Journal of Machine Learning Research 23 (2022) 1-42 Submitted 2/21; Revised 6/22; Published 6/22

Statistical Optimality and Stability of Tangent Transform Algorithms in Logit Models

Indrajit Ghosh INDRAJIT@STAT.TAMU.EDU
Anirban Bhattacharya ANIRBANB@STAT.TAMU.EDU
Debdeep Pati DEBDEEP@STAT.TAMU.EDU
Department of Statistics
Texas A& M University
College Station, TX 77843-3143, USA

Editor: Pierre Alquier

Abstract

A systematic approach to finding variational approximation in an otherwise intractable non-conjugate model is to exploit the general principle of convex duality by minorizing the marginal likelihood that renders the problem tractable. While such approaches are popular in the context of variational inference in non-conjugate Bayesian models, theoretical guarantees on statistical optimality and algorithmic convergence are lacking. Focusing on logistic regression models, we provide mild conditions on the data generating process to derive non-asymptotic upper bounds to the risk incurred by the variational optima. We demonstrate that these assumptions can be completely relaxed if one considers a slight variation of the algorithm by raising the likelihood to a fractional power. Next, we utilize the theory of dynamical systems to provide convergence guarantees for such algorithms in logistic and multinomial logit regression. In particular, we establish local asymptotic stability of the algorithm without any assumptions on the data-generating process. We explore a special case involving a semi-orthogonal design under which a global convergence is obtained. The theory is further illustrated using several numerical studies.

Keywords: Bayesian; Dynamical System; Logistic regression; Rényi divergence; Risk bound; Variational Inference

1. Introduction

Variational Inference (VI) has gained substantial momentum in recent years as an efficient way of performing approximate Bayesian inference. VI seeks to minimize a divergence measure between a tractable family of probability distributions and the posterior distribution, utilizing optimization based techniques to arrive at a minima. In many high dimensional examples where sampling based techniques such as the Markov chain Monte Carlo require expert vigilance and care for scalability, VI provides a viable answer with relatively lower computational cost. Some notable application areas include graphical models (Wainwright et al., 2008; Jordan et al., 1999), hidden markov models (MacKay, 1997), latent class models (Blei et al., 2003), neural networks (Graves, 2011) to name a few. Refer to Chapter 10 in Bishop (2006) and Blei et al. (2017) for excellent reviews on the topic.

The empirical success of VI has prompted researchers to investigate their theoretical properties. Two distinct directions of research seem to have emerged over the last few years.

One line of research concerns the statistical aspects of variational estimators (Alquier et al., 2016; Pati et al., 2018; Yang et al., 2020; Chérief-Abdellatif and Alquier, 2018; Alquier and Ridgway, 2020; Zhang and Gao, 2020; Wang and Blei, 2019a,b) in a general setting, delineating sufficient conditions on the data generation mechanism and the variational family under which the variational estimators have optimal first or second-order statistical properties. Motivated by the robustness properties of a fractional likelihood (Bhattacharya et al., 2019; Alquier and Ridgway, 2020), Yang et al. (2020) proposed a simple modification to the variational objective function, deemed as the α-Variational Bayes (α-VB), that only requires the variational family to be sufficiently flexible and the prior density to be appropriately concentrated around the true parameter to obtain optimal risk bounds.

The other line of research studies the convergence of the algorithms employed to arrive at the variational optimizer. In this aspect, the coordinate ascent variational inference (CAVI) algorithm for mean-field VI (refer to Chapter 10 of Bishop (2006)) has arguably received the most attention due to its simplicity and generality. An early result on algorithmic convergence (and lack thereof) of CAVI in Gaussian mixture models appears in Wang and Titterington (2006). Zhang and Zhou (2020); Mukherjee et al. (2018) analyzed CAVI for stochastic block models, a popular model for networks belonging to conditionally conjugate exponential family (cEXP). Yin et al. (2020) obtained convergence of cluster labels in a stochastic block model by considering a structured variational family which was not possible using mean field VI. Ghorbani et al. (2018) noted instability of naive mean-field VI in latent Dirichlet allocation and provided a remedy by optimizing a different type of free energy (TAP) instead of the standard variational objective. Locatello et al. (2018); Campbell and Li (2019) analyzed convergence of a more flexible class of boosting algorithms which aim to approximate the target class by a mixture of Gaussians rather than a single Gaussian or a product distribution.

Our goal in this article is to explore a popular class of variational approximation technique outside cEXP, called the *tangent-transform approach* (Jaakkola, 1997; Jaakkola and Jordan, 2000). The tangent transform approach is an example of a structured variational approximation, lying on the spectrum between the two extremes given by the restrictive mean-field inference and the highly flexible variational boosting. In this specific instance, the structure exploited is convex duality (Jordan et al., 1999; Wainwright and Jordan, 2003; Wainwright et al., 2005) to minorize the log-likelihood function and provide sharp bounds for the log-partition function in the exponential family of distributions. Assume $p(x, \theta)$ is an exponential family on a discrete space $\mathcal{X}$ indexed by parameter $\theta \in \Theta$,

$$p(x;\theta) = \exp\{\langle \theta, t(x)\rangle - B(\theta)\}, \quad B(\theta) = \log\Big[\sum_{x\in\mathcal{X}} \exp\{\langle \theta, t(x)\rangle\}\Big].$$

The log-partition function $B(\theta)$, a convex function of θ, plays a critical role in computing summary measures of $p(x;\theta)$. Jaakkola and Jordan (2000) exploits the dual representation of the log-partition function in terms of its Fenchel-Legendre conjugate $B(\theta) = \sup_{\mu\in\mathcal{M}}[\langle \theta, t(x)\rangle - \{-H(\mu)\}]$, where $H(\mu)$ is negative entropy of the distribution parameterized by μ and $\mathcal{M}$ is the marginal polytope.

Ideas related to the tangent-transform have found widespread applications ranging from approximate inference in graphical models (Jordan et al., 1999), low-rank approximations (Srebro and Jaakkola, 2003), inference in large scale generalized linear models (Nickisch and

Seeger, 2009), non-conjugate latent Gaussian models (Emtiyaz Khan et al., 2013) to more recently in sparse kernel machines (Shi and Yu, 2019), hierarchical relevance determination (Hirose et al., 2020), online prediction (Konagayoshi and Watanabe, 2019) among others. Jaakkola and Jordan (2000) exploits convex duality to minorize the marginal likelihood by introducing a variational parameter that allows the minorant to be arbitrarily close to the marginal likelihood. Logistic and multinomial logit regression models are notable examples where a clever use of this idea results in a straightforward Expectation-Maximization (EM) algorithm to compute the variational updates.

In this article, we investigate both the statistical and algorithmic aspects of the tangent transform algorithm in logit models. Despite its widespread usage, statistical properties of the point estimate of the regression coefficients resulting from a tangent transform algorithm has not been previously studied. One possible reason is that unlike mean-field VI, where the global objective is to minimize the Kullback—Leibler (or another) divergence between a product distribution and the posterior distribution, the tangent transform algorithm is defined *locally*, without a clear global objective function that is being minimized. A key observation underlying our statistical analysis expresses any stationary point of the EM algorithm as a minimzer of a suitably chosen *global* variational objective function. This observation allows us to extend previously developed variational risk bounds for mean-field VI (Yang et al., 2020; Pati et al., 2018) to the present setting with some non-trivial adaptations. Specifically, two important extensions were made. First, the introduction of variational parameters minorizes the joint density, and the minorant does not integrate to one. This renders the results in Yang et al. (2020) inapplicable in our case which require a probabilistic latent variable augmentation. Second, we accurately characterize the *Jensen gap* of the minorant in terms of the variational parameter which ultimately allows us to establish optimal contraction. We show that with minimal assumptions on the data generating process and the prior density on the regression coefficients, the variational risk bound is minimax optimal (up to logarithmic terms). Moreover, the assumption on the data generating process can be completely relaxed by raising the standard logistic likelihood by a fractional exponent (Bhattacharya et al., 2019).

Next, we investigate the convergence of the EM algorithm to the fixed point of the EM iterations. There has been some previous efforts to shed more light into the EM sequence of tangent-transform algorithms. Hunter and Lange (2004) studied connections between minimization-majorization (MM) in case of logistic likelihood to argue convergence of the coefficient vector updates. However, they do not consider Bayesian inference on the logistic regression coefficients. Durante and Rigon (2019) drew a connection with the Pólya-Gamma data augmentation technique (Polson et al., 2013) to provide a probabilistic interpretation of the EM updates and showed that the optimal evidence lower bound of the tangent transformation approach coincides with the same obtained in a bonafide variational inference with a suitably defined conditionally conjugate exponential family. Although it is possible to use the probabilistic characterization of updates in Durante and Rigon (2019) and leverage on the techniques developed in Yang et al. (2020) to derive variational risk bounds, the theory in Yang et al. (2020) would also require us to impose additional assumptions for optimal concentration of Pólya-Gamma random variables. Moreover, our goal is to generalize the results to the case of multinomial logit regression for which such probabilistic characterization is not readily available. We thus aim to develop a theory to study convergence of the tangent-transform (TT) estimates that is free of such probabilistic characterization. However, a case by case analysis is required for the generalization of the theory to other non-conjugate models.

It may also appear on the surface that the EM algorithm underlying tangent-transforms can be analyzed using the general sufficient conditions for convergence of the EM (we refer to the recent article (Balakrishnan et al., 2017) and the references therein for more on this topic), a careful inspection however reveals that these general-purpose conditions pose significant difficulty to verify in case of the present EM iterates and demand stringent conditions on the design matrix and other data generating parameters. Our approach, on the other hand is to directly analyze the EM sequence without resorting to any high-level results.

By viewing the EM updates as iterations in a discrete time autonomous dynamical system, we show that the EM updates converge to the desired fixed point under suitable initialization, a phenomenon known as local asymptotic stability. While local stability is typically a weaker statement compared to global convergence as it only ensures convergence if the system is initialized in a neighborhood around the fixed point, our stability result is essentially assumption-free and does not require any assumption on the design matrix, the sparsity of the coefficients, the dimension p, and the sample size n. Although the notion of such convergence is local, to the best of our knowledge, this is the first assumption-free result on the stability of a variational algorithm. The main technical contribution is to show that the spectral radius of the Jacobian matrix of the linearized operator of the EM sequence is strictly smaller than one at the fixed point. In the special case when the design matrix is semi-orthogonal, we show that the EM sequence is globally convergent with an exponential rate of convergence (logarithmic run time) independent of the initialization. We also provide a straightforward extension of this result to the case of the multinomial logit model.

1.1 A Summary of Our Contributions

The contributions in this paper are summarized as follows:

1. In §3, we derive frequentist risk bounds for the variational estimates of the model parameters based on the α-Rényi divergence for $\alpha \in (0, 1]$. Also, we argue that the risk bound for the discrepancy measure in (18) is minimax optimal. It is important to mention here that our theory to study the convergence of the tangent-transform estimates does not leverage the probabilistic characterization of the hyperparameters and thus allows for a natural extension to other classes of distributions for which such probabilistic characterization is not available.

2. In §4.1 we show that the variational updates originally developed by Jaakkola and Jordan (2000) are locally asymptotically stable under very mild conditions on the design matrix. This ensures the existence of a neighborhood around the fixed point, within which if an iteration is initialized then the system converges to the fixed point eventually. By invoking results from the dynamical systems literature, we bypass the challenge of verifying the conditions of Lyapnov's stability (Romero et al., 2020) and also the sufficient conditions for convergence of EM algorithm in Balakrishnan et al. (2017).

3. By exploiting the structure of a semi-orthogonal design matrix in a simple hierarchical logistic model, we establish global convergence rates for the variational updates in §4.2.

4. In §5, we explore the convergence of variational updates in the case of multinomial-logit regression and additionally provide algorithmic convergence results. It is worth noting that the updates in multinomial logit extension do not have an immediate probabilistic characterization. Furthermore, we extended the result to achieve global convergence in a special case of multinomial regression.

2. Tangent Transformation Approach

Denote the data by X and the likelihood conditioned on parameter $\theta \in \Theta$ by $p(X \mid \theta)$, where Θ is the parameter space. For a prior density $p(\cdot)$ on Θ, the goal of VI is to approximate the posterior $p(\theta \mid X) \propto p(\theta)\, p(X \mid \theta)$ by a member of a tractable family Γ of densities on Θ with respect to the Kullback—Leibler (KL) divergence.[1] Notationally, VI seeks to find

$$\hat{q} = \operatorname*{arg\,min}_{q \in \Gamma} \mathrm{D}(q \,\|\, p(\cdot \mid X)), \tag{1}$$

equivalently maximizing the evidence lower bound (ELBO), $\mathcal{L}(q) = \int q(\theta) \log\{p(X, \theta)/q(\theta)\}\, d\theta$ with respect to $q \in \Gamma$. Using a component-wise product structure on Γ popularly known as the mean field family (Parisi, 1988), closed-form updates of a coordinate ascent algorithm (CAVI) can be generally derived in conditionally conjugate exponential families (Blei et al., 2017). However, many non-conjugate models such as logistic regression, multinomial logit regression, graphical and, topic models, do not lead to closed-form CAVI updates, necessitating various specialized techniques (Jordan et al., 1999; Blei et al., 2017). One such approach is to introduce variational parameters to minorize the log-marginal likelihood by a tractable family, which when combined with an appropriate prior enjoys conjugate inference. For Bayesian logistic regression models, Jaakkola and Jordan (2000) introduced a tangent-transform of the logistic function using convex duality. By a standard result in convex analysis (Rockafellar, 1970), a convex function $f(\cdot)$ on $\mathbb{R}^d$ can be represented via a *conjugate* or *dual* function f^* as,

$$f(x) = \max_{\lambda}\{\langle \lambda, x\rangle - f^*(\lambda)\}, \quad f^*(\lambda) = \max_{x}\{\langle \lambda, x\rangle - f(x)\}. \tag{2}$$

One simple example of (2) is $x^2 = \max_\lambda\{\lambda x - \lambda^2/4\}$ with equality at $x = \lambda/2$. Similarly, for a concave $f(\cdot)$ we have $f(x) = \min_\lambda\{\langle \lambda, x\rangle - f^*(\lambda)\}$ with the dual being $f^*(\lambda) = \min_x\{\langle \lambda, x\rangle - f(x)\}$. Geometrically, the evaluation of a convex function at any point x can be viewed as the maxima of the uncountable collection of hyperplanes $\langle \lambda, x\rangle - f^*(\lambda)$ indexed by $\lambda \in \mathbb{R}^d$.

