HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2024 March 28.

Published in final edited form as:

J Am Stat Assoc. 2023 ; 118(544): 2315–2328. doi:10.1080/01621459.2022.2044824.

Understanding Implicit Regularization in Over-Parameterized Single Index Model

Jianqing Fan,

Frederick L. Moore '18 Professor of Finance, Professor of Statistics, and Professor of Operations Research and Financial Engineering at the Princeton University.

Zhuoran Yang,

Ph.D. students at Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA.

Mengxin Yu

Ph.D. students at Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA.

Abstract

In this paper, we leverage over-parameterization to design regularization-free algorithms for the high-dimensional single index model and provide theoretical guarantees for the induced implicit regularization phenomenon. Specifically, we study both vector and matrix single index models where the link function is nonlinear and unknown, the signal parameter is either a sparse vector or a low-rank symmetric matrix, and the response variable can be heavy-tailed. To gain a better understanding of the role played by implicit regularization without excess technicality, we assume that the distribution of the covariates is known a priori. For both the vector and matrix settings, we construct an over-parameterized least-squares loss function by employing the score function transform and a robust truncation step designed specifically for heavy-tailed data. We propose to estimate the true parameter by applying regularization-free gradient descent to the loss function. When the initialization is close to the origin and the stepsize is sufficiently small, we prove that the obtained solution achieves minimax optimal statistical rates of convergence in both the vector and matrix cases. In addition, our experimental results support our theoretical findings and also demonstrate that our methods empirically outperform classical methods with explicit regularization in terms of both ℓ_2-statistical rate and variable selection consistency.

Keywords

Implicit Regularization; high-dimensional models; over-parameterization; single-index models

1 Introduction

With the astonishing empirical success in various application domains such as computer vision [Voulodimos et al., 2018], natural language processing [Otter et al., 2020, Torfi et

jqfan@princeton.edu.

al., 2020], and reinforcement learning [Arulkumaran et al., 2017, Li, 2017], deep learning [LeCun et al., 2015, Goodfellow et al., 2016, Fan et al., 2021a] has become one of the most prevalent classes of machine learning methods. When applying deep learning to supervised learning tasks such as regression and classification, the regression function or classifier is represented by a deep neural network, which is learned by minimizing a loss function of the network weights. Here the loss function is defined as the empirical risk function computed based on the training data and the optimization problem is usually solved by gradient-based optimization methods. Due to the nonlinearity of the activation function and the multi-layer functional composition, the landscape of the loss function is highly nonconvex, with many saddle points and local minima [Dauphin et al., 2014, Swirszcz et al., 2016, Yun et al., 2019]. Moreover, oftentimes the neural network is over-parameterized in the sense that the total number of network weights exceeds the number of training data, making the regression or classification problem ill-posed from a statistical perspective. Surprisingly, however, it is often observed empirically that simple algorithms such as (stochastic) gradient descent tend to find the global minimum of the loss function despite non-convexity. Moreover, the obtained solution also generalizes well to unseen data with small test error [Neyshabur et al., 2015, Zhang et al., 2017]. These mysterious observations cannot be fully explained by the classical theory of nonconvex optimization and generalization bounds via uniform convergence.

To understand such an intriguing phenomenon, Neyshabur et al. [2015], Zhang et al. [2017] show empirically that the generalization stems from an “implicit regularization” of the optimization algorithm. Specifically, they observe that, in over-parametrized statistical models, although the optimization problems consist of bad local minima with large generalization error, the choice of optimization algorithm, usually a variant of gradient descent algorithm, usually guard the iterates from bad local minima and prefers the solution that generalizes well. Thus, without adding any regularization term in the optimization objective, the implicit preference of the optimization algorithm itself plays the role of regularization. Implicit regularization has been shown indispensable in training deep learning models [Neyshabur et al., 2015, 2017, Zhang et al., 2017, Keskar et al., 2017, Poggio et al., 2017, Wilson et al., 2017].

With properly designed algorithm, Gunasekar et al. [2017] and Li et al. [2018] provide empirical evidence and theoretical guarantees for the implicit regularization of gradient descent for least-squares regression with a two-layer linear neural network, i.e., low-rank matrix sensing. They show that gradient descent biases towards the minimum nuclear norm solution when the initialization is close to the origin, stepsizes are sufficiently small, and no explicit regularization is imposed. More specifically, when the true parameter is a rank r positive-semidefinite matrix in $\mathbb{R}^{d\times d}$, they rewrite the parameter as $\mathbf{U}\mathbf{U}^{\top}$, where $\mathbf{U}\in\mathbb{R}^{d\times d}$, and propose to estimate the true parameter by updating $\mathbf{U}$ via gradient descent. Li et al. [2018] proves that, with $\widetilde{\mathcal{O}}(r^2d)$ i.i.d. observations of the model, gradient descent provably recovers the true parameter with accuracy, where $\widetilde{\mathcal{O}}(\cdot)$ hides absolute constants and polylogarithmic terms. Thus, in over-parametrized matrix sensing problems, the implicit regularization of gradient descent can be viewed as equivalent to adding a nuclear norm penalty explicitly. See also Arora et al. [2019] for a related topic on deep linear network.

Moreover, Zhao et al. [2019], Vaškevičius et al. [2019] recently design a noval regularization-free algorithm and study the implicit regularization of gradient descent for high-dimensional linear regression with a sparse signal parameter, which is a vector in $\mathbb{R}^p$ with s nonzero entries. They propose to re-parametrize the parameter using two vectors in $\mathbb{R}^p$ via the Hadamard product and estimate the true parameter via un-regularized gradient descent with proper initialization, stepsizes, and the number of iterations. They prove independently that, with $n = \mathcal{O}\left(s^2 \log p\right)$ i.i.d. observations, gradient descent yields an estimator of the true parameter with the optimal statistical accuracy. More interestingly, when the nonzero entries of the true parameter all have sufficiently large magnitude, the proposed estimator attains the oracle $\mathcal{O}(\sqrt{s \log s/n})$ rate that is independent of the ambient dimension p. Hence, for sparse linear regression, the implicit regularization of gradient descent has the same effect as the folded concave penalties [Fan et al., 2014] such as smoothly clipped absolute deviation (SCAD) [Fan and Li, 2001] and minimax concave penalty (MCP) [Zhang et al., 2010].

The aforementioned works all design algorithms and establish theoretical results for linear statistical models with light-tailed noise, which is slightly restricted since linear models with sub-Gaussian noise only comprise a small proportion of the models of interest in statistics. For example, in the field of finance, linear models only bring limited contributions and the datasets are always corrupted by heavy-tailed noise. Thus, one questions is left open:

> Can we leverage over-parameterization and implicit regularization to establish statistically accurate estimation procedures for a more general class of high-dimensional statistical models with possibly heavy-tailed data?

In this work, we focus on the single index model, where the response variable Y and the covariate X satisfy $Y = f(\langle X, \beta^* \rangle) + \epsilon$, with β^* being the true parameter, ϵ being the random noise, and $f : \mathbb{R} \to \mathbb{R}$ being an unknown (nonlinear) link function. Here β^* is either a s-sparse vector in $\mathbb{R}^p$ or a rank r matrix in $\mathbb{R}^{d \times d}$. Since f is unknown, the norm of β is not identifiable. Thus, for the vector and matrix cases respectively, we further assume that the ℓ_2- or Frobenius norms of β^* are equal to one. Our goal is to recover the true parameter β^* given n i.i.d. observations of the model. Such a model can be viewed as the misspecified version of the compressed sensing [Donoho, 2006, Candés, 2008] and phase retrieval [Shechtman et al., 2015, Candés et al., 2015] models, which corresponds to the identical and quadratic link functions respectively.

In a single index model, due to the unknown link function, it is infeasible to directly estimate β^* via nonlinear least-squares. Moreover, jointly minimizing the least-squares loss function with respect to β^* and f is computationally intractable. To overcome these challenges, a recent line of research proposes to estimate β^* by the method of moments when the distribution of X is known. This helps us provide a deep understanding on the implicit regularization induced by over-parameterization in the nonlinear models without excessive technicality and eliminate other complicated factors that convolve insights. Specifically, when X is a standard Gaussian random variable, Stein's identity [Stein et al., 1972] implies that the expectation of $Y \cdot X$ is proportional to β^*. Thus, despite the nonlinear link function, β^* can be accurately estimated by neglecting f and fitting a regularized

least-squares regression. In particular, when β^* is a sparse vector, Plan and Vershynin [2016], Plan et al. [2017] prove that the Lasso estimator achieves the optimal statistical rate of convergence. Subsequently, such an approach has been extended to the cases beyond Gaussian covariates. In particular, Goldstein et al. [2018], Wei [2018], Goldstein and Wei [2019] allow the covariates to follow an elliptically symmetric distribution that can be heavy-tailed. In addition, utilizing a generalized version of Stein's identity [Stein et al., 2004], Yang et al. [2017] extends the Lasso approach to the setting where the covariate X has a known density p_0. Specifically, when p_0 is known, we can define the score function $S_{p_0}(\cdot)$ as $S_{p_0}(\cdot) = -\nabla \log p_0(\cdot)$, which enjoys the property that $\mathbb{E}[Y \cdot S_{p_0}(X)]$ identifies the direction of β^*. Thus, the true parameter can be estimated via M-estimation with $S_{p_0}(X)$ serving as the covariate.

To answer the question given above, in this work, we leverage over-parameterization to design regularization-free algorithms for single index model and provide theoretical guarantees for the induced implicit regularization phenomenon. To be more specific, we first adopt the quadratic loss function in Yang et al. [2017] and rewrite the parameter of interest by over-parameterization. When β^* is a sparse vector in $\mathbb{R}^p$, we adopt a Hadamard product parameterization [Hoff, 2017, Zhao et al., 2019, Vaškevičius et al., 2019] and write β^* as $\mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v}$, where both $\mathbf{w}$ and $\mathbf{v}$ are vectors in $\mathbb{R}^p$. We propose to minimize the loss function as a function of the new parameters via gradient descent, where both $\mathbf{w}$ and $\mathbf{v}$ are initialized near an all-zero vector and the stepsizes are fixed to be a sufficiently small constant $\eta > 0$. Furthermore, when β^* is a low-rank matrix, we similarly represent β^* as $\mathbf{W}\mathbf{W}^\top - \mathbf{V}\mathbf{V}^\top$ and propose to recover β^* by applying the gradient descent algorithm to the quadratic loss function under the new parameterization.

Furthermore, the analysis of our algorithm faces the following two challenges. First, due to over-parameterization, there exist exponentially many stationary points of the population loss function that are far from the true parameter. Thus, it seems that the gradient descent algorithm would be likely to return a stationary point that incurs a large error. Second, both the response Y and the score $S_{p_0}(X)$ can be heavy-tailed random variables. Thus, the gradient of the empirical loss function can deviate significantly from its expectation, which poses an additional challenge to establishing the statistical error of the proposed estimator.

To overcome these difficulties, in our algorithm, instead of estimating $\mathbb{E}[Y \cdot S_{p_0}(X)]$ by its empirical counterpart, we construct robust estimators via proper truncation techniques, which have been widely applied in high-dimensional M-estimation problems with heavy-tailed data [Fan et al., 2021c, Zhu, 2017, Wei and Minsker, 2017, Minsker, 2018, Fan et al., 2021b, Ke et al., 2019, Minsker and Wei, 2020]. These robust estimators are then employed to compute the update directions of the gradient descent algorithm. Moreover, despite the seemingly perilous loss surface, we prove that, when initialized near the origin and sufficiently small stepsizes, the gradient descent algorithm guard the iterates from bad stationary points. More importantly, when the number of iterations is properly chosen, the obtained estimator provably enjoys (near-)optimal $\mathcal{O}(\sqrt{s \log p/n})$ and $\mathcal{O}(\sqrt{rd \log d/n})$ ℓ_2-statistical rates under the sparse and low-rank settings, respectively. Moreover, for sparse β^*, when the magnitude of the nonzero entries is sufficiently large, we prove that our estimator

enjoys an oracle $\mathcal{O}(\sqrt{s \log n/n})$ ℓ_2-statistical rate, which is independent of the dimensionality p. Our proof is based on a jointly statistical and computational analysis of the gradient descent dynamics. Specifically, we decompose the iterates into a signal part and a noise part, where the signal part share the same sparse or low-rank structures as the true signal and the noise part are orthogonal to the true signal. We prove that the signal part converges to the true parameter efficiently whereas the noise part accumulates at a rather slow rate and thus remains small for a sufficiently large number of iterations. Such a dichotomy between the signal and noise parts characterizes the implicit regularization of the gradient descent algorithm and enables us to establish the statistical error of the final estimator.

