THEORETICAL GUARANTEES FOR BAYESIAN GENERALIZED LINEAR REGRESSION AND VARIATIONAL BOOSTING

A Thesis

by

BIRAJ SUBHRA GUHA

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Debdeep Pati
Co-Chairs of Committee,	Anirban Bhattacharya
Committee Members,	Daren Cline
	Raymond Carroll
	Krishna Narayanan
Head of Department,	Brani Vidakovic

August 2021

Major Subject: Statistics

ABSTRACT

This dissertation is the culmination of my research work at Texas A & M Department of Statistics under the supervision and guidance of Dr. Debdeep Pati and Dr. Anirban Bhattacharya. It consists of four chapters, the first of which contains a broad overview of my research topics, detailed literature review and discussions on my motivation to tackle certain unanswered questions in today's Bayesian world. The second chapter presents my project on Variational Boosting, a widely used computation tool for Variational modeling procedures, where I have investigated statistical properties of a variational algorithm. The arXived version is cited: [Guha et al., 2020]. The third chapter deals with posterior convergence and model selection issues in a newly proposed class of Generalized Linear Models, called cGLM, using the popular spike-and-slab prior. This work is currently under re-submission process in *Bayesian Analysis*; the arXived version is cited: Guha and Pati [2021]. The fourth chapter contains a summary of the previous chapters and also a brief discussion on my future research direction. The proofs of theorems, well-known definitions and auxiliary results are deferred to the appendix, Appendix A for second chapter and Appendix B for third chapter.

DEDICATION

To Ma, Sonababa,

My lovely wife Aumma,

My childhood math mentor Subir Guha.

ACKNOWLEDGMENTS

I would like to sincerely thank Dr. Pati, Dr. Bhattacharya and Dr. Cline for their guidance. Not every project I attempted resulted in a manuscript, but I sure learnt a lot from every such endeavor and my mentors' diligent help played a cardinal role in it.

I would specifically like to thank my PhD committee chair, Dr. Debdeep Pati for his consistent engagement in my research progress as well as academic and mental support. I was honored and overjoyed when he agreed to support my summer visit to University of Ottawa to work with Dr. Aaron Smith. Thought processes inculcated in me by Dr. Pati are some of the strongest tools I have in my academic arsenal.

I thank my lovely wife Aumma for her relentless support during these final years of my PhD. I thank my parents for their blessings and encouragement.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Dr. Debdeep Pati (Chair), Dr. Anirban Bhattacharya (Co-Chair), Dr. Daren Cline, Dr. Raymond Carroll and Dr. Krishna Narayanan.

Funding Sources

This work was aided by start-up funds from PhD committee chair Dr. Debdeep Pati.

LIST OF SYMBOLS

$\rightsquigarrow$	Weak convergence / convergence in distribution
$\mapsto$	Mapping / function
$\lesssim, \gtrsim$	Inequalities up to absolute constants
$\\|\cdot\\|_2$	Euclidean norm of vectors
$\\|\cdot\\|_1$	l_1 norm of vectors
$s_{smax}(\cdot)$	Highest singular value of matrices
$KL(a\\|b)$	Kullback–Leibler Divergence of density b from density a
$\chi^2(a\\|b)$	Chi–square Divergence of density b from density a
∇f	Vector derivative of scalr field f
$\partial f(x)$	A sub-gradient of convex function f at point x
$\mathcal{D}_f(y\\|x)$	Bregman divergence of point x from point y under convex function f
$\mathcal{C}_{f,D}$	Curvature of convex function f on domain D
$\zeta_{1,n} = o(\zeta_{2,n})$	$\zeta_{1,n}/\zeta_{2,n} \to 0$ as $n \to \infty$
$\zeta_{1,n} = O(\zeta_{2,n})$	$0 < C_2 \leq \liminf\limits_{n\to\infty} (\zeta_{1,n}/\zeta_{2,n}) \leq \limsup\limits_{n\to\infty} (\zeta_{1,n}/\zeta_{2,n}) \leq C_3$ for absolute constants C_2, C_3
$f \circ g$	Function composition
$\mathrm{supp}\,(\cdot)$	Set of non-zero entries in a Euclidean vector
$\\|\cdot\\|_{(\infty,\infty)}$	Entry-wise absolute maximum of matrices
$a \bigvee b$	maximum of reals a, b
$\mathbf{1}_{\{\cdot\}}$	Set indicator function

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iii

ACKNOWLEDGMENTS iv

CONTRIBUTORS AND FUNDING SOURCES v

NOMENCLATURE vi

TABLE OF CONTENTS vii

LIST OF FIGURES ix

LIST OF TABLES x

1. INTRODUCTION AND OVERVIEW 1

1.1 The issue of computing the posterior 2
1.2 The issue of posterior inference in high-dimensional models 6

2. VARIATIONAL BOOSTING WITH GAUSSIAN MIXTURES 12

2.1 Introduction 12
2.2 Background and target 14
2.3 Statistical properties of the variational optimizer 16
2.3.1 The small bandwidth mixture Gaussian family 16
2.3.2 Comparison of divergences and Bernstein-von Mises Phenomenon 17
2.3.3 Stochastic Boundedness of the Kullback–Leibler Discrepancy 18
2.3.4 Sketch of Proof of Theorem 1 21
2.4 Convergence analysis of the algorithm 22
2.4.1 Steps of the Algorithm 22
2.4.2 Rate of Convergence of Algorithm 1 24
2.4.3 Sketch of Proof of Theorem 2 25
2.5 Concluding remarks 27

3. ADAPTIVE POSTERIOR CONVERGENCE IN SPARSE HIGH DIMENSIONAL CLIPPED GENERALIZED LINEAR MODELS 28

3.1 Introduction 28

3.1.1 Notations ... 31
3.2 Construction of GLM family ... 31
3.2.1 Introduction to clipped GLM ... 32
3.2.2 Connection of cGLM to regression settings ... 36
3.2.3 Numerically understanding clipped Poisson regression ... 37
3.3 Construction of sparsity favoring prior ... 38
3.3.1 Prior on model size and the non-zero coefficients ... 39
3.4 Assumptions on data generating distribution and prior ... 40
3.4.1 Order assumptions on sample size and parameter dimension: ... 40
3.4.2 Identifiability and compatibility assumptions ... 40
3.4.3 Hyper-parameter selection aimed at truth adaptive posterior contraction ... 44
3.5 Adaptive rate-optimal posterior contraction in ℓ_1 norm ... 45
3.5.1 Lower bound of the marginal likelihood ... 46
3.5.2 Posterior dimension and weak model selection ... 47
3.5.3 Truth adaptive posterior contraction in ℓ_1 metric ... 49
3.6 Conclusion ... 49

4. SUMMARY AND FUTURE DIRECTIONS ... 51

4.1 Summary ... 51
4.2 Current and future work ... 51

REFERENCES ... 53

APPENDIX A. SUPPLEMENTARY MATERIAL TO FIRST CHAPTER ... 61

A.1 Revisiting Frank–Wolfe Algorithm ... 61
A.2 Proofs ... 63
A.3 Auxiliary Results ... 69

APPENDIX B. SUPPLEMENTARY MATERIAL TO SECOND CHAPTER ... 73

B.1 Proof of Theorem 1 ... 73
B.2 Proof of Theorem 2 ... 74
B.3 Proof of Theorem 3 ... 78
B.4 Auxiliary results ... 79

LIST OF FIGURES

FIGURE Page

3.1 Graph of $y = 10 - \log(1 + \exp(-x + 10))$ 34

3.2 Average MSE of $\hat{\beta}_{MLE}$ using Clipped Poisson Regression 38

LIST OF TABLES

TABLE Page

3.1 Average MSE of $\hat{\beta}_{MLE}$ over 100 iterations 37

1. INTRODUCTION AND OVERVIEW

Increasing complexity and volume of data in today's world necessitate use of computationally tractable methods for analysis, while their validity needs to be tested through results on statistical guarantees. My research, which can be aptly described as exploratory and innovative, lies at the bridge over this juncture. My studies and works revolve around two broad topics:

- Bayesian methodology and inference
- Machine Learning Algorithms

The driving force behind my research interest is the growing intricacy of real world data, rigorous modeling and analysis of which is essential for data science and statistics to be beneficial for society. Thus my research focuses primarily on novel methodologies that can boast strong theoretical foundations as well as relative computational ease. In what follows, I shall give a brief overview of the Bayesian way of inferring from statistical models, specifically those under the purview of a 'high-dimensional setup', discuss some key ideas that I have borrowed from existing literature on Bayesian statistics and Machine Learning, detail some relevant questions that are either partially answered or unanswered, and finally present how I have endeavored to contribute to these rich and diverse fields.

Bayesian modeling and inference has seen a huge surge of interest in the last few decades and has seen prominent statisticians contribute heavily to diversify this field from perspectives of both novel modeling ideas and cementing their theoretical foundations. Hoff [2009] presents an excellent array of discussions on core Bayesian tools and gave my know-how a huge boost when I started out with my research. Suppose we observe data points $X_1, \dots X_n$ as part of a statistical experiment and wish to employ the model $f(X_{1,\dots n}|\theta)$, where the model-defining parameter $\boldsymbol{\theta}$ is assumed to belong to $\mathbb{R}^d$, often with geometric restrictions relevant to the experiment at hand. Bayesians enforce a 'prior' distribution $\pi(\theta)$ on this θ that reflect the aforementioned geometry and then utilize the celebrated Bayes' formula:

$$\pi(\theta|X_{1,\dots n}) \propto f(X_{1,\dots n}|\theta)\pi(\theta) \tag{1.0.1}$$

to obtain the 'posterior' distribution of $\boldsymbol{\theta}$ by multiplying the likelihood with the prior. My first project deals with the fundamental issue of actually 'computing' this posterior, while my second project aims to contribute to the paradigm of accurately inferring about θ from $\pi(\theta|X_{1,\dots n})$.

1.1 The issue of computing the posterior

The constant of proportionality in (1.0.1) is of fundamental importance when it comes to Bayesian computations. Termed the marginal likelihood, it can only be analytically tractable under assumptions of conjugacy; one can refer to Schlaifer and Raiffa [1961] for the original ideas. For non-conjugate pairs of likelihood and prior one needs to approximate the posterior. The most prevalent idea is to create a Markov Chain of distributions with the transition rule shunning the use of the marginal likelihood, which converges to the true posterior in total variation distance. Markov Chain Monte Carlo (MCMC) methods yields samples from the approximate posterior, whose theoretical accuracy has been studied in seminal works like Roberts et al. [2004] and Neal [1993], while algorithmic extensions have been presented in works like Neal et al. [2011], Damlen et al. [1999], etc. Experts have long identified that MCMC chains are computationally intensive, especially when the chain does not 'mix' well, that is intermediate samples obtained in the process tend to stay dependent; one can look up Brooks and Roberts [1998] for a comprehensive overview of MCMC convergence diagnostics. This drawback of MCMC methods has driven attention towards a wide array of approximate Bayesian computation (ABC) methods, which are often faster in implementation. It is being widely used for biological sciences like population genetics; see Beaumont et al. [2002]. One of the most common one among these approaches is undoubtedly the Laplace approximation of the marginal likelihood (refer to Schervish [1995]) using second order Taylor approximation of the likelihood, which then mimics a Gaussian posterior. Another very common tool is the ABC rejection algorithm, one of the earliest discussions on which can be found in Stigler [2010]. Specific focus on uncertainty of mean and variance estimates in the finite sample setup can be found in works like Huggins et al. [2018]. Variational Bayes is an increasingly popular alternative to Monte Carlo procedures in sampling from intractable posteriors, where a flexible and computationally tenable variational family of distributions is chosen to begin with, from whom we choose a member having least discrepancy with the actual posterior. Blei et al. [2017] serves as an excellent overview of this, and I shall briefly discuss its details in what follows.

Let us start with $\mathbb{R}^p$-valued data points $X_1 \ldots, X_n$ which are independently and identically distributed according to density $f(x;\theta)$, where $\theta \in \mathbb{R}^d$, the parameter space. Given a prior density $\pi(\theta)$ on parameter θ, we denote the posterior of θ as

$$\pi_n(\theta) = \frac{\prod_{i=1}^n f(X_i;\theta)\pi(\theta)}{\int \prod_{i=1}^n f(X_i;\theta)\pi(\theta)d\theta}.$$

Variational Bayes, which has its roots in variational calculus, works with a flexible and rich family $\mathcal{Q}$ called the variational family consisting of densities over the parameter space, within which we search for an approximator to the posterior. The principal aim is to find $q_n^*(\mathcal{Q}) \in \mathcal{Q}$ such that

$$q_n^*(\mathcal{Q}) = \operatorname{argmin}\{q \in \mathcal{Q} \mid KL(q||\pi_n)\}, \quad m_n^*(\mathcal{Q}) = KL(q_n^*(\mathcal{Q})||\pi_n), \tag{1.1.1}$$

where, $KL(a||b) = \int a(\theta)\log(a(\theta)/b(\theta))d\theta$ denotes the Kullback–Leibler divergence of density b from density a, both defined on the parameter space. KL discrepancy is the most widely used measure to quantify the approximation gap, but other measures have been employed, like Hellinger distance in Campbell and Li [2019]. Mean-field variational family, where the approximating posterior factorizes over the co-ordinates (or blocks of them), is computationally fast and widely popular, and have seen a plethora of research through works like Yang et al. [2017], Pati et al. [2018], Alquier and Ridgway [2017], Zhang and Gao [2017], Wang and Blei [2018], Mukherjee et al. [2018], Yang et al. [2018] and Huggins et al. [2020]. The basic structure of such families looks like:

$$\left\{ q(\boldsymbol{\theta};\boldsymbol{\nu}) = \prod_{j=1}^{p} q_j\left(\theta_j;\nu_j\right) : \theta_j \in \Theta, \nu_j \in \mathcal{N} \right\}$$

for parameter space Θ and ν_j's denoting the variational parameters that define the variational distributions. This approach is specifically aimed at recovering the posterior mean accurately, and hence is not suitable when the demand includes posterior covariance recovery. Also, mean-field approximation is tantamount to a Gaussian approximation of the posterior, which does not allow us to deal with deal with posteriors which are not Gaussian.

Let us illustrate with an example: let $\phi(\cdot;\mu_i),\ i=1,2$ denote unit variance standard Gaussian pdfs on $\mathbb{R}$ with corresponding means as parameters. Consider $\theta_1,\theta_2\in\mathbb{R}$, and let

$$X_{1,\ldots n}\overset{i.i.d}{\sim}\frac{1}{2}\phi(\cdot;\mu_1)+\frac{1}{2}\phi(\cdot;\mu_2),\quad \mu_1\sim\phi(\cdot,\theta_1),\ \mu_2\sim\phi(\cdot,\theta_2),\ \mu_1\perp\mu_2 \tag{1.1.2}$$

Model (1.1.2) is a simple mixture model with known weights and independent priors on the two means. By (1.0.1), the posterior $\pi(\mu_1,\mu_2|X_{1,\ldots n})$ now satisfies

$$\pi(\mu_1,\mu_2|X_{1,\ldots n})\propto 2^{-n}\left[\prod_{i=1}^{n}\left(\phi(X_i;\mu_1)+\phi(X_i;\mu_2)\right)\right]\phi(\mu_1,\theta_1)\phi(\mu_2,\theta_2)$$

Irrespective of sample size n, the posterior stays a true mixture of 2^n densities on $\mathbb{R}^2$, so approximating it with a uni-modal density like Gaussian might very well be inferior compared to other approximations, say using a mixture Gaussian variational approximator. Interesting solutions to the problem include modeling the co-dependency through copulas as in Tran et al. [2015] and Han et al. [2016], and employing implicit distribution families as in Huszár [2017], Han et al. [2016], Titsias and Ruiz [2018], Yin and Zhou [2018], Molchanov et al. [2018] and Shi et al. [2017]. We focus on approximating through mixture Gaussians as hinted above, drawing motivation from past works like Wang et al. [2006]. As proposed in Guo et al. [2016] and Miller et al. [2017], the structure of mixture families naturally gives rise to the idea of variational boosting.

Variational boosting is a computation method that iteratively builds a mixture distribution approximation to the posterior by adding simple, new components and re-weighting them. The components of the mixture can thus be considered weak learners in this boosting framework, which are averaged in a weighted, sequential fashion to produce a mixture, the strong learner. This idea is in line with the widely popular ensemble method of boosting used in the machine learning paradigm; one can refer to Zhou [2012] for a comprehensive study. The variational family $\mathcal{Q}$ in this setup is chosen as:

$$\mathcal{Q} = \operatorname{conv}(\Gamma) = \left\{ \sum_{k=1}^{K} \beta_k \phi_k : \phi_k \in \Gamma,\ \boldsymbol{\beta} \in \Delta^K,\ K \geq 1 \right\}, \tag{1.1.3}$$

where Γ is any family of simple, component distributions on $\mathbb{R}^d$ and Δ^K denotes the unit simplex in $\mathbb{R}^K$. To describe the boosting approach, let $\psi_n^{(k)}$ denote the k-th iterate in an algorithm that aims to solve the optimization problem in (1.1.1). Given $\psi_n^{(k)}$, the next iterate $\psi_n^{(k+1)}$ is obtained by

$$\psi_n^{(k+1)} = (1 - \gamma_K)\psi_n^{(k)} + \gamma_k \phi_n^{(k+1)} \tag{1.1.4}$$

where the weight $\gamma_k \in [0, 1]$ and $\phi_n^{(k+1)} \in \Gamma$ depend on the boosting approach employed. The Frank–Wolfe algorithm described in Frank and Wolfe [1956] provides a neat pathway to handle such iterates by choosing the components $\phi_n^{(k+1)}$ in (1.1.4) through a routine called Linear Minimization Oracle (LMO). Jaggi [2013] discusses convergence properties of the Frank–Wolfe algorithm, which has been heavily utilized in my work. Interesting questions under this setup includes the rate of convergence of this algorithm in terms of the number of samples and the statistical properties of the iterates themselves, since they are functions of the posterior. Locatello et al. [2017] serves as the principal motivating work for me in this regard, where necessary assumptions for Frank–Wolfe are enforced through restricting the component distributions in (1.1.4) to compactly supported densities. This results in theoretical issues concerning widely used priors that do not have compact supports, like Gaussians or mixture Gaussians.

My humble contribution to the field of variational inference and variational boosting involves proposing a mixture Gaussian variational family with restricted bandwidths and means varying in a compact set. Consider the following for a fixed c_0 with $1 < c_0 < 2$, and some $M, \sigma_n > 0$:

$$\Gamma_n = \left\{ N\left(\mu, \sigma^2 I_d\right) : \|\mu\|_2 \leq M,\ 0 < \sigma_n \leq \sigma \leq c_0^{1/2} \sigma_n \right\}. \tag{1.1.5}$$

We now define the small-bandwidth mixture Gaussian family as follows:

$$\mathcal{Q}_n = \left\{ \sum_{k=1}^{K} \beta_k \phi_k : \phi_k \in \Gamma_n,\ \boldsymbol{\beta} \in \Delta^K,\ K \geq 1 \right\}.$$

Conditions in (1.1.5) alleviate the issue of support restrictive variational families while simultaneously allowing us to calculate the Frank–Wolfe rate of convergence in terms of the bandwidth. It also allows us to propose stochastic boundedness results of the Kullback–Leibler discrepancy in (1.1.1) and the intermediate iterates described in (1.1.4). My work thus helps to provide a theoretical backing for the widely popular variational boosting computations for general, not-necessarily-Gaussian-like posteriors, which, to the best of my knowledge, is currently lacking in the existing literature on variational inference.

1.2 The issue of posterior inference in high-dimensional models

Bayesian inference banks on properties on the posterior distribution and the estimates of the underlying parameter obtained from it come equipped with automatic uncertainty estimates. This posterior can be considered as a stochastic distribution over the parameter space under the assumption of data being generated from some 'true' distribution. This approach of evaluating the posterior's utility in a Frequentist way is often referred to as Frequentist validation of Bayesian approaches. Assuming $\theta^* \in \mathbb{R}^p$ to be the true value of the parameter for model $f(X_{1,\dots n}|\theta)$, a key idea is to consider the neighborhood

$$\{\theta \in \Theta : d(\theta, \theta^*) \leq \epsilon\}. \tag{1.2.1}$$

If the Bayesian approach needs to be at par with Frequentist estimation ideas, the truth neighborhood in (1.2.1), at least for fixed $\epsilon > 0$ should have very high posterior probability the more samples we observe. The metric employed in (1.2.1) depends on the type of 'consistency' result we wish to achieve using the posterior, and the prior probability of the neighborhood in (1.2.1) plays a crucial role in determining whether such consistency results actually hold. Detailed theoretical discussions underpinning the idea of posterior consistency can be found in Diaconis and Freedman [1986], Choi et al. [2008], etc. Works like Ghosal et al. [1999], Tokdar [2006] and Barron [1988] deal with this issue in the genre of density estimation.

The neighborhood in (1.2.1) can be modified to replace ϵ with ϵ_n in order to capture the sample-size dependent rate at which the posterior concentrates around the truth. Studies surrounding these are often termed as 'posterior contraction' in the Bayesian literature, and aim to find the smallest allowable rate ϵ_n in (1.2.1) so that

$$\pi_n \left(\{\theta \in \Theta : d(\theta, \theta^*) \leq \epsilon_n\} \middle| X_{1,\dots n} \right) \overset{n \to \infty}{\to} 1 \tag{1.2.2}$$

Seminal works Ghosal et al. [1995], Shen et al. [2001], Ghosal et al. [2000] and Ghosal and van der Vaart [2007] provide a strong foundation to deal with posterior contraction in a very general, not necessarily parametric, setup. The method of handling the marginal likelihood introduced in these works carries over to many likelihood-prior setups, and I have strongly leveraged them in my work. Regression and density/function estimation setups have seen a number of research papers utilizing these ideas, like Yang et al. [2015], van der Vaart et al. [2008], Agapiou et al. [2021], Hu [2010], Bhattacharya et al. [2014], Shen et al. [2013], etc.

It is interesting to study high-dimensional parametric regression models using the tools of posterior consistency and contraction, and we first discuss briefly some relevant model and prior setups. A simple linear model:

$$Y = X\beta + \epsilon \tag{1.2.3}$$

moves onto the paradigm of high-dimension if we see less samples than the number of parameters we want to estimate, i.e. $Y : n \times 1, X : n \times p, \beta : p \times 1$ and $n > p$. It is well-known that the model is not identifiable as-is, so both Frequentists and Bayesians proceed with structural imposition on β's. We can assume that the true β^* is 'sparse' in the sense that less than n co-ordinates of it are actually non-zero. This is a standard Frequentist idea that have seen numerous variations in literature , and can be carried over to Bayesian high-dimensional regression for Frequentist validation type results. The book Hastie et al. [2019] provides a comprehensive study of this topic. One of the main ideas to incorporate sparsity in the parameters goes through the concept penalized regression, which includes ridge regression (see Hill [1975], Gruber [2017], etc), least absolute shrinkage and selection operator or LASSO (see Tibshirani [1996], Tibshirani [1997]), elastic net (see Zou and Hastie [2005]), entropy based penalization (Donoho et al. [1992]) etc. This method proceeds by adding a penalty term to the least squares loss function appearing naturally in regression setups, which help dictate the structure of the β estimates as well as take care of the identifiability issue. The Bayesian analog of this regression setup follows by imposing a suitable prior on the parameters that dictate the geometry of the parameter space; Gaussian prior for ridge, Laplace prior for LASSO, etc. It is crucial to note that above methods provide a pathway not only for estimation, but variable selection, i.e. the estimator should try to reflect the 'true' subset of the β's from which the data was generated. Thus in a Bayesian approach, one should also ask the vital question whether the posterior is imposing sufficient probability on the set of non-zero β's. One can refer to the earlier works in George and McCulloch [1997], Mitchell and Beauchamp [1988] and more recent work in O'Hara et al. [2009] for a comprehensive discussion of Bayesian variable selection methods, specifically in regression setups. All these beg the question of inferential power of the estimates obtained, and how the Frequentist approaches might tie in with the analogous Bayesian ones, specifically for priors designed to facilitate sparsity.