The usage of duality is not restricted to linear approximations, i.e., hyperplanes. In fact, Jaakkola and Jordan (2000) used a quadratic bound for the logistic function that induces conjugacy with Gaussian priors. In the following subsection, we discuss the salient features of the tangent transform approach.

[1] The KL divergence between densities f and g, denoted $\mathrm{D}(f \,\|\, g)$, is $D(f \,\|\, g) := \int f \log(f/g) d\mu$, where μ is a common dominating measure.

2.1 Convex Minorant Construction for Logistic Likelihood

We discuss a slightly general version of the tangent transform approach where we raise the usual logistic likelihood to a power $\alpha \in (0,1]$ before combining with the prior. Variational Bayes procedures with fractional likelihoods have been recently considered in Yang et al. (2020); Alquier and Ridgway (2020); Alquier et al. (2016). Note that, the case $\alpha = 1$ recovers the usual tangent transform.

Assuming we observe binary responses y_i corresponding to fixed covariates $\mathbf{x}_i \in \mathbb{R}^p$ $(i = 1,\ldots,n)$, consider the usual logistic regression model,

$$y_i \mid \mathbf{x}_i, \beta \quad \sim \quad \text{Bernoulli}(p_i), \quad p_i = \frac{1}{1+\exp(-\mathbf{x}_i^{\mathrm{T}}\beta)} \quad (i = 1,\ldots,n). \tag{3}$$

Denote by $\mathbf{X}$ the $n \times p$ covariate matrix with ith row $\mathbf{x}_i^{\mathrm{T}}$ $(i = 1,2,\ldots,n)$. Consider a Gaussian prior $\beta \sim \mathrm{N}_p(\mu_\beta, \Sigma_\beta)$, denoted by $\pi(\beta)$.

Denoting $y = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$, call the joint density of (y, β) given $\mathbf{X}$ by $p(y, \beta \mid \mathbf{X})$. For a fixed $\alpha \in (0,1]$, define the fractional likelihood (Walker and Hjort, 2001) by $p^\alpha(y \mid \mathbf{X}, \beta) = \{p(y \mid \mathbf{X}, \beta)\}^\alpha$ and denote with a slight abuse of notation, $p^\alpha(y, \beta \mid \mathbf{X}) = p^\alpha(y \mid \mathbf{X}, \beta)\pi(\beta)$,

$$p^\alpha(y, \beta \mid \mathbf{X}) \propto \exp\Big[\alpha\, y^{\mathrm{T}}\mathbf{X}\beta - \alpha \sum_{i=1}^n \log\left(1 + e^{\mathbf{x}_i^{\mathrm{T}}\beta}\right) - \frac{1}{2}(\beta - \mu_\beta)^{\mathrm{T}}\Sigma_\beta^{-1}(\beta - \mu_\beta)\Big]. \tag{4}$$

Jaakkola and Jordan (2000) begins with the following quadratic duality result that holds for all $x \in \mathbb{R}$:

$$\begin{aligned} &-\log\{1 + \exp(x)\} = \max_{t \in \mathbb{R}}[A(t)x^2 - x/2 + C(t)], \\ &A(t) = -\tanh(t/2)/4t, \quad C(t) = t/2 - \log\{1 + \exp(t)\} + t\tanh(t/2)/4. \end{aligned}$$

We can then bound $\log p^\alpha(y, \beta \mid \mathbf{X})$ from below by $\log p_l^\alpha(y, \beta \mid \mathbf{X}, \xi)$, where

$$\begin{aligned} \log p_l^\alpha(y, \beta \mid \mathbf{X}, \xi) = &-\frac{1}{2}\beta^{\mathrm{T}}\left[\Sigma_\beta^{-1} - 2\alpha\mathbf{X}^{\mathrm{T}}\mathrm{diag}\{A(\xi)\}\mathbf{X}\right]\beta + \Big\{\alpha\Big(y - \frac{1}{2}\mathbb{1}_n\Big)^{\mathrm{T}}\mathbf{X} + \mu_\beta^{\mathrm{T}}\Sigma_\beta^{-1}\Big\}\beta \\ &- \mu_\beta^{\mathrm{T}}\Sigma_\beta^{-1}\mu_\beta + \alpha\mathbb{1}_n^{\mathrm{T}}C(\xi) + \text{Constant}. \end{aligned} \tag{5}$$

In the above display, $\xi = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$ collectively denotes all variational parameters, with ξ_i appearing from applying the previous duality result for $-\log\{1 + \exp(\mathbf{x}_i^{\mathrm{T}}\beta)\}$. Also, $\mathrm{diag}\{A(\xi)\}$ is a $n \times n$ diagonal matrix with diagonal entries $\{A(\xi_1), A(\xi_2), \ldots, A(\xi_n)\}$ and $C(\xi) = \{C(\xi_1), \ldots, C(\xi_n)\}^{\mathrm{T}}$.

Since $p_l^\alpha(y, \beta \mid \mathbf{X}, \xi)$ serves as a lower bound to $p^\alpha(y, \beta \mid \mathbf{X})$ for any $\xi \in \mathbb{R}^n$, similar to Jaakkola and Jordan (2000) we use an empirical Bayes approach to estimate the variational parameters ξ by maximizing $p_l^\alpha(y \mid \mathbf{X}, \xi) = \int p_l^\alpha(y, \beta \mid \mathbf{X}, \xi)d\beta$ with respect to ξ. The true posterior distribution of β in (4) is not available in closed form. However, assuming (5) to be a working (pseudo)-likelihood of y, β given $\mathbf{X}, \xi$, it is straightforward to see that the corresponding conditional posterior distribution of β is $\mathrm{N}(\mu_\alpha(\xi), \Sigma_\alpha(\xi)/\alpha)$ where

$$\Sigma_\alpha^{-1}(\xi) = \Sigma_\beta^{-1}/\alpha - 2\mathbf{X}^{\mathrm{T}}\mathrm{diag}\{A(\xi)\}\mathbf{X}, \quad \mu_\alpha^{\mathrm{T}}(\xi)\Sigma_\alpha^{-1}(\xi) = \Big(y - \frac{1}{2}\mathbb{1}_n\Big)^{\mathrm{T}}\mathbf{X} + \mu_\beta^{\mathrm{T}}\Sigma_\beta^{-1}/\alpha. \tag{6}$$

Treating β as latent variables and augmenting with y to get the complete data, one obtains the E-step,

$$\begin{aligned} Q_\alpha(\xi^{t+1} \mid \xi^t) &= \mathbb{E}_{\beta \mid y, \xi^t, \mathbf{X}} \left[\log p_l^\alpha(y, \beta \mid \xi^{t+1}, \mathbf{X})\right] \qquad (7) \\ &= \mathrm{tr}\left[\alpha \mathbf{X}^{\mathrm{T}} \mathrm{diag}\{A(\xi^{t+1})\} \mathbf{X} \{\Sigma_\alpha(\xi^t)/\alpha + \mu_\alpha(\xi^t)\mu_\alpha^{\mathrm{T}}(\xi^t)\}\right] + \alpha \mathbb{1}_n^{\mathrm{T}} C(\xi^{t+1}) + \text{Constant}, \end{aligned}$$

where $\mathrm{tr}(A)$ denotes the trace of a matrix A. Upon differentiating the above expression with respect to ξ^{t+1} and using the fact that $C'(x) = -x^2 A'(x)$, we get the M-step,

$$(\xi^{t+1})^2 = \mathrm{diag}[\mathbf{X}\{\Sigma_\alpha(\xi^t)/\alpha + \mu_\alpha(\xi^t)\mu_\alpha^{\mathrm{T}}(\xi^t)\}\mathbf{X}^{\mathrm{T}}]. \qquad (8)$$

The square operation in the above display is to be interpreted elementwise. We assume convergence when the increment in $p_l^\alpha(y \mid \mathbf{X}, \xi)$ is negligible which implies convergence of ξ^t by virtue of EM algorithm. The EM sequence in (8) is recognized to be a fixed point iteration corresponding to the fixed point equation given by,

$$(\xi^*)^2 = \mathrm{diag}[\mathbf{X}\{\Sigma_\alpha(\xi^*)/\alpha + \mu_\alpha(\xi^*)\,\mu_\alpha^{\mathrm{T}}(\xi^*)\}\mathbf{X}^{\mathrm{T}}]. \qquad (9)$$

Assuming (8) converges to a fixed point ξ^*, $\mu_\alpha(\xi^*)$ gives the variational estimate of β. We will refer to the above algorithm as α-VB TT henceforth.

3. Statistical Optimality of the Variational Estimate

In this section we develop a rigorous framework to obtain frequentist risk bounds of the variational approximation obtained in (6) at any fixed point ξ^* of (8). Throughout the section, we assume that the data is generated from a logistic regression model

$$p(y \mid \beta^{\mathbf{o}}, \mathbf{X}) = \exp\left[y^{\mathrm{T}}\mathbf{X}\beta^{\mathbf{o}} - \sum_{i=1}^{n} \log\left(1 + e^{\mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}}}\right)\right]. \qquad (10)$$

It is not immediately clear whether the empirical likelihood based inference of ξ as discussed in Section 2 falls into the framework of variational inference in the sense of (1). In the following, we propose an objective function whose minimizer satisfies the fixed point iteration (8). Let our *working model* be

$$p_l^\alpha(y \mid \beta, \mathbf{X}, \xi) \quad = \exp\left\{\alpha\Big(y^{\mathrm{T}}\mathbf{X}\beta + \beta^{\mathrm{T}}\left[\mathbf{X}^{\mathrm{T}}\mathrm{diag}\{A(\xi)\}\mathbf{X}\right]\beta - 0.5\mathbb{1}_n^{\mathrm{T}}\mathbf{X}\beta + \mathbb{1}_n^{\mathrm{T}}C(\xi)\Big)\right\}. \quad (11)$$

It is important to note here that $p_l^\alpha(y \mid \beta, \mathbf{X}, \xi)$ is not a probability density, even when $\alpha = 1$. Let $\mathcal{F}$ be the set of densities on $\mathbb{R}^p$. Define a mapping $\mathcal{L} : \mathcal{F} \times \mathbb{R}^n$ to $\mathbb{R}$ as

$$\mathcal{L}(q, \xi) = -\int \log \frac{p_l^\alpha(y, \beta \mid \mathbf{X}, \xi)}{q(\beta)} q(\beta) d\beta, \qquad (12)$$

where $p_l^\alpha(y, \beta \mid \mathbf{X}, \xi)$ is defined in (5). Observe that $\mathcal{L}(q, \xi)$ is the negative of the evidence lower bound obtained in a variational inference with (11) as the working likelihood, $\mathrm{N}_p(\mu_\beta, \Sigma_\beta)$ the prior on β, and variational family $\mathcal{F} \times \{\delta_\xi : \xi \in \mathbb{R}^n\}$ where δ_ξ is the Dirac delta measure on $\xi \in \mathbb{R}^n$. In Lemma 1, we show that the tangent transform algorithm maximizes $-\mathcal{L}(q, \xi)$.

Lemma 1 *Any minimizer* (q^*, ξ^*) *of* (12) *over* $\mathcal{F} \times \mathbb{R}^n$ *satisfies*

$$q^* = N_p\{\mu_\alpha(\xi^*), \Sigma_\alpha(\xi^*)/\alpha\}, \quad (\xi^*)^2 = diag[\mathbf{X}\{\Sigma_\alpha(\xi^*)/\alpha + \mu_\alpha(\xi^*)\,\mu_\alpha^{\mathrm{T}}(\xi^*)\}\mathbf{X}^{\mathrm{T}}], \tag{13}$$

where $\mu_\alpha(\xi), \Sigma_\alpha(\xi)$ *are defined in* (6).

Proof We start the proof by re-writing (12) as

$$\mathcal{L}(q, \xi) = -\int q(\beta) \log p_l^\alpha(y, \beta \mid \mathbf{X}, \xi) d\beta + \int q(\beta) \log\{q(\beta)\} d\beta. \tag{14}$$

To minimize (14) jointly with respect to (q, ξ), we set up the first order stationarity conditions. We first set the gradient of $\mathcal{L}(q, \xi)$ with respect to ξ to zero holding q fixed. As the second term in (14) is independent of ξ, this is equivalent to setting the gradient of $\mathbb{E}_q\big[\log p_l^\alpha(y, \beta \mid \xi, \mathbf{X})\big]$ with respect to ξ to be zero,

$$\frac{\partial}{\partial \xi} \mathbb{E}_q \left[\log p_l^\alpha(y, \beta \mid \xi, \mathbf{X})\right] = 0. \tag{15}$$

By interchanging the integration and differentiation, (15) is equivalent to

$$\mathbb{E}_q \left[\frac{\partial}{\partial \xi} \log p_l^\alpha(y, \beta \mid \xi, \mathbf{X})\right] = 0. \tag{16}$$

For fixed ξ, to maximize (12), we simply apply Lemma 13 in the appendix. This leads to the optimal choice of $q(\beta)$ being the conditional distribution $p_l^\alpha(\beta \mid y, \xi, \mathbf{X})$ which is $\mathrm{N}_p(\mu_\alpha(\xi), \Sigma_\alpha(\xi)/\alpha)$. This when combined with (16) yields

$$\mathbb{E}_{\mathrm{N}_p(\mu_\alpha(\xi), \Sigma_\alpha(\xi)/\alpha)} \left[\frac{\partial}{\partial \xi} \log p_l^\alpha(y, \beta \mid \xi, \mathbf{X})\right] = 0. \tag{17}$$

To show that the solution of (17) satisfies (9), recall that the first-order stationarity condition for maximizing $Q_\alpha(\xi^{t+1} \mid \xi^t)$ in (7) with respect to ξ^{t+1} is given by

$$\frac{\partial}{\partial \xi^{t+1}} Q_\alpha(\xi^{t+1} \mid \xi^t) = \mathbb{E}_{\beta \mid y, \xi^t, \mathbf{X}} \left[\frac{\partial}{\partial \xi^{t+1}} \log p_l^\alpha(y, \beta \mid \xi^{t+1}, \mathbf{X})\right] = 0,$$

which in turn is equivalent to solving the fixed point iteration $(\xi^{t+1})^2 = \mathrm{diag}[\mathbf{X}\{\Sigma_\alpha(\xi^t)/\alpha + \mu_\alpha(\xi^t)\,\mu_\alpha^{\mathrm{T}}(\xi^t)\}\mathbf{X}^{\mathrm{T}}]$. Thus the solution to (17) satisfies $(\xi^*)^2 = \mathrm{diag}[\mathbf{X}\{\Sigma_\alpha(\xi^*)/\alpha + \mu_\alpha(\xi^*)\,\mu_\alpha^{\mathrm{T}}(\xi^*)\}\mathbf{X}^{\mathrm{T}}]$. ■

Although (12) is reminiscent of the α-variational objective function of Yang et al. (2020), we note a couple of key differences : (a) $p_l^\alpha(y \mid \beta, \xi, \mathbf{X})$ is not a valid probability density, but it is a lower bound to $p^\alpha(y \mid \beta, \mathbf{X})$, (b) The latent variables ξ lack a probabilistic interpretation as in Yang et al. (2020), where one recovers the original likelihood after marginalization over the latent variables. Here, the latent variables instead correspond to tuning parameters appearing from convex duality.

