[Supplementary Material] Submodular Dictionary Learning for Sparse Coding

Zhuolin Jiang[†], Guangxiao Zhang[†§], Larry S. Davis[†]
[†]Institute for Advanced Computer Studies, University of Maryland, College Park, MD, 20742
[§]Global Land Cover Facility, University of Maryland, College Park, MD, 20742
{zhuolin, gxzhang, lsd}@umiacs.umd.edu

1. Proofs of the Monotonicity and Submodularity Properties of Entropy Rate $\mathcal{H}(A)$

Recall our definition of $\mathcal{H}(A)$:

$$\mathcal{H}(A) = -\sum_i \mu_i \sum_j P_{i,j}(A) \log P_{i,j}(A) \tag{1}$$

where μ_i is the stationary probability of v_i in the stationary distribution $\boldsymbol{\mu}$ and $P_{i,j}(A)$ is the transition probability from v_i to v_j with respect to A.

1.1. Monotonicity

We prove that $\mathcal{H}(A)$ is monotonically increasing by showing $\mathcal{H}(A \cup \{a\}) \geq \mathcal{H}(A)$, for all $a \in E \backslash A$ and $A \subseteq E$. Without loss of generality, we assume $a = e_{1,2}$. The weights of the self loops for v_1 and v_2 are given by:

$$w_{1,1} = w_1 - \sum_{j:e_{1,j} \in A \cup \{a\}} w_{1,j}, \tag{2}$$

$$w_{2,2} = w_2 - \sum_{j:e_{2,j} \in A \cup \{a\}} w_{2,j}. \tag{3}$$

By the definition of entropy rate in (1), the increase of entropy rate due to the addition of a to A is computed as:

$$\begin{aligned}
&\mathcal{H}(A \cup \{a\}) - \mathcal{H}(A) \\
&= -\sum_i \mu_i \sum_j P_{i,j}(A \cup \{a\}) \log P_{i,j}(A \cup \{a\}) \\
&\quad + \sum_i \mu_i \sum_j P_{i,j}(A) \log P_{i,j}(A) \qquad (4) \\
&= -\sum_i \sum_j \frac{w_i}{w_{all}} P_{i,j}(A \cup \{a\}) \log P_{i,j}(A \cup \{a\}) \\
&\quad + \sum_i \sum_j \frac{w_i}{w_{all}} P_{i,j}(A) \log P_{i,j}(A) \qquad (5) \\
&= -\sum_i \sum_j \frac{w_i}{w_{all}} P_{i,j}(A \cup \{a\}) \log \frac{w_i P_{i,j}(A \cup \{a\})}{w_{all}} \\
&\quad + \sum_i \sum_j \frac{w_i}{w_{all}} P_{i,j}(A \cup \{a\}) \log \frac{w_i}{w_{all}} \\
&\quad + \sum_i \sum_j \frac{w_i}{w_{all}} P_{i,j}(A) \log \frac{w_i P_{i,j}(A)}{w_{all}} \\
&\quad - \sum_i \sum_j \frac{w_i}{w_{all}} P_{i,j}(A) \log \frac{w_i}{w_{all}} \qquad (6) \\
&= -\sum_i \sum_j \frac{w_i}{w_{all}} P_{i,j}(A \cup \{a\}) \log \frac{w_i P_{i,j}(A \cup \{a\})}{w_{all}} \\
&\quad + \sum_i \sum_j \frac{w_i}{w_{all}} P_{i,j}(A) \log \frac{w_i P_{i,j}(A)}{w_{all}} \\
&\quad + \sum_i \frac{w_i}{w_{all}} \log \frac{w_i}{w_{all}} \left(\sum_j P_{i,j}(A \cup \{a\}) - \sum_j P_{i,j}(A) \right) \qquad (7)
\end{aligned}$$

