Unified New Techniques for NP-Hard Budgeted Problems with Applications in Team Collaboration, Pattern Recognition, Document Summarization, Community Detection and Imaging

Dorit S. Hochbaum[a]
Department of Industrial Engineering and Operations Research, University of California, Berkeley, U.S.A.

Keywords: Parametric Flow, Maximum Diversity, Quadratic Knapsack, Efficient Frontier, Text Summarization.

Abstract: This paper introduces new techniques for any NP-hard problems formulated as monotone integer programming (IPM) with a budget constraint "budgeted IPM". Problems of this type have diverse applications, including maximizing team collaboration, the maximum diversity problem, facility dispersion, threat detection, minimizing conductance, clustering, and pattern recognition.

We present a unified framework for effective algorithms for budgeted IPM problems based on the Langrangian relaxation of the budget constraint. It is shown that all optimal solutions for all values of the Lagrange multiplier are generated very efficiently, and the piecewise linear *concave envelope* (convex, for minimization problems) of these solutions has breakpoints that are optimal solutions for the respective budgets. This is used to derive high quality upper and lower bounds for budgets that do not correspond to breakpoints. We show that for all these problems, the weight "perturbation" concept, that was successful for the problem of maximum diversity in enhancing the number and distribution of breakpoints, is applicable. Furthermore, the insights derived from this efficient frontier of solutions, lead to the result that all the respective ratio problems have a solution at the "first" breakpoint, which generalizes the concept of maximum density subgraph.

1 INTRODUCTION

We explore here NP-hard problems that can be formulated as monotone integer programs with an additional budget constraint. Monotone integer programming problems (IPM) are solvable in polynomial time as a minimum cut on an associated graph, (Hochbaum, 2002). The addition of the budget constraint renders these problems NP-hard. Among such problems are the quadratic knapsack, the minimum conductance problem, the facility dispersion problem, the ratio problem, image segmentation problems and others. The relationship of IPM problems to the minimum cut problem is crucial, as it allows to solve the relaxed budget constraint problem parametrically, and very efficiently, and to generate the concave envelope which has desirable properties: that the breakpoints give optimal solutions and that the solutions at the sequence of breakpoints are *nested*.

Special cases of IPM with budget constraint, the maximum diversity and the facility dispersion problems, were recently addressed in (Hochbaum et al., 2023), by solving efficiently the Lagrangean relaxation of the budget constraint for all values of the Lagrange multiplier. The function of all solutions for each budget value has an upper envelope of the intersection of all lines that lie above all solutions, which is concave, piecewise linear. We refer to this piecewise linear function that maps the budget to an upper bound on the value of the objective, as the *concave envelope*. Because IPM problems are minimum cut problems, the breakpoints in the concave envelope correspond to solutions that are optimal and *nested*. This concave envelope provides an upper bound for the solutions for each budget, but is also used to generate high quality feasible solutions as lower bounds (for maximization). The feasible solutions are generated with the *breakpoints algorithm* that utilizes breakpoints with close budget values to the value in the input, to append or remove elements, using a fast heuristic, (Hochbaum et al., 2023). Since increased density of the breakpoints enhances the quality of the solutions, it was shown that a *perturbation* method on the utility values lead to tighter lower bounds (for maximization). This resulted in the ability to solve the problem, even for small budget values, which are

[a] https://orcid.org/0000-0002-2498-0512

Hochbaum, D.
Unified New Techniques for NP-Hard Budgeted Problems with Applications in Team Collaboration, Pattern Recognition, Document Summarization, Community Detection and Imaging.
DOI: 10.5220/0012207200003598
In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023) - Volume 1: KDIR*, pages 365-372
ISBN: 978-989-758-671-2; ISSN: 2184-3228

particularly challenging, either to optimality, or very close to optimality. All this is done within a small fraction of the running time required by competing approaches, such as an integer programming solver, or by state-of-the-art meta-heuristics.

The Lagrangean relaxation of the budget constraint has been previously used in the context of the quadratic knapsack problem, most recently by (Spiers et al., 2023). The quadratic knapsack, which is identical to the maximum diversity and maximum facility dispersion problems, can be formulated as IPM with a budget constraint. However, the methods developed for the quadratic knapsack were ad-hoc and suitable for this case only, and the general structure has not been recognized up till now. In addition, all the literature that uses this approach, going back to (Chaillou et al., 1989; Pisinger, 2006) can solve the problems to optimality for at most a few hundred variables, and for selected, relatively high, values of the budget (accommodating 10%-50% of the variables), whereas the harder problems, that are more prevalent in applications, are for smaller budgets. In addition, the running times of the current approaches do not scale well. Indeed in the reported results of (Spiers et al., 2023) the largest instances contain up to 2000 nodes of the GKD benchmark. These instances were shown to be particularly easy, in (Hochbaum et al., 2023). In contrast, in (Hochbaum et al., 2023), new insights as to how to deal with harder problems, with perturbation, were able to provide, often optimal, or very close to optimal, solutions, within a tiny fraction of the running time of competing approaches, including integer programming software.

