Nearly Non-Expansive Bounds for Mahalanobis Hard Thresholding

Xiao-Tong Yuan XTYUAN1980@GMAIL.COM
Cognitive Computing Lab, Baidu Research
No. 10 Xibeiwang East Road, Beijing 100085, China

Ping Li PINGLI98@GMAIL.COM
Cognitive Computing Lab, Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Given a vector $w \in \mathbb{R}^p$ and a positive semi-definite matrix $A \in \mathbb{R}^{p\times p}$, we study the expansion ratio bound for the following defined Mahalanobis hard thresholding operator of w:

$$\mathcal{H}_{A,k}(w) := \arg\min_{\|\theta\|_0 \le k} \frac{1}{2}\|\theta - w\|_A^2,$$

where $k \le p$ is the desired sparsity level. The core contribution of this paper is to prove that for any $\bar{k}$-sparse vector $\bar{w}$ with $\bar{k} < k$, the estimation error $\|\mathcal{H}_{A,k}(w) - \bar{w}\|_A$ satisfies

$$\|\mathcal{H}_{A,k}(w) - \bar{w}\|_A^2 \le \left(1 + \mathcal{O}\left(\kappa(A, 2k)\sqrt{\frac{\bar{k}}{k-\bar{k}}}\right)\right)\|w - \bar{w}\|_A^2,$$

where $\kappa(A, 2k)$ is the restricted strong condition number of A over $(2k)$-sparse subspace. This estimation error bound is nearly non-expansive when k is sufficiently larger than $\bar{k}$. Specially when A is the identity matrix such that $\kappa(A, 2k) \equiv 1$, our bound recovers the previously known nearly non-expansive bounds for Euclidean hard thresholding operator. We further show that such a bound extends to an approximate version of $\mathcal{H}_{A,k}(w)$ estimated by Hard Thresholding Pursuit (HTP) algorithm. We demonstrate the applicability of these bounds to the mean squared error analysis of HTP and its novel extension based on preconditioning method. Numerical evidence is provided to support our theory and demonstrate the superiority of the proposed preconditioning HTP algorithm.

Keywords: hard thresholding pursuit, Mahalanobis distance, compressed sensing, preconditioning.

1. Introduction

We study the following generalized hard thresholding estimator for truncating a vector $w \in \mathbb{R}^p$ with respect to a positive semi-definite matrix $A \in \mathbb{R}^{p\times p}$:

$$\mathcal{H}_{A,k}(w) := \arg\min_{\|\theta\|_0 \le k} \frac{1}{2}\|\theta - w\|_A^2, \tag{1}$$

where $\|\cdot\|_A$ is the Mahalanobis distance associated with A and $k \le p$ is the sparsity level of truncation. We call $\mathcal{H}_{A,k}(\cdot)$ as Mahalanobis hard thresholding (MHT) estimator. Specially for the identity matrix $A = I_{p\times p}$, the MHT estimator reduces to the conventional Euclidean hard thresholding (HT) operator $\mathcal{H}_k(w) = \arg\min_{\|\theta\|_0\le k} \frac{1}{2}\|\theta - w\|^2$ which plays an important role underlying a large body of greedy pursuit algorithms for compressed sensing (Blumensath and Davies, 2009; Needell and

Tropp, 2009; Foucart, 2011; Wang and Li, 2017) and sparse learning (Bahmani et al., 2013; Blumensath, 2013; Yuan and Zhang, 2013). In general, MHT can be regarded as a projection operator that projects a vector w onto an ℓ_0-ball with respect to Mahalanobis distance. In this paper, we are interested in the following fundamental problem associated with MHT for estimating an unknown sparse vector:

Given a $\bar{k}$-sparse vector $\bar{w}$ with $\bar{k} \leq k$, how close is $\|\mathcal{H}_{A,k}(w) - \bar{w}\|_A$ to $\|w - \bar{w}\|_A$?

First off, we must have that the expansion ratio between $\|\mathcal{H}_{A,k}(w) - \bar{w}\|_A$ and $\|w - \bar{w}\|_A$ should be no larger than 2, which is directly implied by the definition of MHT such that

$$\|\mathcal{H}_{A,k}(w) - \bar{w}\|_A = \|\mathcal{H}_{A,k}(w) - w + w - \bar{w}\|_A \leq \|\mathcal{H}_{A,k}(w) - w\|_A + \|w - \bar{w}\|_A \leq 2\|w - \bar{w}\|_A.$$

Obviously, the above bound with expansion ratio 2 is far from tight especially when $k \gg \bar{k}$. For example, in the extreme case when $k = p$, we just have $\|\mathcal{H}_{A,k}(w) - \bar{w}\|_A = \|w - \bar{w}\|_A$. Therefore, intuitively speaking, the expansion ratio of the MHT estimator is expected to be close to one when k becomes close to p. This inspires us to raise the following key question:

Can we find a tighter expansion ratio bound for $\frac{\|\mathcal{H}_{A,k}(w) - \bar{w}\|_A}{\|w - \bar{w}\|_A}$ that approaches to 1 as $k \to p$?

Although intuitive, the problem is challenging because the computation of $\mathcal{H}_{A,k}(w)$ is itself a compressed sensing problem and thus highly non-trivial to solve. In the reduced case when $A = I_{p\times p}$ is an identity matrix, $\mathcal{H}_k(w)$ has a close-form solution that preserves the top k (in magnitude) entries of w, of which the expansion ratio bound has been well understood in Li et al. (2016); Shen and Li (2018). In sharp contrast, for general $A \succeq 0$, the estimator has no close-form expression and usually needs to be approximately estimated using sparsity recovery algorithms such as those iterative greedy pursuit methods (Pati et al., 1993; Blumensath and Davies, 2009). Therefore, the existing hard thresholding analysis does not readily extend to MHT and we need to propose new treatments to analyze its expansion ratio guarantee even if the estimator is assumed to be exactly known, no menton when $\mathcal{H}_{A,k}(w)$ is approximately estimated by complex sparsity recovery algorithms. Particularly, denote $\mathcal{H}^{\text{HTP}}_{A,k}(w)$ the estimation of MHT via the popularly applied hard thresholding pursuit (HTP) algorithm (Foucart, 2011). We will investigate the closely relevant problem of how to bound the expansion ratio $\|\mathcal{H}^{\text{HTP}}_{A,k}(w) - \bar{w}\|_A / \|w - \bar{w}\|_A$ as tight as possible.

1.1. Motivation

The importance of offering a nearly non-expansive error bound of $\mathcal{H}_k(w)$ for high-dimensional sparse recovery analysis was extensively justified in Shen and Li (2018). The main motivation of studying the generic MHT with metric matrix $A \succeq 0$ comes from the mean squared error analysis of sparsity-constrained least squared regression. Assume that the data sample $D_n = \{x_i, y_i\}_{i=1}^n$ obeys the linear model $y_i = \bar{w}^\top x_i + \varepsilon_i$ where $\bar{w}$ is a k-sparse parameter vector and ε_i's are n i.i.d. zero-mean sub-Gaussian random variables with parameter σ^2. The model can be compactly expressed as $Y = X\bar{w} + \varepsilon$ where $X \in \mathbb{R}^{n\times p}$ is the design matrix and $Y, \varepsilon \in \mathbb{R}^n$ are respectively vectors of response and random noise. In compressed sensing, the following sparsity-constrained least squares regression model is commonly considered for estimating the true sparse signal $\bar{w}$:

$$\hat{w}^{\ell_0} = \underset{\|w\|_0 \leq k}{\arg\min} \left\{ F(w) := \frac{1}{2n} \sum_{i=1}^n (y_i - w^\top x_i)^2 = \frac{1}{2n} \|Y - Xw\|^2 \right\}. \tag{2}$$

In the fixed design setting where X is deterministic, the mean squared error (MSE) is often used as the prediction performance measurement of a sparse estimator $\hat{w}$, which is defined by

$$\texttt{MSE}(\hat{w}, \bar{w}; X) = \frac{1}{n}\|X(\hat{w} - \bar{w})\|^2 = \|\hat{w} - \bar{w}\|_H^2,$$

where $H = \frac{1}{n}X^\top X$ is the sample covariance matrix. It is well understood (see, e.g., Raskutti et al., 2011; Rigollet, 2015) that the following MSE bound of w^{ℓ_0} holds with probability at least $1 - \delta$:

$$\texttt{MSE}(\hat{w}^{\ell_0}, \bar{w}; X) \leq \mathcal{O}\left(\frac{\sigma^2 k \log(p/k)}{n} + \frac{\sigma^2 k}{n} + \frac{\sigma^2 \log(1/\delta)}{n}\right). \tag{3}$$

The above MSE bound immediately implies a parameter estimation error bound of $\hat{w}^{\ell_0}$ given by

$$\|\hat{w}^{\ell_0} - \bar{w}\|^2 \leq \mathcal{O}\left(\frac{1}{\lambda_{\min}(H, 2k)}\texttt{MSE}(\hat{w}^{\ell_0}, \bar{w}; X)\right),$$

where $\lambda_{\min}(H, 2k)$ is the minimal $(2k)$-parse eigenvalue of H. For comparison, the parameter estimation error bound based on objective value sub-optimality analysis scales as $\|\hat{w}^{\ell_0} - \bar{w}\|^2 \leq \mathcal{O}\left(n^{-1}\lambda_{\min}^{-2}(H, 2k)\sigma^2 k \log(p/k))\right)$ (Yuan et al., 2016; Shen and Li, 2017), which is inferior to the above MSE implied bound in the sense of an additional factor $\lambda_{\min}^{-1}(H, 2k)$. Moreover, in ill-conditioned problems where $\lambda_{\min}(H, 2k)$ would be fairly small, the MSE bound itself will be much tighter than the parameter estimation error bounds in terms of signal reconstruction and prediction.

It is, however, not clear so far whether the MSE bound in (3) extends to sparsity recovery algorithms often used for approximately estimating $\hat{w}^{\ell_0}$. As will be shown shortly in this paper, the expansion ratio analysis of MHT turns out to be a useful tool for analyzing the MSE performance of the HTP algorithm for estimating (2). To gain an intuition, let $\hat{w}^{\texttt{HTP}}$ be the corresponding HTP estimator of $\hat{w}^{\ell_0}$ and consider an index set $S = \text{supp}(\bar{w}) \cup \text{supp}(\hat{w}^{\ell_0}) \cup \text{supp}(\hat{w}^{\texttt{HTP}})$. Let $\tilde{w} = \arg\min_{\text{supp}(w)\subseteq S} F(w)$. Then we can verify that $\hat{w}^{\texttt{HTP}} = \mathcal{H}^{\texttt{HTP}}_{H_{SS},k}(\tilde{w})$ and $\|\tilde{w} - \bar{w}\|_{H_{SS}}^2 \leq \mathcal{O}\left(n^{-1}\sigma^2 k \log(p/k)\right)$. Provided that we can well bound the ratio $\|\mathcal{H}^{\texttt{HTP}}_{H_{SS},k}(\tilde{w}) - \bar{w}\|_{H_{SS}}^2 / \|\tilde{w} - \bar{w}\|_{H_{SS}}^2$ from above, it follows directly that $\|\hat{w}^{\texttt{HTP}} - \bar{w}\|_H^2 = \|\hat{w}^{\texttt{HTP}} - \bar{w}\|_{H_{SS}}^2 \leq \mathcal{O}\left(n^{-1}\sigma^2 k \log(p/k)\right)$. This shows that the MSE upper bound of HTP is nearly identical to that of $\hat{w}^{\ell_0}$. A detailed MSE analysis of HTP can be found in Section 4.1. Interestingly, the MHT expansion bound has also been found beneficial for analyzing a preconditioning HTP method which provably enjoys superior computational efficiency to HTP in large-scale problems.

1.2. Main results

As the main result of this paper, we establish in Theorem 3 the following error expansion bound of MHT for estimating any $\bar{k}$-sparse vector $\bar{w}$ and sparsity level $k > \bar{k}$:

$$\|\mathcal{H}_{A,k}(w) - \bar{w}\|_A^2 \leq \min\left\{4, 1 + 3\kappa(A, 2k)\sqrt{\frac{3\bar{k}}{k - \bar{k}}}\right\}\|w - \bar{w}\|_A^2.$$

When A is identity matrix with $\kappa(A, 2k) \equiv 1$, our bound reduces to be identical to those non-expansive error ratio bounds of HT established in Li et al. (2016); Shen and Li (2018). For general $A \succeq 0$, the above result shows that the sparse estimation error of MHT is still nearly non-expansive

when k is large enough. Furthermore, we show in Theorem 5 that a similar error expansion bound to the above one holds for a stationary estimation $\mathcal{H}^{\mathrm{HTP}}_{A,k}(w)$ after $\tilde{\mathcal{O}}\left(k\kappa(A,2k)\right)$ rounds of HTP iteration. Here we have used big o notation $\tilde{\mathcal{O}}$ to hide logarithmic factors. In view of these bounds of MHT, we then provide an MSE analysis for HTP for solving the sparsity-constrained linear regression problem (2). Our result in Theorem 7 shows that for any $k \geq \mathcal{O}(\kappa^2(H,2k)\bar{k})$, a nearly identical MSE bound to (3) can be established for HTP after sufficient rounds of iteration. Particularly, the MSE bound of HTP then implies a stronger parameter estimation error bound of HTP than the prior results. As another significant contribution, we develop a novel preconditioning HTP algorithm and show that the new algorithm enjoys a similar strong MSE guarantee to HTP but with substantially reduced computational complexity in large-scale problems. Preliminary numerical results are provided to verify our theoretical findings and demonstrate the efficiency of our algorithm.