Furthermore, our method has several merits compared with classical regularized methods. From the theoretical perspective, our strengths are two-fold. First, as we mentioned in the last paragraph, under mild conditions, our estimator enjoys oracle statistical rate whereas the most commonly used ℓ_1-regularized method always results in large bias. In this case, our method is equivalent with adding folded-concave regularizers (e.g. SCAD, MCP) to the loss function. Second, for all estimators inside the wide optimal time interval, our range of choosing the truncating parameter to achieve variable selection consistency (rank consistency) is much wider than classical regularized methods. Thus, our method is more robust than all regularized methods in terms of selecting the truncating parameter. Meanwhile, from the aspect of applications, our strengths are three-fold. First, in terms of ℓ_2-statistical rate, numerical studies show that our method generalizes even better than adding folded-concave penalties. Second, from the aspect of variable selection, experimental results also show that the robustness of our method helps reduce false positive rates greatly. Last but not least, as we only need to run gradient descent and the gradient information is able to be efficiently transferred among different machines, our method is easier to be paralleled and generalized to large-scale problems. Thus, our method can be applied to modern machine learning applications such as federated learning.

To summarize, our contribution is several-fold. First, for sparse and low-rank single index models where the random noise is possible heavy-tailed, we employ a quadratic loss function based on a robust estimator of $\mathbb{E}[Y \cdot S_{p_0}(X)]$ and propose to estimate β^* by combining over-parameterization and regularization-free gradient descent. Second, we prove that, when the initialization, stepsizes, and stopping time of the gradient descent algorithm are properly chosen, the proposed estimator achieves optimal statistical rates of convergence up to logarithm terms under both the sparse and low-rank settings. This captures the implicit regularization phenomenon induced by our algorithm. Third, in order to corroborate our theories, we did extensive numerical studies. The experimental results support our theoretical findings and also show that our method outperforms classical regularized methods in terms of both ℓ_2-statistical rates and variable selection consistency.

1.1 Related Works

Our work belongs to the recent line of research on understanding the implicit regularization of gradient-based optimization methods in various statistical models. In addition, our work is also closely related to the large body of literature on single index models. Due to the space limit, we defer the discussions on related works to Appendix A in the supplement.

1.2 Notation

In this subsection, we give an introduction to our notations. Throughout this work, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. For a subset S in $[n]$ and a vector $\mathbf{u}$, we use $\mathbf{u}_S$ to denote the vector whose i-th entry is u_i if $i \in S$ and 0 otherwise. For any vector $\mathbf{u}$ and $q \geqslant 0$, we use $\|\mathbf{u}\|_{\ell_q}$ to represent the vector ℓ_q norm. In addition, the inner product $\langle \mathbf{u}, \mathbf{v} \rangle$ between any pair of vectors $\mathbf{u}$, $\mathbf{v}$ is defined as the Euclidean inner product $\mathbf{u}^\top \mathbf{v}$. Moreover, we define $\mathbf{u} \odot \mathbf{v}$ as the Hardmard product of vectors $\mathbf{u}$, $\mathbf{v}$. For any given matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, we use $\|\mathbf{X}\|_{\mathrm{op}}$, $\|\mathbf{X}\|_F$ and $\|\mathbf{X}\|_*$ to represent the operator norm, Frobenius norm and nuclear norm of matrix $\mathbf{X}$ respectively. In addition, for any two matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d_1 \times d_2}$, we define their inner product $\langle \mathbf{X}, \mathbf{Y} \rangle$ as $\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{tr}(\mathbf{X}^\top \mathbf{Y})$. Moreover, if we write $\mathbf{X} \succcurlyeq 0$ or $\mathbf{X} \preccurlyeq 0$, then the matrix $\mathbf{X}$ is meant to be positive semidefinite or negative semidefinite. We let $\{a_n, b_n\}_{n \geqslant 1}$ be any two positive series. We write $a_n \lesssim b_n$ if there exists a universal constant C such that $a_n \leqslant C \cdot b_n$ and we write $a_n \ll b_n$ if $a_n / b_n \to 0$. In addition, we write $a_n \asymp b_n$, if we have $a_n \lesssim b_n$ and $b_n \lesssim a_n$ and the notations of $a_n = \mathcal{O}(b_n)$ and $a_n = o(b_n)$ share the same meaning with $a_n \lesssim b_n$ and $a_n \ll b_n$. Moreover, $a_n = \widetilde{\mathcal{O}}(b_n)$ means $a_n \leqslant C b_n$ up to some logarithm terms. Finally, we use $a_n = \Omega(b_n)$ if there exists a universal constant $c > 0$ such that $a_n / b_n \geqslant c$ and we use $a_n = \Theta(b_n)$ if $c \leqslant a_n / b_n \leqslant C$ where $c, C > 0$ are universal constants.

1.3 Roadmap

The organization of our paper is as follows. We introduce the background knowledge in §2. In §3 and §4 we investigate the implicit regularization effect of gradient descent in over-parameterized SIM under the vector and matrix settings, respectively. Extensive simulation studies are presented in §B to corroborate our theory.

2 Preliminaries

In this section, we introduce the phenomenon of implicit regularization via over-parameterization, high dimensional single index model, and generalized Stein's identity [Stein et al., 2004].

2.1 Related Works on Implicit Regularization

Both Gunasekar et al. [2017] and Li et al. [2018] have studied least squares objectives over positive semidefinite matrices $\beta \in \mathbb{R}^{d \times d}$ of the following form

$$\min_{\beta \succcurlyeq 0} F(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \mathbf{X}_i, \beta \rangle)^2, \tag{1}$$

where the labels $\{y_i\}_{i=1}^n$ are generated from linear measurements $y_i = \langle \mathbf{X}_i, \beta^* \rangle$, $i \in [n]$, with $\beta^* \in \mathbb{R}^{d \times d}$ being positive semidefinite and low rank. Here β^* is of rank r where r is much smaller than d.

Instead of working on parameter β directly, they write β as $\beta = \mathbf{U}\mathbf{U}^\top$ where $\mathbf{U} \in \mathbb{R}^{d \times d}$, and study the optimization problem related to $\mathbf{U}$,

$$\min_{\mathbf{U} \in \mathbb{R}^{d \times d}} f(\mathbf{U}) = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \left\langle \mathbf{X}_i, \mathbf{U}\mathbf{U}^\top \right\rangle \right)^2. \tag{2}$$

The least-squares problem in (2) is over-parameterized because here β is parameterized by $\mathbf{U}$, which has d^2 degrees of freedom, whereas β^*, being a rank-r matrix, has $\mathcal{O}(rd)$ degrees of freedom. Gunasekar et al. [2017] proves that when $\{\mathbf{X}_i\}_{i=1}^m$ are commutative and $\mathbf{U}$ is properly initialized, if the gradient flow of (2) converges to a solution $\widehat{\mathbf{U}}$ such that $\hat{\beta} = \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top$ is a globally optimal solution of (1), then $\widehat{\mathbf{U}}$ has the minimum nuclear norm over all global optima. Namely,

$$\begin{aligned} &\hat{\beta} \in \operatorname{argmin} \|\beta\|_*, \\ &\quad \beta \succcurlyeq 0 \\ &\text{subject to } \left\langle \mathbf{X}_i, \hat{\beta} \right\rangle = y_i, \quad \forall i \in [n]. \end{aligned}$$

Subsequently, Li et al. [2018] assumes $\{\mathbf{X}_i\}_{i=1}^n$ satisfy the restricted isometry property (RIP) condition [Candés, 2008] and proves that by applying gradient descent to (2) with the initialization close to zero and sufficiently small fixed stepsizes, the near exact recovery of β^* is achieved.

Recently, Li et al. [2021] proves that the algorithm of gradient flow with infinitesimal initialization on the general covariate of (2) tends to be equivalent to the Greedy Low-Rank Learning (GLRL) algorithm, which is a greedy rank minimization algorithm. Results in Gunasekar et al. [2017] with commutable $\{\mathbf{X}_i\}_{i=1}^m$ serves as a special case to Li et al. [2021].

As for noisy statistical model, both Zhao et al. [2019] and Vaškevičius et al. [2019] study over-parameterized high dimensional noisy linear regression problem independently. Specifically, here the response variables $\{y_i\}_{i=1}^n$ are generated from a linear model

$$y_i = \mathbf{x}_i^\top \beta^* + \epsilon_i, i \in [n], \tag{3}$$

where $\beta^* \in \mathbb{R}^p$ and $\{\epsilon_i\}_{i=1}^n$ are i.i.d. sub-Gaussian random variables that are independent with the covariates $\{\mathbf{x}_i\}_{i=1}^n$. Moreover, here β^* has only s nonzero entries where $s \ll p$. Instead of adding sparsity-enforcing penalties, they propose to estimate β^* via gradient descent with respect to $\mathbf{w}$, $\mathbf{v}$ on a loss function L,

$$\min_{\mathbf{w} \in \mathbb{R}^p, \mathbf{v} \in \mathbb{R}^p} L(\mathbf{w}, \mathbf{v}) = \frac{1}{2n} \sum_{i=1}^{n} \left[\mathbf{x}_i^\top (\mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v}) - y_i \right]^2,$$

(4)

where the parameter β is over-parameterized as $\beta = \mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v}$. Under the restricted isometry property (RIP) condition on the covariates, these works prove that, when the hyperparameters is proper selected, gradient descent on (4) finds an estimator of β^* with optimal statistical rate of convergence.

2.2 High Dimensional Single Index Model

In this subsection, we first introduce the score functions associated with random vectors and matrices, which are utilized in our algorithms. Then we formally define the high dimensional single index model (SIM) in both the vector and matrix settings.

Definition 1. *Let* $\mathbf{x} \in \mathbb{R}^p$ *be a random vector with density function* $p_0(\mathbf{x}): \mathbb{R}^p \to \mathbb{R}$. *The score function* $S_{p_0}(\mathbf{x}): \mathbb{R}^p \to \mathbb{R}^p$ *associated with* $\mathbf{x}$ *is defined as*

$$S_{p_0}(\mathbf{x}) := -\nabla_{\mathbf{x}} \log p_0(\mathbf{x}) = -\nabla_{\mathbf{x}} p_0(\mathbf{x}) / p_0(\mathbf{x}).$$

Here the score function $S_{p_0}(\mathbf{x})$ relies on the density function $p_0(\mathbf{x})$ of the covariate $\mathbf{x}$. In order to simplify the notations, in the rest of the paper, we omit the subscript p_0 from S_{p_0} when the underlying distribution of $\mathbf{x}$ is clear to us.

Remark: If the covariate $\mathbf{X} \in \mathbb{R}^{d \times d}$ is a random matrix whose entries are i.i.d. with a univariate density $p_0(x): \mathbb{R} \to \mathbb{R}$, we then define the score function $S(\mathbf{X}) \in \mathbb{R}^{d \times d}$ entrywisely. In other words, for any $\{i, j\} \in [d] \times [d]$, we obtain

$$S(\mathbf{X})_{i,j} := -p_0'(\mathbf{X}_{i,j}) / p_0(\mathbf{X}_{i,j}).$$

(5)

Next, we introduce the first-order general Stein's identity.

Lemma 1. *(First-Order General Stein's Identity*, [Stein et al., 2004]*) We assume that the covariate* $\mathbf{x} \in \mathbb{R}^p$ *follows a distribution with density function* $p_0(\mathbf{x}): \mathbb{R}^p \to \mathbb{R}$ *which is differentiable and satisfies the condition that* $|p_0(\mathbf{x})|$ *converges to zero as* $\|\mathbf{x}\|_2$ *goes to infinity. Then for any differentiable function* $f(\mathbf{x})$ *with* $\mathbb{E}[|f(\mathbf{x})S(\mathbf{x})|] \vee \mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x})\|_2] < \infty$, *it holds that,*

$$\mathbb{E}[f(\mathbf{x})S(\mathbf{x})] = \mathbb{E}[\nabla_{\mathbf{x}} f(\mathbf{x})],$$

where $S(\mathbf{x}) = -\nabla_{\mathbf{x}} p_0(\mathbf{x}) / p_0(\mathbf{x})$ *is the score function with respect to* $\mathbf{x}$ *defined in* Definition 1.