One of the most common class of priors employed to achieve sparsity and variable selection in regressions is the spike-and-slab prior and its variations. The principal spike-and-slab idea relevant to my work is to separate out the β co-ordinates that are zero (noise co-ordinate) and non-zero (signal co-ordinate), and then weight them based on the cardinality of the non-null subset. Details on this types of prior can be found in Ishwaran et al. [2005], Malsiner-Walli and Wagner [2018], Andersen et al. [2014], etc. The structure of spike-and-slab that we utilize is induced through a prior on the duo (S, β), where S denotes a subset of $\{1, \ldots d_n\}$ and p denotes the dimension of the ambient parameter space. First, the prior on the dimension $0 \leq s \leq p$ is chosen to be $\omega_n(s) = C_n p^{-a_n s}, \ s = 0, \ldots, p$ with hyper-parameter $a_n > 0$, where C_n is chosen to normalize the distribution. For any β and S mentioned above, recall that β_S denotes the same vector β, but co-ordinates in S^c set to 0. With hyper-parameter $\lambda_n > 0$, the full prior is taken to be of the form

$$\begin{aligned}\Pi_n\left(S,\beta\right) &:= \omega_n(|S|).\binom{p}{|S|}^{-1}.\left(\frac{\lambda_n}{2}\right)^{|S|}.\ \exp(-\lambda_n\|\beta_S\|_1).\ \delta_0\left(\beta_{S^c}\right)\\ &= C_n.\binom{p}{|S|}^{-1}.\left(\frac{\lambda_n}{2p^{a_n}}\right)^{|S|}.\ \exp(-\lambda_n\|\beta_S\|_1).\ \delta_0\left(\beta_{S^c}\right),\end{aligned} \tag{1.2.4}$$

where $\|.\|_1$ denotes ℓ_1-norm of Euclidean vectors, $|S|$ denotes cardinality of the set S and δ_0 denotes the degenerate distribution. The prior on the main parameter of interest, β, is given by

$$\Pi_n(\beta) := \sum_{S\subset\{1,\ldots n\}} \Pi_n\left(S,\beta\right),$$

and the posterior probability of a general $B \subset \mathbb{R}^p$ is

$$\Pi_n(B \mid Y^{(n)}) := \frac{\int_B \exp[L_n(\eta,\eta^*)]\Pi_n(\beta)d\beta}{\int \exp[L_n(\eta,\eta^*)]\Pi_n(\beta)d\beta}.$$

The signal co-ordinates have thus been assigned Laplace priors, which is reminiscent of LASSO regression. The choice of prior on the dimension follows the idea of the so-called complexity priors, where the aim is to down-weight models (or subset of β co-ordinates) that have higher cardinality, thus enforcing sparsity.

Majority of posterior based inference for parametric regression deals with linear regression setups. Castillo and van der Vaart [2012] and Castillo et al. [2015] describe posterior convergence rates and variable selection results in detail for the sequence model and standard linear model respectively, using priors similar to (1.2.4). These works form the central core of my research on this topic, and contain theorems concerning the recovery of the true signal β^* through ℓ_1 and ℓ_2 distances at minimax-optimal rates, as well as results on the posterior selecting the true signal subset. Mini-max rate analysis of sparse linear regression presented in Van De Geer et al. [2009] is leveraged to construct the so-called ‘local invertibility’ conditions of the Gram matrix X^TX. However, the issue of these assumptions and results carrying over to setups beyond linear regression has not seen a lot of investigation in the Bayesian literature, with the possible exception of Jiang et al. [2007]. Jiang et al. [2007] operated in a high dimensional setting where the use of a Gaussian prior leads to a restrictive assumption on the growth of the true coefficients; refer to the assumptions of Theorem 1 in pg. 1493. Atchadé [2017] considered a Laplace-type prior for the coefficients which obviated the need for such a restriction, but their results are specific to logistic regression. This motivated me to delve into the widely popular Generalized Linear Models(GLM), and study the effects of spike-and-slab type priors on obtaining minimax-optimal convergence rates in these models.

I have endeavored to contribute to this topic by proposing a novel class of GLM models, which deviates slightly from the standard GLM construction. Standard GLM (see McCullagh [2018]) starts with the exponential family

$$f(y \mid \theta) = h(y) \exp\left[\theta T(y) - A(\theta)\right], \; y \in \mathcal{Y} \subset \mathbb{R},$$

where $\theta \in \Theta \subset \mathbb{R}$ is the native parameter and $A(\cdot)$ is the log-partition function. Data is assumed to come from such an exponential family member, and the method models a function of the mean through a linear function of a covariate, i.e. as $x^{\mathrm{T}}\beta$, where x represents a covariate and β is the new parameter of interest. The said function, denoted by $g(\cdot) : \mathrm{range}[A'(\cdot)] \to \mathbb{R}$, is termed as the link function. With n data points and p covariates, $X : n \times p$ makes up the design matrix corresponding to (1.2.3), whose i-th row is denoted by x_i^{T}. Thus, for every $i = 1, \ldots n$, GLM prescribes the transition θ to β as

$$g^{-1}\left(x_i^{\mathrm{T}}\beta\right) = A'(\theta), \text{ equivalently } \theta = \left(g \circ A'\right)^{-1}\left(x_i^{\mathrm{T}}\beta\right).$$

It is clear from the right hand side that GLM actually models the original parameter of the exponential family, but it does so indirectly, through the link function and $A'(\cdot)$. This motivates modeling the original parameter θ using $A''(\cdot)$, and not through $A'(\cdot)$, leading to the definition of a newly proposed clipping function $\eta(\cdot)$ and clipped GLM family. The deviation from standard GLM is in that we allow the effect of linear term $x^{\mathrm{T}}\beta$ in the argument of the log-partition function to 'clip' away from the singularities of the log-partition function. This plays a crucial role in transferring the necessary assumptions made in Castillo et al. [2015] for linear models to the non-linear geometry of GLM's and allows application of the marginal likelihood approximation presented in Ghosal [2000] to this generalized setup. I present in my work how all the standard GLM models, like logistic, Poisson and negative binomial regressions tally with this novel 'clipped' GLM setup using the following property of clipping function $\eta(\cdot)$:

Clipping function condition: There exists constant $\mathcal{M}_0(A) > 0$ depending on $A(\cdot)$, so that $\eta(\cdot)$ satisfies

$$\eta(\cdot) : \mathbb{R} \to \mathcal{I}_A\left(\frac{\mathcal{M}_0^2(A)}{2}\right), \text{ Lipschitz, injective.} \tag{1.2.5}$$

where $\mathcal{I}_A(b) := \{t \in \mathbb{R} : 0 \leq A''(t) \leq b\}$ is an interval on the real line for any $b \in (0, \infty]$. Assumptions in current literature (see dissertation by Seonghyun Jeong at NC State University, Chapter 4) require growth conditions on $\|\beta^*\|_1$ and Bi-Lipschitz bounds on the link function to simplify likelihood calculations for GLM to that of linear models. Both of these are subverted in my work, resulting in a ℓ_1 norm posterior contraction result, which is simultaneously adaptive to a large set of possible true β^*'s and minimax optimal. I also present a weak model selection result which guarantees that models strictly larger than the true non-zero subset of β^* have posterior probability tending to zero with increasing sample size.

2. VARIATIONAL BOOSTING WITH GAUSSIAN MIXTURES

2.1 Introduction

Variational Bayes has gained popularity in recent years as an alternative to Markov chain Monte Carlo procedures to approximate analytically intractable posterior distributions; refer to Blei et al. [2017] for a comprehensive overview. Variational inference formulates the problem of approximating the posterior as an optimization routine by minimizing a measure of discrepancy between probability densities in an approximating class and the posterior density. The variational solution refers to the closest member of the approximating class to the posterior, with closeness measured through divergences or metrics, usually Kullback–Leibler divergence. Other discrepancy measures for approximating the posterior have been studied, like the Wasserstein distance and Rényi divergence in Huggins et al. [2020], Fisher distance in Huggins et al. [2018] and Hellinger metric in Campbell and Li [2019].

The approximating class or the domain of optimization, commonly referred to as the variational family, plays a central role in these methods. It is chosen to strike a balance between computational tractability and approximation power. A richer, more flexible family allows better approximation of the posterior, while a simpler class of distributions facilitate calculations and computation speed. The Gaussian family is a popular example of a parametric variational family, where the optimization effectively takes place over a finite-dimensional parameter space. For a semi-parametric approach, one can use the popular mean-field family, which only assumes that the variational density factorizes over pre-specified sub-blocks of the parameter, with the factors otherwise unrestricted.

Statistical guarantees, frequentist validation as well as convergence issues focusing on mean-field appear in works like Yang et al. [2017], Pati et al. [2018], Alquier and Ridgway [2017], Zhang and Gao [2017], Wang and Blei [2018], Mukherjee et al. [2018], Yang et al. [2018] and Huggins et al. [2020]. However, mean-field approximations can only hope to recover the center of the posterior and fails to capture posterior co-dependence, so need for more general families arise. Copula modelling has been used in Tran et al. [2015] and Han et al. [2016], while implicit distribution families have been used in Huszár [2017], Han et al. [2016], Titsias and Ruiz [2018], Yin and Zhou [2018], Molchanov et al. [2018] and Shi et al. [2017]. Another recent approach to gain modelling flexibility is to use mixture distributions as variational families, which is the focus of this paper. Wang et al. [2006] is an early theoretical work on Gaussian mixtures as variational family, focusing on conjugate priors. As proposed in Guo et al. [2016] and Miller et al. [2017], the structure of mixture families naturally gives rise to the idea of variational boosting. This computation method iteratively builds a mixture distribution approximation to the posterior by adding simple, new components and re-weighting them. The components of the mixture can thus be considered weak learners in this boosting framework, which are averaged in a weighted, sequential fashion to produce a mixture, the strong learner.

Variational boosting offers better computational efficacy due to iterative fitting, while simultaneously improving approximation prowess owing to the more flexible mixture distribution class. Guo et al. [2016] modify the boosting method based on L_2–regularized variational objective. Miller et al. [2017] incorporate covariance structure to modify the variational family. Locatello et al. [2017] provide some theoretical basis of this computational method using truncated densities as mixture components. However, their result is limited to compactly supported densities only, thus technically not including even Gaussian distributions. This idea is extended in Locatello et al. [2018] for black box variational inference; refer to Ranganath et al. [2014] for the original work on black box variational inference. Wang [2016] uses gradient boosting technique, and suffers from a drawback similar to Locatello et al. [2017]. Campbell and Li [2019] note that the domain of mixture families does not allow Kullback–Leibler divergence to be sufficiently smooth, hence switches to Hellinger metric to provide algorithm and theoretical study for boosting. We address the above problem with the boosting method by providing a pathway to work with mixture families and simultaneously maintaining the use of Kullback–Leibler divergence.

Our contribution to variational boosting revolves around frequentist properties of the variational solution. We study this by proposing a small bandwidth Gaussian mixture variational family and using a functional version of the Frank–Wolfe algorithm (refer to Frank and Wolfe [1956] for the original formulation) for the variational optimization routine. Our method relaxes the assumption in Locatello et al. [2017] regarding compact support of variational distributions, allows working with the standard choice of Kullback–Leibler divergence in contrast to Campbell and Li [2019], as well as makes assumptions that are strictly milder than Local Asymptotic Normality (LAN) type assumptions in Wang and Blei [2018]. Our first result is on understanding statistical properties of the global optimizer of the boosting algorithm. In particular, we show that the Kullback–Leibler divergence of the posterior from the optimal variational solution is bounded in probability, a phenomenon that is similar to what is observed in Bernstein-von Mises theorems for regular parametric models. Our second result pertains to convergence analysis of the algorithm. Our findings are less than encouraging, much along the conjecture of Campbell and Li [2019]. Specifically, we show that the number of iterations required for the boosting algorithm to converge is exponential in the inverse bandwidth, which is a parameter crucial to the definition of our small bandwidth mixture variational family. We provide intutive justification for this in sections 4.2 and 4.3.

2.2 Background and target

We start with $\mathbb{R}^p$-valued data points $X_1 \ldots, X_n$ which are independently and identically distributed according to density $f(x; \theta)$, where $\theta \in \mathbb{R}^d$, the parameter space. Given a prior density $\pi(\theta)$ on parameter θ, we denote the posterior of θ as

$$\pi_n(\theta) = \frac{\prod_{i=1}^n f(X_i; \theta)\pi(\theta)}{\int \prod_{i=1}^n f(X_i; \theta)\pi(\theta) d\theta}.$$

Variational boosting works with the following type of distribution family on $\mathbb{R}^d$:

$$\mathcal{Q} = \mathrm{conv}(\Gamma) = \left\{ \sum_{k=1}^{K} \beta_k \phi_k : \phi_k \in \Gamma,\ \boldsymbol{\beta} \in \Delta^K,\ K \geq 1 \right\},$$

where Γ is any family of simple distributions on $\mathbb{R}^d$ and Δ^K denotes the unit simplex in $\mathbb{R}^K$. The variational family for boosting framework is thus a flexible mixture family. The main aim is to find $q_n^*(\mathcal{Q}) \in \mathcal{Q}$ such that

$$q_n^*(\mathcal{Q}) = \operatorname{argmin}\{q \in \mathcal{Q} \mid KL(q||\pi_n)\}, \quad m_n^*(\mathcal{Q}) = KL(q_n^*(\mathcal{Q})||\pi_n), \tag{2.2.1}$$

where, $KL(a||b) = \int a(\theta)\log(a(\theta)/b(\theta))d\theta$ denotes the Kullback–Leibler divergence of density b from density a, both defined on the parameter space. This proceeds through an optimization routine, called the boosting technique. To describe the algorithm, let $\psi_n^{(k)}$ denote the k-th iterate in the algorithm for $k \geq 0$. Given $\psi_n^{(k)}$, the next iterate $\psi_n^{(k+1)}$ is obtained by

$$\psi_n^{(k+1)} = (1-\gamma_K)\psi_n^{(k)} + \gamma_k\phi_n^{(k+1)},$$

where the weight $\gamma_k \in [0,1]$ and $\phi_n^{(k+1)} \in \Gamma$ depend on the boosting approach employed. Locatello et al. [2017] proposed the use of Frank–Wolfe algorithm to tackle boosting technique iterates and our setup bears similarity to theirs. Iterates of this kind are also used in Guo et al. [2016] and Locatello et al. [2018]. Observe that, if $\psi_n^{(k)}$ is already a mixture distribution, every iteration just adds a new component, namely $\phi_n^{(k+1)}$, to the mixture. The Frank–Wolfe algorithm (see appendix; also refer to Jaggi [2013] and Frank and Wolfe [1956]) handles optimization by proceeding exactly in this fashion, and hence is a natural choice as variational algorithm in this case. A quantity $\mathcal{C}_n$ called curvature (see appendix for definition), that depends only on the the objective map $q \mapsto KL(q||\pi_n)$ and domain of optimization $\mathcal{Q}$, plays a crucial role in this algorithm. After initializing with some $\phi_n^{(0)} \in \Gamma$, the k-th step of the routine involves finding the new component $\phi^{(k+1)}$ to be added, through a linear minimization routine called linear minimization oracle (LMO). This intermediate step is carried out by solving the LMO approximately in terms of the derivative of the objective map and curvature $\mathcal{C}_n$. Now note that, $q_n^*(\mathcal{Q})$ defined in (1) is a random quantity with respect to data $X_1, \ldots X_n$, and so is each iterate of the boosting routine. We aim to provide statistical properties of the random quantities $m_n^*(\mathcal{Q})$ and $\psi_n^{(k)}$ with respect to the true data generating distribution. As a first step, we show stochastic boundedness of $m_n^*(\mathcal{Q})$ in our theorem 1. We next use theorem 1 from Jaggi [2013] and our result on upper bounding the curvature $\mathcal{C}_n$ (theorem 2) to upper bound the decrements of objective value in the boosting algorithm in terms of the variational family hyper-parameters and sample size n. Finally, we tie up the above two results to gain parity of the theoretical minimum and the algorithm. We end with a result (corollary 2) on the order of the required number of boosting updates for a certain degree of error.

2.3 Statistical properties of the variational optimizer

2.3.1 The small bandwidth mixture Gaussian family

Recall the definition of $m_n^*(\mathcal{Q})$, the minimum of the objective map $q \mapsto KL(q||\pi_n)$ over domain $\mathcal{Q}$. Since the function $a \mapsto KL(a||b)$ has closed sub-level sets, this minimum is attained, i.e. $m_n^*(\mathcal{Q}) = KL(q_n^*(\mathcal{Q})||\pi_n)$ is the minimum corresponding to the domain $\mathcal{Q}$. $m_n^*(\mathcal{Q})$ is a random quantity with respect to data $X_1 \ldots X_n$, so we can make probability statements about it with respect to the true data generating distribution. Before we state our stochastic boundedness theorem (section 3.3), we discuss our setup and introduce our variational family. Consider the following restricted Gaussian family for a fixed c_0 with $1 < c_0 < 2$, and some $M, \sigma_n > 0$:

$$\Gamma_n = \left\{ N\left(\mu, \sigma^2 I_d\right) : \|\mu\|_2 \leq M, \; 0 < \sigma_n \leq \sigma \leq c_0^{1/2}\sigma_n \right\}. \tag{2.3.1}$$

Denoting by $\text{conv}(\Gamma_n)$ the set of all finite affine combinations of members of Γ_n, we define $\mathcal{Q}_n = \text{conv}(\Gamma_n)$ as the following restricted mixture Gaussian family:

$$\mathcal{Q}_n = \left\{ \sum_{k=1}^{K} \beta_k \phi_k : \phi_k \in \Gamma_n, \; \boldsymbol{\beta} \in \Delta^K, \; K \geq 1 \right\}. \tag{2.3.2}$$

This domain $\mathcal{Q}_n$ is our variational family of choice, which we call the small bandwidth mixture Gaussian family. Observe that, the components of the mixtures are isotropic Gaussians with means lying in the radius-M compact Euclidean ball around zero in $\mathbb{R}^d$, while the variance σ is constrained to be of the same order as σ_n, the bandwidth parameter. The specific constraint on the constant c_0

plays a crucial role in our analysis and will be justified when we study the employed algorithm and its convergence in detail in section 4.

2.3.2 Comparison of divergences and Bernstein-von Mises Phenomenon

The Bernstein-von Mises (BvM) theorem is a well-known frequentist phenomenon for Bayesian posteriors; refer to Van der Vaart [2000] for an overview of the BvM phenomenon. It encompasses results about the asymptotic normality of appropriately scaled posterior distributions under regularity conditions on the likelihood and the prior. Local Asymptotic Normality (LAN) assumptions on the likelihood is a pathway to achieving BvM, and Wang and Blei [2018] employ it in the context of variational inference. However, BvM results use total variation distance (d_{TV}) as the metric, and we wish to focus on Kullback–Leibler discrepancy. In this context, we give a brief comparison of Kullback–Leibler divergence (KL) with total variation distance and Hellinger distance (d_H). We choose to work with a simplified setting in order to help emphasize our point. Suppose we use a single normal distribution to approximate the posterior. Say $X_1, \ldots, X_n$ are d-vectors, which are independently and identically distributed as $N(\theta, \Sigma)$, with $\theta \sim N(\mu_0, \Sigma_0)$ and μ_0, Σ, Σ_0 known. Let $\overline{X}_n$ denote the sample mean, and μ_n, Σ_n the posterior normal's mean and variance respectively. Let θ_0 denote the true parameter. Then the following statements follow from straightforward calculations for Gaussians,

Results:

$$KL\left(N\left(\theta_0, n^{-1}\Sigma\right)||N\left(\mu_n, \Sigma_n\right)\right) \rightsquigarrow \frac{1}{2}\chi_d^2, \tag{2.3.3}$$

$$d_{TV}\left(N\left(\mu_n, \Sigma_n\right), N\left(\theta_0, n^{-1}\Sigma\right)\right) \to 0 \quad \text{a.s.}, \tag{2.3.4}$$

$$d_H\left(N\left(\mu_n, \Sigma_n\right), N\left(\theta_0, n^{-1}\Sigma\right)\right) \to 0 \quad \text{a.s.}, \tag{2.3.5}$$

$$KL\left(N\left(\overline{X}_n, n^{-1}\Sigma\right)||N\left(\mu_n, \Sigma_n\right)\right) \to 0 \quad \text{a.s.}, \tag{2.3.6}$$

where '$\rightsquigarrow$' denotes weak convergence and a.s stands for almost sure validity with respect to the data generating distribution. Refer to the appendix for proofs of these statements. Since d_{TV}

and d_H are equivalent distances, (2.3.4) and (2.3.5) imply each other, so we just compare (2.3.3) with (2.3.6). The result in (2.3.4) says the posterior comes close to a single Gaussian distribution in total variation, a.s with respect to the data. However the very same distributions are not close in Kullback–Leibler divergence, even in the simplest Gaussian case, as pointed out by (2.3.3). This suggests that under Kullback–Leibler divergence, which is a stronger measure of discrepancy, the divergence between the posterior and a deterministic approximator of it should not go down to zero. We shall see this property in play in our theorem 1.

Now note the comparison of (2.3.3) and (2.3.6), where, just by changing the centering from the truth (a deterministic quantity) to the sample mean (a random data-dependent quantity), we achieve convergence to zero under Kullback–Leibler divergence. However, this phenomenon is very special to this case, as the correct centering may be computationally impossible to find for complicated posteriors. Hence, (2.3.3) is of more practical importance to us than (2.3.6) as a statistical statement. In the next section, we present a result similar in flavor to (2.3.3), but milder and applicable much more generally.

In general, for the BvM phenomenon to hold, strong regularity conditions are required, which guarantee posterior shape (Gaussian) with high probability with respect to data. However, we wish to include those posteriors in our analysis as well whose shapes are non-Gaussian, making our analysis more general. Works in Kruijer et al. [2010] and Shen et al. [2013] establish approximations of deterministic densities, suitably smooth and exponentially tailed, using Gaussian mixtures. However, if we wish to apply such results to the posterior, which is a random density, we need high probability statements about the smoothness and tail of the posterior, which might be tantamount to using hypotheses that the BvM phenomenon demands.

2.3.3 Stochastic Boundedness of the Kullback–Leibler Discrepancy

We now state a theorem about the theoretical minimum Kullback–Leibler divergence $m_n^*(\mathcal{Q}_n)$; see (2.2.1) for definition of m_n^* and (2.3.2) for definition of $\mathcal{Q}_n$. Recall that $X_1, \ldots X_n$ are independently and identically distributed data points following density $x \mapsto f(x;\theta)$ and $\pi(\theta)$ is the prior on the parameter $\theta \in \mathbb{R}^d$. With $\theta_0 \in \mathbb{R}^d$ denoting the true parameter value, define the log-likelihood ratio $\ell_i(\theta, \theta_0) = \log\left(f(X_i;\theta)/f(X_i;\theta_0)\right)$ for $i = 1, \ldots, n$, $L_n(\theta, \theta_0) = \sum_{i=1}^{n} \ell_i(\theta, \theta_0)$, $KL(\theta_0||\theta) = -E(\ell_1(\theta, \theta_0))$, $\mu_2(\theta_0||\theta) = E(\ell_1(\theta, \theta_0))^2$ and $U(\theta) = -\log(\pi(\theta))$ Also, denote by $KL^{(j)}(\theta_0||\theta)$ and $\mu_2^{(j)}(\theta_0||\theta)$ for $j = 1, 2$ the respective derivatives of the maps with respect to the second argument. Let $s_{\max}(A)$ stand for the highest singular value of square matrix A. $\|.\|_2$ stands for the l_2 norm on $\mathbb{R}^d$. Denote by $\lesssim, \gtrsim$ the corresponding inequalities up to absolute constants. The following assumptions will be required for the theorem:

Assumption 1: The truth θ_0 satisfies $\|\theta_0\|_2 \leq M$.

Assumption 2: The variance bound σ_n satisfies $\sigma_n \leq n^{-1/2} \leq c_0^{1/2}\sigma_n$ for all $n \geq 1$.

Assumption 3: The quantities $KL(\theta_0||\theta), \mu_2(\theta_0||\theta)$ are finite for every $\theta \in \mathbb{R}^d$.

Assumption 4: Matrices $KL^{(2)}(\theta_0||\theta), \mu_2^{(2)}(\theta_0||\theta)$ and $U_2^{(2)}(\theta)$ exist on $\mathbb{R}^d$ and satisfy for any $\theta, \theta' \in \mathbb{R}^d$:

$$\begin{aligned} s_{max}\left(KL^{(2)}(\theta_0||\theta) - KL^{(2)}(\theta_0||\theta')\right) &\lesssim \|\theta - \theta'\|_2^{\alpha_1}, \\ s_{max}\left(\mu_2^{(2)}(\theta_0||\theta) - \mu_2^{(2)}(\theta_0||\theta')\right) &\lesssim \|\theta - \theta'\|_2^{\alpha_2}, \\ s_{max}\left(U_2^{(2)}(\theta) - U_2^{(2)}(\theta')\right) &\lesssim \|\theta - \theta'\|_2^{\alpha_3}, \end{aligned} \tag{2.3.7}$$

for some $\alpha_1, \alpha_2, \alpha_3 \geq 0$.

Assumption 5: $KL(\theta_0||\theta) \gtrsim \|\theta - \theta_0\|_2^2$.

Theorem 1: *Under assumptions 1-5, it holds that $m_n^*(\mathcal{Q}_n)$ is bounded in probability with respect to the data generating distribution, i.e. given any $\epsilon \in (0, 1)$, there exists $M_\epsilon, N_\epsilon > 0$ such that for all $n \geq N_\epsilon$*, we have with probability greater than $1 - \epsilon$

$$m_n^*(\mathcal{Q}_n) < M_\epsilon. \tag{2.3.8}$$

Remark 1: Assumption 2 dictates the exact order of σ_n, and also allows $N_\epsilon = 1$. It is at par with the order of the bandwidth expected in parametric estimation. Finiteness of $\mu_2(\theta_0||\theta)$ in assumption 3 is crucial for concentration inequalities to be applied. The smoothness assumption i.e. assumption 4, helps dictate posterior moments, but not the shape of the entire posterior. The final assumption is the standard identifiability condition for using Kullback–Leibler divergence. Assumptions of these types are quite common in the literature; refer to the moment assumptions for the posterior in Huggins et al. [2020] in the context of distributional bounds in variational inference and section 5 in Ghosal et al. [2000] in the context of posterior contraction. An important observation is the fact that we do not intend to recover the posterior mean as in Pati et al. [2018], where assumptions are aimed at studying variational point estimates.

