FPT APPROXIMATIONS FOR CONSTRAINED CLUSTERING PROBLEMS

DISHANT GOYAL

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

MARCH 2023

FPT APPROXIMATIONS FOR CONSTRAINED CLUSTERING PROBLEMS

by

DISHANT GOYAL

Department of Computer Science and Engineering

Submitted

in fulfillment of the requirements of the degree of **Doctor of Philosophy**

to the

Indian Institute of Technology Delhi

March 2023

Certificate

This is to certify that the thesis titled **FPT APPROXIMATIONS FOR CONSTRAINED CLUSTERING PROBLEMS** being submitted by **Mr. DISHANT GOYAL** for the award of **Doctor of Philosophy** in **Computer Science and Engineering** is a record of bona fide work carried out by him under my guidance and supervision at the **Department of Computer Science and Engineering, Indian Institute of Technology Delhi**. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

Ragesh Jaiswal

Associate Professor

Department of Computer Science and Engineering

Indian Institute of Technology Delhi

New Delhi- 110016

Amit Kumar

Professor

Department of Computer Science and Engineering

Indian Institute of Technology Delhi

New Delhi- 110016

Acknowledgements

I am highly indebted to my advisors Prof. Ragesh Jaiswal and Prof. Amit Kumar, for their constant support and guidance during my PhD and my life in general. I am highly thankful to them for believing in me and encouraging me. A very special thanks to Prof. Ragesh Jaiswal for tolerating my grammar mistakes. My heartfelt gratitude to Prof. Ragesh Jaiswal for keeping a constant check on my well-being all this time.

A very special thanks to Prof. Anup Bhattacharya for his invaluable encouragement and discussion sessions. I am also grateful to my research committee members Prof. Naveen Garg, Prof. Bhawani Sankar Panda, and Prof. Preeti Ranjan Panda, for their valuable and constructive suggestions during the development of this research work. I am deeply fortunate to have been taught and guided by Prof. Venkatesh Raman during my B.Tech. at IIT Jodhpur.

I extend my gratitude to Tata Consultancy Services (TCS) for their support through the TCS Research Scholar Program Fellowship. I am also thankful to the theoretical computer science stack exchange community for the invaluable discussions on the clustering problems.

A special thanks to my SIT 309 friends – Vinayak Gupta, Sandeep Kumar, Arindam Bhattacharya, Dilpreet Kaur, Omais Shafi, and Ovia Seshadri for their constant support via poker nights, cricket games, food hunting trips, and innumerable coffee breaks. My heartfelt thanks to Vinayak Gupta for his realistic life advises. I am also thankful to have immense support from Neetu Jindal during the last phase of my degree. Lastly, I thank my parents and relatives for

standing by me during the course of this degree. I thank my sister for being there with me in all my good and bad times. This thesis is dedicated to my sister – Neha! :)

Dishant Goyal

Abstract

In this work, we study a wide range of *constrained* clustering problems in offline and streaming settings. We study these problems corresponding to three clustering objectives: k-median, k-means, and k-supplier. The (unconstrained) k-median problem is defined as follows. We are given a set of clients C in a metric space $\mathcal{X}$, with distance function $d(.,.)$. We are also given a set of feasible facility locations $L \subseteq \mathcal{X}$. The goal is to open a set $F \subseteq L$ of k facilities that minimizes the objective function: cost$(F, C) \equiv \sum_{j \in C} d(F, j)$, where $d(F, j)$ is the distance of client j to the closest facility in F. The k-means problem is defined in similar manner by replacing the distances with squared distances in the cost function, i.e., cost$(F, C) \equiv \sum_{j \in C} d(F, j)^2$. On the other hand, the k-supplier objective is defined as: cost$(F, C) \equiv \max_{j \in C} \{d(F, j)\}$. Furthermore, for $L = C$, the k-supplier problem is known as the k-center problem.