Paper organization. In Section 2 we briefly review the related literature. In Section 3 we present the estimation error ratio bound analysis for exact MHT and its approximate estimation by HTP. In Section 4 we show two implications of the established bounds in MSE analysis of HTP and a novel algorithm extension of HTP via preconditioning method. A numerical study for theory verification and algorithm evaluation is presented in Section 5. The concluding remarks are made in Section 6. Finally, all the technical proofs are relegated to the appendix.

2. Related Work

The problem of learning parsimonious models under sparsity constraint has long been studied with a vast body of beautiful theoretical results and efficient practical algorithms established in signal processing, statistics and machine learning (Bach et al., 2012; Hastie et al., 2015). Early efforts mainly lie in high-dimensional sparse signal recovery, or compressed sensing (Donoho, 2006), for which a bunch of low-complexity greedy approximation methods have been developed including orthogonal matching pursuit (OMP) (Pati et al., 1993), iterative hard thresholding (IHT) (Blumensath and Davies, 2009), compressed sampling matching pursuit (CoSaMP) (Needell and Tropp, 2009) and subspace pursuit (SP) (Dai and Milenkovic, 2009). These compressed sensing algorithms have later been extended to broader class of sparse machine learning problems with loss functions beyond least squared error (Shalev-Shwartz et al., 2010; Bahmani et al., 2013; Yuan et al., 2018). Statistical consistency of learning with sparsity constraint are now well understood for some popular statistical learning models including least squares regression, logistic regression and principle component analysis (Ma, 2013; Rigollet, 2015; Foucart and Rauhut, 2017). The out-of-sample generalization theory of sparsity-constrained/regularized learning models was studied in Chen and Lee (2018); Abramovich and Grinshtein (2019); Yuan and Li (2020). Particularly, the mean squared error bound of the ℓ_0-estimator for linear regression models, which is mostly closely relevant to our work, has been analyzed in Raskutti et al. (2011); Rigollet (2015).

Among others, the IHT-style methods are popularly studied as they have been witnessed to offer attractive efficiency and scalability in many cases (Yuan and Zhang, 2013; Jain et al., 2014; Li et al., 2016). The rate of convergence and parameter estimation error of IHT-style methods were initially analyzed under proper restricted isometry property (RIP) (Candes and Tao, 2005), or restricted strong condition number, bounding conditions (Blumensath and Davies, 2009; Foucart, 2011; Yuan et al., 2018). The RIP-type conditions, however, tend to be too stringent to hold in real-world high dimensional data analysis problems. It is noteworthy that the dependence of IHT sparsity recovery analysis on RIP-type conditions mainly attributes to the HT operator which is

non-convex and expansive in most cases. This is contrast to the convex soft thresholding operator used by ℓ_1-estimators which is non-expansive (Boyd and Vandenberghe, 2004). To remedy this deficiency, the following HT sub-optimality bound was proved in Jain et al. (2014)

$$\|\mathcal{H}_k(w) - w\|^2 \leq \frac{p-k}{p-\bar{k}}\|w - \bar{w}\|^2.$$

Based on this bound, it was shown in that paper that under proper sparsity level relaxation, the high-dimensional estimation consistency of IHT can be established without imposing RIP-type conditions on the objective function. One shortcoming of the above bound lies in that the expansion factor relies on feature dimension which could be huge. Later, the following dimension-free bound of HT has been essentially independently established in Li et al. (2016); Shen and Li (2018):

$$\|\mathcal{H}_k(w) - \bar{w}\|^2 \leq \left(1 + \mathcal{O}\left(\sqrt{\frac{\bar{k}}{k-\bar{k}}}\right)\right)\|w - \bar{w}\|^2. \tag{4}$$

Such a nearly non-expansive estimation error bound has been demonstrated beneficial for analyzing the parameter estimation and sparsity recovery accuracy of IHT and its stochastic/distributed variants without assuming RIP-type conditions (Li et al., 2016; Shen and Li, 2018; Zhou et al., 2018; Liu et al., 2019; Yuan and Li, 2020). Despite the remarkable success achieved in understanding the benefit of tighter HT bounds for analyzing the IHT-style algorithms, these results are mostly relevant to the objective value sub-optimality and parameter estimation accuracy. In contrast, the theoretical understanding of the MHT operator, which is expected to be a backbone for the MSE analysis of IHT-style methods, yet still remains an open issue that we aim to study in this paper.

3. Estimation Error Analysis of MHT

In this section, we analyze the expansion ratio of MHT for estimating a sparse vector. We distinguish our analysis in two settings with regard to the exactness of MHT: in the first ideal setting we assume that the MHT operator is exactly known, while in the second inexact but more realistic setting we focus on the case that MHT is approximately estimated by the HTP algorithm.

3.1. A key lemma

We start by presenting a key lemma which lays the foundation for deriving the error expansion bounds of MHT and its approximate estimation via HTP. In the following analysis, we denote $\lambda_{\max}(A,k) = \max_{\|x\|=1,\|x\|_0\leq k} x^\top Ax$ the largest k-sparse eigenvalue of a positive semi-definite matrix A. The smallest k-sparse eigenvalue $\lambda_{\min}(A,k)$ is defined analogously. We denote $\text{supp}(u)$ the index set of nonzero entries of a vector u. A full list of notation can be found in Appendix A.

Lemma 1 *Consider a given vector $w \in \mathbb{R}^p$, a $\bar{k}$-sparse vector $\bar{w} \in \mathbb{R}^p$, and a positive semi-definite matrix $A \in \mathbb{R}^{p\times p}$. Let u be a k-sparse vector and $S_u = supp(u)$. Assume that*

$$A_{S_u,:}(u - w) = 0, \;\; [u]_{\min} \geq \frac{\|A(u-w)\|_\infty}{\nu},$$

where $\nu > 0$ is some scalar. Then for any $k \geq \left(1 + 3\nu^2\lambda_{\min}^{-2}(A, 2k)\right)\bar{k}$, the following estimation error expansion bound holds:

$$\|u - \bar{w}\|_A^2 \leq \left(1 + \frac{3\nu}{\lambda_{\min}(A,2k)}\sqrt{\frac{3\bar{k}}{k-\bar{k}}}\right)\|w - \bar{w}\|_A^2.$$

Proof The key ingredient of the proof argument is to bound the numerator and denominator of the ratio $(\|u-\bar{w}\|_A^2 - \|w-\bar{w}\|_A^2)/\|w-\bar{w}\|_A^2$ respectively from above based on the orthogonal condition $A_{S_u,:}(u-w)=0$ and from below using the strong-signal condition $[u]_{\min} \geq \frac{\|A(u-w)\|_\infty}{\nu}$. See Appendix B.1 for a detailed proof of this result. ■

Remark 2 *We comment on the conditions required in the lemma. The condition $A_{S_u,:}(u-w)=0$ implies that $u^\top A(u-w)=0$, i.e., u and $A(u-w)$ are orthogonal. The $[u]_{\min} \geq \|A(u-w)\|_\infty/\nu$ basically requires that the magnitude of the non-zero entries of u should be significantly larger than those of the vector $A(u-w)$. As we will shortly see in the subsequent sections that these two conditions can be fulfilled by MHT and its HTP estimation as well. Finally, the condition $k \geq \left(1+3\nu^2\lambda_{\min}^{-2}(A,2k)\right)\bar{k}$ guarantees that the expansion ratio is no larger than 4.*

3.2. Results for exact MHT

We first consider an ideal setting where the MHT operator $\mathcal{H}_{A,k}(w)$ in (1) is exactly known. The following is our main result on the estimation error expansion bound of exact MHT.

Theorem 3 (Expansion bound of exact MHT) *Let A be a positive semi-definite matrix. Consider a given vector w and a target $\bar{k}$-sparse vector $\bar{w}$. Then for any $k > \bar{k}$, the following estimation error bound of $\mathcal{H}_{A,k}(w)$ holds:*

$$\|\mathcal{H}_{A,k}(w)-\bar{w}\|_A^2 \leq \min\left\{4, 1+3\kappa(A,2k)\sqrt{\frac{3\bar{k}}{k-\bar{k}}}\right\}\|w-\bar{w}\|_A^2.$$

Proof It is straightforwardly known that $\|\mathcal{H}_{A,k}(w)-\bar{w}\|_A^2 \leq 4\|w-\bar{w}\|_A^2$. Let $S=\mathrm{supp}(\mathcal{H}_{A,k}(w))$. Based on a standard result in Lemma 19 we can show that

$$A_{S,:}(\mathcal{H}_{A,k}(w)-w)=0, \quad [\mathcal{H}_{A,k}(w)]_{\min} \geq \frac{\|A(\mathcal{H}_{A,k}(w)-w)\|_\infty}{\lambda_{\max}(A,2k)}. \tag{5}$$

Then by substantializing Lemma 1 with $u=\mathcal{H}_{A,k}(w)$ and $\nu=\lambda_{\max}(A,2k)$ we immediately get the other part of the bound. See Appendix B.2 for a full proof of this result. ■

Remark 4 *To compare with the nearly non-expansive expansion ratio bound of HT in (4), our bound in Theorem 3 for generic MHT enjoys an almost identical near non-expansion property but at the cost of an additional factor of $\kappa(A,2k)$. We remark that the universal constants in our bound are by no means optimal and might be further improved with more careful treatment.*

As discussed previously that the MHT operator by definition is a compressed sensing problem which is generally non-convex and NP-hard. This means that unlike the Euclidean HT operator which has close-form expression, it is hopeless to find an exact estimation of MHT in polynomial time and one must instead seek approximate solutions. The blessing here is that our analysis of exact MHT does not directly hinge the global optimality of the operator. Rather, we only need to make use of the first-order optimality condition of MHT over its own supporting set and a strong-signal property of the operator as listed in (5). This offers the potential to extend the near non-expansion guarantee on exact MHT to its inexact counterpart approximately estimated via HTP-style methods.

Algorithm 1: Hard Thresholding Pursuit (HTP) for approximately solving $\min_{\|\theta\|_0 \le k} f(\theta)$

Input : Learning rate $\eta > 0$.
Output: $\theta^{(t)}$.
Initialization: $\theta^{(0)}$ *with* $\|\theta^{(0)}\|_0 \le k$ *(typically* $\theta^{(0)} = 0$*),* $t = 1$.
repeat
  (**S1**) Compute $\tilde{\theta}^{(t)} = \mathcal{H}_k\left(\theta^{(t-1)} - \eta \nabla f(\theta^{(t-1)})\right)$;
  (**S2**) Compute $\theta^{(t)} = \arg\min f(\theta)$ subject to $\mathrm{supp}(\theta) \subseteq \mathrm{supp}(\tilde{\theta}^{(t)})$;
  $t = t + 1$;
until $S^{(t)} = S^{(t-1)}$;

3.3. Results for inexact MHT

We now move to study a more realistic case where MHT is approximately estimated via the HTP algorithm (Foucart, 2011) as outlined in Algorithm 1. The key observation here is that HTP can find a stationary solution of MHT with identical properties to those in (5) at exponentially fast rate of convergence. Based on such an observation, we can establish the following main result on the estimation error expansion bound of the inexact MHT estimator $\mathcal{H}^{\mathrm{HTP}}_{A,k}(w)$ estimated by HTP.

