Tree-projected gradient descent for estimating gradient-sparse parameters on graphs

Sheng Xu   Zhou Fan   Sahand Negahban

Yale University
Department of Statistics and Data Science

COLT 2020

Model Setup

- **Data:** Samples $Z_1^n := (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ are drawn from an unknown distribution $\mathcal{P}$.
- **Loss Function:** $\mathcal{L} : \mathbb{R}^p \times \mathcal{Z}^n \to \mathbb{R}$ is convex and differentiable.

Model Setup

- **Data:** Samples $Z_1^n := (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ are drawn from an unknown distribution $\mathcal{P}$.
- **Loss Function:** $\mathcal{L} : \mathbb{R}^p \times \mathcal{Z}^n \to \mathbb{R}$ is convex and differentiable.
- **Goal:** Find estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^* \in \mathbb{R}^p$ where

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathbb{E}_{\mathcal{P}}\big[\mathcal{L}(\boldsymbol{\theta}; Z_1^n)\big].$$

Model Setup

- **Data:** Samples $Z_1^n := (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ are drawn from an unknown distribution $\mathcal{P}$.
- **Loss Function:** $\mathcal{L} : \mathbb{R}^p \times \mathcal{Z}^n \to \mathbb{R}$ is convex and differentiable.
- **Goal:** Find estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^* \in \mathbb{R}^p$ where

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg\min}\, \mathbb{E}_{\mathcal{P}}\big[\mathcal{L}(\boldsymbol{\theta}; Z_1^n)\big].$$

- Example: Linear models

$$y_i = \mathbf{x}_i^\top \boldsymbol{\theta}^* + e_i,$$

where $Z_i = (\mathbf{x}_i, y_i)$ and $\mathcal{L}(\boldsymbol{\theta}; Z_1^n) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$.

Model Setup

- Identify $\boldsymbol{\theta}^* \in \mathbb{R}^p$ with the vertices of a known graph $G = (V, E)$ where $|V| = p$.

Model Setup

- Identify $\boldsymbol{\theta}^* \in \mathbb{R}^p$ with the vertices of a known graph $G = (V, E)$ where $|V| = p$.
- **Discrete gradient operator** $\nabla_G : \mathbb{R}^p \to \mathbb{R}^{|E|}$:

$$\nabla_G \boldsymbol{\theta} = \big(\theta_i - \theta_j : (i, j) \in E\big).$$

Model Setup

- Identify $\boldsymbol{\theta}^* \in \mathbb{R}^p$ with the vertices of a known graph $G = (V, E)$ where $|V| = p$.
- **Discrete gradient operator** $\nabla_G : \mathbb{R}^p \to \mathbb{R}^{|E|}$:

$$\nabla_G \boldsymbol{\theta} = \big(\theta_i - \theta_j : (i, j) \in E\big).$$

- Assume the **gradient sparsity**

$$s^* := \|\nabla_G \boldsymbol{\theta}^*\|_0$$

is small relative to $|E|$.

Model Setup

- Identify $\boldsymbol{\theta}^* \in \mathbb{R}^p$ with the vertices of a known graph $G = (V, E)$ where $|V| = p$.
- **Discrete gradient operator** $\nabla_G : \mathbb{R}^p \to \mathbb{R}^{|E|}$:

$$\nabla_G \boldsymbol{\theta} = \big(\theta_i - \theta_j : (i,j) \in E\big).$$

- Assume the **gradient sparsity**

$$s^* := \|\nabla_G \boldsymbol{\theta}^*\|_0$$

  is small relative to $|E|$.
- Find graph-sparse $\widehat{\boldsymbol{\theta}}$ with small $\mathcal{L}(\boldsymbol{\theta})$

Motivating Examples

- Statistical changepoint detection

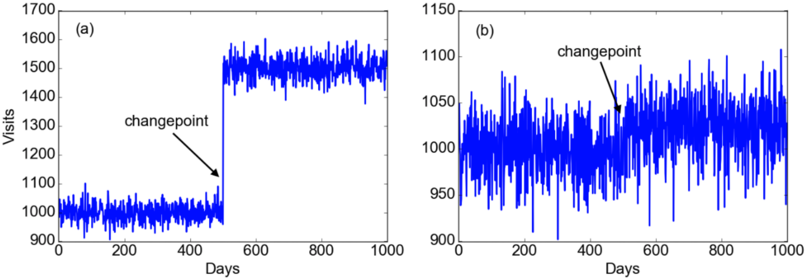


Figure 1: History of visits on a website for 1000 days.