In many applications, there are additional constraints imposed on the clusters. For example, to balance the load among the facilities in resource allocation problems, a capacity u is imposed on every cluster. That is, no more than u clients can be assigned to any facility/cluster. This problem is known as the *capacitated* clustering problem. Likewise, various other applications have different constraints, which give rise to different *constrained* versions of the problem. In the past, the constrained versions of clustering problems were studied separately as independent problems. Recently, Ding and Xu [72] gave a unified framework for these problems that they called the *constrained clustering* framework. They proposed this framework in the context

of the k-median and k-means objectives in the continuous Euclidean space where $L = \mathbb{R}^p$ (p-dimensional Euclidean space) and C is a finite subset of $\mathbb{R}^p$. In this work, we extend this framework to the k-supplier objective and general metric spaces. The unified framework allows us to obtain results simultaneously for the following constrained versions of the problem: r-gather, r-capacity, balanced, chromatic, fault-tolerant, strongly private, ℓ-diversity, and fair clustering problems. We also study the *outlier* versions of these problems. In the outlier version, a clustering is obtained over at least $|C| - m$ clients instead of the entire client set.

For the constrained k-supplier and k-center problems, we obtain the following results:

(1) We give 3 and 2 approximation algorithms for the constrained k-supplier and k-center problems, respectively, with FPT (fixed-parameter tractable) running time $k^{O(k)} \cdot n^{O(1)}$, where $n = |C \cup L|$. Moreover, we note that the obtained approximation guarantees are tight. That is, for any constant $\varepsilon > 0$, no algorithm can achieve $(3-\varepsilon)$ and $(2-\varepsilon)$ approximation guarantees for the constrained k-supplier and k-center problems, respectively, in FPT time parameterized by k, assuming FPT $\neq$ W[2].

(2) For the outlier versions of the constrained k-supplier and k-center problems, we give 3 and 2 approximation guarantees with FPT running time $(k + m)^{O(k)} \cdot n^{O(1)}$, where $n = |C \cup L|$ and m is the number of outliers. Moreover, we note that the obtained approximation guarantees are tight. That is, for any constant $\varepsilon > 0$, no algorithm can achieve $(3 - \varepsilon)$ and $(2 - \varepsilon)$ approximation guarantees for the constrained k-supplier and k-center problems, respectively, in FPT time parameterized by k and m, assuming FPT $\neq$ W[2].

For the constrained k-median and k-means problems, we obtain the following results:

(3) We give $(3 + \varepsilon)$ and $(9 + \varepsilon)$ approximation algorithms for the constrained k-median and k-means problems, respectively, with FPT running time $(k/\varepsilon)^{O(k)} \cdot n^{O(1)}$, where

$n = |C \cup L|$. For the outlier version of the constrained k-median and k-means problems, we give $(3+\varepsilon)$ and $(9+\varepsilon)$ approximation algorithms, respectively, with FPT running time $\left((k+m)/\varepsilon\right)^{O(k)} \cdot n^{O(1)}$, where $n = |C \cup L|$ and m is the number of outliers.

(4) We also study the problems when $C \subseteq L$, i.e., a facility can be opened at a client location as well. For this special case, we design $(2+\varepsilon)$ and $(4+\varepsilon)$-approximation algorithms for the constrained k-median and k-means problems, respectively, with FPT running time $(k/\varepsilon)^{O(k)} \cdot n^{O(1)}$, where $n = |L|$. For the outlier version, we obtain the same approximation guarantees with FPT running time $\left((k+m)/\varepsilon\right)^{O(k)} \cdot n^{O(1)}$, where $n = |L|$ and m is the number of outliers. Note that the case $C \subseteq L$ subsumes the case $C = L$. Therefore, this result also holds for the case when $C = L$.

(5) We show that the analysis of our algorithm is tight. That is, there are instances for which our algorithm does not provide better than $(3-\delta)$ and $(9-\delta)$ approximation guarantee corresponding to k-median and k-means objectives, respectively, for any arbitrarily small constant $\delta > 0$. Similarly, the analysis of our algorithm is tight for the special case $C \subseteq L$.

(6) Our algorithms are based on a simple sampling-based approach. This approach allows us to convert these algorithms to constant-pass log-space streaming algorithms.

(7) We also study the constrained k-median/means problem in continuous Euclidean space where $L = \mathbb{R}^p$ and C is a finite subset of $\mathbb{R}^p$. We design $(1+\varepsilon)$-approximation algorithm for the outlier version of these problems with FPT running time $O\left(np \cdot ((k+m)/\varepsilon)^{O(k/\varepsilon^{O(1)})}\right)$, where $n = |C|$ and m is the number of outliers. We also convert these algorithms to constant-pass log-space streaming algorithms.