The usage of the fractional likelihood for $\alpha \in (0, 1)$ results in only minor changes from a methodological and implementation perspective. However, from a theoretical perspective, like Yang et al. (2020), $\alpha \in (0, 1)$ requires fewer assumptions to deliver optimal risk bounds.

3.1 Variational Risk Bounds

In the following, we develop risk bounds for the variational estimator separately for the case $\alpha \in (0,1)$ and $\alpha = 1$. In the former case, to quantify the discrepancy between the variational estimate and the true parameter, we use an α-Rényi divergence

$$\mathrm{D}_\alpha(\beta, \beta^{\mathbf{o}}) = \frac{1}{n(\alpha-1)} \log \sum_{y \in \{0,1\}} \left\{ \frac{p(y \mid \beta, \mathbf{X})}{p(y \mid \beta^{\mathbf{o}}, \mathbf{X})} \right\}^{\alpha} p(y \mid \beta^{\mathbf{o}}, \mathbf{X}). \tag{18}$$

Refer to Bhattacharya et al. (2019) for more on posterior risk bounds under the α-Rényi divergence. The factor $(1/n)$ is used to measure *average discrepancy* per observation. We can further simplify (18) to

$$\mathrm{D}_\alpha(\beta, \beta^{\mathbf{o}}) = \frac{1}{n(\alpha-1)} \sum_{i=1}^{n} \log \left[p_{i,\beta}^{\alpha} p_{i,\beta^{\mathbf{o}}}^{1-\alpha} + (1-p_{i,\beta})^{\alpha}(1-p_{i,\beta^{\mathbf{o}}})^{1-\alpha} \right],$$

where $p_{i,\beta} = 1/\{1+\exp(-\mathbf{x}_i^{\mathrm{T}}\beta)\}$. The next theorem derives an upper bound to the risk obtained by integrating the α-Rényi divergence with respect to the optimal variational solution. Denote by $\phi_p(x; \mu, \Sigma)$ the p-dimensional multivariate Gaussian density evaluated at $x \in \mathbb{R}^p$, with mean μ and variance covariance matrix Σ. Let $\|\mathbf{X}\|_{2,\infty} = \max\{\|\mathbf{x}_i\|, i = 1, \ldots, n\}$ and $\|\mathbf{X}\|_\infty := \max\{|x_{ij}|, i = 1, \ldots, n, j = 1, \ldots, p\}$. Let $L(\beta^{\mathbf{o}}, \mathbf{X}) = \max\{4\|\mathbf{X}\|_{2,\infty}, 8\|\mathbf{X}\|_{2,\infty}^2 \|\beta^{\mathbf{o}}\|_2\}$.

Theorem 2 *For any $\varepsilon \in (0,1)$, with probability $(1-\varepsilon) - 1/\{(D-1)^2\, n\, \varepsilon^2\}$ under* (10)

$$\begin{aligned}(1-\alpha)\int D_\alpha(\beta, \beta^{\mathbf{o}}) \phi_p\{\beta; \mu_\alpha(\xi^*), \Sigma_\alpha(\xi^*)\} d\beta \quad &\leq \quad D\alpha\,\varepsilon^2 + \frac{p}{n}\log\left\{\frac{L(\beta^{\mathbf{o}}, \mathbf{X})}{\varepsilon^2}\right\} + \\ & \quad\quad C_n(\beta^{\mathbf{o}}, \mu_\beta, \Sigma_\beta) + \frac{1}{n}\log\left(\frac{1}{\varepsilon}\right)\end{aligned}$$

for some constant $D > 0$, where

$$C_n(\beta^{\mathbf{o}}, \mu_\beta, \Sigma_\beta) = \frac{1}{2n}(\beta^{\mathbf{o}} - \mu_\beta)^{\mathrm{T}} \Sigma_\beta^{-1} (\beta^{\mathbf{o}} - \mu_\beta).$$

The proof of Theorem 2 can be found in §A.1 in the appendix.

Remark 3 *(a) It is important to mention here that α in* (18) *and* (12) *need not be same. One can use the inequality $\alpha(1-\widetilde{\alpha})/\{\widetilde{\alpha}(1-\alpha)\}D_{\widetilde{\alpha}} \leq D_\alpha \leq D_{\widetilde{\alpha}}$ for $0 < \alpha \leq \widetilde{\alpha} < 1$ (Van Erven and Harremos, 2014), to generalize the above theorem for any D_γ such that $\gamma \in (0,1)$.(b) Setting $\varepsilon^2 = p\log n/n$, the risk bound for discrepancy D_α is p/n up to logarithmic terms which is near-minimax optimal. Please refer to Bhattacharya et al. (2019) and the references therein for further readings on this topic. The explicit bound is non-asymptotic and depends on prior parameters, the covariate matrix $\mathbf{X}$ and the true data generating density.*

Next, we separately deal with the case $\alpha = 1$. In doing so, we work with a limiting metric of α-Rényi divergence as α tends to 1. Let $a(t) = \log(1+e^t)$ and $a^{(1)}$ and $a^{(2)}$ denote the first and second derivatives. a satisfies $a(t+h) \geq a(t) + h\, a^{(1)}(t) + \mathrm{r}(|h|)\, a^{(2)}(t)/2$ for all t, h, where $r(h) = h^2/(\mathrm{r}_1 h + 1)$ for $\mathrm{r}_1 > 0$. Define

$$\mathrm{D}(\beta^{\mathbf{o}}, \beta) := \frac{1}{n}\mathbb{E}_{\beta^{\mathbf{o}}}\Big\{ \log \frac{p(y \mid \beta^{\mathbf{o}}, \mathbf{X})}{p(y \mid \beta, \mathbf{X})} \Big\} = \frac{1}{n}\sum_{i=1}^{n} \big\{ a(\mathbf{x}_i^{\mathrm{T}}\beta) - a(\mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}}) - a^{(1)}(\mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}})\mathbf{x}_i^{\mathrm{T}}(\beta - \beta^{\mathbf{o}}) \big\}.$$

The last term in the above display follows from the fact that, $\mathbb{E}(y) = a^{(1)}(\mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}})$ under (10). $\mathrm{D}(\beta^{\mathbf{o}}, \beta)$ is the KL divergence between $p(\cdot \mid \beta^{\mathbf{o}}, \mathbf{X})$ and $p(\cdot \mid \beta, \mathbf{X})$. Let $W = \mathrm{diag}\{a^{(2)}(\mathbf{x}_1^{\mathrm{T}}\beta^{\mathbf{o}}), \ldots, a^{(2)}(\mathbf{x}_n^{\mathrm{T}}\beta^{\mathbf{o}})\}$ and let $\kappa_1 = \lambda_p(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n)$, $\kappa_2 = \lambda_1(\mathbf{X}^{\mathrm{T}}W\mathbf{X}/n)$, where $\lambda_j(A)$ denotes the jth largest eigen value of a positive definite matrix A and $\zeta_p = \|\mathbf{X}\|_\infty \sqrt{np\log p}/2$.

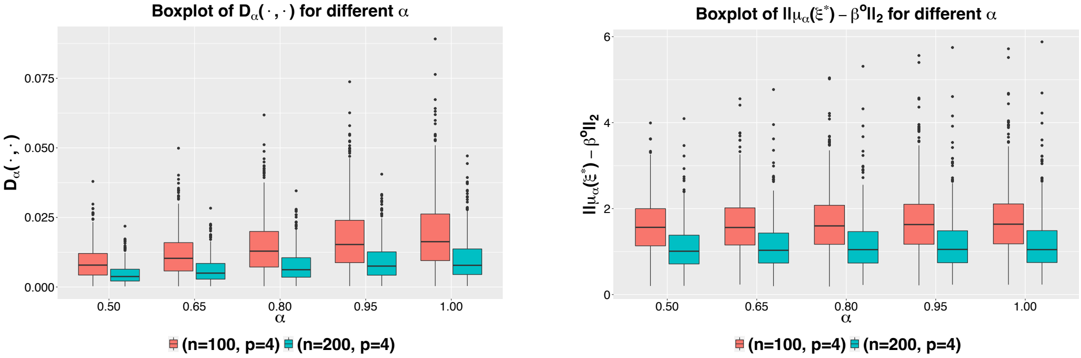
Theorem 4 *Fix $\gamma \in (0,1)$ and set $\widetilde{\varepsilon} = (n\kappa_2 + 2\zeta_p \mathrm{r}_1 \sqrt{p}\|\mathbf{X}\|_\infty)/\{\mathrm{r}_1\|\mathbf{X}\|_\infty\sqrt{p}\,(n\kappa_2 - 2\zeta_p \mathrm{r}_1\|\mathbf{X}\|_\infty\sqrt{p})\}$, for any $\varepsilon \geq 2\widetilde{\varepsilon}$, set $\epsilon = \kappa_2\varepsilon/(8\mathrm{r}_1\|\mathbf{X}\|_\infty\sqrt{p})$. If $\sqrt{n} \geq \mathrm{r}_1 p\sqrt{\log p}\|\mathbf{X}\|_\infty^2/\kappa_2$ and both $\kappa_1, \kappa_2 > 0$, then with probability $1 - 2/p - e^{-n\gamma\epsilon/8} - e^{-n\kappa_1\varepsilon^2/32} - 1/\{(D-1)^2\, n\,\varepsilon^2\}$ under* (10),

$$\begin{aligned}\frac{\gamma}{4}\int D(\beta^{\mathbf{o}}, \beta)\phi_p\{\beta; \mu(\xi^*), \Sigma(\xi^*)\}d\beta \leq \frac{1}{n}\log 3 \quad &+ \quad (D + \gamma\kappa_1/16)\,\varepsilon^2 + \frac{p}{n}\log\Big\{\frac{L(\beta^{\mathbf{o}}, \mathbf{X})}{\varepsilon^2}\Big\} \\ &+ \quad C_n(\beta^{\mathbf{o}}, \mu_\beta, \Sigma_\beta).\end{aligned}$$

The proof of Theorem 4 can be found in §A.2 in the appendix.

Remark 5 *(a) To obtain a risk bound, we keep γ to be a fixed number in $(0,1)$ and set $\varepsilon^2 = p\log n/n$. Then KL divergence risk is p/n (up to logarithmic terms) which is again minimax optimal (́Bhattacharya et al., 2019). As opposed to Theorem 2, Theorem 4 requires the eigenvalues of $\mathbf{X}^{\mathrm{T}}W\mathbf{X}/n$ to be bounded from below and the eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}/n$ to be bounded from above. (b) Raising the likelihood with a fractional power is an effective theoretical tool to reduce the complexity of assumptions in deriving the variational risk. In a well-specified model, choosing $\alpha = 1$ will lead to the best risk bound on average. However, in absence of the knowledge of the data generation mechanism, one can resort to optimizing appropriate inferential goals to choose α. For instance, one can choose α so that the corresponding credible region achieves the nominal frequentist coverage probability; refer to Syring and Martin (́2019) for a data-driven procedure to set α.*

Now we conduct a numerical study to empirically support the conclusions of the theorems above. For fixed (n,p) we construct a $n \times p$ design matrix $\mathbf{X}$ where $\mathbf{x}_i^{\mathrm{T}}$ $(i = 1, \ldots, n)$ are independently drawn from $\mathrm{N}_p\big(0, (0.5\,\mathbb{I}_p + 0.5\,\mathbb{1}_p\mathbb{1}_p^{\mathrm{T}})\big)$. We then normalize each row of $\mathbf{X}$ by $\sqrt{p}$. We fix $\beta^{\mathbf{o}}$ to be $\{-4, 4, 4, -4\}$ and generate $y_i \sim \mathrm{Bernoulli}(p_i)$ with $p_i = 1/\{1 + \exp(-\mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}})\}$, independently for $i \in \{1, 2, \ldots, n\}$. We place a zero-mean Gaussian prior $\beta \sim \mathrm{N}_p(0_p, \Sigma_\beta)$ and set $\Sigma_\beta = 5^2\mathbb{I}_p$. Given a dataset $(y, \mathbf{X})$ and fixed $\alpha \in \{0.50, 0.65, 0.80, 0.95, 1.00\}$ we calculate the fixed point solution ξ^* using (8) with tolerance 10^{-5}. We use $p(y \mid \mu_\alpha(\xi^*), \mathbf{X})$ as the final estimated density and calculate the discrepency $\mathrm{D}_\alpha(\mu_\alpha(\xi^*), \beta^{\mathbf{o}})$ with $p(y \mid \beta^{\mathbf{o}}, \mathbf{X})$. In panel (a) we plot $\mathrm{D}_\alpha(\mu_\alpha(\xi^*), \beta^{\mathbf{o}})$ for $\alpha \in \{0.50, 0.65, 0.80, 0.95\}$ along with $\mathrm{D}(\beta^{\mathbf{o}}, \mu(\xi^*))$ that corresponds to $\alpha = 1$. In panel (b) we plot the ℓ_2 norm between $\mu_\alpha(\xi^*)$ and $\beta^{\mathbf{o}}$. We repeat this process for 500 independent samples with $(n = 100, p = 4)$ and $(n = 200, p = 4)$. Clearly, increasing the sample size leads to improved estimation as seen from either panel of Figure 1. Also, $\mathrm{D}_\alpha(\mu_\alpha(\xi^*), \beta^{\mathbf{o}})$ slightly increases as α increases to 1. Since α-Rényi divergence counterbalances the effect of the misspecified likelihood and reinforces concentration around the truth, this behavior is expected. It is important to mention here that the discrepancy measure $\mathrm{D}_\alpha(\mu_\alpha(\xi^*), \beta^{\mathbf{o}})$ changes with α, and it can not be used to measure the accuracy of the variational estimates themselves. However, this issue is addressed in panel (b) where we use ℓ_2-norm as a measure of the discrepancy.