Here we demonstrate that the breakpoints algorithm is applicable to a vast collection of hard problems, with the potential of providing optimal or very close to optimal, solutions One example explored here is the *text summarization* problem, aka *multi-document summarization*, which is modeled, under the MMR criterion, as combining one goal of maximizing the sum of dissimilarities in the selected set (of sentences), to enhance the diversity of the selected set and eliminate redundancy, with the second goal maximizing the similarities between the selected set and its complement. This combined objective is NP-hard to solve even without the budget constraint on the total size of the sentences selected. A straightforward formulation of this optimization problem, given in (Lin and Bilmes, 2010), is reported to be solved for an instance of the problem of size 178 sentences, in 17 hours, using an integer programming software. Our approach is to model the problem as budgeted IPM simply by replacing the similarity weights by dis-similarity weights, e.g. by taking the reciprocal of the similarities. Once the problem is modeled as IPM with a budget constraint, the framework presented here can utilize the concave envelope to solve the problem effectively, and with a highly scalable algorithm. This is discussed in detail in Section 3.

There is a close relationship between *ratio problems* and budgeted IPM problems. This relationship is reflected in the concave envelope, where the optimal solution to the respective budget problem is the first breakpoint, for the smallest budget value, that corresponds to a generalization of the *maximum density subgraph* problem.

Our contributions here include:

- Introducing a large class of NP-hard problems that are formulated as monotone integer programming with a budget constraint: *budgeted IPM*.
- Demonstrating that for all budgeted IPM problems the concave envelope (for maximization, convex for minimization) related to the Lagrangean relaxation of the budget constraint is constructed as the output of a (parametric) minimum cut procedure on a respective graph.
- The breakpoints in the concave envelope are shown to be optimal solutions for the respective budget values, and correspond to nested solutions.
- The perturbation concept, of (Hochbaum et al., 2023), applies to the class of budgeted IPM problems, and can increase the number of breakpoints and enhance their distribution around the budget values of interest.
- Relationship of budgeted IPM problems to the respective IPM ratio problems, that are polynomial time solvable with a parametric cut procedure, showing that the first (leftmost for maximization) breakpoint, solves the ratio problem optimally. The first breakpoint is shown to generalize the concept of maximum density subgraph.
- The newly introduced, **procedure incremental-para**, that solves IPM ratio problems, with a given initial feasible solution, in the complexity of a single minimum cut procedure and generates *sequentially* all breakpoints.
- Show how all budgeted IPM problems are amenable to the breakpoints algorithm of (Hochbaum et al., 2023) which bodes well to the chances of being able to use a scalable algorithm that delivers high quality solutions.
- Demonstrating a new formulation for the text summarization problem that renders it a budgeted IPM problem with the potential of new scalable methods for the problem.

The paper is structured as follows: the following Section 2 provides the basic graph notation used in the rest of the paper, the formulation of a general IPM problem, and the construction of the associated graph and several examples of IPM problems and ratio problems that are IPM. Section 3 introduces the text summarization problem in the known form of non-IPM problem, and shows that it can be modeled as a budgeted IPM problem. The key insights to the concave envelope and the implications for budgeted IPM, and ratio IPM problem are presented in Section 4. Section 5 presents a new form of fully parametric cut procedure, that is more efficient than the parametric cut approach in that it only computes the solutions at the breakpoints. Section 6 includes conclusions and pointers to future research.

It is noted that although most of the results are presented in the setting of maximization problems, they all apply analogously to minimization problems as well.

2 NOTATION AND IPM PROBLEMS

Firstly, we introduce graph notation used here: Let the input graph be an undirected graph $G = (V,E)$ with the weights of the edges be w_{ij} for $[i,j] \in E$. A bipartition of the graph is called a *cut*, $(S,\bar{S}) = \{[i,j] \in E | i \in S, j \in \bar{S}\}$, where $\bar{S} = V \setminus S$. Given two subsets of nodes, $A \subseteq V$, $B \subseteq V$ let the sum of weights of the edges, with one endpoint in A and the other in B be $C(A,B) = \sum_{i\in A, j\in B} w_{ij}$. For the cut $(S,\bar{S})$, the *capacity* of this cut is $C(S,\bar{S})$, and the sum of weights inside the set A is $C(A,A) = \sum_{i,j\in A} w_{ij}$.

If the edges have two sets of weights, these will be denoted by w^1_{ij} and w^2_{ij}. For inputs with two sets of edge weights we let $C_1(A,B) = \sum_{i\in A, j\in B} w^1_{ij}$ and $C_2(A,B) = \sum_{i\in A, j\in B} w^2_{ij}$.

We denote by d_i the weighted degree of node i in G: $d_i = \sum_{j|[i,j]\in E} w_{ij}$. The sum of the weighted degrees of nodes in a set $H \subset V$ is $d(H) = \sum_{i\in H} d_i$, and is also referred to as the *volume* of the set H. We also allow nodes to be associated with arbitrary weights, which are denoted by q_i for $i \in V$, and the sum of weights of nodes in the set H is $q(H) = \sum_{i\in H} q_i$.