Motivating Examples

- Image denoising and compressed sensing

Figure 2: Four cameraman images.

Motivating Examples

- Anomaly detection

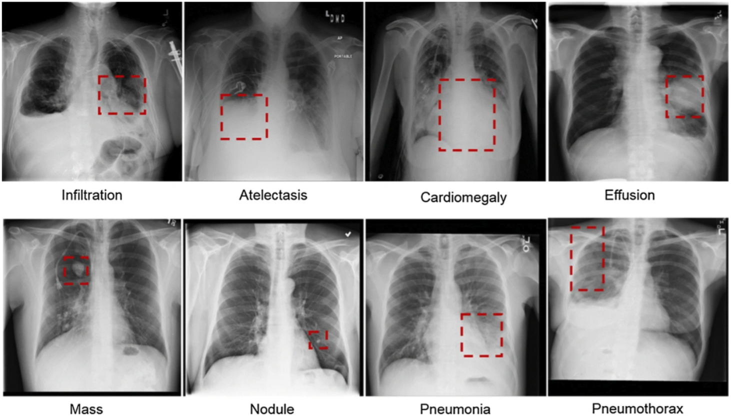


Figure 3: Eight common diseases observed in the chest radiographs. Retrieved from https://doi.org/10.1186/s12938-018-0544-y. Copyright by Qin, C., Yao, D., Shi, Y. et al. Computer-aided detection in chest radiography based on artificial intelligence: a survey. BioMed Eng OnLine 17, 113 (2018).

Tree-Projected Gradient Descent

- Estimation guarantee for the linear model is

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq C \cdot \sqrt{\frac{s^*}{n} \log\left(1 + \frac{p}{s^*}\right)}$$

$$\downarrow$$

independent of G

Tree-Projected Gradient Descent

- Estimation guarantee for the linear model is

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq C \cdot \sqrt{\frac{s^*}{n} \log\left(1 + \frac{p}{s^*}\right)}$$

$$\downarrow$$

independent of G

- Comparison with convex approaches:
  - Well conditioned discrete gradient matrix $\nabla_G \in \mathbb{R}^{|E| \times p}$ (Hütter and Rigollet '16)
  - For line graph $\mathbf{X} = \mathbf{I}$

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \sqrt{s^* \log(p)}$$

  - Improved rate with minimum spacing requirement between changepoints of $\boldsymbol{\theta}^*$ (Dalalyan et al. '17, Guntuboyina et al. '17)

Tree-Projected Gradient Descent

- **Idea:** non-convex projected gradient descent
- IHT (Blumensath and Davies '08 and Jain et al. '14), CoSaMP (Needell and Tropp '09), HTP (Foucart '11).

$$\boldsymbol{\theta}_t = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^p : \|\nabla_G \boldsymbol{\theta}\|_0 \leq S} \|\boldsymbol{\theta} - \mathbf{u}_t\|_2,$$

where $\mathbf{u}_t = \boldsymbol{\theta}_{t-1} - \eta \cdot \nabla \mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$.

Tree-Projected Gradient Descent

- **Idea:** non-convex projected gradient descent
- IHT (Blumensath and Davies '08 and Jain et al. '14), CoSaMP (Needell and Tropp '09), HTP (Foucart '11).

$$\boldsymbol{\theta}_t = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^p : \|\nabla_G \boldsymbol{\theta}\|_0 \leq S} \|\boldsymbol{\theta} - \mathbf{u}_t\|_2,$$

where $\mathbf{u}_t = \boldsymbol{\theta}_{t-1} - \eta \cdot \nabla \mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$.

- Performing projection step is intractable in general!

Tree-Projected Gradient Descent

- **Idea:** non-convex projected gradient descent
- IHT (Blumensath and Davies '08 and Jain et al. '14), CoSaMP (Needell and Tropp '09), HTP (Foucart '11).

$$\boldsymbol{\theta}_t = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^p : \|\nabla_G \boldsymbol{\theta}\|_0 \leq S} \|\boldsymbol{\theta} - \mathbf{u}_t\|_2,$$

where $\mathbf{u}_t = \boldsymbol{\theta}_{t-1} - \eta \cdot \nabla \mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$.