Remark: In the case of having matrix covariate, we are able to achieve the same conclusion by simply replacing $\mathbf{x} \in \mathbb{R}^p$ by $\mathbf{X} \in \mathbb{R}^{d \times d}$ in Lemma 1 with the definition of matrix score function in (5).

In the sequel, we introduce the single index models considered in this work. We first define sparse vector single index models as follows.

Definition 2. *(Sparse Vector SIM) We assume the response* $Y \in \mathbb{R}$ *is generated from model*

$$Y = f(\langle \mathbf{x}, \beta^* \rangle) + \epsilon, \tag{6}$$

with unknown link $f : \mathbb{R} \to \mathbb{R}$, *p-dimensional covariate* $\mathbf{x}$ *as well as signal* β^* *which is the parameter of interest. Here, we let* $\epsilon \in \mathbb{R}$ *be an exogenous random noise with mean zero. In addition, if not particularly indicated, we assume the entries of* $\mathbf{x}$ *are i.i.d. random variables with a known univariate density* $p_0(x)$. *As for the underlying true signal* β^*, *it is assumed to be s-sparse with* $s \ll p$. *Note that the length of* β^* *can be absorbed by the unknown link f, we then let* $\|\beta^*\|_2 = 1$ *for model identifiability.*

By the definition of sparse vector SIM, we notice that many well-known models are included in this category, such as linear regression $y_i = \mathbf{x}_i^\top \beta^* + \epsilon$, phase retrieval $y_i = (\mathbf{x}_i^\top \beta^*)^2 + \epsilon$, as well as one-bit compressed sensing $y = \mathrm{sign}(\mathbf{x}_i^\top \beta^*) + \epsilon$.

Finally, we define the low rank matrix SIM as follows.

Definition 3. *(Symmetric Low Rank Matrix SIM) For the low rank matrix SIM, we assume the response* $Y \in \mathbb{R}$ *is generated from*

$$Y = f(\langle \mathbf{X}, \beta^* \rangle) + \epsilon, \tag{7}$$

in which $\beta^* \in \mathbb{R}^{d \times d}$ *is a low rank symmetric matrix with rank* $r \ll d$ *and the link function f is unknown. For the covariate* $\mathbf{X} \in \mathbb{R}^{d \times d}$, *we assume the entries of* $\mathbf{X}$ *are i.i.d. with a known density* $p_0(x)$. *Besides, since* $\|\beta^*\|_F$ *can be absorbed in the unknown link function f, we further assume* $\|\beta^*\|_F = 1$ *for model identifiability. In addition, the noise term* ϵ *is also assumed additive and mean zero.*

As we have discussed in the introduction, almost all existing literature designs algorithms and studies the corresponding implicit regularization phenomenon in linear models with sub-Gaussian data. The scope of this work is to leverage over-parameterization to design regularization-free algorithms and delineate the induced implicit regularization phenomenon for a more general class of statistics models with possibly heavy-tailed data. Specifically, in §3 and §4, we design algorithms and capture the implicit regularization induced by the gradient descent algorithm for over-parameterized vector and matrix SIMs, respectively.

3 Main Results for Over-Parameterized Vector SIM

Leveraging our conclusion from Lemma 1 as well as our definition of sparse vector SIM in Definitions 2, we have

$$\mathbb{E}[Y \cdot S(\mathbf{x})] = \mathbb{E}[f(\langle \mathbf{x}, \beta^* \rangle) \cdot S(\mathbf{x})] = \mathbb{E}[f'(\langle \mathbf{x}, \beta^* \rangle)] \cdot \beta^* := \mu^* \beta^*,$$

which recovers our true signal β^* up to scaling. Here we define $\mu^* = \mathbb{E}[f'(\langle \mathbf{x}, \beta^* \rangle)]$, which is assumed nonzero throughout this work. Hence, $Y \cdot S(\mathbf{x})$ serves as an unbiased estimator of $\mu^* \beta^*$, and we can correctly identify the direction of β^* by solving a population level optimization problem:

$$\min_{\beta} L(\beta) := \langle \beta, \beta \rangle - 2\langle \beta, \mathbb{E}[Y \cdot S(\mathbf{x})] \rangle .$$

Since we only have access to finite data, we replace $\mathbb{E}[Y \cdot S(\mathbf{x})]$ by its sample version estimator $\frac{1}{n}\sum_{i=1}^{n} y_i S(\mathbf{x}_i)$, and plug the sample-based estimator into the loss function. In a high dimensional SIM given in Definition 2, where the true signal β^* is assumed to be sparse, various works [Plan and Vershynin, 2016, Plan et al., 2017, Yang et al., 2017] have shown that the ℓ_1-regularized estimator $\hat{\beta}$ given by

$$\hat{\beta} \in \operatorname*{argmin}_{\beta} L(\beta) := \langle \beta, \beta \rangle - 2\left\langle \beta, \frac{1}{n}\sum_{i=1}^{n} y_i S(\mathbf{x}_i) \right\rangle + \lambda \|\beta\|_1 \tag{8}$$

attains the optimal statistical rate of convergence rate to $\mu^* \beta^*$.

In contrast, instead of imposing an ℓ_1-norm regularization term, we propose to obtain an estimator by minimizing the loss function L directly, with β re-parameterized using two vectors $\mathbf{w}$ and $\mathbf{v}$ in $\mathbb{R}^p$. Specifically, we write β as $\beta = \mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v}$ and thus equivalently write the loss function $L(\beta)$ as $L(\mathbf{w}, \mathbf{v})$, which is given by

$$L(\mathbf{w}, \mathbf{v}) = \langle \mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v}, \mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v} \rangle - 2\left\langle \mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v}, \frac{1}{n}\sum_{i=1}^{n} y_i S(\mathbf{x}_i) \right\rangle . \tag{9}$$

Note that the way of writing β in terms of $\mathbf{w}$ and $\mathbf{v}$ is not unique. In particular, β has p degrees of freedom but we use $2p$ parameters to represent β. Thus, by using $\mathbf{w}$ and $\mathbf{v}$ instead of β, we employ over-parameterization in (9).

We briefly describe our motivation on over-parameterizing β by $\mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v}$. Suppose that β is sparse, an explicit regularization is to use ℓ_1-penalty. Note that $\|\beta\|_1 = \min_{\gamma \odot \delta = \beta} \{\|\gamma\|^2 + \|\delta\|^2\}/2$, where $\odot$ denotes the Hadamard (componentwise) product. Thus, an explicit regularization is to $\min_{\gamma, \delta} \sum_{i=1}^{n} \{Y_i - f(\mathbf{x}_i^T \gamma \odot \delta)\}^2 + \lambda\left\{\|\gamma\|^2 + \|\delta\|^2\right\}$ for a penalty parameter λ, following the method in Hoff [2017]. To gain understanding on implicit regularization by over parametrization, we let $\mathbf{w} = (\gamma + \delta)/2$ and $\mathbf{v} = (\gamma - \delta)/2$. Then $\beta = \gamma \odot \delta = \mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v}$ with $2p$ new parameters $\mathbf{w}$ and $\mathbf{v}$ that over parameterize the problem.

This leads to the empirical loss $L(\mathbf{w}, \mathbf{v}) = \sum_{i=1}^{n} \{Y_i - f(\mathbf{x}_i^T(\mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v}))\}^2$. Following the neural network training, we drop the explicit penalty and run the gradient decent to minimize $L(\mathbf{w}, \mathbf{v})$.

To be more specific, for the sparse SIM, we propose to construct an estimator of β^* by applying gradient descent to L in (9) with respect to $\mathbf{w}$ and $\mathbf{v}$, without any explicit regularization. Such an estimator, if achieves desired statistical accuracy, demonstrates the efficacy of implicit regularization of gradient descent in over-parameterized sparse SIM. Specifically, the gradient updates for the vector $(\mathbf{w}^\top, \mathbf{v}^\top)^\top$ for solving (9) are given by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_w L(\mathbf{w}_t, \mathbf{v}_t) = \mathbf{w}_t - \eta \left(\mathbf{w}_t \odot \mathbf{w}_t - \mathbf{v}_t \odot \mathbf{v}_t - \frac{1}{n} \sum_{i=1}^{n} S(\mathbf{x}_i) y_i \right) \odot \mathbf{w}_t, \tag{10}$$

$$\mathbf{v}_{t+1} = \mathbf{v}_t + \eta \nabla_v L(\mathbf{w}_t, \mathbf{v}_t) = \mathbf{v}_t + \eta \left(\mathbf{w}_t \odot \mathbf{w}_t - \mathbf{v}_t \odot \mathbf{v}_t - \frac{1}{n} \sum_{i=1}^{n} S(\mathbf{x}_i) y_i \right) \odot \mathbf{v}_t. \tag{11}$$

Here $\eta > 0$ is the stepsize. By the parameterization of β, $\{\mathbf{w}_t, \mathbf{v}_t\}_{t \geq 0}$ leads to a sequence of estimators $\{\beta_t\}_{t \geq 0}$ given by

$$\beta_{t+1} = \mathbf{w}_{t+1} \odot \mathbf{w}_{t+1} - \mathbf{v}_{t+1} \odot \mathbf{v}_{t+1}. \tag{12}$$

Meanwhile, in terms of chooisng initial values, since the zero vector is a stationary point of the algorithm, we cannot set the initial values of $\mathbf{w}$ and $\mathbf{v}$ to the zero vector. To utilize the structure of β^*, ideally we would like to initialize $\mathbf{w}$ and $\mathbf{v}$ such that they share the same sparsity pattern as β^*. That is, we would like to set the entries in the support of β^* to nonzero values, and set those outside of the support to zero. However, such an initialization scheme is infeasible since the support of β^* is unknown. Instead, we initialize $\mathbf{w}_0$ and $\mathbf{v}_0$ as $\mathbf{w}_0 = \mathbf{v}_0 = \alpha \cdot \mathbf{1}_{p \times 1}$, where $\alpha > 0$ is a small constant and $\mathbf{1}_{p \times 1}$ is an all-one vector in $\mathbb{R}^p$. By setting $\mathbf{w}_0 = \mathbf{v}_0$, we equivalently set β_0 to the zero vector. And more importantly, such a construction provides a good compromise: zero components get nearly zero initializations, which are the majority under the sparsity assumption, and nonzero components get nonzero initializations. Even though we initialize every component at the same value, the nonzero components move quickly to their stationary component, while zero components remain small. This is how over-parameterization differentiate active components from inactive components. We illustrate this by a simulation experiment.

A simulation study.

In this simulation, we fix sample size $n = 1000$, dimension $p = 2000$, number of non-zero entries $s = 5$. Let $S := \{i : |\beta_i^*| > 0\}$. The responses $\{y_i\}_{i=1}^n$ are generated from $y_i = f(\langle \mathbf{x}, \beta^* \rangle) +$

ϵ_i, $i \in [n]$ with link functions $f_1(x) = x$ (linear regression) and $f_2(x) = \sin(x)$. Here we assume β^* is s-sparse with $\beta_i = 1/\sqrt{s}$, $i \in S$, and $\{\mathbf{x}_i\}_{i=1}^n$ are standard Gaussian random vectors. We over-parameterize β as $\mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v}$ and set $\mathbf{w}_0 = \mathbf{v}_0 = 10^{-5} \cdot \mathbf{1}_{p \times 1}$. Then we update $\mathbf{w}$, $\mathbf{v}$ and β regarding equations (10), (11), and (12) with stepsize $\eta = 0.01$. The evolution of the distance between our unnormalized iterates β_t and $\mu^*\beta^*$, trajectories of $\beta_{j,t}$ for $j \in S$ and $\max_{j \in S^c} |\beta_{j,t}|$ are depicted in Figures 1 and 2.

From the simulation results given in Figure 1-(a) and Figure 2-(a), we notice that there exists a time interval, where we can nearly recover $\mu^*\beta^*$. From plots (b) in Figures 1 and 2, we can see with over-parameterization, five nonzero components all increase rapidly and converge quickly to their stationary points. Meanwhile, the maximum estimation error for inactive component, represented by $\|\beta_{S^c,t}\|_\infty$, still remains small, as shown in Figure 1-(c) and Figure 2-(c). In other words, running gradient descent with respect to over-parameterized parameters helps us distinguish non-zero components from zero components, while applying gradient descent to the ordinary loss can not.

