Journal of Machine Learning Research X (XXXX) X-XX Submitted X/XX; Published XX/XX

High-Dimensional Model Selection via Chebyshev Greedy Algorithms

You-Lin Chen YOULINCHEN@UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago, IL 60615, USA

Chi-Shian Dai CDAI39@WISC.EDU
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706, USA

Ching-Kang Ing CKING@STAT.NTHU.EDU.TW
Institute of Statistics
National Tsing Hua University
Hsinchu, 30013, Taiwan

Editor: XXX

Abstract

We propose a two-stage method CGA+HDAIC, which can automatically adapt under any $\ell_{1/r}$-sparsity($r \in [1, \infty]$), to select the best models with respect to the prediction error in high-dimensional weakly sparse generalized linear models. Furthermore, CGA+HDAIC can achieve the minimax prediction rate $(\log p/n)^{1-1/2r}$ under regularity condition. The applications include logistic regression, poisson regression, quantile regression.

Keywords: Chebyshev Greedy Algorithms, Orthogonal Matching Pursuit, high-dimensional generalized linear models, high-dimensional information criterion, minimax prediction rate

1. Introduction

Past decade have brought about a flurry of work on models selection for high-dimensional statistical linear or nonlinear models which have broad application in a variety of important fields such as bioinformatics, quantitative fiance, image process and advanced manufacturing. Let data $\{y_t, \boldsymbol{x}_t\}_{t=1}^n$ be drawn form a distribution, $P_{\boldsymbol{\beta}^*}$, parameterized by a p-dimensional vector of unknown parameters $\boldsymbol{\beta}^*$, where y_t is the response variable and $\boldsymbol{x}_t = (x_{t,1}, \ldots, x_{t,p})$ represents the predictor/explanatory variables, and p is allowed to be much larger than n. To make inference about $\boldsymbol{\beta}^*$ or predict the future values of y_t, we consider a empirical loss function:

$$\ell_n(\boldsymbol{\beta}|\boldsymbol{y}_t, \boldsymbol{x}_t, t = 1, \ldots, n) \equiv \frac{1}{n}\sum_{t=1}^{n} \gamma(\boldsymbol{\beta}, y_t, \boldsymbol{x}_t), \tag{1}$$

in which $\gamma(\cdot)$ is assumed to be convex. This empirical loss function should be viewed as a surrogate to the population risk function:

$$\ell(\boldsymbol{\beta}) = \mathrm{E}_{(y,\boldsymbol{x})\sim P_{\boldsymbol{\beta}^*}}[\gamma(\boldsymbol{\beta}, y, \boldsymbol{x})]$$

The γ in (1) can be negative log-likelihood functions (Negahban et al. (2012)), quantile regression loss functions (Belloni and Chernozhukov (2011)), or hinge loss of the support vector machine (Peng, Wang, and Wu (2016)).

When $p \gg n$, it is impossible to solve this problem directly due to the identifiability and computation issues. The key insight alleviating the difficulty is that extra structures on true parameters and corresponding convex conditions in the loss function might expected in practice. In this paper, we focus on discussing the case of "sparsity". Typically, $\boldsymbol{\beta}^*$ is assumed to be weakly sparse with the sparsity level $r \in [1, \infty]$ such that

$$\|\boldsymbol{\beta}^*\|_{1/r}^r := \sum |\beta_j^*|^{1/r} < \infty$$

We also consider a more general weak sparse condition in which the case $\beta_i = i^{-r}, i = 1, \ldots$ satisfy the sparsity level r. See Assumption (A2) for precise definition. Moreover, since finding such solution is NP hard problem in general, it is necessary to impose some regular conditions on loss function such as Restricted Strong Convexity/Smoothness Property () which ensure the identifiability of sparse solutions on a restricted domain and is known as Nullspace Property, Restricted Eigenvalue Property, Restricted Isometry Property in the linear case. See Jain and Kar (2017) and the reference therein. However, Convexity/Smoothness Property is the key to generalize methods in high-dimensional statistics to nonlinear models.

High-dimensional statistical models have been discussed for many years and most of all focus on selecting the correct variables (Ing and Lai (2011)) or screening out the irrelevant variables (Fan and Lv (2008)). However, in the weakly sparse models, none of model is correct and most of the variables are relevant. Hence, finding a model to achieve minimax prediction becomes the crucial goal in this field of study. In the study of weakly sparsity, "what the minimax rate is" and "how to attain it" have been discussed for years. In the high-dimensional linear weakly sparse models, Raskutti, Wainwright, and Yu (2012) derives the minimax rate $(\log p/n)^{1-1/2r}$, and Negahban et al. (2012) claims Lasso have ability to achieve that rate. Furthermore, Ing (2019) use different approach (OGA) to get the same prediction rate under weaker sparsity assumption (same as Assumption (A2) in this paper). In nonlinear weakly sparse models, Abramovich and Grinshtein (2016) proposes ℓ_0-constraint have minimax rate in the generalized linear exponential family models under strong sparsity.

In this paper, we are interested in the weakly sparsity as well as more general nonlinear models. Inspiring by Ing (2019) who propose a stepwise method, called the orthogonal greedy algorithm (OGA), in high-dimensional linear models to achieve minimax prediction rate in weakly sparse case and model selection consistency in strong sparse case, we propose the corresponding method Chebyshev Greedy Algorithms (CGA) in the non-linear case, which can apply on wider class of field. We show that CGA with high-dimensional information criterion, whose details is in Section 2, can automatically adapt under high-dimensional nonlinear models with the unknown sparsity level and achieve the same minimax rate as the linear regression. Furthermore, models selection consistency is also provided in our theory.

Chebyshev Greedy Algorithms (CGA) a.k.a. Orthogonal Matching Pursuit (OMP) is a pursuit-style algorithm working by beginning with an empty set and adding the feature or variable with the largest negative partial derivative at each step. The greedy search rule is also referred to as Gauss-Southwell rule Luo and Tseng (1992) in the literature, and OMP is called Chebyshev greedy algorithm in approximation theory Temlyakov (2015). There are two main methodologies to solve the high-dimensional problems: convex relaxation (regularization methods) and non-convex method (greedy methods or projected method). The former like lasso adds the penalty which is convex relaxation of L_0 norm to lose function.

It becomes very popular in the last decades due to its unified theory and easy manipulation: the practitioner can apply to any applications by adding any type of convex penalty depending on the structure of data. From the view of statistics, it combines estimation and model selection in one procedure. However, distorting the original loss function makes interpretation more obscured. For the theoretical side, studying consistency property in the convex relaxation method is the troublesome mission. For the computational side, introducing a nonsmooth regularizer bring the difficulty in optimization and the advanced optimization methods to solve convex relaxation problems usually is slow in large scale data. On the other hand, non-convex methods become more and more popular recently due to the reasons that non-convex methods is suitable for large scale data and its solutions are more easy to interpret. Pursuit-style algorithms enjoy better consistency properties as well as benefits of nonconvex methods such as interpretation and have the same convergence rate as convex relaxation methods. Via promising theoretical results and strong simulation evidence in our paper, we believe CGA have great potential for future research.

2. Methodology

With setting (1), we state the sample CGA. Let $J \subset \{1, 2, \dots, p\}$ and $J^c = \{1, 2, \dots, p\} \setminus J$. Define

$$\hat{\boldsymbol{\beta}}_J = \arg \min_{\boldsymbol{\beta}=(\beta_1,\dots,\beta_p)^\top \in \mathbb{B},\ \boldsymbol{\beta}(J^c)=\mathbf{0}} \ell_n(\boldsymbol{\beta}),$$

where $\mathbb{B} \subset R^p$ is the domain of $\ell_n(\cdot)$ and $\boldsymbol{\beta}(J) = (\beta_j, j \in J)^\top$.

Algorithm 1 Chebyshev Greedy Algorithm

```
procedure Chebyshev Greedy Algorithm
    $\hat{J}_0 \leftarrow \varnothing$, $\hat{\boldsymbol{\beta}}_{\hat{J}_0} \leftarrow 0$
    for $m$ in $1 : K_n$ do                                        ▷ CGA
        $\hat{j}_m \leftarrow \arg\max_{1 \le j \le p_n} |\nabla_j \ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_{m-1}})|$
        $\hat{J}_m \leftarrow \hat{J}_{m-1} \cup \{\hat{j}_m\}$
        $\hat{\boldsymbol{\beta}}_{\hat{J}_m} \leftarrow \arg\min_{\boldsymbol{\beta}_{\hat{J}^c_m = 0}} \ell_n(\boldsymbol{\beta})$,
    end for
end procedure
```

The algorithm greedily chooses the direction in which the sample loss function ℓ_n decreases most rapidly. In order to avoid over fitting, it is crucial to determine the number of CGA iterations. We therefore introduce the high-dimensional Akaike information criterion(HDAIC)

$$\text{HDAIC}(J) = \ell_n(\hat{\boldsymbol{\beta}}_J) + |J|\omega\frac{\log p}{n}, \tag{2}$$

where $J \subset \{1, \dots, p\}$, $|J|$ is the cardinality of set J, ω is a tuning parameter, and $\log p/n$ is the penalty term. Our data-driven selector of the number of CGA iterations based on the HDAIC is given as follows:

$$\hat{m}_n = \arg\min_{1\le m\le K_n} \text{HDAIC}(\hat{J}_m), \tag{3}$$

where K_n is a prescribed upper bound for the number of iterations.

