Stability and Risk Bounds of Iterative Hard Thresholding

Xiao-Tong Yuan and Ping Li
Cognitive Computing Lab
Baidu Research
No. 10 Xibeiwang East Road, Beijing 100193, China
10900 NE 8th St. Bellevue, Washington 98004, USA
{xtyuan1980, pingli98}@gmail.com

Abstract

The Iterative Hard Thresholding (IHT) algorithm is one of the most popular and promising greedy pursuit methods for high-dimensional statistical estimation under cardinality constraint. The existing analysis of IHT mostly focuses on parameter estimation and sparsity recovery consistency. From the perspective of statistical learning theory, another fundamental question is how well the IHT estimation would perform on unseen samples. The answer to this question is important for understanding the generalization ability of IHT yet has remaind elusive. In this paper, we investigate this problem and develop a novel generalization theory for IHT from the viewpoint of algorithmic stability. Our theory reveals that: 1) under natural conditions on the empirical risk function over n samples of dimension p, IHT with sparsity level k enjoys an $\tilde{\mathcal{O}}(n^{-1/2}\sqrt{k\log(n)\log(p)})$ rate of convergence in sparse excess risk; and 2) a fast rate of order $\tilde{\mathcal{O}}(n^{-1}k(\log^3(n)+\log(p)))$ can be derived for strongly convex risk function under certain strong-signal conditions. The results have been substantialized to sparse linear regression and logistic regression models along with numerical evidence provided to support our theory.

1 Introduction

We consider in this paper the following problem of high-dimensional stochastic risk minimization under hard sparsity constraint:

$$\min_{w\in\mathcal{W}} F(w) := \mathbb{E}_{\xi\sim D}[\ell(w;\xi)] \quad \text{s.t. } \|w\|_0 \le k,$$

where $w \in \mathcal{W} \subseteq \mathbb{R}^p$ is the model parameter vector, $\ell(w;\xi)$ is a non-negative convex function that measures the loss of w at a data instance $\xi \in \mathcal{X}$, D represents a random distribution over $\mathcal{X}$. The cardinality constraint $\|w\|_0 \le k$ is imposed for enhancing learnability and interpretability of model. In realistic problems, the mathematical formulation of D is typically unknown and thus it is hopeless to directly optimize such a stochastic formulation. Alternatively, given a set of i.i.d. training samples $S = \{\xi_i\}_{i=1}^n \in \mathcal{X}^n$ drawn from D, the following sparsity-constrained empirical risk minimization problem is often considered for learning sparse models in high-dimensional settings (Donoho, 2006; Bach et al., 2012; Hastie et al., 2015):

$$\min_{w\in\mathcal{W}} F_S(w) := \frac{1}{n}\sum_{i=1}^{n}\ell(w;\xi_i) \quad \text{s.t. } \|w\|_0 \le k. \tag{1}$$

Here the cardinality constraint is crucial for accurate estimation especially when $p \gg n$ which is usually the case in big data era. The above sparse M-estimation model will be referred to as ℓ_0-ERM in this work.

Due to the presence of cardinality constraint, the ℓ_0-ERM estimator is simultaneously non-convex and NP-hard even when the loss function is quadratic (Natarajan, 1995), which makes it computationally intractable to solve the problem exactly in general cases. Therefore, one must seek approximate solutions instead of carrying out combinatorial search over all possible models. Among others, the Iterative Hard Thresholding (IHT) (Blumensath and Davies, 2009) is a family of first-order greedy selection methods popularly used and studied for solving ℓ_0-ERM with outstanding practical efficiency and scalability witnessed in many applications (Jain et al., 2014; Yuan et al., 2018; Zhou et al., 2018). The common theme of IHT-style algorithms is

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

to iterate between gradient descent and hard thresholding to decrease the objective value while maintaining sparsity of solution. In the considered problem setting, a plain IHT algorithm is given by the following recursion for all $t \geq 1$ with learning rate $\eta > 0$:

$$w_{S,k}^{(t)} := \mathrm{H}_k \left(w_{S,k}^{(t-1)} - \eta \nabla F_S(w_{S,k}^{(t-1)}) \right), \qquad (2)$$

where $\mathrm{H}_k(\cdot)$ is the truncation operator that preserves the top k (in magnitude) entries of input and sets the remaining to be zero, with ties broken arbitrarily. The procedure is typically initialized with all-zero vector. The IHT-style algorithms have been shown to converge linearly towards certain nominal sparse model with optimal estimation accuracy under proper regularity conditions (Bahmani et al., 2013; Yuan et al., 2014; Yuan and Li, 2020).

1.1 Problem and motivation

The main motivation of our work is to answer the following fundamental question about the generalization performance of the IHT estimator $w_{S,k}^{(t)}$:

How well the corresponding population risk $F(w_{S,k}^{(t)}) = \mathbb{E}_{\xi \sim D}[\ell(w_{S,k}^{(t)}; \xi)]$ *can approximate the optimal population risk* $F(\bar{w}) = \min_{\|w\|_0 \leq \bar{k}} F(w)$ *for* $\bar{k} \leq k$*?*

The value $F(w_{S,k}^{(t)}) - F(\bar{w})$ is referred to as the $\bar{k}$-*sparse excess risk* of IHT. Our primary goal is to answer this question by deriving a suitable law of large numbers, i.e., a sample size vanishing rate γ_n such that $F(w_{S,k}^{(t)}) - F(\bar{w}) \leq \gamma_n$ holds with high probability. Let $w^* = \arg\min_{w \in \mathcal{W}} F(w)$ be the conventional dense minimizer. The sparse excess risk bound immediately gives arise to an oracle inequality in terms of w^*:

$$F(w_{S,k}^{(t)}) - F(w^*) \leq \min_{\|w\|_0 \leq \bar{k}} \{F(w) - F(w^*)\} + \gamma_n.$$

Therefore, the sparse excess risk bound is also crucial for understanding the excess risk of IHT when misspecified model sparsity is allowed. The classical uniform convergence analysis has been used for bounding excess risk (Bartlett et al., 2006; Shalev-Shwartz et al., 2009). Although showing to be general (e.g., applicable to non-convex losses) and tight in some restricted settings (Kakade et al., 2009), uniform risk bounds tend to suffer from the polynomial dependence on data dimensionality and thus are less satisfactory for high-dimensional learning algorithms. Ideally in well-specified setting where the underlying statistical model for generating the data samples is truly sparse, i.e., $\nabla F(\bar{w}) = 0$ for some $\bar{k}$-sparse vector $\bar{w}$, classical sparse parameter estimation error bounds (Jain et al., 2014; Yuan et al., 2018) can be shown to imply strong sparse excess risk bounds for IHT (see Section 3.2 for more detailed discussions). However, when applied to misspecified sparse models, the existing parameter estimation error bounds will lead to sub-optimal risk bounds having undesirable dependence on the residual term $\|\nabla F(\bar{w})\|_\infty \neq 0$.

