Electronic Journal of Statistics
Vol. 18 (2024) 395–428
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2204

# A tradeoff between false discovery and true positive proportions for sparse high-dimensional logistic regression

**Jing Zhou**[1] **and Gerda Claeskens**[2]

[1]*School of Mathematics, UEA, Norwich Research Park, Norwich, NR4 7TJ, UK*
*e-mail:* J.Zhou6@uea.ac.uk

[2]*ORStat and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium*
*e-mail:* gerda.claeskens@kuleuven.be

**Abstract:** The logistic regression model is a simple and classic approach to binary classification, where in sparse high-dimensional settings, one believes that only a small proportion of the predictive variables are relevant to the response variable with nonnull regression coefficients. We focus on regularized logistic regression models and the analysis is valid for a large group of regularizers, including folded-concave regularizers such as MCP and SCAD. For finite samples, the discrepancy between the estimated and true nonnull coefficients is evaluated by the false discovery and true positive rates. We show that the false discovery rate can be described using a nonlinear tradeoff function of power asymptotically using a system of equations with six parameters. The analysis is conducted in an "average-over-components" fashion for the unknown parameter and follows the conventional assumptions of the literature in the relevant field. More specifically, we assume a linear growth rate $n/p \to \delta > 0$ covering not only the typical high dimensional settings where $p \geq n$ but also for $n > p$. Further, we propose two applications of this tradeoff function that improve the reproducibility of variable selection: (1) a sample size calculation procedure to achieve a certain power under a prespecified level of false discovery rate using the tradeoff; (2) calibration of the false discovery rate for variable selection taking power into consideration. A similar asymptotic analysis for the model-X knockoff, which provides a selection with a controlled false discovery rate, is investigated to show how to compare two selection methods by comparing the tradeoff curves. We illustrate the tradeoff analysis and its corresponding applications using simulated and real data.

**MSC2020 subject classifications:** Primary 62J12; secondary 62F99.
**Keywords and phrases:** High-dimensional data, FDR control, knockoff, logistic regression, sparsity, false discovery rate.

Received October 2022.

## 1. Introduction

The classic maximum likelihood estimators (MLE) possess properties such as asymptotic normality and consistency under regularity conditions in the fixed $p$, $n \to \infty$ setting. When $p$ is considerably large and grows simultaneously with $n$ following a linear order, the MLE for a high dimensional logistic regression model

was studied in [44, 42] when the parameter vector $\beta \in \mathbb{R}^p$ is unstructured (without assuming it sparse, block-sparse, etc.). One of their findings is that the MLE of the logistic regression coefficients is biased due to high dimensionality, and the bias can be obtained by solving a system of nonlinear equations with three unknown parameters relying on the approximate message passing (AMP) analysis.

However, the ratio $n/p$ is often less than 1 and nonnegligible for high dimensional data, for which the regularized estimators are popular options. [40] developed a system of equations with six parameters to characterize the limiting behavior of the regularized logistic regression estimators. This work covers both $n > p$ and $n \leq p$ settings with structured parameter vectors, e.g., sparse, block-sparse vectors. The system of equations with six parameters resembles the one obtained by the AMP analysis in [44, 42], but was derived through an alternative framework called *convex Gaussian min-max theorem* (CGMT) [27, 28, 45] in the compressed sensing field. An attractive feature of the CGMT framework is that the proof technique is less stringent on the i.i.d. Gaussian assumption on the components of the design matrices compared to the AMP analysis. Literature on the AMP analysis has attempted to relax such assumption mainly from two aspects: (1) by generalizing the limiting analysis from i.i.d. Gaussian random matrices to other random matrices, for example, [22, 24, 39, 38]. Although the Gaussian distribution assumption can be relaxed, the i.i.d. assumption is more challenging to handle; (2) incorporate a general Gaussian matrix $\boldsymbol{X} \sim N(0, \boldsymbol{\Sigma})$ with an arbitrary covariance matrix [14, 44, 56]. This is often performed by assuming the covariance matrix $\boldsymbol{\Sigma}$ is nonsingular, the inverse of the covariance matrix exists, and the parameter vector $\beta$ is unstructured. The analysis is then performed through $(\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2})(\boldsymbol{\Sigma}^{1/2}\beta)$, where $\boldsymbol{\Sigma}^{1/2}$ serves as a "preconditioner" to reduce the correlations between predictive variables. However, preconditioning is far more complex for sparse high-dimensional regression, which is the main topic of this paper. The issues were systematically discussed in [31] using dependency graphs of the distributions of $\boldsymbol{X}$. [31] first explained that it is possible in theory to "precondition" $\boldsymbol{X}$ when the dependency graph of the distribution of $\boldsymbol{X}$ has low treewidth. The treewidth describes the size of the largest subset of connected $\boldsymbol{\Sigma}_{ij}$ that is the smallest among all possible tree decomposition; intuitively, it refers to the size of group correlated variables. The preconditioner is constructed through a *sparse* Cholesky decomposition of $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Sigma}^{1/2})^\top$ to guarantee that *both* $\boldsymbol{\Sigma}^{1/2}$ *and* $\boldsymbol{\Sigma}^{-1/2}$ *are sparse preserving*. That is, the sparsity of the covariance matrix of the preconditioned $\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ and the parameter $\boldsymbol{\Sigma}^{1/2}\beta$ are preserved. In practice, when $\boldsymbol{\Sigma}$ is unknown, the preconditioner can be approximated by algorithms such as block Cholesky factorizations while carefully controlling errors from the factorization. "Precondition" for $\boldsymbol{X}$ with high treewidth will fail for sure. One obvious reason is that the sparse structure of $\beta$ could hardly be kept after the transformation $\boldsymbol{\Sigma}^{1/2}\beta$, which intuitively would cause regularization to fail. The limiting analysis in the CGMT framework can be easily extended to random matrices with elliptical distributions or isotropically random orthogonal matrices, which are row orthogonal satisfying $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{I}_n$, see [46, 45]. The generalization to other random design matrices

in the CGMT framework is natural and suitable for dealing with sparse high dimensional regression when $p > n$. Abundant literature has been devoted to using the AMP framework to analyze high-dimensional linear regression models. One major research line is the asymptotic distribution-related work, including bias correction and asymptotic mean squared error for linear regression [see for example, 15, 4, 8, 57, 13, 12]. The AMP-based analysis for linear regression was extended to evaluate the falsely selected null components of the Lasso estimator in high dimensional linear regression models when $p > n$, see [41, 48, 49, 50].

Another research line investigating the property of variable selection mostly has requirements on the sparsity, i.e., the number of nonnull coefficients satisfies certain orders of $p, n$. For instance, this is the case for the selection consistency of regularized estimators, which states that the nonnull coefficients are correctly identified in probability [see for example 16, 19, 58, 55, 32]. Although asymptotic selection consistency is mathematically elegant, it is hard to achieve for finite samples, which motivates the study of the discrepancy between selected and the true nonnulls, where the true positive rate (FDR) and false discovery rate (FDR) play an important role.

In ultra-high dimensional settings, when the number of parameters $p$ is much larger than $n$, selection by regularization-based methods becomes unstable. Then, auxiliary model-based variable screening methods are proposed, starting from [17] for linear models using Pearson correlation to the subsequent serial work such as [20, 18, 29, 51, 53]. Due to the concern for model-misspecification, model-free screening methods attracted attention leading to methods such as [33] using distance correlation, [35] based on Kolmogorov-distance for binary classification, [37] using ball correlation.

Driven by the two concerns mentioned above, model-free variable selection with simultaneous FDP control gradually became a new research interest. A popular FDP control approach builds on conditional independence, for instance, [47, 34, 10, 11]. Especially, the model-X knockoff [11] will be compared in this paper with variable selection by a regularized estimator. A variable is selected when its feature statistic, which compares the importance of the original variable and the contrast knockoff variables, exceeds a certain threshold. Furthermore, the choice of the threshold is determined by controlling the estimated FDP. The flexible choice of the feature statistic is a big advantage of the model-X knockoff method. [2, 11] proposed several example options of the feature statistic such as the *Lasso Signed Max* statistic and *Lasso Coefficient Difference* statistic; but the potential choice is not only limited to these. This paper considers logistic regression, where the predictive variables are correlated with the response variable through the regression coefficient vector. We consider taking as the feature statistic the *Lasso Coefficient Difference* statistic due to the connection with the regularized logistic regression. Further, the model-X knockoff has been extended to FDP control in hypothesis testing for the generalized linear model since one can easily turn the variable selection into testing if the components of the regression coefficient vector are nulls. Some selected references on FDP control in the domain of hypothesis testing are [25, 5, 6, 23, 26], we refer to [25] for a more thorough literature review since hypothesis testing is less the focus of this paper.

In this paper, see Section 2, we follow the logic of [41, 48, 49, 50] by investigating the true positive proportion (TPP) and false discovery proportion (FDP) in the limit. In addition, we consider regularized logistic regression estimators, which initiate future work on generalized linear models. The analysis is valid for a wide class of regularization functions, and we showcase the $\ell_1$-regularized logistic regression for simplicity. We incorporate the modern data structure in which the number of variables $p$ is nontrivial compared to the sample size $n$ by assuming a ratio $n/p \to \delta \in (0, \infty)$. Especially our work includes the case where $n > p$ but $p$ is nontrivial, where variable selection by regularized estimators is often used in practice, but the selection performance is not thoroughly discussed in the relevant literature. Based on the system of six equations in [40] using the $\ell_1$-regularizer, we obtain the limiting expressions of the TPP and FDP.

A tradeoff curve for $\ell_1$-regularized logistic regression can be constructed using the limiting expressions, clearly illustrating the nonlinear association between TPP and FDP. Further, similar to classical hypothesis testing, the tradeoff curve can be used to compare the selection power between multiple selection methods. For comparing two selection methods, we consider an FDR-controlled selection called model-X knockoff [11], see Section 3 since its construction is based on $\ell_1$-regularized logistic regression, which allows obtaining the limiting expressions of FDP and TPP similarly, see also [49, 50] for a linear model. The tradeoff curve is first utilized to investigate the impact of the averaged signal strength and the sparsity on selection power for a given FDR level in Section 2.3. We propose a sample size calculation based on the impact of the ratio $n/p$ and FDR level calibration. This is applied in Section 4 using the `Wisconsin Breast Cancer` dataset for potential practical use of the tradeoff curves. The Appendix contains the technical results.

## 2. High dimensional logistic regression

For a $p$-dimensional vector $X = (X_{\cdot 1}, \ldots, X_{\cdot p})$ consisting of $p$ predictive variables $X_{\cdot j}$, $j = 1, \ldots, p$, a binary response variable $Y \in \{0, 1\}$ is modelled by a logistic regression model using $X$ as follows

$$P(Y = 1 \mid X) = \rho'\big(X\beta^*\big), \tag{1}$$

where the function $\rho'(t) = \frac{1}{1+e^{-t}}$ is the first derivative of the logistic link function

$$\rho(t) = \log\big(1 + e^t\big). \tag{2}$$

The $X_{\cdot j}$'s, with $j = 1, \ldots, p$, are combined using the coefficient vector $\beta^* \in \mathbb{R}^p$, where the absolute value $|\beta_j^*|$ indicates the importance of $X_{\cdot j}$ to $Y$. Due to advances in data collection, we often obtain large datasets with excessive predictive variables, among which most are irrelevant to the response variable of interest. Thus, we assume a sparse vector, i.e., $\beta_j^* = 0$ with probability $1 - s$ to describe this scenario, see Assumption (A2) for details. The sample pairs $(Y_i, X_{i\cdot})$, $i = 1, \ldots, n$ are i.i.d. copies of $(Y, X)$ satisfying Assumption (A1).

A strong point of this paper is that it investigates a linear growth framework where $n/p \to \delta \in (0, \infty)$ when $n, p \to \infty$ which covers both $n > p$ and $n \leq p$ cases. We state the following assumptions. However, see our introduction and [31] for a possible extension of (A1) to correlated designs.

(A1) Standard Gaussian design: the vectors $X_{i\cdot} \sim N(0, \frac{1}{p}\boldsymbol{I}_p), i = 1, \ldots, n$ are independent and identically distributed (i.i.d.).

(A2) The components of the $p$-vector $\beta^*$ are i.i.d. samples of a random variable $B$ with signal strength $E(B^2) = \kappa^2$ and p.d.f

$$f_B(\beta) = (1 - s) \cdot \delta_0(\beta) + s \cdot f_{B'}(\beta),$$

where $f_{B'}$ denotes the p.d.f. of the nonnull components $B'$ of $B$.

Further, we clarify the sparsity parameter $s$. The convention of such "average-over-components" analysis does not address sparsity assumptions and often assumes $\|\beta\|_0/p \to s$ when $p \to \infty$. Unlike [58, 21], which discuss the performance of the vector $\hat{\beta}$ and impose assumptions on $s$ in order to bound $\hat{\beta} - \beta$ when deriving theoretical results, the asymptotic results in the CGMT framework are derived for the components of $\beta$ on average and $s$ is not involved in the derivation. In fact, the result in [1] is valid for general $\beta$ where a sparse structure stated in (A2) is a special case. This sparsity parameter simply indicates that we let a small proportion of $\beta$ be nonzero. We specifically choose the mixture distribution in Assumption (A2) such that the impact of the sparsity level $s$ can be visualized. Also, Figure 2 shows that our theoretical analysis of the mean-squared error still holds when $s = 0.5$. This obviously violates the sparsity assumptions in [58, 21] but does not affect our analysis. The limiting performance analysis based on the system of equations still holds when $\beta$ is unstructured/dense when $s = 1$ and the tuning parameter $\lambda$ of the regularization term is set to zero. This has been discussed in detail in [1, Section 4.1], and the results agree with [42] in the AMP framework.