3. Asymptotic Theory of CGA under Weak Sparsity

In this section, we derive convergence rates for the population and sample versions of Chebyshev greedy algorithm under weak sparsity, which are detailed in Sections 3.1 and 3.2, respectively.

3.1 Convergence Rates of the Population CGA

Let $\ell(\cdot)$ be a convex function on $\mathbb{B} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_1 \le M_1\}$, where M_1 is a large positive constant and $\|\boldsymbol{\nu}\|_1$ denotes the L_1 norm of vector $\boldsymbol{\nu}$. Define

$$\boldsymbol{\beta}^* = \arg\inf_{\boldsymbol{\beta}\in\mathbb{B}} \ell(\boldsymbol{\beta}),\ \boldsymbol{\beta}_J = \arg\inf_{\boldsymbol{\beta}\in\mathbb{B}:\boldsymbol{\beta}(J^c)=0} \ell(\boldsymbol{\beta}),$$

where $J \subset \{1, \dots, p\}$ and $J \ne \emptyset$. Also, define $\boldsymbol{\beta}_J = \mathbf{0} \in R^p$, if $J = \emptyset$. Let $0 < \xi \le 1$. The population weak CGA is an iterative scheme that chooses indices $j_1, j_2, \dots$ sequentially from $\{1, \dots, p\}$ according to the recursive relation:

$$|\nabla_{j_m}\ell(\boldsymbol{\beta}_{J_{m-1}}) \ge \xi \max_{1\le j\le p} |\nabla_j \ell(\boldsymbol{\beta}_{J_{m-1}})|, \tag{4}$$

where $J_0 = \varnothing$, $J_m = J_{m-1} \cup \{j_m\}$.

To analyze the convergence of $\ell(\boldsymbol{\beta}_{J_m}) - \ell(\boldsymbol{\beta}^*)$ with $m \le K$ for some positive integer K, we impose the following conditions. Let $|J|$ denotes the cardinality of set J.

(A1) $\ell(\cdot)$ is continuous on $\mathbb{B}$ and differentiable at any interior point of $\mathbb{B}$. In addition, $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}_J$ are interior points of $\mathbb{B}$, for any $|J| \leq K$.

(A2) (Weak Sparsity) There is a constant $r \in [1, \infty]$ and a positive number C_r dependent r such that for all $J \subset \{1, \ldots, p_n\}$

$$\|\boldsymbol{\beta}_J^*\|_1 < C_r \{\|\boldsymbol{\beta}_J^*\|_2^2\}^{\frac{r-1}{2r-1}}, \tag{5}$$

where $\|\cdot\|_2$ denotes the Euclidean norm and for $r = \infty$, the exponent on the right-hand side of (5) is set to 1/2.

(A3) There exists $M_2 > 0$ such that for any $J \subset \{1, \ldots, p\}$ with $1 \leq |J| \leq p - 1$, and any $\lambda \in R$,

$$\max_{1 \leq j \leq p} \{\ell(\boldsymbol{\beta}_J + \lambda e_j) - \ell(\boldsymbol{\beta}_J) - \lambda \nabla^\top \ell(\boldsymbol{\beta}_J) e_j\} \leq \frac{M_2 \lambda^2}{2}, \tag{6}$$

where e_j is the j-th coordinate unit vector.

(A4) There exist small positive numbers ϵ and δ and an integer K such that $B_\epsilon(\boldsymbol{\beta}^*) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 < \epsilon\} \subset \mathbb{B}$ and $B_\epsilon(\boldsymbol{\beta}_J) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_J\|_2 < \epsilon\} \subset \mathbb{B}$, for any $J \subset \{1, \ldots, p\}$ with $|J| \leq K$. Moreover, for any $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ in $B_\epsilon(\boldsymbol{\beta}_J)$ or in $B_\epsilon(\boldsymbol{\beta}^*)$,

$$\ell(\boldsymbol{\nu}_2) - \ell(\boldsymbol{\nu}_1) - \nabla^\top \ell(\boldsymbol{\nu}_1)(\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1) \geq \frac{\delta}{2} \|\boldsymbol{\nu}_2 - \boldsymbol{\nu}_1\|_2^2. \tag{7}$$

Remark 3.1 *Discuss (A2)-(A4) here.*

Theorem 3.1 *Assume that (A1)-(A4) hold. Then, for all $1 \leq m \leq K$,*

$$\ell(\boldsymbol{\beta}_{J_m}) - \ell(\boldsymbol{\beta}^*) \leq C_r^* m^{1-2r}, \ \textit{provided } 1 \leq r < \infty, \tag{8}$$

and

$$\ell(\boldsymbol{\beta}_{J_m}) - \ell(\boldsymbol{\beta}^*) \leq C_1^* \exp(-C_2^* m), \ \textit{provided } r = \infty. \tag{9}$$

Here, C_r^* (dependent on r), C_1^* and C_2^* are some positive constants.

Proof We first consider the case of $1 \leq r < \infty$. Let $0 < s \leq \frac{\epsilon}{2M_1}$. Then,

$$\|s(\boldsymbol{\beta}_{J_{m-1}} - \boldsymbol{\beta}^*)\|_2 \leq s\|\boldsymbol{\beta}_{J_{m-1}} - \boldsymbol{\beta}^*\|_1 \leq \epsilon,$$

and hence by (7) in (A4),

$$\ell(\boldsymbol{\beta}^* + s(\boldsymbol{\beta}_{J_{m-1}} - \boldsymbol{\beta}^*)) - \ell(\boldsymbol{\beta}^*) - s\nabla^\top \ell(\boldsymbol{\beta}^*)(\boldsymbol{\beta}_{J_{m-1}} - \boldsymbol{\beta}^*) \geq \frac{\delta s^2}{2}\|\boldsymbol{\beta}_{J_{m-1}} - \boldsymbol{\beta}^*\|_2^2. \tag{10}$$

In addition, the convexity of $\ell(\cdot)$ implies

$$\ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^* + s(\boldsymbol{\beta}_{J_{m-1}} - \boldsymbol{\beta}^*)) \geq 0,$$

which, together with (10), yields

$$\begin{aligned}
&\ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*) = \ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*) - \nabla^\top \ell(\boldsymbol{\beta}^*)(\boldsymbol{\beta}_{J_{m-1}} - \boldsymbol{\beta}^*) \\
&= \ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^* + s(\boldsymbol{\beta}_{J_{m-1}} - \boldsymbol{\beta}^*)) - (1-s)\nabla^\top \ell(\boldsymbol{\beta}^*)(\boldsymbol{\beta}_{J_{m-1}} - \boldsymbol{\beta}^*) \\
&+ \ell(\boldsymbol{\beta}^* + s(\boldsymbol{\beta}_{J_{m-1}} - \boldsymbol{\beta}^*)) - \ell(\boldsymbol{\beta}^*) - s\nabla^\top \ell(\boldsymbol{\beta}^*)(\boldsymbol{\beta}_{J_{m-1}} - \boldsymbol{\beta}^*) \\
&\geq \frac{\delta s^2}{2}\|\boldsymbol{\beta}_{J_{m-1}} - \boldsymbol{\beta}^*\|_2^2 \geq \frac{\delta s^2}{2}\|\boldsymbol{\beta}^*_{J^c_{m-1}}\|_2^2.
\end{aligned} \tag{11}$$

In addition, it follows from (A1), (A2), the convexity of $\ell(\cdot)$, and (11) that

$$\begin{aligned}
&\ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*) \leq |\nabla^\top \ell(\boldsymbol{\beta}_{J_{m-1}})(\boldsymbol{\beta}_{J_{m-1}} - \boldsymbol{\beta}^*)| = |\nabla^\top \ell(\boldsymbol{\beta}_{J_{m-1}})\boldsymbol{\beta}^*_{J^c_{m-1}}| \\
&\leq \|\nabla \ell(\boldsymbol{\beta}_{J_{m-1}})\|_\infty \|\boldsymbol{\beta}^*_{J^c_{m-1}}\|_1 \leq \|\nabla \ell(\boldsymbol{\beta}_{J_{m-1}})\|_\infty C_r \|\boldsymbol{\beta}^*_{J^c_{m-1}}\|_2^{(2r-2)/(2r-1)} \\
&\leq C_r \|\nabla \ell(\boldsymbol{\beta}_{J_{m-1}})\|_\infty \{\delta^{*^{-1}}(\ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*))\}^{\frac{r-1}{2r-1}},
\end{aligned} \tag{12}$$

where $\delta^* = \delta s^2/2$. Consequently,

$$\|\nabla \ell(\boldsymbol{\beta}_{J_{m-1}})\|_\infty \geq A_r(\ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*))^{\frac{r}{2r-1}}, \tag{13}$$

where $A_r = \delta^{*^{(r-1)/(2r-1)}}/C_r$. Relations (13) and (4) imply

$$|\nabla_{j_m} \ell(\boldsymbol{\beta}_{J_{m-1}})| \geq \xi \max_{1 \leq j \leq p} |\nabla_j \ell(\boldsymbol{\beta}_{J_{m-1}})| \geq \xi A_r(\ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*))^{\frac{r}{2r-1}}. \tag{14}$$

