Laplace Approximation for Bayesian Variable Selection via Le Cam's One-Step Procedure

Tianrui Hou
Boston University

Yves Atchade
Boston University

Abstract

The selection of relevant features in high-dimensional settings is a central challenge in modern scientific research and decision-making. Although many existing methods offer strong statistical guarantees, they are often computationally intractable in high-dimensional problems. To address this issue, we introduce a novel Laplace approximation method based on Le Cam's one-step procedure, termed OLAP. This approach is specifically designed to alleviate computational burdens while maintaining statistical rigor. Under standard high-dimensional assumptions, we establish that OLAP achieves consistent variable selection. Moreover, the method yields a posterior distribution that can be efficiently explored in polynomial time with a Gibbs sampling algorithm. We demonstrate the effectiveness of OLAP through applications to logistic and Poisson regression models, using both simulated and real data.

1 INTRODUCTION

Feature/variable selection methods are crucial tools in the advancement of many data-driven problems (O'Hara and Sillanpää, 2009; Fan and Lv, 2010). However, existing methods that are known to be statistically consistent in high-dimensional regimes are computationally expensive to deploy. As a result, these methods are often implemented using greedy searches or other crude approximations lacking convergence guarantees (Freedman et al., 1992; Leng et al., 2006; Chen and Chen, 2008; Luo and Chen, 2011; Barber et al., 2016). We propose a new Laplace approximation that yields a variable selection procedure that is both consistent and computationally inexpensive to implement. Although our methodology applies more generally, we focus the presentation on a class of generalized linear regression (GLM) models with known dispersion parameters. Specifically, we assume that we have a data set $\mathcal{D} \stackrel{\text{def}}{=} \{(y_i, \mathbf{x}_i),\ 1 \le i \le n\}$, with random response $y_i \in \mathbb{R}$, and non-random explanatory variable $\mathbf{x}_i \in \mathbb{R}^p$ collected from n independent units. We consider a GLM model where, up to an additive constant that we ignore, the log-likelihood function is given by

$$\ell(\theta; \mathcal{D}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} y_i \langle \theta, \mathbf{x}_i \rangle - \psi\left(\langle \theta, \mathbf{x}_i \rangle\right), \quad \theta \in \mathbb{R}^p, \tag{1}$$

for some known function $\psi : \mathbb{R} \to \mathbb{R}$, and where $\theta \in \mathbb{R}^p$ is the parameter of interest. We assume throughout that the model is well-specified, with a true value of the parameter $\theta_\star \in \mathbb{R}^p$ that satisfies

$$\mathbb{E}\left(y_i - \psi'(\langle \mathbf{x}_i, \theta_\star \rangle)\right) = 0,$$

where ψ' denotes the derivative of ψ. We consider the classical high-dimensional scenario where p is potentially much larger than the sample size n, but $\theta_\star$ is believed to be sparse. The variable selection problem in this context corresponds to finding the support of $\theta_\star$. We take a Bayesian approach and introduce an additional variable $\delta \in \Delta \stackrel{\text{def}}{=} \{0,1\}^p$ that encodes the support of θ, with prior distribution

$$\pi(\delta) \propto \left(\frac{1}{p^{\mathsf{u}}}\right)^{\|\delta\|_0}, \quad \delta \in \Delta, \tag{2}$$

for some user-defined parameter $\mathsf{u} > 0$. Furthermore, in our prior specification, we assume that given δ the components of θ are independent, and $\theta_j | \{\delta_j = 1\}$ follows a Gaussian prior $\mathbf{N}(0,1)$, whereas $\theta_j | \{\delta_j = 0\}$ follows a degenerate point-mass distribution with full mass at 0. The variable selection problem therefore boils down to sampling from (or computing the mode

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

of) the marginal posterior distribution of δ given the data $\mathcal{D}$ given by

$$\Pi(\delta|\mathcal{D}) \propto \left(\frac{1}{p^{\mathsf{u}}}\sqrt{\frac{1}{2\pi}}\right)^{\|\delta\|_0} \int_{\mathbb{R}^{\|\delta\|_0}} \exp\left(-\frac{1}{2}\|\vartheta\|_2^2 + \ell^\delta(\vartheta;\mathcal{D})\right) \mathrm{d}\vartheta,$$
$$\text{where} \quad \ell^\delta(\vartheta;\mathcal{D}) \stackrel{\text{def}}{=} \ell((\vartheta,0)_\delta;\mathcal{D}), \quad (3)$$

and where for $w \in \mathbb{R}^{\|\delta\|_0}$, the notation $(w,0)_\delta$ denotes the vector of $\mathbb{R}^p$ obtained from w by adding 0 at components where $\delta_j = 0$ (see Section 1.3 below for a more precise definition). When the log-likelihood function ℓ is quadratic, the integrals in (3) have an explicit form, and it is possible to sample from (3) using off-the-shelves MCMC methods on Δ. This approach, sometimes with slightly different priors, has been a cornerstone of Bayesian variable selection for several decades (George and McCulloch, 1993; O'Hara and Sillanpää, 2009; Guan and Stephens, 2011; Yang et al., 2016).

The challenge is that this marginalization strategy is generally not applicable beyond the quadratic loss case. One common solution is data augmentation that samples jointly a model δ together with its corresponding parameter, using trans-dimensional MCMC algorithms (Green, 2003; Lamnisos et al., 2009). However, this strategy is known to scale poorly. A related idea developed in (Carlin and Chib, 1995; Atchade and Bhattacharyya, 2019), consists in complementing the parameter set of each model δ with additional variables drawn from the so-called pseudo-prior distribution. In the same vein, several carefully constructed priors that are absolutely continuous with respect to the Lebesgue measure have also been proposed and studied in the literature (Polson and Scott, 2011; Narisetty and He, 2014; Rockova and George, 2018); see (Bhadra et al., 2019) for a review. These different approaches circumvent the intractable integrals in (3), however, their implementations require more sophisticated and costly MCMC algorithms.

Another line of work that addresses the intractability of the integrals in (3) is the Laplace approximation. Specifically, the idea is to replace the function $\ell^\delta(\cdot;\mathcal{D})$ by its second-order Taylor expansion around the MLE estimator[1]

$$\hat{\theta}^\delta \stackrel{\text{def}}{=} \underset{w\in\mathbb{R}^{\|\delta\|_0}}{\mathsf{Argmax}} \left[-\frac{1}{2}\|w\|_2^2 + \ell^\delta(w;\mathcal{D})\right]. \quad (4)$$

[1]The addition of a quadratic penalty in (4) is mainly for improved numerical stability. Our method and analysis remain valid if $\hat{\theta}^\delta$ is replaced by the actual MLE $\mathsf{Argmax}_{w\in\mathbb{R}^{\|\delta\|_0}} \ell^\delta(w;\mathcal{D})$, provided that the Hessian matrices of $\ell^\delta(\cdot;\mathcal{D})$ remain always positive definite. For this reason, we will use the term MLE, even though (4) is strictly speaking a penalized MLE.

The modes of the resulting posterior approximation are known to be equivalent to BIC (Schwarz, 1978), or e-BIC (Chen and Chen, 2008). In (Barber et al., 2016) the authors show that under some regularity conditions, and for a class of generalized linear models, replacing the integrals in (3) with their Laplace approximations still produces consistent variable selection. The major limitation of this approach is the high per-iteration cost, as one must compute the estimator $\hat{\theta}^\delta$ (typically using an iterative optimization solver) as well as the Hessian matrix of ℓ^δ at $\hat{\theta}^\delta$. Since a large number of MCMC iterations is needed, the overall cost of this approach is substantial.

In the frequentist realm, the high-dimensional variable selection problem in a GLM model with likelihood such as (1) is often framed as the best-subset selection problem that solves the minimization problem

$$\min -\ell(\theta;\mathcal{D}) \quad \text{subject to} \quad \|\theta\|_0 \le s,$$

or some relaxation thereof, for a specified sparsity level s (Beck and Eldar, 2013; Hastie et al., 2015). These estimators have seen renewed interest due to recent optimization advances to tackle an otherwise NP-hard problem (Bertsimas et al., 2016; Hazimeh and Mazumder, 2020). One drawback of this approach is the limited understanding of the statistical properties of the resulting estimators and the lack of effective and adaptive methods to select the sparsity level s.

1.1 Main Contributions

We revisit the Laplace approximation method. We develop a variant of the method – based on Le Cam's one-step procedure – that circumvents the aforementioned challenges. The one-step procedure due to L. Le Cam (Lecam, 1969; Vaart, 1998) is a well-known scheme to improve a given estimator through a one-step Newton-Raphson update.

Hence, starting from an initial estimator $\widetilde{\theta}$ that is computed once, we propose fast approximations of the estimators $\hat{\theta}^\delta$ using Le Cam's one-step procedure. This leads to a new Laplace approximation that we call a one-step Laplace approximation (OLAP). Under some basic high-dimension regularity conditions (basically restricted strong convexity and limited coherence of Hessian matrices) we show that *the method achieves the same statistical performance as the standard Laplace approximation with knowledge of the estimators* $\hat{\theta}^\delta$, while being an order of magnitude faster to implement. Furthermore, we analyze the mixing time of a simple Gibbs sampler to sample from the posterior distribution of OLAP. Under the same regularity assumptions mentioned above, we show that the algorithm has a mixing time that scales polynomially

with (n, p), effectively leading to a polynomial-time solution to the feature selection problem. We conduct an extensive numerical comparison that highlights the effectiveness of OLAP.

Our method is closely related to Rossell et al. (2021), which was brought to our attention after the release of an initial version of this paper. Rossell et al. (2021) tackled the same intractable marginal likelihood in model/variable selection as us and proposed a very similar solution, devising a Laplace approximation from few iterations of the Newton method started from an initial estimate. However Rossell et al. (2021) did not make the connection to Le Cam's one-step method, and their initial estimator is actually the zero vector (null model). This leads to a very different estimator and a very different analysis than ours. Indeed, in Le Cam's one-step framework the quality of the initial estimator is crucial, and this rules out the zero vector. Another key contribution of our work not addressed by Rossell et al. (2021) is the mixing time analysis of the MCMC algorithm. Our numerical results include a comparison to Rossell et al. (2021).

1.2 Outline

We end this introduction with some general notation. The one-step Laplace approximation of (3) and some basic statistical and computational properties are derived in Section 2. Numerical illustrations of the methods are presented in Section 3, and all technical proofs are collected in the Supplement. We close the paper with a brief discussion in Section 4.

1.3 Notations

Throughout our parameter space is $\mathbb{R}^p$ equipped with its Euclidean norm $\|\cdot\|_2$ and inner product $\langle\cdot,\cdot\rangle$. We also use $\|\cdot\|_0$ which counts the number of non-zero elements, and $\|\cdot\|_\infty$ which returns the largest absolute value. We set $\Delta \stackrel{\text{def}}{=} \{0,1\}^p$ for the model space. For $\delta, \delta' \in \Delta$, we define $\min(\delta, \delta')$ as the component-wise minimum of the vectors δ and δ'. And we write $\delta \subseteq \delta'$ if $\min(\delta, \delta') = \delta$, and we write $\delta \supseteq \delta'$ if $\delta' \subseteq \delta$.

Given $\delta \in \Delta$, and $\theta \in \mathbb{R}^p$, we write θ_δ to denote the component-wise product of θ and δ, and $\delta^c \stackrel{\text{def}}{=} 1 - \delta$. Assuming $\|\delta\|_0 = s$, we will use the notation $[\theta]_\delta$ to denote the vector $(\theta_{j_1}, \ldots, \theta_{j_s})$, where j_i is the i-th component of δ that is non-zero. Conversely, for $w \in \mathbb{R}^s$, we define $(w, 0)_\delta$ as the element of $\mathbb{R}^p$ such that $[(w, 0)_\delta]_\delta = w$. At times we will abuse the notation and use θ_δ and $[\theta]_\delta$ interchangeably.

2 THE ONE-STEP LAPLACE APPROXIMATION

With ℓ^δ as in (3), we define

$$\bar{\ell}^\delta(w; \mathcal{D}) \stackrel{\text{def}}{=} \ell^\delta(w; \mathcal{D}) - \frac{1}{2}\|w\|_2^2, \;\; w \in \mathbb{R}^{\|\delta\|_0}.$$

The Laplace approximation of the posterior distribution $\Pi(\cdot|\mathcal{D})$ is formed by replacing $\bar{\ell}^\delta$ in the posterior (3) by its second-order Taylor approximation around the mode $\hat{\theta}^\delta$, where $\hat{\theta}^\delta$ is as defined in (4). The resulting Gaussian integral has an explicit form, which leads to the approximation of (3) given by

$$\hat{\Pi}(\delta|\mathcal{D}) \propto \left(\frac{1}{p^{\mathsf{u}}}\right)^{\|\delta\|_0} \frac{e^{\bar{\ell}^\delta(\hat{\theta}^\delta; \mathcal{D})}}{\sqrt{\det\left(\hat{\mathcal{H}}^\delta\right)}},$$
$$\text{where} \quad \hat{\mathcal{H}}^\delta \stackrel{\text{def}}{=} -\nabla^{(2)}\bar{\ell}^\delta(w; \mathcal{D})|_{w=\hat{\theta}^\delta}. \quad (5)$$

(Barber et al., 2016) studied the statistical properties of $\hat{\Pi}$ and showed that it is consistent (in the sense that it puts a high probability on the true model) under some regularity conditions, such as $n, p \to \infty$. However, as observed above, in general, the estimator $\hat{\theta}^\delta$ is computationally expensive to obtain, making the approach difficult to use in large-scale applications.

To derive our method OLAP, we first note that for any $v_0 \in \mathbb{R}^{\|\delta\|_0}$, using a second-order Taylor approximation of the function $\bar{\ell}^\delta(\cdot; \mathcal{D})$ around v_0, we can approximate the integral $\int_{\mathbb{R}^{\|\delta\|_0}} e^{\bar{\ell}^\delta(w;\mathcal{D})} \mathrm{d}w$ by

$$\frac{(2\pi)^{\frac{\|\delta\|_0}{2}}}{\sqrt{\det(\mathcal{H}^{v_0})}} \exp\left(\bar{\ell}(v_0; \mathcal{D}) + \frac{1}{2}(\mathcal{G}^{v_0})^{\mathsf{T}}(\mathcal{H}^{v_0})^{-1}\mathcal{G}^{v_0}\right). \quad (6)$$

In the above expression, $\mathcal{G}^{v_0} \stackrel{\text{def}}{=} \nabla\bar{\ell}^\delta(w; \mathcal{D})|_{w=v_0}$, and $\mathcal{H}^{v_0} \stackrel{\text{def}}{=} -\nabla^{(2)}\bar{\ell}^\delta(w; \mathcal{D})|_{w=v_0}$. The approximation performs at its best when v_0 is near $\hat{\theta}^\delta$, and this is what the standard Laplace approximation does. Let $\widetilde{\theta} \in \mathbb{R}^p$ be some initial estimator of $\theta_\star$. The estimator $\widetilde{\theta}$ is computed only once, and for any $\delta \in \Delta$, we can take $\widetilde{\theta}^\delta \stackrel{\text{def}}{=} [\widetilde{\theta}]_\delta$ (where $[\widetilde{\theta}]_\delta$ is the sub-vector of $\tilde{\theta}$ that collects the components of $\tilde{\theta}$ for which $\delta_j = 1$) as an initial approximation of $\hat{\theta}^\delta$, which we then improve upon using the one-step procedure. Hence, our proposed approximation of $\hat{\theta}^\delta$ is

$$\check{\theta}^\delta \stackrel{\text{def}}{=} \tilde{\theta}^\delta + (\tilde{\mathcal{H}}^\delta)^{-1}\tilde{\mathcal{G}}^\delta, \quad (7)$$

where

$$\tilde{\mathcal{G}}^\delta \stackrel{\text{def}}{=} \nabla\bar{\ell}^\delta(w; \mathcal{D})|_{w=\tilde{\theta}^\delta}$$
$$\tilde{\mathcal{H}}^\delta \stackrel{\text{def}}{=} -\nabla^{(2)}\bar{\ell}^\delta(w; \mathcal{D})|_{w=\tilde{\theta}^\delta}. \quad (8)$$

The estimator $\check{\theta}^\delta$ is obtained from $\tilde{\theta}^\delta$ by a single Newton-Raphson (NR) update. In statistical estimation theory, it is well known that a single-step NR update is enough to improve any rate-optimal initial estimator into an efficient rate-optimal estimator (Vaart, 1998; Brouste et al., 2021). Versions of the procedure have also appeared in the recent statistics literature to de-bias and otherwise improve upon classical high-dimensional estimators (van de Geer et al., 2014; Javanmard and Montanari, 2014; Xia et al., 2020).

Using $\check{\theta}^\delta$ and (6) leads to the approximation of Π given by

$$\delta \mapsto \left(\frac{1}{p^{\mathsf{u}}}\right)^{\|\delta\|_0} \frac{\exp\left(\bar{\ell}^\delta(\check{\theta}^\delta;\mathcal{D}) + \frac{1}{2}(\check{\mathcal{G}}^\delta)^\intercal(\check{\mathcal{H}}^\delta)^{-1}\check{\mathcal{G}}^\delta\right)}{\sqrt{\det\left(\check{\mathcal{H}}^\delta\right)}}, \quad \delta \in \Delta, \quad (9)$$

where $\check{\mathcal{G}}^\delta$ and $\check{\mathcal{H}}^\delta$ are defined as in (8), but with $\tilde{\theta}^\delta$ replaced by $\check{\theta}^\delta$. We note that if, instead of a single update, we iterate (7) a large number of times, then naturally $\check{\theta}^\delta \approx \hat{\theta}^\delta$, and $\check{\mathcal{G}}^\delta \approx 0$, and (9) becomes (5). Hence, we propose to drop the term $(\check{\mathcal{G}}^\delta)^\intercal(\check{\mathcal{H}}^\delta)^{-1}\check{\mathcal{G}}^\delta$, as well as the log-determinant term in (9), leading to OLAP, our proposed Laplace approximation distribution over Δ:

$$\check{\Pi}(\delta|\mathcal{D}) \propto \left(\frac{1}{p^{\mathsf{u}}}\right)^{\|\delta\|_0} \exp\left(\bar{\ell}^\delta(\check{\theta}^\delta;\mathcal{D})\right). \quad (10)$$

We then solve the variable selection problem by sampling from the OLAP distribution (10).

Remark 2.1. It is very important to reiterate here that we have designed OLAP specifically with the high-dimensional regime ($p \geq n$ or $p \approx n$) in mind. In more classical settings where n is much larger than p, the log-determinant term $\log\det(\check{\mathcal{H}}^\delta)/2 \approx \|\delta\|_0 \log(n)/2$ that we have ignored in (10) becomes the correct penalty term, and should be included in (10). In the high-dimensional regime that is our focus, the penalty term needed is $\mathsf{u}\|\delta\|_0 \log(p)$.