Alternatively, a useful and popular proxy for analyzing the generalization performance is the *stability* of learning algorithms to changes in the training dataset (Bousquet and Elisseeff, 2002). By hinging the optimality of ERM, stability has been extensively demonstrated to beget strong generalization bounds for ERM solutions with convex loss functions (Mukherjee et al., 2006; Shalev-Shwartz et al., 2009) and for iterative learning algorithms (such as SGD) as well (Hardt et al., 2016; Charles and Papailiopoulos, 2018; Kuzborskij and Lampert, 2018). Specially, the state-of-the-art generalization results for strongly convex ERM are offered by approaches based on the notion of uniform stability (Feldman and Vondrak, 2018, 2019; Bousquet et al., 2020). Inspired by the remarkable success of stability theory, we aim at deriving sparse excess risk bounds for IHT in view of the uniform stability arguments, which to our knowledge has not been systematically treated elsewhere in literature.

Yet, the traditional uniform stability arguments of regularized convex ERM do not naturally extend to IHT. The crux here is that the stability of IHT relies heavily on the stability of its recovered supporting set, $\mathrm{supp}(w_{S,k}^{(t)})$, which could be highly non-trivial to guarantee even that the risk function is strongly convex. In contrast, the convectional dense ERM is supported over the entire range of feature dimension and thus its supporting set is by nature unique and stable.

1.2 Our work and main results

The idea of our solution to address the above mentioned stability issue about the sparsity pattern of IHT is intuitive in principle: *If the empirical risk* F_S *has restricted strong convexity and smoothness, then based on the uniform stability of ERM restricted over any feature index set of cardinality* k*, we can establish a high probability generalization bound for IHT via applying union probability arguments to all the possible* k*-sparse supporting sets.* A main technical obstacle we need to overcome for this strategy is that in many statistical learning problems the restricted strong convexity of the empirical risk usually holds with high probability over data sample rather than uniformly. As a new element of our analysis for dealing with such a small failure probability of strong convexity, we propose to analyze IHT when applied to a regularized variant of ℓ_0-ERM with a penalty term

Result	Risk Bound	Model Sparsity	Key Condition	
			Empirical risk	Population risk
Theorem 1	$\tilde{\mathcal{O}}\left(\sqrt{\frac{k\log(n)\log(p/k)}{n}}\right)$	Misspecified	RSC/RLS	—
Theorem 2	$\tilde{\mathcal{O}}\left(\frac{k(\log^3(n)+\log(p))}{n}\right)$	Misspecified	RSC/RLS	RLS
Theorem 3	$\tilde{\mathcal{O}}\left(\frac{k\log(p)}{n}\right)$	Well-specified	RSC/RLS	—

Table 1: Overview of our main results on the sparse excess risk bounds of IHT. The big $\tilde{\mathcal{O}}$ notation hides the logarithmic factors on tail bound. RSC and RLS respectively stand for Restricted Strongly Convexity and Restricted Lipschitz Smoothness (see Definition 2).

$\mathcal{O}(n^{-1/2}\|w\|^2)$ added to guarantee restricted uniform stability, and consequently show that the stability-induced risk bound of the regularized IHT estimator can be inherited by the original IHT with high chance. The corresponding main result in Theorem 1 sows that the sparse excess risk of IHT can be upper bounded by $\tilde{\mathcal{O}}\left(n^{-1/2}\sqrt{k\log(n)\log(ep/k)}\right)$ with high probability over data sample.

The $\tilde{\mathcal{O}}(n^{-1/2}\sqrt{k})$-type rate of convergence established in Theorem 1 is usually referred to as slow rates in statistical learning theory. For strongly convex risk minimization problems, we further derive in Theorem 2 a fast rate of order $\tilde{\mathcal{O}}\left(n^{-1}k(\log^3(n)+\log(p))\right)$ for IHT under additional strong-signal conditions. The key observation is that when the signal strength of the target sparse optimal solution is sufficiently strong, then the support of the target solution can be recovered as a subset of that of the IHT estimation. Consequently, the desired fast rate of convergence can be derived via invoking the theory of local Rademacher complexities (Bartlett et al., 2005) over the supporting set of IHT. Further, specially for well-specified sparse learning models such as sparse generalized linear models, we show through Theorem 3 that an $\tilde{\mathcal{O}}(n^{-1}k\log(p))$ fast rate of convergence can be more directly derived based on the existing parameter estimation error bounds of IHT under mild conditions (Jain et al., 2014; Yuan et al., 2018). To demonstrate the applicability of our theory, we have substantialized these risk bounds to the widely used sparse linear regression and logistic regression models, along with numerical evidences provided to support the theoretical predictions.

In a nutshell, this paper establishes a set of uniform stability induced sparse excess risk bounds for IHT without imposing any distribution-specific assumptions on the data generation model. As a side contribution, we have also derived a fast rate of convergence for IHT when the data is assumed to be generated by a well-specified sparse learning model. Our main results and the related key model assumptions and technical conditions are highlighted in Table 1.

1.3 Paper organization

The paper proceeds with the material organized as follows: In Section 2 and Section 3 we respectively present a set of slow and fast sparse excess risk bounds of IHT via uniform stability arguments. In Section 4, we briefly review the related literature. A comparison of our results to some prior relevant results is provided in Section 5. A preliminary numerical study for theory verification is provided in Section 6. The concluding remarks are made in Section 7. All the technical proofs are relegated to the appendix sections which can be found in the supplementary document.

2 Sparse Excess Risk Bounds of IHT

In this section, we analyze the generalization performance of IHT through the lens of algorithmic stability theory. Particularly, we establish an excess risk bound of IHT induced by the uniform stability of strongly convex ERM restricted over arbitrary feature set of cardinality k.

2.1 Preliminaries

We begin by introducing some definitions and basic assumptions that will be used frequently in the analysis to follow. The concept of uniform stability, as formally defined in below, is a powerful tool for analyzing generalization bounds of M-estimators and their learning algorithms (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2009; Hardt et al., 2016).