Since for any $\lambda \in R$, $\ell(\boldsymbol{\beta}_{J_{m-1}} + \lambda e_{jm}) \geq \ell(\boldsymbol{\beta}_{J_m})$, this, together with (A3) and $\nabla^\top \ell(\boldsymbol{\beta}_{J_{m-1}})e_{jm} = \nabla_{j_m}\ell(\boldsymbol{\beta}_{J_{m-1}})$, yields

$$\begin{aligned}
&\ell(\boldsymbol{\beta}_{J_m}) - \ell(\boldsymbol{\beta}^*) \leq \ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*) + \ell(\boldsymbol{\beta}_{J_{m-1}} + \lambda e_{jm}) - \ell(\boldsymbol{\beta}_{J_{m-1}}) \\
&\leq \ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*) + \lambda \nabla_{j_m} \ell(\boldsymbol{\beta}_{J_{m-1}}) + M_2\lambda^2/2.
\end{aligned} \tag{15}$$

Setting $\lambda = S \equiv (M_2)^{-1}\xi A_r(\ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*))^{r/(2r-1)}$ if $\nabla_{j_m}\ell(\boldsymbol{\beta}_{J_{m-1}}) \le 0$, and $\lambda = -S$ if $\nabla_{j_m}\ell(\boldsymbol{\beta}_{J_{m-1}}) > 0$, we obtain by (15) and (14) that

$$\begin{aligned} \ell(\boldsymbol{\beta}_{J_m}) - \ell(\boldsymbol{\beta}^*) &\le \ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*) - S|\nabla_{j_m}\ell(\boldsymbol{\beta}_{J_{m-1}})| + M_2S^2/2 \\ &\le \ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*) - S\xi A_r(\ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*))^{\frac{r}{2r-1}} + M_2S^2/2 \\ &\le \left\{\ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*)\right\}\left\{1 - \frac{(\xi A_r)^2}{2M_2}(\ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*))^{\frac{1}{2r-1}}\right\}. \end{aligned} \tag{16}$$

By (16) and Lemma 1 of Gao, Ing, and Yang (2013), we obtain for $1 \le m \le K$,

$$\ell(\boldsymbol{\beta}_{J_m}) - \ell(\boldsymbol{\beta}^*) \le C_r^* m^{1-2r},$$

where $C_r^* = \max\{2^{(2r-1)^2}\{(\xi A_r)^2/(2M_2(2r-1))\}^{1-2r}, \ell(0) - \ell(\boldsymbol{\beta}^*)\}$. Thus, (8) follows.

Next we deal with the case of $r = \infty$. By an argument similar to those used to derive (16), we have

$$\ell(\boldsymbol{\beta}_{J_m}) - \ell(\boldsymbol{\beta}^*) \le (\ell(\boldsymbol{\beta}_{J_{m-1}}) - \ell(\boldsymbol{\beta}^*))\left\{1 - \frac{(\xi A_\infty)^2}{2M_2}\right\},$$

where $A_\infty = {\delta^*}^{1/2}/C_\infty$, and hence

$$\ell(\boldsymbol{\beta}_{J_m}) - \ell(\boldsymbol{\beta}^*) \le (\ell(0) - \ell(\boldsymbol{\beta}^*))\left\{1 - \frac{(\xi A_\infty)^2}{2M_2}\right\}^m,$$

where, without loss of generality, we assume $(\xi A_\infty)^2/(2M_2) < 1$. As a result, (9) holds with $C_1^* = \ell(0) - \ell(\boldsymbol{\beta}^*)$ and $C_2^* = -\log\{1 - (\xi A_\infty)^2/(2M_2)\}$, with log denoting the natural logarithm. ■

Remark 3.2 *Discuss (A2)–(A4) here.*

3.2 Convergence Rates of the Sample CGA

The convergence rate of population version WCGA strongly depend on a decent path of variables by which the population gradient decreasing with a suitable rate (4). We assume the following two conditions.

(U) $\ell_n(\boldsymbol{\beta})$ is (a.s.) continuous on $\mathbb{B}$ and differentiable at any interior point of $\mathbb{B}$. Moreover,

$$P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}} \|\nabla\ell_n(\boldsymbol{\beta}) - \nabla\ell(\boldsymbol{\beta})\|_\infty \le s_0\sqrt{\mathcal{R}_{n,p}}\right) \to 1, \tag{17}$$

and

$$P\left(|\ell_n(\boldsymbol{\beta}^*) - \ell(\boldsymbol{\beta}^*)| \leq s_0\sqrt{\mathcal{R}_{n,p}}\right) \to 1, \tag{18}$$

where s_0 is some positive constant and $\mathcal{R}_{n,p}$, depending only on the sample size, n, and the number of covariates, p, converges to 0 at a certain rate.

(C) For any $|J| \leq K_n$, $\nabla\ell(\cdot)$ is continuously differentiable on $B_\epsilon(\boldsymbol{\beta}_J)$, where $B_\epsilon(\boldsymbol{\beta}_J)$ is defined in (A4) and $K_n = O(\mathcal{R}_{n,p}^{-1/2})$. In addition, there is an $M_3 > 0$ such that

$$\max_{|J|\leq K_n, i\notin J} \sup_{\boldsymbol{\beta}\in B_\epsilon(\boldsymbol{\beta}_J)} \left\| \left(\int_0^1 \nabla^2_{JJ}\ell(t\boldsymbol{\beta} + (1-t)\boldsymbol{\beta}_J)dt\right)^{-1} \left(\int_0^1 \nabla_{Ji}\ell(t\boldsymbol{\beta} + (1-t)\boldsymbol{\beta}_J)dt\right)\right\|_1 \leq M_3. \tag{19}$$

Remark 3.3 *$\mathcal{R}_{n,p}$ in Condition (U) is the uniformly rate of Law of Large number. In Section 5, we show that under bounded covariate $\mathcal{R}_{n,p} = \frac{\log p}{n}$ in the generalized linear model. Condition (C) describe the relationship between the selected variables J and the unselected variables J^c. If the dimension p is fixed, Condition (C) is automatically satisfied since there are only finite many combinations of J and J^c. Hence, we can view Condition (C) as controlling ”the future models”, which make sure even the dimension is growing the correlation between J and J^c cannot be too crazy. In the linear model, Condition (C) can be written as a beautify form, which is the same as Assumption (A5) in Ing (2019).*

Theorem 3.2 *Assume that* (A1)–(A3), (A4) *with* $K = K_n$, (C) *and* (U) *hold true. Then, for* $1 \leq r < \infty$,

$$\frac{\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}^*)}{m^{1-2r} + m\mathcal{R}_{n,p}} = O_p(1), \tag{20}$$

and for $r = \infty$,

$$\frac{\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}^*)}{\exp(-C^*m) + m\mathcal{R}_{n,p}} = O_p(1), \tag{21}$$

where C^* is some positive constant.

Proof .We only prove (20) because the proof of (21) is similar. Our proof is divided into three steps.

Step 1: Bias Analysis

In this step, we establish a bound for $\ell(\boldsymbol{\beta}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}^*)$ uniform over $1 \leq m \leq K_n$ First define

$$\begin{aligned} A_n(m) &= \Big\{ \max_{|J| \leq m-1} \|\nabla \ell_n(\hat{\boldsymbol{\beta}}_J) - \nabla \ell(\boldsymbol{\beta}_J)\|_\infty \leqslant s^* \mathcal{R}_{n,p}^{1/2} \Big\}, \\ B_n(m) &= \Big\{ \min_{0 \leqslant i \leqslant m-1} \|\nabla \ell(\boldsymbol{\beta}_{\hat{J}_i})\|_\infty > \bar{\xi} s^* \mathcal{R}_{n,p}^{1/2} \Big\}, \end{aligned}$$

where $m \geq 1$, s^* is a positive constant defined in Lemma **??**, and $\bar{\xi} = 2/(1-\xi)$ with $0 < \xi < 1$ being arbitrarily chosen. Note first that by Lemma **??**,