(a)

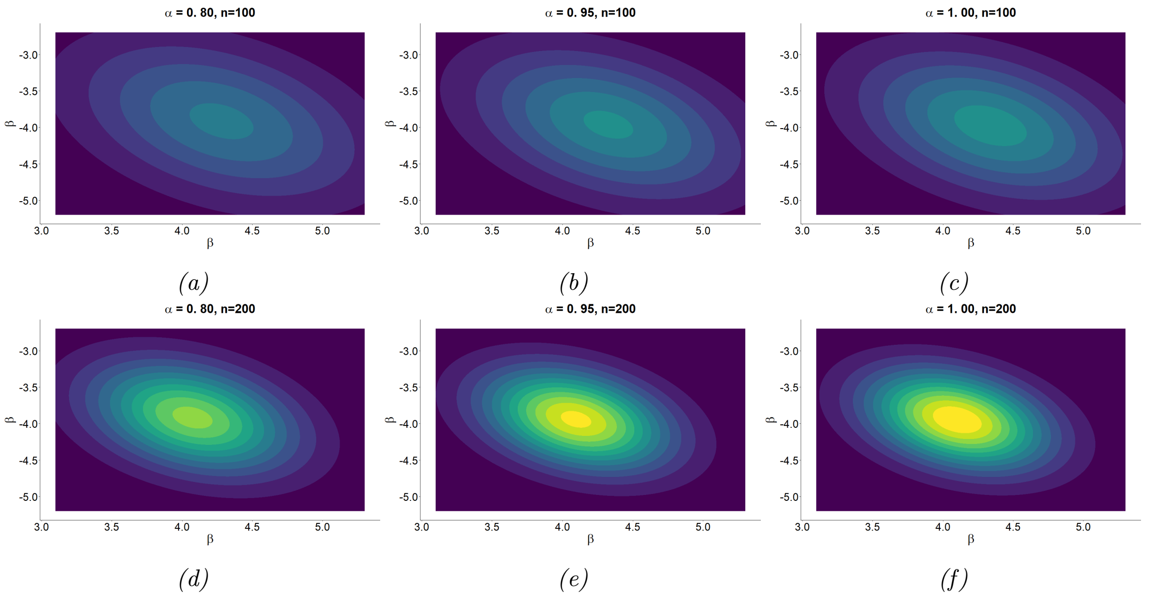
(b)

Figure 1: (a) Boxplot of $D_\alpha(\mu_\alpha(\xi^*), \beta^{\mathbf{o}})$ for $\alpha \in (0, 1)$ and $D(\beta^{\mathbf{o}}, \mu(\xi^*))$ for $\alpha = 1$ (b) Boxplot of $\|\mu_\alpha(\xi^*) - \beta^{\mathbf{o}}\|_2$ for different values of $\alpha \in (0, 1]$

In Figure 2 we show the contour plots of the marginal of (β_2, β_4) obtained from the variational approximation $q^* = \mathrm{N}_p\{\mu_\alpha(\xi^*), \Sigma_\alpha(\xi^*)\}$ for a given dataset $(y, \mathbf{X})$. The upper and lower panels correspond to $(n = 100, p = 4)$ and $(n = 200, p = 4)$ respectively with different $\alpha \in \{0.80, 0.95, 1.00\}$. Clearly the concentration of the approximate posterior around the truth $(4, -4)$ increases as α tends to 1. Also concentration increases with the increase in the sample size. Further, the variational approximations appear to be almost similar for $\alpha = 0.95$ and $\alpha = 1$. This is important since the conditions required to achieve the variational risk for $\alpha < 1$ are much milder than at $\alpha = 1$.

3.2 Related Work

PAC-Bayesian inequalities (McAllester, 1999; Seeger, 2002; Catoni, 2003; Maurer, 2004); see Alquier (2021) for a comprehensive review; and Theorem 2 both use the variational inequality to provide sharp bounds to the posterior / variational risk. Traditional PAC-Bayes inequalities, e.g. Theorem 2.1 in Alquier and Ridgway (2020) or Proposition 2.1 in [Catoni, "Lecture notes for the IFCAM Summer School on Applied Mathematics", Indian Institute of Science. Bangalore, 2014] are not applicable here since tangent transformation (TT) involves optimizing over an additional parameter ξ which is not present in the risk $\mathrm{D}_\alpha(\beta, \beta^{\mathbf{o}})$, but plays a critical role in minimizing the upper bound. Realizing the fact that $p_l(y \mid \beta, \xi, \mathbf{X})$ minorizes $p(y \mid \beta, \mathbf{X})$ for any $\xi \in \mathbb{R}^+$, we connect the variational risk with a suitably defined variational objective function in (12) involving ξ. Our primary contribution is to recognize that this objective function when minimized over variational family $\mathcal{F} \times \{\delta_\xi : \xi \in \mathbb{R}^n\}$ leads to the TT algorithm. From a theoretical point of view,



(a) *(b)* *(c)*

(d) *(e)* *(f)*

Figure 2: Contour plot of posterior distribution of (β_2, β_4) *given by* $N_p\{\mu_\alpha(\xi^*), \Sigma_\alpha(\xi^*)\}$ *for different* α *and* n*. The upper row ((a)—(c)) corresponds to* $(n = 100, p = 4)$ *and the lower row ((d)—(f)) corresponds to* $(n = 200, p = 4)$.

careful selection of ξ aids in controlling the Jensen gap in the in the upper bound which in turn delivers an optimal variational risk bound. However, Alquier and Ridgway (2020) studied contraction in logit models using *PAC-Bayes* inequalities (Corollary 3.3). While Theorem 2 is developed under a fixed design, Alquier and Ridgway (2020) assumes the $\mathbf{X}$ to be random. Moreover, Alquier and Ridgway (2020) restricts the variational family to be Gaussian whereas no such condition is imposed in our case. One can analyze the above theorem in a random design setting as well. In that case Theorem 2 remains unchanged except $L(\beta^{\mathbf{o}}, \mathbf{X})$ on the upper bound of risk would be replaced by $\tilde{\mathcal{E}}_{\beta^{\mathbf{o}},\mathbf{X}} := \max\{2\mathbf{E}^{1/4}(\|\mathbf{X}\|_2^4), 8\mathbf{E}^{1/2}(\|\mathbf{X}\|_2^4)\,\|\beta^{\mathbf{o}}\|_2, 4\mathbf{E}^{1/2}(\|\mathbf{X}\|_2^2), 2\mathbf{E}^{1/8}(\|\mathbf{X}\|_2^8), 4\mathbf{E}^{1/4}(\|\mathbf{X}\|_2^8)\,\|\beta^{\mathbf{o}}\|_2\}$, which follows from working with $\mathbf{E}_{y,\mathbf{X}}[\Delta(\beta, \beta^{\mathbf{o}})]$ and $V_{y,\mathbf{X}}[\Delta(\beta, \beta^{\mathbf{o}})]$ in (37). One can show that $\mathbf{E}_{y,\mathbf{X}}[\Delta(\beta, \beta^{\mathbf{o}})] \leq n\varepsilon^2$ and $V_{y,\mathbf{X}}[\Delta(\beta, \beta^{\mathbf{o}})] \leq \mathbf{E}_{y,\mathbf{X}}[\Delta^2(\beta, \beta^{\mathbf{o}})] \leq n\varepsilon^2$ whenever $\tilde{\mathcal{E}}_{\beta^{\mathbf{o}},\mathbf{X}}\|\beta - \beta^{\mathbf{o}}\|_2 \leq \varepsilon^2$. This immediately imposes a finite eigth order moment condition on $\mathbf{X}$ whereas Corollary 3.3 of Alquier and Ridgway (2020) only requires a finite second order moment condition. As a special case, assuming the entries $\mathbf{X} \sim \mathrm{N}(0, 1)$, $\mathbf{E}\|\mathbf{X}\|_2^{2k} = 2^k\Gamma(k + p/2)/\Gamma(p/2)$. Hence, $\tilde{\mathcal{E}}_{\beta^{\mathbf{o}},\mathbf{X}} \leq \max\{4\sqrt{p}, 16p\,\|\beta^{\mathbf{o}}\|_2\}$ which leads to an upper bound $\mathcal{O}(p\log(p\|\beta^{\mathbf{o}}\|_2)/n)$ in Theorem 2, while Alquier and Ridgway (2020) achieved a rate $\mathcal{O}(p\log(n^2\sqrt{p})/n)$. In both cases it matches the minimax rate of contraction p/n upto logarithmic terms. Furthermore, both approaches lead to an upper bound that scales similarly with respect to $\|\beta^{\mathbf{o}}\|_2$ given by $\mathcal{O}(\|\beta^{\mathbf{o}}\|_2^2/n)$.

Atchadé (2017) studied contraction of quasi-posterior distribution in sparse logistic regression and hence could not be directly applied to prove the Theorem above. A more direct comparison is possible with a recent work by Bhattacharya and Pati (2020) which provides posterior contraction in generalized linear models (Theorem 2) and as a consequence imposes restrictions on the data generating process that is similar to Theorem 4.

4. Stability and Convergence of Tangent Transform Algorithm

We provide a brief review of stability of dynamical systems here; a more detailed review and relevant references can be found in §B of the appendix. Consider the following discrete-time autonomous system,

$$\psi^{t+1} = f(\psi^t), \quad t \in \mathbb{N}, \tag{19}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ (or, $f : \mathbb{D} \to \mathbb{R}^n, \mathbb{D} \subseteq \mathbb{R}^n$) is a twice continuously differentiable function and ψ^t is the iteration at the t^{th} time-point. Any $\psi^* \in \mathbb{R}^n$ satisfying $\psi^* = f(\psi^*)$ is called a fixed point for this system. A fixed point ψ^* of (19) is called *locally asymptotically stable* if given any $\epsilon > 0$, there exists $\delta = \delta(\epsilon)$ such that whenever $\|\psi^0 - \psi^*\| < \delta$, we have $\|f(\psi^t) - \psi^*\| < \epsilon$ for all t and $\lim_{t\to\infty} \|\psi^t - \psi^*\| = 0$.

The following well-known result is instrumental to show that a fixed point is locally asymptotically stable. Denote by $\rho(\mathbf{J})$ the spectral radius of a square matrix $\mathbf{J}$, the largest eigenvalue of $\mathbf{J}$ in absolute value. Refer to Theorem 4 in Barbarossa (2011) for the following lemma.

Lemma 6 *Let ψ^* be a fixed point solution to the discrete-time autonomous system given by $\psi_{t+1} = f(\psi_t)$. Suppose, $f : \mathbb{D} \to \mathbb{R}^n(\mathbb{D} \subseteq \mathbb{R}^n)$ is a twice continuously differentiable function around a neighborhood $\mathbb{D}$ of ψ^*. Let $\mathbf{J} = [\partial_i f(\psi)/\partial\psi_j]_{\psi=\psi^*}$ be the Jacobian matrix of f evaluated at ψ^*. Then, ψ^* is locally asymptotically stable if $\rho(\mathbf{J})$ is less than 1.*

4.1 Asymptotic Stability of Tangent Transform EM

In this subsection, we study the EM sequence of ξ from equation (8) viewed as a discrete time dynamical system in ξ^2. As noted above, the convergence and stability aspects of the system depends crucially on the properties of the Jacobian of the map. Since the function $A(\xi) := -\tanh(\xi/2)/4\xi = \{1-\exp(\xi)\}/[4\xi\{1+\exp(\xi)\}]$ is symmetric around 0 (follows from the definition), and $\Sigma_\alpha(\xi)$ and $\mu_\alpha(\xi)$ are dependent on ξ through $A(\cdot)$, only the magnitude of ξ is relevant and hence we will discuss the nature of EM iterates on $\mathbb{R}^+$. The properties of the function $A(\cdot)$ play a crucial role in such an analysis. In particular, it can be shown that the function $A : \mathbb{R}^+ \to \mathbb{R}^-$ is monotonically increasing and twice continuously differentiable with $A(0) = -1/8$ and $A(\xi)+\xi A'(\xi) < 0$ for all $\xi \in \mathbb{R}^+$ (see Proposition 26 in the appendix).

In Theorem 7 below, we show that the EM sequence in equation (8) is locally asymptotically stable.

Theorem 7 *Suppose the design matrix $\mathbf{X}$ does not have any row equal to the zero vector. For any $\alpha \in (0, 1]$ and positive definite Σ_β, any fixed point solution ξ^* of the EM sequence in (8) is locally asymptotically stable.*

Proof In light of Lemma 6, one needs to check the spectral radius of the Jacobian of the system at the fixed point to prove Theorem 7. We present an outline of the proof here; refer to §C.1 in the appendix for a complete proof. Given positive semi-definite matrices A, B of the same dimension, we follow the usual convention to denote $B \prec A$ (resp. $B \precsim A$) to mean $(A - B)$ is positive definite (resp. positive semi-definite).