Remark 2: The hypothesis for the theorem is milder than what can guarantee a weak convergence type result, like (2.3.3). There are no assumptions that allow local quadratic nature of the posterior, as we impose conditions only on the expected log-likelihood. This makes our assumptions more general than a BvM type setup. Contrast this with the stochastic LAN assumption in Wang and Blei [2018], which approximates the likelihood with a stochastic linear term and a deterministic quadratic term.

Remark 3: Theorem 1 establishes $m_n^*(\mathcal{Q}_n)$ to be an $O_p(1)$ quantity with respect to the true data generating distribution. From the proof in appendix, one can further conclude $M_\epsilon \gtrsim \epsilon^{-1/2}$ for small ϵ.

We shall state a corollary which is a simplification of theorem 1 in the case the density $x \mapsto f(x : \theta)$ belongs to the K-parameter exponential family on $\mathbb{R}^p$. Let $\theta \in \mathbb{R}^d$ be the canonical parameter, $T_l : \mathbb{R}^p \mapsto \mathbb{R},\ l = 1, \dots K$ be the sufficient statistics and $A : \mathbb{R}^d \mapsto \mathbb{R}$ be the log-partition function. The form of the density is given by

$$f(x;\theta) = \exp\left(\sum_{l=1}^{K} \theta_l T_l(x) - A(\theta)\right).$$

Let $A^{(j)}(\theta), j = 1, 2$ denote the respective derivatives of $A(\theta)$. We shall use the notion of strong convexity in the corollary that follows; refer to the appendix for a general definition. We also need to note down the definition of α-Lipschitz functions:

Definition: Vector or square-matrix valued functions f defined on $D \subset \mathbb{R}^d$ are said to be α-Lipschitz for an $\alpha > 0$, if there exists a constant $C > 0$ such that for all $x, y \in D$

1. $\|f(x) - f(y)\|_2 \leq C\|x - y\|_2^\alpha$ for vector valued functions, and

2. $s_{\max}(f(x) - f(y)) \leq C\|x - y\|_2^\alpha$ for square-matrix valued functions.

Corollary 1: *Assume that $A(\theta)$ is twice differentiable and strongly convex on $\mathbb{R}^d$. Also assume that the vectors $A^{(1)}(\theta), A^{(2)}(\theta)\theta$ and the square matrices $A^{(1)}(\theta)A^{(1)}(\theta)^T, A^{(2)}(\theta)$ are α-Lipschitz functions of $\theta \in \mathbb{R}^d$ for some $\alpha \geq 0$. Under these conditions, assumptions on the expected likelihood (assumption 3 and 4) in theorem 1 hold.*

2.3.4 Sketch of Proof of Theorem 1

We now briefly discuss the proof technique of theorem 1, shedding more light on the importance of the assumptions made; refer to appendix for a detailed proof of theorem 1. Note that $m_n^*(\mathcal{Q}_n)$ is bounded above by the objective map $q \mapsto KL(q||\pi_n)$ evaluated at any member of $\mathcal{Q}_n$. We choose that member to be q_0, the d-dimensional Gaussian density centered at the truth θ_0 and variance $\sigma_n^2 I_d$. Here, I_d stands for the d-dimensional identity matrix and σ_n satisfies assumption 2. Along with assumption 1, we have, $q_0 \in \mathcal{Q}_n$ and hence $m_n^*(\mathcal{Q}_n) \leq KL(q_0||\pi_n)$. Thus it is enough to show $KL(q_0||\pi_n)$ is bounded in probability. This Kullback–Leibler discrepancy can now be broken down in a sum to give two deterministic and two random quantities. The stochastic part of the sum is given by

$$\log\left(m(X_n)\right) - \left(\int L_n(\theta, \theta_0) q_0(\theta) d\theta\right),$$

where

$$m(X_n) = \int \exp\left(L_n(\theta, \theta_0)\right) \pi(\theta) d\theta.$$

Under true data generating distribution, the integrand defining $m(X_n)$ has expectation 1, which helps upper bound in probability the first stochastic term above. For the second, we notice that $L_n(\theta, \theta_0)$ has expectation $-nKL(\theta_0||\theta)$ under true θ_0. We then utilize the smoothness and identifiability assumptions on $KL(\theta_0||\theta)$ to obtain its Taylor series expansion around θ_0, that helps lower bound in probability the second stochastic term through Chebyshev's inequality. We can now conclude our result by noting that sum of $O_p(1)$ quantities is again $O_p(1)$.

2.4 Convergence analysis of the algorithm

2.4.1 Steps of the Algorithm

We opt for a functional version of the Frank–Wolfe algorithm as our variational boosting algorithm (algorithm 1), which bears analogy to variant 0 of algorithm 1 in Locatello et al. [2017]. Refer to Jaggi [2013] for the general algorithm and our discussion in the appendix for a brief overview and notations. Recall that we aim to perform the optimization (1) with domain $\mathcal{Q} = \mathcal{Q}_n = \mathrm{conv}(\Gamma_n)$. Thus our objective map is $q \mapsto KL(q||\pi_n), q \in \mathcal{Q}_n$, which is convex on its domain. Let us generically denote members of Γ_n, which are single Gaussians, by ϕ and those of $\mathcal{Q}_n$, which are mixture Gaussians, by ψ. Superscripts stand for iterate numbers and the subscript n denotes dependence on sample size. Say $\psi_n^{(k)}$ is the mixture obtained as the k-th step iterate. We use the notation $\mathcal{D}_n$ to denote the Bregman divergence of our convex objective map (refer to appendix for general definition of Bregman divergence). For $\psi_1, \psi_2 \in \mathcal{Q}_n$, the Bregman divergence $\mathcal{D}_n$ of ψ_1 from ψ_2, at ψ_1, under the objective map, is given by

$$\mathcal{D}_n(\psi_2||\psi_1) = KL(\psi_2||\pi_n) - KL(\psi_1||\pi_n) - \int (\psi_2 - \psi_1)\left(\log(\psi_1) - \log(\pi_n)\right) d\theta. \qquad (2.4.1)$$

The last term above derives from the fact that our domain of optimization, $\mathcal{Q}_n$, lies embedded in the L_2 inner product function space on $\mathbb{R}^d$. The term $\log(\psi_1) - \log(\pi_n)$, appearing within the integrand in (10), is the sub-gradient (see appendix for general definition of subgradient) of the objective map at ψ_1, and $\psi_n^{(k)}$ plays the role of ψ_1 at the k-th step of the algorithm. We use the notation $\mathcal{C}_n$ to denote the curvature of the objective map for the domain $\mathcal{Q}_n$ (see section 4.3 for details on $\mathcal{C}_n$).

That $s_n^{(k)} = \log(\psi_n^{(k)}) - \log(\pi_n)$ in algorithm 1 is indeed a valid subgradient of the objective map at $\psi_n^{(k)}$, follows from the following lemma on Bregman divergence $\mathcal{D}_n$ and the non-negativity of Kullback–Leibler divergence. It basically says calculating $\mathcal{D}_n$ and Kullback–Leibler divergence of the objective map are one and the same.

Algorithm 1 Functional Frank–Wolfe Algorithm with small bandwidth mixture Gaussian variational family

```
1. Initialize with $\psi_n^{(0)} = \phi_n^{(0)} \in \Gamma_n$.
2. At $k$-th step, calculate the subgradient $s_n^{(k)} = \log(\psi_n^{(k)}) - \log(\pi_n)$ at the $k$-th iterate $\psi_n^{(k)}$.
3. Set $\gamma_k = 2/(k+2)$, solve LMO approximately i.e. find $\phi_n^{(k+1)} \in \Gamma_n$ such that $\int s_n(\theta)\phi_n^{(k+1)} d\theta \leq \min\left\{\phi \in \Gamma_n \mid \int s_n^{(k)}\phi(\theta)d\theta\right\} + \gamma_k \mathcal{C}_n/2$.
4. Update $\psi_n^{(k+1)} = (1-\gamma_k)\psi_n^{(k)} + \gamma_k \phi_n^{(k+1)}$ to get the $(k+1)$-th iterate.
```

Lemma 1: *For any densities ψ_1 and ψ_2 defined on $\mathbb{R}^d$, we have*

$$\mathcal{D}_n(\psi_2||\psi_1) = KL(\psi_2||\psi_1).$$

In algorithm 1, the target is to greedily fit single Gaussian components to build a mixture of Gaussians, that is close to the posterior in Kullback–Leibler divergence. Let the optimal approximating mixture be denoted by $\psi_n^* = q_n^*(\mathcal{Q}_n)$ (see (1) for definition of q_n^*). Note that, both $\psi_n^{(k)}$ and ψ_n^* are random with respect to data. But the practitioner is given the data $X_1, \dots X_n$, hence she runs algorithm 1 deterministically and upper bounds, point-wise on the sample space, the random quantity

$$KL(\psi_n^{(k)}||\pi_n) - KL(\psi_n^*||\pi_n).$$

We want to find the aforementioned upper bound in terms of number of iterations k and sample size n.

Step 3 of algorithm 1 is an approximate linear minimization routine for which we shall use the parametric structure of Gaussians and optimize over the parameters. The practitioner starts with an initial guess of $\mu^{(0)}, \sigma^{(0)}$ such that $N\left(\mu^{(0)}, (\sigma^{(0)})^2 I_d\right) \in \Gamma_n$. This gives her the Gaussian $\phi_n^{(0)} = N\left(\mu^{(0)}, (\sigma^{(0)})^2 I_d\right)$. At the beginning of the k-th step, she has the mixture $\psi_n^{(k-1)} = \sum_{j=1}^{k-1} \beta_j \phi_n^{(j)}$, where the vector $\beta \in \Delta^{k-1}$ and are positive functions of γ_l's with $\gamma_l = 2/(l+2), l = 1 \dots k-1$.

In order to obtain the Gaussian $\phi_n^{(k)}$, she finds approximate $\mu^{(k)}, \sigma^{(k)}$ through the LMO optimization routine

$$\operatorname{argmin}\left\{ \|\mu\|_2 \leq M,\ 0 < \sigma_n \leq \sigma \leq \sqrt{c_0}\sigma_n : \int \phi(\theta; \mu, \sigma^2) \log\left(\frac{\psi_n^{(k-1)}(\theta)}{\pi_n(\theta)}\right) d\theta \right\}. \qquad (2.4.2)$$

By lemma 5 of Jaggi [2013], the practitioner is allowed to use any algorithm at her disposal for the above routine, as long as she is able to perform the optimization of this k-th step with error $\leq \gamma_k \mathcal{C}_n/2$. A very important observation is that knowing the normalizing constant of the posterior is not necessary for the above routine. Thus, using the $\phi_n^{(k)}$ obtained, we update $\psi_n^{(k)} = (1-\gamma_k)\psi_n^{(k-1)} + \gamma_k \phi_n^{(k)}$.

2.4.2 Rate of Convergence of Algorithm 1

We now state our main theorem on rate of convergence:

Theorem 2:

$$KL(\psi_n^{(k)}||\pi_n) - KL(\psi_n^{*}||\pi_n) \leq \frac{8(2-c_0)^{-d/2}\exp\left(\frac{2M^2}{(2-c_0)\sigma_n}\right)}{k+2}. \qquad (2.4.3)$$

This theorem upper bounds the gap in value of the Kullback–Leibler objective map, between the k-th boosting iterate and the optimal approximator to the posterior. It depends upon the sample size, number of iterations, dimension of parameter space and hyper-parameters of the domain $\mathcal{Q}_n$. We make the following important remarks:

Remark 4: The above convergence-rate holds point-wise with respect to the data generating distribution, and the upper bound is deterministic.

Remark 5: The rate is sub-linear in the number of iterates k and exponential in inverse bandwidth σ_n^{-1}. Sub-linearity follows from the use of Frank–Wolfe, while exponentiality is an artifact of enforcing Kullback–Leibler divergence to be strongly smooth over our small bandwidth domain.

We now combine theorems 1 and 2 to obtain a novel probability statement about the random k-th iterate of our boosting algorithm. It is important to know how many iterations we need in order to obtain a certain error in the algorithm, and that depends on the sample size n. So we let k vary with sample size n, i.e. take $k \equiv k_n$.

Corollary 2: *If* $k_n \gtrsim \exp(n^{1/2})$, *then for any* $\epsilon \in (0,1)$, *there exists constant* $C > 0$, *such that with probability greater than* $1-\epsilon$ *we have*

$$KL(\psi_n^{(k_n)}||\pi_n) < C\epsilon^{-1/2}. \tag{2.4.4}$$

Remark 6: Corollary 2 shows the the required order of k_n, the number of iterations, in terms of sample size n that maintains order parity while combining theorems 1 and 2. It shows after how many runs of the boosting algorithm the random iterates are guaranteed to be bounded in probability with respect to the data generating distribution.

2.4.3 Sketch of Proof of Theorem 2

We briefly discuss how we arrived at theorem 2. For k-th step, theorem 1 of Jaggi [2013] allows the following upper bound regarding our objective map:

$$KL(\psi_n^{(k)}||\pi_n) - KL(\psi_n^{*}||\pi_n) \leq \frac{4\mathcal{C}_n}{k+2}, \tag{2.4.5}$$

where $\mathcal{C}_n$ denotes the curvature of the objective map $q \mapsto KL(q||\pi_n)$ over domain of optimization $\mathcal{Q}_n$. It is given by

$$\mathcal{C}_n = \sup\left\{\frac{2}{\alpha^2}\mathcal{D}_n(\psi_2||\psi_1) : \psi_1 \in \mathcal{Q}_n, \phi \in \Gamma_n, \alpha \in [0,1], \psi_2 = \psi_1 + \alpha(\phi - \psi_1)\right\}. \tag{2.4.6}$$

Curvature is thus essentially the maximum scaled Bregman divergence between densities in domain $\mathcal{Q}_n$ and their perturbations through mixtures. It is entirely determined by n, σ_n and the variational family hyper-parameters M, c_0, d. Refer to the appendix for a general definition of curvature. We prove the following lemma that upper bounds $\mathcal{C}_n$ on the small bandwidth mixture Gaussian domain $\mathcal{Q}_n$:

Lemma 2:

$$\mathcal{C}_n \leq 2(2 - c_0)^{-d/2} \exp\left(\frac{2M^2}{(2 - c_0)\sigma_n}\right). \tag{2.4.7}$$

Remark 6: Note the exponential dependence of the bound on inverse bandwidth σ_n^{-1}. Whether $\mathcal{C}_n$ is a finite quantity depends on the constraints in the definition of Γ_n (see section (2.3.1)) through which it is defined. We are enforcing finiteness of $\mathcal{C}_n$ through utilizing proposed small bandwidth mixture Gaussian family. which is the essence of the constraint $1 < c_0 < 2$ in the definition of Γ_n.

Remark 7: Locatello et al. [2017] and our work share the central motive of ensuring strong smoothness (see definition in appendix) of Kullback–Leibler divergence, which essentially dictates finiteness of $\mathcal{C}_n$. It is a property of both the objective function to be minimized, and the domain over which it is minimized. However, Locatello et al. [2017] assume Γ_n to consist of truncated, lower bounded densities, which although works from a practical point of view, excludes the very basic Gaussian distribution. Their theory only accommodates compactly supported densities, which is too restrictive and is remedied through the above lemma. Thus lemma 2 is a cardinal contribution of this work; it shows that the curvature of Kullback–Leibler discrepancy is bounded over the domain of small bandwidth Gaussian mixture family.

Remark 8: If $c_0 \geq 2$, then upper bound on $\mathcal{C}_n$ improves with smaller bandwidth σ_n, which is evidently untrue, as smaller bandwidth Gaussian mixtures are spikier and worse approximators through Kullback–Leibler divergence. If $c_0 < 1$, the bound improves for higher dimension, which contradicts curse of dimensionality. This intuitively justifies the technical importance of the constraint: $1 < c_0 < 2$.

We now state a straightforward formula that shall play a crucial role in the proof of lemma 2, as well as provide insight into why the proposed family of small bandwidth Gaussian mixtures is a good choice as a variational family.

Lemma 3: *Whenever $2\sigma_1^2 > \sigma_2^2 > 0$, χ^2 distance between two Gaussians is given by*

$$\chi^2\left(\mathrm{N}\left(\mu_2, \sigma_2^2\mathbf{I}_d\right) \,||\, \mathrm{N}\left(\mu_1, \sigma_1^2\mathbf{I}_d\right)\right) = -1 + \left(\frac{\sigma_1^2}{\sigma_2\sqrt{2\sigma_1^2 - \sigma_2^2}}\right)^d \exp\left(\frac{\|\mu_2 - \mu_1\|_2^2}{2\sigma_1^2 - \sigma_2^2}\right). \tag{2.4.8}$$

Remark 8: The χ^2 distance (refer to appendix for definition) arises during the calculations of curvature of Kullback–Leibler divergence, and one can clearly see the right hand side above is defined as a real number only for a certain interval of values of the variances. This calculation was an important motivator for the small bandwidth variational family. It is also the reason for the exponential dependence of the upper bound, in theorem 2, on the inverse bandwidth.

We can now directly plug into lemma 3 the domain parameters of $\mathcal{Q}_n$ to calculate the proposed upper bound in lemma 2. In turn, lemma 2 directly gives us theorem 2 in combination with (2.4.5).

2.5 Concluding remarks

To the best of our knowledge, we provide, for the first time, statistical properties of the iterates in a variational boosting algorithm. As part of frequentist validation of this variational method, we assume regularity conditions similar to ones widely used in Bayesian contraction literature. Our hypotheses are general enough to include most likelihoods and priors; they do not demand Gaussian like posteriors. Regarding the boosting algorithm itself, we employ the non-compact domain of mixture Gaussian densities as our variational family, much in contrast to the use of compact domain in Locatello et al. [2017]. Convergence analysis of our algorithm shows how smoothness properties of Kullback–Leibler discrepancy lead to exponential dependence of iterate number on the inverse bandwidth, which can be contrasted to faster convergence rates for weaker metrics like Hellinger distance [Campbell and Li, 2019].

3. ADAPTIVE POSTERIOR CONVERGENCE IN SPARSE HIGH DIMENSIONAL CLIPPED GENERALIZED LINEAR MODELS

3.1 Introduction

The GLM [McCullagh, 2019] is a flexible generalization of ordinary linear regression that allows for response variables to accommodate error distributions which are non-additive and non-Gaussian. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function. Although primarily restricted to a lower dimensional setting, Bayesian approaches for GLM has been very popular from the 90's with the advent of Markov chain Monte Carlo [Dey et al., 2000].

The emergence of more sophisticated data acquisition techniques in gene expression microarray, among many other fields, triggered the development of innovative statistical methods [Friedman et al., 2001, Bühlmann and Van De Geer, 2011, Hastie et al., 2015] in the last decade, that help in analyzing large scale datasets. The overarching goal is to identify relevant predictors associated with a response out of a large number of predictors, but only with a smaller number of samples. This *large p, small n* paradigm is arguably the most researched topic in the last decade. Primarily focusing on the linear models, statisticians have devised a number of penalized regression techniques for estimating β in $p \gg n$ setting under the assumption of sparsity, with accompanying theoretical justification of optimal estimation, prediction and selection consistency; refer to Tibshirani [1996], Fan and Li [2001], Efron et al. [2004], Zou and Hastie [2005], Candes et al. [2007], Zou [2006], Belloni et al. [2011]. Pioneering extensions of penalization based methods have been made for generalized linear models [Friedman et al., 2010], but existing results on theoretical guarantees for high dimensional GLMs are relatively few. Van de Geer et al. [2008] studied the oracle rate of the empirical risk minimizer with the lasso penalty in high dimensional GLMs. More recently, Abramovich and Grinshtein [2016] derived convergence rates with respect to the Kullback–Leibler risk with a wide class of penalizing functions, which can be translated into convergence rates relative to the ℓ_2-norm under certain conditions.

From a Bayesian standpoint, sparsity favoring mixture priors with separate control on the signal and noise coefficients have been proposed [Leamer, 1978, Mitchell and Beauchamp, 1988, George and McCulloch, 1995, 1997, Scott et al., 2010, Johnson and Rossell, 2010, Narisetty et al., 2014, Yang et al., 2016, Ročková and George, 2018]. Although in principle such methods can be used for generalized linear models, accompanying theoretical justification on optimal estimation in the high dimensional case is primarily available in the context of linear models [Castillo and van der Vaart, 2012, Castillo et al., 2015, Gao et al., 2015].

To the best of our knowledge, analogous results for generalized linear models in the high dimensional case are comparatively sparse, with the exception of Jiang et al. [2007]. However, special cases from the GLM family including high dimensional logistic regression using a pseudo likelihood [Atchadé, 2017] and high dimensional logistic regression using shrinkage priors [Wei and Ghosal, 2020] are available. Jiang et al. [2007] operated in a high dimensional setting where the use of a Gaussian prior leads to a restrictive assumption on the growth of the true coefficients; refer to the assumptions of Theorem 1 in pg. 1493. Atchadé [2017] considered a Laplace-type prior for the coefficients which obviated the need for such a restriction, but their results are specific to logistic regression.

In this article, we develop a framework to study posterior contraction in high dimensional clipped generalized linear models using complexity priors that involve a Laplace prior on the non-zero coefficients. The clipped GLM class deviates slightly from the standard GLM construction in that we allow the effect of linear term $x^{\mathrm{T}}\beta$ in the argument of the log-partition function to "clip" away from the singularities of the function. Our clipped GLM directly subsumes high dimensional linear, polynomial and logistic regression, while also incorporating variants of Poisson, negative Binomial (and similar) regressions, which are identical from a practical standpoint to the standard Poisson/negative binomial regressions.

Our sufficient conditions are grouped into two categories: i) a set of identifiability and compatibility conditions based on the geometry of the clipped GLM, specified by the log-partition function that allows separation between models, and ii) an appropriate growth rate of scale parameter of the Laplace distribution that imposes appropriate penalty on the non-zero coefficients, along with an appropriate decay rate for the model weights that penalizes larger models. Existing literature [Jiang et al., 2007] on posterior contraction in GLMs requires growth rate assumptions on the true coefficient vector. The crucial feature of our methodology is achieving adaptive, rate-optimal posterior contraction with respect to the data generation mechanism, while simultaneously avoiding any growth assumptions on the true coefficients.

While our article was in final stages, we came across a dissertation by Seonghyun Jeong at NC State University under the supervision of Prof. Subhashis Ghosal, which considers posterior contraction in GLMs using complexity priors on the model space in Chapter 4. Their results make use of the same identifiability and compatibility assumptions as in [Castillo et al., 2015] to deliver optimal posterior contraction rates, albeit with a growth restriction on the true coefficient vector. On the other hand, we do not require any growth assumption on the true coefficient vector. Our assumptions for obtaining adaptive rate-optimal posterior contraction are specifically designed for the clipped GLMs which can be viewed as appropriate generalization of the identifiability and compatibility assumptions of Castillo et al. [2015] in the linear model case. Finally, the prior dependence on the true parameter can be completely eliminated making our results rate-adaptive.

The remaining of the article is organized as follows. Section 3.2 introduces the construction of the clipped GLM family. Section 3.3 details the sparsity favoring prior construction while section 3.4 entails the identifiability and compatibility assumptions on the data generating process and the choice of hyperparameters. Section 3.5 states our main results on adaptive rate-optimal posterior contraction. This is divided into three parts: a lower bound on the marginal likelihood, a result on controlling the effective sparsity of the posterior distribution and finally a truth-adaptive contraction rate theorem. The proofs are deferred to the Appendices B.1-B.3 with the auxiliary results in Appendix B.4.

3.1.1 Notations

For reals ζ_1, ζ_2, $\zeta_1 \lesssim \zeta_1$ denotes $\zeta_1 \leq C_1 \zeta_2$ for an absolute constant C_1. Similarly, we define $\zeta_1 \gtrsim \zeta_2$. For sequences of real numbers $\{\zeta_{1,n}\}$ and $\{\zeta_{2,n}\}$, we say $\zeta_{1,n} = o(\zeta_{2,n})$ if $\zeta_{1,n}/\zeta_{2,n} \to 0$, and $\zeta_{1,n} = O(\zeta_{2,n})$ if we have $0 < C_2 \leq \liminf_{n\to\infty} (\zeta_{1,n}/\zeta_{2,n}) \leq \limsup_{n\to\infty} (\zeta_{1,n}/\zeta_{2,n}) \leq C_3$ for absolute constants C_2, C_3.

