JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2021.1990769

Statistical Inference for High-Dimensional Generalized Linear Models With Binary Outcomes

T. Tony Cai[a], Zijian Guo[b], and Rong Ma[c]

[a]Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA; [b]Department of Statistics, Rutgers University, Piscataway, NJ; [c]Department of Statistics, Stanford University, Stanford, CA

ABSTRACT

This article develops a unified statistical inference framework for high-dimensional binary generalized linear models (GLMs) with general link functions. Both unknown and known design distribution settings are considered. A two-step weighted bias-correction method is proposed for constructing confidence intervals (CIs) and simultaneous hypothesis tests for individual components of the regression vector. Minimax lower bound for the expected length is established and the proposed CIs are shown to be rate-optimal up to a logarithmic factor. The numerical performance of the proposed procedure is demonstrated through simulation studies and an analysis of a single cell RNA-seq dataset, which yields interesting biological insights that integrate well into the current literature on the cellular immune response mechanisms as characterized by single-cell transcriptomics. The theoretical analysis provides important insights on the adaptivity of optimal CIs with respect to the sparsity of the regression vector. New lower bound techniques are introduced and they can be of independent interest to solve other inference problems in high-dimensional binary GLMs.

ARTICLE HISTORY
Received May 2020
Accepted September 2021

KEYWORDS
Adaptivity; Confidence interval; Hypothesis testing; Link functions; Optimality; Weighting

1. Introduction

Generalized linear models (GLMs) with binary outcomes are ubiquitous in modern data-driven scientific research, as binary outcome variables arise frequently in many applications such as genetics, metabolomics, finance, and econometrics, and play important roles in many observational studies. With rapid technological advancements in data collection and processing, it is often needed to analyze massive and high-dimensional data where the number of variables is much larger than the sample size. In such high-dimensional settings, most of the classical inferential procedures such as the maximum likelihood are no longer valid, and there is a pressing need to develop new principles, theories and methods for parameter estimation, hypothesis testing, and confidence intervals (CIs).

1.1. Problem Formulation

This article aims to develop a unified statistical inference framework for high-dimensional GLMs with binary outcomes. We assume the observations $(X_i, y_i) \in \mathbb{R}^p \times \{0, 1\}, i = 1, ..., n$, are independently generated from

$$y_i|X_i \sim \text{Bernoulli}(f(X_i^\top \beta)), \quad X_i \sim P_X, \tag{1}$$

where $f : \mathbb{R} \to (0, 1)$ is a known link function, $\beta \in \mathbb{R}^p$ is a high-dimensional sparse regression vector with sparsity k and P_X is some probability distribution on $\mathbb{R}^p$. The goal of the present paper is threefold:

1. Construct optimal CIs for the individual components of β;
2. Conduct simultaneous hypothesis testing for the individual components of β;
3. Establish the minimum sample size requirement for constructing CIs which are adaptive to the sparsity level k of β.

Throughout, we consider a general class of link functions f, which can be characterized by a set of mild regularity conditions specified later in Section 3.1. The following are a few important examples of this general class. Among them, logistic regression is perhaps the most commonly used methods for analyzing datasets with binary-outcomes. However, in many applications, alternative link functions have been adopted due to their specific interpretations with respect to the applications (Razzaghi 2013).

Example 1. Logistic link function. Problems related to high-dimensional logistic regression with the link function $f(x) = \exp(x)/(1 + \exp(x))$ have been extensively studied in literature. See Section 1.3 for the existing works.

Example 2. Probit link function. In probit regression model, the link function is the standard Gaussian cumulative distribution function (cdf). This model is also widely used in practice and well-understood in the classical low-dimensional settings. However, comparing to the logistic regression, much less is known for the high-dimensional probit regression model.

Example 3. Latent variable model. Generalizing the above link functions, one may consider the class of link functions induced by a latent variable model. Consider an auxiliary random variable $y_i^* = X_i^\top \beta + \epsilon_i$ with $\epsilon_i \sim P_\epsilon$ for $1 \le i \le n$. Then

CONTACT Zijian Guo zijguo@stat.rutgers.edu Department of Statistics, Rutgers University, Piscataway, NJ 08854.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

the observed binary outcome variable $y_i = \mathbf{1}(y_i^* \geq 0)$ can be reformulated as a binary GLM with $y_i|X_i \sim \text{Bernoulli}(f(X_i^\top \beta))$, where $f(\cdot)$ is the cdf of $-\epsilon_i$. Besides the logistic and the probit link functions, examples include the cdfs of the generalized logistic distribution, where $f(x) = \frac{1}{2}\tanh^\gamma(\varphi x) + \frac{1}{2}$ for any $\varphi > 0$ and $\gamma \geq 1$, and the Student's t_ν-distributions with any degrees of freedom $\nu \in \mathbb{N}$.

1.2. Main Results and Contributions

We propose a unified two-step procedure for constructing CIs and performing statistical tests for the regression coefficients in the high-dimensional binary GLM (1). A penalized maximum-likelihood estimator (MLE) is implemented to estimate the high-dimensional regression vector and then a link-specific weighting (LSW) method is proposed to correct the bias of the penalized estimator. CIs and statistical tests are constructed by quantifying the uncertainty of the proposed LSW estimator. The asymptotic normality of the proposed LSW estimator is established and the validity of the constructed CIs and statistical tests are justified.

Comparing to the existing methods for logistic models (van de Geer et al. 2014; Belloni, Chernozhukov, and Wei 2016; Ning and Liu 2017; Ma, Cai, and Li 2020; Guo et al. 2020; Shi et al. 2020), a key methodological advancement is the construction of the link-specific weights. With this novel weight construction, the proposed LSW method is shown to be effective for a general class of link functions, including both the canonical and noncanonical binary GLMs. Furthermore, the proposed LSW method is effective for the general unknown sub-Gaussian design with a regular population design covariance matrix. To the best of our knowledge, the proposed method is the first inference procedure that works for such a general class of link functions and designs; see the discussion after Theorem 1 for a detailed comparison. In contrast to the equal-weight methods for bias-correction in the linear models (Zhang and Zhang 2014; van de Geer et al. 2014; Javanmard and Montanari 2014a), our results show that a careful weight construction is essential to debiasing for the binary GLMs. This idea can be of independent interest to study other inference problems in high-dimensional GLMs.

The minimax optimality of CIs for a single regression coefficient of the binary GLMs with general link functions is established, and our proposed CIs are shown to achieve the optimal expected length up to a logarithmic factor over the sparse regime with $k = ||\beta||_0 = o(\frac{n}{\log n \log p})$. The analysis provides important insights on the adaptivity of the optimal CIs with respect to a collection of nested parameter spaces indexed by the sparsity k of β. It is shown that the possible region of constructing adaptive CIs for the individual components of β is the ultra-sparse regime with $k = o(\frac{\sqrt{n}}{\sqrt{\log n}\log p})$. The minimaxity and adaptivity results are illustrated in Figure 1. New lower bound techniques are developed, which can be of independent interest for other high-dimensional binary GLM inference problems. Moreover, for both theoretical and practical interests, we study the optimal CIs and statistical tests in the case of known design distributions.

Simulation studies indicate several practical advantages of the LSW method over the existing ones. Specifically, our proposed method is flexible with respect to the underlying link function and efficient in terms of computational costs. The proposed CIs have more precise empirical coverage probabilities and shorter lengths. As for hypothesis testing, under the sparse setting, the proposed test is more powerful than the likelihood ratio test of Sur, Chen, and Candès (2019), which is well defined only for the moderate-dimensional settings with $p < n/2$. In addition, an analysis of a real single-cell RNA-seq dataset yields interesting biological insights that integrate well into the current literature on the cellular immune response mechanisms as characterized by single-cell transcriptomics.

Our proposed method has been included in the R package `SIHR`, which is available from CRAN. More details about using the R package `SIHR` can be found in Rakshit, Cai, and Guo (2021).

1.3. Related Work

The estimation problem in the high-dimensional GLMs has been extensively studied in the literature (van de Geer 2008; Meier, van de Geer, and Bühlmann 2008; Negahban et al. 2010; Bach 2010; Huang and Zhang 2012; Plan and Vershynin 2013). However, for the high-dimensional binary GLMs, most of aforementioned papers focus on the logistic link function. In the present paper, we establish precise estimation bounds for the high-dimensional binary GLMs with general link functions, including both canonical and noncanonical links.

There is a paucity of methods and fundamental theoretical results on statistical inference including hypotheses testing and CIs in the high-dimensional GLMs. van de Geer et al. (2014) constructed CIs and statistical tests for β_j with $1 \leq j \leq p$

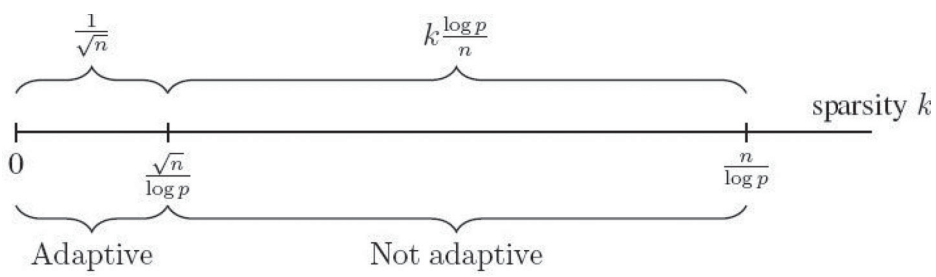

Figure 1. An illustration of the optimality and adaptivity of the CIs with respect to the sparsity k of β for the unknown design setting. On the top of the figure, we report the minimax expected lengths of the CIs, which can be attained by our proposed LSW method (up to a $\log n$ factor for the rate $k\frac{\log p}{n}$). On the bottom of the figure, the possibility of being adaptive to the sparsity k is presented.

in high-dimensional GLMs. Belloni, Chernozhukov, and Wei (2016) constructed confidence regions for β_j with $1 \le j \le p$ in the GLMs based on the construction of an instrument that immunizes against model selection mistake. Ning and Liu (2017) proposed a general framework for hypothesis testing and confidence regions for low-dimensional components in high-dimensional models based on the generic penalized M-estimators. More recently, Zhu, Shen, and Pan (2020) proposed a constrained MLE method for hypothesis testing involving unspecific nuisance parameters. Focusing on the high-dimensional binary regression with sparse design matrix, Mukherjee, Pillai, and Lin (2015) studied detection boundary for the minimax hypothesis testing. For the high-dimensional logistic regression model, Sur, Chen, and Candès (2019) and Sur and Candès (2019) studied the likelihood ratio test under the setting where $p/n \to \kappa$ for some $\kappa < 1/2$; Ma, Cai, and Li (2020) proposed testing procedures for both the global null hypothesis and the large-scale simultaneous hypotheses for the regression coefficients under the $p \gg n$ setting; Guo et al. (2020) studied inference for the case probability, which is a transformation of a linear combination of the regression coefficients. Shi et al. (2020) proposed an inference procedure based on a recursive online-score estimation approach. The articles van de Geer et al. (2014), Ma, Cai, and Li (2020), Guo et al. (2020), and Shi et al. (2020) focused on the logistic link and impose certain stringent assumptions, such as the bounded individual probability condition, or sparse inverse population Hessian/precision matrix. In contrast, we propose a novel weighting method for general link functions that produces optimal CIs without requiring these stringent assumptions. The numerical advantages of our proposed method over the existing methods are demonstrated in Section 5.

Statistical inference for high-dimensional linear regression has been well studied in the literature. Specifically, Zhang and Zhang (2014), van de Geer et al. (2014), and Javanmard and Montanari (2014a, 2014b) considered CIs and testing for individual regression coefficients of the high-dimensional linear model, and the minimaxity and adaptivity of the confidence set construction has been studied in Nickl and van de Geer (2013), Cai and Guo (2017), and Cai and Guo (2018).

1.4. Organization and Notation

The rest of the article is organized as follows. We finish this section with notation. In Section 2, we construct CIs and statistical tests for single regression coefficients in high-dimensional binary GLMs with unknown design distribution. We then study in Section 3 the theoretical properties of the proposed CIs and statistical tests, and establish their minimax optimality and adaptivity. Optimal CIs and statistical tests in the setting of known design distributions are considered in Section 4. The numerical performance of the proposed methods is evaluated in Section 5. In Section 6, the methods are illustrated through an analysis of a real single cell RNA-seq dataset. Further discussions are presented in Section 7. The proofs of the theorems together with some additional discussions are collected in the supplementary material.