It is worth noting that, with over-parameterization, there are $\Omega(2^p)$ stationary points of L satisfying $\nabla_{\mathbf{w}} L(\mathbf{w}, \mathbf{v}) = \nabla_{\mathbf{v}} L(\mathbf{w}, \mathbf{v}) = \mathbf{0}_{p \times 1}$, where $\mathbf{0}_{p \times 1}$ is the zero vector. To see this, for any subset $I \subseteq [p]$, we define vectors $\overline{\mathbf{w}}$ and $\overline{\mathbf{v}}$ as follows. For any $j \notin I$, we set the j-th entries of $\overline{\mathbf{w}}$ and $\overline{\mathbf{v}}$ to zero. Meanwhile, for any $j \in I$, we choose $\overline{\mathbf{w}}_j$ and $\overline{\mathbf{v}}_j$ such that $\overline{\mathbf{w}}_j^2 - \overline{\mathbf{v}}_j^2 = n^{-1} \sum_{i=1}^n S(\mathbf{x}_i)_j y_i$, where $\overline{\mathbf{w}}_j$, $\overline{\mathbf{v}}_j$, and $S(\mathbf{x}_i)_j$ are the j-th entries of $\overline{\mathbf{w}}$, $\overline{\mathbf{v}}$, and $S(\mathbf{x}_i)$, respectively. By direct computation, it can be shown that $(\overline{\mathbf{w}}, \overline{\mathbf{v}})$ is a stationary point of L, and thus there are at least 2^p stationary points. However, our numerical results demonstrate that not all of these stationary points are likely to be found by the gradient descent algorithm — gradient descent favors the stationary points that correctly recover $\mu^*\beta^*$. Such an intriguing observation captures the implicit regularization induced by the optimization algorithm and over-parameterization.

3.1 Gaussian Design

In this subsection, we discuss over-parameterized SIM with Gaussian covariates. In this subsection, we assume the distribution of $\mathbf{x}$ in (6) is $N(\mu, \Sigma)$, where both μ and Σ are assumed known. Moreover, only in this subsection, we slightly modify the identifiability condition in Definition 2 from assuming $\|\beta^*\|_2 = 1$ to $\|\Sigma^{1/2}\beta^*\|_2 = 1$.

3.1.1 Theoretical Results for Gaussian Covariates—

We first introduce an structural assumption on the SIM.

Assumption 1. *Assume that* $\mu^* = \mathbb{E}[f'(\langle \mathbf{x}, \beta^* \rangle)] \neq 0$ *is a constant and the following two conditions hold.*

- **a.** *Covariance matrix* Σ *is positive-definite and has bounded spectral norm. To be more specific, there exist constants* $C_{\min}$ *and* $C_{\max}$ *such that* $C_{\min} \mathbb{I}_{p \times p} \preccurlyeq \Sigma \preccurlyeq C_{\max} \mathbb{I}_{p \times p}$ *holds, where* $\mathbb{I}_{p \times p}$ *is the identity matrix.*
- **b.** *Both* $\{f(\langle \mathbf{x}_i, \beta^* \rangle)\}_{i=1}^n$ *and* $\{\epsilon_i\}_{i=1}^n$ *are i.i.d. sub-Gaussian random variables, with the sub-Gaussian norms denoted by* $\|f\|_{\psi_2} = \mathcal{O}(1)$ *and* $\sigma = \mathcal{O}(1)$ *respectively. Here we*

let $\|f\|_{\psi_2}$ *denote the sub-Gaussian norm of* $f(\langle \mathbf{x}_i, \beta^* \rangle)$. *In addition, we further assume that* $|\mu^*|/\|f\|_{\psi_2} = \Theta(1)$, $|\mu^*|/\sigma = \Omega(1)$.

The score function for the Gaussian distribution $N(\mu, \Sigma)$ is $S(\mathbf{x}) = \Sigma^{-1}(\mathbf{x} - \mu)$ and Assumption 1-(a) makes the Gaussian distributed covariates non-degenerate. Assumption 1-(b) enables the the empirical estimator $n^{-1}\sum_{i=1}^{n} y_i S(\mathbf{x}_i)$ to concentrate to its expectation $\mu^*\beta^*$, and also sets a lower bound to the signal noise ratio $|\mu^*|/\sigma$. Note that this assumption is quite standard and easy to be satisfied by a broad class of models as long as there exists a lower bound on the signal noise ratio, which include models with link functions $f(x) = x$, $\sin x$, $\tanh(x)$, and etc. In addition, in §3.2, the assumption that both $f(\langle x, \beta^* \rangle)$ and the noise ϵ are sub-Gaussian random variables will be further relaxed to simply assuming they have bounded finite moments with perhaps heavy-tailed distributions.

We present the details of the proposed method for the Gaussian case in Algorithm 1. In the following, we present the statistical rates of convergence for the estimator constructed by Algorithm 1. Let us divide the support seta $S = \{i : |\beta_i^*| > 0\}$ into $S_0 = \{i : |\beta_i| \geq C_s\sqrt{\log p/n}\}$ and $S_1 = \{i : 0 < |\beta^*| < C_s\sqrt{\log p/n}\}$, which correspond to the sets of strong and weak signals, respectively. Here C_s is an absolute constant. We let s_0 and s_1 be the cardinalities of S_0 and S_1, respectively. In addition, we let $s_m = \min_{i \in S_0}|\beta_i^*|$ be the smallest value of strong signals.

Theorem 1. *Apart from* Assumption 1, *if we further let our initial value* α *satisfy* $0 < \alpha \leqslant M_0^2/p$ *and set stepsize* η *as* $0 < \eta \leqslant 1/(12(|\mu^*| + M_0))$ *in* Algorithm 1 *with* M_0 *being a constant proportional to* $\max\{\|f\|_{\psi_2}, \sigma\}$, *there exist absolute constants* a_1, $a_2 > 0$ *such that, with probability at least* $1 - 2p^{-1} - 2n^{-2}$, *we have*

$$\|\beta_{T_1} - \mu^*\beta^*\|_2^2 \lesssim \frac{s_0 \log n}{n} + \frac{s_1 \log p}{n},$$

for all $T_1 \in \left[a_1 \log(1/\alpha)/\left(\eta\left(|\mu^*|s_m - M_0\sqrt{\log p/n}\right)\right), a_2 \log(1/\alpha)\sqrt{n/\log p}/(\eta M_0)\right]$. *Meanwhile, the statistical rate of convergence for the normalized iterates are given by*

$$\left\| \frac{\beta_{T_1}}{\left\|\Sigma^{1/2}\beta_{T_1}\right\|_2} - \frac{\mu^*\beta^*}{|\mu^*|} \right\|_2^2 \lesssim \frac{s_0 \log n}{n} + \frac{s_1 \log p}{n}.$$

Algorithm 1: Algorithm for Vector SIM with Gaussian Design

Data: Training covariates $\{\mathbf{x}_i\}_{i=1}^n$, response variables $\{y_i\}_{i=1}^n$, initial value α, step size η;

Initialize variables $\mathbf{w}_0 = \alpha \cdot \mathbf{1}_{p\times 1}$, $\mathbf{v}_0 = \alpha \cdot \mathbf{1}_{p\times 1}$ and set iteration number $t = 0$;

while $t < T_1$ **do**

$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\left[\mathbf{w}_t \odot \mathbf{w}_t - \mathbf{v}_t \odot \mathbf{v}_t - \frac{1}{n}\sum_{i=1}^n \Sigma^{-1}(\mathbf{x}_i - \mu)y_i\right] \odot \mathbf{w}_t$;

$\mathbf{v}_{t+1} = \mathbf{v}_t + \eta\left[\mathbf{w}_t \odot \mathbf{w}_t - \mathbf{v}_t \odot \mathbf{v}_t - \frac{1}{n}\sum_{i=1}^n \Sigma^{-1}(\mathbf{x}_i - \mu)y_i\right] \odot \mathbf{v}_t$;

$\beta_{t+1} = \mathbf{w}_t \odot \mathbf{w}_t - \mathbf{v}_t \odot \mathbf{v}_t$;

$t = t + 1$;

end

Result: Output the final estimate $\widehat{\beta}^* = \beta_{T_1}$.

Theorem 2. *(Variable Selection Consistency) Under the setting of* Theorem 1, *for all*

$$T_1 \in \left[a_1 \log(1/\alpha)/\left(\eta\left(|\mu^*|s_m - M_0\sqrt{\log p/n}\right)\right), a_2 \log(1/\alpha)\sqrt{n/\log p}/(\eta M_0)\right],$$

we let $\left[\tilde{\beta}_{T_1}\right]_i = [\beta_{T_1}]_i \cdot \mathbb{I}_{\{|\beta_{T_1}|_i \geq \lambda\}}$, *for all* $i \in [p]$. *Then, with probability at least* $1 - 2p^{-1} - 2n^{-2}$, *for all* $\lambda \in \left[\alpha, (C_s|\mu^*| - 2M_0)\sqrt{\log p/n}\right]$, *we have* $supp(\tilde{\beta}_{T_1}) \subset supp(\beta^*)$. *Moreover, when there only exists strong signals in* S_0. *We further have* $supp(\tilde{\beta}_{T_1}) = supp(\beta^*)$ *and* $sign(\tilde{\beta}_{T_1}) = sign(\beta^*)$.

Theorem 1 shows that if we just have strong signals, then with high probability, for any $T_1 \in \left[a_1 \log(1/\alpha)/\left(\eta\left(|\mu^*|s_m - M_0\sqrt{\log p/n}\right)\right), a_2 \log(1/\alpha)\sqrt{n/\log p}/(\eta M_0)\right]$, we get the oracle statistical rate $\mathcal{O}(\sqrt{s \log n/n})$ in terms of the ℓ_2-norm, which is independent of the ambient dimension p. Besides, when β^* also consists of weak signals, we achieve $\mathcal{O}(\sqrt{s \log p/n})$ statistical rate in terms of the ℓ_2-norm, where s is the sparsity of β^*. Such a statistical rate matches the minimax rate of sparse linear regression [Raskutti et al., 2011] and is thus minimax optimal. Notice that the oracle rate is achievable via explicit regularization using folded concave penalties [Fan et al., 2014] such as SCAD [Fan and Li, 2001] and MCP [Zhang et al., 2010]. Thus, Theorem 1 shows that, with over-parameterization, the implicit regularization of gradient descent has the same effect as adding a folded concave penalty function to the loss function in (9) explicitly.

Furthermore, comparing our work to Plan and Vershynin [2016], Plan et al. [2017], which study high dimensional SIM with ℓ_1-regularization, thanks to the implicit regularization phenomenon, we avoid bias brought by the ℓ_1-penalty and attain the oracle statistical rate. Moreover, our another advantage over regularized methods is shown in Theorem 2. It shows that by properly truncating β_{T_1} when T_1 falls in the optimal time interval, we are able to recover the support of β^* with high probability. Comparing to existing literatures on support recovery via using explicit regularization on single index model [Neykov et al., 2016], our method offers a wider range for choosing tuning parameter λ with a known left boundary α, instead of only using $\lambda = \Theta(\sqrt{\log p/n})$. This efficiently reduces false discovery rate, see §D.1 for more details. Last but not least, as we only need to run gradient descent, comparing to regularized methods, it is easier to parallel our algorithm since the gradient information is able to be efficiently transferred among different machines. The use of implicit regularization allows our methodology to be generalized to large-scale problems easily [McMahan et al., 2017, Richards and Rebeschini, 2020, Richards et al., 2020]. The detailed discussions are given in §C.5.

Theorem 1 and Theorem 2 generalizes the results in Zhao et al. [2019] and Vaškevičius et al. [2019] for the linear model to high-dimensional SIMs. In addition, to satisfy the RIP condition, their sample complexity is at least $\mathcal{O}(s^2 \log p)$ if their covariate $\mathbf{x}$ follows the Gaussian distribution. Whereas, by using the loss function in (9) motivated by the Stein's identity [Stein et al., 1972, 2004], the RIP condition is unnecessary in our analysis. Instead, our theory only requires that $n^{-1}\sum_{i=1}^{n} S(\mathbf{x}_i) \cdot y_i$ concentrates at a fast rate. As a result, our sample complexity is $\mathcal{O}(s \log p)$ for ℓ_2-norm consistency, which is better than $\mathcal{O}(s^2 \log p)$.