3.2 Construction of GLM family

For both univariate and multivariate observations, one of the most widely used and well-structured family of models is the exponential family. One can refer to Koopman [1936] and Pitman [1936] for the initial works on exponential families. We discuss this briefly with the example of univariate observations and real valued parameter. The exponential family takes the form

$$f(y \mid \theta) = h(y) \exp\left[\theta T(y) - A(\theta)\right], \ y \in \mathcal{Y} \subset \mathbb{R}, \tag{3.2.1}$$

where $\theta \in \Theta \subset \mathbb{R}$ is the parameter of interest, $h(\cdot) : \mathcal{Y} \to \mathbb{R}$ is called the base measure, $A(\cdot) : \Theta \to \mathbb{R}$ is the convex log-partition function and $T(\cdot) : \mathcal{Y} \to \mathbb{R}$ is called the sufficient statistic for estimating parameter θ. This form is known as the canonical form of an exponential family. Many standard distributions, like the Bernoulli and Gaussian with known variance, Poisson, negative Binomial, among many others, follow model (3.2.1). It is well known that the mean and variance of the sufficient statistic is given in terms of $A(\cdot)$, namely $\mathbb{E}[T(Y)] = A'(\theta), \text{Var}(T(Y)) = A''(\theta)$. $A'(\cdot)$ and $A''(\cdot)$ are thus known as the mean and variance functions respectively, and $A'(\cdot)$ can be assumed to strictly increasing on its domain. An interesting property of exponential families is that it affords a neat expression of Kullback–Leibler(KL) divergence in terms of the Bregman divergence of log-partition function $A(\cdot)$. The Bregman divergence of convex function $A(\cdot)$ at θ_0 from θ is given by $A(\theta) - A(\theta_0) - (\theta - \theta_0)\, A'(\theta_0)$, which turns out to be the same expression for KL divergence of θ_0 from θ, which we denote by $\mathcal{D}(\theta_0 || \theta)$. These properties play a major role in dealing with exponential family distributions.

A generalized linear model (GLM) assumes that the observation comes from an exponential family member as above, and models a function of the mean through a linear function of a covariate, i.e. as $x^{\mathrm{T}}\beta$, where x represents a covariate and β is the new parameter vector of interest. The said function, denoted by $g(\cdot) : \mathrm{range}[A'(\cdot)] \to \mathbb{R}$, is termed as the link function. With n data points and d_n covariates, $X : n \times d_n$ makes up the design matrix, whose i-th row is denoted by x_i^{T}. Thus, for every $i = 1, \dots n$, GLM prescribes the transition θ to β as

$$g^{-1}\left(x_i^{\mathrm{T}}\beta\right) = A'(\theta), \text{ equivalently } \theta = \left(g \circ A'\right)^{-1}\left(x_i^{\mathrm{T}}\beta\right). \tag{3.2.2}$$

It is clear from the right hand side of (3.2.2) that GLM actually models the original parameter of the exponential family, but it does so indirectly, through the link function and $A'(\cdot)$. As we shall see in our next section 3.2.1, (3.2.2) motivates modeling the original parameter θ using $A''(\cdot)$, and not through $A'(\cdot)$, leading to the definition of clipping function $\eta(\cdot)$ and clipped GLM family.

3.2.1 Introduction to clipped GLM

We now discuss in detail the clipped Generalized Linear Model (cGLM), which includes, but are not limited to, the distributions like Bernoulli, binomial with known number of trials, Poisson, negative binomial with known number of failures, exponential, Pareto with known minimum, Weibull with known shape, Laplace with known mean, chi-squared and Gaussian with known known variance. We start with the canonical rank-one exponential family of distributions, where the canonical parameter θ is expressed through a function of covariates. However, in contrast to GLM, we choose to represent

$$\theta = \eta\left(x^{\mathrm{T}}\beta\right),$$

where $\eta(\cdot)$, termed as the clipping function, depends only on $A''(\cdot)$. In cGLM, we consider log-partition functions $A(\cdot)$ that satisfy

- $A''(\cdot)$ exists everywhere in the domain of $A(\cdot)$,
- $\mathcal{I}_A(b) := \{t \in \mathbb{R} : 0 \leq A''(t) \leq b\}$ is an interval on the real line for any $b \in (0, \infty]$.

All the standard examples of exponential families we discuss satisfy these simple properties. We now turn to the clipping functions we use in cGLM, which play an intermediary role, sitting between $A(\cdot)$ and the i-th linear term $x_i^{\mathrm{T}}\beta$. We motivate the choice of clipping functions by describing some examples. Since we work with $\beta \in \mathbb{R}^{d_n}$, the linear term $x_i^{\mathrm{T}}\beta$ belongs to $\mathbb{R}$, whereas the log-partition function A can have strict interval subsets of the real line as their support. These types of log-partition functions have a single *pole* (r_0 such that $\lim_{x\to r_0} A(x) = \infty$) on the real line. Examples include:

- *Negative Binomial:* $A(t) = -q\log\left(1-\exp(t)\right), t < 0$ with q denoting known number of failures. This shows $r_0 = 0$.

- *Exponential:* $A(t) = -\log(-t), t < 0$ so that $r_0 = 0$.

- *Pareto:* $A(t) = -\log(-1-t) + (1+t)\log q_{\min}, t < -1$ with $q_{\min}$ denoting known minimum value. This shows $r_0 = -1$.

- *Laplace:* $A(t) = -\log(-t/2), t < 0$ so that $r_0 = 0$. Mean is assumed to be zero.

Distributions like Bernoulli (or Binomial with known number of trials), Poisson and Gaussian (with known variance) have log-partition functions with entire real line as support. The clipping function's first role is to ensure that $\eta_i \equiv \eta\left(x_i^{\mathrm{T}}\beta\right)$, which acts as an argument to $A(\cdot)$ to have the same range as the domain of $A(\cdot)$. Its second role, which turns out to be the central point of our hyper-parameter assumption, is to control the growth of $A''(\cdot)$, specifically to allow a local quadratic majorizability of $A(\cdot)$. Bernoulli and Gaussian (with known variance) already enjoy the special status of having a universal bound on $A''(\cdot)$. Hence, for Poisson, which has $A(t) = \exp(t), t \in \mathbb{R}$, as well as the distributions that have a pole in their log-partition function, $\eta(\cdot)$ should be assumed to be playing the role of clipping the linear term $x_i^{\mathrm{T}}\beta$ away from $+\infty$ and r_0 respectively, or $\pm\infty$ and poles in general cGLM members. We illustrate one possible set of choices of clipping function $\eta(\cdot)$ in the following examples. It is important to note their connection to the popular regression settings, which we shall delve into in (3.2.2).

- *Bernoulli:* $\eta(t) = t$, due to universal bound on $A''(\cdot)$.

- *Negative binomial with known number of failures:* $\eta(t) = -\delta - \log(1 + \exp(-t - \delta))$, where δ is a small positive absolute constant.

- *Poisson:* $\eta(t) = \mathcal{C}_0 - \log(1 + \exp(-t + \mathcal{C}_0))$, where $\mathcal{C}_0$ is large positive absolute constant (see figure (3.1), where $\mathcal{C}_0 = 10$).

- *Exponential:* $\eta(t) = -\delta - \log(1 + \exp(-t - \delta))$, where δ is a small positive absolute constant.

- *Gaussian with known variance:* $\eta(t) = t$, due to universal bound on $A''(\cdot)$.

- *Pareto with known minimum value:* $\eta(t) = -(1 + \delta) - \log(1 + \exp(-1 - t - \delta))$, where δ is a small positive absolute constant.

- *Laplace with known mean:* $\eta(t) = -\delta - \log(1 + \exp(-t - \delta))$, where δ is a small positive absolute constant.

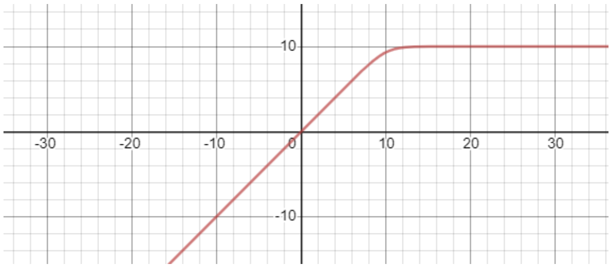

Figure 3.1: Graph of $y = 10 - \log(1 + \exp(-x + 10))$

Two points are crucial to note here. The clipping function $\eta(\cdot)$ can be defined as injective and Lipschitz, as all our examples show. These two properties play an important role in identifiability of the model, as is discussed in the next section. Secondly, the constants $\mathcal{C}_0, \delta$ are absolute, meaning that the practitioner should choose them before-hand, and their choice is totally independent of the observed data or the true value of the parameter in question. An example of such a choice would be $\delta = 10^{-4}$ and $\mathcal{C}_0 = 10^3$. We now summarize the defining properties of clipping functions $\eta(\cdot)$ used in cGLM, and their connection to $A(\cdot)$ through a the following simple and mild condition:

Clipping function condition: There exists constant $\mathcal{M}_0(A) > 0$ depending on $A(\cdot)$, so that $\eta(\cdot)$ satisfies

$$\eta(\cdot) : \mathbb{R} \to \mathcal{I}_A\left(\frac{\mathcal{M}_0^2(A)}{2}\right), \text{ Lipschitz, injective.} \tag{3.2.3}$$

We now describe our data-generating model. For $i = 1, \dots n$, $y_i \in \mathcal{Y} \subset \mathbb{R}$ are independent data points with $x_i \in \mathbb{R}^{d_n}$ as the covariate, $\beta \in \mathbb{R}^{d_n}$ as the parameter of interest and β^* denoting the true parameter value. Let $\eta_i \equiv \eta\left(x_i^{\mathrm{T}}\beta\right), \eta_i^* \equiv \eta\left(x_i^{\mathrm{T}}\beta^*\right)$ and let X denote the covariate matrix or design matrix, with the vector x_i^{T} representing the i-th row of X. The sufficient statistic is $T_i \equiv T\left(y_i\right)$, the base measure by $h(y_i)$ and the density for the i-th data point is denoted by $f\left(y_i \mid \eta_i\right)$. The i-th log-partition function is denoted by $A\left(\eta_i\right)$. We denote by S^* the true model, the non-zero co-ordinates of β^*. Also, we shall denote by $\operatorname{supp}(\beta)$ the set of non-zero entries in β, and by β_S the same vector as β with the co-ordinates in S^c set to zero. $L_n(\eta, \eta^*)$ stands for the log-likelihood ratio, which is expressed in terms of its two parts; $Z_n(\eta, \eta^*)$ is the centered stochastic term and while $\mathcal{D}_n(\eta^* || \eta)$ denotes the Kullback–Leibler(KL) divergence, both based on $y^{(n)}$. Thus we have the following:

$$\begin{aligned}
& f\left(y_i \mid \eta_i\right) = h\left(y_i\right) \exp\left(T_i \eta_i - A\left(\eta_i\right)\right), \ i = 1, \dots n, \\
& \mathcal{D}_i(\eta_i^* || \eta_i) := A\left(\eta_i\right) - A\left(\eta_i^*\right) - \left(\eta_i - \eta_i^*\right) A'\left(\eta_i^*\right), \ \mathcal{D}_n(\eta^* || \eta) := \sum_{i=1}^n \mathcal{D}_i(\eta_i^* || \eta_i), \\
& Z_i(\eta_i, \eta_i^*) := \left(T_i - \mathbb{E}\, T_i\right)\left(\eta_i - \eta_i^*\right), \ Z_n(\eta, \eta^*) := \sum_{i=1}^n Z_i(\eta_i, \eta_i^*), \\
& L_n(\eta, \eta^*) := Z_n(\eta, \eta^*) - \mathcal{D}_n(\eta^* || \eta).
\end{aligned} \tag{3.2.4}$$

3.2.2 Connection of cGLM to regression settings

We briefly discuss how cGLM incorporates more commonly used high dimensional linear and non-linear regression settings. As we shall see, GLM and cGLM are interchangeable from the standpoint of practical implementation. Recall from (3.2.1) that we model the canonical parameter θ of the exponential family underlying cGLM as $\theta = \eta\left(x_i^{\mathrm{T}}\beta\right)$.

- *Linear regression with Gaussian error:* Since the native parameter, which is the mean, is the same as the canonical parameter for Gaussian, the choice of normal distribution with known variance for the exponential family in cGLM, alongside the valid choice of $\eta(t) = t,\ t \in \mathbb{R}$ as the clipping function, leads us to classical high dimensional linear regression (large d_n and small n) with Gaussian errors. Here, $\mathcal{Y} = \mathbb{R}$.

- *Logistic regression:* The native parameter here is probability of success $p \in (0, 1)$, while the canonical parameter is $\theta = \log(p/(1-p)) \in \mathbb{R}$. Thus, choosing Bernoulli for the exponential family and then, similar to linear regression, taking $\eta(t) = t,\ t \in \mathbb{R}$ as the clipping function, gives us the standard logistic regression setup. Here, $\mathcal{Y} = \{0, 1\}$.

- *Poisson regression:* Denoting the native parameter in Poisson as $\nu > 0$, we see that the canonical parameter takes the form $\theta = \log \nu \in \mathbb{R}$. Standard Poisson regression would demand of us an identity clipping function alongside the choice of Poisson for the exponential family. However, because of (3.2.3), we can allow $\eta(t) = \mathcal{C}_0 - \log\left(1 + \exp(-t + \mathcal{C}_0)\right),\ t \in \mathbb{R}$ for a large $\mathcal{C}_0 > 0$ of the practitioner's choosing. Refer to (3.1) for a graph of this clipping function when $\mathcal{C}_0 = 10$. As is clear, we are allowing $\eta\left(x_i^{\mathrm{T}}\beta\right)$ to be approximately $x_i^{\mathrm{T}}\beta$ i.e. linear, on $t < \mathcal{C}_0$, which is desired in Poisson regression, but clipping it to almost the constant value $\mathcal{C}_0$ on $t \geq \mathcal{C}_0$. Intuitively for Poisson regression, $A(t) = \exp(t), t \in \mathbb{R}$ is already very large for moderately large t, hence allowing $t = \eta\left(x_i^{\mathrm{T}}\beta\right)$ to be large for large $\|\beta\|_1$ serves no extra purpose from a modelling perspective. Our choice of $\eta(\cdot)$ reflects this, maintaining negligible difference of GLM and cGLM from implementation perspective. Here, $\mathcal{Y} = \{0, 1, \ldots\}$.

- *Negative binomial regression with known number of failures:* The native parameter here is probability of success $p \in (0,1)$, while the canonical parameter is $\theta = \log p \in (-\infty, 0)$. In contrast to the regression setups described above, standard negative binomial regression would require of us the clipping function $\eta(t) = -\log(1 + q.\exp(t))$, where $q \geq 1$ is the known number of failures, alongside choosing negative binomial for the exponential family. However, such a choice is unwarranted owing to (3.2.3). Instead, we can go with $\eta(t) = -\delta - \log(1 + \exp(-t - \delta))\,,\ t \in \mathbb{R}$, $\delta > 0$ being a pre-fixed, small constant the practitioner decides upon. Our cGLM based choice of $\eta(\cdot)$, which almost completely mimics the GLM dictated choice, appropriately reflects the presence of pole at $r_0 = 0$ for negative binomial's log-partition function $A(t) = -q\log(1 - \exp(t))\,, t < 0$. Here, $\mathcal{Y} = \{0, 1, \ldots\}$.

3.2.3 Numerically understanding clipped Poisson regression

To gain a geometrical understanding of the likelihood of cGLM, we demonstrate in the low-dimension case, i.e. $\beta \in \mathbb{R}^{d_n},\ d_n < n$, the effect of the clipping function on the Poisson GLM likelihood. As mentioned in Section (3.2.2), $\mathcal{C}_0$ can be chosen by the practitioner, but it is preferable to choose it large enough to mimic the practical properties of standard Poisson regression more closely. To describe the setup of our simulation, we first chose sample size $n = 100$, dimension $d_n = 10$ and generated covariate matrix $X : n \times d_n$ with standard Gaussian entries. Using this X, we next generated $Y \in \{0, 1, 2, \ldots\}$ from the standard Poisson GLM assuming true $\beta^* = (1, 1, \ldots 1)$ and $\beta^* = (0, 0, \ldots 0)$. Here, our target is to compare how the Maximum Likelihood Estimate (MLE) of β under the clipped Poisson model tallies with the true β^* as we vary the value of $\mathcal{C}_0$ and hence vary the likelihood:

MSE of $\hat{\beta}_{MLE}$ for $n = 100, d_n = 10$			
True β^*	$\mathcal{C}_0 = 10$	$\mathcal{C}_0 = 3$	$\mathcal{C}_0 = 0.5$
$(0, 0, \ldots 0)$	0.00973	0.013473	0.05956
$(1, 1, \ldots 1)$	0.02324	4.27605	5.79374

Table 3.1: Average MSE of $\hat{\beta}_{MLE}$ over 100 iterations

It is clear that smaller values of $\mathcal{C}_0$ results in higher Mean Squared Error (MSE) of the MLE. Intuitively, the original data generating model allowed incorporation of the effect of the parameter β growing in magnitude (ℓ_1, ℓ_2 etc.), but the clipping function *clips* this growth more and more as $\mathcal{C}_0$ decreases. The Y values that were generated from the standard Poisson GLM now dictates that the MLE from the clipped version must lie at a higher magnitude to compensate for the growth dampening of the likelihood. This phenomenon is perhaps even better demonstrated in the univariate case, i.e. $d_n = 1$. It is evident from the following plot that for fixed true β^*, we get larger MSE values for smaller values $\mathcal{C}_0$:

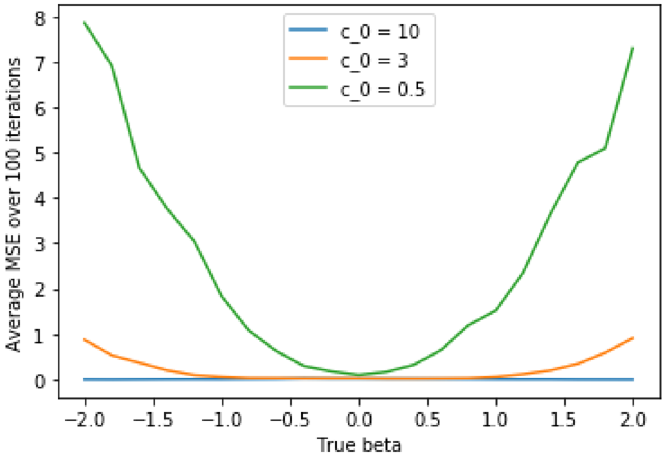

Figure 3.2: Average MSE of $\hat{\beta}_{MLE}$ using Clipped Poisson Regression

3.3 Construction of sparsity favoring prior

The sparsity favoring prior on the high dimensional β is motivated by Castillo and van der Vaart [2012], Castillo et al. [2015] and follows the construction of spike-and-slab prior proposed in the early references [Leamer, 1978, Mitchell and Beauchamp, 1988, George and McCulloch, 1995, 1997]. The crucial difference is in the slab part; we use a Laplace prior as in Castillo and van der Vaart [2012], Castillo et al. [2015] instead of the more commonly used Gaussian slab. More recently, Johnson and Rossell [2010] advocated the use of spike-and-non-local prior which has a better subset selection property compared to the spike-and-slab priors. However, our primary focus is in consistent estimation of β and a spike-and-Laplace suffices in achieving this goal.

The prior on parameter β is induced through a prior on the duo (S, β), where S denotes a subset of $\{1, \dots d_n\}$. First, the prior on the dimension $0 \leq s \leq d_n$ is chosen to be $\omega_n(s) = C_n d_n^{-a_n s}, \ s = 0, \dots, d_n$ with hyper-parameter $a_n > 0$, where C_n is chosen to normalize the distribution. For any β and S mentioned above, recall that β_S denotes the same vector β, but co-ordinates in S^c set to 0. With hyper-parameter $\lambda_n > 0$, the full prior is taken to be of the form

$$\begin{aligned}\Pi_n\left(S, \beta\right) &:= \omega_n(|S|). \binom{d_n}{|S|}^{-1} . \left(\frac{\lambda_n}{2}\right)^{|S|} . \ \exp(-\lambda_n \|\beta_S\|_1). \ \delta_0\left(\beta_{S^c}\right) \\ &= C_n . \binom{d_n}{|S|}^{-1} . \left(\frac{\lambda_n}{2 d_n^{a_n}}\right)^{|S|} . \ \exp(-\lambda_n \|\beta_S\|_1). \ \delta_0\left(\beta_{S^c}\right),\end{aligned} \tag{3.3.1}$$

where $\|.\|_1$ denotes ℓ_1-norm of Euclidean vectors, $|S|$ denotes cardinality of the set S and δ_0 denotes the degenerate distribution. The prior on the main parameter of interest, β, is given by

$$\Pi_n(\beta) := \sum_{S \subset \{1, \dots n\}} \Pi_n\left(S, \beta\right),$$

and the posterior probability of a general $B \subset \mathbb{R}^{d_n}$ is

$$\Pi_n(B \mid Y^{(n)}) := \frac{\int_B \exp[L_n(\eta, \eta^*)] \Pi_n(\beta) d\beta}{\int \exp[L_n(\eta, \eta^*)] \Pi_n(\beta) d\beta}.$$

3.3.1 Prior on model size and the non-zero coefficients

Our choice for model weights $\omega_n(\cdot)$ is special case of what is known as a complexity prior in Castillo et al. [2015]. The prior is designed to down-weight models based on their larger sizes, and weight decrease is geometric in model dimension. We thus induce sparsity in the posterior through our prior choice. We point out that there are multiple ways of specifying and generalizing the prior we have used, specifically as in Castillo and van der Vaart [2012] and Castillo et al. [2015], and they all share the central theme of exponential down-weighting of bigger models, and have the same effect on the posterior as our prior. We place independent Laplace signals for the non-zero coordinates. One can find dependent priors in the literature in this setup, for example in Castillo and van der Vaart [2012], but we choose to work with independent signals aiming to make our analysis neater.

3.4 Assumptions on data generating distribution and prior

Our assumptions on the likelihood stem from that on the KL divergence term, while assumptions about the prior come from assumptions on the hyper-parameters λ_n and a_n. These assumptions also dictate the possible values of true β^*, uniformly over which we shall state our results. In the first subsection, we present identifiability and compatibility (IC) conditions, and connect them to uniformly adaptive statements about the posterior. The second subsection is concerned with the choice of hyper-parameters that avoid any dependence of the prior on true β^*. We start by describing some order conditions, which shall help us define the rest of the assumptions.

3.4.1 Order assumptions on sample size and parameter dimension:

Since we work with a high dimensional problem, a natural condition is $d_n > n$ where $n, d_n \to \infty$. Now define a deterministic sequence of positive reals $\{b_n\}$, such that

$$b_n = o\left(\frac{n}{\log d_n}\right). \tag{3.4.1}$$

We shall focus on those true β^*'s whose sparsity s_n^* satisfies $1 \leq s_n^* \leq b_n$. This gives us, among other things, the important relation: $(s_n^* \log d_n)/n \to 0$ as $n \to \infty$. It is also important that $s_n^* \nrightarrow 0$, which first gives us $s_n^* \log d_n \to \infty$, and second, forces us to have $\log d_n = o(n)$. This shows that $b_n = O(1)$ is a valid choice, satisfying (3.4.1). We work with n large enough so that $b_n \log d_n < n$ for all our calculations. Also, note that $d_n > n$ implies $3b_n < d_n$ for large enough n.

3.4.2 Identifiability and compatibility assumptions

The ability of the log-likelihood term $L_n(\eta, \eta^*)$ to create a separation between the true value of β^* from any other β is a fundamental criterion in posterior contraction analysis, and is termed as the identifiability criterion. Again, the natural measure of discrepancy in cGLM model is the Kullback–Leibler(KL) divergence $\mathcal{D}_n(\eta^*||\eta)$, and since we work with Laplace signals in our prior, it is a natural demand to connect $\mathcal{D}_n(\eta^*||\eta)$ with the ℓ_1 distance, making them compatible. The requirements of compatibility and identifiability are simultaneously met by enforcing a lower bound on the KL divergence in terms of ℓ_1 distance between the β's i.e. $\|\beta_2 - \beta_1\|_1$ for $\beta_1, \beta_2 \in \mathbb{R}^{d_n}$. We express this through the IC (Model) and IC (Dimension) assumptions, essentially requiring existence of a model S and a dimension s, where $S \subset \{1, \dots d_n\}$ and $s = 3b_n, \dots d_n$ and they satisfy a certain lower bound property through the KL term. These assumptions not only generalize the compatibility assumptions made in Castillo et al. [2015], but also link them to identifiability of the truth.

IC (Model) Assumption: There exists at least one non-null model $S \subset \{1, \dots d_n\}$ and the corresponding quantity $\phi_1(A, X, S) > 0$, such that for any $\beta_1, \beta_2 \in \mathbb{R}^{d_n}$, we have

$$\beta_{1S} \neq \beta_{2S},\ \|\beta_{2S^c} - \beta_{1S^c}\|_1 < 7\|\beta_{2S} - \beta_{1S}\|_1 \quad \Rightarrow \quad \mathcal{D}_n(\eta_1||\eta_2) \geq \frac{n\phi_1^2(A, X, S)}{|S|}\|\beta_{2S} - \beta_{1S}\|_1^2.$$

The suffix 1 of ϕ_1 emphasizes we are working with constraints in the ℓ_1 distance, as seen above. Intuitively, a general β, that is close to the truth β^* in ℓ_1 norm, will tend to have smaller absolute values in the true noise co-ordinates S^{*c}, and hence such a β will tend to satisfy $\|\beta_{S^{*c}} - \beta^*_{S^{*c}}\|_1 < 7\|\beta_{S^*} - \beta^*_{S^*}\|_1$ or equivalently $\|\beta_{S^{*c}}\|_1 < 7\|\beta_{S^*} - \beta^*\|_1$. It is precisely in this scenario we shall need the IC (Model) assumption, i.e., $\phi_1(A, X, S^*) > 0$ so that the KL term creates a separation of the true and non-true β's that are close in ℓ_1 distance. The IC (Model) assumption will be crucially used in our proof of Theorem 2.

