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

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

- $F(w)$ is a convex loss such as least square, logistic loss, ... 
- $\mathcal{M}(\mathcal{M})$ models structured sparsity such as connected subgraphs, dense subgraphs, and subgraphs isomorphic to a query graph, ...
Algorithm 1 \textbf{GraphStoIHT}

1: \textbf{Input:} $\eta_t, F(\cdot), \mathcal{M}_H, \mathcal{M}_T$
2: \textbf{Initialize:} $w^0$ and $t = 0$
3: \textbf{for} $t = 0, 1, 2, \ldots$ \textbf{do}
4: \hspace{1em} Choose $\xi_t$ from $[n]$ with prob. $p_{\xi_t}$
5: \hspace{1em} $b^t = P(\nabla f_{\xi_t}(w^t), \mathcal{M}_H)$
6: \hspace{1em} $w^{t+1} = P(w^t - \eta_t b^t, \mathcal{M}_T)$
7: \textbf{end for}
8: \textbf{Return} $w^{t+1}$

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

$$P(w, \mathcal{M}) = \arg \min_{w' \in \mathcal{M}(\mathcal{M})} \|w - w'\|^2$$

- $s$-sparse set
- Weighted Graph Model

Two differences from \textbf{StoIHT}:
- project the gradient $\nabla f_{\xi_t}(\cdot)$
- projects the proxy onto $\mathcal{M}(\mathcal{M}_T)$. 

Why projection $b^t = P(\nabla f_{\xi_t}(w^t), \mathcal{M}_H)$?
- 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 $\mathcal{M}(\mathcal{M})$:

1. $f_i(w)$: $\beta$-Restricted Strong Smoothness
2. $F(w)$: $\alpha$-Restricted Strong Convexity

Efficient Approximated projections:
- $P(\cdot, \mathcal{M}_H)$ with approximation factor $c_H$
- $P(\cdot, \mathcal{M}_T)$ with approximation factor $c_T$

Theorem 1 (Linear Convergence)

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

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

where

$$\kappa = (1 + c_T) \left( \sqrt{\alpha \beta \eta^2 - 2 \alpha \eta + 1} + \sqrt{1 - \alpha_0^2} \right), \quad \alpha_0 = c_H \alpha \tau - \sqrt{\alpha \beta \tau^2 - 2 \alpha \tau + 1}, \quad \beta_0 = (1 + c_H) \tau$$

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

\[ \mathbf{X} \in \mathbb{R}^{m \times p}, \; \mathbf{\epsilon} \sim \mathcal{N}(0, I_m) \]
\[ \mathbf{y} = \mathbf{Xw}^* + \mathbf{\epsilon} \]

Consider the least square loss

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

Graph Logistic Regression

\[ \mathbf{x}_i \in \mathbb{R}^p, \; y_i \in \{+1, -1\} \]
\[ (1 + e^{-y_i \cdot \langle \mathbf{w}^*, \mathbf{x}_i \rangle})^{-1} \]

Consider the logistic loss

\[ \arg \min_{\text{supp}(\mathbf{w}) \in \mathcal{M}(\mathcal{M})} F(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} \frac{n}{m} \sum_{j=1}^{m/n} h(\mathbf{w}, i_j) + \frac{\lambda}{2} \| \mathbf{w} \|^2, \]
where \( h(\mathbf{w}, i_j) = \log(1 + \exp(-y_{i_j} \cdot \langle \mathbf{x}_{i_j}, \mathbf{w} \rangle)) \).

Contraction factor

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>( \kappa )</th>
</tr>
</thead>
<tbody>
<tr>
<td>GraphIHT</td>
<td>((1 + c_T)(\sqrt{\delta} + 2\sqrt{1 - \delta})\sqrt{\delta})</td>
</tr>
<tr>
<td>GraphStoIHT</td>
<td>((1 + c_T)(\sqrt{\frac{2}{1+\delta}} + \frac{2\sqrt{2(1-\delta)}}{1+\delta})\sqrt{\delta})</td>
</tr>
</tbody>
</table>

- For GraphIHT, \( \delta \leq 0.0527 \)
- For GraphStoIHT, \( \delta \leq 0.0142 \)

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

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

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

- each entry $\sqrt{m}X_{ij} \sim \mathcal{N}(0, 1)$
- $\text{supp}(\mathbf{w}^*)$ is generated by random walk
- Entries of $\mathbf{w}^*$ 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.
See you at Poster #92

Thank you!