We also study the *socially fair k-median/k-means problem*, which is a generalization of the k-supplier and k-median/means problems. The problem is defined as follows. We are given a set of clients C in a metric space $\mathcal{X}$ with a distance function $d(.,.)$. There are ℓ groups:

$C_1, \ldots, C_\ell \subseteq C$. We are also given a set L of feasible centers in $\mathcal{X}$. The goal in the socially fair k-median problem is to find a set $F \subseteq L$ of k centers that minimizes the maximum average cost over all the groups. That is, find F that minimizes the objective function: fair-cost$(F, C) \equiv \max_j \left\{ \sum_{x \in C_j} d(F, x)/|C_j| \right\}$, where $d(F, x)$ is the distance of x to the closest center in F. The socially fair k-means problem is defined similarly by using squared distances, i.e., $d^2(., .)$ instead of $d(., .)$. We obtain the following results for this problem:

(8) We design $(3+\varepsilon)$ and $(9+\varepsilon)$ approximation algorithms for the socially fair k-median and k-means problems, respectively, in FPT time $f(k, \varepsilon) \cdot n^{O(1)}$, where $f(k, \varepsilon) = (k/\varepsilon)^{O(k)}$ and $n = |C \cup L|$.

(9) Furthermore, these approximation guarantees are tight; that is, for any constant $\varepsilon > 0$, no algorithm can achieve $(3 - \varepsilon)$ and $(9 - \varepsilon)$ approximation guarantees for the socially fair k-median and k-means problems in FPT time parametrized by k, assuming FPT $\neq$ W[2].

Lastly, we give hardness of approximation result for the k-median problem in the continuous Euclidean space where $L = \mathbb{R}^p$ and C is a finite subset of $\mathbb{R}^p$. This solves an open problem posed explicitly in the work of Awasthi *et al.* [19]. More precisely, we obtain the following result:

(10) There exists a constant $\varepsilon > 0$ such that the Euclidean k-median problem in $O(\log k)$ dimensional space cannot be approximated to a factor better than $(1 + \varepsilon)$, assuming the Unique Games Conjecture.

Furthermore, we study the hardness of approximation for the Euclidean k-means/k-median problems in the *bi-criteria setting*. In the bi-criteria setting, algorithms are allowed to output βk centers (for some constant $\beta > 1$), and the approximation ratio is computed with respect to the optimal k-means/k-median cost. We show the following results:

(11) For any constant $1 < \beta < 1.015$, there exists a constant $\varepsilon > 0$ such that there is no $(1 + \varepsilon)$ bi-criteria approximation algorithm for the Euclidean k-median problem in $O(\log k)$ dimensional space assuming the Unique Games Conjecture.

(12) For any constant $1 < \beta < 1.28$, there exists a constant $\varepsilon > 0$ such that there is no $(1 + \varepsilon)$ bi-criteria approximation algorithm for the Euclidean k-means problem in $O(\log k)$ dimensional space assuming the Unique Games Conjecture.

सार

इस काम में, हम ऑफ़लाइन और स्ट्रीमिंग समायोजन में बाध्य क्लस्टरिंग समस्याओं की एक विस्तृत श्रृंखला का अध्ययन करते हैं। हम तीन क्लस्टरिंग उद्देश्यों के अनुरूप इन समस्याओं का अध्ययन करते हैं: k-माध्यिका, k-माध्य, और k-आपूर्तिकर्ता। (अप्रतिबंधित) k-माध्यिका समस्या को निम्नानुसार परिभाषित किया गया है। हमें मेट्रिक जगह X में ग्राहक C का एक सेट दिया गया है, जिसमें डिस्टेंस फंक्शन d(., .) है। हमें सुविधा स्थानों का एक सेट L ⊆ X भी दिया जाता है। हमें लक्ष्य के सुविधाओं का एक सेट F⊆ L खोलना है जो उद्देश्य लागत को कम करता है: लागत (F, C) = ∑j∈C d(F, j), जहां d(F, j) दूरी है ग्राहक j कि F में निकटतम सुविधा के लिए। k-माध्य समस्या को कुछ इसी तरह परिभाषित किया गया है- लागत फलन में दूरियों की जगह वर्ग दूरियों का प्रयोग किया जाता है, अर्थात् d(F,j) की जगह d(F,j)^2। दूसरी ओर, k-आपूर्तिकर्ता उद्देश्य निम्नानुसार परिभाषित परिभाषित किया गया है: लागत(F, C) = max_j∈C {d(F, j)}। इसके अलावा, L = C के लिए, k-आपूर्तिकर्ता समस्या को k-केंद्र समस्या के रूप में जाना जाता है।