Now consider the following subset of the parameter space:

$$\mathcal{B}_{1,n} := \left\{\beta \in \mathbb{R}^{d_n} : \phi_1\left(A, X, \text{supp}\left(\beta\right)\right) > 0\right\}.$$

Based on the previous discussion, we would need true $\beta^* \in \mathcal{B}_{1,n}$, and due to IC(Model) assumption, $\mathcal{B}_{1,n}$ is non-null. Also, given any A and X, the quantity $\phi_1(A, X, S)$ can only take finitely many values as S varies over subsets of $\{1, \ldots d_n\}$, all of those values being positive for $S = S^*$. This gives us the quantity, for any non-null $\mathcal{B} \subset \mathcal{B}_{1,n}$,

$$\phi_{\mathcal{B}}(A, X) := \inf \Big\{ \phi_1(A, X, S^*) : \beta^* \in \mathcal{B} \Big\} > 0. \tag{3.4.2}$$

This quantity, with a special choice of $\mathcal{B}$ as laid out in the ensuing discussion, plays an important role in both Corollary 1 and Theorem 3. We now turn our attention to the IC(Dimension) assumption.

IC (Dimension) Assumption: There exists at least one $s \in \{3b_n, \ldots d_n\}$, and a corresponding quantity $\phi_0(A, X, s) > 0$, such that for any $\beta_1, \beta_2 \in \mathbb{R}^{d_n}$, we have

$$\beta_1 \neq \beta_2, |\operatorname{supp}(\beta_2 - \beta_1)| \leq s \quad \text{implies} \quad \mathcal{D}_n(\eta_1 || \eta_2) \geq \frac{n\phi_0^2(A, X, s)}{|\operatorname{supp}(\beta_2 - \beta_1)|} . \|\beta_2 - \beta_1\|_1^2.$$

The suffix 0 of ϕ_0 emphasizes we are working with constraints in the ℓ_0 distance. Similar to IC(Model), the intuition behind IC(Dimension) is to guarantee that whenever a general β matches on most of the co-ordinates with true β^*, i.e. their ℓ_0 distance is small, the KL term should be able to separate them.

The IC(Dimension) assumption, coupled with the IC(Model) assumption, form one of the central conditions in the proof of our posterior contraction statement, and we shall call it the IC (Joint) condition. First, consider the set

$$\mathcal{B}_{0,n} := \left\{ \beta \in \mathbb{R}^{d_n} : \overline{\phi}_0 \left(A, X, 3|\operatorname{supp}(\beta)| \right) > 0 \right\},$$

where, for any $s \in \{1, \dots d_n\}$,

$$\overline{\phi}_0(A, X, s) := \inf \left\{ \frac{\sqrt{s\mathcal{D}_n(\eta^* || \eta)}}{\sqrt{n}\|\beta - \beta^*\|_1} : |\operatorname{supp}(\beta - \beta^*)| \leq s, \ \beta \neq \beta^* \right\}. \tag{3.4.3}$$

Now observe that $\overline{\phi}_0(A, X, s)$ is decreasing in s, by definition, for any fixed A and X. Now, by IC(Dimension), we have $\overline{\phi}_0(A, X, 3b_n) > 0$, which shows $\overline{\phi}_0(A, X, 3|\operatorname{supp}(\beta)|) > 0$ whenever $|\operatorname{supp}(\beta)| \leq b_n$. We thus have

$$\mathcal{B}_{0,n} \supset \left\{\beta \in \mathbb{R}^{d_n} : 0 < |\operatorname{supp}(\beta)| \leq b_n\right\} =: \mathcal{B}_{2,n}, \tag{3.4.4}$$

which is a desirable relation based on the discussion at the start of this section. We are now ready to state

IC (Joint) Assumption:

$$\mathcal{B}_n := \mathcal{B}_{1,n} \cap \mathcal{B}_{2,n} \text{ is non-empty.}$$

A direct and vital consequence of this assumption is $\phi_{\mathcal{B}_n}(A, X) > 0$, as seen from (3.4.2) by choosing $\mathcal{B} = \mathcal{B}_n$. As we shall see, the statements of our results in Theorem 2 and Theorem 3 are uniformly adaptive over $\beta^* \in \mathcal{B}_n$. More precisely, our posterior contraction statement will have the form

$$\sup_{\beta^* \in \mathcal{B}_n} \mathbb{P}\left(\|\beta - \beta^*\|_1 > \varepsilon_{n,1} \mid Y^{(n)}\right) \to 0 \quad \text{as} \quad n \to \infty.$$

with $\varepsilon_{n,1} > 0$ generically denoting the optimal radius of posterior contraction.

We end this section with the pivotal role of clipping function $\eta(\cdot)$ in the IC assumptions. We require the geometries of the likelihood and the prior to match up in terms of the parameter β; $\mathcal{D}_n(\eta_1 || \eta_2)$ captures the discrepancy among β's in the likelihood, while the ℓ_1 gap does the same for the Laplace signals in the prior. To have a posterior contraction statement in ℓ_1 distance, it is necessary for the $\mathcal{D}_n(\eta_1 || \eta_2)$ to grow with $\|\beta_2 - \beta_1\|_1$, at least in sparsity restricted sense, and that is what the IC(Model) and IC(Dimension) assumptions reflect. Clipping function $\eta(\cdot)$, being an intermediary of $A(\cdot)$ and linear term $x_i^{\mathrm{T}}\beta$, must also reflect this growth, and hence has to be necessarily injective. The Lipschitz nature of $\eta(\cdot)$ allows us to translate gaps between η's to gaps between β's.

3.4.3 Hyper-parameter selection aimed at truth adaptive posterior contraction

Since we aim to avoid prior dependence on the truth, choosing the hyper-parameter λ_n, a_n should only take into account the sample size n, parameter dimension d_n, covariate matrix X and log-partition function $A(\cdot)$. Our assumptions must allow us to forgo use of any prior knowledge of the truth β^* while hyper-parameter selection. Choice of λ_n is significantly inter-twined with the log-partition function $A(\cdot)$ as well as the clipping function $\eta(\cdot)$. As in Castillo et al. [2015], λ_n needs to scale with some function of the design matrix X, and since covariate information from X is fed into the log-partition function through $\eta(\cdot)$, choice of λ_n depends on $A(\cdot), \eta(\cdot)$ and X. Based on this, consider the bound

$$\sup_{\beta^*\in\mathcal{B}_{2,n}} \max_{1\leq i\leq n} \sup \left\{ A^{''}(\gamma) : |\gamma - \eta_i^*| \leq \sqrt{\frac{s_n^*\log d_n}{n}} \right\} \leq \mathcal{M}_0^2(A),$$

which essentially gives us local control over $A''(\eta_i)\ \forall\ i = 1, \ldots n$, uniformly over $\beta^* \in \mathbb{R}^{d_n}$. The proof of this statement is detailed in Lemma 3 in the appendix, and basically uses two main points. Firstly, since $s_n^* \leq b_n$ for $\beta^* \in \mathcal{B}_n$, we have $(s_n^* \log d_n)/n \to 0$ by (3.4.1), which allows us to have shrinking neighborhoods around every η_i^*. Secondly, based on the behavior of $A''(\cdot)$, the clipping function $\eta(\cdot)$ restricts the set of arguments passed to $A(\cdot)$, thus controlling the growth of $A''(\cdot)$.

Now, define the quantities

$$\begin{aligned} \mathcal{M}_1(A) &:= \left(1 \wedge \mathcal{M}_0^{-1}(A)\right)^{-1}, \|X\|_{(\infty,\infty)} := \max\left\{X_{i.j} : i = 1, \ldots n, j = 1. \ldots . d_n\right\}, \\ \mathcal{M}(A, X) &:= \|X\|_{(\infty,\infty)}\mathcal{M}_1(A). \end{aligned} \tag{3.4.5}$$

We can now state our assumption on the hyper-parameter λ_n:

Assumption $\mathcal{L}_0$:

$$\frac{\mathcal{M}(A,X)}{d_n} \leq \lambda_n \leq \mathcal{M}(A,X)\sqrt{\log d_n}.$$

This bound, which we utilize in all our Theorems, generalizes the hyper-parameter bounds mentioned in Castillo et al. [2015], as well as avoids prior dependence on the truth. Existence of $\mathcal{M}_0(A) > 0$, through which $\mathcal{M}(A,X)$ is defined in (3.4.5), is guaranteed by (3.2.3), and it acts as a pre-fixed constant quantity that the practitioner can choose based solely on $A(\cdot)$, and then choose clipping function $\eta(\cdot)$. This, in turn, shows that the choice of hyper-parameter λ_n depends solely on the three quantities (A, X, b_n). This makes our hyper-parameter choice of λ_n free of the truth.

We turn our attention to hyper-parameter a_n, which controls how fast the the model weights $\omega_n(\cdot)$ decay. First, define

$$\mathcal{E}_1 := 8\left(1 + \frac{49\mathcal{M}^2(A,X)}{8\phi^2_{\mathcal{B}_n}(A,X)}\right), \tag{3.4.6}$$

which is an adaptive choice, as well as free of any knowledge of the truth, owing to (3.4.2) and IC (Joint) assumption. For mild demands, like in Theorem 2, $a_n > 1$ suffices. On the contrary, for the weak model selection result in Corollary 1, we need to choose a_n that supports very strong down-weighting of larger models, namely $a_n \geq 1 + 2b_n\mathcal{E}_1$. One can note from (3.4.1) why this choice of a_n heaviliy penalizes larger models. Lastly, the choice $a_n \geq 1 + \mathcal{E}_1$, which is much milder than our previous choice, is sufficient for the posterior contraction result in Theorem 3. It is crucial to note that just like λ_n, our choice of hyper-parameter a_n avoids any knowledge of true β^*.

3.5 Adaptive rate-optimal posterior contraction in ℓ_1 norm

In this section, we provide the statements of our results and lay out sketches of how we arrive at them, putting under spotlight the use of the assumptions.

3.5.1 Lower bound of the marginal likelihood

Starting from (3.2.4), the marginal likelihood is defined as

$$\int \exp\left(L_n(\eta, \eta^*)\right) \Pi_n(\beta) d\beta, \tag{3.5.1}$$

which appears as the denominator in calculating the posterior through Bayes' Theorem. Theorem 1 provides a high probability lower bound to this quantity in terms of the parameter dimension d_n and the true model size s_n^*.

Theorem 1. *Let $a_n > 0$ and λ_n satisfy assumption $\mathcal{L}_0$. Let $n, d_n \to \infty$ and $d_n > n$. Based on* (3.4.1)*, consider large enough n so that $b_n \log d_n < n$. Let the true β^* belong to $\mathcal{B}_{2,n}$ as in* (3.4.4)*. Then, with probability $1 - (s_n^* \log d_n)^{-1}$ with respect to the data generating distribution, the marginal likelihood defined in* (3.5.1) *satisfies for all sufficiently large n*

$$\int \exp\left(L_n(\eta, \eta^*)\right) \Pi_n(\beta) d\beta \gtrsim C_n d_n^{-(a_n+6)s_n^*} \exp(-\lambda_n \|\beta^*\|_1).$$

The fact that $s_n^* \log d_n \to \infty$ as $n \to \infty$ makes this a high probability statement about the marginal likelihood. Since majority of the mass under the integral should lie around the truth β^*, it is natural that the lower bound should contain information about that truth. Theorem 1 quantifies that relation. One generic tool for reaching such a bound has been described in Ghosal et al. [2000], which we modify to suit our needs. We provide a small sketch of our method here, while the full proof is given in the appendix.

Consider the set

$$D_n := \left\{\beta \in \mathbb{R}^{d_n} : \left[-\mathbb{E}\left[L_n(\eta, \eta^*)\right] \bigvee \text{Var}\left[L_n(\eta, \eta^*)\right] \right] \leq s_n^* \log d_n \right\}, \tag{3.5.2}$$

noting that $-\mathbb{E}\left[L_n(\eta, \eta^*)\right] = \mathcal{D}_n(\eta^* || \eta)$ and $\text{Var}\left[L_n(\eta, \eta^*)\right] = \mathbb{E}\, Z_n^2(\eta, \eta^*)$. Let $\Pi_{D_n}(\beta)$ denote the restriction of prior $\Pi_n(\beta)$ to D_n. Then the denominator satisfies

$$\int \exp\left(L_n(\eta, \eta^*)\right) \Pi_n(\beta) d\beta \geq \int_{D_n} \exp\left(L_n(\eta, \eta^*)\right) \Pi_n(\beta) d\beta$$

$$= \Pi_n\left(D_n\right) \int \exp\left(L_n(\eta, \eta^*)\right) \Pi_{D_n}(\beta) d\beta.$$

This method of restricting the integral of the marginal likelihood to a neighborhood of the truth is reminiscent of the original method found in Ghosal et al. [2000]. The radius of such a neighborhood, here given by $s_n^* \log d_n$, signifies the order of allowable growth in both the expectation and variance of the log-likelihood ratio. We now have two terms to deal with, the prior probability of D_n and the restricted integral. First, we use the variance of $L_n(\eta, \eta^*)$ in a Chebyshev inequality to obtain the lower bound

$$\int \exp\left(L_n(\eta, \eta^*)\right) \Pi_{D_n}(\beta) d\beta \geq d_n^{-2s_n^*}. \tag{3.5.3}$$

with high probability. Next, as detailed in Lemma 1 in the appendix, we bound from below the prior probability of D_n as

$$\Pi_n(D_n) \gtrsim C_n \exp(-\lambda_n \|\beta^*\|_1) d_n^{-(a_n+4)s_n^*}.$$

Theorem 1 now follows by combining Lemma 1 with (3.5.3).

3.5.2 Posterior dimension and weak model selection

We work with complexity priors that put increasingly higher penalty, or lower weight, on models that have larger sizes. It is expected that the posterior would reflect this prior property, which is tantamount to the posterior having vanishingly low probability of exceeding a certain dimension. Theorem 2 does exactly that, showing that the posterior should be at least as sparse as the true β^*, up to multiplicative constants. Sparsity is quantified using $|\operatorname{supp}(\beta)|$ and is compared with s_n^*, the true level of sparsity in β^*.

Theorem 2. *Let $a_n > 1$ and λ_n satisfy assumption $\mathcal{L}_0$. Let $n, d_n \to \infty$ and $d_n > n$. Based on*

(3.4.1)*, consider large enough n so that $b_n \log d_n < n$. Let assumptions IC(Model) and IC (Joint) hold, and consider the non-null set $\mathcal{B}_n$. Then, with quantity $\phi_1(A, X, S)$ given by IC(Model), and $\mathcal{M}(A, X)$ as in* (3.4.5)*, we have for all sufficiently large n,*

$$\sup_{\beta^* \in \mathcal{B}_n} \mathbb{E}\left[\Pi_n\left(|\operatorname{supp}(\beta)| > s_n^*\left[1 + \frac{8}{a_n - 1}\left(1 + \frac{49\mathcal{M}^2(A, X)}{8\phi_1^2(A, X, S^*)}\right)\right]\middle| Y^{(n)}\right)\right] \to 0 \quad \text{as} \quad n \to \infty.$$

The statement of the theorem is presented in an asymptotic fashion, but is true for every n large enough, satisfying the order assumptions. For simplicity, let us define the quantity $\mathcal{E}_1^* := 8\left(1 + 49\mathcal{M}^2(A, X)/8\phi_1^2(A, X, S^*)\right)$ so that Theorem 2 is a statement about the posterior probability of the set $\{|\operatorname{supp}(\beta)| > s_n^*\left(1 + \mathcal{E}_1^*/(a_n - 1)\right)\}$. It is important to note that we have used $\beta^* \in \mathcal{B}_n$ implies $\phi_1(A, X, S^*) > 0$. Owing to IC (Joint), (3.4.2) and the choice $\mathcal{B} = \mathcal{B}_n$, we can have from Theorem 2,

$$\sup_{\beta^* \in \mathcal{B}_n} \mathbb{E}\left[\Pi_n\left(|\operatorname{supp}(\beta)| > s_n^*\left[1 + \frac{8}{a_n - 1}\left(1 + \frac{49\mathcal{M}^2(A, X)}{8\phi_{\mathcal{B}_n}^2(A, X)}\right)\right]\middle| Y^{(n)}\right)\right] \to 0 \quad \text{as} \quad n \to \infty.$$

By the definition of $\mathcal{E}_1$ in (3.4.6) and its analogy with $\mathcal{E}_1^*$, we now work with the posterior probability of $\{|\operatorname{supp}(\beta)| > s_n^*\left(1 + \mathcal{E}_1/(a_n - 1)\right)\}$. This allows to us to choose the hyper-parameter a_n as $a_n \geq 1 + 2b_n\mathcal{E}_1$, which is a truth-free choice, and leads to the following corollary:

Corollary 1. *With $\mathcal{E}_1$ as in* (3.4.6)*, if hyper-parameter a_n in the prior satisfies $a_n \geq 1 + 2b_n\mathcal{E}_1$ in addition to the hypotheses of Theorem 2, we have*

$$\sup_{\beta^* \in \mathcal{B}_n} \mathbb{E}\left[\Pi_n\left(\operatorname{supp}(\beta) \supsetneqq S^* \middle| Y^{(n)}\right)\right] \to 0 \quad \text{as} \quad n \to \infty.$$

This statement is a straightforward consequence of Theorem 2, the fact that $s_n^* \leq b_n$ for $\beta^* \in \mathcal{B}_n$, and the observation that $\{\operatorname{supp}(\beta) \supsetneqq S^*\} \subset \{|\operatorname{supp}(\beta)| > s_n^* + 1/2\}$. Thus, Corollary 1 is a weak statement on model selection consistency. It ensures vanishingly small posterior probability attached to models that are strict super sets of the true model S^*.

3.5.3 Truth adaptive posterior contraction in ℓ_1 metric

We now turn our attention to the central result of our article, which is a truth adaptive statement about ℓ_1-contraction of the posterior distribution. Essentially, it gives the radius of the smallest possible ℓ_1 ball around true β^*, whose posterior probability vanishes with large n. Define the quantity

$$\mathcal{E}_2 := 6 + \frac{12\mathcal{M}^2(A, X)}{\overline{\phi}_0^2(A, X, 3b_n)}, \tag{3.5.4}$$

which can be observed to be truth-free. By describing the aforementioned radius in terms of $a_n, \mathcal{E}_2, d_n$ and n, we have the following.

Theorem 3. *Let hyper-parameter a_n satisfy $a_n \geq 1 + \mathcal{E}_1$ for $\mathcal{E}_1$ as in* (3.5.4)*, and hyper-parameter λ_n satisfy assumption $\mathcal{L}_0$. Let $n, d_n \to \infty$ and $d_n > n$. Based on* (3.4.1)*, consider large enough n so that $b_n \log d_n < n$. Let assumptions IC(Model), IC(Dimension) and IC (Joint) hold, and consider the non-null set $\mathcal{B}_n$. Then, with quantity $\mathcal{E}_2$ given by* (3.5.4) *and $\mathcal{M}(A, X)$ as in* (3.4.5)*, we have for all sufficiently large n,*

$$\sup_{\beta^* \in \mathcal{B}_n} \mathbb{E}\left[\Pi_n\left(\|\beta - \beta^*\|_1 > \frac{2s_n^*(1 + a_n + \mathcal{E}_2)}{\mathcal{M}(A, X)}\sqrt{\frac{\log d_n}{n}} \,\Big|\, Y^{(n)}\right)\right] \to 0. \tag{3.5.5}$$

It is important to note that the contraction rate linearly increases with a_n and as long as a_n is chosen to be a constant larger than $1 + \mathcal{E}_1$, the rate is unaffected. However, if one chooses a stronger penalty on the model space to achieve weak model selection consistency as in Corollary 1, the rate of contraction in ℓ_1 norm becomes slower unless the upper bound b_n on the number of true non-zero coefficients is assumed to be a constant.

3.6 Conclusion

To summarize, we introduced a new family of GLMs and developed sufficient conditions for obtaining posterior contraction rates that are adaptive rate-optimal. From an implementation point of view, the new family does not bring in additional challenges, but from a theoretical point of view, it allows us to obtain adaptivity, while simultaneously obviating the need to enforce growth restriction on the true coefficient vector. Our analysis is restricted to the use of Laplace prior on the regression coefficients primarily due to the clarity and ease of calculations. More general priors, including compactly supported distributions, heavier tailed family or non-local priors can be considered. As a topic of immediate future research, strong model selection consistency is deemed important. As already demonstrated in Theorem 2, the posterior does not concentrate on subsets which are larger than the true subset with a stronger complexity prior. With more identifiability conditions, one can ensure that the posterior does not concentrate on subsets that miss one or more non-zero true coordinates, thereby ensuring strong model selection consistency.

4. SUMMARY AND FUTURE DIRECTIONS

4.1 Summary

For the first project, we provide statistical guarantees for Bayesian variational boosting by proposing a novel small bandwidth Gaussian mixture variational family. We employ a functional version of Frank-Wolfe optimization as our variational algorithm and study frequentist properties of the iterative boosting updates. Comparisons are drawn to the recent literature on boosting, describing how the choice of the variational family and the discrepancy measure affect both convergence and finite-sample statistical properties of the optimization routine. Specifically, we first demonstrate stochastic boundedness of the boosting iterates with respect to the data generating distribution. We next integrate this within our algorithm to provide an explicit convergence rate, ending with a result on the required number of boosting updates.

For the second project, we develop a framework to study posterior contraction rates in sparse high dimensional generalized linear models (GLM). We introduce a new family of GLMs, denoted by clipped GLM, which subsumes many standard GLMs and makes minor modification of the rest. With a sparsity inducing prior on the regression coefficients, we delineate sufficient conditions on true data generating density that leads to minimax optimal rates of posterior contraction of the coefficients in ℓ_1 norm. Our key contribution is to develop sufficient conditions commensurate with the geometry of the clipped GLM family, propose prior distributions which do not require any knowledge of the true parameters and avoid any assumption on the growth rate of the true coefficient vector.

4.2 Current and future work

I aim to build a research career aimed at bridging the gap between cutting-edge ML/AI tools and their applicability in heavily data-driven fields like digital medicine, automated market making and computer vision. I shall conclude with a brief a list of topics that I believe lie well inside my radar of interest and attainable expertise:

- Reinforcement learning techniques for sequential data analysis, dynamic treatment regimes
- Using Belief propagation and its connection with variational algorithms in the context of marginal approximation for intractable posteriors, application in singular models.
- Extending variational families using Copulas, studying their posterior convergence properties and algorithmic issues
- Theoretical frameworks for Variational Auto-encoders, Generative Adversarial Networks in Deep learning, etc.
- Extending existing high-performance classifiers and clustering techniques to their scalable counter-parts.

REFERENCES

Felix Abramovich and Vadim Grinshtein. Model selection and minimax estimation in generalized linear models. *IEEE Transactions on Information Theory*, 62(6):3721–3730, 2016.

Sergios Agapiou, Masoumeh Dashti, and Tapio Helin. Rates of contraction of posterior distributions based on p-exponential priors. *Bernoulli*, 27(3):1616–1642, 2021.

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *arXiv preprint arXiv:1706.09293*, 2017.

Michael R Andersen, Ole Winther, and Lars K Hansen. Bayesian inference for structured spike and slab priors. *Advances in Neural Information Processing Systems*, 27:1745–1753, 2014.

Y.A. Atchadé. On the contraction properties of some high-dimensional quasi-posterior distributions. *The Annals of Statistics*, 45(5):2248–2273, 2017.

Andrew R Barron. *The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions*. Department of Statistics, University of Illinois Champaign, IL, 1988.

Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Anirban Bhattacharya, Debdeep Pati, and David Dunson. Anisotropic function estimation using multi-bandwidth gaussian processes. *Annals of statistics*, 42(1):352, 2014.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Stephen P Brooks and Gareth O Roberts. Assessing convergence of markov chain monte carlo algorithms. *Statistics and Computing*, 8(4):319–335, 1998.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Trevor Campbell and Xinglong Li. Universal boosting variational inference. *arXiv preprint arXiv:1906.01235*, 2019.

Emmanuel Candes, Terence Tao, et al. The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007.

I. Castillo and A.W. van der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101, 2012.

Ismaël Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

Taeryon Choi, RV Ramamoorthi, et al. Remarks on consistency of posterior distributions. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pages 170–186. Institute of Mathematical Statistics, 2008.

P Damlen, John Wakefield, and Stephen Walker. Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):331–344, 1999.

Dipak K Dey, Sujit K Ghosh, and Bani K Mallick. *Generalized linear models: A Bayesian perspective*. CRC Press, 2000.