Definition 1 (Uniform Stability). *Let $A : \mathcal{X}^n \mapsto \mathcal{W}$ be a learning algorithm that maps a dataset $S \in \mathcal{X}^n$ to a model $A(S) \in \mathcal{W}$. A is said to have uniform stability γ with respect to a loss function $\ell : \mathcal{W} \times \mathcal{X} \mapsto \mathbb{R}$ if for any pair of datasets $S, S' \in \mathcal{X}^n$ that differ in a single element and all $x \in \mathcal{X}$, $|\ell(A(S); x) - \ell(A(S'); x)| \leq \gamma$.*

For an instance, conventional ERM estimators with λ-strongly convex loss functions have uniform stability of order $\mathcal{O}(\frac{1}{\lambda n})$ (Bousquet and Elisseeff, 2002). This fundamental result then gives rise to the ℓ_2-norm regularized ERM which introduces a penalty term $\frac{\lambda}{2}\|w\|^2$ to the convex loss with optimal choice $\lambda = \mathcal{O}(n^{-1/2})$ to balance empirical loss and generalization gap (Shalev-Shwartz et al., 2009).

Our analysis also relies on the conditions of Restricted Strong Convexity/Lipschitz Smoothness (RSC/RLS) which extend the concept of strong convexity and smoothness to the analysis of sparsity recovery methods (Bahmani et al., 2013; Blumensath and Davies, 2009; Jain et al., 2014; Yuan et al., 2018).

Definition 2 (Restricted Strong Convexity/Lipschitz Smoothness). *For any sparsity level $1 \le s \le p$, we say a function f is restricted μ_s-strongly convex and L_s-smooth if there exist $\mu_s, L_s > 0$ such that $\frac{\mu_s}{2}\|w - w'\|^2 \le f(w) - f(w') - \langle \nabla f(w'), w - w'\rangle \le \frac{L_s}{2}\|w - w'\|^2$, $\forall \|w - w'\|_0 \le s$. Particularly, we say f is L-smooth (μ-strongly convex) if f is L_p-smooth (μ_p-strongly convex).*

The ratio number L_s/μ_s will be referred to as *restricted strong condition number* in this paper. By definition we have $L_s \le L_{s'}$ and $\mu_s \ge \mu_{s'}$ for all $s \le s'$. We say that a function f is G-Lipschitz over $\mathcal{W}$ if for all $w, w' \in \mathcal{W}$, $|f(w) - f(w')| \le G\|w - w'\|$. We denote $[p] = \{1, ..., p\}$. The following basic assumptions will be made in different combinations in our analysis.

Assumption 1. *The convex loss function ℓ is G-Lipschitz continuous with respect to its first argument and $\ell(\cdot;\xi) \le M$ for all $\xi \in \mathcal{X}$.*

Assumption 2. *The empirical risk F_S is L_{4k}-smooth and μ_{4k}-strongly convex with probability at least $1 - \delta'_n$ over sample S for some $\delta'_n \in (0, 1)$.*

Assumption 3. *Consider $\bar{w} = \arg\min_{\|w\|_0 \le \bar{k}} F(w)$ and set the sparsity level $k \ge \frac{32L_{4k}^2}{\mu_{4k}^2}\bar{k}$ for IHT.*

Assumption 4. *The population risk F is L-smooth and ρ-strongly convex and without loss of generality $\|w\| \le 1, \forall w \in \mathcal{W}$.*

2.2 A uniform-stability induced risk bound

We now analyze the excess risk of IHT based on the uniform stability of strongly convex ERM. In order to make sure that the output $w_{S,k}^{(T)}$ *at the end of iteration* is uniformly stable, we propose to slightly modify it as $\tilde{w}_{S,k}^{(T)}$ which just minimizes F_S over the support of $\mathrm{supp}(w_{S,k}^{(T)})$, i.e.,

$$\tilde{w}_{S,k}^{(T)} := \arg\min_{w \in \mathcal{W}} F_S(w) \quad \text{s.t. } \mathrm{supp}(w) = \mathrm{supp}(w_{S,k}^{(T)}).$$

Unless otherwise stated, in what follows we will work on the above modified IHT algorithm and assume that $w_{S,k}^{(0)} = 0$. The following result is our main result on the sparse excess risk bound of IHT.

Theorem 1. *Suppose that Assumptions 1, 2, 3 hold. Set the step-size $\eta = \frac{2}{3L_{4k}}$. For any $\delta \in (0, 1 - \delta'_n)$, with probability at least $1 - \delta - \delta'_n$ over the random draw of sample set S, after sufficiently large $T \ge \mathcal{O}\left(\frac{L_{4k}}{\mu_{4k}}\log\left(\frac{nM}{k\log(n)\log(p/k)}\right)\right)$ rounds of IHT iteration, the $\bar{k}$-sparse excess risk of IHT is upper bounded as $F(\tilde{w}_{S,k}^{(T)}) - F(\bar{w}) \le$*

$$\mathcal{O}\left(\frac{G^{\frac{3}{2}}M^{\frac{1}{4}}}{\mu_{4k}^{\frac{3}{4}}}\sqrt{\frac{\log(n)(\log(\frac{1}{\delta}) + k\log(\frac{p}{k}))}{n}} + M\sqrt{\frac{\log(\frac{1}{\delta})}{n}}\right).$$

Proof in sketch. The basic idea is to show that a nearly identical bound holds for ERM restricted over a supporting set of size k and that bound can be extended to IHT in light of Lemma 2 (in Appendix A) and union probability. More precisely, for a given feature index set $J \subseteq [p]$ with $|J| = k$, we first establish a generalization gap bound for the restrictive estimator over J defined by $w_{S|J} := \arg\min_{\mathrm{supp}(w) \subseteq J} F_S(w)$. Since F_S is only assumed to have strong convexity over J with high probability, $w_{S|J}$ is not necessarily uniformly stable. To handle this issue, we propose to alternatively study an ℓ_2-regularized variant of $w_{S|J}$ defined by

$$w_{\lambda,S|J} := \operatorname*{arg\,min}_{\mathrm{supp}(w) \subseteq J}\left\{F_{\lambda,S}(w) := F_S(w) + \frac{\lambda}{2}\|w\|^2\right\},$$