Theorem 5 (Expansion bound of inexact MHT) *Let A be a positive semi-definite matrix. Consider a given vector w and a target $\bar{k}$-sparse vector $\bar{w}$. Set the learning rate $\eta = \frac{1}{2\lambda_{\max}(A,2k)}$ for HTP invoked to $f(\theta) = \frac{1}{2}\|\theta - w\|_A^2$. Then for any $k \ge (1 + 12\kappa^2(A, 2k))\bar{k}$, the following estimation error bound holds after $\tilde{\mathcal{O}}\left(k\kappa(A, 2k)\right)$ rounds of HTP iteration:*

$$\|\mathcal{H}^{HTP}_{A,k}(w) - \bar{w}\|_A^2 \le \left(1 + 6\kappa(A, 2k)\sqrt{\frac{3\bar{k}}{k - \bar{k}}}\right)\|w - \bar{w}\|_A^2.$$

Proof As a key step, we first prove Lemma 20 which tells that for restricted strongly convex functions, HTP with proper learning rate converges exponentially fast to a stationary point with the desirable first-order optimality and strong-signal properties. By substantializing this lemma to MHT we can show that the following conditions hold after $\tilde{\mathcal{O}}\left(k\kappa(A, 2k)\right)$ rounds of HTP iteration:

$$A_{S,:}(\mathcal{H}^{\mathrm{HTP}}_{A,k}(w) - w) = 0, \quad [\mathcal{H}^{\mathrm{HTP}}_{A,k}(w)]_{\min} \ge \frac{\|A(\mathcal{H}^{\mathrm{HTP}}_{A,k}(w) - w)\|_\infty}{2\lambda_{\max}(A, 2k)}, \tag{6}$$

where $S = \mathrm{supp}(\mathcal{H}^{\mathrm{HTP}}_{A,k}(w))$. The desired bound then follows immediately by applying Lemma 1 to $u = \mathcal{H}^{\mathrm{HTP}}_{A,k}(w)$ and $\nu = 2\lambda_{\max}(A, 2k)$. A full proof is provided in Appendix B.3. ■

Remark 6 *We remark that the required $\tilde{\mathcal{O}}\left(k\kappa(A, 2k)\right)$ iteration complexity to find a stationary sparse solution satisfying (6) is higher by a factor k than the corresponding $\tilde{\mathcal{O}}\left(\kappa(A, 2k)\right)$ complexity of HTP for sparse parameter estimation and loss minimiztion (Jain et al., 2014; Yuan et al., 2018). It is an interesting open issue to explore the opportunity of tightening such a complexity bound in our considered problem regime, which we will leave for future investigation.*

4. Implications

In this section, we discuss two implications of our main results in the theoretical understanding and algorithm extension of HTP for the sparsity-constrained least squares regression problem (2). We first analyze the MSE performance of the conventional HTP based on Theorem 5. Then we present a novel preconditioning HTP method along with its MSE and computational complexity analysis.

4.1. MSE analysis of HTP

The sparsity recovery performance of HTP is typically measured by parameter estimation error and loss value sub-optimality (Foucart, 2011; Jain et al., 2014; Li et al., 2016; Yuan et al., 2018). The bounds of these measurements usually depend on the restricted strong convexity parameter of objective function. In contrast, as discussed in the motivation section that the MSE bound of $\hat{w}^{\ell_0}$ in (3) is free of the dependency on restricted strong convexity which can lead to tighter bounds on parameter estimation error. However, it remains an open question whether such an attractive MSE bound can be generalized to HTP for estimating $\hat{w}^{\ell_0}$. In view of the expansion ratio bound in Theorem 5, we can establish the following result which shows that an almost identical MSE bound still holds for HTP, and thus answer the question affirmatively.

Theorem 7 (MSE bound of HTP) *Let $w^{(t)}$ be the output of HTP with learning rate $\eta = \frac{1}{2\lambda_{\max}(H,2k)}$ for estimating the sparse least squares regression problem (2) after $t = \tilde{\mathcal{O}}(k\kappa(H,2k))$ rounds of sufficient iteration. Then for any $k \geq (1 + 12\kappa^2(H,2k))\bar{k}$, the following bound holds with probability at least $1-\delta$:*

$$\mathit{MSE}(w^{(t)}, \bar{w}; X) \leq \mathcal{O}\left(\kappa(H,2k)\sqrt{\frac{\bar{k}}{k-\bar{k}}}\left(\frac{\sigma^2 k\log(p/k)}{n} + \frac{\sigma^2 k}{n} + \frac{\sigma^2\log(1/\delta)}{n}\right)\right).$$

Proof A proof of this result is provided in Section C.1. ■

Remark 8 *We comment that the scale factor $\kappa(H,2k)\sqrt{\bar{k}/(k-\bar{k})} = \mathcal{O}(1)$ under the condition of $k \geq (1 + 12\kappa^2(H,2k))\bar{k}$. As a direct consequence of the MSE bound in Theorem 7, the squared parameter estimation error of $w^{(t)}$ can be bounded with probability at least $1-\delta$ as*

$$\|w^{(t)} - \bar{w}\|^2 \leq \mathcal{O}\left(\frac{1}{\lambda_{\min}(H,2k)}\left(\frac{\sigma^2 k\log(p/k)}{n} + \frac{\sigma^2 k}{n} + \frac{\sigma^2\log(1/\delta)}{n}\right)\right),$$

which is substantially tighter than the existing bounds (Jain et al., 2014; Shen and Li, 2018; Yuan et al., 2018) in the considered setting in terms of the dependence on $\lambda_{\min}(H,2k)$. Also, the MSE bound in Theorem 7 essentially reveals that the MSE lower bound of Zhang et al. (2014) for polynomial-time sparse linear regression estimators can be attained efficiently by HTP.

4.2. A preconditioning HTP method

We further show an application of the near non-expansion bounds of MHT to analyzing a novel preconditioning variant of HTP, namely PC-HTP, for solving the sparsity-constrained least squares regression problem (2). As outlined in Algorithm 2, the PC-HTP algorithm contains two nested

Algorithm 2: Preconditioning Hard Thresholding Pursuit (PC-HTP) for solving the ℓ_0-estimator (2)

Input : Hyper parameter $\gamma > 0$.
Output: $w^{(t)}$.
Initialization *Sample* $\tilde{D}_m \subseteq D_n$ *to form* $\tilde{F}(w) := \frac{1}{2m}\sum_{(x_i,y_i)\in\tilde{D}_m}(y_i - w^\top x_i)^2$. *Set* $w^{(0)} = 0$.
for $t = 1, 2, ...$ **do**
(**S1**) Construct a quadratic function

$$P^{(t-1)}(w) := \langle \nabla F(w^{(t-1)}) - \nabla\tilde{F}(w^{(t-1)}), w\rangle + \frac{\gamma}{2}\|w - w^{(t-1)}\|^2 + \tilde{F}(w); \tag{7}$$

(**S2**) Estimate k-sparse estimator $w^{(t)} = \arg\min_{\|w\|_0\le k} P^{(t-1)}(w)$ via HTP.
end

loops: 1) in the outer-loop we construct a quadratic function $P^{(t-1)}$ as expressed in (7) based on a stochastic approximate function $\tilde{F}$ and the previous full gradient $\nabla F(w^{(t-1)})$; and 2) in the subsequent inner-loop we minimize $P^{(t-1)}$ under the same sparsity constraint using HTP to obtain the updated estimator $w^{(t)}$.

4.2.1. THE PRECONDITIONING BEHAVIOR

We first roughly justify the preconditioning behavior of PC-HTP for quadratic objective functions in the unconstrained case with $k = p$ such that the cardinality constraint is inactive. Let $\tilde{H}$ denote the Hessian of $\tilde{F}$. In this case, since by definition $w^{(t)}$ minimizes $P^{(t-1)}(w)$, we must have $0 = \nabla P^{(t-1)}(w^{(t)}) = \nabla\tilde{F}(w^{(t)}) - \nabla\tilde{F}(w^{(t-1)}) + \nabla F(w^{(t-1)}) + \gamma(w^{(t)} - w^{(t-1)}) = (\tilde{H} + \gamma I)(w^{(t)} - w^{(t-1)}) + \nabla F(w^{(t-1)})$, which then implies $w^{(t)} = w^{(t-1)} - \left(\tilde{H} + \gamma I\right)^{-1}\nabla F(w^{(t-1)})$. This is essentially an approximate Newton iteration form in the sense that $\tilde{H} + \gamma I$ is a stochastic approximate to the full Hessian H. To see further the preconditioning effect, we note $\nabla F(w^*) = 0$ and thus the above leads to $w^{(t)} - w^* = \left(I - \left(\tilde{H} + \gamma I\right)^{-1} H\right)(w^{(t-1)} - w^*)$. The effect of preconditioning comes from the fact that when the stochastic approximation $\tilde{F}(w)$ is sufficiently close to $F(w)$ and γ is sufficiently small, the condition number of $(\tilde{H} + \gamma I)^{-1}H$ would be close to 1, which then implies the contraction factor $\left\|I - \left(\tilde{H} + \gamma I\right)^{-1} H\right\|$ would be much smaller than one. Therefore, $\tilde{F}(w)$ essentially serves as a preconditioner that is expected to potentially improve the rate of convergence. From the perspective of algorithmic framework, PC-HTP shares a similar spirit of preconditioning to the distributed inexact Newton pursuit (DINPS) method for distributed learning with sparsity (Liu et al., 2019), although our method is designed in the context of single-machine compressed sensing. Nevertheless, as we will shortly addressed that the analysis of PC-HTP is substantially different from DINPS: we focus on the MSE analysis of PC-HTP which in turn will imply a tighter parameter estimation error bound than that of DINPS.

4.2.2. MSE ANALYSIS

The following is our main result on the MSE performance of PC-HTP.

Theorem 9 (MSE bound of PC-HTP) *Let H and $\tilde{H}$ be the Hessian matrices of F and $\tilde{F}$, respectively. Assume that $\|\tilde{H} - H\| \leq \gamma$. Let $\bar{w}$ be a $\bar{k}$-sparse vector. Let $s = 2k + \bar{k}$. If $k \geq \left(1 + \mathcal{O}\left(\frac{(\lambda_{\max}(H,s)+\gamma)^2(\lambda_{\min}(H,s)+\gamma)^2}{\lambda^4_{\min}(H,s)}\right)\right)\bar{k}$, then with probability at least $1-\delta$, Algorithm 2 will output $w^{(t)}$ satisfying*

$$\mathtt{MSE}(w^{(t)}, \bar{w}; X) \leq \mathcal{O}\left(\left(\frac{\lambda_{\min}(H,s)+\gamma}{\lambda_{\min}(H,s)}\right)^2\left(\frac{\sigma^2 s\log(p/s)}{n} + \frac{\sigma^2 s}{n} + \frac{\sigma^2\log(1/\delta)}{n}\right)\right)$$

after $t \geq \tilde{\mathcal{O}}\left(\frac{\lambda_{\min}(H,s)+\gamma}{\lambda_{\min}(H,s)}\right)$ rounds of iteration.

Proof See Appendix C.2 for a proof of this result. ■

Remark 10 *For the sake of simplicity, we focus on the ideal case where the inner-loop compressed sensing problem is solved exactly. Our analysis can be easily extended to the setting where the inner-loop sub-problem is solved approximately via HTP.*

The following result is a corollary of Theorem 9 when the subset $\tilde{D}_m$ for constructing $\tilde{F}$ is uniformly randomly sampled from the entire sample D_n with $m = \tilde{\mathcal{O}}\left(\lambda^{-2}_{\min}(H,s)\right)$.

Corollary 11 *Assume that $\|x_i\| \leq 1, \forall i \in [n]$ and that $\tilde{D}_m$ is a uniform random subset of D_n. Assume the conditions in Theorem 9 hold. For any $\delta \in (0,1)$, set $m = \mathcal{O}\left(\lambda^{-2}_{\min}(H,s)\log(p/\delta)\right)$. Then with probability at least $1-\delta$ over the randomness associated with model noise and $\tilde{D}_m$, Algorithm 2 with $\gamma = \mathcal{O}\left(\lambda_{\min}(H,s)\right)$ will output $w^{(t)}$ satisfying*

$$\mathtt{MSE}(w^{(t)}, \bar{w}; X) \leq \mathcal{O}\left(\frac{\sigma^2 s\log(p/s)}{n} + \frac{\sigma^2 s}{n} + \frac{\sigma^2\log(1/\delta)}{n}\right)$$

after $t \geq \tilde{\mathcal{O}}\left(1\right)$ rounds of iteration.

Proof See Appendix C.3 for a proof of this result. ■

Remark 12 *The MSE bound in Corollary 11 readily implies that the squared parameter estimation error $\|w^{(t)} - \bar{w}\|^2$ can be bounded as $\tilde{\mathcal{O}}\left(\lambda^{-1}_{\min}(H,s)n^{-1}\sigma^2 s\right)$ which is tighter than those $\tilde{\mathcal{O}}\left(\lambda^{-2}_{\min}(H,s)n^{-1}\sigma^2 s\right)$ bounds of DINPS (Liu et al., 2019) in terms of the dependence on $\lambda^{-1}_{\min}(H,s)$.*

4.3. Computational complexity

We next analyze the computation complexity of PC-HTP to understand its overall computational efficiency. We consider using conjugate gradient method to solve the debiasing step (see the step **S2** of Algorithm 1) of HTP when invoked to the sparse estimator $w^{(t)} = \arg\min_{\|w\|_0 \leq k} P^{(t-1)}(w)$. The amount of computation is measured by matrix-vector product for gradient evaluation. As a consequence of Theorem 5 and Corollary 11, the following result summaries the computational complexity of PC-HTP in the considered setting.

Corollary 13 (Computational complexity of PC-HTP) *Assume the conditions in Corollary 11 hold. For any $\delta \in (0,1)$, set $m = \mathcal{O}(\lambda_{\min}^{-2}(H,s)\log(p/\delta))$ and $\gamma = \mathcal{O}(\lambda_{\min}(H,s))$. Then with probability at least $1-\delta$, the computational complexity of PC-HTP for attaining*

$$MSE(w^{(t)}, \bar{w}; X) \leq \mathcal{O}\left(\frac{\sigma^2 s\log(p/s)}{n} + \frac{\sigma^2 s}{n} + \frac{\sigma^2\log(1/\delta)}{n}\right)$$

is dominantly bounded by $\tilde{\mathcal{O}}\left(nk + k\kappa(H,2k)\left(mk + k^2\sqrt{\kappa(H,2k)}\right)\right)$.

