Notes on Graph-structured Sparsity Optimization

Fei Jie
hfut_jf@aliyun.com

April 25, 2019

Abstract

This note is a summary for works on graph-structured nonconvex optimization by Prof. Feng Chen's lab. They borrows ideas from structured sparse learning, which only focuses on several limited score functions and specific structures. Graph-structured nonconvex optimization generalizes the structured sparse learning to generic graph structure and optimize any differentiable score function. The works' were deployed in subgraph detection tasks to validate their effectiveness and efficiency.

1 Introduction

2 Preliminary

2.1 Approximation algorithms for the projection oracle $\mathrm{P}(\mathbf{x})$

There are two nearly-linear time approximation algorithms [3] that have the following properties:

- **Tail approximation** ($\mathrm{T}(\mathbf{x})$): Find a $S \subseteq \mathbb{V}$ such that

$$\|\mathbf{x} - \mathbf{x}_S\|_2 \leq c_T \cdot \min_{S \in \mathbb{M}(\mathbb{G}, k_T)} \|\mathbf{x} - \mathbf{x}_{S'}\|_2 \tag{1}$$

  where $c_T = \sqrt{7}, k_T = 5s$, and $\mathbf{x}_S$ is the restriction of $\mathbf{x}$ to indices in S: we have $(\mathbf{x}_S)_i = x_i$ for $i \in S$ and $(\mathbf{x}_S)_i = 0$ otherwise.

- **Head approximation** ($\mathrm{H}(\mathbf{x})$): Find a $S \subseteq \mathbb{V}$ such that

$$\|\mathbf{x}_S\|_2 \geq \cdot \max_{S' \in \mathbb{M}(\mathbb{G}, k_H)} \|\mathbf{x}_{S'}\|_2 \tag{2}$$

  where $c_H = \sqrt{1/14}$ and $k_H = 2s$.

3 Graph-Mp

This section is a summary for IJCAI 2016 paper [1], which introduces Graph-Mp algorithm.

3.1 Problem Statement

Given an underlying graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ defined on the coefficients of the unknown vector $\mathbf{x}$, where $\mathbb{V} = [n]$,$\mathbb{E} \subseteq \mathbb{V} \times \mathbb{V}$. The sparsity model of connected subgraphs in $\mathbb{G}$ is defined as

$$\mathbb{M}(\mathbb{G}, s) = \{S \subseteq \mathbb{V} || S| \leq s, S \text{ is connected}\} \tag{3}$$

The problem to be studied is formulated as:

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad \text{supp} \in \mathbb{M}(\mathbb{G}, s) \tag{4}$$

The key idea is to decompose this problem to sub-problems that are easier to solve. These sub-problems include and optimization sub-problem of $f(\mathbf{x})$ that is independent on $\mathbb{G}(\mathbb{G}, s)$ and projection approximations for $\mathbb{M}(\mathbb{G}, s)$, including $\mathrm{H}(\mathbf{x})$ and $\mathrm{T}(\mathbf{x})$.

Algorithm 1 GRAPH-MP

Input: Input graph $\mathbb{G}$, maximum subgraph size s, and step size η (1 by default).
Output: The estimated vector $\hat{\mathbf{x}}$ and the corresponding connected subgraph S.
$i \leftarrow 0$, $\mathbf{x}^i \leftarrow \mathbf{0}$, $S^i \leftarrow \emptyset$
repeat
    $\Gamma \leftarrow \mathrm{H}(\nabla f(\mathbf{x}))$
    $\Omega \leftarrow \Gamma \cup \mathrm{supp}(\mathbf{x}^i)$
    $\mathbf{b} \leftarrow \arg\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad \mathrm{supp}(\mathbf{x}) \in \Omega$
    $S^{i+1} \leftarrow \mathrm{T}(\mathbf{b})$
    $\mathbf{x}^{i+1} \leftarrow \mathbf{b}_{S^{i+1}}$
    $i \leftarrow i + 1$
until halting condition holds
return $\hat{\mathbf{x}} = \mathbf{x}^i$ and $S = \mathbb{G}_{S^i}$

3.2 Algorithm

3.3 Theoretical Analysis

Definition 1 (Weak Restricted Strong Convexity Property (WRSC))**.** *A function $f(\mathbf{x})$ has the $(\xi, \delta, \mathbb{M})$-WRSC if $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\forall S \in \mathbb{M}$ with $\mathrm{supp}(\mathbf{x}) \cup \mathrm{supp}(\mathbf{y}) \subseteq S$, the following inequality holds for some $\xi > 0$ and $0 < \delta < 1$:*

$$\|\mathbf{x} - \mathbf{y} - \xi\nabla_S f(\mathbf{x}) + \xi\nabla_S f(\mathbf{y})\|_2 \leq \delta\|\mathbf{x} - \mathbf{y}\|_2 \tag{5}$$

The WRSC is weaker than the *Restricted Strong Convexity/Smoothness* (RSC/RSS) conditions that are used in theoretical analysis of convex optimization algorithms. The RSC/RSS conditions imply condition WRSC, which indicates that WRSC is no stronger than RSC/RSS.

Theorem 1. *Consider the graph-structured sparsity model $\mathbb{M}(\mathbb{G}, s)$ for some $s \in \mathbb{N}$ and a cost function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies condition $(\xi, \delta, \mathbb{M}(\mathbb{G}, 8s))$-WRSC. If $\eta = c_H(1-\delta) - \delta > 0, \rho = \xi(1 + c_H)$, then for any true $\mathbf{x} \in \mathbb{R}^n$ with $\mathrm{supp}(\mathbf{x}) \in \mathbb{M}(\mathbb{G}, s)$, the iterations of Algorithm 1 obey*

$$\|\mathbf{x}^{i+1} - \mathbf{x}\|_2 \leq \alpha\|\mathbf{x}^i - \mathbf{x}\|_2 + \beta\|\nabla_I f(\mathbf{x})\|_2 \tag{6}$$

where $\beta = \frac{\xi(1+c_T)}{1-\delta}\left[\frac{(1+c_H)}{\eta} + \frac{\eta(1+c_H)}{\sqrt{1-\eta^2}} + 1\right] = \frac{(1+c_T)}{1-\delta}\left[\frac{\rho}{\eta} + \frac{\eta\rho}{\sqrt{1-\eta^2}} + \xi\right]$, $\alpha = \frac{(1+c_T)}{1-\delta}\sqrt{1-\eta^2}$, and $I = \arg\max_{S\in\mathbb{M}(\mathbb{G},8s)} \|\nabla_S f(\mathbf{x})\|_2$.

Proof. A proof of this result can be found in Appendix. □

Theorem 2. *Let $\mathbf{x} \in \mathbb{R}^n$ be a true optimum such that $\mathrm{supp}(\mathbf{x}) \in \mathbb{M}(\mathbb{G}, s)$, and $f : \mathbb{R}^n \to \mathbb{R}$ be a cost function that satisfies condition $(\xi, \delta, \mathbb{M}(\mathbb{G}, 8s))$-WRSC. Assuming that $\alpha < 1$,* GRAPH-MP *returns a $\hat{\mathbf{x}}$ such that $\mathrm{supp}(\hat{\mathbf{x}}) \in \mathbb{M}(\mathbb{G}, 5s)$ and $\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq c\|\nabla_I f(\mathbf{x})\|_2$, where $c = 1 + \frac{\beta}{1-\alpha}$ is a fixed constant. Moreover,* GRAPH-MP *runs in time*

$$O((T + |\mathbb{E}|\log^3 n)\log(\|\mathbf{x}\|_2/\|\nabla_I f(\mathbf{x})\|_2)) \tag{7}$$

where T is the time complexity of one execution of the sub problem in Line 7. In particular, if T scales linearly with n, then GRAPH-MP *scales nearly linearly with n.*

Proof. A proof of this result can be found in Appendix. □

4 GRAPH-IHT and GRAPH-GHTP

This section introduce Baojian Zhou's 2016 ICDM work [5], in which he proposes two algorithms: GRAPH-IHT and GRAPH-GHTP.

4.1 Problem Statement

These two algorithms solve the same problem as problem (4).