which can be shown to have uniform stability for any $\lambda > 0$. Then according to the result from Bousquet et al. (2020, Corollary 8) its generalization gap is upper bounded by $\tilde{\mathcal{O}}\left(\frac{\log(n)}{\lambda n} + \frac{1}{\sqrt{n}}\right)$. The next key step is to bound the discrepancy between $w_{S|J}$ and $w_{\lambda,S|J}$ as $\|w_{S|J} - w_{\lambda,S|J}\| \le \mathcal{O}\left(\frac{\lambda}{\mu_k + \lambda}\right)$ in view of the (high probability) restricted strong convexity of F_S, which consequently indicates that the generalization guarantee of $w_{\lambda,S|J}$ can be handed over to $w_{S|J}$ with a small overhead of $\mathcal{O}\left(\frac{\lambda}{\mu_k + \lambda}\right)$. Under optimal selection of λ, applying union probability arguments over all the possible J yields a generalization gap bound for ℓ_0-ERM. The final step is to show, according to Lemma 2, that such a generalization gap bound of ℓ_0-ERM leads to the desired sparse excess risk bound of IHT after sufficient iteration with proper sparsity relaxation. A full proof of this result is provided in Appendix B.1. □

Remark 1. *Theorem 1 shows that under proper relaxation of sparsity level, the $\bar{k}$-sparse excess risk of IHT converges at a rate of $\tilde{\mathcal{O}}\left(n^{-1/2}\sqrt{k\log(n)\log(p/k)}\right)$, which matches those of the ℓ_0-penalized binary prediction estimators (Chen and Lee, 2018, 2020) up to logarithmic factors. Such a sparse excess risk bound* *immediately implies an oracle inequality*

$$F(\tilde{w}^{(T)}_{S,k}) - F(w^*)$$
$$\leq \min_{\|w\|_0 \leq \bar{k}} (F(w) - F(w^*)) + \tilde{\mathcal{O}}\left(\sqrt{\frac{k\log(n)\log(p/k)}{n}}\right).$$

Implication for sparse logistic regression. Let us substantialize Theorem 1 to binary logistic regression model with loss function $\ell(w;\xi) = \log(1 + \exp(-2yw^\top x))$ at a labeled data sample $\xi = (x, y) \in \mathbb{R}^p \times \{-1, 1\}$. Given a set of n independently drawn data samples $\{(x_i, y_i)\}_{i=1}^n$, sparse logistic regression learns the parameters so as to minimize the logistic loss function under sparsity constraint:

$$\min_{\|w\|_0 \leq k} F_S(w) = \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(-2y_i w^\top x_i)).$$

Let $X = [x_1, ..., x_n] \in \mathbb{R}^{d\times n}$ be the design matrix and $s(z) = \frac{1}{1+\exp(-z)}$ be the sigmoid function. It can be shown that $\nabla F_S(w) = Xa(w)/n$ in which the vector $a(w) \in \mathbb{R}^n$ is given by $[a(w)]_i = -2y_i(1 - s(2y_i w^\top x_i))$, and the Hessian $\nabla^2 F_S(w) = X\Lambda(w)X^\top/n$ where $\Lambda(w)$ is an $n \times n$ diagonal matrix whose diagonal entries are $[\Lambda(w)]_{ii} = 4s(2y_i w^\top x_i)(1 - s(2y_i w^\top x_i))$. Then we have the following corollary as an application of Theorem 1 to the above sparsity-constrained logistic regression.

Corollary 1. *Assume that x_i are i.i.d. zero-mean sub-Gaussian distribution with covariance matrix $\Sigma \succ 0$ and $\Sigma_{jj} \leq \frac{\sigma^2}{32}$. Suppose that $\|x_i\| \leq 1$ for all i and $\mathcal{W} \subset \mathbb{R}^p$ is bounded by R. Then there exist universal constants $c_0, c_1 > 0$ such that when $n \geq \frac{4kc_1\log(p)}{\lambda_{\min}(\Sigma)}$, for any $\delta \in (0, 1 - \exp\{-c_0 n\})$, with probability at least $1 - \delta - \exp\{-c_0 n\}$ the sparse excess risk of IHT is upper bounded as $F(\tilde{w}^{(T)}_{S,k}) - F(\bar{w}) \leq$*

$$\mathcal{O}\left(\frac{\exp(R)}{\lambda^{3/4}_{\min}(\Sigma)}\sqrt{\frac{\log(n)(\log(\frac{1}{\delta}) + k\log(\frac{p}{k}))}{n}} + R\sqrt{\frac{\log(\frac{1}{\delta})}{n}}\right)$$

after sufficiently large rounds of iteration, i.e.,

$$T \geq \mathcal{O}\left(\frac{\exp(R)}{\lambda_{\min}(\Sigma)}\log\left(\frac{nR}{k\log(n)\log(p/k)}\right)\right).$$

3 Fast Rates for Strongly Convex Optimization

In consistency with statistical learning theory (Bartlett et al., 2005; Srebro et al., 2010; Foster and Syrgkanis, 2019b), we regard the $\tilde{\mathcal{O}}(n^{-1/2})$ rates of convergence established so far as *slow* rates in terms of sample size. In the absence of sparsity constraint, it is well known that convergence rates of order $\tilde{\mathcal{O}}(n^{-1})$ are possible for finite dimensional strongly convex function classes, for instance, via local Rademacher complexities (Bartlett et al., 2005; Koltchinskii, 2006). Inspired by such type of *fast* rates for strongly convex dense ERM, we further show in this section that the $\tilde{\mathcal{O}}(n^{-1})$ sparse excess risk bounds can also be derived for IHT under additional regularity conditions on population risk and signal strength. Moreover, specially for well-specified sparse learning models such as sparse generalized linear models, we show that the $\tilde{\mathcal{O}}(n^{-1})$-type of fast rates can be derived much more directly in view of the conventional parameter estimation error bounds of IHT.

3.1 Fast rates under strong-signal conditions

In what follows, we denote $w_{\min} := \min_{i\in\text{supp}(w)} |w_i|$ as the smallest (in modulus) non-zero entry of a sparse vector w.