Proof See Section C.4 for a proof of this result. ■

Remark 14 *For comparison, it can be verified that the computational complexity of the conventional HTP to achieve a similar level of MSE is dominated by*

$$\tilde{\mathcal{O}}\left(k\kappa(H,2k)\left(nk + k^2\sqrt{\kappa(H,2k)}\right)\right) = \tilde{\mathcal{O}}\left(nk^2\kappa(H,2k) + k^3\kappa^{1.5}(H,2k)\right).$$

If $k \gg m\kappa^{-1/2}(H,2k)$, then the leading terms in the complexity bound of PC-HTP are of the order $\tilde{\mathcal{O}}\left(nk + k^3\kappa^{1.5}(H,2k)\right)$. Then in the big sample regime where $n \gg k\sqrt{\kappa(H,2k)}$, the complexity of PC-HTP would be considerably cheaper than that of HTP to achieve comparable MSE.

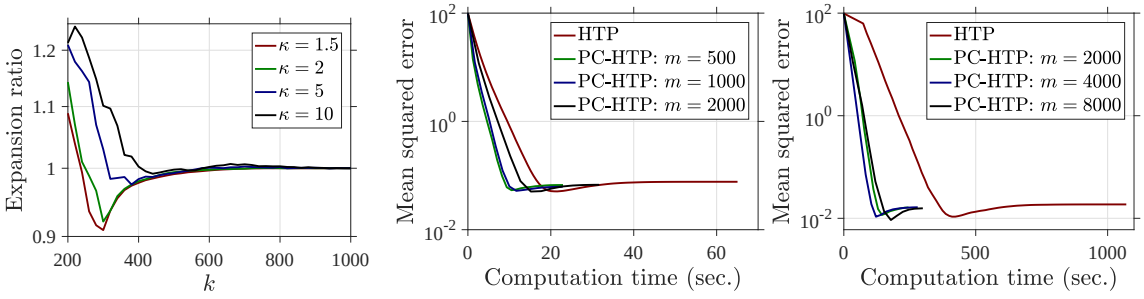
5. Numerical Study

In this section, we carry out a preliminary numerical study to verify the nearly non-expansive bounds of MHT established in Section 3 and evaluate the actual computational performance of PC-HTP as presented in Section 4.2.

5.1. Theory verification

The result in Theorem 5 suggests that given w and $\bar{w}$, the expansion ratio bound of the HTP-based MHT estimator $\mathcal{H}_{A,k}^{\text{HTP}}(w)$ relies on the (restricted) condition number of A and the truncation sparsity level k. To verify this theory, we consider a $\bar{k}$-sparse vector $\bar{w} \in \mathbb{R}^p$ whose non-zero entries are sampled from Gaussian distribution $\mathcal{N}(10,1)$, and construct a dense vector $w = \bar{w} + \bar{w}' + \varepsilon$ where $\bar{w}'$ is a $\bar{k}$-sparse vector satisfying $\bar{w}^\top\bar{w}' = 0$ and its non-zero entries are sampled from Gaussian distribution $\mathcal{N}(0,10^2)$, and ε is a standard Gaussian noise vector. We fix $p = 1000, \bar{k} = 200$ and test how the expansion ratio $\|\mathcal{H}_{A,k}^{\text{HTP}}(w) - \bar{w}\|_A / \|w - \bar{w}\|_A$ evolves under different A with condition number $\kappa \in \{1.5, 2, 5, 10\}$[1] and varying truncation sparsity level k with $k/p \in [0.2, 1]$. Figure 1(*a*) shows the corresponding expansion ratio evolving curves. From this set of results we can make the following two observations: 1) for each fixed A with condition number κ, the expansion ratio of MHT is relatively large when k is relatively small and the ratio converges to one as k approaches to p; and 2) for each fixed sparsity level k, the expansion ratio grows larger as the condition number κ increases. These numerical evidences well support the theoretical prediction in Theorem 5.

1. We first generate a semi-positive definite matrix $A' \succeq 0$ with $\lambda_{\min}(A') = 0$, and then we set $A = A' + \beta I$ with $\beta = \frac{\lambda_{\max}(A')}{\kappa - 1}$. It can be verified that the condition number of A equals to κ.



(*a*) MHT bound verification (*b*) PC-HTP evaluation wiht $n = p$ (left) and $n = 4p$ (right)

Figure 1: Theory verification for MHT and computational efficiency evaluation for PC-HTP.

5.2. Algorithm evaluation for PC-HTP

We now turn to evaluate the computational efficiency of the proposed PC-HTP algorithm for estimating the sparse least squares regression model in (2). The feature points $\{x_i\}_{i=1}^n$ are sampled from standard multivariate Gaussian distribution and the responses $\{y_i\}_{i=1}^n$ are generated according to a linear model $y_i = \bar{w}^\top x_i + \varepsilon_i$ where $\bar{w}$ is a $\bar{k}$-sparse parameter vector whose non-zero entries are sampled from standard Gaussian distribution, and ε_i are a standard Gaussian noises. For this experiment, we set feature dimension $p = 10000$ and sparsity level $\bar{k} = 100$. We evaluate the MSE performance of PC-HTP under varying subset sample size m with $m/n \in \{0.05, 0.1, 0.2\}$ and take the conventional HTP as a baseline for comparison. For each m, we set the hyper-parameter $\gamma = 1/\sqrt{m}$ for PC-HTP. Figure 1(*b*) shows the MSE evolving curves of HTP and PC-HTP as functions of wall-clock computation time (in second) for $n = p$ (left panel) and $n = 4p$ (right panel). From this group of results we can observe that: 1) for all the configurations of sample size n and subset size m, PC-HTP is consistently faster than HTP in MSE convergence and the margin becomes more significant for relatively larger n; and 2) for each considered n, the most efficient implementation of PC-HTP occurs when using relatively smaller m. These observations are consistent with the computational complexity result in Corollary 13 and the related discussions in Remark 14.

6. Conclusions

In this paper, we studied the expansion ratio bound of the MHT operator which plays a fundamental role in analyzing the MSE performance of compressed sensing algorithms such as HTP. Traditional expansion analysis for hard thresholding, however, does not readily extend to MHT due to its non-convexity and NP-hardness. For an ideal case where the operator is assumed to be exactly solved, we established a nearly non-expansive bound for MHT when the truncation sparsity level is sufficiently large. Then in a more realistic regime where MHT is approximately estimated by HTP, we show that such a near non-expansion property extends to the stationary output of HTP with proper learning rate. We have substantialized our theoretical results to the MSE analysis of HTP and its novel extension with preconditioning method. Particularly, our MSE bounds for HTP and its preconditioning extension nearly match the known lower bounds for polynomial-time sparse least squares regression estimators. More importantly, the MSE bounds imply tighter parameter estimation error bounds of HTP than the existing ones. Preliminary numerical results support our theoretical findings and demonstrate the computational advantage of the proposed preconditioning HTP method over the plane HTP.

Acknowledgments

The authors sincerely thank the anonymous reviewers for their constructive comments on this work. Xiao-Tong Yuan would also like to acknowledge the partial support from National Major Project of China for New Generation of AI under Grant No.2018AAA0100400 and Natural Science Foundation of China (NSFC) under Grant No.61876090.

References

Felix Abramovich and Vadim Grinshtein. High-dimensional classification by sparse logistic regression. *IEEE Transactions on Information Theory*, 65(5):3068–3079, 2019.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

Sohail Bahmani, Bhiksha Raj, and Petros T Boufounos. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 14:807–841, 2013.

Thomas Blumensath. Compressed sensing with nonlinear observations and related nonlinear optimization problems. *IEEE Transactions on Information Theory*, 59(6):3466–3474, 2013.

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

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Le-Yu Chen and Sokbae Lee. Best subset binary prediction. *Journal of Econometrics*, 206(1): 39–56, 2018.

Wei Dai and Olgica Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, 2009.

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

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

Simon Foucart and Holger Rauhut. A mathematical introduction to compressive sensing. *Bull. Am. Math*, 54:151–165, 2017.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Daniel Hsu, Sham Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

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.

Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Jarvis Haupt. Stochastic variance reduced optimization for nonconvex sparse learning. In *International Conference on Machine Learning*, pages 917–925, 2016.

Bo Liu, Xoa-Tong Yuan, Lezi Wang, Qingshan Liu, Junzhou Huang, and Dimitris N. Metaxas. Distributed inexact newton-type pursuit for non-convex sparse learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Zongming Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 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.

Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Annual Asilomar Conference on Signals, Systems, and Computers*, pages 40–44, 1993.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10): 6976–6994, 2011.

Philippe Rigollet. 18. s997: High dimensional statistics. *Lecture Notes, Cambridge, MA, USA: MIT Open-CourseWare*, 2015.

Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International Conference on Machine Learning*, 2014.

Jie Shen and Ping Li. On the iteration complexity of support recovery via hard thresholding pursuit. In *International Conference on Machine Learning*, pages 3115–3124, 2017.

Jie Shen and Ping Li. A tight bound of hard thresholding. *Journal of Machine Learning Research*, 18(208):1–42, 2018.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Jian Wang and Ping Li. Recovery of sparse signals using multiple orthogonal least squares. *IEEE Transactions on Signal Processing*, 65(8):2049–2062, 2017.

Xiao-Tong Yuan and Ping Li. Generalization bounds for high-dimensional m-estimation under sparsity constraint. *arXiv preprint arXiv:2001.07212*, 2020.

Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14:899–925, 2013.

Xiao-Tong Yuan, Ping Li, and Tong Zhang. Exact recovery of hard thresholding pursuit. In *Advances in Neural Information Processing Systems*, pages 3558–3566, 2016.

Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18(166):1–43, 2018.

Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948, 2014.

Pan Zhou, Xiao-Tong Yuan, and Jiashi Feng. Efficient stochastic gradient hard thresholding. In *Advances in Neural Information Processing Systems*, pages 1988–1997, 2018.

Appendix A. Preliminaries and Technical Lemmas

Notation. In the following, u is a vector, A is a matrix, and $S \subseteq \{1, ..., q\}$ is an index set. The following notations will be used in this article.

- $[u]_i$: the ith entry of vector u.
- u_S: the restriction of u on S, i.e., $[u_S]_i = [u]_i$ if $i \in S$, and $[u_S]_i = 0$ otherwise.
- $\|u\| = \sqrt{u^\top u}$: the Euclidean norm of u.
- $\|u\|_\infty = \max_i |[u]_i|$: the ℓ_∞-norm of u.
- $\|u\|_0$: the number of nonzero entries of u.
- $\|u\|_A = \sqrt{u^\top A u}$: the Mahalanobis distance of u with respect to $A \succeq 0$.
- $\mathrm{supp}(u)$: the index set of nonzero entries of u.
- $\mathrm{supp}(u, k)$: the index set of the top k (in modulus) entries of u.
- $[u]_{\min} = \min_{i \in \mathrm{supp}(u)} |[u]_i|$: the smallest absolute value of nonzero element of u.
- $[A]_{ij}$: the element on the ith row and jth column of matrix A.
- $\|A\| = \sup_{\|x\|\leq 1} \|Ax\|$: the spectral norm of matrix A.
- $A_{S,S'}$: the restriction of A on row index set S and column index set S'.
- $A_{S,:}$ ($A_{:,S}$): the restriction of A on row (column) index set S.
- $\lambda_{\max}(A, k) = \max_{\|x\|=1, \|x\|_0 \leq k} x^\top A x$: the largest k-sparse eigenvalue of a positive semi-definite matrix A. Particularly, we denote $\lambda_{\max}(A)$ the largest eigenvalue of A.
- $\lambda_{\min}(A, k) = \min_{\|x\|=1, \|x\|_0 \leq k} x^\top A x$: the smallest k-sparse eigenvalue of a positive semi-definite matrix A. Particularly, we denote $\lambda_{\min}(A)$ the smallest eigenvalue of A.
- $\kappa(A, k) = \lambda_{\max}(A, k)/\lambda_{\min}(A, k)$: the k-sparse condition number of A.

In our analysis, we will use the following defined concepts of restricted strong convexity and smoothness which are conventionally used in the analysis of sparsity recovery methods (Shalev-Shwartz et al., 2010; Bahmani et al., 2013; Jain et al., 2014).