2.1 Initial Estimator $\tilde{\theta}$

Throughout we take the initial estimator $\tilde{\theta} \in \mathbb{R}^p$ as either lasso, the ridge regression estimator, or more generally the elastic-net estimator defined as

$$\underset{\theta\in\mathbb{R}^p}{\mathsf{Argmin}} \left[\sum_{i=1}^n -y_i \langle\theta, \mathbf{x}_i\rangle + \psi\left(\langle\theta, \mathbf{x}_i\rangle\right) + \lambda_1\|\theta\|_1 + \frac{\lambda_2}{2}\|\theta\|_2^2\right],$$

for some well-tuned regularization parameters $\lambda_1 \geq 0$, $\lambda_2 \geq 0$. We refer the reader to (Hastie et al., 2015) and the extensive literature on these high-dimensional estimators.

2.2 Statistical Properties

We recall that we have assumed that our data $\mathcal{D} = \{(y_i, \mathbf{x}_i),\ 1 \leq i \leq n\}$ is generated from a well-specified GLM model, with the true value of the parameter $\theta_\star$ with support $\delta_\star$, and we set $s_\star \stackrel{\text{def}}{=} \|\theta_\star\|_0$. In this section, we search for conditions under which the OLAP posterior distribution introduced in (10) concentrates around $\delta_\star$. We start with the following general definition. Let $\{\hat{\beta}^\delta,\ \delta \in \Delta\}$ be some arbitrary family of estimators, where for $\delta \in \Delta$, $\hat{\beta}^\delta \in \mathbb{R}^{\|\delta\|_0}$. We say that the family $\{\hat{\beta}^\delta,\ \delta \in \Delta\}$ is variable selection consistent if there exist positive constants $c_1, c_2 < \infty$ such that the following two properties hold.

1. For all $\delta \in \Delta$, and $\delta_0 \subseteq \delta$, with $\min(\delta, \delta_\star) = \delta_0$, it holds
$$\bar{\ell}(\hat{\beta}^\delta;\mathcal{D}) \leq \bar{\ell}(\hat{\beta}^{\delta_0};\mathcal{D}) + c_1(\|\delta\|_0 - \|\delta_0\|_0)\log(p), \quad (11)$$
2. and for all $\delta_0, \delta \in \Delta$ such that $\delta_0 \subseteq \delta \subseteq \delta_\star$, it holds
$$\bar{\ell}(\hat{\beta}^\delta;\mathcal{D}) \geq \bar{\ell}(\hat{\beta}^{\delta_0};\mathcal{D}) + c_2(\|\delta\|_0 - \|\delta_0\|_0)n. \quad (12)$$

Remark 2.2. This definition basically says that $\{\hat{\beta}^\delta,\ \delta \in \Delta\}$ is model selection consistent if increasing a model δ_0 by adding a non-relevant variable increases the log-likelihood $\bar{\ell}(\hat{\beta}^{\delta_0};\mathcal{D})$ only by a factor at most $\log(p)$, whereas adding a relevant variable increases the log-likelihood by a factor n. The behavior depicted in this definition is generally the expected behavior of the (maximum) log-likelihood ratio. However, establishing that this behavior prevails for a given model can sometimes be very challenging, particularly in the high-dimensional setting.

Our first result shows that if the one-step family $\{\check{\theta}^\delta,\ \delta \in \Delta\}$ is variable selection consistent as defined above, then the OLAP distribution (10) puts most of its probability mass on $\delta_\star$. For $j \geq 0$, we define

$$\mathcal{A}_j \stackrel{\text{def}}{=} \{\delta \in \Delta :\ \delta \supseteq \delta_\star,\ \text{ and } \ \|\delta\|_0 \leq s_\star + j\}.$$

The following result is proved in the Supplement.

Theorem 2.3. *Assume $p \geq 2$. Suppose that the family of one-step estimators $\{\check{\theta}^\delta,\ \delta \in \Delta\}$ introduced in (7) is model selection consistent with constants c_1, c_2. Consider the posterior distribution (10), and suppose that the sparsity parameter* u *satisfies*

$$\mathsf{u} \geq 2(1 + c_1), \quad (13)$$

and the sample size n satisfies

$$c_2 n \geq 2(\mathsf{u}+1)\log(p). \tag{14}$$

Then for all $j \geq 0$, we have

$$\check{\Pi}\left(\mathcal{A}_j|\mathcal{D}\right) \geq 1 - 2\left(\frac{1}{p^{\frac{\mathsf{u}(j+1)}{2}}} + \frac{2}{e^{c_2 n/4}}\right). \tag{15}$$

Proof. See Supplement Section A. □

Remark 2.4. We note that the result holds true for $j = 0$ and yields

$$\check{\Pi}\left(\delta_\star|\mathcal{D}\right) \geq 1 - 2\left(\frac{1}{p^{\frac{\mathsf{u}}{2}}} + \frac{2}{e^{c_2 n/4}}\right).$$

The theorem assumes that the family of one-step estimators $\{\check{\theta}^\delta,\ \delta \in \Delta\}$ is model selection consistent. We show below that if the initial estimator $\widetilde{\theta}$ is rate optimal (lasso is known to be rate optimal under mild conditions), then the one-step family $\{\check{\theta}^\delta,\ \delta \in \Delta\}$ actually inherits the variable selection consistency property of the MLE family $\{\hat{\theta}^\delta,\ \delta \in \Delta\}$, while being an order of magnitude faster to compute. This result mirrors the way in which the one-step estimator inherits the efficiency of the MLE in regular statistical models (Vaart, 1998, Section 5.7).

The fact that the right-hand side of (15) decreases as p increases does *not* mean that the statistical problem becomes easier as one adds more covariates at a fixed sample size. Rather, the bound reflects two features of our theoretical regime: (i) we use an informative sparsity-inducing prior (which turns out to be correct) and whose strength scales with $\log p$, so increasing p further downweights large models; and (ii) the theorem is stated under high-dimensional scaling conditions (see Eq. 14), which require n to grow at least on the order of $\log p$.

2.3 MCMC Sampling and Mixing Time

One of the main advantages of OLAP is that the distribution (10) can be easily explored using a simple Gibbs sampler in the discrete space Δ. We note that the conditional distribution of δ_j given δ_{-j} is the Bernoulli distribution $\mathbf{Ber}(q_j(\delta))$, with the probability of success given by

$$q_j(\delta) \stackrel{\text{def}}{=} \Big(1 + \exp\Big(\mathsf{u}\log(p) + \overline{\ell}^{\delta^{(j,0)}}(\check{\theta}_{\delta^{(j,0)}};\mathcal{D}) - \overline{\ell}^{\delta^{(j,1)}}(\check{\theta}_{\delta^{(j,1)}};\mathcal{D})\Big)\Big)^{-1}. \tag{16}$$

where $\delta^{(j,0)}$ (resp. $\delta^{(j,1)}$) is equal to δ except for the component j where $\delta_j^{(j,0)} = 0$ (resp. $\delta_j^{(j,1)} = 1$). The resulting Gibbs sampler for OLAP is presented in Algorithm 1.

The cost of computing $q_j(\delta)$ is driven by the cost of forming the matrix $\tilde{\mathcal{H}}^\delta$ in (7), plus the cost of the Cholesky factorization needed to perform the Newton-Raphson update. Hence, the computational cost of $q_j(\delta)$ is of order $n\|\delta\|_0^2 + \|\delta\|_0^3$. As a result, we see that the computation cost of the k-th iteration of Algorithm 1 is of order

$$J \times \max\left(n\|\delta^{(k)}\|_0^2,\ \|\delta^{(k)}\|_0^3\right).$$

We analyze the mixing time of OLAP. We focus on the case $J = 1$, where the resulting Markov chain is reversible and positive. In the case $J > 1$, a similar analysis can be developed for the lazy version of the algorithm, but we will not pursue this. Assuming the data $\mathcal{D}$ are fixed, let $\mathbb{P}(\cdot|\delta)$ denote the probability measure of the Markov chain generated by OLAP started from δ. The following result is proved in the supplement Section B.

Theorem 2.5. *Assume $p \geq 2$. Let $\{\delta^{(k)},\ k \geq 0\}$ be the Markov chain generated by OLAP with $J = 1$ and initial state $\delta^{(0)}$. Suppose that the one-step family $\{\check{\theta}^\delta,\ \delta \in \Delta\}$ is variable selection consistent with constants c_1, c_2 such that (13) and (14) hold. Suppose also that there exists α, $0 \leq \alpha \leq \mathsf{u}(p - s_\star)/4$ such that $\delta^{(0)}$ satisfies*

$$\check{\Pi}(\delta^{(0)}|\mathcal{D}) \geq \frac{1}{p^\alpha}. \tag{17}$$

Fix $\zeta_0 \in (0,1)$. Then there exists an absolute constant C such that if the number of iterations satisfies

$$N \geq C\left(s_\star + \frac{4\alpha}{\mathsf{u}}\right)\left(\log\left(\frac{1}{\zeta_0}\right) + \alpha\log(p)\right)p, \tag{18}$$

then it holds

$$\|\mathbb{P}(\delta^{(N)} = \cdot\,|\delta^{(0)}) - \check{\Pi}(\cdot)\|_{\text{tv}} \leq \max\left(\zeta_0, \frac{4}{p^{\frac{\mathsf{u}}{4}}}\right).$$

The main conclusion of the theorem is that under the warm-start condition (17), if the one-step family $\{\check{\theta}^\delta,\ \delta \in \Delta\}$ is variable selection consistent, then ζ_0-approximate sampling from (10) is achieved after at most $N = Cp(s_\star + 4\alpha/\mathsf{u})(\log(1/\zeta_0) + \alpha\log(p))$ iterations of Algorithm 1. More discussion on the benefits of a warm start and how to check the condition (17) is provided in the Supplement Section B.3.

Numerical illustrations of Theorem 2.5 and comparisons between Algorithm 1 and the exact Metropolis-within-Gibbs sampler of (Atchade and Bhattacharyya, 2019) (denoted as Exact) using the L-lag coupling method of (Biswas et al., 2019; Jacob et al., 2020) are given in Supplement Section B.4.

Algorithm 1 Gibbs sampler for OLAP

Input: Iinitial estimator $\tilde{\theta}$, dimension p, block size J, data $\mathcal{D}$
Initialize: pick $\delta^{(0)} \in \Delta$
for $k = 0, 1, 2, \ldots$ **do**
  Set $\bar{\delta} \leftarrow \delta^{(k)}$
  Randomly and uniformly select an ordered subset $\mathsf{J} \subset \{1, \ldots, p\}$ of size J
  for $\iota = 1$ **to** J **do**
    Draw $V_\iota \sim \mathrm{Ber}\big(q_{\mathsf{J}_\iota}(\bar{\delta})\big)$ ▷ q_j as in (16)
    Set $\bar{\delta}_{\mathsf{J}_\iota} \leftarrow V_\iota$
  end for
  Set $\delta^{(k+1)} \leftarrow \bar{\delta}$
end for

2.4 Variable Selection Consistency of the One-Step Family

Since the cost per iteration of OLAP is polynomial in (n, p), Theorem 2.3 and Theorem 2.5 together show that in problems where the one-step family $\{\check{\theta}^\delta,\ \delta \in \Delta\}$ is variable selection consistent, the variable selection problem can be provably solved with a computational cost that is polynomial in (n, p) using OLAP. But what can be said about the variable selection consistency of the one-step family $\{\check{\theta}^\delta,\ \delta \in \Delta\}$? In this section, we show that if the initial estimator $\widetilde{\theta}$ is rate optimal (lasso is known to be rate optimal under mild conditions), then the one-step family $\{\check{\theta}^\delta,\ \delta \in \Delta\}$ inherits the variable selection consistency of the MLE family $\{\hat{\theta}^\delta,\ \delta \in \Delta\}$.

We work under the following basic set of assumptions on the data generating process.

*H*1. 1. There exists an absolute constant $b < \infty$ such that

$$\max_{1 \le i \le n} \max_{1 \le j \le p} |\mathbf{x}_{ij}| \le b. \tag{19}$$

2. There exists an absolute constant $M_0 < \infty$ such that

$$\left\| \sum_{i=1}^n \left(y_i - \psi'(\langle \theta_\star, \mathbf{x}_i \rangle)\right) \mathbf{x}_i \right\|_\infty \le M_0 \sqrt{n \log(p)}.$$

3. The function ψ is strictly convex and there exists an absolute constant $c_3 > 0$ such that for all $x \in \mathbb{R}$,

$$\left| \frac{\mathrm{d}^3 \psi(x)}{\mathrm{d}x^3} \right| \le c_3 \frac{\mathrm{d}^2 \psi(x)}{\mathrm{d}x^2}.$$

4. There exists $0 < \underline{\kappa} \le \bar{\kappa}$ such that for all $\delta_0 \subseteq \delta_\star$, for all $w \in \mathbb{R}^p$, with $\|w\|_0 \le s_\star$ with $\|w\|_2 = 1$, it holds

$$\underline{\kappa} n \le w^{\mathsf{T}} \left(\sum_{i=1}^n \psi'' \left(\left\langle \mathbf{x}_i, \tilde{\theta}^{\delta_0} \right\rangle \right) \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} \right) w \le \bar{\kappa} n.$$

5. There exist constants c, and M such that for all $1 \le j \ne k \le p$,

$$\sup_{\substack{\vartheta \in \mathbb{R}^p,\ \|\vartheta\|_0 \le s_\star, \\ \|\vartheta - \theta_\star\|_2 \le c\sqrt{\frac{s_\star \log(p)}{n}}}} \left| \sum_{i=1}^n \psi''\left(\langle \vartheta, \mathbf{x}_i \rangle\right) \mathbf{x}_{ij} \mathbf{x}_{ik} \right| \le M \sqrt{n \log(p)}.$$

Remark 2.6. H1-(2) holds if $y_i - \psi'(\langle \theta_\star, \mathbf{x}_i \rangle)$ is mean-zero and sub-Gaussian, which holds for many GLM models, including linear and logistic regression. Indeed, if data $\mathcal{D} \stackrel{\text{def}}{=} \{(y_i, \mathbf{x}_i),\ 1 \le i \le n\}$ is a sequence of independent and identically distributed random variables, and the sub-Gaussian norm of $y_i - \psi'(\langle \theta_\star, \mathbf{x}_i \rangle)$ is σ^2, then by Hoeffding's inequality (Vershynin, 2018, Theorem 2.6.2), and a union bound inequality, there exists an absolute constant c_0 such that

$$\mathbb{P}\left(\left\| \sum_{i=1}^n \left(y_i - \psi'(\langle \theta_\star, \mathbf{x}_i \rangle)\right) \mathbf{x}_i \right\|_\infty > c_0 \sigma \sqrt{bn \log(p)} \right) \le \frac{2}{p}. \tag{20}$$

Assumption H1-(3) is the self-concordance assumption of (Bach, 2010) that is satisfied by commonly used functions including in linear, logistic and Poisson regression models. For instance, for logistic regression, $\psi(x) = \log(1 + e^x)$, and

$$\left| \frac{\mathrm{d}^3 \psi(x)}{\mathrm{d}x^3} \right| = \left| \frac{e^x (e^x - 1)}{(1 + e^x)^3} \right| \le \frac{e^x}{(1 + e^x)^2} = \frac{\mathrm{d}^2 \psi(x)}{\mathrm{d}x^2}.$$

Hence H1-(3) holds with $c_3 = 1$. For Poisson regression $\psi(x) = e^x$, hence H1-(3) also holds with $c_3 = 1$.

H1-(4) can be shown to follow a restricted eigenvalue assumption on the sample covariance matrix $X^{\mathsf{T}} X / n$, where $X \in \mathbb{R}^{n \times p}$ is the matrix of covariates with i-th row given by $\mathbf{x}_i'$. To see this, note that for most reasonable initial estimators, with high probability, it holds $\|\tilde{\theta}\|_1 \le b_2$ for some constant b_2 that can be assumed to be independent of (n, p). Therefore, in that

case, using H1-(1), $|\left\langle \mathbf{x}_i, \tilde{\theta}^{\delta_0} \right\rangle| \leq b\|\tilde{\theta}\|_1 \leq B$ for some constant B. Thus, letting $\tau = \inf_{x\in[-B,B]} \psi''(x) > 0$, we see that with high probability we have

$$w^\top \left(\sum_{i=1}^n \psi''(\left\langle \mathbf{x}_i, \tilde{\theta}^{\delta_0} \right\rangle) \mathbf{x}_i \mathbf{x}_i^\top \right) w \tag{21}$$

$$= \sum_{i=1}^n \psi''(\left\langle \mathbf{x}_i, \tilde{\theta}^{\delta_0} \right\rangle) \langle \mathbf{x}_i, w \rangle^2$$

$$\geq \tau\, w^\top (X^\top X)\, w. \tag{22}$$

The restricted eigenvalue assumption on the sample covariance matrix $X^{\mathsf{T}}X/n$ is known to hold in setting where the matrix X can be viewed as a realization of a random matrix with i.i.d. rows drawn from a sub-Gaussian distribution. We refer the reader for instance to (Wainwright, 2019) Chapter 9 for more details.

H1-(5) imposes a bound on the off-diagonal elements of the Hessian matrices of the log-likelihood. In the linear case, this corresponds to the limited coherence assumption of the design matrix commonly imposed (Candès and Plan, 2009). See also (Barber and Drton, 2015) for a similar assumption in a logistic regression setting.

The following result is proved in the Supplement Section C.

Theorem 2.7. *Assume H1, with $p \geq 4$. Suppose that the initial estimator $\widetilde{\theta}$ and the oracle MLE $\hat{\theta}^{\delta_\star}$ satisfy*

$$\max\left(\|\hat{\theta}^{\delta_\star} - \theta_\star\|_2,\ \|\tilde{\theta} - \theta_\star\|_2 \right) \leq C\sqrt{\frac{s_\star \log(p)}{n}}, \tag{23}$$

for some constant C. If the MLE family $\{\hat{\theta}^\delta,\ \delta \in \Delta\}$ is model selection consistent, then we can find a constant C' such that with a sample size

$$n \geq C' s_\star^4 \log(p), \tag{24}$$

the one-step family $\{\check{\theta}^\delta,\ \delta \in \Delta\}$ is also model selection consistent.

