Stochastic Iterative Hard Thresholding for Graph-Structured Sparsity Optimization

Baojian Zhou[1], Feng Chen[1], and Yiming Ying[2]

[1]Department of Computer Science,
[2]Department of Mathematics and Statistics,
University at Albany, NY, USA

06/13/2019
Poster # 92

Graph structure information as a prior often have:

- better classification, regression performance
- stronger interpretation

Current limitations:

- only focus on specific loss
- expensive full-gradient calculation
- cannot handle complex structure

Our goals propose/provide:

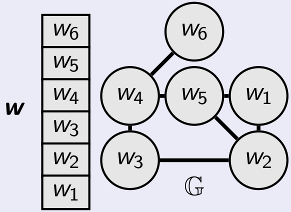
- an algo. for general loss under stochastic setting
- convergence analysis
- real-world applications

Structured sparse learning

Given $\mathcal{M}(\mathbb{M}) = \{\boldsymbol{w} : \text{supp}(\boldsymbol{w}) \in \mathbb{M}\}$, the structured sparse learning problems can be formulated as

$$\min_{\boldsymbol{w} \in \mathcal{M}(\mathbb{M})} F(\boldsymbol{w}) := \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{w}), \text{ where}$$

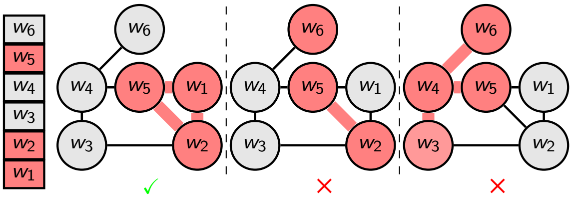
- $F(\boldsymbol{w})$ is a convex loss such as least square, logistic loss, ...
- $\mathcal{M}(\mathbb{M})$ models structured sparsity such as connected subgraphs, dense subgraphs, and subgraphs isomophic to a query graph, ...



Inspired by two recent works Hegde et al. (2016); Nguyen et al. (2017)

Algorithm 1 GraphStoIHT

1: **Input**: $\eta_t, F(\cdot), \mathbb{M}_{\mathcal{H}}, \mathbb{M}_{\mathcal{T}}$
2: **Initialize**: $\boldsymbol{w}^0$ and $t = 0$
3: **for** $t = 0, 1, 2, \ldots$ **do**
4: Choose ξ_t from $[n]$ with prob. p_{ξ_t}
5: $\boldsymbol{b}^t = \mathrm{P}(\nabla f_{\xi_t}(\boldsymbol{w}^t), \mathbb{M}_{\mathcal{H}})$
6: $\boldsymbol{w}^{t+1} = \mathrm{P}(\boldsymbol{w}^t - \eta_t \boldsymbol{b}^t, \mathbb{M}_{\mathcal{T}})$
7: **end for**
8: **Return** $\boldsymbol{w}^{t+1}$



Weighted Graph Model
$\mathbb{M} = \{S : |S| \le 3, S$ is connected $\}$ (Hegde et al., 2015a)

Orthogonal Projection Operator $\mathrm{P}(\cdot, \mathbb{M}) : \mathbb{R}^p \to \mathbb{R}^p$ defined as

$$\mathrm{P}(\boldsymbol{w}, \mathbb{M}) = \operatorname*{arg\,min}_{\boldsymbol{w}' \in \mathcal{M}(\mathbb{M})} \|\boldsymbol{w} - \boldsymbol{w}'\|^2$$

- s-sparse set
- Weighted Graph Model

Two differences from StoIHT:

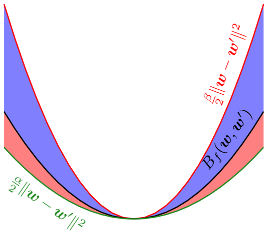
- project the gradient $\nabla f_{\xi_t}(\cdot)$
- projects the proxy onto $\mathcal{M}(\mathbb{M}_{\mathcal{T}})$.