Definition 15 (Restricted Strong Convexity/Smoothness) *For any integer $s > 0$, we say f is restricted μ_s-strongly convex and L_s-smooth if there exist $\mu_s, L_s > 0$ such that*

$$\frac{\mu_s}{2}\|u - v\|^2 \leq f(u) - f(v) - \langle \nabla f(v), u - v\rangle \leq \frac{L_s}{2}\|u - v\|^2, \quad \forall \|u - v\|_0 \leq s. \tag{8}$$

The ratio number $\kappa_s := L_s/\mu_s$ that measures the curvature of the loss function over sparse subspaces is referred to as *restricted strong condition number*. Specially for quadratic objective function with Hessian A, we have $L_s = \lambda_{\max}(A, s)$, $\mu_s = \lambda_{\min}(A, s)$ and $\kappa_s = \lambda_{\max}(A, s)/\lambda_{\min}(A, s)$.

The following simple lemma is useful in our analysis.

Lemma 16 *Consider a real value function $g(x) = bx/(ax^2 + c)$ where $a, b, c > 0$. Then $g(x) \leq b/(2\sqrt{ac})$ for all $x \in \mathbb{R}$.*

Proof Obviously, the inequality holds when $x < 0$. Next we only consider $x \geq 0$. The non-negative stationary point of g is $x^* = \sqrt{\frac{c}{a}}$. It can be directly computed that $g''(x^*) = -8bc^2\sqrt{ac} < 0$, which indicates that x^* is the maximizer of g. Then we must have $g(x) \leq g(x^*) = b/(2\sqrt{ac})$ for all $x \in \mathbb{R}$. ■

We also need the following technical lemma for spectral analysis.

Lemma 17 *Let A and B be two symmetric and positive definite matrices and $B \succeq \mu I$ for some $\mu > 0$. If $\|A - B\| \leq \gamma$, then $(A + \gamma I)^{-1}B$ is diagonalizable and*

$$\lambda_{\max}((A + \gamma I)^{-1}B) \leq 1, \quad \lambda_{\min}((A + \gamma I)^{-1}B) \geq \frac{\mu}{\mu + 2\gamma}.$$

Moreover, the following spectral norm bound holds:

$$\left\| I - (A + \gamma I)^{-1/2}B(A + \gamma I)^{-1/2} \right\| \leq \frac{2\gamma}{\mu + 2\gamma}.$$

Proof Since both $A + \gamma I$ and B are symmetric and positive definite, it is known that the eigenvalues of $(A + \gamma I)^{-1}B$ are positive real numbers and identical to those of $(A + \gamma I)^{-1/2}B(A + \gamma I)^{-1/2}$. Let us consider the following eigenvalue decomposition of $(A + \gamma I)^{-1/2}B(A + \gamma I)^{-1/2}$:

$$(A + \gamma I)^{-1/2}B(A + \gamma I)^{-1/2} = Q^\top \Lambda Q,$$

where $Q^\top Q = I$ and Λ is a diagonal matrix with eigenvalues as diagonal entries. The above decomposition then implies

$$(A + \gamma I)^{-1}B = (A + \gamma I)^{-1/2}Q^\top \Lambda Q(A + \gamma I)^{1/2},$$

which is a diagonal eigenvalue decomposition of $(A + \gamma I)^{-1}B$. Thus $(A + \gamma I)^{-1}B$ is diagonalizable.

To prove the eigenvalue bounds of $(A + \gamma I)^{-1}B$, it suffices to prove the same bounds for $(A + \gamma I)^{-1/2}B(A + \gamma I)^{-1/2}$. Since $\|A - B\| \leq \gamma$, we have $B \preceq A + \gamma I$ which implies $(A + \gamma I)^{-1/2}B(A + \gamma I)^{-1/2} \preceq I$ and hence $\lambda_{\max}((A + \gamma I)^{-1/2}B(A + \gamma I)^{-1/2}) \leq 1$. Moreover, since $B \succeq \mu I$, it holds that $\frac{2\gamma}{\mu}B - \gamma I \succeq \gamma I \succeq A - B$. Then we obtain $(A + \gamma I)^{-1/2}B(A + \gamma I)^{-1/2} \succeq \frac{\mu}{\mu + 2\gamma}I$ which implies $\lambda_{\min}((A + \gamma I)^{-1/2}B(A + \gamma I)^{-1/2}) \geq \frac{\mu}{\mu + 2\gamma}$. Therefore we obtain that $\left\| I - (A + \gamma I)^{-1/2}B(A + \gamma I)^{-1/2} \right\| \leq 1 - \frac{\mu}{\mu + 2\gamma} = \frac{2\gamma}{\mu + 2\gamma}$. ■

The following standard result will also be used in our analysis.

Lemma 18 *If A and B are $p \times p$ symmetric matrices such that $\|A - B\| \leq \gamma$, then for any positive integer $s \leq p$*

$$\lambda_{\min}(B, s) - \gamma \leq \lambda_{\min}(A, s) \leq \lambda_{\max}(A, s) \leq \lambda_{\max}(B, s) + \gamma.$$

Proof Let $\tilde{x}$ be a largest s-sparse eigenvector of A. Since $\|A - B\| \leq \gamma$, we must have $A \preceq B + \gamma I$. Therefore,

$$\lambda_{\max}(A, s) = \tilde{x}A\tilde{x} \leq \tilde{x}(B + \gamma I)\tilde{x} \leq \lambda_{\max}(B, s) + \gamma,$$

which shows the right hand side of the inequality. The other side of the bound can be proved similarly. ■

Appendix B. Proofs of Main Results in Section 3

B.1. Proof of Lemma 1

Proof Let us define $g = A(u-w)$. Let $S_g = \text{supp}(g)$ and $\bar{S} = \text{supp}(\bar{w})$. By definition we have $|S_u| = k$ and $|\bar{S}| = \bar{k}$. It follows immediately from the condition $A_{S_u,:}(u-w) = 0$ that $S_u \cap S_g = \emptyset$, and thus $\langle u, g\rangle = 0$. We start by showing the following equality:

$$\|u-\bar{w}\|_A^2 - \|w-\bar{w}\|_A^2 = \|u\|_A^2 - \|w\|_A^2 + 2\langle \bar{w}, A(w-u)\rangle = -2\langle \bar{w}, g\rangle - \|w-u\|_A^2,$$

where in the last "=" we have used $u^\top g = 0$ which then implies $\|w\|_A^2 = \|w-u+u\|_A^2 = \|w-u\|_A^2 + \|u\|_A^2$. Denote $S = S_u \cup \bar{S}$, $\bar{S}_u = S_u \cap \bar{S}$ and $\bar{S}_g = S_g \cap \bar{S}$. Then the previous equality leads to the following bound:

$$\|u-\bar{w}\|_A^2 - \|w-\bar{w}\|_A^2 = -2\langle \bar{w}, g\rangle - \|w-u\|_A^2 = -2\langle \bar{w}_{\bar{S}_g}, g_{\bar{S}_g}\rangle - \|w-u\|_A^2 \le 2\|\bar{w}_{\bar{S}_g}\|\|g_{\bar{S}_g}\|.$$

Next we bound $\|w-\bar{w}\|_A^2$ from below. In order to avoid heavy notation, we abbreviate $\mu_{2k} = \lambda_{\min}(A, 2k)$. Then we can derive that

$$\begin{aligned}
\|w-\bar{w}\|_A^2 =& \|w-u+u-\bar{w}\|_A^2 = \|w-u\|_A^2 + \|u-\bar{w}\|_A^2 - 2\langle g, u-\bar{w}\rangle \\
\overset{\zeta_1}{=}& \|w-u\|_A^2 + \|u-\bar{w}\|_A^2 + 2\langle g, \bar{w}\rangle \\
\ge& \|u-\bar{w}\|_A^2 - 2\|g_{\bar{S}_g}\|\|\bar{w}_{\bar{S}_g}\| \ge \mu_{2k}\|u-\bar{w}\|^2 - 2\|g_{\bar{S}_g}\|\|\bar{w}_{\bar{S}_g}\| \\
\overset{\zeta_2}{\ge}& \mu_{2k}\|u_{S_u\setminus\bar{S}_u}\|^2 + \mu_{2k}\|\bar{w}_{\bar{S}_g}\|^2 - 2\|g_{\bar{S}_g}\|\|\bar{w}_{\bar{S}_g}\|,
\end{aligned}$$

where "ζ_1" is due to the fact $\langle g, u\rangle = 0$ and in "ζ_2" we have used $\bar{w}_{S_u\setminus\bar{S}_u} = 0$ and $u_{\bar{S}_g} = 0$. Let us now distinguish the two complementary cases of $\|\bar{w}_{\bar{S}_g}\| \le \frac{3\|g_{\bar{S}_g}\|}{\mu_{2k}}$ and $\|\bar{w}_{\bar{S}_g}\| > \frac{3\|g_{\bar{S}_g}\|}{\mu_{2k}}$ respectively in the following analysis.

Case I: Assume that $\|\bar{w}_{\bar{S}_g}\| \le \frac{3\|g_{\bar{S}_g}\|}{\mu_{2k}}$. In this case, we can further show the following bound:

$$\begin{aligned}
\|w-\bar{w}\|_A^2 \ge& \mu_{2k}\|u_{S_u\setminus\bar{S}_u}\|^2 + \mu_{2k}\|\bar{w}_{\bar{S}_g}\|^2 - 2\|g_{\bar{S}_g}\|\|\bar{w}_{\bar{S}_g}\| \\
\overset{\zeta_1}{\ge}& \mu_{2k}\|u_{S_u\setminus\bar{S}_u}\|^2 - \frac{\|g_{\bar{S}_g}\|^2}{\mu_{2k}} \\
\overset{\zeta_2}{\ge}& \mu_{2k}(k-\bar{k})[u]_{\min}^2 - \frac{\bar{k}\|g\|_\infty^2}{\mu_{2k}} \overset{\zeta_3}{\ge} \mu_{2k}(k-\bar{k})\frac{\|g\|_\infty^2}{\nu^2} - \frac{\bar{k}\|g\|_\infty^2}{\mu_{2k}} \\
=& \frac{\mu_{2k}\|g\|_\infty^2}{\nu^2}\left(k-\bar{k}-\frac{\nu^2}{\mu_{2k}^2}\bar{k}\right) > 0,
\end{aligned}$$

where in the inequality "ζ_1" we have used the basic fact $ax^2 - bx \ge -\frac{b^2}{4a}$ for any $a, b > 0$, in "ζ_2" we have used $\sqrt{\|x\|_0}[x]_{\min} \le \|x\| \le \sqrt{\|x\|_0}\|x\|_\infty$, in "$\zeta_3$" we have used the condition of $[u]_{\min} \ge \frac{\|g\|_\infty}{\nu}$, and the last inequality sign "$>$" is due to the condition $k \ge \bar{k} + \frac{3\nu^2}{\mu_{2k}^2}\bar{k}$. Therefore,

$$\begin{aligned}
\frac{\|u-\bar{w}\|_A^2 - \|w-\bar{w}\|_A^2}{\|w-\bar{w}\|_A^2} \le& \frac{2\|\bar{w}_{\bar{S}_g}\|\|g_{\bar{S}_g}\|}{\frac{\mu_{2k}}{\nu^2}\|g\|_\infty^2\left(k-\bar{k}-\frac{\nu^2}{\mu_{2k}^2}\bar{k}\right)} \overset{\zeta_1}{\le} \frac{\frac{6\|g_{\bar{S}_g}\|^2}{\mu_{2k}}}{\frac{\mu_{2k}}{\nu^2}\|g\|_\infty^2\left(k-\bar{k}-\frac{\nu^2}{\mu_{2k}^2}\bar{k}\right)} \\
\le& \frac{6\bar{k}\|g\|_\infty^2}{\frac{\mu_{2k}^2}{\nu^2}\|g\|_\infty^2\left(k-\bar{k}-\frac{\nu^2}{\mu_{2k}^2}\bar{k}\right)} = \frac{\frac{6\nu^2}{\mu_{2k}^2}\bar{k}}{k-\bar{k}-\frac{\nu^2}{\mu_{2k}^2}\bar{k}},
\end{aligned}$$

where in the inequality "ζ_1" we have again used the assumption $\|\bar{w}_{\bar{S}_g}\| \leq \frac{3\|g_{\bar{S}_g}\|}{\mu_{2k}}$. Since $k \geq \bar{k} + \frac{3\nu^2}{\mu_{2k}^2}\bar{k}$, the above inequality immediately implies that

$$\frac{\|u-\bar{w}\|_A^2 - \|w-\bar{w}\|_A^2}{\|w-\bar{w}\|_A^2} \leq \frac{\frac{6\nu^2}{\mu_{2k}^2}\bar{k}}{k-\bar{k}-\frac{\nu^2}{\mu_{2k}^2}\bar{k}} \leq \frac{9\nu^2\bar{k}}{\mu_{2k}^2(k-\bar{k})} \leq \frac{3\nu}{\mu_{2k}}\sqrt{\frac{3\bar{k}}{k-\bar{k}}},$$

where in the last inequality we have used $\frac{3\nu^2\bar{k}}{\mu_{2k}^2(k-\bar{k})} < 1$ which implies $\frac{3\nu^2\bar{k}}{\mu_{2k}^2(k-\bar{k})} \leq \sqrt{\frac{3\nu^2\bar{k}}{\mu_{2k}^2(k-\bar{k})}}$.

