Submitted to the Annals of Statistics

USING SYNTHETIC DATA TO REGULARIZE MAXIMUM LIKELIHOOD ESTIMATION

BY WEIHAO LI[1,a] AND DONGMING HUANG[2,b]

[1]*Department of Statistics & Data Science, National University of Singapore,* [a]*weihao.li@u.nus.edu*

[2]*Department of Statistics & Data Science, National University of Singapore,* [b]*stahd@nus.edu.sg*

To overcome challenges in fitting complex models with small samples, catalytic priors have recently been proposed to stabilize the inference by supplementing observed data with synthetic data generated from simpler models. Based on a catalytic prior, the Maximum A Posteriori (MAP) estimator is a regularized estimator that maximizes the weighted likelihood of the combined data. This estimator is straightforward to compute, and its numerical performance is superior or comparable to other likelihood-based estimators. In this paper, we study several theoretical aspects regarding the MAP estimator in generalized linear models, with a particular focus on logistic regression. We first prove that under mild conditions, the MAP estimator exists and is stable against the randomness in synthetic data. We then establish the consistency of the MAP estimator when the dimension of covariates diverges slower than the sample size. Furthermore, we utilize the convex Gaussian min-max theorem to characterize the asymptotic behavior of the MAP estimator as the dimension grows linearly with the sample size. These theoretical results clarify the role of the tuning parameters in a catalytic prior, and provide insights in practical applications. We provide numerical studies to confirm the effective approximation of our asymptotic theory in finite samples and to illustrate adjusting inference based on the theory.

1. Introduction. In statistical modeling, using auxiliary samples—such as pseudo data, historical data, data from related studies, or synthetic data—can often significantly improve the inference. Traditional Bayesian perspective interprets conjugate priors[1] for exponential families as supplements to the observed data with some "prior data" (Birnbaum, 1962; Pratt, Raiffa and Schlaifer, 1964; Dempster, 1968). In modern scientific research, there often exist massive datasets different but related to the data to be analyzed, and one of the important goals is to leverage these datasets for improved predictions and inferences. For example, Chen, Ibrahim and Shao (2000) proposed the *power prior* distributions for Bayesian models to incorporate historical data, and Li, Cai and Li (2022) investigated the integration of gene expression datasets measured in different issues to understand the gene regulations for a specific tissue.

Recently, Huang et al. (2020) considered using synthetic data generated from a fitted simpler model to construct the *catalytic prior* for a model that is challenging to stably fit due to the limited sample size. The class of catalytic priors has broad applicability and offers straightforward interpretation via synthetic data. When using a catalytic prior, the resulting posterior distribution can be easily formulated and computing the Maximum A Posteriori (MAP) estimate is as simple as computing a maximum weighted likelihood estimate based on the observed data and the ancillary data. This class of priors has been applied to generalized linear models (GLMs) and Cox proportional hazard models (Huang et al., 2020;

MSC2020 subject classifications: Primary 00X00, 00X00; secondary 00X00.

Keywords and phrases: synthetic data, logistic regression, exact asymptotics, regularization.

[1]Throughout the paper, we use the compact word "priors" in place of the term "prior distributions".

Li and Huang, 2023), and empirical results based on simulation studies and applications to real-world datasets suggest that the estimation accuracy and predictive performance of the resulting inference are superior or comparable to those of traditional priors (Huang et al., 2022).

Despite the aforementioned empirical exploration of the catalytic prior method, there is little theoretical investigation, especially regarding the frequentist properties of the resulting inference. This paper aims to bridge this gap through a comprehensive theoretical analysis of the resulting estimations derived from these priors. We first develop several theoretical results for logistic regression, a special GLM for binary regression, and we then extend these results to some other GLMs.

1.1. *MAP estimation under catalytic prior.* For a parametric model with parameter $\boldsymbol{\theta}$, suppose the likelihood function is $L(\boldsymbol{\theta};\mathcal{D})$ where $\mathcal{D}$ denotes the observed dataset. The catalytic prior is based on a synthetic dataset $\mathcal{D}^*$ and is formulated as a weighted likelihood based on $\mathcal{D}^*$, i.e., $\pi(\boldsymbol{\theta}) = L(\boldsymbol{\theta};\mathcal{D}^*)^{\frac{\tau}{M}}$ where τ is a positive tuning parameter that for down-weighting the impact of synthetic data and M is the size of $\mathcal{D}^*$. The idea of catalytic priors is based on the *data-centric perspective*, which emphasizes that data are real while models are human constructs for analyzing data. The Maximum A Posteriori (MAP) estimator (i.e. the mode of this posterior density) derived from the catalytic prior can be expressed as

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\theta} \left[L(\boldsymbol{\theta};\mathcal{D}) L(\boldsymbol{\theta};\mathcal{D}^*)^{\frac{\tau}{M}} \right]. \tag{1}$$

This estimator is a regularized maximum likelihood estimator whose regularization is formulated using the likelihood based on the synthetic dataset. Compared to other regularization such as L_q-norm, an advantage of this regularization via synthetic data is its invariance to the affine group of the parameter space.

Consider logistic regression models for example. Let $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ be n independent pairs of observed data, where $Y_i \in \{0,1\}$ is a response and $\boldsymbol{X}_i$ is a p-dimensional covariate vector. The conditional distribution of $Y_i|X_i$ is specified as $\mathbb{P}[Y_i = 1 \mid \boldsymbol{X}_i] = [1 + \exp(-\boldsymbol{X}_i^T \boldsymbol{\beta}_0)]^{-1}$, where $\boldsymbol{\beta}_0$ is the vector of true regression coefficients. The likelihood function given the observed data is $L(\boldsymbol{\beta}) = \exp\left(\sum_{i=1}^n \left[Y_i \boldsymbol{X}_i^\top \boldsymbol{\beta} - \rho\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}\right)\right]\right)$, where $\rho(t) = \log\left(1 + e^t\right)$ is the log-partition function.

The catalytic prior is formulated as a weighted likelihood function evaluated on synthetic data. Given a synthetic dataset of size M, denoted by $\{(Y_i^*, \boldsymbol{X}_i^*)\}_{i=1}^M$, the catalytic prior on the logistic regression coefficients is defined as

$$\pi_{cat,M}(\boldsymbol{\beta} \mid \tau) \propto \exp\left\{ \frac{\tau}{M} \sum_{i=1}^{M} \left[Y_i^* \boldsymbol{X}_i^{*\top} \boldsymbol{\beta} - \rho\left(\boldsymbol{X}_i^{*\top} \boldsymbol{\beta} \right) \right] \right\}, \tag{2}$$

where τ is a positive tuning parameter for down-weighting the impact of synthetic data. Various generation schemes for synthetic datasets can be found in Huang et al. (2020). One simple example is to generate synthetic covariates by independently resampling each coordinate of the observed covariates, and then generate synthetic responses from a symmetric Bernoulli distribution.

Under the above catalytic prior, the posterior density of $\boldsymbol{\beta}$ is proportional to

$$\exp\left\{ \sum_{i=1}^{n} \left[Y_i \boldsymbol{X}_i^\top \boldsymbol{\beta} - \rho\left(\boldsymbol{X}_i^\top \boldsymbol{\beta} \right) \right] + \frac{\tau}{M} \sum_{i=1}^{M} \left[Y_i^* \boldsymbol{X}_i^{*\top} \boldsymbol{\beta} - \rho\left(\boldsymbol{X}_i^{*\top} \boldsymbol{\beta} \right) \right] \right\}.$$

The MAP estimator is defined as the mode of this posterior density, which can be expressed as the following M-estimator:

$$\widehat{\boldsymbol{\beta}}_M = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \left[Y_i \boldsymbol{X}_i^\top \boldsymbol{\beta} - \rho\left(\boldsymbol{X}_i^\top \boldsymbol{\beta} \right) \right] + \frac{\tau}{M} \sum_{i=1}^{M} \left[Y_i^* \boldsymbol{X}_i^{*\top} \boldsymbol{\beta} - \rho\left(\boldsymbol{X}_i^{*\top} \boldsymbol{\beta} \right) \right] \right\}. \tag{3}$$

This estimator is the focus of our paper and will be referred to as "the MAP estimator" for short whenever there is no confusion. Numerically, it can be computed as a maximum likelihood estimator with existing software by supplementing the observed data with weighted synthetic data. Empirical studies demonstrated that the MAP estimator provides superior estimations and predictions compared to the standard maximum likelihood estimator (MLE), particularly when the dimension p of $\boldsymbol{\beta}$ is large relative to the observed sample size n.

Since we have full control over the generation of synthetic data, we can consider the limiting case with M diverging to ∞. This leads to the definition of the *population catalytic prior* that replaces the average in (2) with an expectation. Consequently, the resulting MAP estimator $\widehat{\boldsymbol{\beta}}_\infty$ is given by

$$\widehat{\boldsymbol{\beta}}_\infty = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \sum_{i=1}^{n} \left[Y_i \boldsymbol{X}_i^\top \boldsymbol{\beta} - \rho\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}\right) \right] + \tau \mathbb{E}\left[Y^* \boldsymbol{X}^{*\top} \boldsymbol{\beta} - \rho(\boldsymbol{X}^{*\top} \boldsymbol{\beta}) \right] \right\}, \tag{4}$$

where the expectation is taken over the synthetic data-generating distribution. This estimator is mainly of theoretical interest rather than practical usage, as we never have access to infinitely many synthetic data. Huang et al. (2020) has established upper bounds on some distributional distances between a catalytic prior and its population counterpart, which decrease fast when the synthetic sample size M increases. It is also intuitively clear that $\widehat{\boldsymbol{\beta}}_M$ converges to $\widehat{\boldsymbol{\beta}}_\infty$ as $M \to \infty$, but a theoretical justification is missing. The development of such a theoretical result can provide insights into how the stability of $\widehat{\boldsymbol{\beta}}_M$ can be controlled by M. This is important for deciding the value of M in practice, as practitioners often want the resulting inference to be stable against the randomness of the synthetic data.

Although the MAP estimator $\widehat{\boldsymbol{\beta}}_M$ exhibits superior finite-sample performance, large sample properties remain unexplored. Intuitively, the superior performance is achieved because the MAP estimator undergoes a regularization induced by the likelihood based on the synthetic data and the tuning parameter τ can be chosen to optimize the bias-variance tradeoff. Huang et al. (2020) recommended to choose τ proportional to the dimension p based on a heuristic argument about the Hessian matrix of the log posterior density, but there is no rigorous theoretical justification for this choice.

To be more concrete, we want to answer the following questions: 1. How stable is the MAP estimator against the randomness of the synthetic data? 2. Is the MAP estimator consistent even though it relies on synthetic data that are artificially generated? 3. When the dimension p of covariates diverges along with the sample size n, does the MAP estimator behave preferably? 4. What is the asymptotic behavior of the MAP estimator and what are the effects of the tuning parameters τ and M? In this work, we develop the theoretical results for answering these questions regarding the properties of the MAP estimator.

1.2. *Our contribution.* In this paper, we investigate the theoretical properties of the MAP estimator resulting from the catalytic prior and provide several related applications using our theories. We begin with the logistic regression model and then discuss the extension to some other GLMs. In our analysis, the dimension p of covariates is allowed to grow along with the sample size n, and it can scale linearly with n.

Our theoretical contributions are summarized as follows:

1. We show the MAP estimator exists and is unique, even when the dimension p is larger than the sample size n, under verifiable conditions. Furthermore, the squared difference between $\widehat{\boldsymbol{\beta}}_M$ and $\widehat{\boldsymbol{\beta}}_\infty$ can be bounded by C/M, where C depends on the eigenvalues of the Hessian matrix of the negative log-likelihood function and M is the synthetic sample size. This result establishes a guarantee of the stability of the MAP estimator w.r.t. the random synthetic data.

2. When p is allowed to diverge along with n and the ratio p/n converges to zero, we show the MAP is consistent and $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0\|_2^2 = O_p(p/n)$ provided that the tuning parameter satisfies $\tau = O(p)$. This result clarifies that when compared with the MLE, using artificially generated synthetic data does not impede the large sample performance of the MAP estimator while it improves the stability significantly in small samples.
3. When p is allowed to grow as fast as or even faster than n, we prove that $\|\widehat{\boldsymbol{\beta}}_M\|_2 = O_p(1)$ and the estimation error $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0\|_2$ is of constant order provided that $\tau \propto p$. This result together with the last property on consistency establishes the minimax rate optimality of the MAP estimator for any diverging p and n. Note that this optimality is not achieved by the MLE, which may not exist or may not be bounded when p is as large as n.
4. We establish a precise theoretical characterization of the performance of the MAP estimator, which provides exact limits of the performance rather than just upper bounds on the errors. Our analyses cover both the case where the synthetic data are non-informative and the case where the synthetic data may contain information about the true regression coefficients. The results for the former case enhance our understanding on the impacts of the synthetic sample size M and the tuning parameter τ on the MAP estimation. The latter case connects the catalytic prior with the power prior (Chen, Ibrahim and Shao, 2000), where some informative auxiliary data (for example, data from different but similar studies) are used in place of the synthetic data in the formulation (see Section 1.3). Our results for the latter case clarify the role of the similarity between the informative auxiliary data and the observed data.

Our precise characterization is based on a novel application of the Convex Gaussian Minimax Theorem (CGMT) (Thrampoulidis, Oymak and Hassibi, 2014). This technique allows us to characterize the asymptotic behavior of the MAP estimator in terms of the optimal solutions to a convex-concave problem with a few scalar variables. Although CGMT has been applied to study the behavior of regularized M-estimators with separable regularization[2] in the literature (see Section 1.3), there are significant technical challenges in the application of CGMT in the analysis of the MAP estimator under catalytic priors. Specifically, the regularization in (3) is not separable and the existing technique fails to apply. Furthermore, the analysis involves the projection of the MAP estimator onto the two-dimensional space spanned by the true regression coefficients and the coefficients for the synthetic data generation, rather than a rank-one projection as in other existing works. From a technical perspective, this paper contributes to the literature with a novel application of CGMT to regularized M-estimators with non-separable regularization, and our proving strategy may shed light on future studies on related problems.

1.3. *Related literature.* The standard estimation for logistic regression is the Maximum Likelihood Estimation (MLE), whose feasibility depends on the geometry presented in the data. Albert and Anderson (1984) proved that MLE does not exist when the observed data are separable, i.e. there is a vector $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $(2Y_i - 1)\boldsymbol{X}_i^\top \boldsymbol{\beta} \geq 0$ for every i, and several studies considered using linear programming to identify separation (Albert and Anderson, 1984; Silvapulle and Burridge, 1986; Konis, 2007). Recently, Candès and Sur (2020) applied conic integral geometry theory to reveal that the existence of MLE undergoes a sharp phase transition when the dimension scales with sample size, and Sur and Candès (2019) provided explicit expressions for the asymptotic bias and variance of the MLE, showing that even when the MLE exists, it may be biased upward and have larger variability than

[2] A regularization function $h(\boldsymbol{b})$ is said to be separable if $h(\boldsymbol{b}) = \sum_{j=1}^p h\left(b_j\right)$ for some convex function $h(\cdot)$. E.g.: $h(\boldsymbol{b}) = \|\boldsymbol{b}\|_1 = \sum_i |b_i|$ and $h(\boldsymbol{b}) = \|\boldsymbol{b}\|_2^2 = \sum_i b_i^2$ are separable regularization functions.

classically estimated. The class of catalytic priors provides a remedy to these issues with the likelihood-based inference (Huang et al., 2020). However, the existence and stability of MAP estimations utilizing catalytic priors in logistic regression remain unexplored.

The current paper fills in the aforementioned gap and extends to study the consistency and asymptotic behavior of the MAP estimation. In the classical setting, the consistency and asymptotic behavior of MLE are thoroughly examined in fixed dimensional settings (p is fixed) and in low-dimensional settings (p grows along with n at a slower order) in several foundational studies (Fahrmeir and Kaufmann, 1985; Portnoy, 1984, 1988; He and Shao, 2000). However, in the linear asymptotic regime where p and n are of the same scale, the classical consistency theory of MLE is not valid. This gap has spurred researchers to develop new theoretical frameworks for understanding the asymptotic nature of MLE and other regularized estimators. These frameworks, which offer precise characterizations of the limiting distributions of estimators, have been successfully employed in both linear models (Bayati and Montanari, 2011; El Karoui et al., 2013; Thrampoulidis, Oymak and Hassibi, 2015; El Karoui, 2018) and binary regression models (Sur and Candès, 2019; Salehi, Abbasi and Hassibi, 2019; Taheri, Pedarsani and Thrampoulidis, 2020; Deng, Kammoun and Thrampoulidis, 2022). The main technical tools for the development of these frameworks include approximate message passing (Donoho, Maleki and Montanari, 2009; Bayati and Montanari, 2011), Convex Gaussian Min-Max Theorem (CGMT) (Thrampoulidis, Oymak and Hassibi, 2015; Thrampoulidis, Abbasi and Hassibi, 2018), and the leave-one-out analysis (El Karoui et al., 2013; El Karoui, 2018). Specifically, the precise characterization of the MAP estimator developed in the current paper is based on CGMT. Although CGMT is a powerful tool for reducing the analysis of a min-max optimization to a much simpler optimization with the same optimum, the analysis of the reduced optimization problem is problem-specific and often challenging. To take into account the use of synthetic data in the MAP estimation, novel probabilistic analyses have to be developed.