$\Sigma_\alpha(\xi^*) = [\Sigma_\beta^{-1}/\alpha - 2\mathbf{X}^{\mathrm{T}}\mathrm{diag}\{A(\xi^*)\}\mathbf{X}]^{-1}$ is positive definite since $\Sigma_\beta^{-1}/\alpha - 2\mathbf{X}^{\mathrm{T}}\mathrm{diag}A(\xi^*)\}\mathbf{X}$ is positive definite as $A(\xi) := -\tanh(\xi/2)/4\xi \in \mathbb{R}^-$ for all $\xi \in \mathbb{R}^+$ and Σ_β is positive definite. Now, for any $\mathbf{x}_i^{\mathrm{T}} \neq 0$ $(i = 1, 2, \ldots, n)$, one can conclude $\xi_i^* > 0$ from (9). Next, we show that the Jacobian matrix evaluated at the fixed point ξ^* can be analytically expressed as

$$\mathbf{J}_\alpha = [\mathbf{X}\Sigma_\alpha(\xi^*)\mathbf{X}^{\mathrm{T}} \circ \mathbf{X}\{\Sigma_\alpha(\xi^*)/\alpha + 2\mu_\alpha(\xi^*)\,\mu_\alpha^{\mathrm{T}}(\xi^*)\}\mathbf{X}^{\mathrm{T}}]\,\mathrm{D}, \tag{20}$$

where $\circ$ denotes the Hadamard (or, elementwise) product, $\mu_\alpha(\xi), \Sigma_\alpha(\xi)$ are defined in (6), and $\mathrm{D} = \mathrm{diag}\{A'(\xi^*)/\xi^*\}$, where the / operation is to be interpreted elementwise. By similarity, $\mathbf{J}_\alpha$ and

$$\tilde{\mathbf{J}}_\alpha = \mathrm{D}^{1/2}\,[\mathbf{X}\Sigma_\alpha(\xi^*)\mathbf{X}^{\mathrm{T}} \circ \mathbf{X}\{\Sigma_\alpha(\xi^*)/\alpha + 2\mu_\alpha(\xi^*)\,\mu_\alpha^{\mathrm{T}}(\xi^*)\}\mathbf{X}^{\mathrm{T}}]\,\mathrm{D}^{1/2},$$

have the same set of eigenvalues. Clearly, $\tilde{\mathbf{J}}_\alpha$ is real symmetric and positive semi-definite by the Schur product theorem. Therefore, $\tilde{\mathbf{J}}_\alpha$, and hence $\mathbf{J}_\alpha$, has non-negative eigenvalues. Hence, the spectral radius $\rho(\mathbf{J}_\alpha)$ is simply the largest eigenvalue of $\mathbf{J}_\alpha$, which we proceed to bound next.

Using the fact that, $\mathrm{D}^{1/2}[\mathbf{X}\Sigma_\alpha(\xi^*)\mathbf{X}^{\mathrm{T}} \circ \mathbf{X}\Sigma_\alpha(\xi^*)\mathbf{X}^{\mathrm{T}}]\mathrm{D}^{1/2}/\alpha$ is a positive semi-definite matrix, we have,

$$\tilde{\mathbf{J}}_\alpha \precsim 2\,\mathrm{D}^{1/2}\,[\mathbf{X}\Sigma_\alpha(\xi^*)\mathbf{X}^{\mathrm{T}} \circ \mathbf{X}\{\Sigma_\alpha(\xi^*)/\alpha + \mu_\alpha(\xi^*)\,\mu_\alpha^{\mathrm{T}}(\xi^*)\}\mathbf{X}^{\mathrm{T}}]\,\mathrm{D}^{1/2}. \tag{21}$$

Denote $\Lambda_\alpha = \Sigma_\alpha(\xi^*)/\alpha + \mu_\alpha(\xi^*)\,\mu_\alpha^{\mathrm{T}}(\xi^*)$ and $\mathbf{X}\Lambda_\alpha\mathbf{X}^{\mathrm{T}} = \Delta(\xi^*) \circ \Gamma_\alpha$ where $[\Delta(\xi^*)]_{ij} = \xi_i^*\xi_j^*$, $[\Gamma_\alpha]_{ij} = \mathbf{x}_i^{\mathrm{T}}\Lambda_\alpha\mathbf{x}_j/(\xi_i^*\xi_j^*)$. Then the matrix on the right hand side of the (21) can be written as,

$$2\,\mathrm{D}^{1/2}\,\{\mathbf{X}\Sigma_\alpha(\xi^*)\mathbf{X}^{\mathrm{T}} \circ \Delta(\xi^*)\}\,\mathrm{D}^{1/2} \circ \Gamma_\alpha.$$

A result from Horn and Johnson (1994) (see Lemma 23 in the appendix) provides bounds on the largest eigenvalues of $\mathrm{M} \circ \mathrm{N}$ as a product of the largest eigenvalue of M and largest diagonal of the N. The diagonals of Γ_α are 1 and the largest eigenvalue of $2\,\mathrm{D}^{1/2}\{\mathbf{X}\Sigma_\alpha(\xi^*)\mathbf{X}^{\mathrm{T}} \circ \Delta(\xi^*)\}\mathrm{D}^{1/2}$ which is equal to $2\,\mathrm{diag}[\{\xi^*\,A'(\xi^*)\}^{1/2}]\mathbf{X}\Sigma_\alpha(\xi^*)\mathbf{X}^{\mathrm{T}}\mathrm{diag}[\{\xi^*\,A'(\xi^*)\}^{1/2}]$ is the same as that of $2\,\Sigma_\alpha(\xi^*)\mathbf{X}^{\mathrm{T}}\mathrm{diag}\{\xi^*\,A'(\xi^*)\}\mathbf{X}$. Since $A(x) + xA'(x) < 0$ for all $x \in \mathbb{R}$, $2\,\mathbf{X}^{\mathrm{T}}\mathrm{diag}\{\xi^*\,A'(\xi^*)\}\mathbf{X} \prec \Sigma_\alpha^{-1}(\xi^*)$. Lemma 24 shows that the largest eigenvalue of $\mathrm{M}^{-1/2}\,\mathrm{N}\,\mathrm{M}^{-1/2}$ is strictly less than 1 where $\mathrm{N} \prec \mathrm{M}$ and M, N are positive definite and positive semi-definite matrices respectively. This delivers the proof that $\rho(\mathbf{J}_\alpha) < 1$.

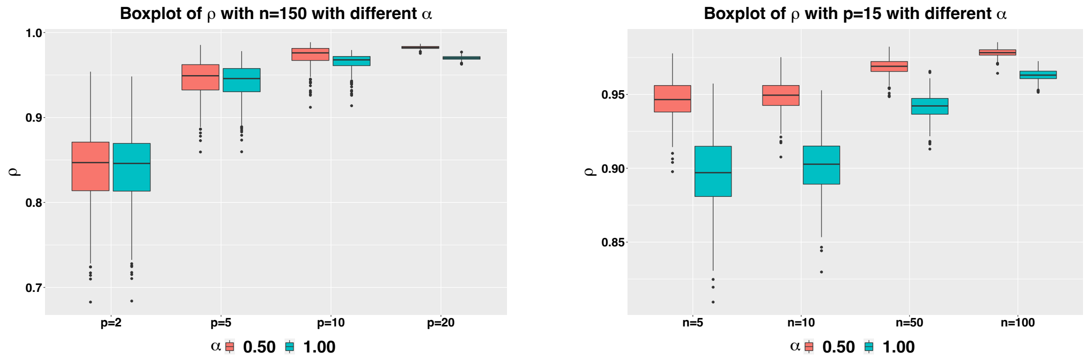
In the special case when $p = 1$, we can make substantial simplifications and show that (see §C.2 in the appendix for details),

$$\begin{aligned}\rho(\mathbf{J}_\alpha) &= \frac{2\sum_i^n x_i^2 A'(\xi_i^*)\xi_i^*}{\{\sigma_\beta^{-2}/\alpha - \sum_{i=1}^n 2x_i^2 A(\xi_i^*)\}} - \frac{\sum_i^n x_i^4 A'(\xi_i^*)/\xi_i^*}{\alpha\,\{\sigma_\beta^{-2}/\alpha - \sum_{i=1}^n 2x_i^2 A(\xi_i^*)\}^2}, \\ &< \frac{2\sum_{i=1}^n x_i^2 A'(\xi_i^*)\xi_i^*}{\sigma_\beta^{-2}/\alpha - \sum_{i=1}^n 2x_i^2 A(\xi_i^*)} < 1,\end{aligned} \tag{22}$$

where the first inequality follows from the fact that the second term in (22) is positive as $A'(x)/x > 0$ for all $x \in \mathbb{R}$. The second inequality follows from the fact that $A(x)+xA'(x) < 0$ for all $x \in \mathbb{R}$. ■

One of the critical elements in the proof is to show that the diagonals of Γ_α are equal to 1 which is achieved due to the fixed point equation evaluated at ξ^*. One can achieve global convergence if the proof could be generalized for every point in the domain. However, the invalidity of (9) at any point other than ξ^* does not allow us to show that. Also, It is important to note that Theorem 7 places minimal restriction on the design matrix $\mathbf{X}$.

We conduct a replicated numerical study to empirically demonstrate some of these features. We use the same simulation design corresponding to Figure 1 except now for a fixed (n, p), we provide a sufficiently flat prior $\Sigma_\beta = 10^2\mathbb{I}_p$ while fixing the first $p/2$ (resp. $[p/2]+1$) entries of $\beta^{\mathbf{o}}$ to be -4, and the remaining $p/2$ (resp. $[p/2]$) to be 4 when p is even (resp. odd). To remain faithful to the assumptions of Theorem 7, we do not normalize $\mathbf{X}$ with $\sqrt{p}$. We compute the spectral radius $\rho := \rho(\mathbf{J}_\alpha)$ of the Jacobian matrix $\mathbf{J}_\alpha$ for $\alpha \in \{0.50, 1.00\}$ at the fixed point ξ^* for different values of (n, p) over 500 independent replicates, with summary boxplots shown in Figure 3. In panel (a), we fix $n = 150$ and vary $p \in \{2, 5, 10, 20\}$. In panel (b), we fix p at 15 and vary $n \in \{5, 10, 50, 100\}$. It is evident that ρ remains less than 1 for all combinations of (n, p). Observe also that the first two cases in panel (b) correspond to $p > n$, and as predicted by the theory, the spectral radius continues to be smaller than 1. It can be seen from either panel that on an average ρ at $\alpha = 0.5$ is higher than the corresponding value at $\alpha = 1$.

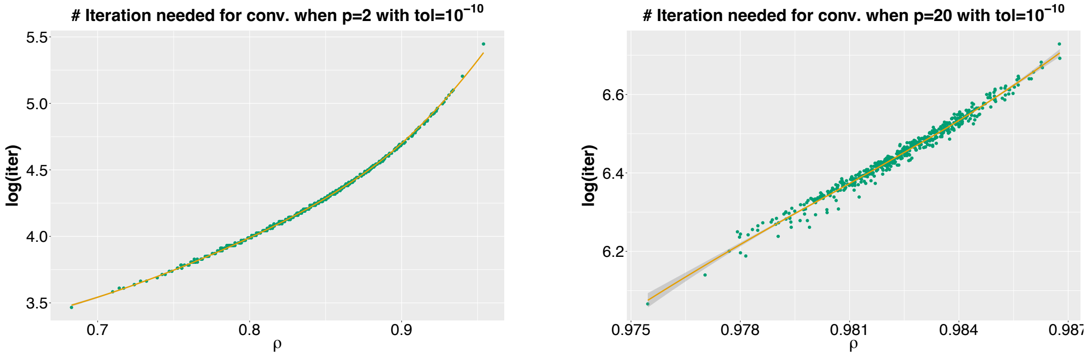


(a) *(b)*

Figure 3: (a) Boxplot of ρ with n fixed and varying p (b) Boxplot of ρ with p fixed and different values of n. Both the plots are produced with 500 replications using the same data $(y, \mathbf{X})$. It can be clearly seen that, for both the $\alpha \in \{0.50, 1.00\}$ the spectral radius is strictly less than 1 irrespective of n and p.

It is worth noting that local asymptotic stability does not provide any information other than the existence of a δ - neighborhood around ξ^* such that, if the system is initialized in that region the iterates converge to ξ^* as $t \to \infty$. Also, the definition does not say anything about the rate of convergence. In the following, we provide a heuristic argument to connect the notion of the rate of convergence with the spectral radius.

For simplicity, consider the one-dimensional system $x^{t+1} = g(x^t)$ for some function $g : \mathbb{R} \to \mathbb{R}$ which is twice continuously differentiable. If x^* is a fixed point of this system, using Taylor's theorem we have for some $T_0 > 0$, $(x^{t+1} - x^*) \approx g(x^*)(x^t - x^*)$ for all $t \geq T_0$. Recall that the linear rate of convergence (Romero et al., 2019) is given by, $\gamma = \lim_{t\to\infty} \|x^{t+1} - x^*\| / \|x^t - x^*\|$, provided the limit exists. In the above scenario, the iterates converge when $g(x^*) < 1$ and the rate of convergence is $g(x^*)$. For a general d-dimensional linear system $\alpha^{(t+1)} = A\alpha^{(t)}$ with fixed point $\alpha^* = 0$, it can be shown that,



(a) *(b)*

*Figure 4: For each replicated dataset $i \in \{1, 2, \ldots, 500\}$, we observe the number of iterations (*log-scale*) required by the algorithm for convergence and calculate the ρ at the fixed point solution for $\alpha = 1$. We plot $(\rho_i, \log(iter_i))$ for all $i \in \{1, 2, \ldots, 500\}$ and fitted with LOWESS line to explore the relationship between these two dependent variables. In (a), we generated the data $(y, \mathbf{X})$ with $(n = 150, p = 2)$. In (b), data $(y, \mathbf{X})$ is generated with $(n = 150, p = 20)$. It can be seen that number of iteration grows almost exponentially with the increasing ρ. Also, for fixed n, bigger p leads to higher ρ and as a consequence more iteration are required for convergence.*

$\|\alpha^{(t)}\|_2 = \|A^t \alpha^{(0)}\|_2 \leq \{\rho(A)\}^t \|\alpha^{(0)}\|_2$ where A is a square matrix and $\|\cdot\|_2$ is the Euclidean norm. Hence $\rho(A)$ acts as a rate of convergence for this case. Figure 4 is an illustration of the number of iterations needed for the system given by (8) to converge to the fixed point as a function of $\rho(\mathbf{J}_{\alpha=1})$. It is evident that the number of iterations increase exponentially as $\rho(\mathbf{J}_{\alpha=1})$ tends to 1.

4.2 A Special Case of Semi-Orthogonal Design

In this section, we shall consider a simple hierarchical logistic regression model given by,

$$p(y_{ij} = 1 \mid \beta) = 1/\{1 + \exp(-\beta_j)\} \quad (i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, p), \tag{23}$$

We assume a prior $\beta \sim \mathrm{N}_p(0, \sigma_\beta^2 \mathbb{I}_p)$. In this case, the results of Section 4.1 can be strengthened to obtain a global convergence rate of the EM sequence in (8). One key advantage here is the ability to decouple the EM sequence into independent coordinate-wise updates. This is illustrated in Lemma 8.