Theorem 2. *Suppose that Assumptions 1, 2, 3, 4 hold. Set the step-size $\eta = \frac{2}{3L_{4k}}$. For any given $\delta \in (0, 1)$, assume that $\delta'_n \leq \frac{\delta}{4}$ for large enough n and*

$$\bar{w}_{\min} > \frac{2\sqrt{2k}\|\nabla F(\bar{w})\|_\infty}{\mu_{4k}} + \frac{3G}{\mu_{4k}}\sqrt{\frac{k\log(4p/\delta)}{n}}.$$

Then after $T \geq \mathcal{O}\left(\frac{L_{4k}}{\mu_{4k}}\log\left(\frac{n\mu_{4k}}{kG\log(n)\log(p/k)}\right)\right)$ rounds of IHT iteration, the following upper bound of the $\bar{k}$-sparse excess risk $F(\tilde{w}^{(T)}_{S,k}) - F(\bar{w})$ holds with probability at least $1 - \delta$ over the random draw of S:

$$\mathcal{O}\left(\frac{G^2(\log^3(\rho n) + \log(ep/k))}{\rho}\left(\frac{k}{n}\right) + \frac{\log(1/\delta)}{n}\right).$$

Proof in sketch. For any fixed indices set $J \subseteq [p]$ with $|J| = k$, let us denote $w^*_J := \arg\min_{\text{supp}(w)\subseteq J} F(w)$. A core observation here is that under the strong-signal condition on $\bar{w}_{\min}$, we can show via Lemma 9 (in Appendix C.1) that $\text{supp}(\bar{w}) \subseteq \tilde{J} := \text{supp}(w^{(T)}_{S,k})$ holds with high probability, and thus so does $F(w^*_{\tilde{J}}) \leq F(\bar{w})$. As another key ingredient, we then show through Lemma 8 (in Appendix C.1) that $\sup_{J\subseteq[p],|J|=k}\left\{F(w_{S|J}) - F(w^*_J)\right\}$ is uniformly upper bounded as $\tilde{\mathcal{O}}(k/n)$ with high probability. In view of this supporting-set-wise uniform excess risk bound, the desired bound follows directly by noting $F(w^{(T)}_{S,k}) - F(\bar{w}) \leq F(w^{(T)}_{S,k}) - F(w^*_{\tilde{J}}) = F(w_{S|\tilde{J}}) - F(w^*_{\tilde{J}})$. A full proof of this result is provided in Appendix C.1. □

Remark 2. *Consider the well-specified setting where $\nabla F(\bar{w}) = 0$, i.e., the minimizer of the population risk is truly sparse. In this case, under the signal-strength* *condition $\bar{w}_{\min} = \tilde{\Omega}\left(\sqrt{k/n}\right)$, Theorem 2 suggests that the sparse excess risk bound of IHT decays as fast as $\tilde{\mathcal{O}}(k/n)$ with high probability. A benefit of the result in Theorem 2 is that it allows for misspecified sparse models. More precisely, even if the risk F does not have zero gradient at $\bar{w}$, the sparse excess risk of IHT can still converge as fast as $\tilde{\mathcal{O}}(k/n)$ provided that $\bar{w}_{\min}$ significantly outweighes $\tilde{\Omega}(\sqrt{k}\|\nabla F(\bar{w})\|_\infty + \sqrt{k/n})$.*

3.2 Fast rates for well-specified sparse learning models

The sparse excess risk bounds derived so far are essentially for misspecified sparse learning models. In this subsection, we further study the risk bounds of IHT in well-specified scenarios where the data is assumed to be generated according to a truly sparse model. Such a statistical treatment is conventional in the theoretical analysis of high-dimensional sparsity recovery approaches (Agarwal et al., 2012; Mei et al., 2018; Yuan et al., 2018). More specifically, we assume that there exists a k-sparse parameter vector $\bar{w}$ such that, roughly speaking, the population risk function is minimized exactly at $\bar{w}$ with $\nabla F(\bar{w}) = 0$. Formally, we impose the following assumption on the loss function which basically requires the gradient of loss at $\bar{w}$ obeys a light tailed distribution.

Assumption 5 (Sub-Gaussian gradient at the true model)**.** *For each $j \in \{1, ..., p\}$, we assume that $\nabla_j \ell(\bar{w}; \xi)$ is σ^2-sub-Gaussian with zero mean, namely, $\mathbb{E}_\xi[\nabla_j \ell(\bar{w}; \xi)] = 0$ and there exists a constant $\sigma > 0$ such that for any real number τ,*

$$\mathbb{E}_\xi\left[\exp\left\{\tau(\nabla_j \ell(\bar{w}; \xi))\right\}\right] \leq \exp\left\{\frac{\sigma^2\tau^2}{2}\right\}.$$

Remark 3. *The zero-mean assumption directly implies $\nabla F(\bar{w}) = 0$. As we will show shortly, this assumption can be fulfilled by linear regression and logistic regression models.*

As a side contribution of this work, we present in the following theorem a sharper excess risk bound of IHT for well-specified sparse learning models under less stringent conditions. A proof of this theorem is deferred to Appendix C.2.

Theorem 3. *Assume that $\bar{w}$ is a $\bar{k}$-sparse vector satisfying Assumption 5. Suppose that Assumptions 2, 3, 4 hold. Then for any $\delta \in (0, 1 - \delta'_n)$ and any $\epsilon > 0$, IHT with step-size $\eta = \frac{2}{3L_{4k}}$ and sufficiently large $T \geq \mathcal{O}\left(\frac{L_{4k}}{\mu_{4k}} \log\left(\frac{n\mu_{4k}}{k\sigma^2 \log(p/\delta)}\right)\right)$ rounds of iteration will output $w_{S,k}^{(T)}$ such that the following sparse excess risk bound holds with probability at least $1 - \delta - \delta'_n$ over S,*

$$F(w_{S,k}^{(T)}) - F(\bar{w}) \leq \mathcal{O}\left(\frac{L}{\mu_{4k}^2}\left(\frac{k\sigma^2 \log(p/\delta)}{n}\right)\right).$$

Remark 4. *In comparison to the risk bound established in Theorem 2 that allows for misspecified models, the above fast rate of convergence for well-specified models is sharper in the sense that it is not dependence on $\log(n)$-factors and it is valid without needing to assume Lipschitz-loss and strong-signal conditions.*

Remark 5. *We comment on the tightness of the excess risk bounds in Theorem 3 in the minimax sense. It is well known (see, e.g., Rigollet, 2015; Zhang et al., 2014) that, up to logarithmic factors, the high probability bound $\tilde{\mathcal{O}}\left(n^{-1}k\log(p)\right)$ is minimax optimal for the squared estimation error $\|w_{S,k}^{(T)} - \bar{w}\|^2$, which immediately implies that the same bound should be minimax optimal for excess risk provided that the population function F is strongly convex.*

We next showcase how to apply the bounds in Theorem 3 to the widely used sparse linear regression and logistic regression models.