To estimate $\beta^*$, we consider regularized logistic regression estimators obtained by solving a minimization problem as follows

$$\widehat{\beta}(\lambda) = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \{\rho(X_{i\cdot}\beta) - Y_i(X_{i\cdot}\beta)\} + \frac{1}{p} \sum_{j=1}^{p} R(\beta_j; \lambda), \tag{3}$$

where $\lambda \in \mathbb{R}_+$ is the tuning parameter for the regularizer $R(\cdot; \lambda)$.

Since we allow for a potential sparse structure of $\beta^*$, the following regularizers are great examples for performing selection:

1. $\ell_1$-regularizer: $R_{\ell_1}(\beta_j; \lambda) = \lambda|\beta_j|$.
2. Bridge: $R_{\mathrm{B}}(\beta_j; \lambda, a) = \lambda|\beta_j|^a$, for $a > 0$.
3. Smoothly clipped absolute deviation (SCAD [16]):

$$R_{\mathrm{SCAD}}(\beta_j; \lambda, a) = \begin{cases} \lambda|\beta_j| & |\beta_j| \leq \lambda \\ \frac{2a\lambda|\beta_j| - \beta_j^2 - \lambda^2}{2(a-1)} & \lambda < |\beta_j| < a\lambda \\ \lambda^2(a+1)/2 & |\beta_j| \geq a\lambda, \end{cases}$$

for $a > 0$. The value $a = 3.7$ is suggested in [16].

4. Minimax concave penalty (MCP [54]):

$$R_{\text{MCP}}(\beta_j; \lambda, a) = \begin{cases} \lambda|\beta_j| - \beta_j^2/2a & |\beta_j| \le a\lambda \\ a\lambda^2/2 & |\beta_j| > a\lambda, \end{cases}$$

for $a > 1$ with recommended value $a = 2$ [54].

The SCAD and MCP regularizers are known to be "folded-concave"; that is, the regularization functions are symmetric around zero and concave on $\mathbb{R}_+$ and $\mathbb{R}_-$. They possess desirable properties such as "unbiasedness", "sparsity", and "continuity" [16]. The bridge regression, when $a \in (0, 1)$, achieves "unbiasedness" and "sparsity" but is less favored due to the discontinuity of the estimator as a function of the data. Our investigation focuses on the variable selection properties of the regularizers. Thus, we do not discuss $a > 1$ for the bridge regularizer since "sparsity" is not well achieved, implying that parameters for irrelevant predictors might not be set to zero properly.

### *2.1. A system of nonlinear equations*

We discuss a linear growth framework where $n/p \to \delta \in (0, \infty)$ when $n, p \to \infty$; meanwhile, Assumption (A2) gives the limiting representation of the components of $\beta^*$ in an "average-over-components" sense. Since $n, p$ grow simultaneously following a linear growth rate, we denote $p, n \to \infty$ as $p_n \to \infty$ or simply $p \to \infty$ throughout this paper. The limiting behavior of the estimator $\widehat{\beta}$ in (3) averaged over components is determined by the ratio $\delta$, the signal strength $\kappa \in \mathbb{R}_+$, and the tuning parameter $\lambda$. Given $\kappa, \delta$, for any $\lambda$, [40] characterizes the limiting performance of $\widehat{\beta}$ by a system of equations consisting of six parameters $(\alpha, \sigma, \gamma, \theta, \tau, r)$. The equations incorporate the logistic link function and the $\ell_1$-regularizer separately by their corresponding proximal operators. For any deterministic function $\tilde{f}$, its proximal operator is defined as

$$\text{Prox}_{t\tilde{f}(x)}(v) = \arg\min_x \left\{ \frac{1}{2t}\|x - v\|_2^2 + \tilde{f}(x) \right\}, \tag{4}$$

where $t$ is the parameter. The proximal operator guarantees unique minimizers of the $l_2$-regularized $\tilde{f}$ function, where $\tilde{f}$ can be non-differentiable. Since [40] imposes convexity on the function $\tilde{f}$ in (4), we consider investigating the adaptive Lasso [58] which incorporates a weighting scheme for the regularizer by taking $R(\beta_j; \lambda) = w_j(\lambda)|\beta_j|$. The rationale is as follows: the $\ell_1$-norm guarantees that the regularizer is convex, and it is obvious that the $\ell_1$-regularizer is a special case of the adaptive Lasso by taking equal weights $w_j(\lambda) = \lambda$. Further, the adaptive Lasso can be used to approximate nonconvex regularizers, including the bridge, SCAD, and MCP, by the local linear approximation (LLA) algorithm, see [21]. This linear approximation is achieved by taking the weights in the adaptive Lasso to be the first derivative of the regularization functions. For

sparse logistic regression, it was shown by [21] that the LLA algorithm converges after two iterations to an oracle estimator with a high probability.

Specifically, the proximal operator for $\tilde{f}(x) = w|x|$ with parameter $t$ follows

$$\text{Prox}_{tw|x|} = \eta_{tw}(x) = \text{sgn}(x)(|x| - wt)_+. \tag{5}$$

The tuning parameter $w$ varies for each component of $\beta$. When approximating the nonconvex regularizers, the weights take $w_j = R'(|\beta_j|; \lambda)$ where $R'$ is the first derivative of the regularizer $R(|\beta_j|; \lambda)$. In practice, we plug in an initial estimator $\beta$, such as the $\ell_1$-regularized estimator, to obtain the estimated weights.

Then, the weight vectors used to approximate the $\ell_1$, SCAD, and MCP regularizers are as follows

1. ($\ell_1$-regularizer) $w_{\ell_1}(x) = \lambda$.
2. (SCAD) $w_{\text{SCAD}}(x) = \lambda I\{x \le \lambda\} + \frac{(a\lambda - x)_+}{a-1} I\{x > \lambda\}$.
3. (MCP) $w_{\text{MCP}}(x) = (\lambda - \frac{x}{a})_+$.

By (A2), the components of the weight vector, when assigned $1/p$ weight, converge weakly to $w(B)$ by continuous mapping theorem. While plugging in an initial estimator in practice, the asymptotic analysis of the weight vector is much more complicated and requires a second-stage analysis based on the system of equations (6). However, the estimated weights $\hat{w}_j = w(\hat{\beta}_j)$ can still be expressed as a function of $B$ in the limit as a composite function of $w(\cdot)$ and the proximal operator of the regularization function of the initial estimator.

The system of equations is obtained by rewriting (3) as two min-max optimization problems using the Lagrange multiplier method resulting in parameters $\theta$, $r$ among $(\alpha, \sigma, \gamma, \theta, \tau, r)$. The Lagrange multiplier incorporates equality constraints in the optimization equations; see [40, Eq. (54) and above Eq. (45)] for technical details.

Then the system of equations with six parameters $(\alpha, \sigma, \gamma, \theta, \tau, r)$ for the regularized logistic regression [40], given $\delta, \kappa, \lambda$, is as follows

$$\begin{cases} \kappa^2\alpha = E\left[B \cdot \eta_{w(B)\sigma\tau}\left(\sigma\tau \cdot \left(\theta B + \frac{r}{\sqrt{\delta}} Z\right)\right)\right] \\ \gamma = \frac{1}{r\sqrt{\delta}} E\left[Z \cdot \eta_{w(B)\sigma\tau}\left(\sigma\tau \cdot \left(\theta B + \frac{r}{\sqrt{\delta}} Z\right)\right)\right] \\ \kappa^2\alpha^2 + \sigma^2 = E\left[\eta_{w(B)\sigma\tau}\left(\sigma\tau \cdot \left(\theta B + \frac{r}{\sqrt{\delta}} Z\right)\right)^2\right] \\ \gamma^2 = \frac{2}{r^2} E\left[\rho'(-\kappa Z_1) \cdot \left(\kappa\alpha Z_1 + \sigma Z_2 - \text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\right)^2\right] \\ \theta\gamma = -2E\left[\rho''(-\kappa Z_1) \cdot \text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\right] \\ 1 - \frac{\gamma}{\sigma\tau} = E\left[2\rho'(-\kappa Z_1) \cdot \left(1 + \gamma\rho''\left(\text{Prox}_{\gamma\rho(\cdot)}(\kappa\alpha Z_1 + \sigma Z_2)\right)\right)^{-1}\right], \end{cases} \tag{6}$$

where the first three equations involve the proximal operator of the regularizer in (5), the last three expressions involve the proximal operator of the logistic link function $\rho$ in (2). The random variables $Z, Z_1, Z_2 \sim N(0, 1)$ are mutually independent and independent of $B$; and $Z_1, Z_2$ come from decomposing $(X_{1\cdot}, \ldots, X_{n\cdot})^\top$ to orthogonal matrices [40, see p17 and later]. Importantly, the

first three equations, when conditioning on $B = \beta_j^*$, provide an approximation for the components of the regularized logistic regression using the soft-thresholding function [40, Eq. (90)]

$$\widehat{\beta}_j = \eta_{w(\beta_j^*)\sigma\tau}\left(\sigma\tau \cdot \left(\theta\beta_j^* + \frac{r}{\sqrt{\delta}}Z\right)\right),\ j = 1\ldots, p. \tag{7}$$

The right-hand-side of (7) suggests that $\widehat{\beta}_j$ can be understood as a thresholded random variable which is a convolution of the true signal $B$ and a scaled standard normal random variable $Z$. Then, the limiting performance of the regularized logistic regression estimator can be investigated in an averaged-over-components of $\beta^*$ fashion which will be stated in detail in Lemma 2.1.

The six parameters which are critical for describing $\widehat{\beta}_j$ in the limit, refer to

- $\alpha$: denotes the correlation and is relevant for the bias of the estimator $\widehat{\beta}$. It is defined as in [40, Corollary 1], where $\xrightarrow{P}$ denotes convergence in probability,

$$\frac{1}{\|\beta^*\|_2}\widehat{\beta}\beta^* \xrightarrow{P} \alpha. \tag{8}$$

- $\sigma$: relevant to the MSE. It is defined in [40, Corollary 2] using the following expression $\frac{1}{p}\|\frac{\widehat{\beta}}{\alpha} - \beta^*\|_2^2 \xrightarrow{P} \frac{\sigma^2}{\alpha^2}$.
- $\gamma$: parameter of the proximal operator of the logistic link function $\rho(\cdot)$.
- $\theta$: Lagrange multiplier in [40, Eq. 54].
- $\tau$: part of the parameter $(\lambda\sigma\tau)$ of the proximal operator of the regularizer.
- $r$: square-root of the average of a Lagrange multiplier [40, see above Eq. (45)].

The six parameters depend on the tuning scheme $w$, but the dependence is not explicitly indicated to simplify the notation. Specifically, when taking the tuning function $w(\beta;\lambda) = \lambda$, the proximal operator in (5) corresponds to the soft-thresholding operator and the system of equations in (6) can be used to analyze the limiting performance of the $\ell_1$-regularized logistic regression model. As a clarification, we make a comparison of the above six equations to Conjecture 4.3 of [43], which states a set of four equations to characterize regularized estimation in logistic regression models. Six equations are required to guarantee unique solutions of the six parameters $(\alpha, \sigma, \gamma, \theta, \tau, r)$. The six parameters arise from the proof technique in the CGMT framework. In contrast, Conjecture 4.3 is derived from the AMP framework, where a set of four equations suffices. The reason for considering the CGMT framework instead of AMP has been elaborated in the Introduction section.

We present example curves of the fixed point solutions against the tuning parameter $\lambda$ for different ratios $\delta = 0.5, 1, 5$ in Figure 1. We observe that the fixed point solutions stabilize at 0 for large $\lambda$, which typically leads to $\widehat{\beta}$ with all-zero components. By observing (8), it is reasonable that $\alpha$ converges to 0 when the components of $\widehat{\beta}$ are all zero since the inner product $\widehat{\beta}\beta^* = 0$ in that case. The MSE curves (see (9) for the definition) in the upper-left panel stabilize at the true signal strength when $\widehat{\beta}$ has all-zero components.

### 2.2. Limiting expressions of performance measures

The investigation of the limiting performance of the $\ell_1$-regularized logistic regression relies on [40, Theorem 1]. This theorem is similar to [3, Theorem 2] in the sense that it provides a powerful tool to analyze the averaged performance of the estimator $\widehat{\beta}$ in the limit. For the completeness of this paper, we state Theorem 1 of [40] in Lemma 2.1. Convergence in probability is obtained by applying the strong law of large numbers on the $\beta_j^*$'s and the Gaussian min-max theorem [27, 28].

**Lemma 2.1.** *Take $\widehat{\beta}(\lambda)$ as in (3), where $X_{i\cdot}$ satisfies (A1) and the components of $\beta^*$ satisfy (A2) with density function $f_B$. Assume for given parameters $\kappa$, $\delta$, $\lambda$, that the system of equations in (6) has a unique solution for the six parameters $(\alpha, \sigma, \gamma, \theta, \tau, r)$. Then, as $p \to \infty$, for any locally-Lipschitz function $\Psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$,*

$$\frac{1}{p}\sum_{j=1}^{p} \Psi\big(\widehat{\beta}_j, \beta_j^*\big) \xrightarrow{P} E\Psi\left(\eta_{w(B)\sigma\tau}\left(\sigma\tau \cdot \left(\theta B + \frac{r}{\sqrt{\delta}} Z\right)\right), B\right),$$

*where $Z \sim N(0,1)$ is independent of $B$.*

By choosing different functions $\Psi$, we are able to investigate various performance measures in the limit. We introduce here two special choices of the function $\Psi$ leading to the mean-squared-error (MSE) and out-of-sample classification accuracy.

#### 2.2.1. Asymptotic MSE

By choosing $\Psi(x, y) = (x - y)^2$ in Lemma 2.1 and defining the MSE of $\widehat{\beta}$ as

$$\mathrm{MSE}(\widehat{\beta}) = \frac{1}{p}\sum_{j=1}^{p}\big(\widehat{\beta}_j - \beta_j^*\big)^2,$$

the asymptotic MSE expression $E[\{\eta_{\lambda\sigma\tau}(\sigma\tau \cdot (\theta B + \frac{r}{\sqrt{\delta}} Z)) - B\}^2]$ follows immediately. By using the system of equations (6), Proposition 2.2 provides a simplified expression, its proof is in Appendix A.1.1.