Throughout, for a vector $a = (a_1, ..., a_n)^\top \in \mathbb{R}^n$, we define the ℓ_p norm $||a||_p = \left(\sum_{i=1}^n |a_i|^p\right)^{1/p}$, the ℓ_0 norm $||a||_0 = \sum_{i=1}^n 1\{a_i \neq 0\}$, and the ℓ_∞ norm $||a||_\infty = \max_{1\le j\le n} |a_i|$, and let $a_{-j} \in \mathbb{R}^{n-1}$ stand for the subvector of a without the j-th component. For a matrix $A \in \mathbb{R}^{p\times q}$, $\lambda_i(A)$ stands for the i-th largest singular value of A and $\lambda_{\max}(A) = \lambda_1(A)$, $\lambda_{\min}(A) = \lambda_{\min\{p,q\}}(A)$. For a smooth function $f(x)$ defined on $\mathbb{R}$, we denote $f'(x) = df(x)/dx$ and $f''(x) = d^2f(x)/dx^2$. For any positive integer n, we denote the set $\{1, 2, ..., n\}$ as $[1 : n]$. For any $a, b \in \mathbb{R}$, we denote $I_x(a, b) = B(x; a, b)/B(a, b)$ as the regularized incomplete beta function, where $B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$ is the incomplete beta function. We define $\phi(x)$ and $\Phi(x)$ as the density function and cdf of the standard Gaussian random variable, respectively. We denote $\to_d$ as convergence in distribution. For positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = o(b_n)$, $a_n \ll b_n$ or $b_n \gg a_n$ if $\lim_n a_n/b_n = 0$, and write $a_n = O(b_n)$, $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists a constant C such that $a_n \le Cb_n$ for all n. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

2. Statistical Inference and the Weighting Method

We use $\ell_f(\beta)$ to denote the negative log-likelihood function associated to the GLM in Equation (1),

$$\ell_f(\beta) = -\frac{1}{n}\sum_{i=1}^n y_i \log\left[\frac{f(X_i^\top\beta)}{1-f(X_i^\top\beta)}\right] - \frac{1}{n}\sum_{i=1}^n \log(1-f(X_i^\top\beta)). \tag{2}$$

We define the penalized negative log-likelihood estimator for GLM,

$$\hat\beta = \arg\min_\beta\{\ell_f(\beta) + \lambda||\beta||_1\}, \tag{3}$$

with $\lambda \asymp \sqrt{\log p/n}$. Although $\hat\beta$ achieves the optimal rate of convergence (Negahban et al. 2010; Huang and Zhang 2012), it cannot be directly used for CI construction due to its bias. As in the high-dimensional linear regression (Javanmard and Montanari 2014a, 2014b; van de Geer et al. 2014; Zhang and Zhang 2014), bias correction is needed to make statistical inference for β_j with $1 \le j \le p$. An important extra step for high-dimensional GLM is to introduce link-specific weights to carry out the bias correction.

For technical reasons, we split the samples such that the initial estimation step and the bias-correction step are conducted on independent datasets. Without loss of generality, we assume there are $2n$ samples $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{2n}$, divided into two disjoint subsets $\mathcal{D}_1 = \{(X_i, y_i)\}_{i=1}^n$ and $\mathcal{D}_2 = \{(X_i, y_i)\}_{i=n+1}^{2n}$. The initial estimator $\hat\beta$ is obtained by applying (3) to $\mathcal{D}_2$ while the bias-correction step (detailed in Section 2.1) is based on $\hat\beta$ and the samples in $\mathcal{D}_1$. Importantly, the sample splitting procedure is used only to facilitate the theoretical analysis, which does not make it a restriction for practical applications. Numerically, we show in Section 5 that, without sample splitting the proposed methods could also perform well, and is statistically more efficient than the alternative methods. See also Section 4.2 in the supplementary material for additional numerical comparisons, and Section 7 for more discussions.

2.1. A Weighting Method for Bias Correction

For a given $j \in [1:p]$, we consider the following generic form of the bias-correction estimator:

$$\tilde{\beta}_j = \hat{\beta}_j + u^\top \frac{1}{n} \sum_{i=1}^n W_i \cdot X_i(y_i - f(X_i^\top \hat{\beta})), \tag{4}$$

where $\hat{\beta}$ is defined in (3), $W_i \in \mathbb{R}$ for $1 \le i \le n$ and $u \in \mathbb{R}^p$ denote respectively the data-dependent weights and projection direction to be constructed. For the link function $f : \mathbb{R} \to (0,1)$ in (1), we will construct the link-specific weights $\{W_i\}_{i=1}^n$ and a projection vector $u \in \mathbb{R}^p$ such that $u^\top \frac{1}{n} \sum_{i=1}^n W_i \cdot X_i(y_i - f(X_i^\top \hat{\beta}))$ is an accurate estimator of the bias $\hat{\beta}_j - \beta_j$.

We now derive the error decomposition of the generic estimator in (4), which provides intuitions on the construction of $\{W_i\}_{i=1}^n$ and $u \in \mathbb{R}^p$. Rewrite the model (1) as $y_i = f(X_i^\top \beta) + \epsilon_i$ with ϵ_i satisfying $\mathbb{E}[\epsilon_i | X_i] = 0$ and $\mathrm{var}(\epsilon_i | X_i) = f(X_i^\top \beta)(1 - f(X_i^\top \beta))$. We apply Taylor expansion of f near $X_i^\top \hat{\beta}$ and obtain

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n W_i X_i (y_i - f(X_i^\top \hat{\beta})) \\
&\quad = \frac{1}{n} \sum_{i=1}^n W_i X_i \epsilon_i + \frac{1}{n} \sum_{i=1}^n W_i f'(X_i^\top \hat{\beta}) X_i X_i^\top (\beta - \hat{\beta}) \\
&\qquad + \frac{1}{n} \sum_{i=1}^n W_i X_i \Delta_i,
\end{aligned} \tag{5}$$

with $\Delta_i = f''(X_i^\top \hat{\beta} + t X_i^\top (\beta - \hat{\beta})) \cdot [X_i^\top (\hat{\beta} - \beta)]^2$ for some $t \in (0,1)$. Combining Equations (4) and (5), the estimation error $\tilde{\beta}_j - \beta_j$ is expressed as follows:

$$\underbrace{u^\top \frac{1}{n} \sum_{i=1}^n W_i X_i \epsilon_i}_{\text{Stochastic error}} + \underbrace{\left(u^\top \frac{1}{n} \sum_{i=1}^n W_i f'(X_i^\top \hat{\beta}) X_i X_{i\cdot}^\top - e_j^\top \right)(\beta - \hat{\beta})}_{\text{Remaining bias}} + \underbrace{u^\top \frac{1}{n} \sum_{i=1}^n W_i X_i \Delta_i}_{\text{Approximation error}} \tag{6}$$

where $\{e_j\}_{j=1}^p$ is the canonical basis of the Euclidean space $\mathbb{R}^p$. In the expression (6), the first term is the stochastic error due to the model error ϵ_i, the second term is the remaining bias due to the penalized estimator $\hat{\beta}$, and the last term is the approximation error due to the nonlinearity of f.

Our goal is to construct $\{W_i\}_{i=1}^n$ and $u \in \mathbb{R}^p$ such that a) the stochastic error in Equation (6) is asymptotically normal and its standard error is minimized; b) the remaining bias and approximation errors in Equation (6) are negligible in comparison to the stochastic error. If these two properties hold, then we can establish the asymptotic normality of the bias-corrected estimator $\tilde{\beta}_j$ in Equation (4) and justify certain efficiency properties of our proposed estimator.

In the following, we first discuss the weight construction and then turn to the construction of the projection direction. The conditional variance of the stochastic error in Equation (6) is

$$\mathrm{var}\left(u^\top \frac{1}{n} \sum_{i=1}^n W_i \cdot X_i \epsilon_i \,\middle|\, \{X_i\}_{i=1}^n \right)$$

$$= u^\top \frac{1}{n^2} \sum_{i=1}^n W_i^2 \cdot f(X_i^\top \beta)(1 - f(X_i^\top \beta)) X_i X_i^\top u, \tag{7}$$

and the Hölder's inequality implies an upper bound for the remaining bias in Equation (6),

$$\begin{aligned}
&\left| \left(u^\top \frac{1}{n} \sum_{i=1}^n W_i f'(X_i^\top \hat{\beta}) X_i X_{i\cdot}^\top - e_j^\top \right)(\beta - \hat{\beta}) \right| \\
&\quad \le \left\| \frac{1}{n} \sum_{i=1}^n W_i f'(X_i^\top \hat{\beta}) X_i X_i^\top u - e_j \right\|_\infty ||\beta - \hat{\beta}||_1.
\end{aligned} \tag{8}$$

We construct the weights $\{W_i\}_{i=1}^n$ such that

$$W_i^2 \cdot f(X_i^\top \beta)(1 - f(X_i^\top \beta)) \approx W_i \cdot f'(X_i^\top \hat{\beta}). \tag{9}$$

In other words, we let the entries of the matrix $W_i^2 \cdot f(X_i^\top \beta)(1 - f(X_i^\top \beta)) X_i X_i^\top$ in Equation (7) be approximately equal to the corresponding entries of the matrix $W_i \cdot f'(X_i^\top \hat{\beta}) X_i X_i^\top$ in Equation (8). The relation (9) motivates the weight construction

$$W_i = w(X_i^\top \hat{\beta}) \quad \text{with} \quad w(z) = \frac{f'(z)}{f(z)(1 - f(z))}. \tag{10}$$

Such weight construction is rooted in the bias-variance tradeoff: together with our proposed construction of $\hat{u}$ detailed in Equations (11) and (12), the weight in Equation (10) ensures that the stochastic error in Equation (6) is the dominating term; see Remark 1 for more details. Examples of link functions and their corresponding weight functions are given in Table 1.

With the weight function $w(\cdot)$ defined in Equation (10), we construct the projection vector $\hat{u} \in \mathbb{R}^p$ as

$$\hat{u} = \arg\min_{u \in \mathbb{R}^p} u^\top \left[n^{-1} \sum_{i=1}^n w(X_i^\top \hat{\beta}) \cdot f'(X_i^\top \hat{\beta}) X_i X_i^\top \right] u, \tag{11}$$

subject to

$$\begin{aligned}
&\left\| n^{-1} \sum_{i=1}^n w(X_i^\top \hat{\beta}) \cdot f'(X_i^\top \hat{\beta}) X_i X_i^\top u - e_j \right\|_\infty \le \lambda_n \\
&\text{and} \quad \max_{1 \le i \le n} |X_i^\top u| \le \tau_n,
\end{aligned} \tag{12}$$

with $\hat{\beta}$ defined in (3), $\lambda_n = C_1 \sqrt{\log p / n}$ and $\tau_n = C_2 \sqrt{\log n}$ for some constants $C_1, C_2 > 0$. The construction of the projection vector $\hat{u}$ adopts the idea (Zhang and Zhang 2014; Javanmard and Montanari 2014a) of minimizing the variance of the stochastic error in (6) while constraining the remaining bias and the approximation error in (6).

We propose the link-specific bias-corrected estimator

$$\tilde{\beta}_j = \hat{\beta}_j + \hat{u}^\top \frac{1}{n} \sum_{i=1}^n w(X_i^\top \hat{\beta}) \cdot (y_i - f(X_i^\top \hat{\beta})) X_i. \tag{13}$$

where $\hat{u}$ is defined in Equation (11) and the weight function $w(\cdot)$ is defined in Equation (10).

Table 1. Examples of link functions and their corresponding weight functions

Link function	$f(x)$	Weight function $w(x)$
Logistic	$\frac{\exp(x)}{1+\exp(x)}$	1
Probit	$\Phi(x)$	$\frac{\phi(x)}{\Phi(x)(1-\Phi(x))}$
cdf of Student's t_ν	$1-\frac{1}{2}I_{\frac{\nu}{x^2+\nu}}(\frac{\nu}{2},\frac{1}{2}), \nu\in\mathbb{N}$	$\frac{2\Gamma(\frac{\nu+1}{2})(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})I_{\frac{\nu}{x^2+\nu}}(\frac{\nu}{2},\frac{1}{2})(1-\frac{1}{2}I_{x(t)}(\frac{\nu}{2},\frac{1}{2}))}, \nu\in\mathbb{N}$
Generalized logistic	$\frac{1}{2}\tanh^{\gamma}(\varphi x)+\frac{1}{2}, \varphi>0, \gamma\geq 1$	$\frac{2\varphi\gamma\tanh^{\gamma-1}(\varphi x)\mathrm{sech}^2(\varphi x)}{1-\tanh^{2\gamma}(\varphi x)}, \varphi>0, \gamma\geq 1$

Remark 1. In contrast to the equal weights used in the linear model (Zhang and Zhang 2014; Javanmard and Montanari 2014a), we have to carefully construct the weights for the binary outcome models. Particularly, the approximate equivalence in (9), together with the projection direction constructed in Equations (11) and (12), guarantee that the remaining bias in (6) is negligible in comparison to the stochastic error in Equation (6). If other weights were applied, then it is possible that the stochastic error is no longer dominant, and the asymptotic variance of the corresponding estimator could be larger than that based on our proposed weight.