Implication for sparse linear regression. We assume the samples $S = \{x_i, y_i\}$ obey the linear model $y_i = \bar{w}^\top x_i + \varepsilon_i$ where $\bar{w}$ is a k-sparse parameter vector, the random feature vectors x_i are drawn i.i.d. from a zero-mean sub-Gaussian distribution with covariance matrix $\Sigma \succ 0$, and ε_i are n i.i.d. zero-mean sub-Gaussian random variables with parameter σ^2. The sparsity-constrained least squares regression model is then written by

$$\min_{\|w\|_0 \leq k} F_S(w) = \frac{1}{2n}\sum_{i=1}^{n} \|y_i - w^\top x_i\|^2.$$

We present the following corollary as a consequence of Theorem 3 to the considered linear regression model with bounded design. See Appendix C.3 for its proof.

Corollary 2. *Assume that ε_i are i.i.d. zero-mean σ^2-sub-Gaussian and x_i are i.i.d. zero-mean sub-Gaussian distribution with covariance matrix $\Sigma \succ 0$ and $\Sigma_{jj} \leq 1$. Then there exist universal constants $c_0, c_1 > 0$ such that when $n \geq \frac{4kc_1 \log(p)}{\lambda_{\min}(\Sigma)}$, for any $\delta \in (0, 1 - \exp\{-c_0 n\})$, with probability at least $1 - \delta - \exp\{-c_0 n\}$ the sparse excess risk of IHT with step-size $\eta = \mathcal{O}\left(\frac{1}{\lambda_{\max}(\Sigma)}\right)$ is bounded as*

$$F(w_{S,k}^{(T)}) - F(\bar{w}) \leq \mathcal{O}\left(\frac{\lambda_{\max}(\Sigma)\sigma^2 \log(p/\delta)}{\lambda_{\min}^2(\Sigma)}\left(\frac{k}{n}\right)\right)$$

after $T \geq \mathcal{O}\left(\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \log\left(\frac{n\lambda_{\min}(\Sigma)}{k\sigma^2 \log(p/\delta)}\right)\right)$ rounds of iteration.

Implication for sparse logistic regression. Let us further consider a well-specified binary logistic regression model in which the relation between the random feature vector $x \in \mathbb{R}^p$ and its associated random binary label $y \in \{-1, +1\}$ is determined by the conditional probability $\mathbb{P}(y|x;\bar{w}) = \frac{\exp(2y\bar{w}^\top x)}{1+\exp(2y\bar{w}^\top x)}$, where $\bar{w}$ is a k-sparse parameter vector. Then we have the following corollary as an application of Theorem 3 to this well-specified sparse logistic regression model. A proof of this result is provided in Appendix C.3.

Corollary 3. *Assume that x_i are i.i.d. zero-mean sub-Gaussian distribution with covariance matrix $\Sigma \succ 0$ and $\Sigma_{jj} \le \frac{\sigma^2}{32}$. Suppose that $\|x_i\| \le 1$ for all i and $\mathcal{W} \subset \mathbb{R}^p$ is bounded by R. Then there exist universal constants $c_0, c_1 > 0$ such that when $n \ge \frac{4kc_1\log(p)}{\lambda_{\min}(\Sigma)}$, for any $\delta \in (0, 1-\exp\{-c_0 n\})$, with probability at least $1-\delta-\exp\{-c_0 n\}$ the sparse excess risk of IHT with step-size $\eta = \mathcal{O}(1)$ is upper bounded by*

$$F(w^{(t)}_{S,k}) - F(\bar{w}) \le \mathcal{O}\left(\frac{\exp(R)\sigma^2\log(p/\delta)}{\lambda^2_{\min}(\Sigma)}\left(\frac{k}{n}\right)\right)$$

after $T \ge \mathcal{O}\left(\frac{\exp(R)}{\lambda_{\min}(\Sigma)}\log\left(\frac{n\lambda_{\min}(\Sigma)}{k\exp(R)\sigma^2\log(p/\delta)}\right)\right)$ rounds of iteration.

4 Other Related Work

The problem regime considered in this paper lies at the intersection of high-dimensional sparse M-estimation and statistical learning theory, both of which have long been studied with a vast body of beautiful theoretical results established in literature. Next we will incompletely connect our research to several closely relevant lines of study in this context. We refer the interested readers to Cesa-Bianchi and Lugosi (2006); Hastie et al. (2015); Wainwright (2019) and the references there in for a more comprehensive coverage of the related topics.

Consistency and generalization of M-estimation with sparsity. Statistical consistency of learning with sparsity models is now well understood for some popular sparse M-estimators including ℓ_0-ERM (Foucart and Rauhut, 2017; Rigollet, 2015), ℓ_1-penalized estimations (Lasso) (Meinshausen and Yu, 2009; Wainwright, 2009) and folded concave penalization (Fan and Li, 2001; Zhang and Zhang, 2012). The generalization ability of sparsity-inducing learning models is relatively less understood but has gained recent significant attention. The excess risk of Lasso for generalized linear models was investigated by Van de Geer (2008). Later with almost no assumptions imposed on the design matrix, the least squares Lasso estimator was still shown to be consistent in out-of-sample predictive risk (Chatterjee, 2013). For a class of ℓ_1-penalized high dimensional M-estimators with non-convex loss functions, uniform convergence bounds with polynomial dependence on the sparsity level of certain nominal model were established by Mei et al. (2018). The misclassification excess risk of sparsity-penalized binary logistic regression has been investigated by Abramovich and Grinshtein (2019) with near-optimal high probability bounds established. For linear prediction models, a data dependent generalization error bound was derived for a class of risk minimization algorithms with structured sparsity constraints (Maurer and Pontil, 2012). Particularly concerning the generalization of ℓ_0-ERM, a set of uniform excess risk bounds were derived by Chen and Lee (2018, 2020) for binary loss functions under proper regularity conditions. More recently, based on the arguments of localized Rademacher complexity (Bartlett et al., 2005), tighter risk bounds for ℓ_0-ERM have been established over bounded liner prediction classes (Foster and Syrgkanis, 2019b). The existing uniform convergence implied excess risk bounds for ℓ_0-ERM, however, rely largely on its optimality which is NP-hard to be estimated exactly in high-dimensional setting. It is not yet clear if these results can be extended to approximate sparsity recovery algorithms such as IHT considered in this work.

