Submodular Maximization with Multi-Knapsack Constraints and its Applications in Scientific Literature Recommendations

Qilian Yu\textsuperscript{1} Easton Li Xu\textsuperscript{2} Shuguang Cui\textsuperscript{1}

\textsuperscript{1}Department of Electrical and Computer Engineering
University of California, Davis

\textsuperscript{2}Department of Electrical and Computer Engineering
Texas A&M University

2016 IEEE Global Conference on Signal and Information Processing
Dec. 7, 2016
Outline

1. Introduction

2. Formulation and Main Results
   - Problem Formulation
   - Related Work and Main Results

3. Streaming Algorithm for Maximizing Monotone Submodular Functions
   - Algorithm
   - Theoretical Guarantee

4. Applications
Introduction

Background

- Main Problems in Big Data Era
  - Unprecedented large datasets.
  - Heterogenous data sources.
- Submodular Optimization
  - Rich theoretical and practical features to preprocess massive data [Liu et. al. 2013].
- Streaming Algorithms
  - Memory required for a small portion of data.
  - Solution provided at the end of data stream.
Prerequisites

- Ground set: $V = \{1, 2, \ldots, n\}$.
- Set function: $f : 2^V \rightarrow [0, \infty)$.
- Characteristic vector: $x_S = (x_{S,1}, x_{S,2}, \ldots, x_{S,n})$, where for $1 \leq j \leq n$, $x_{S,j} = 1$, if $j \in S$; $x_{S,j} = 0$, otherwise.
- Marginal gain: $\Delta_f(r|S) \triangleq f(S \cup \{r\}) - f(S)$.
  - Submodularity: $\Delta_f(r|B) \leq \Delta_f(r|A)$, for $A \subseteq B \subseteq V$ and $r \in V \setminus B$.
  - Monotone: $\Delta_f(r|S) \geq 0$, for any $S \subseteq V$ and $r \in V$. 
Formulation

■ Motivation: scientific literature recommendations, new recommendations, etc.

■ **d-MASK**: Aim to **MA**ximize a monotone **S**ubmodular set function subject to a **d-K**napsack constraint.

\[
\text{maximize } \quad f(S) \\
\text{subject to } \quad Cx_S \leq b.
\]  

■ \( b = (b_1, b_2, \ldots, b_d)^T \): \(d\)-dimension knapsack constraint vector.

■ \( C = (c_{i.j}) \): \( c_{i.j} > 0 \) is the weight of the element \( j \) with respect to the \( i \)-th knapsack resource constraint.

■ **d-MASK** can be easily standardized such that \( c_{i.j} \geq 1 \) and \( b_i = b \), for \( 1 \leq i \leq d, 1 \leq j \leq n \).
### Formulation and Main Results

#### Related Work and Main Results

<table>
<thead>
<tr>
<th>Best Performance Known Algorithms</th>
<th>Proposed Streaming Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Knapsack Constraint</td>
<td>$1 - \frac{e^{-1}}{n^5}$ [Sviridenko, 2004]</td>
</tr>
<tr>
<td>$d$-Knapsack Constraint</td>
<td>$1 - \frac{e^{-1} - \epsilon}{n^5}$ [Kulik et. al., 2009]</td>
</tr>
</tbody>
</table>

First to propose an efficient streaming algorithm for $d$-MASK, with

- a constant-factor approximation guarantee;
- no assumption on full access to the dataset;
- execution of a single pass;
- $O(b \log b)$ memory requirement;
- $O(\log b)$ computation complexity per element;
- only assumption on monotonicity and submodularity of the objective function.
Algorithm 1 $d$-KNAPSACK-STREAMING

