Understanding Implicit Regularization in Over-Parameterized Single Index Model

Jianqing Fan, Zhuoran Yang, and Mengxin Yu

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ; Department of Statistics and Data Science, Yale University, New Haven, CT

ABSTRACT

In this article, 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 $\ell_2$-statistical rate and variable selection consistency. Supplementary materials for this article are available online.

1. Introduction

With the astonishing empirical success in various application domains such as computer vision (Voulodimos et al. 2018), natural language processing (Otter, Medina, and Kalita 2020; Torfi et al. 2020), and reinforcement learning (Arulkumaran et al. 2017; Li 2017), deep learning (LeCun, Bengio, and Hinton 2015; Goodfellow, Bengio, and Courville 2016; Fan, Ma, and Zhong 2021) 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, Czarnecki, and Pascanu 2016; Yun, Sra, and Jadbabaie 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 nonconvexity. Moreover, the obtained solution also generalizes well to unseen data with small test error (Neyshabur, Tomioka, and Srebro 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, Tomioka, and Srebro (2015) and 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, Tomioka, and Srebro 2015; Neyshabur et al. 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, Ma, and Zhang (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, that is, low-rank matrix sensing. They show that gradient descent biases toward the minimum nuclear norm solution when the initialization is close to the origin, step sizes
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 $UU^\top$, where $U \in \mathbb{R}^{d \times d}$, and propose to estimate the true parameter by updating $U$ via gradient descent. Li, Ma, and Zhang (2018) proves that, with $\tilde{O}(r^2d)$ iid observations of the model, gradient descent provably recovers the true parameter with accuracy, where $\tilde{O}(\cdot)$ hides absolute constants and poly-logarithmic 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, Yang, and He (2019) and Vaskevičius, Kanade, and Rebeschini (2019) recently design a novel 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 reparametrize the parameter using two vectors in $\mathbb{R}^d$ via the Hadamard product and estimate the true parameter via unregularized gradient descent with proper initialization, stepsizes, and the number of iterations. They prove independently that, with $n = \tilde{O}(s^2 \log p)$ iid 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 $\tilde{O}(\sqrt{s \log s/n})$ 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, Xue, and Zou 2014) such as smoothly clipped absolute deviation (SCAD) (Fan and Li 2001) and minimax concave penalty (MCP) (Zhang 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 $Y = f(X, \beta^*) + \epsilon$, with $\beta^*$ being the true parameter, $\epsilon$ being the random noise, and $f : \mathbb{R}^d \to \mathbb{R}$ being an unknown (nonlinear) link function. Here $\beta^*$ 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 $\beta$ is not identifiable. Thus, for the vector and matrix cases, respectively, we further assume that the $\ell_2$-or Frobenius norms of $\beta^*$ are equal to one. Our goal is to recover the true parameter $\beta^*$ given $n$ iid observations of the model. Such a model can be viewed as the misspecified version of the compressed sensing (Donoho 2006; Candès 2008) and phase retrieval (Shechtman et al. 2015; Candès 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 $\beta^*$ via nonlinear least-squares. Moreover, jointly minimizing the least-squares loss function with respect to $\beta^*$ and $f$ is computationally intractable. To overcome these challenges, a recent line of research proposes to estimate $\beta^*$ 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 1972) implies that the expectation of $Y \cdot X$ is proportional to $\beta^*$. Thus, despite the nonlinear link function, $\beta^*$ can be accurately estimated by neglecting $f$ and fitting a regularized least-squares regression. In particular, when $\beta^*$ is a sparse vector, Plan and Vershynin (2016) and Plan, Vershynin, and Yudovina (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, Minsker, and Wei (2018), Wei (2018), and Goldstein and Wei (2019) allow the covariates to follow an elliptically symmetric distribution that can be heavy-tailed. In addition, using a generalized version of Stein's identity (Stein et al. 2004), Yang, Balasubramanian, and Liu (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 $\beta^*$. 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, Balasubramanian, and Liu (2017) and rewrite the parameter of interest by over-parameterization. When $\beta^*$ is a sparse vector in $\mathbb{R}^p$, we adopt a Hadamard product parameterization (Hoff 2017; Zhao, Yang, and He 2019; Vaskevičius, Kanade, and Rebeschini 2019) and write $\beta^*$ as $w \odot v - v \odot w$, where both $w$ and $v$ 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 $w$ and $v$ are initialized near an all-zero vector and the stepsizes are fixed to be a sufficiently small constant $\eta > 0$. Furthermore, when $\beta^*$ is a low-rank matrix, we similarly represent $\beta^*$ as $WW^\top - VV^\top$ and propose to recover $\beta^*$ 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 $\mathbb{E}[Y \cdot S_p(X)]$ 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 (Zhu 2017; Wei and Minsker 2017; Minsker 2018; Ke et al. 2019; Minsker and Wei 2020; Fan et al. 2021; Fan, Wang, and Zhu 2021). 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 $\mathcal{O}(\sqrt{\log p/n})$ and $\mathcal{O}(\sqrt{\kappa d \log d/n})$ $\ell_2$-statistical rates under the sparse and low-rank settings, respectively. Moreover, for sparse $\beta^*$, when the magnitude of the nonzero entries is sufficiently large, we prove that our estimator enjoys an oracle $\mathcal{O}(\sqrt{\log n/n})$ $\ell_2$-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 2-fold. First, as we mentioned in the last paragraph, under mild conditions, our estimator enjoys oracle statistical rate whereas the most commonly used $\ell_1$-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 3-fold. First, in terms of $\ell_2$-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 $\mathbb{E}[Y \cdot S_p(X)]$ and propose to estimate $\beta^*$ 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 $\ell_2$-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 supplementary materials.

### 1.2. Notation

In this section, 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 $u$, we use $u_S$ to denote the vector whose $i$th entry is $u_i$ if $i \in S$ and 0 otherwise. For any vector $u$ and $q \geq 0$, we use $\|u\|_{\ell_q}$ to represent the vector $\ell_q$ norm. In addition, the inner product $\langle u, v \rangle$ between any pair of vectors $u, v$ is defined as the Euclidean inner product $u^T v$. Moreover, we define $u \odot v$ as the Hadamard product of vectors $u, v$. For any given matrix $X \in \mathbb{R}^{d_1 \times d_2}$, we use $\|X\|_{\text{op}}, \|X\|_F$ and $\|X\|_*$ to represent the operator norm, Frobenius norm and nuclear norm of matrix $X$, respectively. In addition, for any two matrices $X, Y \in \mathbb{R}^{d_1 \times d_2}$, we define their inner product $\langle X, Y \rangle$ as $\langle X, Y \rangle = \text{tr}(X^T Y)$. Moreover, if we write $X \geq 0$ or $X \preceq 0$, then the matrix $X$ is meant to be positive semidefinite or negative semidefinite. We let $(a_n)_{n \geq 1}$ be any two positive series. We write $a_n \lesssim b_n$ if there exists a universal constant $C$ such that $a_n \leq C \cdot b_n$ and we write $a_n \ll b_n$ if $a_n/b_n \to 0$. In addition, we write $a_n = b_n$ 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 $a_n = \omega(b_n)$ share the same meaning with $a_n \lesssim b_n$ and $a_n \ll b_n$. Moreover, $a_n = \mathcal{O}(b_n)$ means $a_n \leq C b_n$ up to some logarithm terms. Finally, we use $a_n = \Theta(b_n)$ if there exists a universal constant $c > 0$ such that $a_n/b_n \geq c$ and we use $a_n = \Omega(b_n)$ if $c \leq a_n/b_n \leq C$ where $c, C > 0$ are universal constants.

