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

Sheng Xu SHENG.XU@YALE.EDU
Zhou Fan ZHOU.FAN@YALE.EDU
Sahand Negahban SAHAND.NEGAHBAN@YALE.EDU
Department of Statistics and Data Science, Yale University

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We study estimation of a gradient-sparse parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$, having strong gradient-sparsity $s^* := \|\nabla_G \boldsymbol{\theta}^*\|_0$ on an underlying graph G. Given observations $Z_1, \ldots, Z_n$ and a smooth, convex loss function $\mathcal{L}$ for which $\boldsymbol{\theta}^*$ minimizes the population risk $\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}; Z_1, \ldots, Z_n)]$, we propose to estimate $\boldsymbol{\theta}^*$ by a projected gradient descent algorithm that iteratively and approximately projects gradient steps onto spaces of vectors having small gradient-sparsity over low-degree spanning trees of G. We show that, under suitable restricted strong convexity and smoothness assumptions for the loss, the resulting estimator achieves the squared-error risk $\frac{s^*}{n} \log(1 + \frac{p}{s^*})$ up to a multiplicative constant that is independent of G. In contrast, previous polynomial-time algorithms have only been shown to achieve this guarantee in more specialized settings, or under additional assumptions for G and/or the sparsity pattern of $\nabla_G \boldsymbol{\theta}^*$. As applications of our general framework, we apply our results to the examples of linear models and generalized linear models with random design.

Keywords: structured sparsity, changepoint models, piecewise-constant signals, compressed sensing, graph signal processing, approximation algorithms

1. Introduction

We study estimation of a piecewise-constant or gradient-sparse parameter vector on a given graph. This problem may arise in statistical changepoint detection (Killick et al., 2012; Fryzlewicz, 2014), where an unknown vector on a line graph has a sequential changepoint structure. In image denoising (Rudin et al., 1992) and compressed sensing (Candès et al., 2006a; Donoho, 2006), this vector may represent a gradient-sparse image on a 2D or 3D lattice graph, as arising in medical X-rays and CT scans. For applications of epidemic tracking and anomaly detection on general graphs and networks, this vector may indicate regions of infected or abnormal nodes (Arias-Castro et al., 2011).

We consider the following general framework: Given observations $Z_1^n := (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ with distribution $\mathcal{P}$, we seek to estimate a parameter $\boldsymbol{\theta}^* \in \mathbb{R}^p$ associated to $\mathcal{P}$. The coordinates of $\boldsymbol{\theta}^*$ are identified with the vertices of a known graph $G = (V, E)$, where the number of vertices is $|V| = p$. Denoting by $\nabla_G : \mathbb{R}^p \to \mathbb{R}^{|E|}$ the discrete gradient operator

$$\nabla_G \boldsymbol{\theta} = \big(\theta_i - \theta_j : (i, j) \in E\big), \tag{1}$$

we assume that the gradient sparsity $s^* := \|\nabla_G \boldsymbol{\theta}^*\|_0$ is small relative to the total number of edges in G. For example, when G is a line or lattice graph, s^* measures the number of changepoints or the total boundary size between the constant pieces of an image, respectively. For a given convex and differentiable loss function $\mathcal{L} : \mathbb{R}^p \times \mathcal{Z}^n \to \mathbb{R}$, we assume that $\boldsymbol{\theta}^*$ is related to the data distribution $\mathcal{P}$ as the minimizer of the population risk,

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

Important examples include linear and generalized linear models for $Z_i = (\mathbf{x}_i, y_i)$, where $\boldsymbol{\theta}^*$ is the vector of regression coefficients and $\mathcal{L}$ is the usual squared-error or negative log-likelihood loss.

Our main result implies that, under suitable restricted strong convexity and smoothness properties of the loss (Negahban et al., 2012) and subgaussian assumptions on the noise, a polynomial-time projected gradient descent algorithm yields an estimate $\widehat{\boldsymbol{\theta}}$ which achieves the squared-error guarantee

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 \leq C \cdot \frac{s^*}{n} \log\left(1 + \frac{p}{s^*}\right) \tag{2}$$

with high probability. Here, $C > 0$ is a constant independent of the graph G, and depends only on the loss $\mathcal{L}$ and distribution $\mathcal{P}$ via their convexity, smoothness, and subgaussian constants.

Despite the simplicity of the guarantee (2) and its similarity to results for estimating *coordinate-sparse* parameters $\boldsymbol{\theta}^* \in \mathbb{R}^p$, to our knowledge, our work is the first to establish this guarantee in polynomial time for estimating *gradient-sparse* parameters on general graphs, including the 1D line. In particular, (2) is not necessarily achieved by convex approaches which constrain or regularize the ℓ_1 (total-variation) relaxation $\|\nabla_G \boldsymbol{\theta}^*\|_1$, for the reason that an ill-conditioned discrete gradient matrix $\nabla_G \in \mathbb{R}^{|E| \times p}$ contributes to the restricted convexity and smoothness properties of the resulting convex problem (Hütter and Rigollet, 2016; Fan and Guan, 2018). We discuss this further below, in the context of related literature.

Our work instead analyzes an algorithm that iteratively and approximately computes the projected gradient update over a sequence of low-degree spanning trees $T_1, T_2, \ldots$ of G.[1]

$$\boldsymbol{\theta}_t \approx \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^p : \|\nabla_{T_t} \boldsymbol{\theta}\|_0 \leq S} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1} + \eta \cdot \nabla \mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)\|_2 \tag{3}$$

For graphs G that do not admit spanning trees of low degree, we apply an idea of Padilla et al. (2017) and construct T_t using a combination of edges in G and additional edges representing backtracking paths along a depth-first-search traversal of G.

Our algorithm and analysis rely on an important insight from Jain et al. (2014), which is to perform each projection using a target sparsity-level S that is larger than the true gradient-sparsity s^* by a constant factor. This idea was applied in Jain et al. (2014) to provide a statistical analysis of iterative thresholding procedures such as IHT, CoSaMP, and HTP for estimating coordinate-sparse parameters (Blumensath and Davies, 2009; Needell and Tropp, 2009; Foucart, 2011). A key ingredient in our proof, Lemma 8 below, is a combinatorial argument which compares the errors of approximating any vector $\mathbf{u}$ by vectors $\mathbf{u}^S$ and $\mathbf{u}^*$ that are gradient-sparse over a tree, with two different sparsity levels S and s^*. This extends a central lemma of Jain et al. (2014) from the simpler setting of coordinate-sparsity to a setting of gradient-sparsity on trees.

1.1. Related literature

Existing literature on this and related problems is extensive, and we provide here a necessarily partial overview.

[1] Here, $\nabla \mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$ is the gradient of $\mathcal{L}(\boldsymbol{\theta}; Z_1^n)$ with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta}_{t-1}$, and $\nabla_{T_t} \boldsymbol{\theta}$ is the discrete gradient operator (1) over the edges in T_t instead of G.

Convex approaches: Estimating a piecewise-constant vector $\boldsymbol{\theta}^*$ in both the direct-measurements model $y_i = \theta_i^* + e_i$ and the indirect linear model $y_i = \mathbf{x}_i^\top \boldsymbol{\theta}^* + e_i$ has been of interest since early work on the fused lasso (Tibshirani et al., 2005; Rinaldo, 2009) and compressed sensing (Candès et al., 2006b,a; Donoho, 2006). A natural and commonly-used approach is to constrain or penalize the total-variation semi-norm $\|\nabla_G \boldsymbol{\theta}^*\|_1$ (Rudin et al., 1992). Statistical properties of this approach have been extensively studied, including estimation guarantees over signal classes of either bounded variation or bounded exact gradient-sparsity (Mammen and van de Geer, 1997; Hütter and Rigollet, 2016; Sadhanala et al., 2016; Dalalyan et al., 2017; Lin et al., 2017; Ortelli and van de Geer, 2018); exact or robust recovery guarantees in compressed sensing contexts (Needell and Ward, 2013a,b; Cai and Xu, 2015); and correct identification of changepoints or of the discrete gradient support (Harchaoui and Lévy-Leduc, 2010; Sharpnack et al., 2012). Extensions to higher-order trend-filtering methods have been proposed and studied in Kim et al. (2009); Wang et al. (2016); Sadhanala et al. (2017); Guntuboyina et al. (2017). These works have collectively considered settings of both direct and indirect linear measurements, for the 1D line, 2D and 3D lattices, and more general graphs.

In the above work, statistical guarantees analogous to (2) have only been obtained under restrictions for either G or $\boldsymbol{\theta}^*$, which we are able to remove using a non-convex approach. Hütter and Rigollet (2016) established a guarantee analogous to (2) when certain compatibility and inverse-scaling factors of G are $O(1)$; a sufficient condition is that G has constant maximum degree, and the Moore-Penrose pseudo-inverse $\nabla_G^\dagger$ has constant $\ell_1 \to \ell_2$ operator norm. This notably does not include the 1D line or 2D lattice. Dalalyan et al. (2017), Lin et al. (2017), and Guntuboyina et al. (2017) developed complementary results, showing that (2) can hold for the 1D line provided that the s^* changepoints of $\boldsymbol{\theta}^*$ have minimum spacing $\gtrsim p/(s^*+1)$. An extension of this to tree graphs was proven in Ortelli and van de Geer (2018). Roughly speaking, $\nabla_G^\dagger$ is an effective design matrix for an associated sparse regression problem, and the spacing condition ensures that the *active* variables in the regression model are weakly correlated, even if the full design $\nabla_G^\dagger$ has strong correlations.

Synthesis approach: A separate line of work focuses on the *synthesis* approach, which uses a sparse representation of $\boldsymbol{\theta}^*$ in an orthonormal basis or more general dictionary. Such methods include wavelet approaches in 1D (Daubechies, 1988; Donoho and Johnstone, 1994, 1995), curvelet and ridgelet frames in 2D (Candès, 1998; Candès and Donoho, 2000, 2004), and tree-based wavelets for more general graphs (Gavish et al., 2010; Sharpnack et al., 2013). Elad et al. (2007) and Nam et al. (2013) compare and discuss differences between the synthesis and analysis approaches. Note that in general, an s^*-gradient-sparse signal $\boldsymbol{\theta}^*$ may not admit a $O(s^*)$-sparse representation in an orthonormal basis. For example, $\boldsymbol{\theta}^*$ having s^* changepoints on the line may have up to $s^* \log_2 p$ non-zero coefficients in the Haar wavelet basis, and (2) would be inflated by an additional log factor using Haar wavelets.

Our contributions: In contrast to this first line of work on convex methods, our current work is most closely related to a third line of literature on methods that penalize or constrain the exact non-convex gradient-sparsity $\|\nabla_G \boldsymbol{\theta}^*\|_0$, rather than its convex ℓ_1 relaxation (Mumford and Shah, 1989; Boykov et al., 2001; Boysen et al., 2009; Fan and Guan, 2018). This direct method enables theoretical guarantees that remove the spectral conditions on the graph G as well as the minimum spacing requirements of the work alluded to above.

Our results extend those of Fan and Guan (2018), which established similar guarantees to (2) for direct measurements $y_i = \theta_i^* + e_i$. Our projected gradient algorithm is similar to the proximal-gradient method recently studied in Xu and Fan (2019), which considered indirect linear measurements $y_i = \mathbf{x}_i^\top \boldsymbol{\theta}^* + e_i$ in a compressed sensing context. In contrast to Xu and Fan (2019), which considered deterministic measurement errors and a restrictive RIP-type condition on the measurement design, we provide guarantees in the statistical setting of random noise, with much weaker conditions for the regression design, and for a general convex loss. These statistical guarantees are based on a novel tree-projection algorithm that approximates the graph at every iteration. The analysis leverages a new bound that controls the approximation error of tree projections, which is presented in Lemma 8.

2. Tree-projected gradient descent algorithm

Our proposed algorithm, tree-projected gradient descent (tree-PGD), consists of two main steps:

1. For a specified vertex degree $d_{\max} \geq 2$ and iteration count $\tau \geq 1$, we construct a sequence of trees $T_1, \ldots, T_\tau$ on the same vertices as G, such that each tree T_t has maximum degree $\leq d_{\max}$, and any gradient-sparse vector on G remains gradient-sparse on T_t.

2. For a specified step size $\eta > 0$ and sparsity level $S > 0$, we compute iterates $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_\tau$ where each $\boldsymbol{\theta}_t$ solves the projected gradient-descent step (3) over a discretized domain—see (5) and (6) below.

For simplicity, we initialize the algorithm at $\boldsymbol{\theta}_0 = 0$. The main tuning parameter is the projection sparsity S, which controls the bias-variance trade-off and the gradient sparsity of the final estimate $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_\tau$. The additional parameters of the algorithm are $d_{\max}$, τ, η, and the discretization (5) specified by $(\Delta_{\min}, \Delta_{\max}, \delta)$. We discuss these two steps in detail below.

For our theoretical guarantees, it is sufficient to choose $d_{\max} = 2$ and to fix the same tree in every iteration. However, we observe in Section 5 that using both larger values of $d_{\max}$ and a different random tree in each iteration can yield substantially lower recovery error in practice, so we will state our algorithm and theory to allow for these possibilities.