Our theoretical analyses of the MAP estimation using catalytic prior are also related to prior elicitation and transfer learning. In Bayesian analysis, Ibrahim and Chen (2000) proposed the class of power priors for incorporating information from historical data or data from previous similar studies, and Chen, Ibrahim and Shao (2000); Ibrahim, Chen and Sinha (2003) studied some of the theoretical properties of these priors. More generally, it is often of great interest to improve the estimation or prediction of a "target model" by borrowing information from auxiliary samples that are generated using a different but possibly related "source model." This is the goal of transfer learning (Torrey and Shavlik, 2010), which is possible when lots of auxiliary samples are available and the difference between the target model and the source model is sufficiently small. Many methods for transfer learning have been investigated recently from the statistical perspectives; see for example Bastani (2021); Reeve, Cannings and Samworth (2021); Li, Cai and Li (2022, 2023); Tian and Feng (2023); Li et al. (2023); Zhang and Li (2023). The MAP estimation using a power prior can be considered as a transfer learning method, and its theoretical properties have not been explored in the literature. Since this estimation uses the same expression as in (1) with $\mathcal{D}^*$ being an auxiliary dataset, the results in the current paper also apply to this estimation.

Finally, we point out that there is a vast literature on regularization methods in statistics. For a selective overview of common regularization, see Bickel et al. (2006) and the accompanying discussion papers. In high-dimensional problems, structured regularization is often employed to achieve statistical and computational efficiency; see the survey Wainwright (2014) for an overview of these developments. Such a classical regularization method usually involves a prechosen penalty function of the model parameter values, while the regularization considered in this paper is based on the idea of supplementing the actual observed data with generated synthetic data or informative auxiliary data. Despite this difference, both methods

can improve estimation and prediction by trading off bias and variance. Classical regularization methods are often useful in utilizing structured assumptions such as sparsity, and there is a connection between many regularization methods and the MAP estimator; see Huang et al. (2022, Section 4) for more details. We emphasize that regularization via synthetic data is not a replacement but an interesting supplement to classical regularization methods, and the purpose of this work is not to compare these methods but to provide a theoretical foundation for regularization via synthetic data that has not been fully studied.

1.4. *Organization and Notation.* The remainder of this paper is organized as follows: In Section 2, we establish the existence of the MAP estimator (i.e. $\widehat{\boldsymbol{\beta}}_M$ defined in (3)) and investigate its stability against the random synthetic data. Section 3 proves the consistency of the MAP estimator when the ratio of the dimension p over the sample size n converges to 0. Section 4 focuses on the theoretical properties of the MAP estimator as p/n converges to a positive constant. Importantly, we establish a precise characterization of the asymptotic behavior of the MAP estimator using non-informative synthetic data and informative auxiliary data in Section 4.2 and Section 4.3 respectively. Section 5 studies the estimation of the parameters that determine the aforementioned asymptotic behaviors and provides some numerical studies on the application of our theories. These sections focus on logistic regression models, and Section 6 extends the developed theoretical framework to other generalized linear models. We discuss some related problems and future directions in Section 7. All proofs are provided in the appendix.

Throughout the paper, for a vector $\boldsymbol{v} \in \mathbb{R}^p$, we write $\|\boldsymbol{v}\|_q$ with $q \geq 1$ for the standard ℓ_q norm of $\boldsymbol{v}$, i.e., $\|\boldsymbol{v}\|_q = (\sum_i |v_i|^q)^{1/q}$. For a positive definite matrix A, we use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote its minimum and maximum eigenvalues, respectively. For a general real matrix B, we use $\|B\|_{\text{op}}$ to denote its operator norm. For a positive integer n, we use $[n]$ as a shorthand for the set $\{1, 2, \cdots, n\}$. For a statement $\mathcal{E}$, we use $\mathbf{1}\{\mathcal{E}\}$ to denote the indicator function that is equal to 1 if the statement $\mathcal{E}$ holds and is equal to 0 otherwise. The function $\rho(t) = \log(1+\exp(t))$ is the log-partition function of the Bernoulli distribution. $Y \sim \text{Bern}(\theta)$ means that the random variable Y follows the Bernoulli distribution with success probability θ. For any real number x, χ_x denotes the point mass at x. The symbols $\rightsquigarrow$ and $\xrightarrow{\mathbb{P}}$ denote weak convergence and convergence in probability, respectively.

2. Existence and Stability of MAP. In this section, we focus on finite-sample properties of the MAP, namely, the existence and the stability. To ease the notation, the first coordinate of a covariate vector is set to be 1 so that the first coordinate of $\boldsymbol{\beta}$ corresponds to the intercept (constant) term in the regression model. We first introduce a condition on the synthetic data generation.

CONDITION 1. The synthetic data are i.i.d. copies of $(\boldsymbol{X}^*, Y^*)$ such that the followings hold:

- The synthetic covariate vector $\boldsymbol{X}^* = \left(X_1^*, X_2^*, \cdots, X_p^*\right)$ satisfies (1) $X_1^* \equiv 1$; (2) $\mathbb{E}X_j^* = 0$, $\text{Var}\left(X_j^*\right) = 1$, and $\left|X_j^*\right| \leq B_1$, a.s., for $j = 2, \cdots, p$; (3) $X_2^*, \cdots, X_p^*$ are independent.
- For the synthetic response Y^*, there is some $q \in (0, 1)$ such that $q \leq \mathbb{P}(Y^* = 1 \mid \boldsymbol{X}^*) \leq 1 - q$.

Condition 1 is mild. The first two requirements on $\boldsymbol{X}^*$ require the coordinates to be standardized and bounded, and the last one requires the coordinates to be independent. These requirements will be satisfied if the coordinates of $\boldsymbol{X}^*$ are resampled independently from the

coordinates of observed covariate data (and historical data if available). The requirement on Y^* is also mild. In particular, if we generate synthetic responses independently from a non-informative Bernoulli distribution with a success probability of 0.5, the condition is satisfied with $q = 0.5$. Generally, if the synthetic responses are generated from a fitted simpler model, the condition will often be satisfied. For example, if $\mathbb{P}(Y^* = 1 \mid \boldsymbol{X}^*) = [1+\exp(\boldsymbol{X}^{*\top}\boldsymbol{\beta}_s)]^{-1}$, where the coefficient $\boldsymbol{\beta}_s$ satisfied $\|\boldsymbol{\beta}_s\|_1 \leq C$, then the requirement on the synthetic covariate implies that $|\boldsymbol{X}^{*\top}\boldsymbol{\beta}_s| \leq CB_1$ and thus the conditional probability $\mathbb{P}(Y^* = 1 \mid \boldsymbol{X}^*)$ is uniformly bounded away from zero and one.

2.1. *Existence of the MAP Estimate.* In high-dimensional logistic regression, the MLE will often be infinite (Candès and Sur, 2020), especially when $p/n > 1/2$. In contrast, the MAP estimate of the catalytic prior exists under some mild conditions, as will be shown in this section.

A dataset $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ is said to be separable if and only if there exists a hyperplane in the covariate space that separates the covariate vectors $\boldsymbol{X}_i$ with $Y_i = 0$ and those with $Y_i = 1$. It is well-known that the MLE does not exist if the observed data are separable (Albert and Anderson, 1984). In this case, there exists some $\overline{\boldsymbol{\beta}} \in \mathbb{R}^p$ s.t. $(2Y_i - 1)\boldsymbol{X}_i^\top \overline{\boldsymbol{\beta}} \geq 0$ for all i, and the likelihood at $\boldsymbol{\beta}$ can be continuously increased to 1 if we choose $\boldsymbol{\beta} = c\overline{\boldsymbol{\beta}}$ with c being a positive constant increasing to infinity. On the other hand, the MAP estimation resulting from the catalytic prior takes into account the synthetic data and behaves much better. When the synthetic data are not separable, the MAP estimate is finite because for any $\overline{\boldsymbol{\beta}} \in \mathbb{R}^p$, the likelihood at $c\overline{\boldsymbol{\beta}}$ is not increasing in c. Additionally, the strong concavity of the log posterior density guarantees the uniqueness of the MAP estimate. This observation leads to the following result.

THEOREM 2.1. *If the synthetic data $\{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M$ are not separable and the synthetic covariate matrix has full column rank, then MAP estimate in (3) exists and is unique.*

Theorem 2.1 guarantees the existence and uniqueness of the MAP estimate resulting from a catalytic prior under the condition that the synthetic covariate matrix is of full rank and the synthetic data are non-separable. This condition can be verified straightforwardly by checking the existence of the MLE based on the synthetic data. Furthermore, this condition can be facilitated since we have full control over the generation of synthetic data. For example, if the synthetic covariates satisfy Condition 1, the synthetic responses are generated from a sub-model, and the ratio M/p is sufficiently large, then with probability converging to 1, the MLE based on the synthetic data exists and the condition in Theorem 2.1 is satisfied (Liang and Du, 2012, Theorem 1).

2.2. *Stability of the MAP against finite M.* In this section, we study the influence of the synthetic sample size M on the stability of the MAP estimator. Specifically, we establish a bound on the distance between the estimate $\widehat{\boldsymbol{\beta}}_M$ based on M synthetic samples defined in (3) and the estimate $\widehat{\boldsymbol{\beta}}_\infty$ based on the population catalytic prior defined in (4). This bound decays to 0 linearly in M.

Here we treat the observed data as fixed and consider the synthetic data the only source of randomness. For any $L > 0$, we define $\mathcal{B}_L := \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_2 \leq L\}$.

THEOREM 2.2. *Suppose that $\tau > 0$ and the following holds*

(a) *the synthetic data are generated according to Condition 1;*
(b) *there is some constant $L > 0$, such that both $\widehat{\boldsymbol{\beta}}_M$ and $\widehat{\boldsymbol{\beta}}_\infty$ lie in $\mathcal{B}_L$.*

Let $\lambda_n \geq 0$ be a constant such that for any $\boldsymbol{\beta} \in \mathcal{B}_L$, the smallest eigenvalue value of $\sum_{i=1}^n \rho''(\boldsymbol{X}_i^\top \boldsymbol{\beta}) \boldsymbol{X}_i \boldsymbol{X}_i^\top$ is lower bounded by λ_n. The following hold:

(i) *There is a positive constant γ that only depends on the constants L and B_1 in Condition 1 such that the smallest eigenvalue value of $\mathbb{E}\left(\rho''(\boldsymbol{X}^{*\top}\boldsymbol{\beta})\boldsymbol{X}^*\boldsymbol{X}^{*\top}\right)$ is lower bounded by γ for all $\boldsymbol{\beta} \in \mathcal{B}_L$.*

(ii) *For any $\epsilon \in (0,1)$, it holds with probability at least $1-\epsilon$ that*

$$\|\widehat{\boldsymbol{\beta}}_M - \widehat{\boldsymbol{\beta}}_\infty\|^2 \leq \left(2 + \log\frac{1}{\epsilon}\right)\left(\frac{4\, p\, \tau^2 B_1^2}{M\,(\lambda_n + \tau\gamma)^2}\right).$$

In particular, since $\lambda_n \geq 0$, we have $\|\widehat{\boldsymbol{\beta}}_M - \widehat{\boldsymbol{\beta}}_\infty\|^2 = O_p\left(\frac{p}{M\gamma^2}\right)$.

Theorem 2.2 shows that $\|\widehat{\boldsymbol{\beta}}_M - \widehat{\boldsymbol{\beta}}_\infty\|^2$ decays linearly in the synthetic sample size M. Numerical illustrations in Appendix C.1 show this decay rate in more general settings. The theorem also indicates that the influence of the random synthetic data on the MAP estimator can be effectively mitigated by increasing the value of M. In a classical setting where the model is considered fixed, the smallest eigenvalue λ_n as defined in Theorem 2.2 typically grows linearly with the sample size n. Consequently, an upper bound deduced from Theorem 2.2 is $O\left(\frac{\tau^2}{n^2 M}\right)$, which suggests that a moderately large M is sufficient to ensure the MAP estimate stays close to $\widehat{\boldsymbol{\beta}}_\infty$ since we can take τ to be negligible relative to n. In a high-dimensional setting where the model dimension increases along with n, λ_n could be as small as 0 so that the upper bound in Theorem 2.2 becomes $O\left(\frac{p}{M\gamma^2}\right)$. In this case, we require M/p sufficiently large to guarantee that the MAP estimate is not far away from $\widehat{\boldsymbol{\beta}}_\infty$.

Theorem 2.2 requires that $\|\widehat{\boldsymbol{\beta}}_M\|_2$ and $\|\widehat{\boldsymbol{\beta}}_\infty\|_2$ are bounded, without which the term $\rho''(t)$ in the Hessian matrix of the objective function could be too small with a large $|t|$. This boundedness condition on $\|\widehat{\boldsymbol{\beta}}_M\|_2$ and $\|\widehat{\boldsymbol{\beta}}_\infty\|_2$ is considered mild since we will show in Theorem 3.1 and Theorem 4.1 that both estimators are either consistent or have bounded norms in a broad range of scenarios.

3. Consistency of MAP when p diverges. In this section, we show that the MAP estimator defined in (3) is consistent in the regime that the dimension p is allowed to diverge to infinity with the sample size n in the order of $p = o(n)$. The asymptotic behavior of the MAP estimator when p grows as fast as n is studied in Section 4.

To obtain consistency, we impose the following conditions on the observed covariates and the underlying true regression coefficients:

CONDITION 2. $\mathbb{E}\left(\|\boldsymbol{X}_i\|_2^2\right) \leq C_2 p$ for all $i \in \{1, 2, \cdots n\}$.

CONDITION 3. There exist positive constants c_1, c_2, ζ, and N_0 such that for any $n > N_0$ and any subset $S \subseteq \{1, 2, \cdots, n\}$ with $|S| \geq (1-\zeta)n$, the following inequality holds:

$$c_1|S| \leq \lambda_{\min}\left(\sum_{i\in S} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right) \leq \lambda_{\max}\left(\sum_{i\in S} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right) \leq c_2|S|.$$

CONDITION 4. There exists a positive constant C_3 such that the true regression coefficients $\boldsymbol{\beta}_0$ is bounded as $\|\boldsymbol{\beta}_0\|_2 \leq C_3$.

Condition 2 is a moment condition on the observed covariate vectors and is weaker than the common condition in the literature on M-estimators with diverging dimension. For example, Portnoy (1984) assumes that $\max_{i\le n}\|\boldsymbol{X}_i\|^2 = O(n^2)$ and Liang and Du (2012) assumes $\sup_{i\le n, j\le p}|X_{i,j}| < \infty$.

Condition 3 is slightly stronger than the usual condition regarding the upper bound for the sample covariance matrix. The common condition is $c_1 \le \lambda_{\min}\left(n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top\right) \le \lambda_{\max}\left(n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top\right) \le c_2$, which has been adopted in studies with both fixed dimensions (Chen, Hu and Ying, 1999; Lai and Wei, 1982) and increasing dimensions (Portnoy, 1984; Wang, 2011; Liang and Du, 2012). This common condition on its own is not enough to ensure the good behavior of the Hessian matrix for diverging dimension, unless an extra condition is made, which assumes that the conditional probability $\mathbb{P}(Y_i = 1|\boldsymbol{X}_i)$ is uniformly bounded away from 0 and 1 for all $i \in [n]$; see (Wang, 2011; Liang and Du, 2012). However, this extra condition on the conditional probabilities is too strong to hold, even when $\boldsymbol{X}_i$ is standard Gaussian. Condition 3 mitigates the need for stringent conditions on the conditional probabilities and it will hold under various designs such as sub-Gaussian designs.

Condition 4 is intended to avoid degenerate scenarios where as p increases, the size of the log-odds ratio $\boldsymbol{X}_i^\top\boldsymbol{\beta}_0$ becomes unbounded and the true conditional probability $\mathbb{P}(Y_i = 1|\boldsymbol{X}_i)$ often becomes either 0 or 1. This condition is also commonly assumed in the literature.

The following theorem establishes the consistency of the MAP estimator with diverging p.