Even for the logistic model, the bias-corrected estimators constructed in van de Geer et al. (2014) and Ma, Cai, and Li (2020) do not coincide with our proposed estimator. Specifically, different projection vectors have been proposed in van de Geer et al. (2014) and Ma, Cai, and Li (2020) based on the nodewise regression approach. Moreover, the simulation results in Section 5 indicate that our proposed method leads to more precise and flexible CIs and more powerful statistical tests.

Our proposed weight in Equation (10) has some interesting connection to the most efficient influence function. We shall emphasize that, in our construction, the weight in Equation (10) is proposed for the purpose of balancing the bias and variance, which is a different perspective from variance minimization in the construction of the most influence function. Theorem 3.5 of Tsiatis (2007) implies that, under (1), we may construct an estimator $\tilde{\beta}_j$ such that $\sqrt{n}(\tilde{\beta}_j-\beta_j)=\frac{1}{n}\sum_{i=1}^n\psi^{\mathrm{eff}}(X_i,y_i)+o_p(1)$, where $\psi^{\mathrm{eff}}(X_i,y_i)=e_j^\intercal\left[\mathbb{E}w(X_i^\intercal\beta)f'(X_i^\intercal\beta)X_iX_i^\intercal\right]^{-1}w(X_i^\intercal\beta)X_i\left[y_i-f(X_i^\intercal\beta)\right]$ is the most efficient influence function; see Section 7 of the supplementary material for the detailed derivation. It is interesting to compare the above decomposition with the error decomposition in Equation (6). Although the same weighting function is used, the inference problem in high-dimensional sparse GLMs is much more challenging as we have to construct the weight and projection direction $\hat{u}$ simultaneously such that the stochastic error in Equation (6) dominates the corresponding remaining bias term. Furthermore, we consider a practical setting where $\left[\mathbb{E}f'(X_i^\intercal\beta)w(X_i^\intercal\beta)X_iX_i^\intercal\right]^{-1}e_j$ might be dense. In such a case, it is hard to construct an accurate estimator of $\left[\mathbb{E}f'(X_i^\intercal\beta)w(X_i^\intercal\beta)X_iX_i^\intercal\right]^{-1}e_j$ in high dimensions. The proposed $\hat{u}$ is not necessarily an accurate estimate of $\left[\mathbb{E}f'(X_i^\intercal\beta)w(X_i^\intercal\beta)X_iX_i^\intercal\right]^{-1}e_j$ but guarantees that the remaining bias in Equation (6) is negligible in comparison to the stochastic error.

2.2. CIs and Statistical Tests

Under mild regularity conditions, we will show in Theorem 1 that, conditioning on $\mathcal{D}_2$ and the design covariates $\{X_i\}_{i=1}^n$ in $\mathcal{D}_1$, the asymptotic variance of $\tilde{\beta}_j$ in Equation (13) has the expression

$$v_j=\hat{u}^\intercal\left[\frac{1}{n}\sum_{i=1}^n w^2(X_i^\intercal\hat{\beta})f(X_i^\intercal\beta)(1-f(X_i^\intercal\beta))X_iX_i^\intercal\right]\hat{u}, \tag{14}$$

which can be estimated by

$$\hat{v}_j=\hat{u}^\intercal\left[\frac{1}{n}\sum_{i=1}^n\frac{[f'(X_i^\intercal\hat{\beta})]^2}{f(X_i^\intercal\hat{\beta})(1-f(X_i^\intercal\hat{\beta}))}X_iX_i^\intercal\right]\hat{u}. \tag{15}$$

Hence, we construct the $(1-\alpha)$-level CI for the regression coefficient β_j as

$$\mathrm{CI}_\alpha^*(\beta_j,\mathcal{D})=\left[\tilde{\beta}_j-\tilde{\rho}_j,\tilde{\beta}_j+\tilde{\rho}_j\right]\quad\text{with}$$
$$\tilde{\rho}_j=\max\left\{\frac{z_{\alpha/2}\hat{v}_j^{1/2}}{\sqrt{n}},C\frac{\tau_n k\log p}{n}\right\}, \tag{16}$$

where τ_n is defined after Equation (12), $z_{\alpha/2}=\Phi^{-1}(1-\alpha/2)$, and $C>0$ is a constant. In Equation (16), the proposed $\mathrm{CI}_\alpha^*(\beta_j,\mathcal{D})$ relies on the underlying sparsity k over certain regions. For the ultra-sparse regime $k\ll\frac{\sqrt{n}}{\log p\sqrt{\log n}}$, the above definition in (16) reduces to $\tilde{\rho}_j=\frac{z_{\alpha/2}\hat{v}_j^{1/2}}{\sqrt{n}}$, which does not depend on k or C (see Corollary 1). In practice, we will use $\tilde{\rho}_j=\frac{z_{\alpha/2}\hat{v}_j^{1/2}}{\sqrt{n}}$, since it is shown in Section 3.3 that the ultra-sparsity is in fact necessary for constructing any adaptive CIs.

As a direct consequence of the proposed CI, we construct a test for the null hypothesis $H_0:\beta_j=\beta_j^0$, for any given $j\in[1:p]$. In light of the debiased estimator (13) and $\mathrm{CI}_\alpha^*(\beta_j,\mathcal{D})$ in (16), we can construct the test statistic $R_j=\sqrt{n}(\tilde{\beta}_j-\beta_j^0)/\hat{v}_j^{1/2}$, and define an α-level test as $T_\alpha(R_j)=\mathbf{1}\{|R_j|\geq z_{\alpha/2}\}$. The null hypothesis H_0 is rejected whenever $T_\alpha(R_j)=1$.

2.3. Simultaneous Inference

Our proposed method can be extended to make simultaneous inference for a subset of regression coefficients such as testing multiple hypotheses with family-wise error rate (FWER) or false discovery rate (FDR) control. Suppose we are interested in simultaneously testing the null hypotheses $H_{0j}:\beta_j=0, j\in J$ where $J\subseteq[1:p]$. Let $\{\tilde{\beta}_j\}_{j\in J}$ and $\{\hat{v}_j\}_{j\in J}$ be the proposed bias-corrected estimators in Equation (13) and their corresponding

variance estimators in (14), respectively. When the goal is to control the FWER at a significance level $0 < \alpha < 1$, the classical Bonferroni correction can be applied so that H_{0j} is rejected whenever $|R_{0j}| \geq z_{\alpha/(2|J|)}$ with $R_{0j} = \sqrt{n}\tilde{\beta}_j/\hat{v}_j^{1/2}$ for $j \in J$. In this case, if we denote $H_0 = \cap_{j\in J} H_{0j}$, then asymptotically the FWER can be controlled by

$$\begin{aligned}\text{FWER} &= P_{H_0}\left(\bigcup_{j\in J}\left\{|R_{0j}| \geq z_{\alpha/(2|J|)}\right\}\right)\\ &\leq |J| \cdot P_{H_0}(|R_{0j}| \geq z_{\alpha/(2|J|)}) \leq \alpha.\end{aligned}$$

If $|J|$ is large, one may compute the threshold value in a more refined way by adopting bootstrap approaches in Chernozhukov, Chetverikov, and Kato (2017), Dezeure, Bühlmann, and Zhang (2017), and Zhang and Cheng (2017).

As is well known, when $|J|$ is large, controlling FWER with Bonferroni correction is often too conservative and controlling for the FDR is more desirable. To this end, one can apply the modified BH procedure (Liu 2013; Javanmard and Javadi 2019; Ma, Cai, and Li 2020), where one rejects the null hypothesis H_{0j} if $|R_{0j}| \geq t$ for a certain carefully chosen threshold t. A good choice for the threshold t can be seen as follows. Note that, if $J_0 \subseteq J$ is the set of true nulls contained in J, the FDR is expressed as

$$\text{FDR}(t) = \mathbb{E}\left[\frac{\sum_{j\in J_0}\mathbf{1}\{|R_{0j}| \geq t\}}{\max\{\sum_{j\in J}\mathbf{1}\{|R_{0j}| \geq t\}, 1\}}\right],$$

where $\sum_{j\in J}\mathbf{1}\{|R_{0j}| \geq t\}$ is the total number of rejected hypotheses. By assuming that the true alternatives are sparse, we approximate $|J_0|$ by $|J|$, and use the standard normal tail $2 - 2\Phi(t)$ to approximate the proportion of nulls falsely rejected among all the true nulls at the threshold level t, namely, $\frac{1}{|J_0|}\sum_{j\in J_0}\mathbf{1}\{|R_{0j}| \geq t\}$. As a consequence, for a prespecified significance level $0 < \alpha < 1$, the proposed threshold level $\hat{t}$ is defined by

$$\begin{aligned}\hat{t} = \inf\Bigg\{0 \leq t \leq \sqrt{2\log|J| - 2\log\log|J|} :\\ \frac{|J|\{2-2\Phi(t)\}}{\max\{\sum_{j\in J}\mathbf{1}(|R_{0j}| \geq t), 1\}} \leq \alpha\Bigg\},\end{aligned} \tag{17}$$

and we set $\hat{t} = \sqrt{2\log|J|}$ if $\hat{t}$ in Equation (17) does not exist. In particular, the condition $0 \leq t \leq \sqrt{2\log|J| - 2\log\log|J|}$ in (17) is determined by a careful analysis, which reflects the range of applicability of the above approximations (Liu 2013). It can be shown by using similar techniques as those in Javanmard and Javadi (2019) and Ma, Cai, and Li (2020) that the above procedure controls the FDR at level α in probability under mild conditions as $n \to \infty$ and $|J| \to \infty$.

The knockoff methods such as Candès et al. (2018) could potentially be applied to control the FDR in more restrictive settings. The knockoff approach does not require a prespecified link function, but it requires the design distribution to be known. In comparison, our approach does not require knowledge of the design distribution and can be applied to a large class of GLMs with binary outcomes. As observed in Section 4.3 of the supplementary material, our testing procedure based on the debiased estimators can be more powerful than the knockoff method; this is likely because our method takes advantage of the underlying sparse structure.

3. Theoretical Properties

3.1. Asymptotic Normality and Inference Properties

We begin with the regularity conditions for the general link function $f : \mathbb{R} \to (0, 1)$.

(L1). The link function f is twice differentiable, monotonic increasing, Lipschitz on $\mathbb{R}$, and concave on $\mathbb{R}_+$; and for any $x \in \mathbb{R}$, it holds that $f(x) + f(-x) = 1$.

(L2). There exist some constants $C_1, C_2 > 0$ such that, for all $x \geq 0, f(x) \leq \Phi(C_1 x)$ where $\Phi(x)$ is the standard Gaussian cdf, and $\max\{\frac{f'(x)}{x(1-f(x))}, x^2 f'(x)\} < C_2$.

(L3). There exist some constants $c_1, c_2 > 0$ such that $\sup_{x\in\mathbb{R}}|xf''(x+\omega)| \leq c_1$ and $|\omega| < c_2$.

(L4). For $\ell_f(\beta)$ defined in (2), there exists some constant $C > 1$ such that the Hessian matrix $\ell_f''(\beta)$ can be expressed as $\ell_f''(\beta) = \frac{1}{n}\sum_{i=1}^n h(\beta; y_i, X_i)X_i X_i^\top$ for some $h(\beta; y_i, X_i) > 0$ satisfying

$$\begin{aligned}&\max_{1\leq i\leq n}\left|\log h(\beta + b; y_i, X_i) - \log h(\beta; y_i, X_i)\right|\\ &\quad\leq C(|X_i^\top\beta|^2 + |X_i^\top b|^2 + |X_i^\top b|).\end{aligned} \tag{18}$$

The above regularity conditions are mild as they are satisfied by a large class of link functions, including but not limited to the examples we listed at the beginning of Section 1. Specifically, for the logistic link function in Example 1, conditions (L1) to (L3) are easy to verify and condition (L4) follows from Example 8 of Huang and Zhang (2012). For the probit link in Example 2, conditions (L1) to (L3) follow directly from the properties of the Gaussian distribution, though condition (L4) is less straightforward. Similarly, these conditions can also be verified for the class of generalized logistic functions for any $\varphi > 0$ and $\gamma \geq 1$, as well as the cdfs of Student's t_ν-distributions with any $\nu \in \mathbb{N}$ in Example 3. See the detailed proofs of these statements in Section 3 of the Supplement.

For the random design covariates and their distribution P_X, we assume

(A). $\{X_i\}_{1\leq i\leq 2n}$ are independent and identically distributed sub-Gaussian random vectors, that is, there exists a constant $c \in \mathbb{R}$ satisfying $\mathbb{E}\exp\{v^\top X\} \leq e^{||v||_2^2 c^2/2}$ for all $v \in \mathbb{R}^p$.