2.1. Tree construction

We construct a tree T on the vertices $V = \{1, \ldots, p\}$ by the following procedure.

1. Compute any spanning tree $\tilde{T}$ of G. If $\tilde{T}$ has maximum degree $\leq d_{\max}$, then set $T = \tilde{T}$.

2. Otherwise, let $\mathcal{O}_{DFS}$ be the ordering of unique vertices and edges visited in any depth-first-search (DFS) traversal of $\tilde{T}$. For each vertex v whose degree exceeds $d_{\max}$ in $\tilde{T}$, keep its first $d_{\max}$ edges in this ordering, and delete its remaining edges from $\tilde{T}$. Note that the deleted edges are between v and its children.

3. For each such deleted edge (v, w) where w is a child of v, let w' be the vertex preceding w in the ordering $\mathcal{O}_{DFS}$, and add to $\tilde{T}$ the edge (w', w). Let T be the final tree.

This procedure is illustrated in Figure 1. We repeat this construction to obtain each tree $T_1, \ldots, T_\tau$.

If G itself has maximum degree $\leq d_{\max}$, then Steps 2 and 3 above are not necessary, and the guarantee (4) below may be trivially strengthened to $\|\nabla_T \boldsymbol{\theta}\|_0 \leq \|\nabla_G \boldsymbol{\theta}\|_0$. For graphs G of larger maximum degree, the idea in Steps 2 and 3 above and the associated guarantee (4) are drawn from Lemma 1 of Padilla et al. (2017), which considered the case of a line graph for T (where $d_{\max} = 2$).

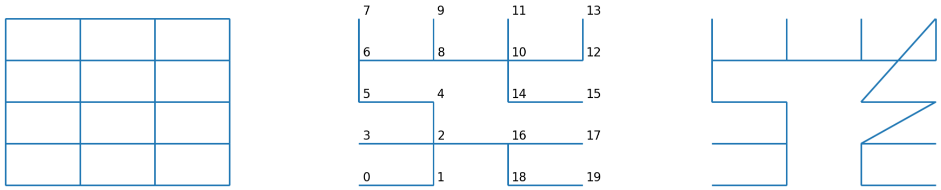

Figure 1: An illustration of the tree construction method. Left: Original lattice graph G. Middle: A spanning tree $\tilde{T}$ of G, with vertices numbered in DFS ordering. Right: The final tree T with $d_{\max} = 3$, which changes edge $(2, 16)$ to $(15, 16)$, and edge $(10, 14)$ to $(13, 14)$, thus replacing the two edges adjacent to the degree-4 vertices of T.

Lemma 1 *Let $G = (V, E)$ be any connected graph with p vertices, and let T be as constructed above. Then T is a tree on V with maximum degree $\leq d_{\max}$. Furthermore, for any $\boldsymbol{\theta} \in \mathbb{R}^p$,*

$$\|\nabla_T \boldsymbol{\theta}\|_0 \leq 2\|\nabla_G \boldsymbol{\theta}\|_0. \tag{4}$$

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

2.2. Projected gradient approximation

The exact minimizer of (3) is the projection of $\mathbf{u}_t := \boldsymbol{\theta}_{t-1} - \eta \cdot \nabla \mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$ onto the space of S-gradient-sparse vectors over T_t. This space is a union of $\binom{p-1}{S}$ linear subspaces, and naively iterating over these subspaces is intractable for large S. We instead propose to approximate the projection by taking a discrete grid of values

$$\Delta := \left\{\Delta_{\min}, \Delta_{\min} + \delta, \Delta_{\min} + 2\delta, \ldots, \Delta_{\max} - \delta, \Delta_{\max}\right\} \tag{5}$$

and performing the minimization over $\boldsymbol{\theta} \in \Delta^p$. Thus, our tree-PGD algorithm sets

$$\boldsymbol{\theta}_t = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Delta^p : \|\nabla_{T_t} \boldsymbol{\theta}\|_0 \leq S} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1} + \eta \cdot \nabla \mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)\|_2 \tag{6}$$

Each $\boldsymbol{\theta}_t$ may be computed by a dynamic-programming recursion over T_t.[2]

In detail, fix any target vector $\mathbf{u} \in \mathbb{R}^p$ and a tree T on the vertices $\{1, \ldots, p\}$. To compute

$$\operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Delta^p : \|\nabla_T \boldsymbol{\theta}\|_0 \leq S} \|\boldsymbol{\theta} - \mathbf{u}\|_2, \tag{7}$$

pick any vertex $o \in \{1, \ldots, p\}$ with degree 1 in T as the root. For each vertex v of T, let T_v be the sub-tree consisting of v and its descendants. Let $|T_v|$ be the number of vertices in T_v and $\mathbf{u}_{T_v} \in \mathbb{R}^{|T_v|}$ be the coordinates of $\mathbf{u}$ belonging to T_v. Define $f_v : \Delta \times \{0, 1, \ldots, S\} \to \mathbb{R}$ by

$$f_v(c, s) = \min\left\{\|\boldsymbol{\theta} - \mathbf{u}_{T_v}\|_2^2 : \boldsymbol{\theta} \in \Delta^{|T_v|},\ \|\nabla_{T_v} \boldsymbol{\theta}\|_0 \leq s,\ \theta_v = c\right\}. \tag{8}$$

These values $f_v(c, s)$ may be computed recursively from the leaves to the root, as follows.

[2] For the case where T_t is a line graph, an alternative non-discretized algorithm with complexity $O(p^2 S)$ is presented in Auger and Lawrence (1989).

1. For each leaf vertex v of T and each $(c,s) \in \Delta \times \{0,1,\ldots,S\}$, set $f_v(c,s) = (c-u_v)^2$.
2. For each vertex v of T with children $(w_1,\ldots,w_k)$, given $f_w(c,s)$ for all $w \in \{w_1,\ldots,w_k\}$ and $(c,s) \in \Delta \times \{0,1,\ldots,S\}$:
   (a) For each $s \in \{0,1,\ldots,S\}$ and $w \in \{w_1,\ldots,w_k\}$, compute $m_w(s) = \min_{c\in\Delta} f_w(c,s)$.
   (b) For each $(c,s) \in \Delta \times \{0,1,\ldots,S\}$ and $w \in \{w_1,\ldots,w_k\}$, compute $g_w(c,s) = \min\{f_w(c,s), m_w(s-1)\}$, where this is taken to be $f_w(c,s)$ if $s=0$.
   (c) For each $(c,s) \in \Delta \times \{0,1,\ldots,S\}$, set

$$f_v(c,s) = (c-u_v)^2 + \min_{\substack{s_1,\ldots,s_k \geq 0 \\ s_1+\ldots+s_k=s}} \Big(g_{w_1}(c,s_1) + \ldots + g_{w_k}(c,s_k) \Big). \tag{9}$$

The following then produces the vector $\boldsymbol{\theta}$ which solves (7).

3. For the root vertex o, set $\theta_o = \arg\min_{c\in\Delta} f_o(c,S)$ and $S_o = S$.
4. For each other vertex v, given θ_v and S_v: Let $w_1,\ldots,w_k$ be the children of v and let $s_1,\ldots,s_k$ be the choices which minimized (9) for $f_v(\theta_v,S_v)$. For each $i=1,\ldots,k$, if $g_{w_i}(\theta_v,s_i) = f_{w_i}(\theta_v,s_i)$, then set $\theta_{w_i} = \theta_v$ and $S_{w_i} = s_i$. If $g_{w_i}(\theta_v,s_i) = m_{w_i}(s_i-1)$, then set $\theta_{w_i} = \arg\min_{c\in\Delta} f_{w_i}(c,s_i-1)$ and $S_{w_i} = s_i - 1$.

The update $\boldsymbol{\theta}_t$ in (6) is computed by applying this algorithm to $\mathbf{u} \equiv \mathbf{u}_t = \boldsymbol{\theta}_{t-1} - \eta \cdot \nabla\mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$.

Lemma 2 *This algorithm minimizes (7). Letting $d_{\max}$ be the maximum vertex degree of T and $|\Delta|$ be the cardinality of Δ, its computational complexity is $O(d_{\max}p|\Delta|(S+d_{\max})^{d_{\max}-1})$.*

Let us compute the total complexity of this tree-PGD algorithm, under parameter settings that yield a rate-optimal statistical guarantee for the linear model discussed in Section 4.1. We set $d_{\max}$ as a small integer and S as a constant multiple of s^*. Evaluating $\nabla\mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$ in the linear model requires two matrix-vector multiplications of complexity $O(np)$, where n is the sample size. Let us assume that the number of graph edges is $|E| = O(p)$, and that the entries of $\boldsymbol{\theta}^*$ and the noise $\mathbf{e}$ are both of constant order. Then Corollary 10 indicates that we may take $\Delta_{\max} - \Delta_{\min} = O(\sqrt{p})$, $\delta = O(\sqrt{s^*/np})$, and $\tau = O(\log np)$. Under these settings, the total complexity of tree-PGD is $O\Big((np + p^2\sqrt{n}(s^*)^{d_{\max}-3/2})\log np\Big)$. Setting $d_{\max} = 2$ (i.e. taking $T_1,\ldots,T_\tau$ to be line graphs) yields the lowest complexity.

3. Main theorem

We introduce the following notation which identifies gradient-sparse vectors, partitions of the vertices $\{1,\ldots,p\}$, and subspaces of $\mathbb{R}^p$.

Definition 3 *Let T be a connected graph on the vertices $V = \{1,\ldots,p\}$, and let $\boldsymbol{\theta} \in \mathbb{R}^p$. The* **partition induced by $\boldsymbol{\theta}$ over** *T is the partition of V whose sets are the connected components of $\{(i,j) \in T : \theta_i = \theta_j\}$ in T. For such a partition $\mathcal{P}$ having k sets, the* **subspace associated to** *$\mathcal{P}$ is the dimension-k subspace of vectors in $\mathbb{R}^p$ taking a constant value over each set. The* **boundary of** *$\mathcal{P}$* **over** *T, denoted by $\partial_T\mathcal{P}$, is the set of edges $(i,j) \in T$ where i,j belong to different sets of $\mathcal{P}$.*

Thus, the sets of the partition $\mathcal{P}$ induced by $\boldsymbol{\theta}$ over T are the "pieces" of the graph T where $\boldsymbol{\theta}$ takes a constant value. If $\mathcal{P}$ is induced by $\boldsymbol{\theta}$ over T, and K is the associated subspace, then $\boldsymbol{\theta} \in K$. Furthermore, $\partial_T \mathcal{P}$ is exactly the edge set where $\nabla_T \boldsymbol{\theta}$ is non-zero, and $\|\nabla_T \boldsymbol{\theta}\|_0 = |\partial_T \mathcal{P}|$.

We introduce two properties for the loss, defined for pairs of connected graphs (T_1, T_2) on the same vertices V. We will apply these to consecutive pairs of trees generated by tree-PGD.

Definition 4 (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 convexity and smoothness constants $\alpha, L > 0$, if the following holds: For any partitions $\mathcal{P}_1, \mathcal{P}_2$ of $\{1, \ldots, p\}$ where $|\partial_{T_1} \mathcal{P}_1| \leq S$ and $|\partial_{T_2} \mathcal{P}_2| \leq S$, and any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in K := K_1 + K_2$ where K_1, K_2 are the subspaces associated to $\mathcal{P}_1, \mathcal{P}_2$,*

$$f(\boldsymbol{\theta}_2) \geq f(\boldsymbol{\theta}_1) + \langle \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1, \nabla f(\boldsymbol{\theta}_1) \rangle + \frac{\alpha}{2} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2^2, \tag{10}$$

$$f(\boldsymbol{\theta}_2) \leq f(\boldsymbol{\theta}_1) + \langle \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1, \nabla f(\boldsymbol{\theta}_1) \rangle + \frac{L}{2} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2^2. \tag{11}$$

Definition 5 (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 the following holds: For any partitions $\mathcal{P}_1, \mathcal{P}_2$ of $\{1, \ldots, p\}$ where $|\partial_{T_1} \mathcal{P}_1| \leq S$ and $|\partial_{T_2} \mathcal{P}_2| \leq S$, letting K_1, K_2 be their associated subspaces and $\mathbf{P}_K$ be the orthogonal projection onto $K := K_1 + K_2$,*

$$\|\mathbf{P}_K \nabla f(\boldsymbol{\theta}^*)\|_2 \leq \Phi(S). \tag{12}$$

To provide some interpretation, the below lemma gives an example for this function Φ in the important setting where $\mathbf{w}^\top \nabla \mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)$ is subgaussian for any $\mathbf{w} \in K$.

Lemma 6 *Let $S \geq 1$, let T_1, T_2 be trees on $\{1, \ldots, p\}$, and let $\boldsymbol{\theta}^* \in \mathbb{R}^p$. Suppose, for any subspace K as defined in Definition 5 and any $\mathbf{w} \in K$, that $\mathbf{w}^\top \nabla \mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)$ is σ^2/n-subgaussian.[3] Then for any $k > 0$ and a constant $C_k > 0$ depending only on k, with probability at least $1 - p^{-k}$, the loss $\mathcal{L}(\cdot\,; Z_1^n)$ has the cPGB*

$$\Phi(S) = C_k \sigma \sqrt{\tfrac{S}{n} \log\left(1 + \tfrac{p}{S}\right)}$$

with respect to (T_1, T_2), at $\boldsymbol{\theta}^*$ and sparsity level S.