THEOREM 3.1. *Consider the logistic regression model and the MAP estimators $\widehat{\boldsymbol{\beta}}_M$ and $\widehat{\boldsymbol{\beta}}_\infty$ defined in Section 1.1. Suppose $p = o(n)$ and the tuning parameter is chosen such that $\tau \le C_4 p$ for a constant C_4. Under Conditions 2, 3, and 4, the followings hold:*

(i) *Suppose there is a constant Λ such that $\|\frac{1}{M}\sum_{i=1}^M \boldsymbol{X}_i^*\boldsymbol{X}_i^{*\top}\| \le \Lambda$, then*

$$\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0\|_2^2 = O_p\left(\frac{p}{n}\right).$$

(ii) *Under Condition 1, we have*

$$\|\widehat{\boldsymbol{\beta}}_\infty - \boldsymbol{\beta}_0\|_2^2 = O_p\left(\frac{p}{n}\right).$$

REMARK 1. The condition that $\|M^{-1}\sum_{i=1}^M \boldsymbol{X}_i^*\boldsymbol{X}_i^{*\top}\| \le \Lambda$ is very mild since we have full control over the generation of $\boldsymbol{X}_i^*$. For example, if M/p is sufficiently large and Condition 1 holds, the results from random matrix theory (e.g., Theorem 5.44 in Vershynin (2010)) imply that with probability 1 this condition will be eventually met.

Theorem 3.1 shows the consistency of both estimators $\widehat{\boldsymbol{\beta}}_M$ in (3) and $\widehat{\boldsymbol{\beta}}_\infty$ in (4) when the dimension p diverges to infinity along with the sample size n but of a slower order. The consistency requires that the tuning parameter τ is at largest the order of p, so that the influence of the synthetic data does not overwhelm the information derived from the observed data. This requirement about τ aligns with the empirical suggestion that τ should be chosen proportional to p (Huang et al., 2020).

Based on Theorem 3.1, we know that when p grows slower than n, both $\widehat{\boldsymbol{\beta}}_M$ and $\widehat{\boldsymbol{\beta}}_\infty$ will converge to $\boldsymbol{\beta}_0$ in probability. In this case, it is not only straightforward to see that $\|\widehat{\boldsymbol{\beta}}_M - \widehat{\boldsymbol{\beta}}_\infty\|_2 = O_p(\frac{p}{n})$, but we can also justify the condition (b) required in Theorem 2.2, which allows for precise quantification of the stability of the MAP estimator against the random synthetic data.

4. Characterization in the linear asymptotic regime . In this section, we study the behavior of the MAP estimator in the regime where the dimension p of the regression coefficients grows as fast as the sample size n. This scenario is often referred to as the linear asymptotic regime, which has attracted significant interest recently.

Specifically, we assume that p grows along with n such that $\lim n/p = \delta$ for some $\delta \in (0, \infty)$. Without imposing any additional structural condition on the true regression coefficients such as sparsity, no estimation can achieve consistency in this regime. Nevertheless, it is of our interest to understand the asymptotic behavior of the MAP estimator in this regime. In Section 4.1, we show that the estimation error of the MAP estimator is bounded in probability, regardless of the value of δ. This result indicates that the behavior of the MAP estimator in the linear asymptotic regime is quite different from that of the MLE, which usually fails to exist for small δ. In Section 4.2, we consider a special case where synthetic responses are generated from a logistic regression model with coefficients $\boldsymbol{\beta}_s = \mathbf{0}$ and we characterize the precise asymptotic behavior of the MAP estimator using the Convex Gaussian Min-max Theorem (CGMT). In Section 4.3, we consider a general case where synthetic responses are generated with coefficients $\boldsymbol{\beta}_s$ and provide a precise characterization in terms of the cosine similarity between $\boldsymbol{\beta}_s$ and $\boldsymbol{\beta}_0$. Section 4.4 discusses a conjecture on the precise characterization of the asymptotic behavior of the MAP estimator based on the population catalytic prior.

4.1. *Nonasymptotic results.* We will show that the norm of the MAP estimator can be uniformly bounded even when the dimension p is as large as n. Establishing this boundedness of the MAP estimator is not trivial and our proof has to make use of some desirable properties of catalytic priors for GLMs that have been studied by Huang et al. (2020). In addition, our result does not require $n > p$ or put any strong condition on observed data, but instead merely requires some mild conditions on synthetic data, on which we have full control. This stands in contrast to the requirement for bounding the norm of the MLE, as proved by Sur, Chen and Candès (2019), which requires that $n > 2p$ and puts the normality condition on the covariates.

To introduce our result, we impose the following condition on the tuning parameter τ:

CONDITION 5. There is a positive constant c_* such that the tuning parameter τ is chosen such that $\tau \geq c_* p$.

Condition 5 guarantees that τ is not too small so that the catalytic prior provides sufficient regularization to the estimation. This aligns with the general principle that a model with more parameters usually requires more regularization to prevent overfitting (Hastie, Tibshirani and Friedman, 2009). In the linear asymptotic regime, the lack of such a condition will cause difficulties in the MAP estimation: if $\tau = o(p)$, then the regularization term will be negligible relative to the log-likelihood of the observed data and the performance of the MAP estimator will be similar to the MLE, which may often be unbounded if n/p is small (Candès and Sur, 2020).

The following result considers the scenario where p is not negligible relative to n and shows that the MAP estimator is bounded if sufficient regularization has been imposed.

THEOREM 4.1. *Consider the logistic regression model and the MAP estimators defined in Section 1.1. Suppose Conditions 1 and 5 hold and $p > \omega_* n$ for some positive constant ω_*. Let C_* be the constant $1 + c_* \omega_*$. There are some positive constants $\tilde{c}, \tilde{C}, \eta_0, \nu$ that only depend on the constants B_1 and q in Condition 1 such that the followings hold:*

(i) *If $M \geq \tilde{C}p$, the estimator $\widehat{\boldsymbol{\beta}}_M$ defined in (3) satisfies that*

$$\|\widehat{\boldsymbol{\beta}}_M\|_2 \leq \frac{4C_* \log(2)}{\eta_0 \nu}$$

with probability at least $1 - 2\exp(-\tilde{c}M)$.

(ii) *The estimator $\widehat{\boldsymbol{\beta}}_\infty$ defined in (4) satisfies that*

$$\|\widehat{\boldsymbol{\beta}}_\infty\|_2 \leq \frac{C_* \log(2)}{\eta_0 \nu}.$$

REMARK 2. Theorem 4.1 requires sufficient regularization that $\tau \geq c_* p$ for some positive c_* as stated in Condition 5. In contrast, Theorem 3.1 requires that $\tau \leq C_4 p$ for some C_4 in order to achieve the consistency. These two conditions do not contradict each other, as they can be met simultaneously if the tuning parameter τ is chosen as $\tau \propto p$. This choice of τ aligns with the empirical recommendation reported in Huang et al. (2020).

Theorem 4.1 has several implications. First, under the condition of Theorem 4.1, both $\|\widehat{\boldsymbol{\beta}}_M\|_2$ and $\|\widehat{\boldsymbol{\beta}}_\infty\|_2$ have bounded norms for sufficiently large M, which justifies the condition required by Theorem 2.2 for bounding $\|\widehat{\boldsymbol{\beta}}_M - \widehat{\boldsymbol{\beta}}_\infty\|_2$ in terms of M. Second, under Condition 4 about the true parameter values, the boundedness of the MAP estimator immediately implies the following corollary regarding the estimation error.

COROLLARY 4.2. Suppose the conditions in Theorem 4.1 hold. Under Condition 4, there are positive constants $\tilde{C}_1, \tilde{C}_2$, and $\tilde{c}$ that depend on c_*, B_1, and q only such that

$$\|\widehat{\boldsymbol{\beta}}_\infty - \boldsymbol{\beta}_0\|_2^2 \leq \tilde{C}_1,$$

and if $M \geq \tilde{C}_2 p$, then the following holds

$$\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0\|_2^2 \leq \tilde{C}_1$$

with probability at least $1 - 2\exp(-\tilde{c}M)$.

Corollary 4.2 shows that the MAP estimator can achieve a constant error rate even when p grows as fast as or faster than n. This turns out to be the best possible error rate. Without imposing any stringent conditions on $\boldsymbol{\beta}_0$ such as sparsity, the minimax lower bound on the estimation risk using quadratic loss in generalized linear models is typically at the order $\min(\frac{p}{n}, 1)$ (see, for example, Chen et al. (2016) and Lee and Courtade (2020)). Based on Corollary 4.2 and Theorem 3.1, we can conclude that the MAP estimator (with tuning parameter $\tau \propto p$) can achieve the optimal convergence rate for estimation regardless how fast p diverges. This contrasts sharply with the MLE, whose error becomes unbounded when the ratio p/n is large.

4.2. *Exact asymptotics with non-informative synthetic data.* In this and the next sections, we will focus on the asymptotic behavior of the MAP estimator in the linear asymptotic regime, i.e., $\lim n/p = \delta$. Informally, in this section, we demonstrate that

$$\widehat{\boldsymbol{\beta}}_M \approx \alpha_* \boldsymbol{\beta}_0 + p^{-1/2} \sigma_* \boldsymbol{Z}, \tag{5}$$

where $\boldsymbol{Z}$ is a standard normal vector, and (α_*, σ_*) are constants that depend on δ, τ, and the data generation process. This suggests that asymptotically the MAP estimator is centered around $\alpha_* \boldsymbol{\beta}_0$ with some additive Gaussian noise.

To state the rigorous result, we begin with some scaling parameters and conditions.

CONDITION 6. There are constants $\tau_0 \in (0,1)$, $m \in (0,\infty)$, and $\delta \in (1,\infty)$ such that the tuning parameter τ and the synthetic sample size M satisfy $\tau/n = \tau_0$, $M/n = m$, and $p/n = 1/\delta$.

Condition 6 requires that both the tuning parameter τ and the synthetic sample size M scale linearly to n. This condition aligns with the results developed in the previous sections; Section 3 and Section 4.1 suggest that choosing τ proportional to p achieves the optimal rate of estimation, and Section 2.2 suggests that if M/p is sufficiently large, the variability becomes negligible. This condition also echos the empirical recommendation of the choice of τ and M made in Huang et al. (2020).

CONDITION 7. $\{\boldsymbol{X}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $Y_i \mid X_i \sim \text{Bern}(\rho'(X_i^\top \boldsymbol{\beta}_0)$ and as $n \to \infty$, the empirical distribution of coordinates of $\sqrt{p}\boldsymbol{\beta}_0$ converges weakly to a distribution Π with a finite second moment (i.e., $\frac{1}{p}\sum_{j=1}^p \chi_{\sqrt{p}\beta_{0,j}} \rightsquigarrow \Pi$). Furthermore, there is a constant $\kappa_1 > 0$, such that $\lim_{p\to\infty} \|\boldsymbol{\beta}_0\|^2 = \kappa_1^2$.

Condition 7 imposes a strong condition on the covariate matrix, assuming that its entries are independent standard Gaussian random variables. Without a Gaussian design, we have shown that $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0\|^2$ is of constant order, but we cannot further characterize the behavior of the MAP in the linear asymptotic regime. Standard Gaussian designs condition are common in such regimes; see, e.g., Bayati and Montanari (2011); Thrampoulidis, Oymak and Hassibi (2015); Donoho and Montanari (2016); Thrampoulidis, Abbasi and Hassibi (2018); Sur and Candès (2019); Salehi, Abbasi and Hassibi (2019); Deng, Kammoun and Thrampoulidis (2022); Dai et al. (2023a) for an incomplete list of related works. Some recent works attempt to relax the standard Gaussian design condition in various settings to allow general covariance structural (Zhao, Sur and Candes, 2022; Celentano, Montanari and Wei, 2023) and replace the normality by moment conditions (El Karoui, 2018; Han and Shen, 2023). We expect that it is possible to relax the Gaussian design condition for our result to hold and we provide empirical justification in Section C.7, which suggests that the same convergence seems to hold if the entries of $\boldsymbol{X}_i$'s are independent with zero mean, unit variance, and a finite fourth moment. However, the development will be much more complicated than the current work and we leave it for future study.

In Condition 7, the constant κ_1 can be understood as the signal strength of $\boldsymbol{\beta}_0$, because the inner product $\boldsymbol{X}_i^\top \boldsymbol{\beta}_0$ has variance κ_1^2. As a result, this condition guarantees that the value of $\rho'\left(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0\right)$ does not degenerate to either 0 or 1 when p increases. In Candès and Sur (2020), κ_1 is an important parameter to determine the existence of MLE: if κ_1 is above a certain threshold determined by $\delta = n/p$, the MLE does not exist with high probability.

The MAP estimator of our interest here is based on the catalytic prior whose synthetic data generation satisfies the following condition.

CONDITION 8. $\{\boldsymbol{X}_i^*\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and $\{Y_i^*\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} Bern(0.5)$.

Condition 8 can always be met since we have full control of the synthetic data generation. This condition essentially assumes that the synthetic responses are generated from the logistic regression with coefficient $\boldsymbol{\beta}_s = \mathbf{0}$. In the next section, we will consider a more general case where the coefficient $\boldsymbol{\beta}_s$ for the synthetic data is nonzero and can be correlated with $\boldsymbol{\beta}_0$.

The constants α_* and σ_* in (5) are related to the following important system of equations in three variables (α, σ, γ):

$$
\begin{cases}
\dfrac{\sigma^2}{2\delta} = \mathbb{E}\left[\rho'\left(-\kappa_1 Z_1\right)\left(\kappa_1\alpha Z_1 + \sigma Z_2 - \mathrm{Prox}_{\gamma\rho(\cdot)}\left(\kappa_1\alpha Z_1 + \sigma Z_2\right)\right)^2\right] \\
\qquad\qquad + m\mathbb{E}\left[\dfrac{1}{2}\left(\kappa_1\alpha Z_1 + \sigma Z_2 - \mathrm{Prox}_{\gamma_0\rho(\cdot)}\left(\kappa_1\alpha Z_1 + \sigma Z_2\right)\right)^2\right], \\
1 - \dfrac{1}{\delta} = \mathbb{E}\left[\dfrac{2\rho'\left(-\kappa_1 Z_1\right)}{1 + \gamma\rho''\left(\mathrm{Prox}_{\gamma\rho(\cdot)}\left(\kappa_1\alpha Z_1 + \sigma Z_2\right)\right)}\right] \\
\qquad\qquad - \mathbb{E}\left[\dfrac{\gamma\tau_0\rho''\left(\mathrm{Prox}_{\gamma_0\rho(\cdot)}\left(\kappa_1\alpha Z_1 + \sigma Z_2\right)\right)}{1 + \gamma_0\rho''\left(\mathrm{Prox}_{\gamma_0\rho(\cdot)}\left(\kappa_1\alpha Z_1 + \sigma Z_2\right)\right)}\right], \\
-\dfrac{\alpha}{2\delta} = \mathbb{E}\left[\rho''\left(-\kappa_1 Z_1\right)\mathrm{Prox}_{\gamma\rho(\cdot)}\left(\kappa_1\alpha Z_1 + \sigma Z_2\right)\right],
\end{cases}
\tag{6}
$$

where Z_1, Z_2 are independent standard Gaussian variables, the scaling parameters (τ_0, m, δ) are defined in Condition 6, $\gamma_0 = \gamma\tau_0/m$ is a shorthand, κ_1 is defined in Condition 7, and proximal mapping operator $\mathrm{Prox}_{\lambda\rho}(z)$ is defined via

$$
\mathrm{Prox}_{\lambda\rho}(z) = \arg\min_{t\in\mathbb{R}}\left\{\lambda\rho(t) + \frac{1}{2}(t-z)^2\right\}.
$$

Based on the system of equations (6), we are able to make the statement in (5) rigorous and precisely characterize the asymptotic behavior of the MAP estimator.

THEOREM 4.3. *Consider the logistic regression model and the MAP estimator $\widehat{\boldsymbol{\beta}}_M$ defined in Section 1.1. Suppose Conditions 6, 7, and 8 hold and $m\delta > 2$. Assume the parameters $(\kappa_1, \delta, \tau_0, m)$ are such that the system of equations (6) has a unique solution $(\alpha_*, \sigma_*, \gamma_*)$. Then, as $p \to \infty$, for any locally-Lipschitz function*[3] $\Psi : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ *or for the indicator function $\Psi(a,b) = \mathbf{1}\{|a/\sigma_*| \le t\}$ with any fixed $t > 0$, we have*

$$
\frac{1}{p}\sum_{j=1}^{p}\Psi\left(\sqrt{p}(\widehat{\boldsymbol{\beta}}_{M,j} - \alpha_*\boldsymbol{\beta}_{0,j}), \sqrt{p}\boldsymbol{\beta}_{0,j}\right) \xrightarrow{\mathbb{P}} \mathbb{E}[\Psi(\sigma_* Z, \beta)],
$$

where $Z \sim N(0,1)$ independent of $\beta \sim \Pi$.

REMARK 3. The condition $m\delta > 2$ is equivalent to $M > 2p$, and it ensures that the non-informative synthetic data $\{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M$ are not separable with high probability. This condition also ensures that the MAP estimator lies in a compact set, which is a technical requirement for applying CGMT in our proof.