Remark 2.8. Theorem 2.7 establishes that OLAP inherits the variable selection consistency of the MLE family. The variable selection consistency of the MLE family itself is a more classical problem that has been considered in the literature. For instance, for a class of sparse GLM models, it is shown to hold in (Barber and Drton, 2015) Theorem 2.2 with a high probability over the data $\mathcal{D}$ under some additional sparsity assumptions.

3 NUMERICAL ILLUSTRATION

We evaluate OLAP on simulated logistic/Poisson regressions and two gene expression applications. Unless stated otherwise, we set $\mathsf{u} = 1$ in the posterior (10). For the Gibbs sampler in Algorithm 1, we set $J = 100$, and take $\delta^{(0)}$ as the support of the lasso estimator unless stated otherwise.

All experiments were conducted using MATLAB and R on Boston University's Shared Computing Cluster (SCC) using CPU cores; no GPUs were used.

Simulation design (same for Logistic and Poisson). We generate $X \in \mathbb{R}^{n\times p}$ with i.i.d. rows generated from $\mathcal{N}_p(0, \Sigma)$, where $\Sigma_{kj} = \varrho^{|j-k|}$ with $\varrho \in \{0, 0.9\}$ (*low* vs. *high* correlation). The true $\theta_\star$ is sparse: its first 10 non-zero points are drawn uniformly from $(-3,-2)\cup(2,3)$. Logistic responses use $Y_i \sim \text{Ber}(p_i)$, $p_i = \{1+\exp(-\langle \mathbf{x}_i, \theta_\star\rangle)\}^{-1}$; Poisson responses use $Y_i \sim \text{Poi}(\lambda_i)$, $\lambda_i = \exp(\langle \mathbf{x}_i, \theta_\star\rangle)$. **Performance metric.** For a sampled model δ and truth $\delta_\star$, we use the F1 score to measure the performance of support recovery

$$\text{F1}(\delta) = \frac{2\|\min(\delta_\star, \delta)\|_0}{\|\delta_\star\|_0 + \|\delta\|_0}.$$

3.1 Illustration Using Logistic Regression

Effect of the initial estimator $\widetilde{\theta}$. We compare different initial estimators, lasso, ridge and elasticNet. We set p=1000 and vary $n \in \{200, \ldots, 1000\}$. The results are given in Supplement Figure 6 and show that lasso initialization is best at small n, and all initializations lead to similar results as n grows (Supplement Figure 6). For this reason, we focus on the lasso initialization for the remaining experiments.

Comparison in terms of accuracy and running time. We compare OLAP against the following benchmarks: the exact Bayesian method Exact (Atchade and Bhattacharyya, 2019), ALA-BB and ALA-Complex (Rossell et al., 2021), Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), MCP (Zhang, 2010), ℓ_0, $\ell_0+\ell_1$, $\ell_0+\ell_2$ (Hazimeh et al., 2023), mRMR (Peng et al., 2005), ABESS (Zhu et al., 2020), S-Gibbs (Narisetty et al., 2018), and SparseVB (Ray et al., 2020). See Section D for the descriptions of these methods.

We set $p = 1000$, vary n and the design correlation. We report both medians and standard errors of F1 scores over 50 replications. Against these benchmarks, OLAP attains top-tier F1 with markedly lower and more stable runtimes (Tables 1–4).

The evaluation of the computing times require some care, since some of the methods are MCMC-based while others are optimization-based with hyper-parameter tuning. For MCMC-based methods (OLAP, Exact), we evaluate runtimes using the unbiased MCMC framework (Biswas et al., 2019; Jacob et al., 2020), and performance are measured at the coupling event. In contrast, methods such as SCAD,

MCP, ℓ_0-based selections, and ABESS require dataset-specific hyper-parameters via cross-validation, so their reported runtime includes this overhead.

3.2 Illustration with Poisson Regression

Effect of the initial estimator $\widetilde{\theta}$. The initial estimators behave similarly as in the logistic regression setting. We observe that Poisson regression models require a larger sample size n for comparable F1 (Supplement Figures 7).

Comparison in terms of accuracy and running time. We compare the performance of OLAP with the methods in the benchmarks. We drop ℓ_0, $\ell_0 + \ell_1$, $\ell_0 + \ell_2$, S-Gibbs and SparseVB, as the provided software does not handle Poisson models.

The median and standard deviation over 50 replications are presented in Tables 5–8 in the supplement. We observe again that OLAP remains very competitive in terms of F1 score while having a modest running time; ABESS and Exact can match, or, in many settings, perform better than OLAP but become computationally expensive for large n (Supplement Tables 5–8).

3.3 Real-Data Illustrations

Breast cancer. Accurate prediction of distant metastasis development in breast cancer can help guide treatments and ultimately save life. In (van de Vijver et al., 2002), using gene expression data, the authors developed a prognostic score based on 70 identified genes to predict the advent of metastasis of breast cancer cells in distant organs within 10 years of diagnosis. In this illustration, we reanalyze the data and compare their 70-genes algorithm with a high-dimensional logistic regression model. We compare SparseVB, 70-genes, OLAP and Exact using a 50-fold cross-validation procedure. OLAP gives higher mean/median F1 score: SparseVB (0.362/0.509), 70-genes (0.593/0.609), OLAP (0.688/0.688) and Exact (0.647/0.655). We refer to Supplement Section G for more details and results.

Mouse PCR (GPAT). As another illustration, following (Lan et al., 2006; Narisetty et al., 2018), we analyze a mouse PCR data where the objective is to predict the level (low or high) of glycerol-3-phosphate acyltransferase (GPAT) using gene expression data. The level of GPAT in the body is important, as reduced GPAT levels have been linked to decreased hepatic steatosis, a disease commonly associated with obesity. We compare by 30-fold cross-validation SparseVB, S-Gibbs, OLAP and Exact in fitting the resulting high-dimensional logistic regression. OLAP gives a lower median RMSE and a higher median F1 score (0.308/0.889) compared to SparseVB (0.500/0.536), S-Gibbs (0.524/0.364) and a comparable RMSE and F1 score to Exact (0.293/0.889). We refer to Section Supplement H for more details and results.

4 CONCLUDING REMARKS

Variable selection is an NP-hard problem (Welch, 1982), and algorithms that can solve all versions in polynomial times are unlikely to exist. Therefore, identifying instances of the problem (and corresponding algorithms) that can be solved in polynomial time is of practical importance. Our work in this paper contributes to this literature. We develop a novel Laplace approximation algorithm that is applicable to a large class of generalized linear models (GLMs) and beyond. The resulting algorithm is fast and performs well in practice. Furthermore, we established that whether our algorithm provably solves the variable selection problem reduces to the variable selection consistency of the underlying class of MLE estimators, a more classical problem. The numerical experiments illustrate the competitiveness of OLAP against several existing methods.

One limitation of OLAP is the computational cost of the Newton update. In the current implementation, each component update in the Gibbs sampler is performed at the cost of a complete re-forming and Cholesky re-factorization of the matrix $\tilde{\mathcal{H}}^{\delta}$. Finding a recursive update to these calculations will significantly improve the speed of the algorithm. Another useful direction for further research is the extension of the theory of OLAP beyond GLM models.

ACKNOWLEDGEMENT

The last author was supported in part by NSF Grant DMS 2210664.

References

Atchade, Y. and Bhattacharyya, A. (2019). An approach to large-scale quasi-bayesian inference with spike-and-slab priors.

Atchadé, Y. F. (2021). Approximate spectral gaps for markov chain mixing times in high dimensions. *SIAM Journal on Mathematics of Data Science*, 3(3):854–872.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electron. J. Statist.*, 4:384–414.

Barber, R. and Drton, M. (2015). High-dimensional ising model selection with bayesian information criteria. *Electronic Journal of Statistics*, 9:567–6–7.

	n=200		n=300		n=400		n=500		n=1000	
Method	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**
OLAP	0.778	0.094	**1.000**	0.052	**1.000**	0.000	**1.000**	0.007	**1.000**	0.000
Exact	0.750	0.220	**1.000**	0.035	**1.000**	0.000	**1.000**	0.000	**1.000**	0.000
ALA-BB	0.462	0.127	0.824	0.096	0.947	0.046	1.000	0.030	**1.000**	0.000
ALA-Complex	0.000	0.055	0.182	0.115	0.333	0.108	0.667	0.121	**1.000**	0.014
Lasso	0.300	0.080	0.294	0.083	0.280	0.098	0.288	0.097	0.339	0.122
SCAD	0.526	0.110	0.625	0.107	0.656	0.127	0.785	0.160	0.976	0.251
MCP	0.800	0.082	0.909	0.079	0.931	0.083	0.952	0.083	**1.000**	0.201
l0	**1.000**	0.123	**1.000**	0.016	**1.000**	0.020	**1.000**	0.019	**1.000**	0.141
l0+l1	0.952	0.110	**1.000**	0.036	**1.000**	0.040	**1.000**	0.026	**1.000**	0.144
l0+l2	**1.000**	0.070	**1.000**	0.042	**1.000**	0.039	**1.000**	0.026	**1.000**	0.143
mMRM-5	0.667	0.058	0.667	0.000	0.667	0.000	0.667	0.000	0.667	0.000
mMRM-10	0.700	0.114	0.900	0.087	0.900	0.062	**1.000**	0.053	**1.000**	0.000
mMRM-20	0.533	0.070	0.667	0.041	0.667	0.032	0.667	0.018	0.667	0.000
ABESS	0.870	0.064	0.833	0.082	0.800	0.115	0.952	0.127	**1.000**	0.142
SparseVB	0.918	0.086	**1.000**	0.011	**1.000**	0.022	**1.000**	0.018	**1.000**	0.023
S-Gibbs	0.791	0.109	0.952	0.035	0.952	0.024	0.952	0.014	0.952	0.009

Table 1: F1 scores for logistic model feature selections. $p = 1000$, low correlation.

	n=200		n=300		n=400		n=500		n=1000	
Method	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**
OLAP	0.471	0.188	**0.842**	0.164	**0.900**	0.101	**1.000**	0.070	**1.000**	0.000
Exact	**0.572**	0.184	**0.842**	0.136	**0.900**	0.105	0.900	0.105	**1.000**	0.037
ALA-BB	0.182	0.131	0.462	0.168	0.750	0.172	0.824	0.234	**1.000**	0.045
ALA-Complex	0.000	0.060	0.000	0.123	0.308	0.134	0.462	0.195	0.800	0.123
Lasso	0.229	0.064	0.265	0.070	0.281	0.076	0.314	0.057	0.310	0.052
SCAD	0.424	0.153	0.628	0.146	0.734	0.138	0.833	0.166	0.952	0.188
MCP	0.490	0.164	0.667	0.147	0.762	0.159	0.762	0.122	0.870	0.159
l0	0.526	0.191	0.683	0.166	0.762	0.145	0.842	0.123	0.928	0.098
l0+l1	0.547	0.206	0.612	0.232	0.714	0.220	0.800	0.169	0.900	0.154
l0+l2	0.522	0.193	0.698	0.161	0.743	0.145	0.800	0.147	0.900	0.124
mMRM-5	0.400	0.158	0.400	0.148	0.533	0.124	0.533	0.100	0.667	0.076
mMRM-10	0.400	0.159	0.550	0.162	0.600	0.153	0.700	0.130	0.900	0.125
mMRM-20	0.333	0.116	0.400	0.104	0.467	0.102	0.467	0.081	0.600	0.076
ABESS	0.522	0.163	0.700	0.128	0.821	0.127	0.900	0.094	**1.000**	0.076
SparseVB	0.500	0.201	0.718	0.161	0.800	0.186	0.800	0.223	**1.000**	0.076
S-Gibbs	0.167	0.186	0.429	0.212	0.706	0.188	0.737	0.225	0.952	0.042

Table 2: F1 scores for logistic model feature selections. $p = 1000$, high correlation.

Method	**n=200**	**n=300**	**n=400**	**n=500**	**n=1000**
OLAP	1.11	1.77	1.09	0.81	1.38
Exact	0.89	1.74	2.19	2.40	8.56
ALA-BB	0.48	0.32	0.40	0.44	0.71
ALA-Complex	0.10	0.19	0.27	0.35	0.53
Lasso	0.23	0.34	0.48	0.63	1.69
SCAD	1.32	1.85	2.71	3.44	2.74
MCP	0.97	1.55	2.13	3.68	9.40
l0	5.16	9.09	15.00	21.80	70.40
l0+l1	41.76	70.64	100.47	134.34	301.24
l0+l2	25.90	35.59	43.89	48.00	65.75
mMRM-5	0.01	0.01	0.01	0.01	0.02
mMRM-10	0.01	0.01	0.01	0.01	0.02
mMRM-20	0.02	0.02	0.02	0.03	0.04
ABESS	0.06	0.14	0.26	0.43	2.29
SparseVB	1.20	2.20	2.88	3.68	8.20
S-Gibbs	12.36	12.82	13.38	13.50	15.35

Table 3: Running time (seconds) for logistic model feature selection. $p = 1000$, low correlation.

Method	**n=200**	**n=300**	**n=400**	**n=500**	**n=1000**
OLAP	1.50	2.28	1.57	1.39	1.38
Exact	1.30	3.17	2.40	2.81	8.56
ALA-BB	0.56	1.46	21.27	20.38	34.29
ALA-Complex	0.11	0.17	3.64	3.55	4.99
Lasso	0.33	0.40	0.55	0.68	1.71
SCAD	1.46	2.67	3.45	3.68	2.34
MCP	0.93	1.83	3.17	4.41	3.25
l0	5.24	9.94	16.09	22.69	72.04
l0+l1	41.81	74.50	113.59	149.20	326.55
l0+l2	26.55	39.30	51.88	60.32	88.66
mMRM-5	0.01	0.01	0.01	0.01	0.02
mMRM-10	0.01	0.01	0.01	0.01	0.02
mMRM-20	0.02	0.02	0.02	0.03	0.04
ABESS	0.07	0.16	0.29	0.48	2.89
SparseVB	1.21	1.93	2.53	3.38	7.94
S-Gibbs	11.53	11.41	12.89	12.25	12.82

Table 4: Running time (seconds) for logistic model feature selection. $p = 1000$, high correlation.

Barber, R. F., Drton, M., and Tan, K. M. (2016). Laplace approximation in high-dimensional bayesian regression. In *Statistical Analysis for High-Dimensional Data: The Abel Symposium 2014*, pages 15–36. Springer.

Beck, A. and Eldar, Y. C. (2013). Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509.

Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813 – 852.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019). Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):pp. 405–427.

Biswas, N., Jacob, P. E., and Vanetti, P. (2019). Estimating convergence of markov chains with l-lag couplings. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Brouste, A., Dutang, C., Mieniedou, D. N., et al. (2021). Onestep: Le cam's one-step estimation procedure. *R J.*, 13(1):366.

Candès, E. J. and Plan, Y. (2009). Near-ideal model selection by L1 minimization. *The Annals of Statistics*, 37(5A):2145 – 2177.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3):473–484.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of markov chains. *The Annals of Applied Probability*, 1:36–61.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, pages 101–148.

Freedman, L. S., Pee, D., and Midthune, D. N. (1992). The problem of underestimating the residual error variance in forward stepwise regression. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41(4):405–412.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Green, P. J. (2003). Trans-dimensional markov chain monte carlo. *Oxford Statistical Science Series*, pages 179–198.

Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780 – 1815.

Guo, J. (2018). *Some Contributions to High Dimensional Mixed Effects Logistic Regression Models*. PhD thesis, University of Michigan.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC.

Hazimeh, H. and Mazumder, R. (2020). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Oper. Res.*, 68(5):1517–1537.

Hazimeh, H., Mazumder, R., and Nonet, T. (2023). L0learn: A scalable package for sparse learning using l0 regularization.

Jacob, P. E., O'Leary, J., and Atchade, Y. F. (2020). Unbiased Markov Chain Monte Carlo Methods with Couplings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):543–600.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15(1):2869–2909.

Lamnisos, D., Griffin, J. E., and Steel, M. F. J. (2009). Transdimensional sampling algorithms for bayesian variable selection in classification problems with many more variables than observations. *Journal of Computational and Graphical Statistics*, 18(3):592–612.

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T.-K., Flowers, M. T., Schueler, K. L., Manly, K. F., Williams, R. W., Kendziorski, C., and Attie, A. D. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLOS Genetics*, 2(1):1–11.

Lecam, L. (1969). *Theorie Asymptotique De La Decision Statistique, Par Lucien M. Lecam*. Seminaire De Mathematiques Superieures, 33. Les presses de l'Université de Montreal.

Leng, C., Lin, Y., and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, pages 1273–1284.

Luo, S. and Chen, Z. (2011). Extended bic for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. *Journal of Statistical Planning and Inference*, 143.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817.

Narisetty, N. N., Shen, J., and He, X. (2018). Skinny gibbs: A consistent and scalable gibbs sampler for model selection. *Journal of the American Statistical Association*.

O'Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1):85 – 117.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency,max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence*, 27:1226–38.

Polson, N. G. and Scott, J. G. (2011). 501Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction. In *Bayesian Statistics 9*. Oxford University Press.

Ray, K., Szabó, B., and Clara, G. (2020). Spike and slab variational bayes for high dimensional logistic regression. *Advances in Neural Information Processing Systems*, 33:14423–14434.

Rockova, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Rossell, D., Abril, O., and Bhattacharya, A. (2021). Approximate laplace approximations for scalable model selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4):853–879.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464.

Sinclair, A. (1992). Improved bounds for mixing rates of markov chains and multicommodity flow. *Combinatorics, Probability and Computing*, 1(4):351–370.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166 – 1202.

van de Vijver, M. J., He, Y. D., van 't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009. PMID: 12490681.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Welch, W. J. (1982). Algorithmic complexity: three np- hard problems in computational statistics. *Journal of Statistical Computation and Simulation*, 15(1):17–25.

Xia, L., Nan, B., and Li, Y. (2020). A revisit to de-biased lasso for generalized linear models. *arXiv preprint arXiv:2006.12778*.

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016). On the computational complexity of high-dimensional bayesian variable selection. *Ann. Statist.*, 44(6):2497–2532.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2).

Zhu, J., Wen, C., Zhu, J., Zhang, H., and Wang, X. (2020). A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences*, 117(52):33117–33123.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Sections 2.2 to 2.3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] See Theorems 2.3, 2.5, and 2.7.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Assumptions 1 and 2 and Sections 2.2 to 2.3.
   (b) Complete proofs of all theoretical results. [Yes] See Appendix Sections A to C.
   (c) Clear explanations of any assumptions. [Yes] See Remark 2.6 and the discussion following Assumptions 1 and 2.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not yet.]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Section 3 and the data generation description therein. Hyperparameters are fixed across experiments ($\mathsf{u} = 0.8$, $J = 100$, lasso initialization).

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] F1 scores are reported as medians and standard deviations over 50 replications; see Section 3.