The following is our main result, which provides a deterministic estimation guarantee when tree-PGB is applied with an appropriate choice of the projection sparsity $S = \kappa s^*$. This result yields the same type of guarantee for any choice of $d_{\max} \geq 2$ and any sequence of trees.

Theorem 7 *Suppose $\|\nabla_G \boldsymbol{\theta}^*\|_0 \leq s^*$, where $s^* > 0$. Set $S = \kappa s^*$ in tree-PGD for a constant $\kappa > 1$. Let $\tau \geq 1$ and $d_{\max} \geq 2$, let $T_1, \ldots, T_\tau$ be the sequence of trees generated by tree-PGD, and denote $T_0 = T_1$ and $S' = S + 2s^* + \max(\sqrt{S}, d_{\max})$. Suppose, for all $1 \leq t \leq \tau$, that*

1. *$\mathcal{L}(\cdot\,; Z_1^n)$ satisfies cRSC and cRSS with respect to (T_{t-1}, T_t), at sparsity level S' and with convexity and smoothness constants $\alpha, L > 0$.*
2. *$\mathcal{L}(\cdot\,; Z_1^n)$ has the cPGB $\Phi(S')$ with respect to (T_{t-1}, T_t), at the point $\boldsymbol{\theta}^*$ and sparsity level S'.*

[3] This means that for any $t > 0$, $\mathbb{P}[|\mathbf{w}^\top \nabla \mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)| > t] \leq 2e^{-nt^2/(2\sigma^2)}$.

Define

$$\gamma=\sqrt{\tfrac{(d_{\max}-1)(2s^*+\sqrt{S}+1)+1}{S-2s^*-\sqrt{S}}},\quad \Gamma=(1+\gamma)\sqrt{1-\tfrac{\alpha}{L}},\quad \Lambda=\tfrac{1}{1-\Gamma}\left(\tfrac{4(1+\gamma)}{\alpha}\cdot\Phi(S')+\delta\sqrt{p}\right),$$

and suppose κ is large enough such that $S>\sqrt{S}+2s^*$ and $\Gamma<1$. Take $\eta=\frac{1}{L}$, $\boldsymbol{\theta}_0=0$, and $-\Delta_{\min},\Delta_{\max}\geq\frac{1}{L}\|\nabla\mathcal{L}(\boldsymbol{\theta}^*;Z_1^n)\|_\infty+3\|\boldsymbol{\theta}^*\|_2+2\Lambda$ in tree-PGD. Then the τ^{th} iterate $\boldsymbol{\theta}_\tau$ of tree-PGD satisfies

$$\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}^*\|_2\leq\Gamma^\tau\cdot\|\boldsymbol{\theta}^*\|_2+\Lambda.$$

Note that since $\gamma\to0$ as $\kappa\to\infty$, for any value $\alpha/L\in(0,1]$, there is a choice of constant $\kappa\equiv\kappa(\alpha,L)$ sufficiently large to ensure $\Gamma<1$.

3.1. Proof overview

The proof of Theorem 7 adopts an induction argument. For simplicity, let us suppose here that $\boldsymbol{\theta}_t$ exactly minimizes (3). Then for each iteration, we wish to prove

$$\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\|_2\leq\Gamma\cdot\|\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\|_2+\tfrac{4(1+\gamma)}{\alpha}\cdot\Phi(S'). \tag{13}$$

The proof of (13) contains two main steps. First, we construct a subspace K which contains $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}^*$ and write $\|\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\|_2\leq\|\mathbf{P}_K\mathbf{u}_t-\boldsymbol{\theta}_t\|_2+\|\mathbf{P}_K\mathbf{u}_t-\boldsymbol{\theta}^*\|_2$. Using the following key lemma, we show that there exists such a subspace K for which $\|\mathbf{P}_K\mathbf{u}_t-\boldsymbol{\theta}_t\|_2\leq\gamma\|\mathbf{P}_K\mathbf{u}_t-\boldsymbol{\theta}^*\|_2$, and the vectors in K have gradient-sparsity not much larger than $S+s^*$.

Lemma 8 *Let T be a tree on the vertices $\{1,\ldots,p\}$ with maximum vertex degree $d_{\max}$. Let $s^*>0$ and $S=\kappa s^*$, where $\kappa>1$ and $S>\sqrt{S}+s^*$. Let $\mathbf{u}\in\mathbb{R}^p$ be arbitrary, let $\mathbf{u}^*\in\mathbb{R}^p$ be any vector satisfying $\|\nabla_T\mathbf{u}^*\|_0\leq s^*$, and set*

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

Denote by (K^S,K^*) the subspaces associated to the partitions induced by $(\mathbf{u}^S,\mathbf{u}^*)$ over T. Then there exists a partition $\mathcal{P}$ of $\{1,\ldots,p\}$ with associated subspace K, such that K contains K^S+K^*,

$$|\partial_T\mathcal{P}|\leq S+s^*+\sqrt{S}, \tag{14}$$

and the orthogonal projection $\mathbf{P}_K\mathbf{u}$ of $\mathbf{u}$ onto K satisfies

$$\|\mathbf{P}_K\mathbf{u}-\mathbf{u}^S\|_2^2\leq\frac{(d_{\max}-1)(s^*+\sqrt{S}+1)+1}{S-s^*-\sqrt{S}}\|\mathbf{P}_K\mathbf{u}-\mathbf{u}^*\|_2^2. \tag{15}$$

Then, in the second step, we bound $\|\mathbf{P}_K\mathbf{u}_t-\boldsymbol{\theta}^*\|_2$ by introducing $\mathbf{v}=\arg\min_{\boldsymbol{\theta}\in K}\mathcal{L}(\boldsymbol{\theta};Z_1^n)$. Using a property of the gradient mapping (Lemma 13) and the cRSC and cRSS conditions, we show that $\|\mathbf{P}_K\mathbf{u}_t-\mathbf{v}\|_2\leq\sqrt{1-\alpha/L}\cdot\|\boldsymbol{\theta}_{t-1}-\mathbf{v}\|_2$. Applying the triangle inequality, this implies $\|\mathbf{P}_K\mathbf{u}_t-\boldsymbol{\theta}^*\|_2\leq\sqrt{1-\alpha/L}\cdot\|\boldsymbol{\theta}_{t-1}-\boldsymbol{\theta}^*\|_2+2\|\mathbf{v}-\boldsymbol{\theta}^*\|_2$. Finally, we show that $\|\mathbf{v}-\boldsymbol{\theta}^*\|_2\leq(2/\alpha)\Phi(S')$ using the cRSC and cPGB properties of the loss, and combining gives (13).

The use of Lemma 8 is inspired by an analogous argument of Jain et al. (2014) for coordinate-sparse parameter estimation. However, the analysis for coordinate-sparsity is simpler, due to a key structural property that if $\mathbf{u}^S$ and $\mathbf{u}^*$ are the best (coordinate-) S-sparse and s^*-sparse approximations of $\mathbf{u}$, then the sparse subspace of $\mathbf{u}^*$ is contained inside that of $\mathbf{u}^S$. This nested subspace structure does not hold for gradient-sparsity, and thus our proofs of both Lemma 8 and Theorem 7 follow different arguments from those of Jain et al. (2014).

4. Examples

4.1. Gradient-Sparse Linear Regression

Consider the example of $Z_i = (\mathbf{x}_i, y_i)$ satisfying a linear model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\theta}^* + e_i \tag{16}$$

for independent design vectors $\mathbf{x}_i \in \mathbb{R}^p$ and mean-zero residual errors e_i. Let us write this as $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \mathbf{e}$ where $\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{e} = (e_1, \ldots, e_n)$, and $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the random design matrix with rows $\mathbf{x}_i^\top$. Then $\boldsymbol{\theta}^*$ is the minimizer of $\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}; Z_1^n)]$ for the squared-error loss $\mathcal{L}(\boldsymbol{\theta}; Z_1^n) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$. The gradient of the loss is given by $\nabla\mathcal{L}(\boldsymbol{\theta}; Z_1^n) = \mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})/n$.

We assume that

$$\mathrm{Cov}(\mathbf{x}_i) = \boldsymbol{\Sigma}, \qquad \lambda_{\max}(\boldsymbol{\Sigma}) = \lambda_1, \qquad \lambda_{\min}(\boldsymbol{\Sigma}) = \lambda_p, \qquad \|\mathbf{x}_i\|_{\psi_2}^2 \leq D\lambda_p \tag{17}$$

$$\mathbb{E}[e_i] = 0, \qquad \|e_i\|_{\psi_2}^2 \leq \sigma^2 \tag{18}$$

for constants $\lambda_1, \lambda_p, D, \sigma^2 > 0$, where $\|\cdot\|_{\psi_2}$ denotes the scalar or vector subgaussian norm. Then the cRSC, cRSS, and cPGB conditions hold according to the following proposition.

Proposition 9 *Suppose (17) and (18) hold, and let $S' \geq 1$. Define*

$$g(S') = S' \log(1 + \tfrac{p}{S'}). \tag{19}$$

Let $T_1, \ldots, T_\tau$ be the trees generated by tree-PGD, and let $T_0 = T_1$. For any $k > 0$, and some constants $C_1, C_2, C_3 > 0$ depending only on k and D, if

$$n \geq C_1 g(S')$$

then with probability at least $1 - \tau \cdot p^{-k}$, for every $1 \leq t \leq \tau$,

1. *$\mathcal{L}(\cdot\,; Z_1^n)$ satisfies cRSC and cRSS with respect to (T_{t-1}, T_t) at sparsity level S' and with convexity and smoothness constants $\alpha = \lambda_p/2$ and $L = 3\lambda_1/2$.*
2. *$\mathcal{L}(\cdot\,; Z_1^n)$ has the cPGB $\Phi(S') = C_2\sigma\sqrt{\lambda_1 g(S')/n}$ with respect to (T_{t-1}, T_t), at $\boldsymbol{\theta}^*$ and sparsity level S'.*
3. *$\|\nabla\mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)\|_\infty \leq C_3\sigma\sqrt{(\lambda_1 \log p)/n}$.*

Applying this and Theorem 7, we obtain the following immediate corollary.

Corollary 10 *Suppose (17) and (18) hold, and $\|\nabla_G \boldsymbol{\theta}^*\|_0 \leq s^*$ and $\|\boldsymbol{\theta}^*\|_2 \leq c_0\sqrt{p}$ for some $s^* \geq 1$ and $c_0 > 0$. Set $S = c_1(\lambda_1/\lambda_p)^2 s^*$, $\eta = 2/(3\lambda_1)$, $\omega = \sigma\lambda_1^{3/2}/\lambda_p^2$, $-\Delta_{\min} = \Delta_{\max} = c_2(\sqrt{p} + \omega\sqrt{(s^* \log p)/n})$, $\delta = \omega\sqrt{s^*/np}$, and $\tau = c_3 \log(np/\omega^2 s^*)$ in tree-PGD, for sufficiently large constants $c_1 > 0$ depending on $d_{\max}, D$ and $c_2, c_3 > 0$ depending on $d_{\max}, D, c_0$.*

Then for any $k > 0$ and some constants $C_1, C_2 > 0$ depending only on $k, d_{\max}, D$, if $n \geq C_1(\lambda_1/\lambda_p)^2 s^* \log(1 + p/s^*)$, then with probability at least $1 - \tau \cdot p^{-k}$,

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}^*\|_2^2 \leq C_2 \cdot \frac{\sigma^2\lambda_1^3}{\lambda_p^4} \cdot \frac{s^*}{n} \log\left(1 + \frac{p}{s^*}\right).$$

4.2. Gradient-Sparse GLM

Consider the example of $Z_i = (\mathbf{x}_i, y_i)$ satisfying a generalized linear model (GLM)

$$P(y_i|\mathbf{x}_i, \boldsymbol{\theta}^*, \phi) = \exp\left\{\frac{y_i\mathbf{x}_i^\top\boldsymbol{\theta}^* - b(\mathbf{x}_i^\top\boldsymbol{\theta}^*)}{\phi}\right\} \cdot h(y_i, \phi)$$

for independent design vectors $\mathbf{x}_i \in \mathbb{R}^p$. Here $\phi > 0$ is a constant scale parameter, and h and b are the base measure and cumulant function of the exponential family, where $\mathbb{E}(y_i|\mathbf{x}_i) = b'(\mathbf{x}_i^\top\boldsymbol{\theta}^*)$. Then $\boldsymbol{\theta}^*$ minimizes the population risk $\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}; Z_1^n)]$ for the negative log-likelihood loss $\mathcal{L}(\boldsymbol{\theta}; Z_1^n) = \frac{1}{n}\sum_{i=1}^n \left(b(\mathbf{x}_i^\top\boldsymbol{\theta}) - y_i\mathbf{x}_i^\top\boldsymbol{\theta}\right)$. The gradient of this loss is $\nabla\mathcal{L}(\boldsymbol{\theta}; Z_1^n) = \frac{1}{n}\sum_{i=1}^n (b'(\mathbf{x}_i^\top\boldsymbol{\theta}) - y_i)\mathbf{x}_i$.