Monotone integer programming problems, IPM, also referred to as *Monotone IP3* problems, are integer programming problems on at most 3 variables per constraint, where two of the variables, the x-variables, appear with opposite sign coefficients, and a third variable, a z-variable, if included, can appear in at most one constraint. The coefficient of the third variable, in the objective function, must be non-negative for minimization problems, or non-positive for maximization problems. The formulation of a general maximization (monotone) IP3, for a set of n x-variables, those that can appear in multiple constraints, and a set of constraints involving a collection of pairs of variables A and a respective set of z-variables is

$$\begin{aligned} \text{(IPM)} \quad \max \quad & \sum_{i=1}^{n} w_i x_i - \sum_{(i,j)\in A} u_{ij} z_{ij} \\ \text{s.t.} \quad & a_{ij}x_i - b_{ij}x_j \le c_{ij} + z_{ij} \quad \forall\,(i,j) \in A \\ & \ell_i \le x_i \le u_i, \ \text{integer} \quad \forall\, i \in V \\ & z_{ij} \ge 0, \ \text{integer} \quad \forall\,(i,j) \in A. \end{aligned}$$

Here all a_{ij} and b_{ij} and u_{ij} are all non-negative. A constraint may appear in the form $a_{pq}x_p - b_{pq}x_q \le c_{pq}$, without a third variable. In that case, for the sake of streamlined presentation, we can assume that z_{pq}'s coefficient is $u_{pq} = \infty$.

(Hochbaum, 2002) showed that any IPM can be written as the *maximum s-excess problem*, in $U = \sum_{i=1}^{n}(u_i - \ell_i)$ binary variables. (Solving the problem (IPM) independently of U was proved to be NP-hard.) The *maximum s-excess problem* is formulated as a binary optimization problem, where $x_i = 1$ iff node i is in the optimal set S:

$$\begin{aligned} (s\text{-excess}) \quad \max \quad & \textstyle\sum_{j\in V} w_i x_i - \sum_{(i,j)\in A} u_{ij} z_{ij} \\ \text{subject to} \quad & x_i - x_j \le z_{ij} \quad \text{for } (i,j) \in A \\ & x_j \quad \text{binary } j = 1,\ldots,n \\ & z_{ij} \quad \text{binary } (i,j) \in A. \end{aligned}$$

Although the constraints of the type $x_i - x_j \le 0$ do not appear explicitly in this formulation, these can be written as $x_i - x_j \le z_{ij}$ where the cost coefficient u_{ij} of the respective variable z_{ij} (and the capacity of the corresponding arc) is infinite. We now show that the maximum s-excess set in a graph G is the source set of a minimum cut in an associated graph G_{st}, constructed as follows, (Hochbaum, 2002): We add nodes s and t to the graph G, with an arc from s to every positive weight node i, of capacity $u_{si} = w_i$, and an arc from every negative weight node j to t of capacity $u_{jt} = -w_j$. Let this added set of arcs, adjacent to s and t (source node and sink node respectively) be denoted by A_{st}. The arcs of A each carry the capacity u_{ij}. The graph G_{st} is then $(V \cup \{s,t\}, A \cup A_{st})$.

Lemma 1. *S^* is a set of maximum s-excess capacity in the original graph G if and only if S^* is the source set of a minimum s,t-cut in the associated graph G_{st}.*

Proof. Let $V^+ \equiv \{i \in V | w_i > 0\}$, and let $V^- \equiv \{j \in V | w_j < 0\}$. Let $(s \cup S, t \cup T)$ be a minimum s,t cut on

G_{st}. Then the capacity of this cut is given by

$$\begin{aligned} & C(s\cup S, t\cup T) \\ = & \sum_{(s,i)\in A_s, i\in T} u_{s,i} + \sum_{(j,t)\in A_t, j\in S} u_{j,t} + \sum_{i\in S, j\in T} u_{ij} \\ = & \sum_{i\in T\cap V^+} w_i + \sum_{j\in S\cap V^-} -w_j + \sum_{i\in S,\ j\in T} u_{ij} \\ = & W^+ - [\sum_{j\in S} w_j - \sum_{i\in S,\ j\in T} u_{ij}] \end{aligned}$$

Where $W^+ = \sum_{i\in V^+} w_i$ is the sum of all positive weights in G, which is a constant. Therefore, minimizing $C(s\cup S, t\cup T)$ is equivalent to maximizing $\sum_{j\in S} w_j - \sum_{i\in S,\ j\in T} u_{ij}$, and we conclude that the source set of a minimum s,t cut on G_{st} is also a maximum s-excess set of G.

□

Examples of IPM Problems. For the following problems, the discovery of the IPM model for the problem led to very fast, high quality solutions.

1. The co-segmentation problem: This is to identify the same feature in two distinct images. The problem is modeled as increased similarity between histogram buckets of the foregrounds and decreased binary MRF (Markov Random Field - equivalent to distinctness of the image) in image 1 and in image 2 that determine the foregrounds (co-segmentation), (Hochbaum and Singh, 2009).
2. Identifying nuclear threat alerts regions: The model is to identify a region with increased number and intensity of alerts within; decreased length of region boundary and decreased number of no-alerts within the region, contributing to alert concentration. (Hochbaum and Fishbain, 2011).