### 1.3. Roadmap

The organization of our article is as follows. We introduce the background knowledge in Section 2. In Sections 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 Section B, Supplementary materials 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, Ma, and Zhang (2018) have studied least squares objectives over positive semidefinite matrices $\beta \in \mathbb{R}^{d \times d}$ of the following form

$$
\min_{\beta \succeq 0} F(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle X_i, \beta \rangle)^2,
$$

(1)

where the labels $\{y_i\}_{i=1}^{n}$ are generated from linear measurements $y_i = \langle X_i, \beta^* \rangle, i \in [n]$, with $\beta^* \in \mathbb{R}^{d \times d}$ being positive semidefinite and low rank. Here $\beta^*$ is of rank $r$ where $r$ is much smaller than $d$. Instead of working on parameter $\beta$ directly, they write $\beta$ as $\beta = UU^T$ where $U \in \mathbb{R}^{d \times d}$, and study the optimization problem related to $U$,

$$
\min_{U \in \mathbb{R}^{d \times d}} f(U) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle X_i, UU^T \rangle)^2.
$$

(2)

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

$$
\hat{\beta} = \arg\min_{\beta \succeq 0} \|\beta\|_*,
$$

subject to $\langle X_i, \hat{\beta} \rangle = y_i, \forall i \in [n]$.

Subsequently, Li, Ma, and Zhang (2018) assumes $\{X_i\}_{i=1}^{m}$ satisfy the restricted isometry property (RIP) condition (Candès 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 $\beta^*$ is achieved.

Recently, Li, Luo, and Lyu (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 commutative $\{X_i\}_{i=1}^{m}$ serves as a special case to Li, Luo, and Lyu (2021).

As for noisy statistical model, both Zhao, Yang, and He (2019) and Vaskevičius, Kanade, and Rebeschini (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

$$
y_i = x_i^T \beta^* + \epsilon_i, i \in [n],
$$

(3)

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

$$
\min_{w \in \mathbb{R}^p, v \in \mathbb{R}^p} L(w, v) = \frac{1}{2n} \sum_{i=1}^{n} \left( y_i - (w \odot v) - (x_i^T \beta^* - v) \right)^2,
$$

(4)

where the parameter $\beta$ is over-parameterized as $\beta = w \odot w - v \odot v$. 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 $\beta^*$ with optimal statistical rate of convergence.

2.2. High-Dimensional Single Index Model

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

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

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

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

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

$$
S(X)_{ij} := -p_0'(X_{ij})/p_0(X_{ij}).
$$

(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 $x \in \mathbb{R}^p$ follows a distribution with density function $p_0(x) : \mathbb{R}^p \rightarrow \mathbb{R}$ which is differentiable and satisfies the condition that $|p_0(x)|$ converges to zero as $|x|^2$ goes to infinity. Then for any differentiable function $f(x)$ with $E[|f(x)S(x)|] < \infty$, it holds that,

$$
E[f(x)S(x)] = E[\nabla_x f(x)],
$$

where $S(x) = -\nabla_x p_0(x)/p_0(x)$ is the score function with respect to $x$ defined in Definition 1.

**Remark.** In the case of having matrix covariate, we are able to achieve the same conclusion by simply replacing $x \in \mathbb{R}^p$ by $X \in \mathbb{R}^{d \times d}$ 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 function \( f : \mathbb{R} \rightarrow \mathbb{R} \), \( p \)-dimensional covariate \( \mathbf{x} \) as well as signal \( \beta^* \) which is the parameter of interest. Here, we let \( \epsilon \in \mathbb{R} \) be an exogenous noisy response with mean zero. In addition, if not particularly indicated, we assume the entries of \( \mathbf{x} \) are iid random variables with a known univariate density \( p_0(x) \). As for the underlying true signal \( \beta^* \), it is assumed to be \( s \)-sparse with \( s \ll p \). Note that the length of \( \beta^* \) can be absorbed by the unknown link function, 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 \( y = \text{sign}(\mathbf{x}_i^\top \beta^*) + \epsilon \).

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 matrix \( \mathbf{X} \in \mathbb{R}^{n \times d} \), we assume the entries of \( \mathbf{X} \) are iid 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 \( \epsilon \) 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 overparameterization 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 Sections 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
\[
\mathbb{E}[Y \cdot S(x)] = \mathbb{E}[f(\langle \mathbf{x}, \beta^* \rangle) \cdot S(x)] = \mathbb{E}[f'(\langle \mathbf{x}, \beta^* \rangle)] \cdot \beta^* =: \mu^* \beta^*,
\]
which recovers our true signal \( \beta^* \) up to scaling. Here we define \( \mu^* = \mathbb{E}[f'(\langle \mathbf{x}, \beta^* \rangle)] \), which is assumed nonzero throughout this work. Hence, \( Y \cdot S(x) \) serves as an unbiased estimator of \( \mu^* \beta^* \), and we can correctly identify the direction of \( \beta^* \) by solving a population-level optimization problem:
\[
\min_{\beta} L(\beta) := \langle \beta, \beta \rangle - 2\beta, \mathbb{E}[Y \cdot S(x)].
\]
Since we only have access to finite data, we replace \( \mathbb{E}[Y \cdot S(x)] \) by its sample version estimator \( \frac{1}{n} \sum_{i=1}^n y_i S(x_i) \), and plug the sample-based estimator into the loss function. In a high dimensional SIM given in Definition 2, where the true signal \( \beta^* \) is assumed to be sparse, various works (Plan and Vershynin 2016; Plan, Vershynin, and Yudovina 2017; Yang, Balasubramanian, and Liu 2017) have shown that the \( \ell_1 \)-regularized estimator \( \hat{\beta} \) given by
\[
\hat{\beta} \in \text{argmin}_{\beta} L(\beta) := \langle \beta, \beta \rangle - 2\beta, \frac{1}{n} \sum_{i=1}^n y_i S(x_i) + \lambda \|\beta\|_1
\]
attains the optimal statistical rate of convergence rate to \( \mu^* \beta^* \).

In contrast, instead of imposing an \( \ell_1 \)-norm regularization term, we propose to obtain an estimator by minimizing the loss function \( L \) directly, with \( \beta \) reparameterized using two vectors \( w \) and \( v \) in \( \mathbb{R}^p \). Specifically, we write \( \beta = w \odot w - v \odot v \) and thus, equivalently write the loss function \( L(\beta) \) as \( L(w, v) \), which is given by
\[
L(w, v) = \langle w \odot w - v \odot v, w \odot w - v \odot v \rangle
- 2\langle w \odot w - v \odot v, \frac{1}{n} \sum_{i=1}^n y_i S(x_i) \rangle. \tag{9}
\]
Note that the way of writing \( \beta \) in terms of \( w \) and \( v \) is not unique. In particular, \( \beta \) has \( p \) degrees of freedom but we use \( 2p \) parameters to represent \( \beta \). Thus, by using \( w \) and \( v \) instead of \( \beta \), we employ over-parameterization in (9).

We briefly describe our motivation on over-parameterizing \( \beta \) by \( w \odot w - v \odot v \). Suppose that \( \beta \) is sparse, an explicit regularization is to use \( \ell_1 \)-penalty. Note that \( \|\beta\|_1 = \min_{\gamma, \delta} \sum_{i=1}^p (Y_i - f(x_i^\top \gamma \odot \delta)) + \lambda \|\gamma\|_2^2 + \|\delta\|_2^2 \) where \( \odot \) denotes the Hadamard (componentwise) product. Thus, an explicit regularization is to minimize \( \sum_{i=1}^n (Y_i - f(x_i^\top \gamma \odot \delta))^2 + \lambda \|\gamma\|_2^2 + \|\delta\|_2^2 \) for a penalty parameter \( \lambda \), following the method in Hoff (2017). To gain understanding on implicit regularization by over parameterization, we let \( w = (\gamma + \delta)/2 \) and \( v = (\gamma - \delta)/2 \). Then \( \beta = (\gamma \odot \delta) = w \odot w - v \odot v \) with \( 2p \) new parameters \( w \) and \( v \) that over-parameterize the problem. This leads to the empirical loss \( L(w, v) := \sum_{i=1}^n (Y_i - f(x_i^\top (w \odot w - v \odot v)))^2 \). Following the neural network training, we drop the explicit penalty and run the gradient descent to minimize \( L(w, v) \).