कई अनुप्रयोगों में, क्लस्टर पर अतिरिक्त प्रतिबंध लगाए गए हैं। उदाहरण के लिए, संमाध्य आवंटन समस्याओं में सुविधाओं के बीच भार को संतुलित करने के लिये, एक क्षमता हर क्लस्टर पर थोपा गया है। इस समस्या को कैपेसिटेटेड क्लस्टरिंग समस्या के रूप में जाना जाता है। इसी तरह, विभिन्न अन्य अनुप्रयोग, अलग-अलग बाधाएँ लगाते हैं, जो समस्या के विभिन्न बाध्य संस्करणों को जन्म देती हैं। अतीत में, क्लस्टरिंग समस्याओं के बाध्य संस्करणों का स्वतंत्र रूप से अलग से अध्ययन किया गया था। हाल ही में, डिंग और जू [68] ने इन समस्याओं के लिए एक एकीकृत ढांचा दिया जोकि बाध्य क्लस्टरिंग ढांचा कहा जाता है। उन्होंने इस ढांचे को यूक्लिडियन अंतरिक्ष में k-माध्यिका और k- माध्य उद्देश्यों में प्रस्तावित किया जहां L = R^p (p-डिमेंशनल यूक्लिडियन स्पेस) और C एक परिमित उपसमुच्चय है R^p का। इस काम में, हम इसका विस्तार करते हैं k-आपूर्तिकर्ता उद्देश्य और सामान्य मीट्रिक स्थान के लिए। एकीकृत ढांचा समस्या के निम्नलिखित बाध्य संस्करणों के लिए एक साथ परिणाम प्राप्त करने की अनुमति देता है: r-इकट्ठा, r-क्षमता, संतुलित, रंगीन, दोष-सहिष्णु, दृढ़ता से निजी, ℓ-विविधता, और निष्पक्ष क्लस्टरिंग समस्याएं। हम इन समस्याओं के बाहरी

संस्करणों का भी अध्ययन करते हैं। बाहरी संस्करण में, एक क्लस्टरिंग में कम से कम |C| - m ग्राहकों की प्राप्त की जाती है बजाय संपूर्ण ग्राहक सेट के। बाध्य k-आपूर्तिकर्ता और k- केंद्र समस्याओं के लिए, हम निम्नलिखित परिणाम प्राप्त करते हैं:

(1) हम बाध्य k-आपूर्तिकर्ता और k- केंद्र समस्याओं के लिए 3 और 2 सन्निकटन एल्गोरिदम देते हैं तथा FPT (फिक्स्ड-पैरामीटर ट्रैक्टेबल) चलने के समय O(f(k))·n^O(1) के साथ, जहां n = |C ∪L|। इसके अलावा, ध्यान दें कि प्राप्त सन्निकटन गारंटी तंग हैं। अर्थात्, किसी भी स्थिरांक ε > 0 के लिए, कोई भी एल्गोरिथ्म (3−ε) और (2−ε) अनुकरण गारंटी प्राप्त नहीं कर सकता है- बाध्य k-आपूर्तिकर्ता और k- केंद्र समस्याओं के लिए FPT समय में, FPT ≠ W[2] मानकर।