Theorem 4.3 is based on the general result in the subsequent section and we leave the discussion of the proof there.

Theorem 4.3 reveals that in the linear asymptotic regime with Gaussian design, the MAP estimator $\widehat{\boldsymbol{\beta}}_M$ does not concentrate around the true coefficient vector $\boldsymbol{\beta}_0$; instead it is roughly equal to the scaled true coefficient vector $\alpha_*\boldsymbol{\beta}_0$ plus a Gaussian noise vector, as expressed in (5). The solutions α_* and σ_* from the system of equations (6) characterize the bias and the variance of the MAP estimator $\widehat{\boldsymbol{\beta}}_M$, respectively, in the asymptotic sense.

[3] A function $\Psi : \mathbb{R}^m \to \mathbb{R}$ is said to be locally-Lipschitz if there exists a constant $L > 0$ such that for all $\boldsymbol{t}_0, \boldsymbol{t}_1 \in \mathbb{R}^m$, $\|\Psi(\boldsymbol{t}_0) - \Psi(\boldsymbol{t}_1)\| \le L\left(1 + \|\boldsymbol{t}_0\| + \|\boldsymbol{t}_1\|\right)\|\boldsymbol{t}_0 - \boldsymbol{t}_1\|$.

Theorem 4.3 implies various asymptotic relationships between $\widehat{\boldsymbol{\beta}}_M$ and the true coefficients by varying the locally-Lipschitz function Ψ. Here are some examples:

1. Centering of the MAP estimator. By taking $\Psi(a, b) = a$, we obtain
$$\frac{1}{\sqrt{p}} \sum_{j=1}^{p} \left(\widehat{\boldsymbol{\beta}}_{M,j} - \alpha_* \boldsymbol{\beta}_{0,j} \right) \xrightarrow{\mathbb{P}} 0,$$
which suggests that $\widehat{\boldsymbol{\beta}}_M$ is centered at $\alpha_* \boldsymbol{\beta}_0$.
2. Squared error of the MAP estimator. By taking $\Psi(a, b) = (a + (\alpha_* - 1)b)^2$, we have
$$\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0\|^2 \xrightarrow{\mathbb{P}} \sigma_*^2 + (\alpha_* - 1)^2 \kappa_1^2. \tag{7}$$
The limit is the summation of the variance term and the squared bias, which are affected by the tuning parameter τ_0 implicitly. In Section 4.5.1, we plot the theoretical limit against the value of τ_0 and reveal a bias and variance trade-off phenomenon for the regularization using synthetic data.
3. Cosine similarity between true coefficients and the MAP estimator. Together with (7) and Slutsky's theorem, we have
$$\frac{\langle \widehat{\boldsymbol{\beta}}_M, \boldsymbol{\beta}_0 \rangle}{\|\widehat{\boldsymbol{\beta}}_M\|_2 \|\boldsymbol{\beta}_0\|_2} \xrightarrow{\mathbb{P}} \frac{\alpha_* \kappa_1}{\sqrt{\alpha_*^2 \kappa_1^2 + \sigma_*^2}}. \tag{8}$$
4. Confidence intervals for the regression coefficients. For each $j \in [p]$, consider the interval estimate for $\boldsymbol{\beta}_{0,j}$ given by
$$\mathrm{CI}_j = \left[\frac{\widehat{\boldsymbol{\beta}}_{M,j} - 1.96\sigma_*/\sqrt{p}}{\alpha_*}, \frac{\widehat{\boldsymbol{\beta}}_{M,j} + 1.96\sigma_*/\sqrt{p}}{\alpha_*} \right]. \tag{9}$$
By taking $\Psi(a, b) = \mathbf{1}\{-1.96 \le a/\sigma_* \le 1.96\}$, we have $\frac{1}{p}\sum_{j=1}^{p} \mathbf{1}\left\{\boldsymbol{\beta}_{0,j} \in \mathrm{CI}_j\right\} \xrightarrow{\mathbb{P}} 0.95$, which indicates that CI_j's are asymptotically valid on average.

These choices of Ψ have previously been explored in the literature and we remark that these results continue to hold without the condition that $\frac{1}{p}\sum_{j=1}^{p} \chi_{\sqrt{p}\beta_{0,j}} \rightsquigarrow \Pi$. Apart from the above examples, Theorem 4.3 also suggests the convergence of two quantities regarding the prediction performance of the MAP estimator—specifically, the generalization error and the predictive deviance. Let $(\boldsymbol{X}_T, Y_T)$ be a pair of future data sampled from the same population as the observed data. Given the covariate vector $\boldsymbol{X}_T$ and the MAP estimator $\widehat{\boldsymbol{\beta}}_M$, the binary prediction is $\widehat{Y} = \mathbf{1}\{\boldsymbol{X}_T^\top \widehat{\boldsymbol{\beta}}_M \ge 0\}$. The following convergence of the generalization error holds:

$$\mathbb{E}_T[\mathbf{1}\{\widehat{Y} \ne Y_T\}] \xrightarrow{\mathbb{P}} \mathbb{E}[\mathbf{1}\{Y_1 \ne Y_2\}],$$

where $\mathbb{E}_T$ is averaging over the randomness in $(\boldsymbol{X}_T, Y_T)$ and $Y_1 = \mathbf{1}\{\sigma_* Z_1 + \alpha_* \kappa_1 Z_2 \ge 0\}, Y_2 \sim \mathrm{Bern}(\rho'(\kappa_1 Z_2))$ for i.i.d. standard normal variables Z_1 and Z_2. Furthermore, the predictive probability for Y_T is $\rho'(\boldsymbol{X}_T^\top \widehat{\boldsymbol{\beta}}_M)$ and we have the following convergence of the predictive deviance:

$$\mathbb{E}_T\left[D(Y_T, \rho'(\boldsymbol{X}_T^\top \widehat{\boldsymbol{\beta}}_M)) \right] \xrightarrow{\mathbb{P}} \mathbb{E}\left[D(\rho'(\kappa_1 Z_2), \rho'(\sigma_* Z_1 + \alpha_* \kappa_1 Z_2)) \right],$$

where the deviance is $D(a, b) = a \log(a/b) + (1-a)\log((1-a)/(1-b))$ with the convention that $0 \log(0) := 0$. The details of the proof are in Appendix A.7, and we provide a numerical illustration in Appendix C.2.1.

The condition of Theorem 4.3 that $\mathrm{Cov}(\boldsymbol{X}) = \mathbb{I}_p$ can be relaxed to allow for a general covariance matrix, as stated in the following corollary. This result can be proved by combining the proof of Theorem 4.3 with the argument in Zhao, Sur and Candes (2022) and we omit the details here.

COROLLARY 4.4. Consider the logistic regression model and the MAP estimator $\widehat{\boldsymbol{\beta}}_M$ defined in Section 1.1 under Condition 6 and the condition $m\delta > 2$. Suppose $\boldsymbol{X}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for $i \in [n]$ and $\boldsymbol{X}_i^* \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for $i \in [m]$, where $\boldsymbol{\Sigma}$ is a positive definite matrix. Let $v_j^2 = \mathrm{Var}\left(X_{i,j} \mid \boldsymbol{X}_{i,-j}\right)$ denote the conditional variance of $X_{i,j}$ given all other covariates. Furthermore, assume that the empirical distribution $\frac{1}{p}\sum_{j=1}^p \chi_{\sqrt{p}v_j\beta_{0,j}}$ converges weakly to a distribution Π with a finite second moment, $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_0\|^2 \xrightarrow{\mathbb{P}} \kappa_1^2$, and $\sum_{j=1}^p v_j^2\beta_{0,j}^2 \xrightarrow{\mathbb{P}} \mathbb{E}\left[\beta^2\right]$ for $\beta \sim \Pi$. Assume the parameters $(\kappa_1, \delta, \tau_0, m)$ are such that the system of equations (6) has a unique solution $(\alpha_*, \sigma_*, \gamma_*)$. Then, for any locally-Lipschitz function $\Psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ or for the indicator function $\Psi(a, t) = \mathbf{1}\{|a/\sigma_*| \le t\}$ with any fixed $t > 0$, we have

$$\frac{1}{p}\sum_{j=1}^p \Psi\left(\sqrt{p}v_j(\widehat{\boldsymbol{\beta}}_{M,j} - \alpha_*\boldsymbol{\beta}_{0,j}), \sqrt{p}v_j\boldsymbol{\beta}_{0,j}\right) \xrightarrow{\mathbb{P}} \mathbb{E}[\Psi(\sigma_* Z, \beta)].$$

where $Z \sim N(0,1)$ independent of $\beta \sim \Pi$.

4.3. *Exact asymptotics with informative auxiliary data.* In this section, we consider the general case where the synthetic data are sampled from a logistic regression with coefficient $\boldsymbol{\beta}_s$, which can be correlated with the true coefficient $\boldsymbol{\beta}_0$. As discussed in Section 1.3, this setting connects the catalytic prior and the power prior and the MAP estimator can be considered as a transfer learning method. In this context, we use the term *informative auxiliary data* in place of *synthetic data* since these data may either come from similar but different studies or may be generated using estimates reported by previous studies.

Denoting by ξ the cosine similarity between $\boldsymbol{\beta}_s$ and $\boldsymbol{\beta}_0$, our result informally states that

$$\widehat{\boldsymbol{\beta}}_M \approx \alpha_{1*}\boldsymbol{\beta}_0 + \frac{\alpha_{2*}}{\sqrt{1-\xi^2}}(\boldsymbol{\beta}_s - \xi\frac{\|\boldsymbol{\beta}_s\|}{\|\boldsymbol{\beta}_0\|}\boldsymbol{\beta}_0) + p^{-1/2}\sigma_*\boldsymbol{Z}, \tag{10}$$

where $(\alpha_{1*}, \alpha_{2*}, \sigma_*)$ depends on δ, τ, M and the data generation process. Compared with (5), the MAP estimator is not centered at scaled $\boldsymbol{\beta}_0$ but a linear combination of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_s$.

To be specific, we continue to assume Conditions 6 and 7, and we suppose the auxiliary data satisfies the following condition.

CONDITION 9. The covariate vector $\{\boldsymbol{X}_i^*\} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and the auxiliary response $Y_i^* \mid \boldsymbol{X}_i^* \sim Bern\left(\rho'(\boldsymbol{X}_i^\top\boldsymbol{\beta}_s)\right)$. There is a constant $\kappa_2 > 0$, and $\xi \in [0, 1)$, such that $\lim_{p\to\infty}\|\boldsymbol{\beta}_s\|^2 = \kappa_2^2$ and $\lim_{p\to\infty}\frac{1}{\|\boldsymbol{\beta}_0\|\|\boldsymbol{\beta}_s\|}\langle\boldsymbol{\beta}_0, \boldsymbol{\beta}_s\rangle = \xi$.

Similar to (6), we introduce an important system of equations in four variables $(\alpha_1, \alpha_2, \sigma, \gamma)$, which includes an extra variable α_2 to track the influence of informative auxiliary data. To present the new system of equations, let Z_1, Z_2, Z_3 be i.i.d. standard normal random variables. The variable W is defined as a linear combination of Z_1, Z_2 and Z_3, specifically $W := \kappa_1\alpha_1 Z_1 + \kappa_2\alpha_2 Z_2 + \sigma Z_3$. Additionally, we adopt the shorthand notation $\gamma_0 := \tau_0\gamma/m$. The system of equations is given as follows.

$$
(11)\quad \begin{cases} \dfrac{\sigma^2}{2\delta} = \mathbb{E}\left[\rho'\left(-\kappa_1 Z_1\right)\left(W - \mathrm{Prox}_{\gamma\rho(\cdot)}\left(W\right)\right)^2\right] \\ \qquad\qquad + m\mathbb{E}\left[\rho'\left(-\kappa_2\xi Z_1 - \kappa_2\sqrt{1-\xi^2}Z_2\right)\left(W - \mathrm{Prox}_{\gamma_0\rho(\cdot)}\left(W\right)\right)^2\right], \\ 1 - \dfrac{1}{\delta} + m = \mathbb{E}\left[\dfrac{2\rho'\left(-\kappa_1 Z_1\right)}{1+\gamma\rho''\left(\mathrm{Prox}_{\gamma\rho(\cdot)}\left(W\right)\right)}\right] \\ \qquad\qquad + m\mathbb{E}\left[\dfrac{2\rho'\left(-\kappa_2\xi Z_1 - \kappa_2\sqrt{1-\xi^2}Z_2\right)}{1+\gamma_0\rho''\left(\mathrm{Prox}_{\gamma_0\rho(\cdot)}\left(W\right)\right)}\right], \\ -\dfrac{\alpha_1}{2\delta} = \mathbb{E}\left[\rho''\left(-\kappa_1 Z_1\right)\mathrm{Prox}_{\gamma\rho(\cdot)}\left(W\right)\right] \\ \qquad\qquad + m\xi\dfrac{\kappa_2}{\kappa_1}\mathbb{E}\left[\rho''\left(-\kappa_2\xi Z_1 - \kappa_2\sqrt{1-\xi^2}Z_2\right)\mathrm{Prox}_{\gamma_0\rho(\cdot)}\left(W\right)\right], \\ -\dfrac{\alpha_2}{2\delta} = m\sqrt{1-\xi^2}\mathbb{E}\left[\rho''\left(-\kappa_2\xi Z_1 - \kappa_2\sqrt{1-\xi^2}Z_2\right)\mathrm{Prox}_{\gamma_0\rho(\cdot)}\left(W\right)\right]. \end{cases}
$$

We are now ready to make the statement in (10) rigorous.

THEOREM 4.5. *Consider the MAP estimator defined in* (3)*. Suppose Conditions 6, 7, and 9 hold and the auxiliary data are not separable. Assume the parameters* $\delta, \kappa_1, \kappa_2, \tau_0$*, and* ξ *are such that the system of equations* (11) *has a unique solution* $(\alpha_{1*}, \alpha_{2*}, \sigma_*, \gamma_*)$*. Then, as* $p \to \infty$*, Then, for any locally-Lipschitz function* $\Psi : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ *or for the indicator function* $\Psi(a,t) = \mathbf{1}\{|a/\sigma_*| \le t\}$ *with any fixed* $t > 0$*, we have*

(12)
$$
\frac{1}{p}\sum_{j=1}^{p}\Psi\left(\sqrt{p}[\widehat{\boldsymbol{\beta}}_{M,j} - \alpha_{1*}\boldsymbol{\beta}_{0,j} - \frac{\alpha_{2*}}{\sqrt{1-\xi^2}}(\boldsymbol{\beta}_{s,j} - \xi\frac{\kappa_2}{\kappa_1}\boldsymbol{\beta}_{0,j})], \sqrt{p}\boldsymbol{\beta}_{0,j}\right) \xrightarrow{\mathbb{P}} \mathbb{E}[\Psi(\sigma_* Z, \beta)],
$$

where $Z \sim N(0,1)$ *is independent of* $\beta \sim \Pi(\beta)$*. In the special case that* $\xi = 1$*,* (12) *continues to hold if the left-hand side is replaced by* $\frac{1}{p}\sum_{j=1}^{p}\Psi\left(\sqrt{p}(\widehat{\boldsymbol{\beta}}_{M,j} - \alpha_{1*}\boldsymbol{\beta}_{0,j}), \sqrt{p}\boldsymbol{\beta}_{0,j}\right)$.

Our proof of Theorem 4.5 is based on an application of CGMT and a novel orthogonal decomposition of the optimum on the space spanned by $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_s}$. To apply CGMT, it is generally necessary to reduce the optimization problem to an ancillary optimization (AO) over compact sets of variables and then analyze the optima of the AO. However, for our optimization (3), we need to project $\boldsymbol{\beta}$ into a space spanned by $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_s$. To the best of our knowledge, previous analyses using CGMT only proceeded through a rank-one projection matrix $\boldsymbol{\beta}_0\boldsymbol{\beta}_0^\top/|\boldsymbol{\beta}_0|_2^2$, which cannot accommodate our scenario.

We provide a brief overview of our proof. To proceed with our analysis, we utilize the Gram-Schmidt process to find two orthonormal vectors $\boldsymbol{e}_1, \boldsymbol{e}_2$ that span our target space and then decompose the MAP estimator as follows:

$$
\widehat{\boldsymbol{\beta}}_M = (\boldsymbol{e}_1^\top\widehat{\boldsymbol{\beta}}_M)\boldsymbol{\beta}_0 + (\boldsymbol{e}_2^\top\widehat{\boldsymbol{\beta}}_M)\boldsymbol{\beta}_s + \mathbf{P}^\perp\widehat{\boldsymbol{\beta}},
$$

where $\mathbf{P}^\perp$ is the projection matrix onto the orthogonal complement of the space spanned by $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_s$. Next, we develop a novel reduction of the AO problem to track the limits of $\boldsymbol{e}_1^\top\widehat{\boldsymbol{\beta}}_M$ and $\boldsymbol{e}_2^\top\widehat{\boldsymbol{\beta}}_M$. Finally, we demonstrate that $\mathbf{P}^\perp\widehat{\boldsymbol{\beta}}_M$ will be asymptotically equal to $\sigma_*\boldsymbol{Z}$. We believe this novel decomposition could be of independent interest and applicable in other

analyses where it is necessary to project the optimization variable into a multidimensional space.