To be more specific, for the sparse SIM, we propose to construct an estimator of \( \beta^* \) by applying gradient descent to \( L \) in (9) with respect to \( w \) and \( v \), 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 \( (w^\top, v^\top)^\top \) solving (9) are given by
\[
w_{t+1} = w_t - \eta \nabla_w L(w_t, v_t)
= w_t - \eta \left( w_t \odot w_t - v_t \odot v_t - \frac{1}{n} \sum_{i=1}^n S(x_i) y_i \right) \odot w_t, \tag{10}
\]
\[
v_{t+1} = v_t + \eta \nabla_v L(w_t, v_t)
= v_t + \eta \left( w_t \odot w_t - v_t \odot v_t - \frac{1}{n} \sum_{i=1}^n S(x_i) y_i \right) \odot v_t. \tag{11}
\]
Here $\eta > 0$ is the stepsize. By the parameterization of $\beta$, 
\{w_t, v_t\}_{t \geq 0}$ leads to a sequence of estimators $\{\beta_t\}_{t \geq 0}$ given by
\begin{equation}
\beta_{t+1} = w_{t+1} \odot w_{t+1} - v_{t+1} \odot v_{t+1}.
\end{equation}

Meanwhile, in terms of choosing initial values, since the zero vector is a stationary point of the algorithm, we cannot set the initial values of $w$ and $v$ to the zero vector. To use the structure of $\beta^*$, ideally we would like to initialize $w$ and $v$ such that they share the same sparsity pattern as $\beta^*$. That is, we would like to set the entries in the support of $\beta^*$ to nonzero values, and set those outside of the support to zero. However, such an initialization scheme is infeasible since the support of $\beta^*$ is unknown. Instead, we initialize $w_0$ and $v_0$ as $w_0 = v_0 = \alpha \cdot 1_{p \times 1}$, where $\alpha > 0$ is a small constant and $1_{p \times 1}$ is an all-one vector in $\mathbb{R}^p$. By setting $w_0 = v_0$, we equivalently set $\beta_0$ 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 nonzero entries $s = 5$. Let $S := \{i : |\beta^*_i| > 0\}$. The responses $\{y_i\}_{i=1}^n$ are generated from $y_i = f((x, \beta^*_i)) + \epsilon_i$, $i \in [n]$ with link functions $f_1(x) = x$ (linear regression) and $f_2(x) = \sin(x)$. Here we assume $\beta^*$ is $s$-sparse with $\beta_i = 1/\sqrt{s}, i \in S$, and $\{x_i\}_{i=1}^n$ are standard Gaussian random vectors. We over-parameterize $\beta$ as $w \odot w - v \odot v$ and set $w_0 = v_0 = 10^{-5} \cdot 1_{p \times 1}$. Then we update $w, v$ and $\beta$ regarding Equations (10)–(12) with stepsize $\eta = 0.01$. The evolution of the distance between our unnormalized iterates $\beta_t$ and $\mu^* \beta^*$, trajectories of $\beta_{j,t}$ for $j \in S$ and $\max_{j \in S} |\beta_{j,t}|$ are depicted in Figures 1 and 2.

From the simulation results given in Figures 1(a) and 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 $|\beta_{S, t}|$, still remains small, as shown in Figures 1(c) and 2(c). In other words, running gradient descent with respect to over-parameterized parameters helps us distinguish nonzero components from zero components, while applying gradient descent to the ordinary loss cannot.

It is worth noting that, with over-parameterization, there are $\Omega(2^p)$ stationary points of $L$ satisfying $\nabla_v L(w, v) = 0$.\[
\text{Figure 1.} \text{ With link function } f(x) = x, \text{ (a) characterizes the evolution of distance } \|\beta_t - \mu^* \beta^*\|_2^2 \text{ against iteration number } t; \text{ (b) depicts the trajectories } \beta_{j,t} (j \in S) \text{ for five nonzero components, and (c) presents the trajectory } \max_{j \in S} |\beta_{j,t}|.
\]

\[
\text{Figure 2.} \text{ With link function } f(x) = \sin(x), \text{ similar to Figure 1, here (a) characterizes the evolution of distance } \|\beta_t - \mu^* \beta^*\|_2^2 \text{ against iteration number } t; \text{ (b) depicts the trajectories } \beta_{j,t} (j \in S) \text{ for five nonzero components, and (c) presents the trajectory } \max_{j \in S} |\beta_{j,t}|.
\]
\[\nabla_v L(w, v) = 0_{p \times 1}, \text{ where } 0_{p \times 1} \text{ is the zero vector. To see this, for any subset } I \subseteq [p], \text{ we define vectors } w \text{ and } v \text{ as follows. For any } j \neq I, \text{ we set the } j \text{th entries of } w \text{ and } v \text{ to zero. Meanwhile, for any } j \in I, \text{ we choose } w_j \text{ and } v_j \text{ such that } w_j^2 - v_j^2 = n^{-1} \sum_{i=1}^n S(x_i)y_i, \text{ where } w_j, v_j, \text{ and } S(x_i) \text{ are the } j \text{th entries of } w, v, \text{ and } S(x), \text{ respectively. By direct computation, it can be shown that } (w, v) \text{ is a stationary point of } L, \text{ and thus, there are at least } 2^p \text{ 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^*. \text{ Such an intriguing observation captures the implicit regularization induced by the optimization algorithm and over-parameterization.}

### 3.1. Gaussian Design

In this section, we discuss over-parameterized SIM with Gaussian covariates. In this section, we assume the distribution of \( x \) in (6) is \( N(\mu, \Sigma) \), where both \( \mu \) and \( \Sigma \) are assumed known. Moreover, only in this section, 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 \( \mu^* = \mathbb{E}[f'(\langle x, \beta^* \rangle)] \neq 0 \) is a constant and the following two conditions hold.

(a) Covariance matrix \( \Sigma \) is positive-definite and has bounded spectral norm. To be more specific, there exist constants \( C_{\min} \) and \( C_{\max} \) such that \( C_{\min} I_p \leq \Sigma \leq C_{\max} I_p \) holds, where \( I_p \) is the identity matrix.

(b) Both \( \{f'(\langle x, \beta^* \rangle)\}_{i=1}^n \) and \( \{e_i\}_{i=1}^n \) are iid sub-Gaussian random variables, with the sub-Gaussian norms denoted by \( \|f\|_{\psi_2} = O(1) \) and \( \sigma = O(1) \), respectively. Here we let \( \|f\|_{\psi_2} \) denote the sub-Gaussian norm of \( f'(\langle x, \beta^* \rangle) \). In addition, we further assume that \( \|\mu^*\|/\|f\|_{\psi_2} = \Theta(1), \|\mu^*\|/\sigma = \Theta(1) \).

The score function for the Gaussian distribution \( N(\mu, \Sigma) \) is \( S(x) = \Sigma^{-1}(x - \mu) \) and **Assumption 1(a)** makes the Gaussian distributed covariates non-degenerate. **Assumption 1(b)** enables the empirical estimator \( n^{-1} \sum_{i=1}^n y_i S(x_i) \) to concentrate its expectation \( \mu^* \beta^* \) and also sets a lower bound on 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 \( f(x) = x, \sin x, \tan(x) \), etc. In addition, in Section 3.2, the assumption that both \( f'(\langle x, \beta^* \rangle) \) and the noise \( \epsilon \) 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 set \( S = \{i : |\beta_i| > 0\} \) into \( S_0 = \{i : |\beta_i| \geq C_s \sqrt{\log p/n}\} \) and \( S_1 = \{i : 0 < |\beta_i| < C_s \sqrt{\log p/n}\} \), 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 \( s_m = \min(s_0, s_1) \) be the smallest value of strong signals.