Let us assume that (17) holds for the design vectors $\mathbf{x}_i$. Setting $e_i = y_i - b'(\mathbf{x}_i^\top\boldsymbol{\theta}^*)$, let us assume also that for some constants $\alpha_b, L_b, D_1, D_2 > 0$ and $\beta \in [1, 2]$,

$$\frac{\alpha_b}{2}(x_2 - x_1)^2 \le b(x_2) - b(x_1) - b'(x_1)(x_2 - x_1) \le \frac{L_b}{2}(x_2 - x_1)^2 \quad \text{for all } x_1, x_2 \in \mathbb{R}, \tag{20}$$

$$\mathbb{P}(|e_i| > \zeta) \le D_1 \exp(-D_2\zeta^\beta) \quad \text{for all } \zeta > 0. \tag{21}$$

Then the cRSC, cRSS, and cPGB conditions hold according to the following proposition.

Proposition 11 *Suppose that (17), (20), and (21) hold. Let $S' \ge 1$ and $g(S')$ be as in (19). Let $T_1, \ldots, T_\tau$ be the trees generated by tree-PGD, and let $T_0 = T_1$. For any $k > 0$ and some constants $C_1, C_2, C_3 > 0$ depending only on k, D, D_1, D_2, β, if $n \ge C_1 g(S')$, then with probability at least $1 - \tau \cdot p^{-k}$, for every $1 \le t \le \tau$,*

1. *$\mathcal{L}(\cdot\,; Z_1^n)$ satisfies cRSC and cRSS with respect to (T_{t-1}, T_t) at sparsity levels S' with convexity and smoothness constants $\alpha = \frac{\alpha_b\lambda_p}{2}$ and $L = \frac{3L_b\lambda_1}{2}$.*

2. *$\mathcal{L}(\cdot\,; Z_1^n)$ has the cPGB with respect to (T_{t-1}, T_t), at $\boldsymbol{\theta}^*$ and sparsity level S'.*

$$\Phi(S') = \begin{cases} C_2\sqrt{\lambda_1/n} \cdot g(S')^{1/\beta} & \textit{if} \quad 1 < \beta \le 2 \\ C_2 \log n\sqrt{\lambda_1/n} \cdot g(S') & \textit{if} \quad \beta = 1 \end{cases}$$

3. $\|\nabla\mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)\|_\infty \le \begin{cases} C_3(\log p)^{1/\beta}\sqrt{\lambda_1/n} & \textit{if} \quad 1 < \beta \le 2 \\ C_3(\log n)(\log p)\sqrt{\lambda_1/n} & \textit{if} \quad \beta = 1 \end{cases}$

Under suitable settings of the tree-PGD parameters, similar to Corollary 10 and which we omit for brevity, when $n \ge C's^* \log(1 + p/s^*)$, this yields the estimation rate

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}^*\|_2^2 \le C \cdot \frac{(s^* \log(1 + p/s^*))^{2/\beta}}{n}$$

in models where $1 < \beta \le 2$, and this rate with an additional $(\log n)^2$ factor in models where $\beta = 1$. (Here, these constants C, C' depend on $\lambda_1, \lambda_p, D, D_1, D_2, \beta$.)

We note that this result may be established under a relaxed condition (20) that only holds over a sufficiently large bounded region for x_1, x_2, following a more delicate analysis and ideas of Negahban et al. (2012). For simplicity, we will not pursue this direction in this work.

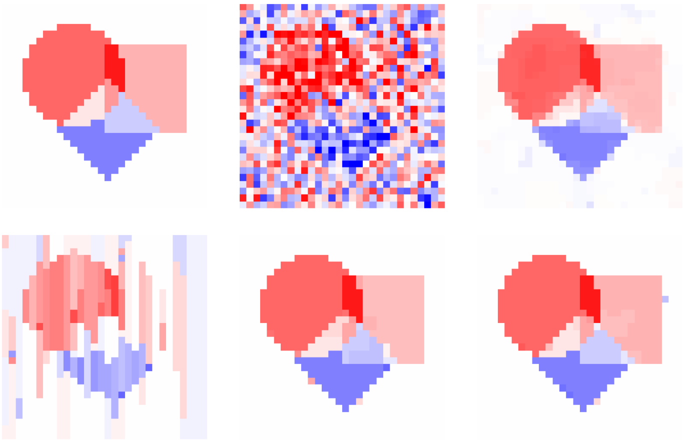

Figure 2: Top-left: True image $\boldsymbol{\theta}^*$, with values between -0.5 (blue) and 0.9 (red). Top-middle: Noisy image $\frac{1}{n}\mathbf{X}^\top\mathbf{y}$, for $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \mathbf{e}$ with Gaussian design and noise standard deviation $\sigma = 1.5$. Top-right: Best total-variation penalized estimate $\widehat{\boldsymbol{\theta}}$. Bottom row: Best tree-PGD estimate $\widehat{\boldsymbol{\theta}}$ for a fixed line graph T_t in every iteration (zig-zagging vertically through G, bottom left), a different random tree with $d_{\max} = 2$ in each iteration (bottom middle), and a different random tree with $d_{\max} = 4$ in each iteration (bottom right).

5. Simulations

Theorem 7 applies for any choices of trees $T_1, \ldots, T_\tau$ in tree-PGD, with any maximum degree $d_{\max} \geq 2$. We perform a small simulation study in the linear model (16) to compare the empirical estimation accuracy of tree-PGD using different tree constructions.

We recover the image $\boldsymbol{\theta}^*$ depicted in Figure 2 on a 30×30 lattice graph G, using $n = 500$ linear measurements with $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I})$ and $e_i \sim \mathcal{N}(0, \sigma^2)$. For $\sigma = 1.5$, a noisy image $\frac{1}{n}\mathbf{X}^\top\mathbf{y} = \boldsymbol{\theta}^* + (\frac{1}{n}\mathbf{X}^\top\mathbf{X} - \mathbf{I})\boldsymbol{\theta}^* + \frac{1}{n}\mathbf{X}^\top\mathbf{e}$ is also depicted.

Tree construction: We applied tree-PGD in two settings: First, we constructed T_t using a deterministic DFS over G, fixed across all iterations. This resulted in T_t being a line graph that zig-zags vertically through G. Second, we constructed T_t using a different spanning tree $\tilde{T}_t$ generated by random DFS in each iteration. The DFS procedure started at a uniform random node and, at each forward step, chose a uniform random unvisited neighbor. We tested restricting to $d_{\max} = 2$ or $d_{\max} = 3$ for T_t, or letting $T_t = \tilde{T}_t$ (corresponding to $d_{\max} = 4$). In all experiments, we used $\tau = 80$, $\eta = 1/5$, and $(\Delta_{\min}, \Delta_{\max}, \delta) = (-0.6, 1.0, 0.05)$.

Results for a single experiment at $\sigma = 1.5$ are depicted in Figure 2, and average MSE across 20 experiments for varying σ are reported in Table 1. These results correspond to the best choices $S = \kappa s^*$ across a range of tested values. Estimation accuracy is substantially better using different and random trees than using the same fixed line graph. We observe small improvements using $d_{\max} = 3$ or $d_{\max} = 4$ over random line graphs with $d_{\max} = 2$, especially in the higher signal-to-noise settings. For comparison, we display in Figure 2 and Table 1 also the total-variation (TV) regularized estimate $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\nabla_G\boldsymbol{\theta}\|_1$ and its average MSE, corresponding to the best choices of λ. We observe that tree-PGD, which targets the exact gradient-sparsity rather

Noise std. dev. σ	1.0	1.5	2.0	2.5	3.0
Fixed line	0.0372	0.0373	0.0383	0.0388	0.0407
Random, $d_{\max} = 2$	0.0005	0.0009	0.0020	0.0040	0.0058
Random, $d_{\max} = 3$	0.0003	0.0008	0.0014	0.0028	0.0052
Random, $d_{\max} = 4$	0.0003	0.0007	0.0013	0.0032	0.0055
Total variation	0.0006	0.0013	0.0023	0.0036	0.0052

Table 1: MSE $\frac{1}{p}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2$ for recovering the image of Figure 2 (under best tuning of S), averaged across 20 independent simulations. For tree-PGD, using a different random tree T_t per iteration yields a sizeable improvement over using a fixed line graph across all iterations, and small improvements are observed for increasing $d_{\max}$. Average MSE for the total-variation penalized estimate is provided for comparison (under best tuning of λ).

than a convex surrogate, is more accurate in high signal-to-noise settings, and becomes less accurate in comparison with TV as signal strength decreases. This agrees with previous observations made in similar contexts in Hastie et al. (2017); Mazumder et al. (2017); Fan and Guan (2018).

6. Discussion

We have shown linear convergence of gradient descent with projections onto the non-convex space of gradient-sparse vectors on a graph. Our results show that this method achieves strong statistical guarantees in regression models, without requiring a matching between the underlying graph and design matrix. We do this by introducing a careful comparison between gradient-sparse approximations at different sparsity levels, which generalizes previous results for coordinate-sparse vectors.

Our theory is presented in such a way that allows the approximation trees to vary at each iteration. However, this is not required and the tree can be fixed with $d_{\max} = 2$ at the start of the algorithm. Nevertheless, we observe experimentally that using a different random tree in each iteration substantially improves the practical performance. Our intuition for the improvement with random trees is that the gradient-sparsity of the signal on the original graph G may be better captured by the average sparsity with respect to a randomly chosen sub-tree of G, than by the sparsity with respect to any fixed sub-tree. By using a different random tree in each iteration, the algorithm is better targeting this average sparsity. This observation will be studied in future work.

Another interesting direction for future work is to explore the connections between this work and computationally tractable sparse linear regression problems with highly correlated designs. For instance, some work Bühlmann et al. (2013); Dalalyan et al. (2017) discuss various ways to overcome correlated designs. In our setting, the tree projection step enables a computationally efficient method, and it is of interest to understand more general settings where one may overcome the correlated structure of the problem using a computationally efficient procedure.

Acknowledgments

This research is supported in part by NSF Grant DMS-1916198 and DMS-1723128.

References

Ery Arias-Castro, Emmanuel J Candès, and Arnaud Durand. Detection of an anomalous cluster in a network. *The Annals of Statistics*, 39(1):278–304, 2011.

Ivan E Auger and Charles E Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of mathematical biology*, 51(1):39–54, 1989.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239, 2001.

Leif Boysen, Angela Kempe, Volkmar Liebscher, Axel Munk, and Olaf Wittich. Consistencies and rates of convergence of jump-penalized least squares estimators. *The Annals of Statistics*, 37(1): 157–183, 2009.

Peter Bühlmann, Philipp Rütimann, Sara van de Geer, and Cun-Hui Zhang. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143 (11):1835–1858, 2013.

Jian-Feng Cai and Weiyu Xu. Guarantees of total variation minimization for signal recovery. *Information and Inference: A Journal of the IMA*, 4(4):328–353, 2015.

Emmanuel J Candès. *Ridgelets: Theory and applications*. PhD thesis, Stanford University Stanford, 1998.

Emmanuel J Candès and David L Donoho. Curvelets: A surprisingly effective nonadaptive representation for objects with edges. Technical report, Stanford University Dept of Statistics, 2000.

Emmanuel J Candès and David L Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Communications on Pure and Applied Mathematics*, 57(2):219–266, 2004.

Emmanuel J Candès, Justin K Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489, 2006a.

Emmanuel J Candès, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006b.

Arnak S Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581, 2017.

Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 41(7):909–996, 1988.

David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

David L Donoho and Iain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

David L Donoho and Iain M Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

Michael Elad, Peyman Milanfar, and Ron Rubinstein. Analysis versus synthesis in signal priors. *Inverse Problems*, 23(3):947, 2007.

Zhou Fan and Leying Guan. Approximate ℓ_0-penalized estimation of piecewise-constant signals on graphs. *The Annals of Statistics*, 46(6B):3217–3245, 2018.

Simon Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, 2011.

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

Matan Gavish, Boaz Nadler, and Ronald R Coifman. Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. In *ICML*, pages 367–374, 2010.

Adityanand Guntuboyina, Donovan Lieu, Sabyasachi Chatterjee, and Bodhisattva Sen. Adaptive risk bounds in univariate total variation denoising and trend filtering. *arXiv preprint arXiv:1702.05113*, 2017.

Zaid Harchaoui and Céline Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, 2010.

Trevor Hastie, Robert Tibshirani, and Ryan J Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*, 2017.

Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603–1618, 2008.

Jan-Christian Hütter and Philippe Rigollet. Optimal rates for total variation denoising. In *Conference on Learning Theory*, pages 1115–1146, 2016.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional M-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM review*, 51(2):339–360, 2009.

Kevin Lin, James L Sharpnack, Alessandro Rinaldo, and Ryan J Tibshirani. A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems*, pages 6884–6893, 2017.

Enno Mammen and Sara van de Geer. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.

Rahul Mazumder, Peter Radchenko, and Antoine Dedieu. Subset selection with shrinkage: Sparse linear modeling when the SNR is low. *arXiv preprint arXiv:1708.03288*, 2017.

David Mumford and Jayant Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 42(5):577–685, 1989.