Case II: Assume that $\|\bar{w}_{\bar{S}_g}\| > \frac{3\|g_{\bar{S}_g}\|}{\mu_{2k}}$. In this regime we can show that

$$\|w-\bar{w}\|_A^2 \geq \mu_{2k}\|u_{S_u\setminus\bar{S}_u}\|^2 + \mu_{2k}\|\bar{w}_{\bar{S}_g}\|^2 - 2\|g_{\bar{S}_g}\|\|\bar{w}_{\bar{S}_g}\| \geq \mu_{2k}\|u_{S_u\setminus\bar{S}_u}\|^2 + \frac{\mu_{2k}}{3}\|\bar{w}_{\bar{S}_g}\|^2$$

where in the last "$\geq$" we have used the assumption $\|\bar{w}_{\bar{S}_g}\| > \frac{3\|g_{\bar{S}_g}\|}{\mu_{2k}}$ which implies $\|g_{\bar{S}_g}\|\|\bar{w}_{\bar{S}_g}\| \leq \frac{\mu_{2k}}{3}\|\bar{w}_{\bar{S}_g}\|^2$. Therefore, we obtain

$$\begin{aligned}
&\frac{\|u-\bar{w}\|_A^2 - \|w-\bar{w}\|_A^2}{\|w-\bar{w}\|_A^2} \leq \frac{2\|\bar{w}_{\bar{S}_g}\|\|g_{\bar{S}_g}\|}{\mu_{2k}\|u_{S_u\setminus\bar{S}_u}\|^2 + \frac{\mu_{2k}}{3}\|\bar{w}_{\bar{S}_g}\|^2} \\
&\overset{\zeta_1}{\leq} \frac{\sqrt{3}\|g_{\bar{S}_g}\|}{\mu_{2k}\|u_{S_u\setminus\bar{S}_u}\|} \leq \frac{\sqrt{3\bar{k}}\|g\|_\infty}{\mu_{2k}\sqrt{k-\bar{k}}[u]_{\min}} \overset{\zeta_2}{\leq} \frac{\nu\sqrt{3\bar{k}}\|g\|_\infty}{\mu_{2k}\sqrt{k-\bar{k}}\|g\|_\infty} = \frac{\nu}{\mu_{2k}}\sqrt{\frac{3\bar{k}}{k-\bar{k}}} < \frac{3\nu}{\mu_{2k}}\sqrt{\frac{3\bar{k}}{k-\bar{k}}},
\end{aligned}$$

where in the inequality "ζ_1" we have invoked Lemma 16 with $a = \mu_{2k}/3, b = 2\|g_{\bar{S}_g}\|, c = \mu_{2k}\|u_{S_u\setminus\bar{S}_u}\|^2$, and in "$\zeta_2$" we have used the condition of $[u]_{\min} \geq \frac{\|g\|_\infty}{\nu}$. By combining the results in the above two cases we can see that the following bound holds when $k \geq \bar{k} + \frac{3\nu^2}{\mu_{2k}^2}\bar{k}$:

$$\|u-\bar{w}\|_A^2 \leq \left(1 + \frac{3\nu}{\mu_{2k}}\sqrt{\frac{3\bar{k}}{k-\bar{k}}}\right)\|w-\bar{w}\|_A^2.$$

This concludes the proof. ■

B.2. Proof of Theorem 3

The following lemma from Yuan et al. (2018) gives a necessary condition on the k-sparse minimizer of an objective function with restrictive smoothness.

Lemma 19 *If f is L_{2k}-smooth, then for the global k-sparse minimizer $\theta^* = \arg\min_{\|\theta\|_0\leq k} f(\theta)$ we have*

$$[\nabla f(\theta^*)]_{S^*} = 0, \quad [\theta^*]_{\min} \geq \frac{\|\nabla f(\theta^*)\|_\infty}{L_{2k}},$$

where $S^* = supp(\theta^*)$.

Proof [of Theorem 3] Let $S = \mathrm{supp}(\mathcal{H}_{A,k}(w))$. By substantializing Lemma 19 with $f(\theta) = \frac{1}{2}\|\theta - w\|_A^2$ and $L_{2k} = \lambda_{\max}(A, 2k)$ we obtain

$$A_{S,:}(\mathcal{H}_{A,k}(w) - w) = 0, \quad [\mathcal{H}_{A,k}(w)]_{\min} \geq \frac{\|A(\mathcal{H}_{A,k}(w) - w)\|_\infty}{\lambda_{\max}(A, 2k)}.$$

Then invoking Lemma 1 to $u = \mathcal{H}_{A,k}(w)$ with $\nu = \lambda_{\max}(A, 2k)$ yields

$$\|\mathcal{H}_{A,k}(w) - \bar{w}\|_A^2 \leq \left(1 + \frac{3\lambda_{\max}(A, 2k)}{\lambda_{\min}(A, 2k)}\sqrt{\frac{3\bar{k}}{k - \bar{k}}}\right)\|w - \bar{w}\|_A^2.$$

Since we always have $\|\mathcal{H}_{A,k}(w) - \bar{w}\|_A \leq 2\|w - \bar{w}\|_A$ and $1 + \frac{3L_{2k}}{\mu_{2k}}\sqrt{\frac{3\bar{k}}{k-\bar{k}}} > 4$ when $\bar{k} < k < \bar{k} + \frac{3M_{2k}^2}{\mu_{2k}^2}\bar{k}$, the following bound naturally holds for any $k > k$:

$$\|\mathcal{H}_{A,k}(w) - \bar{w}\|_A^2 \leq \min\left\{4, 1 + \frac{3L_{2k}}{\mu_{2k}}\sqrt{\frac{3\bar{k}}{k - \bar{k}}}\right\}\|w - \bar{w}\|_A^2.$$

This completes the proof. ■

B.3. Proof of Theorem 5

The following lemma gives a necessary condition on the k-sparse output of HTP algorithm after sufficient iteration. It is essentially a counterpart of Lemma 19 for the stationary output of HTP.

Lemma 20 *Assume that function f is L_{2k}-smooth and μ_{2k}-strongly convex. Let $\theta^{(t)}$ be the output of HTP when applied to estimating $\min_{\|\theta\|_0 \leq k} f(\theta)$ with learning rate $\eta = \frac{1}{2L_{2k}}$. Let $S^{(t)} = supp(\theta^{(t)})$. Then*

$$[\nabla f(\theta^{(t)})]_{S^{(t)}} = 0, \quad [\theta^{(t)}]_{\min} \geq \frac{\|\nabla f(\theta^{(t)})\|_\infty}{2L_{2k}}$$

after at most

$$t = \left\lfloor \frac{4kL_{2k}}{\mu_{2k}} \log\left(\frac{\Delta^{(0)}}{\Delta^{-*}}\right)\right\rfloor + 1$$

rounds of iteration, where $\Delta^{(0)} = f(\theta^{(0)}) - f(\theta^*)$ and

$$\Delta^{-*} = \min_{\|\theta\|_0 \leq k, supp(\theta) \neq supp(\theta^*), f(\theta) > f(\theta^*)} [f(\theta) - f(\theta^*)].$$

Proof Let $\theta^{(t)}$ be the output of HTP when Algorithm 1 terminates at time instance t with $S^{(t)} = S^{(t-1)}$. The stationary equation $[\nabla f(\theta^{(t)})]_{S^{(t)}} = 0$ follows immediately from the debiasing step **S2** of Algorithm 1. Assume otherwise that $[\theta^{(t)}]_{\min} < \frac{\|\nabla f(\theta^{(t)})\|_\infty}{2L_{2k}}$. Since f is μ_{2k}-strongly convex, the optimality of $\theta^{(t)}$ over $S^{(t)}$ together with $S^{(t)} = S^{(t-1)}$ implies that $\theta^{(t)} = \theta^{(t-1)}$. Then the following holds for $\theta^{(t-1)}$:

$$[\nabla f(\theta^{(t-1)})]_{S^{(t-1)}} = 0, \quad [\theta^{(t-1)}]_{\min} < \frac{\|\nabla f(\theta^{(t-1)})\|_\infty}{2L_{2k}}.$$

Then according to the truncated gradient descent step **S1** of Algorithm 1, $S^{(t-1)}$ must differ from $S^{(t)}$ in at least one element with value $[\theta^{(t-1)}]_{\min}$, which contradicts the assumption. Therefore, it must be true that $[\theta^{(t)}]_{\min} \geq \frac{\|\nabla f(\theta^{(t)})\|_\infty}{2L_{2k}}$ when HTP terminates at time instance t.

Next we bound the total number of iteration steps required to achieve $S^{(t)} = S^{(t-1)}$. To this end, we assume that $S^{(t)} \neq S^{(t-1)}$ for all $t = 0, ...,$ before termination. From the step **S2** we know that $[\nabla f(\theta^{(t-1)})]_{S^{(t-1)}} = 0$. Let $l := |S^{(t)} \setminus S^{(t-1)}| = |S^{(t-1)} \setminus S^{(t)}| \leq k$. By the step **S1** we have that $S^{(t)} \setminus S^{(t-1)}$ contains the top l (in magnitude) entries in $\nabla f(\theta^{(t-1)})$ while $S^{(t-1)} \setminus S^{(t)}$ contains the bottom l entries in $\theta^{(t-1)}$. Since $S^{(t)} \neq S^{(t-1)}$, we have $l \geq 1$. Then from the definition of $\tilde{\theta}^{(t)}$ the following inequality holds:

$$\|\tilde{\theta}^{(t)} - \theta^{(t-1)}\| \geq \eta \|[\nabla f(\theta^{(t-1)})]_{S^{(t)} \setminus S^{(t-1)}}\|. \tag{9}$$

According to the definition of $\theta^{(t)}$, we have $f(\theta^{(t)}) \leq f(\tilde{\theta}^{(t)})$. Since f is L_{2k}-smooth, it follows that

$$\begin{aligned} & f(\theta^{(t)}) - f(\theta^{(t-1)}) \leq f(\tilde{\theta}^{(t)}) - f(\theta^{(t-1)}) \\ \leq & \langle \nabla f(\theta^{(t-1)}), \tilde{\theta}^{(t)} - \theta^{(t-1)} \rangle + \frac{L_{2k}}{2} \|\tilde{\theta}^{(t)} - \theta^{(t-1)}\|^2 \\ \overset{\xi_1}{\leq} & -\frac{1}{2\eta} \|\tilde{\theta}^{(t)} - \theta^{(t-1)}\|^2 + \frac{L_{2k}}{2} \|\tilde{\theta}^{(t)} - \theta^{(t-1)}\|^2 = -\frac{1 - \eta L_{2k}}{2\eta} \|\tilde{\theta}^{(t)} - \theta^{(t-1)}\|^2, \end{aligned} \tag{10}$$

where ξ_1 follows from the fact that $\tilde{\theta}^{(t)}$ is the best k-support approximation to $\theta^{(t-1)} - \eta \nabla f(\theta^{(t-1)})$ such that

$$\|\tilde{\theta}^{(t)} - \theta^{(t-1)} + \eta \nabla f(\theta^{(t-1)})\|^2 \leq \|\theta^{(t-1)} - \theta^{(t-1)} + \eta \nabla f(\theta^{(t-1)})\|^2 = \|\eta \nabla f(\theta^{(t-1)})\|^2,$$

which implies $2\eta \langle \nabla f(\theta^{(t-1)}), \tilde{\theta}^{(t)} - \theta^{(t-1)} \rangle \leq -\|\tilde{\theta}^{(t)} - \theta^{(t-1)}\|^2$. By combining (10) and (9) we get

$$f(\theta^{(t)}) - f(\theta^{(t-1)}) \leq -\frac{(1 - \eta L_{2k})\eta}{2} \|[\nabla f(\theta^{(t-1)})]_{S^{(t)} \setminus S^{(t-1)}}\|^2. \tag{11}$$

Let us now consider $S^* = \text{supp}(\theta^*)$. From the μ_{2k}-strong convexity of f we have

$$\begin{aligned} \frac{\mu_{2k}}{2} \|\theta^* - \theta^{(t-1)}\|^2 & \leq f(\theta^*) - f(\theta^{(t-1)}) - \langle \theta^* - \theta^{(t-1)}, \nabla f(\theta^{(t-1)}) \rangle \\ & \overset{\xi_1}{\leq} f(\theta^*) - f(\theta^{(t-1)}) + \frac{\mu_{2k}}{2} \|\theta^* - \theta^{(t-1)}\|^2 + \frac{1}{2\mu_{2k}} \|[\nabla f(\theta^{(t-1)})]_{F^* \setminus S^{(t-1)}}\|^2, \end{aligned}$$

where ξ_1 follows from Cauchy-Schwartz inequality, $ma^2/2 + b^2/(2m) \geq ab$ for any $m > 0$, and $\nabla_{S^{(t-1)}} f(\theta^{(t-1)}) = 0$. This implies

$$\|[\nabla f(\theta^{(t-1)})]_{F^* \setminus S^{(t-1)}}\|^2 \geq 2\mu_{2k} \left[f(\theta^{(t-1)}) - f(\theta^*) \right].$$