**Theorem 1.** Apart from **Assumption 1**, if we further let our initial value \( \alpha \) satisfy \( 0 < \alpha \leq M_0 / 2 \) and set stepsize \( \eta \leq n \leq 1/(12(\|\mu^*\| + M_0)) \) 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

\[
\frac{\|\beta_{T_1} - \mu^* \beta^*\|_{\psi_2}}{\|\Sigma^{1/2} \beta^*\|_{\psi_2}} \leq \frac{s_0 \log n}{n} + \frac{s_1 \log p}{n},
\]

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

\[
\frac{\|\beta_{T_1} - \mu^* \beta^*\|}{\|\Sigma^{1/2} \beta^*\|_{\psi_2}} \leq \frac{s_0 \log n}{n} + \frac{s_1 \log p}{n}.
\]

**Theorem 2 (Variable Selection Consistency).** Under the setting of **Theorem 1**, for all \( T_1 \in \left[a_1 \log(1/\alpha)/\eta(\|\mu^*\|s_m - M_0 \sqrt{\log p/n})\right], a_2 \log(1/\alpha) \sqrt{n/\log p}/(\eta M_0)\right] \), we let \( \tilde{\beta}_{T_1} = [\beta_{T_1}] \cdot I[|\beta_{T_1}| > 0] \), for all \( i \in [p] \). Then, with probability at least \( 1 - 2p^{-1} - 2n^{-2} \), for all \( \lambda \in [\alpha, (C_s \|\mu^*\| - 2M_0) \sqrt{\log p/n}] \), we have \( \text{supp}(\tilde{\beta}_{T_1}) \subseteq \text{supp}(\beta^*) \). Moreover, when there then only exists strong signals in \( S_0 \). We further have \( \text{supp}(\tilde{\beta}_{T_1}) = \text{supp}(\beta^*) \) and \( \text{sign}(\tilde{\beta}_{T_1}) = \text{sign}(\beta^*) \).

**Theorem 1** shows that if we just have strong signals, then with high probability, for any \( T_1 \in \left[a_1 \log(1/\alpha)/\eta(\|\mu^*\|s_m - M_0 \sqrt{\log p/n})\right], a_2 \log(1/\alpha) \sqrt{n/\log p}/(\eta M_0)\right] \), we get the oracle statistical rate \( O(\sqrt{\log n/n}) \) in terms of the \( \ell_2 \)-norm, which is independent of the ambient dimension \( p \). Besides, when \( \beta^* \)}
also consists of weak signals, we achieve $O(\sqrt{s \log p / n})$ statistical rate in terms of the $\ell_2$-norm, where $s$ is the sparsity of $\beta^*$. Such a statistical rate matches the minimax rate of sparse linear regression (Raskutti, Wainwright, and Yu 2011) and is thus, minimax optimal. Notice that the oracle rate is achievable via explicit regularization using folded concave penalties (Fan, Xue, and Zou 2014) such as SCAD (Fan and Li 2001) and MCP (Zhang 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) and Plan, Vershynin, and Yudovina (2017), which study high dimensional SIM with $\ell_1$-regularization, thanks to the implicit regularization phenomenon, we avoid bias brought by the $\ell_1$-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 $\beta_{1j}$ when $T_1$ falls in the optimal time interval, we are able to recover the support of $\beta^*$ with high probability. Comparing to existing literatures on support recovery via using explicit regularization on single index model (Neykov, Liu, and Cai 2016), our method offers a wider range for choosing tuning parameter $\lambda$ with a known left boundary, instead of only using $\lambda = \Theta(\sqrt{\log p / n})$. This efficiently reduces false discovery rate, see Section D.1, supplementary materials 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, Rebeschini, and Rosasco 2020). The detailed discussions are given in Section C.5, supplementary materials.

Theorems 1 and 2 generalizes the results in Zhao, Yang, and He (2019) and Vaškevičius, Kanade, and Rebeschini (2019) for the linear model to high-dimensional SIMs. In addition, to satisfy the RIP condition, their sample complexity is at least $O(s^2 \log p)$ if their covariate $x$ follows the Gaussian distribution. Whereas, by using the loss function in (9) motivated by the Stein’s identity (Stein 1972; Stein et al. 2004), the RIP condition is unnecessary in our analysis. Instead, our theory only requires that $n^{-1} \sum_{i=1}^n S(x_i) \cdot y_i$ concentrates at a fast rate. As a result, our sample complexity is $O(s \log p)$ for $\ell_2$-norm consistency, which is better than $O(s^2 \log p)$.

The ideas of proof behind Theorems 1 and 2 are as follows. First, we are able to control the strengths of error component, denoted by $\|\beta^* \cap S_0\|_\infty$, at the same order with the square root of their initial values until $O(\log(1/\alpha) \cdot \sqrt{n \log p / (\eta M_0)}$ steps. This gives us the right boundary of the stopping time $T_1$. Meanwhile, every entry of strong signal part $\beta_j \cap S_0$ grows at exponential rates to $\epsilon = O(\sqrt{\log n / n})$ accuracy around $\mu^* \beta^* \cap S_0$, within $O(\log(1/\alpha) / (\eta (\mu^* s_0 - M_0 \sqrt{\log p / n}))$ steps, which offers us the left boundary of the stopping time $T_1$.

Finally, we prove for weak signals, their strengths will not exceed $O(\sqrt{\log p / n})$ 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 $\beta_j$ properly, since we are able to control the error component of $\beta_j$ uniformly as mentioned above. See Appendix E.1, supplementary materials 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 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 $\beta_{1j}$, 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 use the training data to implement Algorithm 1 and get the estimator $\beta_1$ as well as the value of the training loss (9) at step $t$. We notice $\beta_j$ 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 flattened 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 $\beta_1 / \|\Sigma^{1/2} \beta_1\|_2$, $j \in [m]$ to fit the link function $f$. Let the obtained estimator be $\tilde{f}_j$. For the testing dataset, we perform out-of-sample prediction and get $m$ prediction losses:

$$l_j = \frac{1}{m_{\text{test}}} \sum_{i=1}^{m_{\text{test}}} \left| Y_i - \tilde{f}_j(\langle x_i, \beta_j \rangle / \|\Sigma^{1/2} \beta_j\|_2) \right|^2,$$

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

We remark that each $\tilde{f}_j$ can be obtained by any nonparametric regression methods. To show case our method, in the following, we apply univariate kernel regression to obtain each $\tilde{f}_j$ 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 $\beta$ and $n$ iid observations $\{y_i, x_i\}_{i=1}^n$ of the model. For simplicity of the technical analysis, we assume that $\hat{\beta}$ is independent of $\{y_i, x_i\}_{i=1}^n$, which can be achieved by data-splitting. Moreover, we assume that $\hat{\beta}$ is an estimator of $\beta^*$ such that

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

(13)
Our goal is to construct an estimate the regression function \( f((\cdot, \beta^*)) \) based on \( \hat{\beta} \) and \( \{y_i, x_i\}_i \).

Note that, when \( \beta^* \) is known, we can directly estimate \( f \) based on \( y_i \) and \( Z_i^* := x_i^T \beta^*, i \in [n] \) via standard nonparametric regression. When \( \hat{\beta} \) is accurate, a direct idea is to replace \( Z_i^* \) by \( Z_i := x_i^T \hat{\beta} \) and follow the similar route. For a new observation \( x \), we define \( Z := x^T \hat{\beta} \) and \( Z^* := x^T \beta^* \), respectively.