Sangnam Nam, Mike E Davies, Michael Elad, and Rémi Gribonval. The cosparse analysis model and algorithms. *Applied and Computational Harmonic Analysis*, 34(1):30–56, 2013.

Deanna Needell and Joel A Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and computational harmonic analysis*, 26(3):301–321, 2009.

Deanna Needell and Rachel Ward. Near-optimal compressed sensing guarantees for total variation minimization. *IEEE transactions on image processing*, 22(10):3941–3949, 2013a.

Deanna Needell and Rachel Ward. Stable image reconstruction using total variation minimization. *SIAM Journal on Imaging Sciences*, 6(2):1035–1058, 2013b.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Francesco Ortelli and Sara van de Geer. On the total variation regularized estimator over a class of tree graphs. *Electronic Journal of Statistics*, 12(2):4517–4570, 2018.

Oscar Hernan Madrid Padilla, James G Scott, James Sharpnack, and Ryan J Tibshirani. The DFS fused lasso: Linear-time denoising over general graphs. *Journal of Machine Learning Research*, 18(1):6410–6445, 2017.

Alessandro Rinaldo. Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5B): 2922–2952, 2009.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, and Ryan J Tibshirani. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Advances in Neural Information Processing Systems*, pages 3513–3521, 2016.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, James L Sharpnack, and Ryan J Tibshirani. Higher-order total variation classes on grids: Minimax theory and trend filtering methods. In *Advances in Neural Information Processing Systems*, pages 5800–5810, 2017.

James Sharpnack, Aarti Singh, and Alessandro Rinaldo. Sparsistency of the edge lasso over graphs. In *Artificial Intelligence and Statistics*, pages 1028–1036, 2012.

James Sharpnack, Aarti Singh, and Akshay Krishnamurthy. Detecting activations over graphs using spanning tree wavelet bases. In *Artificial Intelligence and Statistics*, pages 536–544, 2013.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Yu-Xiang Wang, James Sharpnack, Alexander J Smola, and Ryan J Tibshirani. Trend filtering on graphs. *The Journal of Machine Learning Research*, 17(1):3651–3691, 2016.

Sheng Xu and Zhou Fan. Iterative Alpha Expansion for estimating gradient-sparse signals from linear measurements. *arXiv preprint arXiv:1905.06097*, 2019.

Appendix A. Correctness and complexity of algorithm

We prove Lemmas 1 and 2 on basic guarantees for the two steps of the tree-PGD algorithm.

Proof [Lemma 1] For the first statement, since $d_{\max} \geq 2$, the vertex w corresponding to each deleted edge (v, w) must be a child of v which is not its first child in the ordering $\mathcal{O}_{DFS}$. Then its preceding vertex w' must be a leaf vertex of $\tilde{T}$. Each such w corresponds to a different such leaf w', so deleting these edges (v, w) and adding (w', w) preserves the connectedness and tree structure. By construction, each non-leaf vertex of $\tilde{T}$ has degree at most $d_{\max}$ in T. Each leaf vertex of $\tilde{T}$ has degree at most $2 \leq d_{\max}$ in T, so T has maximum degree $\leq d_{\max}$.

For the second statement, since the edges of $\tilde{T}$ are a subset of those of G,

$$\|\nabla_{\tilde{T}} \boldsymbol{\theta}\|_0 \leq \|\nabla_G \boldsymbol{\theta}\|_0.$$

Let the root vertex of T be 1. For each other vertex $i \geq 2$, denote its parent in T by $p(i)$. Then

$$\|\nabla_T \boldsymbol{\theta}\|_0 = \sum_{i=2}^{p} \mathbf{1}\{\theta_i \neq \theta_{p(i)}\}. \tag{22}$$

Now consider two cases: If the edge $(i, p(i))$ exists in $\tilde{T}$, then it is a forward edge in the DFS of $\tilde{T}$, and $\mathbf{1}\{\theta_i \neq \theta_{p(i)}\}$ contributes to $\|\nabla_{\tilde{T}} \boldsymbol{\theta}\|_0$. If $(i, p(i))$ is not an edge of $\tilde{T}$, then $p(i)$ is a leaf node in $\tilde{T}$, and there is path of backward edges $(p_1, p_2, \ldots, p_r)$ in the DFS of $\tilde{T}$ where $p_1 = p(i)$ and $p_r = i$. The triangle inequality then implies

$$\mathbf{1}\{\theta_i \neq \theta_{p(i)}\} \leq \sum_{j=1}^{r-1} \mathbf{1}\{\theta_{p_j} \neq \theta_{p_{j+1}}\},$$

where each term on the right contributes to $\|\nabla_{\tilde{T}} \boldsymbol{\theta}\|_0$. Applying this to each term on the right of (22), and invoking the fundamental property that DFS visits each edge of $\tilde{T}$ exactly twice, we get

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

■

Proof [Lemma 2] It is clear that Step 1 computes (8) at the leaf vertices v. For Step 2, assume inductively that $f_w(c, s)$ is the value (8) for all children w of v. The value $g_w(c, s)$ represents the minimum value of $\|\boldsymbol{\theta} - \mathbf{u}_{T_w}\|_2^2$, if $\theta_v = c$ and the gradient-sparsity of $\boldsymbol{\theta}$ on T_w *and* the additional edge (v, w) is at most s—we have either $\theta_w = c$ and $g_w(c, s) = f_w(c, s)$, or $\theta_w \neq c$, in which case $\theta_w = \arg\min_{c \in \Delta} f_w(c, s-1)$ and $g_w(c, s) = m_w(s-1)$. Then (9) computes (8) at v by partitioning the gradient-sparsity s across its k children, and summing the costs $g_{w_i}(c, s_i)$ and the additional cost $(c - u_v)^2$ for the best such partition. Thus Step 2 correctly computes (8) for each vertex v. In particular, the minimum value for (7) is given by $\min_{c \in \Delta} f_o(c, S)$. The minimizer $\boldsymbol{\theta}$ is obtained by examining the minimizing choices in Steps 1 and 2, which is carried out in Steps 3 and 4: Each θ_v is the value of $\boldsymbol{\theta}$ at v, and each S_v is (an upper-bound for) the value of $\|\nabla_{T_v} \boldsymbol{\theta}\|_0$ at the minimizer $\boldsymbol{\theta}$.

For each vertex v, Step 1 has complexity $(S+1)|\Delta|$, Steps 2(a) and 2(b) both have complexity $(S+1)k|\Delta|$, and Step 2(c) has complexity $(S+1)|\Delta|k\binom{S+k-1}{k-1}$, as there are $\binom{s+k-1}{k-1} \leq \binom{S+k-1}{k-1}$

partitions of s into $s_1, \ldots, s_k$. Note that $k \leq d_{\max} - 1$, where this holds also for the root vertex o because we chose it to have degree 1 in T. Then $\binom{S+k-1}{k-1} = O((S + d_{\max})^{d_{\max}-2})$. Storing the relevant minimizers in Steps 1 and 2, the complexity of Steps 3 and 4 is $O(1)$ per vertex. So the total complexity is $O(d_{\max} p|\Delta|(S + d_{\max})^{d_{\max}-1})$. ■

Appendix B. Proof of Lemma 8

Proof Let $\mathcal{P}^S$ be the partition of $\{1, \ldots, p\}$ induced by $\mathbf{u}^S$ over T. We have $|\partial_T \mathcal{P}^S| \leq S$. If $|\partial_T \mathcal{P}^S| < S$, then let us arbitrarily split some vertex sets in $\mathcal{P}^S$ along edges of T, until $|\partial_T \mathcal{P}^S| = S$. Thus, we may assume henceforth that $|\partial_T \mathcal{P}^S| = S$.

We construct another partition $\mathcal{P}'$ of $\{1, \ldots, p\}$ into the (disjoint) vertex sets $(V_1, \ldots, V_B, R)$, such that each set of $\mathcal{P}'$ is connected over T, and $\mathcal{P}'$ satisfies the following properties:

1. For each $b = 1, \ldots, B$, the number of edges (i, j) in T where both $i, j \in V_b$, but i and j do not belong to the same set of $\mathcal{P}^S$, is greater than or equal to $s^* + \sqrt{\kappa s^*}$.

2. B has the upper and lower bounds

$$\frac{S - s^* - \sqrt{S}}{(d_{\max} - 1)(s^* + \sqrt{S} + 1) + 1} \leq B \leq \sqrt{S} \tag{23}$$

We construct this partition $\mathcal{P}'$ in the following way: Initialize $\tilde{T} = T$ and pick any degree-1 vertex of T as its root. Assign to each edge (i, j) of $\tilde{T}$ a "score" of 1 if i and j belong to the same set of $\mathcal{P}^S$, and 0 otherwise. Repeat the following steps for all vertices i of T, in reverse-breadth-first-search order (starting from a vertex i farthest from the root):

- Let $\tilde{T}_i$ be the sub-tree of $\tilde{T}$ rooted at i and consisting of the descendants of i in $\tilde{T}$.
- If the total score of edges in $\tilde{T}_i$ is at least $s^* + \sqrt{\kappa s^*}$, then add the vertices of $\tilde{T}_i$ as a set V_b to the partition $\mathcal{P}'$, and remove $\tilde{T}_i$ (including the edge from i to its parent) from $\tilde{T}$.

This terminates when the remaining tree $\tilde{T}$ has total score less than $s^* + \sqrt{\kappa s^*}$. Take the last set R of $\mathcal{P}'$ to be the vertices of this remaining tree.

By construction, each set $V_1, \ldots, V_B, R$ is connected on T, and property 1 above holds. To verify the bounds in property 2, note that the total score of the starting tree $\tilde{T} = T$ is S, and the total score of the final tree belongs to the range $[0, s^* + \sqrt{\kappa s^*})$. Each time we remove a sub-tree $\tilde{T}_i$, the score of $\tilde{T}$ decreases by at least $s^* + \sqrt{\kappa s^*}$. We claim that the score also decreases by at most $(d_{\max} - 1)(s^* + \sqrt{\kappa s^*} + 1) + 1$: This is because i has at most $d_{\max} - 1$ children, and if $\tilde{T}_i$ has total score $\geq (d_{\max} - 1)(s^* + \sqrt{\kappa s^*} + 1)$, then some sub-tree rooted at one of its children j would have total score $\geq s^* + \sqrt{\kappa s^*}$. (The additional $+1$ accounts for a possible $+1$ score on the edge (i, j).) This sub-tree $\tilde{T}_j$ would have been removed under the above reverse-breadth-first-search ordering, so this is not possible. Thus, $\tilde{T}_i$ has total score $< (d_{\max} - 1)(s^* + \sqrt{\kappa s^*} + 1)$, verifying our claim. Then the total number B of sub-trees removed must satisfy

$$\frac{S - (s_* + \sqrt{\kappa s^*})}{(d_{\max} - 1)(s^* + \sqrt{\kappa s^*} + 1) + 1} \leq B \leq \frac{S}{s^* + \sqrt{\kappa s^*}}.$$

Recalling $S = \kappa s^*$, this implies (23) as desired.

Now let $\mathcal{P}^*$ be the partition of $\{1, \ldots, p\}$ induced by $\mathbf{u}^*$ over T, and let $\mathcal{P}$ be the common refinement of $\mathcal{P}^S$, $\mathcal{P}^*$, and $\mathcal{P}'$ constructed above: Each edge of T which connects two different sets of $\mathcal{P}$ must connect two different sets of at least one of $\mathcal{P}^S$, $\mathcal{P}^*$, and $\mathcal{P}'$. Then the subspace K associated to $\mathcal{P}$ contains K^S and K^*, and furthermore

$$|\partial_T \mathcal{P}| \leq |\partial_T \mathcal{P}^S| + |\partial_T \mathcal{P}^*| + |\partial_T \mathcal{P}'| \leq S + s^* + B \leq S + s^* + \sqrt{S}.$$

Here, we have used $|\partial_T \mathcal{P}'| = B$ because $\mathcal{P}'$ consists of $B + 1$ connected sets over T.

For each $b = 1, \ldots, B$, recall the set V_b of $\mathcal{P}'$, and construct a vector $\mathbf{v}^b \in \mathbb{R}^p$ whose coordinates are

$$(\mathbf{v}^b)_i = \begin{cases} (\mathbf{u}^*)_i & \text{if } i \in V_b \\ (\mathbf{P}_K \mathbf{u})_i & \text{if } i \notin V_b. \end{cases}$$

That is, $\mathbf{v}^b$ is equal to $\mathbf{u}^*$ on V_b and equal to $\mathbf{P}_K \mathbf{u}$ outside V_b. Then

$$\|\mathbf{P}_K \mathbf{u} - \mathbf{u}^*\|_2^2 \geq \sum_{b=1}^{B} \sum_{i \in V_b} |(\mathbf{P}_K \mathbf{u})_i - (\mathbf{u}^*)_i|^2 = \sum_{b=1}^{B} \|\mathbf{P}_K \mathbf{u} - \mathbf{v}^b\|_2^2. \tag{24}$$

We claim that $\|\nabla_T \mathbf{v}^b\|_0 \leq S$: Indeed, the edges (i, j) of T where $(\mathbf{v}^b)_i \neq (\mathbf{v}^b)_j$ are contained in the union of $\partial_T \mathcal{P}^*$, $\partial_T \mathcal{P}'$, and the edges of $\partial_T \mathcal{P}^S$ whose endpoints both belong to the complement of V_b. Since $|\partial_T \mathcal{P}^S| = S$, and of these S edges, at least $s^* + \sqrt{\kappa s^*}$ have both endpoints in V_b by property 1 of our construction of $\mathcal{P}'$, this implies $\|\nabla_T \mathbf{v}^b\|_0 \leq s^* + B + (S - s^* - \sqrt{\kappa s^*}) \leq S$.