Algorithm 2 GRAPH-IHT

Input: Input graph $\mathbb{G}$, maximum subgraph size s, and step size η (1 by default).
Output: The estimated vector $\hat{\mathbf{x}}$ and the corresponding connected subgraph S.
$i \leftarrow 0$, $\mathbf{x}^i \leftarrow \mathbf{0}$, $S^i \leftarrow \emptyset$
repeat
  $\Omega \leftarrow \mathrm{H}(\nabla f(\mathbf{x}))$
  $\mathbf{b} \leftarrow \mathbf{x}^i - \eta \cdot \nabla_\Omega f(\mathbf{x}^i)$
  $S^{i+1} \leftarrow \mathrm{T}(\mathbf{b})$
  $\mathbf{x}^{i+1} \leftarrow \mathbf{b}_{S^{i+1}}$
  $i \leftarrow i + 1$
until halting condition holds
return $\hat{\mathbf{x}} = \mathbf{x}^i$ and $S = \mathbb{G}_{S^i}$

Algorithm 3 GRAPH-GHTP

Input: Input graph $\mathbb{G}$, maximum subgraph size s, and step size η (1 by default).
Output: The estimated vector $\hat{\mathbf{x}}$ and the corresponding connected subgraph S.
$i \leftarrow 0$, $\mathbf{x}^i \leftarrow \mathbf{0}$, $S^i \leftarrow \emptyset$
repeat
  $\Omega \leftarrow \mathrm{H}(\nabla f(\mathbf{x}))$
  $\Psi \leftarrow \mathrm{supp}(\mathbf{x}^i - \eta \cdot \nabla_\Omega f(\mathbf{x}^i))$
  $\mathbf{b} \leftarrow \arg\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad \mathrm{supp}(\mathbf{x}) \subseteq \Psi$
  $S^{i+1} \leftarrow \mathrm{T}(\mathbf{b})$
  $\mathbf{x}^{i+1} \leftarrow \mathbf{b}_{S^{i+1}}$
  $i \leftarrow i + 1$
until halting condition holds
return $\hat{\mathbf{x}} = \mathbf{x}^i$ and $S = \mathbb{G}_{S^i}$

4.2 Algorithm

4.3 Theoretical Analysis of GRAPH-IHT

Theorem 3. *Consider the sparsity model of connected subgraphs $\mathbb{M}(\mathbb{G}, s)$ for some $s \in \mathbb{N}$ and a cost function $f : \mathbb{R}^n \to \mathbb{N}$ that satisfies the $(\xi, \delta, \mathbb{M}(\mathbb{G}, 5s))$-WRSC condition. If $\eta = c_H(1-\delta)-\delta$, $\rho = \delta(1+c_H)$, then for any $\mathbf{x} \in \mathbb{R}^n$ such that $\mathrm{supp}(\mathbf{x}) \in \mathbb{M}(\mathbb{G}, s)$, with $\eta > 0$ the iterations of Algorithm 2 obey*

$$\|\mathbf{x}^{i+1} - \mathbf{x}\|_2 \leq \alpha \|\mathbf{x}^i - \mathbf{x}\|_2 + \beta \|\nabla_I f(\mathbf{x})\|_2 \tag{8}$$

where

$$\alpha = (1 + c_T)\left(\sqrt{1 - \eta^2} + 1 - \frac{\eta}{\xi} + (2 - \frac{\eta}{\xi})\delta\right),$$

$$\beta = (1 + c_T)\left(\eta + \frac{\rho}{\eta} + \frac{\eta\rho}{\sqrt{1 - \eta^2}}\right),$$

and $I = \arg\max_{S \in \mathbb{M}(\mathbb{G}, 8k)} \|\nabla_S f(\mathbf{x})\|_2$

Proof. A proof of this result can be found in Appendix. □

4.4 Theoretical Analysis of GRAPH-GHTP

Theorem 4. *Consider the sparsity model of connected subgraphs $\mathbb{M}(\mathbb{G}, s)$ for some $s \in \mathbb{N}$ and a cost function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies the $(\xi, \delta, \mathbb{M}(\mathbb{G}, 5s))$-WRSC. If $\eta = c_H(1 - \delta) - \delta, \rho = \delta(1 + c_H)$, then for any $\mathbf{x} \in \mathbb{R}^n$ such that $\mathrm{supp}(\mathbf{x}) \in \mathbb{M}(\mathbb{G}, s)$, with $\eta > 0$ the iterations of Algorithm 3 obey*

$$\|\mathbf{x}^{i+1} - \mathbf{x}\|_2 \leq \alpha \|\mathbf{x}^i - \mathbf{x}\|_2 + \beta \|\nabla_I f(\mathbf{x})\|_2 \tag{9}$$

where

$$\alpha = \frac{\sqrt{2}(1 + c_T)}{1 - \delta}\left(\sqrt{1 - \eta^2} + \left((2 - \frac{\eta}{\xi})\delta + 1 - \frac{\eta}{\xi}\right)\right),$$

$$\beta = \frac{1+c_T}{1-\delta}\left((1+2\sqrt{2})\xi + (2-2\sqrt{2})\eta + \frac{\sqrt{2}\rho}{\eta} + \frac{\sqrt{2}\eta\rho}{\sqrt{1-\eta^2}}\right),$$

and $I = \arg\max_{S\in\mathbb{M}(\mathbb{G},8s)} \|\nabla_S f(\mathbf{x})\|_2$.

Proof. A proof of this result can be found in Appendix. □

Theorem 5. *Let* $\mathbf{x} \in \mathbb{R}^n$ *such that* $\text{supp}(\mathbf{x}) \in \mathbb{M}(\mathbb{G}, s)$*, and* $f : \mathbb{R}^n \to \mathbb{R}$ *be cost function that satisfies condition* $(\xi, \delta, \mathbb{M}(\mathbb{G}, 8s))$*-WRSC. Assuming that* $\alpha < 1$*,* Graph-*GHTP (or* Graph-*IHT) returns a* $\hat{\mathbf{x}}$ *such that,* $\text{supp}(\hat{\mathbf{x}}) \in \mathbb{M}(\mathbb{G}, 5s)$ *and* $\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq c\|\nabla_I f(\mathbf{x})\|_2$*, where* $c = 1 + \frac{\beta}{1-\alpha}$ *is a fixed constant. Moreover,* Graph-*GHTP (or* Graph-*IHT) runs in time*

$$O\left((T + |E|\log^3 n)\log\left(\frac{\|\mathbf{x}\|_2}{\|\nabla_I f(\mathbf{x})\|_2}\right)\right) \tag{10}$$

where T *is the time complexity of one execution of the subproblem in Step 7 in* Graph-*GHTP (or Step 6 in* Graph-*IHT). In particular, if* T *scales linearly with* n*, then* Graph-*GHTP (or* Graph-*IHT) scales nearly linearly with* n*.*

5 SG-Pursuit

5.1 Problem Statement

We consider a multi-attributed network that is defined as $\mathbb{G} = (\mathbb{V}, \mathbb{E}, \mathbf{w})$, where $\mathbb{V} = \{1, \ldots, n\}$ is the ground set of nodes of size n, $\mathbb{E} \subseteq \mathbb{V} \times \mathbb{V}$ is the set of edges, and the function $\mathbf{w} : \mathbb{V} \to \mathbb{R}^p$ defines a vector of attributes of size p for each node $v \in \mathbb{V} : \mathbf{w}(v) \in \mathbb{R}^p$. For simplicity, we denote the attribute vector $\mathbf{w}(v)$ by $\mathbf{w}_v$.