To predict \( Y \), we estimate function \( g(z) \) using kernel regression with data \( \{(y_i, x_i^T \hat{\beta})\}_i \).

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
\[
\hat{g}(Z) = \begin{cases} 
\frac{\sum_{i=1}^{n} y_i K_h(Z-Z_i)}{\sum_{i=1}^{n} K_h(Z-Z_i)}, & |Z - \mu^T \hat{\beta}| \leq R, \\
0, & \text{otherwise},
\end{cases}
\tag{14}
\]
where we follow the convention that \( 0/0 = 0 \). In what follows, we consider the \( \ell_2 \)-prediction risk of \( \hat{g} \), which is given by
\[
\mathbb{E} \left[ \left\{ \hat{g}(\langle x, \hat{\beta} \rangle) - f(\langle x, \beta^* \rangle) \right\}^2 \right],
\]
where the expectation is taken with respect to \( x \) and \( \{y_i, x_i\}_i \).

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 \( h = n^{-1/3} \) in (14), under Assumption 2, the \( \ell_2 \)-prediction risk of \( \hat{g} \) defined in (14) is given by
\[
\mathbb{E} \left[ \left\{ \hat{g}(\langle x, \hat{\beta} \rangle) - f(\langle x, \beta^* \rangle) \right\}^2 \right] \leq \frac{\text{polylog}(n)}{n^{2/3}},
\]
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 \( \hat{\beta} = \beta_{T_1} / \|\Sigma^{1/2} \beta_{T_1}\|_2 \) 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 Section E.3, supplementary materials. Note that it is possible to refine the analysis on the prediction risk for \( f \) with higher order derivatives by using higher order kernels (see Tsybakov 2008, therein) this is not the key message of our article.

### 3.2. General Design

In this section, we extend our methodology to the setting with covariates generated from a general distribution. Following our discussions at the beginning of Section 3, ideally we aim at solving the loss function with over-parameterized variable given in (9). However, when the distribution of \( x \) has density \( p_0 \), the score \( S(x) \) can be heavy-tailed such that \( \mathbb{E}[Y \cdot S(x)] \) 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 \( S(x) \) by their truncated (Winsorized) version \( \tilde{y}_i \) and \( \tilde{S}(x_i) \), respectively. Specifically, we propose to apply gradient descent to the following modified loss function with respect to \( u \) and \( v \):
\[
\min_{w,v} L(w, v) := \langle w \odot w - v \odot v, w \odot w - v \odot v \rangle + \frac{2}{n} \sum_{i=1}^{n} \tilde{y}_i \langle w \odot w - v \odot v, \tilde{S}(x_i) \rangle.
\tag{15}
\]

Let \( \tilde{a} \in \mathbb{R}^d \) denote the truncated version of vector \( a \in \mathbb{R}^d \) based on a parameter \( \tau \) (Fan et al. 2021). That is, its entries are given by \( [\tilde{a}]_j = [a]_j \) if \( |a|_j \leq \tau \) and \( \tau \) otherwise. Applying elementwise truncation to \( \{|y_i\}_i \) and \( \{|S(x_i)\}_i \) in (15), we allow the score \( S(x) \) and the response \( Y \) to both have heavy-tailed distributions. By choosing a proper threshold \( \tau \), such a truncation step ensures \( n^{-1} \sum_{i=1}^{n} \tilde{y}_i \tilde{S}(x_i) \) converge to \( \mathbb{E}[Y \cdot S(x)] \) with a desired rate in \( \ell_2 \)-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 Section E.5, supplementary materials.

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
\[
\mathbb{E}[Y^4] \leq M, \quad \mathbb{E}[S(x)^4] \leq M, \quad \forall j \in [p].
\]

Here \( S(x)_j \) is the \( j \)-th entry of \( S(x) \). Moreover, recall that we denote \( \mu^* = \mathbb{E}[f(\langle x, \beta^* \rangle)] \). We assume that \( \mu^* \) 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, Liu, and Wang 2018; Fan, Wang, and Zhong 2019; Fan, Wang, and Zhu 2021). 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 \( \tau = (M \cdot n/\log p)^{1/2} \), let the initialization parameter \( \alpha \) satisfy \( 0 < \alpha \leq M^2/\mu^2 / p \), and set the stepsize \( \eta \) such that \( 0 < \eta \leq 1/(12(|\mu^*| + M_\theta)) \) in Algorithm 3 given in Section E.5, supplementary materials where \( M_\theta \) is a constant proportional to \( M \). There exist absolute constants \( a_3, a_4 \) such that, with probability at least \( 1 - 2 \rho^{-2} \),
\[
\|\beta_{T_1} - \mu^* \beta^*\|_2^2 \leq \frac{8 \log p}{n}.
\]
holds for all $T_1 \in [a_2 \log(1/\alpha)/\eta(|\mu^*| s_m - M_g \sqrt{\log p/n})]$, $a_1 \log(1/\alpha)/\eta(|\mu^*| s_m - M_g \sqrt{\log p/n})$. Here $s$ is the cardinality of the support set $S$ and $s_m = \min_{i \in S_0} |\beta_i^*|$, where $S_0 = \{i \in [p] : |\beta_i| \geq C_\epsilon \sqrt{\log p/n}\}$ is the set of strong signals. In addition, for the normalized iterates, we further have

$$\frac{\|\beta_{T_1} - \beta_{T_1}^*\|_2}{\|\beta_{T_1}^*\|_2}^2 \leq \frac{\log p}{n},$$

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

Compared with Theorem 1 for the Gaussian design, here we achieve the $O(\sqrt{s \log p/n})$ statistical rate of convergence in terms of the $\ell_2$-norm. These rates are the same of those achieved by adding an $\ell_1$-norm regularization explicitly (Plan and Vershynin 2016; Plan, Vershynin, and Yudovina 2017; Yang, Balasubramanian, and Liu 2017) and are minimax optimal (Raskutti, Wainwright, and Yu 2011). Moreover, we note that here $S(x)$ 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 $C_\epsilon$ can be set as a sufficiently large absolute constant, and the statistical rates established in Theorem 4 holds for all choices of $C_\epsilon$. In addition, for heavy-tailed case, we also let $[\beta_{T_1}^*]_i = [\beta_{T_1}]_i - [1/\beta_{T_1}]_i \epsilon_{\beta_T}^\alpha$ for all $i \in [p]$. Then for all $\alpha \in [\alpha, (C_\epsilon|\mu^*| - 2M_g \sqrt{\log p/n})$, 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 Section 3, here we also focus on matrix SIM with first-order links, that is, we assume that $\mu^* = E[f'(\langle X, \beta^* \rangle)] \neq 0$, where $\beta^*$ is a low rank matrix with rank $r$. Note that we assume that the entries of covariate $X \in \mathbb{R}^{d \times d}$ are iid with a univariate density $p_0$. Also recall that we define the score function $S(\beta) \in \mathbb{R}^{d \times d}$ in (3). Then, similar to the loss function in (9), we consider the loss function

$$L(\beta) = \langle \beta, \beta \rangle - 2 \langle \beta, \frac{1}{n} \sum_{i=1}^{n} y_i S(X_i) \rangle,$$

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

$$L(W, V) := \langle WW^T - VV^T, WW^T - VV^T \rangle - 2 \langle WW^T - VV^T, \frac{1}{n} \sum_{i=1}^{n} y_i S(X_i) \rangle.$$  

Since the rank of $\beta^*$ is unknown, we initialization $W_0$ and $V_0$ as $W_0 = V_0 = \alpha \cdot I_{d \times d}$ for a small $\alpha > 0$ and construct a sequence of iterates $\{W_t, V_t, \beta_t\}_{t \geq 0}$ via the gradient descent method as follows:

$$W_{t+1} = W_t - \eta (WW_t^T - VV_t^T - \frac{1}{2n} \sum_{i=1}^{n} S(X_i)y_i)$$

$$- \frac{1}{2n} \sum_{i=1}^{n} S(X_i)y_i) W_t,$$

$$V_{t+1} = V_t + \eta (WW_t^T - VV_t^T - \frac{1}{2n} \sum_{i=1}^{n} S(X_i)y_i)$$

$$- \frac{1}{2n} \sum_{i=1}^{n} S(X_i)y_i) V_t,$$

$$\beta_{t+1} = W_t W_t^T - V_t V_t^T,$$

where $\eta$ 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 $\beta^*$ with near-optimal statistical rates of convergence.