Since $\sum_j P_{i,j}(A \cup \{a\}) = \sum_j P_{i,j}(A) = 1$, the last term in (7) becomes zero. Hence we have

$$\begin{aligned}
&\mathcal{H}(A \cup \{a\}) - \mathcal{H}(A) \\
&= -\sum_i \sum_j \frac{w_i}{w_{all}} P_{i,j}(A \cup \{a\}) \log \frac{w_i P_{i,j}(A \cup \{a\})}{w_{all}} \\
&\quad + \sum_i \sum_j \frac{w_i}{w_{all}} P_{i,j}(A) \log \frac{w_i P_{i,j}(A)}{w_{all}} \qquad (8)
\end{aligned}$$

We notice that in (8), all the terms associated with vertices other than v_1 and v_2 are canceled out if $a = e_{1,2}$. Thus,

$$\begin{aligned}
&\mathcal{H}(A \cup \{e_{1,2}\}) - \mathcal{H}(A) \\
&= -\Big\{ \frac{w_1}{w_{all}} P_{1,1}(A \cup \{e_{1,2}\}) \log \frac{w_1 P_{1,1}(A \cup \{e_{1,2}\})}{w_{all}} \\
&\qquad + \frac{w_1}{w_{all}} P_{1,2}(A \cup \{e_{1,2}\}) \log \frac{w_1 P_{1,2}(A \cup \{e_{1,2}\})}{w_{all}} \Big\} \\
&\quad + \Big\{ \frac{w_1}{w_{all}} P_{1,1}(A) \log \frac{w_1 P_{1,1}(A)}{w_{all}} \\
&\qquad + \frac{w_1}{w_{all}} P_{1,2}(A) \log \frac{w_1 P_{1,2}(A)}{w_{all}} \Big\} \\
&\quad - \Big\{ \frac{w_2}{w_{all}} P_{2,1}(A \cup \{e_{1,2}\}) \log \frac{w_2 P_{2,1}(A \cup \{e_{1,2}\})}{w_{all}} \\
&\qquad + \frac{w_2}{w_{all}} P_{2,2}(A \cup \{e_{1,2}\}) \log \frac{w_2 P_{2,2}(A \cup \{e_{1,2}\})}{w_{all}} \Big\} \\
&\quad + \Big\{ \frac{w_2}{w_{all}} P_{2,1}(A) \log \frac{w_2 P_{2,1}(A)}{w_{all}} \\
&\qquad + \frac{w_2}{w_{all}} P_{2,2}(A) \log \frac{w_2 P_{2,2}(A)}{w_{all}} \Big\} \qquad (9)
\end{aligned}$$

Recall the definition of the transition probability:

$$P_{i,j}(A) = \begin{cases} 1 - \frac{\sum_{j:e_{i,j} \in A} w_{i,j}}{w_i} & \text{if } i = j, \\ \frac{w_{i,j}}{w_i} & \text{if } i \neq j,\ e_{i,j} \in A, \\ 0 & \text{if } i \neq j,\ e_{i,j} \notin A. \end{cases} \tag{10}$$

Note that $P_{i,j}(A) = 0$ if there is no edge connecting v_i and v_j. Hence, $P_{1,2}(A) = P_{2,1}(A) = 0$. From (2), (3) and the definition of $P_{i,j}$, (9) becomes:

$$\begin{aligned}
&\mathcal{H}(A \cup \{e_{1,2}\}) - \mathcal{H}(A) \\
&= \frac{w_{1,1} + w_{1,2}}{w_{all}} \log \frac{w_{1,1} + w_{1,2}}{w_{all}} - \frac{w_{1,1}}{w_{all}} \log \frac{w_{1,1}}{w_{all}} - \frac{w_{1,2}}{w_{all}} \log \frac{w_{1,2}}{w_{all}} \\
&\quad + \frac{w_{2,2} + w_{2,1}}{w_{all}} \log \frac{w_{2,2} + w_{2,1}}{w_{all}} - \frac{w_{2,2}}{w_{all}} \log \frac{w_{2,2}}{w_{all}} - \frac{w_{2,1}}{w_{all}} \log \frac{w_{2,1}}{w_{all}} \qquad (11) \\
&= f\Big(\frac{w_{1,1}}{w_{all}} + \frac{w_{1,2}}{w_{all}}\Big) - f\Big(\frac{w_{1,1}}{w_{all}}\Big) - f\Big(\frac{w_{1,2}}{w_{all}}\Big) \\
&\quad + f\Big(\frac{w_{2,2}}{w_{all}} + \frac{w_{2,1}}{w_{all}}\Big) - f\Big(\frac{w_{2,2}}{w_{all}}\Big) - f\Big(\frac{w_{2,1}}{w_{all}}\Big) \qquad (12) \\
&\geq 0 \qquad (13)
\end{aligned}$$