$$\lim_{n \to \infty} P(A_n^c(K_n)) = 0. \tag{22}$$

On the set $A_n(m) \bigcap B_n(m)$, we have for $1 \leqslant k \leqslant m$,

$$\begin{aligned} |\nabla_{\hat{j}_k} \ell(\boldsymbol{\beta}_{\hat{J}_{k-1}})| &\geqslant -|\nabla_{\hat{j}_k} \ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_{k-1}}) - \nabla_{\hat{j}_k} \ell(\boldsymbol{\beta}_{\hat{J}_{k-1}})| + |\nabla_{\hat{j}_k} \ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_{k-1}})| \\ &\geqslant -\max_{|J| \leq m-1} \|\nabla \ell_n(\hat{\boldsymbol{\beta}}_J) - \nabla \ell(\boldsymbol{\beta}_J)\|_\infty + \|\nabla \ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_{k-1}})\|_\infty \\ &\geqslant -s^* \mathcal{R}_{n,p}^{1/2} - \|\nabla \ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_{k-1}}) - \nabla \ell(\boldsymbol{\beta}_{\hat{J}_{k-1}})\|_\infty + \|\nabla \ell(\boldsymbol{\beta}_{\hat{J}_{k-1}})\|_\infty \\ &\geqslant -2s^* \mathcal{R}_{n,p}^{1/2} + \|\nabla \ell(\boldsymbol{\beta}_{\hat{J}_{k-1}})\|_\infty \geq \xi \|\nabla \ell(\boldsymbol{\beta}_{\hat{J}_{k-1}})\|_\infty, \end{aligned}$$

showing that the sample CGA is indeed a population weak CGA on $A_n(m) \bigcap B_n(m)$. Therefore, Theorem 3.1 ensures that on $A_n(m) \bigcap B_n(m)$,

$$\ell(\boldsymbol{\beta}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}^*) \leqslant C_r^* m^{1-2r}. \tag{23}$$

On the other hand, we obtain from (13) that

$$\begin{aligned} &\ell(\boldsymbol{\beta}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}^*) \leq \min_{0 \leqslant i \leqslant m-1} \ell(\boldsymbol{\beta}_{\hat{J}_i}) - \ell(\boldsymbol{\beta}^*) \\ &\leq (A_r)^{-\frac{2r-1}{r}} \min_{0 \leqslant i \leqslant m-1} \|\nabla \ell(\boldsymbol{\beta}_{\hat{J}_i})\|_\infty^{\frac{2r-1}{r}} \leq \left(\frac{\bar{\xi} s^*}{A_r} \right)^{\frac{2r-1}{r}} \mathcal{R}_{n,p}^{\frac{2r-1}{2r}} \quad \text{on} \quad B_n^c(m). \end{aligned} \tag{24}$$

Combining (23) and (24) yields

$$\ell(\boldsymbol{\beta}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}^*) \leqslant V_r^* (m^{1-2r} + \mathcal{R}_{n,p}^{1-1/2r}) \;\; \text{on} \;\; A_n(m), \tag{25}$$

where $V_r^* = \max\{C_r^*, (\bar{\xi} s^*/A_r)^{\frac{2r-1}{r}}\}$. This, together with (22) and $A_n(K_n) \subset A_n(m)$ for $1 \leq m \leq K_n$, gives

$$\lim_{n \to \infty} P\Big(\max_{1 \leqslant m \leqslant K_n} \frac{\ell(\boldsymbol{\beta}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}^*)}{m^{1-2r} + \mathcal{R}_{n,p}^{1-1/2r}} \leq V_r^* \Big) = 1. \tag{26}$$

Step2: Variance Analysis

In this step, we provide a bound for $\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}_{\hat{J}_m})$ uniform over $1 \le m \le K_n$. By making use of (A4), we show in Lemma **??** that for any $\boldsymbol{\theta}, \boldsymbol{\eta} \in B_\epsilon(\boldsymbol{\beta}_J)$ with $|J| \le K_n$,

$$\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\eta}) \le \nabla^\top \ell(\boldsymbol{\eta})(\boldsymbol{\theta} - \boldsymbol{\eta}) + \frac{1}{2\delta}\|\nabla \ell(\boldsymbol{\theta}) - \nabla \ell(\boldsymbol{\eta})\|_2^2. \tag{27}$$

Therefore, on the set $H_n^* \bigcap I_n$, where $H_n^* = \{\max_{|J| \le K_n} \|\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J\|_2 < \epsilon\}$ and $I_n = \{\sup_{\boldsymbol{\beta} \in \mathbb{B}} \|\nabla \ell_n(\boldsymbol{\beta}) - \nabla \ell(\boldsymbol{\beta})\|_\infty \le s_0\sqrt{\mathcal{R}_{n,p}}\}$, (27) yields that for all $1 \le m \le K_n$,

$$\begin{aligned} 0 \le \ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}_{\hat{J}_m}) &\le \frac{1}{2\delta}\|\nabla \ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m}) - \nabla \ell(\boldsymbol{\beta}_{\hat{J}_m})\|_2^2 \\ = \frac{1}{2\delta}\|\nabla_{\hat{J}_m} \ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m})\|_2^2 &= \frac{1}{2\delta}\|\nabla_{\hat{J}_m} \ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m}) - \nabla_{\hat{J}_m} \ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_m})\|_2^2 \\ \le \frac{m}{2\delta} \sup_{\boldsymbol{\beta} \in \mathbb{B}} \|\nabla \ell_n(\boldsymbol{\beta}) - \nabla \ell(\boldsymbol{\beta})\|_\infty^2 &\le \frac{m s_0^2}{2\delta}\mathcal{R}_{n,p} \end{aligned} \tag{28}$$

By Lemma **??** and condition (U), $\lim_{n\to\infty} P(H_n^* \bigcap I_n) = 1$, which, together with (28), gives

$$\lim_{n\to\infty} P\Big(\max_{1 \leqslant m \leqslant K_n} \frac{\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}_{\hat{J}_m})}{m\mathcal{R}_{n,p}} \le \frac{s_0^2}{2\delta}\Big) = 1. \tag{29}$$

Step3: Combining the Bias and Variance

By

$$\begin{aligned} &\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}^*) = (\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}_{\hat{J}_m})) - (\ell(\boldsymbol{\beta}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}^*)), \\ &\mathcal{R}_{n,p}^{1-1/2r} \le m^{1-2r} + m\mathcal{R}_{n,p} \quad \text{for all} \quad 1 \le m \le K_n, \end{aligned}$$

(26) and (29), there exists some $D > 0$ such that

$$\lim_{n\to\infty} P\Big(\max_{1 \leqslant m \leqslant K_n} \frac{\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}^*)}{m^{1-2r} + m\mathcal{R}_{n,p}} \le D\Big) = 1, \tag{30}$$

leading directly to (20). ■

Remark 3.4 *Discuss Thm 3.1 here.*

4. Analysis of CGA+HDAIC

In this section, we provide the convergence rate of prediction error for CGA+HDAIC under several types of sparsity. The main result are stated and proved in Section 4.1.

4.1 Convergence rate of prediction error for CGA+HDAIC

We need a slightly strengthened version of condition (C).

(C*) Condition (C) holds. Moreover, there is an $M_4 > 0$ such that

$$\max_{|J|\le K_n, i\notin J} \sup_{\boldsymbol{\beta}\in B_\epsilon(\boldsymbol{\beta}^*)} \left\| \left[\int_0^1 \nabla_{JJ}\ell(t\boldsymbol{\beta}+(1-t)\boldsymbol{\beta}^*)dt\right]^{-1} \left[\int_0^1 \nabla_{Ji}\ell(t\boldsymbol{\beta}+(1-t)\boldsymbol{\beta}^*)dt\right]\right\|_1 \le M_4, \tag{31}$$

where $B_\epsilon(\boldsymbol{\beta}^*)$ is defined in (A4) and $K_n = O(\mathcal{R}_{n,p}^{-1/2})$ with $\mathcal{R}_{n,p}$ defined in (U).

Remark 4.1 *Discuss (C*) here.*

Consider a variant of (2)

$$\text{HDAIC}(J) = \ell_n(\hat{\boldsymbol{\beta}}_J) + |J|\omega\mathcal{R}_{n,p}, \tag{32}$$

in which $\mathcal{R}_{n,p}$ is used in place of $\log p/n$. Our data-driven selector of the number of CGA iterations based on the HDAIC is given as follows:

$$\hat{k}_n = \arg\min_{1\le m\le K_n} \text{HDAIC}(\hat{J}_m). \tag{33}$$

Theorem 4.1 *Let the assumptions of Theorem 3.2 hold except that* (C) *is strengthened to* (C*)*. Then, for ω in* (32) *satisfying*

$$\omega > \frac{s_0^2}{\delta}, \tag{34}$$

where s_0 is defined in (U) and δ is defined in (A4), the data-driven selector $\hat{k}_n$ in (33) *possesses the following properties:*

$$\frac{\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_{\hat{k}_n}}) - \ell(\boldsymbol{\beta}^*)}{\mathcal{R}_{n,p}^{1-1/2r}} = O_p(1), \quad \textit{provided} \;\; 1 \le r < \infty, \tag{35}$$

$$\frac{\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_{\hat{k}_n}}) - \ell(\boldsymbol{\beta}^*)}{(-\log\mathcal{R}_{n,p})\mathcal{R}_{n,p}} = O_p(1), \quad \textit{provided} \;\; r = \infty, \tag{36}$$

and

$$\frac{\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_{\hat{k}_n}}) - \ell(\boldsymbol{\beta}^*)}{\mathcal{R}_{n,p}} = O_p(1), \quad \textit{provided} \tag{37}$$

$$\min_{j \in N_n} |\beta_j^*| > \underline{\theta}, \tag{38}$$

where $N_n = \{1 \leqslant j \leqslant p : \beta_j^* \neq 0\}$ *and* $\underline{\theta}$ *is an arbitrarily small positive constant.*