The ideas of proof behind Theorem 1 and Theorem 2 are as follows. First, we are able to control the strengths of error component, denoted by $\|\beta_t \odot \mathbf{1}_{S^c}\|_\infty$, at the same order with the square root of their initial values until $\mathcal{O}\left(\log(1/\alpha) \cdot \sqrt{n/\log p}/(\eta M_0)\right)$ steps. This gives us the right boundary of the stopping time T_1. Meanwhile, every entry of strong signal

part $\beta_t \odot \mathbf{1}_{s_0}$ grows at exponential rates to $\epsilon = \mathcal{O}(\sqrt{\log n/n})$ accuracy around $\mu^*\beta^* \odot \mathbf{1}_{s_0}$ within $\mathcal{O}\big(\log(1/\alpha)/\big(\eta\big(|\mu^*|s_m - M_0\sqrt{\log p/n}\big)\big)\big)$ steps, which offers us the left boundary of the stopping time T_1. Finally, we prove for weak signals, their strengths will not exceed $\mathcal{O}(\sqrt{\log p/n})$ for all steps as long as we properly choose the stepsize. Thus, by letting the stopping time T_1 be in the interval given in Theorem 1, we obtain converged signal component and well controlled error component. The final statistical rates are obtained by combining the results on the active and inactive components together. Moreover, the conclusion of Theorem 2 holds by truncating the β_t properly, since we are able to control the error component of β_t uniformly as mentioned above. See Appendix §E.1 for the detail. As shown in the proof, we observe that with small initialization and over-parameterized loss function, the signal component converges rapidly to the true signal, while the the error component grows in a relatively slow pace. Thus, gradient descent rapidly isolates the signal components from the noise, and with a proper stopping time, finds a near-sparse solution with high statistical accuracy. Thus, with proper initialization, over-parameterization plays the role of an implicit regularization by favoring approximately sparse saddle points of the loss function in (9).

Finally, we remark that Theorem 1 establishes optimal statistical rates for the estimator β_{T_1}, where T_1 is any stopping time that belongs to the interval given in Theorem 1. However, in practice, such an interval is infeasible to compute as it depends on unknown constants. To make the proposed method practical, in the following, we introduce a method for selecting a proper stopping time T_1.

3.1.2 Choosing the Stopping Time T_1—We split the dataset into training data and testing data. We utilize the training data to implement Algorithm 1 and get the estimator β_t as well as the value of the training loss (9) at step t. We notice β_t varies slowly inside the optimal time interval specified in Theorem 1, so that the fluctuation of the training loss (9) can be smaller than a threshold. Based on that, we choose m testing points on the flatted curve of the training loss (9) and denote their corresponding number of iterations as $\{t_j\}, j \in [m]$. For each $j \in [m]$, we then reuse the training data and normalized estimator $\beta_{t_j}/\left\|\Sigma^{1/2}\beta_{t_j}\right\|_2, j \in [m]$ to fit the link function f. Let the obtained estimator be $\hat{f}_j$. For the testing dataset, we perform out-of-sample prediction and get m prediction losses:

$$l_j = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left[Y_i - \hat{f}_j\left(\left\langle \mathbf{x}_i, \beta_{t_j}/\left\|\Sigma^{1/2}\beta_{t_j}\right\|_2 \right\rangle\right)\right]^2, \quad \forall j \in [m].$$

Next, we choose T_1 as t_{j^*} where we define $j^* = \operatorname{argmin}_{j \in [m]} l_j$.

We remark that each $\hat{f}_j$ can be obtained by any nonparametric regression methods. To show case our method, in the following, we apply univariate kernel regression to obtain each $\hat{f}_j$ and establish its theoretical guarantee.

3.1.3 Prediction Risk—We now consider estimating the nonparametric component and the prediction risk. Suppose we are given an estimator $\hat{\beta}$ of β and n i.i.d. observations $\{y_i, \mathbf{x}_i\}_{i=1}^n$ of the model. For simplicity of the technical analysis, we assume that $\hat{\beta}$ is

independent of $\{y_i, \mathbf{x}_i\}_{i=1}^n$, which can be achieved by data-splitting. Moreover, we assume that $\hat{\beta}$ is an estimator of β^* such that

$$\|\hat{\beta} - \beta^*\|_2 = o\left(n^{-1/3}\right), \quad \|\Sigma^{1/2}\hat{\beta}\|_2 = 1, \quad \text{and} \quad \|\Sigma^{1/2}\beta^*\|_2 = 1. \tag{13}$$

Our goal is to construct an estimate the regression function $f(\langle \cdot, \beta^* \rangle)$ based on $\hat{\beta}$ and $\{y_i, \mathbf{x}_i\}_{i=1}^n$.

Note that, when β^* is known, we can directly estimate f based on y_i and $Z_i^* := \mathbf{x}_i^\top \beta^*$, $i \in [n]$ via standard non-parametric regression. When $\hat{\beta}$ is accurate, a direct idea is to replace Z_i^* by $Z_i := \mathbf{x}_i^\top \hat{\beta}$ and follow the similar route. For a new observation $\mathbf{x}$, we define Z as $Z := \mathbf{x}^\top \hat{\beta}$ and Z^* as $Z^* := \mathbf{x}^\top \beta^*$ respectively.

To predict Y, we estimate function $g(z)$ using kernel regression with data $\left\{\left(y_i, \mathbf{x}_i^\top \hat{\beta}\right)\right\}_{i=1}^n$. Specifically, we let the function $K_h(u)$ be $K_h(u) := 1/h \cdot K(u/h)$, in which $K: \mathbb{R} \to \mathbb{R}$ is a kernel function with $K(u) = \mathbb{I}_{\{|u| \leq 1\}}$ and h is a bandwidth. By the definitions of Z^*, Z, and Z_i, $i \in [n]$ given above, the prediction function $\hat{g}(Z)$ is defined as

$$\hat{g}(Z) = \begin{cases} \frac{\sum_{i=1}^n y_i K_h(Z - Z_i)}{\sum_{i=1}^n K_h(Z - Z_i)}, & \left|Z - \mu^\top \hat{\beta}\right| \leq R, \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

where we follow the convention that 0/0 = 0. In what follows, we consider the ℓ_2-prediction risk of $\hat{g}$, which is given by

$$\mathbb{E}\left[\left\{\hat{g}\left(\langle \mathbf{x}, \hat{\beta} \rangle\right) - f(\langle \mathbf{x}, \beta^* \rangle)\right\}^2\right],$$

where the expectation is taken with respect to $\mathbf{x}$ and $\{\mathbf{x}_i, y_i\}_{i=1}^n$. Before proceeding to the theoretical guarantees, we make the following assumption on the regularity of f.

Assumption 2. *There exists an* $\alpha_1 > 0$ *and a constant* $C > 0$ *such that* $|f(x)|$, $|f'(x)| \leq C + |x|^{\alpha_1}$.

For the rationality of the Assumption 2, we note that the constraint on $f'(x)$ and $f(x)$ given above is weaker than assuming $f'(x)$ and $f(x)$ are bounded functions directly. Next, we present Theorem 3 which characterizes the convergence rate of mean integrated error of our prediction function $\hat{g}(Z)$.

Theorem 3. *If we set* $R = 2\sqrt{\log(n)}$ *and* $h \asymp n^{-1/3}$ *in* (14), *under* Assumption 2, *the* ℓ_2-*prediction risk of* $\hat{g}$ *defined in* (14) *is given by*

$$\mathbb{E}\left[\left\{\hat{g}\left(\left\langle \mathbf{x}, \hat{\beta}\right\rangle\right) - f(\langle \mathbf{x}, \beta^* \rangle)\right\}^2\right] \lesssim \frac{polylog(n)}{n^{2/3}},$$

where $\hat{\beta}$ *is any vector that satisfies* (13) *and* $polylog(n)$ *contains terms that are polynomials of* $\log n$.

It is worth noting that the estimator $\hat{\beta} = \beta_{T_1} / \left\| \Sigma^{1/2} \beta_{T_1} \right\|_2$ constructed in Theorem 1 with any T_1 belongs to the optimal time interval given in Theorem 1 satisfy (13). Thus, under such regimes, Theorem 3 also holds. The proof of Theorem 3 is given in §E.3. Note that it is possible to refine the analysis on the prediction risk for f with higher order derivatives by utilizing higher order kernels (see Tsybakov [2008] therein) this is not the key message of our paper.

3.2 General Design

In this subsection, we extend our methodology to the setting with covariates generated from a general distribution. Following our discussions at the beginning of §3, ideally we aim at solving the loss function with over-parameterized variable given in (9). However, when the distribution of $\mathbf{x}$ has density p_0, the score $S(\mathbf{x})$ can be heavy-tailed such that $\mathbb{E}[Y \cdot S(\mathbf{x})]$ and its empirical counterpart may not be sufficiently close.

To remedy this issue, we modify the loss function in (9) by replacing y_i and $S(\mathbf{x}_i)$ by their truncated (Winsorized) version $\check{y}_i$ and $\check{S}$, respectively. Specifically, we propose to apply gradient descent to the following modified loss function with respect to $\mathbf{u}$ and $\mathbf{v}$:

$$\min_{\mathbf{w},\mathbf{v}} L(\mathbf{w}, \mathbf{v}) := \langle \mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v}, \mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v} \rangle - \frac{2}{n} \sum_{i=1}^{n} \check{y}_i \left\langle \mathbf{w} \odot \mathbf{w} - \mathbf{v} \odot \mathbf{v}, \check{S}(\mathbf{x}_i) \right\rangle. \tag{15}$$

Let $\check{\mathbf{a}} \in \mathbb{R}^d$ denote the truncated version of vector $\mathbf{a} \in \mathbb{R}^d$ based on a parameter τ [Fan et al., 2021b]. That is, its entries are given by $[\check{\mathbf{a}}]_j = [\mathbf{a}]_j$ if $|\mathbf{a}_j| \leqslant \tau$ and τ otherwise. Applying elementwise truncation to $\{y_i\}_{i=1}^n$ and $\{S(\mathbf{x}_i)\}_{i=1}^n$ in (15), we allow the score $S(x)$ and the response Y to both have heavy-tailed distributions. By choosing a proper threshold τ, such a truncation step ensures $n^{-1} \sum_{i=1}^n \check{y}_i \check{S}(\mathbf{x}_i)$ converge to $\mathbb{E}[Y \cdot S(\mathbf{x})]$ with a desired rate in ℓ_∞-norm. Compared with Algorithm 1, here we only modify the definition of the loss function. Thus, we defer the details of the proposed algorithm for this setting to Algorithm 3 in §E.5.

Before stating our main theorem, we first present an assumption on the distributions of the covariate and the response variables.

Assumption 3. *Assume there exists a constant* M *such that*

$$\mathbb{E}\left[Y^4\right] \leqslant M, \quad \mathbb{E}\left[S(\mathbf{x})_j^4\right] \leqslant M, \quad \forall j \in [p].$$

Here $S(\mathbf{x})_j$ *is the j-th entry of* $S(\mathbf{x})$. *Moreover, recall that we denote* $\mu^* = \mathbb{E}[f'(\langle \mathbf{x}, \beta^* \rangle)]$. *We assume that* μ^* *is a nonzero constant such that* $M/|\mu^*| = \Theta(1)$.

Assuming the fourth moments exist and are bounded is significant weaker than the sub-Gaussian assumption. Moreover, such an assumption is prevalent in robust statistics literature [Fan et al., 2021c, 2018, 2019]. Now we are ready to introduce the theoretical results for the setting with general design.

Theorem 4. *Under* Assumption 3, *we set the thresholding parameter* $\tau = (M \cdot n/\log p)^{1/4}/2$, *let the initialization parameter* α *satisfy* $0 < \alpha \leqslant M_g^2/p$, *and set the stepsize* η *such that* $0 < \eta \leqslant 1/(12(|\mu^*| + M_g))$ *in* Algorithm 3 *given in* §E.5 *where* M_g *is a constant proportional to* M. *There exist absolute constants* a_3, a_4, *such that, with probability at least* $1 - 2p^{-2}$,

$$\|\beta_{T_1} - \mu^* \beta^*\|_2^2 \lesssim \frac{s \log p}{n}$$

holds for all $T_1 \in \left[a_3 \log(1/\alpha)/\left(\eta\left(|\mu^*| s_m - M_g\sqrt{\log p/n}\right)\right), a_4 \log(1/\alpha)\sqrt{n/\log p}/(\eta M_g)\right]$. *Here* s *is the cardinality of the support set* S *and* $s_m = \min_{i \in S_0} |\beta_j^*|$, *where* $S_0 = \left\{j \in i : |\beta_i| \geqslant C_s\sqrt{\log p/n}\right\}$ *is the set of strong signals. In addition, for the normalized iterates, we further have*

$$\left\| \frac{\beta_{T_1}}{\|\beta_{T_1}\|_2} - \frac{\mu^* \beta^*}{|\mu^*|} \right\|_2^2 \lesssim \frac{s \log p}{n},$$

with probability at least $1 - 2p^{-2}$.