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] All experiments were conducted using MATLAB and R on Boston University's Shared Computing Cluster (SCC) using CPU cores; no GPUs were used.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator if your work uses existing assets. [Yes] The breast cancer dataset and mouse PCR dataset are cited; see Sections 3.3 and Appendix Sections D.

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A PROOF OF THEOREM (2.3)

Proof. We partition the model space Δ as

$$\Delta = \bigcup_{\delta_0 \in \mathcal{C}} \Delta(\delta_0),$$

where

$$\mathcal{C} \stackrel{\text{def}}{=} \{\delta \in \Delta :\ \delta \subseteq \delta_\star\}, \quad \text{and} \quad \Delta(\delta_0) \stackrel{\text{def}}{=} \{\delta' \in \Delta :\ \delta_0 \subseteq \delta', \ \text{ and } \ \min(\delta', \delta_\star) = \delta_0\}.$$

We claim that under the conditions of the theorem the following holds true. For all $\delta_0 \in \mathcal{C}$, and all $M \geq 1$,

$$\sum_{\delta \in \Delta(\delta_0)} \frac{\check{\Pi}(\delta|\mathcal{D})}{\check{\Pi}(\delta_0|\mathcal{D})} \leq 2 \quad \text{and} \quad \check{\Pi}\left(\|\delta\|_0 \geq s_\star + M|\mathcal{D}\right) \leq 2\left(\frac{1}{p^{\mathsf{u}/2}}\right)^M. \tag{25}$$

We give a proof below. To use the claim in (25) to prove the theorem we observe that $\delta \notin \mathcal{A}_j$ means that either $\|\delta\|_0 > s_\star + j$, or $\min(\delta, \delta_\star)$ is strictly a sub-model of $\delta_\star$. We can rewrite this statement as

$$\mathcal{A}_j^c = \{\delta \in \Delta :\ \|\delta\|_0 > s_\star + j\} \ \cup \bigcup_{\delta_0 \in \mathcal{C}\setminus\{\delta_\star\}} \{\delta \in \Delta(\delta_0) :\ \|\delta\|_0 \leq s_\star + j\}.$$

Therefore, using (25),

$$\begin{aligned}
\check{\Pi}(\mathcal{A}_j^c|\mathcal{D}) \leq & \frac{2}{p^{\mathsf{u}(j+1)/2}} + \check{\Pi}(\delta_\star|\mathcal{D}) \sum_{\delta_0 \in \mathcal{C}\setminus\{\delta_\star\}} \frac{\check{\Pi}(\delta_0|\mathcal{D})}{\check{\Pi}(\delta_\star|\mathcal{D})} \sum_{\delta \in \Delta(\delta_0)} \frac{\check{\Pi}(\delta|\mathcal{D})}{\check{\Pi}(\delta_0|\mathcal{D})} \\
\leq & \frac{2}{p^{\mathsf{u}(j+1)/2}} + 2 \sum_{\delta_0 \in \mathcal{C}\setminus\{\delta_\star\}} \frac{\check{\Pi}(\delta_0|\mathcal{D})}{\check{\Pi}(\delta_\star|\mathcal{D})}.
\end{aligned}$$

Take $\delta_0 \in \mathcal{C}$, with $\|\delta_0\|_0 = s_\star - k$, for some $k > 0$. Then from (10) of the main document, we have

$$\frac{\check{\Pi}(\delta_0|\mathcal{D})}{\check{\Pi}(\delta_\star|\mathcal{D})} = p^{\mathsf{u}k} \exp\left(\bar{\ell}^{\delta_0}(\check{\theta}_{\delta_0}; \mathcal{D}) - \bar{\ell}^{\delta_\star}(\check{\theta}_{\delta_\star}; \mathcal{D})\right).$$

Since the family $\{\check{\theta}_\delta,\ \delta \in \Delta\}$ is variable selection consistent, $\bar{\ell}^{\delta_0}(\check{\theta}^{\delta_0}; \mathcal{D}) - \bar{\ell}^{\delta_\star}(\check{\theta}_{\delta_\star}; \mathcal{D}) \leq -c_2 kn$. It follows that

$$\check{\Pi}(\mathcal{A}_j^c|\mathcal{D}) \leq \frac{2}{p^{\mathsf{u}(j+1)/2}} + 2\sum_{k=1}^{s_\star} \binom{s_\star}{k} \exp\left(\mathsf{u}k\log(p) - c_2 kn\right).$$

Hence for $c_2 n \geq 2(\mathsf{u}+1)\log(p)$ as assumed in (14) of the main document, we obtain

$$\check{\Pi}(\mathcal{A}_j^c|\mathcal{D}) \leq \frac{2}{p^{\mathsf{u}(j+1)/2}} + 2\sum_{k=1}^{s_\star} e^{-c_2 nk/2} \leq \frac{2}{p^{\mathsf{u}(j+1)/2}} + 4e^{-c_2 n/2},$$

which yields the stated inequality.

It remains to prove (25). Given $\delta_0 \in \mathcal{C}$, $\delta \in \Delta(\delta_0)$, and setting $\|\delta\|_0 - \|\delta_0\|_0 = j$, we have

$$\frac{\check{\Pi}(\delta|\mathcal{D})}{\check{\Pi}(\delta_0|\mathcal{D})} = \left(\frac{1}{p^{\mathsf{u}}}\right)^j \exp\left(\bar{\ell}^{\delta}(\check{\theta}^{\delta}; \mathcal{D}) - \bar{\ell}^{\delta_0}(\check{\theta}^{\delta_0}; D)\right).$$

The variable selection consistency of $\{\check{\theta}_\delta,\ \delta \in \Delta\}$ yields $\bar{\ell}^{\delta}(\check{\theta}^{\delta}; \mathcal{D}) - \bar{\ell}^{\delta_0}(\check{\theta}^{\delta_0}; \mathcal{D}) \leq c_1 j\log(p)$, so that,

$$\frac{\check{\Pi}(\delta|\mathcal{D})}{\check{\Pi}(\delta_0|\mathcal{D})} \leq \exp\left(-\mathsf{u}j\log(p) + c_1 j\log(p)\right), \tag{26}$$

and we conclude that

$$\sum_{\delta\in\Delta(\delta_0)} \frac{\check{\Pi}(\delta|\mathcal{D})}{\check{\Pi}(\delta_0|\mathcal{D})} = \sum_{j\geq 0} \sum_{\delta\in\Delta(\delta_0):\ \|\delta\|_0=\|\delta_0\|_0+j} \frac{\check{\Pi}(\delta|\mathcal{D})}{\check{\Pi}(\delta_0|\mathcal{D})}$$
$$\leq \sum_{j\geq 0} \binom{p-s_0}{j} e^{-(\mathsf{u}-c_1)j\log(p)} \leq \sum_{j\geq 0} e^{-(\mathsf{u}-c_1-1)j\log(p)} \leq 2,$$

provided that $\mathsf{u} \geq 2 + c_1$, and $p \geq 2$, which is satisfied by taking u as in (13) of the main document. The second part of (25) follows a similar argument. Since $\Delta = \cup_{\delta_0\in\mathcal{C}}\Delta(\delta_0)$, for any $M \geq 1$, we get

$$\check{\Pi}\left(\|\delta\|_0 \geq s_\star + M|\mathcal{D}\right) = \sum_{\delta_0\in\mathcal{C}} \check{\Pi}(\delta_0|\mathcal{D}) \sum_{\delta\in\Delta(\delta_0):\ \|\delta\|_0\geq s_\star+M} \frac{\check{\Pi}(\delta|\mathcal{D})}{\check{\Pi}(\delta_0|\mathcal{D})}. \tag{27}$$

Fix $\delta_0 \in \mathcal{C}$, and set $\|\delta_0\|_0 = s_0$. For $k \geq s_\star$, we have

$$\sum_{\delta\in\Delta(\delta_0):\ \|\delta\|_0\geq k} \frac{\check{\Pi}(\delta|\mathcal{D})}{\check{\Pi}(\delta_0|\mathcal{D})} = \sum_{j\geq k-s_0} \sum_{\delta\in\Delta(\delta_0):\ \|\delta\|_0=s_0+j} \frac{\check{\Pi}(\delta|\mathcal{D})}{\check{\Pi}(\delta_0|\mathcal{D})}$$
$$\leq \sum_{j\geq k-s_0} \binom{p-s_0}{j} \exp\left(-\mathsf{u}j\log(p) + c_1 j\log(p)\right)$$
$$\leq \sum_{j\geq k-s_0} \exp\left(-\left(\mathsf{u}-c_1-1\right)j\log(p)\right).$$

Hence for $\mathsf{u}/2 \geq (c_1 + 1)$ as assumed in (13) of the main document we obtain for $k > s_0$, that

$$\sum_{\delta\in\Delta(\delta_0):\ \|\delta\|_0\geq k} \frac{\check{\Pi}(\delta|\mathcal{D})}{\check{\Pi}(\delta_0|\mathcal{D})} \leq \sum_{j\geq k-s_0} \left(\frac{1}{p^{\mathsf{u}/2}}\right)^j \leq 2\left(\frac{1}{p^{\mathsf{u}/2}}\right)^{k-s_0}.$$

Hence, the last display together with (27) implies that for any $M \geq 1$, we have

$$\check{\Pi}\left(\|\delta\|_0 \geq s_\star + M|\mathcal{D}\right) \leq 2\sum_{\delta_0\in\mathcal{C}} \check{\Pi}(\delta_0|\mathcal{D}) \left(\frac{1}{p^{\mathsf{u}/2}}\right)^M \leq 2\left(\frac{1}{p^{\mathsf{u}/2}}\right)^M.$$

□

B PROOF OF THEOREM 2.5

We organize this section in four parts. To make the proofs self-contained, we first recall some basic facts on the mixing times of Markov chains. The second part contains the actual proof of the theorem. In the third, we further explore the warm-start condition and its implications on the mixing of Algorithm 1, connecting the theorem with the practical performance of the algorithm. In the last part, we present some numerical comparison between the mixing time of Algorithm 1 and the exact algorithm of (Atchade and Bhattacharyya, 2019).

B.1 Preliminary on Markov Chains Mixing Times

Let π be a density on some measure space $\mathcal{X}$ equipped with a sigma-algebra $\mathcal{B}$ and a reference sigma-finite measure $\mathrm{d}x$. We will also write π to denote the induced probability measure: $\pi(\mathrm{d}x) = \pi(x)\mathrm{d}x$. For a function $f:\ \mathcal{X} \to \mathbb{R}$, we write $f \in \mathcal{B}$ to say that f is $\mathcal{B}$-measurable. We also define

$$\|f\|_\infty \stackrel{\text{def}}{=} \sup_{x\in\mathcal{X}} |f(x)|, \quad \text{and} \quad \mathtt{osc}(f) \stackrel{\text{def}}{=} \sup_{y,z\in\mathcal{X}} |f(y) - f(z)|.$$

We let $L^2(\pi)$ denote the Hilbert space of all real-valued square-integrable functions (wrt π) on $\mathcal{X}$, equipped with the inner product $\langle f, g\rangle \stackrel{\text{def}}{=} \int_{\mathcal{X}} f(x)g(x)\pi(\mathrm{d}x)$ with associated norm $\|\cdot\|_2$. We will also make use of the essential

supremum (resp. essential infimum) of f with respect to π defined as $\texttt{ess-sup}(f) \stackrel{\text{def}}{=} \inf\{M \geq 0 :\ \pi(\{x \in \mathcal{X} : |f(x)| > M\}) = 0\}$ (resp. $\texttt{ess-inf}(f) \stackrel{\text{def}}{=} \sup\{M \geq 0 :\ \pi(\{x \in \mathcal{X} :\ |f(x)| < M\}) = 0\}$).

If K is a Markov kernel on $\mathcal{X}$, and $n \geq 1$ an integer, K^n denotes the n-th iterate of K, defined recursively as $K^n(x, A) \stackrel{\text{def}}{=} \int_{\mathcal{X}} K^{n-1}(x, \mathrm{d}z)K(z, A)$, $x \in \mathcal{X}$, A measurable. For $f \in \mathcal{B}$, we define $Kf :\ \mathcal{X} \to \mathbb{R}$ as $Kf(x) \stackrel{\text{def}}{=} \int_{\mathcal{X}} K(x, \mathrm{d}z)f(z)$, $x \in \mathcal{X}$, whenever the integral is well defined. And if μ is a probability measure on $\mathcal{X}$, then μK is the probability on $\mathcal{X}$ defined as $\mu K(A) \stackrel{\text{def}}{=} \int_{\mathcal{X}} \mu(\mathrm{d}z)K(z, A)$, $A \in \mathcal{B}$. The total variation distance between two probability measures μ, ν is defined as

$$\begin{aligned}\|\mu - \nu\|_{\text{tv}} &\stackrel{\text{def}}{=} 2 \sup_{A \in \mathcal{B}} (\mu(A) - \nu(A)) \\ &= \sup_{f \in \mathcal{B}:\ \|f\|_\infty \leq 1} \left(\int_{\mathcal{X}} f(x)\, \mu(\mathrm{d}x) - \int_{\mathcal{X}} f(x)\, \nu(\mathrm{d}x) \right).\end{aligned}$$

If the Markov kernel K has invariant distribution π, without changing the notation, we will also view K as the linear operator on $L^2(\pi)$ that transforms f to Kf as defined above. We write $K^\star$ to denote the adjoint of K, that is, the linear operator on $L^2(\pi)$ such that $\langle Kf, g\rangle = \langle f, K^\star g\rangle$ for all $f, g \in L^2(\pi)$. We say that K is reversible with respect to π (π-reversible for short) if $K^\star = K$, and we say that K is positive if it is π-reversible and $\langle f, Kf\rangle \geq 0$ for all $f \in L^2(\pi)$. For $f \in L^2(\pi)$, we set $\pi(f) \stackrel{\text{def}}{=} \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x)$, $\mathsf{Var}_\pi(f) \stackrel{\text{def}}{=} \|f - \pi(f)\|_2^2$, and

$$\begin{aligned}\mathcal{E}_K(f, f) &\stackrel{\text{def}}{=} \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} (f(y) - f(x))^2\, \pi(\mathrm{d}x)\, K(x, \mathrm{d}y) \\ &= \langle f, f\rangle \ - \ \langle f, Kf\rangle\ . \end{aligned} \tag{28}$$

For $\zeta \in [0, 1)$, the ζ-spectral gap of K is

$$\lambda_\zeta(K) \stackrel{\text{def}}{=} \inf\left\{ \frac{\mathcal{E}_K(f, f)}{\mathsf{Var}_\pi(f) - \frac{\zeta}{2}},\ f \in L^2(\pi) \text{ s.t. } \|f\|_\infty \leq 1, \mathsf{Var}_\pi(f) > \zeta \right\}.$$

We recover the classical spectral gap (denoted $\lambda(K)$) by taking $\zeta = 0$, and replacing the infinity norm by the L^2-norm[2]. We note that when K is positive, then $0 \leq \lambda_\zeta(K) \leq 2$. This quantity can be used to quantify the convergence rate of K toward a total variation ball of radius $O(\sqrt{\zeta})$ around π. Specifically, the following is a slight modification of Lemma 2.1 of (Atchadé, 2021). We provide a proof for completeness.

Lemma B.1. *Suppose that K is π-reversible and positive, and let $\zeta \in [0, 1)$. Let $\pi_0(\mathrm{d}x) = f_0(x)\pi(\mathrm{d}x)$, for some bounded function f_0. For all integer $N \geq 1$, we have*

$$\|\pi_0 K^N - \pi\|_{\text{tv}}^2 \leq \max\Big(\zeta\|f_0\|_\infty^2,\ \big[1 - \tfrac{\lambda_\zeta(K)}{2}\big]^N \times \mathsf{Var}_\pi(f_0)\Big). \tag{29}$$

Proof. By reversibility, for all $j \geq 1$, the density of $\pi_0 K^j$ with respect to π is $f_j = K^j f_0$. Therefore,

$$\|\pi_0 K^j - \pi\|_{\text{tv}} = \int |f_j(x) - 1|\pi(\mathrm{d}x) \leq \sqrt{\mathsf{Var}_\pi(f_j)}.$$

Take $f \in L^2(\pi)$. Since $\pi(f) = \pi(Kf)$, and $\mathsf{Var}_\pi(f) = \langle f, f\rangle_\pi - \pi(f)^2$, we have

$$\mathsf{Var}_\pi(Kf) - \mathsf{Var}_\pi(f) = \langle Kf, Kf\rangle_\pi - \langle f, f\rangle_\pi = \langle f, K_\star Kf\rangle_\pi - \langle f, f\rangle_\pi = -\mathcal{E}_{K^\star K}(f, f).$$

If K is positive (it is therefore π- reversible), then $K^\star K = K^2$, and K admits a square root: there exists a bounded π-reversible operator S such that $S^2 = K$, and S commutes with K. Furthermore, with $\mathbb{I}$ denoting the identity operator, $\mathbb{I} - K$ is also a positive operator: $\mathbb{I} - K$ is clearly π-reversible, and $\langle f, (\mathbb{I} - K)f\rangle_\pi = \|f\|_2^2 - \langle f, Kf\rangle_\pi \geq 0$, using the fact that the operator norm of K is smaller than or equal to one. Hence

$$\langle f, Kf\rangle_\pi - \langle f, K^\star Kf\rangle_\pi = \big\langle f, (K - K^2)f\big\rangle_\pi = \langle f, S(\mathbb{I} - K)Sf\rangle_\pi = \langle Sf, (\mathbb{I} - K)Sf\rangle_\pi \geq 0.$$

[2] In the definition of $\lambda_\zeta(K)$ one can replace the infinity norm $\|\cdot\|_\infty$ by any other norm that satisfies $\|Kf\| \leq \|f\|$ for all $f \in L^2(\pi)$. We use the infinity norm here mainly for convenience.