(2) बाध्य k-आपूर्तिकर्ता और k-केंद्र समस्याओं के बाहरी संस्करणों के लिए, हम 3 और 2 सन्निकटन गारंटी देते हैं, FPT चलने के समय (k + m)^O(k) · n^O(1) के साथ, जहां n = |C ∪L| और m बाहरी ग्राहकों (आउटलेर्स) की संख्या है। इसके अलावा, ध्यान दें कि प्राप्त सन्निकटन गारंटी तंग हैं। अर्थात्, किसी भी स्थिरांक ε> 0 के लिए, कोई एल्गोरिदम (3 - ε) और (2 - ε) सन्निकटन गारंटी प्राप्त नहीं कर सकता है सीमित k-आपूर्तिकर्ता और k- केंद्र समस्याओं के लिए, FPT समय में FPT ≠ W[2] मानकर।

बाध्य k-माध्यिका और k-माध्य समस्याओं के लिए, हम निम्नलिखित परिणाम प्राप्त करते हैं:

(3) हम बाध्य k-माध्यिका और k-माध्य समस्याएं के लिए (3 + ε) और (9 +ε) सन्निकटन एल्गोरिदम देते हैं, (k/ε)^O(k)·n^O(1) FPT चलने के समय के साथ, जहां n = |C ∪L|। बाध्य k-माध्यिका और k-माध्य समस्याओं के बाहरी संस्करण के लिए, हम (3 + ε) और (9 + ε) सन्निकटन एल्गोरिदम देते हैं, ((k + m)/ε)O(k) · n^O(1) FPT चलने के समय के साथ, जहां n = |C ∪ L| और m बाहरी ग्राहकों (आउटलेर्स) की संख्या है।

(4) हम उन समस्याओं का भी अध्ययन करते हैं जब C ⊆ L, यानी, ग्राहक स्थान पर एक सुविधा भी खोली जा सकती है। इस विशेष मामले के लिए, हम (2 + ε) और (4 + ε) - सन्निकटन एल्गोरिदम देते हैं, बाध्य

k-माध्यिका और k-माध्य समस्याओं के लिए, (k/ε)^O(k) · n^O(1) FPT चलने के समय के साथ, जहां n = |L|. बाहरी संस्करण के लिए, हम वही सन्निकटन गारंटी प्राप्त करते हैं, FPT चलने के समय के साथ ((k + m)/ε)^O(k) · n^O(1), जहां n = |L| और m बाहरी ग्राहकों (आउटलेर्स) की संख्या है। ध्यान दें कि स्थिति C ⊆ L, स्थिति C = L को समाहित करती है। इसलिए, यह परिणाम उस स्थिति के लिए भी मान्य है जब C = L.

(5) हम दिखाते हैं कि हमारे एल्गोरिथ्म का विश्लेषण कड़ा है। अर्थात्, ऐसे उदाहरण हैं जिनके लिए हमारा एल्गोरिथ्म k-माध्यिका और k-माध्य उद्देश्यों के अनुरूप (3 - ε) और (9 - ε) सन्निकटन गारंटी से बेहतर प्रदान नहीं करता है किसी भी मनमाने ढंग से छोटे ε > 0 के लिए. इसी तरह, विशेष मामले के लिए हमारे एल्गोरिथ्म का विश्लेषण तंग है।

(6) हमारे एल्गोरिदम एक साधारण नमूना-आधारित दृष्टिकोण पर आधारित हैं। यह दृष्टिकोण इन एल्गोरिदम को निरंतर-पास लॉग-स्पेस स्ट्रीमिंग एल्गोरिदम में बदलने के लिए हमें अनुमति देता है।

(7) हम निरंतर यूक्लिडियन अंतरिक्ष में बाध्य k-माध्यिका/माध्य समस्या का भी अध्ययन करते हैं जहां L = R^p और C एक परिमित उपसमुच्चय है R^p का। हम इन समस्याओं के बाहरी संस्करण के लिए (1 + ε) - सन्निकटन एल्गोरिथ्म O(np· ((k + m)/ε)^O(k/ε^O(1))) FPT चलने के समय के साथ रचना करते हैं, जहां n = |C| और m बाहरी ग्राहकों (आउटलेर्स) की संख्या है। हम इन एल्गोरिदम को निरंतर-पास लॉग-स्पेस स्ट्रीमिंग एल्गोरिदम में भी परिवर्तित करते हैं।