**Proposition 2.2.** *Take $\widehat{\beta}(\lambda)$ as in (3), where $X_{i\cdot}$ satisfies (A1) and the components of $\beta^*$ satisfy (A2) with density function $f_B$. For given parameters $\kappa$, $\delta$, $\lambda$, assume that the six parameters $(\alpha, \sigma, \gamma, \theta, \tau, r)$ are a unique solution to the system of equations in (6). Then, when $p \to \infty$,*

$$\begin{aligned} \mathit{MSE}(\widehat{\beta}) &= \frac{1}{p}\sum_{j=1}^{p}\big(\widehat{\beta}_j - \beta_j^*\big)^2 \\ &\xrightarrow{P} E\left[\eta_{\lambda\sigma\tau}\left(\sigma\tau \cdot \left(\theta B + \frac{r}{\sqrt{\delta}} Z\right)\right) - B\right]^2 = \kappa^2(\alpha - 1)^2 + \sigma^2, \end{aligned} \tag{9}$$

*where $Z \sim N(0,1)$ is independent of $B$.*

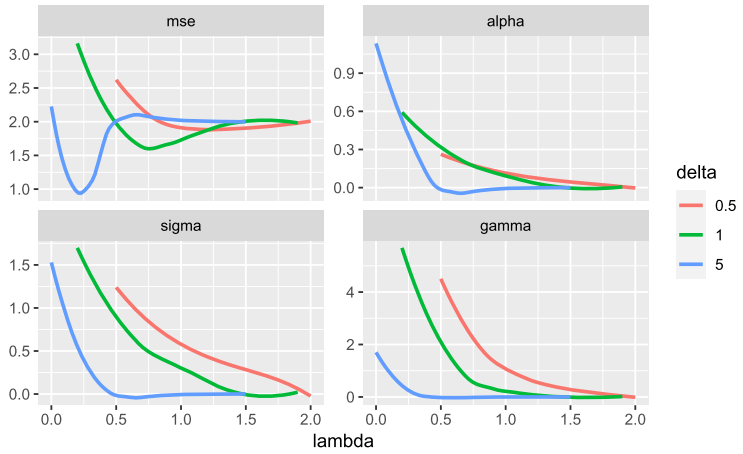


FIG 1. *Tuning parameter $\lambda$ against asymptotic MSE (upper-left), fixed point solutions of $\alpha$ (upper-right), $\sigma$ (bottom-left), $\gamma$ (bottom-right). We take the signal strength $\kappa^2 = 2$, the sparsity $s = 0.2$, the nonnull component $B' \sim N(0, \kappa^2/s)$.*

We plot the obtained asymptotic MSE expression in (9) against $\lambda$ in the upper-left panel of Figure 1, where we clearly see that there is a unique minimum of the MSE curves.

By Proposition 2.2, we propose to tune the parameter $\lambda \in \mathbb{R}_+$ by minimizing the asymptotic MSE expression, i.e.,

$$\lambda_{\text{opt}} = \arg\min_{\lambda} E\left[\left\{\eta_{\lambda\sigma\tau}\left(\sigma\tau\cdot\left(\theta B + \frac{r}{\sqrt{\delta}}Z\right)\right) - B\right\}^2\right] = \arg\min_{\lambda}\{\kappa^2(\alpha-1)^2 + \sigma^2\}, \tag{10}$$

where given $\kappa, \delta$, the parameters of the system of equations $(\alpha, \sigma, \gamma, \theta, \tau, r)$ vary for different values of the tuning parameter $\lambda$. The proposed optimal tuning parameter in (10) is important for Section 3 using selection by knockoffs, where we first tune $\lambda$. Figure 2 presents the minimum MSEs obtained by (9) for the following settings: (Left panel) $s = 0.1$, $\kappa^2 = 2$, $\delta = 0.5, 1$; (Right panel) $s = 0.5$, $\kappa^2 = 2$, $\delta = 2, 5$. Further, the limiting minimum MSEs obtained by (9) shown by solid triangles are compared with the boxplots based on $R = 300$ times finite sample replications. In each replication, a new dataset is randomly generated and $\widehat{\beta}$ is obtained using the R package `glmnet` with $\lambda$ obtained by leave-one-out cross-validation (CV). The solid triangle-shaped points refer to the minimum MSE value from (9) and the horizontal lines refer to the median minimum MSE by `glmnet` over 300 replications. We observe that the theoretical minimum MSE is close to the median minimum MSE by leave-one-out CV. This agrees with the conclusion from [52] for high dimensional linear models and suggests that our theoretical tuning well approximates the optimal tuning by CV in practice.

The tuning in (10), see Figure 2, is important for the following reasons: (1) Corollary 3.2 provides the limiting expressions for the tradeoff curves, which requires tuning $\lambda$ on the basis of limiting expressions. (2) Figure 2 shows that

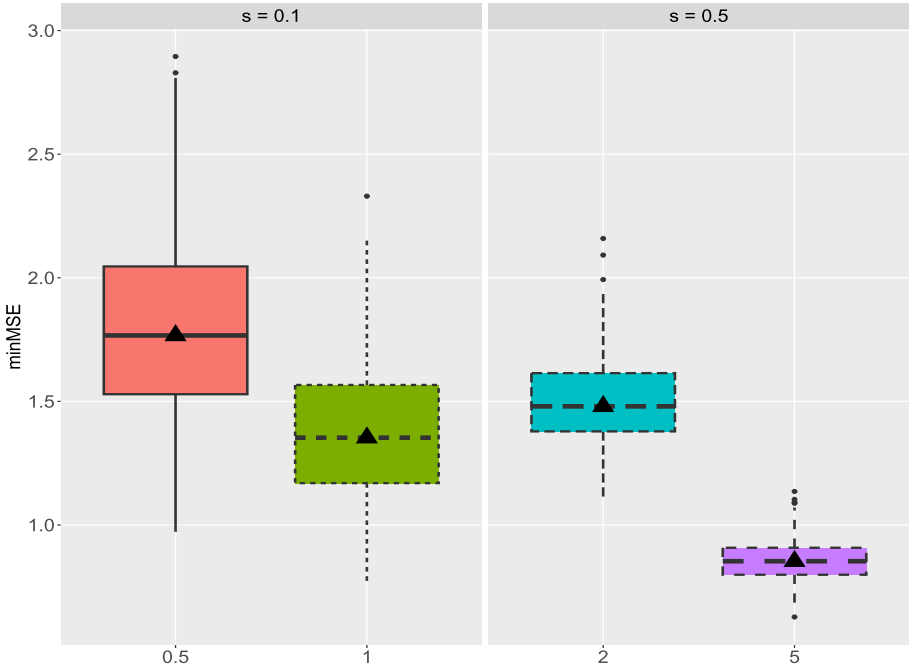


FIG 2. *Minimum MSE from (9) (solid triangle) and boxplots for finite samples. The boxplots are based on $R = 300$ times replication. In each replication, a new dataset is randomly generated, and $\widehat{\beta}$ is obtained using the R package* `glmnet` *with $\lambda$ set by leave-one-out cross-validation. We present the following settings: (Left panel) sparsity $s = 0.1$, $\kappa^2 = 2$, $\delta = 0.5, 1$; (Right panel) sparsity $s = 0.5$, $\kappa^2 = 2$, $\delta = 2, 5$.*

the minimum MSE by Corollary 3.2 is very close to tuning by leave-one-out CV which provides optimal tuning of $\lambda$ and minimizes the estimation risk [52]. This suggests that when tuning by Corollary 3.2 for limiting curves; the corresponding limiting curves approximate the optimal situation for finite sample curves using CV-tuned $\lambda$.

### *2.2.2. Classification accuracy*

Since logistic regression is a powerful classification tool, it is interesting to formally investigate classification accuracy. Assume we have new samples $(Y_{i.}', X_{i.}')$, $i = 1, \ldots, n_{\text{new}}$ for prediction. By the logistic link function in (1), $\widehat{Y}_{i.} = 1$ is classified to have label 1 if $X_{i.}\widehat{\beta} \geq 0$ and 0 if $X_{i.}\widehat{\beta} < 0$. Consequently, we define the classification accuracy (CA) by

$$\text{CA}(\widehat{\beta}) = \sum_{i=1}^{n_{\text{new}}} I\left\{\left\{Y_i' - I\left\{\sum_{j=1}^{p} X_{i,j}'\widehat{\beta}_j \geq 0\right\} = 0\right\}\right\}/n_{\text{new}}.$$

When $n_{\text{new}} \to \infty$, we obtain the following convergence by the strong law of large numbers.

**Proposition 2.3.** *For a fixed regularization parameter $\lambda > 0$, let $X_i$. satisfy (A1) for all $i = 1, \ldots, n$ and the entries of $\beta^*$ satisfy (A2) with density function $f_B$. Given parameters $\kappa$, $\delta$, $\lambda$, assume that the system of equations in (6) has a unique solution to the six parameters $(\alpha, \sigma, \gamma, \theta, \tau, r)$. The following almost sure convergence holds for the out-of-sample prediction accuracy*

$$CA(\widehat{\beta}) \xrightarrow{a.s.} P\left(\left\{Y - I\left\{\left[\sum_{j=1}^{p} X_{\cdot j} \cdot \eta_{\lambda\sigma\tau}\left(\sigma\tau \cdot \left(\theta\beta_j^* + \frac{r}{\sqrt{\delta}}Z\right)\right)\right] \geq 0\right\}\right\} = 0\right). \tag{11}$$

*Proof.* By [40, Eq. (90)] taking

$$\widehat{\beta}_j = \eta_{\lambda\sigma\tau}\left(\sigma\tau \cdot \left(\theta\beta_j^* + \frac{r}{\sqrt{\delta}}Z\right)\right),\ j = 1, \ldots, p$$

and applying the strong law of large numbers letting $n_{\text{new}} \to \infty$, (11) follows immediately. □

By observing the almost sure convergence in Proposition 2.3, it is obvious that the classification accuracy is determined by the distribution of $Y, X$, and the estimator $\widehat{\beta}$. Further, the estimator $\widehat{\beta}$, for a given tuning parameter $\lambda$, can be expressed using the six parameters $(\alpha, \sigma, \gamma, \theta, \tau, r)$ which are uniquely determined by the ratio $n/p \to \delta$ and the signal strength $\kappa$. Hence, the classification accuracy for the data generating process in (1) is determined by $\delta, \kappa$ as a function of $\lambda$. That is, for a system with fixed $\delta, \kappa$, for given $\lambda$, we obtain a set of fixed point solutions $(\alpha, \sigma, \gamma, \theta, \tau, r)$, which determines the limiting classification accuracy in (11) for logistic regression in (2).

### *2.3. A tradeoff curve based on selection by $\widehat{\beta} \neq 0$*

Emerging literature has attempted to analyze the tradeoff between true and false discoveries, i.e., the correctly and incorrectly estimated nonnulls for high dimensional linear models using the limiting AMP analysis, see for example [49, 41, 9, 48]. The tradeoff curve helps us to understand the connection between the falsely estimated nonnulls and the correctly estimated nulls. We first show Lemma 2.4 for the *Gaussian min-max theorem* framework for a smoother transition to obtain the limiting proportions of $\text{FDP}^{\text{est}}$ and $\text{TPP}^{\text{est}}$. Lemma 2.4 can be obtained by taking the locally-Lipschitz function $\Psi$ in the general result in Lemma 2.1 to be an indicator function for events such as $\{j : \beta_j = 0, \widehat{\beta}_j(\lambda) = 0\}$. Lemma 2.4 agrees with [49, Lemma 2.1] and [41, Lemma A.1] for the case of linear regression, which are obtained similarly using [7, Theorem 1] and [4, Theorem 1.5] in the approximate message passing framework. Further, the system of equations (6) specifies certain parameters that are relevant to the true parameter vector $\beta^*$, such as $B'$, $B$, $\kappa^2$, $s$; this structure has both pros and cons.

Due to this specification, we can investigate the impact of sparsity $s$ and signal strength $\kappa^2$ on the selection tradeoff. The downside is that plugging in the estimators of these parameters brings bias to the tradeoff curve for practical use.

To discuss the selection tradeoff, we first introduce the "false discovery proportion" (FDP) and "true positive proportion" (TPP) for variable selection. In addition, specifying the sparsity $s$ is unavoidable since FDP and TPP are computable only when the estimated nulls (nonnulls) are compared with the true nulls (nonnulls). We denote the index set $\mathcal{H} = \{1, \ldots, p\}$.

For any $\lambda > 0$ with a corresponding estimator $\widehat{\beta}(\lambda)$ in (3), we denote $V^{\text{est}}(\lambda) = \#\{j \in \mathcal{H} : \widehat{\beta}_j(\lambda) \neq 0, \beta_j = 0\}$ the number of false discoveries, $T^{\text{est}}(\lambda) = \#\{j \in \mathcal{H} : \widehat{\beta}_j(\lambda) \neq 0, \beta_j \neq 0\}$ the number of true discoveries, and $\#\{j \in \mathcal{H} : \beta_j \neq 0\} \to s \cdot p$ the number of true nonnulls.

The "false discovery proportion" (FDP) and the "true positive proportion" (TPP) for the estimator $\widehat{\beta}(\lambda)$ are defined as

$$\text{FDP}^{\text{est}} = \frac{V^{\text{est}}(\lambda)}{\max(\#\{j \in \mathcal{H} : \widehat{\beta}_j(\lambda) \neq 0\}, 1)}, \tag{12}$$

$$\text{TPP}^{\text{est}} = \frac{T^{\text{est}}(\lambda)}{\max(\#\{j \in \mathcal{H} : \beta_j \neq 0\}, 1)}. \tag{13}$$

By the system of equations in (6), we obtain the limiting expressions of $\text{FDP}^{\text{est}}$ and $\text{TPP}^{\text{est}}$ in Lemma 2.4, letting $p_n \to \infty$. The proof is in Appendix A.1.2.