Therefore when K is positive we can have

$$\mathsf{Var}_\pi(Kf) \leq \mathsf{Var}_\pi(f) - \mathcal{E}_K(f,f).$$

In particular, for all $j \geq 1$,

$$\mathsf{Var}_\pi(f_j) \leq \mathsf{Var}_\pi(f_{j-1}) - \mathcal{E}_K(f_{j-1}, f_{j-1}). \tag{30}$$

First, we observe that the last display implies that $\{\mathsf{Var}_\pi(f_k),\ k \geq 0\}$ is non-increasing. Fix $j \geq 1$, and suppose now that $\mathsf{Var}_\pi(f_j) > \zeta\|f_0\|_\infty^2$. Since the sequence $\{\mathsf{Var}_\pi(f_k),\ k \geq 0\}$ is non-increasing, for all $0 \leq i \leq j$, $\mathsf{Var}_\pi(f_i) > \zeta\|f_0\|_\infty^2$. Note also that $\|f_k\|_\infty \leq \|f_0\|_\infty$ for all $k \geq 0$. Therefore, for all $1 \leq i \leq j$,

$$\begin{aligned}
\mathsf{Var}_\pi(f_i) &\leq \mathsf{Var}_\pi(f_{i-1}) - \mathcal{E}_K\left(\frac{f_{i-1}}{\|f_0\|_\infty}, \frac{f_{i-1}}{\|f_0\|_\infty}\right)\|f_0\|_\infty^2 \\
&\leq \mathsf{Var}_\pi(f_{i-1}) - \|f_0\|_\infty^2 \lambda_\zeta(K)\left(\mathsf{Var}_\pi\left(\frac{f_{i-1}}{\|f_0\|_\infty}\right) - \frac{\zeta}{2}\right), \\
&\leq \mathsf{Var}_\pi(f_{i-1}) - \frac{\lambda_\zeta(K)}{2}\mathsf{Var}_\pi(f_{i-1}) \\
&\leq \left(1 - \frac{\lambda_\zeta(K)}{2}\right)^i \mathsf{Var}_\pi(f_0).
\end{aligned}$$

We conclude that for all $j \geq 1$,

$$\mathsf{Var}_\pi(f_j) \leq \max\left[\zeta\|f_0\|_\infty^2, \left(1 - \frac{\lambda_\zeta(K)}{2}\right)^j \mathsf{Var}_\pi(f_0)\right].$$

This ends the proof. □

When $\mathcal{X}$ is a discrete set, one can easily adapt the canonical path arguments of (Sinclair, 1992) (see also (Diaconis and Stroock, 1991)) to obtain lower bounds on $\lambda_\epsilon(K)$. Toward that, we recall some basic definitions from graph theory. A graph $(\mathcal{V}, \mathcal{E})$ with the vertex set $\mathcal{V}$ and the edge set $\mathcal{E}$ is a set $\mathcal{V}$ together with a subset $\mathcal{E}$ of $\mathcal{V} \times \mathcal{V}$. We say that $(\mathcal{V}, \mathcal{E})$ is undirected if for all $(x,y) \in \mathcal{V} \times \mathcal{V}$, $(x,y) \in \mathcal{E}$ if and only if $(y,x) \in \mathcal{E}$. We write an edge as (e_-, e_+), where e_-, e_+ denote its two incident nodes. We say that the graph is connected if for all $(x,y) \in \mathcal{V} \times \mathcal{V}$, $x \neq y$, we can find a sequence of edges $(z_0, z_1), \ldots, (z_{\ell-1}, z_\ell)$ such that $z_0 = x$, $z_\ell = y$, and $(z_{k-1}, z_k) \in \mathcal{E}$ for $k = 1, \ldots, \ell$. The integer ℓ is the length of the path. There may exist many paths linking any two points. For each pair $(x,y) \in \mathcal{V}$, $x \neq y$, we assume, given a special path denoted γ_{xy} linking (x,y) that we call a canonical path. We impose the additional restriction that an edge can appear only once along a given canonical path. We write $|\gamma_{xy}|$ to denote the length of the canonical path γ_{xy}, and $\Gamma \stackrel{\text{def}}{=} \{\gamma_{xy},\ x,y \in \mathcal{V}\}$ for the set of all canonical paths.

We make the following assumption.

H2. There exists a subset $\mathcal{X}_0 \subseteq \mathcal{X}$ such that $\pi(\mathcal{X}_0) > 0$, and a connected undirected graph $(\mathcal{X}_0, \mathcal{E}_0)$ with canonical paths $\Gamma \stackrel{\text{def}}{=} \{\gamma_{xy},\ x,y \in \mathcal{X}_0\}$ such that for all $(x,y) \in \mathcal{E}_0$, $\pi(x)K(x,y) > 0$.

We define

$$\mathsf{m}(\mathcal{X}_0) \stackrel{\text{def}}{=} \max_{e \in \mathcal{E}_0} \sum_{\gamma_{xy}:\ \gamma_{xy} \ni e} \frac{|\gamma_{xy}|\pi(x)\pi(y)}{\pi(e_-)K(e_-,e_+)}. \tag{31}$$

When $\mathcal{X}_0 = \mathcal{X}$, $\mathsf{m}(\mathcal{X}_0)$ is the geometric measure of bottleneck of (Sinclair, 1992) (see also (Diaconis and Stroock, 1991)). In many cases, by carefully choosing $\mathcal{X}_0$, $\mathsf{m}(\mathcal{X}_0)$ scales better than $\mathsf{m}(\mathcal{X})$. The next result is analogous to Proposition 1 of (Diaconis and Stroock, 1991).

Proposition B.2. *Assume H2. Given $\epsilon \in [0, 1/2)$, if $\pi(\mathcal{X}_0) \geq 1 - \epsilon/8$, then $\lambda_\epsilon(K) \geq \mathsf{m}(\mathcal{X}_0)^{-1}$.*

Proof. Take a function $f:\ \mathcal{X}\to\mathbb{R}$ such that $\|f\|_\infty\le 1$, and $\mathsf{Var}_\pi(f)>\zeta$. We have

$$\begin{aligned}2\mathsf{Var}_\pi(f) &= \sum_{x\in\mathcal{X}_0}\sum_{y\in\mathcal{X}_0}(f(y)-f(x))^2\pi(x)\pi(y) + 2\sum_{x\in\mathcal{X}_0}\sum_{y\in\mathcal{X}\setminus\mathcal{X}_0}(f(y)-f(x))^2\pi(x)\pi(y)\\ &\quad+\sum_{x\in\mathcal{X}\setminus\mathcal{X}_0}\sum_{x\in\mathcal{X}\setminus\mathcal{X}_0}(f(y)-f(x))^2\pi(x)\pi(y)\\ &\le \sum_{x\in\mathcal{X}_0}\sum_{y\in\mathcal{X}_0}(f(y)-f(x))^2\pi(x)\pi(y)+8\pi(\mathcal{X}\setminus\mathcal{X}_0).\end{aligned}$$

Using $\pi(\mathcal{X}_0)\ge 1-(\zeta/8)$, we get

$$2\left(\mathsf{Var}_\pi(f)-\frac{\zeta}{2}\right)\le\sum_{x\in\mathcal{X}_0}\sum_{y\in\mathcal{X}_0}(f(y)-f(x))^2\pi(x)\pi(y).$$

Hence

$$\frac{\mathcal{E}(f,f)}{\mathsf{Var}_\pi(f)-\frac{\zeta}{2}}\ge\frac{\sum_{x\in\mathcal{X}_0}\sum_{y\in\mathcal{X}_0}(f(y)-f(x))^2\pi(x)K(x,y)}{\sum_{x\in\mathcal{X}_0}\sum_{y\in\mathcal{X}_0}(f(y)-f(x))^2\pi(x)\pi(y)}.$$

For $x\neq y$ in $\mathcal{X}_0$, let γ_{xy} be a canonical path linking x,y so that we can write, using the Cauchy-Schwarz inequality:

$$\begin{aligned}(f(y)-f(x))^2 &= \left(\sum_{e\in\gamma_{xy}}f(e_-)-f(e_+)\right)^2\\ &\le\sum_{e\in\gamma_{xy}}\frac{|\gamma_{xy}|}{\pi(e_-)K(e_-,e_+)}\times\pi(e_-)K(e_-,e_+)(f(e_-)-f(e_+))^2.\end{aligned}$$

It follows that

$$\begin{aligned}\sum_{x\neq y}(f(y)-f(x))^2\pi(x)\pi(y) &\le\sum_{x\neq y}\sum_{e\in\gamma_{xy}}\frac{\pi(x)\pi(y)|\gamma_{xy}|}{\pi(e_-)K(e_-,e_+)}\left\{\pi(e_-)K(e_-,e_+)(f(e_-)-f(e_+))^2\right\}\\ &=\sum_{e\in\mathcal{E}_0}\sum_{\gamma_{xy}\ni e}\left\{\frac{\pi(x)\pi(y)|\gamma_{xy}|}{\pi(e_-)K(e_-,e_+)}\right\}\left\{\pi(e_-)K(e_-,e_+)(f(e_-)-f(e_+))^2\right\}\\ &\le\mathsf{m}(\mathcal{X}_0)\times\sum_{e\in\mathcal{E}_0}\pi(e_-)K(e_-,e_+)(f(e_-)-f(e_+))^2\\ &\le\mathsf{m}(\mathcal{X}_0)\times\sum_{x\in\mathcal{X}_0}\sum_{y\in\mathcal{X}_0}(f(y)-f(x))^2\pi(x)K(x,y).\end{aligned}$$

As a result we conclude that for all $f:\ \mathcal{X}\to\mathbb{R}$, with $\|f\|_\infty\le 1$, and $\mathsf{Var}_\pi(f)>\zeta$

$$\frac{\mathcal{E}(f,f)}{\mathsf{Var}_\pi(f)-\frac{\zeta}{2}}\ge\frac{1}{\mathsf{m}(\mathcal{X}_0)}.$$

□

B.2 Proof of Theorem 2.5

The theorem is proved by applying Proposition B.2 and Lemma B.1 to the Gibbs sampler chain produced by OLAP. Indeed, when $J=1$, Algorithm 1 is also known as a random scan Gibbs sampler. It generates a reversible and positive Markov chain with invariant distribution $\check{\Pi}$ that fits the framework presented in Section B.1. We denote K its transition kernel. We prove the result by applying Lemma B.1 and Proposition B.2. The initial distribution of Algorithm 1 has a density f_0 with respect to $\check{\Pi}$ given by: $f_0(\delta)=1/\check{\Pi}(\delta^{(0)}|\mathcal{D})$, if $\delta=\delta^{(0)}$, and zero everywhere else. Hence,

$$\|f_0\|_\infty=\frac{1}{\check{\Pi}(\delta^{(0)}|\mathcal{D})}\le p^\alpha.$$

Fix $\Delta_0 \subseteq \Delta$ to be determined later such that $\check{\Pi}(\Delta_0|\mathcal{D}) \geq 15/16$, and set $\zeta = 8(1 - \check{\Pi}(\Delta_0|\mathcal{D}))$. If we can build canonical paths on Δ_0 such that H2 holds, then by Proposition B.2, we get $\lambda_\zeta(K) \geq 1/\mathsf{m}(\Delta_0)$. And since $\log(1-x) \leq -x$ for all $x \in (0,1]$, it follows that for $N \geq 1$ large enough such that

$$N \geq 2\mathsf{m}(\Delta_0) \log\left(\frac{\|f_0\|_\infty^2}{\zeta_0^2}\right),$$

we have

$$\left(1 - \frac{\lambda_\zeta(K)}{2}\right)^N = \exp\left(N \log\left(1 - \frac{\lambda_\zeta(K)}{2}\right)\right) \leq \exp\left(-\frac{N}{2\mathsf{m}(\Delta_0)}\right) \leq \frac{\zeta_0^2}{\|f_0\|_\infty^2}.$$

Therefore, by Lemma B.1,

$$\|K^N(\delta^{(0)}, \cdot) - \check{\Pi}\|_{\mathrm{tv}}^2 \leq \max\left(\zeta\|f_0\|_\infty^2, \left(1 - \frac{\lambda_\zeta(K)}{2}\right)^N \|f_0\|_\infty^2\right) \leq \max\left(\zeta\|f_0\|_\infty^2, \zeta_0^2\right).$$

We apply this result with the choice $\Delta_0 = \{\delta \in \Delta : \ \|\delta\|_0 \leq s_\star + J_0\}$. By (25),

$$\zeta\|f_0\|_\infty^2 = 8\|f_0\|_\infty^2(1 - \check{\Pi}(\Delta_0|\mathcal{D}) \leq \frac{16\|f_0\|_\infty^2}{p^{\mathsf{u}(J_0+1)/2}} \leq \frac{16}{p^{\mathsf{u}/2}} \frac{\|f_0\|_\infty^2}{p^{\mathsf{u}J_0/2}} \leq \frac{16}{p^{\mathsf{u}/2}},$$

by taking $J_0 = 4\alpha/\mathsf{u}$ Hence, for

$$N \geq 4\mathsf{m}(\Delta_0)\left(\log(1/\zeta_0) + \alpha\log(p)\right),$$

$$\|K^N(\delta^{(0)}, \cdot) - \check{\Pi}\|_{\mathrm{tv}} \leq \max\left(\frac{4}{p^{\mathsf{u}/4}}, \zeta_0\right). \tag{32}$$

We now use the canonical path argument to upper bound the term $\mathsf{m}(\Delta_0)$. First, we build a graph on Δ_0 by putting an edge between δ and δ' if $\|\delta - \delta'\|_0 = 1$. Clearly, H2 holds. We build the canonical paths by removing or adding variables as follows. For any δ, we first build a path between δ and $\delta_\star$. First, we set to zero (1 term at the time and in their decreasing index order) the components j of δ for which $\delta_j = 1$ and $\delta_{\star j} = 0$. We then reach $\underline{\delta} \stackrel{\text{def}}{=} \min(\delta, \delta_\star)$. This corresponds to removing nonrelevant variables one by one from the model δ. Then we set to one (one at the time, and in their increasing index order) the component j of $\underline{\delta}$ for which $\underline{\delta}_j = 0$, and $\delta_{\star j} = 1$. We then reach $\delta_\star$. This corresponds to adding one-by-one relevant variables not already contained in $\underline{\delta}$.

Given two arbitrary points δ and δ', we build the canonical path between them as follows. Let ϑ denote the point where the canonical path from δ to $\delta_\star$ and the canonical path from δ' to $\delta_\star$ meet for the first time. The canonical path from δ to δ' is obtained by following the canonical path from δ to $\delta_\star$ until ϑ is reached, then we follow the canonical path from δ' to $\delta_\star$ in reverse direction starting from ϑ until δ'. We have

$$|\gamma_{\delta,\delta'}| \leq 2\left(s_\star + J_0\right). \tag{33}$$

Let x, y denote the generic elements of Δ_0. Fix an edge $e = (\delta', \delta)$. Let $\Lambda(e)$ be the set of all elements of Δ_0 whose path to $\delta_\star$ goes through e. If $\gamma_{xy} \ni e$ then $x \in \Lambda(e)$ or $y \in \Lambda(e)$ (but not both). Therefore,

$$\begin{aligned}\sum_{\gamma_{xy}:\, \gamma_{xy} \ni e} \frac{|\gamma_{xy}|\check{\Pi}(x|\mathcal{D})\check{\Pi}(y|\mathcal{D})}{\check{\Pi}(\delta'|\mathcal{D})K(\delta',\delta)} &\leq 2\sum_{x\in\Lambda(e)}\sum_{y\in\mathcal{X}_0} \frac{|\gamma_{xy}|\check{\Pi}(x|\mathcal{D})\check{\Pi}(y|\mathcal{D})}{\check{\Pi}(\delta'|\mathcal{D})K(\delta',\delta)} \\ &\leq 2\sum_{x\in\Lambda(e)} \frac{|\gamma_{xy}|\check{\Pi}(x|\mathcal{D})}{\check{\Pi}(\delta'|\mathcal{D})K(\delta',\delta)}.\end{aligned}$$

Using the last display with (33), we conclude that

$$\mathsf{m}(\Delta_0) \leq 4\left(s_\star + J_0\right) \max_{(\delta',\delta)\in\mathcal{E}_0} \frac{1}{K(\delta',\delta)} \sum_{x\in\Lambda(e)} \frac{\Pi(x|\mathcal{D})}{\Pi(\delta'|\mathcal{D})}.$$

Case 1: $\delta = \delta^{(j,0)}$, and $\delta' = \delta^{(j,1)}$, for some j such that $\delta_{\star j} = 0$. In this case,

$$\begin{aligned}
K(\delta', \delta) =& \frac{1}{p}\left(1 - q_j(\delta')\right) \\
=& \frac{1}{p}\left(1 + \exp(-\mathsf{u}\log(p) + \bar{\ell}^{\delta^{(j,1)}}(\check{\theta}_{\delta^{(j,1)}}; \mathcal{D}) - \bar{\ell}^{\delta^{(j,0)}}(\check{\theta}_{\delta^{(j,0)}}; \mathcal{D}))\right)^{-1} \\
\geq& \frac{1}{p}\left(1 + \frac{1}{p^{\mathsf{u}/2}}\right)^{-1} \geq \frac{1}{p}\left(1 - \frac{1}{p^{\mathsf{u}/2}}\right) \geq \frac{1}{2p},
\end{aligned}$$

where we use (11) of the main document to claim that $\bar{\ell}^{\delta^{(j,1)}}(\check{\theta}_{\delta^{(j,1)}}; \mathcal{D}) - \bar{\ell}^{\delta^{(j,0)}}(\check{\theta}_{\delta^{(j,0)}}; \mathcal{D}) \leq c_1 \log(p)$, and for $\mathsf{u} > 2c_1$, and $p^{\mathsf{u}/2} \geq 2$. Furthermore, in this case, note that $\delta' \supset \delta$, and differs from δ only at some j such that $\delta_{\star j} = 0$. Therefore, $\Lambda(e) = \Delta(\delta')$, that is the elements $\vartheta \in \Delta_{s_\star + J_0}$ contains δ' and differs from δ' only at components at which $\delta_{\star k} = 0$. Therefore, as seen in the proof of Theorem 2.2,

$$\sum_{x \in \Lambda(e)} \frac{\check{\Pi}(x|\mathcal{D})}{\check{\Pi}(\delta'|\mathcal{D})} \leq 2.$$

and it follows that

$$\frac{1}{K(\delta', \delta)} \sum_{x \in \Lambda(e)} \frac{\Pi(x|\mathcal{D})}{\Pi(\delta'|\mathcal{D})} \leq 4p.$$

Case 2: $\delta' \subset \delta_\star$, $\delta' = \delta^{(j,0)}$, and $\delta = \delta^{(j,1)}$, for some j such that $\delta_{\star j} = 1$. In this case, using (12) of the main document,

$$\begin{aligned}
K(\delta', \delta) =& \frac{q_j(\delta')}{p} = \frac{1}{p}\left(1 + \exp(\mathsf{u}\log(p) - \bar{\ell}^{\delta^{(j,1)}}(\check{\theta}_{\delta^{(j,1)}}; \mathcal{D}) + \bar{\ell}^{\delta^{(j,0)}}(\check{\theta}_{\delta^{(j,0)}}; \mathcal{D}))\right)^{-1} \\
\geq& \frac{1}{p}\left(1 + \exp\left(\mathsf{u}\log(p) - c_2 n\right)\right)^{-1} \geq \frac{1}{p}\left(1 + \frac{1}{e^{c_2 n/2}}\right) \geq \frac{1}{2p},
\end{aligned}$$

for $c_2 n \geq 2\mathsf{u}\log(p)$. In this configuration, we see that

$$\Lambda(e) = \bigcup_{\vartheta \subseteq \delta'} \Delta(\vartheta),$$

therefore,

$$\sum_{x \in \Lambda(e)} \frac{\check{\Pi}(x|\mathcal{D})}{\check{\Pi}(\delta'|\mathcal{D})} = \sum_{\vartheta \subseteq \delta'} \frac{\check{\Pi}(\vartheta|\mathcal{D})}{\check{\Pi}(\delta'|\mathcal{D})} \sum_{x \in \Delta(\vartheta)} \frac{\check{\Pi}(x|\mathcal{D})}{\check{\Pi}(\vartheta|\mathcal{D})} \leq 2 \sum_{\vartheta \subseteq \delta'} \frac{\check{\Pi}(\vartheta|\mathcal{D})}{\check{\Pi}(\delta'|\mathcal{D})}.$$

We proceed similarly as in the proof of Theorem 2.2 to show that

$$\sum_{\vartheta \subseteq \delta'} \frac{\check{\Pi}(\vartheta|\mathcal{D})}{\check{\Pi}(\delta'|\mathcal{D})} \leq 1 + 4e^{-c_2 n/2} \leq 2,$$

for $c_2 n \geq 2(\mathsf{u}+1)\log(p)$. We conclude that for some absolute constant C.

$$\mathsf{m}(\Delta_0) \leq C(s_\star + J_0)p.$$

We combine this bound with (32) and $J_0 = 4\alpha/\mathsf{u}$, to conclude that

$$\|K^N(\delta^{(0)}, \cdot) - \check{\Pi}\|_{\mathrm{tv}} \leq \max\left(\frac{4}{p^{\mathsf{u}/4}}, \zeta_0\right), \text{ for } N \geq C\left(s_\star + \frac{4\alpha}{\mathsf{u}}\right)\left(\log(1/\zeta_0) + \alpha\log(p)\right)p,$$

for some absolute constant C. This concludes the proof.