Persi Diaconis and David Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

David L Donoho, Iain M Johnstone, Jeffrey C Hoch, and Alan S Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1): 41–67, 1992.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

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

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Chao Gao, Aad W van der Vaart, and Harrison H Zhou. A general framework for Bayes structured linear models. *The Annals of Statistucs*, 2015. (to appear).

Edward I George and Robert E McCulloch. Stochastic search variable selection. *Markov chain Monte Carlo in practice*, 68:203–214, 1995.

Edward I George and Robert E McCulloch. Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.

S. Ghosal. Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *Journal of Multivariate Analysis*, 74(1):49–68, 2000.

S. Ghosal and A.W. van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223, 2007.

Subhashis Ghosal, Jayanta K Ghosh, Tapas Samanta, et al. On convergence of posterior distributions. *The Annals of Statistics*, 23(6):2145–2152, 1995.

Subhashis Ghosal, Jayanta K Ghosh, and RV Ramamoorthi. Posterior consistency of dirichlet mixtures in density estimation. *Annals of Statistics*, pages 143–158, 1999.

Subhashis Ghosal, Jayanta K Ghosh, Aad W Van Der Vaart, et al. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

Marvin HJ Gruber. *Improving efficiency by shrinkage: the James-Stein and ridge regression estimators*. Routledge, 2017.

Biraj Subhra Guha and Debdeep Pati. Adaptive posterior convergence in sparse high dimensional clipped generalized linear models. *arXiv preprint arXiv:2103.08092*, 2021.

Biraj Subhra Guha, Anirban Bhattacharya, and Debdeep Pati. Statistical guarantees and algorithmic convergence issues of variational boosting. *arXiv preprint arXiv:2010.09540*, 2020.

Fangjian Guo, Xiangyu Wang, Kai Fan, Tamara Broderick, and David B Dunson. Boosting variational inference. *arXiv preprint arXiv:1611.05559*, 2016.

Shaobo Han, Xuejun Liao, David Dunson, and Lawrence Carin. Variational gaussian copula inference. In *Artificial Intelligence and Statistics*, pages 829–838, 2016.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2019.

Bruce M Hill. A simple general approach to inference about the tail of a distribution. *The annals of statistics*, pages 1163–1174, 1975.

Peter D Hoff. *A first course in Bayesian statistical methods*, volume 580. Springer, 2009.

Yuao Hu. On the convergence of bayesian regression models. *arXiv preprint arXiv:1010.1049*, 2010.

Jonathan Huggins, Mikolaj Kasprzak, Trevor Campbell, and Tamara Broderick. Validated variational inference via practical posterior error bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 1792–1802. PMLR, 2020.

Jonathan H Huggins, Trevor Campbell, Mikołaj Kasprzak, and Tamara Broderick. Practical bounds on the error of bayesian posterior approximations: A nonasymptotic approach. *arXiv preprint arXiv:1809.09505*, 2018.

Ferenc Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

Hemant Ishwaran, J Sunil Rao, et al. Spike and slab variable selection: frequentist and bayesian strategies. *Annals of statistics*, 33(2):730–773, 2005.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

Wenxin Jiang et al. Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. *The Annals of Statistics*, 35(4):1487–1511, 2007.

Valen E Johnson and David Rossell. On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170, 2010.

Bernard Osgood Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409, 1936.

Willem Kruijer, Judith Rousseau, Aad Van Der Vaart, et al. Adaptive bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.

Edward E Leamer. Regression selection strategies and revealed priors. *Journal of the American Statistical Association*, 73(363):580–587, 1978.

Francesco Locatello, Rajiv Khanna, Joydeep Ghosh, and Gunnar Rätsch. Boosting variational inference: an optimization perspective. *arXiv preprint arXiv:1708.01733*, 2017.

Francesco Locatello, Gideon Dresdner, Rajiv Khanna, Isabel Valera, and Gunnar Rätsch. Boosting black box variational inference. In *Advances in Neural Information Processing Systems*, pages 3401–3411, 2018.

Gertraud Malsiner-Walli and Helga Wagner. Comparing spike and slab priors for bayesian variable selection. *arXiv preprint arXiv:1812.07259*, 2018.

Peter McCullagh. *Generalized linear models*. Routledge, 2018.

Peter McCullagh. Generalized linear models. 2019.

Andrew C Miller, Nicholas J Foti, and Ryan P Adams. Variational boosting: Iteratively refining posterior approximations. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2420–2429. JMLR. org, 2017.

Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

Dmitry Molchanov, Valery Kharitonov, Artem Sobolev, and Dmitry Vetrov. Doubly semi-implicit variational inference. *arXiv preprint arXiv:1810.02789*, 2018.

Soumendu Sundar Mukherjee, Purnamrita Sarkar, YX Rachel Wang, and Bowei Yan. Mean field for the stochastic blockmodel: optimization landscape and convergence issues. In *Advances in Neural Information Processing Systems*, pages 10694–10704, 2018.

Naveen Naidu Narisetty, Xuming He, et al. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817, 2014.

Radford M Neal. *Probabilistic inference using Markov chain Monte Carlo methods*. Department of Computer Science, University of Toronto Toronto, Ontario, Canada, 1993.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Robert B O'Hara, Mikko J Sillanpää, et al. A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117, 2009.

Debdeep Pati, Anirban Bhattacharya, and Yun Yang. On statistical optimality of variational bayes. In *International Conference on Artificial Intelligence and Statistics*, pages 1579–1588, 2018.

Edwin James George Pitman. Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the cambridge Philosophical society*, volume 32, pages 567–579. Cambridge University Press, 1936.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

Gareth O Roberts, Jeffrey S Rosenthal, et al. General state space markov chains and mcmc algorithms. *Probability surveys*, 1:20–71, 2004.

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

Mark J Schervish. Decision theory. In *Theory of Statistics*, pages 144–213. Springer, 1995.

Robert Schlaifer and Howard Raiffa. *Applied statistical decision theory*. 1961.

James G Scott, James O Berger, et al. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619, 2010.

Weining Shen, Surya T Tokdar, and Subhashis Ghosal. Adaptive bayesian multivariate density estimation with dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.

Xiaotong Shen, Larry Wasserman, et al. Rates of convergence of posterior distributions. *Annals of Statistics*, 29(3):687–714, 2001.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. Kernel implicit variational inference. *arXiv preprint arXiv:1705.10119*, 2017.

Stephen M Stigler. Darwin, galton and the statistical enlightenment. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(3):469–482, 2010.

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

Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

Michalis K Titsias and Francisco JR Ruiz. Unbiased implicit variational inference. *arXiv preprint arXiv:1808.02078*, 2018.

Surya T Tokdar. Posterior consistency of dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, pages 90–110, 2006.

Dustin Tran, David Blei, and Edo M Airoldi. Copula variational inference. In *Advances in Neural Information Processing Systems*, pages 3564–3572, 2015.

Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

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

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Aad W van der Vaart, J Harry van Zanten, et al. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Bo Wang, DM Titterington, et al. Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650, 2006.

Xiangyu Wang. *Boosting variational inference: theory and examples*. PhD thesis, PhD thesis, Duke

University, 2016.

Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, pages 1–15, 2018.

Ran Wei and Subhashis Ghosal. Contraction properties of shrinkage priors in logistic regression. *Journal of Statistical Planning and Inference*, 2020.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. *arXiv preprint arXiv:1802.05438*, 2018.

Yun Yang, Surya T Tokdar, et al. Minimax-optimal nonparametric regression in high dimensions. *Annals of Statistics*, 43(2):652–674, 2015.

Yun Yang, Martin J Wainwright, Michael I Jordan, et al. On the computational complexity of high-dimensional bayesian variable selection. *The Annals of Statistics*, 44(6):2497–2532, 2016.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *arXiv preprint arXiv:1710.03266*, 2017.

Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. *arXiv preprint arXiv:1805.11183*, 2018.

Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *arXiv preprint arXiv:1712.02519*, 2017.

Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

APPENDIX A

SUPPLEMENTARY MATERIAL TO FIRST CHAPTER

A.1 Revisiting Frank–Wolfe Algorithm

Algorithm 2 Frank–Wolfe algorithm with approximate Linear Minimization Oracle:

```
1. Initialize with $x^{(0)} \in D$.
2. For the $k$-th step, set $\gamma_k = 2/(k+2)$ and calculate sub-gradient $\nabla f(x^{(k)})$.
3. Solve the linear minimization oracle (LMO) approximately, i.e. find $y^{(k+1)} \in D$ such that
   $\langle \nabla f(x^{(k)}), y^{(k+1)} \rangle \leq \min \left\{ y \in D \mid \langle \nabla f(x^{(k)}), y \rangle \right\} + \gamma_k \mathcal{C}_{f,D}/2$.
4. Update $x^{(k+1)} = (1-\gamma_k)x^{(k)} + \gamma_k y^{(k+1)}$.
```

We note down the basics of Frank–Wolfe algorithm in this section. The reader is referred to Jaggi [2013] for further details and to Frank and Wolfe [1956] for the original formulation. We start by reviewing the notation. In what follows, Y is an inner product space with $\langle y_1, y_2 \rangle$ denoting the inner product of $y_1, y_2 \in Y$ and $\|y\| = \langle y, y \rangle$ the norm induced by the inner product. D shall denote a compact, convex subset of Y, which is our optimization domain. We shall work with a convex function f defined on D, which is our objective function for the optimization routine. We start by revising the notion of a subgradient.

Definition 1: A sub-gradient of f at $x \in D$, denoted by $\nabla f(x)$, is a member of $\partial f(x) \subset D$, given by

$$\partial f(x) = \left\{ y \in D \mid f(z) - f(x) - \langle y, z - x \rangle \geq 0, \ \forall \, z \in D \right\}.$$

It is easy to note that, if $Y = \mathbb{R}^d$, and convex function f is differentiable at $x \in Y$, then the gradient at x satisfies $f'(x) \in \partial f(x)$. Subgradients are useful when the notion of differentiability is untenable. Next, we note down the definition of Bregman divergence.

Definition 2: For any $x, y \in D$, the Bregman divergence of y from x under function f is defined as

$$\mathcal{D}_f(y||x) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle. \tag{A.1.1}$$

With this definition in hand, we now define the curvature of f on domain D.

Definition 3: The curvature $\mathcal{C}_{f,D}$ of f on domain D is given by

$$\mathcal{C}_{f,D} = \sup \left\{ \frac{2}{\alpha^2} \mathcal{D}_f(x_2||x_1) : y, x_1 \in D, \alpha \in [0,1], x_2 = x_1 + \alpha(y - x_1) \right\}. \tag{A.1.2}$$

One can think of the curvature as the maximum scaled Bregman divergence between points in D and their perturbations through mixtures. We now recall the definition and significance of strong smoothness and strong convexity of f.

Definition 4: If for any $x, y \in D$ and some $C_1, C_2 > 0$ (possibly depending on f and D)

1. $\mathcal{D}_f(y||x) \leq C_1 \|y - x\|^2$, then f is strongly smooth on D,

2. $\mathcal{D}_f(y||x) \geq C_2 \|y - x\|^2$, then f is strongly convex on D.

Convex functions f on D that satisfy strong smoothness allow calculations of rate of convergence. The most basic Frank–Wolfe algorithm minimizes convex function f, defined on domain D. We note down a version with approximately solved subproblem, following Jaggi [2013].

Let x^* denote the minimum point in domain D. The above algorithm, by theorem 1 in Jaggi [2013], satisfies

$$f(x^{(k)}) - f(x^*) \leq 4C_{f,D}/(k+2) \tag{A.1.3}$$

This gives us the rate of convergence of this algorithm, in terms of the curvature $\mathcal{C}_{f,D}$. Note that such a rate of convergence with respect of number of iterations k is called sub-linear. For statistical problems, $\mathcal{C}_{f,D}$ is typically a function of the sample size and the parameter dimension, and can be quite large for densities having non-compact support.

A.2 Proofs

Proof of Theorem 1

In what follows, $\gtrsim, \lesssim$ respectively stand for greater than and less than up to an absolute constant and $s_{\max}(A)$ denotes the highest singular value of square matrix A. Let q_0 denote the d-dimensional Gaussian density, centered at the truth θ_0, and variance $\sigma_n^2 I_d$, where σ_n satisfies assumption 2. Along with assumption 1, we have $q_0 \in \mathcal{Q}_n$ and hence $m_n^*(\mathcal{Q}_n) \leq KL(q_0||\pi_n)$, so that it is enough to show $KL(q_0||\pi_n)$ is bounded in probability. We decompose $KL(q_0||\pi_n)$ as

$$\begin{aligned}\int q_0(\theta)\log\frac{q_0(\theta)}{\pi_n(\theta)}d\theta =& -d\left[\log(\sqrt{2\pi})+\frac{1}{2}\right]+\log\left(m(X_n)\right)\\ &-\left(\int L_n(\theta,\theta_0)q_0(\theta)d\theta\right)+\int U(\theta)q_0(\theta)d\theta.\end{aligned} \tag{A.2.1}$$

Since sum of $O_p(1)$ quantities are again $O_p(1)$, we can stochastically bound (A.2.1) term by term. The first and last terms are already constants. We will also prove $\int U(\theta)q_0(\theta)d\theta \lesssim U(\theta_0)$. So, it is enough to show $\log\left(m(X_n)\right)$ and $\int L_n(\theta,\theta_0)q_0(\theta)d\theta$ are stochastically bounded from above and below, respectively. We have

$$\text{pr}\left(\log\left(m(X_n)\right) > -\log\epsilon\right) = \text{pr}\left(\left(m(X_n)\right) > \frac{1}{\epsilon}\right) \leq \epsilon.E\left(m(X_n)\right) = \epsilon \quad \forall n. \tag{A.2.2}$$

We now employ Taylor expansion around θ_0. Using assumption 4 for $U(\theta)$, observe that

$$\begin{aligned}\int q_0(\theta)U(\theta)d\theta \lesssim\, & U(\theta_0)+s_{max}^2\left(U^{(2)}(\theta_0)\right)\left(\int\|\theta-\theta_0\|^2 q_0(\theta)d\theta\right)\\ &+\left(\int\|\theta-\theta_0\|_2^{2+2\alpha_3}q_0(\theta)d\theta\right)\lesssim U(\theta_0),\end{aligned} \tag{A.2.3}$$

and for $\mu_2(\theta_0||\theta)$, observe that

$$\begin{aligned}\int q_0(\theta)\mu_2(\theta_0||\theta)d\theta \leq & s_{max}^2\left(\mu_2^{(2)}(\theta_0||\theta_0)\right)\left(\int \|\theta-\theta_0\|^2 q_0(\theta)d\theta\right) \\ & + \left(\int \|\theta-\theta_0\|_2^{2+2\alpha_2} q_0(\theta)d\theta\right) \leq C_1\sigma_n^2,\end{aligned} \tag{A.2.4}$$

for constant $C_1 > 0$. Again, by assumption 4 for $KL(\theta_0||\theta)$, we get the upper bound (similar to previous step)

$$\int q_0(\theta)KL(\theta_0||\theta)d\theta \leq C_u\sigma_n^2, \tag{A.2.5}$$

and using assumption 5, we get the lower bound

$$C_l\sigma_n^2 \leq \int q_0(\theta)KL(\theta_0||\theta)d\theta, \tag{A.2.6}$$

where $0 < C_l < C_u$ are absolute constants. Now, for $\delta > 0$ to be chosen later, we have

$$\begin{aligned}&\mathrm{pr}\left[\int L_n(\theta,\theta_0)q_0(\theta)d\theta \leq -C_u(1+\delta)n\sigma_n^2\right] \\ &\leq \mathrm{pr}\left[\int L_n(\theta,\theta_0)q_0(\theta)d\theta \leq -(1+\delta)n\int KL(\theta_0||\theta)q_0(\theta)d\theta\right] \\ &\leq \mathrm{pr}\left[\int \frac{1}{\sqrt{n}}\zeta_n(\theta,\theta_0)q_0(\theta)d\theta \leq -\delta\sqrt{n}\int KL(\theta_0||\theta)q_0(\theta)d\theta\right] \\ &\leq \frac{E\left(\int l(\theta,\theta_0)q_0(\theta)d\theta\right)^2}{\delta^2 n\left(\int KL(\theta_0||\theta)q_0(\theta)d\theta\right)^2} \leq \frac{\int q_0(\theta)\mu_2(\theta_0||\theta)d\theta}{\delta^2 n\left(\int KL(\theta_0||\theta)q_0(\theta)d\theta\right)^2} \leq \frac{C_1\sigma_n^2}{\delta^2 C_l^2 n\sigma_n^4} = \frac{C_1}{\delta^2 C_l n\sigma_n^2}.\end{aligned} \tag{A.2.7}$$

From assumption 2, we have $c_0^{-1/2}n^{-1/2} \leq \sigma_n \leq n^{-1/2}$ and hence, given $\epsilon > 0$, we can choose $\delta := \left(C_1c_0^{1/2}/\epsilon C_l\right)^{1/2}$ to get

$$\mathrm{pr}\left[\int L_n(\theta,\theta_0)q_0(\theta)d\theta \leq -C_u\left(1+\left(C_1c_0^{1/2}/\epsilon C_l\right)^{1/2}\right)\right] \leq \epsilon, \tag{A.2.8}$$

and the stochastic boundedness result is complete.

Proof of Corollary 1

For this specific family of densities, let $\nu_l = E\left(T_l|\theta_0\right),\ l = 1 \ldots K$ denote the expectations of the sufficient statistics, and let σ_{ℓ_1,l_2} denote $Cov(T_{\ell_1}, T_{l_2}|\theta_0)$ for $\ell_1, l_2 = 1, \ldots K$. Let $\nu_0 = (\nu_1, \ldots \nu_K)$ and $\Sigma_0 = ((\sigma_{\ell_1,l_2}))_{\ell_1,l_2}$, so that for the K-vector $T = (T_1, \ldots T_K)$, $E(T|\theta_0) = \nu_0$ and $Cov(T|\theta_0) = \Sigma_0$. Direct calculations shows

$$KL(\theta_0||\theta) = A(\theta) - A(\theta_0) - (\theta - \theta_0)^T \nu_0, \tag{A.2.9}$$

and we know that for Exponential families, $\nu_0 = A^{(1)}(\theta_0)$. Thus (A.2.9) becomes (see Definition 2)

$$KL(\theta_0||\theta) = D_A(\theta_0||\theta). \tag{A.2.10}$$

It is worthwhile to observe the analogy to lemma 2. $KL(\theta_0||\theta)$ is finite for all $\theta \in \Theta$ as $D_A(\theta_0||\theta)$ is. Now by hypothesis, $A(\theta)$ is strongly convex, and hence by part 2 of Definition 4 and (A.2.9), we have $KL(\theta_0||\theta) \geq C\|\theta - \theta_0\|_2^2$ for some constant $C > 0$. Next, calculation shows

$$\mu_2(\theta_0||\theta) = (\theta - \theta_0)^T \Sigma_0 (\theta - \theta_0) + \left(D_A(\theta_0||\theta)\right)^2, \tag{A.2.11}$$

and we know that for Exponential families, $\Sigma_0 = A^{(2)}(\theta_0)$. Thus (A.2.11) becomes

$$\mu_2(\theta_0||\theta) = (\theta - \theta_0)^T A^{(2)}(\theta_0)(\theta - \theta_0) + \left(D_A(\theta_0||\theta)\right)^2. \tag{A.2.12}$$

Since, by hypothesis, $A^{(2)}(\theta)$ exists and $D_A(\theta)$ is finite for all $\theta \in \Theta$, we conclude $\mu_2(\theta_0||\theta)$ is finite for all $\theta \in \Theta$. This verifies assumptions 3 and 5 starting from the hypothesis. Using (A.2.9), we see that the first and second derivatives of D_A, with respect to the second argument, satisfy

$$D_A^{(1)}(\theta_0||\theta) = A^{(1)}(\theta) - A^{(1)}(\theta_0), \quad D_A^{(2)}(\theta_0||\theta) = A^{(2)}(\theta). \tag{A.2.13}$$

Thus, combining (A.2.10), (A.2.12) and (A.2.13), we have

$$
\begin{gathered}
KL^{(2)}(\theta_0||\theta) = A^{(2)}(\theta), \\
\mu_2^{(2)}(\theta_0||\theta) = 2\left(A^{(2)}(\theta_0) + D_A(\theta_0||\theta)A^{(2)}(\theta)\right) + \left(A^{(1)}(\theta) - A^{(1)}(\theta_0)\right)\left(A^{(1)}(\theta) - A^{(1)}(\theta_0)\right)^T.
\end{gathered}
\tag{A.2.14}
$$

Since sum of α-Lipschitz functions is again α-Lipschitz, (A.2.14) shows why the hypothesis of the corollary suffices to conclude that assumption 4 of theorem 1 holds.

Proof of lemma 1

$KL(\psi_2||\pi_n) = \int \psi_2 \log \psi_2 - \int \psi_2 \log \pi_n$, $KL(\psi_1||\pi_n) = \int \psi_1 \log \psi_1 - \int \psi_1 \log \pi_n$, so that $KL(\psi_2||\pi_n) - KL(\psi_1||\pi_n) = KL(\psi_2||\psi_1) + \int(\psi_2 - \psi_1)\log \psi_1 - \int(\psi_2 - \psi_1)\log \pi_n = KL(\psi_2||\psi_1) + \int(\psi_2 - \psi_1)(\log \psi_1 - \log \pi_n)$ and we are done.

Proof of lemma 2

Let $\phi, \phi_1, \ldots \phi_k \in \Gamma_n, \psi_1 = \sum_{j=1}^n \beta_j \phi_j$ for $\boldsymbol{\beta} \in \Delta^k$ and $\psi_2 = \psi_1 + \alpha(\phi - \psi_1)$ for $\alpha \in [0, 1]$. Starting with lemma 1, we have the Taylor expansion

$$
\begin{aligned}
\mathcal{D}_n(\psi_2||\psi_1) &= KL(\psi_1 + \alpha(\phi - \psi_1)||\psi_1) \\
&= \alpha \frac{\partial}{\partial \alpha'}\bigg|_{\alpha'=0} KL(\psi_1 + \alpha'(\phi - \psi_1)||\psi_1) + \frac{\alpha^2}{2}\frac{\partial^2}{\partial \alpha'^2}\bigg|_{\alpha'=\beta} KL(\psi_1 + \alpha'(\phi - \psi_1)||\psi_1) \\
&=: \alpha T_1 + \frac{\alpha^2}{2} T_2,
\end{aligned}
\tag{A.2.15}
$$

where β lies in between 0 and α, and T_1, T_2 denote the first and second derivatives respectively. For T_1 we have

$$
\begin{aligned}
\frac{\partial}{\partial \alpha} KL(\psi_1 + \alpha(\phi - \psi_1)||\psi_1) &= \frac{\partial}{\partial \alpha}\left[\int (\psi_1 + \alpha(\phi - \psi_1)) \log\left(1 + \alpha\left(\frac{\phi}{\psi_1} - 1\right)\right)\right] \\
&= \int (\phi - \psi_1) \log\left(1 + \alpha\left(\frac{\phi}{\psi_1} - 1\right)\right),
\end{aligned}
\tag{A.2.16}
$$

where in the last step of (A.2.16), we have used the fact $\int \psi_1 = \int \phi = 1$. This directly shows $T_1 = 0$. Also, for any $\alpha \in [0,1]$, we have

$$\frac{\partial^2}{\partial \alpha^2} KL(\psi_1 + \alpha(\phi - \psi_1)||\psi_1) = \int \frac{(\phi - \psi_1)^2}{\psi_1 + \alpha(\phi - \psi_1)} \leq \chi^2(\psi_1||\phi) + \chi^2(\phi||\psi_1). \tag{A.2.17}$$

Now, by Cauchy–Schwartz inequality,

$$\begin{aligned} \chi^2(\psi_1||\phi) &= \int \frac{\left(\sum_{j=1}^k \beta_j(\phi_j - \phi)\right)^2}{\sum_{j=1}^k \beta_j \phi} \leq \int \sum_{j=1}^k \beta_j \frac{(\phi_j - \phi)^2}{\phi} = \sum_{j=1}^k \beta_j \chi^2(\phi_j||\phi), \\ \chi^2(\phi||\psi_1) &= \int \frac{\left(\sum_{j=1}^k \beta_j\left(\phi - \phi_j\right)\right)^2}{\sum_{j=1}^k \beta_j \phi_j} \leq \int \sum_{j=1}^k \beta_j \frac{(\phi - \phi_j)^2}{\phi_j} = \sum_{j=1}^k \beta_j \chi^2(\phi||\phi_j). \end{aligned} \tag{A.2.18}$$

Adding the two inequalities in (A.2.18) and combining with (A.2.17), we have

$$T_2 \leq \sum_{j=1}^k \beta_j \left(\chi^2(\phi||\phi_j) + \chi^2(\phi_j||\phi)\right).$$

Plugging in the results for T_1, T_2 in (A.2.15), we can conclude that

$$\mathcal{D}_n(\psi_2||\psi_1) \leq \frac{\alpha^2}{2} \sum_{j=1}^k \beta_j \left(\chi^2(\phi||\phi_j) + \chi^2(\phi_j||\phi)\right).$$

This gives

$$\mathcal{C}_n \leq \sup \left\{ \sum_{j=1}^k \beta_j \left(\chi^2(\phi||\phi_j) + \chi^2(\phi_j||\phi)\right) : \phi, \phi_1 \ldots \phi_k \in \Gamma_n, \boldsymbol{\beta} \in \Delta^k \right\}, \tag{A.2.19}$$

where this upper bound is a function of only M, c_0 and σ_n, appearing in the definition of Γ_n. We have reduced the upper bound calculation to that of χ^2 divergence of single Gaussians, so we can now apply lemma 3. For any two members of Γ_n, say ϕ_2 and ϕ_1 , we get

$$\chi^2\left(\phi_2||\phi_1\right) \leq (2 - c_0)^{-d/2} \exp\left(\frac{2M^2}{(2 - c_0)\sigma_n}\right). \tag{A.2.20}$$

Identical bound holds if we swap ϕ_1 and ϕ_2. Now this translates to

$$\mathcal{C}_n \leq 2(2-c_0)^{-d/2} \exp\left(\frac{2M^2}{(2-c_0)\sigma_n}\right). \tag{A.2.21}$$

Proof of lemma 3

We first define the χ^2 divergence between densities, which is a discrepancy measure comparable to Kullback–Leibler divergence and stronger than it. Refer to Van Erven and Harremos [2014] for more details.