Such a general characterization of the design covariates includes the special case where $X_{i1} = 1$ for all $1 \leq i \leq 2n$ so that β_1 represents the intercept. Define $\mathbf{\Sigma} = \mathbb{E}[X_i X_i^\top] \in \mathbb{R}^{p\times p}$. We focus on the following parameter space indexed by the sparsity level k,

$$\begin{aligned}\Theta(k) = \Big\{\theta = (\beta, \mathbf{\Sigma}) : ||\beta||_0 \leq k, ||\beta||_2 \leq C, M^{-1}\\ \leq \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) \leq M\Big\}\end{aligned} \tag{19}$$

for constants $M > 1$ and $C > 0$ independent of n and p.

The following theorem establishes the asymptotic property of the bias-corrected estimator $\tilde{\beta}_j$.

Theorem 1. Suppose that Conditions (L1)–(L4), and (A) hold, and $(\beta, \mathbf{\Sigma}) \in \Theta(k)$. For any $j \in [1 : p]$, if $k \ll \frac{n}{\log n \log p}$, then we have $\tilde{\beta}_j - \beta_j = A_n + B_n$, where conditioning on $\mathcal{D}_2 = \{(X_i, y_i)\}_{i=n+1}^{2n}$ and $\{X_i\}_{i=1}^n$, $\sqrt{n}A_n/v_j^{1/2} \to_d N(0, 1)$ with v_j defined in (14) and $|B_n| \lesssim k\log p\sqrt{\log n/n}$ with probability

at least $1 - p^{-c} - n^{-c}$ for some constant $c > 0$. Additionally, if $k \ll \frac{\sqrt{n}}{\log p \sqrt{\log n}}$, then $\sqrt{n}(\tilde{\beta}_j - \beta_j)/v_j^{1/2} \to_d N(0,1)$.

A few remarks are in order for the above theorem. First, we have removed several stringent but commonly used assumptions for the high-dimensional GLM inference. For a general class of link functions satisfying (L1) to (L4), Theorem 1 only requires the general sub-Gaussian design with parameters $(\beta, \mathbf{\Sigma}) \in \Theta(k)$, which includes many important cases such as Gaussian, bounded, and binary designs, or any combinations of them. This makes our proposed LSW estimator applicable to many practical settings. As a comparison, the existing inference methods (van de Geer 2008; van de Geer et al. 2014; Ning and Liu 2017; Ma, Cai, and Li 2020; Guo et al. 2020; Shi et al. 2020) require in general the bounded individual probability condition that $P(y_i = 1|X_i) \in (c, 1-c)$ for all $1 \le i \le n$ and some $c \in (0, 1/2)$, or the bounded design assumption. In contrast, under our assumptions (A) and $(\beta, \mathbf{\Sigma}) \in \Theta(k)$, we have $\mathbb{E}|X_i^\intercal \beta| \le \sqrt{\mathbb{E}|X_i^\intercal \beta|^2} \le \sqrt{M}C$. Together with (L1), this implies $c \le P(y_i = 1) \le 1 - c$ for $c = f(-\sqrt{M}C)$. That is, in Theorem 1, we have relaxed the stringent bounded individual probability condition $P(y_i = 1|X_i) \in (c, 1-c)$ for all $1 \le i \le n$ to the balanced outcome assumption $P(y_i = 1) \in (c, 1-c)$, which can be directly verified for any given dataset.

Second, the removal of the bounded individual probability condition is not just a technical innovation but has profound practical implications. We consider in Section 5 a setting where part of the observations do not satisfy the bounded probability condition $P(y_i = 1|X_i) \in (c, 1-c)$ but the outcome variable is balanced, and demonstrate that the proposed procedure outperforms the state-of-the-art methods in the literature; see Section 5.1 for details.

Third, commonly used theoretical assumptions such as the sparse inverse population Hessian condition (van de Geer et al. 2014; Belloni, Chernozhukov, and Wei 2016; Ning and Liu 2017; Janková and van de Geer 2018) and the sparse precision matrix condition (van de Geer et al. 2014; Ma, Cai, and Li 2020), are completely removed from our analysis, as they are also difficult to verify in practice, and can potentially limit the applicability of the methods in practical settings (Xia, Nan, and Li 2020).

The proof of Theorem 1, concerning a large class of high-dimensional GLMs, is involved and consists of a careful analysis of the debiased Lasso estimator (13) as well as the projection vector $\hat{u}$ defined by Equations (11) and (12). A key step is to establish the asymptotic normality of the stochastic error in Equation (6), which is obtained under mild conditions with the sample splitting procedure. See Section 7 for more discussions. In addition, the validity of these arguments relies on certain theoretical properties of $\hat{\beta}$ given by Equation (3), which is summarized by the following theorem.

Theorem 2. Under Conditions (L1), (L2), (L4), and (A), suppose $(\beta, \mathbf{\Sigma}) \in \Theta(k)$ and $k \lesssim n/\log p$. Then the event $\mathcal{B} = \cap_{i=1}^3 \mathcal{B}_i$ holds with probability at least $1 - p^{-c}$, where $\mathcal{B}_1 = \{||\hat{\beta} - \beta||_1 \lesssim k\sqrt{\frac{\log p}{n}}\}$, $\mathcal{B}_2 = \{||\hat{\beta} - \beta||_2 \lesssim \sqrt{\frac{k \log p}{n}}\}$ and $\mathcal{B}_3 = \{\frac{1}{n}\sum_{i=1}^n [X_i^\top(\hat{\beta} - \beta)]^2 \lesssim \frac{k \log p}{n}\}$.

Theorem 2 establishes the rate of convergence for the GLM Lasso estimator $\hat{\beta}$ under the sub-Gaussian random design specified by the condition (A) for a general class of link functions satisfying the conditions (L1), (L2), and (L4). This theorem is novel and can be of independent interest. Importantly, building upon the convex analysis and the empirical process theory, a new analytical framework was developed so that a large class of binary outcome GLMs can be simultaneously analyzed. The result extends those of Negahban et al. (2010) and Huang and Zhang (2012), which focused on the GLMs with canonical links. In addition, Theorem 2 also generalizes the results of van de Geer (2008), which focuses on the bounded design. Comparing to the weaker condition that $k \lesssim n/\log p$ in Theorem 2, the slightly more stringent condition $k \ll \frac{n}{\log n \log p}$ in Theorem 1 is to ensure that $v_j \asymp 1$.

Remark 2. It can be seen from the proof of Theorem 1 that, any initial estimator satisfying the properties given by Theorem 2 can be used to replace $\hat{\beta}$ in (3) for constructing the bias-corrected estimator $\tilde{\beta}_j$, without altering the asymptotic properties described in Theorem 1.

Building upon Theorem 1, we obtain the following theorem concerning the asymptotic coverage probability of the proposed $\mathrm{CI}_\alpha^*(\beta_j, \mathcal{D})$ as well as an upper bound for its expected length.

Theorem 3. Suppose that Conditions (L1)–(L4), and (A) hold, and $\theta = (\beta, \mathbf{\Sigma}) \in \Theta(k)$. If $k \ll \frac{n}{\log n \log p}$, then for any constant $0 < \alpha < 1$ and any $j \in [1:p]$, the $\mathrm{CI}_\alpha^*(\beta_j, \mathcal{D})$ defined in (16) satisfies

$$\underline{\lim}_{n,p\to\infty} \inf_{\theta \in \Theta(k)} P_\theta(\beta_j \in \mathrm{CI}_\alpha^*(\beta_j, \mathcal{D})) \ge 1 - \alpha. \tag{20}$$

$$\sup_{\theta \in \Theta(k)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha^*(\beta_j, \mathcal{D})) \lesssim \frac{1}{\sqrt{n}} + \frac{k \log p \sqrt{\log n}}{n}, \tag{21}$$

where $L(\mathrm{CI}_\alpha^*(\beta_j, \mathcal{D}))$ denotes the length of $\mathrm{CI}_\alpha^*(\beta_j, \mathcal{D})$.

Compared with the CIs proposed by van de Geer et al. (2014) and Belloni, Chernozhukov, and Wei (2016), which only have guaranteed coverage when $k \ll \frac{\sqrt{n}}{\log p}$, the proposed CIs have guaranteed coverage for all $k \ll \frac{n}{\log n \log p}$, including the moderately sparse region $\frac{\sqrt{n}}{\log p} \lesssim k \ll \frac{n}{\log n \log p}$.

The next result concerns the behavior of the proposed CIs over the ultra-sparse region.

Corollary 1. Suppose that Conditions (L1)–(L4), and (A) hold, and $(\beta, \mathbf{\Sigma}) \in \Theta(k)$. If $k \ll \frac{\sqrt{n}}{\sqrt{\log n \log p}}$, then for any constant $0 < \alpha < 1$ and any $j \in [1:p]$, the $\mathrm{CI}_\alpha^*(\beta_j, \mathcal{D})$ defined in Equation (16) admits the expression $\left[\tilde{\beta}_j - z_{\alpha/2}\hat{v}_j^{1/2}/\sqrt{n}, \tilde{\beta}_j + z_{\alpha/2}\hat{v}_j^{1/2}/\sqrt{n}\right]$, and satisfies $\underline{\lim}_{n,p\to\infty} \inf_{\theta\in\Theta(k)} P_\theta(\beta_j \in \mathrm{CI}_\alpha^*(\beta_j, \mathcal{D})) \ge 1 - \alpha$, and $\sup_{\theta\in\Theta(k)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha^*(\beta_j, \mathcal{D})) \lesssim 1/\sqrt{n}$.

The following corollary, as a result of Corollary 1, concerns the Type I error of the proposed test $T_\alpha(R_j)$ and its statistical power under some local alternatives.

Corollary 2. Suppose that Conditions (L1)–(L4), and (A) hold, and $(\beta, \mathbf{\Sigma}) \in \Theta(k)$. If $k \ll \frac{\sqrt{n}}{\sqrt{\log n \log p}}$, then for any constant $0 < \alpha < 1$ and $j \in [1:p]$, we have $\overline{\lim}_{n,p\to\infty} \sup_{\theta\in H_0} P_\theta(T_\alpha(R_j) = 1) \le \alpha$ where $H_0 = \{\theta \in \Theta(k) : \beta_j = \beta_j^0\}$. Moreover, for any $0 < q < 1$, there exists some $c > 0$ such that, for any $\phi \ge cn^{-1/2}$, we have $\underline{\lim}_{n,p\to\infty} \inf_{\theta\in H_1(\phi)} P_\theta(T_\alpha(R_j) = 1) \ge 1 - q$ where $H_1(\phi) = \{\theta \in \Theta(k) : |\beta_j - \beta_j^0| \ge \phi\}$.

3.2. Optimal Expected Lengths and Efficiency

We now study the minimax optimality of CIs in the high-dimensional GLM with binary outcomes. For any $0 < \alpha < 1$, $j \in [1:p]$ and a given parameter space Θ of $\theta = (\beta, \mathbf{\Sigma})$, we denote by $\mathcal{I}_\alpha(\Theta, \beta_j)$ the set of all $(1-\alpha)$ level CIs for β_j over Θ,

$$\mathcal{I}_\alpha(\Theta, \beta_j) = \Big\{\mathrm{CI}_\alpha(\beta_j, \mathcal{D}) : \inf_{\theta\in\Theta} P_\theta(\beta_j \in \mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \ge 1 - \alpha\Big\}. \tag{22}$$

The following theorem establishes the minimax lower bound for the CI's expected length for a large class of link functions over the parameter space $\Theta(k)$ under the Gaussian design.

Theorem 4. Suppose that the link function f satisfies Conditions (L1) and (L2), $\{X_i\}_{i=1}^{2n} \overset{\text{iid}}{\sim} N(0, \mathbf{\Sigma})$, $0 < \alpha < 1/2$ and $k \lesssim \min\{p^c, \frac{n}{\log p}\}$ for some $0 \le c < 1/2$. Then for any $j \in [1:p]$,

$$\inf_{\mathrm{CI}_\alpha(\beta_j,\mathcal{D})\in\mathcal{I}_\alpha(\Theta(k),\beta_j)} \sup_{\theta\in\Theta(k)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \gtrsim \frac{1}{\sqrt{n}} + k\frac{\log p}{n}, \tag{23}$$

where $L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \in \mathbb{R}$ is the length of $\mathrm{CI}_\alpha(\beta_j, \mathcal{D})$.

The proof of Theorem 4 requires a careful construction of two hypotheses belonging to the parameter space $\Theta(k)$, and a nontrivial calculation of the chi-squared divergence between the associated two probability measures. The following two lemmas play a key role in the proof and can be of independent interest for establishing lower bounds for other GLM problems. The first lemma reduces the calculation of chi-squared divergence to some link-specific nonlinear moment quantity.