हम सामाजिक रूप से निष्पक्ष k-माध्यिका/k-माध्य समस्या का भी अध्ययन करते हैं जिसे निम्नानुसार परिभाषित किया गया है। हमें मेट्रिक स्पेस X में ग्राहक C का एक सेट, डिस्टेंस फंक्शन d(., .) के साथ दिया गया है। समूह हैं: C1, . . . , Cℓ . हमें X में व्यवहार्य केंद्रों का एक सेट L भी दिया गया है। सामाजिक रूप से लक्ष्य निष्पक्ष k-माध्यिका समस्या के केंद्रों का एक सेट F ⊆ L ढूंढना है जो अधिकतम समूहों पर औसत लागत को कम करता है। यही है, F खोजें जो उद्देश्य लागत को कम करता है: लागत (F, C) = max{ x∈C_j d(F,x)/|C_j |} , जहां d(F, x), F में निकटतम केंद्र से x की दूरी है। सामाजिक रूप से निष्पक्ष k- माध्य समस्या को समान रूप से

वर्ग दूरी का उपयोग करके परिभाषित किया जाता है, अर्थात, d(।.।) के बजाय d(|.|)^2। हम इस समस्या के लिए निम्नलिखित परिणाम प्राप्त करते हैं:

(8) हम सामाजिक रूप से निष्पक्ष k-माध्यिका और k-माध्य के लिए (3+ε) और (9+ε) सन्निकटन एल्गोरिदम डिज़ाइन करते हैं, FPT समय में f (k,ε).n^O(1), जहाँ f(k,ε) = (k/ε)^O(k) और n = |C ∪ L|।

(9) इसके अलावा, ये सन्निकटन गारंटी तंग हैं; अर्थात्, किसी अचर ε > 0 के लिए, एल्गोरिदम सामाजिक रूप से निष्पक्ष k-माध्यिका और k-माध्य के लिए (3 -ε) और (9 - ε) सन्निकटन गारंटी प्राप्त नहीं कर सकता है, FPT= W[2] मानते हुए।

आखिर में हम k-माध्यिका समस्या के लिए सन्निकटन परिणाम की कठोरता देते हैं, यूक्लिडियन अंतरिक्ष में जहां L= R^p और C ⊆ R^p। यह एक खुली समस्या का समाधान करता है जो अवस्थी आदि [18] के कार्यों में स्पष्ट रूप से प्रस्तुत किया गया है। हम निम्नलिखित प्राप्त करते हैं:

(10) एक स्थिर ε> 0 मौजूद है, जैसे कि O(log k) यूक्लिडियन k-माध्यिका समस्या को (1 + ε) से बेहतर कारक के रूप में अनुमानित नहीं किया जा सकता है, अद्वितीय खेल अनुमान (UGC) मानते हुए।

इसके अलावा, हम यूक्लिडियन k-माध्य/k-माध्यिका के लिए सन्निकटन की कठोरता का अध्ययन करते हैं, द्वि-मानदंड सेटिंग में। द्वि-मानदंड सेटिंग में, एल्गोरिदम को βk केंद्र (कुछ स्थिर β> 1 के लिए) आउटपुट करने की अनुमति है और सन्निकटन अनुपात की गणना इष्टतम k-माध्य/k-माध्यिका लागत के संबंध में की जाती है। हम निम्नलिखित परिणाम दिखाते हैं:

(11) किसी भी स्थिरांक $1 < \beta < 1.015$ के लिए, एक अचर $\varepsilon > 0$ मौजूद है जिस्से कि कोई $(1+\varepsilon)$ द्वि-मानदंड सन्निकटन एल्गोरिथम नहीं मौजूद हो सकता है, $O(\log k)$ यूक्लिडियन k-माध्यिका समस्या के लिए, अद्वितीय खेल अनुमान (UGC) मानते हुए।

(12) किसी भी अचर $1 < \beta < 1.28$ के लिए, एक अचर $\varepsilon > 0$ मौजूद है जिस्से कि कोई $(1+\varepsilon)$ द्वि-मानदंड सन्निकटन एल्गोरिथम नहीं मौजूद हो सकता है, $O(\log k)$ यूक्लिडियन k-माध्यिका समस्या के लिए, अद्वितीय खेल अनुमान (UGC) मानते हुए।

Contents

List of Figures

List of Tables