Why projection $\boldsymbol{b}^t = \mathrm{P}(\nabla f_{\xi_t}(\boldsymbol{w}^t), \mathbb{M}_{\mathcal{H}})$?

- Both of them solve the same projection problem
- Intuitively, sparsity is both in primal and dual space
- Remove some noisy directions at the first stage

Two assumptions in $\mathcal{M}(\mathbb{M})$:

1. $f_i(\boldsymbol{w})$: β-Restricted Strong Smoothness
   $F(\boldsymbol{w})$: α-Restricted Strong Convexity
2. Efficient Approximated projections:
   - $\mathrm{P}(\cdot, \mathbb{M}_{\mathcal{H}})$ with approximation factor $c_{\mathcal{H}}$
   - $\mathrm{P}(\cdot, \mathbb{M}_{\mathcal{T}})$ with approximation factor $c_{\mathcal{T}}$



$B_f(\boldsymbol{w}, \boldsymbol{w}') = f(\boldsymbol{w}) - f(\boldsymbol{w}') - \langle \nabla f(\boldsymbol{w}'), \boldsymbol{w} - \boldsymbol{w}' \rangle$

Theorem 1 (Linear Convergence)

Let $\boldsymbol{w}^0$ be the start point and choose $\eta_t = \eta$, then $\boldsymbol{w}^{t+1}$ of Algorithm 1 satisfies

$$\mathbb{E}_{\xi_{[t]}} \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\| \leq \kappa^{t+1} \|\boldsymbol{w}^0 - \boldsymbol{w}^*\| + \frac{\sigma}{1-\kappa},$$

where

$$\kappa = (1 + c_{\mathcal{T}})\left(\sqrt{\alpha\beta\eta^2 - 2\alpha\eta + 1} + \sqrt{1 - \alpha_0^2}\right), \alpha_0 = c_{\mathcal{H}}\alpha\tau - \sqrt{\alpha\beta\tau^2 - 2\alpha\tau + 1}, \beta_0 = (1 + c_{\mathcal{H}})\tau$$

$$\sigma = \left(\frac{\beta_0}{\alpha_0} + \frac{\alpha_0\beta_0}{\sqrt{1-\alpha_0^2}}\right)\mathbb{E}_{\xi_t}\|\nabla_I f_{\xi_t}(\boldsymbol{w}^*)\| + \eta\mathbb{E}_{\xi_t}\|\nabla_I f_{\xi_t}(\boldsymbol{w}^*)\|, \text{ and } \eta, \tau \in (0, 2/\beta).$$

Graph Linear Regression

$\boldsymbol{X} \in \mathbb{R}^{m\times p}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$ $\xrightarrow{\boldsymbol{w}^*:}$  $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}^* + \boldsymbol{\epsilon}$

Consider the least square loss

$$\arg\min_{\text{supp}(\boldsymbol{w})\in\mathcal{M}(\mathbb{M})} F(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n}\frac{n}{2m}\|\boldsymbol{X}_{B_i}\boldsymbol{w} - \boldsymbol{y}_{B_i}\|^2.$$

Contraction factor

Algorithm	κ
GraphIHT	$(1+c_{\mathcal{T}})\left(\sqrt{\delta}+2\sqrt{1-\delta}\right)\sqrt{\delta}$
GraphStoIHT	$(1+c_{\mathcal{T}})\left(\sqrt{\frac{2}{1+\delta}}+\frac{2\sqrt{2(1-\delta)}}{1+\delta}\right)\sqrt{\delta}$

- For GraphIHT, $\delta \leq 0.0527$
- For GraphStoIHT, $\delta \leq 0.0142$

Graph Logistic Regression

$\boldsymbol{x}_i \in \mathbb{R}^p, y_i \in \{+1, -1\}$ $\xrightarrow{\boldsymbol{w}^*:}$  $(1 + e^{-y_i\cdot\langle \boldsymbol{w}^*, \boldsymbol{x}_i\rangle})^{-1}$

Consider the logistic loss

$$\arg\min_{\text{supp}(\boldsymbol{w})\in\mathcal{M}(\mathbb{M})} F(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n}\frac{n}{m}\sum_{j=1}^{m/n} h(\boldsymbol{w}, i_j) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2,$$

where $h(\boldsymbol{w}, i_j) = \log(1+\exp(-y_{i_j}\cdot\langle \boldsymbol{x}_{i_j}, \boldsymbol{w}\rangle))$.