- Performing projection step is intractable in general!
- Approximate G with a tree T_t at each iteration

Tree-Projected Gradient Descent

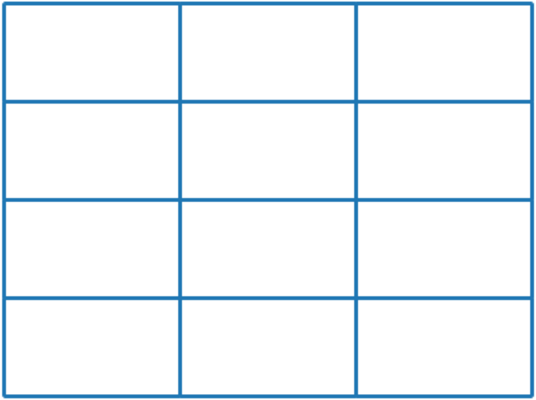
- **Step 1:** Tree Construction

  Construct a sequence of spanning trees $T_1, T_2, \ldots$ with maximum degree $d_{\max}$ such that $\boldsymbol{\theta}^*$ remains gradient-sparse over these trees

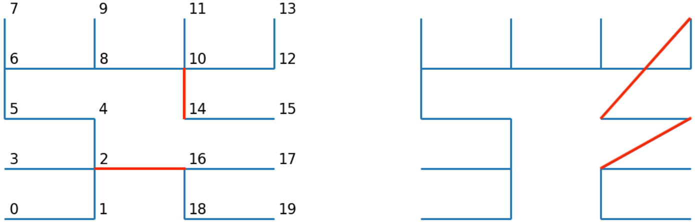
- **Step 2:** Projected Gradient Approximation

  Perform a single projected gradient descent step on each tree in this sequence over a discrete domain

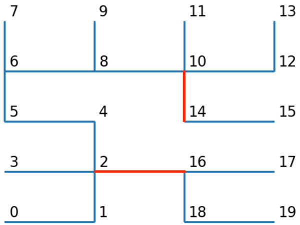
Tree Construction

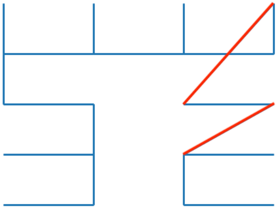


Tree Construction



Tree Construction





Lemma (Padilla et al '17)

Let T be as constructed above. Then T is a tree on V with maximum degree $\leq d_{\max}$. Furthermore, for any $\boldsymbol{\theta} \in \mathbb{R}^p$,

$$\|\nabla_T \boldsymbol{\theta}\|_0 \leq 2\|\nabla_G \boldsymbol{\theta}\|_0.$$

The computational complexity for constructing T is $O(|E|)$.

Projected Gradient Approximation

- Iteration step

$$\boldsymbol{\theta}_t \approx \underset{\boldsymbol{\theta} \in \mathbb{R}^p : \|\nabla_{T_t} \boldsymbol{\theta}\|_0 \leq S}{\arg\min} \|\boldsymbol{\theta} - \mathbf{u}_t\|_2,$$

where $\mathbf{u}_t = \boldsymbol{\theta}_{t-1} - \eta \cdot \nabla \mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$.

- Optimize over $\boldsymbol{\theta}$ in a discrete domain Δ^p rather than $\mathbb{R}^p$ where

$$\Delta := \left\{\Delta_{\min}, \Delta_{\min} + \delta, \Delta_{\min} + 2\delta, \ldots, \Delta_{\max} - \delta, \Delta_{\max}\right\}.$$

Computational Complexity for Linear Model

Total computational complexity for the linear model:

$$O\left(\left(np + p^2\sqrt{n}(s^*)^{d_{\max}-3/2}\right)\log np\right)$$

Cut-Restricted Strong Convexity/Smoothness

Definition (cRSC and cRSS)

A differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ satisfies **cut-restricted strong convexity (cRSC) and smoothness (cRSS)** with respect to (T_1, T_2), at sparsity level S and with constants $\alpha, L > 0$, if the following holds: For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in K := K_1 + K_2$ where K_i is the subspace of all S-gradient-sparse vectors with respect to T_i,

$$f(\boldsymbol{\theta}_2) \geq f(\boldsymbol{\theta}_1) + \langle \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1, \nabla f(\boldsymbol{\theta}_1) \rangle + \frac{\alpha}{2} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2^2 \quad \text{(cRSC)},$$

$$f(\boldsymbol{\theta}_2) \leq f(\boldsymbol{\theta}_1) + \langle \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1, \nabla f(\boldsymbol{\theta}_1) \rangle + \frac{L}{2} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2^2 \quad \text{(cRSS)}.$$

