ARTICLE IN PRESS

Discrete Optimization ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Discrete Optimization

www.elsevier.com/locate/disopt

On the approximability of the maximum interval constrained coloring problem

Stefan Canzar[a], Khaled Elbassioni[b,*], Amr Elmasry[c], Rajiv Raman[d]

[a] *Gene Center, Ludwig-Maximilians-Universität München, Munich, Germany*
[b] *Masdar Institute (part of Khalifa University of Science and Technology), Abu Dhabi, United Arab Emirates*
[c] *Alexandria University, Egypt*
[d] *Indraprastha Institute of Information Technology, Delhi, India*

ARTICLE INFO

Article history:
Received 2 September 2016
Received in revised form 6 September 2017
Accepted 6 September 2017
Available online xxxx

Keywords:
Approximation algorithms
APX-hardness
dynamic programming
protein structure
Interval constrained coloring
partially ordered set

ABSTRACT

In the Maximum Interval Constrained Coloring problem, we are given a set of vertices and a set of intervals on a line and a k-dimensional requirement vector for each interval, specifying how many vertices of each of k colors should appear in the interval. The objective is to color the vertices of the line with k colors so as to maximize the total weight of intervals for which the requirement is satisfied. This $\mathcal{NP}$-hard combinatorial problem arises in the interpretation of data on protein structure emanating from experiments based on hydrogen/deuterium exchange and mass spectrometry. For constant k, we give a factor $O(\sqrt{|\textsc{Opt}|})$-approximation algorithm, where Opt is the *smallest cardinality* maximum-weight solution. We show further that, even for $k = 2$, the problem remains APX-hard.

1. Introduction

The Interval Constrained Coloring (Icc) problem was introduced by Althaus et al. [2,3] as the mathematical abstraction of a problem appearing in the interpretation of experimental data in biochemistry. Monitoring exchange rates via mass spectrometry is a method used to obtain information about the 3-dimensional structure of proteins. ICC captures the problem of increasing the resolution of the exchange data from peptic fragments to single residues. We refer the interested reader to Althaus et al. [3] for more on the biochemical background. Interval Constrained Coloring is a decision problem that asks for a coloring of an ordered sequence of n vertices $V = [n]$ using k colors such that a given set of requirements is satisfied. Each requirement is made up of a closed interval $I \subseteq [n]$, and a complete specification of how many elements in I should be colored with each color.

* Corresponding author.
E-mail addresses: canzar@genzentrum.lmu.de (S. Canzar), kelbassioni@masdar.ac.ae (K. Elbassioni), elmasry@alexu.edu.eg (A. Elmasry), rajiv@iiitd.ac.in (R. Raman).

http://dx.doi.org/10.1016/j.disopt.2017.09.002
1572-5286/

More formally, let $\mathcal{I}$ be a set of m intervals defined on $V = [n]$, let $[k]$ be a set of color classes, and $r : \mathcal{I} \times [k] \to \mathbb{Z}_+$ be a requirement function such that

$$\sum_{c \in [k]} r(I, c) = |I| \quad \text{for all } I \in \mathcal{I}. \tag{1}$$

The given set of intervals $\mathcal{I}$ is said to be *feasible* (or *colorable*) if there exists a coloring $\chi : V \to [k]$ such that for every $I \in \mathcal{I}$, we have

$$N_\chi(I, c) = r(I, c), \text{ for all } c \in [k], \tag{2}$$

where $N_\chi(I, c) = |\{i \in I : \chi(i) = c\}|$ denotes the number of vertices in I assigned color c by χ. Clearly, not all sets of intervals are colorable.

In the biochemical application from which our problem arises, the requirement function models exchange data collected in real experiments which usually contain some noise. Hence, we might not obtain a colorable instance, even if the real underlying instance is colorable. This motivates the study of the problem of extracting the largest subset $\mathcal{I}' \subseteq \mathcal{I}$ that is colorable. More precisely, we consider a more general version of the problem where each interval also has a weight, in addition to its color requirements, given by $w : \mathcal{I} \to \mathbb{R}_+$, and we wish to find a maximum weight colorable subset of the intervals and to produce a feasible coloring for this subset of intervals. We define the problem formally below.

Definition 1 (Max-Feasible-Coloring *(*Mfc*)*)**.** Given a set of intervals $\mathcal{I}$ with non-negative weights $w : \mathcal{I} \to \mathbb{R}_+$, a requirement function $r : \mathcal{I} \times [k] \to \mathbb{Z}^+$ satisfying (1), problem Max-Feasible-Coloring (Mfc) is to find a maximum weight colorable subset $\mathcal{I}' \subseteq \mathcal{I}$.

The Mfc problem can also be cast as a problem on linear systems, as observed by Byrka et al. [4]. A 0/1 matrix A is said to have the *consecutive* 1*s property*, if there is an ordering of the columns of the matrix A such that in each row of A, the 1s appear consecutively in that order. Consider the following system, where $A \in \{0, 1\}^{m \times n}$ is the consecutive 1s matrix derived in the natural manner from the Mfc problem (where the order on the columns corresponds to the order of points on the line), and $r^i \in \mathbb{Z}_+^m$ is a column vector of requirements for the ith color from the m intervals, and let I be the $n \times n$ identity matrix:

$$\begin{bmatrix} A & 0 & \dots & 0 \\ 0 & A & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & & A \\ I & \dots & & I \end{bmatrix} \begin{bmatrix} x^1 \\ x^2 \\ x^3 \\ \vdots \\ x^k \end{bmatrix} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^k \\ 1 \end{bmatrix}.$$

The Interval Constrained Coloring problem asks for a feasible solution to the above system with $x^i \in \{0, 1\}^n$, $i = 1, \dots, k$. The Mfc problem can now be stated as the problem of computing the maximum subset of rows of the matrix A for which the above system has a feasible solution. In this light, the problem is similar to the Maximum Feasible Subsystem problem (Mfs), where the problem is given by an infeasible linear system $Ax = b$, and we wish to find the largest subsystem that has a feasible solution. However, in the Maximum Feasible Subsystem problem, we are usually allowed a rational vector x; see also [5] for a version of Mfs with solution restricted to be 0/1.

1.1. Our results

We study the approximability of Mfc and present the first non-trivial approximation algorithm for the problem. In particular, we give a factor $O(\sqrt{|\text{Opt}|})$-approximation algorithm, where we denote throughout