**Lemma 2.4.** *For a fixed regularization parameter $\lambda > 0$, for $i = 1, \ldots, n$ let $X_i$. satisfy (A1) and the components of $\beta^*$ satisfy (A2) with density function $f_B$. In addition, assume for given parameters $\kappa$, $\delta$, $\lambda$, the system of equations in (6) has a unique solution for the six parameters $(\alpha, \sigma, \gamma, \theta, \tau, r)$. Then, as $p \to \infty$, the $\ell_1$-regularized logistic regression coefficient estimator $\widehat{\beta}(\lambda)$ complies with*

$$\frac{V^{\text{est}}(\lambda)}{p} \xrightarrow{P} 2(1-s) \cdot \Phi\left(-\frac{\lambda}{\frac{r}{\sqrt{\delta}}}\right), \tag{14}$$

$$\frac{T^{\text{est}}(\lambda)}{p} \xrightarrow{P} s \cdot P\left(\left|\theta B' + \frac{r}{\sqrt{\delta}} Z\right| > \lambda\right), \tag{15}$$

*where $Z \sim N(0,1)$ independent of $B'$, and $\theta, r$ are unique solutions of the system of equations (6) with six parameters $(\alpha, \sigma, \gamma, \theta, \tau, r)$. Consequently, the limiting expressions of "false discovery proportion" (FDP) and "true positive proportion" (TPP) based on $\widehat{\beta}$ are as follows*

$$FDP^{\text{est}}(\lambda) \xrightarrow{P} \frac{2(1-s) \cdot \Phi(-\frac{\lambda}{\frac{r}{\sqrt{\delta}}})}{2(1-s) \cdot \Phi(-\frac{\lambda}{\frac{r}{\sqrt{\delta}}}) + s \cdot P(|\theta B' + t_2 Z| > \lambda)} = fdp^{\text{est}}(\lambda), \tag{16}$$

$$TPP^{\text{est}}(\lambda) \xrightarrow{P} P\left(|\theta B' + \frac{r}{\sqrt{\delta}} Z| > \lambda\right) = tpp^{\text{est}}(\lambda). \tag{17}$$

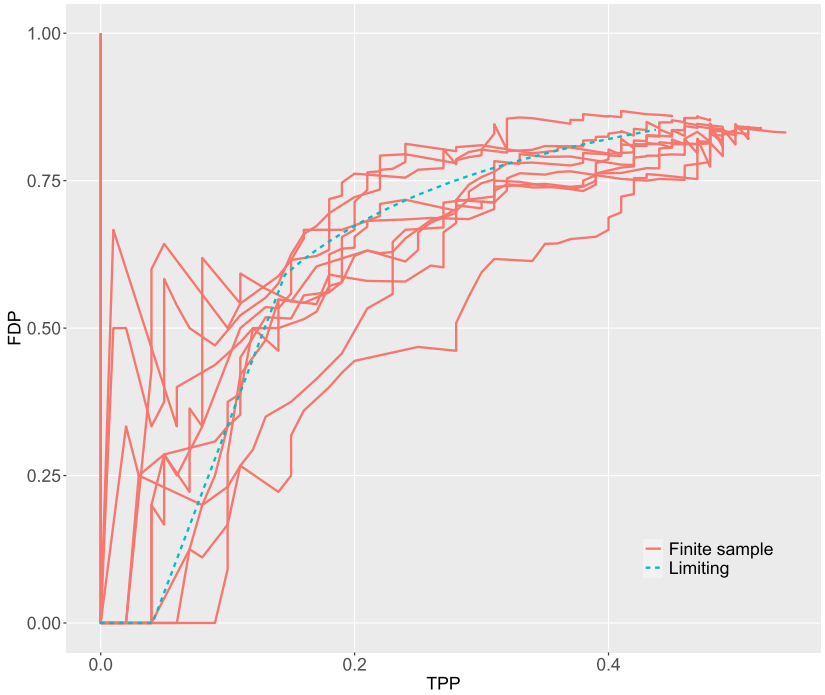


FIG 3. $q^{B'}(TPP^{\text{est}}(\lambda); s, \delta, \kappa^2) = FDP^{\text{est}}(\lambda)$ *for* $\delta = 0.5$, $s = 0.1$, $\kappa^2 = 2$. *The blue dotted curve* $q^{B'}(tpp^{\text{est}}(\lambda); s, \delta, \kappa^2) = fdp^{\text{est}}(\lambda)$ *is obtained by the limiting expressions in* (16), (17). *The red curves are from 10 times finite sample realization by* `glmnet`.

The limiting expressions (16), (17) clearly show that $\text{tpp}^{\text{est}}$ is a component of the denominator of $\text{fdp}^{\text{est}}$. This complies with the conclusion in [41] that $\text{FDP}^{\text{est}}$ is a function of $\text{TPP}^{\text{est}}$ and relevant parameters. We denote this function by $q^{B'}(\text{TPP}^{\text{est}}(\lambda); s, \delta, \kappa^2) = \text{FDP}^{\text{est}}(\lambda)$ for which the limiting curve $q^{B'}(\text{tpp}^{\text{est}}(\lambda); s, \delta, \kappa^2) = \text{fdp}^{\text{est}}(\lambda)$. Figure 3 shows example limiting and 10 times finite sample tradeoff curves for $\delta = 0.5$, $s = 0.1$ and $\kappa^2 = 2$. The finite sample curves are generated using $p = 500$ fitted by the R package `glmnet`. The dashed turquoise curve is obtained by the limiting expressions in (16) and (17). The overall trend of the finite samples curves is captured by the limiting curve. Due to sampling variability, we observe the difficulty of selection, especially for low TPP.

The derived limiting curve resembles the type-I and type-II error tradeoff curves for classical hypothesis testing. It is foreseeable since, in principle, the variable selection problem classifies the $\beta_j^*$'s to be nulls and nonnulls. Then the FDP and TPP are calculated by comparing the discrepancy between the estimated and the true nonnull vector. For hypothesis testing, the hypothesis is either true or false, which is also a binary outcome similar to a variable being null or nonnull. The derived limiting function $q^{B'}(\text{TPP}^{\text{est}}(\lambda); s, \delta, \kappa^2)$ validates the

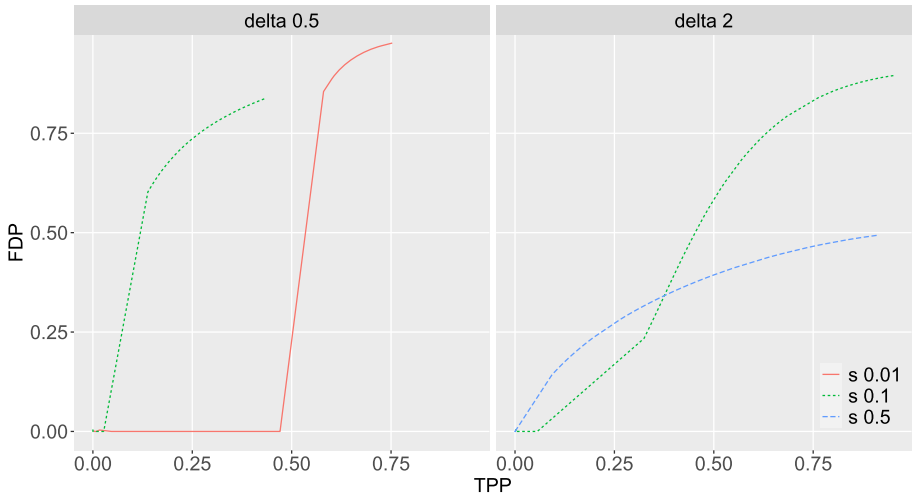


FIG 4. *Comparison of the impact of sparsity on the limiting tradeoff curve* $q^{B'}(tpp^{\text{est}}(\lambda); s, \delta, \kappa^2) = fdp^{\text{est}}(\lambda)$. *The sparsity levels* $s = 0.01, 0.1$ *are considered for the setting where* $\kappa^2 = 2$, $\delta = 0.5$ *(left panel), while the sparsity levels* $s = 0.1, 0.5$ *are considered for the setting where* $\kappa^2 = 2$, $\delta = 2$ *(right panel).*

tradeoff of type-I (FDP) and type-II (1-TPP) variable selection errors. Type-I and type-II errors cannot be both low; focusing on minimizing the type-I error would inflate the type-II error. For convenience, we follow the convention of [41] and name the curve $q^{B'}(\text{TPP}^{\text{est}}(\lambda); s, \delta, \kappa^2)$ a "tradeoff" curve. Since the tradeoff function $q^{B'}(\text{TPP}^{\text{est}}(\lambda); s, \delta, \kappa^2)$ has three deterministic parameters $s$, $\delta$, and $\kappa^2$, we investigate the individual impact of the three parameters on the tradeoff curve.

Figure 4 compares the tradeoff curve influenced by different sparsity values for two settings: (1) $\kappa^2 = 2$, $\delta = 0.5$; (2) $\kappa^2 = 2$, $\delta = 2$. Since the consistency of the $\ell_1$-regularized estimator requires certain sparsity, for the setting where $\delta = 0.5$, i.e. $p \geq n$, we consider high sparsity $s = 0.01$ and medium sparsity $s = 0.1$. For $\delta = 2$, i.e. $p < n$, we consider $s = 0.1, 0.5$. From Figure 4, we observe that for given FDP level, high sparsity causes TPP loss, which in plain words is: high sparsity induces difficulty in selecting relevant variables.

In addition, Figure 5 investigates the impact of the signal strength $\kappa^2$ in various settings where $\delta = 2, 5$ and $s = 0.1, 0.5$. Two different signal strengths $\kappa^2 = 2, 4$ are considered. It is not surprising that, given FDP, we observe higher TPP for stronger signal strength $\kappa^2$.

Further, we propose two practical uses of the tradeoff curve for variable selection—sample size calculation in Section 2.3.1 and FDR level calibration in Section 2.3.2.

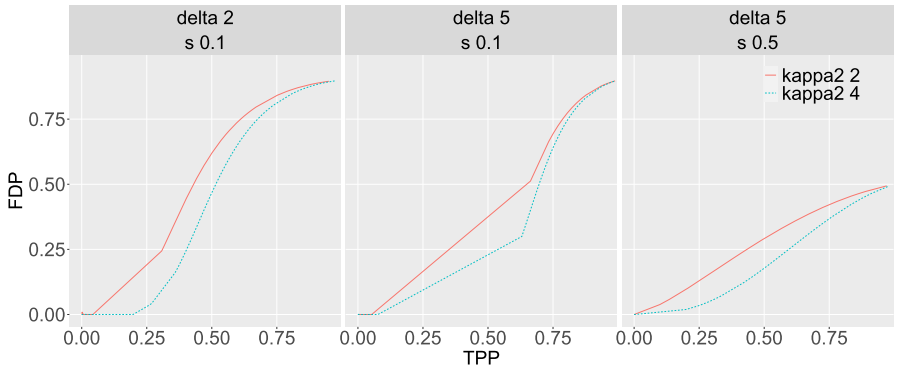


FIG 5. *Comparison of the impact of the signal strength* $\kappa^2$ *on the limiting tradeoff curve* $q^{B'}(tpp^{\text{est}}(\lambda); s, \delta, \kappa^2) = fdp^{\text{est}}(\lambda)$. *Two different signal strengths* $\kappa^2 = 2, 4$ *are considered for settings where* $\delta = 2, 5$ *and* $s = 0.1, 0.5$.

### *2.3.1. Sample size calculation*

Similar to the hypothesis testing procedure, the FDP (type-I error) is often predetermined. We manually set the FDP to be under a certain level $q$. By using the limiting tradeoff curve $q^{B'}(\text{TPP}^{\text{est}}(\lambda); s, \delta, \kappa)$, a corresponding $\text{tpp}_{\text{est}}$ can be determined by the inverse of $q^{B'}$. In experimental design, we expect the experimenters to collect a certain number of variables in experiments indicating unchanged $p$, $s$, $\kappa^2$. Thus, for predetermined $\text{fdp}_{\text{est}}$, one could consider multiple values of $\delta$ corresponding to different sample sizes $n$, to achieve the desired $\text{tpp}_{\text{est}}$.

Figure 6 shows an example simulation by fixing the sparsity $s = 0.1$, $\kappa^2 = 2$ and for different ratios $\delta = 0.2, 0.5, 1, 2$. As expected, given FDP, larger values of the ratio $\delta$ provide a higher TPP suggesting a higher percentage of correctly selected true relevant variables. In this example, if the experiment requires FDP $\leq 0.2$ and TPP $\geq 0.2$, only $\delta = 2$ suffices. Since Assumption (A1) suggests independence of the predictive variables providing an optimal (uncorrelated) design for selection, the TPPs obtained in this scenario provide an upper bound for TPPs for fixed $\kappa^2$, $\delta$, and $s$. In practice, when the $X_{\cdot j}$'s are correlated, we expect a TPP loss resulting in a steeper tradeoff curve. This phenomenon is relevant to the "irrepresentable condition" discussed in [36, 55]. In general, to guarantee asymptotic selection consistency, the small eigenvalues of the population covariance matrix of the submatrix consisting of the relevant variables should be bounded away from zero. Some example population covariance matrix structures such as constant positive correlation, power decay correlation, and bounded correlation are discussed in [55]. But as mentioned in the Introduction, sparsity pattern recovery of $\widehat{\beta}$ is merely an asymptotic property, and the "irrepresentable condition" is often violated in practice making estimation more challenging. When the desired TPP cannot be achieved for certain $\delta$ under this uncorrelated design, this selection TPP is not achievable using this $\delta$ for

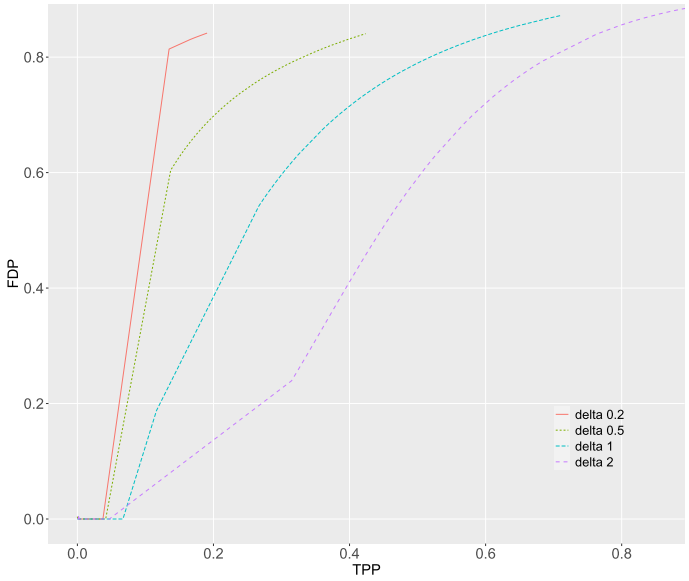


FIG 6. *The impact of the ratio $\delta$ on the limiting tradeoff curve $q^{B'}(tpp^{\text{est}}(\lambda); s, \delta, \kappa^2) = fdp^{\text{est}}(\lambda)$. The curves are obtained for different ratios $\delta = 0.2, 0.5, 1, 2$ while fixing $s = 0.2$, $\kappa^2 = 2$.*

any other correlation structure. Further, we observe a sharp decrease of TPP for the $\ell_1$-regularized logistic regression model compared to the $\ell_1$-regularized least-squares estimator for the linear regression model [41]. The detailed reason for the sharp decrease is unknown but possibly due to the binary response for logistic regression or the link function, which builds a nonlinear association between $X_{\cdot j}$'s and $Y$. This aspect deserves further investigation.