Similar to the vector case, for theoretical analysis, here we also divide eigenvalues of $\beta^*$ into different groups by their strengths. We let $\tau_i^*; i \in [d]$ be the $i$th eigenvalue of $\beta^*$. The support set $R$ of the eigenvalues is defined as $R := \{i : |\tau_i^*| > 0\}$, whose cardinality is $r$. We then divide the support set $R$ into $R_0 := \{i : |\tau_i^*| > C_{m_2}(\sqrt{\log d/n})\}$ and $R_1 := \{i : 0 < |\tau_i^*| < C_{m_2}(\sqrt{\log d/n})\}$, which correspond to collections of strong and weak signals with cardinality denoting by $r_0$ and $r_1$, respectively. Here $C_{m_2} > 0$ 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, that is, $r_m = \min_{i \in R} |\tau_i^*|$.  

4.1. Gaussian Design

In this section, we focus on the model in (7) with the entries of covariate $X$ being iid $N(0, 1)$ random variables. In this case, $S(X_i) = X_i$. This leads to Algorithm 4 given in Section F1, supplementary materials, where we place $S(X_i)$ by $X_i$ in (16)-(18).

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

Assumption 4. We assume that $\mu^* = E[f'(\langle X, \beta^* \rangle)]$ is a nonzero constant. Moreover, we assume that both $\{f'(\langle X_i, \beta^* \rangle)\}_{i=1}^{n}$ and $\{\epsilon_i\}_{i=1}^{n}$ are iid sub-Gaussian random variables, with sub-Gaussian norm denoted by $\|f\|_{\psi_2} = \Theta(1)$ and $\sigma = O(1)$, respectively. Here we let $\|f\|_{\psi_2}$ denote the sub-Gaussian norm of $f(X, \beta^*)$. 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 $0 < \alpha \leq M_m^2/d$ and stepsize $0 < \eta \leq 1/[12(\mu^* + M_m)]$ in Algorithm 4, where $M_m$ is a constant proportional to $\max(|\mu|, \sigma)$. Under Assumption 4, there exist
constants $a_5, a_6$ such that, with probability at least $1 - 1/(2d) - 3/n^2$, we have
\[
\|\beta_T - \mu^* \beta^*\|_F^2 \leq \frac{rd \log d}{n}
\]
for all $T_1 \in [a_5 \log(1/\alpha)/\eta(|\mu^*| r_m - M_m \sqrt{d \log d/n}), a_6 \log(1/\alpha) \sqrt{n/(d \log d)/|\eta M_m|}]$. Moreover, for the normalized iterates $\beta_{T_1}/\|\beta_{T_1}\|_F$, we have
\[
\frac{\|\beta_{T_1} - \mu^* \beta^*\|_F^2}{\|\beta_{T_1}\|_F^2} \leq \frac{rd \log d}{n}.
\]

Similar to the vector case given in Section 3.1, as shown in the proof in Appendix F, supplementary materials, here we require $C_{ms}$ to satisfy $C_{ms} \geq \max\{a_5/a_6 + 1\}M_m/|\mu^*|, 2M_m/|\mu^*|\}$ 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 $\alpha$, stepsize $\eta$, 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 minmax lower bound (Roehde 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, Vershynin, and Yudovina 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 $\ell_2$-statistical rate, we are able to recover the true rank with high-probability by properly truncating the eigenvalues of $\beta_{T_1}$ for all $T_1 \in [a_5 \log(1/\alpha)/\eta(|\mu^*| r_m - M_m \sqrt{d \log d/n}), a_6 \log(1/\alpha) \sqrt{n/(d \log d)/|\eta M_m|}]$. Comparing with the literature Lee, Sun, and Taylor (2015) which studies the rank consistency via $\ell_1$-regularization, we offer a wider range for choosing the tuning parameter with known left boundary $\alpha$, instead of only setting the nuclear tuning parameter $\lambda = \bar{\Theta}(\sqrt{rd/n})$.

**Theorem 6 (Rank Consistency).** Under the setting of Theorem 5, for all
\[
T_1 \in [a_5 \log(1/\alpha)/\eta(|\mu^*| r_m - M_m \sqrt{d \log d/n}), a_6 \log(1/\alpha) \sqrt{n/(d \log d)/|\eta M_m|}],
\]
we let $\tilde{\beta}_{T_1} = \sum_{i=1}^d u_i u_i^* \lambda_i \cdot \mathbb{1}_{\{\lambda_i(\tilde{\beta}_{T_1}) \geq \lambda\}},$ for all $i \in [d]$. Here $u_i, k \in [d]$ are eigenvectors of $\beta_{T_1}$. Then, with probability at least $1 - 2d^{-1} - 3n^{-2}$, for all $\lambda \in [\alpha, (C_{ms})^t|\mu^*| - 2M_m \sqrt{d \log d/n}]$, we have $\tilde{\beta}_{T_1}$ enjoys the conclusion of Theorem 5, and rank$(\tilde{\beta}_{T_1}) \leq$ rank$(\beta^*)$. Moreover, when there only exists strong signals in $R_0$, we further have rank$(\tilde{\beta}_{T_1}) = $ 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, Ma, and Zhang 2018; Arora et al. 2019; Gidel, Bach, and Lacoste-Julien 2019). Specifically, we allow a more general class of (noisy) models and symmetric signal matrices. Compared with Li, Ma, and Zhang (2018), our methodology possesses several strengths, which include achieving low sample complexity ($\tilde{O}(rd)$ instead of $O(r^2d)$), allowing weak signals (min$_{i \in R} |\beta_i|^q \geq \tilde{O}((1/n)^{1/2})$ instead of min$_{i \in R} |\beta_i|^q \geq O((1/n)^{1/6})$), getting tighter statistical rate under noisy models ($\tilde{O}(dr/n)$ instead of $\tilde{O}(krd/n)$), 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 Section 3.1.2. The only difference between them is that here we replace $x^T \beta^*$ by tr$(X^T \beta^*)$. Indeed, as we assume $|\Sigma_i^{1/2} \beta_i|^2 = 1$ in vector SIM and $|\Sigma_i^{1/2} \beta_i|^2 = 1$ in matrix version for model identifiability, both $x^T \beta_i$ and tr$(X^T \beta_i)$ follow the standard normal distribution. Thus, our results on the prediction risk in Section 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 $X$ are iid random variables with a known density function $p_0: \mathbb{R} \to \mathbb{R}$. Recall that, according to the results following Definition 1, the score function $S(X) \in \mathbb{R}^{d \times d}$ is defined as
\[
S(X)_{jk} := S(X_{jk}) = -p_0'(X_{jk})/p_0(X_{jk}),
\]
where $S(X)_{jk}$ and $X_{jk}$ are the $(j,k)$th entries of $S(X)$ and $X$ for all $j, k \in [d]$. However, similar to the results in Section 3.2, the entries of $S(X)$ can have heavy-tailed distributions and thus, $n^{-1} \sum_{i=1}^n y_i \cdot S(X_i)$ may not converge its expectation $\mathbb{E}[Y \cdot S(X)]$ efficiently in terms of spectral norm. Here $X_i$ is the $i$th observation of the covariate $X$. To tackle such a challenge, we employ a shrinkage approach (Catoni 2012; Minsker 2018; Fan, Wang, and Zhu 2021) to construct a robust estimator of $\mathbb{E}[Y \cdot S(X)]$. Specifically, we let
\[
\phi(x) = \begin{cases} 
\log(1 - x + x^2/2), & x \leq 0, \\
\log(1 + x + x^2/2), & x > 0,
\end{cases}
\]
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 2012; Brownlees, Joly, and Lugosi 2015). For any matrix $X \in \mathbb{R}^{d \times d}$, we apply spectral decomposition to its Hermitian dilation and obtain
\[
X^*: = \begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix} = Q \Sigma^* Q^T,
\]
where $\Sigma^* \in \mathbb{R}^{2d \times 2d}$ is a diagonal matrix. Based on such a decomposition, we define $\bar{X} = \phi(\Sigma^*) Q^T$, where $\phi$ applies
elementwisely to $\Sigma^*$. Then we write $\tilde{X}$ as a block matrix as