Proof We only prove (35). The proof of (36) is omitted because it is similar to that of (35). The proof of (37) is deferred to Appendix. Without loss of generality, assume $K_n > \mathcal{R}_{n,p}^{-1/2r} \equiv m_n^*$. Define

$$\mathrm{BA}_n = \left\{ \max_{1 \leqslant m \leqslant K_n} \frac{\ell(\boldsymbol{\beta}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}^*)}{m^{1-2r} + \mathcal{R}_{n,p}^{1-1/2r}} \leq V_r^* \right\}, \ \mathrm{VA}_n = \left\{ \max_{1 \leqslant m \leqslant K_n} \frac{\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}_{\hat{J}_m})}{m\mathcal{R}_{n,p}} \leq \frac{s_0^2}{2\delta} \right\},$$
$$\mathrm{TA}_n = \left\{ \max_{1 \leqslant m \leqslant K_n} \frac{\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}^*)}{m^{1-2r} + m\mathcal{R}_{n,p}} \leq D \right\},$$

where V_r^* and D are defined in the proof of Theorem 3.2. It is shown in the proof of Theorem 3.2 that $\lim_{n\to\infty} P(\mathrm{BA}_n \bigcap \mathrm{VA}_n \bigcap \mathrm{TA}_n \bigcap I_n \bigcap H_n^*) = 1$, recalling that I_n and H_n^* are defined after (27). By an argument used in (11), one obtains $\max_{m_n^* \leq k \leq K_n} \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_{\hat{J}_k}\|_2^2 \leq {\delta^*}^{-1}(\ell(\boldsymbol{\beta}_{\hat{J}_{m_n^*}}) - \ell(\boldsymbol{\beta}^*)) \leq 2{\delta^*}^{-1} V_r^* \mathcal{R}_{n,p}^{1-1/2r}$ on the set BA_n, yielding $\lim_{n\to\infty} P\big(L_n\big) = 1$, where $L_n = \{\max_{m_n^* \leq k \leq K_n} \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_{\hat{J}_k}\|_2^2 \leq \epsilon\}$ and ϵ is defined in (A4). In addition, we show in Lemma **??** that there is some $D^* > 0$ such that

$$\lim_{n\to\infty} P\big(G_n^*\big) = 1, \tag{39}$$

where $G_n^* = \{\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_{\hat{J}_m}\|_1 \leq D^* \|\boldsymbol{\beta}^*_{\hat{J}_m^c}\|_1 \text{ for all } 1 \leq m \leq K_n\}$. Let $G > 2V_r^*$ be a large constant. Define

$$\tilde{k}_n = \min\{1 \leq m \leq K_n : \ell(\boldsymbol{\beta}_{\hat{J}_k}) - \ell(\boldsymbol{\beta}^*) \leq G\mathcal{R}_{n,p}^{1-1/2r}\} \ (\min \emptyset = K_n). \tag{40}$$

Then $\lim_{n\to\infty} P(\tilde{k}_n \leq m_n^*) = 1$ due to $\ell(\boldsymbol{\beta}_{\hat{J}_{m_n^*}}) - \ell(\boldsymbol{\beta}^*) \leq 2V_r^* \mathcal{R}_{n,p}^{1-1/2r}$ on BA_n. As a result,

$$\lim_{n\to\infty} P(S_n^*) = 1, \tag{41}$$

where $S_n^* = \mathrm{BA}_n \cap \mathrm{VA}_n \cap \mathrm{TA}_n \cap I_n \cap H_n^* \cap G_n^* \cap L_n \cap \{\tilde{k}_n \leq m_n^*\}$.

Our proof of (35) is divided into three steps.

Step 1: Prove

$$\lim_{n\to\infty} P(\hat{k}_n < \tilde{k}_n,\ S_n^*) = 0. \tag{42}$$

By (41), it suffices for (42) to show that

$$\lim_{n\to\infty} P\big(\min_{1\le k<\tilde{k}_n} \mathrm{HDAIC}(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) > \mathrm{HDAIC}(\hat{\boldsymbol{\beta}}_{\hat{J}_{m_n^*}}),\ S_n^*\big) = 1. \tag{43}$$

We prove (43) by first decomposing $\ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) - \ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_{m_n^*}})$ as follows:

$$\begin{aligned}
&\ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) - \ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_{m_n^*}}) = \{\ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) - \ell_n(\boldsymbol{\beta}_{\hat{J}_k}) - [\ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_{m_n^*}}) - \ell_n(\boldsymbol{\beta}_{\hat{J}_{m_n^*}})]\} \\
&+ \{\ell_n(\boldsymbol{\beta}_{\hat{J}_k}) - \ell_n(\boldsymbol{\beta}^*) - (\ell(\boldsymbol{\beta}_{\hat{J}_k}) - \ell(\boldsymbol{\beta}^*)) - [\ell_n(\boldsymbol{\beta}_{\hat{J}_{m_n^*}}) - \ell_n(\boldsymbol{\beta}^*) - (\ell(\boldsymbol{\beta}_{\hat{J}_{m_n^*}}) - \ell(\boldsymbol{\beta}^*))]\} \\
&+ \{\ell(\boldsymbol{\beta}_{\hat{J}_k}) - \ell(\boldsymbol{\beta}^*) - [\ell(\boldsymbol{\beta}_{\hat{J}_{m_n^*}}) - \ell(\boldsymbol{\beta}^*)]\} \equiv \{\mathrm{I}_k\} + \{\mathrm{II}_k\} + \{\mathrm{III}_k\} \\
&\ge \{\mathrm{I}_k'\} + \{\mathrm{II}_k\} + \{\mathrm{III}_k\},
\end{aligned} \tag{44}$$

where $(\mathrm{I}_k') = \ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) - \ell_n(\boldsymbol{\beta}_{\hat{J}_k})$. It is straightforward to see that on the set S_n^*,

$$\mathrm{III}_k \ge (1 - 2V_r^*/G)(\ell(\boldsymbol{\beta}_{\hat{J}_k}) - \ell(\boldsymbol{\beta}^*)),\ 1 \le k < \tilde{k}_n. \tag{45}$$

By (A4), the mean value theorem, and the Cauchy-Schwarz inequality, we have on the set S_n^*,

$$\begin{aligned}
|\mathrm{I}_k'| &\le \sup_{\boldsymbol{\beta}\in\mathbb{B}} \|\nabla\ell_n(\boldsymbol{\beta}) - \nabla\ell(\boldsymbol{\beta})\|_\infty \sqrt{k} \|\hat{\boldsymbol{\beta}}_{\hat{J}_k} - \boldsymbol{\beta}_{\hat{J}_k}\|_2 + \ell(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) - \ell(\boldsymbol{\beta}_{\hat{J}_k}) \\
&\le \sqrt{k\mathcal{R}_{n,p}} s_0 \sqrt{2/\delta} (\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) - \ell(\boldsymbol{\beta}_{\hat{J}_k}))^{1/2} + \ell(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) - \ell(\boldsymbol{\beta}_{\hat{J}_k}) \\
&\le \frac{3s_0^2}{2\delta} k\mathcal{R}_{n,p} \le \frac{3s_0^2}{2\delta} \mathcal{R}_{n,p}^{1-1/2r} \le \frac{3s_0^2}{2\delta} G^{-1}(\ell(\boldsymbol{\beta}_{\hat{J}_k}) - \ell(\boldsymbol{\beta}^*)),\ 1 \le k < \tilde{k}_n.
\end{aligned} \tag{46}$$

By (A2), (A4), and an argument similar to that used in (46), we have on the set S_n^*,

$$\begin{aligned}
&|\mathrm{II}_k| \le s_0 \sqrt{\mathcal{R}_{n,p}} (\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_{\hat{J}_k}\|_1 + \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_{\hat{J}_{m_n^*}}\|_1) \\
&\le s_0 D^* \sqrt{\mathcal{R}_{n,p}} (\|\boldsymbol{\beta}^*_{\hat{J}_k^c}\|_1 + \|\boldsymbol{\beta}^*_{\hat{J}_{m_n^*}}\|_1) \\
&\le s_0 D^* \sqrt{\mathcal{R}_{n,p}} C_r \{(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^*_{\hat{J}_k}\|_2^2)^{\frac{r-1}{2r-1}} + (\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^*_{\hat{J}_{m_n^*}}\|_2^2)^{\frac{r-1}{2r-1}}\} \\
&\le s_0 D^* \sqrt{\mathcal{R}_{n,p}} C_r \big\{[{\delta^*}^{-1}(\ell(\boldsymbol{\beta}_{\hat{J}_k}) - \ell(\boldsymbol{\beta}^*))]^{\frac{r-1}{2r-1}} + [(2/\delta)(\ell(\boldsymbol{\beta}_{\hat{J}_{m_n^*}}) - \ell(\boldsymbol{\beta}^*))]^{\frac{r-1}{2r-1}}\big\} \\
&\le s_0 D^* C_r ({\delta^*}^{-\frac{r-1}{2r-1}} G^{-\frac{r}{2r-1}} + (4V_r^*/\delta)^{\frac{r-1}{2r-1}} G^{-1})(\ell(\boldsymbol{\beta}_{\hat{J}_k}) - \ell(\boldsymbol{\beta}^*)),\ 1 \le k < \tilde{k}_n.
\end{aligned} \tag{47}$$

In view of (41) and (44)–(47), the desired result (43) follows by taking G in (40) large enough.

