \mid C)) \geq I_f(A; B \mid C) \geq f(A \cap B \mid C)$.

Proof: First, we prove the lower bound. Conditional submodular mutual information is defined as

$$I_f(A; B \mid C) = f(A \mid C) + f(B \mid C) - f(A \cup B \mid C)$$

It is given that f is monotone submodular. $I_f(A; B \mid C)$ can be written as $I_g(A; B)$ where $g(A) = f(A \mid C)$ is also monotone submodular. So, we have

$$f(A \mid C) + f(B \mid C) \geq f(A \cup B \mid C) + f(A \cap B \mid C)$$

$$\implies f(A \mid C) + f(B \mid C) - f(A \cup B \mid C) \geq f(A \cap B \mid C)$$

Hence, $I_f(A; B \mid C) \geq f(A \cap B \mid C)$.

Now, we prove the upper bound. Conditional submodular mutual information can be written as

$$I_f(A; B \mid C) = f(A \mid C) - f(A \mid B; C) = f(B \mid C) - f(B \mid A; C)$$

It is given that f is monotone submodular. $I_f(A; B \mid C)$ can be written as $I_g(A; B)$ where $g(A) = f(A \mid C)$ is also monotone submodular. So, we have

$$f(A \mid B; C) \geq 0 \text{ and } f(B \mid A; C) \geq 0 \ \forall A, B, C \subseteq \Omega$$

$$\implies I_f(A; B \mid C) \leq f(A \mid C) \text{ and } I_f(A; B \mid C) \leq f(B \mid C)$$

$$\implies I_f(A; B \mid C) \leq \min(f(A \mid C), f(B \mid C))$$

Hence, $\min(f(A \mid C), f(B \mid C)) \geq I_f(A; B \mid C)$.

(5) Show that $I_f(A; B)$ can be lower bounded by $f(A) - \sum_{j \in A \setminus B} f(j \mid B) \leq I_f(A; B)$ and upper bounded by $I_f(A; B) \leq f(A) - \sum_{j \in A \setminus B} f(j \mid \Omega \setminus j) \leq f(A)$. Are upper/lower bounds submodular?

Proof: First, we prove the lower bound $f(A) - \sum_{j \in A \setminus B} f(j \mid B) \leq I_f(A; B)$. This can be written as

$$f(A) - I_f(A; B) \leq \sum_{j \in A \setminus B} f(j \mid B)$$

$$\implies f(A \mid B) \leq \sum_{j \in A \backslash B} f(j \mid B)$$

Suppose the set $A \backslash B$ contain k elements: $\{a_1, \ldots, a_k\}$. Now, we construct a chain of sets $X_0, X_1, X_2, \ldots, X_k$ such that $X_0 = B$ and $X_i = X_{i-1} \cup \{a_i\}$. Thus, $X_k = A \cup B$. We know that

$$\begin{aligned} f(A \mid B) &= f(A \cup B) - f(B) \\ &= f\left(X_k\right) - f\left(X_0\right) \\ &= \sum_{i=1}^{k} f\left(X_i\right) - f\left(X_{i-1}\right) \\ &= \sum_{i=1}^{k} f\left(X_{i-1} \cup \{a_i\}\right) - f\left(X_{i-1}\right) \\ &= \sum_{i=1}^{k} f\left(\{a_i\} \mid X_{i-1}\right) \end{aligned}$$

Also, $B \subseteq X_i \; \forall i = 0, 1, \ldots, k$. It is given that f is submodular. So, we have

$$f\left(\{a_i\} \mid X_{i-1}\right) \leq f\left(\{a_i\} \mid B\right) \;\; \forall i = 0, 1, \ldots, k$$

$$\implies f(A \mid B) = \sum_{i=1}^{k} f\left(\{a_i\} \mid X_{i-1}\right) \leq \sum_{i=1}^{k} f\left(\{a_i\} \mid B\right) = \sum_{j \in A \backslash B} f(j \mid B)$$

Hence, $f(A) - \sum_{j \in A \backslash B} f(j \mid B) \leq I_f(A; B)$.

Now, we prove the upper bound $I_f(A; B) \leq f(A) - \sum_{j \in A \backslash B} f(j \mid \Omega \backslash j)$. This can be written as

$$\sum_{j \in A \backslash B} f(j \mid \Omega \backslash j) \leq f(A) - I_f(A; B)$$

$$\implies \sum_{j \in A \backslash B} f(j \mid \Omega \backslash j) \leq f(A \mid B)$$

Suppose the set $A \backslash B$ contain k elements: $\{a_1, \ldots, a_k\}$. Now, we construct a chain of sets $X_0, X_1, X_2, \ldots, X_k$ such that $X_0 = B$ and $X_i = X_{i-1} \cup \{a_i\}$. Thus, $X_{i-1} \subseteq \Omega \backslash \{a_i\}$ because $\{a_i\}$ does not lie in both the sets. It is given that f is submodular. So, we have

$$f\left(\{a_i\} \mid X_{i-1}\right) \geq f\left(\{a_i\} \mid \Omega \backslash \{a_i\}\right) \forall i = 0, 1, \ldots, k$$

$$\implies f(A \mid B) = \sum_{i=1}^{k} f\left(\{a_i\} \mid X_{i-1}\right) \geq \sum_{i=1}^{k} f\left(\{a_i\} \mid \Omega \backslash \; \{a_i\}\right) = \sum_{j \in A \backslash B} f(j \mid \Omega \backslash j)$$

Hence, $I_f(A; B) \leq f(A) - \sum_{j \in A \backslash B} f(j \mid \Omega \backslash j)$. Also, $f(A) - \sum_{j \in A \backslash B} f(j \mid \Omega \backslash j) \leq f(A)$. So, $I_f(A; B) \leq f(A) - \sum_{j \in A \backslash B} f(j \mid \Omega \backslash j) \leq f(A)$.

Yes, the upper and lower bounds are submodular.

Q 3. We have a regression function for the variable x, given by $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$. There are n instances. The coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by solving the given optimization problem -:

$$\hat{\beta}_0, \hat{\beta}_1 = \text{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{n} w_i(x) \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

where the weight factors $w_i(x)$ do not depend on either of the coefficients.