Let $l' = |F^* \setminus S^{(t-1)}|$. Obviously, we have $l' \leq k$. Based on the above arguments, it can be verified that

$$\begin{aligned} k \|[\nabla f(\theta^{(t-1)})]_{S^{(t)} \setminus S^{(t-1)}}\|^2 & \geq (l'/l) \|[\nabla f(\theta^{(t-1)})]_{S^{(t)} \setminus S^{(t-1)}}\|^2 \\ & \geq \|[\nabla f(\theta^{(t-1)})]_{F^* \setminus S^{(t-1)}}\|^2 \geq 2\mu_{2k} \left[f(\theta^{(t-1)}) - f(\theta^*) \right]. \end{aligned}$$

By setting $\eta = \frac{1}{2L_{2k}}$ and using (11) and (12) we get that

$$f(\theta^{(t)}) \leq f(\theta^{(t-1)}) - \frac{\mu_{2k}}{4kL_{2k}}\left[f(\theta^{(t-1)}) - f(\theta^*)\right].$$

Therefore, we get

$$f(\theta^{(t)}) - f(\theta^*) \leq \left(1 - \frac{\mu_{2k}}{4kL_{2k}}\right)(f(\theta^{(t-1)}) - f(\theta^*)).$$

Note that $f(\theta^{(t)}) \geq f(\theta^*)$. By recursively using the above inequality we obtain.

$$f(\theta^{(t)}) - f(\theta^*) \leq \left(1 - \frac{\mu_{2k}}{4kL_{2k}}\right)^t (f(\theta^{(0)}) - f(\theta^*)).$$

Therefore $f(\theta^{(t)}) - f(\theta^*) \leq \Delta^{-*}$ when $t \geq \frac{4kL_{2k}}{\mu_{2k}} \log\left(\frac{\Delta^{(0)}}{\Delta^{-*}}\right)$ (note that $\Delta^{-\star} > 0$). After that, we have $f(\theta^{(t)}) < f(\theta^{-*})$ and thus $f(\theta^{(t)}) = f(\theta^*)$. Then by invoking Lemma 19 we have

$$[\nabla f(\theta^{(t)})]_{S^{(t)}} = 0, \quad [\theta^{(t)}]_{\min} \geq \frac{\|\nabla f(\theta^{(t)})\|_\infty}{2L_{2k}}$$

which in turn leads to $S^{(t+1)} = S^{(t)}$ so that the algorithm terminates at $t+1$. ■

Now we are ready to prove the main result.

Proof [of Theorem 5] Let $S = \mathrm{supp}(\mathcal{H}^{\mathtt{HTP}}_{A,k}(w))$. By substantializing Lemma 20 with $f(\theta) = \frac{1}{2}\|\theta - w\|_A^2$ and $L_{2k} = \lambda_{\max}(A, 2k)$ we can show that the following two conditions hold

$$A_{S,:}(\mathcal{H}^{\mathtt{HTP}}_{A,k}(w) - w) = 0, \quad [\mathcal{H}^{\mathtt{HTP}}_{A,k}(w)]_{\min} \geq \frac{\|A(\mathcal{H}^{\mathtt{HTP}}_{A,k}(w) - w)\|_\infty}{2\lambda_{\max}(A, 2k)}$$

after $\tilde{\mathcal{O}}\left(\frac{k\lambda_{\max}(A,2k)}{\lambda_{\min}(A,2k)}\right)$ rounds of HTP iteration. Then by invoking Lemma 1 to $u = \mathcal{H}^{\mathtt{HTP}}_{A,k}(w)$ with $\nu = 2\lambda_{\max}(A, 2k)$ we have the following holds for any $k \geq \bar{k} + \frac{12\lambda^2_{\max}(A,2k)}{\lambda^2_{\min}(A,2k)}\bar{k}$:

$$\|\mathcal{H}_{A,k}(w) - \bar{w}\|_A^2 \leq \left(1 + \frac{6\lambda_{\max}(A, 2k)}{\lambda_{\min}(A, 2k)}\sqrt{\frac{3\bar{k}}{k - \bar{k}}}\right)\|w - \bar{w}\|_A^2.$$

This concludes the proof. ■

Appendix C. Proofs of Main Results in Section 4

C.1. Proof of Theorem 7

Lemma 21 *Consider the sparse least squares regression model in (2). Let S be an index set such that $|S| \leq s$ and $\bar{w} \subseteq S$. Let $w_S = \arg\min_{supp(w)\subseteq S} F(w)$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$MSE(w_S, \bar{w}; X) \leq \mathcal{O}\left(\frac{\sigma^2 s \log(p/s)}{n} + \frac{\sigma^2 s}{n} + \frac{\sigma^2 \log(1/\delta)}{n}\right).$$

Proof The lemma can be proved by using similar arguments to those for (Rigollet, 2015, Theorem 2.6) and thus we omit the details. ■

Now we can prove the main result.

Proof [of Theorem 7] Let $S^{(t)} = \text{supp}(w^{(t)})$ and $S = \text{supp}(\bar{w}) \cup \text{supp}(\hat{w}^{\ell_0}) \cup S^{(t)}$. Define $w_S = \arg\min_{\text{supp}(w)\subseteq S} F(w)$. Using Taylor expansion we can re-express the quadratic function F as

$$F(w) = F(w_S) + \langle \nabla F(w_S), w - w_S \rangle + \frac{1}{2}(w - w_S)^\top H (w - w_S).$$

Based on Lemma 20 and the optimality of w_S over $S \supseteq S^{(t)}$ we have

$$\nabla_{S^{(t)}} F(w^{(t)}) = 0, \quad [w^{(t)}]_{\min} \geq \frac{\|\nabla F(w^{(t)})\|_\infty}{2\lambda_{\max}(H, 2k)} \geq \frac{\|\nabla_S F(w^{(t)})\|_\infty}{2\lambda_{\max}(H, 2k)} = \frac{\|H_{S,:}(w^{(t)} - w_S)\|_\infty}{2\lambda_{\max}(H, 2k)}.$$

Let us consider the quadratic form $\frac{1}{2}(w - w_S)\tilde{H}(w - w_S)$ in which $\tilde{H} := H_{SS}$ over the supporting set S. Then the above implies that

$$\tilde{H}_{S^{(t)},:}(w^{(t)} - w_S) = 0, \quad [w^{(t)}]_{\min} \geq \frac{\|\tilde{H}(w^{(t)} - w_S)\|_\infty}{2\lambda_{\max}(H, 2k)}.$$

Then by invoking Lemma 1 to $u = w^{(t)}$ with $\nu = 2\lambda_{\max}(H, 2k)$ we can shwo that the following holds for any $k \geq \bar{k} + \frac{12\lambda^2_{\max}(H,2k)}{\lambda^2_{\min}(H,2k)}\bar{k} \geq \bar{k} + \frac{12\lambda^2_{\max}(H,2k)}{\lambda^2_{\min}(\tilde{H},2k)}\bar{k}$:

$$\|w^{(t)} - \bar{w}\|_H^2 = \|w^{(t)} - \bar{w}\|_{\tilde{H}}^2 \leq \left(1 + \frac{6\lambda_{\max}(H, 2k)}{\lambda_{\min}(\tilde{H}, 2k)}\sqrt{\frac{3\bar{k}}{k - \bar{k}}}\right) \|w_S - \bar{w}\|_{\tilde{H}}^2$$

$$\leq \left(1 + \frac{6\lambda_{\max}(H, 2k)}{\lambda_{\min}(H, 2k)}\sqrt{\frac{3\bar{k}}{k - \bar{k}}}\right) \|w_S - \bar{w}\|_H^2$$

$$\leq \mathcal{O}\left(\left(1 + \kappa(H, 2k)\sqrt{\frac{\bar{k}}{k - \bar{k}}}\right)\left(\frac{\sigma^2 k \log(p/k)}{n} + \frac{\sigma^2 k}{n} + \frac{\sigma^2 \log(1/\delta)}{n}\right)\right),$$

where in the last inequality we have used Lemma 21. This then implies the desired bound. ■

C.2. Proof of Theorem 9

The following simple lemma, which is an application of the tail bound from (Hsu et al., 2012), controls the norm of a sub-Gaussian random vector.

Lemma 22 *Let $u = (u_1, ..., u_s)$ be an s-dimensional zero-mean σ^2-sub-Gaussian random vector. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$*

$$\|u\|^2 \leq \sigma^2\left(s + 2\sqrt{s \log\left(\frac{1}{\delta}\right)} + 2\log\left(\frac{1}{\delta}\right)\right).$$

Proof By invoking (Hsu et al., 2012, Theorem 2.1) to u with identity scaling matrix we have that for all $t > 0$,

$$\mathbb{P}\left(\|u\|^2 > \sigma^2(s + 2\sqrt{st} + 2t)\right) \leq \exp(-t).$$

The desired bound then follows readily from setting $t = \log(1/\delta)$ in the above bound. ■

The following lemma is useful to our statistical analysis.

Lemma 23 *For any $\delta \in (0,1)$, the following bound holds with probability at least $1-\delta$ for all index set $S \subseteq \{1,...,p\}$ with $|S| \leq s$:*

$$\|\nabla_S F(\bar{w})\|^2_{H^{-1}_{SS}} \leq \mathcal{O}\left(\frac{\sigma^2 s \log(p/s)}{n} + \frac{\sigma^2 s}{n} + \frac{\sigma^2 \log(1/\delta)}{n}\right).$$

Proof Let S be an index set such that $|S| \leq s$. Recall that $Y = X\bar{w} + \varepsilon$ and $\nabla_S F(\bar{w}) = \frac{1}{n} X^\top_{:,S}\varepsilon$. Then we can write

$$\|\nabla_S F(\bar{w})\|^2_{H^{-1}_{SS}} = (\nabla_S F(\bar{w}))^\top H^{-1}_{SS} \nabla_S F(\bar{w}) = \frac{1}{n^2}\varepsilon^\top X_{:,S} H^{-1}_{SS} X^\top_{:,S}\varepsilon = \frac{1}{n}\|\tilde{\varepsilon}_S\|^2, \tag{12}$$

where $\tilde{\varepsilon}_S = U_S^\top \varepsilon$ and $U_S = \frac{1}{\sqrt{n}} X_{:,S} H^{-1/2}_{SS} \in \mathbb{R}^{n\times s}$ is an orthonormal matrix as $U_S^\top U_S = I_{s\times s}$. Note that for any $\|v\| = 1$, $\|U_S v\| = \|v\| = 1$. Thus, for any $r \in \mathbb{R}$,

$$\mathbb{E}\left[\exp(r\tilde{\varepsilon}_S^\top v)\right] = \mathbb{E}\left[\exp(r\varepsilon_S^\top U_S v)\right] \leq \exp\left(\frac{r^2\sigma^2}{2}\right),$$

which indicates that $\tilde{\varepsilon}_S$ is a s-dimensional zero-mean σ^2-sub-Gaussian vector. Then based on Lemma 22, the following norm bound holds with probability at least $1-\delta$

$$\|\tilde{\varepsilon}_S\|^2 \leq \sigma^2\left(s + 2\sqrt{s\log\left(\frac{1}{\delta}\right)} + 2\log\left(\frac{1}{\delta}\right)\right).$$

It follows from (12) that with probability at least $1-\delta$

$$\|\nabla_S F(\bar{w})\|^2_{H^{-1}_{SS}} \leq \frac{\sigma^2}{n}\left(s + 2\sqrt{s\log\left(\frac{1}{\delta}\right)} + 2\log\left(\frac{1}{\delta}\right)\right).$$

Let $\mathcal{S} = \{S \subseteq \{1,...,p\} : |S| = s\}$ be the set of index set of cardinality s. It is standard to know $|\mathcal{S}| = \binom{p}{s} \leq \left(\frac{ep}{s}\right)^s$. Thus, by simple union probability and preserving leading terms we obtain that

$$\begin{aligned}\|\nabla_S F(\bar{w})\|^2_{H^{-1}_{SS}} \leq& \frac{\sigma^2}{n}\left(s + 2\sqrt{s^2\log\left(\frac{ep}{s}\right) + s\log\left(\frac{1}{\delta}\right)} + 2s\log\left(\frac{ep}{s}\right) + 2\log\left(\frac{1}{\delta}\right)\right)\\ \leq& \mathcal{O}\left(\frac{\sigma^2 s\log(p/s)}{n} + \frac{\sigma^2 s}{n} + \frac{\sigma^2\log(1/\delta)}{n}\right)\end{aligned}$$

holds with probability at least $1-\delta$. ■

The following key lemma characterizes the progress made in step of iteration.