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

where each block of $\tilde{X}$ 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 $\phi_1(X) := \tilde{X}_{12}$, which is a regularized version of $X$. Given data $y_i, X_i$, we finally define $H(\cdot)$ as

$$
H(y_i S(X_i), \kappa) := 1 / \kappa \cdot \phi_1(\kappa y_i \cdot S(X_i)), \quad \forall \kappa > 0, \tag{19}
$$

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

$$
L(W, V) := \langle WW^T - VV^T, WW^T - VV^T \rangle \\
- \frac{2}{n} \sum_{i=1}^{n} \langle WW^T - VV^T, H(y_i S(X_i), \kappa) \rangle. \tag{20}
$$

After over-parameterizing $\beta$ as $WW^T - VV^T$, we propose to construct an estimator of $\beta^*$ by applying gradient descent on the following loss function in (20) with respect to $W, V$. See Algorithm 5 in Section E.5, supplementary materials 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 $S(X)$ have bounded fourth moments. Specifically, there exists an absolute constant $M$ such that

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

Moreover, we assume that $\mu^* = \mathbb{E}[f'(\langle X, \beta^* \rangle)]$ 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 $\kappa$ in (19) as $\kappa = \sqrt{\log(4d)/(nd - M)}$ and let the initialization parameter $\alpha$ and the stepsize $\eta$ satisfy $0 < \alpha \leq M_{mg}^2 / d$ and $0 < \eta \leq 1/[12(\|\mu^*\| + M_{mg})]$, where $M_{mg}$ is a constant proportional to $M$. Then, under Assumption 5, there exist absolute constants $\alpha_7, \alpha_8$ such that, with probability at least $1 - (4d)^{-2}$, we have

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

for all $T_1 \in [\alpha_7 \log(1/\alpha)/(\eta(\|\mu^*\| r_m - M_{mg} \sqrt{d \log d/n}))$, $\alpha_8 \log(1/\alpha)/(\sqrt{d \log d/n}(\eta M_{mg}))$]. Moreover, for the normalized iterate $\beta_t/\|\beta_t\|_F$, we have

$$
\frac{\|\beta_{T_1} - \mu^* \|_F}{\|\beta_{T_1}\|_F} \leq \frac{rd \log d}{n}.
$$

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 $H$ in (19) can also be viewed as a data preprocessing 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 $\ell_2$-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, Balasubramanian, and Liu (2017), Goldstein, Minsker, and Wei (2018) and 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 $\tilde{\beta}_{T_1} = \sum_{i=1}^d \mu_i \hat{X}_{i} \cdot \lambda_i(\tilde{\beta}_{T_1}) \cdot \|\lambda_i(\tilde{\beta}_{T_1})\|_2$, then for all $\lambda \in [\alpha, (C_{\mu} |\mu^*| - 2M_{mg}) \sqrt{d \log d/n}]$, we achieve the same results with Theorem 6.

### 5. Conclusion

In this article, 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 $\ell_2$-statistical rate and variable selection consistency.

### Supplementary Materials

The supplementary material provides all technical proofs of main theorems as well as corresponding numerical studies.

### Funding

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.
**ORCID**

Mengxin Yu http://orcid.org/0000-0002-6818-4083

**References**

Arora, S., Cohen, N., Hu, W., and Luo, Y. (2019), “Implicit Regularization in Deep Matrix Factorization,” in Advances in Neural Information Processing Systems, pp. 7411–7422. [2316,2325]

Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. (2017), “Deep Reinforcement Learning: A Brief Survey,” IEEE Signal Processing Magazine, 34, 26–38. [2315]

Brownlees, C., Joly, E., and Lugosi, G. (2015), “Empirical Risk Minimization for Heavy-Tailed Losses,” Annals of Statistics, 43, 2507–2536. [2325]

Candès, E. J. (2008), “The Restricted Isometry Property and its Implications for Compressed Sensing,” Comptes rendus Mathematique, 9, 589–592. [2316,2318]

Candès, E. J., Liu, H., and Wang, W. (2018), “Large Covariance Estimation Through Elliptical Factor Models,” in Advances in Neural Information Processing Systems, pp. 2939–2941. [2315]

Donoho, D. L. (2006), “Compressed Sensing,” IEEE Transactions on Information Theory, 52, 1289–1306. [2316]

Fan, J., and Li, R. (2001), “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties,” Journal of the American Statistical Association, 96, 1348–1360. [2316,2322]

Fan, J., Xue, L., and Zou, H. (2014), “Strong Oracle Optimality of Folded Concave Penalized Estimation,” Annals of Statistics, 42, 819–849. [2316,2322]

Fan, J., Liu, H., and Wang, W. (2018), “Large Covariance Estimation Through Elliptical Factor Models,” Annals of Statistics, 46, 1383–1414. [2323]

Fan, J., Wang, W., and Zhong, Y. (2019), “Robust Covariance Estimation for Approximate Factor Models,” Journal of Econometrics, 208, 5–22. [2323]

Fan, J., Ma, C., and Zhong, Y. (2021), “A Selective Overview of Deep Learning,” Statistical Science, 36, 264–290. [2315]

Fan, J., Wang, K., Zhong, Y., and Zhu, Z. (2021), “Robust High Dimensional Factor Models with Applications to Statistical Machine Learning,” Statistical Science, 36, 303–327. [2317,2323]

Fan, J., Wang, W., and Zhu, Z. (2021), “A Shrinkage Principle for Heavy-Tailed Data: High-dimensional Robust Low-Rank Matrix Recovery,” Annals of Statistics, 49, 1239–1266. [2317,2323,2325]

Gidel, G., Bach, F., and Lacoste-Julien, S. (2019), “Implicit Regularization of Discrete Gradient Dynamics in Linear Neural Networks,” in Advances in Neural Information Processing Systems, pp. 3196–3206. [2325]

Goldstein, L., and Wei, X. (2019), “Non-Gaussian Observations in Non-linear Compressed Sensing via Stein Discrepancies,” Information and Inference: A Journal of the IMA, 8, 125–159. [2316]

Goldstein, L., Minsker, S., and Wei, X. (2018), “Structured Signal Recovery from Non-linear and Heavy-tailed Measurements,” IEEE Transactions on Information Theory, 64, 5513–5530. [2316,2326]

Goodfellow, I., Bengio, Y., and Courville, A. (2016), Deep Learning. Cambridge, MA: MIT Press. [2315]

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2017), “Implicit Regularization in Matrix Factorization,” in Advances in Neural Information Processing Systems, pp. 6151–6159. [2315,2318,2325]