There are many ratio problems that are IPM, but were not always recognized as such. As explained in Section 4, a ratio problem is IPM if the "linearized" λ-question is an IPM problem. One such problem, thought to be NP-hard (Sharon et al., 2006), is to find a subset in the graph, that maximizes the ratio of the sum of similarities (edge weights) in the set, divided by the cut capacity separating the set from its complement: $\max_{S\subset V} \frac{C(S,S)}{C(S,\bar{S})}$. This problem was stated in (Sharon et al., 2006) to be equivalent to the normalized cut problem, (Shi and Malik, 2000), which is NP-hard. However, recognizing that the problem is IPM led to a polynomial time algorithm (Hochbaum, 2010). This problem is also called HNC. A variant of HNC, $\max_{S\subset V} \frac{C_1(S,S)}{C_2(S,\bar{S})}$, where the set of weights used in the numerator is different from the set of weights used in the denominator, is also IPM, (Hochbaum, 2010).

Another problem, related to the *conductance problem*, is to minimize the cut separating the set and its complement divided by the "volume" of the set: $\min_{S\subset V} \frac{C(S,\bar{S})}{d(S)}$. This problem is in fact equivalent to HNC as shown in (Hochbaum, 2010). It is also a relaxation of conductance and Cheeger constant, (Cheeger, 1970) where the "budget" constraint $d(S) \le \frac{1}{2} d(V)$ is relaxed.

The well known problem, the *maximum density subgraph* problem, $\max_{S\subset V} \frac{C(S,S)}{d(S)}$, is IPM. We refer to it in its most general form where the nodes can carry arbitrary weights q_i: $\max_{S\subset V} \frac{C(S,S)}{q(S)}$.

In all the above problems, the formulations have binary variables and do not require the *binarization* transformation of a general IPM.

3 TEXT SUMMARIZATION-AS BUDGETED IPM

The *text summarization* problem, aka, the multi-document summarization problem has been studied extensively, with heuristics and approximations. The problem is modeled as the *Maximal Marginal Relevance* (MMR) criterion, introduced in (Carbonell and Goldstein, 1998), which strives to reduce redundancy while maintaining query relevance in re-ranking retrieved documents and in selecting appropriate passages for text summarization. This MMR model was formulated as integer programming in (Lin and Bilmes, 2010), as shown next.

The formulation uses the variables: Let x_i be the binary variable which takes the value 1 if node (sentence) i is selected, and 0 otherwise,

$$x_i = \begin{cases} 1 & \text{if } i \in S \\ 0 & \text{if } i \in \bar{S}. \end{cases}$$

$z_{ij} = 1$ if exactly one of i and j is in S; $y_{ij} = 1$ if both With these variables the following formulation of (MMR) $\max_{S\subset V} C(S,\bar{S}) - \alpha C(S,S)$ was given in (Lin and Bilmes, 2010):

$$\begin{array}{lll} \text{(MMR) max} & \sum_{(i,j)\in E} w_{ij} z_{ij} - \alpha \sum_{[i,j]\in E} w_{ij} y_{ij} & \\ \text{subject to} & x_i - x_j \le z_{ij} & \text{for all } (i,j)\in E \\ & z_{ij} \le x_i & \\ & 1 + z_{ij} \ge x_j & \\ & y_{ij} \le x_i & \text{for all } [i,j]\in E \\ & y_{ij} \le x_j & \\ & x_i + x_j \le 1 - y_{ij} & \\ & x_j, z_{ij}, y_{ij} \text{ binary} . & \end{array}$$

This formulation is not IPM since the "third" variables of some of the constraints, also appear in other constraints. We maintain the spirit of this MMR

objective and change the weights to mean the *dissimilarity* between every pair of nodes. This results in a polynomial time solvable problem, since it isIPM,

$$\text{(MMR*)} \quad \max_{S\subset V} C(S,S) - \beta C(S,\bar{S}).$$

This objective function is $f(\mathbf{x})$ where $x_i = 1$ if sentence i is selected in the "summary" set S. Here β corresponds to $1/\alpha$ in the maximization formulation. The problem formulation as IPM is:

$$\begin{array}{lll} \text{(MMR*)} \ \max & \sum w'_{ij} y_{ij} - \beta \sum w_{ij} z_{ij} & \\ \text{subject to} & x_i - x_j \le z_{ij} & \text{for all } [i,j] \in E \\ & x_j - x_i \le z_{ji} & \text{for all } [i,j] \in E \\ & y_{ij} \le x_i & \text{for all } [i,j] \in E \\ & y_{ij} \le x_j & \\ & x_j, z_{ij}, y_{ij} \text{ binary} . & \end{array}$$

To verify the validity of the formulation notice that the objective function drives the values of z_{ij} to be as small as possible, and the values of y_{ij} to be as large as possible. With the constraints, z_{ij} cannot be 0 unless both endpoints i and j are in the same set. On the other hand y_{ij} cannot be equal to 1 unless both endpoints i and j are in S. In this formulation, both x_i and y_{ij} are the x-variables and there is a node for each in the associated flow graph. Next we show that there is a more compact formulation that has one node in the graph for each candidate sentence.