Lemma 24 *Consider an index set S with $|S| \leq s$ and a vector w with supp$(w) \subseteq S$. Let $w' = w - \left(\tilde{H}_{SS} + \gamma I\right)^{-1} \nabla_S F(w)$. Let $\bar{w}$ be a sparse vector such that supp$(\bar{w}) \subseteq S$. If $\|H - \tilde{H}\| \leq \gamma$, then the following inequality holds:*

$$\|w' - \bar{w}\|_{\tilde{H}+\gamma I} \leq \left(1 - \frac{\mu_s}{\mu_s + 2\gamma}\right) \|w - \bar{w}\|_{\tilde{H}+\gamma I} + \|\nabla_S F(\bar{w})\|_{H_{SS}^{-1}}.$$

Proof From the definition of w' we have

$$\begin{aligned} w' - \bar{w} =& w - \bar{w} - \left(\tilde{H}_{SS} + \gamma I\right)^{-1} \nabla_S F(w) \\ =& w - \bar{w} - \left(\tilde{H}_{SS} + \gamma I\right)^{-1} (\nabla_S F(w) - \nabla_S F(\bar{w})) + \left(\tilde{H}_{SS} + \gamma I\right)^{-1} \nabla_S F(\bar{w}) \\ =& \left(I - \left(\tilde{H}_{SS} + \gamma I\right)^{-1} H_{SS}\right)(w - \bar{w}) + \left(\tilde{H}_{SS} + \gamma I\right)^{-1} \nabla_S F(\bar{w}). \end{aligned}$$

By multiplying $(\tilde{H}_{SS} + \gamma I)^{1/2}$ on both sides of the above recurrent form we get

$$\begin{aligned} (\tilde{H}_{SS} + \gamma I)^{1/2}(w' - \bar{w}) = &\left(I - (\tilde{H}_{SS} + \gamma I)^{-1/2} H_{SS} (\tilde{H}_{SS} + \gamma I)^{-1/2}\right)(\tilde{H}_{SS} + \gamma I)^{1/2}(w - \bar{w}) \\ &+ \left(\tilde{H}_{SS} + \gamma I\right)^{-1/2} \nabla_S F(\bar{w}). \end{aligned}$$

It follows readily from the triangle and Cauchy-Schwarz inequalities that

$$\begin{aligned} &\|w' - \bar{w}\|_{\tilde{H}_{SS}+\gamma I} \\ \leq &\left\|I - (\tilde{H}_{SS} + \gamma I)^{-1/2} H_{SS} (\tilde{H}_{SS} + \gamma I)^{-1/2}\right\| \|w - \bar{w}\|_{\tilde{H}_{SS}+\gamma I} + \left\|\left(\tilde{H}_{SS} + \gamma I\right)^{-1/2} \nabla_S F(\bar{w})\right\| \\ \leq &\left\|I - (\tilde{H}_{SS} + \gamma I)^{-1/2} H_{SS} (\tilde{H}_{SS} + \gamma I)^{-1/2}\right\| \|w - \bar{w}\|_{\tilde{H}_{SS}+\gamma I} + \|\nabla_S F(\bar{w})\|_{H_{SS}^{-1}} \\ \overset{\zeta_1}{\leq} &\left(1 - \frac{\mu_s}{\mu_s + 2\gamma}\right) \|w - \bar{w}\|_{\tilde{H}_{SS}+\gamma I} + \|\nabla_S F(\bar{w})\|_{H_{SS}^{-1}}, \end{aligned}$$

where in the inequality "ζ_1" we have used Lemma 17 in view of $\|\tilde{H}_{SS} - H_{SS}\| \leq \|\tilde{H} - H\| \leq \gamma$. Since $w, w', \bar{w}$ are all vectors with supporting set inside S, we must have $\|w' - \bar{w}\|_{\tilde{H}_{SS}+\gamma I} = \|w' - \bar{w}\|_{\tilde{H}+\gamma I}$ and $\|w - \bar{w}\|_{\tilde{H}_{SS}+\gamma I} = \|w - \bar{w}\|_{\tilde{H}+\gamma I}$. Then based on the above inequality we further obtain that

$$\|w' - \bar{w}\|_{\tilde{H}+\gamma I} \leq \left(1 - \frac{\mu_s}{\mu_s + 2\gamma}\right) \|w - \bar{w}\|_{\tilde{H}+\gamma I} + \|\nabla_S F(\bar{w})\|_{H_{SS}^{-1}}.$$

This proves the desired bound. ■

Now we are in the position to prove the main result.

Proof [of Theorem 9] Let $S^{(t)} = \text{supp}(w^{(t)})$ and $\bar{S} = \text{supp}(\bar{w})$. Consider $S = S^{(t-1)} \cup S^{(t)} \cup \bar{S}$. Let $s = 2k + \bar{k}$, $L_s = \lambda_{\max}(H, s)$ and $\mu_s = \lambda_{\min}(H, s)$. Let us define $\hat{w}^{(t)} = w^{(t-1)} - \left(\tilde{H}_{SS} + \gamma I\right)^{-1} \nabla_S F(w^{(t-1)})$. By invoking Lemma 24 we get

$$\|\hat{w}^{(t)} - \bar{w}\|_{\tilde{H}+\gamma I} \leq \left(1 - \frac{\mu_s}{\mu_s + 2\gamma}\right) \|w^{(t-1)} - \bar{w}\|_{\tilde{H}+\gamma I} + \|\nabla_S F(\bar{w})\|_{H_{SS}^{-1}}. \tag{13}$$

Since $\hat{w}^{(t)}$ is the minimizer of the quadratic function $P^{(t-1)}(w)$ restricted on the supporting set S, it is easy to verify that for any w with $\text{supp}(w) \subseteq S$,

$$\begin{aligned} P^{(t-1)}(w) =& \frac{1}{2}(w - \hat{w}^{(t)})^\top \left(\tilde{H}_{SS} + \gamma I_{SS}\right)(w - \hat{w}^{(t)}) + constant \\ =& \frac{1}{2}(w - \hat{w}^{(t)})^\top \left(\tilde{H} + \gamma I\right)(w - \hat{w}^{(t)}) + constant, \end{aligned}$$

where the term $constant$ is not relying on $w^{(t-1)}$. Then, the definition of $w^{(t)}$ implies that

$$w^{(t)} = \mathop{\arg\min}_{\|w\|_0 \le k} P^{(t-1)}(w) = \mathop{\arg\min}_{\|w\|_0 \le k, \text{supp}(w) \subseteq S} \frac{1}{2}\|w - \hat{w}^{(t)}\|^2_{\tilde{H}+\gamma I}.$$

Applying Theorem 3 to the above quadratic form over S we get

$$\begin{aligned} \|w^{(t)} - \bar{w}\|_{\tilde{H}+\gamma I} \le& \min\left\{2, \sqrt{1 + \frac{3(\lambda_{\max}(\tilde{H}, s) + \gamma)}{\lambda_{\min}(\tilde{H}, s) + \gamma}\sqrt{\frac{3\bar{k}}{k - \bar{k}}}}\right\}\|\hat{w}^{(t)} - \bar{w}\|_{\tilde{H}+\gamma I} \\ \overset{\zeta_1}{\le}& \min\left\{2, \sqrt{1 + \frac{3(L_s + 2\gamma)}{\mu_s}\sqrt{\frac{3\bar{k}}{k - \bar{k}}}}\right\}\|\hat{w}^{(t)} - \bar{w}\|_{\tilde{H}+\gamma I} \\ \le& \rho\left(1 - \frac{\mu_s}{\mu_s + 2\gamma}\right)\|w^{(t-1)} - \bar{w}\|_{\tilde{H}+\gamma I} + \rho\|\nabla_S F(\bar{w})\|_{H^{-1}_{SS}}, \end{aligned}$$

where $\rho = \sqrt{1 + \frac{3(L_s+2\gamma)}{\mu_s}\sqrt{\frac{3\bar{k}}{k-\bar{k}}}}$, in "$\zeta_1$" we have used Lemma 18 and in the last inequality we have used (13). By choosing $k \ge \left(1 + \frac{108(L_s+2\gamma)^2(\mu_s+2\gamma)^2}{\mu_s^4}\right)\bar{k}$, we have $\rho \le \sqrt{1 + \frac{\mu_s}{2(\mu_s+2\gamma)}} \le 1 + \frac{\mu_s}{2(\mu_s+2\gamma)} \le 1.5$. Then it follows from the previous inequality and Lemma 23 that the following holds with probability at least $1 - \delta$:

$$\|w^{(t)} - \bar{w}\|_{\tilde{H}+\gamma I} \le \left(1 - \frac{\mu_s}{2(\mu_s + 2\gamma)}\right)\|w^{(t-1)} - \bar{w}\|_{\tilde{H}+\gamma I} + \mathcal{O}\left(\Delta(n, s, p, \delta)\right),$$

where the quantity

$$\Delta(n, s, p, \delta) := \sigma\sqrt{\frac{s\log(p/s)}{n}} + \sigma\sqrt{\frac{s}{n}} + \sigma\sqrt{\frac{\log(1/\delta)}{n}}$$

abbreviates the error term. The above inequality then leads to that with probability at least $1 - \delta$

$$\|w^{(t)} - \bar{w}\|_{\tilde{H}+\gamma I} \le \left(1 - \frac{\mu_s}{2(\mu_s + 2\gamma)}\right)^t \|w^{(0)} - \bar{w}\|_{\tilde{H}+\gamma I} + \mathcal{O}\left(\frac{(\mu_s + \gamma)}{\mu_s}\Delta(n, s, p, \delta)\right).$$

Based on fact $(1 - x)^t \le \exp\{-xt\}$ and $H \preceq \tilde{H} + \gamma I$ we can show that

$$\|w^{(t)} - \bar{w}\|_H \le \|w^{(t)} - \bar{w}\|_{\tilde{H}+\gamma I} \le \mathcal{O}\left(\frac{\mu_s + \gamma}{\mu_s}\Delta(n, s, p, \delta)\right)$$

when

$$t \geq \mathcal{O}\left(\frac{(\mu_s+\gamma)}{\mu_s}\log\left(\frac{\mu_s\|w^{(0)}-\bar{w}\|_{\tilde{H}+\gamma I}}{(\mu_s+\gamma)\Delta(n,s,p,\delta)}\right)\right).$$

The above then implies the desired result. ■

C.3. Proof of Corollary 11

The following lemma, which is based on a matrix concentration bound in Tropp (2012), shows that the Hessian of the quadratic function $\tilde{F}$ is close to that of F when the subset size m is sufficiently large. A similar result appears in Shamir et al. (2014).

Lemma 25 *Assume that $\|x_i\| \leq L$ holds for all $i \in [n]$. Then with probability at least $1-\delta$ over the m data points drawn to construct $\tilde{F}$, the following bound holds:*

$$\|\tilde{H}-H\| \leq L\sqrt{\frac{32\log(p/\delta)}{m}}.$$

Based on this lemma, we can prove the corollary.

Proof [of Corollary 11] Since $\|x_i\| \leq 1$, we know from Lemma 25 that $\|\tilde{H}-H\| \leq \gamma = \sqrt{\frac{32\log(2p/\delta)}{m}}$ holds with probability at least $1-\delta/2$. Provided that $m = \mathcal{O}\left(\lambda_{\min}^{-2}(H,s)\log(p/\delta)\right)$, we have $\gamma = \mathcal{O}(\lambda_{\min}(H,s))$. Then Theorem 9 shows that with probability at least $1-\delta/2$, Algorithm 2 will output $w^{(t)}$ satisfying

$$\mathtt{MSE}(w^{(t)},\bar{w};X) \leq \mathcal{O}\left(\frac{\sigma^2 s\log(p/s)}{n}+\frac{\sigma^2 s}{n}+\frac{\sigma^2\log(1/\delta)}{n}\right)$$

after $t \geq \tilde{\mathcal{O}}(1)$ rounds of iteration. The desired result then follows readily by union probability. ■

C.4. Proof of Corollary 13

Proof From Corollary 11 we know that with probability $1-\delta$, the outer-loop iteration complexity of PC-HTP is of the order $\tilde{\mathcal{O}}(1)$. For each outer-loop iteration, two gradient vectors of F and $\tilde{F}$ are computed via sparse matrix-vector product with complexity $\mathcal{O}(nk)$ to construct the quadratic form in (7), which is then solved via HTP. From Theorem 5 we know that HTP needs $\tilde{\mathcal{O}}\left(k\kappa(\tilde{H}+\gamma I,2k)\right) \leq \tilde{\mathcal{O}}\left(k\kappa(H,2k)\right)$ rounds of iteration to converge. In each iteration, the computational complexity is dominated by local batch gradient computation (see step **S1** in Algorithm 1) and solving a linear system (see step **S**2 in Algorithm 1), which are respectively of the order $\mathcal{O}(mk) = \mathcal{O}\left(k\lambda_{\min}^{-2}(H,s)\log(p/\delta)\right)$ and $\mathcal{O}\left(k^2\sqrt{\kappa(H,2k)}\right)$. Combing the inner-loop and outer-loop complexity bounds yields the following overall computational complexity bound

$$\tilde{\mathcal{O}}\left(nk+k\kappa(H,2k)\left(mk+k^2\sqrt{\kappa(H,2k)}\right)\right),$$

which holds with probability at least $1-\delta$. ■