B.3 Further Discussions on the Warm-Start in Algorithm 1

Suppose that we slightly strengthen the definition of model selection consistency of $\{\check{\theta}^{\delta},\ \delta \in \Delta\}$ by replacing (11) and (12) respectively by

$$\bar{\ell}(\check{\theta}^{\delta_0};\mathcal{D}) \leq \bar{\ell}(\check{\theta}^{\delta};\mathcal{D}) \leq \bar{\ell}(\check{\theta}^{\delta_0};\mathcal{D}) + c_1(\|\delta\|_0 - \|\delta_0\|_0)\log(p), \tag{34}$$

and

$$\bar{\ell}(\check{\theta}^{\delta};\mathcal{D}) + c_3(\|\delta_\star\|_0 - \|\delta\|_0)n \ \ \geq \ \ \bar{\ell}(\check{\theta}_{\delta_\star};\mathcal{D}) \geq \bar{\ell}(\check{\theta}^{\delta};\mathcal{D}) + c_2(\|\delta_\star\|_0 - \|\delta\|_0)n. \tag{35}$$

for some constant $c_3 \geq c_2$. Now, suppose that we run Algorithm 1 from some initial state $\delta^{(0)} \supseteq \delta_\star$, with $\|\delta^{(0)}\|_0 = k_0$. In other words, the initial state has no false-negative, but $k_0 - s_\star$ false-positives. Then, noting that $\check{\Pi}(\delta_\star|\mathcal{D}) \geq 1/2$, for n and p large enough as a consequence of Theorem 2.3, we have:

$$\check{\Pi}(\delta^{(0)}|\mathcal{D}) = \check{\Pi}(\delta_\star|\mathcal{D})\frac{\check{\Pi}(\delta^{(0)}|\mathcal{D})}{\check{\Pi}(\delta_\star|\mathcal{D})} \tag{36}$$

$$\geq \frac{1}{2p^{\mathsf{u}(k_0-s_\star)}}\exp\Big(\bar{\ell}(\check{\theta}^{\delta^{(0)}};\mathcal{D}) - \bar{\ell}(\check{\theta}^{\delta_\star};\mathcal{D})\Big). \tag{37}$$

By the first inequality of (34), $\bar{\ell}(\check{\theta}^{\delta^{(0)}};\mathcal{D}) - \bar{\ell}(\check{\theta}^{\delta_\star};\mathcal{D}) \geq 0$. Hence, we conclude that

$$\check{\Pi}(\delta^{(0)}|\mathcal{D}) \geq \frac{1}{2p^{\mathsf{u}(k_0-s_\star)}}.$$

This inequality means that we can apply Theorem 2.5 with $\alpha = \mathsf{u}(k_0 - s_\star)/2$, and it follows that OLAP has a mixing time that is at most $k_0^2\log(p) \times p$. Crucially, in this case, the mixing time of the algorithm does not directly depend on the sample size n.

Consider now the case where the initial state $\delta^{(0)}$ contains some false negatives. Specifically, suppose that we start OLAP from $\delta^{(0)}$ such that $\|\delta^{(0)}\|_0 = k_0$, but $\delta_\star^{(0)} \stackrel{\text{def}}{=} \min(\delta^{(0)}, \delta_\star)$ is a strict sub-model of $\delta_\star$, say $\|\delta_\star^{(0)}\|_0 = s_0 < s_\star$. In that case, using the definition of $\check{\Pi}$ in (10), and $\check{\Pi}(\delta_\star|\mathcal{D}) \geq 1/2$,

$$\check{\Pi}(\delta^{(0)} \mid \mathcal{D}) = \check{\Pi}(\delta_\star \mid \mathcal{D})\frac{\check{\Pi}(\delta_\star^{(0)} \mid \mathcal{D})}{\check{\Pi}(\delta_\star \mid \mathcal{D})}\frac{\check{\Pi}(\delta^{(0)} \mid \mathcal{D})}{\check{\Pi}(\delta_\star^{(0)} \mid \mathcal{D})} \tag{38}$$

$$\geq \frac{1}{2p^{\mathsf{u}(k_0-s_\star)}}\exp\Big(\bar{\ell}(\check{\theta}^{\delta_\star^{(0)}};\mathcal{D}) - \bar{\ell}(\check{\theta}^{\delta_\star};\mathcal{D}) + \bar{\ell}(\check{\theta}^{\delta^{(0)}};\mathcal{D}) - \bar{\ell}(\check{\theta}^{\delta_\star^{(0)}};\mathcal{D})\Big). \tag{39}$$

By the first inequality of (34), $\bar{\ell}(\check{\theta}^{\delta^{(0)}};\mathcal{D}) - \bar{\ell}(\check{\theta}^{\delta_\star^{(0)}};\mathcal{D}) \geq 0$, and the first inequality of (35) yields $\bar{\ell}(\check{\theta}^{\delta_\star^{(0)}};\mathcal{D}) - \bar{\ell}(\check{\theta}^{\delta_\star};\mathcal{D}) \geq -c_3(s_\star - s_0)n$. Hence, it follows that

$$\check{\Pi}(\delta^{(0)}|\mathcal{D}) \geq \frac{\exp\left(-c_3(s_\star - s_0)n\right)}{2p^{\mathsf{u}(k_0-s_\star)}} \geq \frac{1}{2p^{\alpha}},$$

with $\alpha = \mathsf{u}(k_0 - s_\star) + \frac{c_3(s_\star - s_0)n}{\log(p)}$. In that case, it follows from Theorem 2.5 that OLAP has a mixing time that is at most

$$\left(k_0 + \frac{(s_\star - s_0)n}{\log(p)}\right)^2 \log(p) \times p.$$

Here, the mixing time is still polynomial in (n, p), however, the result suggests that the mixing time is negatively impacted by the sample size n. This is similar to the worst-case mixing time bound of $ns_0^2p\log(p)$ obtained by (Yang et al., 2016) for sampling from a version of (5) for variable selection in a high-dimensional linear regression model, where, using their notation, s_0 denotes an upper-bound on $s_\star$. Although we have a worst dependence on n, our work improves on (Yang et al., 2016) in two ways. Firstly, instead of a worst-case analysis, our result describes more precisely the effect of initializations. Secondly, our analysis does not require an a priori bound on

$s_\star$. We stress however that our results (as well as (Yang et al., 2016)) only provide upper bounds on the mixing times. It is possible that OLAP actually converges faster than what is described in Theorem 2.5.

We illustrate these findings with an example of Poisson regression simulation. The data generation process is described in Section 3. Here we set $p = 1000$ and we vary $n \in \{1000, 1500, 2000\}$. The true model has 10 relevant variables. For each (n, p) we generate 50 datasets, leading to 50 OLAP distributions $\check{\Pi}(\cdot|\mathcal{D})$.

For each posterior, we estimate the mixing time of OLAP using the L-lag coupling method of (Biswas et al., 2019) (see Section B.4 for more details), employing 50 coupled chains. We compare the mixing times of OLAP when started from the null model, and a version started from $\delta^{(0)}$ that contains the true model plus 10 false-positives. The boxplots of the distributions of the mixing times are given in Figure 1. The results are consistent with our theory and show that with an initialization without false-negatives, the mixing times is independent of the sample size. These findings highlight the importance of good MCMC initialization (warm-start).

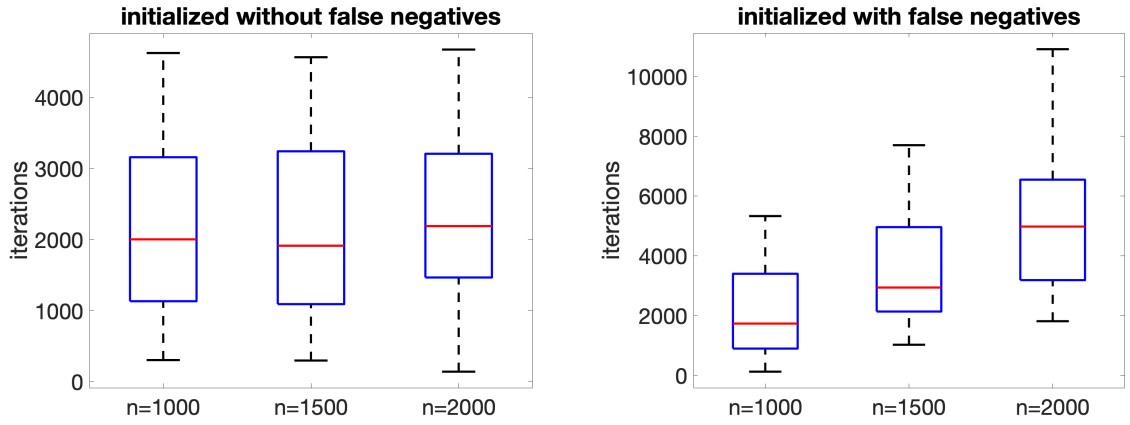

Figure 1: Estimated mixing times of OLAP under different initialization $\delta^{(0)}$ and for different data size n. Left: $\delta^{(0)}$ contains all relevant variables and 10 false-positive. Right: $\delta^{(0)}$ is null model.

B.4 Further Mixing Times Comparison Between Algorithm 1 and the Exact Sampler of (Atchade and Bhattacharyya, 2019)

To further illustrate the advantage of OLAP we compare the mixing times of Algorithm 1 with the Metropolis-within-Gibbs sampler targeting the augmented posterior of (Atchade and Bhattacharyya, 2019) described in Section D.1. We estimate mixing times using the L-lag coupling method (Biswas et al., 2019; Jacob et al., 2020). We run coupled versions of the samplers until a coupling event and relate mixing to a moment of the coupling time. In OLAP we use maximal coupling of the Bernoulli draws. For the Metropolis-within-Gibbs sampler targeting the augmented posterior of (Atchade and Bhattacharyya, 2019), we construct the coupling as in (Jacob et al., 2020).

For logistic regression, the results are shown in Figure 2 (low correlation setting) and 3 (high correlation setting). In this logistic regression setting, $n = p = 1000$. The results show that OLAP is an order of magnitude faster than the exact algorithm, both in terms of number of iterations and real time.

For Poisson regression, the results are shown in Figure 4 and 5. Here we set $n = p = 1000$ in the low correlation case, and $n = 2000$, $p = 1000$ in the high correlation case. The results are similar, except in the high-correlation setting where the convergence of OLAP is slower.

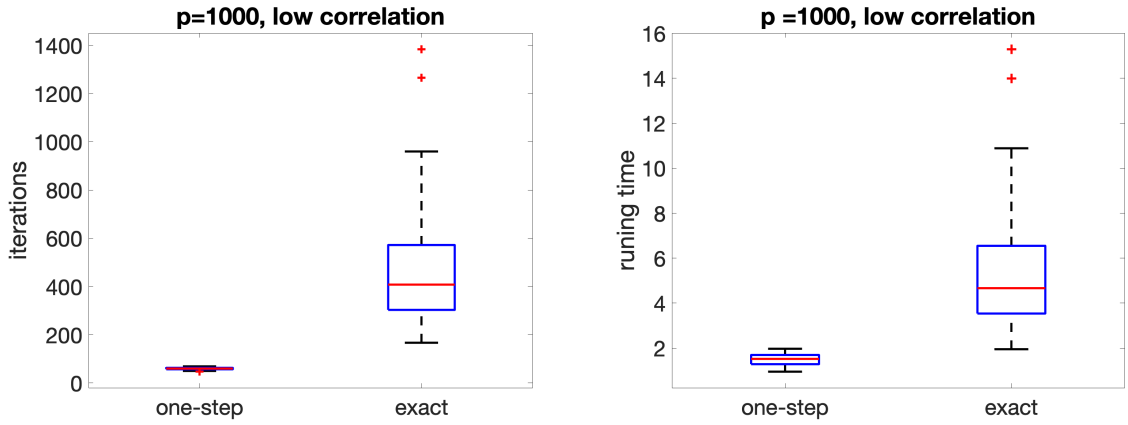

Figure 2: Iterations and mixing time (seconds), logistic regression, low correlation.

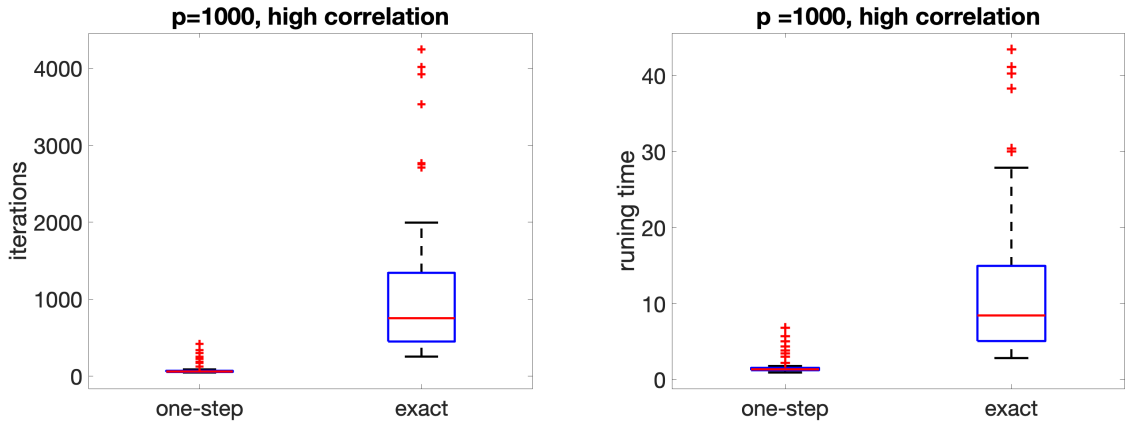

Figure 3: Iterations and mixing time (seconds), logistic regression, high correlation.

C PROOF OF THEOREM 2.7

Proof. Let $q^\delta(\cdot;\mathcal{D})$ denote the quadratic approximation of $\bar{\ell}^\delta(\cdot;\mathcal{D})$ around $\tilde{\theta}^\delta$. Specifically, for $u \in \mathbb{R}^{\|\delta\|_0}$,

$$q^\delta(u;\mathcal{D}) \stackrel{\text{def}}{=} \bar{\ell}^\delta(\tilde{\theta}^\delta;\mathcal{D}) + \left\langle \tilde{\mathcal{G}}_\delta, u - \tilde{\theta}^\delta \right\rangle - \frac{1}{2}(u-\tilde{\theta}^\delta)'\tilde{\mathcal{H}}_\delta(u-\tilde{\theta}^\delta),$$

and

$$\eta^\delta(u;\mathcal{D}) \stackrel{\text{def}}{=} \bar{\ell}^\delta(u;\mathcal{D}) - q^\delta(u;\mathcal{D}).$$

An important step in the argument is to upper bound the remainder $|\eta^\delta(u;\mathcal{D})|$ for u close to $\theta_\star$. This can be easily done in the setting where the third derivative of ψ is uniformly bounded. However, this will leave out the Poisson model. Following (Bach, 2010), we use the self-concordance assumption in H1-(3) to handle a larger class of link functions ψ. Specifically, under assumption H1, the following holds. We can find a constant C that depends only on c_3, $\bar{\kappa}$ and b in H1 such that for all $\delta \subseteq \delta_\star$,

$$\left|\eta^\delta(u;\mathcal{D})\right| \leq Cne^{c_3 b\|u-\tilde{\theta}^\delta\|_1} \times \|u-\tilde{\theta}^\delta\|_1 \times \|u-\tilde{\theta}^\delta\|_2^2, \quad u \in \mathbb{R}^{\|\delta\|_0}. \tag{40}$$

This claim is proved below. For the time being, assume that (40) holds and fix $\delta, \delta_0 \in \Delta$. Since $\hat{\theta}^\delta$ maximizes $\bar{\ell}^\delta(\cdot;\mathcal{D})$, we have

$$\bar{\ell}^\delta(\check{\theta}^\delta;\mathcal{D}) \leq \bar{\ell}^\delta(\hat{\theta}^\delta;\mathcal{D}). \tag{41}$$

But since $\check{\theta}^\delta$ maximizes $q^\delta(\cdot;\mathcal{D})$, we have

$$\bar{\ell}^\delta(\check{\theta}^\delta;\mathcal{D}) \;=\; q^\delta(\check{\theta}^\delta;\mathcal{D}) + \eta^\delta(\check{\theta}^\delta;\mathcal{D}) \;\geq\; q^\delta(\hat{\theta}^\delta;\mathcal{D}) + \eta^\delta(\check{\theta}^\delta;\mathcal{D}) \;=\; \bar{\ell}^\delta(\hat{\theta}^\delta;\mathcal{D}) + \eta^\delta(\check{\theta}^\delta;\mathcal{D}) - \eta^\delta(\hat{\theta}^\delta;\mathcal{D}). \tag{42}$$

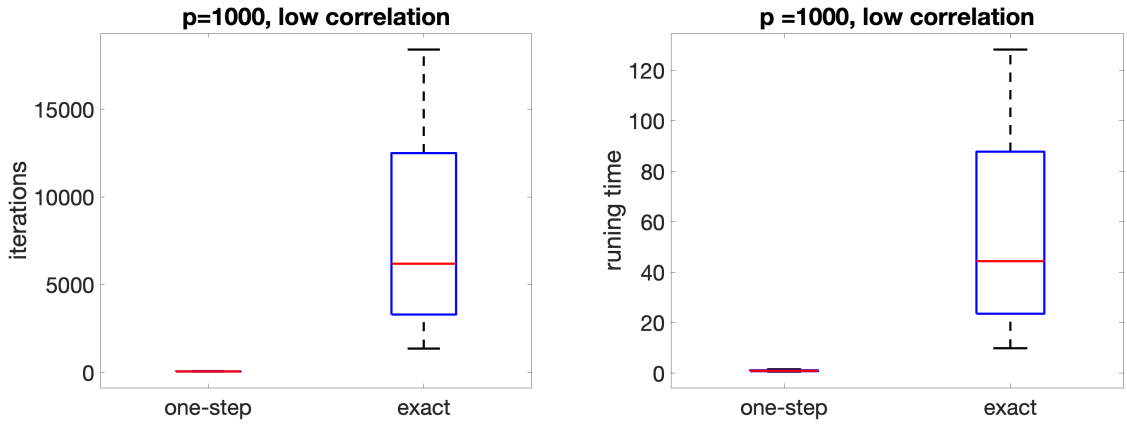

Figure 4: Iterations and mixing time (seconds), Poisson regression, low correlation.