3.1 A Compact IPM MMR* Formulation

We first need additional notation: Let the input graph be an undirected graph $G = (V,E)$ with the weights of the edges be w_{ij} for $[i,j] \in E$. Let d_i be the weighted degree of node i in G which is the sum of the weights of the edges adjacent to j: $d_i = \sum_{j|[i,j]\in E} w_{ij}$. For a subset of nodes $D \subset V$ let $d(D) = \sum_{i\in D} d_i$ which is also known as the "volume" of the set D.

The compact formulation is presented as the following Lemma:

Lemma 2. *Solving* $\max_{S\subset V} C(S,S) - \beta C(S,\bar{S})$ *is equivalent to solving* $\min_{S\subset V} C(S,\bar{S}) - \frac{1}{1+2\beta} d(S)$.

Proof. Using the equality

$$2C(S,S) + C(S,\bar{S}) = d(S) \qquad (1)$$

$\max_{S\subset V} C(S,S) - \beta C(S,\bar{S}) = \max_{S\subset V} \frac{1}{2} d(S) - \frac{1}{2} C(S,\bar{S}) - \beta C(S,\bar{S})$
$= -\frac{1}{2} \min_{S\subset V} (1+2\beta) C(S,\bar{S}) - d(S)$.

The latter is equivalent to $\min_{S\subset V} C(S,\bar{S}) - \frac{1}{1+2\beta} d(S)$. □

In the formulation of $\min_{S\subset V} C(S,\bar{S}) - \frac{1}{1+2\beta} d(S)$ the variables y_{ij} do not appear. Therefore the number of nodes the respective graph is n, where n is the number of sentences, or x-variables.

4 LINEARIZING RATIOS AND ENVELOPES

We consider a generic IPM problem $\max_{\mathbf{x}\in\mathcal{F}} f(\mathbf{x})$, for $\mathbf{x}$ an integer vector. The entire discussion applies to minimization problems as well. Suppose there is a budget constraint of the form $g(\mathbf{x}) \le B$. The Lagrange relaxation approach is related to minimizing, or maximizing, a fractional (or as it is sometimes called, geometric) objective function over a feasible region $\mathcal{F}$, $\max_{\mathbf{x}\in\mathcal{F}} \frac{f(\mathbf{x})}{g(\mathbf{x})}$. To solve such fractional problem one can reduce it to a sequence of calls to an oracle that provides the yes/no answer to the λ-*question*:
Is there a feasible solution $\mathbf{x} \in \mathcal{F}$ such that $f(\mathbf{x}) - \lambda g(\mathbf{x}) > 0$?
The answer to this question is given by solving the optimization problem

$$\max_{\mathbf{x}\in\mathcal{F}} f(\mathbf{x}) - \lambda g(\mathbf{x}).$$

It is yes if the maximum value is greater than 0, no if the maximum value is less than 0. Suppose the answer to the λ-question, for maximization, is *yes*, then the optimal solution has value greater than λ. Otherwise, the optimal value is less than or equal to λ. If the answer is 0 then the optimum has been found. Note that this maximization problem is still an IPM problem since adding any term that is linear in the x variables, retains the form of the s-excess problem.

The λ-*question* problem is also the Lagrangean relaxation of the budget constraint for the budgeted IPM problem, $\max_{\mathbf{x}\in\mathcal{F}} \{f(\mathbf{x}) | g(\mathbf{x}) \le B\}$.

Since the λ-*question* is formulated as IPM, (to make it simpler assume the variables are binary, though everything applies to the general integer case) the x-variables of the problem correspond to nodes in the associated graph that are set to 1 if they belong to the source set, and 0 if belong to the sink set. Furthermore, the terms that depend on λ appear only in the coefficients of the x variables, and therefore the associated flow graph is *parametric*.

An s,t-graph with source and sink adjacent capacities that depend on a parameter, is said to be a *parametric* flow graph if source adjacent capacities are monotone nondecreasing in λ, and sink adjacent capacities monotone nonincreasing in λ (or vice versa). A parametric flow algorithm solves the maximum flow and minimum cut on a parametric flow

graph, for all values of the parameter, in the complexity of a single max-flow, or min-cut. There are two parametric flow algorithms known that solve the problem, for all values of λ, in strongly polynomial time in the complexity of a single cut, $O(mn\log\frac{n^2}{m})$, for m the number of edges in the graph, and n the number of nodes ((Gallo et al., 1989; Hochbaum, 2008) (the improved run time of the latter is given in (Hochbaum and Orlin, 2013)).

First we show the properties of the minimum cut function of λ in a parametric flow network.