Theorem 4.5 is more general than Theorem 4.3. Note that the generation of non-informative synthetic data corresponds to $\boldsymbol{\beta}_s = \mathbf{0}$ and $\kappa_2 = 0$. In this case, the system of equation (11) reduces to the system of equations (6), and the convergence in Theorem 4.3 is implied by Theorem 4.5.

When κ_2 is nonzero, the difference between Theorem 4.5 and Theorem 4.3 lies in the extra term $\tilde{\beta} - \xi\frac{\kappa_2}{\kappa_1}\beta$, which results from the redundant information contained in the auxiliary data that is irrelevant to the estimand. Intuitively, the relevant information contained in the auxiliary data comes from the similarity between the auxiliary coefficients $\boldsymbol{\beta}_s$ and true coefficient vector $\boldsymbol{\beta}_0$, which can be quantified as the projection of $\boldsymbol{\beta}_s$ onto the direction of $\boldsymbol{\beta}_0$. The remaining part orthogonal to $\boldsymbol{\beta}_0$ is $\boldsymbol{\beta}_s - \frac{\langle \boldsymbol{\beta}_s, \boldsymbol{\beta}_0\rangle \boldsymbol{\beta}_0}{\|\boldsymbol{\beta}_0\|^2} \approx \boldsymbol{\beta}_s - \xi\frac{\kappa_2}{\kappa_1}\boldsymbol{\beta}_0$ and contributes to the extra term above. In Section 5.3, we utilize the limit in Theorem 4.5 to illustrate that when the cosine similarity ξ is above a certain level, the MAP estimator based on informative auxiliary data can be substantially better than the one with non-informative synthetic data.

Similar to Theorem 4.3, Theorem 4.5 also implies various asymptotic relationships between the MAP estimator and the true coefficients with different choices of Ψ as discussed in Section 4.2. In particular, for squared error, we have

$$\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0\|_2^2 \xrightarrow{\mathbb{P}} (\alpha_{1*} - 1)^2\kappa_1^2 + \alpha_{2*}^2\kappa_2^2 + \sigma_*^2; \tag{13}$$

for cosine similarity, we have

$$\frac{\langle \widehat{\boldsymbol{\beta}}_M, \boldsymbol{\beta}_0\rangle}{\|\widehat{\boldsymbol{\beta}}_M\|_2\|\boldsymbol{\beta}_0\|_2} \xrightarrow{\mathbb{P}} \frac{\alpha_{1*}\kappa_1}{\sqrt{\alpha_{1*}^2\kappa_1^2 + \alpha_{2*}^2\kappa_2^2 + \sigma_*^2}} \tag{14}$$

4.4. *Exact asymptotics under population catalytic prior.* In this section, we investigate the asymptotic behaviour of the MAP estimator under the population catalytic prior, that is the estimator $\widehat{\boldsymbol{\beta}}_\infty$ defined in (4). The difficulty of the analysis lies in the fact that the regularization term $\mathbb{E}\left[Y^*\boldsymbol{X}^{*\top}\boldsymbol{\beta} - \rho(\boldsymbol{X}^{*\top}\boldsymbol{\beta})\right]$ does not have a simple expression that permits the use of our tools developed in Section 4.2 for $\widehat{\boldsymbol{\beta}}_M$ or the tools developed in Salehi, Abbasi and Hassibi (2019) for M-estimators under separable regularization (i.e., the regularization term can be written as $f(\boldsymbol{\beta}) = \sum_{i=1}^p f_i(\beta_i)$).

In the following, we provide a conjecture for the exact asymptotics of $\widehat{\boldsymbol{\beta}}_\infty$ with non-informative synthetic data; a similar conjecture can also be obtained in the case with informative auxiliary data (see Appendix A.6).

Recall that the asymptotic behavior of $\widehat{\boldsymbol{\beta}}_M$ is tracked by the solution of the system of equations (6). As a heuristic argument to derive the asymptotic characterization of $\widehat{\boldsymbol{\beta}}_\infty$, we take $m \to \infty$ in (6) and consider the following approximation. Let $Q := \kappa_1\alpha Z_1 + \sigma Z_2$ to be a shorthand notation. Using the Taylor expansion that $\operatorname{Prox}_{\frac{\tau_0\gamma}{m}\rho(\cdot)}(Q) = Q + \rho'(Q)\frac{\tau_0\gamma}{m^2} + o(\frac{1}{m})$, we have

$$\lim_{m\to\infty} \frac{-\tau_0\gamma\rho''\left(\operatorname{Prox}_{\frac{\tau_0\gamma}{m}\rho(\cdot)}(Q)\right)}{1 + \frac{\tau_0\gamma}{m}\rho''\left(\operatorname{Prox}_{\frac{\tau_0\gamma}{m}\rho(\cdot)}(Q)\right)} = -\tau_0\gamma\rho''\left(\operatorname{Prox}_{\frac{\tau_0\gamma}{m}\rho(\cdot)}(Q)\right),$$

$$\lim_{m\to\infty} m(Q - \operatorname{Prox}_{\frac{\tau_0\gamma}{m}\rho(\cdot)}(Q))^2 = 0.$$

Suppose the convergence and expectation are interchangeable, the limit of (6) becomes the following system of equations, whose solution characterizes the asymptotic behavior of $\widehat{\boldsymbol{\beta}}_\infty$:

$$
(15)\qquad \begin{cases} \dfrac{\sigma^2}{2\delta} = \mathbb{E}\left[\rho'\left(-\kappa_1 Z_1\right)\left(\kappa_1\alpha Z_1 + \sigma Z_2 - \mathrm{Prox}_{\gamma\rho(\cdot)}\left(\kappa_1\alpha Z_1 + \sigma Z_2\right)\right)^2\right], \\ 1 - \dfrac{1}{\delta} = \mathbb{E}\left[\dfrac{2\rho'\left(-\kappa_1 Z_1\right)}{1 + \gamma\rho''\left(\mathrm{Prox}_{\gamma\rho(\cdot)}\left(\kappa_1\alpha Z_1 + \sigma Z_2\right)\right)}\right] \\ \qquad\qquad - \gamma\tau_0\mathbb{E}\left[\rho''(\kappa_1\alpha Z_1 + \sigma Z_2)\right], \\ -\dfrac{\alpha}{2\delta} = \mathbb{E}\left[\rho''\left(-\kappa_1 Z_1\right)\mathrm{Prox}_{\gamma\rho(\cdot)}\left(\kappa_1\alpha Z_1 + \sigma Z_2\right)\right]. \end{cases}
$$

CONJECTURE 4.6. Consider the MAP estimator $\widehat{\boldsymbol{\beta}}_\infty$ defined in (4). Suppose Conditions 6, 7, and 8 hold. Assume the parameters $(\kappa_1, \delta, \tau_0)$ are such that the system of equations (15) has a unique solution $(\alpha_*, \sigma_*, \gamma_*)$. Then, as $p \to \infty$, for any locally-Lipschitz function $\Psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ or for the indicator function $\Psi(a, t) = \mathbf{1}\{|a/\sigma_*| \le t\}$ with any fixed $t > 0$, we have

$$
\frac{1}{p}\sum_{j=1}^{p} \Psi\left(\sqrt{p}(\widehat{\boldsymbol{\beta}}_{\infty,j} - \alpha_*\boldsymbol{\beta}_{0,j}), \sqrt{p}\boldsymbol{\beta}_{0,j}\right) \xrightarrow{\mathbb{P}} \mathbb{E}[\Psi(\sigma_* Z, \beta)],
$$

where $Z \sim N(0,1)$ independent of $\beta \sim \Pi$.

Conjecture 4.6 provides a precise characterization of the asymptotic behavior of the estimator $\widehat{\boldsymbol{\beta}}_\infty$ that is similar to the one of $\widehat{\boldsymbol{\beta}}_M$ in Theorem 4.3. Although we are not able to prove this result due to the difficulty induced by the regularization term, we provide numerical verification for the convergence of the squared error and cosine similarity in Section 4.5.3.

4.5. *Numerical illustration.* In this section, through some simulation experiments, we test the finite-sample accuracy of our theoretical results on the MAP estimator in Theorem 4.3 and Theorem 4.5. We focus on the squared error $\|\widehat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0\|_2^2$ and the cosine similarity $\frac{\langle \widehat{\boldsymbol{\beta}}_M, \boldsymbol{\beta}_0\rangle}{\|\widehat{\boldsymbol{\beta}}_M\|_2\|\boldsymbol{\beta}_0\|_2}$ and we compare the theoretical prediction on these quantities with the finite-sample counterparts. Throughout the section the synthetic sample size is set to be $M = 20p$ in all experiments and the MAP estimator is computed with tuning parameter $\tau = p\tau_0\delta$ for some sequence of values for τ_0. To get the solutions from system of equations (6), (11), and (15), we use the fixed-point iterative method (Berinde and Takens, 2007, Ch 1.2).

4.5.1. *Non-informative synthetic data.* We consider the setting in Section 4.2 where the MAP estimator is constructed with non-informative synthetic data. In the experiments, we pick different combinations of parameters δ and κ_1, and fix p at 250 so that n is 250δ. The observed data $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$ and the synthetic data $\{\boldsymbol{X}_i^*, Y_i^*\}_{i=1}^M$ are generated following the condition of Theorem 4.3. For the true coefficients $\boldsymbol{\beta}_0$, we first generate $T_j \sim t_3$ independently for each $j \in [p]$ and then set $\beta_{0j} = \frac{\kappa_1}{\sqrt{3p}}T_j$. The limiting values of the squared error and the cosine similarity are given in (7) and (8) respectively.

For $\kappa_1 = 0.5$ and $\kappa_1 = 1.5$, we plot the finite-sample averaged squared error and cosine similarity as points and we draw the limiting values as curves in Figure 1, where the x-axis shows the value of τ_0. Results for $\kappa_1 = 1$ and 2 are provided in Appendix C.2.1. In these plots, the points align well with the curves, which demonstrates that our asymptotic theory has desirable finite sample accuracy. Furthermore, the U-shaped curve of the squared error suggests that for bias-variance tradeoff, the optimal value of τ should have the same order as the dimension p, which aligns with the practical suggestion in Huang et al. (2022).

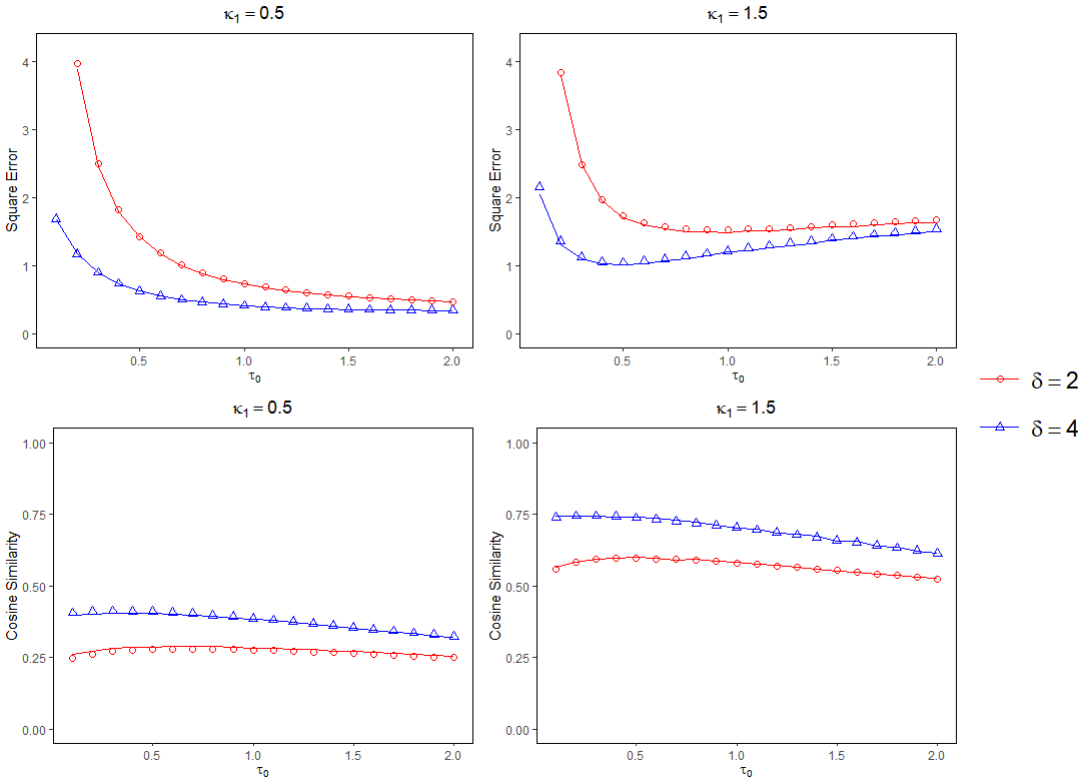

FIG 1. *Performance of the MAP estimator with non-informative synthetic data as a function of $\tau_0 = \tau/n$. Each point is obtained by calculating the performance metrics of the MAP estimator averaging over 50 simulation replications. The solid lines represent the corresponding theoretical prediction.*

4.5.2. *Informative auxiliary data.* We consider the setting in Section 4.3 where the auxiliary data are generated using regression coefficients $\boldsymbol{\beta}_s$ that have nonzero cosine similarity ξ with the true regression coefficients $\boldsymbol{\beta}_0$. In the experiments, we pick different combinations of parameters δ and κ_1, and fix $\kappa_2 = 1$, $\xi = 0.9$, and $p = 250$ so that $n = p\delta$. The observed data and true regression coefficients $\boldsymbol{\beta}_0$ are generated as in Section 4.5.1. We set $\boldsymbol{\beta}_s = \xi\frac{\kappa_2}{\kappa_1}\boldsymbol{\beta}_0 + \kappa_2\sqrt{1-\xi^2}\tilde{\boldsymbol{\varepsilon}}$ with $\xi = 0.9$, where $\tilde{\boldsymbol{\varepsilon}}$ is a random vector independent of $\boldsymbol{\beta}_0$ and the entries of $\tilde{\boldsymbol{\varepsilon}}$ are independently generated from the scale t-distribution with 3 degrees of freedom and mean zero and variance $1/p$. This particular choice guarantees that $\lim_{p\to\infty}\|\boldsymbol{\beta}_s\|_2^2 = \kappa_2^2$ and $\lim_{p\to\infty}\frac{1}{\|\boldsymbol{\beta}_0\|_2\|\boldsymbol{\beta}_s\|_2}\langle\boldsymbol{\beta}_0, \boldsymbol{\beta}_s\rangle = \xi$. Then we generate informative auxiliary data as in Condition 9. The limiting values of the squared error and the cosine similarity are given in (13) and (14) respectively.

For $\kappa_1 = 0.5$ and $\kappa = 1.5$, we plot the finite-sample averaged squared error and cosine similarity as points and we draw the limiting values as curves in Figure 1, where the x-axis shows the value of τ_0. Results for $\kappa_1 = 1$ and 2 are provided in Appendix C.2.2. In these plots, the points align well with the curves, which demonstrates that our asymptotic theory has desirable finite sample accuracy.

When compared with the experiments in Section 4.5.1, Figure 2 demonstrates that incorporating additional informative auxiliary data can significantly reduce estimation errors. For example, consider the case with parameters $(\delta = 2, \kappa_1 = 1.5)$. In Figure 1, the lowest MSE is approximately 1.5. In contrast, Figure 2 shows a reduction in this value to below 1. Similarly, we observe that the maximum cosine similarity improves from 0.6 to 0.8. These observations

indicate the effectiveness of transferring valuable information from informative auxiliary data in enhancing the estimation accuracy of the MAP estimator.