Definition 5: For densities ϕ_1, ϕ_2 on $\mathbb{R}^d$, the χ^2 divergence of ϕ_2 from ϕ_1 is defined as

$$\chi^2(\phi_2||\phi_1) = -1 + \int \frac{\phi_2^2}{\phi_1}. \tag{A.2.22}$$

Choosing $\phi_i = \mathrm{N}\left(\mu_i, \sigma_i^2 \mathbf{I}_d\right), i = 1, 2$, we have

$$\frac{\phi_2^2}{\phi_1}(y) = \left(\sqrt{2\pi}\frac{\sigma_2^2}{\sigma_1}\right)^{-d} \exp\left[\left(-\frac{1}{2}\right)\left(\frac{2\|y-\mu_2\|_2^2}{\sigma_2^2} - \frac{\|y-\mu_1\|_2^2}{\sigma_1^2}\right)\right]. \tag{A.2.23}$$

Let the term inside the exponent in (A.2.23) be $-B(y)/2$. We re-write

$$B(y) = \left(\frac{2\sigma_1^2 - \sigma_2^2}{\sigma_1^2\sigma_2^2}\right)\left\|y - \frac{\frac{2\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}}{\frac{2\sigma_1^2-\sigma_2^2}{\sigma_1^2\sigma_2^2}}\right\|_2^2 - \frac{2\left\|\mu_2 - \mu_1\right\|_2^2}{2\sigma_1^2 - \sigma_2^2}, \tag{A.2.24}$$

which shows

$$\begin{aligned}\int_{y\in\mathbb{R}^d} \left(\sqrt{2\pi}\right)^{-d} \exp\left[\left(-\frac{1}{2}\right)\left(\frac{2\|y-\mu_2\|_2^2}{\sigma_2^2} - \frac{\|y-\mu_1\|_2^2}{\sigma_1^2}\right)\right] dy \\ = \exp\left(\frac{\|\mu_2-\mu_1\|_2^2}{2\sigma_1^2-\sigma_2^2}\right)\left(\frac{\sqrt{2\sigma_1^2-\sigma_2^2}}{\sigma_1\sigma_2}\right)^{-d}.\end{aligned} \tag{A.2.25}$$

Combining (A.2.23) and (A.2.25), and then using (A.2.22), we get lemma 3.

A.3 Auxiliary Results

In this section we provide proofs of the results $(3)-(6)$ in the main part of Chapter 2. Since (4) follows directly from Bernstein-von-Mises theorem and its equivalence with (5) follows from

$$d_H^2 \leq d_{TV} \leq \sqrt{2} d_H, \tag{A.3.1}$$

we only focus on

Proposition 1:

$$\begin{aligned} KL\left(N\left(\theta_0, n^{-1}\Sigma\right) || N\left(\mu_n, \Sigma_n\right)\right) &\rightsquigarrow \frac{1}{2}\chi_d^2, \\ KL\left(N\left(\overline{X}_n, n^{-1}\Sigma\right) || N\left(\mu_n, \Sigma_n\right)\right) &\to 0 \quad \text{a.s.}, \end{aligned} \tag{A.3.2}$$

where '$\rightsquigarrow$' denotes weak convergence and a.s stands for almost sure validity with respect to the data generating distribution.

Proof of proposition 1

We start with noting that

$$\mu_n = \frac{n\overline{X}_n + \mu_0}{n+1}, \quad \Sigma_n^{-1} = n\Sigma^{-1} + \Sigma_0^{-1}, \quad v_{1,n} := \mu_n - \theta_0, \quad w_{1,n} := \mu_n - \overline{X}_n. \tag{A.3.3}$$

Observe the difference between $v_{1,n}$ and $w_{1,n}$ by noting

$$\begin{aligned} \sqrt{n} v_{1,n} &= \frac{\sqrt{n}\left(\overline{X}_n - \theta_0\right)}{1+\frac{1}{n}} + \frac{\sqrt{n}}{n+1}\left(\mu_0 - \theta_0\right) =: v_{2,n} + v_{3,n}, \\ \sqrt{n} w_{1,n} &= \frac{\sqrt{n}}{n+1}\left(\mu_0 - \overline{X}_n\right) =: w_{2,n}. \end{aligned} \tag{A.3.4}$$

We now have from (A.3.3) and (A.3.4)

$$\begin{aligned} v_{1,n}^T \Sigma_n^{-1} v_{1,n} &= (v_{2,n} + v_{3,n})^T \left[\Sigma^{-1} + n^{-1}\Sigma_0^{-1}\right](v_{2,n} + v_{3,n}) \\ &= (v_{2,n} + v_{3,n})^T \Sigma^{-1} (v_{2,n} + v_{3,n}) + n^{-1}(v_{2,n} + v_{3,n})^T \Sigma_0^{-1} (v_{2,n} + v_{3,n}) \qquad \text{(A.3.5)} \\ &=: E_{1,n} + n^{-1} E_{2,n}. \end{aligned}$$

Let us deal with $E_{2,n}$ first. Breaking down further, we see

$$E_{2,n} = v_{2,n}^T \Sigma_0^{-1} v_{2,n} + 2 v_{2,n}^T \Sigma_0^{-1} v_{3,n} + v_{3,n}^T \Sigma_0^{-1} v_{3,n}. \qquad \text{(A.3.6)}$$

The last term in the right-hand-side of (A.3.6) is non-stochastic, and Σ_0, μ_0 are free of n. Hence, we get

$$v_{3,n}^T \Sigma_0^{-1} v_{3,n} = \frac{n}{(n+1)^2} (\mu_0 - \theta_0)^T \Sigma_0^{-1} (\mu_0 - \theta_0) \to 0. \qquad \text{(A.3.7)}$$

The second term in the right-hand-side of (A.3.6) satisfies

$$v_{2,n}^T \Sigma_0^{-1} v_{3,n} = \left(\frac{n}{n+1}\right)^2 \left[(\mu_0 - \theta_0)^T \Sigma_0^{-1}\right] \left(\overline{X}_n - \theta_0\right) \to 0 \quad a.s., \qquad \text{(A.3.8)}$$

using Strong Law of Large Numbers(SLLN). Now for the first term on the right-hand-side of (A.3.6), we have

$$n^{-1} v_{2,n}^T \Sigma_0^{-1} v_{2,n} = \left(\frac{n}{n+1}\right)^2 \left(\overline{X}_n - \theta_0\right)^T \Sigma_0^{-1} \left(\overline{X}_n - \theta_0\right) \to 0 \quad a.s., \qquad \text{(A.3.9)}$$

where we have again used SLLN. Putting together (A.3.3), (A.3.4), (A.3) and (A.3.6), we get that

$$n^{-1} E_{2,n} \to 0 \quad a.s. \qquad \text{(A.3.10)}$$

We now deal with $E_{1,n}$ in the right-hand-side of (A.3), observing

$$E_{1,n} = v_{2,n}^T \Sigma^{-1} v_{2,n} + 2 v_{2,n}^T \Sigma^{-1} v_{3,n} + v_{3,n}^T \Sigma^{-1} v_{3,n}. \qquad \text{(A.3.11)}$$

Similar to (A.3.7) and (A.3.8), we have

$$v_{3,n}^T \Sigma^{-1} v_{3,n} \to 0, \quad v_{2,n}^T \Sigma^{-1} v_{3,n} \to 0 \; a.s., \tag{A.3.12}$$

while the first term on the right-hand-side of (A.3.11) satisfies

$$v_{2,n}^T \Sigma^{-1} v_{2,n} = \left(\frac{n}{n+1}\right)^2 \left\| \Sigma_0^{-1/2} \sqrt{n} \left(\overline{X}_n - \theta_0\right) \right\|_2^2 \rightsquigarrow \chi_d^2. \tag{A.3.13}$$

by an application of Slutsky's theorem, as the random vector within norm signs in (A.3.13) follows a standard Gaussian distribution in d-dimensions under the true vector θ_0. Another application of Slutsky's theorem to combine (A.3.11), (A.3.12) and (A.3.13) allows us to get from (A.3)

$$v_{1,n}^T \Sigma_n^{-1} v_{1,n} \rightsquigarrow \chi_d^2. \tag{A.3.14}$$

We next turn our attention to

$$w_{1,n}^T \Sigma_n^{-1} w_{1,n} = w_{2,n}^T \left[\Sigma^{-1} + n^{-1} \Sigma_0^{-1}\right] w_{2,n} = w_{2,n}^T \Sigma^{-1} w_{2,n} + n^{-1} w_{2,n}^T \Sigma_0^{-1} w_{2,n} := F_{1,n} + n^{-1} F_{2,n}, \tag{A.3.15}$$

to get

$$\begin{aligned} F_{1,n} &= \frac{n}{(n+1)^2} \left(\overline{X}_n - \mu_0\right)^T \Sigma^{-1} \left(\overline{X}_n - \mu_0\right) \to 0 \quad a.s., \\ F_{2,n} &= \frac{n}{(n+1)^2} \left(\overline{X}_n - \mu_0\right)^T \Sigma_0^{-1} \left(\overline{X}_n - \mu_0\right) \to 0 \quad a.s. \end{aligned} \tag{A.3.16}$$

Note that the limit is driven by the preceding factor of $n/(n+1)^2$ rather than SLLN. We thus have

$$w_{1,n}^T \Sigma_n^{-1} w_{1,n} \to 0 \quad a.s. \tag{A.3.17}$$

We are now prepared for the final part of the proof where we use the well-known formula for KL divergence between Gaussians to get

$$\begin{aligned}
&KL\left(N\left(\theta_0, n^{-1}\Sigma\right)||N\left(\mu_n, \Sigma_n\right)\right)\\
=&\frac{1}{2}\left[\text{trace}\left(n^{-1}\Sigma_n^{-1}\Sigma\right) - d - \log\det\left(n^{-1}\Sigma_n^{-1}\Sigma\right) + v_{1,n}^T\Sigma_n^{-1}v_{1,n}\right],\\
&KL\left(N\left(\overline{X}_n, n^{-1}\Sigma\right)||N\left(\mu_n, \Sigma_n\right)\right)\\
=&\frac{1}{2}\left[\text{trace}\left(n^{-1}\Sigma_n^{-1}\Sigma\right) - d - \log\det\left(n^{-1}\Sigma_n^{-1}\Sigma\right) + w_{1,n}^T\Sigma_n^{-1}w_{1,n}\right]
\end{aligned} \tag{A.3.18}$$

Most of the expressions are common for both the equations in (A.3.18), and we evaluate them term by term. We have

$$\begin{aligned}
\text{trace}\left(n^{-1}\Sigma_n^{-1}\Sigma\right) - d &= \text{trace}\left(I_d + n^{-1}\Sigma_0^{-1}\Sigma\right) - d \to 0,\\
\log\det\left(n^{-1}\Sigma_n^{-1}\Sigma\right) &= \log\det\left(I_d + n^{-1}\Sigma_0^{-1}\Sigma\right) \to 0.
\end{aligned} \tag{A.3.19}$$

Proposition 1 now follows by combining (A.3.14), (A.3.17), (A.3.18) and (A.3.19) along with an application of Slutsky's theorem.

APPENDIX B

SUPPLEMENTARY MATERIAL TO SECOND CHAPTER

B.1 Proof of Theorem 1

Recall the following neighborhood of the parameter space

$$D_n = \Big\{ \beta \in \mathbb{R}^{d_n} : \Big[-\mathbb{E}\left[L_n(\eta,\eta^*)\right] \bigvee \text{Var}\left[L_n(\eta,\eta^*)\right] \Big] \leq s_n^* \log d_n \Big\},$$

noting that $-\mathbb{E}\left[L_n(\eta,\eta^*)\right] = \mathcal{D}_n(\eta^*||\eta)$ and $\text{Var}\left[L_n(\eta,\eta^*)\right] = \mathbb{E}\, Z_n^2(\eta,\eta^*)$. Let $\Pi_{D_n}(\beta)$ denote the restriction of prior $\Pi_n(\beta)$ to D_n. Then the denominator satisfies

$$\begin{aligned} \int \exp\left(L_n(\eta,\eta^*)\right)\Pi_n(\beta)d\beta &\geq \int_{D_n} \exp\left(L_n(\eta,\eta^*)\right)\Pi_n(\beta)d\beta \\ &= \Pi_n\left(D_n\right) \int \exp\left(L_n(\eta,\eta^*)\right)\Pi_{D_n}(\beta)d\beta. \end{aligned} \tag{B.1.1}$$

We shall separately lower bound the the prior probability term and the integral term. First, we work with the integral in the above display. Rewrite $d_n^{-2s_n^*} = \exp\left(-2s_n^* \log d_n\right)$. Then consider following the tail event and inclusions, corresponding to the integral above:

$$\begin{aligned} &\left\{ \int \exp\left(L_n(\eta,\eta^*)\right)\Pi_{D_n}(\beta)d\beta \leq d_n^{-2s_n^*} \right\} \\ \subset &\left\{ \int \Big[Z_n(\eta,\eta^*) - \mathcal{D}_n(\eta^*||\eta) \Big] \Pi_{D_n}(\beta)d\beta \leq -2s_n^* \log d_n \right\} \\ \subset &\left\{ \int Z_n(\eta,\eta^*)\Pi_{D_n}(\beta) \leq -s_n^* \log d_n \right\}. \end{aligned}$$

The first inclusion follows from Jensen's inequality, while the second uses that $\mathcal{D}_n(\eta^*||\eta) \leq s_n^* \log d_n$ on D_n. We can now make the following probability statement about the integral,

$$\begin{aligned} &\mathbb{P}\left[\int \exp\left(L_n(\eta,\eta^*)\right)\Pi_{D_n}(\beta)d\beta \leq d_n^{-2s_n^*}\right] \leq \mathbb{P}\left[\int Z_n(\eta,\eta^*)\Pi_{D_n}(\beta) \leq -s_n^*\log d_n\right] \\ \leq &\frac{\mathbb{E}\left[\int Z_n(\eta,\eta^*)\Pi_{D_n}(\beta)d\beta\right]^2}{\left(s_n^*\log d_n\right)^2} \leq \frac{\int \mathbb{E}\, Z_n^2(\eta,\eta^*)\Pi_{D_n}(\beta)d\beta}{\left(s_n^*\log d_n\right)^2} \leq \left(s_n^*\log d_n\right)^{-1}, \end{aligned} \tag{B.1.2}$$

where we have used Chebysev's inequality for the second inequality, the variance inequality in the third, and the fact that $\mathbb{E}\, Z_n^2(\eta,\eta^*) \leq s_n^*\log d_n$ on D_n for the final inequality. Note that (B.1.2) makes sense asymptotically because $s_n^*\log d_n \to \infty$ with $n \to \infty$. Now recall the definition of $\mathcal{B}_{2,n}$ in (3.4.4). Due to the hypothesis of Theorem 2, we can use Lemma 1 to have

$$\Pi_n(D_n) \gtrsim C_n \exp(-\lambda_n\|\beta^*\|_1)d_n^{-(a_n+4)s_n^*}.$$

Combining this with (B.1.1) and (B.1.2) shows that with probability greater than $1-\left(s_n^*\log d_n\right)^{-1}$ w.r.t the data generating distribution for $\beta^* \in \mathcal{B}_{2,n}$, we have

$$\int \exp\left(L_n(\eta,\eta^*)\right)\Pi_n(\beta)d\beta \gtrsim C_n \exp(-\lambda_n\|\beta^*\|_1)d_n^{-(a_n+6)s_n^*},$$

concluding the proof.

B.2 Proof of Theorem 2

We work with the quantity

$$\mathbb{E}\left[\Pi_n(B \mid Y^{(n)})\right] = \mathbb{E}\left[\frac{\int_B \exp\left(L_n(\eta,\eta^*)\right)\Pi_n(\beta)d\beta}{\int \exp\left(L_n(\eta,\eta^*)\right)\Pi_n(\beta)d\beta}\right],$$

where the expectation is taken with respect to the true data generating distribution and set B has the form $B = \left\{\beta \in \mathbb{R}^{d_n} : |\operatorname{supp}(\beta)| > \varepsilon_n\right\}$ for some constant $\varepsilon_n > 0$. Thus, Theorem 2 is concerned with the dimensionality of the posterior vector, specifically the posterior probability that the sparsity of β does not fall below a certain threshold. Start by defining $U_n :=$

$\mathcal{M}(A,X)\sqrt{n\log d_n}$, so that we have $\lambda_n \le U_n$ from the hyper-parameter bounds assumption. Consider $\Omega_n := \left\{Y^{(n)} : Z_n(\eta,\eta^*) \le U_n\|\beta-\beta^*\|_1\right\}$, where Ω_n^c represents a tail event of the centered log-likelihood ratio $Z_n(\eta,\eta^*)$. To find the probability of this event, observe that similar to calculations in Lemma 1,

$$\text{Var}\left(Z_n(\eta,\eta^*)\right) = \sum_{i=1}^{n}\left(\eta_i-\eta_i^*\right)^2 A''\left(\eta_i^*\right) \le n\mathcal{M}^2(A,X)\|\beta-\beta^*\|_1^2.$$

This shows, with the use of the definition of U_n and Chebysev's inequality,

$$\mathbb{P}\left(\Omega_n^c\right) \le \frac{\text{Var}\left(Z_n(\eta,\eta^*)\right)}{U_n^2\|\beta-\beta^*\|_1^2} \le \left(\log d_n\right)^{-1}.$$

Also, Theorem 1 claims existence of event $\overline{\Omega}_n$ so that we have $\mathbb{P}\left(\overline{\Omega}_n\right) \ge 1-\left(s_n^*\log d_n\right)^{-1}$ and $\int \exp\left(L_n(\eta,\eta^*)\right)\Pi_n(\beta)d\beta \gtrsim C_n d_n^{-(a_n+6)s_n^*}\exp(-\lambda_n\|\beta^*\|_1)$ on $\overline{\Omega}_n$, simultaneously. Thus, using $\Pi_n(B \mid Y^{(n)}) \le 1$ and the union bound for probabilities, we have

$$\begin{aligned}
&\mathbb{E}\left[\Pi_n(B \mid Y^{(n)})\right] \le \mathbb{E}\left[\Pi_n(B \mid Y^{(n)})\mathbf{1}_{\Omega_n\cap\overline{\Omega}_n}\right] + \mathbb{P}\left(\Omega_n^c\right) + \mathbb{P}\left(\overline{\Omega}_n^c\right)\\
&\lesssim C_n^{-1}d_n^{(a_n+6)s_n^*}\exp(\lambda_n\|\beta^*\|_1)\,\mathbb{E}\int_B \exp[L_n(\eta,\eta^*)]\mathbf{1}_{\Omega_n}\Pi_n(\beta)d\beta + \left(\log d_n\right)^{-1} + \left(s_n^*\log d_n\right)^{-1},
\end{aligned} \tag{B.2.1}$$

since $\mathbf{1}_{\Omega_n\cap\overline{\Omega}_n} \le \mathbf{1}_{\Omega_n}$. Since $\log d_n \to \infty$ by the order assumptions, it now suffices to work with the expectation term on the right hand side. Due to restriction to Ω_n, we get

$$\begin{aligned}
&\mathbb{E}\int_B \exp[L_n(\eta,\eta^*)]\mathbf{1}_{\Omega_n}\Pi_n(\beta)d\beta,\\
&\le \int_B \mathbb{E}\exp\left[\left(1-\frac{\lambda_n}{2U_n}\right)Z_n(\eta,\eta^*) + \frac{\lambda_n}{2}\|\beta-\beta^*\|_1 - \mathcal{D}_n(\eta^*||\eta)\right]\Pi_n(\beta)d\beta, \qquad \text{(B.2.2)}\\
&= \int_B \exp\left[\frac{\lambda_n}{2}\|\beta-\beta^*\|_1 - \mathcal{D}_n(\eta^*||\eta)\right].\mathbb{E}\left(\exp\left[\left(1-\frac{\lambda_n}{2U_n}\right)Z_n(\eta,\eta^*)\right]\right)\Pi_n(\beta)d\beta.
\end{aligned}$$

To calculate the expectation in the above display, we shall use Lemma 2, which is concerned with the connection of the KL divergence $\mathcal{D}_n(\eta^*||\eta)$ with the cumulant generating function(cgf) of the centered log-likelihood ratio $Z_n(\eta, \eta^*)$. We use $\alpha = 1 - \lambda_n/(2U_n)$ in Lemma 2, obtaining

$$\mathbb{E}\left(\exp\left[\left(1 - \frac{\lambda_n}{2U_n}\right) Z_n(\eta, \eta^*)\right]\right) \leq \left(1 - \frac{\lambda_n}{2U_n}\right) \mathcal{D}_n(\eta^* || \eta). \tag{B.2.3}$$

The fact $0 < \lambda_n/(2U_n) < 1$, implied by assumption $\mathcal{L}_0$, has been crucially used here. Combining (B.2.2) and (B.2.3), we have for the expectation in (B.2.1)

$$\begin{aligned}
&\mathbb{E}\int_B \exp[L_n(\eta, \eta^*)]\mathbf{1}_{\Omega_n}\Pi_n(\beta)d\beta \\
&\leq \int_B \exp\left[\frac{\lambda_n}{2}\|\beta - \beta^*\|_1 - \mathcal{D}_n(\eta^*||\eta)\right].\exp\left[\left(1 - \frac{\lambda_n}{2U_n}\right)\mathcal{D}_n(\eta^*||\eta)\right]\Pi_n(\beta)d\beta \\
&\leq \int_B \exp\left[\frac{\lambda_n}{2}\|\beta - \beta^*\|_1 - \frac{\lambda_n}{2U_n}\mathcal{D}_n(\eta^*||\eta)\right]\Pi_n(\beta)d\beta,
\end{aligned}$$

and hence

$$\begin{aligned}
&\exp(\lambda_n\|\beta^*\|_1).\,\mathbb{E}\int_B \exp[L_n(\eta, \eta^*)]\mathbf{1}_{\Omega_n}\Pi_n(\beta)d\beta \\
&\lesssim \int_B \exp\left[\lambda_n\|\beta^*\|_1 + \frac{\lambda_n}{2}\|\beta - \beta^*\|_1 - \frac{\lambda_n}{2U_n}\mathcal{D}_n(\eta^*||\eta)\right]\Pi_n(\beta)d\beta.
\end{aligned} \tag{B.2.4}$$

We now work with the exponent inside the integrand in (B.2.4). First, $\|\beta^*\|_1 + (1/2)\|\beta - \beta^*\|_1 \leq \|\beta_{S^*}\|_1 + (3/2)\|\beta_{S^*} - \beta^*\|_1 + \frac{1}{2}\|\beta_{S^{*c}}\|_1$. If $\|\beta_{S^{*c}}\|_1 \geq 7\|\beta_{S^*} - \beta^*\|_1$, then

$$\|\beta_{S^*}\|_1 + \frac{3}{2}\|\beta_{S^*} - \beta^*\|_1 + \frac{1}{2}\|\beta_{S^{*c}}\|_1 \leq -\frac{1}{4}\|\beta - \beta^*\|_1 + \|\beta\|_1, \tag{B.2.5}$$

and if $\|\beta_{S^{*c}}\|_1 < 7\|\beta_{S^*} - \beta^*\|_1$, then we use the IC(Model) assumption to get

$$\begin{aligned}
&\|\beta^*\|_1 + \frac{1}{2}\|\beta - \beta^*\|_1 \leq \|\beta_{S^*}\|_1 + \frac{7}{2}\|\beta_{S^*} - \beta^*\|_1 - 2\|\beta_{S^*} - \beta^*\|_1 + \frac{1}{2}\|\beta_{S^{*c}}\|_1 \\
&\leq \frac{7}{2}\frac{\sqrt{\mathcal{D}_n(\eta^*||\eta)s_n^*}}{\sqrt{n}\phi_1(A, X, S^*)} - \frac{1}{4}\|\beta - \beta^*\|_1 + \|\beta\|_1 \\
&\leq \frac{49U_n s_n^*}{8n\phi_1^2(A, X, S^*)} + \frac{1}{2U_n}\mathcal{D}_n(\eta^*||\eta) - \frac{1}{4}\|\beta - \beta^*\|_1 + \|\beta\|_1.
\end{aligned} \tag{B.2.6}$$