Compared with Theorem 1 for the Gaussian design, here we achieve the $\mathcal{O}(\sqrt{s \log p/n})$ statistical rate of convergence in terms of the ℓ_2-norm. These rates are the same of those achieved by adding an ℓ_1-norm regularization explicitly [Plan and Vershynin, 2016, Plan et al., 2017, Yang et al., 2017] and are minimax optimal [Raskutti et al., 2011]. Moreover, we note that here $S(\mathbf{x})$ and Y can be both heavy-tailed and our truncation procedure successfully tackles such a challenge without sacrificing the statistical rates. Moreover, similar to the Gaussian case, here C_s can be set as a sufficiently large absolute constant, and the statistical rates established in Theorem 4 holds for all choices of C_s. In addition, for heavy-tailed case, we also let $\left[\tilde{\beta}_{T_1}\right]_i = [\beta_{T_1}]_i \cdot \mathbb{1}_{\left\{|[\beta_{T_1}]_i| \geqslant \lambda\right\}}$, for all $i \in [p]$. Then for all $\lambda \in \left[\alpha, (C_s|\mu^*| - 2M_g)\sqrt{\log p/n}\right]$, we obtain similar theoretical guarantees as in Theorem 2.

4 Main Results for Over-Parametrized Low Rank SIM

In this section, we present the results for over-parameterized low rank matrix SIM introduced in Definition 3 with both standard Gaussian and generally distributed covariates. Similar to the results in §3, here we also focus on matrix SIM with first-order links, i.e., we assume that $\mu^* = \mathbb{E}[f'(\langle \mathbf{X}, \beta^* \rangle)] \neq 0$, where β^* is a low rank matrix with rank r. Note that we assume that the entries of covariate $\mathbf{X} \in \mathbb{R}^{d \times d}$ are i.i.d. with a univariate density p_0. Also recall that we define the score function $S(\mathbf{X}) \in \mathbb{R}^{d \times d}$ in (5). Then, similar to the loss function in (9), we consider the loss function

$$L(\beta) := \langle \beta, \beta \rangle - 2\left\langle \beta, \frac{1}{n}\sum_{i=1}^{n} y_i S(\mathbf{X}_i) \right\rangle,$$

where $\beta \in \mathbb{R}^{d\times d}$ is a symmetric matrix. Hereafter, we rewrite β as $\mathbf{W}\mathbf{W}^\top - \mathbf{V}\mathbf{V}^\top$, where both $\mathbf{W}$ and $\mathbf{V}$ are matrices in $\mathbb{R}^{d\times d}$. The intuitions of re-parameterizing $\beta = \mathbf{W}\mathbf{W}^\top - \mathbf{V}\mathbf{V}^\top$ are as follows. Any (low rank) symmetric matrix is able to be written as the difference of two positive semidefinite matrices, namely $\mathbf{W}\mathbf{W}^\top - \mathbf{V}\mathbf{V}^\top$ with $\mathbf{W}, \mathbf{V} \in \mathbb{R}^{d\times d}$. Re-parameterizing the symmetric matrix this way is a generalization of re-parameterizing its eigenvalues by the Hadamard products. Thus this can be regarded as an extension of the re-parameterization mechanism from the vector case to the spectral domain. With such an over-parameterization, we propose to estimate β^* by applying gradient descent to the loss function

$$L(\mathbf{W}, \mathbf{V}) := \left\langle \mathbf{W}\mathbf{W}^\top - \mathbf{V}\mathbf{V}^\top, \mathbf{W}\mathbf{W}^\top - \mathbf{V}\mathbf{V}^\top \right\rangle - 2\left\langle \mathbf{W}\mathbf{W}^\top - \mathbf{V}\mathbf{V}^\top, \frac{1}{n}\sum_{i=1}^{n} y_i S(\mathbf{X}_i) \right\rangle. \tag{16}$$

Since the rank of β^* is unknown, we initialization $\mathbf{W}_0$ and $\mathbf{V}_0$ as $\mathbf{W}_0 = \mathbf{V}_0 = \alpha \cdot \mathbb{I}_{d\times d}$ for a small $\alpha > 0$ and construct a sequence of iterates $\{\mathbf{W}_t, \mathbf{V}_t, \beta_t\}_{t\geq 0}$ via the gradient decent method as follows:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta\left(\mathbf{W}_t\mathbf{W}_t^\top - \mathbf{V}_t\mathbf{V}_t^\top - \frac{1}{2n}\sum_{i=1}^{n} S(\mathbf{X}_i) y_i - \frac{1}{2n}\sum_{i=1}^{n} S(\mathbf{X}_i)^\top y_i\right)\mathbf{W}_t, \tag{17}$$

$$\begin{aligned} \mathbf{V}_{t+1} &= \mathbf{V}_t + \eta\left(\mathbf{W}_t\mathbf{W}_t^\top - \mathbf{V}_t\mathbf{V}_t^\top - \frac{1}{2n}\sum_{i=1}^{n} S(\mathbf{X}_i) y_i - \frac{1}{2n}\sum_{i=1}^{n} S(\mathbf{X}_i)^\top y_i\right)\mathbf{V}_t, \\ \beta_{t+1} &= \mathbf{W}_t\mathbf{W}_t^\top - \mathbf{V}_t\mathbf{V}_t^\top, \end{aligned} \tag{18}$$

where η in (17) and (18) is the stepsize. Note that here the algorithm does not impose any explicit regularization. In the rest of this section, we show that such a procedure yields an estimator of the true parameter β^* with near-optimal statistical rates of convergence.

Similar to the vector case, for theoretical analysis, here we also divide eigenvalues of β^* into different groups by their strengths. We let r_i^*, $i \in [d]$ be the i-th eigenvalue of β^*. The support set R of the eigenvalues is defined as $R := \{i : |r_i^*| > 0\}$, whose cardinality is r. We then divide the support set R into $R_0 := \{i : |r_i^*| \geq C_{ms}\sqrt{d \log d/n}\}$ and $R_1 := \{i : 0 < |r_i^*| < C_{ms}\sqrt{d \log d/n}\}$, which correspond to collections of strong and weak signals with cardinality denoting by r_0 and r_1,

respectively. Here $C_{ms} > 0$ is an absolute constant and we have $R = R_0 \cup R_1$. Moreover, we use r_m to denote the minimum strong eigenvalue in magnitude, i.e. $r_m = \min_{i \in R_0} |r_i^*|$.

4.1 Gaussian Design

In this subsection, we focus on the model in (7) with the entries of covariate $\mathbf{X}$ being i.i.d. $N(0, 1)$ random variables. In this case, $S(\mathbf{X}_j) = \mathbf{X}_j$. This leads to Algorithm 4 given in §F.1, where we place $S(\mathbf{X}_j)$ by $\mathbf{X}_j$ in (16)–(18).

Similar to the case in §3.1, here we also impose the following assumption for the function class of the low rank SIM.

Assumption 4. *We assume that* $\mu^* = \mathbb{E}[f'(\langle \mathbf{X}, \beta^* \rangle)]$ *is a nonzero constant. Moreover, we assume that both* $\{f(\langle \mathbf{X}_i, \beta^* \rangle)\}_{i=1}^n$ and $\{\epsilon_i\}_{i=1}^n$ *are i.i.d. sub-Gaussian random variables, with sub-Gaussian norm denoted by* $\|f\|_{\psi_2} = \mathcal{O}(1)$ and $\sigma = \mathcal{O}(1)$ *respectively. Here we let* $\|f\|_{\psi_2}$ *denote the sub-Gaussian norm of* $f(\langle \mathbf{X}, \beta^* \rangle)$. *In addition, we further assume* $|\mu^*|/\|f\|_{\psi_2} = \Theta(1)$, $|\mu^*|/\sigma = \Omega(1)$.

The following theorem establishes the statistical rates of convergence for the estimator constructed by Algorithm 4.

Theorem 5. *We set and stepsize* $0 < \alpha \leqslant M_m^2/d$ and stepsize $0 < \eta \leqslant 1/[12(|\mu^*| + M_m)]$ *in* Algorithm 4, *where* M_m *is a constant proportional to* $\max\{\|f\|_{\psi_2}, \sigma\}$. *Under* Assumption 4, *there exist constants* a_5, a_6 *such that, with probability at least* $1 - 1/(2d) - 3/n^2$, *we have*

$$\|\beta_{T_1} - \mu^* \beta^*\|_F^2 \lesssim \frac{rd \log d}{n}$$

for all $T_1 \in \left[a_5 \log(1/\alpha)/\left(\eta\left(|\mu^*| r_m - M_m\sqrt{d \log d/n}\right)\right), a_6 \log(1/\alpha)\sqrt{n/(d \log d)}/(\eta M_m)\right]$. *Moreover, for the normalized iterates* $\beta_t/\|\beta_t\|_F$, *we have*

$$\left\| \frac{\beta_{T_1}}{\|\beta_{T_1}\|_F} - \frac{\mu^* \beta^*}{|\mu^*|} \right\|_F^2 \lesssim \frac{rd \log d}{n}.$$

Similar to the vector case given in §3.1, as shown in the proof in Appendix §F, here we require C_{ms} to satisfy $C_{ms} \geqslant \max\{(a_5/a_6 + 1)M_m\{|\mu^*|, 2M_m/|\mu^*|\}$ in order to let the strong signals in R_0 dominate the noise and let the interval for T_1 to exist. The statistical rates hold for all such a C_{ms}. As shown in Theorem 5, with the proper choices of initialization parameter α, stepsize η, and the stopping time T_1, Algorithm 4 constructs an estimator that achieves near-optimal statistical rates of convergence (up to logarithmic factors compared to minimax lower bound [Rohde and Tsybakov, 2011]). Notice that the statistical rates established in Theorem 5 are also enjoyed by the M-estimator based on the least-squares loss function with nuclear norm penalty [Plan and Vershynin, 2016, Plan et al., 2017]. Thus, in terms of statistical estimation, applying gradient descent to the over-parameterized loss function in (16) is equivalent to adding a nuclear norm penalty explicitly, hence demonstrating the implicit regularization

effect. Except for obtaining the optimal ℓ_2-statistical rate, we are able to recover the true rank with high-probability by properly truncating the eigenvalues of β_{T_1} for all $T_1 \in \left[a_5 \log(1/\alpha)/\left(\eta\left(|\mu^*|r_m - M_m\sqrt{d \log d/n}\right)\right), a_6 \log(1/\alpha)\sqrt{n/(d \log d)}/(\eta M_m)\right]$. Comparing with the literature Lee et al. [2015] which studies the rank consistency via ℓ_1-regularization, we offer a wider range for choosing the tuning parameter with known left boundary α, instead of only setting the nuclear tuning parameter $\lambda = \widetilde{\Theta}(\sqrt{rd/n})$.

Theorem 6. *(Rank Consistency) Under the setting of* Theorem 5, *for all*

$$T_1 \in \left[a_5 \log(1/\alpha)/\left(\eta\left(|\mu^*|r_m - M_m\sqrt{d \log d/n}\right)\right), a_6 \log(1/\alpha)\sqrt{n/(d \log d)}/(\eta M_m)\right],$$

we let $\tilde{\beta}_{T_1} = \sum_{i=1}^{d} \mathbf{u}_i \mathbf{u}_i^\top \lambda_i(\beta_{T_1}) \cdot \mathbb{I}_{\{|\lambda_i(\beta_{T_1})| \geq \lambda\}}$, *for all* $i \in [d]$. *Here* $\mathbf{u}_k$, $k \in [d]$ *are eigenvectors of* β_{T_1}. *Then, with probability at least* $1 - 2d^{-1} - 3n^{-2}$, *for all* $\lambda \in \left[\alpha, (C_{ms}|\mu^*| - 2M_m)\sqrt{d \log d/n}\right]$, *we have* $\tilde{\beta}_{T_1}$ *enjoys the conclusion of* Theorem 5, *and* $rank\left(\tilde{\beta}_{T_1}\right) \leqslant rank(\beta^*)$. *Moreover, when there only exists strong signals in* R_0, *we further have* $rank\left(\tilde{\beta}_{T_1}\right) = rank(\beta^*)$.