Finally, we use this to lower-bound the right side of (24): Observe that by construction, $\mathbf{u}^S$ and all of the vectors $\mathbf{v}^b$ for $b = 1, \ldots, B$ belong to the subspace K associated to $\mathcal{P}$. Note that

$$\|\mathbf{u} - \mathbf{v}^b\|_2^2 \geq \|\mathbf{u} - \mathbf{u}^S\|_2^2 \tag{25}$$

by optimality of $\mathbf{u}^S$ and the condition $\|\nabla_T \mathbf{v}^b\|_0 \leq S$ shown above. So, applying the Pythagorean identity for the projection $\mathbf{P}_K$ and its orthogonal projection $\mathbf{P}_K^\perp$,

$$\|\mathbf{P}_K \mathbf{u} - \mathbf{v}^b\|_2^2 = \|\mathbf{u} - \mathbf{v}^b\|_2^2 - \|\mathbf{P}_K^\perp \mathbf{u}\|_2^2 \geq \|\mathbf{u} - \mathbf{u}^S\|_2^2 - \|\mathbf{P}_K^\perp \mathbf{u}\|_2^2 = \|\mathbf{P}_K \mathbf{u} - \mathbf{u}^S\|_2^2.$$

Applying this to (24), we get

$$\|\mathbf{P}_K \mathbf{u} - \mathbf{u}^*\|_2^2 \geq B \cdot \|\mathbf{P}_K \mathbf{u} - \mathbf{u}^S\|_2^2.$$

Combining this with the lower-bound on B in (23) yields the lemma. ■

Appendix C. Proof of Theorem 7

We first extend the result of Lemma 8 to address the discretization error in our approximate projection step (6).

Lemma 12 *In the setting of Lemma 8, suppose that* $\mathbf{u}$ *and* $\mathbf{u}^*$ *are as defined in Lemma 8, but*

$$\mathbf{u}^S = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Delta^p : \|\nabla_T \boldsymbol{\theta}\|_0 \leq S} \|\mathbf{u} - \boldsymbol{\theta}\|_2 \tag{26}$$

where the minimization is over the discrete lattice $\Delta = (\Delta_{\min}, \Delta_{\min} + \delta, \ldots, \Delta_{\max} - \delta, \Delta_{\max})$. If $[-\|\mathbf{u}\|_\infty, \|\mathbf{u}\|_\infty] \subseteq [\Delta_{\min}, \Delta_{\max}]$, then the result of Lemma 8 still holds, with (15) replaced by

$$\|\mathbf{P}_K\mathbf{u} - \mathbf{u}^S\|_2^2 \leq \frac{(d_{\max} - 1)(s^* + \sqrt{S} + 1) + 1}{S - s^* - \sqrt{S}} \|\mathbf{P}_K\mathbf{u} - \mathbf{u}^*\|_2^2 + p\delta^2. \tag{27}$$

Proof The proof is the same as Lemma 8, up until (25) where we used optimality of $\mathbf{u}^S$: We define $\mathcal{P}^S$ and construct $\mathcal{P}$ as in Lemma 8, using this discrete vector $\mathbf{u}^S$. Now let us denote by $\check{\mathbf{u}}^S$ the minimizer of (26) over $\mathbb{R}^p$ rather than over Δ^p. Note that we do not necessarily have $\check{\mathbf{u}}^S \in K^S$, i.e. $\check{\mathbf{u}}^S$ may have a different gradient-sparsity pattern from $\mathbf{u}^S$. However, since $\|\nabla_T \mathbf{v}^b\|_0 \leq S$, we still have the bound $\|\mathbf{u} - \mathbf{v}^b\|_2^2 \geq \|\mathbf{u} - \check{\mathbf{u}}^S\|_2^2$ in place of (25), by optimality of $\check{\mathbf{u}}^S$.

Let $\check{\mathbf{u}}^S_\Delta$ be the vector $\check{\mathbf{u}}^S$ with each entry rounded to the closest value in Δ. Note that the value of $\check{\mathbf{u}}^S$ on each set of its induced partition over T is the average of the entries of $\mathbf{u}$ over this set: This implies that $\|\check{\mathbf{u}}^S\|_\infty \leq \|\mathbf{u}\|_\infty$, and also that the residual $\mathbf{u} - \check{\mathbf{u}}^S$ is orthogonal to $\check{\mathbf{u}}^S - \check{\mathbf{u}}^S_\Delta$. By the given condition on $\Delta_{\min}$ and $\Delta_{\max}$, we have the entrywise bound $\|\check{\mathbf{u}}^S_\Delta - \check{\mathbf{u}}^S\|_\infty \leq \delta$ from the rounding. Then

$$\|\mathbf{u} - \mathbf{v}^b\|_2^2 \geq \|\mathbf{u} - \check{\mathbf{u}}^S\|_2^2 = \|\mathbf{u} - \check{\mathbf{u}}^S_\Delta\|_2^2 - \|\check{\mathbf{u}}^S_\Delta - \check{\mathbf{u}}^S\|_2^2 \geq \|\mathbf{u} - \check{\mathbf{u}}^S_\Delta\|_2^2 - p\delta^2.$$

Since $\check{\mathbf{u}}^S_\Delta \in \Delta^p$ also satisfies $\|\nabla_T \check{\mathbf{u}}^S_\Delta\|_0 \leq S$, optimality of $\mathbf{u}^S$ implies $\|\mathbf{u} - \check{\mathbf{u}}^S_\Delta\|_2^2 \geq \|\mathbf{u} - \mathbf{u}^S\|_2^2$. Substituting above and continuing the proof as in Lemma 8, we get the bound

$$\|\mathbf{P}_K\mathbf{u} - \mathbf{u}^*\|_2^2 \geq B \cdot (\|\mathbf{P}_K\mathbf{u} - \mathbf{u}^S\|_2^2 - p\delta^2),$$

and rearranging and applying the lower-bound for B concludes the proof as before. ■

The second step of the proof is carried out by the following lemma, establishing a key property of the gradient mapping following ideas of Theorem 2.2.7 in (Nesterov, 2013).

Lemma 13 *Let (T_1, T_2) be two trees on $\{1, \ldots, p\}$. Let $(\mathcal{P}_1, \mathcal{P}_2)$ be two partitions of $\{1, \ldots, p\}$, with associated subspaces (K_1, K_2), such that $|\partial_{T_1}\mathcal{P}_1| \leq s$ and $|\partial_{T_2}\mathcal{P}_2| \leq s$ for some sparsity level $s > 0$. Let $K = K_1 + K_2$, and let $\mathbf{P}_K$ be the orthogonal projection onto K.*

Let $\mathcal{L}$ be a loss function satisfying cRSC and cRSS with respect to (T_1, T_2), at sparsity level s and with convexity and smoothness constants $\alpha, L > 0$. Fix $\boldsymbol{\theta}_1 \in K_1$ and define

$$\mathbf{u} = \mathbf{P}_K(\boldsymbol{\theta}_1 - \nabla\mathcal{L}(\boldsymbol{\theta}_1)/L), \qquad \mathbf{v} = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in K} \mathcal{L}(\boldsymbol{\theta}).$$

Then

(a) $\|\mathbf{u} - \mathbf{v}\|_2 \leq \sqrt{1 - \alpha/L} \cdot \|\boldsymbol{\theta}_1 - \mathbf{v}\|_2$*, and*

(b) $\|\boldsymbol{\theta}_1 - \mathbf{v}\|_2 \leq (2/\alpha) \cdot \|\mathbf{P}_K \nabla\mathcal{L}(\boldsymbol{\theta}_1)\|_2$*.*

Proof Denote

$$\mathbf{g} = \mathbf{P}_K \nabla\mathcal{L}(\boldsymbol{\theta}_1).$$

Since $\boldsymbol{\theta}_1 \in K$, we have $\mathbf{u} = \boldsymbol{\theta}_1 - \mathbf{g}/L$. Then

$$\|\mathbf{u} - \mathbf{v}\|_2^2 = \|\boldsymbol{\theta}_1 - \mathbf{v} - \mathbf{g}/L\|_2^2 = \|\boldsymbol{\theta}_1 - \mathbf{v}\|_2^2 + \frac{1}{L^2}\|\mathbf{g}\|_2^2 - \frac{2}{L}\langle \mathbf{g}, \boldsymbol{\theta}_1 - \mathbf{v}\rangle.$$

So part (a) will follow from

$$\langle \mathbf{g}, \boldsymbol{\theta}_1 - \mathbf{v} \rangle \geq \frac{1}{2L}\|\mathbf{g}\|_2^2 + \frac{\alpha}{2}\|\boldsymbol{\theta}_1 - \mathbf{v}\|_2^2. \tag{28}$$

To show (28), observe that $\mathbf{v} \in K = K_1 + K_2$, so we may apply the cRSC condition to $\boldsymbol{\theta}_1$ and $\mathbf{v}$. This gives

$$\mathcal{L}(\mathbf{v}) \geq \mathcal{L}(\boldsymbol{\theta}_1) + \langle \nabla\mathcal{L}(\boldsymbol{\theta}_1), \mathbf{v} - \boldsymbol{\theta}_1 \rangle + \frac{\alpha}{2}\|\mathbf{v} - \boldsymbol{\theta}_1\|_2^2. \tag{29}$$

Then, introducing

$$Q(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}_1) + \langle \nabla\mathcal{L}(\boldsymbol{\theta}_1), \boldsymbol{\theta} - \boldsymbol{\theta}_1 \rangle + \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_1\|_2^2,$$

we get

$$\mathcal{L}(\mathbf{v}) \geq Q(\mathbf{u}) - \frac{L}{2}\|\mathbf{u} - \boldsymbol{\theta}_1\|_2^2 + \langle \nabla\mathcal{L}(\boldsymbol{\theta}_1), \mathbf{v} - \mathbf{u} \rangle + \frac{\alpha}{2}\|\mathbf{v} - \boldsymbol{\theta}_1\|_2^2.$$

Applying $\mathbf{u} - \boldsymbol{\theta}_1 = -\mathbf{g}/L$ and $\mathbf{v} - \mathbf{u} \in K$, this gives

$$\begin{aligned}
\mathcal{L}(\mathbf{v}) &\geq Q(\mathbf{u}) - \frac{1}{2L}\|\mathbf{g}\|_2^2 + \langle \mathbf{g}, \mathbf{v} - \mathbf{u} \rangle + \frac{\alpha}{2}\|\mathbf{v} - \boldsymbol{\theta}_1\|_2^2 \\
&= Q(\mathbf{u}) + \frac{1}{2L}\|\mathbf{g}\|_2^2 + \langle \mathbf{g}, \mathbf{v} - \boldsymbol{\theta}_1 \rangle + \frac{\alpha}{2}\|\mathbf{v} - \boldsymbol{\theta}_1\|_2^2.
\end{aligned}$$

Next, observe that $\mathbf{u} \in K = K_1 + K_2$, so we may apply the cRSS condition to $\boldsymbol{\theta}_1$ and $\mathbf{u}$. This yields $\mathcal{L}(\mathbf{u}) \leq Q(\mathbf{u})$. Since $\mathcal{L}(\mathbf{v}) \leq \mathcal{L}(\mathbf{u})$ by optimality of $\mathbf{v}$, combining these observations gives

$$0 \geq \frac{1}{2L}\|\mathbf{g}\|_2^2 + \langle \mathbf{g}, \mathbf{v} - \boldsymbol{\theta}_1 \rangle + \frac{\alpha}{2}\|\mathbf{v} - \boldsymbol{\theta}_1\|_2^2.$$

Rearranging yields (28), which establishes part (a).

For part (b), let us again apply (29) and the optimality condition $\mathcal{L}(\mathbf{v}) \leq \mathcal{L}(\boldsymbol{\theta}_1)$ to get

$$\begin{aligned}
0 &\geq \langle \nabla\mathcal{L}(\boldsymbol{\theta}_1), \mathbf{v} - \boldsymbol{\theta}_1 \rangle + \frac{\alpha}{2}\|\mathbf{v} - \boldsymbol{\theta}_1\|_2^2 \\
&= \langle \mathbf{g}, \mathbf{v} - \boldsymbol{\theta}_1 \rangle + \frac{\alpha}{2}\|\mathbf{v} - \boldsymbol{\theta}_1\|_2^2 \\
&\geq -\|\mathbf{g}\|_2 \cdot \|\mathbf{v} - \boldsymbol{\theta}_1\|_2 + \frac{\alpha}{2}\|\mathbf{v} - \boldsymbol{\theta}_1\|_2^2.
\end{aligned}$$

Rearranging yields part (b). ■

Proof [Theorem 7] Let $\mathbf{u}_t = \boldsymbol{\theta}_{t-1} - \frac{1}{L}\nabla\mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)$. We claim by induction that

$$[-\|\mathbf{u}_t\|_\infty, \|\mathbf{u}_t\|_\infty] \subseteq [\Delta_{\min}, \Delta_{\max}] \tag{30}$$

and

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2 \leq \Gamma \cdot \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|_2 + \frac{4(1+\gamma)}{\alpha} \cdot \Phi(S') + \delta\sqrt{p} \tag{31}$$

for each $t = 1, \ldots, \tau$.