The fact that $\beta^* \in \mathcal{B}_n$ implies $\phi_1(A, X, S^*) > 0$ is crucially used above. Combining the above two exhaustive cases, namely (B.2.5) and (B.2.6), we get for the integral in (B.2.4)

$$\begin{aligned} &\int_B \exp\left[\lambda_n\|\beta^*\|_1 + \frac{\lambda_n}{2}\|\beta-\beta^*\|_1 - \frac{\lambda_n}{2U_n}\mathcal{D}_n(\eta^*||\eta)\right]\Pi_n(\beta)d\beta \\ &\leq \exp\left(\frac{49U_n^2 s_n^*}{8n\phi_1^2(A,X,S^*)}\right)\int_B \exp\left[-\frac{\lambda_n}{4}\|\beta-\beta^*\|_1 + \lambda_n\|\beta\|_1\right]\Pi_n(\beta)d\beta, \end{aligned} \tag{B.2.7}$$

where we have also used $\lambda_n \leq U_n$. Now consider the integral in (B.2.7) for

$$B = \left\{\beta \in \mathbb{R}^{d_n} : \operatorname{supp}(\beta) > \varepsilon_n\right\}.$$

Writing out the prior fully, we have

$$\begin{aligned} &\int_{|\operatorname{supp}(\beta)|>\varepsilon_n} \exp\left[-\frac{\lambda_n}{4}\|\beta-\beta^*\|_1 + \lambda_n\|\beta\|_1\right]\Pi_n(\beta)d\beta \\ &= \sum_{|S|>\varepsilon_n} C_n\binom{d_n}{|S|}^{-1}\left(\frac{\lambda_n}{2d_n^{a_n}}\right)^{|S|}\int \exp\left[-\frac{\lambda_n}{4}\|\beta_S-\beta^*\|_1\right]d\beta_S \\ &\leq C_n.\sum_{s=\varepsilon_n}^{d_n}\left(4d_n^{-a_n}\right)^s \lesssim C_n\left(4d_n^{-a_n}\right)^{\varepsilon_n} \leq C_n\left(d_n^{-(a_n-1)}\right)^{\varepsilon_n}, \end{aligned} \tag{B.2.8}$$

where, for the first inequality, we have used $\|\beta_S - \beta^*\|_1 \geq \|\beta_S - \beta^*_S\|_1$ before performing the Laplace density integral, and have used $d_n \geq 4$. Now putting together the bounds in (B.2.8), (B.2.7) and (B.2.4), and using $U_n = \mathcal{M}(A, X)\sqrt{n \log d_n}$, the first term in right-hand side of (B.2.1) satisfies

$$\mathbb{E}\left[\Pi_n(B \mid Y^{(n)})\mathbf{1}_{\Omega_n\cap\overline{\Omega}_n}\right] \lesssim \exp\left[\log d_n.\left[(a_n+6)s_n^* - (a_n-1)\varepsilon_n + \frac{49\mathcal{M}^2(A,X)s_n^*}{8\phi_1^2(A,X,S^*)}\right]\right] \to 0,$$

as soon as

$$(a_n+6)s_n^* - (a_n-1)\varepsilon_n + \frac{49\mathcal{M}^2(A,X)s_n^*}{8\phi_1^2(A,X,S^*)} \leq -s_n^*, \iff \varepsilon_n \geq s_n^* + \frac{8s_n^*}{a_n-1}\left[1 + \frac{49\mathcal{M}^2(A,X)}{8\phi_1^2(A,X,S^*)}\right],$$

as $s_n^* \log d_n \to \infty$ as $n \to \infty$. The proof is now completed by observing that the same lower bound on ε_n works for every $\beta^* \in \mathcal{B}_n$.

B.3 Proof of Theorem 3

Theorem 3 deals with the ℓ_1-distance based posterior contraction of β towards β^*, specifically the posterior probability of a set of the form $B_1 = \left\{\beta \in \mathbb{R}^{d_n} : \|\beta - \beta^*\|_1 > \varepsilon_{n,1}\right\}$. With $\mathcal{E}_1$ as in (3.4.6) and the choice $a_n \geq 1 + \mathcal{E}_1$, observe

$$\left\{\beta \in \mathbb{R}^{d_n} : |\operatorname{supp}(\beta)| \leq s_n^* \left(1 + \frac{\mathcal{E}_1}{a-1},\right)\right\} \subset \left\{\beta \in \mathbb{R}^{d_n} : |\operatorname{supp}(\beta)| \leq 2s_n^*\right\} =: B_2. \tag{B.3.1}$$

Now put $\overline{B} := B_1 \cap B_2$. As a result of Theorem 2, we shall can focus only on $\mathbb{E}\left[\Pi_n(\overline{B} \mid Y^{(n)})\right]$. Observe that $|\operatorname{supp}(\beta - \beta^*)| \leq 3s_n^*$ on $\overline{B}$ due to triangle inequality and (B.3.1). Now recall the definitions of Ω_n and $\overline{\Omega}_n$. Since $Z_n(\eta, \eta^*) \leq U_n\|\beta - \beta^*\|_1$ on Ω_n, and $\lambda_n \leq 2U_n$ implies $\lambda_n\|\beta^*\|_1 \leq \lambda_n\|\beta\|_1 + 2U_n\|\beta - \beta^*\|_1$, we have

$$\begin{aligned}\mathbb{E}\left[\Pi_n(\overline{B} \mid Y)\mathbf{1}_{\Omega_n \cap \overline{\Omega}_n}\right] &\lesssim C_n^{-1} d_n^{(a_n+6)s_n^*} \exp(\lambda_n\|\beta^*\|_1)\, \mathbb{E} \int_{\overline{B}} \exp[L_n(\eta, \eta^*)]\mathbf{1}_{\Omega_n}\Pi_n(\beta)d\beta \\ \leq C_n^{-1} d_n^{(a_n+6)s_n^*} &\int_{\overline{B}} \exp\left[4U_n\|\beta - \beta^*\|_1 - \mathcal{D}_n(\eta^*||\eta) - U_n\|\beta - \beta^*\|_1 + \lambda_n\|\beta\|_1\right]\Pi_n(\beta)d\beta,\end{aligned} \tag{B.3.2}$$

where, similar to (B.2.1) we already know $\mathbb{P}\left(\Omega_n \cap \overline{\Omega}_n\right)^c \to 0$ as $n \to \infty$. We shall now require the use IC(Dimension) assumption. Recall the definition of $\overline{\phi}_0(A, X, s)$ in (3.4.3), which shows it is decreasing in s. Hence we have $\overline{\phi}_0\left(A, X, |\operatorname{supp}(\beta - \beta^*)|\right) \geq \overline{\phi}_0\left(A, X, 3s_n^*\right)$ whenever $\beta \in \overline{B}$, and $\overline{\phi}_0\left(A, X, 3s_n^*\right) > 0$ on account of $\beta^* \in \mathcal{B}_n$ and IC(Dimension) assumption. This leads to

$$4U_n\|\beta - \beta^*\|_1 \leq \frac{4U_n\sqrt{3s_n^*\mathcal{D}_n(\eta^*||\eta)}}{\sqrt{n}\overline{\phi}_0\left(A, X, 3s_n^*\right)} \leq \frac{12U_n^2 s_n^*}{n\overline{\phi}_0^2\left(A, X, 3s_n^*\right)} + \mathcal{D}_n(\eta^*||\eta). \tag{B.3.3}$$

Combining (B.3.2) and (B.3.3) with inequalities $\|\beta-\beta^*\|_1 > \varepsilon_{n,1}$ for $\beta \in \overline{B}$, and $U_n \geq \frac{\lambda_n}{4} + \frac{U_n}{2}$, we have

$$\begin{aligned}&\mathbb{E}\left[\Pi_n(\overline{B} \mid Y)\mathbf{1}_{\Omega_n\cap\overline{\Omega}_n}\right]\\ &\lesssim C_n^{-1} d_n^{(a_n+6)s_n^*}\exp\left(\frac{12U_n^2 s_n^*}{n\overline{\phi}_0^2\left(A,X,3s_n^*\right)} - \frac{U_n\varepsilon_{n,1}}{2}\right)\int_{\overline{B}}\exp\left[-\frac{\lambda_n}{4}\|\beta-\beta^*\|_1 + \lambda_n\|\beta\|_1\right]\Pi_n(\beta)d\beta.\end{aligned}$$

Similar to calculations in (B.2), we note

$$\int_{\overline{B}}\exp\left[-\frac{\lambda_n}{4}\|\beta-\beta^*\|_1 + \lambda_n\|\beta\|_1\right]\Pi_n(\beta)d\beta \lesssim \frac{C_n}{1-4d_n^{-a_n}} \lesssim C_n,$$

for large enough d_n, hence sufficiently large n, leading to

$$\mathbb{E}\left[\Pi_n(\overline{B} \mid Y)\mathbf{1}_{\Omega_n\cap\overline{\Omega}_n}\right] \lesssim \exp\left[\log d_n\left((a_n+6)s_n^* - \frac{\sqrt{n}\mathcal{M}(A,X)\varepsilon_{n,1}}{2\sqrt{\log d_n}}\right) + \frac{12U_n^2 s_n^*}{n\overline{\phi}_0^2\left(A,X,3s_n^*\right)}\right]. \tag{B.3.4}$$

Now recall the definition of $\mathcal{E}_2$ in (3.5.4). Since $\beta^* \in \mathcal{B}_n$, we put to use that $s_n^* \leq b_n$ and that $\overline{\phi}_0\left(A,X,s\right)$ is monotonically decreasing in s, to get from (B.3.4)

$$\mathbb{E}\left[\Pi_n(\overline{B} \mid Y)\mathbf{1}_{\Omega_n\cap\overline{\Omega}_n}\right] \lesssim \exp\left[\log d_n\left(s_n^*\left(a_n+\mathcal{E}_2\right) - \frac{\sqrt{n}\mathcal{M}(A,X)\varepsilon_{n,1}}{2\sqrt{\log d_n}}\right)\right] \to 0,$$

as soon as

$$s_n^*\left(a_n+\mathcal{E}_2\right) - \frac{\sqrt{n}\mathcal{M}(A,X)\varepsilon_{n,1}}{2\sqrt{\log d_n}} \leq -s_n^* \Longleftrightarrow \varepsilon_{n,1} \geq \frac{2s_n^*\left(1+a_n+\mathcal{E}_2\right)}{\mathcal{M}(A,X)}\sqrt{\frac{\log d_n}{n}},$$

as $s_n^*\log d_n \to \infty$ with $n \to \infty$. The proof is now completed by observing that the same lower bound on $\varepsilon_{n,1}$ works for every $\beta^* \in \mathcal{B}_n$.

B.4 Auxiliary results

In this section, we note down three Lemmata used in the proofs of our Theorems. Lemma 1 deals with lower bounding $\Pi_n(D_n)$ with D_n defined in (3.5.2). Lemma 2 upper bounds the cumulant generating function of $Z_n(\eta, \eta^*)$ (defined in (3.2.4)), in terms of $\mathcal{D}_n(\eta^*||\eta)$ (also defined in (3.2.4)). Lastly, Lemma 3 deals with a local upper bound on $A''(\cdot)$ that is uniform over all $\beta^* \in \mathcal{B}_{2,n}$ (see (3.4.4) for definition).

Lemma 1. *Let $a_n > 0$ and λ_n satisfy assumption $\mathcal{L}_0$. Let $n, d_n \to \infty$ and $d_n > n$. Based on* (3.4.1)*, consider large enough n so that $b_n \log d_n < n$. With $\mathcal{B}_{2,n}$ defined in* (3.4.4)*, let the true β^* belong to $\mathcal{B}_{2,n}$. Then, for the set D_n defined in* (3.5.2)*, we have for large enough n*

$$\Pi_n(D_n) \geq C_n e^{-1/2} \exp(-\lambda_n \|\beta^*\|_1) d_n^{-(a_n+4)s_n^*}.$$

Proof: Begin by defining

$$\begin{aligned} B_n^*(A, X) &:= \mathcal{M}^{-1}(A, X)\sqrt{\frac{s_n^* \log d_n}{n}}, \\ \Delta_n &:= \left\{\beta \in \mathbb{R}^{d_n} : \|\beta - \beta^*\|_1 \leq B_n^*(A, X)\right\}. \end{aligned} \tag{B.4.1}$$

Consider any $\beta \in \Delta_n$. We have

$$\left|\eta\left(x_i^T\beta\right) - \eta\left(x_i^T\beta^*\right)\right| \leq \|X\|_{(\infty,\infty)}\|\beta - \beta^*\|_1 \leq \frac{1}{\mathcal{M}_1(A)}\sqrt{\frac{s_n^* \log d_n}{n}} \leq \sqrt{\frac{s_n^* \log d_n}{n}},$$

for all $i = 1, \dots n$, since η is a Lipschitz function, and $\mathcal{M}_1(A) \geq 1$ because of (3.4.5). This shows, by Lemma 3, that for all $i = 1, \dots n$, we have $A''(\gamma) \leq \mathcal{M}_0^2(A)$ whenever γ lies between η_i and η_i^*. Now note that

$$\text{Var}\left[L_n(\eta, \eta^*)\right] = \sum_{i=1}^n \left(\eta_i - \eta_i^*\right)^2 A''\left(\eta_i^*\right), \; -\mathbb{E}\left[L_n(\eta, \eta^*)\right] = \frac{1}{2}\sum_{i=1}^n \left(\eta_i - \eta_i^*\right)^2 A''\left(\tilde{\eta}_i\right),$$

where for all $i \in 1, \dots n$, we have $\tilde{\eta}_i$ lying between η_i and η_i^*. Thus we have for any $\beta \in \Delta_n$

$$\begin{aligned} -\mathbb{E}\left[L_n(\eta, \eta^*)\right] \bigvee \text{Var}\left[L_n(\eta, \eta^*)\right] &\leq \mathcal{M}_0^2(A)\sum_{i=1}^n \left(\eta_i - \eta_i^*\right)^2 \leq \mathcal{M}_1^2(A)\sum_{i=1}^n \left(\eta_i - \eta_i^*\right)^2 \\ &\leq n\|X\|_{(\infty,\infty)}^2 \mathcal{M}_1^2(A)\|\beta - \beta^*\|_1^2 \leq s_n^* \log d_n. \end{aligned} \tag{B.4.2}$$

Taken together, (B.4.2) and (3.5.2) imply $\beta \in D_n$, which implies $\Delta_n \subset D_n$, and hence $\Pi_n\left(D_n\right) \geq \Pi_n\left(\Delta_n\right)$. Restricting prior Π_n to the true model S^*, we see

$$\begin{aligned} \Pi_n(\Delta_n) &\geq C_n \binom{d_n}{s_n^*}^{-1} \left(\frac{\lambda_n}{2d_n^{a_n}}\right)^{s_n^*} \int \exp(-\lambda_n \|\beta_{S^*}\|_1) \mathbf{1}_{\{\|\beta_{S^*}-\beta^*\|_1 \leq B_n^*(A,X)\}} d\beta_{S^*} \\ &\geq C_n \binom{d_n}{s_n^*}^{-1} \left(\frac{\lambda_n}{2d_n^{a_n}}\right)^{s_n^*} \exp(-\lambda_n \|\beta^*\|_1) \int \exp(-\lambda_n \|\chi_{S^*}\|_1) \mathbf{1}_{\{\|\chi_{S^*}\|_1 \leq B_n^*(A,X)\}} d\chi_{S^*}, \end{aligned}$$

with the change of variable $\chi_{S^*} := \beta_{S^*} - \beta^*$, applying triangle inequality and noting that $\|\beta^*_{S^*}\|_1 = \|\beta^*\|_1$. To lower bound the above integral, we use it's analogy with Poisson process calculations. If we denote by $\mathcal{P}_j, j \geq 1$ independently and identically distributed exponential random variables with rate parameter λ_n, then the above integral is identical to calculating the probability of the event that at least s_n^* many occurrences of the Poisson process $\left\{\sum_{j=1}^m \mathcal{P}_j\right\}_{m=1}^\infty$ happen before time $B_n^*(A,X)$. This leads us to

$$\begin{aligned} \Pi_n(\Delta_n) &\geq C_n \binom{d_n}{s_n^*}^{-1} d_n^{-a_n s_n^*} \exp(-\lambda_n \|\beta^*\|_1) \exp\Big[-\lambda_n B_n^*(A,X)\Big] \sum_{j=s_n^*}^{\infty} \frac{\Big[\lambda_n B_n^*(A,X)\Big]^j}{j!} \\ &\geq C_n \binom{d_n}{s_n^*}^{-1} d_n^{-a_n s_n^*} \exp(-\lambda_n \|\beta^*\|_1) \exp\Big[-\lambda_n B_n^*(A,X)\Big] \frac{\Big[\lambda_n B_n^*(A,X)\Big]^{s_n^*}}{s_n^*!} \\ &\geq C_n d_n^{-(a_n+1)s_n^*} \exp(-\lambda_n \|\beta^*\|_1) \exp\left[-\frac{\lambda_n \epsilon_n}{\mathcal{M}(A,X)}\right] \left(\frac{\lambda_n \epsilon_n}{\mathcal{M}(A,X)}\right)^{s_n^*}. \end{aligned} \tag{B.4.3}$$

where $\epsilon_n := \sqrt{(s_n^* \log d_n)/n}$, so that $B_n^*(A,X) = \epsilon_n / \mathcal{M}(A,X)$ by (B.4.1), and we have used $\binom{d_n}{s_n^*} s_n^*! \leq d_n^{s_n^*}$ for the last inequality. Note that $\beta^* \in \mathcal{B}_n$ implies $s_n^* \leq b_n$, which, coupled with $b_n \log d_n < n$ implies $\epsilon_n < 1$. Based on assumption $\mathcal{L}_0$ and $d_n > n$, observe that

$$\frac{\lambda_n \epsilon_n}{\mathcal{M}(A,X)} \leq \frac{1}{2} \Rightarrow \exp\left[-\frac{\lambda_n \epsilon_n}{\mathcal{M}(A,X)}\right] \left(\frac{\lambda_n \epsilon_n}{\mathcal{M}(A,X)}\right)^{s_n^*} \geq e^{-\frac{1}{2}}. \left(\sqrt{\frac{s_n^* \log d_n}{n d_n^2}}\right)^{s_n^*} \geq e^{-\frac{1}{2}} d_n^{-\frac{3}{2} s_n^*},$$

and since $d_n \to \infty$ and $\epsilon_n < 1$, we have

$$\frac{\lambda_n \epsilon_n}{\mathcal{M}(A,X)} \geq \frac{1}{2} \implies \exp\left[-\frac{\lambda_n \epsilon_n}{\mathcal{M}(A,X)}\right]\left(\frac{\lambda_n \epsilon_n}{\mathcal{M}(A,X)}\right)^{s_n^*} \geq 2^{-s_n^*} \exp\left[-\sqrt{\log d_n}\right] \geq e^{-\frac{1}{2}} d_n^{-\frac{3}{2}s_n^*},$$

where the last inequality holds for large enough d_n, hence for large enough n. Plugging these back into the lower bound on $\Pi_n(\Delta_n)$ in (B.4.3), we arrive at the statement of the Lemma.

Lemma 2. *Let the centered log-likelihood ratio $Z_n(\eta, \eta^*)$ and the Kullback–Leibler divergence term $\mathcal{D}_n(\eta^*||\eta)$ be defined as in* (3.2.4). *Then, for any $\alpha \in (0,1)$, the cumulant generating function $\psi(\alpha) := \log \mathbb{E}\left[\exp\left(\alpha Z_n(\eta, \eta^*)\right)\right]$ of $Z_n(\eta, \eta^*)$ satisfies*

$$\psi(\alpha) \leq \alpha \mathcal{D}_n(\eta^*||\eta).$$

Proof: This Lemma is concerned with the connection of the KL divergence $\mathcal{D}_n(\eta^*||\eta)$ in cGLM models with the cumulant generating function(cgf) of the centered log-likelihood ratio $Z_n(\eta, \eta^*)$. Start by fixing $i = 1, \dots n$ and let $\alpha \in (0,1)$. Put the cgf at α of $Z_i(\eta_i, \eta_i^*)$ as $\psi_i(\alpha) := \log \mathbb{E}\left[\exp\left(\alpha Z_i(\eta_i, \eta_i^*)\right)\right]$. Since y_i is a draw from the exponential family, we know from standard properties that $\mathbb{E}\, T_i = A'\left(\eta_i^*\right)$ and for any $b \in \mathbb{R}$, $\log \mathbb{E}\left[\exp(bT_i)\right] = A\left(\eta_i^* + b\right) - A\left(\eta_i^*\right)$. Hence, we have

$$\begin{aligned}\psi_i(\alpha) &= \log \mathbb{E}\left(\exp\left[\alpha\left(\eta_i - \eta_i^*\right)\left(T_i - \mathbb{E}\, T_i\right)\right]\right) \\ &= -\alpha\left(\eta_i - \eta_i^*\right)\mathbb{E}\, T_i + \log \mathbb{E}\left(\exp\left[\alpha\left(\eta_i - \eta_i^*\right)T_i\right]\right) \\ &= A\left(\eta^* + \alpha\left(\eta_i - \eta_i^*\right)\right) - A\left(\eta^*\right) - \alpha\left(\eta_i - \eta_i^*\right)A'\left(\eta_i^*\right) = \mathcal{D}_i\left(\eta_i^*||\alpha\eta_i + (1-\alpha)\eta_i^*\right).\end{aligned} \tag{B.4.4}$$

Now, since $\alpha \in (0,1)$, we can use the convexity of KL divergence to obtain for every $i = 1, \dots n$

$$\mathcal{D}_i\left(\eta_i^*||\alpha\eta_i + (1-\alpha)\eta_i^*\right) \leq \alpha\mathcal{D}_i(\eta_i^*||\eta_i). \tag{B.4.5}$$

Since cgf of sum of independent random variables equals sum of their cgf's, we can sum over $i = 1, \dots n$ both the sides of (B.4.5) and use (B.4.4) for each term to obtain the statement of the Lemma.

Lemma 3. *Let $\mathcal{B}_{2,n}$ defined by* (3.4.4)*, while $\eta(\cdot)$ and $\mathcal{M}_0(A)$ are as in* (3.2.3)*. Then, for large enough n, we have*

$$\sup_{\beta^* \in \mathcal{B}_{2,n}} \max_{1 \le i \le n} \sup \left\{ A''(\gamma) : |\gamma - \eta_i^*| \le \sqrt{\frac{s_n^* \log d_n}{n}} \right\} \le \mathcal{M}_0^2(A).$$

Proof: Start with the simpler case, where $\mathcal{M}_0(A)$ can be chosen based on $A(\cdot)$, so that we have $\mathcal{I}_A\left(\mathcal{M}_0^2(A)/2\right) = \mathbb{R}$. This results in $A''(\cdot)$ having the global upper bound $\mathcal{M}_0^2(A)/2$ on its support, and hence the above display holds trivially. Next, assume $\mathcal{I}_A(b)$ is a strict interval subset of $\mathbb{R}$ for any $b > 0$. For our proof, we shall only deal with interval form $\mathcal{I}_A\left(\mathcal{M}_0^2(A)/2\right) = (-\infty, z_2],\ \mathcal{I}_A\left(\mathcal{M}_0^2(A)\right) = (-\infty, z_1]$, where $z_1, z_2 \in \mathbb{R},\ z_1 > z_2$. All other form of intervals can be dealt with essentially the same technique we use.

By the definition of clipping function, we have $\eta_i \in (-\infty, z_2], i = 1, \dots n$ for any $\beta \in \mathbb{R}^{d_n}$, specifically for any $\beta = \beta^* \in \mathcal{B}_{2,n}$. Define the following neighborhood union

$$\mathcal{N}_n := \bigcup_{i=1}^{n} \left\{ \gamma \in \mathbb{R} : |\gamma - \eta_i^*| \le \sqrt{\frac{b_n \log d_n}{n}} \right\}. \tag{B.4.6}$$

(B.4.6) deals with neighborhoods of η_i^*, where $\eta_i^* \equiv \eta\left(x_i^{\mathrm{T}} \beta^*\right), \beta^* \in \mathcal{B}_n, i = 1, \dots n$. By (3.4.1), these neighborhoods shrink to zero with large n. Since $s_n^* \le b_n$ is implied by $\beta^* \in \mathcal{B}_{2,n}$, and $z_1 > z_2$, we have for large enough n,

$$\mathcal{N}_n \subset \mathcal{I}_A\left(\mathcal{M}_0^2(A)\right),$$

implying

$$\max_{1\leq i\leq n}\,\sup\left\{A^{''}(\gamma):|\gamma-\eta_i^*|\leq\sqrt{\frac{s_n^*\log d_n}{n}}\right\}\leq\mathcal{M}_0^2(A). \tag{B.4.7}$$

As (B.4.7) holds for any $\beta^*\in\mathcal{B}_{2,n}$, this concludes the proof.