We introduce two vectors of coefficients, including $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^p$, that will be optimized for detecting the most interesting subspace cluster in $\mathbb{G}$, where $\mathbf{x}$ identifies the cluster (subset) of nodes and $\mathbf{y}$ identifies their relevant attributes. In particular, the vector $\mathbf{x}$ refers to the vector of coefficients of the nodes in $\mathbb{V}$. Each node $i \in \mathbb{V}$ has a coefficient score x_i indicating the importance of this node in the cluster of interest. Similarly, the vector $\mathbf{y}$ refers to the vector of coefficients of the p attributes. Each attribute $j \in \{1, \ldots, p\}$ has a coefficient score y_j indicating the relevance of this attribute to the clusters of interest. Let $\text{supp}(\mathbf{x})$ be the support set of indices of nonzero entries in $\mathbf{x}$: $\text{supp}(\mathbf{x}) = \{i | x_i \neq 0\}$. Then the support set of $\text{supp}(\mathbf{x})$ represents the subset of nodes that belong to the cluster of interest. The support set $\text{supp}(\mathbf{y})$ represents the subset of relevant attributes. We define the feasible space of clusters of nodes as

$$\mathbb{M}(\mathbb{G}, s) = \{S | S \subseteq \mathbb{V}; |S| \leq s; \mathbb{G}_S \text{ satisfies predefined topological constraints}\},$$

where S refers to a subset of nodes in $\mathbb{V}$, $\mathbb{G}_S = \{S, \mathbb{E} \cap S \times S\}$ refers to the subgraph included by S, $|S|$ refers to the total number of nodes in S, and s refers to an upper bound on the size of the cluster. The topological constraints can be any topological constraints on $\mathbb{G}_S$, such as connected subgraphs, dense subgraph, subgraphs that are isomorphic to a query graph, compact subgraphs, trees, and paths, among others.

Based on the above notations, we consider a general form of the subspace cluster detection problem as

$$\max_{\mathbf{x}\in C_{\mathbf{x}}, \mathbf{y}\in C_{\mathbf{y}}} f(\mathbf{x}, \mathbf{y}) \qquad \text{s.t.} \qquad \text{supp}(\mathbf{x}) \in \mathbb{M}(\mathbb{G}, s) \text{ and } \|\mathbf{y}\|_0 \leq k$$

where $f(\mathbf{x}, \mathbf{y}) : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ is a score function that measures the overall level of interestingness of the subspace clusters indicated by $\mathbf{x}$ and $\mathbf{y}$; $C_{\mathbf{x}} \subseteq \mathbb{R}^n$ represents a convex set in the Euclidean space $\mathbb{R}^n$, $C_{\mathbf{y}} \subseteq \mathbb{R}^p$ represents a convex set in the Euclidean space $\mathbb{R}^p$, $\mathbb{M}(\mathbb{G}, s)$ refers to the feasible space of clusters of nodes as defined above, and k refers to an upper bound on the number of attributes relevant to the subspace clusters of interest. The parameters s and k are predefined by the user.

5.2 Algorithm

5.3 Theoretical Analysis

x[2]

Algorithm 4 SG-Pursuit

Input: Input graph $\mathbb{G}$, maximum subgraph size k, maximum selected feature size s, and step size η (1 by default).
Output: The estimated vectors of coefficients of nodes and attributes, including $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, and the identified subspace cluster C.
$i \leftarrow 0$, $\mathbf{x}^i, \mathbf{y}^i \leftarrow$ initial vectors
repeat
  $\Gamma_{\mathbf{x}} = \mathrm{H}(\nabla_{\mathbf{x}} f(\mathbf{x}^i, \mathbf{y}^i))$
  $\Gamma_{\mathbf{y}} = \arg\max_{R \subseteq \{1,\ldots,p\}} \{\|[\nabla_{\mathbf{y}} f(\mathbf{x}^i, \mathbf{y}^i)]_R\|_2^2 : \|R\|_0 \leq 2s\}$
  $\Omega_{\mathbf{x}} = \Gamma_{\mathbf{x}} \cup \mathrm{supp}(\mathbf{x}^i)$
  $\Omega_{\mathbf{y}} = \Gamma_{\mathbf{y}} \cup \mathrm{supp}(\mathbf{y}^i)$
  $(\mathbf{b}_{\mathbf{x}}^i, \mathbf{b}_{\mathbf{y}}^i) = \arg\max_{\mathbf{x} \in C_{\mathbf{x}}, \mathbf{y} \in C_{\mathbf{y}}} f(\mathbf{x}, \mathbf{y}) \quad \text{s.t.} \quad \mathrm{supp}(\mathbf{x}) \subseteq \Omega_{\mathbf{x}}, \mathrm{supp}(\mathbf{y}) \subseteq \Omega_{\mathbf{y}}$
  $\Psi_{\mathbf{x}}^{i+1} = \mathrm{T}(\mathbf{b}_{\mathbf{x}}^i)$
  $\Psi_{\mathbf{y}}^{i+1} = \arg\max_{R \subseteq \{1,\ldots,p\}} \{\|[\mathbf{b}_{\mathbf{y}^i}]_R\|_2^2 : \|R\|_0 \leq s\}$
  $\mathbf{x}^{i+1} = [\mathbf{b}_{\mathbf{x}}^i]_{\Psi_{\mathbf{x}}^{i+1}}$
  $\mathbf{y}^{i+1} = [\mathbf{b}_{\mathbf{y}}^i]_{\Psi_{\mathbf{y}}^{i+1}}$
  $i \leftarrow i + 1$
until halting condition holds
return $\hat{\mathbf{x}} = \mathbf{x}^i$ and $S = \mathbb{G}_{S^i}$

6 Graph Block-Structured Matching Pursuit

6.1 Problem Statement

Suppose we are given such a network $\mathbb{G} = (\mathbb{V}, \mathbb{E}, \mathbf{W})$, where $\mathbb{V} = \{1, \cdots, N\}$ is the ground set of vertices, $\mathbb{E} \subseteq \mathbb{V} \times \mathbb{V}$ is the ground set of edges, $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_N] \in \mathbb{R}^{P \times N}$ is the feature matrix, and $\mathbf{w}_i \in \mathbb{R}^P$ is the feature vector of vertex i. Suppose $\mathbb{V}$ is decomposed to K disjoint subsets (blocks) of vertices: $\mathbb{V} = \mathbb{V}^1 \cup \cdots \cup \mathbb{V}^K$, where $N_k = |\mathbb{V}^k|$ is the size of the subset of vertices $\mathbb{V}^k$. The general subgraph detection problem in multiple blocks can be formulated as following general block-structured optimization problem:

$$\begin{aligned} \min_{\mathbf{x}=(\mathbf{x}^1,\ldots,\mathbf{x}^K)} F(\mathbf{x}) &= f(\mathbf{x}^1, \cdots, \mathbf{x}^K) + \sum_{k=1}^{K} g_k(\mathbf{x}^k), \\ \text{s.t.} \;\; \mathrm{supp}(\mathbf{x}^k) &\in \mathbb{M}_k(\mathbb{G}, s), \quad k = 1, \cdots, K \end{aligned} \tag{11}$$

where the vector $\mathbf{x} \in \mathbb{R}^N$ is partitioned into multiple disjoint blocks $\mathbf{x}^1 \in \mathbb{R}^{N_1}, \cdots, \mathbf{x}^K \in \mathbb{R}^{N_K}$, f is a continuous differentiable and convex function, each g_k is extended-valued and possibly nondifferentiable convex function, $\mathrm{supp}(\mathbf{x}^k)$ denotes the support set of vector $\mathbf{x}^k$, $\mathbb{M}_k(\mathbb{G}, s)$ denotes all possible subsets of vertices in $\mathbb{G}$ that satisfy a certain predefined topological constraint. The functions f and g_k will be defined based on the feature matrix $\mathbf{W}$, and can be used to formulate the cost function and dependencies among blocks.

One example of topological constraint for defining $\mathbb{M}_k(\mathbb{G}, s)$ is connected subgraph, and we can formally define it as follows:

$$\mathbb{M}_k(\mathbb{G}, s) \coloneqq \{S | S \subseteq \mathbb{V}^k; |S| \leq s; \mathbb{G}_S \text{ is connected.}\} \tag{12}$$

where s is a predefined upperbound size of S, $S \subseteq \mathbb{V}^k$, and $\mathbb{G}_S$ refers to the induced subgraph by a set of vertices S. The topological constraints can be any graph structured sparsity constraints on $\mathbb{G}_S$, such as connected subgraphs, dense subgraphs, compact subgraphs [2]. Moreover, we do not restrict all $\mathrm{supp}(\mathbf{x}^1), \cdots, \mathrm{supp}(\mathbf{x}^K)$ satisfy an identical topological constraint.