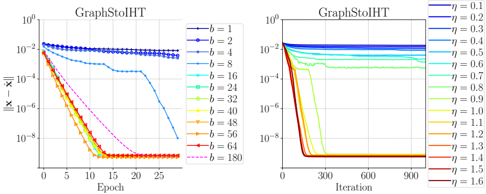
If $\boldsymbol{x}_i$ is normalized, then $F(\boldsymbol{w})$ satisfies λ-RSC and each $f_i(\boldsymbol{w})$ satisfies $(\alpha + (1+\nu)\theta_{max})$-RSS. The condition of $\kappa < 1$ is

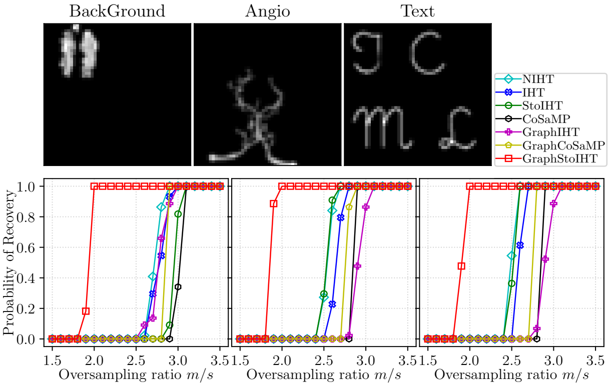
$$\frac{\lambda}{\lambda + n(1+\nu)\theta_{max}/4m} \geq \frac{243}{250},$$

with prob. $1 - p\exp(-\theta_{max}\nu/4)$, where $\theta_{max} = \lambda_{max}(\sum_{j=1}^{m/n}\mathbb{E}[\boldsymbol{x}_{i_j}\boldsymbol{x}_{i_j}^{\mathsf{T}}])$ and $\nu \geq 1$.

Simulation Dataset

- each entry $\sqrt{m}\boldsymbol{X}_{ij} \sim \mathcal{N}(0,1)$
- $\mathrm{supp}(\boldsymbol{w}^*)$ is generated by random walk
- Entries of $\boldsymbol{w}^*$ from $\mathcal{N}(0,1)$
- Weighted Graph Model (Hegde et al., 2015b)





Breast Cancer Dataset

- 295 samples with 78 positives (metastatic) and 217 negatives (non-metastatic) provided in (Van De Vijver et al., 2002).
- PPI network with 637 pathways is provided in (Jacob et al., 2009). We restrict our analysis on 3,243 genes (nodes) with 19,938 edges. These cancer-related genes form a connected subgraph.

Algorithm	Cancer related genes	$\|\boldsymbol{w}^t\|_0$	AUC
GraphStoIHT	BRCA2, CCND2, CDKN1A, ATM, AR, TOP2A	051.7	**0.715**
GraphIHT	ATM, CDKN1A, BRCA2, AR, TOP2A	055.2	0.714
ℓ^1-Path	BRCA1, CDKN1A, ATM, DSC2	061.2	0.675
StoIHT	MKI67, NAT1, AR, TOP2A	059.6	0.708
ℓ^1/ℓ^2-Edge	CCND3, ATM, CDH3	051.4	0.705
ℓ^1-Edge	CCND3, AR, CDH3	039.9	0.698
ℓ^1/ℓ^2-Path	BRCA1, CDKN1A	147.6	0.705
IHT	NAT1, TOP2A	067.9	0.707

See you at Poster #92

Thank you!