### *2.3.2. FDR level calibration*

Often a dataset with fixed $\delta$, $s$, $\kappa^2$ is obtained, and we want to select variables relevant to the response variable. While the selection is performed by $\ell_1$-regularized estimator, the tradeoff curve $q^{B'}(\text{TPP}^{\text{est}}(\lambda); s, \delta, \kappa)$ can be utilized to determine a suitable level $q$ for FDP which correspondingly provides a reference level of TPP. Setting a low target level of FDP would inevitably restrict the level of TPP in high dimensions, suggesting that the selected variables can hardly provide sufficient information for further analysis since not many truly relevant variables are selected. The proposed tradeoff curve could assist in determining a suitable target FDP level for which the corresponding TPP level is adequate. We illustrate this using the `Wisconsin Breast Cancer` dataset obtained from the UCI machine learning repository in Section 4.

## 3. Knockoff calibration

Section 2.3 introduced a TPP-FDP tradeoff curve for $\ell_1$-regularized logistic regression based on correctly and falsely estimated nonnulls. However, variable selection by the regularized estimator $\widehat{\beta}$ has been controversial for long. The selection uncertainty of $\widehat{\beta}$ causes false discovery bias, i.e, true nulls are falsely selected, which causes concerns about the reproducibility of research in different fields. Further, similar to hypothesis testing using the tradeoff curve to compare multiple tests, we want to construct the tradeoff curve for an alternative selection method. By comparing the curves in one plot, one can easily identify the method with a better TPP for a given FDP.

For comparing the tradeoff curves, we consider an FDR-controlled selection method called model-X knockoff [11], of which the selection still relies on regularization hence it does not deviate far from the approach in Section 2.3. Further, we want to investigate if selection by knockoffs has reasonable TPP. Since controlling FDR would inevitably set restrictions on the selection, which intuitively would cause a loss on TPP, see [50] showing the power loss of knockoff as opposed to selection by $\ell_1$-regularized least-squares estimators for the linear model.

Let the knockoff vector $X' = (X'_{\cdot 1}, \ldots, X'_{\cdot p})$ be an independent copy of $X$ with i.i.d components $X'_{\cdot j} \sim N(0, 1/p)$, $j = 1, \ldots, p$. This also indicates that $X'$ is independent of $Y$ since $X'$ is simply an independent random Gaussian vector. By this construction, we consider the optimal selection TPP for knockoffs since the correlation between $X'$ and $X$ can affect the TPP. The samples $X'_{i\cdot}$, $i = 1, \ldots, n$ are i.i.d. copies of the vector $X'$. The importance of the original variables $X_{\cdot j}$'s is measured by an importance statistic in the knockoff framework, which is computed by using the original variables $X_{\cdot j}$'s and their knockoff counterparts $X'_{\cdot j}$'s corresponding to the estimator

$$\widehat{\beta}(\lambda) = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \big(\rho(\widetilde{X}_{i\cdot}\beta) - Y_i(\widetilde{X}_{i\cdot}\beta)\big) + \frac{\lambda}{2p}\|\beta\|_1 \right\}, \tag{18}$$

where $\widetilde{X}_{i\cdot} = (X_{i\cdot}, X'_{i\cdot}), i = 1, \ldots, n$ with true coefficient vector $\widetilde{\beta}^* = ({\beta^*}^\top, \mathbf{0}_p^\top)^\top$, $\mathbf{0}_p = (0, \ldots, 0)^\top \in \mathbb{R}^p$. With the extra $p$ zero components, we denote the number of parameters $\tilde{p} = 2p$, the ratio $n/\tilde{p} \to \tilde{\delta}$, and the sparsity (the limiting probability of $\beta$ being nonnull) $\tilde{s} = s/2$. In addition, the signal strength becomes $\|\widetilde{\beta}^*\|_2^2/\tilde{p} \to \tilde{\kappa}^2 = \kappa^2/2$.

Following the knockoff framework, for any importance statistic $W_j$ and fixed threshold $t$, the *false discovery proportion (FDP)* and *true positive proportion (TPP)* are similar to the expressions in (12) and (13) with the selection adjusted from $\widehat{\beta}_j \neq 0$ to $W_j \geq t$. Then, the *false discovery proportion (FDP)* for selection by knockoffs is defined as

$$\mathrm{FDP}^{\mathrm{kno}}(t) = \frac{\#\{j \in \mathcal{H} : W_j \geq t, \beta_j^* = 0\}}{\#\{j \in \mathcal{H} : W_j \geq t\}}, \tag{19}$$

and the *true positive proportion (TPP)* is defined as

$$\mathrm{TPP}^{\mathrm{kno}}(t) = \frac{\#\{j \in \mathcal{H} : W_j \geq t, \beta_j^* \neq 0\}}{\#\{j \in \mathcal{H} : \beta_j^* \neq 0\}}. \tag{20}$$

To obtain detailed $\mathrm{FDP}^{\mathrm{kno}}$ and $\mathrm{TPP}^{\mathrm{kno}}$ expressions, the feature statistic $W_j$ needs to be specified. Two prevailing options—*Lasso Coefficient Difference* and *Lasso Signed Max*—are investigated thoroughly in various literature [see for example 2, 11, 30]. The *Lasso Signed Max* statistic is defined as

$$W_j^{\mathrm{lsm}} = \mathrm{sign}(|Z_j| - |Z_{j+p}|) \max\{|Z_j|, |Z_{j+p}|\}, \tag{21}$$

where $|Z_j| = \sup\{\lambda : \widehat{\beta}_j(\lambda) \neq 0\}$. The LSM statistic can be simplified to the following two cases

$$W_j^{\mathrm{lsm}} = \begin{cases} Z_j & Z_j \geq Z_{j+p} \\ -Z_{j+p} & Z_j < Z_{j+p}. \end{cases}$$

The absolute value signs are dropped since the $Z_j$'s are values of the tuning parameter $\lambda$ which are nonnegative. Alternatively, the *Lasso Coefficient Difference* statistic, which is shown to be more powerful than the *Lasso Signed Max* statistic in various simulation settings in [11], is defined as

$$W_j^{\mathrm{lcd}} = |\widehat{\beta}_j(\lambda)| - |\widehat{\beta}_{p+j}(\lambda)|. \tag{22}$$

The selection set $\widehat{S}$ containing the selected important parameters is defined as

$$\widehat{S} = \{j : W_j \geq \widehat{t}\},$$

where the estimated threshold $\widehat{t}$ determines the number of variables entering $\widehat{\mathcal{S}}$. We discuss next in Section 3.1 the choices of the threshold $t$.

### *3.1. Choices of the threshold t for FDR and k-FWER control*

An often-seen choice of $t$ in the knockoff literature performs FDR control at level $q$ and is defined as follows

$$\widehat{t} = \min\{t > 0 : \widehat{\mathrm{FDP}}(t) \leq q\}, \widehat{\mathrm{FDP}}(t) = \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}}. \tag{23}$$

Alternatively, the $k$-FWER control in multiple testing is also of practical interest, which considers the following probability

$$P(V \geq k) \leq q,$$

where $V^{\mathrm{est}}(\lambda) = \#\{j \in \mathcal{H} : \widehat{\beta}_j(\lambda) \neq 0, \beta_j = 0\}$ is the number of false discoveries. Similar to [30] for the *Lasso Signed Max* statistic based on the ordered importance statistics $W_{(j)}$'s, we propose the threshold for

$$W_{(1)} \geq W_{(2)} \geq \cdots \geq W_{(p)}$$

defined by

$$t_v = \sup\{t > 0 : \#\{j : W_{(j)} \leq -t\} = v\}, \tag{24}$$

which allows in total $v$ statistics $W_{(j)}$'s exceeding a certain threshold. Define the number of false discoveries $V = \#\{j \in \mathcal{H} : W_{(j)} \geq t_v, \beta_j^* = 0\}$. Next, we explain the choice of using a Poisson distribution with mean $v$, i.e., Pois$(v)$, to approximate $V$. By (24), the threshold $t_v$ is chosen such that the number of $W_{(j)}$'s with values less than $-t_v$ is $v$. Then

$$\#\{j \in \mathcal{H} : W_{(j)} \leq -t_v, \beta_j^* = 0\} = (1-s)\#\{j \in \mathcal{H} : W_{(j)} \leq -t_v\} = (1-s)v, \tag{25}$$

where the first equality holds simply by conditioning and the second equality holds by the definition of $t_v$. In practice, $W_{(j)} \leq -t$ happens mostly for the true nulls $\beta_j^* = 0$ when $\beta_j^* = \beta_{j+p}^* = 0$. While given $\beta_j^* \neq 0$, $W_{(j)} \leq -t$ happens when $\beta_j^*$ is a weak signal which does not deviate far from 0. Further, by [2, Lemma 1], it holds that

$$\#\{j \in \mathcal{H} : W_{(j)} \leq -t_v, \beta_j^* = 0\} \stackrel{d}{=} \#\{j \in \mathcal{H} : W_{(j)} \geq t_v, \beta_j^* = 0\} = V(t_v). \tag{26}$$

Hence, by (25) and (26), the number of false discoveries $V(t_v) = (1-s)v$. In high sparsity settings where $s \to 0$, $V(t_v) \to v$, dropping the unknown scaling factor $s$ would still lead to a satisfactory estimator for $V$. Similar approximation for estimating $V$ is a common technique in the knockoff literature to obtain estimators that can be used in practice; see for example, [11, Section 3.3]. We use a Poisson distribution with parameter $v$ denoted by Pois$(v)$ to model $V(t_v)$, where $t_v$ denotes a continuous threshold/timeline and $(W_{(j)} \leq t_v)$'s are the events. Notice that the intensity parameter $v$ is upward biased, leading to a conservative $k$-FWER control. A similar construction can be found in [30, Theorem 3.1], who investigated the *Lasso Signed Max* statistic and used a negative binomial distribution, while we used a Poisson distribution of the false discoveries $V$.

Then, for any integer $k \geq 1$ for $k$-FWER control and significance level $q \in (0,1)$, let $v$ be the largest integer satisfying

$$\sum_{i_v=k}^{\infty} \frac{v^{i_v} e^{-v}}{i_v!} \leq q. \tag{27}$$

It follows that the knockoff procedure controls $k$-FWER control at significance level $q$, i.e., $P(V \geq k) \leq q$.

### *3.2. Limiting tradeoff curve after knockoff calibration*

The *Lasso Coefficient Difference* statistic calculates the differences of coefficient estimators of the original variables and their knockoff counterparts. In this section while letting $p_n \to \infty$ (or simply $p \to \infty$), we consider an optimal parameter tuning for $\lambda$ by minimizing the asymptotic MSE expressions of $\widehat{\beta}$, see (10). Further, we consider the system of equations (6) to investigate FDP and TPP in

the limit for $p \to \infty$, the latter which are critical for this paper by the extensions of [50, Corollary 3.1 and 3.2].

The following discussion is based on the limiting expressions of performance measures of $\widetilde{\beta}^* = (\beta^{*\top}, \mathbf{0}_p^\top)^\top$ for $\widetilde{X}_{i\cdot} = (X_{i\cdot}, X'_{i\cdot})$'s where $X'_{i\cdot}$ consists of knockoff variables. Hence, we first adjust Lemma 2.1 [40, Theorem 1] for knockoff calibration. We denote the estimator in (18) $\widehat{\beta} = ((\widehat{\beta}^{\text{org}})^\top, \widehat{\beta}'^\top)^\top$ for which the following Theorem holds. The proof is in the Appendix.