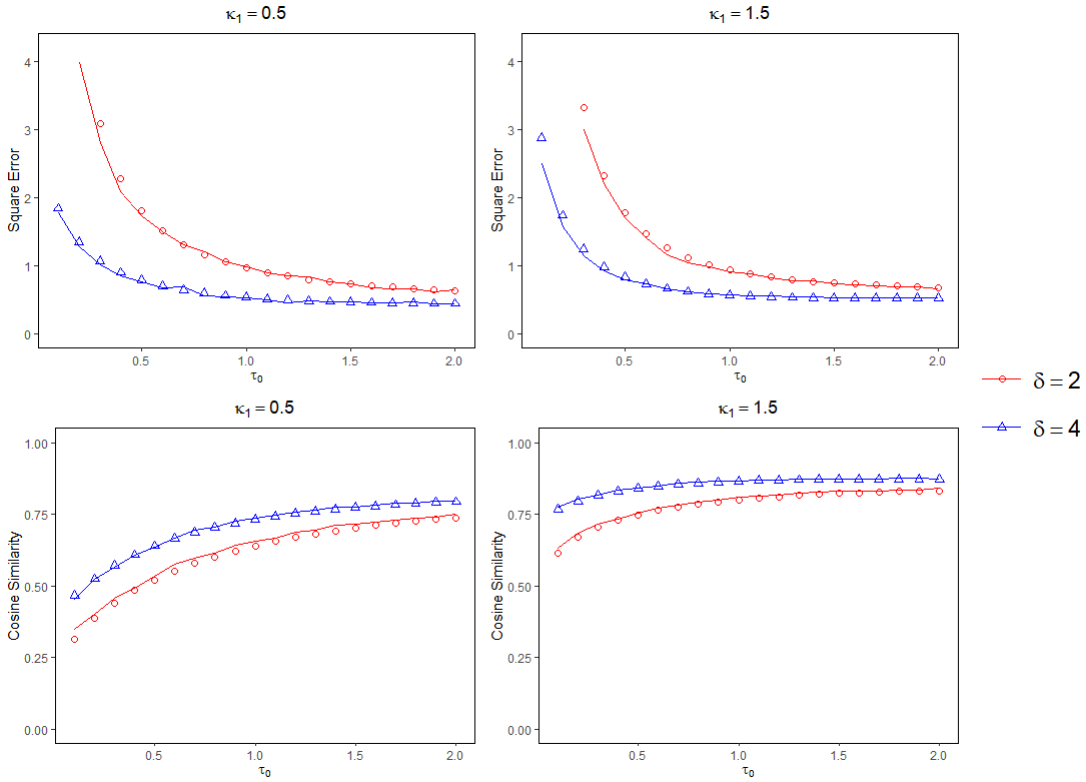

FIG 2. *Performance of the MAP estimator with informative auxiliary data* $(\kappa_2 = 1, \xi = 0.9)$ *as a function of* $\tau_0 = \tau/n$. *Each point is obtained by calculating the performance metrics of the MAP estimator averaging over 50 simulation replications. The solid lines represent the corresponding theoretical prediction.*

4.5.3. *Infinite synthetic data* . We provide a numerical verification for Conjecture 4.6 about the asymptotics of the MAP estimator $\widehat{\boldsymbol{\beta}}_\infty$ with infinite synthetic data. We consider the same experimental setting and data generation as in Section 4.5.1. To compute the $\widehat{\boldsymbol{\beta}}_\infty$, we solve the optimization (4), where the expectation has an explicit form $g(\boldsymbol{\beta}) = -\int_{-\infty}^{\infty} \rho(\|\boldsymbol{\beta}\|_2 z)\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})dz$.

Given parameters $(\kappa_1, \delta, \tau_0)$, let $(\alpha_*, \sigma_*, \gamma_*)$ be the solution of the system of equations (15). According to Conjecture 4.6, the limiting value of the squared error of $\widehat{\boldsymbol{\beta}}_\infty$ is $(\alpha_* - 1)^2\kappa_1^2 + \sigma_*^2$, and the limiting value of the cosine similarity is $\frac{\alpha_*\kappa_1}{\sqrt{\alpha_*^2\kappa_1^2+\sigma_*^2}}$. We plot the finite-sample averaged squared error and cosine similarity as points and we draw the limiting values as curves in Figure 3, where the x-axis shows the value of τ_0 with $\kappa_1 = 0.5$ and 1.5. Results for experiments with $\kappa_1 = 1$ and 2 are in Appendix C.2.3. In these plots, the points align well with the curves, which demonstrates that our conjectures on $\widehat{\boldsymbol{\beta}}_\infty$ may indeed be correct.

5. Adjusting inference based on exact asymptotics. In this section, we consider adjusting the statistical inference in the linear asymptotic regime using the theory developed in Section 4. We propose methods for estimating the unknown signal strength κ_1 and the

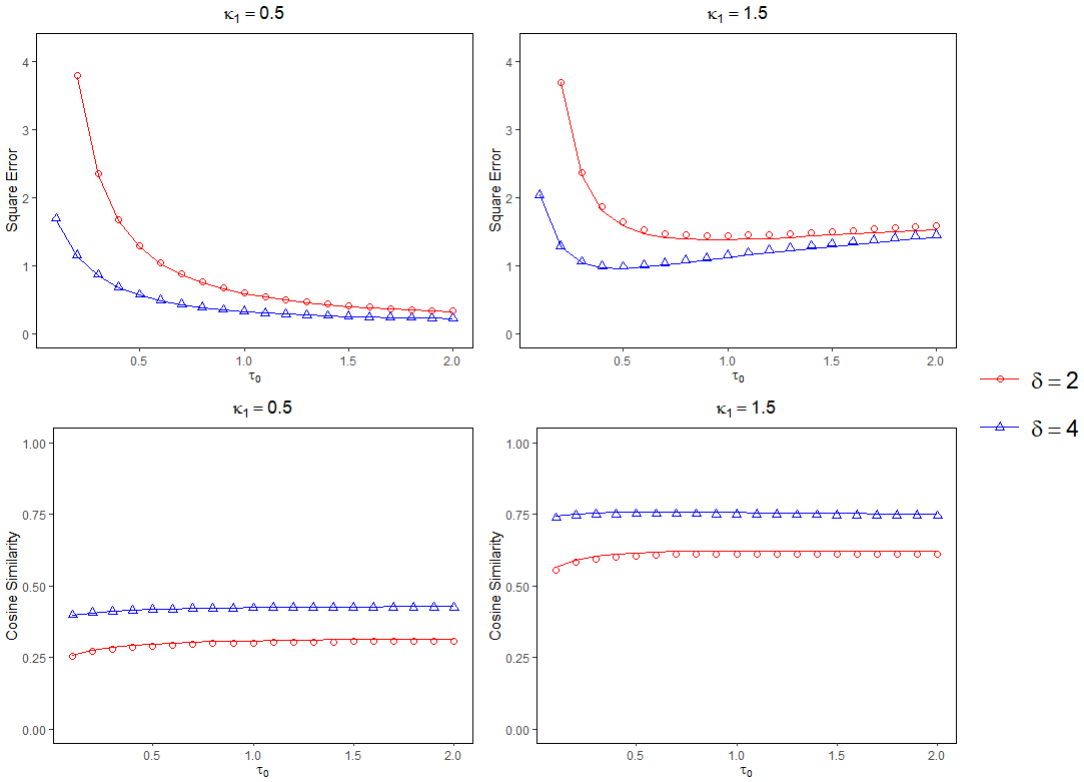

FIG 3. *Performance of the MAP estimator $\widehat{\boldsymbol{\beta}}_\infty$ as a function of $\tau_0 = \tau/n$. Each point is obtained by calculating the performance metrics of the MAP estimator averaging over 50 simulation replications. The solid lines represent the prediction by Conjecture 4.6.*

similarity ξ, and numerically evaluate their accuracy. We then consider adjusting confidence intervals and variable selection procedures using the estimated parameters.

5.1. *Estimation of signal strength.* The precise asymptotic characterization in Theorem 4.3 depends on the unknown signal strength κ_1. This section is devoted to the estimation of this parameter.

Sur and Candès (2019) has proposed a method for estimating the signal strength called *ProbeFrontier* based on an asymptotic theory of the existence of the MLE, but their method only works when $p/n < 1/2$. Our method introduced below works for any value of $p/n \in (0, 1)$.

Our method is based on the precise limit of the MAP estimator. For any given (δ, τ_0, m) and any κ_1, let $\alpha_*(\kappa_1)$ and $\sigma_*(\kappa_1)$ be from the solutions of (6). Intuitively, if the norm of the true coefficients (i.e., signal strength κ_1) increases, the norm of the MAP estimator increases accordingly. This is in light of the result proved in Candès and Sur (2020) that a large κ_1 makes the norm of the MLE unbounded. This intuition can be justified by plotting the limiting value of $\|\widehat{\boldsymbol{\beta}}_M\|_2^2$ with respect to κ_1. Theorem 4.3 suggests that the squared norm of the MAP estimator converges to $\eta_M^2 := \alpha_*^2(\kappa_1)\kappa_1^2 + \sigma_*^2(\kappa_1)$. We illustrate the relationship between η_M^2 and κ_1 in Figure 4, which suggests that η_M^2 is increasing in κ_1. We denote this relationship as $\eta_M = g_\delta(\kappa_1)$, where we omit the dependence on τ_0 and m because the values of τ_0 and m are manipulable and can be pre-chosen. Although it could be challenging to estimate κ_1 directly, it is straightforward to estimate η_M by $\widehat{\eta}_M := \|\widehat{\boldsymbol{\beta}}_M\|_2$, the norm of the MAP estimator with

non-informative synthetic data of size $M = mn$ and with total weight parameter $\tau = \tau_0 n$. Subsequently, κ_1 can be estimated by $\widehat{\kappa}_1$, which is the solution to $g_\delta(\kappa) = \widehat{\eta}_M$. This solution can be computed using a numerical evaluation of g_δ and root-finding algorithms.

Given the value of $\widehat{\kappa}_1$, the corresponding solution to the system of equations (6) will be denoted by $(\widehat{\alpha}_*, \widehat{\sigma}_*, \widehat{\gamma}_*)$. Substituting the unknown parameters in (9) with these estimates, we construct the following 95% adjusted confidence intervals

$$\widehat{\mathrm{CI}}_j = \left[\frac{\widehat{\boldsymbol{\beta}}_{M,j} - 1.96\widehat{\sigma}_*/\sqrt{p}}{\widehat{\alpha}_*}, \frac{\widehat{\boldsymbol{\beta}}_{M,j} + 1.96\widehat{\sigma}_*/\sqrt{p}}{\widehat{\alpha}_*} \right], \quad j \in [p].$$

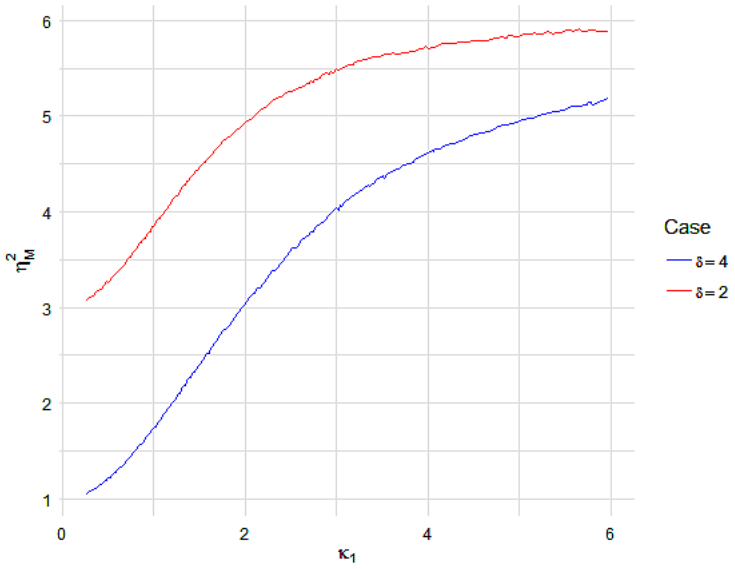

FIG 4. *Relationship between η_M^2 and κ_1 across different values of δ. For each δ, η_M^2 is computed using a grid of κ_1 values, with $\tau_0 = 1/4$ and $m = 20/\delta$.*

We demonstrate the accuracy of our estimation of κ_1 as well as the solutions (α_*, σ_*) via some empirical results. We consider the same setting described in Section 4.5.1 but examine a sequence of dimensions $p = \{100, 400, 1600\}$.

We first investigate the estimation accuracy of κ_1. The results are displayed in Table 1. From the table, it is evident that when δ and κ_1 are held constant, both the estimation error and its standard deviation decrease as p increases. This trend is expected since $\widehat{\eta}_M$ converges to its limit η_M as p increases. Given κ_1 and p, the estimation error is smaller for larger δ, since the sample size is larger. This observation aligns with the curves of $g_\delta(\cdot)$ in Figure 4, where a larger value of δ leads to a steeper slope and thus a more accurate estimate for κ_1, the solution to $g_\delta(\kappa) = \eta_M$.

Next, we investigate the estimation accuracy of (α_*, σ_*). The true values (α_*, σ_*) are presented in Table 3. We observe that the estimation errors for (α_*, σ_*) are relatively small compared to the true values, thus the estimates are quite accurate despite the estimation error of κ_1. Furthermore, we note that the errors decrease as p increases, which aligns with the observed pattern in the estimation of κ_1.

Finally, we investigate the performance of the adjusted confidence intervals based on $\widehat{\boldsymbol{\beta}}_M$. We present in Table 4 the cases with $\delta = 2$ and the cases with $\delta = 4$ are provided in Appendix C.2.4. For $\delta = 2$, the MLE does not exist and the existing methods such as the classical MLE

TABLE 1
Summary of Mean and Standard Deviation (in parentheses) of Error $|\widehat{\kappa}_1 - \kappa_1|$ based on 50 independent replications.

κ_1	p	$\delta = 2$	$\delta = 4$
0.5	100	0.363(0.315)	0.196(0.127)
	400	0.234(0.132)	0.128(0.102)
	1600	0.129(0.102)	0.060(0.045)
1	100	0.397(0.285)	0.228(0.160)
	400	0.227(0.165)	0.134(0.116)
	1600	0.104(0.114)	0.068(0.067)
1.5	100	0.426(0.325)	0.294(0.240)
	400	0.230(0.214)	0.178(0.164)
	1600	0.154(0.159)	0.103(0.091)
2	100	0.678(0.747)	0.396(0.305)
	400	0.329(0.307)	0.209(0.255)
	1600	0.201(0.214)	0.135(0.121)

TABLE 2
Summary of Mean and Standard Deviation (in parentheses) of the estimation error of true solutions of system of equations (α_*, σ_*) based on 50 independent replications.

κ_1	p	$\delta = 2$		$\delta = 4$	
		$\lvert\widehat{\alpha}_* - \alpha_*\rvert$	$\lvert\widehat{\sigma}_* - \sigma_*\rvert$	$\lvert\widehat{\alpha}_* - \alpha_*\rvert$	$\lvert\widehat{\sigma}_* - \sigma_*\rvert$
0.5	100	0.049(0.058)	0.007(0.017)	0.017(0.012)	0.006(0.004)
	400	0.028(0.020)	0.002(0.002)	0.011(0.010)	0.004(0.003)
	1600	0.015(0.011)	0.003(0.002)	0.005(0.004)	0.003(0.002)
1	100	0.066(0.051)	0.015(0.017)	0.027(0.018)	0.006(0.004)
	400	0.040(0.032)	0.009(0.008)	0.018(0.014)	0.003(0.003)
	1600	0.018(0.021)	0.006(0.006)	0.010(0.008)	0.003(0.002)
1.5	100	0.079(0.055)	0.026(0.026)	0.041(0.033)	0.004(0.003)
	400	0.044(0.040)	0.020(0.020)	0.025(0.023)	0.002(0.002)
	1600	0.029(0.029)	0.014(0.015)	0.015(0.012)	0.002(0.002)
2	100	0.110(0.092)	0.051(0.047)	0.052(0.039)	0.005(0.006)
	400	0.058(0.049)	0.031(0.028)	0.029(0.033)	0.004(0.005)
	1600	0.036(0.034)	0.021(0.018)	0.018(0.018)	0.003(0.003)

TABLE 3
Solutions of system of equations (α_*, σ_*) under different settings with non-informative synthetic data.

$\delta \backslash \kappa_1$	0.5	1	1.5	2
2	(1.004, 1.735)	(0.932, 1.726)	(0.833, 1.708)	(0.740, 1.665)
4	(0.890, 1.008)	(0.836, 1.021)	(0.773, 1.030)	(0.701, 1.031)

asymptotic confidence intervals and adjusted confidence intervals based on the MLE do not apply. In contrast, our adjusted confidence intervals achieve desirable average coverage for the true regression coefficients.

TABLE 4
Coverage rate of 95% adjusted confidence intervals based on $\widehat{\boldsymbol{\beta}}_M$ with $\delta = 2$ (MLE does not exist). Average over 50 independent experiments.

p	$\kappa_1 = 0.5$	$\kappa_1 = 1$	$\kappa_1 = 1.5$	$\kappa_1 = 2$
100	0.947	0.948	0.948	0.942
400	0.948	0.950	0.946	0.946

5.2. *Estimation of similarity.* We introduce a methodology for estimating the cosine similarity ξ between the underlying regression coefficients for two datasets. Specifically, suppose we have two independent datasets: target dataset $\{\boldsymbol{X}_{i0}, Y_{i0}\}_{i=1}^{n_0}$ and source dataset $\{\boldsymbol{X}_{is}, Y_{is}\}_{i=1}^{n_s}$, both satisfy Condition 7 with true regression coefficients $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_s$ respectively. Furthermore, we assume $\|\boldsymbol{\beta}_0\|_2 = \kappa_1$, $\|\boldsymbol{\beta}_s\|_2 = \kappa_2$, and $\frac{1}{\|\boldsymbol{\beta}_0\|\|\boldsymbol{\beta}_s\|}\langle \boldsymbol{\beta}_0, \boldsymbol{\beta}_s\rangle = \xi$.