(1) Show that the above function can be written in the form -:

$$(\boldsymbol{y} - \boldsymbol{Ba})^\top g(x)(\boldsymbol{y} - \boldsymbol{Ba})$$

where $\boldsymbol{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^\top, \boldsymbol{a} = [\beta_0 \ \beta_1]^\top, \boldsymbol{B} = [1 \ x_1; 1 \ x_2; \cdots 1 \ x_n]$, and $g(x)$ is a diagonal matrix whose diagonal entries are the weights $w_i(x)$.

Solution: $\boldsymbol{y} - \boldsymbol{Ba}$ can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$= \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \beta_0 + B_1 x_1 \\ \beta_0 + B_1 x_2 \\ \vdots \\ \beta_0 + B_1 x_n \end{bmatrix}$$

$$= \begin{bmatrix} y_1 - \beta_0 - \beta_1 x_1 \\ y_2 - \beta_0 - \beta_1 x_2 \\ \vdots \\ y_n - \beta_0 - \beta_1 x_n \end{bmatrix}$$

Thus, $(\boldsymbol{y} - \boldsymbol{Ba})^\top g(x)(\boldsymbol{y} - \boldsymbol{Ba})$ can be written as

$$\begin{bmatrix} y_1 - \beta_0 - \beta_1 x_1 \\ y_2 - \beta_0 - \beta_1 x_2 \\ \vdots \\ y_n - \beta_0 - \beta_1 x_n \end{bmatrix}^\top \begin{bmatrix} w_1(x) & 0 & \cdots & 0 \\ 0 & w_2(x) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & w_n(x) \end{bmatrix} \begin{bmatrix} y_1 - \beta_0 - \beta_1 x_1 \\ y_2 - \beta_0 - \beta_1 x_2 \\ \vdots \\ y_n - \beta_0 - \beta_1 x_n \end{bmatrix}$$

On multiplying, we get

$$w_1(x) \left(y_1 - \beta_0 - \beta_1 x_1\right)^2 + w_2(x) \left(y_2 - \beta_0 - \beta_1 x_2\right)^2 + \cdots + w_n(x) \left(y_n - \beta_0 - \beta_1 x_n\right)^2$$

$$= \sum_{i=1}^{n} w_i(x) \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

Hence, the given optimization objective function can be written in the form $(\boldsymbol{y} - \boldsymbol{Ba})^\top g(x)(\boldsymbol{y} - \boldsymbol{Ba})$.

(2) Using the formulation above, show that $\hat{f}(x)$ is a linear combination of $\boldsymbol{y}$. In other words, $\hat{f}(x) = \sum_{i=1}^{n} h_i(x)y_i$, where $h_i(x)$ is some function over x. Clearly mention what $h_i(x)$ is.

Solution: We solve the given optimization problem by calculating the gradient of the objective function with respect to $\boldsymbol{a}$.

$$\nabla_{\boldsymbol{a}}(\boldsymbol{y} - \boldsymbol{Ba})^\top g(x)(\boldsymbol{y} - \boldsymbol{Ba}) = -2\boldsymbol{B}^\top g(x)(\boldsymbol{y} - \boldsymbol{Ba})$$

Now, we equate the gradient to zero and solve for $\boldsymbol{a}$.

$$\begin{aligned} -2\boldsymbol{B}^\top g(x)(\boldsymbol{y} - \boldsymbol{Ba}) &= 0 \\ \implies \boldsymbol{B}^\top g(x)(\boldsymbol{y} - \boldsymbol{Ba}) &= 0 \\ \implies \boldsymbol{B}^\top g(x)\boldsymbol{y} &= \boldsymbol{B}^\top g(x)\boldsymbol{Ba} \\ \implies \boldsymbol{a} &= \left(\boldsymbol{B}^\top g(x)\boldsymbol{B}\right)^{-1}\boldsymbol{B}^\top g(x)\boldsymbol{y} \end{aligned}$$

Now, let $\boldsymbol{b} = [1,\ x]$. Hence, we can write $\hat{f}(x)$ in terms of $\boldsymbol{a}$ as

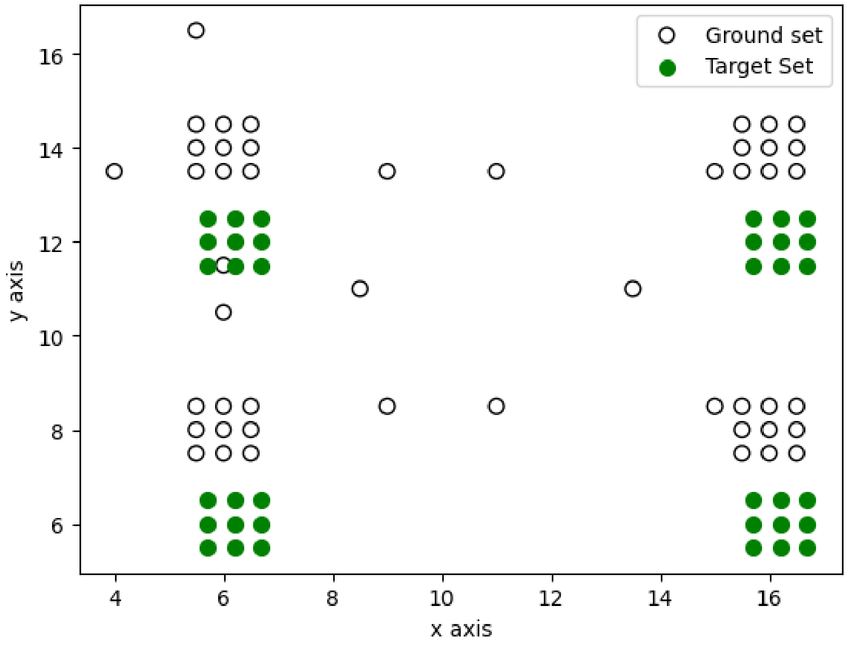
$$\hat{f}(x) = \boldsymbol{ba} = \boldsymbol{b}\left(\boldsymbol{B}^\top g(x)\boldsymbol{B}\right)^{-1}\boldsymbol{B}^\top g(x)\boldsymbol{y}$$

Clearly, $\hat{f}(x)$ is a linear combination of $\mathbf{y}$. $\hat{f}(x) = \sum_{i=1}^{n} h_i(x)y_i$, where $h_i(x)$ corresponds to the i-th column of the $2 \times n$ matrix $\boldsymbol{b}\left(\boldsymbol{B}^\top g(x)\boldsymbol{B}\right)^{-1}\boldsymbol{B}^\top g(x)$.