Note in (12), a convex function $f(x)$ in (0,1) is defined as: $f(x) = x \log x$. It's easy to show that the convex function $f(x)$ is superadditive in (0,1), *i.e.,*

$$\begin{aligned} f(x_1) + f(x_2) &= f\left((x_1+x_2)\frac{x_1}{x_1+x_2}\right) + f\left((x_1+x_2)\frac{x_2}{x_1+x_2}\right) \\ &\le \frac{x_1}{x_1+x_2} f(x_1+x_2) + \frac{x_2}{x_1+x_2} f(x_1+x_2) \\ &= f(x_1+x_2). \end{aligned} \tag{14}$$

Hence, inequality (13) holds, which completes the proof of the monotonically increasing property of $\mathcal{H}(A)$.

1.2. Submodularity

We prove $\mathcal{H}(A)$ is a submodular function by showing

$$\begin{aligned} &\mathcal{H}(A \cup \{a_1\}) - \mathcal{H}(A) \\ &\quad \ge \mathcal{H}(A \cup \{a_1, a_2\}) - \mathcal{H}(A \cup \{a_2\}), \quad \forall a_1, a_2 \in E \backslash A \end{aligned} \tag{15}$$

Based on whether a_1, a_2 have a common vertex or not, we compare the value of $\mathcal{H}(A \cup \{a_1\}) - \mathcal{H}(A)$ with the value of $\mathcal{H}(A \cup \{a_1, a_2\}) - \mathcal{H}(A \cup \{a_2\})$ in two cases.

- **Case1**: a_1, a_2 share no common vertex. Without loss of generality, we assume $a_1 = e_{1,2}$ and $a_2 = e_{3,4}$. According to (9), adding a_1 to A causes the same weight changes as adding a_1 to $A \cup \{a_2\}$ because the addition of a_2 has no effect on the loop weights of v_1 and v_2.

$$\begin{aligned} &\mathcal{H}(A \cup \{a_1, a_2\}) - \mathcal{H}(A \cup \{a_2\}) \\ =& \frac{w_{1,1}+w_{1,2}}{w_{all}} \log \frac{w_{1,1}+w_{1,2}}{w_{all}} - \frac{w_{1,1}}{w_{all}} \log \frac{w_{1,1}}{w_{all}} - \frac{w_{1,2}}{w_{all}} \log \frac{w_{1,2}}{w_{all}} \\ +& \frac{w_{2,2}+w_{2,1}}{w_{all}} \log \frac{w_{2,2}+w_{2,1}}{w_{all}} - \frac{w_{2,2}}{w_{all}} \log \frac{w_{2,2}}{w_{all}} - \frac{w_{2,1}}{w_{all}} \log \frac{w_{2,1}}{w_{all}} \end{aligned} \tag{16}$$

  Thus, $\mathcal{H}(A \cup \{a_1\}) - \mathcal{H}(A)) = \mathcal{H}(A \cup \{a_1, a_2\}) - \mathcal{H}(A \cup \{a_2\})$

- **Case2**: a_1, a_2 share a common vertex. Without loss of generality, We assume $a_1 = e_{1,2}$ and $a_2 = e_{1,3}$. Then the new loop weights for vertex v_1 and v_2 are given by:

$$w'_{1,1} = w_1 - \sum_{j: e_{1,j} \in A \cup \{e_{1,2}, e_{1,3}\}} w_{1,j} = w_{1,1} - w_{1,3} \tag{17}$$

$$w'_{2,2} = w_2 - \sum_{j: e_{2,j} \in A \cup \{e_{1,2}, e_{1,3}\}} w_{2,j} = w_{2,2}, \tag{18}$$