Lemma 1. Under model (1) with any link function $f : \mathbb{R} \to (0,1)$ and $X_i \sim N(0, \mathbf{\Sigma})$, let $p_f(X_i, y_i; \beta, \mathbf{\Sigma})$ be the joint density function of (X_i, y_i). Then for any $(\beta, \mathbf{\Sigma})$ and $(\beta', \mathbf{\Sigma}') \in \Theta(k)$,

$$\int \frac{p_f(X_i, y_i; \beta, \mathbf{\Sigma}) p_f(X_i, y_i; \beta', \mathbf{\Sigma}')}{p_f(X_i, y_i; 0, \mathbf{I})} = \frac{4\det(\mathbf{\Omega})\det(\mathbf{\Omega}')}{\det(\mathbf{\Omega} + \mathbf{\Omega}' - \mathbf{I})} \cdot \mathbb{E} f(Z^\top\beta) f(Z^\top\beta'), \tag{24}$$

where $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, $\mathbf{\Omega}' = (\mathbf{\Sigma}')^{-1}$ and $Z \sim N(0, (\mathbf{\Omega} + \mathbf{\Omega}' - \mathbf{I})^{-1})$.

The second lemma provides a sharp upper bound for $\mathbb{E} f(Z^\top\beta) f(Z^\top\beta')$, which is a special case of the extreme nonlinear correlations studied by Lancaster (1957), Yu (2008), and Guo and Zhang (2019). This inequality is proved in the Supplement using the Wiener-Itô chaotic decomposition theory (Nualart 2006) and properties of the Hermite polynomials.

Lemma 2. For a bivariate vector $(X, Y) \sim N(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ for some $\sigma^2 \le 1$ and $\rho \in [0, 1)$, for any $f : \mathbb{R} \to (0,1)$ satisfying (L1) and (L2), we have $\mathbb{E} f(X) f(Y) \le \frac{1}{4} + C\sigma^2\rho$, for some universal constant $C > 0$.

Combining Theorems 3 and 4, we can establish the following minimax optimal rates for the length of CIs for β_j. Specifically, under the Gaussian design,

$$\inf_{\mathrm{CI}_\alpha(\beta_j,\mathcal{D})\in\mathcal{I}_\alpha(\Theta(k),\beta_j)} \sup_{\theta\in\Theta(k)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \asymp \frac{1}{\sqrt{n}} + k\frac{\log p}{n}, \tag{25}$$

where the second term $k\log p/n$ holds up to a $\sqrt{\log n}$ factor. The optimal rate is attained by the proposed CI (16) with the second term $k\log p/n$ holding up to a $\sqrt{\log n}$ factor. The optimal rate in (25) agrees with the minimax rates for the length of CIs in the high-dimensional linear regression (Cai and Guo 2017), up to a $\sqrt{\log n}$ factor in the second term $k\log p/n$.

Finally, we discuss the efficiency of the proposed estimators. Efficiency in high-dimensional linear regression has been discussed in van de Geer et al. (2014) and Jankova and van de Geer (2018). The next result concerns the lower bound for estimating a single regression coefficient.

Proposition 1. Let $f : \mathbb{R} \to (0,1)$ be any link function satisfying (L1) and (L2), $Z = \{(X_i, y_i)\}_{i=1}^n$ be n independent samples with $X_i \sim N(0, \mathbf{\Sigma})$, and $p_f(Z; \beta, \mathbf{\Sigma})$ be their joint probability density function. Then, for any given $(\beta, \mathbf{\Sigma}) \in \Theta(k)$, $j \in [1:p]$ and any unbiased estimator $\hat\beta_j$ of β_j based on Z, such that $\frac{\partial}{\partial\beta_j}[\int \hat\beta_j p_f(Z; \beta, \mathbf{\Sigma}) dZ] = \int \hat\beta_j [\frac{\partial}{\partial\beta_j} p_f(Z; \beta, \mathbf{\Sigma})] dZ$ whenever the right-hand side exists, we have $\mathrm{Var}(\hat\beta_j) \ge [I(\beta)]_{jj}^{-1}/n$, where $I(\beta) = \mathbb{E}\big[\frac{[f'(X_i^\top\beta)]^2}{f(X_i^\top\beta)(1-f(X_i^\top\beta))} X_i X_i^\top\big]$.

The proof of Theorem 1 (especially Lemma 3) in the Supplement implies that, for $k \ll \frac{\sqrt{n}}{\sqrt{\log n \log p}}$, $\mathrm{Var}(\tilde\beta_j) \le (1 + o(1))[I(\beta)]_{jj}^{-1}/(\xi n)$ for all $j \in [1:p]$, where n is the total sample size and the constant $\xi \in (0,1)$ is the proportion of samples used for the bias correction step. In other words, for $0 < \delta < 1$, the efficiency of the proposed estimator is at least $1 - \delta$ by choosing $\xi > 1 - \delta$. As a consequence, when data splitting is applied, as long as the samples used for the initial Lasso estimator is not too scarce, the more samples used for bias correction, the more efficient the proposed method is; see also Section 4.2 of the supplementary material for numerical evidences.

3.3. Adaptivity of Optimal CIs

Now we study the adaptivity of the optimal CIs over a sequence of parameter spaces indexed by k. We follow the framework of Cai and Guo (2017) and define the following benchmark. For a given $j \in [1:p]$, $k_1 \le k$ and $\Theta(k_1) \subset \Theta(k)$, define

$$L_\alpha^*(\Theta(k_1), \Theta(k), \beta_j) = \inf_{\mathrm{CI}_\alpha(\beta_j,\mathcal{D})\in\mathcal{I}_\alpha(\Theta(k),\beta_j)} \sup_{\theta\in\Theta(k_1)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})), \tag{26}$$

where $\mathcal{I}_\alpha(\Theta(k), \beta_j)$ is defined in Equation (22). For $k_1 \le k$, $L^*_\alpha(\Theta(k_1), \Theta(k), \beta_j)$ characterizes the infimum of the maximum expected length over $\Theta(k_1)$ among all CIs having coverage over $\Theta(k)$. We say that a $\mathrm{CI}_\alpha(\beta_j, \mathcal{D})$ is rate-optimal adaptive over $\Theta(k_1)$ and $\Theta(k)$ if $\mathrm{CI}_\alpha(\beta_j, \mathcal{D}) \in \mathcal{I}_\alpha(\Theta(k), \beta_j)$ and

$$\sup_{\theta\in\Theta(k_1)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \asymp L^*_\alpha(\Theta(k_1), \Theta(k_1), \beta_j),$$
$$\sup_{\theta\in\Theta(k)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \asymp L^*_\alpha(\Theta(k), \Theta(k), \beta_j).$$

That is, $\mathrm{CI}_\alpha(\beta_j, \mathcal{D})$ has the correct coverage over the larger parameter space $\Theta(k)$ and achieves the optimal expected length simultaneously over $\Theta(k_1)$ and $\Theta(k)$.

A comparison of $L^*_\alpha(\Theta(k_1), \Theta(k), \beta_j)$ and $L^*_\alpha(\Theta(k_1), \Theta(k_1), \beta_j)$ can be used to decide whether it is possible to construct rate-optimal adaptive CIs over the nested spaces $\Theta(k_1) \subset \Theta(k)$. For $\mathrm{CI}_\alpha(\beta_j, \mathcal{D}) \in \mathcal{I}_\alpha(\Theta(k), \beta_j)$, we apply the definition of $L^*_\alpha(\Theta(k_1), \Theta(k), \beta_j)$ in Equation (26) and obtain

$$\sup_{\theta\in\Theta(k_1)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \ge L^*_\alpha(\Theta(k_1), \Theta(k), \beta_j).$$

As a consequence, whenever $L^*_\alpha(\Theta(k_1), \Theta(k), \beta_j) \gg L^*_\alpha(\Theta(k_1), \Theta(k_1), \beta_j)$, then the rate-optimal adaptation between $\Theta(k_1)$ and $\Theta(k)$ is impossible since

$$\sup_{\theta\in\Theta(k_1)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D})) \ge L^*_\alpha(\Theta(k_1), \Theta(k), \beta_j) \gg L^*_\alpha(\Theta(k_1), \Theta(k_1), \beta_j).$$

The following theorem establishes the lower bound for $L^*_\alpha(\Theta(k_1), \Theta(k), \beta_j)$.

Theorem 5. Suppose that Conditions (L1) and (L2) hold, $\{X_i\}_{i=1}^{2n} \overset{\text{iid}}{\sim} N(0, \mathbf{\Sigma})$, $0 < \alpha < 1/2$ and $k_1 \le k \lesssim \min\left\{p^c, \frac{n}{\log p}\right\}$ for some constant $0 \le c < 1/2$. Then for any given $j \in [1 : p]$,

$$L^*_\alpha(\Theta(k_1), \Theta(k), \beta_j) \gtrsim \frac{1}{\sqrt{n}} + k\frac{\log p}{n}. \tag{27}$$

Combining the above theorem with the second statement of Theorem 3, we have

$$L^*_\alpha(\Theta(k_1), \Theta(k), \beta_j) \asymp \frac{1}{\sqrt{n}} + k\frac{\log p}{n}, \tag{28}$$

where the second term $k\log p/n$ is up to a $\sqrt{\log n}$ factor. In particular, for $k_1 \ll \min\{\frac{k}{\sqrt{\log n}}, \frac{n}{\log p \log n}\}$ and $\frac{\sqrt{n}}{\log p} \ll k \le \min\{p^c, \frac{n}{\log p}\}$ for some constant $0 \le c < 1/2$, we have

$$L^*_\alpha(\Theta(k_1), \Theta(k), \beta_j) \gtrsim \frac{k\log p}{n} \gg \frac{1}{\sqrt{n}} + k_1\frac{\log p\sqrt{\log n}}{n} \gtrsim L^*_\alpha(\Theta(k_1), \Theta(k_1), \beta_j).$$

This rules out the possibility of constructing rate-optimal adaptive CIs beyond regime $k \lesssim \frac{\sqrt{n}}{\log p}$. When $k_1 \le k \ll \frac{\sqrt{n}}{\sqrt{\log n}\log p}$, we have $L^*_\alpha(\Theta(k_1), \Theta(k), \beta_j) \asymp L^*_\alpha(\Theta(k_1), \Theta(k_1), \beta_j) \asymp 1/\sqrt{n}$ and our proposed CI in (16) achieves the optimal rates simultaneously over $\Theta(k_1)$ and $\Theta(k)$. See Figure 1 for an illustration.

4. Statistical Inference When the Design Distribution Is Known

Similar to the inference theory for the high-dimensional linear regression, the analysis of the high-dimensional GLMs demonstrates the important role of the design distribution P_X in determining the fundamental difficulty of the inference problem. Let us consider the problem of constructing CIs with the prior knowledge of the design distribution. Although such knowledge is usually not readily available in practice, as pointed out in Cai and Guo (2020), insights from such an analysis can be instrumental in semi-supervised learning.

Suppose the design distribution P_X is known and has density function $p(X)$. To make inference about β_j for some given $j \in [1 : p]$, one could start with the joint density function $p(y_i, X_i; \beta) = p(y_i|X_i)p(X_i)$ and, calculate the joint density for $(y_i, X_{ij}, V_i) \equiv (y_i, X_{ij}, \beta_{-j}^\top X_{i,-j})$ as

$$p_{\eta_j}(y_i, X_{ij}, V_i) = p(y_i|X_{ij}, V_i)p_{\zeta_j}(X_{ij}, V_i), \tag{29}$$

where $\eta_j = (\beta_j, \zeta_j)$ and ζ_j only depends on β_{-j} and P_X. Since the variable V_i is not observable, we consider instead the density $p_{\eta_j}(y_i, X_{ij}) = \int p_{\eta_j}(y_i, X_{ij}, V_i)dV_i$ by marginalizing out V_i and define the marginal maximum likelihood estimator (MMLE) of η_j based on the observations $\mathcal{D}' = \{(y_i, X_{ij})\}_{i=1}^n$ as

$$\hat\eta_j^{ML} = (\hat\beta_j^{ML}, \hat\zeta_j^{ML}) = \arg\max_{\eta_j} \sum_{i=1}^n \log p_{\eta_j}(y_i, X_{ij}). \tag{30}$$

Based on the classical large sample theory for the MLEs, for a large class of regular likelihood functions, a $(1-\alpha)$-level CI for the regression coefficient β_j can be constructed as

$$\mathrm{CI}^{**}_\alpha(\beta_j, \mathcal{D}') = \left[\hat\beta_j^{ML} - \frac{z_{\alpha/2}\tilde\nu_j^{1/2}}{\sqrt{n}}, \hat\beta_j^{ML} + \frac{z_{\alpha/2}\tilde\nu_j^{1/2}}{\sqrt{n}}\right], \tag{31}$$

with $\tilde\nu_j = (\hat I_{11} - \hat I_{12}^2/\hat I_{22})^{-1}$, where $\hat I_{11} = \frac{1}{n}\sum_{i=1}^n \left[\frac{\partial \log p_{\eta_j}(y_i, X_{ij})}{\partial \beta_j}\right]^2$, $\hat I_{22} = \frac{1}{n}\sum_{i=1}^n \left[\frac{\partial \log p_{\eta_j}(y_i, X_{ij})}{\partial \zeta_j}\right]^2$, and $\hat I_{12} = \frac{1}{n}\sum_{i=1}^n \left[\frac{\partial \log p_{\eta_j}(y_i, X_{ij})}{\partial \beta_j} \frac{\partial \log p_{\eta_j}(y_i, X_{ij})}{\partial \zeta_j}\right]$.