Q 4. In this problem, we will explore the various types of submodular functions by plotting the data against the optimal set obtained by greedy inference. You are encouraged to use libraries such as submodlib for submodular functions and matplotlib for plotting. You may access the text files here.

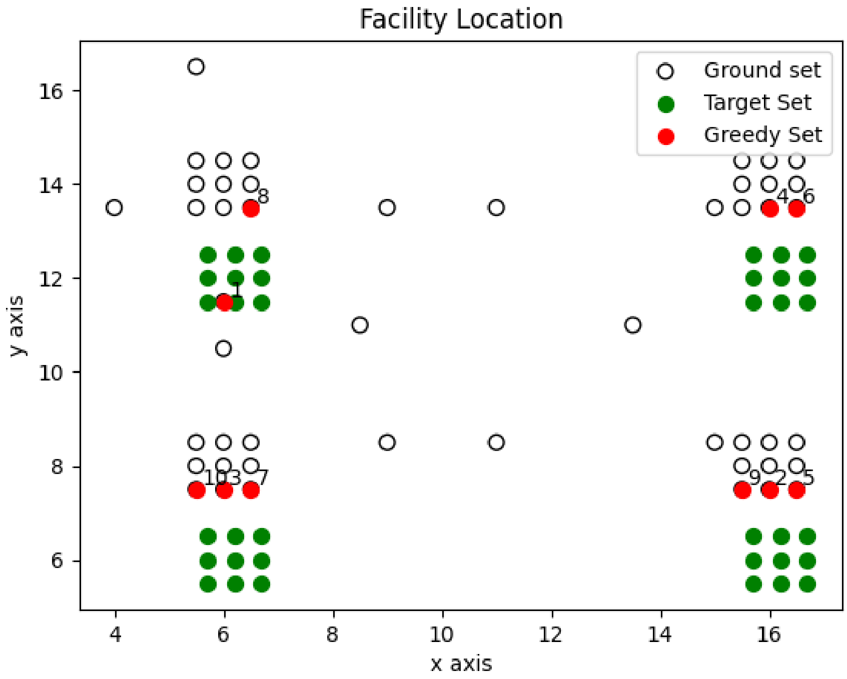
(1) In this part of the problem, you are provided with two sets of files - 'gset_1.txt' and 'rep.txt'. The former contains the ground set of points (whose subset we want) and the latter contains the target set of points (whose representation we want). Each line of a file represents an (x, y)-coordinate.

(a) First, plot the data from both sets of files using differing colors. Legends and axis should be present and labelled properly.

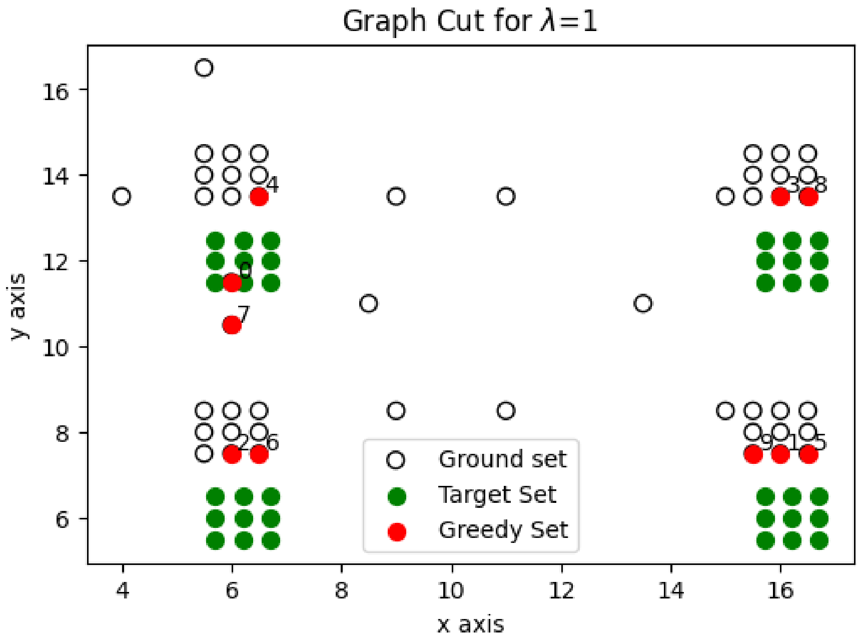


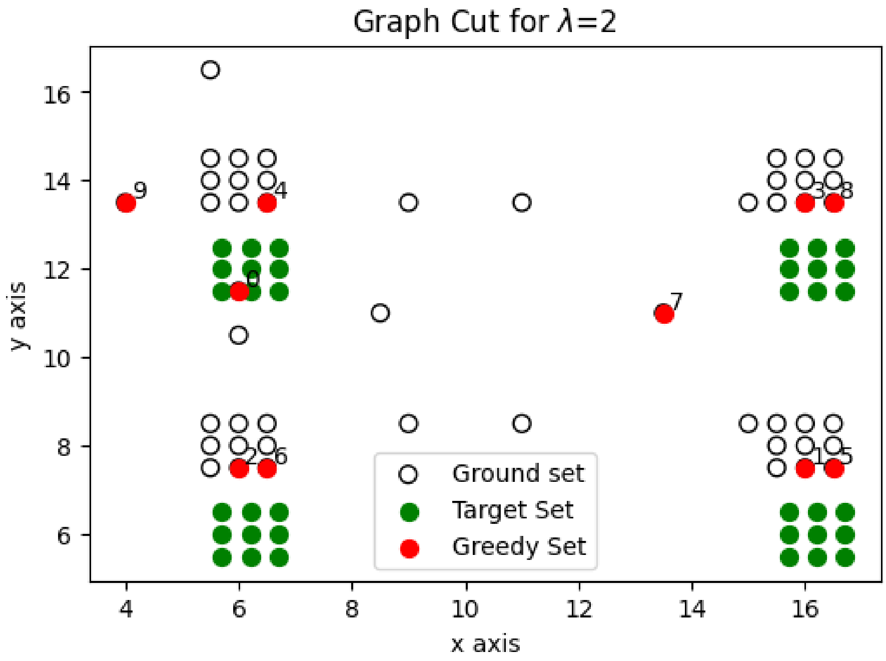
Then, plot the observations using the functions given below for optimal set budget of size 10 . Use the naive greedy algorithm and make sure to plot points from the optimal set obtained in a different color. Report the order in which the points were picked up by your algorithm -:

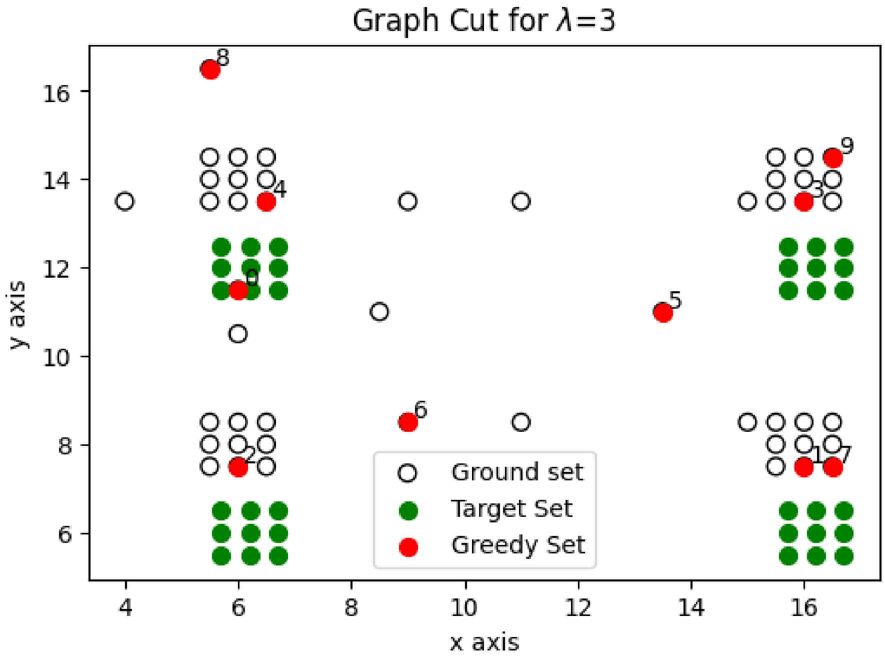
(b) Facility Location



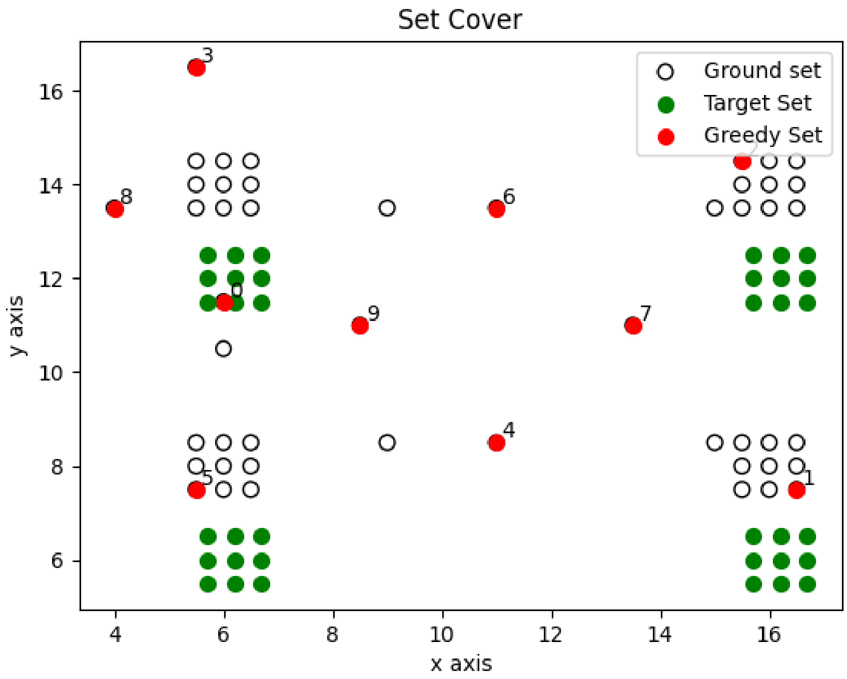
(c) Graph Cut with varying parameter λ



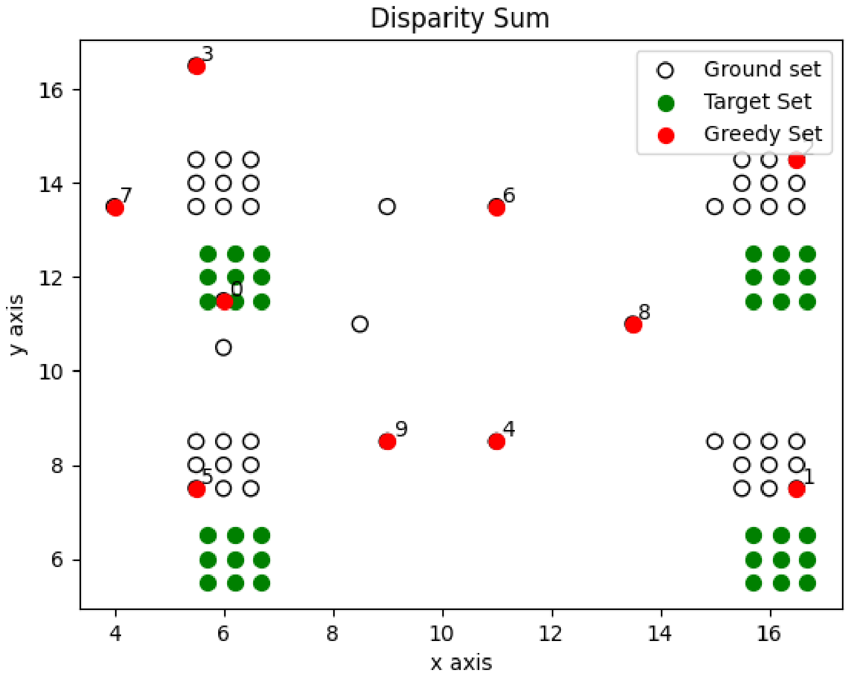

Graph Cut for λ=2
y axis
x axis
16
14
12
10
8
6
4
6
8
10
12
14
16
0
1
2
3
4
5
6
7
8
9
○ Ground set
● Target Set
● Greedy Set