  where $w_{1,1}$ and $w_{2,2}$ here are given by (2) and (3). Hence

$$\begin{aligned} &(\mathcal{H}(A \cup \{a_1\}) - \mathcal{H}(A)) - (\mathcal{H}(A \cup \{a_1, a_2\}) - \mathcal{H}(A \cup \{a_2\}) \\ =& \left\{ \frac{w_{1,1}+w_{1,2}}{w_{all}} \log \frac{w_{1,1}+w_{1,2}}{w_{all}} - \frac{w_{1,1}}{w_{all}} \log \frac{w_{1,1}}{w_{all}} - \frac{w_{1,2}}{w_{all}} \log \frac{w_{1,2}}{w_{all}} \right. \\ &\left. + \frac{w_{2,2}+w_{2,1}}{w_{all}} \log \frac{w_{2,2}+w_{2,1}}{w_{all}} - \frac{w_{2,2}}{w_{all}} \log \frac{w_{2,2}}{w_{all}} - \frac{w_{2,1}}{w_{all}} \log \frac{w_{2,1}}{w_{all}} \right\} \\ -& \left\{ \frac{w'_{1,1}+w_{1,2}}{w_{all}} \log \frac{w'_{1,1}+w_{1,2}}{w_{all}} - \frac{w'_{1,1}}{w_{all}} \log \frac{w'_{1,1}}{w_{all}} - \frac{w_{1,2}}{w_{all}} \log \frac{w_{1,2}}{w_{all}} \right. \\ &\left. + \frac{w_{2,2}+w_{2,1}}{w_{all}} \log \frac{w_{2,2}+w_{2,1}}{w_{all}} - \frac{w_{2,2}}{w_{all}} \log \frac{w_{2,2}}{w_{all}} - \frac{w_{2,1}}{w_{all}} \log \frac{w_{2,1}}{w_{all}} \right\} \end{aligned} \tag{19}$$

$$\begin{aligned} =& \left\{ \frac{w_{1,1}+w_{1,2}}{w_{all}} \log \frac{w_{1,1}+w_{1,2}}{w_{all}} - \frac{w_{1,1}}{w_{all}} \log \frac{w_{1,1}}{w_{all}} \right\} \\ &- \left\{ \frac{w'_{1,1}+w_{1,2}}{w_{all}} \log \frac{w'_{1,1}+w_{1,2}}{w_{all}} - \frac{w'_{1,1}}{w_{all}} \log \frac{w'_{1,1}}{w_{all}} \right\} \end{aligned} \tag{20}$$

$$= g(\frac{w'_{1,1}+w_{1,3}}{w_{all}}) - g(\frac{w'_{1,1}}{w_{all}}) \tag{21}$$

$$\ge 0 \tag{22}$$

  From (20) to (21), the relationship between $w_{1,1}$ and $w'_{1,1}$ given in (17) is employed. And $g(x)$ in (21) is defined as:

$$g(x) = (x+\delta)\log(x+\delta) - x \log x \tag{23}$$

  Here $\delta = \frac{w_{1,2}}{w_{all}}$. By taking advantage of the strictly increasing property of $g(x)$, we arrive at (22).

Showing the two cases above, we conclude that $\mathcal{H}(A)$ is a submodular function.

2. Proofs of the Monotonicity and Submodularity Properties of Discriminative Term $\mathcal{Q}(A)$

Recall our definition,

$$\mathcal{Q}(A) = \frac{1}{C} \sum_{i=1}^{N_A} \max_y N_y^i - N_A \tag{24}$$

where $\max_y N_y^i$ denotes the maximum element of the count vector $\mathbf{N}^i = [N_1^i, ..., N_m^i]^t$ for cluster S_i, N_A is the number of connected components.

2.1. Monotonicity

We prove that $\mathcal{Q}(A)$ is monotonically increasing by showing:

$$\mathcal{Q}(A \cup \{a\}) \ge \mathcal{Q}(A), \tag{25}$$

for all $a \in E \backslash A$ and $A \subseteq E$.

Given any set of selected edges A and its corresponding graph partitioning $\mathcal{S}_A = \{S_1, ..., S_{N_A}\}$, we are only interested in the nontrivial case in which the two vertices of a belong to different clusters. Otherwise the addition of edge a has no impact on the graph partitioning, *i.e.*, $\mathcal{Q}(A \cup \{a\}) - \mathcal{Q}(A) = 0$.