Furthermore, our method extends the existing works that focus on designing algorithms and studying implicit regularization phenomenon in noiseless linear matrix sensing models with positive semidefinite signal matrices [Gunasekar et al., 2017, Li et al., 2018, Arora et al., 2019, Gidel et al., 2019]. Specifically, we allow a more general class of (noisy) models and symmetric signal matrices. Compared with Li et al. [2018], our methodology possesses several strengths, which include achieving low sample complexity ($\widetilde{\mathcal{O}}(rd)$ insted of $\widetilde{\mathcal{O}}(r^2 d)$., allowing weak signals ($\min_{i \in R}|r_i^*| z\mathcal{O}\left((1/n)^{1/2}\right)$ instead of $\min_{i \in R}|r_i^*| \gtrsim \mathcal{O}\left((1/n)^{1/6}\right)$), getting tighter statistical rate under noisy models ($\widetilde{\mathcal{O}}(dr/n)$ instead of $\widetilde{\mathcal{O}}(\kappa rd/n)$), and applying to a more general class of noisy statistical models. These strengths are achieved by the use of score transformation together with a refined trajectory analysis, which involves studying the dynamics of eigenvalues inside the strong signal set elementwisely with multiple stages instead of only studying the dynamics of the minimum eigenvalue with two stages.

The way of choosing stopping time T_1 in the case of matrix SIM is almost the same with our method in §3.1.2. The only difference between them is that here we replace $\mathbf{x}^\top \beta^*$ by $\mathrm{tr}(\mathbf{X}^\top \beta^*)$ Indeed, as we assume $\|\Sigma^{1/2}\beta^*\|_2 = 1$ in vector SIM and $\|\beta^*\|_F = 1$ in matrix version for model identifiability, both $\mathbf{x}^\top \beta_t$ and $\mathrm{tr}(\mathbf{X}^\top \beta_t)$ follow the standard normal distribution. Thus, our results on the prediction risk in §3.1.3 can be applied here directly.

4.2 General Design

In the rest of this section, we focus on the low rank matrix SIM beyond Gaussian covariates. Hereafter, we assume the entries of $\mathbf{X}$ are i.i.d. random variables with a known density function $p_0 : \mathbb{R} \to \mathbb{R}$. Recall that, according to the remarks following Definition 1, the score function $S(\mathbf{X}) \in \mathbb{R}^{d \times d}$ is defined as

$$S(\mathbf{X})_{j,k} \coloneqq S(\mathbf{X}_{j,k}) = -p'_0(\mathbf{X}_{j,k})/p_0(\mathbf{X}_{j,k}),$$

where $S(\mathbf{X})_{j,k}$ and $\mathbf{X}_{j,k}$ are the (j, k)-th entries of $S(\mathbf{X})$ and $\mathbf{X}$ for all $j, k \in [d]$. However, similar to the results in §3.2, the entries of $S(\mathbf{X})$ can have heavy-tailed distributions and thus $n^{-1}\sum_{i=1}^{n} y_i \cdot S(\mathbf{X}_i)$ may not converge its expectation $\mathbb{E}[Y \cdot S(\mathbf{X})]$ efficiently in terms of spectral norm. Here $\mathbf{X}_i$ is the i-th observation of the covariate $\mathbf{X}$. To tackle such a challenge, we employ a shrinkage approach [Catoni et al., 2012, Fan et al., 2021c, Minsker, 2018] to construct a robust estimator of $\mathbb{E}[Y \cdot S(\mathbf{X})]$. Specifically, we let

$$\phi(x) = \begin{cases} \log\left(1 - x + x^2/2\right), & x \leqslant 0, \\ \log\left(1 + x + x^2/2\right), & x > 0 \end{cases},$$

which is approximately x when x is small and grows at logarithmic rate for large x. The rescaled version $\lambda^{-1}\phi(\lambda x)$ for $\lambda \to 0$ behaves like a soft-winsorizing function, which has been widely used in statistical mean estimation with finite bounded moments [Catoni et al., 2012, Brownlees et al., 2015]. For any matrix $\mathbf{X} \in \mathbb{R}^{d \times d}$, we apply spectral decomposition to its Hermitian dilation and obtain

$$\mathbf{X}^* := \begin{bmatrix} 0 & \mathbf{X} \\ \mathbf{X}^\top & 0 \end{bmatrix} = \mathbf{Q}\boldsymbol{\Sigma}^*\mathbf{Q}^\top,$$

where $\Sigma^* \in \mathbb{R}^{2d \times 2d}$ is a diagonal matrix. Based on such a decomposition, we define $\widetilde{\mathbf{X}} = \mathbf{Q}\phi(\boldsymbol{\Sigma}^*)\mathbf{Q}^\top$, where ϕ applies elementwisely to Σ^*. Then we write $\widetilde{\mathbf{X}}$ as a block matrix as

$$\widetilde{\mathbf{X}} := \begin{bmatrix} \widetilde{\mathbf{X}}_{11} & \widetilde{\mathbf{X}}_{12} \\ \widetilde{\mathbf{X}}_{21} & \widetilde{\mathbf{X}}_{22} \end{bmatrix},$$

where each block of $\widetilde{\mathbf{X}}$ is in $\mathbb{R}^{d \times d}$. We further define a mapping $\phi_1 : \mathbb{R}^{d \times d} \to \mathbb{R}^{d \times d}$ by letting $\phi_1(\mathbf{X}) := \widetilde{\mathbf{X}}_{12}$, which is a regularized version of $\mathbf{X}$. Given data y_1, $\mathbf{X}_1$, we finally define $\mathscr{H}(\cdot)$ as

$$\mathscr{H}(y_1 S(\mathbf{X}_1), \kappa) := 1/\kappa \cdot \phi_1(\kappa y_1 \cdot S(\mathbf{X}_1)), \quad \forall \kappa > 0, \tag{19}$$

where κ is a thresholding parameter, converging to zero. This method is in a similar spirit of robustifying the singular value of $\mathbf{X}$. Based on the operator $\mathscr{H}$ defined in (19), we define a loss function $L(\mathbf{W}, \mathbf{V})$ as

$$L(\mathbf{W}, \mathbf{V}) := \left\langle \mathbf{W}\mathbf{W}^\top - \mathbf{V}\mathbf{V}^\top, \mathbf{W}\mathbf{W}^\top - \mathbf{V}\mathbf{V}^\top \right\rangle - \frac{2}{n}\sum_{i=1}^{n} \left\langle \mathbf{W}\mathbf{W}^\top - \mathbf{V}\mathbf{V}^\top, \mathscr{H}(y_i S(\mathbf{X}_i), \kappa) \right\rangle. \tag{20}$$

After over-parameterizing β as $\mathbf{W}\mathbf{W}^\top - \mathbf{V}\mathbf{V}^\top$, we propose to construct an estimator of β^* by applying gradient descent on the following loss function in (20) with respect to $\mathbf{W}$, $\mathbf{V}$. See Algorithm 5 in §F.5 for the details of the algorithm.

In the following, we present the statistical rates of convergence for the obtained estimator. We first introduce the assumption on Y and p_0.

Assumption 5. *We assume that both the response variable Y and entries of $S(\mathbf{X})$ have bounded fourth moments. Specifically, there exists an absolute constant M such that*

$$\mathbb{E}\left[Y^4\right] \leqslant M, \quad \mathbb{E}\left[S(\mathbf{X})_{i,j}^4\right] \leqslant M, \quad \forall (i,j) \in [d] \times [d].$$

Moreover, we assume that $\mu^* = \mathbb{E}[f'(\langle \mathbf{X}, \beta^* \rangle)]$ is a nonzero constant such that $|\mu^*|/M = \Theta(1)$.

Next, we present the main theorem for low rank matrix SIM.

Theorem 7. *In* Algorithm 5, *we set parameter κ in* (19) *as $\kappa = \sqrt{\log(4d)/(nd \cdot M)}$ and let the initialization parameter α and the stepsize η satisfy $0 < \alpha \leqslant M_{mg}^2/d$* and $0 < \eta \leqslant 1/[12(|\mu^*| + M_{mg})]$, *where M_{mg} is a constant proportional to M. Then, under* Assumption 5, *there exist absolute constants a_7, a_8 such that, with probability at least* $1 - (4d)^{-2}$, *we have*

$$\|\beta_{T_1} - \mu^*\beta^*\|_F^2 \lesssim \frac{rd \log d}{n},$$

for all $T_1 \in \left[a_7 \log(1/\alpha)/\left(\eta\left(|\mu^*| r_m - M_{mg}\sqrt{d \log d/n}\right)\right), a_8 \log(1/\alpha)\sqrt{n/(d \log d)}/(\eta M_{mg})\right]$, Moreover, for the normalized iterate $\beta_t/\|\beta_t\|_F$, we have

$$\left\| \frac{\beta_{T_1}}{\|\beta_{T_1}\|_F} - \frac{\mu^*\beta^*}{|\mu^*|} \right\|_F^2 \lesssim \frac{rd \log d}{n}.$$

For low rank matrix SIM, when the hyperparameters of the gradient descent algorithm are properly chosen, we also capture the implicit regularization phenomenon by applying a simple optimization procedure to over-parameterized loss function with heavy-tailed measurements. Here, applying the thresholding operator $\mathscr{H}$ in (19) can also be viewed as a data pre-processing step, which arises due to handling heavy-tailed observations. Note that the way of choosing C_{ms} here is similar with the way in Theorem 5, in order to ensure the convergence rate and existence of a time interval, so we omit the details. Note that the ℓ_2-statistical rate given in Theorem 7 are minimax optimal up to a logarithmic term [Rohde and Tsybakov, 2011]. Similar results were also obtained by Plan and Vershynin [2016], Yang et al. [2017], Goldstein et al. [2018], Na et al. [2019] via adding explicit nuclear norm regularization. Thus, in terms of statistical recovery, when employing the thresholding in (19) and over-parameterization, gradient descent enforces implicit regularization that has the same effect as the nuclear norm penalty. In addition, in terms of the rank consistency

result for the heavy-tailed case, if we also let $\tilde{\beta}_{T_1} = \sum_{i=1}^{d} \mathbf{u}_i \mathbf{u}_i^\top \lambda_i(\beta_{T_1}) \cdot \mathbb{I}_{\{|\lambda_i(\beta_{T_1})| \geq \lambda\}}$, then for all $\lambda \in \left[\alpha, (C_s|\mu^*| - 2M_{mg})\sqrt{d \log d/n}\right]$, we achieve the same results with Theorem 6.

5 Conclusion

In this paper, we leverage over-parameterization to design regularization-free algorithms for single index model and provide theoretical guarantees for the induced implicit regularization phenomenon. We consider the case where the link function is unknown, the distribution of the covariates is known as a prior, and the signal parameter is either a s-sparse vector in $\mathbb{R}^p$ or a rank-r matrix in $\mathbb{R}^{d \times d}$. Using the score function and the Stein's identity, we propose an over-parameterized nonlinear least-squares loss function. To handle the possibly heavy-tailed distributions of the score functions and the response variables, we adopt additional truncation techniques that robustify the loss function. For both the vector and matrix SIMs, we construct an estimator of the signal parameter by applying gradient descent to the proposed loss function, without any explicit regularization. We prove that, when initialized near the origin, gradient descent with a small stepsize finds an estimator that enjoys minimax-optimal statistical rates of convergence. Moreover, for vector SIM with Gaussian design, we further obtain the oracle statistical rates that are independent of the ambient dimension. Furthermore, our experimental results support our theoretical findings and also demonstrate that our methods empirically outperform classical methods with explicit regularization in terms of both ℓ_2-statistical rate and variable selection consistency.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research supported by the NSF grant DMS-1662139 and DMS-1712591, the ONR grant N00014-19-1-2120, and the NIH grant 2R01-GM072611-16.

References

Arora S, Cohen N, Hu W, and Luo Y. Implicit regularization in deep matrix factorization. In Advances in Neural Information Processing Systems, pages 7411–7422, 2019.

Arulkumaran K, Deisenroth MP, Brundage M, and Bharath AA. Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine, 34(6):26–38, 2017.