Lemma 8 *The EM updates for the model* (23) *can be simplified to,*

$$\left(\zeta_j^{t+1}\right)^2 = \frac{1}{\{\sigma_\beta^{-2} - 2n\, A(\zeta_j^t)\}} + \frac{n^2\,(\bar{y}_j - 1/2)^2}{\{\sigma_\beta^{-2} - 2n\, A(\zeta_j^t)\}^2} \quad (j = 1, 2, \ldots, p), \tag{24}$$

where, $\bar{y}_j = \sum_{i=1}^n y_{ij}/n$, for all $j = 1, 2, \ldots, p$ and ζ^t is the update at the t^{th} iteration.

Proof The log-likelihood from (23) is given by,

$$\log p(y \mid \beta) \propto \sum_{j=1}^p n\, \bar{y}_j\, \beta_j - \sum_{j=1}^p n\, \log\{1 + \exp(\beta_j)\}.$$

Following the calculations of (4) and (5) corresponding to $\alpha = 1$,

$$\log p_l(y, \beta \mid \zeta) = -\frac{1}{2}\beta^{\mathrm{T}} \left[\sigma_\beta^{-2}\mathbb{I}_p - 2n\,\mathrm{diag}\{A(\zeta)\}\right] \beta + n\,(\bar{Y} - 1/2\mathbb{1}_p)^{\mathrm{T}}\beta + n\,\mathbb{1}_p^{\mathrm{T}}C(\zeta) + \mathrm{Const.},$$

where, $\bar{Y}^{\mathrm{T}} = [\bar{y}_1, \bar{y}_2, \dots, \bar{y}_p]$. We claim from the above equation that, $\beta_j \mid Y, \zeta_j \sim \mathrm{N}\big(\mu(\zeta_j), \Sigma(\zeta_j)\big)$, independently for all $j = 1, 2, \dots, p$. Here $\Sigma^{-1}(\zeta_j) = \{\sigma_\beta^{-2} - 2n\,A(\zeta_j)\}$ and $\mu(\zeta_j) = n\,\big(\bar{y}_j - 1/2\big)/\{\sigma_\beta^{-2} - 2n\,A(\zeta_j)\}$. Following the calculation similar to (8) we obtain (24). ■

It is important to distinguish between the EM update ξ in (8) and ζ in (24). In the general setting (3), the variational parameter ξ are introduced for each individual $i \in \{1, 2, \dots, n\}$, whereas ζ is introduced here for different groups $j \in \{1, 2, \dots, p\}$. Though we used similar techniques to get the updates, they have different interpretation. Figure 3 and Figure 5 are not comparable in that sense.

The parallelization of the updates of ζ makes the posterior of β independent. Also, since the updates are independent and identical for all $j = 1, 2, \dots, p$ given the initial point, it suffices to study the stability of a single coordinate. The following theorem assures the global asymptotic stability of the EM sequence in (24).

Theorem 9 *The EM updates in* (24) *are globally asymptotically stable assuming* $\beta_j \sim N(0, \sigma_\beta^2)$ *with* $\sigma_\beta = 1$ *for all* $j = 1, 2, \dots, p$ *and* $n \geq 2$. *Moreover, with* $\lambda_j^t := (\zeta_j^t)^2$ $(j = 1, \dots, p)$, *there exists a global constant* $\rho \in (0, 1)$ *such that*

$$|\lambda_j^t - \lambda_j^*| \leq \rho^t |\lambda_j^0 - \lambda_j^*|.$$

Proof The proof of Theorem 9 is provided for $\sigma_\beta = 1$ for technical convenience. Letting $z = \zeta_j^2$ and $u = (\bar{y}_j - 0.5)$, consider,

$$h_{u,\sigma_\beta,n}(z) = \frac{1}{\{\sigma_\beta^{-2} - 2n\,A(\sqrt{z})\}} + \frac{n^2\,u^2}{\{\sigma_\beta^{-2} - 2n\,A(\sqrt{z})\}^2},$$

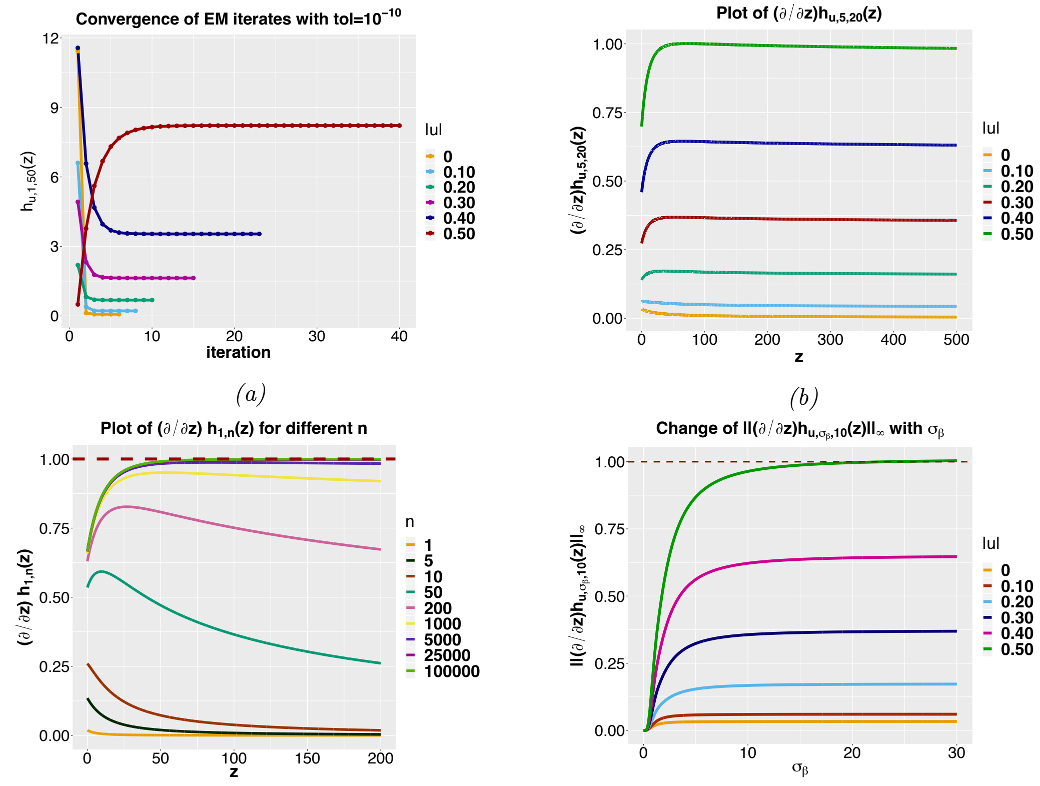
for some fixed $j = 1, 2, \dots, p$. Then one can write (24) by, $z^{t+1} = h_{u,\sigma_\beta,n}(z^t)$. It is easy to see that,

$$\frac{\partial h_{u,\sigma_\beta,n}(z)}{\partial z} = h'_{u,\sigma_\beta,n}(z) = \frac{n\,A'(\sqrt{z})}{\sqrt{z}}\{\sigma_\beta^{-2} - 2n\,A(\sqrt{z})\}^{-2}\left[1 + \frac{2n^2\,u^2}{\{\sigma_\beta^{-2} - 2n\,A(\sqrt{z})\}}\right]. \quad (25)$$

Let us call $\sigma_n = \{\sigma_\beta^{-2}/n - 2\,A(\sqrt{z})\}$. Since $u^2 \leq 1/4$ as $\bar{y}_j \in [0, 1]$, we have the following inequality,

$$h'_{u,\sigma_\beta,n}(z) \leq \frac{A'(\sqrt{z})}{\sqrt{z}}\sigma_n^{-2}\left[\frac{1}{n} + \frac{1}{2\,\sigma_n}\right] := h'_{\sigma_\beta,n}(z).$$

In appendix §C.4, we show that $\sup_{n\geq 1}\|h'_{\sigma_\beta,n}\|_\infty < 1$ when $\sigma_\beta = 1$, where $\|h'_{\sigma_\beta,n}\|_\infty := \sup_{z\in\mathbb{R}^+} h'_{\sigma_\beta,n}(z)$. The proof is then concluded by appealing to Lemma 16 in the appendix with $\rho = \sup_{n\geq 1}\|h'_{1,n}\|_\infty$. ■



(a)

(b)

(c)

(d)

Figure 5: (a) Evolution of λ_1^t with arbitrary initialization when $\sigma_\beta = 1$ and $n = 50$ for different $|u|$ (b) Plot of $h'_{u,\sigma_\beta,n}(z)$ for varying $|u|$ when $\sigma_\beta = 5$ and $n = 20$. (c) Plot of $h'_{1,n}(z)$ for different values of n, (d) We plot $\|h'_{u,\sigma_\beta,n}\|_\infty = \max_{z \in \mathbb{R}^+} h'_{u,\sigma_\beta,n}(z)$ as a function of σ_β, for different $|u|$ and a fixed $n = 10$. Numerically it is seen that $\|h'_{0.5,\sigma_\beta,10}(z)\|_\infty \geq 1$ when $\sigma_\beta \geq 12.894$.

It can be seen from (24) that the updates of ζ_j depend on the $(y, \mathbf{X})$ through $\bar{y}_j$, which is a sufficient statistic for β_j. Therefore if for some $j \neq j'$ we have $\bar{y}_j = \bar{y}_{j'}$, the sequences $\{\zeta_j^t\}$ and $\{\zeta_{j'}^t\}$ converge to the same limit. Figure 5a shows the global convergence of the EM sequence for different $\bar{y}_j$ when $\sigma_\beta = 1$ and $n = 50$ with arbitrary initializations. Numerically we assumed convergence when $|\zeta^{t+1} - \zeta^t| < 10^{-10}$. Interestingly, it is observed that convergence is slower when the data becomes more imbalanced, i.e. $|\bar{y}_j - 0.5| \to \pm 0.5$. A similar behavior for the mixing time of the Pólya-Gamma data augmentation Gibbs sampling in Bayesian logistic regression is observed in Johndrow et al. (2019), which is all the more interesting given the connection between Pólya-Gamma augmentation and tangent transforms established by Durante and Rigon (2019).

Figure 5b shows the behavior of $h'_{u,5,20}(z)$ for different values of u. Barring $u = 0$, in all other cases $h'_{u,5,20}(z)$ increase first before dropping off. Figure 5c shows that for fixed z, $h'_{1,n}(z)$ is an increasing function of n and less than 1. Lemma 27 proves this fact and in addition shows that for fixed z, $\lim_{n\to\infty} h'_{1,n}(z) < 1$. It is important to note that $h'_{\sigma_\beta,n}(z)$ is dependent on σ_β and for large σ_β and fixed z, $h'_{\sigma_\beta,n}(z)$ may not be an increasing function of n. Finally, Figure 5d shows $\|h'_{u,\sigma_\beta,10}\|_\infty$ increases as σ_β increases. It can be easily verified that for fixed u and n, $h'_{u,\sigma_\beta,n}(z)$ is an increasing function of σ_β and also for fixed n and σ_β, an increasing function of $|u|$. Numerically it can be seen that $\|h'_{0.5,\sigma_\beta,10}\|_\infty \geq 1$ when $\sigma_\beta \geq 12.894$. Overall, as the data get more imbalanced, a flatter prior on β increasingly hurts the convergence.

5. Extension to Multinomial Logit

In this section we provide an extension of the results in Section 4.1 to the case of multinomial logit regression where the response is an unordered categorical random variables with K levels. Assume y_i $(i = 1, 2, \ldots, n)$, takes the values in $\{1, 2, \ldots, K\}$ with following probabilities:

$$p[y_i = j \mid \beta_1, \beta_2, \ldots, \beta_{K-1}] = \begin{cases} \frac{\exp(\mathbf{x}_i^T \beta_j)}{1+\sum_{j=1}^{K-1} \exp(\mathbf{x}_i^T \beta_j)} & \text{for } j = 1, 2, \ldots, K-1 \\ \frac{1}{1+\sum_{j=1}^{K-1} \exp(\mathbf{x}_i^T \beta_j)} & \text{for } j = K. \end{cases}$$

Also assume $\beta_j \sim \mathrm{N}_p(\mu_j, \Sigma_j)$ $(j = 1, 2, \ldots, K-1)$. Let us define, $Y_{n\times(K-1)} = [Y_1, Y_2, \ldots, Y_{K-1}]$ with $Y_j^{\mathrm{T}} = [\mathbb{1}(y_1 = j), \mathbb{1}(y_2 = j), \ldots, \mathbb{1}(y_n = j)]$ $(j = 1, 2, \ldots, K-1)$, $\mathbf{X}$ is the design matrix. Specific to each individual i and class j, we introduce a variational parameter denoted by χ_{ij} $(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, K-1)$. Let us call $\chi_j^{\mathrm{T}} = (\chi_{1j}, \chi_{2j}, \ldots, \chi_{nj})$, and $\beta = (\beta_1, \beta_2, \ldots, \beta_{K-1})$.

The multinomial logistic log-likelihood contains a log-sum-exp term which poses the same difficulty of intractability as (4). Moreover, the logistic term can not be optimized straightaway due to sum of exponents inside the logistic function. Various methods have been proposed to circumvent this issue; Taylor approximation to the log-sum-exp term (Braun and McAuliffe, 2010), Quasi-Monte-Carlo (Lawrence et al., 2004), Jensen's inequality (Blei and Lafferty, 2007), quadratic approximation (Bouchard, 2008; Jebara and Choromanska, 2012). We use following inequality based on Bouchard (2008)

$$\sum_{j=1}^{K-1} \log\left(1 + e^{\mathbf{x}_i^{\mathrm{T}}\beta_j}\right) \geq \log\left(1 + \sum_{j=1}^{K-1} e^{\mathbf{x}_i^{\mathrm{T}}\beta_j}\right). \tag{26}$$

In an ideal scenario, if $p[y_i = l] > p[y_i = j]$ for all $j \neq l$ (ensuring a high probabillity that y_i belongs to the l^{th} class), then $\mathbf{x}_i^{\mathrm{T}}\beta_l > \mathbf{x}_i^{\mathrm{T}}\beta_j$ for all $j \neq l$. Hence, the major contribution on the both sides of the above display comes from a single term $\exp(\mathbf{x}_i^{\mathrm{T}}\beta_l)$ making the inequality in (26) tight. In general, however, (26) is not optimal and Bouchard (2008) introduces additional parameters to attain a tighter bound. This leads us to the following Lemma which provides the update equation for the EM sequence in multinomial logistic regression utilizing the Bouchard's bound in (26).