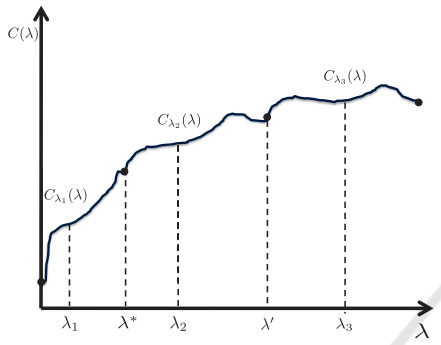

Figure 1: The cut capacity as a function of λ in a parametric cut.

Definition 1. *A function $C(\lambda)$ is* breakpoint-concave *if for $\lambda_1 < \lambda_2$, $C_{\lambda_1}(\lambda) - C_{\lambda_2}(\lambda)$ is a monotone nondecreasing function of λ.*

Recall that in a parametric flow network the source adjacent capacities are monotone nondecreasing and the sink adjacent capacities are monotone non-increasing.

Lemma 1. *In a parametric graph, the cut capacity function $C(\lambda)$ is breakpoint-concave.*

Proof: By the definition, we need to show that for $\lambda_1 < \lambda_2$, $C_{\lambda_1}(\lambda) - C_{\lambda_2}(\lambda)$ is a monotone nondecreasing function of λ.

Let $(\{s\} \cup S_1, \{t\} \cup T_1)$, $(\{s\} \cup S_2, \{t\} \cup T_2)$ be the cuts corresponding to λ_1 and λ_2 respectively. Because of the nestedness property, $S_1 \subseteq S_2$ and $T_1 \supseteq T_2$.

$$
\begin{aligned}
& C_{\lambda_1}(\lambda) - C_{\lambda_2}(\lambda) = \\
& C(S_1, T_1) - C(S_2, T_2) \\
+ \; & C(\{s\}, T_1)(\lambda) - C(\{s\}, T_2)(\lambda) \\
+ \; & C(S_1, \{t\})(\lambda) - C(S_2, \{t\})(\lambda) = \\
& K_{12} + C(\{s\}, T_1 \setminus T_2)(\lambda) - C(S_2 \setminus S_1, \{t\})(\lambda)
\end{aligned}
$$

$K_{1,2}$ is a constant independent of λ. $C(\{s\}, T_1 \setminus T_2)(\lambda)$ is a monotone nondecreasing function, since it is a sum of capacities adjacent to the source, and $C(S_2 \setminus S_1, \{t\})(\lambda)$ is a monotone nonincreasing function, since it is the sum of capacities adjacent to the sink. Therefore the difference between these two terms is monotone nondecreasing. □

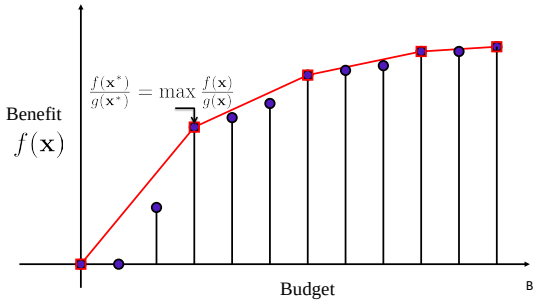

Figure 2: The concave envelope, the breakpoints and the ratio maximizing solution.

We are interested next in the link between $\max_{\mathbf{x}\in\mathcal{F}} f(\mathbf{x}) - \lambda g(\mathbf{x})$ and the **efficient frontier** of the solutions to $\max_{\mathbf{x}\in\mathcal{F}}\{f(\mathbf{x})|g(\mathbf{x}) \le B\}$: Suppose that we graph the optimal solution $\mathbf{x}_B$ to the problem $\mathbf{x}_B = \arg\max_{\mathbf{x}\in\mathcal{F}}\{f(\mathbf{x})|g(\mathbf{x}) \le B\}$, with the horizontal axis the value of the budgets B and the vertical axis the value of the objective $f(\mathbf{x}_B)$. We will refer to B as the "budget" of $\mathbf{x}_B$ and to $f(\mathbf{x})$ as the "benefit" of of $\mathbf{x}$.

Consider the intersection of all the lines that have the entire collection of optimal solutions below them. This upper envelope is concave piecewise linear and the points at which the line segment changes, to lower slope line, are called *breakpoints*, see Figure 2. We note that the first breakpoint is also the optimal solution to the ratio problem, $\max_{\mathbf{x}\in\mathcal{F}} \frac{f(\mathbf{x})}{g(\mathbf{x})}$.

Let $\mathbf{x}^0 \in \mathcal{F}$ be a feasible solution, and let $\lambda_0 = \frac{f(\mathbf{x}^0)}{g(\mathbf{x}^0)}$. We claim that $\max_{\mathbf{x}\in\mathcal{F}} f(\mathbf{x}) - \lambda_0 g(\mathbf{x})$ provides the tangent line to the concave envelope with the slope λ_0 at a budget $\le g(\mathbf{x}^0)$.