1: $m := 0.$
2: $Q := \{ [1 + (1 + 2d)\varepsilon]^l | l \in \mathbb{Z} \}.$
3: for $v \in Q$
4:   $S_v := \emptyset.$
5:   for $j := 1$ to $n$
6:     for $i := 1$ to $d$
7:       $m := \max \{ m, f(\{j\})/c_{i,j} \}.$
8:     end for
9:     $Q := \{ [1 + (1 + 2d)\varepsilon]^l | l \in \mathbb{Z}, \frac{m}{1+(1+2d)\varepsilon} \leq [1 + (1 + 2d)\varepsilon]^l \leq 2bm \}.$
10:    if $c_{i,j} \geq \frac{b}{2}$ and $\frac{f(\{j\})}{c_{i,j}} \geq \frac{2v}{b(1+2d)}$ for some $i \in [1, d]$ then
11:       $S_v := \{ j \}.$
12:       break
13:    end if
14:    if $\sum_{l \in S \cup \{j\}} c_{i,l} \leq b$ and $\frac{\Delta f(\{j\}|S)}{c_{i,j}} \geq \frac{2v}{b(1+2d)}$ for all $i \in [1, d]$ then
15:       $S_v := S_v \cup \{ j \}.$
16:    end if
17:  end for
18: end for
19: $S := \arg\max_{S_v, v \in Q} f(S_v).$
20: return $S.$
Simpler Version

Algorithm 2 $d$-KNAPSACK-STREAMING

1: Initialize: Set $Q$.
2: for $v \in Q$
3:     for $j := 1$ to $n$
4:         Update Set $Q$.
5:         if $j$ is big element then
6:             $S_v := \{j\}$.
7:             break.
8:     end if
9:     if $j$ satisfies criteria($v$) then
10:        $S_v := S_v \cup \{j\}$.
11:    end if
12: end for
13: end for
14: $S := \arg\max_{S_v, v \in Q} f(S_v)$.
15: return $S$. 
**Lemma 1**

Let

\[ Q = \left\{ \left[ 1 + (1 + 2d)\epsilon \right]^l | l \in \mathbb{Z}, \frac{m}{1 + (1 + 2d)\epsilon} \leq \left[ 1 + (1 + 2d)\epsilon \right]^l \leq 2bm \right\} \]

for some \( \epsilon \) with \( 0 < \epsilon < \frac{1}{1+2d} \). Then there exists at least some \( v \in Q \) such that \( \left[ 1 - (1 + 2d)\epsilon \right] \text{OPT} \leq v \leq \text{OPT} \).

**Lemma 2 (Big Element)**

Assume \( v \) satisfies \( \alpha \text{OPT} \leq v \leq \text{OPT} \), and there exists an element \( j \) such that \( c_{i,j} \geq \frac{b}{2} \) and \( \frac{f(\{j\})}{c_{i,j}} \geq \frac{2v}{b(1+d)} \) for some \( i \in [1, d] \).

\[ f(\{j\}) \geq \frac{\alpha}{1 + 2d} \text{OPT}. \]
Submodular Maximization with Multi-Knapsack Constraints and its Applications in Scientific Literature Recommendations

Streaming Algorithm for Maximizing Monotone Submodular Functions

Theoretical Guarantee

**Theorem 3**

Algorithm 1 has the following properties:

- It outputs $S$ that satisfies $f(S) \geq (\frac{1}{1+2d} - \epsilon) OPT$;

- It goes one pass over the dataset, stores at most $O\left(\frac{b\log b}{d\epsilon}\right)$ elements, and has $O\left(\frac{\log b}{\epsilon}\right)$ computation complexity per element.