Algorithm 5 Graph Block-structured Matching Pursuit

Initialization, $i = 0$, $\mathbf{x}^{k,i}$ = initial vectors, k=1,..., K
repeat
  for $k = 1, \cdots, K$ **do**
    $\Gamma_{\mathbf{x}^k} = H(\nabla_{\mathbf{x}^k} F(\mathbf{x}^{1,i}, \ldots, \mathbf{x}^{K,i}))$
    $\Omega_{\mathbf{x}^k} = \Gamma_{\mathbf{x}^k} \cup \mathrm{supp}(\mathbf{x}^{k,i})$
  end for
  Get $(\mathbf{b}^i_{\mathbf{x}^1}, \ldots, \mathbf{b}^i_{\mathbf{x}^K})$ by solving problem (**??**)
  for $k = 1, \cdots, K$ **do**
    $\Psi^{i+1}_{\mathbf{x}^k} = \mathrm{T}(\mathbf{b}^i_{\mathbf{x}^k})$
    $\mathbf{x}^{k,i+1} = [\mathbf{b}^i_{\mathbf{x}^k}]_{\Psi^{i+1}_{\mathbf{x}^k}}$
  end for
  $i = i + 1$
until $\sum_{k=1}^{K} \left\|\mathbf{x}^{k,i+1} - \mathbf{x}^{k,i}\right\| \leq \epsilon$
$C = (\Psi^i_{x^1}, \ldots, \Psi^i_{\mathbf{x}^k})$
return $(\mathbf{x}^{1,i}, \cdots, \mathbf{x}^{K,i}), C$

Algorithm 6 Graph Block-structured Iterative Hard Thresholding (GBIHT)

Input: Input graph $\mathbb{G}$, maximum subgraph size k, and step size η (1 by default).
Output: The estimated vector $\hat{\mathbf{x}}$ and the corresponding connected subgraph S.
Initialization, $i = 0$, $\mathbf{x}^i = (\mathbf{x}^{1,i}, \ldots, \mathbf{x}^{K,i})$, $\mathbf{x}^{k,i}$ = initial vectors, k=1,..., K
$i \leftarrow 0$, $\mathbf{x}^i \leftarrow \mathbf{0}$, $S^i \leftarrow \emptyset$
repeat
  for $k = 1, \ldots, K$ **do**
    $\Omega_k \leftarrow \mathrm{H}(\nabla_{\mathbf{x}^k} f(\mathbf{x}^{1,i}, \ldots, \mathbf{x}^{K,i}))$
  end for
  $\Omega = \bigcup_{k=1}^{K} \Omega_k$
  $\mathbf{b}^i = \mathbf{x}^i - \eta \cdot \nabla_\Omega f(\mathbf{x}^i)$ or $\mathbf{b}^i_{\mathbf{x}^k} = \mathbf{x}^{k,i} - \eta \cdot \nabla_{\Omega_k} f(\mathbf{x}^i), k = 1, \ldots, K$
  for k=1,...,K **do**
    $S^{k,i+1} \leftarrow \mathrm{T}(\mathbf{b}^i_{\mathbf{x}^k})$
    $\mathbf{x}^{k,i+1} \leftarrow \mathbf{b}^i_{S^{k,i+1}}$
  end for
  $i \leftarrow i + 1$
until halting condition holds
return $\hat{\mathbf{x}} = \mathbf{x}^i$ and $S = \mathbb{G}_{S^i}$

6.2 Algorithm

6.3 Theoretical Analysis

7 Graph Block-Structured Iterative Hard Thresholding

7.1 Algorithm

7.2 Theoretical Analysis

Theorem 6. *zzz*

8 Graph Block-Structured Gradient Hard Thresholding Pursuit

8.1 Algorithm

Algorithm 7 Graph Block-structured Gradient Hard Thresholding Pursuit (GBGHTP)

Input: Input graph $\mathbb{G}$, maximum subgraph size k, and step size η (1 by default).
Output: The estimated vector $\hat{\mathbf{x}}$ and the corresponding connected subgraph S.
Initialization, $i = 0$, $\mathbf{x}^i = (\mathbf{x}^{1,i}, \ldots, \mathbf{x}^{K,i})$, $\mathbf{x}^{k,i}$ = initial vectors, k=1,. . . , K
repeat
  for $k = 1, \ldots, K$ **do**
    $\Omega_k \leftarrow \mathrm{H}(\nabla_{\mathbf{x}^k} f(\mathbf{x}^i))$
  end for
  $\Omega = \bigcup_{k=1}^{k} \Omega_k$
  $\Psi \leftarrow \mathrm{supp}(\mathbf{x}^i - \eta \cdot \nabla_\Omega f(\mathbf{x}^i))$
  $\mathbf{b}^i \leftarrow \arg\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad \mathrm{supp}(\mathbf{x}) \subseteq \Psi$
  for k=1,. . . ,K **do**
    $S^{k,i+1} \leftarrow \mathrm{T}(\mathbf{b}^i_{\mathbf{x}^k})$
    $\mathbf{x}^{k,i+1} \leftarrow \mathbf{b}_{S^{k,i+1}}$
  end for
  $i \leftarrow i + 1$
until halting condition holds
return $\hat{\mathbf{x}} = \mathbf{x}^i$ and $S = \mathbb{G}_{S^i}$

8.2 Theoretical Analysis

A Proof

A.1 Proof of Graph-Mp

Lemma 1. *Assume that f is a differentiable function. If f satisfies condition $(\xi, \delta, \mathbb{M})$-WRSC, then $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ with $\text{supp}(\mathbf{x}) \cup \text{supp}(\mathbf{y}) \subseteq S \in \mathbb{M}$, the following two inequalities hold [4]*

$$\frac{1-\delta}{\xi}\|\mathbf{x}-\mathbf{y}\|_2 \leq \|\nabla_S f(\mathbf{x}) - \nabla_S f(\mathbf{y})\|_2 \leq \frac{1+\delta}{\xi}\|\mathbf{x}-\mathbf{y}\|_2$$
$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{x}), \mathbf{x}-\mathbf{y}\rangle + \frac{1+\delta}{2\xi}\|\mathbf{x}-\mathbf{y}\|_2$$

Lemma 2. *Let $\eta = c_H(1-\delta)-\delta, \rho = \xi(1+c_H), \mathbf{r}^i = \mathbf{x}^i - \mathbf{x}$, and $\Omega = \text{H}(\nabla f(\mathbf{x}^i))$. Then*

$$\|\mathbf{r}^i_{\Omega^c}\|_2 \leq \sqrt{1-\eta^2}\|\mathbf{r}^i\|_2 + \left[\frac{\rho}{\eta} + \frac{\eta\rho}{\sqrt{1-\eta^2}}\right]\|\nabla_I f(\mathbf{x})\|_2 \tag{13}$$

where $I = \arg\max_{S\in\mathbb{M}(\mathbb{G},8k)} \|\nabla_S f(\mathbf{x})\|_2$. We assume that c_H and δ are such that $\eta > 0$.

Proof. Denote $\Phi = \text{supp}(\mathbf{x}) \in \mathbb{M}(\mathbb{G}, k), \Omega = \text{H}(\nabla f(\mathbf{x}^i)) \in \mathbb{M}(\mathbb{G}, 2k), \mathbf{r}^i = \mathbf{x}^i - \mathbf{x}$, and $\Gamma = \text{supp}(\mathbf{r}^i) \in \mathbb{M}(\mathbb{G}, 6k)$. The component $\|\nabla_\Omega f(\mathbf{x}^i)\|_2$ can be lower bounded as

$$\begin{aligned}
\|\nabla_\Omega f(\mathbf{x}^i)\|_2 \geq& c_H\|\nabla_\Phi f(\mathbf{x}^i)\|_2 \\
\geq& c_H(\|\nabla_\Phi f(\mathbf{x}^i) - \nabla_\Phi f(\mathbf{x})\|_2 - \|\nabla_\Phi f(\mathbf{x})\|_2) \\
\geq& \frac{c_H(1-\delta)}{\xi}\|\mathbf{r}^i\|_2 - c_H\|\nabla_I f(\mathbf{x})\|_2
\end{aligned}$$

where the first inequality follows from the definition of head approximation and the last inequality follows from Lemma 1. The component $\|\nabla_\Omega f(\mathbf{x}^i)\|_2$ can be also upper bounded as