**Theorem 3.1.** *Take $\widehat{\beta}(\lambda)$ as in (18) and given the adjusted parameters $\tilde{\kappa}$, $\tilde{\delta}$, $\tilde{\lambda}$, assume that the system of equations in (6) has a unique solution to the six parameters $(\tilde{\alpha}, \tilde{\sigma}, \tilde{\gamma}, \tilde{\theta}, \tilde{\tau}, \tilde{r})$. Then, as $\tilde{p} \to \infty$, for any locally-Lipschitz function $\Psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $\widetilde{\Psi}(x_1, x_2, x_3) = \Psi(x_1, x_2) + \Psi(x_3, 0)$,*

$$\begin{aligned}&\frac{1}{p}\sum_{j=1}^{p} \widetilde{\Psi}\big(\widehat{\beta}_j, \beta_j^*, \widehat{\beta}_{j+p}\big) \\ &\quad \xrightarrow{P} E\left[\widetilde{\Psi}\left(\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\left(\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right)\right), B, \eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\left(\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'\right)\right)\right)\right], \end{aligned} \tag{28}$$

*where $Z, Z'$ are independent with $N(0,1)$ distribution and are further independent of $B$.*

Remark: the additional standard Gaussian distributed random variable $Z'$ is crucial for differentiating $\widehat{\beta}_j^{\text{org}}$ and $\widehat{\beta}'_j, j = 1, \ldots, p$, especially for the true nonnulls. Intuitively, $Z$ and $Z'$ come from the original predictive variables $X_{\cdot j}$'s and the corresponding knockoff variables $X'_{\cdot j}$'s; using two independent variables $Z, Z'$ better describes the independence of the knockoff variables which ensures an optimal TPP for the knockoff selection.

We denote the limiting proportions for a fixed threshold $t$ by

$$\begin{aligned}&\text{tpp}^{\text{kno}}(t) = \lim_{p_n \to \infty} \text{TPP}^{\text{kno}}(t), \\ &\text{fdp}^{\text{kno}}(t) = \lim_{p_n \to \infty} \text{FDP}^{\text{kno}}(t), \widehat{\text{fdp}}^{\text{kno}}(t) = \lim_{p_n \to \infty} \widehat{\text{FDP}}^{\text{kno}}(t). \end{aligned} \tag{29}$$

While using the *Lasso Coefficient Difference* statistic in (22) for knockoff calibration, the limiting proportions in (29) are expressed analytically in Corollary 3.2, which is consistent with [50, Corollary 3.1, 3.2] for the *Lasso Coefficient Difference* statistics. Further, while using the *Lasso Signed Max* statistic, the limiting expressions and the bias analysis are also coherent (see Appendix A.2 for details), meaning the upward biased estimator of the FDP can be decomposed into FDP and the bias term in the limit. Further, the bias increases with FDP.

**Corollary 3.2.** *For any $t > 0$, the limiting expressions of the $fdp^{\text{kno}}(t)$ and $tpp^{\text{kno}}(t)$, while using the* Lasso Coefficient Difference *statistic for knockoff calibration, are as follows*

$$fdp^{\text{kno}}(t) = \frac{(1-s) \cdot P(|\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}(\tilde{\sigma}\tilde{\tau} \cdot \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z)| - |\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}(\tilde{\sigma}\tilde{\tau} \cdot \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z')| \geq t)}{P(|\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}(\tilde{\sigma}\tilde{\tau} \cdot (\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z))| - |\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}(\tilde{\sigma}\tilde{\tau} \cdot \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z')| \geq t)}, \tag{30}$$

$$tpp^{\text{kno}}(t) = P\left(\left|\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\cdot\left(\tilde{\theta}B' + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right)\right)\right| - \left|\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\cdot\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'\right)\right| \geq t\right). \quad (31)$$

*Similarly, the estimator $\widehat{fdp}(t)$ has the following limiting expression*

$$\widehat{fdp}^{\text{kno}}(t) = \frac{P(|\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}(\tilde{\sigma}\tilde{\tau}\cdot(\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z))| - |\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}(\tilde{\sigma}\tilde{\tau}\cdot\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z')| \leq -t)}{P(|\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}(\tilde{\sigma}\tilde{\tau}\cdot(\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z))| - |\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}(\tilde{\sigma}\tilde{\tau}\cdot\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z')| \geq t)}. \quad (32)$$

It is known that the knockoff estimator $\widehat{\text{FDP}}^{\text{kno}}$ is larger than the true $\text{FDP}^{\text{kno}}$. The limiting proportions in Corollary 3.2 provide a tool to theoretically evaluate the bias. The bias of the estimator $\widehat{\text{FDP}}^{\text{kno}}$ for any statistic $W_j$ can be rewritten as the true $\text{fdp}^{\text{kno}}$ plus a remainder term $R(t)$. We present here

$$\begin{aligned}
\widehat{\text{FDP}}^{\text{kno}}(t) &= \frac{\#\{j \in \mathcal{H} : W_j \leq -t\}}{\#\{j \in \mathcal{H} : W_j \geq t\}} \\
&= \frac{(1-s)\cdot\#\{j\in\mathcal{H} : W_j \leq -t, \beta_j^* = 0\} + s\cdot\#\{j\in\mathcal{H} : W_j \leq -t, \beta_j^* \neq 0\}}{\#\{j \in \mathcal{H} : W_j \geq t\}} \\
&\approx \frac{(1-s)\cdot\#\{j : W_j \geq t, \beta_j^* = 0\} + s\cdot\#\{j \in \mathcal{H} : W_j \leq -t, \beta_j^* \neq 0\}}{\#\{j \in \mathcal{H} : W_j \geq t\}} \\
&\xrightarrow{P} \text{fdp}^{\text{kno}}(t) + R(t). \quad (33)
\end{aligned}$$

For the *Lasso Coefficient Difference* statistic, the remainder term is obtained as follows

$$R^{\text{lcd}}(t) = \frac{s\cdot P(|\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}(\tilde{\sigma}\tilde{\tau}\cdot(\tilde{\theta}B' + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z))| - |\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}(\tilde{\sigma}\tilde{\tau}\cdot\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z')| \leq -t)}{P(|\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}(\tilde{\sigma}\tilde{\tau}\cdot(\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z))| - |\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}(\tilde{\sigma}\tilde{\tau}\cdot\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z')| \geq t)}, \quad (34)$$

which is expected to be small and becomes negligible when increasing the magnitude of $B'$ (the limiting representation of the nonnull $\beta_j^*$'s). In addition, Figure 7 shows that the estimator $\widehat{\text{fdp}}$ is quite accurate for small fdp (large $t$); this is also reflected later in Figure 8.

### 3.3. Accuracy-error tradeoff curve based on knockoff selection

Similar to Section 2.3, Corollary 3.2 validates that the $\text{TPP}^{\text{kno}}$ and $\text{FDP}^{\text{kno}}$ in (30), (31) still satisfy a tradeoff curve when the selection is by knockoffs using the *Lasso Coefficient Difference* statistic in (22). Figure 8 compares the limiting tradeoff curves $q^{B'}(\text{TPP}^{\text{est}}(\lambda); s, \delta, \kappa)$ for knockoff selection and for selection by $\widehat{\beta}$. We consider two settings where $\delta = 1, 2$, $s = 0.2$, $\kappa^2 = 4$. While performing selection by knockoffs, the fixed point solutions of the six parameters $(\alpha, \sigma, \gamma, \theta, \tau, r)$ are obtained using the adjusted parameters $\widetilde{\delta} = \delta/2$, $\widetilde{s} = s/2$, $\widetilde{\kappa}^2 = \kappa^2/2$. From Figure 8, we observe the loss on TPP for knockoff selection in contrast to selection by the regularized estimator $\widehat{\beta}$ given large values of FDP.

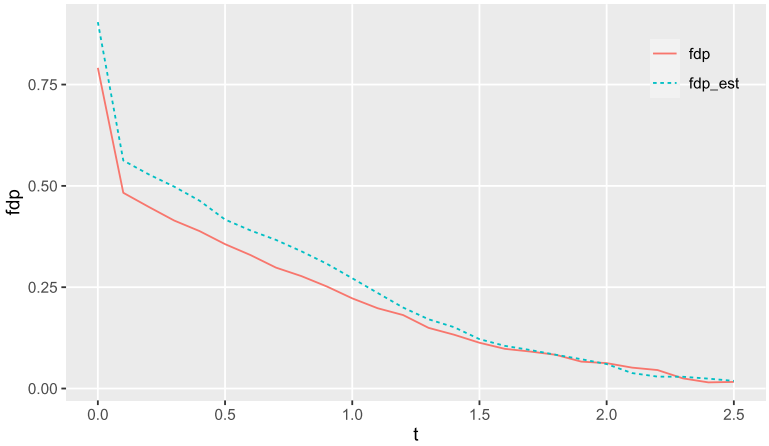


FIG 7. *The red solid curve refers to* $\widehat{fdp}(t)$ *by* (32) *and the turquoise dashed curve refers to* $fdp$ *by* (32). *With a proper choice of* $t$, $\widehat{fdp}(t)$ *is an accurate estimator of* $fdp$.

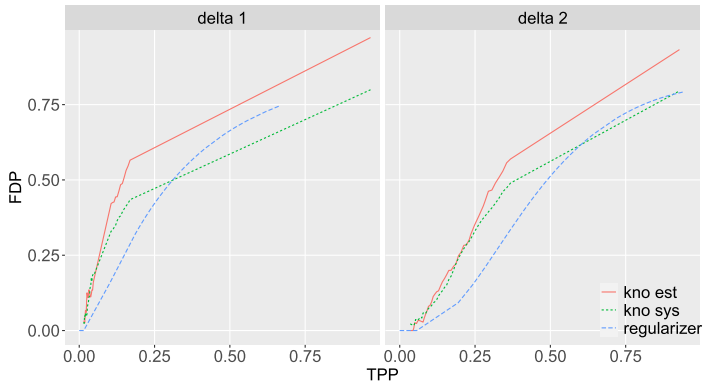


FIG 8. *The limiting tradeoff curves* $q^{B'}(tpp(t); s, \delta, \kappa^2) = fdp(t)$ *for knockoff selection and for selection by* $\widehat{\beta}$. *The red solid curve (legend "kno est") depicts* $q^{B'}(tpp^{\mathrm{kno}}(t); s, \delta, \kappa^2) = \widehat{fdp}^{\mathrm{kno}}(t)$ *using* $\widehat{fdp}^{\mathrm{kno}}(t)$ *in* (32); *the green dotted curve (legend "kno sys") uses* $fdp^{\mathrm{kno}}(t)$ *in* (30). *The blue dashed curve (legend "regularization") corresponds to* $q^{B'}(tpp(\lambda); s, \delta, \kappa^2) = fdp(\lambda)$ *using* (16), (17). *Two settings where* $\delta = 1, 2$, $s = 0.2$, $\kappa^2 = 4$ *are considered.*

However, for a smaller ratio $\delta$, the TPP for knockoff selection is close to selection by $\widehat{\beta}$ for a small threshold of FDP, which is realistic in practice since we often predetermine the FDR level to be 0.05 or 0.1. Notice that constructing the knockoffs in this paper already eliminates the correlation between not only the knockoffs $X' = (X'_{\cdot 1}, \ldots, X'_{\cdot p})$'s and the $X_{\cdot j}$'s, but also the correlation between the $\widetilde{X}_{\cdot j}$'s. Hence, the loss of TPP for selection by knockoff is not caused by the generation of knockoffs, i.e. in [11], the TPP (power) is affected by the correlation between $X'$ and $X$. A similar TPP loss is also observed in [50] for

the $\ell_1$-regularized least-squares estimator, but appears more optimistic with less TPP loss severity.

Similar to Sections 2.3.1 and 2.3.2, the limiting tradeoff curves for selection by knockoff denoted by $q^{B'}(\text{tpp}(t); s, \delta, \kappa^2) = \text{fdp}(t)$, can be incorporated in practice to calculate the sample size or adjust the FDR level. The advantage of using the tradeoff curve for knockoff selection with the *Lasso Coefficient Difference* statistic is that the FDP and TPP are dominated by the threshold $t$ in (23). More importantly, the FDP is estimable for the knockoff selection. The selection is much more robust compared to selection by the estimated nonnull coefficients of $\widehat{\beta}$.

## 4. Wisconsin breast cancer data

We consider the Breast Cancer Wisconsin (diagnostic) dataset on the UCI machine learning repository. The dataset consists of 569 samples, among which 357 are diagnosed as benign and 212 are malignant. The binary response (benign and malignant) is correlated with 30 continuous predictive variables, which are metrics for cell nuclei such as `radius`, `texture`, `perimeter`. Using 30 continuous predictive variables, we construct the corresponding pairwise interactions of all 30 predictive variables, resulting in an expanded dataset with 465 predictive variables and 569 samples. To comply with Assumption (A1) for obtaining the solutions to the six parameters $(\alpha, \sigma, \gamma, \theta, \tau, r)$, we consider incorporating a singular value decomposition

$$X = UDV^\top = \sum_{i=1}^{\min\{n,p\}} d_i \mathbf{u}_i \mathbf{v}_i^\top.$$

The columns of $U \in \mathbb{R}^{n\times\min\{n,p\}}$ are singular vectors of the samples, and the rows of $V^\top \in \mathbb{R}^{p\times\min\{n,p\}}$ are singular vectors of the predictive variables. For further analysis, $X$ is replaced by $UV^\top$, such that the linear predictor $X\beta = UV^\top = UV^\top\breve{\beta}$ where $\breve{\beta} = \text{diag}(D, \mathbf{1}_{\max\{n,p\}-\min\{n,p\}})\beta$. This suggests that the problem changes from estimating the regression coefficient vector $\beta$ to estimating $\breve{\beta}$, which incorporates the eigenvalues of $X$.

Next, we consider a simple initial $\ell_1$-regularized logistic regression estimate obtained by the R package `glmnet` with the tuning parameter chosen by 5-fold cross-validation in order to: (1) obtain the inputs $\delta$, $\kappa^2$ for finding the fixed point solutions $(\alpha, \sigma, \gamma, \theta, \tau, r)$; (2) estimate the distribution of $B$ which is the limiting random variable describing the components of $\beta$ (see Assumption (A2)). The nonnull components $B'$ are modeled by a Gaussian distribution $N(0, \kappa^2/s)$, where $s$ is estimated by the number of nonnull components of the initial estimate. This initial estimate can be largely improved by more advanced numerical estimation, especially more adaptive methods to estimate $\kappa^2$ and the distribution of $B$, which is not in the scope of this paper and is worth further investigation.

We elaborated in Section 2.3.2 about guidance for a proper choice of the FDP level. Figure 9 shows the tradeoff curves for selection by $\widehat{\beta}$ and by knockoffs.