Cut-Projected Gradient Bound

Definition (cPGB)

A differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ has a **cut-projected gradient bound (cPGB)** of $\Phi(S)$ with respect to (T_1, T_2), at a point $\boldsymbol{\theta}^* \in \mathbb{R}^p$ and sparsity level S, if: For any $K := K_1 + K_2$ where K_i is the subspace of all S-gradient-sparse vectors with respect to T_i,

$$\|\mathbf{P}_K \nabla f(\boldsymbol{\theta}^*)\|_2 \leq \Phi(S).$$

Cut-Projected Gradient Bound

Definition (cPGB)

A differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ has a **cut-projected gradient bound (cPGB)** of $\Phi(S)$ with respect to (T_1, T_2), at a point $\boldsymbol{\theta}^* \in \mathbb{R}^p$ and sparsity level S, if: For any $K := K_1 + K_2$ where K_i is the subspace of all S-gradient-sparse vectors with respect to T_i,

$$\|\mathbf{P}_K \nabla f(\boldsymbol{\theta}^*)\|_2 \leq \Phi(S).$$

Lemma (cPGB)

If $\mathbf{w}^\top \nabla \mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)$ *is* σ^2/n*-subgaussian for any* $\mathbf{w} \in K$*. Then* $\Phi(S) \asymp \sigma \sqrt{\frac{S}{n} \log\left(1 + \frac{p}{S}\right)}$ *with high probability.*

Main Theorem

Theorem (Tree-PGD Deterministic Estimation Guarantee)

Suppose $\|\nabla_G \boldsymbol{\theta}^*\|_0 \leq s^*$. *Set* $S = \kappa s^*$ *for a constant* κ. *Suppose, for all* $1 \leq t \leq \tau$ *and* (T_{t-1}, T_t), *that*

1. $\mathcal{L}(\cdot\,; Z_1^n)$ *satisfies cRSC and cRSS with constants* $\alpha, L > 0$ *at sparsity level* S.
2. $\mathcal{L}(\cdot\,; Z_1^n)$ *has the cPGB* $\Phi(S)$ *at the point* $\boldsymbol{\theta}^*$ *and sparsity level* S.

Let $\Gamma \approx \sqrt{1 - \alpha/L} \cdot (1 + \sqrt{2d_{\max}/\kappa})$ *and suppose* κ *is large enough such that* $\Gamma < 1$. *Then the* τ^{th} *iterate* $\boldsymbol{\theta}_\tau$ *of tree-PGD satisfies*

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}^*\|_2 \lesssim \Gamma^\tau \cdot \|\boldsymbol{\theta}^*\|_2 + \Phi(S).$$

For $\hat{\boldsymbol{\theta}} \equiv \boldsymbol{\theta}_\tau$ *and* τ *large enough,this yields*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \lesssim \Phi(S).$$

Main Proof Idea

Construct $K \ni \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}^*$, gradient-sparsity $\approx S + 2s^*$, applying

$$\begin{aligned}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2 &\leq \|\mathbf{P}_K \mathbf{u}_t - \boldsymbol{\theta}_t\|_2 + \|\mathbf{P}_K \mathbf{u}_t - \boldsymbol{\theta}^*\|_2 \\ &\leq \|\mathbf{P}_K \mathbf{u}_t - \boldsymbol{\theta}_t\|_2 + \|\mathbf{P}_K \mathbf{u}_t - \mathbf{v}\|_2 + \|\mathbf{v} - \boldsymbol{\theta}^*\|_2,\end{aligned}$$

$\mathbf{u}_t = \boldsymbol{\theta}_{t-1} - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$ and $\mathbf{v} = \arg\min_{\boldsymbol{\theta}\in K} \mathcal{L}(\boldsymbol{\theta}; Z_1^n)$

Main Proof Idea

Construct $K \ni \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}^*$, gradient-sparsity $\approx S + 2s^*$, applying

$$\begin{aligned}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2 &\le \|\mathbf{P}_K\mathbf{u}_t - \boldsymbol{\theta}_t\|_2 + \|\mathbf{P}_K\mathbf{u}_t - \boldsymbol{\theta}^*\|_2 \\ &\le \|\mathbf{P}_K\mathbf{u}_t - \boldsymbol{\theta}_t\|_2 + \|\mathbf{P}_K\mathbf{u}_t - \mathbf{v}\|_2 + \|\mathbf{v} - \boldsymbol{\theta}^*\|_2,\end{aligned}$$