Hoff, P. D. (2017), “Lasso, Fractional Norm and Structured Sparse Estimation Using a Hadamard Product Parametrization,” Computational Statistics & Data Analysis, 115, 186–198. [2316,2019]

Ke, Y., Minsker, S., Ren, Z., Sun, Q., Zhou, W.-X. (2019), “User-Friendly Covariance Estimation for Heavy-Tailed Distributions,” Statistical Science, 34, 454–471. [2317]

Keskär, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2017), “On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima,” in International Conference on Learning Representations. [2315]

LeCun, Y., Bengio, Y., and Hinton, G. (2015), “Deep Learning,” Nature, 521, 436–444. [2315]

Lee, J.-D., Sun, Y., and Taylor, J. E. (2015), “On Model Selection Consistency of Regularized M-Estimators,” Electronic Journal of Statistics, 9, 608–642. [2325]

Li, Y. (2017), “Deep Reinforcement Learning: An Overview,” arXiv preprint arXiv:1701.07274. [2315]

Li, Y., Ma, T., and Zhang, H. (2018), “Algorithmic Regularization in Over-Parameterized Matrix Sensing and Neural Networks with Quadratic Activations,” in COLT. [2315,2316,2318,2325]

Li, Z., Luo, Y., and Lyu, K. (2021), “Towards Resolving the Implicit Bias of Gradient Descent for Matrix Factorization: Greedy Low-Rank Learning,” in International Conference on Learning Representations. [2318]

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. Y. (2017), “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Volume 54 of Proceedings of Machine Learning Research, eds. A. Singh and J. Zhu, pp. 1273–1282. PMLR. [2322]

Minsker, S. (2018), “Sub-Gaussian Estimators of the Mean of a Random Matrix with Heavy-Tailed Entries,” Annals of Statistics, 46, 2871–2903. [2317,2325]

Minsker, S., and Wei, X. (2020), “Robust Modifications of U-Statistics and Applications to Covariance Estimation Problems,” Bernoulli, 26, 694–727. [2317]

Na, S., Yang, Z., Wang, Z., and Kolar, M. (2019), “High-Dimensional Varying Index Coefficient Models via Stein’s Identity,” Journal of Machine Learning Research, 20, 1–44. [2326]

Neykov, M., Liu, J. S., and Cai, T. (2016), “ε1-Regularized Least Squares for Support Recovery of High Dimensional Single Index Models with Gaussian Designs,” Journal of Machine Learning Research, 17, 2976–3012. [2322]

Neyshabur, B., Tomioka, R., and Srebro, N. (2015), “In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning,” in International Conference on Learning Representations. [2315]

Neyshabur, B., Tomioka, R., Salakhutdinov, R., and Srebro, N. (2017), “Geometry of Optimization and Implicit Regularization in Deep Learning,” arXiv preprint arXiv:1705.03071. [2315]

Otter, D. W., Medina, J. R., and Kalita, J. K. (2020), “A Survey of the Uses of Deep Learning in Natural Language Processing,” IEEE Transactions on Neural Networks and Learning Systems, pp. 1–21. [2315]

Plan, Y., and Vershynin, R. (2016), “The Generalized Lasso with Non-linear Observations,” IEEE Transactions on Information Theory, 62, 1528–1537. [2316,2319,2322,2324,2325,2326]

Plan, Y., Vershynin, R., and Yudovina, E. (2017), “High-Dimensional Estimation with Geometric Constraints,” Information and Inference: A Journal of the IMA, 6, 1–40. [2316,2319,2322,2324,2325]

Poggio, T., Kawaguchi, K., Liao, Q., Miranda, B., Rosasco, L., Sohl-Dickstein, J., Hidary, J., and Mnih, A. (2017), “Theory of Deep Learning III: Explaining the Non-overfitting Puzzle,” arXiv preprint arXiv:1801.00173. [2315]

Raskutti, G., Wainwright, M. J., and Yu, B. (2011), “Minimax Rates of Estimation for High-Dimensional Linear Regression Over ℓ_q-balls,” IEEE Transactions on Information Theory, 57, 6976–6994. [2322,2324]

Richards, D., and Rebeschini, P. (2020), “Graph-Dependent Implicit Regularisation for Distributed Stochastic Subgradient Descent,” Journal of Machine Learning Research, 21, 1–44. [2322]

Richards, D., Rebeschini, P., and Rosasco, L. (2020), “Decentralised Learning with Distributed Gradient Descent and Random Features,” arXiv preprint arXiv:2007.00360. [2322]

Rohde, A., and Tsybakov, A. B. (2011), “Estimation of High-Dimensional Low-Rank Matrices,” Annals of Statistics, 39, 887–930. [2315,2326]

Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., and Segev, M. (2015), “Phase Retrieval with Application to Optical Imaging: A Contemporary Overview,” IEEE Signal Processing Magazine, 32, 87–109. [2316]
Stein, C., Diaconis, P., Holmes, S., and Reinert, G. (2004), “Use of Exchangeable Pairs in the Analysis of Simulations,” in Stein’s Method, pp. 1–25. Institute of Mathematical Statistics. [2316,2318,2322]

Stein, C. (1972), “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 (Vol. 2), The Regents of the University of California. [2316,2322]

Swirszcz, G., Czarnecki, W. M., and Pascanu, R. (2016), “Local Minima in Training of Neural Networks,” in International Conference on Learning Representations. [2315]

Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavvaf, N., and Fox, E. A. (2020), “Natural Language Processing Advancements by Deep Learning: A Survey,” arXiv preprint arXiv:2003.01200. [2315]

Tsybakov, A. B. (2008), Introduction to Nonparametric Estimation, New York: Springer. [2323]

Vaškevičius, T., Kanade, V., and Rebeschini, P. (2019), “Implicit Regularization for Optimal Sparse Recovery,” in Advances in Neural Information Processing Systems, pp. 2972–2983. [2316,2318,2322]

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018), “Deep Learning for Computer Vision: A Brief Review,” Computational Intelligence and Neuroscience, 2018, 1–13. [2315]

Wei, X. (2018), “Structured Recovery with Heavy-Tailed Measurements: A Thresholding Procedure and Optimal Rates,” arXiv preprint arXiv:1804.05959. [2316]

Wei, X., and Minsker, S. (2017), “Estimation of the Covariance Structure of Heavy-Tailed Distributions,” in Advances in Neural Information Processing Systems, pp. 2859–2868. [2317]

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. (2017), “The Marginal Value of Adaptive Gradient Methods in Machine Learning,” in Advances in Neural Information Processing Systems, pp. 4148–4158. [2315]

Yang, Z., Balasubramanian, K., and Liu, H. (2017), “High-Dimensional Non-Gaussian Single Index Models Via Thresholded Score Function Estimation,” in International Conference on Machine Learning, pp. 3851–3860. JMLR. [2316,2319,2324,2326]

Yun, C., Sra, S., and Jadabaie, A. (2019), “Small Nonlinearities in Activation Functions Create Bad Local Minima in Neural Networks,” in International Conference on Learning Representations. [2315]

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017), “Understanding Deep Learning Requires Rethinking Generalization,” in International Conference on Learning Representations. [2315]

Zhang, C.-H. (2010), “Nearly Unbiased Variable Selection Under Minimax Concave Penalty,” Annals of Statistics, 38, 894–942, 2010. [2316,2322]

Zhao, P., Yang, Y., and He, Q.-C. (2019), “Implicit Regularization via Hadamard Product Over-Parametrization in High-Dimensional Linear Regression,” arXiv preprint arXiv:1903.09367. [2316,2318,2322]

Zhu, Z. (2017), “Taming the Heavy-Tailed Features by Shrinkage and Clipping,” arXiv preprint arXiv:1710.09020. [2317]