To estimate ξ, we generate a set of independent non-informative synthetic datasets of size M for each original dataset and then construct the MAP estimator separately. For simplicity, we choose the tuning parameter τ to be the sample size times a fixed positive number τ_0. The resultant estimators are denoted by $\widehat{\boldsymbol{\beta}}_{M,0}$ for the target dataset and $\widehat{\boldsymbol{\beta}}_{M,s}$ for the source dataset. According to Theorem 4.3, asymptotically we have

$$\widehat{\boldsymbol{\beta}}_{M,0} \approx \alpha_{*1}\boldsymbol{\beta}_0 + \sigma_{*1}\boldsymbol{Z}_1,$$
$$\widehat{\boldsymbol{\beta}}_{M,s} \approx \alpha_{*2}\boldsymbol{\beta}_s + \sigma_{*2}\boldsymbol{Z}_2,$$

where $(\alpha_{*1}, \sigma_{*1})$ are solution of system (6) based on parameter $(\delta_0 = n_0/p, \kappa_1, \tau_0, M/n_0)$, and for $(\alpha_{*2}, \sigma_{*2})$ based on parameter $(\delta_s = n_s/p, \kappa_2, \tau_0, M/n_s)$ and $\boldsymbol{Z}_1, \boldsymbol{Z}_2$ are independent Gaussian vectors whose entries are independent and follow $N(0, 1/p)$. Based on this relationship, we have $\langle \widehat{\boldsymbol{\beta}}_{M,0}, \widehat{\boldsymbol{\beta}}_{M,s}\rangle \approx \alpha_{*1}\alpha_{*2} \cdot \langle \boldsymbol{\beta}_1, \boldsymbol{\beta}_2\rangle \approx \alpha_{*1}\alpha_{*2}\kappa_1\kappa_2\xi$. This leads to the following estimator for ξ:

$$\widehat{\xi} = \frac{\langle \widehat{\boldsymbol{\beta}}_{M,0}, \widehat{\boldsymbol{\beta}}_{M,s}\rangle}{\alpha_{*1}\alpha_{*2}\kappa_1\kappa_2}.$$

If κ_1 and κ_2 are unknown, they can be estimated by our method introduced in Section 5.1. In Appendix C.3, we provide a numerical illustration for the accuracy of this estimation.

5.3. *Adjusting Estimation by selection of tuning parameter.* The tuning parameter τ controls the bias-variance tradeoff for the MAP estimator defined in (3). We consider several methods for selecting the value of λ and compare the performance of the resulting estimators.

A universal strategy for selecting τ is cross-validation, which requires data-splitting and recomputing the estimator with subsets of data (Hastie, Tibshirani and Friedman, 2009, Section 7.10). Here we describe the leave-one-out cross-validation and an efficient approximation. The validation error (VE) is measured using the deviance as follows:

$$\mathrm{VE}(\tau) = -\sum_{i=1}^{n}\left\{Y_i\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_{M,-i} - \rho(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_{M,-i})\right\},$$

where $\widehat{\boldsymbol{\beta}}_{M,-i}$ denotes the MAP estimator in (3) with all observed data except the i-th observation. Since computing all $\widehat{\boldsymbol{\beta}}_{M,-i}$ is computationally intensive, it is beneficial to only compute $\widehat{\boldsymbol{\beta}}_M$ once (for each value of τ). Motivated by the leave-one-out estimators in Sur and Candès (2019), we propose an accurate approximation to $\mathrm{VE}(\tau)$. To be concrete, let $\mathcal{I}_{-i} = [n] \setminus \{i\}$ and we approximate $\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_{M,-i}$ by

$$\tilde{l}_i := \boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_M + \boldsymbol{X}_i^\top\left(H_\tau + \rho''\left(\widehat{\boldsymbol{\beta}}_M^\top\boldsymbol{X}_i\right)\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)^{-1}\boldsymbol{X}_i\left(Y_i - \rho'\left(\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}_M\right)\right),$$

where H_τ is the Hessian matrix of the objective in (3), i.e. $H_\tau = -\sum_{j\in[n]}\rho''\left(\widehat{\boldsymbol{\beta}}_M^\top\boldsymbol{X}_j\right)\boldsymbol{X}_j\boldsymbol{X}_j^\top - \frac{\tau}{M}\sum_{j\in[M]}\rho''\left(\widehat{\boldsymbol{\beta}}_M^\top\boldsymbol{X}_j^*\right)\boldsymbol{X}_j\boldsymbol{X}_j^{*\top}$. The matrix inversion in the above display can be computed efficiently using the Sherman-Morrison inverse formula Sherman and Morrison (1950).

Subsequently, we approximate $VE(\tau)$ by $\widetilde{VE}(\tau) := -\sum_{i=1}^{n}\left\{Y_i\tilde{l}_i - \rho(\tilde{l}_i)\right\}$. We provide a detailed derivation and summarize the algorithm for selecting τ by minimizing $\widetilde{VE}(\tau)$ in Appendix C.4. The MAP estimator resulting from this selection of τ is named as the **M**AP with **L**eave-one-out **C**ross **V**alidation (MLCV).

Another way to select the value of τ is to minimize the theoretical limit of the squared error given by Theorem 4.3. Consider the estimator $\widehat{\kappa}_1$ of κ_1 in Section 5.1. The corresponding solutions to the system of equations (6) for any value $\tau_0 = \tau/n$ are denoted by $(\widehat{\alpha}_*(\tau), \widehat{\sigma}_*(\tau), \widehat{\gamma}_*(\tau))$. We can then estimate the limit of the squared error by (7) for a fixed grid of values of τ and select the one that minimizes the estimated limit. The MAP estimator resulting from this selection of τ is named as the **M**AP with **E**stimated **S**quared **E**rror (MESE). For comparison, we also consider the optimal τ that minimizes the limit of the squared error based on the true value of κ_1, and name the resulting estimator as the **M**AP with **T**rue **S**quared **E**rror (MTSE).

These methods are naturally extended to cases where the MAP estimator is constructed using informative auxiliary data. More concretely, we can estimate the limit of the squared error by (13) using the estimation method for $(\kappa_1, \kappa_2, \xi)$ in Section 5.2 and we name the resulting estimator as MESE(I) where (I) represent informative auxiliary data. Similarly, we can select τ that minimizes the limit of the squared error based on the true value of $(\kappa_1, \kappa_2, \xi)$ and name the resulting estimator as MTSE(I). The procedure for leave-one-out cross-validation remains the same as before and the resulting estimator with informative auxiliary data is named MLCV(I).

We provide an experiment to illustrate these methods: MESE, MTSE, and MLCV that are based on observed data and non-informative synthetic data; MESE(I), MTSE(I), and MLCV(I) are based on observed data and informative auxiliary data. We consider the scenarios where $p = 400$, n is either $2p$ or $4p$, and κ_1 is either 1 or 2. The observed covariates and responses are generated according to the observed data generation process described in Section 4.5.1. The non-informative synthetic data are generated according to Condition 8 with $M = 20 \cdot p$. The informative auxiliary data are generated following the procedure described in Section 4.5.2 and we fix $\xi = 0.9$, $\kappa_2 = 1$, and $M = 10 \cdot p$. In each scenario, we repeat the experiments for 50 times and evaluate the squared error of each estimator.

The results across different scenarios are shown in Figure 5. In each scenario, both MESE and MLCV perform on par with the benchmark given by MTSE, which indicates that our selection methods, either using theoretical limits with estimated signal strengths or using leave-one-out cross-validation, are effective in selecting the tuning parameter τ. In addition, the performance of the estimator using informative auxiliary data is significantly superior to that using non-informative synthetic data and there is little difference among MLCV(I), MESE(I), and MTSE(I). This suggests that in the presence of informative auxiliary data, our proposed selection methods can effectively utilize the information from the auxiliary data by selecting a suitable value of τ.

5.4. *Variable selection.* Our precise asymptotic characterization of the MAP estimator can be applied to variable selection with False Discovery Rate (FDR) control using the data-splitting method introduced by Dai et al. (2023a). The original method requires the existence of the MLE on split datasets and is thus restricted. By using the MAP estimator, our extension can apply even when the MLE does not exist.

The index set of null (irrelevant) variables is denoted by S_0 and the index set of relevant variables by S_1; for logistic regression, $S_0 = \{j \in [p] : \boldsymbol{\beta}_{0,j} = 0\}$ and $S_1 = [p] \setminus S_0$. Let $\widehat{S}$ be the index set of selected variables. The False Discovery Proportion (FDP) is defined as $\text{FDP} = \frac{\#(S_0 \cap \widehat{S})}{\#\widehat{S}}$, and FDR is defined as $\mathbb{E}[\text{FDP}]$. Dai et al. (2023a) considered a variable

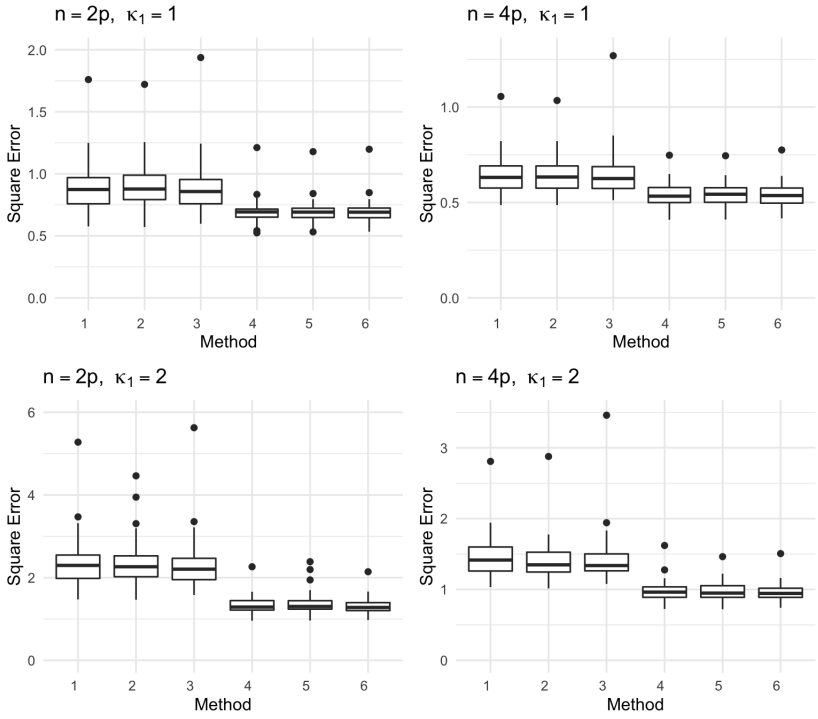

FIG 5. *The box plot is constructed from 50 independent trials. The x-axis, labeled 1,2,3,4,5,6, represents different estimators. Estimators 1 to 3 are based on non-informative synthetic data, specifically MLCV, MESE, and MTSE; while estimators 4 to 6 are based on informative auxiliary data, represented by MLCV(I), MESE(I), and MTSE(I), respectively.*

selection framework based on mirror statistics M_j's that are constructed all $j \in [p]$. A mirror statistic will have two properties: (1) a large magnitude of the mirror statistic often suggests a potentially relevant variable, and (2) the mirror statistic will be symmetric around zero when a variable is null. The first property enables us to rank the importance of variables by their associated mirror statistics so that we select variables with mirror statistics larger than a cutoff value. The second property suggests an estimated upper bound for FDP for each t, which is given by $\frac{\#\{j:M_j<-t\}}{\#\{j:M_j>t\}}$. Following these two intuitions, the cutoff with a preassigned FDR level $q \in (0,1)$ is given by

$$\text{Cutoff}(q, \{M_j\}_{j=1}^p) := \inf\left\{t > 0 : \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\}} \le q\right\},$$

and we select variables with mirror statistics greater than the above cutoff value.

To construct the mirror statistic that satisfies the above two properties, we make use of the theoretical framework in Section 4.2 for $\boldsymbol{X} \sim N(0, \boldsymbol{\Sigma})$ with general covariance. According to Corollary 4.4, for each j we have $v_j \widehat{\boldsymbol{\beta}}_{M,j} \approx v_j \alpha_* \boldsymbol{\beta}_{0,j} + \sigma_* Z_j$, where $Z_j \sim N(0, 1/p)$ and $v_j^2 = \text{Var}(X_j \mid \boldsymbol{X}_{-j})$ is the conditional variance. Adapting the data-splitting method in Dai et al. (2023b), we split the observed data into two equal-sized halves, and compute the MAP estimator for each half with separately generated synthetic data. This leads to

$$v_j \widehat{\boldsymbol{\beta}}_{M,j}^{(1)} \approx v_j \alpha_* \boldsymbol{\beta}_{0,j} + \sigma_* Z_j^{(1)} \quad \text{and} \quad v_j \widehat{\boldsymbol{\beta}}_{M,j}^{(2)} \approx v_j \alpha_* \boldsymbol{\beta}_{0,j} + \sigma_* Z_j^{(2)}, \tag{16}$$

where $(\widehat{\boldsymbol{\beta}}_{M,j}^{(1)}, Z_j^{(1)})$ is independent of $(\widehat{\boldsymbol{\beta}}_{M,j}^{(2)}, Z_j^{(2)})$ due to data splitting. (16) enables us to define the mirror statistic as $M_j := v_j^2 \widehat{\boldsymbol{\beta}}_{M,j}^{(1)} \widehat{\boldsymbol{\beta}}_{M,j}^{(2)}$, which will be large in magnitude when

$\boldsymbol{\beta}_{0,j} \neq 0$ and its distribution will symmetric around 0 when $\boldsymbol{\beta}_{0,j} = 0$. When v_j^2's are unknown, we estimate them using either node-wise regression or the diagonal entries of the inverse of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top$. To overcome the power loss due to data splitting, Dai et al. (2023b) introduced the Multiple Data-Splitting (MDS) procedure that aggregated multiple selection results via repeated sample splits; see Algorithm 2 therein.

In addition to variable selection via mirror statistics, we can consider the adjusted Benjamini-Hochberg (ABH) procedure and the adjusted Benjamini-Yekutieli (ABY) procedure (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). Both procedures rely on the adjusted p-values, which are given by $2 * \Phi(-|\widehat{v}_j \sqrt{p} \widehat{\boldsymbol{\beta}}_{M,j} / \widehat{\sigma}_*|)$ for $j \in [p]$, where $\Phi(\cdot)$ is the cumulative distribution function of standard Gaussian, $\widehat{v}_j^2$ is an estimate of the conditional variance $\mathrm{Var}(X_j \mid \boldsymbol{X}_{-j})$, and $\widehat{\sigma}_*$ is an estimate of σ_* defined in Corollary 4.4; see Appendix C.6 for such an estimation.

We conduct numerical experiments across different settings to compare the performance of the aforementioned variable selection methods based on MAP estimators in terms of FDR and power. See the caption of Figure 6 for details of the experiments. In each simulation, we numerically verified that the MLE does not exist so MLE-based methods are inapplicable in all these experiments. We have the following observations from Figure 6. When the signal strength is fixed and the correlation r of the covariate matrix is varied, the MDS procedure based on the MAP estimator effectively controls the FDR when $r \leq 0.2$ but it suffers from an inflation of FDR when $r \geq 0.3$. This is probably due to the difficulty of estimating v_j's in the presence of high correlations. In addition, ABH is more powerful than MDS in every case, although it lacks of theoretical guarantees on FDR control. On the other hand, ABY comes with a theoretical guarantee but it is too conservative and has the lowest power in every case. When r is fixed at 0.2 and the signal strength is increasing, all three methods have decreasing FDR and increasing power since it becomes easier to distinguish the relevant variables from the null ones.

To compare with the variable selection methods based on the MLE, we also reproduce the numerical experiments in Dai et al. (2023b, Section 5.1.1) where the MLE exists in each case. The results are presented in Appendix C.5 and they reveal that the selection methods based on MAP estimators perform similarly to the MLE-based methods.