Lemma 3. $\max_{\mathbf{x}\in\mathcal{F}} f(\mathbf{x}) - \lambda_0 g(\mathbf{x})$ *provides the tangent line to the concave envelope with the slope λ_0 at a budget $\le g(\mathbf{x}^0)$.*

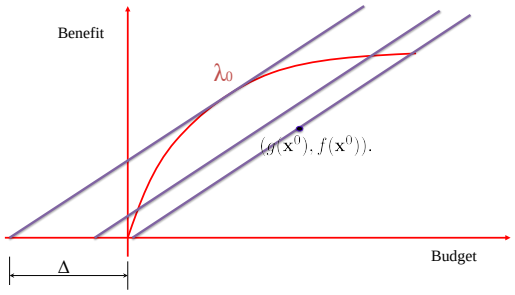

Figure 3: Identifying a breakpoint with λ_0 subgradient.

Proof. Let $-\Delta$ be the intercept of such line, of slope λ_0, on the horizontal axis, as in Figure 3. To find the point of the tangent we want to maximize Δ.
Since $f(\mathbf{x}) = \lambda_0(g(\mathbf{x}) - \Delta)$, it follows that $\Delta = \frac{1}{\lambda_0}[f(\mathbf{x}) - \lambda_0(g(\mathbf{x})]$.

Therefore, maximizing Δ is equivalent to $\max_{\mathbf{x}\in\mathcal{F}} f(\mathbf{x}) - \lambda_0 g(\mathbf{x})$. For such Δ^* the line $f(\mathbf{x}) - \lambda_0(g(\mathbf{x}) - \Delta^*)$ lies above all feasible solutions and is tangential to the concave envelope at breakpoint $\mathbf{x}^1$, where $\mathbf{x}^1 = \arg\max_{\mathbf{x}\in\mathcal{F}} f(\mathbf{x}) - \lambda_0 g(\mathbf{x})$.

$\mathbf{x}^1$ is a breakpoint with a left subgradient equal to λ_1 and right subgradient equal to λ_2, such that $\lambda_1 \geq \lambda_0 \geq \lambda_2$. □

For the convex envelope corresponding to the minimization problem, the tangent is found at a breakpoint to the right, bigger budget, of $g(\mathbf{x}^0)$.

The proof of Lemma 3 leads to an iterative procedure for ratio improvement. Consider the maximization case: From solution $\mathbf{x}^0$ we derive the breakpoint solution $\mathbf{x}^1$ with the left subgradient λ_1. When resolving, for λ_1, $\max_{\mathbf{x}\in\mathcal{F}} f(\mathbf{x}) - \lambda_1 g(\mathbf{x})$ the optimal solution is an adjacent breakpoint on the left, say $\mathbf{x}^2$ that has λ_1 as a right derivative. Also, the nestedness property implies that $\mathbf{x}^2 \leq \mathbf{x}^1$. Therefore, the procedure will scan all the breakpoints, one at a time, till it reaches the maximum "density" point, $\mathbf{x}^*$ that maximizes $\max_{\mathbf{x}\in\mathcal{F}} \frac{f(\mathbf{x})}{g(\mathbf{x})}$. Figure 2 illustrates this breakpoint, which is the leftmost of all breakpoints in the concave envelope. This process leads to the incremental parametric cut procedure described in the next section.

5 THE INCREMENTAL PARAMETRIC CUT AND CLUSTER IMPROVEMENT

In many applications the clustering problem has an NP-hard version that has an additional, budget, constraint on the size or weight of the selected set. Consider for instance the ratio region problem which is a relaxation of the NP-hard graph expander problem, where the minimization of the ratio is constrained by the requirements the size of the set is at most half the number of nodes in the graph. A common heuristic method to address this is *cluster improvement*. One derives an initial solution that satisfies the additional constraint, and then apply the respective ratio optimization to find a subset which is optimizes the ratio within this cluster. Although one can solve for the optimal ratio within the subset with the same parametric cut algorithm after fixing the values of the out-of-initial-cluster nodes to be out of the cluster there is another approach that was used, for instance, in (Lang and Rao, 2004).

PROCEDURE (MAXIMUM) RATIO IMPROVEMENT $(f(), g(), \mathbf{x}^0 \in \mathcal{F})$.

Step 0: Initialize $k = 0$,

Step 1: $\lambda_k = \frac{f(\mathbf{x}^k)}{g(\mathbf{x}^k)}$. {This step can be implemented extra efficiently adding linear time to all iterations.}

Step 2: Solve, with a min-cut algorithm or **procedure incremental-para**, $ratio(\lambda_k) = \max_{\mathbf{x}\in\mathcal{F}} f(\mathbf{x}) - \lambda_k g(\mathbf{x})$, and let $\mathbf{x}^{k+1} = \arg\max_{\mathbf{x}\in\mathcal{F}} f(\mathbf{x}) - \lambda_k g(\mathbf{x})$.

Step 3: If $ratio(\lambda_k) = 0$ stop. Output $\mathbf{x}^* = \mathbf{x}^k$. Else, continue

Step 4: $\{ratio(\lambda_k) < 0\}$ Let $k := k+1$. Go to step 1.

To prove validity need to show that $ratio(\lambda) \leq 0$ (because there is already a feasible solution of the value λ we know that the optimal ratio value can only be better than λ), and that the solution is optimal among those that have their support (valued 1 variables) contained in the support of $\mathbf{x}^0$.

