Greedy Static Max-Min vs Greedy Static Top-1

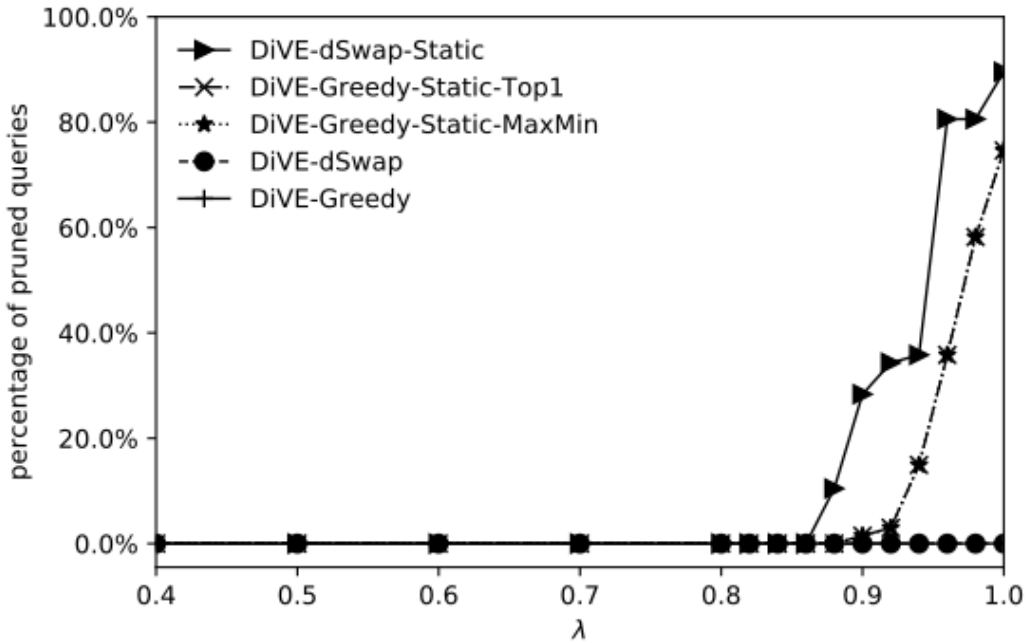
Greedy static Max-min

1. Get two most distant views as the initialization set S.
2. Use √2 as the maximum static bound, this maximum will not be changed until the end.
3. Compute utility U of all views in X using √2 as the max bound and use actual diversity score.
4. Sort views in X based on the highest utility score, store in the list L.
5. Calculate Umax and Umin
6. Prune all views which have Umax < max(Umin)
7. Execute the first query view in L and calculate the real objective function F(S) as the *current_F(S)*.
8. Update max(Umin) = *current_F(S)*.
9. Repeat step 6.
10. Execute the next query view in L, calculate the real objective function F(S), if this F(S) higher than the *current_F(S)*, then use it as the *current_F(S)* and update the max(Umin) = *current_F(S)*.
11. Repeat step 9 and 10 until there is no F(S) that higher than *current_F(S)*.
12. Go to the next iteration.

Greedy Static Top-1

1. Get two most distant views as the initialization set S.
2. Use √2 as the maximum static bound, this maximum will not be changed until the end.
3. Compute utility U of all views in X using √2 as the max bound and use actual diversity score.
4. Sort views in X based on the highest utility score, store in the list L.
5. Execute the first query view in L and calculate the real objective function F(S) as the *current_F(S)*.
6. If U of views in L < *current_F(S)*, those views will be pruned.
7. Execute the next query view, calculate the real objective function F(S), if this F(S) higher than the current, then replace the current.
8. Repeat step 6 and 7 till there is no F(S) that higher than *current_F(S)*.
9. Go to the next iteration.