Brownlees C, Joly E, and Lugosi G. Empirical risk minimization for heavy-tailed losses. Annals of Statistics, 43(6):2507–2536, 2015.

Candés EJ. The restricted isometry property and its implications for compressed sensing. Comptes rendus-Mathematique, 9(346):589–592, 2008.

Candés EJ, Eldar YC, Strohmer T, and Voroninski V. Phase retrieval via matrix completion. SIAM review, 57(2):225–251, 2015.

Catoni O et al. Challenging the empirical mean and empirical variance: A deviation study. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, 48(4):1148–1185, 2012.

Dauphin YN, Pascanu R, Gulcehre C, Cho K, Ganguli S, and Bengio Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Advances in Neural Information Processing Systems, pages 2933–2941, 2014.

Donoho DL. Compressed sensing. IEEE Transactions on information theory, 52(4):1289–1306, 2006.

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

Fan J, Xue L, and Zou H. Strong oracle optimality of folded concave penalized estimation. Annals of Statistics, 42(3):819, 2014. [PubMed: 25598560]

Fan J, Liu H, and Wang W. Large covariance estimation through elliptical factor models. Annals of Statistics, 46(4):1383–1414, 2018. [PubMed: 30214095]

Fan J, Wang W, and Zhong Y. Robust covariance estimation for approximate factor models. Journal of Econometrics, 208(1):5–22, 2019. [PubMed: 30546195]

Fan J, Ma C, and Zhong Y. A selective overview of deep learning. Statistical Science, 36(2):264–290, 2021a. [PubMed: 34305305]

Fan J, Wang K, Zhong Y, and Zhu Z. Robust high dimensional factor models with applications to statistical machine learning. Statistical Science, 36:303–327, 2021b. [PubMed: 34321713]

Fan J, Wang W, and Zhu Z. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. Annals of Statistics, 49(3):1239–1266, 2021c. [PubMed: 34556893]

Gidel G, Bach F, and Lacoste-Julien S. Implicit regularization of discrete gradient dynamics in linear neural networks. In Advances in Neural Information Processing Systems, pages 3196–3206, 2019.

Goldstein L and Wei X. Non-Gaussian observations in nonlinear compressed sensing via Stein discrepancies. Information and Inference: A Journal of the IMA, 8(1):125–159, 2019.

Goldstein L, Minsker S, and Wei X. Structured signal recovery from non-linear and heavy-tailed measurements. IEEE Transactions on Information Theory, 64(8):5513–5530, 2018.

Goodfellow I, Bengio Y, and Courville A. Deep learning. MIT press, 2016.

Gunasekar S, Woodworth BE, Bhojanapalli S, Neyshabur B, and Srebro N. Implicit regularization in matrix factorization. In Advances in Neural Information Processing Systems, pages 6151–6159, 2017.

Hoff PD. Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization. Computational Statistics & Data Analysis, 115:186–198, 2017.

Ke Y, Minsker S, Ren Z, Sun Q, Zhou W-X, et al. User-friendly covariance estimation for heavy-tailed distributions. Statistical Science, 34(3):454–471, 2019.

Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, and Tang PTP. On large-batch training for deep learning: Generalization gap and sharp minima. In International Conference on Learning Representations, 2017.

LeCun Y, Bengio Y, and Hinton G. Deep learning. nature, 521(7553):436–444, 2015. [PubMed: 26017442]

Lee JD, Sun Y, and Taylor JE. On model selection consistency of regularized M-estimators. Electronic Journal of Statistics, 9(1):608–642, 2015.

Li Y. Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274, 2017.

Li Y, Ma T, and Zhang H. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In COLT, 2018.

Li Y Luo, and Lyu K. Towards resolving the implicit bias of gradient descent for matrix factorization:Greedy low-rank learning. In International Conference on Learning Representations, 2021.

McMahan B, Moore E, Ramage D, Hampson S, and Arcas B. A. y.. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh A and Zhu J, editors, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, volume 54 of Proceedings of Machine Learning Research, pages 1273–1282. PMLR, 20–22 Apr 2017.

Minsker S. Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. Annals of Statistics, 46(6A):2871–2903, 2018.

Minsker S and Wei X. Robust modifications of U-statistics and applications to covariance estimation problems. Bernoulli, 26(1):694–727, 2020.

Na S, Yang Z, Wang Z, and Kolar M. High-dimensional varying index coefficient models via stein's identity. Journal of Machine Learning Research, 20(152):1–44, 2019.

Neykov M, Liu JS, and Cai T. ℓ_1-regularized least squares for support recovery of high dimensional single index models with Gaussian designs. Journal of Machine Learning Research, 17(1):2976–3012, 2016. [PubMed: 28503101]

Neyshabur B, Tomioka R, and Srebro N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In International Conference on Learning Representations, 2015.

Neyshabur B, Tomioka R, Salakhutdinov R, and Srebro N. Geometry of optimization and implicit regularization in deep learning. arXiv preprint arXiv:1705.03071, 2017.

Otter DW, Medina JR, and Kalita JK. A survey of the usages of deep learning in natural language processing. IEEE Transactions on Neural Networks and Learning Systems, pages 1–21, 2020.

Plan Y and Vershynin R. The generalized Lasso with non-linear observations. IEEE Transactions on information theory, 62(3):1528–1537, 2016.

Plan Y, Vershynin R, and Yudovina E. High-dimensional estimation with geometric constraints. Information and Inference: A Journal of the IMA, 6(1):1–40, 2017.

Poggio T, Kawaguchi K, Liao Q, Miranda B, Rosasco L, Boix X, Hidary J, and Mhaskar H. Theory of deep learning III: Explaining the non-overfitting puzzle. arXiv preprint arXiv:1801.00173, 2017.

Raskutti G, Wainwright MJ, and Yu B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. TIT, 57(10):6976–6994, 2011.

Richards D and Rebeschini P. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. Journal of Machine Learning Research, 21(34):1–44, 2020. [PubMed: 34305477]

Richards D, Rebeschini P, and Rosasco L. Decentralised learning with distributed gradient descent and random features. 2020.

Rohde A and Tsybakov AB. Estimation of high-dimensional low-rank matrices. Annals of Statistics, 39(2): 887–930, 2011.

Shechtman Y, Eldar YC, Cohen O, Chapman HN, Miao J, and Segev M. Phase retrieval with application to optical imaging: a contemporary overview. IEEE signal processing magazine, 32(3):87–109, 2015.

Stein C, Diaconis P, Holmes S, Reinert G, et al. Use of exchangeable pairs in the analysis of simulations. In Stein's Method, pages 1–25. Institute of Mathematical Statistics, 2004.

Stein C et al. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, volume 2. The Regents of the University of California, 1972.

Swirszcz G, Czarnecki WM, and Pascanu R. Local minima in training of neural networks. In International Conference on Learning Representations, 2016.

Torfi A, Shirvani RA, Keneshloo Y, Tavvaf N, and Fox EA. Natural language processing advancements by deep learning: A survey. arXiv preprint arXiv:2003.01200, 2020.

Tsybakov AB. Introduction to Nonparametric Estimation. Springer, 2008. ISBN 0387790519.

Vaškevičius T, Kanade V, and Rebeschini P. Implicit regularization for optimal sparse recovery. In Advances in Neural Information Processing Systems, pages 2972–2983, 2019.

Voulodimos A, Doulamis N, Doulamis A, and Protopapadakis E. Deep learning for computer vision: A brief review. Computational Intelligence and Neuroscience, 2018, 2018.

Wei X. Structured recovery with heavy-tailed measurements: A thresholding procedure and optimal rates. arXiv preprint arXiv:1804.05959, 2018.

Wei X and Minsker S. Estimation of the covariance structure of heavy-tailed distributions. In Advances in Neural Information Processing Systems, pages 2859–2868, 2017.

Wilson AC, Roelofs R, Stern M, Srebro N, and Recht B. The marginal value of adaptive gradient methods in machine learning. In Advances in Neural Information Processing Systems, pages 4148–4158, 2017.

Yang Z, Balasubramanian K, and Liu H. High-dimensional non-Gaussian single index models via thresholded score function estimation. In International Conference on Machine Learning, pages 3851–3860. JMLR. org, 2017.

Yun C, Sra S, and Jadbabaie A. Small nonlinearities in activation functions create bad local minima in neural networks. In International Conference on Learning Representations, 2019.

Zhang C, Bengio S, Hardt M, Recht B, and Vinyals O. Understanding deep learning requires rethinking generalization. International Conference on Learning Representations, 2017.

Zhang C-H et al. Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics, 38 (2):894–942, 2010.

Zhao P, Yang Y, and He Q-C. Implicit regularization via Hadamard product over-parametrization in high-dimensional linear regression. arXiv preprint arXiv:1903.09367, 2019.

Zhu Z. Taming the heavy-tailed features by shrinkage and clipping. arXiv preprint arXiv:1710.09020, 2017.

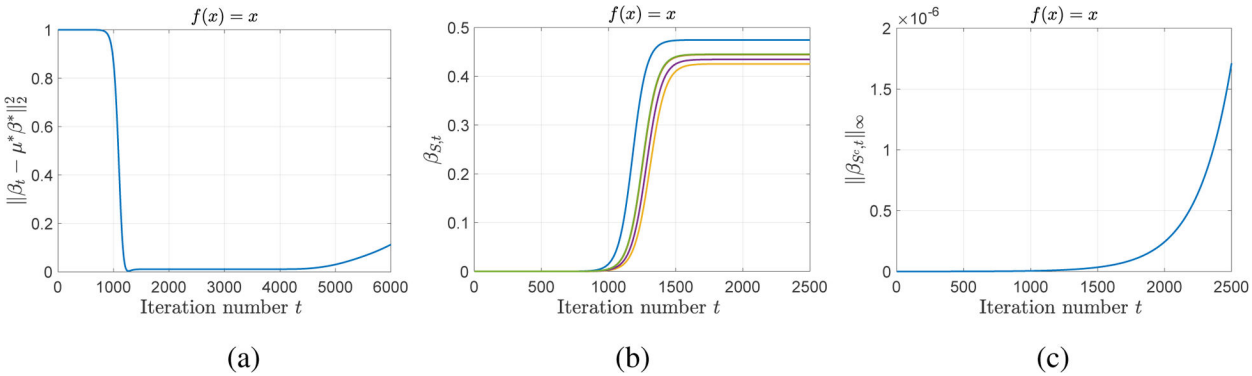
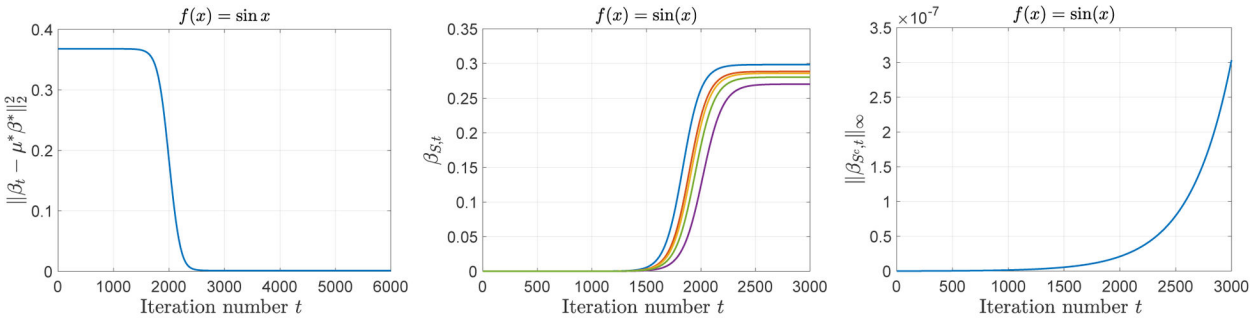


Figure 1:

With link function $f(x) = x$, (a) characterizes the evolution of distance $\|\beta_t - \mu^*\beta^*\|_2^2$ against iteration number t; (b) depicts the trajectories $\beta_{j,t}$ ($j \in S$) for five nonzero components, and (c) presents the trajectory $\max_{j \in S^c}|\beta_{j,t}|$.



(a) (b) (c)

Figure 2:

With link function $f(x) = \sin(x)$, similar to Figure 1, here (a) characterizes the evolution of distance $\|\beta_t - \mu^*\beta^*\|_2^2$ against iteration number t; (b) depicts the trajectories $\beta_{j,t}$ ($j \in S$) for five nonzero components, and (c) presents the trajectory $\max_{j \in S^c}|\beta_{j,t}|$.