The second claim follows from the fact that the problem is monotone IP3, and therefore solved with a min cut procedure. Also need to assume that $g(\mathbf{x}) = \sum q_i x_i$. therefore, when multiplied by λ the source adjacent are monotone increasing and sink adjacent are monotone decreasing and therefore the source sets are nested. That is, for the optimal ratio value $\lambda^* = \frac{f(\mathbf{x})^*}{g(\mathbf{x}^*)}$, $\lambda^* < \lambda$ it is guaranteed that the optimal solution set (the set of nodes with variable value $= 1$) is a strict subset of the set of nodes corresponding to $\mathbf{x}^k$.

While this procedure was used until now with each iteration requiring the running time of one minimum cut procedure, the authors adapted it to **procedure incremental-para** that uses the "continuation" idea of the HPF parametric cut procedure (Chandran and Hochbaum,). With this procedure values of the parameter, computed at each iteration, are used to determine the capacities of the source and the sink adjacent arcs, during the ratio improvement algorithm when a new solution is found. The number of calls to the procedure is the number of breakpoints traversed, till it reaches the leftmost breakpoint. The complexity however of **procedure incremental-para** is that of a single minimum cut procedure on the respective graph, e.g. $O(mn \log \frac{n^2}{m})$.

6 CONCLUSIONS

We present here a general unifying framework for any NP-hard problem, that can be formulated as monotone integer programming problem with a budget constraint (budgeted IPM problem). This framework provides powerful algorithmic methods that work efficiently and deliver very high quality solutions. One example illustrated here is the text summarization problem, for which we introduce here a new model

which is budgeted IPM potentially leading to more efficient and scalable algorithms.

We study here the concave envelope of the Lagrangean relaxation of the budget constraint, and demonstrate, a new parametric cut procedure that identifies, consecutively, all the breakpoints. This *incremental* parametric cut procedure is more easily implemented than the general parametric cut procedure, and reduces the running time on average.

ACKNOWLEDGEMENTS

This research was supported in part by AI Institute NSF Award 2112533.

REFERENCES

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Chaillou, P., Hansen, P., and Mahieu, Y. (1989). Best network flow bounds for the quadratic knapsack problem. In *Combinatorial Optimization: Lectures given at the 3rd Session of the Centro Internazionale Matematico Estivo (CIME) held at Como, Italy, August 25–September 2, 1986*, pages 225–235. Springer.

Chandran, B. and Hochbaum, D. S. Pseudoflow parametric maximum flow solver version 1.0. url-https://riot.ieor.berkeley.edu/Applications/Pseudoflow/parametric.html. Accessed: 2023-04-30.

Cheeger, J. (1970). A lower bound for the smallest eigenvalue of the laplacian, problems in analysis (papers dedicated to salomon bochner, 1969).

Gallo, G., Grigoriadis, M. D., and Tarjan, R. E. (1989). A fast parametric maximum flow algorithm and applications. *SIAM Journal on Computing*, 18(1):30–55.

Hochbaum, D. S. (2002). Solving integer programs over monotone inequalities in three variables: A framework for half integrality and good approximations. *European Journal of Operational Research*, 140(2):291–321.

Hochbaum, D. S. (2008). The pseudoflow algorithm: A new algorithm for the maximum-flow problem. *Operations research*, 56(4):992–1009.

Hochbaum, D. S. (2010). Polynomial time algorithms for ratio regions and a variant of normalized cut. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):889–898.

Hochbaum, D. S. and Fishbain, B. (2011). Nuclear threat detection with mobile distributed sensor networks. *Annals of Operations Research*, 187:45–63.

Hochbaum, D. S., Liu, Z., and Goldschmidt, O. (2023). A breakpoints based method for the maximum diversity and dispersion problems. In *SIAM Conference on Applied and Computational Discrete Algorithms (ACDA23)*, pages 189–200. SIAM.

Hochbaum, D. S. and Orlin, J. B. (2013). Simplifications and speedups of the pseudoflow algorithm. *Networks*, 61(1):40–57.

Hochbaum, D. S. and Singh, V. (2009). An efficient algorithm for co-segmentation. In *2009 IEEE 12th International Conference on Computer Vision*, pages 269–276. IEEE.

Lang, K. and Rao, S. (2004). A flow-based method for improving the expansion or conductance of graph cuts. In *Integer Programming and Combinatorial Optimization: 10th International IPCO Conference, New York, NY, USA, June 7-11, 2004. Proceedings 10*, pages 325–337. Springer.

Lin, H. and Bilmes, J. (2010). Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920.

Pisinger, D. (2006). Upper bounds and exact algorithms for p-dispersion problems. *Computers & operations research*, 33(5):1380–1398.

Sharon, E., Galun, M., Sharon, D., Basri, R., and Brandt, A. (2006). Hierarchy and adaptivity in segmenting visual scenes. *Nature*, 442(7104):810–813.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905.

Spiers, S., Bui, H. T., and Loxton, R. (2023). An exact cutting plane method for the euclidean max-sum diversity problem. *European Journal of Operational Research*.