The following theorem presents the theoretical guarantee for the asymptotic coverage probability of the proposed $\mathrm{CI}^{**}_\alpha(\beta_j, \mathcal{D}')$ and an upper bound for its expected length, under the classical regularity conditions for the MLEs. For reason of space, we delay the explicit statements of these conditions, denoted as (C1) to (C4), to Section 1.6 in the Supplement.

Theorem 6. For any $j \in [1 : p]$, suppose $(\beta, P_X) \in \Theta_P = \{(\beta, P_X) : \text{the density } p_{\eta_j}(y_i, X_{ij}) \text{ satisfies the classical regularity conditions (C1) - (C4)}\}$. Then there exists a sequence $\{\hat\eta_j^{ML}\}$ of estimators satisfying Equation (30) such that the $\mathrm{CI}^{**}_\alpha(\beta_j, \mathcal{D}')$ in Equation (31) satisfies

$$\underline{\lim}_{n,p\to\infty} \inf_{(\beta,P_X)\in\Theta_P} P_{\eta_j}(\beta_j \in \mathrm{CI}^{**}_\alpha(\beta_j, \mathcal{D}')) \ge 1 - \alpha, \tag{32}$$

$$\sup_{(\beta,P_X)\in\Theta_P} \mathbb{E}_{\eta_j} L(\mathrm{CI}^{**}_\alpha(\beta_j, \mathcal{D}')) \lesssim \frac{1}{\sqrt{n}}, \tag{33}$$

where $L(\mathrm{CI}^{**}_\alpha(\beta_j, \mathcal{D}')) = 2z_{\alpha/2}\tilde\nu_j^{1/2}/\sqrt{n}$ is the length of $\mathrm{CI}^{**}_\alpha(\beta_j, \mathcal{D}')$.

As an important consequence, the next result establishes the minimax optimality of the proposed CIs under the Gaussian design where $P_X = N(0, \mathbf{\Sigma}_0)$ for some known $\mathbf{\Sigma}_0$. In this case, we have $p_{\eta_j}(y, X_j) = \int f(X_j\beta_j + V)^y(1 - f(X_j\beta_j + V))^{1-y} p_{\zeta_j}(X_j, V)dV$ where $\eta_j = (\beta_j, \zeta_j)$ and $p_{\zeta_j}(X_j, V)$ is the probability distribution function of a centered bivariate normal random vector whose covariance matrix, parameterized by ζ_j, only depends on β_{-j} and $\mathbf{\Sigma}_0$. To this end, we need the following condition for the link function.

(**L5**) There exists some constant $a > 0$ such that $f(x) \le \frac{1}{2}\tanh(ax) + \frac{1}{2}$ for all $x \ge 0$.

Theorem 7. Suppose Conditions (L1) and (L5) hold and $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} N(0, \mathbf{\Sigma}_0)$. For any $j \in [1 : p]$, define $\Theta_0(k) = \Theta(k) \cap \{(\beta, \mathbf{\Sigma}_0) : \text{the density } p_{\eta_j}(y_i, X_{ij}) \text{ satisfies the classical regularity conditions (C1)–(C4)}\}$. Then for any $k \le p$, we have

$$\inf_{\mathrm{CI}_\alpha(\beta_j, \mathcal{D}') \in \mathcal{I}_\alpha(\Theta_0(k), \beta_j)} \sup_{\theta \in \Theta_0(k)} \mathbb{E}_\theta L(\mathrm{CI}_\alpha(\beta_j, \mathcal{D}')) \asymp \frac{1}{\sqrt{n}}, \quad (34)$$

where the above optimal rate can be attained by $\mathrm{CI}_\alpha^{**}(\beta_j, \mathcal{D}')$ defined by Equation (31).

Remark 3. The above theorem requires the link function $f(x)$, in addition to satisfying the Condition (L1), to be dominated by some affine hyperbolic tangent function on $\mathbb{R}_+$. Again, such requirements are met by a wide range of link functions such as the logistic link function, the generalized logistic functions with any $\varphi > 0$ and $\gamma > 1$, as well as the cdfs of Student's t_ν-distribution for any $\nu \ge 1$ in Example 3. See Section 3 in the Supplement for the proofs.

Lastly, for any given $j \in [1 : p]$, about the null hypothesis $H_0 : \beta_j = \beta_j^0$, the $\mathrm{CI}_\alpha^{**}(\beta_j, \mathcal{D}')$ implies the test statistic $S_j = \frac{\sqrt{n}(\hat{\beta}_j^{ML} - \beta_j^0)}{\tilde{v}_j^{1/2}}$ and the corresponding α-level test $T_\alpha(S_j) = \mathbf{1}\{|S_j| > z_{\alpha/2}\}$. Again, similar theoretical guarantees for the Type I error and the statistical power can be obtained by applying Theorem 6.

5. Simulations

In this section, we evaluate the empirical performance of our proposed method and compare it with some existing inference methods for high-dimensional binary outcome GLMs. Regarding the CI construction, we focus on the coverage probabilities and the lengths of CIs for some regression coefficients; regarding the hypothesis testing, we evaluate the Type I error and the statistical power.

5.1. CIs for High-Dimensional Logistic Regression

We start with the high-dimensional logistic regression model. Specifically, we set $n = 400$ and let p vary from 400 to 1300. The sparsity level k varies from 20 to 35. For the true regression coefficients, given the support $\mathcal{S}$ such that $|\mathcal{S}| = k$, we set $|\beta_j| = \psi 1\{j \in \mathcal{S}\}$ for $j = 1, ..., p$ with equal proportions of ψ and $-\psi$. For the design covariates, we consider two scenarios: X_i's are generated from a multivariate Gaussian distribution with covariance matrix as either (1) $\mathbf{\Sigma} = \mathbf{\Sigma}_M$, where $\mathbf{\Sigma}_M$ is a $p \times p$ blockwise diagonal matrix of 10 identical unit diagonal Toeplitz matrices whose off-diagonal entries descend from 0.6 to 0 (see Section 4 in the supplementary material for the explicit form), or (2) $\mathbf{\Sigma} = r \cdot \mathbf{I}_p$ where we set $r = 0.02$ to ensure the bounded individual probability condition (see the right panel of Figure 2). We consider CIs for a nonzero regression coefficient $\beta_2 = -\psi$ and a zero coefficient $\beta_{100} = 0$, and set the desired confidence level as 95%. We compare our proposed LSW method without sample splitting with some existing methods including (i) the CIs based on the weighted low-dimensional projection method ("wlp") proposed by Ma, Cai, and Li (2020), (ii) the CIs based on the GLM Lasso projection method ("lproj") proposed by van de Geer et al. (2014) and implemented by the function `lasso.proj` in the R package `hdi`, (iii) the CIs based on the GLM Ridge projection method ("rproj") of Bühlmann (2013) and implemented by the function `ridge.proj` in the R package `hdi`, and (iv) the CIs based on the recursive online-score estimation method ("rose") proposed by Shi et al. (2020). The numerical results for the nonzero coefficient are summarized in Tables 2 and 3, where each entry represents an average over 500 rounds of simulations. For reason of space, the results for the zero coefficient are collected in Tables S4.1 and S4.2 in the Supplement.

In Table 2 and Table S4.1 in the supplementary material, we find that, when $\mathbf{\Sigma} = 0.02\mathbf{I}_p$ and $\psi = 1$, the CIs for the nonzero coefficient β_2 and the zero coefficient β_{100} based on "LSW," "wlp," and "lproj" achieve the desired coverage probabilities, with our proposed "LSW" having slightly shorter length in many settings. The CIs based on "rproj" have low coverage probabilities for the nonzero coefficient when p is small, and the CIs based on "rose" have low coverage probabilities for both coefficients. From Table 3 and Table S4.2, we observe that, when $\mathbf{\Sigma} = \mathbf{\Sigma}_M$ and $\psi = 0.5$, both "LSW" and "wlp" achieve the desired coverage probabilities, with "LSW" having shorter length in all settings. In contrast, the CIs based on "lproj" and "rproj" fail to achieve the desirable coverage probabilities for the nonzero coefficient (Table 3), while the CIs based on "rose" fail to achieve the desirable coverage probabilities for the zero coefficient (Table S4.2).

To better understand the discrepancy in performance between Tables 2 and 3, in Figure 2 we compare the individual case probabilities of the samples associated with such two settings, respectively. Figure 2 left panel shows that in the setting of Table 3 there are a significant portion of samples whose case probabilities are close to either 0 or 1, whereas in the right panel, under the setting of Table 2, most of the case probabilities are bounded away from 0 and 1. This may explain why "lproj" and "rproj" failed in Table 3 but not so in Table 2, as both methods rely on the bounded individual probability condition. The removal of such theoretical conditions required by most existing methods is not only a technical innovation but also a methodological improvement.

Moreover, we also perform simulations to compare the efficiency of our proposed LSW estimator with and without data splitting. Specifically, in Section 4.2 of our supplementary material, we show that, while both procedures with and without data splitting produced CIs with desired coverage probability, the method without sample splitting is more efficient (i.e., with

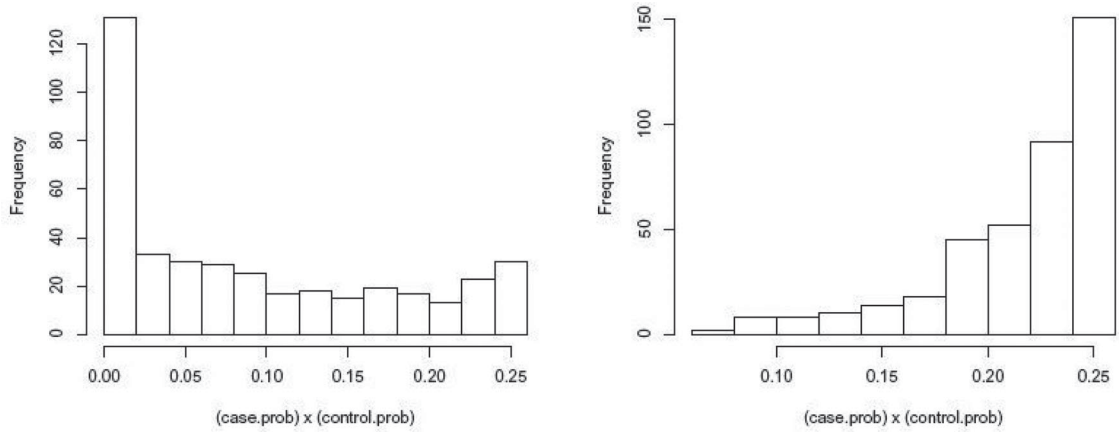

Figure 2. Histograms of $P(y_i = 1|X_i)(1 - P(y_i = 1|X_i))$ associated to the two settings corresponding to Table 2 (left) and Table 3 (right), with $p = 1000$, $n = 400$ and $k = 35$.