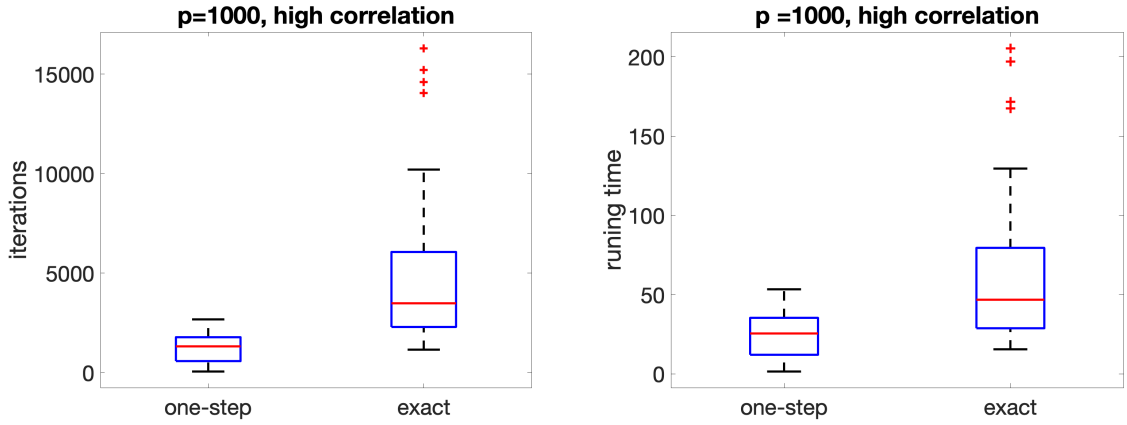

Figure 5: Iterations and mixing time (seconds), Poisson regression, high correlation.

We combine (41) and (42) to conclude that

$$\bar{\ell}^{\delta}(\check{\theta}^{\delta};\mathcal{D}) - \bar{\ell}^{\delta_0}(\check{\theta}^{\delta_0};\mathcal{D}) \leq \left[\bar{\ell}^{\delta}(\hat{\theta}^{\delta};\mathcal{D}) - \bar{\ell}^{\delta_0}(\hat{\theta}^{\delta_0};\mathcal{D})\right] + |\eta^{\delta_0}(\check{\theta}^{\delta_0};\mathcal{D})| + |\eta^{\delta_0}(\hat{\theta}^{\delta_0};\mathcal{D})|. \tag{43}$$

Hence, proving the theorem boils down to showing that there exists a constant $C < \infty$ such that for all $\delta_0 \subseteq \delta_\star$,

$$|\eta^{\delta_0}(\check{\theta}^{\delta_0};\mathcal{D})| + |\eta^{\delta_0}(\hat{\theta}^{\delta_0};\mathcal{D})| \leq C\log(p). \tag{44}$$

Indeed, if the MLE family $\{\hat{\theta}^{\delta},\ \delta \in \Delta\}$ is variable selection consistent with constant c_1, c_2, say, then for $\delta_0 \subseteq \delta$, and $\min(\delta, \delta_\star) = \delta_0$, we can conclude from the above argument that

$$\bar{\ell}^{\delta}(\check{\theta}^{\delta};\mathcal{D}) - \bar{\ell}^{\delta_0}(\check{\theta}^{\delta_0};\mathcal{D}) \leq c_1\left(\|\delta\|_0 - \|\delta_0\|_0\right)\log(p) + C\log(p) \leq (c_1 + C)\left(\|\delta\|_0 - \|\delta_0\|_0\right)\log(p).$$

Similarly, for $\delta_0 \subseteq \delta_\star$, applying again (43) with $\delta_0 \leftarrow \delta_\star$, and $\delta \leftarrow \delta_0$, we get

$$\bar{\ell}^{\delta_\star}(\check{\theta}^{\delta_\star};\mathcal{D}) - \bar{\ell}^{\delta_0}(\check{\theta}^{\delta_0};\mathcal{D}) \geq c_2\left(\|\delta_\star\|_0 - \|\delta_0\|_0\right)n - C\log(p) \geq \frac{c_2}{2}\left(\|\delta_\star\|_0 - \|\delta_0\|_0\right)n,$$

for $c_2 n \geq 2C\log(p)$. This establishes the theorem. It remains to prove (44) and (40).

Proof of (40) Given, $u, h \in \mathbb{R}$, with $h \neq 0$, define

$$g(t) = \psi(u + th), \quad t \in [0, 1].$$

H1-(3) implies that $|g^{'''}(t)| \leq c_3|h|g^{''}(t)$, which in turn implies that for all $t \in [0,1]$,

$$-c_3|h| \leq \frac{\mathrm{d}\log g^{''}(t)}{\mathrm{d}t} \leq c_3|h|.$$

Integrating these two inequalities thrice we obtain for all $t \in [0,1]$,

$$g^{''}(0)e^{-c_3|h|t} \leq g^{''}(t) \leq g^{''}(0)e^{c_3|h|t}, \tag{45}$$

$$g^{''}(0) \times \frac{1-e^{-c_3|h|t}}{c_3|h|} \leq g^{'}(t) - g^{'}(0) \leq g^{''}(0) \times \frac{e^{c_3|h|t}-1}{c_3|h|}, \tag{46}$$

and

$$-g^{''}(0) \times \frac{1-c_3|h|t-e^{-c_3|h|t}}{c_3^2h^2} \leq g(t) - g(0) - tg^{'}(0) \leq g^{''}(0) \times \frac{e^{c_3|h|t}-c_3|h|t-1}{c_3^2h^2}.$$

Subtracting $g^{''}(0)/t^2$ from all sides we get

$$\begin{aligned}-g^{''}(0) \times \frac{1-c_3|h|t+(c_3|h|t)^2/2-e^{-c_3|h|t}}{c_3^2h^2} \leq& g(t) - g(0) - tg^{'}(0) - \frac{t^2}{2}g^{''}(0) \\ \leq& g^{''}(0) \times \frac{e^{c_3|h|t}-(c_3|h|t)^2/2-c_3|h|t-1}{c_3^2h^2}.\end{aligned}$$

We apply this with $t=1$, and we use a Taylor of e^x and e^{-x} to the third order to conclude that

$$\left|\psi(u+h) - \psi(u) - \psi'(u)h - \frac{h^2}{2}\psi^{''}(u)\right| \leq \frac{c_3}{6}|h|^3e^{c_3|h|}\psi^{''}(u). \tag{47}$$

Since for all $u \in \mathbb{R}^{\delta\|_0}$,

$$\eta^\delta(u;\mathcal{D}) = \sum_{i=1}^n \left[\psi\left(\langle u, \mathbf{x}_i\rangle\right) - \psi\left(\left\langle\tilde{\theta}^\delta, \mathbf{x}_i\right\rangle\right) - \psi'\left(\left\langle\tilde{\theta}^\delta, \mathbf{x}_i\right\rangle\right)\left\langle u-\tilde{\theta}^\delta, \mathbf{x}_i\right\rangle - \frac{1}{2}\psi^{''}\left(\left\langle\tilde{\theta}^\delta, \mathbf{x}_i\right\rangle\right)\left\langle u-\tilde{\theta}^\delta, \mathbf{x}_i\right\rangle^2\right].$$

We conclude with (47) that

$$\begin{aligned}\left|\eta^\delta(u;\mathcal{D})\right| \leq& \frac{c_3b}{6}\|u-\tilde{\theta}^\delta\|_1 e^{c_3b\|u-\tilde{\theta}^\delta\|_1}\sum_{i=1}^n \psi^{''}\left(\left\langle\tilde{\theta}^\delta, \mathbf{x}_i\right\rangle\right)\left\langle u-\tilde{\theta}^\delta, \mathbf{x}_i\right\rangle^2 \\ \leq& \frac{c_3b\bar{\kappa}n}{6}e^{c_3b\|u-\tilde{\theta}^\delta\|_1}\|u-\tilde{\theta}^\delta\|_1 \times \|u-\tilde{\theta}^\delta\|_2^2,\end{aligned}$$

which is the claim (40).

Proof of (44) Fix $\delta_0 \subseteq \delta_\star$. Since $\hat{\theta}^{\delta_0} - \tilde{\theta}^{\delta_0} = \hat{\theta}^{\delta_0} - \theta_\star^{\delta_0} + \theta_\star^{\delta_0} - \tilde{\theta}^{\delta_0}$, and $\delta_0 \subseteq \delta_\star$, we get, using (23) of main document,

$$\|\hat{\theta}^{\delta_0} - \tilde{\theta}^{\delta_0}\|_2 \leq \|\hat{\theta}^{\delta_\star} - \theta_\star\|_2 + \|\widetilde{\theta} - \theta_\star\|_2 \leq C\sqrt{\frac{s_\star\log(p)}{n}}.$$

Hence

$$\|\hat{\theta}^{\delta_0} - \tilde{\theta}^{\delta_0}\|_1 \leq C\sqrt{\frac{s_\star^2\log(p)}{n}} \leq C',$$

for $n \geq s_\star^2\log(p)$. As a result, using (40), we conclude that

$$|\eta^{\delta_0}(\check{\theta}^{\delta_0};\mathcal{D})| \leq Cn\sqrt{\frac{s_\star^2\log(p)}{n}} \times \frac{s_\star\log(p)}{n} \leq C'\log(p),$$

under the sample size condition (24) of the main document. By the definition of $\check{\theta}^\delta$ in (7) of main document,

$$\check{\theta}^{\delta_0} - \tilde{\theta}^{\delta_0} = (\tilde{\mathcal{H}}^{\delta_0})^{-1}\tilde{\mathcal{G}}^{\delta_0},$$

and under H1-(4), the smallest eigenvalue of $\tilde{\mathcal{H}}^{\delta_0}$ is at least $\underline{\kappa}n$. Therefore, and using the expression of the gradient $\mathcal{G}^{\delta}$, we have

$$\|\check{\theta}^{\delta_0} - \tilde{\theta}^{\delta_0}\|_2 \leq \frac{1}{\underline{\kappa}n}\left(\left\|[\mathcal{G}_\star]_{\delta_0}\right\|_2 + \left\|\tilde{\mathcal{G}}^{\delta_0} - [\mathcal{G}_\star]_{\delta_0}\right\|_2\right),$$

where we define $\mathcal{G}_\star \stackrel{\text{def}}{=} \nabla\bar{\ell}^{\delta}(u;\mathcal{D})|_{u=\theta_\star}$. According to H1-(2), $\|\mathcal{G}_\star\|_\infty \leq M_0\sqrt{n\log(p)}$. Hence,

$$\frac{\left\|[\mathcal{G}_\star]_{\delta_0}\right\|_2}{\underline{\kappa}n} \leq C\sqrt{\frac{\|\delta_0\|_0\log(p)}{n}} \leq C\sqrt{\frac{s_\star\log(p)}{n}}.$$

On the other hand, a comparison between the j-th component of $\tilde{\mathcal{G}}^{\delta_0}$ and $[\mathcal{G}_\star]_{\delta_0}$ gives

$$\begin{aligned}(\tilde{\mathcal{G}}^{\delta_0})_j - [\mathcal{G}_\star]_j &= -\sum_{i=1}^n\left(\psi'(\left\langle\tilde{\theta}^{\delta_0},\mathbf{x}_i\right\rangle) - \psi'(\langle\theta_\star,\mathbf{x}_i\rangle)\right)[\mathbf{x}_i]_j \\ &= -\sum_{k:\ \delta_{\star,k}=1}(\tilde{\theta}_k - \theta_{\star,k})\sum_{i=1}^n\psi''(\langle\vartheta,\mathbf{x}_i\rangle)\mathbf{x}_{ij}\mathbf{x}_{ik},\end{aligned}$$

for some ϑ on the segment between $\tilde{\theta}^{\delta_0}$ and $\theta_\star$. We can then appeal to H1-(4-5) to conclude that

$$\left|(\tilde{\mathcal{G}}^{\delta_0})_j - [\mathcal{G}_\star]_j\right| \leq \bar{\kappa}n|\tilde{\theta}_j - \theta_{\star,j}| + M\|\widetilde{\theta} - \theta_\star\|_1\sqrt{n\log(p)} \leq \bar{\kappa}n|\tilde{\theta}_j - \theta_{\star,j}| + Cs_\star\log(p).$$

We deduce that

$$\frac{1}{\underline{\kappa}n}\left\|\tilde{\mathcal{G}}^{\delta_0} - [\mathcal{G}_\star]_{\delta_0}\right\|_2 \leq \frac{Cn}{\underline{\kappa}n}\sqrt{\frac{s_\star\log(p)}{n}} + \frac{Cs_\star^{3/2}\log(p)}{n} \leq C'\sqrt{\frac{s_\star\log(p)}{n}}.$$

Therefore, (40), and a similar argument yields

$$|\eta^{\delta_0}(\hat{\theta}^{\delta_0};\mathcal{D})| \leq Cn\sqrt{\frac{s_\star^2\log(p)}{n}} \times \frac{s_\star\log(p)}{n} \leq C'\log(p),$$

which concludes the proof of (44).

□

D NUMERICAL EXPERIMENTS: DESCRIPTION OF THE BENCHMARK METHODS

D.1 An exact full posterior distribution-sampling method: Exact

We consider the data-augmentation strategy proposed in (Atchade and Bhattacharyya, 2019). The method, denoted Exact, consists in sampling jointly (δ,θ) from the distribution

$$\Pi(\delta,\theta|\mathcal{D}) \propto \left(\frac{1}{p^{\mathsf{u}}}\sqrt{\frac{1}{\rho_0}}\right)^{\|\delta\|_0}\exp\left(-\frac{1}{2}\|\theta_\delta\|_2^2 - \frac{\rho_0}{2}\|\theta-\theta_\delta\|_2^2 + \ell(\theta_\delta;\mathcal{D})\right), \tag{48}$$

for some hyper-parameter ρ_0. One can sample from this distribution by Metropolis-within-Gibbs, alternating between a Gibbs update on δ given θ, and a Metropolis-Hastings update on θ given δ (here we use MaLA). We note that the marginal distribution of δ under (48) is precisely (3) in the main document. We use our own implementation of this algorithm with hyper-prameters $\mathsf{u} = 1.5$, $\rho_0 = p$, $\rho_1 = 1$.

D.2 Approximate Laplace Approximations for Scalable Model Selection for Scalable Model Selection: ALA-BB and ALA-Complex

(Rossell et al., 2021) propose the Approximate Laplace approximations method, denoted ALA, which targets the same intractable marginal likelihood arising in Bayesian model/variable selection. Instead of centering the Laplace approximation at the model-specific MLE, ALA performs a second-order Taylor expansion of the log-likelihood (or log-posterior) around a fixed baseline parameter, typically corresponding to the null model, and applies a Newton-type update to obtain a closed-form quadratic surrogate. In our empirical comparisons, we consider two ALA variants differing only in the prior distribution on $\delta \in \{0,1\}^p$, with ALA-BB using a beta-binomial prior and ALA-Complex using a complexity prior.

D.3 Penalized Likelihood Methods: Lasso, SCAD, MCP,ℓ_0, $\ell_0+\ell_1$, $\ell_0+\ell_2$

These methods solve the feature selection problem as the support of the solution of the minimization problem

$$\hat{\beta} = \arg\min_{\beta}\Big\{\ell(\beta;\mathcal{D}) + \sum_{j=1}^{p} R(\beta_j)\Big\},$$

for some penalty function R. The different methods correspond to different penalty functions. With lasso (Hastie et al., 2015) $R_\lambda(x) = \lambda|x|$, with a tuning parameter $\lambda > 0$. SCAD (Smoothly Clipped Absolute Deviation) (Fan and Li, 2001) and MCP (Minimax Concave Penalty) (Zhang, 2010) replace the convex ℓ^1 penalty of the Lasso with piecewise quadratic, non-convex penalties that remain flat beyond a data-adaptive threshold, thereby shrinking small coefficients to *exact* zero while leaving large ones essentially unbiased. Formally,

$$R^{\mathrm{SCAD}}_{\lambda,a}(\beta) = \begin{cases} \lambda|\beta|, & |\beta| \le \lambda, \\ \dfrac{2a\lambda|\beta| - \beta^2 - \lambda^2}{2(a-1)}, & \lambda < |\beta| \le a\lambda, \\ \dfrac{(a+1)\lambda^2}{2}, & |\beta| > a\lambda, \end{cases} \qquad R^{\mathrm{MCP}}_{\lambda,\gamma}(\beta) = \begin{cases} \lambda|\beta| - \dfrac{\beta^2}{2\gamma}, & |\beta| \le \gamma\lambda, \\ \dfrac{\gamma\lambda^2}{2}, & |\beta| > \gamma\lambda, \end{cases}$$

for tuning parameters λ, γ. The best sub-selection methods are denoted ℓ_0, $\ell_0+\ell_1$, $\ell_0+\ell_2$ (Hazimeh and Mazumder, 2020; Hazimeh et al., 2023) and corresponds respectively to $R(x) = \mathbf{1}_{\{x=0\}}$, $R(x) = \mathbf{1}_{\{x=0\}} + \lambda|x|$ and $R(x) = \mathbf{1}_{\{x=0\}} + \lambda x^2/2$.