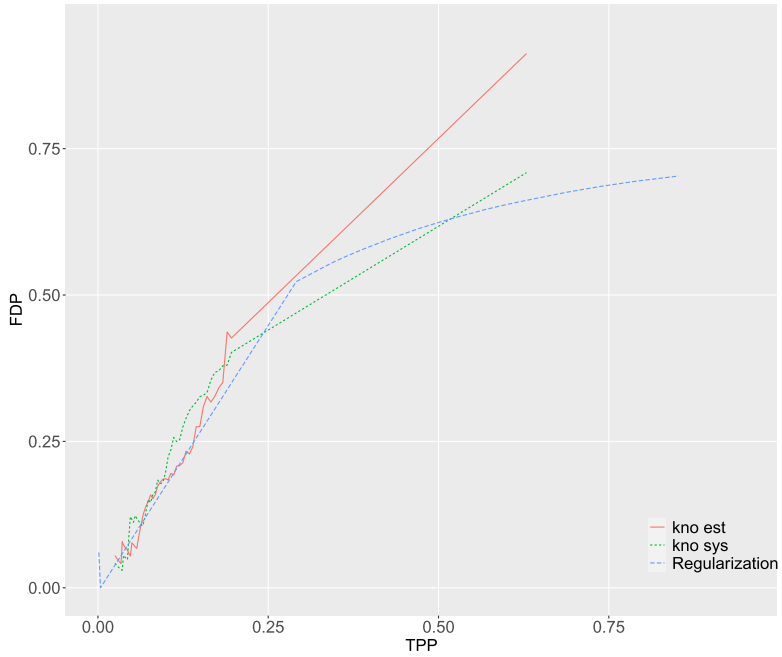


FIG 9. *The limiting tradeoff curves* $q^{B'}(tpp; s, \delta, \kappa^2) = fdp$ *for the* `Wisconsin Breast Cancer dataset`. *The red solid curve (legend "kno est") depicts* $q^{B'}(tpp^{\text{kno}}(t); s, \delta, \kappa^2) = \widehat{fdp}^{\text{kno}}(t)$ *using* $\widehat{fdp}^{\text{kno}}(t)$ *in* (32)*; the green dotted curve (legend "kno sys") uses* $fdp^{\text{kno}}(t)$ *in* (30). *The blue dashed curve (legend "regularization") corresponds to* $q^{B'}(tpp(\lambda); s, \delta, \kappa^2) = fdp(\lambda)$ *using* (16), (17).

Often in practice, the FDP level $q$ is set to 0.05 or 0.1; however, the corresponding TPP is quite low for this dataset with $\delta \approx 1.22$. Figure 9 indicates that with a high probability, we would select less than 5% among all true positive variables for FDP level $q = 0.05$. Based on Figures 4, 6, this steep tradeoff curve is possibly due to $n > p$ and medium level $s$. If the selection objective is to identify the top 10 relevant predictive variables and is sensitive to false discoveries, setting the FDP level $q$ at 5% could be acceptable. However, if the objective is to identify at least 10% of the relevant predictive variables, setting the FDP level at 5% would hardly give a satisfying result; in fact, we have to set the FDP level at 0.25 to achieve the desired TPP.

Since the software in the R package `glmnet` standardizes the predictive variables, there is a mismatch between the $\lambda$'s obtained from `glmnet` and (3). Due to the mismatch, plugging in the value of $\lambda$ obtained from choosing the threshold $q$ and obtaining an estimator becomes infeasible. However, we observe that the curves by the two selection methods are close, which suggests that selection by knockoffs achieves a similar selection TPP as selection by regularization for a given FDP. Thus, we report selection results by knockoffs in the two scenar-

ios mentioned above. In the case where the FDP level is set to be $q = 0.05$, only 11 interaction terms based on 6 main effects `perimeter_mean`, `area_mean`, `area_worst concave.points_se`, `texture_worst`, `perimeter_worst` are selected. However, when raising the FDP level up to $q = 0.25$, among total 45 selected variables, the 6 main effects mentioned above are all selected.

## 5. Discussion

This paper aims at giving insight into aspects of variable selection for regularized logistic regressions in both $n > p$ and $n \leq p$ settings, where $\ell_1$-regularization is investigated in detail. Traditionally, the literature discussing variable selection properties addresses $n \leq p$ with certain sparsity assumptions. Since regularization is also often used in practice where $n > p$ and the sparsity assumptions are not always satisfied, we believe the technical methods used in this paper bring a new perspective for investigating the regularized estimators for future research.

Relevant to the sparsity assumptions, we observe in Figure 4 that a high sparsity for $p \geq n$ in general leads to a better selection TPP when the targeted FDP is low, and the selection TPP for medium sparsity generally has trouble. Unsurprisingly, Figure 6 suggests that a large sample size is, in general, helpful in achieving a higher TPP for a given FDP.

A critique we noticed is that the numerical optimization of the system of equations (6) is computationally demanding and sometimes returns incorrect solutions. Moreover, the solution does not always exist. For example, for $n \leq p$ when $\lambda$ is close to zero, numerical optimization cannot find the solutions to the system of equations. This is realistic in practice since, without regularization, finding the MLE for logistic regression when $p \geq n$ is impossible to our best knowledge. However, this could leave people with the impression that we are "cherry-picking". It is of future research interest to know when the solutions to the six equations exist; this is also a big topic in [42] where the system of equations has no solution when the MLE does not exist. The computation in this paper relies on R and high-performance computing, which is not optimal for practical use. We hope to improve this by developing better algorithms and numerical optimization in other programming languages.

Further, for practical use (see the example application for the `Wisconsin Breast Cancer` dataset), this method requires an initial estimation of the true signal $B$, which appears to be unrealistic at first glance. However, since FDP and TPP expressions also require information on the true nulls and nonnulls, using the estimated signal $B$ to replace the true one is unavoidable for the current paper. In the current stage, we suggest using bootstrap to obtain a more accurate initial estimate of the signal. Deriving estimators for FDP and TPP, as well as a new version of the system of equations, such that these equations do not rely on the true signal $B$ is of future research interest.

In addition, the framework in this paper can also be extended to other loss functions such as used in generalized linear models and robust estimators with a general Gaussian design matrix. Alternatively, the tradeoff curves can be used to

choose the tuning parameter $\lambda$ based on the targeted FDP and TPP. Since the tuning parameter $\lambda$ determines the selection properties for $l_1$-regularized logistic regression, based on the FDP and TPP to be achieved, one could use this to decide if the tuning parameter selected by the available software is suitable.

## Appendix A

### *A.1. Technicalities*

#### *A.1.1. Proof of Proposition 2.2*

*Proof.* By taking the function $\Psi(x, y) = (x - y)^2$ in Lemma 2.1, the following convergence holds, as $n \to \infty$,

$$\mathrm{MSE}(\widehat{\beta}) = \frac{1}{p}\sum_{j=1}^{p}\left(\widehat{\beta}_j - \beta_j^*\right)^2 \xrightarrow{P} E\left[\left(\eta_{\lambda\sigma\tau}\left(\sigma\tau\cdot\left(\theta B + \frac{r}{\sqrt{\delta}}Z\right)\right) - B\right)^2\right]. \quad (35)$$

By simple polynomial expansion, the RHS of (35) can be written as the sum of three expectations which also satisfy the system of equations (6),

$$\begin{aligned}
&E\left[\left(\eta_{\lambda\sigma\tau}\left(\sigma\tau\cdot\left(\theta B + \frac{r}{\sqrt{\delta}}Z\right)\right) - B\right)^2\right] \\
&\quad = E\left[\left(\eta_{\lambda\sigma\tau}\left(\sigma\tau\cdot\left(\theta B + \frac{r}{\sqrt{\delta}}Z\right)\right)\right)^2\right] + E\left[B^2\right] \\
&\qquad - 2E\left[B\cdot\eta_{\lambda\sigma\tau}\left(\sigma\tau\cdot\left(\theta B + \frac{r}{\sqrt{\delta}}Z\right)\right)\right] \\
&\quad = \kappa^2\alpha^2 + \sigma^2 + \kappa^2 - 2\kappa^2\alpha \\
&\quad = \kappa^2(\alpha - 1)^2 + \sigma^2.
\end{aligned}$$

The second equality holds by the first and third expressions in (6), and the assumption on the averaged signal strength. □

#### *A.1.2. Proof of Lemma 2.4*

*Proof.* To obtain the four limiting expressions, we apply Lemma 2.1 by taking the locally Lipschitz function $\Psi$ to be indicator functions. We demonstrate the idea by first presenting the proof of (14). By definition, the number of false discoveries is defined as $V^{\mathrm{est}}(\lambda) = \#\{j \in \mathcal{H} : \widehat{\beta}_j(\lambda) \neq 0, \beta_j = 0\}$, which can further be written as $\sum_{j=1}^{p} I\{\widehat{\beta}_j(\lambda) \neq 0, \beta_j = 0\}$. The indicator function $I\{\widehat{\beta}_j(\lambda) \neq 0, \beta_j = 0\}$ on $\mathbb{R}^2$ is locally Lipschitz since the indicator function is either equal to 0 or 1, the absolute value of the difference of the indicator

function at any two points cannot exceed 1, which can be easily bounded by a sufficiently large Lipschitz constant. Then,

$$\begin{aligned}
\frac{V^{\text{est}}(\lambda)}{p} &= \frac{1}{p}\sum_{j=1}^{p} I\{\widehat{\beta}_j(\lambda)\neq 0, \beta_j = 0\} \\
&\xrightarrow{P} EI\left\{\eta_{\lambda\sigma\tau}(\sigma\tau\cdot\left(\theta B+\frac{r}{\sqrt{\delta}}Z\right)\neq 0, B=0\right\} \\
&= P\left(\eta_{\lambda\sigma\tau}\left(\sigma\tau\cdot\left(\theta B+\frac{r}{\sqrt{\delta}}Z\right)\right)\neq 0, B=0\right) \\
&= P\left(\eta_{\lambda\sigma\tau}\left(\sigma\tau\cdot\left(\theta B+\frac{r}{\sqrt{\delta}}Z\right)\right)\neq 0 \mid B=0\right)\cdot P(B=0) \\
&= P\left(\eta_{\lambda\sigma\tau}\left(\sigma\tau\cdot\left(\frac{r}{\sqrt{\delta}}Z\right)\right)\neq 0\right)\cdot P(B=0) \\
&= P\left(|\sigma\tau\cdot\frac{r}{\sqrt{\delta}}Z|>\lambda\sigma\tau\right)\cdot(1-s) \\
&= P\left(|Z|>\frac{\lambda}{\frac{r}{\sqrt{\delta}}}\right)\cdot(1-s) = 2(1-s)\cdot\Phi\left(-\frac{\lambda}{\frac{r}{\sqrt{\delta}}}\right).
\end{aligned}$$

The fifth equality holds by the definition of soft-thresholding function; the last equality holds since $Z\sim N(0,1)$. The limiting expressions of the other three proportions are obtained similarly. □

*A.1.3. Proof of Theorem 3.1*

*Proof.*

$$\begin{aligned}
&\frac{1}{p}\sum_{j=1}^{p}\widetilde{\Psi}\big(\widehat{\beta}_j,\beta_j^*,\widehat{\beta}_{j+p}\big) \\
&\quad= \frac{1}{p}\sum_{j=1}^{p}\big\{\Psi\big(\widehat{\beta}_j,\beta_j^*\big)+\Psi(\widehat{\beta}_{j+p},0)\big\} = 2\cdot\frac{1}{\tilde{p}}\sum_{j=1}^{\tilde{p}}\Psi\big(\widehat{\beta}_j,\tilde{\beta}_j^*\big) \\
&\quad\xrightarrow{P} 2E\Psi\left(\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\left(\tilde{\theta}\tilde{B}+\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right)\right),\tilde{B}\right) \\
&\quad= 2\Bigg\{E\left[\Psi\left(\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\left(\tilde{\theta}\tilde{B}+\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right)\right),\tilde{B}\right)\mid\tilde{B}=B\right]\cdot P(\tilde{B}=B) \\
&\qquad+E\left[\Psi\left(\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\left(\tilde{\theta}\tilde{B}+\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right)\right),\tilde{B}\right)\mid\tilde{B}\neq B\right]\cdot P(\tilde{B}\neq B)\Bigg\} \\
&\quad= 2\cdot\left\{\frac{1}{2}E\Psi\left(\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\left(\tilde{\theta}B+\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right)\right),B\right)+\frac{1}{2}E\Psi\left(\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\left(\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'\right)\right),\tilde{B}\right)\right\} \\
&\quad= E\widetilde{\Psi}\left(\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\left(\tilde{\theta}B+\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right)\right),B,\eta_{\tilde{\lambda}\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\left(\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'\right)\right)\right).
\end{aligned}$$

The first equality holds by $\tilde{p} = 2p$; the convergence in probability holds by Lemma 2.1. The third equality holds since the true parameters for knockoffs are zero, and integration over $Z$ is equivalent to over $Z'$ since $Z, Z'$ are i.i.d. □

### *A.2. Limiting expressions of FDP and TPP for* Lasso Signed Max *statistic*

To derive similar limiting FDP and TPP expressions, we consider the numerator and denominator in (19) separately. We divide the numerator by the number of the hypotheses $\mathcal{H}$.

$$\frac{\#\{j \in \mathcal{H} : W_j \geq t, j \in \mathcal{H}_0\}}{p} = \frac{(1-s) \cdot \#\{j \in \mathcal{H}_0 : W_j \geq t\}}{p}$$
$$= (1-s)\left\{\frac{\#\{j \in \mathcal{H}_0 : Z_j \geq t, Z_j \geq \widetilde{Z}_j\}}{p} + \frac{\#\{j \in \mathcal{H}_0 : -\widetilde{Z}_j \geq t, \widetilde{Z}_j \geq Z_j\}}{p}\right\}.$$