Graph Cut for λ=3
y axis
x axis
16
14
12
10
8
6
4
6
8
10
12
14
16
0
1
2
3
4
5
6
7
8
9
○ Ground set
● Target Set
● Greedy Set

(d) Set Cover

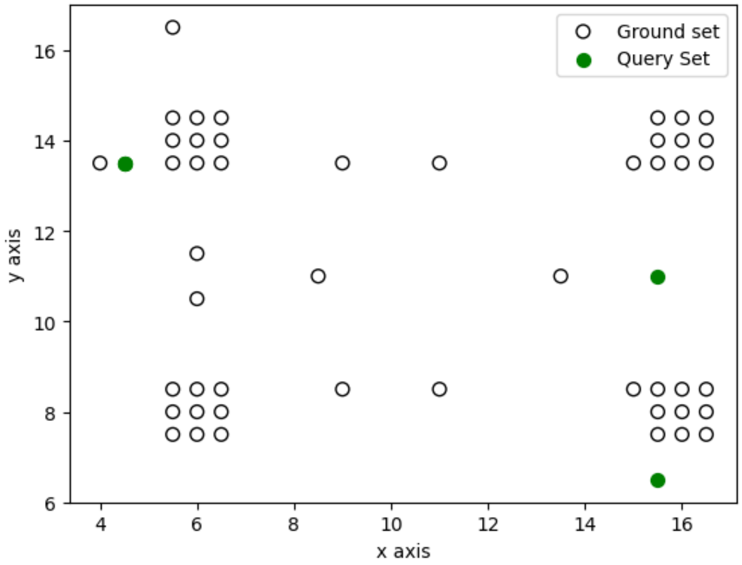


(e) Disparity Sum



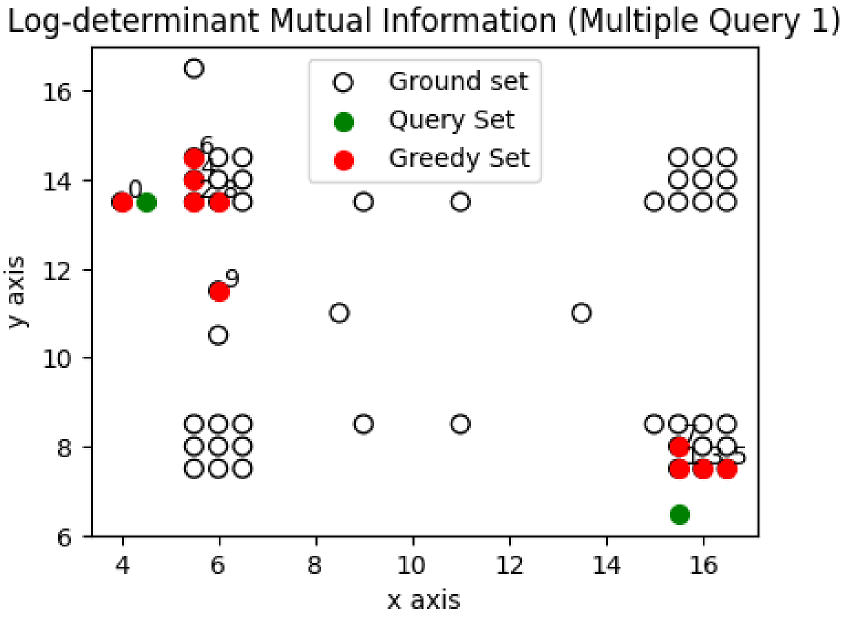
(2) In the next part of the problem, you will explore mutual-information based submodular functions for query-focused summarization. In this case, you are given ground set 'gset_2.txt' and 'qset.txt' (which consists of queries). Note that here is no overlap between the query points and the ground set data points. Multiple queries may be given on a single line, in which case they need to be considered by the function jointly.

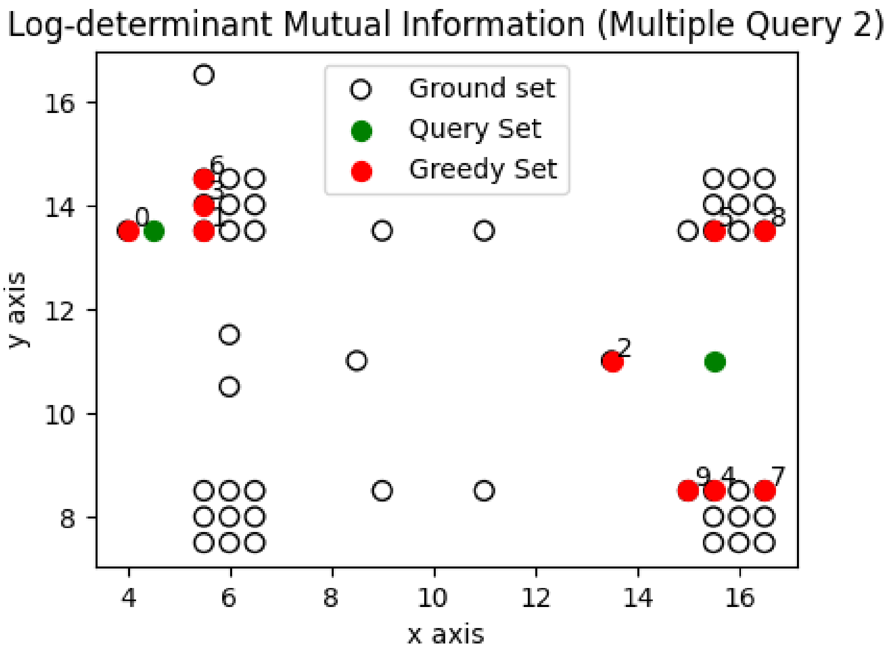
(a) Plot the ground set and query set data using different colors.