Table 2. Empirical performances of CIs for β_2 under $\mathbf{\Sigma} = 0.02 \cdot \mathrm{I}_p$, $\psi = 1, \alpha = 0.05$ and $n = 400$

p	Coverage (%)					Length				
	LSW	wlp	lproj	rproj	rose	LSW	wlp	lproj	rproj	rose
					$k = 20$					
400	95.2	94.6	96.8	75.1	80.7	2.79	2.77	2.81	1.60	1.99
700	95.2	97.1	97.7	91.1	82.0	2.61	2.77	2.79	2.24	1.99
1000	92.6	93.0	98.4	95.5	84.8	2.80	2.79	2.81	2.53	2.01
1300	95.3	95.3	97.1	94.6	79.1	2.70	2.78	2.80	2.65	2.00
					$k = 25$					
400	93.8	95.3	94.7	73.6	82.0	2.81	2.77	2.81	1.60	1.99
700	95.9	95.9	96.9	94.1	85.7	2.63	2.78	2.81	2.25	2.00
1000	95.9	95.6	97.8	92.8	83.2	2.79	2.77	2.80	2.52	2.00
1300	94.3	94.6	97.8	95.0	83.2	2.70	2.77	2.80	2.65	2.00
					$k = 35$					
400	95.5	94.2	94.7	73.6	82.9	2.81	2.77	2.81	1.60	1.99
700	94.4	95.5	96.8	87.5	80.6	2.63	2.78	2.80	2.24	1.99
1000	92.6	91.3	95.9	93.9	86.1	2.78	2.77	2.80	2.52	1.99
1300	95.9	96.3	96.8	94.7	79.1	2.70	2.77	2.80	2.65	2.00

Table 3. Empirical performances of CIs for β_2 under $\mathbf{\Sigma} = \mathbf{\Sigma}_M$, $\psi = 0.5, \alpha = 0.05$ and $n = 400$

p	Coverage (%)					Length				
	LSW	wlp	lproj	rproj	rose	LSW	wlp	lproj	rproj	rose
					$k = 20$					
400	92.9	98.4	57.1	92.6	97.9	1.13	1.56	0.53	0.98	1.04
700	94.6	97.2	45.4	90.3	94.6	1.12	1.32	0.53	0.86	1.08
1000	92 8	97.9	32.9	87.8	96.4	1.13	1.38	0.52	0.77	1.10
1300	94.6	97.2	27.9	81.7	95.8	0.96	1.27	0.53	0.73	1.12
					$k = 25$					
400	93.8	99.2	58.0	93.1	95.4	1.14	1.62	0.54	0.99	1.03
700	95.2	97.6	40.3	90.1	96.0	1.13	1.37	0.53	0.87	1.07
1000	92.8	97.0	31.3	84.0	96.2	1.16	1.43	0.53	0.77	1.09
1300	95.3	93.3	22.2	81.1	96.2	0.97	1.35	0.53	0.73	1.10
					$k = 35$					
400	94.1	97.6	51.8	91.5	94.7	1.15	1.52	0.54	1.00	1.01
700	96.5	99.3	37.3	89.4	92.5	1.14	1.57	0.53	0.87	1.05
1000	96.5	97.3	26.4	79.6	92.8	1.13	1.41	0.53	0.77	1.07
1300	94.7	93.8	22.7	76.7	92.8	0.98	1.30	0.53	0.73	1.09

shorter CIs), whose efficiency relative to its data-split counterparts is roughly $1/\sqrt{\xi}$, with $\xi \in \{0.25, 0.5, 0.75, 0.9\}$ being the proportion of samples used for the bias correction step. Therefore, for practical purpose, we recommend our proposed method without data splitting for better efficiency.

5.2. CIs for Other High-Dimensional Binary GLMs

To show the generality of our proposed LSW, we also evaluate its performance under some other link functions such as the probit link corresponding to the probit regression model, and the inverse cdf of the Student's t_1, or Cauchy distribution (denoted as "Inverse t_1"). In light of the previous results, we only compare the proposed method ("LSW") without sample splitting with the weighted low-dimensional projection method ("wlp") of Ma, Cai, and Li (2020) in various settings. In particular, due to the unavailability of the computational software for the initial Lasso estimators under these non-canonical link functions, we still use the logistic Lasso to obtain the initial estimators for these methods. Similar to the previous settings, we set $n = 400$ and let p vary from 800, 900, to 1000. We choose the sparsity level $k \in \{15, 20, 25\}$ and set the true regression coefficients in the same way as previous simulations with $\psi = 0.4$. The design covariates X_i's are generated in the same way as previous simulations with $\mathbf{\Sigma} = \mathbf{\Sigma}_M$ and again, we set the desired confidence level as $1 - \alpha = 95\%$. The numerical results are summarized in Table 4, with each entry representing an average over 500 rounds of simulations. Table 4 shows that, under both regression models, across all the settings, the proposed method has better coverage probabilities than "wlp," which suggests both preciseness and flexibility of the proposed method with respect to different link functions.

Table 4. Comparison of CIs with $\psi = 0.4, \alpha = 0.05$ and $n = 400$

p	Coverage (%)				Length			
	Probit		Inverse t_1		Probit		Inverse t_1	
	LSW	wlp	LSW	wlp	LSW	wlp	LSW	wlp
				$k = 15$				
800	93.9	93.1	92.0	78.3	0.85	0.72	1.13	0.56
900	94.2	91.6	93.0	76.5	1.26	0.72	1.19	0.56
1000	93.7	90.1	96.0	83.0	1.21	0.68	0.99	0.57
				$k = 20$				
800	93.8	83.1	92.3	66.3	0.60	0.51	1.14	0.53
900	96.1	81.3	93.0	76.5	0.60	0.51	0.25	0.56
1000	92.0	82.8	96.0	62.3	1.06	0.63	0.99	0.54
				$k = 25$				
800	92.7	90.4	94.3	81.0	0.80	0.71	1.15	0.55
900	97.5	87.7	96.7	79.3	1.31	0.68	1.20	0.56
1000	94.1	90.3	95.8	78.2	1.19	0.66	1.02	0.56

Table 5. Comparison of statistical tests with $\psi = 0.4, \alpha = 0.05$ and $n = 800$

p	Type I Errors (%)						Powers (%)					
	$k = 15$		$k = 20$		$k = 25$		$k = 15$		$k = 20$		$k = 25$	
	LSW	lrt	LSW	lrt	LSW	lrt	LSW	lrt	LSW	lrt	LSW	lrt
120	5.75	5.50	5.25	4.28	5.06	6.07	95.7	94.7	93.5	93.5	89.5	89.3
160	5.75	6.75	5.75	5.79	4.66	4.05	93.5	89.7	96.0	92.2	92.3	90.1
200	7.25	7.00	6.50	5.03	7.29	8.50	92.5	87.7	95.2	88.0	90.0	85.4
240	5.75	6.50	7.50	5.54	3.44	6.07	95.0	88.0	92.0	84.3	90.5	83.2

5.3. Hypothesis Testing for High-Dimensional Logistic Regression

It is well known that the construction of CIs and hypothesis testing are dual problems, so any CI considered in Section 5.1 can be converted to a statistical test, and a valid CI with shorter length translates to a valid statistical test with greater power. In this section, we focus on the numerical comparison between our proposed method and a rescaled likelihood ratio test ("lrt") recently proposed and carefully analyzed in Sur, Chen, and Candès (2019) and Sur and Candès (2019) under the modern maximum-likelihood framework with $p \leq n/2$. We compare the empirical performances of the two methods for testing a single regression coefficient, that is, whether a given coefficient is 0. Specifically, we set $n = 800$ and let p vary from 120, 160, 200 to 240. We choose the sparsity level $k \in \{15, 20, 25\}$, set $\alpha = 0.05$, and keep the true regression coefficients and the design covariates the same as those in Section 5.2. The numerical results are summarized in Table 5, with each entry representing an average over 500 rounds of simulations. From Table 5, we find that although both tests have their empirical Type I errors around the nominal level, the proposed method has higher power than "lrt" across all settings, especially when the ratio p/n is large. This again may be explained by the fact that the proposed method takes into account the underlying sparsity structure of the regression coefficients, whereas "lrt" does not.

6. Real Data Analysis

Finally, we analyze a single cell RNA-seq dataset from a recent study (Shalek et al. 2014), containing the expression estimates (transcripts per million) for all the 27,723 UCSC-annotated mouse genes, calculated using RSEM (Li and Dewey 2011), of a total number of 1861 primary mouse bone marrow derived dendritic cells spanning over several experimental conditions, including several isolated stimulations of individual cells in sealed microfluidic chambers. The complete dataset was downloaded from the Gene Expression Omnibus with the accession code GSE48968.

The analysis aims to understand the variability of gene expressions responding to certain stimuli. Specifically, we focus on the subsets of dendritic cells stimulated by one of the three pathogenic components, namely, LPS (a component of gram-negative bacteria), PIC (viralike double stranded RNA) and PAM (synthetic mimic of bacterial lipopeptides), and a set of unstimulated control cells. For the cells subject to one of the above stimulations, the gene expression profiles are obtained along time course (0, 1, 2, 4, and 6 hr) after stimulation. To better meet our purpose, we only consider the expression profiles 6 hours after the stimulation, due to their more significant deviation from the unstimulated cells (see Shalek et al. 2014, fig. 2). Combining the expression profiles of all 96 control cells and one of the three groups of the stimulated cells, namely, 64 PAM stimulated cells, 96 PIC stimulated cells, and 96 LPS stimulated cells, we study the association between the gene expressions and the stimulation status, coded as 0 and 1, corresponding to "unstimulated" and "stimulated," respectively. To reduce the number of genes in the subsequent analysis, for each combined expression matrix, we filter out the genes not expressed in more than 80% of the cells, and only keep the genes whose variance is within the top 10 percentile. The expression levels are log-transformed and normalized to have mean zero and unit variance across the cells. Consequently, for each combination of unstimulated and stimulated cells, we fit a high-dimensional logistic regression and apply the proposed method to obtain 95% CIs for each of the regression coefficients. The constructed CIs under each of the fitted models are illustrated in Figure 3, with an averaged length 1.5.

As a consequence, for each of the stimulations, there is one or more genes whose regression coefficients have CIs that do not cover zero (marked in red in Figure 3), suggesting significant evidence of associations with the stimulation event and therefore potential functional consequences responding to that stimulus. Specifically, for the PAM stimulated cells, our analysis identified the protein coding gene IL6, which encodes a cytokine, interleukin 6, promptly and transiently produced in response to infections and tissue injuries, that contributes to host defense through the stimulation of acute phase responses, hematopoiesis, and immune reactions (Tanaka, Narazaki, and Kishimoto 2014). For the PIC stimulated cells, we identified RSAD2, whose regression coefficient has a CI of (1.26, 2.80). This result has a very interesting connection to the previous experimental findings that RSAD2 is involved in antiviral innate immune response and is a powerful stimulator of adaptive immune response mediated via mature dendritic cells (Jang et al. 2018). Moreover, for the LPS stimulated cells, our analysis highlights the protein coding genes CXCL10 and IL12B. Among them, CXCL10 is known to have antitumor, antiviral, and antifungal activities and is essential for the generation of protective CD8+ T cell responses (Enderlin et al. 2009; Majumder et al. 2012), whereas IL12B encodes a cytokine that acting as a growth factor for activated T and NK cells, which plays a major role in cell-mediated immunity against a variety of pathogens and therefore being host-protective in the context of intracellular bacterial infections (Ymer et al. 2002; Zwiers et al. 2011). These results are interesting and suggest the usefulness of our proposed method in real applications. As a comparison, by applying some other methods considered in Section 5, we observe that, (i) "wlp" tends to produce shorter CIs with averaged length 1.2, among which about 37 CIs in each stimulation scenario do not cover zero, including those of the genes listed above, (ii) "lproj" highlights similar sets of genes whose CI does not cover zero, as the proposed method, yet their CIs are much longer, with averaged length 3.5, and (iii) "rproj" also produces long CIs with averaged length 3.5, and all of them cover zero. However, more experimental and numerical evidences are needed as to determine which method produced

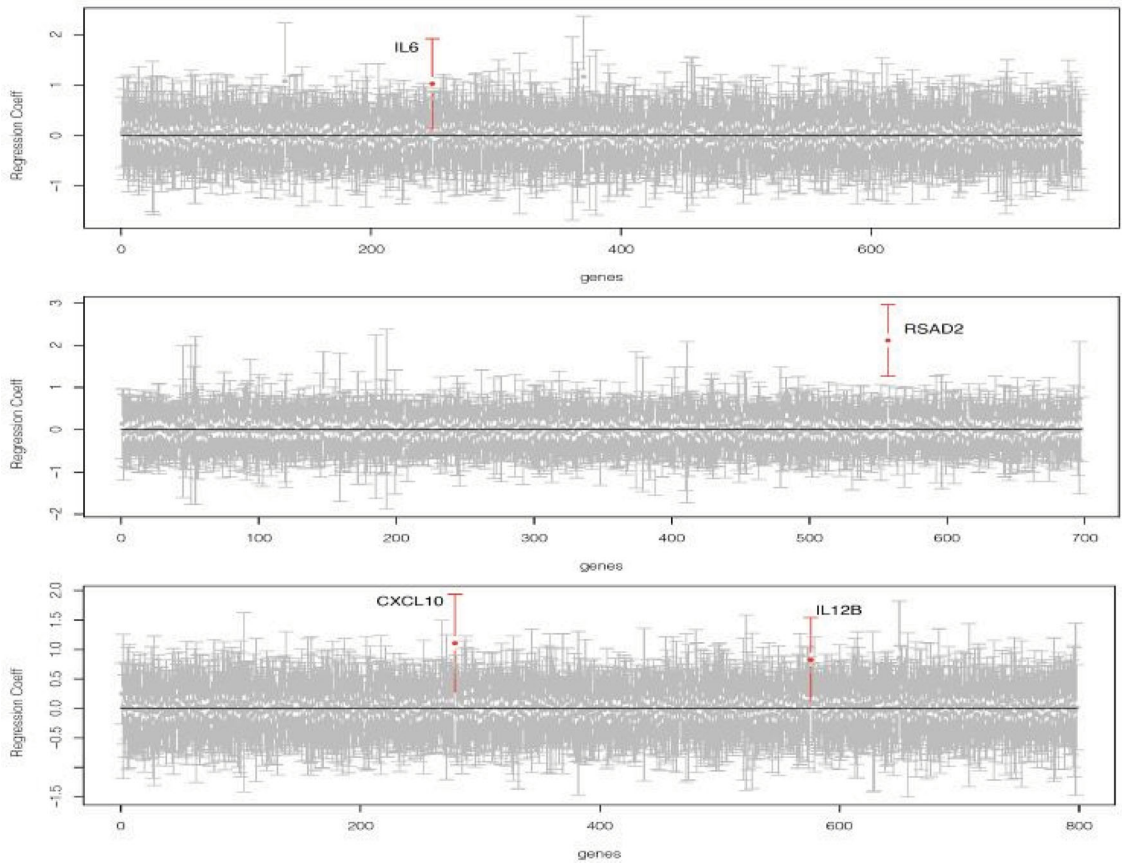

Figure 3. An illustration of the CIs for the high-dimensional logistic regressions corresponding to the stimulations by LPS (top), PIC (middle) and LPS (bottom), respectively. The CIs that do not cover zero are marked in red, with their gene names labeled.

the most precise CIs. See Section 5 in the supplement for more details.