$\mathbf{u}_t = \boldsymbol{\theta}_{t-1} - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$ and $\mathbf{v} = \arg\min_{\boldsymbol{\theta}\in K} \mathcal{L}(\boldsymbol{\theta}; Z_1^n)$

Step 1. Inspired by Jain et al. '14:

$$\|\mathbf{P}_K\mathbf{u}_t - \boldsymbol{\theta}_t\|_2 \le \gamma \cdot \|\mathbf{P}_K\mathbf{u}_t - \boldsymbol{\theta}^*\|_2, \qquad \gamma := \sqrt{2d_{\max}/\kappa}.$$

Main Proof Idea

Construct $K \ni \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}^*$, gradient-sparsity $\approx S + 2s^*$, applying

$$\begin{aligned}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2 &\le \|\mathbf{P}_K\mathbf{u}_t - \boldsymbol{\theta}_t\|_2 + \|\mathbf{P}_K\mathbf{u}_t - \boldsymbol{\theta}^*\|_2 \\ &\le \|\mathbf{P}_K\mathbf{u}_t - \boldsymbol{\theta}_t\|_2 + \|\mathbf{P}_K\mathbf{u}_t - \mathbf{v}\|_2 + \|\mathbf{v} - \boldsymbol{\theta}^*\|_2,\end{aligned}$$

$\mathbf{u}_t = \boldsymbol{\theta}_{t-1} - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$ and $\mathbf{v} = \arg\min_{\boldsymbol{\theta}\in K} \mathcal{L}(\boldsymbol{\theta}; Z_1^n)$

Step 1. Inspired by Jain et al. '14:

$$\|\mathbf{P}_K\mathbf{u}_t - \boldsymbol{\theta}_t\|_2 \le \gamma \cdot \|\mathbf{P}_K\mathbf{u}_t - \boldsymbol{\theta}^*\|_2, \qquad \gamma := \sqrt{2d_{\max}/\kappa}.$$

Step 2. Property of gradient mapping and cRSC/cRSS give

$$\begin{aligned}\|\mathbf{P}_K\mathbf{u}_t - \mathbf{v}\|_2 &\le \sqrt{1 - \alpha/L} \cdot \|\boldsymbol{\theta}_{t-1} - \mathbf{v}\|_2 \\ &\le \sqrt{1 - \alpha/L} \cdot (\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|_2 + \|\mathbf{v} - \boldsymbol{\theta}^*\|_2).\end{aligned}$$

Main Proof Idea

Construct $K \ni \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}^*$, gradient-sparsity $\approx S + 2s^*$, applying

$$\begin{aligned}
\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2 &\leq \|\mathbf{P}_K \mathbf{u}_t - \boldsymbol{\theta}_t\|_2 + \|\mathbf{P}_K \mathbf{u}_t - \boldsymbol{\theta}^*\|_2 \\
&\leq \|\mathbf{P}_K \mathbf{u}_t - \boldsymbol{\theta}_t\|_2 + \|\mathbf{P}_K \mathbf{u}_t - \mathbf{v}\|_2 + \|\mathbf{v} - \boldsymbol{\theta}^*\|_2,
\end{aligned}$$

$\mathbf{u}_t = \boldsymbol{\theta}_{t-1} - \frac{1}{L} \nabla \mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$ and $\mathbf{v} = \arg\min_{\boldsymbol{\theta} \in K} \mathcal{L}(\boldsymbol{\theta}; Z_1^n)$

Step 1. Inspired by Jain et al. '14:

$$\|\mathbf{P}_K \mathbf{u}_t - \boldsymbol{\theta}_t\|_2 \leq \gamma \cdot \|\mathbf{P}_K \mathbf{u}_t - \boldsymbol{\theta}^*\|_2, \qquad \gamma := \sqrt{2 d_{\max}/\kappa}.$$

Step 2. Property of gradient mapping and cRSC/cRSS give

$$\begin{aligned}
\|\mathbf{P}_K \mathbf{u}_t - \mathbf{v}\|_2 &\leq \sqrt{1 - \alpha/L} \cdot \|\boldsymbol{\theta}_{t-1} - \mathbf{v}\|_2 \\
&\leq \sqrt{1 - \alpha/L} \cdot (\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|_2 + \|\mathbf{v} - \boldsymbol{\theta}^*\|_2).
\end{aligned}$$

Step 3. cRSC and cPGB give $\|\mathbf{v} - \boldsymbol{\theta}^*\|_2 \leq C\Phi(S)$.