$$\begin{aligned}
\|\nabla_\Omega f(\mathbf{x}^i)\|_2 \leq& \frac{1}{\xi}\|\xi\nabla_\Omega f(\mathbf{x}^i) - \xi\nabla_\Omega f(\mathbf{x})\|_2 + \|\nabla_\Omega f(\mathbf{x})\|_2 \\
\leq& \frac{1}{\xi}\|\xi\nabla_\Omega f(\mathbf{x}^i) - \xi\nabla_\Omega f(\mathbf{x}) - \mathbf{r}^i_\Omega + \mathbf{r}^i_\Omega\|_2 + \|\nabla_\Omega f(\mathbf{x})\|_2 \\
\leq& \frac{1}{\xi}\|\xi\nabla_{\Gamma\cup\Omega} f(\mathbf{x}^i) - \xi\nabla_{\Gamma\cup\Omega} f(\mathbf{x}) - \mathbf{r}^i_{\Gamma\cup\Omega}\|_2 + \|\mathbf{r}^i_\Omega\|_2 + \|\nabla_\Omega f(\mathbf{x})\|_2 \\
\leq& \frac{\delta}{\xi}\|\mathbf{r}^i\|_2 + \frac{1}{\xi}\|\mathbf{r}^i_\Omega\|_2 + \|\nabla_I f(\mathbf{x})\|_2
\end{aligned}$$

where the fourth inequality follows from condition $(\xi, \delta, \mathbb{M}(\mathbb{G}, 8k))$-WRSC and the fact that $\mathbf{r}^i_{\Gamma\cup\Omega} = \mathbf{r}^i$. Combining the two bounds and grouping terms, we obtain

$$\begin{aligned}
\|\mathbf{r}^i_\Omega\|_2 \geq& (c_H(1-\delta)-\delta)\|\mathbf{r}^i\|_2 - \xi(1+c_H)\|\nabla_I f(\mathbf{x})\|_2 \\
=& \eta\|\mathbf{r}^i\|_2 - \rho\|\nabla_I f(\mathbf{x})\|_2, \eta = c_H(1-\delta)-\delta, \rho = \xi(1+c_H)
\end{aligned} \tag{14}$$

We have $\|\mathbf{r}^i_\Omega\|_2 \geq \eta\|\mathbf{r}^i\|_2 - \rho\|\nabla_I f(\mathbf{x})\|_2$. In order to otain an upper bound of $\|\mathbf{r}^i_{\Omega^c}\|_2$, we consider two cases.

- Case 1: The right hand side of (14) is ≤ 0, i.e., $\alpha_0\|\mathbf{r}^i\|_2 \leq \beta_0\|\nabla_I f(\mathbf{x})\|_2$. Then we have

$$\|\mathbf{r}^i_{\Omega^c}\|_2 \leq \|\mathbf{r}^i\|_2 \leq \frac{\rho}{\eta}\|\nabla_I f(\mathbf{x})\|_2$$

- Case 1: The right hand side of (14) is > 0, i.e., $\eta\|\mathbf{r}^i\|_2 > \rho\|\nabla_I f(\mathbf{x})\|_2$. Then we have

$$\|\mathbf{r}^i_\Omega\|_2 \geq \left(\eta - \frac{\rho\|\nabla_I f(\mathbf{x})\|_2}{\|\mathbf{r}^i\|_2}\right)\|\mathbf{r}^i\|_2$$

Moreover, notice that $\|\mathbf{r}^i_\Omega\|_2^2 = \|\mathbf{r}^i\|_2^2 - \|\mathbf{r}^c_{\Omega^c}\|_2^2$. Then we have obtain

$$\|\mathbf{r}^i_\Omega\|_2^2 = \|\mathbf{r}^i\|_2^2 - \|\mathbf{r}^i_{\Omega^c}\|_2^2$$

$$\|\mathbf{r}^i_{\Omega^c}\|_2 \leq \|\mathbf{r}^i\|_2\sqrt{1-\left(\eta-\frac{\rho\|\nabla_I f(\mathbf{x})\|_2}{\|\mathbf{r}^i\|_2}\right)^2}$$

Denote $\omega_0 = \eta - \rho\|\nabla_I f(\mathbf{x})\|_2/\|\mathbf{r}^i\|_2$. For a given $0 < \omega_0 < 1$ and a free parameter $0 < \omega < 1$, a straight forward calculation yields that $\sqrt{1-\omega_0^2} \leq \frac{1}{\sqrt{1-\omega^2}} - \frac{\omega}{1-\omega^2}\omega_0$. Therefore, substituting into the bound for $\|\mathbf{r}^i_{\Omega^c}\|_2$, we get

$$\begin{aligned}\|\mathbf{r}^i_{\Omega^c}\|_2 \leq& \|\mathbf{r}^i\|_2\left(\frac{1}{\sqrt{1-\omega^2}} - \frac{\omega}{1-\omega^2}\left(\eta-\frac{\rho\|\nabla_I f(\mathbf{x})\|_2}{\|\mathbf{r}^i\|_2}\right)\right)\\ =&\frac{1-\omega\eta}{\sqrt{1-\omega^2}}\|\mathbf{r}^i\|_2 + \frac{\omega\rho}{\sqrt{1-\omega^2}}\|\nabla_I f(\mathbf{x})\|_2\end{aligned}$$

The coefficient preceding $\|\mathbf{r}^i\|_2$ determines the overall convergence rate, and the minimum value of the coefficient is attained by setting $\omega = \eta$ (derived from $(\frac{1-\omega\eta}{\sqrt{1-\omega^2}})' = 0$).

Therefore, by combining the two cases, we obtain

$$\|\mathbf{r}^i_{\Omega^c}\|_2 \leq \sqrt{1-\eta^2}\|\mathbf{r}^i\|_2 + \left[\frac{\rho}{\eta} + \frac{\eta\rho}{\sqrt{1-\eta^2}}\right]\|\nabla_I f(\mathbf{x})\|_2$$

which proves the lemma. □

Proof of Theorem 1

Proof. Let $\mathbf{r}^{i+1} = \mathbf{x}^{i+1} - \mathbf{x}$. $\|\mathbf{r}^{i+1}\|_2$ is upper bounded as

$$\begin{aligned}\|\mathbf{r}^{i+1}\|_2 =&\|\mathbf{x}^{i+1} - \mathbf{x}\|_2\\ &\text{x is the optimum of our problem}\\ \leq&\|\mathbf{x}^{i+1} - \mathbf{b}\|_2 + \|\mathbf{x} - \mathbf{b}\|_2\\ \leq&c_T \cdot \|\mathbf{x}^* - \mathbf{b}\|_2 + \|\mathbf{x} - \mathbf{b}\|_2\\ &\mathbf{x}^* \text{ is the optimum of } \min_{\mathbf{x}'}\|\mathbf{x}' - \mathbf{b}\|_2\\ \leq&c_T\|\mathbf{x} - \mathbf{b}\|_2 + \|\mathbf{x} - \mathbf{b}\|_2\\ &\|\mathbf{x}^* - \mathbf{b}\|_2 \leq \|\mathbf{x} - \mathbf{b}\|_2, \mathbf{x}^*, \mathbf{x} \in \mathbb{M}(\mathbb{G}, 5s)\\ =&(1 + c_T)\|\mathbf{x} - \mathbf{b}\|_2\end{aligned}$$

which follows from the definition of tail approximation. The component $\|(\mathbf{x} - \mathbf{b})_\Omega\|_2^2$ is upper bounded as

$$\begin{aligned}\|(\mathbf{x} - \mathbf{b})_\Omega\|_2^2 =& \langle \mathbf{b} - \mathbf{x}, (\mathbf{b} - \mathbf{x})_\Omega\rangle\\ =& \langle \mathbf{b} - \mathbf{x} - \xi\nabla_\Omega f(\mathbf{b}) + \xi\nabla_\Omega f(\mathbf{x}), (\mathbf{b} - \mathbf{x})_\Omega\rangle - \langle \xi\nabla_\Omega f(\mathbf{x}), (\mathbf{b} - \mathbf{x})_\Omega\rangle\\ \leq&\|\mathbf{b} - \mathbf{x} - \xi\nabla_\Omega f(\mathbf{b}) + \xi\nabla_\Omega f(\mathbf{x})\|_2\|(\mathbf{b} - \mathbf{x})_\Omega\|_2 + \xi\|\nabla_\Omega f(\mathbf{x})\|_2\|(\mathbf{b} - \mathbf{x})_\Omega\|_2\\ \leq&\delta\|\mathbf{b} - \mathbf{x}\|_2\|(\mathbf{b} - \mathbf{x})_\Omega\|_2 + \xi\|\nabla_\Omega f(\mathbf{x})\|_2\|(\mathbf{b} - \mathbf{x})_\Omega\|_2\end{aligned}$$

where the second equality follows from the fact that $\nabla_\Omega f(\mathbf{b}) = 0$ since $\mathbf{b}$ is the solution to the problem in step 7 of Algorithm 1, and the last inequality follows from condition $(\xi, \delta, \mathbb{M}(8k, g))$-WRSC. After simplification, we have

$$\|(\mathbf{x} - \mathbf{b})_\Omega\|_2 \leq \delta\|\mathbf{b} - \mathbf{x}|_2 + \xi\|\nabla_\Omega f(\mathbf{x})\|_2$$