To start the induction, first observe that for every $t \in \{1, \dots, \tau\}$, the following holds: Fix any $i \in \{1, \dots, p\}$ and let $K = K_{t-1} + K^* + \mathrm{span}(\mathbf{e}_i)$ where (K_{t-1}, K^*) are the subspaces associated to the partitions induced by $(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}^*)$ over T_{t-1}, and $\mathrm{span}(\mathbf{e}_i)$ is the 1-dimensional span of the i^{th} standard basis vector $\mathbf{e}_i$. If $\mathcal{P}$ is the partition associated to K, then $|\partial_{T_{t-1}} \mathcal{P}| \leq S + 2s^* + d_{\max} \leq S'$ because $\|\nabla_{T_{t-1}} \boldsymbol{\theta}_{t-1}\|_0 \leq S$, $\|\nabla_{T_{t-1}} \boldsymbol{\theta}^*\|_0 \leq 2s^*$ by Lemma 1, and $\|\nabla_{T_{t-1}} \mathbf{e}_i\|_0 \leq d_{\max}$. Applying the cRSS property for $\mathcal{L}$ with respect to (T_{t-1}, T_t), we get that the loss $\mathcal{L}(\cdot\,; Z_1^n)$ is L-strongly-smooth restricted to K, meaning for all $\mathbf{u}, \mathbf{v} \in K$,

$$\mathcal{L}(\mathbf{u}; Z_1^n) \leq \mathcal{L}(\mathbf{v}; Z_1^n) + \langle \nabla \mathcal{L}(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle + \frac{L}{2} \|\mathbf{u} - \mathbf{v}\|_2^2.$$

Then applying Eq. (2.1.8) of (Nesterov, 2013) to the loss $\mathcal{L}(\cdot\,; Z_1^n)$ restricted to K, we have for all $\mathbf{u}, \mathbf{v} \in K$ that

$$\|\mathbf{P}_K \nabla \mathcal{L}(\mathbf{u}; Z_1^n) - \mathbf{P}_K \nabla \mathcal{L}(\mathbf{v}; Z_1^n)\|_2 \leq L \|\mathbf{u} - \mathbf{v}\|_2,$$

where $\mathbf{P}_K$ is the orthogonal projection onto K. In particular,

$$\left| \langle \mathbf{e}_i, \nabla \mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n) - \nabla \mathcal{L}(\boldsymbol{\theta}^*; Z_1^n) \rangle \right| \leq L \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|_2.$$

This holds for each standard basis vector $\mathbf{e}_i$, so

$$\frac{1}{L} \|\nabla \mathcal{L}(\boldsymbol{\theta}_{t-1}; Z_1^n)\|_\infty \leq \frac{1}{L} \|\nabla \mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)\|_\infty + \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|_2. \tag{32}$$

Then (30) holds for $t = 1$ by the initialization $\boldsymbol{\theta}_0 = 0$ and the given conditions for $\Delta_{\min}, \Delta_{\max}$.

Suppose by induction that (30) holds for t. We apply Lemma 12 to $T = T_t$, $\mathbf{u}^* = \boldsymbol{\theta}^*$, and $\mathbf{u} = \mathbf{u}_t$. Note that by Lemma 1, $\|\nabla_T \boldsymbol{\theta}^*\|_0 \leq 2s^*$. Then by the definition of the update (6), we have $\mathbf{u}^S = \boldsymbol{\theta}_t$ in Lemma 12. Denote by $\mathcal{P}_2$ the partition guaranteed by Lemma 12, with associated subspace K_2. Then the lemma guarantees that

$$|\partial_{T_t} \mathcal{P}_2| \leq S + 2s^* + \sqrt{S} \leq S',$$

and furthermore

$$\|\mathbf{P}_{K_2} \mathbf{u}_t - \boldsymbol{\theta}_t\|_2 \leq \gamma \cdot \|\mathbf{P}_{K_2} \mathbf{u}_t - \boldsymbol{\theta}^*\|_2 + \delta \sqrt{p}.$$

This bound implies

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2 \leq \|\boldsymbol{\theta}_t - \mathbf{P}_{K_2} \mathbf{u}_t\|_2 + \|\mathbf{P}_{K_2} \mathbf{u}_t - \boldsymbol{\theta}^*\|_2 \leq (1 + \gamma) \|\mathbf{P}_{K_2} \mathbf{u}_t - \boldsymbol{\theta}^*\|_2 + \delta \sqrt{p}. \tag{33}$$

Next, let us apply Lemma 13: Take (T_1, T_2) in Lemma 13 to be (T_{t-1}, T_t). Take $\mathcal{P}_1$ to be the common refinement of the partitions induced by $\boldsymbol{\theta}_{t-1}$ and $\boldsymbol{\theta}^*$ over T_{t-1}, and let $\mathcal{P}_2$ be as above. Then $|\partial_{T_{t-1}} \mathcal{P}_1| \leq S + 2s^* < S'$ and $|\partial_{T_t} \mathcal{P}_2| \leq S'$, so the cRSC and cRSS conditions required in Lemma 13 are satisfied. Let K_1, K_2 be the associated subspaces, and set $K = K_1 + K_2$ and

$$\mathbf{v} = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in K} \mathcal{L}(\boldsymbol{\theta}; Z_1^n).$$

First, we take $\boldsymbol{\theta}_1$ to be $\boldsymbol{\theta}_{t-1}$, and apply Lemma 13(a) with $\mathbf{u} = \mathbf{P}_K \mathbf{u}_t$. This gives

$$\|\mathbf{P}_K \mathbf{u}_t - \mathbf{v}\|_2 \leq \sqrt{1 - \frac{\alpha}{L}} \cdot \|\boldsymbol{\theta}_{t-1} - \mathbf{v}\|_2. \tag{34}$$

Note that $\|\mathbf{P}_{K_2}\mathbf{u}_t - \boldsymbol{\theta}^*\|_2 \leq \|\mathbf{P}_K\mathbf{u}_t - \boldsymbol{\theta}^*\|_2$ because $\boldsymbol{\theta}^* \in K_2 \subseteq K$. Applying this and (34) to (33),

$$\begin{aligned}
\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2 &\leq (1+\gamma)\|\mathbf{P}_K\mathbf{u}_t - \boldsymbol{\theta}^*\|_2 + \delta\sqrt{p} \\
&\leq (1+\gamma)\left(\sqrt{1-\frac{\alpha}{L}}\cdot\|\boldsymbol{\theta}_{t-1} - \mathbf{v}\|_2 + \|\mathbf{v} - \boldsymbol{\theta}^*\|_2\right) + \delta\sqrt{p} \\
&\leq (1+\gamma)\left(\sqrt{1-\frac{\alpha}{L}}\cdot\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_*\|_2 + 2\|\mathbf{v} - \boldsymbol{\theta}_*\|_2\right) + \delta\sqrt{p}. \qquad (35)
\end{aligned}$$

Now, let us apply Lemma 13(b) with $\boldsymbol{\theta}_1$ being $\boldsymbol{\theta}^*$. This gives

$$\|\mathbf{v} - \boldsymbol{\theta}_*\|_2 \leq (2/\alpha)\|\mathbf{P}_K\nabla\mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)\|_2 \leq (2/\alpha)\Phi(S'),$$

the second bound holding by the cPGB assumption. Applying this to (35) establishes (31) at the iterate t.

We may apply (31) recursively for $1, \ldots, t$, using $\boldsymbol{\theta}_0 = 0$ and $1 + \Gamma + \Gamma^2 + \ldots = 1/(1-\Gamma)$, to get

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2 \leq \Gamma^t \cdot \|\boldsymbol{\theta}^*\|_2 + \frac{1}{1-\Gamma}\left(\frac{4(1+\gamma)}{\alpha}\cdot\Phi(S') + \delta\sqrt{p}\right) = \Gamma^t \cdot \|\boldsymbol{\theta}^*\|_2 + \Lambda. \qquad (36)$$

In particular,

$$\|\boldsymbol{\theta}_t\|_2 \leq 2\|\boldsymbol{\theta}^*\|_2 + \Lambda.$$

Then, applying also (32),

$$\begin{aligned}
\|\mathbf{u}_{t+1}\|_\infty &\leq \|\boldsymbol{\theta}_t\|_\infty + \frac{1}{L}\|\nabla\mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)\|_\infty + \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_\infty \\
&\leq \frac{1}{L}\|\nabla\mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)\|_\infty + 3\|\boldsymbol{\theta}^*\|_2 + 2\Lambda.
\end{aligned}$$

Then the given condition for $\Delta_{\min}, \Delta_{\max}$ implies that (30) holds for iteration $t+1$, completing the induction. Finally, the theorem follows by applying (36) at $t = \tau$. ■

Appendix D. Proofs for cRSC, cRSS, and cPGB

Proof [Lemma 6] Note that there are $\binom{p-1}{S}$ different partitions $\mathcal{P}_1$ of $V = \{1, \ldots, p\}$ with $|\partial_{T_1}\mathcal{P}_1| = S$, and similarly for $\mathcal{P}_2$, because each such partition corresponds to cutting S of the $p-1$ edges of T_1. Let $g(S) = S\log(1 + p/S)$. Then there are at most $\binom{p-1}{S}\cdot\binom{p-1}{S} \leq e^{2g(S)}$ different combinations of (K_1, K_2), and hence at most this many subspaces K. Taking a union bound over all such K gives, for any $\zeta > 0$,

$$\mathbb{P}(\max_K \|\mathbf{P}_K\nabla\mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)\|_2 \geq \zeta) \leq e^{2g(S)} \cdot \max_K \mathbb{P}(\|\mathbf{P}_K\nabla\mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)\|_2 \geq \zeta).$$

Note that the dimension of K is less than the sum of dimensions of K_1 and K_2, which is at most $2(S+1)$. Applying a covering net argument, we may find a 1/2-net $\mathcal{N}_{1/2}$ for the set $\{\mathbf{v} \in K :$

$\|\mathbf{v}\|_2 = 1\}$ of cardinality at most 5^{2S+2}. Thus,

$$\begin{aligned}\mathbb{P}(\|\mathbf{P}_K \nabla \mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)\|_2 \geq \zeta) &\leq \mathbb{P}(2 \max_{\mathbf{v} \in \mathcal{N}_{1/2}} |\mathbf{v}^\top \nabla \mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)| \geq \zeta) \\ &\leq 5^{2S+2} \cdot \max_{\mathbf{v} \in \mathcal{N}_{1/2}} \mathbb{P}(2|\mathbf{v}^\top \nabla \mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)| \geq \zeta).\end{aligned}$$

Applying the subgaussian assumption on $\mathbf{v}^\top \nabla \mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)$, we get

$$\mathbb{P}(\max_K \|\mathbf{P}_K \nabla \mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)\|_2 \geq \zeta) \leq e^{2g(S)} \cdot 5^{2S+2} \cdot 2e^{-n\zeta^2/8\sigma^2}.$$

Then for any $k > 0$ and some constant $C_k > 0$ depending only on k, setting $\zeta = \sqrt{C_k \sigma^2 g(S)/n}$ and applying $g(S) \geq \log(1+p)$, we get

$$\mathbb{P}(\max_K \|\mathbf{P}_K \nabla \mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)\|_2 \geq \sqrt{C_k \sigma^2 g(S)/n}) \leq p^{-k}.$$

■

Proof [Proposition 9] We will consider a fixed t, and then apply a union bound over $1 \leq t \leq \tau$.

For cRSC and cRSS, note that $\mathcal{L}(\boldsymbol{\theta}; Z_1^n) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$ for the linear model, which gives $\mathcal{L}(\boldsymbol{\theta}_2; Z_1^n) - \mathcal{L}(\boldsymbol{\theta}_1; Z_1^n) - \langle \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1, \nabla \mathcal{L}(\boldsymbol{\theta}_1; Z_1^n)\rangle = \frac{1}{2n}\|\mathbf{X}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|_2^2$. Then the cRSC and cRSS bounds will hold as long as

$$\sup_K \sup_{\mathbf{u} \in K : \|\mathbf{u}\|_2 = 1} \frac{1}{n}\|\mathbf{X}\mathbf{u}\|_2^2 \leq 3\lambda_1/2 \quad \text{and} \quad \inf_K \inf_{\mathbf{u} \in K : \|\mathbf{u}\|_2 = 1} \frac{1}{n}\|\mathbf{X}\mathbf{u}\|_2^2 \geq \lambda_p/2, \tag{37}$$

where the supremum and infimum are over all subspaces $K = K_1 + K_2$ as in Definition 4. This property (37) is invariant under a common rescaling of $\mathbf{X}^\top\mathbf{X}$, λ_1, and λ_p, so we may assume that $\lambda_p = 1$.