Step 2: Prove

$$\lim_{n\to\infty} P(\hat{k}_n > Vm_n^*, S_n^*) = 0, \text{ for a sufficiently large } V. \tag{48}$$

We only consider the case where $K_n > Vm_n^*$, otherwise (48) holds trivially. Note first that for $Vm_n^* < k \leq K_n$,

$$0 \leq \ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_{m_n^*}}) - \ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) \leq \mathrm{IV}_k + \mathrm{V}_k + \mathrm{VI}_k, \tag{49}$$

where $\mathrm{IV}_k = |\ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_{m_n^*}}) - \ell(\hat{\boldsymbol{\beta}}_{\hat{J}_{m_n^*}}) - (\ell_n(\boldsymbol{\beta}^*) - \ell(\boldsymbol{\beta}^*))|$, $\mathrm{V}_k = |\ell_n(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) - \ell(\hat{\boldsymbol{\beta}}_{\hat{J}_k}) - (\ell_n(\boldsymbol{\beta}^*) - \ell(\boldsymbol{\beta}^*))|$, and $\mathrm{VI}_k = \ell(\hat{\boldsymbol{\beta}}_{\hat{J}_{m_n^*}}) - \ell(\boldsymbol{\beta}^*)$. By an argument similar to that used in Step 1, one has on the set S_n^*,

$$\begin{aligned}
\mathrm{IV}_k &\leq \sup_{\boldsymbol{\beta}\in\mathbb{B}} \|\nabla\ell_n(\boldsymbol{\beta}) - \nabla\ell(\boldsymbol{\beta})\|_\infty (\|\hat{\boldsymbol{\beta}}_{\hat{J}_{m_n^*}} - \boldsymbol{\beta}_{\hat{J}_{m_n^*}}\|_1 + \|\boldsymbol{\beta}_{\hat{J}_{m_n^*}} - \boldsymbol{\beta}^*\|_1) \\
&\leq s_0\sqrt{\mathcal{R}_{n,p}}\big\{[(2m_n^*/\delta)(\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_{m_n^*}}) - \ell(\boldsymbol{\beta}_{\hat{J}_{m_n^*}}))]^{1/2} + D^*C_r[(2/\delta)(\ell(\boldsymbol{\beta}_{\hat{J}_{m_n^*}}) - \ell(\boldsymbol{\beta}^*))]^{\frac{r-1}{2r-1}}\big\} \\
&\leq W_r^* m_n^* \mathcal{R}_{n,p} \leq (W_r^*/V)k\mathcal{R}_{n,p},
\end{aligned} \tag{50}$$

where $Vm_n^* < k \leq K_n$ and $W_r^* = s_0[s_0/\delta + D^*C_r(4V_r^*/\delta)^{(r-1)/(2r-1)}]$. Similarly, it can be shown that on the set S_n^*,

$$\mathrm{V}_k \leq (s_0^2/\delta)k\mathcal{R}_{n,p} + s_0 D^* C_r (4V_r^*/\delta)^{\frac{r-1}{2r-1}} V^{-1} k\mathcal{R}_{n,p},\ Vm_n^* < k \leq K_n. \tag{51}$$

Moreover, it is easy to see that on the set S_n^*,

$$\mathrm{VI}_k \leq 2Dm_n^*\mathcal{R}_{n,p} \leq (2D/V)k\mathcal{R}_{n,p},\ Vm_n^* < k \leq K_n. \tag{52}$$

By (41), (49)–(52), and $\omega > s_0^2/\delta$, we obtain

$$\lim_{n\to\infty} P(\min_{Vm_n^* < k \leq K_n} \mathrm{HDAIC}(\hat{J}_k) > \mathrm{HDAIC}(\hat{J}_{m_n^*}), S_n^*) = 1,$$

for a sufficiently large V, leading immediately to the desired conclusion (48).

Step 3: By (41), (42), and (48),

$$\lim_{n\to\infty}(\tilde{k}_n \leq \hat{k}_n \leq Vm_n^*) = 1. \tag{53}$$

Moreover, on the set $\mathrm{VA}_n \bigcap \{\tilde{k}_n \leq \hat{k}_n \leq V m_n^*\}$,

$$\begin{aligned}
&\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_{\hat{k}_n}}) - \ell(\boldsymbol{\beta}^*) \leq \left(\max_{1 \leqslant m \leqslant K_n} \frac{\ell(\hat{\boldsymbol{\beta}}_{\hat{J}_m}) - \ell(\boldsymbol{\beta}_{\hat{J}_m})}{m \mathcal{R}_{n,p}} \right) \hat{k}_n \mathcal{R}_{n,p} + \ell(\boldsymbol{\beta}_{\hat{J}_{\tilde{k}_n}}) - \ell(\boldsymbol{\beta}^*) \\
&\leq \frac{s_0^2 V}{2\delta} \mathcal{R}_{n,p}^{1-1/2r} + G \mathcal{R}_{n,p}^{1-1/2r}.
\end{aligned} \tag{54}$$

Consequently, (35) follows from (53), (54), and $\lim_{n\to\infty} P(\mathrm{VA}_n) = 1$. ■

Remark 4.2 *Discuss Thm 4.1 here.*

5. Applications to High-dimensional generalized linear Exponetial Family

In this section, we provide an application of theorem 4.1 on generalized linear exponetial family models.

$$f(y|\theta(\boldsymbol{x})) = \exp\{\theta(\boldsymbol{x})y - b(\theta(\boldsymbol{x})) + c(y)\}.$$

We model $\theta(\boldsymbol{x})$ as $\boldsymbol{\beta}^T \boldsymbol{x}$. Let loss function ℓ be negative log likelihood functions.

$$\ell(\boldsymbol{\beta}) = E[-\log f(y|\boldsymbol{\beta})] = E[-\boldsymbol{\beta}^T \boldsymbol{x} y + b(\boldsymbol{\beta}^T \boldsymbol{x}) - \eta(y)].$$

(EG1) For $j = 1, \dots, p$, X_j have $\alpha-$ concentration rate ($\alpha > 0$), that means there is a positive constant ξ such that

$$P(|X_j| > t) \leq \exp\{-\xi t^\alpha\},$$

where ξ are independent of j.

(EG2) The first derivative $b'(\cdot)$, is Lipschitz functions with constant L.

(EG3)

$$\sup_{\boldsymbol{\beta} \in \mathbb{B}} \sup_{j=1,\dots,p} E[b''(X^T \boldsymbol{\beta}) X_j^2] \leq M_2$$

(EG4) There exist small positive numbers ϵ and δ and an integer K such that for any $J \subset \{1, \dots, p\}$ with $|J| \leq K$, $\boldsymbol{\beta}$ in $B_\epsilon(\boldsymbol{\beta}_J)$ or in $B_\epsilon(\boldsymbol{\beta}^*)$, the smallest eigenvalue

$$\inf_{\boldsymbol{\beta} \in \mathbb{B}} \lambda_{min} E[b''(X^T \boldsymbol{\beta}) X X^T] \geq \delta$$

(EGC*) There is an $M_4 > 0$ such that

$$\max_{|J|\leq K_n, i\notin J} \sup_{\boldsymbol{\beta}\in B_\epsilon(\boldsymbol{\beta}^*)} \left\| \left[E[b''(X^T\boldsymbol{\beta}_J)X_JX_J^T]\right]^{-1} \left[E[b''(X^T\boldsymbol{\beta}_J)X_JX_i]\right] \right\|_1 \leq M_4, \tag{55}$$

and

$$\max_{|J|\leq K_n, i\notin J} \sup_{\boldsymbol{\beta}\in B_\epsilon(\boldsymbol{\beta}^*)} \left\| \left[E[b''(X^T\boldsymbol{\beta}^*)X_JX_J^T]\right]^{-1} \left[E[b''(X^T\boldsymbol{\beta}^*)X_JX_i]\right] \right\|_1 \leq M_4. \tag{56}$$

Theorem 5.1 *If the exponetial generalized linear model satisfy assumptions (A1)(A2)(EG1)-(EG4)(EGC*), then the Theorem 4.1 hold with rate $\mathcal{R}_{n,p} = \frac{\log p}{n}(\log p)^{2/\alpha}$.*