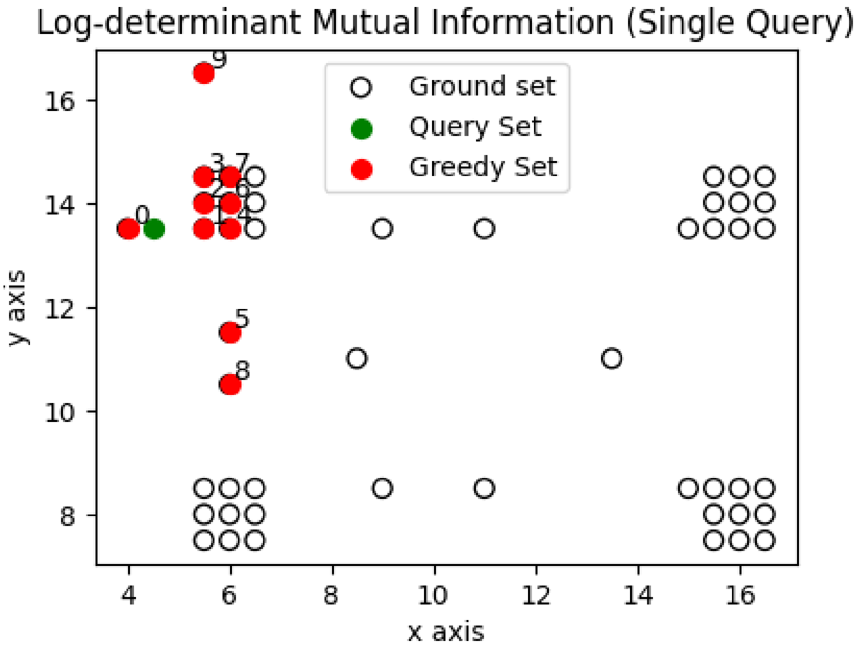


Further, plot the summarization results obtained using the following submodular mutual information functions in a similar manner to the previous part. The optimal set budget continues to be 10 -:

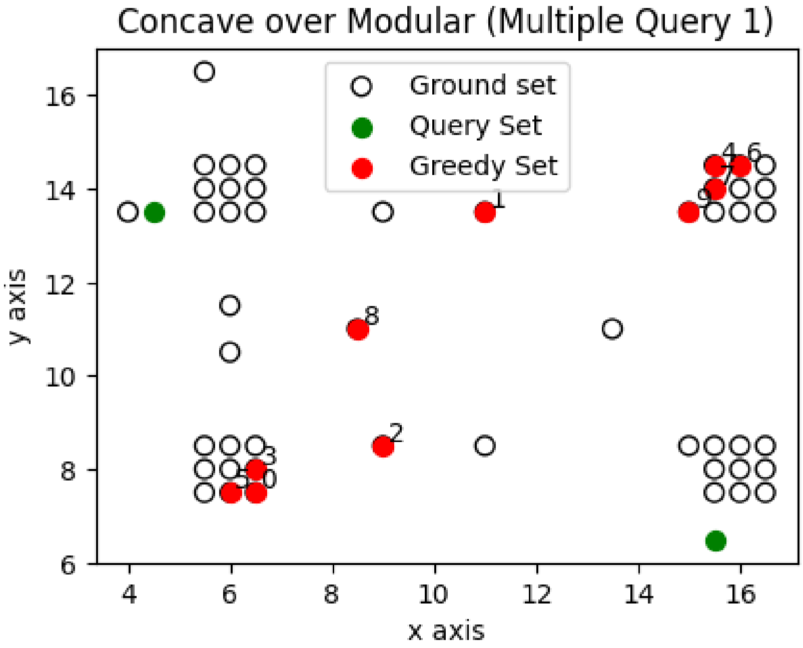
(b) Log-determinant Mutual Information

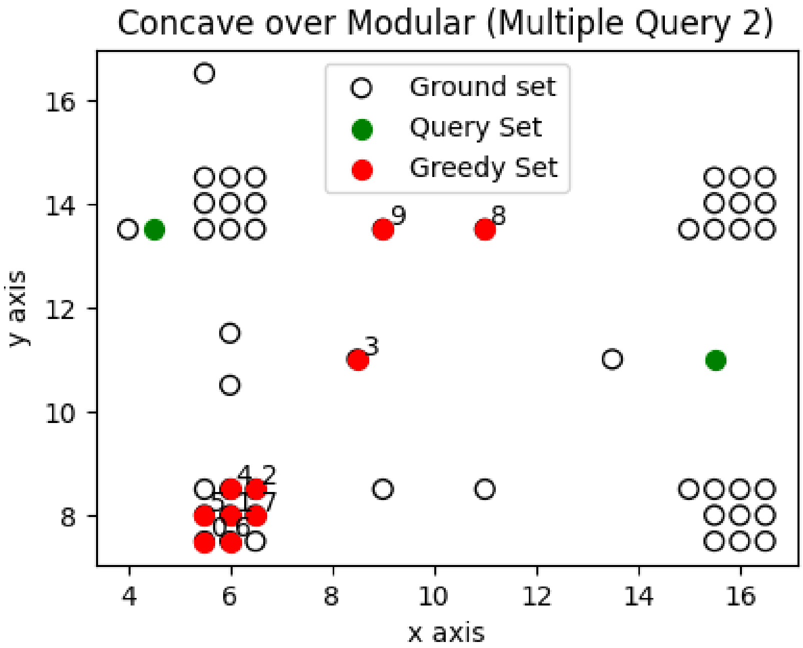


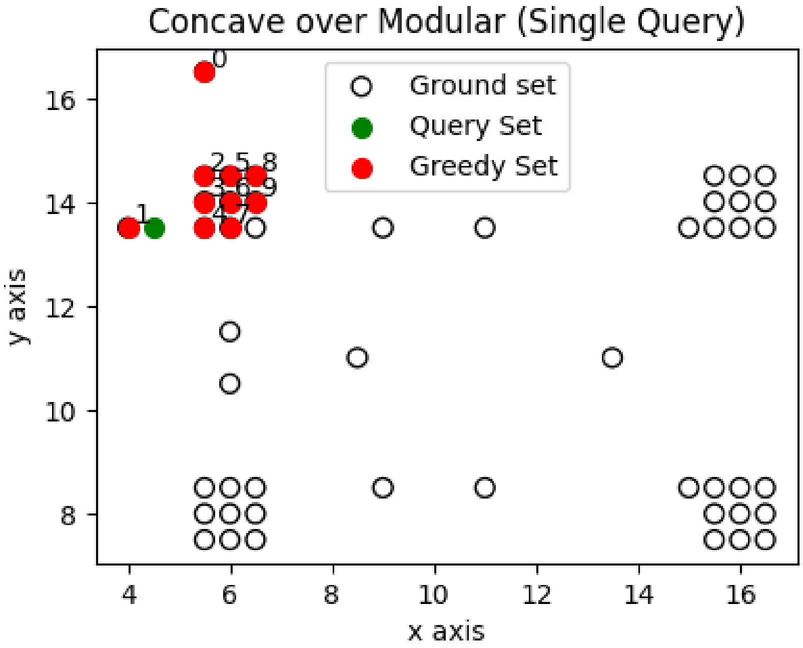

Log-determinant Mutual Information (Multiple Query 2)


Log-determinant Mutual Information (Single Query)

(c) Concave over Modular







Q 5. Recall that the Prox operator of a function h for some argument z is

$$\text{prox}_h(z) = \text{argmin}_x \frac{1}{2\gamma}\|x-z\|^2 + h(x)$$

Compute the Prox operator for $h(x)$ defined as $h(x) = 0$ if $0 \leq x \leq \theta$ (for some fixed $\theta > 0$) and $h(x) = 100$ for any other value of x.

Solution: Let's consider 3 cases:

Case 1: $0 \leq z \leq \theta$

In this case, minimum value occurs when $x = z$. So, the Prox operator reduces to:

$$\text{prox}_h(z) = z$$

Case 2: $z > \theta$ and $\frac{1}{2\gamma}\|z-\theta\|^2 \leq 100$

In this case, the minimum value occurs at $x = \theta$. So, the Prox operator reduces to:

$$\text{prox}_h(z) = \theta$$

Case 3: $z > \theta$ and $\frac{1}{2\gamma}\|z-\theta\|^2 > 100$

In this case, it is better to have $h(x) = 100$ and $x = z$. So, the Prox operator reduces to:

$$\text{prox}_h(z) = z$$

Case 4: $z < 0$ and $\frac{1}{2\gamma}\|z\|^2 > 100$

In this case, it is better to have $h(x) = 100$ and $x = z$. So, the Prox operator reduces to:

$$\text{prox}_h(z) = z$$

Case 5: $z < 0$ and $\frac{1}{2\gamma}\|z\|^2 \leq 100$

In this case, the minimum value occurs at $x = 0$. So, the Prox operator reduces to:

$$\text{prox}_h(z) = 0$$

Hence,

$$\text{prox}_h(z) = \begin{cases} 0 & \text{if } z < 0 \text{ and } \frac{1}{2\gamma}\|z\|^2 \leq 100 \\ \theta & \text{if } z > \theta \text{ and } \frac{1}{2\gamma}\|z-\theta\|^2 \leq 100 \\ z & \text{otherwise} \end{cases}$$

Q 6. Often, in optimization problems in machine learning, we have a simple constraint requiring that parameters lie in a particular interval. Such optimization problems can be effectively solved using the projected gradient descent algorithm discussed in the class.

Derive the exact projection operation of the projected gradient descent algorithm, for the following optimization problem which has the simplest form of such an interval constraint:

$$\begin{array}{ll} \min_{\mathbf{x}\in\Re} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \leq r \text{ and } \quad \mathbf{x} \geq l \end{array}$$

for some fixed given scalars $l < r \in \Re$ and for some machine learning convex loss function $f(\mathbf{x})$. You can use the Karush-Kuhn-Tucker (KKT) conditions for deriving the projection step.

Solution: For deriving the exact projection operation of the projected gradient descent algorithm for the above optimization problem, let's write the Lagrangian function as:

$$\mathrm{L}\,(\mathbf{x}, \lambda_1, \lambda_2) = \mathrm{f}(\mathbf{x}) + \lambda_1(\mathbf{x} - \mathrm{r}) + \lambda_2(l - \mathbf{x})$$

where λ_1 and λ_2 are the Lagrangian multipliers. The corresponding KKT conditions are given below:

1. $\nabla\mathrm{f}(\mathbf{x}) + \lambda_1 - \lambda_2 = 0$
2. $\mathbf{x} - r \leq 0$
3. $l - \mathbf{x} \leq 0$
4. $\lambda_1(\mathbf{x} - r) = 0$
5. $\lambda_2(l - \mathbf{x}) = 0$
6. $\lambda_1 \geq 0$
7. $\lambda_2 \geq 0$

It is clear from KKT condition 4 that $\lambda_1 = 0$ when $\mathbf{x} > r$. Similarly, from KKT condition 5, it is easy to see that $\lambda_2 = 0$ when $\mathbf{x} < l$. Thus, in these cases, we need to project $\mathbf{x}$ onto the interval $[1, r]$. There is no need to project when $l \leq \mathbf{x} \leq r$.

The projection operation can be defined as:

$$P_C(\mathbf{x}) = \begin{cases} l \text{ when } \mathbf{x} < l \\ \mathbf{x} \text{ when } l \leq \mathbf{x} \leq r \\ r \text{ when } \mathbf{x} > r \end{cases}$$

In short,

$$P_C(\mathbf{x}) = \min(\max(\mathbf{x}, l), r)$$

Q 7. Show that the 3-way submodular mutual information $I_f(A; B; C) \geq 0$ if $I_f(A; B)$ is submodular in A for a fixed set B. Similarly show that, $I_f(A; B; C) \leq 0$ if $I_f(A; B)$ is supermodular in A for a fixed set B.