6. Extension to Generalized linear model (GLM). In this section, we extend the theoretical results developed in Sections 2, 3, and 4 from the logistic regression model to the generalized linear model (GLM) with the canonical link. Let $Y \in \mathcal{Y} \subset \mathbb{R}$ be a real-valued (response) variable and $\boldsymbol{X}$ be a covariate vector of dimension p. The conditional density of Y given $\boldsymbol{X}$ is assumed to be

$$p_G(y \mid \boldsymbol{X}, \boldsymbol{\beta}_0) = h(y) \exp\left(y \boldsymbol{X}^\top \boldsymbol{\beta}_0 - b\left(\boldsymbol{X}^\top \boldsymbol{\beta}_0\right)\right), y \in \mathcal{Y}, \tag{17}$$

where $b(\theta)$ and $h(y)$ are Borel functions associated to a particular GLM. Under a catalytic prior with some synthetic data, the MAP estimator for this GLM is given by

$$\widehat{\boldsymbol{\beta}}_M^G = \arg\min_{\boldsymbol{\beta} \in \mathbf{R}^p} \sum_{i=1}^n \ell_G(Y_i, \boldsymbol{X}_i^\top \boldsymbol{\beta}) + \frac{\tau}{M} \sum_{i=1}^M \ell_G(Y_i^*, \boldsymbol{X}_i^{*\top} \boldsymbol{\beta}), \tag{18}$$

where $\ell_G(y, \theta) := b(\theta) - y\theta$ and the subscript (superscript) refers to GLM. Similarly, the MAP estimator with infinite synthetic data is given by

$$\widehat{\boldsymbol{\beta}}_\infty^G = \arg\min_{\boldsymbol{\beta} \in \mathbf{R}^p} \sum_{i=1}^n \ell_G(Y_i, \boldsymbol{X}_i^\top \boldsymbol{\beta}) + \tau \mathbb{E}\left[\ell_G(Y^*, \boldsymbol{X}^{*\top} \boldsymbol{\beta})\right],$$

where the expectation is taken over the synthetic data-generating distribution. To present our theoretical result, we begin with some conditions on the model.

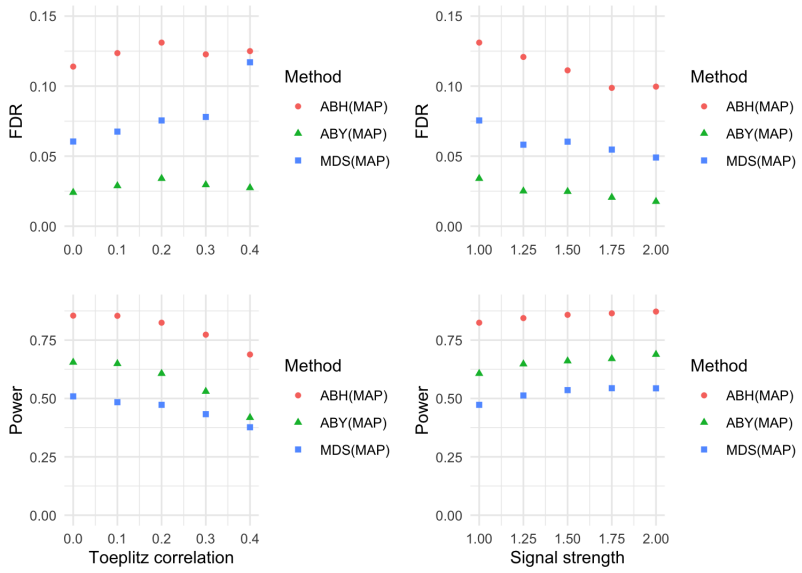

FIG 6. *Empirical FDRs and powers in a logistic regression with* $p = 200$ *and* $n = 500$. *The covariate vectors are sampled from a normal distribution* $N(0, \Sigma)$, *where* Σ *has a Toeplitz correlation structure* ($\Sigma_{ij} = r^{|i-j|}$). *The left panel varies correlation* (r) *while fixing signal strength at* $|\beta_{0j}| = 1$ *for elements in* S_1; *the right panel fixes* $r = 0.2$ *and varies signal strength from* 1 *to* 2. *In each scenario, there are 40 relevant features. The nominal FDR level is* $q = 0.1$. *The power is assessed as the proportion of correctly identified relevant features. Each point represents the average of 100 replications. The MAP estimator is computed using non-informative synthetic data with* $M = 20p$ *and* $\tau = p$.

CONDITION 10. The density function of the GLM satisfies the following:

1. For any $y \in \mathcal{Y}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, $p_G(y \mid \boldsymbol{X}, \boldsymbol{\beta}) \leq C_1$ for some universal constant C_1.
2. For any $y \in \mathcal{Y}$, the function $\ell_G(y, \theta)$ is Lipschitz-L in θ.
3. For any positive value B, there exists $c(B) > 0$ such that $b''(\theta)$ is lower bounded by $c(B)$ for all $|\theta| \leq B$.

REMARK 4. The requirements in Condition 10 are mild and commonly adopted in theoretical analysis on GLMs, as seen in Van de Geer (2008); Fan and Song (2010); Huang et al. (2020). The first requirement states that the probability density function should be bounded. The second and third requirements generalize the properties of the log-likelihood function and log partition function, respectively, in logistic regression.

For the synthetic data generation, we impose the following conditions.

CONDITION 11. The synthetic data are i.i.d. copies of $(\boldsymbol{X}^*, Y^*)$ such that the followings hold:

- The synthetic covariate vector $\boldsymbol{X}^* = \left(X_1^*, X_2^*, \cdots, X_p^*\right)$ satisfies (1) $X_1^* \equiv 1$; (2) $\mathbb{E}X_j^* = 0$, $\mathrm{Var}\left(X_j^*\right) = 1$, and $\left|X_j^*\right| \leq B_1$, a.s., for $j = 2, \cdots, p$; (3) $X_2^*, \cdots, X_p^*$ are independent.
- For the synthetic response Y^*, there are some constants $q \in (0, 1)$ and $\varsigma > 0$ such that $\min\{\mathbb{P}(Y^* \geq b'(0) + \varsigma \mid \boldsymbol{X}^*), \mathbb{P}(Y^* \leq b'(0) - \varsigma \mid \boldsymbol{X}^*)\} \geq q$.

Condition 11 is an extension of Condition 1 with no difference in the generation of synthetic covariates. The requirement on the generation of responses ensures that synthetic re-

sponses do not highly skew towards one side of the domain $\mathcal{Y}$. In logistic regression, this requirement becomes the same as in Condition 1 if we take $\varsigma = 0.5$.

Under these conditions, we can establish the following results for GLM: (1) the MAP estimate $\widehat{\boldsymbol{\beta}}_M^G$ exists and is unique; (2) the MAP estimator is stable against finite M in the sense that $\|\widehat{\boldsymbol{\beta}}_M^G - \widehat{\boldsymbol{\beta}}_\infty^G\|^2 = O_p\left(\frac{p}{M\gamma^2}\right)$; (3) in the regime that p diverges but $p/n \to 0$, the MAP estimator is consistent in the sense that $\|\widehat{\boldsymbol{\beta}}_M^G - \boldsymbol{\beta}_0\|^2 = O_p(p/n)$ under Condition 3, Condition 4, and $\mathbb{E}\|\operatorname{Var}(Y_i|\boldsymbol{X}_i)X_i\|^2 \le C_2 p$ for some universal constant $C_2 > 0$; (4) in the linear asymptotic regime, $\|\widehat{\boldsymbol{\beta}}_M^G\|_2$ is bounded with high probability; (5) under Gaussian design, we precisely characterize the asymptotic behaviour of the MAP estimator, which roughly states that $\widehat{\boldsymbol{\beta}}_M^G \approx \alpha_G \boldsymbol{\beta}_0 + p^{-1/2}\sigma_G \boldsymbol{Z}$, where $\boldsymbol{Z}$ is a standard normal vector and α_G and σ_G are constants similar to the ones in (5). The details and proofs of these results are presented in Appendix B.

7. Discussion. This paper studies the theoretical properties of the maximum a posteriori (MAP) estimator under a catalytic prior. This is a regularized maximum likelihood estimator whose regularization term is based on synthetic data. Our analyses apply to logistic regression and other generalized linear models under some regularity conditions. We establish the existence and uniqueness of the MAP estimator, and show it is stable against the random synthetic data. Additionally, we also investigate the large-sample properties of the MAP estimator with both the sample size n and the dimension p diverging to infinity. We show that it achieves the minimax optimal rate for estimation if the tuning parameter τ is chosen proportional to p. Furthermore, we establish a precise characterization of the asymptotic behavior of the MAP estimator when p/n converges to a constant. This result clarifies the roles of the hyper-parameters of the catalytic prior, namely, the tuning parameter τ and the synthetic sample size M, in the asymptotic performance of the MAP estimator. We extend our analysis to the regularized MLE using informative auxiliary data, which is a simple but effective transfer learning approach. Our results reveal how useful information contained in auxiliary data leads to improved inference. On the methodology front, we propose estimation methods for the key quantities that govern the asymptotic behavior of the MAP estimator and we apply our theory to tune the value of τ, construct confidence intervals, and select important variables.

There are some open questions related to the findings in this paper. Firstly, our estimator of the signal strength is numerically accurate for estimating the true value, but there is no theoretical guarantee yet. It would be of interest to investigate its consistency. Secondly, our numerical experiments suggest that the precise asymptotic characterization given in Theorem 4.3 and Theorem 4.5 continue to hold if the Gaussian design condition is replaced by a moment condition. Although it seems promising to establish this universality thanks to recent developments in the literature, it is technically challenging and beyond the scope of the current work. Thirdly, when informative auxiliary data from multiple sources are available for constructing the estimator, the asymptotic behavior may be characterized by extending Theorem 4.5. Our proving strategy may be useful but a direct extension will make the system of equations too complicated to be useful. An exploration into this extension requires more future efforts. Lastly, the idea of regularizing the MLE by synthetic data can be generalized to other M-estimation for general models. We expect that our analysis will shed light on future developments.

REFERENCES

ALBERT, A. and ANDERSON, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10.

BASTANI, H. (2021). Predicting with proxies: Transfer learning in high dimension. *Management Science* **67** 2964–2984.

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.

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.

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* 1165–1188.

BERINDE, V. and TAKENS, F. (2007). *Iterative approximation of fixed points* **1912**. Springer.

BICKEL, P. J., LI, B., TSYBAKOV, A. B., VAN DE GEER, S. A., YU, B., VALDÉS, T., RIVERO, C., FAN, J. and VAN DER VAART, A. (2006). Regularization in statistics. *Test* **15** 271–344.

BIRNBAUM, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association* **57** 269–306.

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

CANDÈS, E. J. and SUR, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics* **48** 27–42.

CELENTANO, M., MONTANARI, A. and WEI, Y. (2023). The lasso with general gaussian designs with applications to hypothesis testing. *The Annals of Statistics* **51** 2194–2220.

CHEN, K., HU, I. and YING, Z. (1999). Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics* **27** 1155–1163.

CHEN, M.-H., IBRAHIM, J. G. and SHAO, Q.-M. (2000). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference* **84** 121–137.

CHEN, X., ADITYAN, GUNTUBOYINA and ZHANG, Y. (2016). On Bayes Risk Lower Bounds. *Journal of Machine Learning Research* **17** 1–58.

COVER, T. M. and THOMAS, J. A. (2012). *Elements of Information Theory*. John Wiley & Sons.

DAI, C., LIN, B., XING, X. and LIU, J. S. (2023a). A scale-free approach for false discovery rate control in generalized linear models. *Journal of the American Statistical Association* 1–15.

DAI, C., LIN, B., XING, X. and LIU, J. S. (2023b). False discovery rate control via data splitting. *Journal of the American Statistical Association* **118** 2503–2520.

DANSKIN, J. M. (1966). The theory of max-min, with applications. *SIAM Journal on Applied Mathematics* **14** 641–664.

DEMPSTER, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)* **30** 205–232.

DENG, Z., KAMMOUN, A. and THRAMPOULIDIS, C. (2022). A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA* **11** 435–495.

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.

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.

EL KAROUI, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields* **170** 95–175.

EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences* **110** 14557–14562.

FAHRMEIR, L. and KAUFMANN, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* **13** 342–368.

FAN, J. and SONG, R. (2010). SURE INDEPENDENCE SCREENING IN GENERALIZED LINEAR MODELS WITH NP-DIMENSIONALITY. *The Annals of Statistics* **38** 3567–3604.

HAN, Q. and SHEN, Y. (2023). Universality of regularized regression estimators in high dimensions. *The Annals of Statistics* **51** 1799–1823.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

HE, X. and SHAO, Q.-M. (2000). On parameters of increasing dimensions. *Journal of multivariate analysis* **73** 120–135.

HUANG, D., STEIN, N., RUBIN, D. B. and KOU, S. (2020). Catalytic prior distributions with application to generalized linear models. *Proceedings of the National Academy of Sciences* **117** 12004–12010.

HUANG, D., WANG, F., RUBIN, D. B. and KOU, S. (2022). Catalytic Priors: Using Synthetic Data to Specify Prior Distributions in Bayesian Analysis. *arXiv preprint arXiv:2208.14123*.

IBRAHIM, J. G. and CHEN, M.-H. (2000). Power prior distributions for regression models. *Statistical Science* 46–60.

IBRAHIM, J. G., CHEN, M.-H. and SINHA, D. (2003). On optimality properties of the power prior. *Journal of the American Statistical Association* **98** 204–213.

JAVANMARD, A. and SOLTANOLKOTABI, M. (2022). Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics* **50** 2127–2156.

KONIS, K. (2007). Linear programming algorithms for detecting separated data in binary logistic regression models, PhD thesis, University of Oxford.

LAI, T. L. and WEI, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics* **10** 154–166.

LEE, K.-Y. and COURTADE, T. A. (2020). Linear models are most favorable among generalized linear models. In *2020 IEEE International Symposium on Information Theory (ISIT)* 1213–1218. IEEE.

LI, S., CAI, T. T. and LI, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84** 149–173.

LI, S., CAI, T. T. and LI, H. (2023). Transfer learning in large-scale gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association* **118** 2171–2183.

LI, W. and HUANG, D. (2023). Bayesian inference on Cox regression models using catalytic prior distributions. *arXiv preprint arXiv:2312.01411*.

LI, S., ZHANG, L., CAI, T. T. and LI, H. (2023). Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association* 1–12.

LIANG, H. and DU, P. (2012). Maximum likelihood estimation in logistic regression models with a diverging number of covariates. *Electronic Journal of Statistics* **6** 1838–1846.

MCDIARMID, C. et al. (1989). On the method of bounded differences. *Surveys in combinatorics* **141** 148–188.

ORTEGA, J. and RHEINBOLDT, W. (1970). *Iterative Solution of Nonlinear Equations in Several Variables* **30**. SIAM.

PORTNOY, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p 2/n is large. I. Consistency. *The Annals of Statistics* 1298–1309.

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics* 356–366.

PRATT, J. W., RAIFFA, H. and SCHLAIFER, R. (1964). The foundations of decision under uncertainty: An elementary exposition. *Journal of the American statistical association* **59** 353–375.

REEVE, H. W., CANNINGS, T. I. and SAMWORTH, R. J. (2021). Adaptive transfer learning. *The Annals of Statistics* **49** 3618–3649.

ROCKAFELLAR, R. T. and WETS, R. J.-B. (2009). *Variational analysis* **317**. Springer Science & Business Media.

SALEHI, F., ABBASI, E. and HASSIBI, B. (2019). The impact of regularization on high-dimensional logistic regression. *Advances in Neural Information Processing Systems* **32**.

SHERMAN, J. and MORRISON, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics* **21** 124–127.

SILVAPULLE, M. J. and BURRIDGE, J. (1986). Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *Journal of the Royal Statistical Society: Series B (Methodological)* **48** 100–106.

SION, M. (1958). On general minimax theorems. *Pacific J. Math.* **8** 171–176.

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.

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.

TAHERI, H., PEDARSANI, R. and THRAMPOULIDIS, C. (2020). Sharp asymptotics and optimal performance for inference in binary models. In *International Conference on Artificial Intelligence and Statistics* 3739–3749. PMLR.

THRAMPOULIDIS, C. (2016). Recovering structured signals in high dimensions via non-smooth convex optimization: Precise performance analysis, PhD thesis, California Institute of Technology.

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.

THRAMPOULIDIS, C., OYMAK, S. and HASSIBI, B. (2014). The Gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*.

THRAMPOULIDIS, C., OYMAK, S. and HASSIBI, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory* 1683–1709. PMLR.

TIAN, Y. and FENG, Y. (2023). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association* **118** 2684–2697.

TORREY, L. and SHAVLIK, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques* 242–264. IGI global.

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Annals of statistics* **36** 614–645.

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

WAINWRIGHT, M. J. (2014). Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application* **1** 233–253.

WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge university press.

WANG, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *Annals of Statistics* **39** 389–417.

ZHANG, H. and LI, H. (2023). Transfer Learning with Random Coefficient Ridge Regression. *arXiv preprint arXiv:2306.15915*.

ZHAO, Q. (2020). Glmhd: Statistical inference in high-dimensional binary regression R package version 0.0.0.9000.

ZHAO, Q., SUR, P. and CANDES, E. J. (2022). The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance. *Bernoulli* **28** 1835–1861.