Fixing any such subspace K, note that the dimension of K is upper bounded by $2S' + 2$. Let $\mathbf{P}_K$ be the orthogonal projection onto K, and write $\mathbf{P}_K = \mathbf{Q}_K\mathbf{Q}_K^\top$, where $\mathbf{Q}_K$ has orthonormal columns spanning K. Then $\mathbf{X}\mathbf{Q}_K$ also has independent rows $\mathbf{x}_i^\top\mathbf{Q}_K$, where $\|\mathbf{Q}_K^\top\mathbf{x}_i\|_{\psi_2}^2 \leq D$ and $\mathrm{Cov}[\mathbf{Q}_K^\top\mathbf{x}_i] = \mathbf{Q}_K^\top\boldsymbol{\Sigma}\mathbf{Q}_K$. Applying Eq. (5.25) of Vershynin (2010) to $\mathbf{X}\mathbf{Q}_K$, for any $\zeta > 0$ and some constants $C_3, C_4 > 0$ depending only on D,

$$\mathbb{P}\left[\left\|\frac{1}{n}\mathbf{Q}_K^\top\mathbf{X}^\top\mathbf{X}\mathbf{Q}_K - \mathbf{Q}_K^\top\boldsymbol{\Sigma}\mathbf{Q}_K\right\|_{\mathrm{op}} \geq \max(\omega, \omega^2)\right] \leq 2e^{-C_3\zeta^2}, \qquad \omega \equiv \frac{C_4\sqrt{S'} + \zeta}{\sqrt{n}}.$$

Recall $g(S') = S'\log(1+\frac{p}{S'})$. Note that there are at most $\binom{p-1}{S'} \cdot \binom{p-1}{S'} \leq e^{2g(S')}$ different subspaces K. Taking a union bound over K, and noting that any $\mathbf{u} \in K$ may be represented as $\mathbf{u} = \mathbf{Q}_K\mathbf{v}$ for such K, this yields

$$\mathbb{P}\left[\sup_K \sup_{\mathbf{u} \in K : \|\mathbf{u}\|_2 = 1} \left|\frac{1}{n}\mathbf{u}^\top\mathbf{X}^\top\mathbf{X}\mathbf{u} - \mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u}\right| \geq \max(\omega, \omega^2)\right] \leq 2e^{2g(S') - C_3\zeta^2}$$

When $\|\mathbf{u}\|_2 = 1$, $\mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u} \in [\lambda_p, \lambda_1]$. It follows, with probability at least $1 - 2e^{2g(S') - C_3\zeta^2}$ and under our scaling $\lambda_p = 1$, that

$$\sup_K \sup_{\mathbf{u} \in K : \|\mathbf{u}\|_2 = 1} \frac{1}{n}\|\mathbf{X}\mathbf{u}\|_2^2 \leq \lambda_1 + \max(\omega, \omega^2),$$

and

$$\inf_{K} \inf_{\mathbf{u}\in K:\|\mathbf{u}\|_2=1} \frac{1}{n}\|\mathbf{X}\mathbf{u}\|_2^2 \geq (1-\max(\omega,\omega^2))_+.$$

Then, for any $k > 0$ and some constants $C_1, C_5 > 0$ depending only on k, D, assuming $n \geq C_1 g(S')$ and setting $\zeta = \sqrt{C_5 g(S')}$, (37) holds with probability at least $1 - 2e^{-kg(S')}$. Applying $g(S') \geq \log p$, this probability is at least $1 - 2p^{-k}$.

For cPGB, it follows from the first part of the proof that with probability at least $1 - 2p^{-k}$, $\|\mathbf{X}\mathbf{u}\|_2^2/n^2 \leq 3\lambda_1/2n$ for every such subspace K and every $\mathbf{u} \in K$. Applying Lemma 5.9 of Vershynin (2010) and the assumption $\|e_i\|_{\psi_2}^2 \leq \sigma^2$, conditional on $\mathbf{X}$ and this event, $\mathbf{u}^{\mathsf{T}}\mathbf{X}^\top\mathbf{e}/n$ is a subgaussian random variable with subgaussian parameter $C_6\lambda_1\sigma^2/n$, where $C_6 > 0$ is some absolute constant. Noting that $\nabla\mathcal{L}(\boldsymbol{\theta}^*; Z_1^n) = -\mathbf{X}^\top\mathbf{e}/n$ and applying Lemma 6, $\mathcal{L}$ has the cPGB $\Phi(S') = C_2\sigma\sqrt{\lambda_1 g(S')/n}$ with probability at least $1 - 3p^{-k}$.

The bound for $\|\nabla\mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)\|_\infty = \|\mathbf{X}^\top\mathbf{e}/n\|_\infty$ follows from similarly noting that with probability at least $1 - 2p^{-k}$, $\|\mathbf{X}\mathbf{u}\|_2^2/n^2 \leq 3\lambda_1/2n$ for each standard basis vector $\mathbf{u} \in \mathbb{R}^p$. Conditional on $\mathbf{X}$ and this event, $\mathbf{u}^{\mathsf{T}}\mathbf{X}^\top\mathbf{e}/n$ is subgaussian with parameter $C_6\lambda_1\sigma^2/n$ for every standard basis vector $\mathbf{u}$. Then the bound for $\|\mathbf{X}^\top\mathbf{e}/n\|_\infty$ follows from the subgaussian tail bound and a union bound over all such $\mathbf{u}$. Finally, applying a union bound over $1 \leq t \leq \tau$ completes the proof. ■

Proof [Proposition 11] Similar to the proof of Proposition 9, we consider fixed t and then apply a union bound over $1 \leq t \leq \tau$.

For cRSC and cRSS, note that $\mathcal{L}(\boldsymbol{\theta}; Z_1^n) = \frac{1}{n}\sum_{i=1}^n (b(\mathbf{x}_i^\top\boldsymbol{\theta}) - y_i\mathbf{x}_i^\top\boldsymbol{\theta})$, which gives

$$\begin{aligned}&\mathcal{L}(\boldsymbol{\theta}_2; Z_1^n) - \mathcal{L}(\boldsymbol{\theta}_1; Z_1^n) - \langle\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1, \nabla\mathcal{L}(\boldsymbol{\theta}_1; Z_1^n)\rangle \\ &= \frac{1}{n}\sum_{i=1}^n (b(\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\theta}_2) - b(\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\theta}_1) - b'(\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\theta}_1)x_i^{\mathsf{T}}(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)).\end{aligned}$$

Applying the assumption on b,

$$\frac{\alpha_b}{2n}\|\mathbf{X}(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)\|_2^2 \leq \mathcal{L}(\boldsymbol{\theta}_2; Z_1^n) - \mathcal{L}(\boldsymbol{\theta}_1; Z_1^n) - \langle\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1, \nabla\mathcal{L}(\boldsymbol{\theta}_1; Z_1^n)\rangle \leq \frac{L_b}{2n}\|\mathbf{X}(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)\|_2^2.$$

Then cRSC and cRSS hold for (T_{t-1}, T_t) with probability $1 - 2p^{-k}$, by (37) and the same argument as Proposition 9.

For cPGB, note that $\nabla\mathcal{L}(\boldsymbol{\theta}^*; Z_1^n) = -\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i e_i = -\mathbf{X}^{\mathsf{T}}\mathbf{e}/n$ where $\mathbf{e} = (e_1, \ldots, e_n)$. Similar to the proof of Proposition 9, we condition on $\mathbf{X}$ and the probability $1 - 2e^{-kg(S')}$ event $\mathcal{E}$ that $\frac{1}{n}\|\mathbf{X}\mathbf{u}\|_2^2 \leq 3\lambda_1/2$ for every $K = K_1 + K_2$ and every $\mathbf{u} \in K$. Then similar to the proof of Lemma 6, we get for any $\zeta > 0$

$$\begin{aligned}&\mathbb{P}(\sup_K \|\mathbf{P}_K\mathbf{X}^{\mathsf{T}}\mathbf{e}\|_2/\sqrt{n} > \zeta) \\ &\leq e^{2g(S')} \cdot 5^{2S'+2} \cdot \Big(\sup_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbb{P}(\{2|\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{e}|/\sqrt{n} \geq \zeta\} \cap \mathcal{E}) + 2e^{-kg(S')}\Big).\end{aligned}$$

Note that (21) implies $\mathrm{Var}(e_i) \leq C_3$ where $C_3 > 0$ is some constant depending only on D_1, D_2, β. If $1 < \beta \leq 2$, applying Lemma 14,

$$\mathbb{P}(\sup_K \|\mathbf{P}_K\mathbf{X}^{\mathsf{T}}\mathbf{e}\|_2/\sqrt{n} > \zeta) \leq e^{2g(S')} \cdot 5^{2S'+2} \cdot \Big(2e^{-\zeta^\beta/(C_4\sqrt{\lambda_1})^\beta} + 2e^{-kg(S')}\Big),$$

where $C_4 > 0$ is some constant depending only on D_1, D_2, β. Then for any $k > 0$ and some constant $C_2 > 0$ depending only on k, D, D_1, D_2, β, setting $\zeta = C_2\sqrt{\lambda_1} \cdot g(S')^{1/\beta}$ and applying $g(S') \geq \log p$, we have

$$\mathbb{P}(\sup_K \|\mathbf{P}_K \mathbf{X}^\mathsf{T}\mathbf{e}\|_2/n > C_2\sqrt{\lambda_1/n} \cdot g(S')^{1/\beta}) \leq p^{-k}.$$

If $\beta = 1$, applying Lemma 14, we get

$$\mathbb{P}(\sup_K \|\mathbf{P}_K \mathbf{X}^\mathsf{T}\mathbf{e}\|_2/n > C_2\sqrt{\lambda_1/n}\log n \cdot g(S')) \leq p^{-k}.$$

The bound for $\|\nabla\mathcal{L}(\boldsymbol{\theta}^*; Z_1^n)\|_\infty = \|\mathbf{X}^\top\mathbf{e}/n\|_\infty$ is similar to the proof of Proposition 9. Note that with probability at least $1 - 2p^{-k}$, $\|\mathbf{X}\mathbf{u}_i\|_2^2/n \leq 3\lambda_1/2$ for each standard basis vector $\mathbf{u}_i \in \mathbb{R}^p$ with $1 \leq i \leq p$. We condition on $\mathbf{X}$ and this event $\mathcal{E}'$ and get for any $\zeta > 0$

$$\mathbb{P}(\max_{1\leq i\leq p} |\mathbf{u}_i\mathbf{X}^\mathsf{T}\mathbf{e}|/\sqrt{n} > \zeta) \leq p \cdot \Big(\max_{1\leq i\leq p} \mathbb{P}(\{|\mathbf{u}_i\mathbf{X}^\mathsf{T}\mathbf{e}|/\sqrt{n} > \zeta\} \cap \mathcal{E}') + 2p^{-k}\Big).$$

Similarly, if $1 < \beta \leq 2$, applying Lemma 14, for any $k > 0$ and some constant C_3 depending only on k, D, D_1, D_2, β, we get

$$\mathbb{P}(\max_{1\leq i\leq p} |\mathbf{u}_i\mathbf{X}^\mathsf{T}\mathbf{e}|/n > C_3(\log p)^{1/\beta}\sqrt{\lambda_1/n}) \leq p^{-k}.$$

If $\beta = 1$, applying Lemma 14, we get

$$\mathbb{P}(\max_{1\leq i\leq p} |\mathbf{u}_i\mathbf{X}^\mathsf{T}\mathbf{e}|/n > C_3(\log n)(\log p)\sqrt{\lambda_1/n}) \leq p^{-k}.$$

Finally, applying the union bound over $1 \leq t \leq \tau$ completes the proof. ■

Appendix E. Auxilliary Lemma

The following lemma comes from (Huang et al., 2008, Lemma 1).

Lemma 14 *Suppose $X_1, \ldots, X_n$ are i.i.d. random variables with $\mathbb{E}X_i = 0$ and $\mathrm{Var}(X_i) = \sigma^2$. Further suppose, for $1 \leq d \leq 2$ and certain constants $C_1, C_2 > 0$, their tail probabilities satisfy*

$$\mathbb{P}(|X_i| \geq \zeta) \leq C_1 \exp(-C_2\zeta^d),$$

for all $\zeta > 0$. Let $c_1, \ldots, c_n$ be constants satisfying $\sum_{i=1}^n c_i \leq M^2$ and $W = \sum_{i=1}^n c_i X_i$. Then we have

$$\|W\|_{\psi_d} \leq \begin{cases} K_d M\{\sigma + C_3\}, & 1 < d \leq 2 \\ K_1 M\{\sigma + C_4 \log n\}, & d = 1 \end{cases}$$

where K_d is a positive constant depending only on d, C_3 is some positive constant depending only on C_1, C_2, d and C_4 is some positive constant depending only on C_1, C_2. Consequently,

$$\mathbb{P}(|W| > \zeta) \leq \begin{cases} 2\exp\{-(\zeta/(K_d M(\sigma + C_3)))^d\}, & 1 < d \leq 2 \\ 2\exp\{-\zeta/(K_1 M(\sigma + C_4 \log n))\}. & d = 1 \end{cases}$$