Lemma 10 *The EM updates to the above multinomial logit regression under* (26) *are given by,*

$$(\chi_j^{t+1})^2 = diag[\mathbf{X}\{\Sigma_\alpha(\chi_j^t)/\alpha + \mu_\alpha(\chi_j^t)\mu_\alpha^{\mathrm{T}}(\chi_j^t)\}\mathbf{X}^{\mathrm{T}}], \quad j \in \{1, 2, \ldots, K-1\} \tag{27}$$

where, $\Sigma_\alpha^{-1}(\chi_j) = \Sigma_j^{-1}/\alpha - 2\mathbf{X}^{\mathrm{T}} diag\{A(\chi_j)\}\mathbf{X}$ *and* $\mu_\alpha^{\mathrm{T}}(\chi_j)\Sigma_\alpha^{-1}(\chi_j) = \left(y_j - \frac{1}{2}\mathbb{1}_n\right)^{\mathrm{T}}\mathbf{X} + \mu_j^{\mathrm{T}}\Sigma_j^{-1}/\alpha$.

Proof We begin with the log-fractional likelihood,

$$\log p^{\alpha}(y \mid \mathbf{X}, \beta) \;=\; \alpha \sum_{i=1}^{n} \sum_{j=1}^{K-1} x_i^{\mathrm{T}} \beta_j \mathbb{1}[y_i = j] - \alpha \sum_{i=1}^{n} \log\big(1 + \sum_{j=1}^{K-1} e^{x_i^{\mathrm{T}} \beta_j}\big). \tag{28}$$

Now using (26) in (28), we get a lower bound to $p^{\alpha}(y \mid \mathbf{X}, \beta)$ given by,

$$\log p^{\alpha}(y \mid \mathbf{X}, \beta) \;\geq\; \alpha \sum_{j=1}^{K-1} \left\{ \sum_{i=1}^{n} \mathbf{x}_i^{\mathrm{T}} \beta_j \mathbb{1}[y_i = j] - \sum_{i=1}^{n} \log\left(1 + e^{\mathbf{x}_i^{\mathrm{T}} \beta_j}\right) \right\}. \tag{29}$$

Next, we use the quadratic bound proposed by Jaakkola and Jordan (2000) on the right hand side of the above inequality. This leads to a lower bound to $\log p^{\alpha}(y \mid \mathbf{X}, \beta)$ similar to (5) where

$$\begin{aligned} \log p_l^{\alpha}(y, \beta \mid \mathbf{X}, \chi) = & \sum_{j=1}^{K-1} \left[\alpha \left\{ Y_j^{\mathrm{T}} \mathbf{X} \beta_j + (\mathbf{X}\beta_j)^{\mathrm{T}} \mathrm{diag}\{A(\chi_{ij})\}(X\beta_j) - \frac{1}{2} \mathbb{1}_n^{\mathrm{T}} X \beta_j + \mathbb{1}_n^{\mathrm{T}} C(\chi_{ij}) \right\} \right] \\ & - \sum_{j=1}^{K-1} \left[\frac{1}{2} (\beta_j - \mu_j)^{\mathrm{T}} \Sigma_j^{-1} (\beta_j - \mu_j) \right] + \text{Constant}, \end{aligned}$$

for fixed $j \in \{1, 2, \ldots, K-1\}$ the updates are exactly similar to the updates in logistic version. Moreover, updates to $\chi_j^{\mathrm{T}} = [\chi_{1j}, \chi_{2j}, \ldots, \chi_{nj}]^{\mathrm{T}}$ are independent over $j \in \{1, 2, \ldots, K-1\}$. Following the similar E-step and M-step for the logistic version as in (7)-(8), it can be easily seen that for fixed j the update equation is given by (27). ■

As the updates across each level $j \in \{1, 2, \ldots, K-1\}$ are independent and the behavior of the updates is exactly similar to the binary setup in (8), this leads us to the following theorem that guarantees the local asymptotic stability of EM updates in Lemma 10.

Theorem 11 *Suppose the design matrix* $\mathbf{X}$ *does not have any row equal to the zero vector. For any* $\alpha \in (0, 1]$ *and positive definite* Σ_β*, any fixed point solution* χ_j^* *of the EM sequence in* (27) *is locally asymptotically stable.*

Proof For each fixed $j \in \{1, 2, \ldots, K-1\}$, the fixed point equation is,

$$(\chi_j^*)^2 = \mathrm{diag}[\mathbf{X}\{\Sigma_\alpha(\chi_j^*) + \mu_\alpha(\chi_j^*)\mu_\alpha^{\mathrm{T}}(\chi_j^*)\}\mathbf{X}^{\mathrm{T}}]. \tag{30}$$

Call $\xi^* = \chi_j^*$ and $\xi^t = \chi_j^t$. Then, (27) and (30) reduces to (8) and (9) respectively. Now, we directly apply Theorem 7 to conclude that the updates in (27) are locally asymptotically stable. ■

The variational estimates in case of Multinomial logit regression incorporate two levels of approximation, first the log-*sum* term in probability density is bounded from above by *sum*-log term using Bouchard's technique (Bouchard, 2008) in (26) and then the tangent transformation is applied on the lower bound to the density in (29). Theoretically the first approximation is sub-optimal and the risk can not be made arbitrarily small here. However, from a practical perspective the variational approximator in the current framework performs well which we illustrate subsequently.

Here, we conduct simulations to verify the algorithmic convergence and provide empirical evidence of the statistical accuracy of the variational estimates. We do not include any plot for convergence since the fixed point iterations in (27) are essentially equivalent to the logistic case. In case of multinomial logistic regression, the accuracy of the variational approximation poses a more interesting question. We start the investigation by defining

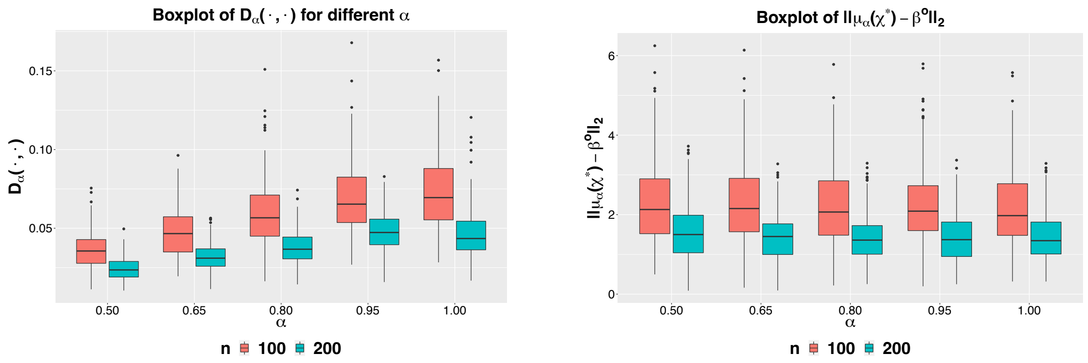
$$\mathrm{D}_\alpha(\beta,\beta^{\mathbf{o}}) = \frac{1}{n(\alpha-1)}\log\int\left\{\frac{p(y\mid\beta,\mathbf{X})}{p(y\mid\beta^{\mathbf{o}},\mathbf{X})}\right\}^{\alpha}p(y\mid\beta^{\mathbf{o}},\mathbf{X})dy,$$

for any $\alpha\in(0,1)$. Under the multinomial distribution, the above display simplifies to

$$\mathrm{D}_\alpha(\beta,\beta^{\mathbf{o}}) = \frac{1}{n(\alpha-1)}\sum_{i=1}^{n}\log\left[\frac{1+\sum_{j=1}^{K-1}\exp\{\alpha\mathbf{x}_i^{\mathrm{T}}\beta_j+(1-\alpha)\mathbf{x}_i^{\mathrm{T}}\beta_j^{\mathbf{o}}\}}{\{1+\sum_{j=1}^{K-1}\exp(\mathbf{x}_i^{\mathrm{T}}\beta_j)\}^{\alpha}\{1+\sum_{j=1}^{K-1}\exp(\mathbf{x}_i^{\mathrm{T}}\beta_j^{\mathbf{o}})\}^{1-\alpha}}\right].$$

Also define,

$$\begin{aligned}\mathrm{D}(\beta^{\mathbf{o}},\beta) &= \frac{1}{n}\int\log\left\{\frac{p(y\mid\beta^{\mathbf{o}},\mathbf{X})}{p(y\mid\beta,\mathbf{X})}\right\}p(y\mid\beta^{\mathbf{o}},\mathbf{X})dy,\\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{K-1}\mathbf{x}_i^{\mathrm{T}}(\beta_j^{\mathbf{o}}-\beta_j)\frac{\exp(\mathbf{x}_i^{\mathrm{T}}\beta_j^{\mathbf{o}})}{1+\sum_{j=1}^{K-1}\exp(\mathbf{x}_i^{\mathrm{T}}\beta_j^{\mathbf{o}})}+\log\frac{1+\sum_{j=1}^{K-1}\exp(\mathbf{x}_i^{\mathrm{T}}\beta_j^{\mathbf{o}})}{1+\sum_{j=1}^{K-1}\exp(\mathbf{x}_i^{\mathrm{T}}\beta_j)}\right].\end{aligned}$$



(a)

(b)

Figure 6: (a) Boxplot of $D_\alpha(\mu_\alpha(\chi^*),\beta^{\mathbf{o}})$ for $\alpha\in(0,1)$ and $D(\beta^{\mathbf{o}},\mu(\chi^*))$ for $\alpha=1$ (b) Boxplot of $\|\mu_\alpha(\chi^*)-\beta^{\mathbf{o}}\|_2$ for different values of $\alpha\in(0,1]$

In order to begin our investigation, we fix $K=4$, $p=5$ and $\beta_1^{\mathbf{o}}=(3,-1,0,-2,0)^{\mathrm{T}}$, $\beta_2^{\mathbf{o}}=(-2,4,1,-1,-2)^{\mathrm{T}}$, $\beta_3^{\mathbf{o}}=(0,1,-2,2,-1)^{\mathrm{T}}$ and varied $n\in\{100,200\}$. Then, each row $\mathbf{x}_i^{\mathrm{T}}$ of the $n\times p$ design matrix $\mathbf{X}$ is generated from $\mathrm{N}(0,\mathbb{I}_p)$. We normalize each row of $\mathbf{X}$ by $\sqrt{p}$. The response $y_i\,(i=1,2,\ldots,n)$ is generated from a multinomial distribution with class probabilities given by $p_i=\exp(\mathbf{x}_i^{\mathrm{T}}\beta_j^{\mathbf{o}})/\{1+\exp(\mathbf{x}_i^{\mathrm{T}}\beta_j^{\mathbf{o}})\}$ $(j=1,2,\ldots,K-1)$ and $p_K=$

$1/\{1+\exp(\mathbf{x}_i^{\mathrm{T}}\beta_j^{\mathbf{o}})\}$. Next, we let the iterations converge and took $\{\mu_\alpha(\chi_1^*), \mu_\alpha(\chi_2^*), \mu_\alpha(\chi_3^*)\}$ to be the variational estimates of $\{\beta_1^{\mathbf{o}}, \beta_2^{\mathbf{o}}, \beta_3^{\mathbf{o}}\}$. We repeat the above process for 250 times. For notational convenience, we call $\mu_\alpha(\chi^*)$ to be collection of $\{\mu_\alpha(\chi_1^*), \mu_\alpha(\chi_2^*), \mu_\alpha(\chi_3^*)\}$. In the left panel of Figure 6, we plot $\mathrm{D}_\alpha(\mu_\alpha(\chi^*), \beta^{\mathbf{o}})$ for $\alpha \in \{0.50, 0.65, 0.80, 0.95\}$ and $\mathrm{D}(\beta^{\mathbf{o}}, \mu(\chi^*))$ for $\alpha = 1$ and on the right panel, we plot $\|\mu_\alpha(\chi^*) - \beta^{\mathbf{o}}\|_2$ where $\|\mu_\alpha(\chi^*) - \beta^{\mathbf{o}}\|_2^2 = \{\|\mu_\alpha(\chi_1^*) - \beta_1^{\mathbf{o}}\|_2^2 + \|\mu_\alpha(\chi_2^*) - \beta_2^{\mathbf{o}}\|_2^2 + \|\mu_\alpha(\chi_3^*) - \beta_3^{\mathbf{o}}\|_2^2\}$. From both these plots, it is evident that the estimates improve with the increase in the sample size. For reasons similar to the case of logistic regression, $\mathrm{D}_\alpha(\cdot,\cdot)$ increases with $\alpha \in \{0.5, 0.65, 0.80, 0.95\}$. In Figure 7, we plot $\|\mu_\alpha(\chi_i^*) - \beta_i^{\mathbf{o}}\|_2$ $(i = 1, 2, 3)$ for $\alpha \in \{0.5, 0.65, 0.80, 0.95, 1\}$. One can see that the ℓ_2 norm between the variational estimate and the truth decreases with the increase in the sample size for each class as well.

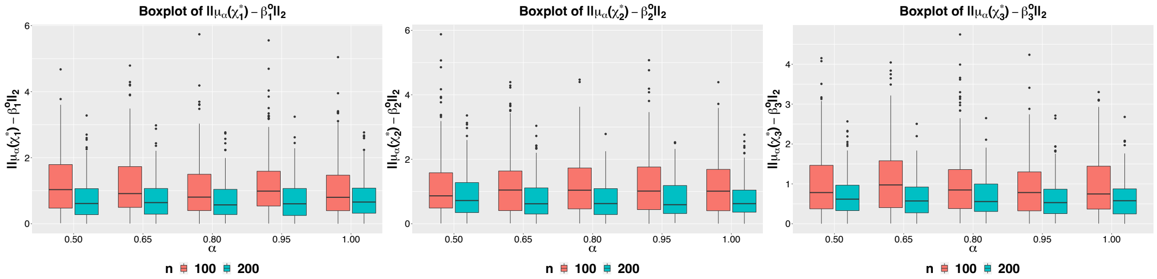


Figure 7: Boxplot of $\|\mu_\alpha(\chi_i^*) - \beta_i^{\mathbf{o}}\|_2$ *for* $i = 1, 2, 3$ *from left to right respectively with varying* n.