Without loss of generality, we assume $a = e_{1,2}$, v_1 and v_2 belong to S_i and S_j, respectively. The new graph partitioning $\mathcal{S}_{A\cup\{e_{1,2}\}}$ for $A \cup \{e_{1,2}\}$ is similar to the graph partitioning $\mathcal{S}_A$ for A except one thing: clusters S_i and S_j are merged into one cluster S_*. Hence,

$$\begin{aligned}\mathcal{Q}(A\cup\{a=e_{1,2}\}) - \mathcal{Q}(A) &= (\frac{1}{C}\sum_{k=1}^{N_A-1}\max_y N_y^k - (N_A-1)) \\ &- (\frac{1}{C}\sum_{k=1}^{N_A}\max_y N_y^k - N_A) \\ &= \frac{1}{C}(\max_y[N_y^i + N_y^j] - \max_y N_y^i - \max_y N_y^j) + 1 \\ &= \frac{1}{C}(\max_y N_y^* - \max_y N_y^i - \max_y N_y^j) + 1\end{aligned} \quad (26)$$

By definition,

$$C = \sum_i \sum_y N_y^i \geq \max_y N_y^i + \max_y N_y^j \quad (27)$$

and with

$$\max_y N_y^* \geq 0,$$

so (26) becomes

$$\mathcal{Q}(A\cup\{a\}) - \mathcal{Q}(A) \geq \frac{1}{C}(0 - C) + 1 = 0 \quad (28)$$

This completes the proof of monotonically increasing property of $\mathcal{Q}(A)$.

2.2. Submodularity

Before starting the proof of submodularity, we want to introduce the following two properties of a count vector $\mathbf{N}^i = [N_1^i, ..., N_m^i]^t$.

(1) (Nonnegative) The elements of the count vector are all nonnegative, $\mathbf{N}^i \geq 0, i = 1, ..., N$.

(2) (Subadditivity) Given the new cluster S_* by merging clusters S_i and S_j, the count of the dominating class for cluster S_* is less than the sum of the counts of the dominating class for S_i and S_j, *i.e.*,

$$\max_y[N_y^i + N_y^j] \leq \max_y N_y^i + \max_y N_y^j, \quad (29)$$

with equality holds only when

$$\arg\max_y N_y^i = \arg\max_y N_y^j$$

Now we prove that $\mathcal{Q}(A)$ is submodular by showing

$$\mathcal{Q}(A\cup\{a_1\}) - \mathcal{Q}(A) \geq \mathcal{Q}(A\cup\{a_1,a_2\}) - \mathcal{Q}(A\cup\{a_2\}) \quad \forall a_1, a_2 \in E\backslash A \quad (30)$$

Again we consider only the nontrivial case in which edge a_1 combines two different subsets S_i and S_j. When $i = j$, $\mathcal{Q}(A\cup\{a_1\}) - \mathcal{Q}(A) = \mathcal{Q}(A\cup\{a_1,a_2\}) - \mathcal{Q}(A\cup\{a_2\}) = 0$.

Suppose the vertices of a_2 belong to clusters S_m, S_n, respectively. Based on the relationship among i, j, m, n ($i \neq j$), we discuss in the following four cases.

- **Case1** (trivial): $m = n$, *i.e.*, the vertices of a_2 belong to the same cluster. Then adding a_2 has no effect on the graph:

$$\begin{aligned}&\mathcal{Q}(A\cup\{a_2\}) = \mathcal{Q}(A), \\ &\mathcal{Q}(A\cup\{a_1,a_2\}) = \mathcal{Q}(A\cup\{a_1\})\end{aligned} \quad (31)$$

Thus $\mathcal{Q}(A\cup\{a_1\}) - \mathcal{Q}(A) = \mathcal{Q}(A\cup\{a_1,a_2\}) - \mathcal{Q}(A\cup\{a_2\})$.