Proof: 3-way submodular mutual information $I_f(A; B; C)$ can be written in terms of 2-way mutual information as

$$I_f(A; B; C) = I_f(A; B) + I_f(C; B) - I_f(A \cup C; B)$$

If $I_f(A; B)$ is submodular in A for a fixed set B, then by union intersection definition of submodularity

$$I_f(A; B) + I_f(C; B) \geq I_f(A \cup C; B) + I_f(A \cap C; B)$$

$$I_f(A; B) + I_f(C; B) - I_f(A \cup C; B) \geq I_f(A \cap C; B)$$

$$\implies I_f(A; B; C) \geq 0$$

Thus, 3-way submodular mutual information $I_f(A; B; C)$ is non negative if $I_f(A; B)$ is submodular in A for a fixed set B. Similarly, if $I_f(A; B)$ is supermodular in A for a fixed set B, then

$$I_f(A; B; C) \leq 0$$

Thus, 3-way submodular mutual information $I_f(A; B; C)$ is non positive if $I_f(A; B)$ is supermodular in A for a fixed set B.

Q 8. Given a monotone submodular function f, does the inequality: $I_f(A_1; A_2; \cdots; A_k) \leq \min(f(A_1), \cdots, f(A_k))$ always hold for any $k \in \mathbb{N}$?

Solution: No, the inequality $I_f(A_1; A_2; \cdots; A_k) \leq \min(f(A_1), \cdots, f(A_k))$ does not hold for all $k \in \mathbb{N}$. It holds for $k = 1, 2, 3$ and 4. However, it doesn't hold for $k = 5$. Hence, it doesn't necessarily hold for $k \geq 5$.

For $k = 1$, the inequality is trivial and for $k = 2$, we have already proved it in part (3) of Q 2.

First, we prove the inequality for $k = 3$. 3-way submodular mutual information is defined as

$$I_f(A; B; C) = f(A) + f(B) + f(C) - f(A \cup B) - f(B \cup C) - f(A \cup C) + f(A \cup B \cup C)$$

$$\implies I_f(A; B; C) = f(A) - (-f(B) - f(C) + f(A \cup B) + f(B \cup C) + f(A \cup C) - f(A \cup B \cup C))$$

$$\implies I_f(A; B; C) = f(A) - (f(A \mid C) + f(C \mid B) - f(C \mid A \cup B))$$

We know that conditional gain $f(A \mid C)$ is non-negative. Also, it is given that f is submodular. So, $f(C \mid B) - f(C \mid A \cup B) \geq 0$. Thus, we have

$$f(A \mid C) + f(C \mid B) - f(C \mid A \cup B) \geq 0$$

$$\implies I_f(A; B; C) \leq f(A)$$

Similarly, we can get

$$I_f(A; B; C) \leq f(B)$$
$$I_f(A; B; C) \leq f(C)$$

Thus, $I_f(A; B; C) \leq \min(f(A), f(B), f(C))$.

Hence, we have proved the inequality for $k = 3$. Now, we prove it for $k = 4$.

We know that the 3-way submodular mutual information is monotone in all its arguments.

$$I_f(A; B; C) \leq I_f(A \cup D; B \cup D; C \cup D)$$

$$\implies I_f(A; B; C; D) = f(D) + I_f(A; B; C) - I_f(A \cup D; B \cup D; C \cup D) \leq f(D)$$

Similarly, we can get

$$I_f(A; B; C; D) \leq f(A)$$
$$I_f(A; B; C; D) \leq f(B)$$
$$I_f(A; B; C; D) \leq f(C)$$

Thus, $I_f(A; B; C; D) \leq \min(f(A), f(B), f(C), f(D))$.

We can not use this proof technique for the k-set case because the k-way submodular mutual information is not necessarily monotone in all its variables. This does not hold for $k = 5$ and thus is not guaranteed to hold for $k \geq 5$. This is proven by the following counter-example.

Let $f(A) = \min(|A|, 3k)$ and let A, B, C, D, E be disjoint sets with $|A| = |B| = |C| = |D| = k$ and $|E| = 1$. Here, the value of RHS is 1. The expansion of $I_f(A; B; C; D; E)$ consists of 5 singleton terms like $f(A)$, 10 pairwise terms like $f(A \cup B)$, 10 triplet terms like $f(A \cup B \cup C)$, 5 terms like $f(A \cup B \cup C \cup D)$ and one term $f(A \cup B \cup C \cup D \cup E)$. Using our definition, the terms $f(A \cup B \cup C \cup D) = 3k$ and similarly, $f(A \cup B \cup C \cup D \cup E) = 3k$. Also, in the formula of I_f, singleton terms have positive signs, pairwise terms have negative signs, triplet terms have positive signs, quadruplet terms have negative signs and $f(A \cup B \cup C \cup D \cup E)$ has positive sign. The contribution of singleton terms is $4k + 1$, pairwise terms is $16k + 4$ and triplet terms is $24k + 6$ (because they are all not saturated). The contribution of quadruplet terms is $15k$ (because it is saturated) and the term with all five sets contributes $3k$. This gives

$$I_f(A; B; C; D; E) = 4k + 1 - 16k - 4 + 24k + 6 - 15k + 3k = 3$$

$$\min(f(A), f(B), f(C), f(D), f(E)) = 1$$
$$\implies I_f(A; B; C; D; E) > \min(f(A), f(B), f(C), f(D), f(E))$$

Thus, this counter-example shows that the upper bound does not hold for $k = 5$.

So, the inequality $I_f(A_1; A_2; \cdots; A_k) \leq \min(f(A_1), \cdots, f(A_k))$ does not hold for all $k \in \mathbb{N}$. It holds for $k = 1, 2, 3$ and 4. However, it doesn't hold for $k = 5$. Hence, it doesn't necessarily hold for $k \geq 5$.

Note: Since $I_f\left(A_1;A_2;\ldots,A_{k+1}\right)=I_f\left(A_1;\ldots;A_k\right)-I_f\left(A_1;\ldots,A_k\mid A_{k+1}\right)$, there exists an upper bound whenever the submodular conditional multi-set information is non-negative because, without loss of generality, if we assume $f\left(A_1\right)\leq f\left(A_2\right)\leq\cdots\leq f\left(A_{k+1}\right)$, it can be shown by induction that $I_f\left(A_1;\ldots,A_k\right)\leq f\left(A_1\right)-I_f\left(A_1;\ldots;A_k\mid A_{k+1}\right)\leq\min_i f\left(A_i\right)$.