7. Discussion

We presented a unified framework for constructing CIs and statistical tests for the regression coefficients in high-dimensional binary GLMs with a range of link functions. Both minimax optimality and adaptivity are studied. For technical reasons, sample splitting was used to establish the theoretical properties. As demonstrated in our proof of Theorem 1 in the Supplement, we essentially need to prove the asymptotic normality of the stochastic error in (6), that is, conditional on the covariates $\{X_i\}_{i=1}^n$, it holds that $v_j^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^n w(X_i^\top\hat{\beta})\hat{u}^\top X_i\epsilon_i \to_d N(0,1)$. It is possible to establish this asymptotic normality without sample splitting, by imposing similar conditions as those in van de Geer et al. (2014) and by obtaining an estimate $\hat{u}$ through the nosewise regression. However, as discussed in Section 3.1, these stringent conditions limit the applicability of the proposed method, and can be avoided by splitting the samples to create independence between $(\hat{u},\hat{\beta})$ and $\{\epsilon_i\}_{i=1}^n$. Hence, after weighing the pros and cons, we decide to present our main theoretical results by keeping the sample splitting while removing other strong assumptions. Nevertheless, given the fact that the proposed methods perform well numerically without sample splitting, it is of interest to develop novel technical tools to yield theoretical guarantees for the inference procedures without splitting the samples or imposing additional stringent conditions.

Recently, Ning and Cheng (2020) proposed to construct sparse confidence sets for sparse normal mean vectors. Unlike our proposal which focuses on the individual regression coefficient, they aim to construct sparse confidence sets for a vector with certain false positive rate control. It is interesting to construct sparse confidence sets for β under the high-dimensional GLMs.

Acknowledgements

We would like to thank the editor, associate editor, and two anonymous referees for helpful suggestions that significantly improved the presentation of the results. This work was completed while Rong Ma was a PhD student in the biostatistics program at the University of Pennsylvania.

Funding

Tony Cai's research was supported in part by NSF grant DMS-2015259 and NIH grants R01-GM129781 and R01-GM123056. Zijian Guo's research was supported in part by NSF grants DMS-1811857 and DMS-2015373 and NIH grants R01-GM140463 and R01-LM013614.

Supplementary Materials

In the supplement, we prove all the main theorems and the technical lemmas. Some additional discussions about assumptions and numerical studies are also included.

References

Bach, F. (2010), "Self-Concordant Analysis for Logistic Regression," *Electronic Journal of Statistics*, 4, 384–414. [2]

Belloni, A., Chernozhukov, V., and Wei, Y. (2016), "Post-Selection Inference for Generalized Linear Models With Many Controls," *Journal of Business & Economic Statistics*, 34, 606–619. [2,3,7]

Bühlmann, P. (2013), "Statistical Significance in High-Dimensional Linear Models," *Bernoulli*, 19, 1212–1242. [10]

Cai, T. T., and Guo, Z. (2017), "Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity," *The Annals of Statistics*, 45, 615–646. [3,8]

——— (2018), "Accuracy Assessment for High-Dimensional Linear Regression," *The Annals of Statistics*, 46, 1807–1836. [3]

——— (2020), "Semi-Supervised Inference for Explained Variance in High-Dimensional Regression and Its Applications," *Journal of the Royal Statistical Society*, Series B, 82, 391–419. [9]

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018), "Panning for Gold:'Model-x'Knockoffs for High Dimensional Controlled Variable Selection," *Journal of the Royal Statistical Society*, Series B, 80, 551–577. [6]

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017), "Central Limit Theorems and Bootstrap in High Dimensions," *Annals of Probability*, 45, 2309–2352. [6]

Dezeure, R., Bühlmann, P., and Zhang, C.-H. (2017), "High-Dimensional Simultaneous Inference With the Bootstrap," *Test*, 26, 685–719. [6]

Enderlin, M., Kleinmann, E., Struyf, S., Buracchi, C., Vecchi, A., Kinscherf, R., Kiessling, F., Paschek, S., Sozzani, S., and Rommelaere, J. (2009), "Tnf-α and the ifn-γ-Inducible Protein 10 (ip-10/cxcl-10) Delivered by Parvoviral Vectors Act in Synergy to Induce Antitumor Effects in Mouse Glioblastoma," *Cancer Gene Therapy*, 16, 149–160. [12]

Guo, Z., Rakshit, P., Herman, D. S., and Chen, J. (2020), "Inference for the Case Probability in High-Dimensional Logistic Regression," arXiv:2012.07133. [2,3,7]

Guo, Z., and Zhang, C.-H. (2019), "Extreme Nonlinear Correlation for Multiple Random Variables and Stochastic Processes With Applications to Additive Models," arXiv:1904.12897. [8]

Huang, J., and Zhang, C.-H. (2012), "Estimation and Selection Via Absolute Penalized Convex Minimization and Its Multistage Adaptive Applications," *Journal of Machine Learning Research*, 13, 1839–1864. [2,3,6,7]

Jang, J.-S., Lee, J.-H., Jung, N.-C., Choi, S.-Y., Park, S.-Y., Yoo, J.-Y., Song, J.-Y., Seo, H. G., Lee, H. S., and Lim, D.-S. (2018), "Rsad2 is Necessary for Mouse Dendritic Cell Maturation Via the irf7-Mediated Signaling Pathway," *Cell Death & Disease*, 9, 1–11. [12]

Janková, J., and van de Geer, S. (2018), "De-biased Sparse PCA: Inference and Testing for Eigenstructure of Large Covariance Matrices," arXiv:1801.10567. [7]

Jankova, J., and van de Geer, S. (2018), "Semiparametric Efficiency Bounds for High-Dimensional Models," *The Annals of Statistics*, 46, 2336–2359. [8]

Javanmard, A., and Javadi, H. (2019), "False Discovery Rate Control Via Debiased Lasso," *Electronic Journal of Statistics*, 13, 1212–1253. [6]

Javanmard, A. and Montanari, A. (2014a), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [2,3,4,5]

——— (2014b), "Hypothesis Testing in High-Dimensional Regression Under the Gaussian Random Design Model: Asymptotic Theory," *IEEE Transactions on Information Theory*, 60, 6522–6554. [3]

Lancaster, H. O. (1957), "Some Properties of the Bivariate Normal Distribution Considered in the Form of a Contingency Table," *Biometrika*, 44, 289–292. [8]

Li, B., and Dewey, C. N. (2011), "RSEM: Accurate Transcript Quantification From RNA-Seq Data With or Without a Reference Genome," *BMC Bioinformatics*, 12, 323. [12]

Liu, W. (2013), "Gaussian Graphical Model Estimation With False Discovery Rate Control," *The Annals of Statistics*, 41, 2948–2978. [6]

Ma, R., Cai, T. T., and Li, H. (2020), "Global and Simultaneous Hypothesis Testing for High-Dimensional Logistic Regression Models," *Journal of the American Statistical Association*, 1–15. [2,3,5,6,7,10,11]

Majumder, S., Bhattacharjee, S., Chowdhury, B. P., and Majumdar, S. (2012), "Cxcl10 is Critical for the Generation of Protective cd8 t Cell Response Induced by Antigen Pulsed cpg-odn Activated Dendritic Cells," *PLoS One*, 7. [12]

Meier, L., van de Geer, S., and Bühlmann, P. (2008), "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society*, 70, 53–71. [2]

Mukherjee, R., Pillai, N. S., and Lin, X. (2015), "Hypothesis Testing for High-Dimensional Sparse Binary Regression," *The Annals of Statistics*, 43, 352–381. [3]

Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. (2010), "A Unified Framework for High-Dimensional Analysis of M-Estimators With Decomposable Regularizers," Technical Report Number 979. [2,3,7]

Nickl, R., and van de Geer, S. (2013), "Confidence Sets in Sparse Regression," *The Annals of Statistics*, 41, 2852–2876. [3]

Ning, Y., and Cheng, G. (2020), "Sparse Confidence Sets for Normal Mean Models," arXiv:2008.07107. [13]

Ning, Y., and Liu, H. (2017), "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models," *The Annals of Statistics*, 45, 158–195. [2,3,7]

Nualart, D. (2006), *The Malliavin Calculus and Related Topics*, Springer Science & Business Media. Berlin: Springer. [8]

Plan, Y., and Vershynin, R. (2013), "Robust 1-bit Compressed Sensing and Sparse Logistic Regression: A Convex Programming Approach," *IEEE Transactions on Information Theory*, 59, 482–494. [2]

Rakshit, P., Cai, T. T., and Guo, Z. (2021), "Sihr: An r Package for Statistical Inference in High-Dimensional Linear and Logistic Regression Models," arXiv:2109.03365. [2]

Razzaghi, M. (2013), "The Probit Link Function in Generalized Linear Models for Data Mining Applications," *Journal of Modern Applied Statistical Methods*, 12, 19. [1]

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N. (2014), "Single-Cell RNA-seq Reveals Dynamic Paracrine Control of Cellular Variation," *Nature*, 510, 363–369. [12]

Shi, C., Song, R., Lu, W., and Li, R. (2020), "Statistical Inference for High-Dimensional Models Via Recursive Online-Score Estimation," *Journal of the American Statistical Association*, 1–12. [2,3,7,10]

Sur, P., and Candès, E. J. (2019), "A Modern Maximum-Likelihood Theory for High-Dimensional Logistic Regression," *Proceedings of the National Academy of Sciences*, 116, 14516–14525. [3,12]

Sur, P., Chen, Y., and Candès, E. J. (2019), "The Likelihood Ratio Test in High-Dimensional Logistic Regression is Asymptotically a Rescaled Chi-Square," *Probability Theory and Related Fields*, 175, 487–558. [2,3,12]

Tanaka, T., Narazaki, M., and Kishimoto, T. (2014), "Il-6 in Inflammation, Immunity, and Disease," *Cold Spring Harbor Perspectives in Biology*, 6, a016295. [12]

Tsiatis, A. (2007), *Semiparametric Theory and Missing Data*, Springer Science & Business Media. New York: Springer-Verlag. [5]

van de Geer, S. (2008), "High-Dimensional Generalized Linear Models and the Lasso," *The Annals of Statistics*, 36, 614–645. [2,7]

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [2,3,5,7,8,10,13]

Xia, L., Nan, B., and Li, Y. (2020), "A Revisit to De-Biased Lasso for Generalized Linear Models," arXiv:2006.12778. [7]

Ymer, S., Huang, D., Penna, G., Gregori, S., Branson, K., Adorini, L., and Morahan, G. (2002), "Polymorphisms in the il12b Gene Affect Structure and Expression of il-12 in Nod and Other Autoimmune-Prone Mouse Strains," *Genes & Immunity*, 3, 151–157. [12]

Yu, Y. (2008), "On the Maximal Correlation Coefficient," *Statistics & Probability Letters*, 78, 1072–1075. [8]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [2,3,4,5]

Zhang, X., and Cheng, G. (2017), "Simultaneous Inference for High-Dimensional Linear Models," *Journal of the American Statistical Association*, 112, 757–768. [6]

Zhu, Y., Shen, X., and Pan, W. (2020), "On High-Dimensional Constrained Maximum Likelihood Inference," *Journal of the American Statistical Association*, 115, 217–230. [3]

Zwiers, A., Fuss, I. J., Seegers, D., Konijn, T., Garcia-Vallejo, J. J., Samsom, J. N., Strober, W., Kraal, G., and Bouma, G. (2011), "A Polymorphism in the Coding Region of il12b Promotes il-12p70 and il-23 Heterodimer Formation," *The Journal of Immunology*, 186, 3572–3580. [12]