- **Case2** (trivial): $m \neq n$, $\{m,n\} = \{i,j\}$, *i.e.*, adding a_2 to the graph has the same effect as adding a_1. Thus,

$$\mathcal{Q}(A\cup\{a_2\}) = \mathcal{Q}(A\cup\{a_1,a_2\}) \quad (32)$$

Together with monotonically increasing property in (28), we have

$$(\mathcal{Q}(A\cup\{a_1\}) - \mathcal{Q}(A)) \geq \mathcal{Q}(A\cup\{a_1,a_2\}) - \mathcal{Q}(A\cup\{a_2\}) = 0 \quad (33)$$

- **Case3**: $\{m,n\} \cap \{i,j\} = \emptyset$, *i.e.*, a_2 combines two clusters S_m, S_n that are not S_i, S_j.

$$\begin{aligned}&\mathcal{Q}(A\cup\{a_1,a_2\}) - \mathcal{Q}(A\cup\{a_2\}) \\ &= \frac{1}{C}\Big\{\max_y[N_y^i + N_y^j] + \max_y[N_y^m + N_y^n] \\ &\quad - \max_y N_y^i - \max_y N_y^j - \max_y[N_y^m + N_y^n]\Big\} + 1 \quad (34) \\ &= \frac{1}{C}\Big\{\max_y[N_y^i + N_y^j] - \max_y N_y^i - \max_y N_y^j\Big\} + 1 \\ &= \mathcal{Q}(A\cup\{a_1\}) - \mathcal{Q}(A) \quad (35)\end{aligned}$$

- **Case4**: $m \in \{i,j\}$, and $n \notin \{i,j\}$. Without loss of generality, we assume $m = i, n = k \neq i, j$, *i.e.*, a_2 combines two subsets S_i, S_k.

$$\begin{aligned}&(\mathcal{Q}(A\cup\{a_1\}) - \mathcal{Q}(A)) - (\mathcal{Q}(A\cup\{a_1,a_2\}) - \mathcal{Q}(A\cup\{a_2\})) \\ &= \frac{1}{C}\Big\{\max_y[N_y^i + N_y^j] - \max_y N_y^i - \max_y N_y^j\Big\} + 1 \\ &\quad - \frac{1}{C}\Big\{\max_y[N_y^i + N_y^j + N_y^k] - \max_y[N_y^i + N_y^k] - \max_y N_y^j\Big\} - 1 \\ &= \frac{1}{C}\Big\{\Big(\max_y[N_y^i + N_y^j] - \max_y N_y^i\Big) \\ &\quad - \Big(\max_y[N_y^i + N_y^k + N_y^j] - \max_y[N_y^i + N_y^k]\Big)\Big\}\end{aligned} \quad (36)$$

Based on the dominating class labels of cluster S_i, S_j and S_k, we compare the values of $\big(\max_y[N_y^i + N_y^j] - \max_y N_y^i\big)$ and $\big(\max_y[N_y^i + N_y^k + N_y^j] - \max_y[N_y^i + N_y^k]\big)$ in the following three situations:

(a) $\arg\max_y N_y^i = \arg\max_y N_y^j$

In this case, S_i and S_j share the same dominating class. From (29) we have,

$$\max_y[N_y^i + N_y^j] - \max_y N_y^i = \max_y N_y^j \quad (37)$$

$$\max_y[N_y^i + N_y^k + N_y^j] - \max_y[N_y^i + N_y^k] \leq \max_y N_y^j \quad (38)$$

This implies that (36) is ≥ 0.

(b) $\arg\max_y N_y^i \neq \arg\max_y N_y^j$, $\max_y N_y^i \geq \max_y N_y^j$

In this case S_i, S_j do not share the same dominating class, and the dominating class label of S_i will become the dominating class label for the merged cluster of S_i, S_j. Therefore, $\max_y[N_y^i + N_y^j] - \max_y N_y^i = 0$.

Note that the dominating class labels of S_j and S_k must not be the same. If S_j, S_k shares the same dominating class label and S_i has a different dominating class label, then according to the proposed greedy algorithm, the edge connecting S_j, S_k (referred to as a_3) must already exist in A before considering a_1, a_2. However, cycle-free constraint requires that a_1, a_2, a_3 cannot exist at the same time. By this contradiction, we conclude that S_j, S_k must have different dominating class labels.