Cross-validation is used to select the optimal regularization parameters or the optimal sparsity level for these methods. We compute Lasso with the R package glmnet, SCAD and MCP with R package ncvreg, and the best sub-selection methods with R package L0Learn.

D.4 ABESS: Adaptive Best Subset Selection

ABESS (Zhu et al., 2020) is a polynomial-time, provably consistent algorithm that incrementally grows the active set. Starting from the null model, ABESS adds (and occasionally swaps) variables via sure-screening, solves the restricted least-squares at each subset size, and uses warm starts to approximate the global best-subset path. Cross-validation is used to select the optimal regularization parameter of ABESS. We compute ABESS using the R package abess.

D.5 mRMR: Minimum Redundancy-Maximum Relevance

mRMR (Peng et al., 2005) is an information-theoretic filter. For a candidate subset S, it maximizes

$$\mathrm{mRMR}(S) = \frac{1}{|S|}\sum_{j\in S} I(X_j;Y) \;-\; \frac{1}{|S|^2}\sum_{j,k\in S} I(X_j;X_k),$$

where $I(\cdot;\cdot)$ denotes mutual information, thus favoring features that are individually informative about Y yet minimally redundant with each other; a greedy forward search is typically used. We compute mRMR using the R package mRMRe.

We manually choose the sparsity level $|S| = 5$, $|S| = 10$ and $|S| = 20$, namely mRMR-5, mRMR-10 and mRMR-20.

D.6 Sparse VB: Variational Bayes

SparseVB (Ray et al., 2020) assumes a spike-and-slab prior to the coefficients and uses a family of factorized variational Bayes distributions ("mean field ") to approximate the posterior distribution; the Coordinated ascend variational inference (CAVI) returns posterior inclusion probabilities that can be used to obtain a sparse estimate of coefficients. We compute SparseVB using the R package svb.

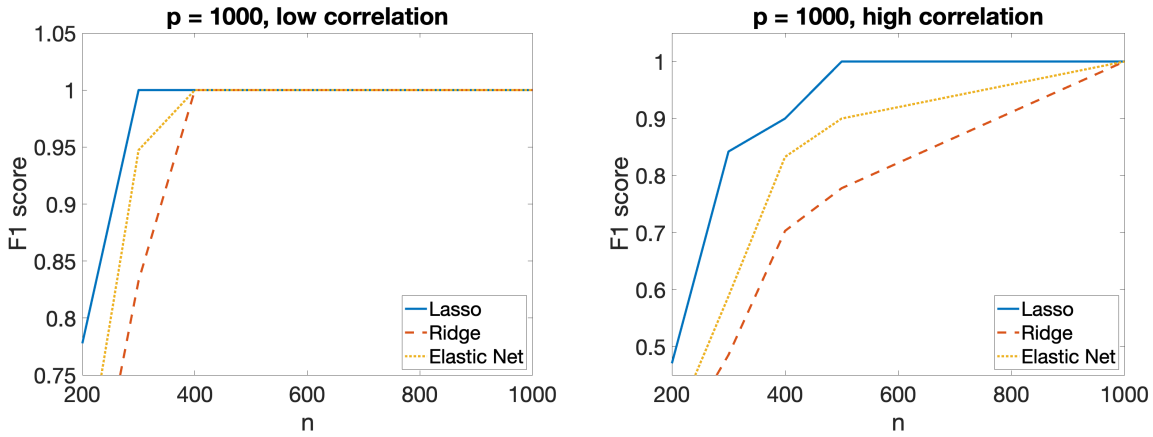

Figure 6: Comparison of F1 scores for different initial estimators for logistic models.

D.7 Skinny Gibbs

This method (Narisetty et al., 2018) accelerates classic spike-and-slab MCMC by updating only a shrinking or "skinny" set of promising variables at each sweep; its per iteration cost is $O(np)$, but it retains strong consistency of model selection even when $p \gg n$. We compute S-Gibbs using the R package skinnybasad.

E NUMERICAL EXPERIMENTS: LOGISTIC REGRESSION RESULTS

E.1 Effect of the Initial Estimator $\widetilde{\theta}$

We compare lasso, ridge and elasticNet. We set p=1000 and vary $n \in \{200, \ldots, 1000\}$. The results are given on Figure 6 and show that lasso initialization is best at small n, and all initializers become similar as n grows.

We compare three different initializations: lasso, ridge, and elasticNet, utilizing R package glmnet. We focus on the case where $p = 1000$ and increase the sample size n from 200 to 1000, in both the low and high correlation settings. In each simulation set up, we again run Algorithm 1 50 times and report the median F1 scores of the 50 chains after mixing.

We observe in Figure 6 that the F1 score is an increasing function of the sample size, and in all scenarios, the lasso initialization produces the highest F1 scores. We also observe that, given enough sample size, all three initializations perform well, which is consistent with our theoretical findings.

E.2 Comparison in Terms of Accuracy and Running Time

The results are presented in the main body Tables 1–4, which shows when $\varrho = 0$, OLAP, Exact, and SparseVB achieve comparable F1 scores, outperforming most methods, while ℓ_0 and its variants perform slightly better at $n = 200$. However, OLAP maintains consistently low runtime, unlike Exact, ℓ_0-based methods, and SparseVB, whose costs grow with n. Under high correlation, OLAP matches Exact in accuracy but retains stable runtime, whereas the mixing time of Exact increases sharply with sample size. In general, these findings highlight OLAP's advantage in achieving competitive accuracy while substantially reducing computational cost, making it superior in terms of the accuracy-cost trade-off.

F NUMERICAL EXPERIMENTS: POISSON REGRESSION RESULTS

F.1 Effect of the Initial Estimator $\widetilde{\theta}$

The results is reported on Figure 7. We observe that the Poisson regression requires comparatively larger sample size for comparable F1 scores to the logistic regression. The effect of initial estimators is similar to that in the logistic regression setting, with lasso outperforming the others.

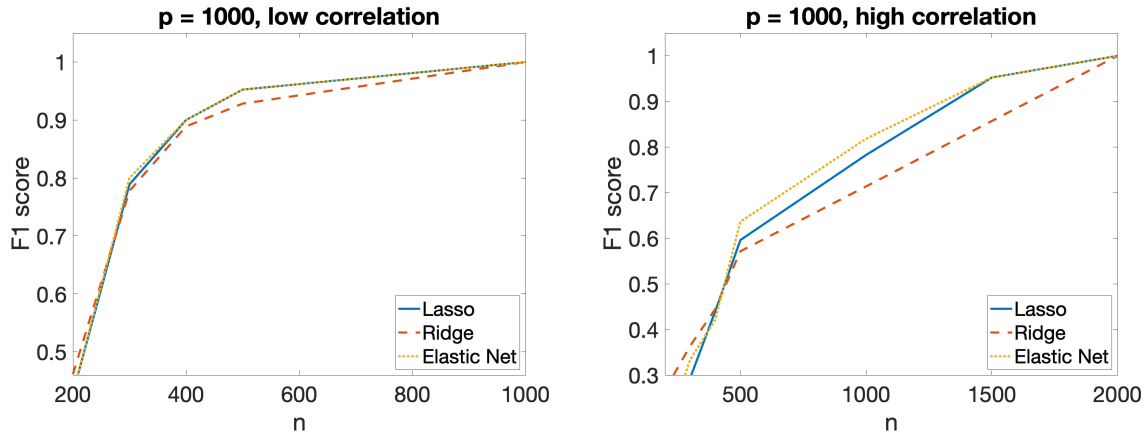

Figure 7: F1 vs. initial estimator for Poisson models, low (left) and high (right) correlation.

	n=200		n=300		n=400		n=500		n=1000	
Method	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**
OLAP	0.429	0.135	0.789	0.133	0.900	0.107	0.952	0.057	**1.000**	0.007
Exact	0.533	0.170	0.833	0.131	0.947	0.197	**1.000**	0.144	**1.000**	0.016
ALA BB	0.000	0.000	0.333	0.044	0.462	0.075	0.667	0.127	**1.000**	0.092
ALA Complex	0.000	0.104	0.000	0.177	0.000	0.151	0.182	0.144	0.824	0.035
Lasso	0.462	0.241	0.606	0.114	0.597	0.136	0.714	0.138	0.817	0.145
SCAD	0.321	0.090	0.297	0.061	0.292	0.060	0.317	0.062	0.500	0.153
MCP	0.509	0.129	0.563	0.084	0.563	0.095	0.606	0.126	0.833	0.144
mMRM-5	0.533	0.136	0.667	0.075	0.667	0.026	0.667	0.000	0.667	0.000
mMRM-10	0.500	0.128	0.700	0.125	0.800	0.095	0.900	0.090	**1.000**	0.044
mMRM-20	0.400	0.096	0.533	0.076	0.600	0.062	0.667	0.035	0.667	0.009
ABESS	**0.625**	0.149	**0.889**	0.115	**0.952**	0.065	**1.000**	0.032	**1.000**	0.027

Table 5: F1 scores for Poisson feature selection. $p = 1000$, low correlation.

F.2 Comparison in Terms of Accuracy and Running Time

We again consider different sample sizes n with low and high correlations. We show the F1 results in Table 5 and 6, and the runtime results in Table 7 and 8. We observe that the Poisson problems are more difficult than logistic problems.

In the low-correlation setting, we see that, generally, OLAP only underperformed Exact and ABESS, but Exact takes much longer to mix, and the cost of ABESS also increases much faster than OLAP along with n. The observation is similar under the high-correlation setting, again reaffirming our claim that OLAP has an advantage in terms of the trade-off between accuracy and costs.

G A BREAST CANCER DATA EXAMPLE

We illustrate the method with a high-profile breast cancer data example taken from (van de Vijver et al., 2002). Accurate prediction of distant metastasis development in breast cancer can help guide treatments and ultimately save life while preserving quality of life. In (van de Vijver et al., 2002), using gene expression data, the authors developed a prognostic score based on 70 identified genes to predict the advent of breast cancer cell metastasis in distant organs within 10 years of diagnostics. In this illustration, we re-analyzed the data collected by the same research group in (van de Vijver et al., 2002) and we compare their 70-genes profile rule with a logistic regression model for predicting the advent of distant metastasis within 5 years.

The data contains 295 women, all younger than 53 years old, with stage I or II breast cancer. Gene intensity measurements of 24496 genes along with 13 clinical variables were collected from the patients. Of the 295 patients, 151 had lymph-node-negative disease, and 144 had lymph-node-positive disease. 10 of the lymph-node-negative, and 120 of the lymph-node-positive had received more aggressive therapy, including chemotherapy, or hormonal therapy, or both. Clearly, the advent of distant metastasis depends on the stage of the cancer when detected, and on initial treatment received. However, despite the cases of lymph-node diseases, most of the cancer cases appear

Method	n=200	n=300	n=400	n=500	n=1000
OLAP	0.43	0.37	0.37	0.36	0.88
Exact	0.45	4.03	6.6	8.46	14.35
ALA BB	2.06	2.94	3.89	4.69	4.85
ALA Complex	0.88	0.97	1.09	1.14	1.68
Lasso	0.41	0.63	0.89	1.21	5.45
SCAD	4.16	9.81	17.07	24.8	39.11
MCP	2.7	8.04	16.41	23.85	73.06
mMRM-5	0.01	0.01	0.01	0.01	0.01
mMRM-10	0.01	0.01	0.01	0.01	0.02
mMRM-20	0.02	0.02	0.02	0.03	0.04
ABESS	0.05	0.12	0.23	0.41	2.51

Table 6: Running time (seconds) for Poisson feature selection. $p = 1000$, low correlation.

	n=200		n=300		n=400		n=500		n=1000		n=1500		n=2000	
Method	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**	**Median**	**Std. Error**
OLAP	0.222	0.132	0.293	0.16	0.44	0.168	0.596	0.189	0.783	0.089	0.952	0.062	**1**	0.057
Exact	0.286	0.164	**0.556**	0.189	0.594	0.198	**0.778**	0.13	**0.894**	0.115	**1**	0.076	**1**	0.07
ALA BB	0.000	0.000	0.182	0.065	0.182	0.077	0.462	0.102	0.743	0.118	**1**	0.155	**1**	0.155
ALA Complex	0.000	0.083	0.000	0.139	0.000	0.124	0.000	0.152	0.571	0.099	0.778	0.076	0.800	0.052
Lasso	0.229	0.064	0.265	0.07	0.281	0.076	0.314	0.057	0.31	0.052	0.286	0.046	0.299	0.066
SCAD	0.268	0.076	0.312	0.085	0.277	0.12	0.415	0.091	0.508	0.169	0.769	0.181	0.851	0.153
MCP	0.325	0.12	0.464	0.121	0.4	0.156	0.5	0.156	0.64	0.166	0.857	0.151	0.909	0.162
mMRM-5	0.267	0.16	0.4	0.14	0.267	0.185	0.533	0.141	0.533	0.123	0.667	0.058	0.667	0.037
mMRM-10	0.3	0.162	0.4	0.166	0.4	0.21	0.6	0.178	0.8	0.131	0.8	0.122	0.9	0.116
mMRM-20	0.267	0.111	0.333	0.124	0.333	0.157	0.433	0.11	0.533	0.088	0.6	0.083	0.6	0.078
ABESS	**0.375**	0.153	0.556	0.187	**0.6**	0.182	0.7	0.154	0.857	0.087	**1**	0.08	**1**	0.086

Table 7: F1 scores for Poisson feature selection. $p = 1000$, high correlation.

to be at similar stages, and following (van de Vijver et al., 2002), we will not account for these interactions in the analysis. As response variable y, we consider the advent of distant metastasis within five years, as a binary response $\{0, 1\}$.

To reduce the size of the covariates, we adopted the pre-processing method implemented by (Guo, 2018), with one-at-the-time initial logistic regressions that keeps only genes with p-value less than 0.01. To this initial set of genes, we then add the 70 genes identified by (van de Vijver et al., 2002), if they are not already selected by the individual T-tests. This pruning process generates a dataset with 295 patients and 1083 genes.

We compare the performance of SparseVB, OLAP, and 70-genes, the predictive model of (van de Vijver et al., 2002) based on the 70 genes they have identified. For the comparison, we use a 50-fold cross-validation procedure. In each cross-validation replication, the test sample size is 30, and the remaining 265 samples are used for training. To avoid distortion, when sampling the test set, we require that the number of 1 and 0 be approximately equal. As a performance metric, we compute the F1 score in correctly predicting the outcome variable on the test set. Table 9 shows the mean, median and standard deviation of the F1 score from the 50 cross-validation replications. The results clearly show a better performance of OLAP both in terms of accuracy and stability.

H A MOUSE PCR DATA EXAMPLE

As another illustration, we analyze the mouse PCR data also previously analyzed in (Lan et al., 2006; Narisetty et al., 2018). The dataset comprises expression levels of 22575 genes obtained from 31 female and 29 male mice, totaling 60 arrays. In addition, the physiological phenotype glycerol-3-phosphate acyltransferase (GPAT) was measured using quantitative real-time PCR. These gene expression and phenotypic data are publicly accessible in the GEO database (http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330).

We want to predict whether a mouse has low GPAT levels given its genetic expression. The level of GPAT in the body is important, as reduced levels of GPAT have been associated with a decrease in liver steatosis, a disease commonly associated with obesity.

Similarly to (Narisetty et al., 2018), we derive the binary response variable based on GPAT levels, defined as $y = I(GPAT < Q(0.4))$, where $Q(0.4)$ represents the quantile of 40% of GPAT. And, given the extensive number of genes, we also did a similar gene pruning as in the previous example, by conducting marginal simple logistic regression of the response y against individual genes. But unlike (Narisetty et al., 2018), and in order to make our experiment more challenging, we select 500 genes that have the most marginally significant p-values, instead of 99 in the original work. Together with the gender variable, this results in $p = 501$ covariates.

Method	n=200	n=300	n=400	n=500	n=1000	n=1500	n=2000
OLAP	2.27	4.41	6.00	5.39	4.34	10.02	25.51
Exact	4.39	9.62	16.28	11.27	12.56	13.69	46.83
ALA BB	3.22	5.15	6.84	9.61	7.07	5.74	5.31
ALA Complex	0.91	1.03	1.14	1.33	2.37	2.22	2.07
Lasso	0.68	1.10	1.66	2.16	7.67	57.80	33.33
SCAD	4.44	6.60	9.42	12.51	15.02	150.73	214.133
MCP	3.94	7.09	10.91	15.62	22.30	107.90	150.493
mMRM-5	0.01	0.01	0.01	0.01	0.01	0.02	0.026
mMRM-10	0.01	0.01	0.01	0.01	0.03	0.04	0.0475
mMRM-20	0.02	0.02	0.03	0.03	0.05	0.07	0.0885
ABESS	0.06	0.15	0.29	0.50	3.42	12.82	33.0305

Table 8: Running time (seconds) for Poisson feature selection. $p = 1000$, high correlation.

	Sparse VB	70-identifiers	OLAP	Exact
Mean	0.362	0.593	0.688	0.647
Median	0.509	0.609	0.688	0.655
Std. Error	0.294	0.109	0.076	0.1013

Table 9: F1 for the breast cancer data.

	Sparse VB	Skinny Gibbs	OLAP	Exact
Mean	0.4825	0.5751	0.3172	0.2926
Median	0.5000	0.5236	0.3080	0.2754
Std. Error	0.0903	0.1279	0.0561	0.1227

Table 10: RMSE for the mouse PCR example.

	Sparse VB	Skinny Gibbs	OLAP	Exact
Mean	0.4724	0.3786	0.8430	0.8542
Median	0.5357	0.3636	0.8889	0.8889
Std. Error	0.3223	0.1993	0.1806	0.1487

Table 11: F1 scores for the mouse PCR example.

For comparison, we apply SparseVB, S-Gibbs and OLAP. Similarly to the breast cancer data experiment, we randomly select 30 different pairs of training and test data sets. To have results comparable to (Narisetty et al., 2018), we report in Table 10 the square root of the mean squared error (RMSE) as a performance measure, which is $RMSE = \sqrt{\frac{1}{n}\sum_i^n (y_i - \hat{\pi}_i)^2}$, where $\hat{\pi}_i$ is the probability predicted by the logistic model. We also report in Table 11 the F1 score in correctly predicting the outcome variable on the test set. The results here again show that OLAP works better than S-Gibbs and SparseVB, in terms of both the performance and stability in these measures.