Proof Check the conditions in Theorem 4.1. Note that $\nabla^2\ell(\boldsymbol{\beta}) = E[b''(X^T\boldsymbol{\beta})XX^T]$; hence, (EG3) (EG4) (EGC*) imply (A3)(A4)(C*) respectively. And Lemma **??** shows that (EG1) (EG2) imply (U) with the rate $\mathcal{R}_{n,p} = \frac{\log p}{n}(\log p)^{1/\alpha}$. ■

Remark 5.1 *Wang et al. (2014) claim $\mathcal{R}_{n,p} = \frac{\log p}{n}$ is the minimax rate for linear model under bounded design($\alpha = \infty$). Theorem 5.1 claim that even in generalized linear exponential famaliy, CGA+HDAIC also can reach minimax rate.*

Remark 5.2 *In linear models,Ing (2019) claim $\mathcal{R}_{n,p} = \frac{\log p}{n}$ is also the minimax rate under sub-Gaussian design($\alpha = 2$). In generalized linear sub-Gaussian design models we cannot reach the same minimax rate theoretically; however, our experimental result shows that it also has decent outcome when we use $\mathcal{R}_{n,p} = \frac{\log p}{n}$ rather than $\mathcal{R}_{n,p} = \frac{(\log p)^2}{n}$.*

Remark 5.3 *Assuming X is bounded, that is $\alpha = \infty$, $b''(\cdot)$ is bound above and away from zero since $\sup_{\boldsymbol{\beta}\in\mathbb{B}} \sup_X |X^T\boldsymbol{\beta}|$ is bounded. In this case, traditional assumptions likes finite second moments and positivity of minimal eigenvalue of design X would be sufficient for (EG3) and (EG4) since*

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}} \sup_{j=1,\ldots,p} E[b''(X^T\boldsymbol{\beta})X_j^2] \leq \sup b''(\cdot) \sup_{j=1,\ldots,p} E[X_j^2] := M_2$$

$$\lambda_{min}E[b''(X^T\boldsymbol{\nu})XX^T] \geq \inf b''(\cdot)\lambda_{min}E[XX^T] := \delta$$

Particualrly, in the logistic and poison model $b''(u)$ *are* $\frac{e^u}{(1+e^u)^2}$ *and* e^u *respectively. Note that in both case* b"(u) *are strictly positive continuous function. Hence, if* X *is bounded,* $\sup\limits_{X}\sup\limits_{\boldsymbol{\beta}\in\mathbb{B}} b''(X^T\boldsymbol{\beta})$ *is bounded above. Also,* $\inf\limits_{X}\inf\limits_{\boldsymbol{\beta}\in\mathbb{B}} b''(X^T\boldsymbol{\beta})$ *is bounded away from zero.*

5.1 Some comparisons with existing results.

Elenberg, E. R., et al. (2018) discuss several types of Greedy algorithm, including our method CGA. Their result claim $k-$steps CGA should't be too larger than $\arg\min\limits_{\|\boldsymbol{\beta}\|_0\leq k}\ell(\boldsymbol{\beta})$ in the population sense. Not only talking about fixed k, our result Theorem 3.1 show that CGA convergence in decent rate when the iterate time k increase. Furthermore, we use information criteria (HDIC) to decide the iterate time which is overlooked in greedy types algorithm. Elenberg, E. R., et al. (2018), Bahmani, S. et al. (2013), Tewari, A. et al (2011).

The maximal eigenvalue of the covariance matrix of $\boldsymbol{X}$, or maximal eigenvalue of Hassian matrix of $\ell(\cdot)$ can been seen as the bound of maximal gradient. Intuitively, larger gradient can improve the converge rate when using the gradient decent method for the optimal. However, in several literature Elenberg, E. R., et al. (2018), Fan and Song (2010), Barut,E. et al. (2016), let the bound of maximal eigenvalue being the necessary conditions.

6. Numerical Studies and Real Data Illustration

Consider i.i.d samples generated from the logistic regression model:

$$\Pr(Y_t = y_t|\boldsymbol{\beta}^*, \boldsymbol{X}_t) = \left(\frac{e^{\boldsymbol{X}_t'\boldsymbol{\beta}^*}}{1+e^{\boldsymbol{X}_t'\boldsymbol{\beta}^*}}\right)^{y_t}\left(\frac{1}{1+e^{\boldsymbol{X}_t'\boldsymbol{\beta}^*}}\right)^{1-y_t}, \quad t = 1,\ldots,n, \tag{57}$$

with some true coefficient $\boldsymbol{\beta}^*$ and p predictor variables $\boldsymbol{X}_t = (x_{t1},\ldots,x_{tp})'$. In this section, we report the simulations studies of the performance. In particular, we present results for negative log-likelihood and the number of selected variables, i.e. $\hat{\beta} \neq 0$, corresponding to CGA+HDAIC and Lasso (L-1 regularized) and Iterative Hard Thresholding (IHT) also known as projected gradient descent with projection to an s-sparse set) with 5-fold cross-validation. CGA and IHT aim to maximize the log-likelihood function:

$$\max_{\boldsymbol{\beta}}\frac{1}{n}\sum_{t=1}^{n} y_t\boldsymbol{X}_t'\boldsymbol{\beta} - \log(1+\exp(\boldsymbol{X}_t'\boldsymbol{\beta}))$$

while using the scikit-learn, L-1 regularized logistic regression solves the following optimization problem:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 + C\sum_{t}^{n} \log(\exp(-y_t\boldsymbol{X}_t'\boldsymbol{\beta}) + 1)$$

where C is the inverse of regularization strength and y_t takes the values in the set $\{-1, 1\}$.

To examine their performance, we generate training and testing data which are same size and independent. Training data are used to train models while testing data are used to compute the testing negative log-likelihood (NLL). We also report the number of non-zeros parameters. Three cases of sparsity are considered as following:

1. Polynomial Sparsity: $\beta_j^* = bj^{-r}$
2. Exponential Sparsity: $\beta_j^* = bc^{-j}$
3. Strong Sparsity: $\beta_1^* = 3, \beta_2^* = 4, \beta_3^* = 3, \beta_4^* = 5$ and $\beta_j^* = 0$ for $j > 4$

In all cases, we chose $p = 5n, b = 5, r = 1.5, c = 2, K_n = 3\sqrt{\log p/n}$, $\omega = 2$ for HDAIC and regularization hyper-parameter $C = c\sqrt{\log p/n}$ with 10 points c uniformly in $[0.1, 10]$ for Lasso and $s = [3, 4, 5, 6, 7]$ for IHT. Cross-validation is applied to Lasso and IHT for adjusting best hyper-parameter c, s.

Example 1 *Consider (57) and assume*

$$x_{tj} = \sqrt{1-\eta^2}d_{tj} + \eta w_t, \tag{58}$$

where $\eta \geqslant 0$ *and* $(d_{t1}, \ldots, d_{tp}, w_t)'$, $1 \leqslant t \leqslant n$, *are i.i.d. multivariate normal with zero mean and identity covariance matrix. The case* $\eta = 0, 0.3$, *which is in table 1 and table 2, respectively. We can see in table 1 CGA+HDAIC achieve the best NLL in all cases using smallest amount of variables. Moreover, in strong sparsity, CGA+HDAIC almost select the correct number of variables. Note that we do not tune the hyper-parameter of HDAIC carefully and computation cost of CV increase dramatically as the number of hyper-parameters increasing. In table 2, the equal correlation improve the NLL. Lasso+CV have minimum NLL under the polynomial sparsity when* $n = 400, 800$, *but CGA+HDAIC have comparable performance. Apart from this, CGA+HDAIC have best results in term of both NLL and the size of nonzero variables. IHT+CV have similar performance with CGA+HDAIC under the* *polynomial and exponential sparsity. Indeed, IHT enjoy the fastest computation due to its simpleness. However, IHT requires more careful selection for the sparsity level without any prior information and including more possible hyper-parameters result in long computation time when CV is applied, which implies losing the speed advantage. Furthermore, under the strong sparsity, there is a significant gap between CGA and other methods.*

Table 1: The mean and standard deviation of negative log-likelihood (NLL) and the number of variables of three methods in 100 simulations under (57) and (58) with $\eta = 0$