It follows that

$$\begin{aligned}\|(\mathbf{x} - \mathbf{b})\|_2 \leq&\|(\mathbf{x} - \mathbf{b})_\Omega\|_2 + \|(\mathbf{x} - \mathbf{b})_{\Omega^c}\|_2\\ \leq&\delta\|\mathbf{b} - \mathbf{x}\|_2 + \xi\|\nabla_\Omega f(\mathbf{x})\|_2 + \|(\mathbf{x} - \mathbf{b})_{\Omega^c}\|_2\end{aligned}$$

After rearrangement, we obtain

$$\begin{aligned}\|\mathbf{b} - \mathbf{x}\|_2 \leq&\frac{\|(\mathbf{b} - \mathbf{x})_{\Omega^c}\|_2}{1-\delta} + \frac{\xi\|\nabla_\Omega f(\mathbf{x})\|_2}{1-\delta}\\ =&\frac{\|\mathbf{x}_{\Omega^c}\|_2}{1-\delta} + \frac{\xi\|\nabla_\Omega f(\mathbf{x})\|_2}{1-\delta}\end{aligned}$$

$$
\begin{aligned}
=&\frac{\|(\mathbf{x}-\mathbf{x}^i)_{\Omega^c}\|_2}{1-\delta}+\frac{\nabla_\Omega f(\mathbf{x})}{1-\delta}\\
=&\frac{\|\mathbf{r}^i_{\Omega^c}\|_2}{1-\delta}+\frac{\xi\|\nabla_\Omega f(\mathbf{x})\|_2}{1-\delta}\\
\leq&\frac{\|\mathbf{r}^i_{\Gamma^c}\|_2}{1-\delta}+\frac{\xi\|\nabla_\Omega f(\mathbf{x})\|_2}{1-\delta}
\end{aligned}
$$

where the first equality follows from the fact that $\mathrm{supp}(\mathbf{b})\subseteq\Omega$, the second and last inequalities follow from the fact that $\Omega=\Gamma\cup\mathrm{supp}(\mathbf{x}^i)$. Combining above inequalities, we obtain

$$
\|\mathbf{r}^{i+1}_{\Gamma^c}\|_2\leq(1+c_T)\frac{\|\mathbf{r}^i_{\Omega^c}\|_2}{1-\delta}+(1+c_T)\frac{\xi\|\nabla_I f(\mathbf{x})\|_2}{1-\delta}
$$

From Lemma 2, we have

$$
\|\mathbf{r}^i_{\Omega^c}\|_2\leq\|\mathbf{r}^i_{\Gamma^c}\|_2\leq\sqrt{1-\eta^2}\|\mathbf{r}^i\|_2+\left[\frac{\rho}{\eta}+\frac{\eta\rho}{\sqrt{1-\eta}}\right]\|\nabla_I f(\mathbf{x})\|_2
$$

Combining the above inequalities, we prove the theorem. □

Proof of Theorem 2

Proof. The i-the iteration of Algorithm 1 satisfies

$$
\|\mathbf{x}-\mathbf{x}^i\|_2\leq\alpha^i\|\mathbf{x}\|_2+\frac{\beta}{1-\alpha}\|\nabla_I f(\mathbf{x})\|_2 \tag{15}
$$

After $t=\left\lceil\log\left(\frac{\|\mathbf{x}\|_2}{\|\nabla_I f(\mathbf{x})\|_2}\right)/\log\frac{1}{\alpha}\right\rceil$ iterations, Algorithm 1 returns an estimate $\hat{\mathbf{x}}$ satisfying $\|\mathbf{x}-\hat{\mathbf{x}}\|_2\leq(1+\frac{\beta}{1-\alpha})\|\nabla_I f(\mathbf{x})\|_2$. The time complexities of both head and tail approximations are $O(|\mathbb{E}|\log^3 n)$. The time complexity of one iteration in Algorithm 1 is $(T+|\mathbb{E}|\log^3 n)$, and the total number of iterations is $\left\lceil\log\left(\frac{\|\mathbf{x}\|_2}{\|\nabla_I f(\mathbf{x})\|_2}\right)\right\rceil$, and the overall time complexity follows. □

A.2 Proof of Graph-IHT

Proof of Theorem 3

Proof. From the triangle inequality, we have

$$
\begin{aligned}
\|\mathbf{r}^{i+1}\|_2=&\|\mathbf{x}^{i+1}-\mathbf{x}\|_2\\
=&\|\mathbf{x}^{i+1}-\mathbf{b}+\mathbf{b}-\mathbf{x}\|_2\\
\leq&\|\mathbf{x}^{i+1}-\mathbf{b}\|_2+\|\mathbf{b}-\mathbf{x}\|_2\\
\leq&(1+c_T)\|\mathbf{b}-\mathbf{x}\|_2\\
=&(1+c_T)\|\mathbf{x}^i-\eta\nabla_\Gamma f(\mathbf{x}^i)-\mathbf{x}\|_2\\
=&(1+c_T)\|\mathbf{r}^i-\eta\nabla_\Gamma f(\mathbf{x}^i)\|_2
\end{aligned}
$$

where $\nabla_\Gamma f(\mathbf{x}^i)$ is the projected vector of $f(\mathbf{x}^i)$ in which the entries outside Ω are set to zero and the entries in Ω are unchanged. $\|\mathbf{r}^i-\eta\nabla_\Omega f(\mathbf{x}^i)\|_2$ has the inequalities

$$
\begin{aligned}
\|\mathbf{r}^i-\eta\nabla_\Omega f(\mathbf{x}^i)\|_2=&\|\mathbf{r}^i_{\Omega^c}+\mathbf{r}^i_\Omega-\eta\nabla_\Omega f(\mathbf{x}^i)\|_2\\
\leq&\|\mathbf{r}^i_{\Omega^c}\|_2+\|\mathbf{r}^i_\Omega-\eta\nabla_\Omega f(\mathbf{x}^i)+\eta\nabla_\Omega f(\mathbf{x})-\eta\nabla_\Omega f(\mathbf{x})\|_2\\
\leq&\|\mathbf{r}^i_{\Omega^c}\|_2+\|\mathbf{r}^i_\Omega-\eta\nabla_\Omega f(\mathbf{x}^i)+\eta\nabla_\Omega f(\mathbf{x})\|_2+\|\eta\nabla_\Omega f(\mathbf{x})\|_2\\
\leq&\|\mathbf{r}^i_{\Omega^c}\|_2+\|\mathbf{r}^i_\Omega-\xi\nabla_\Omega f(\mathbf{x}^i)+\xi\nabla_\Omega f(\mathbf{x})\|_2\\
&+(\xi-\eta)\|\nabla_\Omega f(\mathbf{x}^i)-\nabla_\Omega f(\mathbf{x})\|_2+\|\eta\nabla_\Omega f(\mathbf{x})\|_2\\
\leq&\|\mathbf{r}^i_{\Omega^c}\|_2+\delta\|\mathbf{r}^i\|_2+\frac{(1+\delta)(\xi-\eta)}{\xi}\|\mathbf{r}^i\|_2+\eta\|\nabla_I f(\mathbf{x})\|_2\\
=&\|\mathbf{r}^i_{\Omega^c}\|_2+\frac{\xi+2\xi\delta-\eta-\eta\delta}{\xi}\|\mathbf{r}^i\|_2+\eta\|\nabla_I f(\mathbf{x})\|_2\\
=&\|\mathbf{r}^i_{\Omega^c}\|_2+(1-\frac{\eta}{\xi}+(2-\frac{\eta}{\xi})\delta)\|\mathbf{r}^i\|_2+\eta\|\nabla_I f(\mathbf{x})\|_2
\end{aligned}
$$

where the last inequality follows from condition $(\xi, \delta, \mathbb{M})$-WRSC and Lemma 1. From Lemma 2, we have

$$\|\mathbf{r}^i_{\Omega^c}\|_2 \leq \sqrt{1-\eta^2}\|\mathbf{r}^i\|_2 + \left[\frac{\rho}{\eta} + \frac{\eta\rho}{\sqrt{1-\eta^2}}\right]\|\nabla_I f(\mathbf{x})\|_2$$