Statistical guarantees on IHT-style algorithms. The IHT-style algorithms have been popularly applied and studied in compressed sensing and sparse learning (Blumensath and Davies, 2009; Foucart, 2011). Recent works have demonstrated that by imposing certain assumptions such as restricted strong convexity/smothness and restricted isometry property (RIP) over the risk function, IHT and its variants converge linearly towards certain nominal sparse model with near-optimal estimation accuracy (Bahmani et al., 2013; Yuan et al., 2014). It was later shown by Jain et al. (2014); Shen and Li (2017b) that with proper relaxation of sparsity level, high-dimensional estimation consistency can be established for IHT without assuming RIP conditions. The sparsity recovery performance of IHT-style methods was investigated by Yuan et al. (2016); Shen and Li (2017a) to understand when the algorithm can exactly recover the support of a sparse signal from its compressed measurements. The excess risk analysis of IHT, however, still remains an open challenge that we aim to attack in this work.

Stability and generalization of ERM. The idea of using stability of the algorithm with respect to changes in the training set for generalization error analysis dates back to the seventies (Rogers and Wagner, 1978; Devroye and Wagner, 1979). Since the seminal work of Bousquet and Elisseeff (2002), stability has been extensively studied with a bunch of applications to establishing generalization bounds for strongly convex ERM estimators (Mukherjee et al., 2006; Shalev-Shwartz et al., 2009). Recently, it was shown that the solution obtained via (stochastic) gradient descent is expected to be stable and generalize well for smooth convex and non-convex loss functions (Hardt et al., 2016). Later, a set of data-dependent generalization bounds for SGD were derived based on the stability of algorithm (Kuzborskij and Lampert, 2018). More broadly, generalization bounds for stable learning algorithms (e.g., GD, SGD and SVRG) that converge to global minima were established by Charles and Papailiopoulos (2018). There is a recent renewed interest in the use of uniform stability for deriving high probability risk bounds of strongly convex ERM and optimization algorithms (Feldman and Vondrak, 2018, 2019; Bousquet et al., 2020). We highlight that our generalization analysis of IHT is a novel extension of the uniform stability theory in the direction of non-convex sparse learning under hard sparsity constraint.

5 Discussions

In this section, we discuss the connections and differences between our results established in the previous sections and a number of existing risk bounds for the ℓ_0-ERM and Lasso-type estimators. For each estimator, we distinguish the comparison in two settings of fast and slow convergence rates respectively.

5.1 Comparison with the risk bounds for ℓ_0-ERM

We begin with comparing our results with existing risk bounds for the ℓ_0-ERM.

Fast rates. Given that the ℓ_0-ERM estimator is exactly solved, an essentially $\tilde{\mathcal{O}}(n^{-1}k\log(p))$ sparse excess risk bound has been established for its output over bounded liner prediction classes (Foster and Syrgkanis, 2019a, Example 2). That bound, however, is more of pure theoretical interest than practical usage due to the computational hardness of ℓ_0-ERM. Contrastingly, Theorem 2 shows that an about the same fast rate of convergence can also be derived for IHT which is computationally tractable and efficient for sparsity recovery.

Slow rates. We comment on the difference between the $\tilde{\mathcal{O}}\left(n^{-1/2}\sqrt{k\log(n)\log(p)}\right)$ rate in Theorem 1 and a comparable result established via uniform concentration bounds (Chen and Lee, 2018, Theorem 1) for sparsity constrained binary prediction problems. First, our bound holds for real-valued Lipschitz continuous convex loss functions while that bound was tailored for binary loss functions with linear models. Second, regarding the regularization condition, the result in (Chen and Lee, 2018, Theorem 1) requires $p \vee n \gtrsim k^8$ which could be fairly unrealistic even when k is moderate in real problems. In contrast, we impose much more natural conditions (like $\frac{n}{\log(n)} \gtrsim k\log(p)$ for logistic regression) on data scale.

5.2 Comparison with the risk bounds for Lasso estimators

We further compare the excess risk bounds of IHT to those of the following ℓ_1-regularized ERM (Lasso) estimator (Tibshirani, 1996; Wainwright, 2009):

$$w^{\ell_1}_{S,\lambda} := \arg\min_{w\in\mathcal{W}} F_S(w) + \lambda\|w\|_1$$

which is popularly used as a convex surrogate of the ℓ_0-ERM estimator.

Fast rates. For high-dimensional generalized linear models, the oracle inequality of Van de Geer (2008, Theorem 2.1) suggests that if the target solution $w^* = \arg\min_{w\in\mathcal{W}} F(w)$ is exactly k-sparse, then it holds with high probability that

$$F(w^{\ell_1}_{S,\lambda}) - F(w^*) \le \tilde{\mathcal{O}}\left(\frac{k\log(p)}{n}\right)$$

under $\lambda \asymp n^{-1/2}\sqrt{\log(p)}$. Specially for the well-specified sparse linear regression models, a similar fast rate of convergence can be implied under natural conditions by the parameter estimation error bounds of Negahban et al. (2012, Corollary 2). In Theorem 3, we have shown that the $\tilde{\mathcal{O}}(n^{-1}k\log(p))$ rate also applies to IHT for well-specified sparse learning with sub-Gaussian noises. In comparison to these fast rates for well-specified sparsity models, the fast rate established in Theorem 2 applies to misspecified sparsity models and the analysis allows for non-linear models.

Slow rates. For well-specified linear regression models with the ℓ_1-norm of parameter vector upper bounded by K, it has been shown by Chatterjee (2013) that the expected excess risk of a constrained Lasso estimator scales as $\tilde{\mathcal{O}}(n^{-1/2}K^2\sqrt{\log(p)})$ under mild conditions on the design matrix. To compare it with the $\tilde{\mathcal{O}}(n^{-1/2}\sqrt{k\log(n)\log(p)})$ bound of IHT established in Theorem 1, we remark that 1) these two rates are comparable (up to logarithmic factors) when $K^2 \asymp \sqrt{k}$; 2) the former holds in expectation while the latter (ours) holds in high probability; and 3) more importantly, our bound is applicable to a broader range of learning problems beyond well-specified sparse linear regression, yet at the price of imposing more stringent assumptions on the risk function.