	Sample Size	200		400		800	
Cases	Method	NLL	# of Variables	NLL	# of Variables	NLL	# of Variables
Poly.	GCA+HDAIC	0.3220 (0.0532)	1.6 (0.5099)	0.2826 (0.0298)	2.15 (0.4092)	0.2621 (0.0214)	2.87 (0.4394)
	lasso+CV	0.4404 (0.0824)	42.69 (56.109)	0.3092 (0.0245)	28.0 (19.740)	0.2769 (0.0185)	159.85 (13.007)
	IHT+CV	0.3417 (0.0658)	4.42 (1.4365)	0.2902 (0.0357)	4.09 (1.2735)	0.2689 (0.0255)	4.68 (1.4274)
Exp.	GCA+HDAIC	0.2763 (0.0411)	2.13 (0.3648)	0.2470 (0.0337)	2.69 (0.4624)	0.2308 (0.0209)	3.1 (0.3605)
	lasso+CV	0.4199 (0.0627)	37.83 (52.671)	0.2897 (0.0194)	23.63 (5.3228)	0.2620 (0.0346)	149.66 (28.577)
	IHT+CV	0.2947 (0.0620)	4.28 (1.3422)	0.2625 (0.0397)	4.23 (1.3773)	0.2382 (0.0193)	4.48 (1.4455)
Str.	GCA+HDAIC	0.1672 (0.0339)	3.99 (0.0994)	0.1511 (0.0237)	4.0 (0.0)	0.1489 (0.0160)	4.0 (0.0)
	lasso+CV	0.3994 (0.0583)	38.48 (49.818)	0.2659 (0.0154)	18.02 (4.7180)	0.2057 (0.0116)	115.19 (13.519)
	IHT+CV	0.2347 (0.1642)	5.3 (1.2529)	0.1571 (0.0290)	4.76 (1.0594)	0.1509 (0.0161)	4.8 (0.9486)

Example 2 *Consider (57) and assume $x_{t1}, \ldots, x_{t4}$ are i.i.d. standard normal, and*

$$x_{tj} = \sqrt{1-\eta^2} d_{tj} + \frac{\eta}{2} \sum_{k=1}^{4} x_{tk}, \tag{59}$$

where $\eta \geqslant 0$ and $(d_{t1}, \ldots, d_{tp})'$, $1 \leqslant t \leqslant n$, are i.i.d. multivariate normal with zero mean and identity covariance matrix. Table 3 summarize the results. We can see the performance of IHT get worse significantly than example 1 while CGA maintain the similar performance. This example illustrate an inherent difficulty when irrelevant varaibles have substantial correlations with relevant ones. CGA overcome this difficulty and get the best result in both NLL and the size of nonzero coefficients under all sparsity conditions.

Table 2: The mean and standard deviation of negative log-likelihood (NLL) and the number of variables of three methods in 100 simulations under (57) and (58) with $\eta = 0.3$

	Sample Size	200		400		800	
Case	Method	NLL	# of Variables	NLL	# of Variables	NLL	# of Variables
Poly.	GCA+HDAIC	0.3272 (0.0547)	1.81 (0.5778)	0.2875 (0.0381)	2.61 (0.6464)	0.2581 (0.0230)	3.66 (0.6514)
	lasso+CV	0.4221 (0.0841)	54.87 (56.275)	0.2872 (0.0201)	47.87 (19.813)	0.2424 (0.0161)	138.51 (11.942)
	IHT+CV	0.3523 (0.0697)	5.22 (1.5529)	0.3124 (0.0443)	5.21 (1.5446)	0.2800 (0.0253)	5.36 (1.4732)
Exp.	GCA+HDAIC	0.2577 (0.0436)	2.1 (0.3316)	0.2275 (0.0318)	2.72 (0.4707)	0.2200 (0.0174)	3.11 (0.3128)
	lasso+CV	0.4086 (0.0706)	46.45 (56.025)	0.2735 (0.0175)	24.87 (5.0786)	0.2433 (0.0151)	125.53 (13.589)
	IHT+CV	0.2854 (0.0762)	4.81 (1.5536)	0.2514 (0.0350)	4.8 (1.5748)	0.2274 (0.0264)	4.99 (1.5588)
Str.	GCA+HDAIC	0.1534 (0.0505)	3.96 (0.1959)	0.1355 (0.0237)	4.0 (0.0)	0.1359 (0.0142)	4.0 (0.0)
	lasso+CV	0.3876 (0.0391)	13.06 (25.823)	0.2463 (0.0156)	18.26 (4.9348)	0.1881 (0.0107)	94.36 (11.923)
	IHT+CV	0.2350 (0.1045)	5.51 (1.2688)	0.1530 (0.0440)	5.24 (1.0688)	0.1384 (0.0146)	4.66 (0.9404)

Table 3: The mean and standard deviation of negative log-likelihood (NLL) and the number of variables of three methods in 100 simulations under (57) and (59) with $\eta = 0.3$

	Sample Size	200		400		800	
Case	Method	NLL	# of Variables	NLL	# of Variables	NLL	# of Variables
Poly.	GCA+HDAIC	0.3239 (0.0727)	2.07 (0.7107)	0.2360 (0.0385)	3.48 (0.7138)	0.2089 (0.0192)	3.99 (0.0994)
	lasso+CV	0.4348 (0.1018)	95.47 (58.303)	0.3140 (0.0527)	103.5 (82.641)	0.2464 (0.0179)	154.89 (11.183)
	IHT+CV	0.4040 (0.0597)	5.3 (1.5132)	0.3772 (0.0516)	5.28 (1.5171)	0.3478 (0.0411)	5.8 (1.2806)
Exp.	GCA+HDAIC	0.2718 (0.0524)	2.25 (0.4974)	0.2507 (0.0342)	2.7 (0.4582)	0.2270 (0.0188)	3.26 (0.4386)
	lasso+CV	0.4321 (0.0859)	64.62 (63.559)	0.2963 (0.0365)	64.78 (37.276)	0.2522 (0.0155)	149.56 (11.336)
	IHT+CV	0.4118 (0.0731)	5.19 (1.5918)	0.3364 (0.0636)	5.37 (1.4117)	0.2910 (0.0411)	5.24 (1.5041)
Str.	GCA+HDAIC	0.2290 (0.2190)	3.99 (0.8887)	0.1527 (0.0230)	4.0 (0.0)	0.1509 (0.0147)	4.0 (0.0)
	lasso+CV	0.3741 (0.0820)	168.52 (33.550)	0.3324 (0.0499)	308.93 (126.39)	0.2879 (0.0544)	515.57 (241.78)
	IHT+CV	0.4938 (0.0828)	5.91 (1.2656)	0.3831 (0.0669)	5.89 (1.2953)	0.3268 (0.0606)	5.74 (1.3388)

References

Abramovich, F. and Grinshtein, V. (2014). Model selection and minimax estimation in generalized linear models. *IEEE Transactions on Information Theory*, **62**, 3721–3730.

Bahmani, S., Raj, B., and Boufounos, P. T. (2013). Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, **14(Mar)**, 807-841.

Barut, E., Fan, J., and Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association*, **111**, 1266-1277.

Belloni, A. and Chernozhukov, V. (2009). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*, **39**, 82–130.

Bartlett, Peter L and Mendelson, Shahar. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, **3**, 463–482.

Bousquet, O (2002). A Bennet concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, **334**, 495-550.

Elenberg, E. R., Khanna, R., Dimakis, A. G., amd Negahban, S. (2018). Restricted strong convexity implies weak submodularity. *The Annals of Statistics*,**46**, 3539-3568.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **70**, 849–911.

Fan, J. and Song,R.(2010) Sure independence screening in generalized linear models with NP-dimensionality.*Annals of Statistics*,**38**,3567–3604.

Gao, F., Ing, C.-K., and Yang, Y. (2013). Metric entropy and sparse linear approximation of l_q-hulls for $0 < q \leq 1$. *Journal of Approximation Theory*, **166**, 42–55.

Ing, C.-K. (2019). MModel selection for high-dimensional linear regression with dependent observations, *to appear in Annals of Statistics.*

Ing, C.-K. and Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica*, **21**, 1473–1513.

Prateek Jain and Purushottam Kar. (2017). Non-convex optimization for machine learning. Foun-dations and Trends® in Machine Learning, 10(3-4):142–336.

Ledoux, M., Talagrand, M.(1991). Probability in Banach Spaces-Isoperimetry and Processes. *Springer*

Luo, Z.-Q. and Tseng, P. (1992). On the convergence of the coordinate descent method forconvex differentiable minimization. *Journal of Optimization Theory and Applications*, **72, 7–35.**

Massart, P. (2000). About the constants in talagrand's concentration inequalities for empirical processes. *Annals of Probability*, **28, 863–884.**

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of m**-estimators with decomposable regularizers.** *Statistical Science*, **27, 538–557.**

Peng, B., Wang, L., and Wu, Y. (2016). An error bound for l_1**-norm support vector machine coefficients in ultra-high dimension.** *Journal of Machine Learning Research*, **17, 1-26.**

Peskir, G. (2000). From uniform laws of large numbers to uniform ergodic theorems. Lecture Note.

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

Raskutti, G., Wainwright, M. J., and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, **13, 389–427.**

Temlyakov, V. N. (2015). Greedy approximation in convex optimization. *Constructive Approximation*, **41, 269–296.**

Tewari, A., Ravikumar, P. K., and Dhillon, I. S. (2011). Greedy algorithms for structurally constrained high dimensional problems. *In Advances in Neural Information Processing Systems*, **(pp. 882-890).**

Wang, Z.,Paterlini, S.,Gao,F. and Yang, Y.(2014) Adaptive Minimax Regressioin Estimation over Saprse ℓ_q-Halls. *Journal of Machine Learning Research*, **15, 1675–1711.**

Zhao, P.,Yu, B.(2006) On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, **7, 2541–2563.**