**Theorem 4**

Consider a subset $S \subseteq V$. For $1 \leq i \leq d$, let $r_{i,s} = \Delta f(s|S)/c_{i,s}$, and $s_{i,1}, \ldots, s_{i,|V\setminus S|}$ be the sequence such that $r_{i,s_{i,1}} \geq r_{i,s_{i,2}} \geq \cdots \geq r_{i,s_{i,|V\setminus S|}}$. Let $k_i$ be the integer such that $\sum_{j=1}^{k_i-1} c_{i,s_i,j} \leq b$ and $\sum_{j=1}^{k_i} c_{i,s_i,j} > b$. And let

$$\lambda_i = \left(b - \sum_{j=1}^{k_i-1} c_{i,s_i,j}\right) / c_{i,s_i,k_i}.$$  

Then we have

$$OPT \leq f(S) + \min_{1 \leq i \leq d} \left[ \sum_{j=1}^{k_i-1} \Delta f(s_{i,j}|S) + \lambda_i \Delta f(s_{i,k_i}|S) \right].$$
Scientific Literature Recommendations

Problem Setup

- **Problem setting**
  - A directed acyclic graph $G = (V, E)$ with $V = \{1, 2, \ldots, n\}$.
  - Vertex in $V$: an article.
  - Arc $(i, j) \in E$: paper $i$ cites paper $j$.
  - $A$: the collection of the source papers.

- **Objective**
  - Select a subset $S$ out of $V$ to quickly detect the information spreading of $A$. 
Problem Formulation

- **Measurements**
  - Length of the shortest directed path from $s$ to $a$: $T(s, a)$.
  - The shortest path length from any vertex in $S$ to $a$: $T(S, a) \triangleq \min_{s \in S} T(s, a)$.
  - Pre-assigned weight to each vertex $a \in A$: $W(a)$, such that $\sum_{a \in A} W(a) = 1$.
  - A given maximum penalty: $T_{\text{max}}$.
  - The expected penalty: $\pi(S) \triangleq \sum_{a \in A} W(a) \min\{T(S, a), T_{\text{max}}\}$.

- **Formulation**

\[
\begin{align*}
\text{maximize} \quad & R(S) \triangleq \sum_{a \in A} W(a) [T_{\text{max}} - T(S, a)]^+ \\
\text{subject to} \quad & Cx_S \leq b.
\end{align*}
\]
Experiment Setup

- Constraints Design
  - Recency
  - Biased PageRank Score [Gori & Pucci, 2006]
  - Reference Number

- Experiment Dataset [Joseph & Radev, 2007]
  - Over 20,000 papers in the Association of Computational Linguistics.
  - Citation network provided.
Applications

Experimental Results

- Sensitive Analysis Setup
  - Randomly select five nodes as the source papers.
  - Set $T_{\text{max}} = 50$ and $W(a) = 0.2$ for each source paper $a$.

<table>
<thead>
<tr>
<th>Fixed Constraints</th>
<th>Recency Knapsack Constraint</th>
<th>Biased PageRank Knapsack Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_2 = 10$, $b_3 = 20$.</td>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
</tr>
<tr>
<td>$b_1 = 20$, $b_3 = 20$.</td>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
</tr>
<tr>
<td>$b_1 = 20$, $b_2 = 10$.</td>
<td><img src="image5.png" alt="Graph" /></td>
<td><img src="image6.png" alt="Graph" /></td>
</tr>
</tbody>
</table>
Summary

- The first streaming algorithm for $d$-MASK problem.

- Only a single pass through the dataset required.

- Approximation solution with a \( \left( \frac{1}{1+2d} - \epsilon \right) \) factor guaranteed with much lower computation cost.

- Practical and efficient way to solve related combinatorial problem, e.g., scientific literature recommendations.
Y. Liu, K. Wei, K. Kirchhoff, Y. Song, and J. Bilmes,
“Submodular feature selection for high-dimensional acoustic score spaces,”

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher,
“An analysis of approximations for maximizing submodular set functions–I,”

M. Sviridenko,
“A note on maximizing a submodular set function subject to a knapsack constraint,”

A. Kulik, H. Shachnai, and T. Tamir,
“Maximizing submodular set functions subject to multiple linear constraints,”

M. Gori and A. Pucci,
“Research paper recommender systems: A random-walk based approach,”

M. T. Joseph and D. R. Radev,
“Citation analysis, centrality, and the ACL anthology,”