6 Simulation Study

In this section, we carry out a set of numerical experiments on synthetic sparse logistic regression tasks to verify the IHT generalization theory presented in

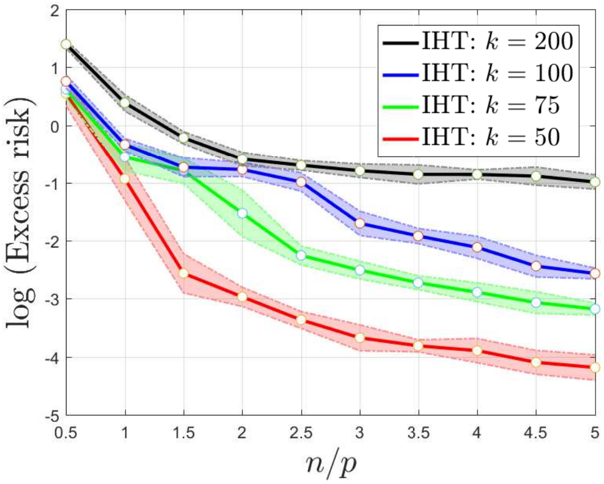

Figure 1: The impact of sample size and sparsity level on the excess risk of sparse logistic regression. The y-axis represents the logarithmic scale of excess risk.

Section 2 and Section 3. Throughout our numerical study, we initialize $w^{(0)} = 0$ for IHT and replicate each individual experiment 10 times over the random generation of training data for generalization performance evaluation.

Experiment setup. We consider the binary logistic regression model with loss function $\ell(w; x_i, y_i) = \log\left(1 + \exp(-y_i w^\top x_i)\right)$. In this set of simulation study, each data feature x_i is sampled from standard multivariate Gaussian distribution and its binary label $y_i \in \{-1, +1\}$ is determined by the conditional probability $\mathbb{P}(y_i|x_i; \bar{w}) = \frac{\exp(2y_i\bar{w}^\top x)}{1+\exp(2y_i\bar{w}^\top x_i)}$ with a $\bar{k}$-sparse parameter vector $\bar{w}$. We test with feature dimension $p = 1000, \bar{k} = 50$ and aim to show the impact of varying ratio $n/p \in (0.5, 5)$ and sparsity level $k \in \{50, 75, 100, 200\}$ on the actual generalization performance of IHT. Since for logistic loss the population risk function F has no close-form expression, we approximate the population value $F(w)$ by its empirical version with sufficient sampling. In order to compute the excess risk, we need to estimate the optimal population risk which in view of the proof of Corollary 1 is given by $\min_{\|w\|_0 \le k} F(w) = F(\bar{w})$ for any $k \ge \bar{k}$.

Numerical results. The evolving curves of excess risk as functions of sample size under different sparsity levels are illustrated in Figure 1, which is plot in semi-log layout with y-axis representing the logarithmic scale of sparse excess risk. For each fixed k, it can be observed that the sparse excess risk of IHT decrease as sample size n increases, while for each fixed n, the same performance measurement increases as k increases. These observations are consistent with the implications of Theorem 1 (and Theorem 2 as well) to sparse binary logistic regression.

7 Conclusions

In this paper, we established a set of novel sparse excess risk bounds for the widely applied IHT method from the perspective of unform stability. Specifically, we have shown that the sparse excess risk of IHT converges at the rate of $\tilde{\mathcal{O}}(n^{-1/2}\sqrt{k\log(n)\log(p)})$ with high probability under natural regularity conditions. Under additional strong-signal conditions, we further proved faster rates of order $\tilde{\mathcal{O}}(n^{-1}k(\log^3(n) + \log(p)))$ for strongly convex risk minimization problems. These sparse excess risk bounds immediately give rise to oracle excess risk inequalities of IHT over cardinality constraint. As a side contribution, we have further shown a fast rate of order $\tilde{\mathcal{O}}(n^{-1}k\log(p))$ for IHT for well-specified sparse learning models with sub-Gaussian noises.

We expect that the theory developed in this paper will fuel future investigation on the generalization bounds of IHT for non-convex loss functions such as those used in the common practice of deep neural nets pruning (Frankle and Carbin, 2019; Han et al., 2016), yet rarely studied in theory. Actually, based on the standard $\tilde{\mathcal{O}}\left(\sqrt{p/n}\right)$ uniform convergence bound for dense models (see, e.g., Shalev-Shwartz et al., 2009), using the arguments of this paper it is more or less straightforward to derive a generalization bound of order $\tilde{\mathcal{O}}\left(\sqrt{k\log(p)/n}\right)$ for IHT which is applicable to the non-convex regime. Also, it is interesting to further explore the structure information such as the deep and wide architectures to hopefully obtain stronger generalization bounds for deep learning with sparsity.

Acknowledgements

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 and No.61936005.

References

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

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, 2012.

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 (Mar):807–841, 2013.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101 (473):138–156, 2006.

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

Károly Böröczky and Gergely Wintsche. Covering the sphere by equal spherical balls. In *Discrete and Computational Geometry*, pages 235–251. Springer, 2003.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626, 2020.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 744–753, 2018.

Sourav Chatterjee. Assumptionless consistency of the lasso. *arXiv preprint arXiv:1303.5817*, 2013.

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

Le-Yu Chen and Sokbae Lee. Binary classification with covariate selection through ℓ_0-penalized empirical risk minimization. *The Econometrics Journal*, pages 1–16, 2020.

Luc Devroye and Terry Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979.

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

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems*, pages 9747–9757, 2018.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279, 2019.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019a.

Dylan J Foster and Vasilis Syrgkanis. Statistical learning with a nuisance component. In *Conference on Learning Theory*, pages 1346–1348, 2019b.

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.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*, 2016.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

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

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.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In

Advances in Neural Information Processing Systems, pages 793–800, 2009.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2820–2829, 2018.

Andreas Maurer and Massimiliano Pontil. Structured sparsity and generalization. *Journal of Machine Learning Research*, 13(Mar):671–690, 2012.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24 (2):227–234, 1995.

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.

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

William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *Conference on Learning Theory*, 2009.

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, 2017a.

Jie Shen and Ping Li. A tight bound of hard thresholding. *The Journal of Machine Learning Research*, 18(1):7650–7691, 2017b.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Xiao-Tong Yuan and Ping Li. Nearly non-expansive bounds for mahalanobis hard thresholding. In *Conference on Learning Theory*, pages 3787–3813, 2020.

Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *International Conference on Machine Learning*, pages 127–135, 2014.

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*, 18:1–43, 2018.

Cun-Hui Zhang and Tong Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

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 1984–1993, 2018.