Combining the above inequalities, we prove the theorem.

$$\begin{aligned}
\|\mathbf{r}^{i+1}\|_2 =&(1+c_T)\|\mathbf{r}^i - \eta\nabla_\Gamma f(\mathbf{x}^i)\|_2 \\
\leq&(1+c_T)\left(\|\mathbf{r}^i_{\Omega^c}\|_2 + (1-\frac{\eta}{\xi}+(2-\frac{\eta}{\xi})\delta)\|\mathbf{r}^i\|_2 + \eta\|\nabla_I f(\mathbf{x})\|_2\right) \\
\leq&(1+c_T)\left(\sqrt{1-\eta^2}\|\mathbf{r}^i\|_2 + \left[\frac{\rho}{\eta} + \frac{\eta\rho}{\sqrt{1-\eta^2}}\right]\|\nabla_I f(\mathbf{x})\|_2\right. \\
&\left. + (1-\frac{\eta}{\xi}+(2-\frac{\eta}{\xi})\delta)\|\mathbf{r}^i\|_2 + \eta\|\nabla_I f(\mathbf{x})\right) \\
=&(1+c_T)\left(\sqrt{1-\eta^2}+1-\frac{\eta}{\xi}+(2-\frac{\eta}{\xi})\delta\right)\|\mathbf{r}^i\|_2 + (1+c_T)\left(\eta+\frac{\rho}{\eta}+\frac{\eta\rho}{\sqrt{1-\eta^2}}\right)\|\nabla_I f(\mathbf{x})\|_2
\end{aligned}$$

□

A.3 Proof of Graph-GHTP

Proof of Theorem 4

Proof. Denote $\Omega = \mathrm{H}(f(\mathbf{x}^i))$ and $\Psi = \mathrm{supp}(\mathbf{x}^i - \eta\cdot\nabla_\Omega f(\mathbf{x}^i))$. Let $\mathbf{r}^{i+1} = \mathbf{x}^{i+1} - \mathbf{x}$. $\|\mathbf{r}^{i+1}\|_2$ is bounded as

$$\begin{aligned}
\|\mathbf{r}^{i+1}\|_2 =&\|\mathbf{x}^{i+1} - \mathbf{x}\|_2 \\
=&\|\mathbf{x}^{i+1} - \mathbf{b} + \mathbf{b} - \mathbf{x}\|_2 \\
\leq&\|\mathbf{x}^{i+1} - \mathbf{x}\|_2 + \|\mathbf{x} - \mathbf{b}\|_2 \\
\leq&c_T\|\mathbf{x} - \mathbf{b}\|_2 + \|\mathbf{x} - \mathbf{b}\|_2 \\
\leq&(1+c_T)\|\mathbf{x} - \mathbf{b}\|_2
\end{aligned} \tag{16}$$

where the second inequality follows from the definition of tail approximation. The component $\|(\mathbf{x}-\mathbf{b})_\Psi\|_2^2$ is bounded as

$$\begin{aligned}
\|(\mathbf{x}-\mathbf{b})_\Psi\|_2^2 =&\langle \mathbf{b}-\mathbf{x}, (\mathbf{b}-\mathbf{x})_\Psi\rangle \\
=&\langle \mathbf{b}-\mathbf{x}-\xi\nabla_\Psi f(\mathbf{b}) + \xi\nabla_\Psi f(\mathbf{x}), (\mathbf{b}-\mathbf{x})_\Psi\rangle - \langle \xi\nabla_\Psi f(\mathbf{x}), (\mathbf{b}-\mathbf{x})_\Psi\rangle \\
\leq&\|\mathbf{b}-\mathbf{x}-\xi\nabla_\Psi f(\mathbf{b}) + \xi\nabla_\Psi f(\mathbf{x})\|_2\|(\mathbf{b}-\mathbf{x})_\Psi\| + \xi\|\nabla_\Psi f(\mathbf{x})\|_2\|(\mathbf{b}-\mathbf{x})_\Psi\|_2 \\
\leq&\delta\|\mathbf{b}-\mathbf{x}\|_2\|(\mathbf{b}-\mathbf{x})_\Psi\|_2 + \xi\|\nabla_\Psi f(\mathbf{x})\|_2\|(\mathbf{b}-\mathbf{x})_\Psi\|_2
\end{aligned}$$

where the second equality follows from the fact that $\nabla_S f(\mathbf{b}) = \mathbf{0}$ since $\mathbf{b}$ is the solution to the problem in the third step (Line 7) of Graph-GHTP, and the last inequality can be derived from condition $(\xi, \delta, \mathbb{M}(\mathbb{G}, 8k))$-WRSC. After simplification, we have

$$\|(\mathbf{x}-\mathbf{b})_\Psi\|_2 \leq \delta\|\mathbf{b}-\mathbf{x}\|_2 + \xi\|\nabla_\Psi f(\mathbf{x})\|_2$$

It follows that

$$\begin{aligned}
\|\mathbf{x}-\mathbf{b}\|_2 \leq&\|(\mathbf{x}-\mathbf{b})_\Psi\|_2 + \|(\mathbf{x}-\mathbf{b})_{\Psi^c}\|_2 \\
\leq&\delta\|\mathbf{b}-\mathbf{x}\|_2 + \xi\|\nabla_\Psi f(\mathbf{x})\|_2 + \|(\mathbf{x}-\mathbf{b})_{\Psi^c}\|_2
\end{aligned}$$

After rearrangement, we obtain

$$\|\mathbf{b}-\mathbf{x}\|_2 \leq \frac{\|(\mathbf{b}-\mathbf{x})_{\Psi^c}\|_2}{1-\delta} + \frac{\xi\|\nabla_\Psi f(\mathbf{x})\|_2}{1-\delta} \tag{17}$$

where this equality follows from the fact that $\mathrm{supp}(\mathbf{b}) \subseteq S$. Let $\Phi = \mathrm{supp}(\mathbf{x}) \in \mathbb{M}(\mathbb{G}, k)$.

$$\|(\mathbf{x}^i - \eta\nabla_\Omega f(\mathbf{x}^i))_\Phi\|_2 \leq \|(\mathbf{x}^i - \eta\nabla_\Omega f(\mathbf{x}^i))_\Psi\|_2$$

as $\Psi = \text{supp}(\mathbf{x}^i - \eta\nabla_\Omega f(\mathbf{x}^i))$. By eliminating the contribution on $\Phi \cap \Psi$, we derive

$$\|(\mathbf{x}^i - \eta\nabla_\Omega f(\mathbf{x}^i))_{\Phi\setminus\Psi}\|_2 \leq \|(\mathbf{x}^i - \eta\nabla_\Omega f(\mathbf{x}^i))_{\Psi\setminus\Phi}\|_2$$