Main Proof Idea

Construct $K \ni \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}^*$, gradient-sparsity $\approx S + 2s^*$, applying

$$\begin{aligned}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2 &\le \|\mathbf{P}_K \mathbf{u}_t - \boldsymbol{\theta}_t\|_2 + \|\mathbf{P}_K \mathbf{u}_t - \boldsymbol{\theta}^*\|_2 \\ &\le \|\mathbf{P}_K \mathbf{u}_t - \boldsymbol{\theta}_t\|_2 + \|\mathbf{P}_K \mathbf{u}_t - \mathbf{v}\|_2 + \|\mathbf{v} - \boldsymbol{\theta}^*\|_2,\end{aligned}$$

$\mathbf{u}_t = \boldsymbol{\theta}_{t-1} - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$ and $\mathbf{v} = \arg\min_{\boldsymbol{\theta}\in K} \mathcal{L}(\boldsymbol{\theta}; Z_1^n)$

Step 1. Inspired by Jain et al. '14:

$$\|\mathbf{P}_K \mathbf{u}_t - \boldsymbol{\theta}_t\|_2 \le \gamma \cdot \|\mathbf{P}_K \mathbf{u}_t - \boldsymbol{\theta}^*\|_2, \qquad \gamma := \sqrt{2d_{\max}/\kappa}.$$

Step 2. Property of gradient mapping and cRSC/cRSS give

$$\begin{aligned}\|\mathbf{P}_K \mathbf{u}_t - \mathbf{v}\|_2 &\le \sqrt{1 - \alpha/L} \cdot \|\boldsymbol{\theta}_{t-1} - \mathbf{v}\|_2 \\ &\le \sqrt{1 - \alpha/L} \cdot (\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|_2 + \|\mathbf{v} - \boldsymbol{\theta}^*\|_2).\end{aligned}$$

Step 3. cRSC and cPGB give $\|\mathbf{v} - \boldsymbol{\theta}^*\|_2 \le C\Phi(S)$.

Combining above gives $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2 \le \Gamma \cdot \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|_2 + C'\Phi(S)$.

Simulations

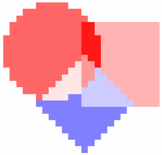


Figure 4: The true image $\boldsymbol{\theta}^*$ with values between -0.5 (blue) and 0.9 (red) on a 30×30 lattice graph G.

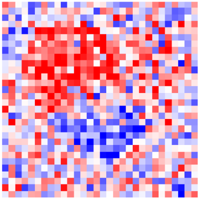


Figure 5: Noisy image $\frac{1}{n}\mathbf{X}^\top \mathbf{y}$, for $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \mathbf{e}$ with Gaussian design and noise standard deviation $\sigma = 1.5$.

Simulations

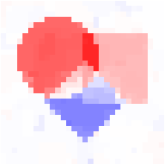


Figure 6: Best total-variation penalized estimate $\widehat{\boldsymbol{\theta}}$.

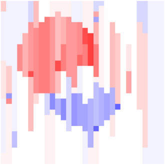


Figure 7: Best tree-PGD estimate $\widehat{\boldsymbol{\theta}}$ for a fixed line graph T_t in every iteration (zig-zagging vertically through G).

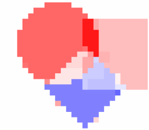


Figure 8: Best tree-PGD estimate $\widehat{\boldsymbol{\theta}}$ for a different random tree with $d_{\max} = 2$ in each iteration.

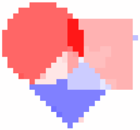


Figure 9: Best tree-PGD estimate $\widehat{\boldsymbol{\theta}}$ for a different random tree with $d_{\max} = 4$ in each iteration.

Conclusions

- Tree-PGD achieves strong statistical guarantees in regression models, without requiring a matching between the underlying graph and design matrix;
- Tree-PGD is a polynomial-time algorithm which approximately solves a non-convex objective;
- Tree-PGD allows for a different random tree in each iteration, which better targets the average sparsity.

Conclusions

- Tree-PGD achieves strong statistical guarantees in regression models, without requiring a matching between the underlying graph and design matrix;
- Tree-PGD is a polynomial-time algorithm which approximately solves a non-convex objective;
- Tree-PGD allows for a different random tree in each iteration, which better targets the average sparsity.

Thank you!