Moreover, taking $\arg\max_y N_y^i \neq \arg\max_y N_y^j$ into consideration, the dominating class, after merging S_i, S_j, and S_k, can only be either $\arg\max_y N_y^i$ or $\arg\max_y N_y^k$, in both case of which we have

$$\max_y[N_y^i + N_y^k + N_y^j] = \max_y[N_y^i + N_y^k], \quad (39)$$

which yields that $\max_y[N_y^i + N_y^j] - \max_y N_y^i = \max_y[N_y^i + N_y^k + N_y^j] - \max_y[N_y^i + N_y^k] = 0$. Thus, (36) is $= 0$.

(c) $\arg\max_y N_y^i \neq \arg\max_y N_y^j, \max_y N_y^i < \max_y N_y^j$

In this case, the dominating class label for S_j becomes the dominating class label for the merged cluster of S_i, S_j, *i.e.*,

$$\max_y[N_y^i + N_y^j] - \max_y N_y^i = \max_y N_y^j \quad (40)$$

Again according to (29),

$$\max_y[N_y^i + N_y^k + N_y^j] - \max_y[N_y^i + N_y^k] \leq \max_y N_y^j \quad (41)$$

which implies (36) is ≥ 0.

From the discussion above we prove that (36) is always ≥ 0, *i.e.*,

$$\mathcal{Q}(A \cup \{a_1\}) - \mathcal{Q}(A) \geq \mathcal{Q}(A \cup \{a_1, a_2\}) - \mathcal{Q}(A \cup \{a_2\}).$$

Summarizing the four cases above, we conclude that $\mathcal{Q}(A)$ is a submodular function.

3. Proof of Matroid

We claim that the cycle free constraint and the connected component constraint induce a matroid $\mathcal{M} = (E, \mathcal{I})$, where E is the edge set, and $\mathcal{I}$ is the collection of subsets $A \subseteq E$ which satisfies (a) A is cycle-free, and (b) the graph partition from A has more than K connected components, *i.e.*, $N_A \geq K$.

$\mathcal{M}$ satisfies the following three conditions:

- $\emptyset \in \mathcal{I}$: It's obvious that the empty set $\emptyset$ induces no cycles. The graph associated with $\emptyset$ has $N_\emptyset = |V|$ connected components, where the total number of nodes $|V| \geq K$. Therefore $\emptyset \in \mathcal{I}$.

- (Hereditary property): Assume $A \in \mathcal{I}$, and $B \subseteq A$. Denote the the graphs associated with edge set A, B as G_A and G_B respectively. Under our constraints (*i.e.,* $A \in \mathcal{I}$) G_A is cycle free, and $N_A \geq K$. B also satisfies $B \in \mathcal{I}$ because: (a) G_B is cycle free, since removing edges from G_A cannot create cycles. (b) $N_B \geq K$, since removing edges from G_A cannot decrease the number of connected components.

- (Exchange property): Suppose $A \in \mathcal{I}$, $B \in \mathcal{I}$, and $|A| < |B|$. Denote the the graphs associated with A, B as G_A, G_B respectively. Clearly G_A has $N_A = |V| - |A|$ connected components, and G_B has $N_B = |V| - |B|$ connected components, where $N_A > N_B$. This means G_B has fewer connected components than G_A, *i.e.*, G_B must contain some connected components, S_i, whose vertices are in two different connected components in G_A. Moreover, since S_i is connected, there must exist an edge $x \in B$ such that x connects two vertices in two different components in G_A. We can add that edge x without creating a cycle. Since $N_A \geq K$, $N_B \geq K$, and $N_A > N_B$, it must be true that $N_A \geq K + 1$. Moreover, adding one edge to a graph decreases the number of connected components by at most one. Hence $N_{A\cup\{x\}} \geq K$, which satisfies the connected component constraint. With that being said, for A, $B \in \mathcal{I}$, and $|A| < |B|$, there exists an element $x \in B - A$ such that $A \cup \{x\} \in \mathcal{I}$

With the three conditions satisfied, we conclude that $\mathcal{M} = (E, \mathcal{I})$ is a matroid.