Greedy Static MaxMin algorithm is equal to Greedy Static Top-1

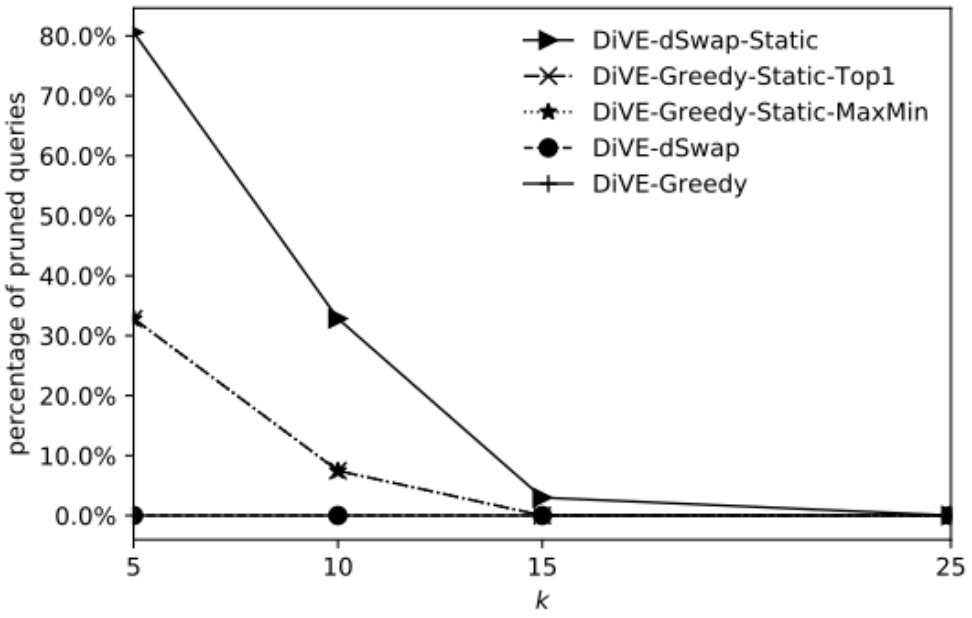


This experiment using Flights dataset, I have checked that the result set S of MaxMin is equal to Top1. For the rest, we just call Greedy Maxmin as Greedy top1.

The reason why we need swap is because Greedy starts with small number of views in the initialization. Only two views as the set S. Hence, while calculating the utility score of each views in X using static maximum bound and actual diversity score, there are a lot of views have same utility score. Due to of this, the chance of pruning is low.

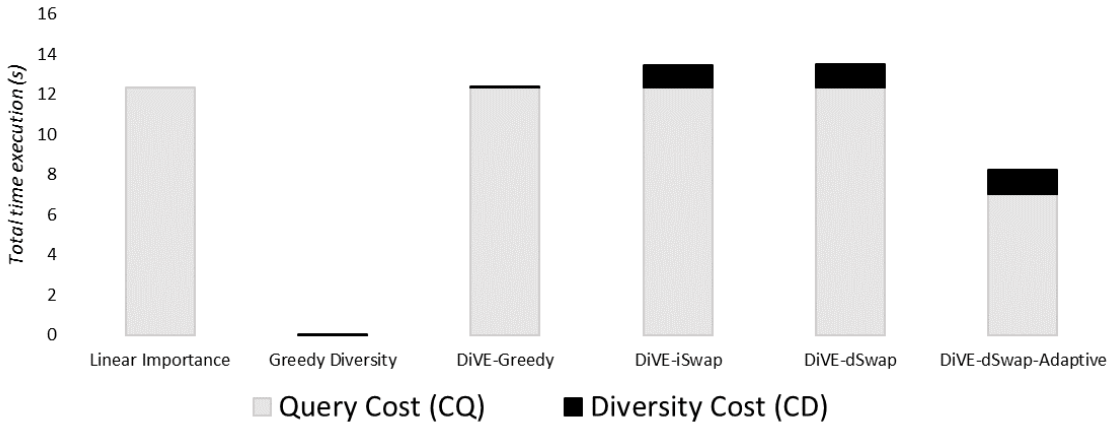
However, Swap has bigger number of views in the initialization (e.g. five views). While calculating utility score of views in X, using static maximum bound and actual diversity score, the chance of views have same utility score is lower than Greedy.

Impact of k to pruning performance



Both Greedy static and Swap static, the pruning happens while $\lambda > 0.8$. In this experiment, we use $\lambda = 0.95$ to see whether the number of *k* has an impact to the pruning performance or not.

Time execution on Flights dataset



In Figure above, DiVE-dSwap-Adaptive uses $\lambda = 0.5$ and PI = 0.97. In this figure there is no DiVE-Greedy-Adaptive because it was not finished yet.

Distance Functions

I have been looking for the bounded distance function to compare between two probability distributions, there are several distance functions that I have studied:

1. Euclidean distance, currently is used and the maximum bound is $\sqrt{2}$, it's proven below.
2. Kullback-Leiber (KL) distance, this distance is not bounded, the mathematically proof has been sent on the previous report.
3. Earth Mover Distance (EMD), this distance is widely used and very good for comparing two probability distributions. Mostly this distance used in computer vision application to compare between two histograms. However, based on my finding, this distance is not bounded as well.

Euclidean distance:

For the general case, Euclidean distance is defined as following:

$$d = \sqrt{\sum (x - y)^2}$$

Which is:

$$\sum (x - y)^2 = \sum x^2 + \sum y^2 - 2\sum xy$$

Where, all values are between 0 and 1 (sum up 1), $\sum x = \sum y = 1$

The theoretical maximum is attained when $2\sum xy = 0$ while $\sqrt{2}$ as the maximum theoretical bound can be proven as following:

$$\sqrt{\sum (x - y)^2} \leq \sqrt{\sum x^2 + \sum y^2}$$

$$\sqrt{\sum (x - y)^2} \leq \sqrt{1 + 1}$$

$$\sqrt{\sum (x - y)^2} \leq \sqrt{2}$$

Update:

Currently I am working on:

1. Change the diversity function to max-min and compare the result to the current result (max-sum)
2. Working on DiVE-Greedy-Adaptive using top-1
3. Trying to implement check points to rectifying bound on the Adaptive algorithm.
4. I found one interesting dataset which is Airbnb dataset. Now I am importing to my db engine and will see if that dataset may contains some interesting insights.
5. One distance function that has explicit maximum bound is Kolmogorov distance, this distance is commonly used to compare between two probability distributions as well. This distance often used as hypothesis test on statistics. I am still studying whether this distance is applicable to our work or not.