For the right-hand side, we have

$$\|(\mathbf{x}^i - \eta\nabla_\Omega f(\mathbf{x}^i))_{\Psi\setminus\Phi}\|_2 \leq \|(\mathbf{x}^i - \mathbf{x} - \eta\nabla_\Omega f(\mathbf{x}^i) + \eta\nabla_\Omega f(\mathbf{x}))_{\Psi\setminus\Phi}\|_2 + \eta\|\nabla_{\Omega\cup\Psi} f(\mathbf{x})\|_2$$

where the inequality falls from the fact that $\Phi = \text{supp}(\mathbf{x})$. From the left-hand side, we have

$$\|(\mathbf{x}^i - \eta\nabla_\Omega f(\mathbf{x}^i))_{\Phi\setminus\Psi}\|_2 \leq -\eta\|\nabla_{\Omega\cup\Phi} f(\mathbf{x})\|_2 + \|(\mathbf{x}^i - \mathbf{x} - \eta\nabla_\Omega f(\mathbf{x}^i) + \eta\nabla_\Omega f(\mathbf{x}))_{\Phi\setminus\Psi} + (\mathbf{x} - \mathbf{b})_{\Psi^c}\|_2$$

where the inequality follows from the fact that $\mathbf{b}_{\Psi^c} = \mathbf{0}, \mathbf{x}_{\Phi\setminus\Psi} = \mathbf{x}_{\Psi^c}$, and $-\mathbf{x}_{\Phi\setminus\Psi} + (\mathbf{x} - \mathbf{b})_{\Psi^c} = \mathbf{0}$. Let $\Phi\Delta\Psi$ be the symmetric difference of the set Φ and Ψ. It follows that

$$\begin{aligned}
\|(\mathbf{b} - \mathbf{x})_{\Psi^c}\|_2 \leq& \sqrt{2}\|(\mathbf{x}^i - \mathbf{x} - \eta\nabla_\Omega f(\mathbf{x}^i) + \eta\nabla_\Omega f(\mathbf{x}))_{\Phi\Delta\Psi}\|_2 + 2\eta\|\nabla_I f(\mathbf{x})\|_2 \\
\leq& \sqrt{2}\|(\mathbf{x}^i - \mathbf{x} - \eta\nabla_\Omega f(\mathbf{x}^i) + \eta\nabla_\Omega f(\mathbf{x}))_{\Phi\Delta\Psi}\|_2 + 2\eta\|\nabla_I f(\mathbf{x})\|_2 \\
\leq& \sqrt{2}\|(\mathbf{x}^i - \mathbf{x} - \xi\nabla_\Omega f(\mathbf{x}^i) + \xi\nabla_\Omega f(\mathbf{x}))_{\Phi\Delta\Psi}\|_2 \\
&+ \sqrt{2}(\xi - \eta)\|(\nabla_\Omega - \nabla_\Omega f(\mathbf{x}))_{\Phi\Delta\Psi}\|_2 + 2\eta\|\nabla_I f(\mathbf{x})\|_2 \\
=& \sqrt{2}\|(\mathbf{r}^i_{\Omega^c} + \mathbf{r}^i_\Omega - \xi\nabla_\Omega f(\mathbf{x}^i) + \xi\nabla_\Omega f(\mathbf{x}))_{\Phi\Delta\Psi}\|_2 \\
&+ \sqrt{2}(\xi - \eta)\|(\nabla_\Omega f(\mathbf{x}^i) - \nabla_\Omega f(\mathbf{x}))_{\Phi\Delta\Psi}\|_2 + 2\eta\|\nabla_I f(\mathbf{x})\|_2 \\
\leq& \sqrt{2}\|\mathbf{r}^i_{\Omega^c}\|_2 + \sqrt{2}\|(\mathbf{r}^i_\Omega - \xi\nabla_\Omega f(\mathbf{x}^i) + \xi\nabla_\Omega f(\mathbf{x}))_{\Phi\Delta\Psi}\|_2 \\
&+ \sqrt{2}(\xi - \eta)\|(\nabla_\Omega f(\mathbf{x}^i) - \nabla_\Omega f(\mathbf{x}))_{\Phi\Delta\Psi}\|_2 + 2\eta\|\nabla_I f(\mathbf{x})\|_2 \\
\leq& \sqrt{2}\|\mathbf{r}^i_{\Omega^c}\|_2 + \sqrt{2}\|\mathbf{r}^i - \xi\nabla_{\Omega\cup\Psi\cup\Phi} f(\mathbf{x}^i) + \xi\nabla_{\Omega\cup\Psi\cup\Phi} f(\mathbf{x})\|_2 \\
&+ \sqrt{2}(\xi - \eta)\|(\nabla_{\Omega\cup\Psi\cup\Phi} f(\mathbf{x}^i) - \nabla_{\Omega\cup\Psi\cup\Phi} f(\mathbf{x}))_{\Phi\Delta\Psi}\|_2 + 2\eta\|\nabla_I f(\mathbf{x})\|_2 \\
\leq& \sqrt{2}\|\mathbf{r}^i_{\Omega^c}\|_2 + \sqrt{2}\left(\left(2 - \frac{\eta}{\xi}\right)\delta + 1 - \frac{\eta}{\xi}\right)\|\mathbf{r}^i\|_2 + 2(\sqrt{2}\xi + (1 - \sqrt{2})\eta)\|\nabla_I f(\mathbf{x})\|_2
\end{aligned}$$

where the first inequality follows from the fact that

$$\eta\|\nabla_{\Omega\cup\Phi} f(\mathbf{x})\|_2 + \eta\|\nabla_{\Psi\cup\Phi\cup\Omega} f(\mathbf{x})\|_2 \leq 2\eta\|\nabla_I f(\mathbf{x})\|_2$$

the third inequality follows as $\mathbf{x}^i - \mathbf{x} = \mathbf{r}^i = \mathbf{r}^i_{\Omega^c} + \mathbf{r}^i_\Omega$, the fourth inequality follows from the fact that $\|(\mathbf{r}^i_{\Omega^c})_{\Phi\Delta\Psi}\|_2 \leq \|\mathbf{r}^i_{\Omega^c}\|_2$, the sixth inequality follows as $\mathbf{r}^i \subseteq \Omega \cup \Psi \cup \Phi$, and the last inequality follows from condition $(\xi, \delta, \mathbb{M}(\mathbb{G}, 8k))$-WRSC and Lemma 1. From Lemma 2, we have

$$\|\mathbf{r}^i_{\Omega^c}\|_2 \leq \sqrt{1 - \alpha_0^2}\|\mathbf{r}\|_2 + \left[\frac{\beta_0}{\alpha_0} + \frac{\alpha_0\beta_0}{\sqrt{1 - \alpha_0^2}}\right]\|\nabla_I f(\mathbf{x})\|_2$$

Combining (13) and above inequalities, we prove the theorem. □

Proof of Theorem 5

A.4 Proof of SG-Pursuit

A.5 Proof of GBMP

A.6 Proof of GBIHT

Proof of Theorem 6

Proof. From the triangle inequality, we have

$$\begin{aligned}
\|\mathbf{r}^{i+1}\|_2 =& \|\mathbf{x}^{i+1} - \mathbf{x}\|_2 \\
& \mathbf{x} \text{ is the optimum of our problem} \\
\leq& \|\mathbf{x}^{i+1} - \mathbf{b}\|_2 + \|\mathbf{x} - \mathbf{b}\|_2 \\
\leq& c_T \cdot \|\mathbf{x}^* - \mathbf{b}\|_2 + \|\mathbf{x} - \mathbf{b}\|_2 \\
& \mathbf{x}^* \text{ is the optimum of } \min_{\mathbf{x}'} \|\mathbf{x}' - \mathbf{b}\|_2
\end{aligned}$$

$$
\begin{aligned}
&\leq c_T\|\mathbf{x}-\mathbf{b}\|_2+\|\mathbf{x}-\mathbf{b}\|_2\\
&\quad \|\mathbf{x}^*-\mathbf{b}\|_2\leq\|\mathbf{x}-\mathbf{b}\|_2, \mathbf{x}^*, \mathbf{x}\in\mathbb{M}(\mathbb{G}, ???)\\
&=(1+c_T)\|\mathbf{x}-\mathbf{b}\|_2
\end{aligned}
$$

□

A.7 Proof of GBGHTP

References

[1] Chen, F., and Zhou, B. A generalized matching pursuit approach for graph-structured sparsity. In *IJCAI* (2016), pp. 1389–1395.

[2] Chen, F., Zhou, B., Alim, A., and Zhao, L. A generic framework for interesting subspace cluster detection in multi-attributed networks. In *2017 IEEE International Conference on Data Mining (ICDM)* (2017), IEEE, pp. 41–50.

[3] Hegde, C., Indyk, P., and Schmidt, L. A nearly-linear time framework for graph-structured sparsity. In *International Conference on Machine Learning* (2015), pp. 928–937.

[4] Yuan, X., Li, P., and Zhang, T. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *International Conference on Machine Learning* (2014), pp. 127–135.

[5] Zhou, B., and Chen, F. Graph-structured sparse optimization for connected subgraph detection. In *2016 IEEE 16th International Conference on Data Mining (ICDM)* (2016), IEEE, pp. 709–718.