For any $t > 0$, the second term $\#\{j \in \mathcal{H}_0 : -\widetilde{Z}_j \geq t, \widetilde{Z}_j \geq Z_j\}/p$ is equal to zero, since $\widetilde{Z}_j = \sup\{\lambda : \widehat{\beta}_{j+p}(\lambda) \neq 0\}$ is nonnegative hence $-\widetilde{Z}_j \geq t$ does not hold for $t > 0$. Further, for the first term, we argue that for any fixed $t > 0$, $Z_j \geq t$ is equivalent to stating that $\widehat{\beta}_j(t) \neq 0$ stays in the path. The main challenge in this proof is developing an alternative representation for the event $Z_j \geq \widetilde{Z}_j$, taking into account the unknown relationship between $\widetilde{Z}_j$ and $t$. Recall that $Z_j = \sup\{\lambda : \widehat{\beta}_j(\lambda) \neq 0\}$, notice that $Z_j \geq \widetilde{Z}_j$ indicates $\widetilde{Z}_j = Z_j - \varepsilon_{\widetilde{Z}_j}$ where $\varepsilon_{\widetilde{Z}_j} > 0$ can be infinitely small. This implies that $\widehat{\beta}_j(\widetilde{Z}_j) \neq 0$, since $Z_j$ is the smallest tuning parameter to guarantee $\widehat{\beta}_j$ leave the path but $Z_j > \tilde{Z}_j$. Similarly, $\widehat{\beta}_{j+p}(\widetilde{Z}_j) = 0$ holds. Following the reasoning above,

$$\frac{\#\{j \in \mathcal{H}_0 : Z_j \geq t, Z_j \geq \widetilde{Z}_j\}}{p}$$
$$= \frac{\#\{j \in \mathcal{H}_0 : \widehat{\beta}_j(t) \neq 0, \widehat{\beta}_j(\widetilde{Z}_j) \neq 0, \widehat{\beta}_{j+p}(\widetilde{Z}_j) = 0\}}{p}$$
$$\xrightarrow{P} P\left(\eta_{t\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau} \cdot \left(\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right)\right) \neq 0; \text{ there exists } t', \text{ such that}\right.$$
$$\left.\eta_{t'\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau} \cdot \left(\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right)\right) \neq 0, \eta_{t'\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau} \cdot \left(\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'\right)\right) = 0 \,|B = 0\right)$$
$$= P\left(\left|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right| > t, \left|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right| > t' \geq \left|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'\right|\right)$$
$$= P\left(\left|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right| > t, \left|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right| > \left|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'\right|\right).$$

We collect all pieces and obtain the numerator of (19) divided by $p$ as follows

$$\frac{\#\{j \in \mathcal{H} : W_j \geq t, j \in \mathcal{H}_0\}}{p} = \frac{(1-s) \cdot \#\{j \in \mathcal{H}_0 : W_j \geq t\}}{p}$$

$$\xrightarrow{P} (1-s)\cdot P\left(\left|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right| > t, \left|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right| > \left|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'\right|\right). \tag{36}$$

Similarly, for the denominator of (19) divided by $p$ it follows that

$$\begin{aligned}\frac{\#\{j\in\mathcal{H}: W_j \geq t\}}{p} &= \frac{\#\{j\in\mathcal{H}: Z_j \geq t, Z_j \geq \widetilde{Z}_j\}}{p} \\ &\xrightarrow{P} P\left(\left|\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right| > t, \left|\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right| > \left|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'\right|\right). \end{aligned} \tag{37}$$

Replacing the numerator and denominator in the general FDP expression (19) by (36) and (37), and letting $p_n \to \infty$, we obtain the limiting expression $\mathrm{fdp}(t)$,

$$\mathrm{fdp}^{\mathrm{lsm}}(t) = \frac{(1-s)\cdot P(|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z| > t, |\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z| > |\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'|)}{P(|\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z| > t, |\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z| > |\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'|)}. \tag{38}$$

Similarly, the limiting tpp expression can be obtained as

$$\begin{aligned}\mathrm{tpp}^{\mathrm{kno}}(t) &= \lim_{p_n\to\infty} \frac{\#\{j\in\mathcal{H}: W_j \geq t, j\notin\mathcal{H}_0\}}{\#\{j\in\mathcal{H}: j\notin\mathcal{H}_0\}} \\ &= P\left(\eta_{t\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\cdot\left(\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right)\right) \neq 0;\ \text{there exists } t', \text{such that}\right. \\ &\quad \left.\eta_{t'\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\cdot\left(\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right)\right) \neq 0, \eta_{t'\tilde{\sigma}\tilde{\tau}}\left(\tilde{\sigma}\tilde{\tau}\cdot\left(\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'\right)\right) = 0 \mid B\neq 0\right) \\ &= P\left(\left|\tilde{\theta}B' + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right| > t, \left|\tilde{\theta}B' + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z\right| \geq \left|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'\right|\right). \end{aligned} \tag{39}$$

And the limiting expression of the fdp estimator is

$$\begin{aligned}\widehat{\mathrm{fdp}}^{\mathrm{kno}}(t) &= \lim_{p\to\infty} \frac{\#\{j: -\widetilde{Z}_j \leq -t, \widetilde{Z}_j \geq Z_j\}/p}{\#\{j: Z_j \geq t, Z_j \geq \widetilde{Z}_j\}/p} \\ &= \frac{P(|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'| > t, |\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z| \leq |\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'|)}{P(|\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z| > t, |\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z| \geq |\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'|)}. \end{aligned} \tag{40}$$

Recall that the $\widehat{\mathrm{fdp}}^{\mathrm{kno}}$ is upward biased for $\mathrm{fdp}^{\mathrm{kno}}$. Similar to (34) for the *Lasso Coefficient difference* statistic, the remainder term, bias of the estimator $\widehat{\mathrm{fdp}}^{\mathrm{kno}}$ for the *Lasso Signed Max* statistic, is given by

$$R^{\mathrm{lsm}}(t) = \frac{s\cdot P(|\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'| > t, |\tilde{\theta}B' + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z| \leq |\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'|)}{P(|\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z| > t, |\tilde{\theta}B + \frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z| \geq |\frac{\tilde{r}}{\sqrt{\tilde{\delta}}}Z'|)}. \tag{41}$$

## Acknowledgments

The authors thank the editor, associate editor, and reviewers for their useful comments, which helped improve the paper.

## Disclosure

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by a Postdoc Fellowship of the Research Foundation Flanders and KU Leuven internal fund C16/20/002. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation-Flanders (FWO) and the Flemish Government.

## References

[1] Abbasi, E. (2020). Universality Laws and Performance Analysis of the Generalized Linear Models, PhD thesis, California Institute of Technology. MR4639752

[2] Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43** 2055–2085. MR3375876

[3] Bayati, M. and Montanari, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory* **57** 764–785. MR2810285

[4] Bayati, M. and Montanari, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory* **58** 1997–2017. MR2951312

[5] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57** 289–300. MR1325392

[6] Blanchard, G. and Roquain, É. (2009). Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research* **10**. MR2579914

[7] Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W. and Candès, E. J. (2015). SLOPE—adaptive variable selection via convex optimization. *The Annals of Applied Statistics* **9** 1103. MR3418717

[8] Bradic, J. (2016). Robustness in sparse high-dimensional linear models: Relative efficiency and robust approximate message passing. *Electronic Journal of Statistics* **10** 3894–3944. MR3581957

[9] Bu, Z., Klusowski, J., Rush, C. and Su, W. (2019). Algorithmic analysis and statistical estimation of slope via approximate message passing. *Advances in Neural Information Processing Systems* **32** 9366–9376. MR4231969

[10] Cai, Z., Li, R. and Zhang, Y. (2022). A distribution free conditional independence test with applications to causal discovery. *Journal of Machine Learning Research* **23** 1–41. MR4576670

[11] 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 (Statistical Methodology)* **80** 551–577. MR3798878

[12] Celentano, M. and Montanari, A. (2021). CAD: Debiasing the Lasso with inaccurate covariate model. *arXiv preprint arXiv:2107.14172*.

[13] Celentano, M. and Montanari, A. (2022). Fundamental barriers to high-dimensional regression with convex penalties. *The Annals of Statistics* **50** 170–196. MR4382013

[14] Donoho, D. and Montanari, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields* **166** 935–969. MR3568043

[15] Donoho, D. L., Maleki, A. and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences* **106** 18914–18919. MR4158199

[16] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

[17] 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. MR2530322

[18] Fan, J., Ma, Y. and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association* **109** 1270–1284. MR3265696

[19] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32** 928–961. MR2065194

[20] Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38** 3567–3604. MR2766861

[21] Fan, J., Xue, L. and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics* **42** 819. MR3210988

[22] Fan, Z. (2022). Approximate Message Passing algorithms for rotationally invariant matrices. *The Annals of Statistics* **50** 197–224. MR4382014

[23] Farcomeni, A. (2006). More powerful control of the false discovery rate under dependence. *Statistical Methods and Applications* **15** 43–73. MR2281214

[24] Feng, O. Y., Venkataramanan, R., Rush, C. and Samworth, R. J. (2022). A unifying tutorial on approximate message passing. *Foundations and Trends® in Machine Learning* **15** 335–536.

[25] Fithian, W. and Lei, L. (2020). Conditional calibration for false discovery rate control under dependence. *arXiv preprint arXiv:2007.10438*. MR4524490

[26] Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *The Annals of Statistics* **32** 1035–1061. MR2065197

[27] Gordon, Y. (1985). Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics* **50** 265–289. MR0800188
[28] Gordon, Y. (1988). On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric Aspects of Functional Analysis* 84–106. Springer. MR0950977
[29] He, X., Wang, L. and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics* **41** 342–369. MR3059421
[30] Janson, L. and Su, W. (2016). Familywise error rate control via knockoffs. *Electronic Journal of Statistics* **10** 960–975. MR3486422
[31] Kelner, J. A., Koehler, F., Meka, R. and Rohatgi, D. (2022). On the power of preconditioning in sparse linear regression. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)* 550–561. IEEE. MR4399714
[32] Lee, J. D., Sun, Y. and Taylor, J. E. (2015). On model selection consistency of regularized M-estimators. *Electronic Journal of Statistics* **9** 608–642. MR3331852
[33] Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107** 1129–1139. MR3010900
[34] Liu, W., Ke, Y., Liu, J. and Li, R. (2022). Model-free feature screening and FDR control with knockoff features. *Journal of the American Statistical Association* **117** 428–443. MR4399096
[35] Mai, Q. and Zou, H. (2015). The fused Kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics* **43** 1471–1497. MR3357868
[36] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **3** 1436-1462. https://doi.org/10.1214/009053606000000281. MR2278363
[37] Pan, W., Wang, X., Xiao, W. and Zhu, H. (2018). A generic sure independence screening procedure. *Journal of the American Statistical Association*. MR3963192
[38] Rangan, S., Schniter, P., Fletcher, A. K. and Sarkar, S. (2019). On the convergence of approximate message passing with arbitrary matrices. *IEEE Transactions on Information Theory* **65** 5339–5351. MR4009237
[39] Rangan, S., Schniter, P., Riegler, E., Fletcher, A. K. and Cevher, V. (2016). Fixed points of generalized approximate message passing with arbitrary matrices. *IEEE Transactions on Information Theory* **62** 7464–7474. MR3599094
[40] Salehi, F., Abbasi, E. and Hassibi, B. (2019). The impact of regularization on high-dimensional logistic regression. *Advances in Neural Information Processing Systems* **32**.
[41] Su, W., Bogdan, M. and Candes, E. (2017). False discoveries occur early on the lasso path. *The Annals of Statistics* **45** 2133–2150. MR3718164
[42] 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. MR3984492
[43] Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression, PhD thesis, Stanford University. MR4197622
[44] 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. MR4009715
[45] Thrampoulidis, C., Abbasi, E. and Hassibi, B. (2018). Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory* **64** 5592–5628. MR3832326
[46] Thrampoulidis, C. and Hassibi, B. (2015). Isotropically random orthogonal matrices: Performance of lasso and minimum conic singular values. In *2015 IEEE International Symposium on Information Theory (ISIT)* 556–560. IEEE.
[47] Tong, Z., Cai, Z., Yang, S. and Li, R. (2022). Model-free conditional feature screening with FDR control. *Journal of the American Statistical Association* 1–13. MR4681605
[48] Wang, S., Weng, H. and Maleki, A. (2020). Which bridge estimator is the best for variable selection? *The Annals of Statistics* **48** 2791–2823. https://doi.org/10.1214/19-AOS1906. MR4152121
[49] Weinstein, A., Barber, R. and Candes, E. (2017). A power and prediction analysis for knockoffs with lasso statistics. *arXiv preprint arXiv: 1712.06465*.
[50] Weinstein, A., Su, W. J., Bogdan, M., Barber, R. F. and Candès, E. J. (2020). A power analysis for knockoffs with the lasso coefficient-difference statistic. *arXiv preprint arXiv:2007.15346*.
[51] Wu, Y. and Yin, G. (2015). Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika* **102** 65–76. MR3335096
[52] Xu, J., Maleki, A., Rad, K. R. and Hsu, D. (2021). Consistent risk estimation in moderately high-dimensional linear regression. *IEEE Transactions on Information Theory* **67** 5997–6030. MR4345048
[53] Yang, G., Yu, Y., Li, R. and Buu, A. (2016). Feature screening in ultrahigh dimensional Cox's model. *Statistica Sinica* **26** 881. MR3559935
[54] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894-942. https://doi.org/10.1214/09-AOS729. MR2604701
[55] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research* **7** 2541–2563. MR2274449
[56] Zhao, Q., Sur, P. and Candès, E. J. (2023). The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance. *Bernoulli* **28**. MR4411513
[57] Zhou, J., Claeskens, G. and Bradic, J. (2020). Detangling robustness in high dimensions: composite versus model-averaged estimation. *Electronic Journal of Statistics* **14** 2551–2599. MR4122516
[58] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469