5.1 An example of global convergence

Similar to §4.2 one can achieve global convergence in case of multinomial logit regression as well. Following a similar setup to the model in (23) one can define the following for any $j \in \{1, 2, \ldots, p\}$ and $i \in \{1, 2, \ldots, n\}$,

$$p[y_{ij} = l \mid \beta_1, \beta_2, \ldots, \beta_{K-1}] = \begin{cases} \frac{\exp\left(\beta_l^{(j)}\right)}{1+\sum_{l=1}^{K-1}\exp\left(\beta_l^{(j)}\right)} & \text{for } l = 1, 2, \ldots, K-1, \\ \frac{1}{1+\sum_{l=1}^{K-1}\exp\left(\beta_l^{(j)}\right)} & \text{for } l = K. \end{cases}$$

Where, $\beta_l^{(j)}$ is the j^{th} component of the vector β_l corresponding to the l^{th} class. Let us assume that, $\beta_l \sim \mathrm{N}(0, \sigma_{\beta_l}^2 \mathbb{I}_p)$ for $l = 1, 2, \ldots, K-1$. Then, we have the following lemma.

Lemma 12 *The EM updates for the model above can be simplified to,*

$$\left(\Upsilon_{j,l}^{t+1}\right)^2 = \frac{1}{\{\sigma_{\beta_l}^{-2} - 2n\, A(\Upsilon_{j,l}^t)\}} + \frac{n^2\, (\bar{y}_{j,l} - 1/2)^2}{\{\sigma_{\beta_l}^{-2} - 2n\, A(\Upsilon_{j,l}^t)\}^2},$$

where, $\bar{y}_{j,l} = \sum_{i=1}^n \mathbb{1}(y_{ij} = l)/n$, *for all* $j = 1, 2, \ldots, p$; $l = 1, 2, \ldots, K-1$ *and* Υ^t *is the update at the* t^{th} *iteration.*

Proof The log-likelihood can be written as,

$$\log p(y \mid \beta) = \sum_{i=1}^{n} \sum_{j=1}^{p} \left[\sum_{l=1}^{K-1} \beta_l^{(j)} \mathbb{1}(y_{ij} = l) - \log \left\{ 1 + \sum_{l=1}^{K-1} \exp\left(\beta_l^{(j)}\right) \right\} \right].$$

Now, applying the inequality in (26) on the last term of the right hand side of the above equation, we get

$$\begin{aligned} \log p(y \mid \beta) &\geq \sum_{l=1}^{K-1} \sum_{j=1}^{p} \sum_{i=1}^{n} \left[\beta_l^{(j)} \mathbb{1}(y_{ij} = l) - \log \left\{ 1 + \exp\left(\beta_l^{(j)}\right) \right\} \right] \\ &= \sum_{l=1}^{K-1} \sum_{j=1}^{p} \left[n\bar{y}_{j,l}\, \beta_l^{(j)} - n \log \left\{ 1 + \exp\left(\beta_l^{(j)}\right) \right\} \right]. \end{aligned}$$

Now, for any fixed $l \in \{1, 2, \ldots, K-1\}$, the rest follows from proof of Lemma 8. ■

Using a combination of the Lemma 12 and Theorem 9, the global convergence follows.

6. Discussion

In this article, we are able to provide statistical and computational guarantees for the tangent transformation approach of Jaakkola and Jordan (2000) in the context of Bayesian logistic regression problem. We showed that for $\alpha \in (0, 1]$ the variational estimates arising out of the α-VB TT in (8) converge to the true parameter at minimax optimal rate. Next, we prove that the algorithm converges to a fixed point under minimal assumptions on the data generation. However, our result on algorithmic convergence pertains to *local asymptotic stability*, which ensures convergence only if the iteration is initialized within a certain neighborhood around a fixed point. Accurately characterizing the neighborhood of initialization is an interesting future work. In simulations, we have empirically observed good convergence behavior for a wide range of initializations, which suggests such an exercise might be possible, perhaps with additional conditions on the design matrix.

One example where we have managed to show global convergence is in the context of a hierarchical logit model. We extend our result on algorithmic convergence using the tangent transformation approach in case of multinomial logistic regression and numerically study the statistical accuracy of the variational estimates. Due to sub-optimality of the inequality (26), the variational risk can not be made arbitrarily small in this case. As an ongoing research, we are exploring the implications of introducing an additional variational parameter to close this gap.

Acknowledgments

Dr. Pati and Dr. Bhattacharya acknowledge support from NSF DMS (1613156, 1854731, 1916371) and NSF CCF 1934904 (HDR-TRIPODS). In addition, Dr. Bhattacharya acknowledges NSF CAREER 1653404 award for supporting this project.

Appendix A. Proof of Statistical Optimality Results in Section 3

In the following, we first provide the proofs of Theorems 2 and 4 in §A.1 and A.2 respectively and then provide the proofs of some of the auxiliary results used in subsequent §A.3.

A.1 Proof of Theorem 2

The proof consists of two major steps.

Risk majorization. In this first step, we obtain an upper bound to the integrated risk in terms of easily controllable quantities. We denote $\mathbb{E}_{\beta^{\mathbf{o}}}$ as taking expectation under (10). From the definition of the α-Renyi divergence and the fact that p_l lower bounds $p(y \mid \beta, \mathbf{X})$

$$\mathbb{E}_{\beta^{\mathbf{o}}} \exp\left\{\alpha \log \frac{p_l(y \mid \beta, \xi, \mathbf{X})}{p(y \mid \beta^{\mathbf{o}}, \mathbf{X})}\right\} \leq \mathbb{E}_{\beta^{\mathbf{o}}} \exp\left\{\alpha \log \frac{p(y \mid \beta, \mathbf{X})}{p(y \mid \beta^{\mathbf{o}}, \mathbf{X})}\right\} = e^{-n(1-\alpha)\mathrm{D}_\alpha(\beta, \beta^{\mathbf{o}})}.$$

Thus, for any $\varepsilon \in (0,1)$, we have

$$\mathbb{E}_{\beta^{\mathbf{o}}} \exp\left[\alpha \log \frac{p_l(y \mid \beta, \xi, \mathbf{X})}{p(y \,|\, \beta^{\mathbf{o}}, \mathbf{X})} + n(1-\alpha)\mathrm{D}_\alpha(\beta, \beta^{\mathbf{o}}) - \log(1/\varepsilon)\right] \leq \varepsilon.$$

Integrating both side of this inequality with respect to the prior π_β and interchanging the integrals using Fubini's theorem, we obtain

$$\mathbb{E}_{\beta^{\mathbf{o}}} \int \exp\left[\alpha \log \frac{p_l(y \,|\, \beta, \xi, \mathbf{X})}{p(y \,|\, \beta^{\mathbf{o}}, \mathbf{X})} + n(1-\alpha)\mathrm{D}_\alpha(\beta, \beta^{\mathbf{o}}) - \log(1/\varepsilon)\right] \pi_\beta(\beta)\, d\beta \leq \varepsilon.$$

Now, recall the *variational inequality* for a probability measure μ and for h such that e^h is integrable,

$$\log \int e^h d\mu = \sup_{\rho \ll \mu} \left[\int h d\rho - D(\rho || \mu)\right]. \tag{31}$$

Using (31),

$$\mathbb{E}_{\beta^{\mathbf{o}}} \exp \sup_{q \ll \pi_\beta} \left[\int \left\{\alpha \log \frac{p_l(y \mid \beta, \xi, \mathbf{X})}{p(y \mid \beta^{\mathbf{o}}, \mathbf{X})} + n(1-\alpha)\mathrm{D}_\alpha(\beta, \beta^{\mathbf{o}}) - \log(1/\varepsilon)\right\} q(\beta)\, d\beta \right.$$
$$\left. - \mathrm{D}(q \,||\, \pi_\beta)\right] \leq \varepsilon.$$

If we choose $\rho = q_\beta^* \equiv \phi_p\big\{\beta; \mu_\alpha(\xi^*), \Sigma_\alpha(\xi^*)\big\}$ as the variational approximation and set $\xi = \xi^*$

$$\mathbb{E}_{\beta^{\mathbf{o}}} \exp\left[\int \left\{\alpha \log \frac{p_l(y \mid \beta, \xi^*, \mathbf{X})}{p(y \mid \beta^{\mathbf{o}}, \mathbf{X})} + n(1-\alpha)\mathrm{D}_\alpha(\beta, \beta^{\mathbf{o}}) - \log(1/\varepsilon)\right\} q_\beta^*(\beta)\, d\beta - \mathrm{D}(q_\beta^* \,||\, \pi_\beta)\right]$$
$$\leq \varepsilon. \tag{32}$$

By applying Markov's inequality, we further obtain that with $\mathbb{P}_{\beta^{\mathbf{o}}}$ probability at least $(1-\varepsilon)$,

$$n(1-\alpha) \int \mathrm{D}_\alpha(\beta, \beta^{\mathbf{o}})\, q_\beta^*(\beta)\, d\beta \leq -\alpha \int_\beta \log \frac{p_l(y \mid \beta, \xi^*, \mathbf{X})}{p(y \,|\, \beta^{\mathbf{o}}, \mathbf{X})}\, q_\beta^*(\beta)\, d\beta + \mathrm{D}(q_\beta^* \,||\, \pi_\beta) + \log(1/\varepsilon).$$

Now using the Lemma 1,

$$\begin{aligned}
&- \alpha \int_\beta \log \frac{p_l(y \mid \beta, \xi^*, \mathbf{X})}{p(y \mid \beta^\mathbf{o}, \mathbf{X})}\, q^*_\beta(\beta)\, d\beta + \mathrm{D}(q^*_\beta \,||\, \pi_\beta) \\
&= \inf_{q,\xi} \bigg\{ - \alpha \int_\beta \log \frac{p_l(y \mid \beta, \xi, \mathbf{X})}{p(y \,|\, \beta^\mathbf{o}, \mathbf{X})}\, q(\beta)\, d\beta + \mathrm{D}(q \,||\, \pi_\beta) \bigg\}. \qquad (33)
\end{aligned}$$

Optimizing the majorized risk. Our second step consists of optimizing the term obtained in (33) by choosing suitable candidates for q and ξ. We refer to them as $\tilde{q}$ and $\tilde{\xi}$. The idea is to choose $\tilde{q}$ and $\tilde{\xi}$ so that $\tilde{q}$ places almost all its mass into a small neighborhood around truth $\mathbf{X}\beta^\mathbf{o}$, so that the first term in the right hand side of (33) becomes small; on the other hand, the neighborhood is large enough so that the second regularization term $n^{-1}\, D(q \,||\, \pi_\beta)$ is not too large. We choose $\tilde{q}$ first and $\tilde{\xi}$ later. Let $\tilde{q}$

$$\tilde{q}(\beta) = \frac{\pi_\beta(\beta)}{\pi_\beta\big[\mathcal{B}_n(\beta^\mathbf{o},\, \varepsilon)\big]}\, I_{\mathcal{B}_n(\beta^\mathbf{o},\, \varepsilon)}(\beta), \quad \forall \beta \in \mathbb{R}^p, \qquad (34)$$

be the restriction of the prior density π_β into the KL neighborhood $\mathcal{B}_n(\beta^\mathbf{o},\, \varepsilon)$ around $\beta^\mathbf{o}$ with radius ε defined as

$$\mathcal{B}_n(\beta^\mathbf{o}, \varepsilon) = \Big\{ n^{-1}\, \tilde{\mathrm{D}}\big[p(\cdot \mid \beta^\mathbf{o}, \mathbf{X}) \,\big|\big|\, p_l(\cdot \mid \beta, \xi, \mathbf{X})\big] \le \varepsilon^2,\ n^{-1}\, \mathrm{V}\big[p(\cdot \mid \beta^\mathbf{o}, \mathbf{X}) \,\big|\big|\, p_l(\cdot \mid \beta, \xi, \mathbf{X})\big] \le \varepsilon^2 \Big\}, \qquad (35)$$

where for two non-negative functions f, g, $\widetilde{\mathrm{D}}(f \,\|\, g) = \int f |\log(f/g)|$ and $\mathrm{V}(f, g) := \int f \log^2(f/g) - \widetilde{\mathrm{D}}^2(f \,\|\, g)$. Note that $\tilde{\mathrm{D}}(f \,|\, g)$ is an extension of the usual KL distance for probability measures to positive functions which may not integrate to one. With this substitution, the second term in (33) becomes the negative log prior mass $[n\,(1-\alpha)]^{-1}\, \log\big\{\pi_\beta[\mathcal{B}_n(\beta^\mathbf{o},\, \varepsilon)]\big\}^{-1}$ and it remains to provide a high-probability bound for the first term and an upper bound for the log-prior concentration term $\log\big\{\pi_\beta[\mathcal{B}_n(\beta^\mathbf{o},\, \varepsilon)]\big\}$.

i) High probability upper bound for the first term in (33). By applying Fubini's theorem and invoking the definition of $\mathcal{B}_n(\beta^\mathbf{o},\, \varepsilon)$, we have

$$\begin{aligned}
&\mathbb{E}_{\beta^\mathbf{o}}\bigg[\int_\beta \tilde{q}(\beta)\, \log \frac{p_l(y \mid \beta, \xi, \mathbf{X})}{p(y \mid \beta^\mathbf{o}, \mathbf{X})}\, d\beta\bigg] = \int_\beta \mathbb{E}_{\beta^\mathbf{o}}\bigg[\log \frac{p_l(y \mid \xi, \beta, \mathbf{X})}{p(y \mid \beta^\mathbf{o}, \mathbf{X})}\bigg]\, \tilde{q}(\beta)\, d\beta \\
&\le \int_{\mathcal{B}_n(\beta^\mathbf{o},\, \varepsilon)} \widetilde{\mathrm{D}}\big[p(\cdot \mid \beta^\mathbf{o}, \mathbf{X}) \,\big|\big|\, p_l(\cdot \mid \beta, \xi, \mathbf{X})\big]\, \tilde{q}(\beta)\, d\beta \le n\, \varepsilon^2.
\end{aligned}$$

Similarly, we have the following bound for the second moment by applying the Cauchy-Schwarz inequality,

$$\mathrm{Var}_{\beta^\mathbf{o}}\bigg[\int_\beta \tilde{q}(\beta)\, \log \frac{p_l(y \mid \beta, \xi, \mathbf{X})}{p(y \mid \beta^\mathbf{o}, \mathbf{X})}\, d\beta\bigg] \le \int_{\mathcal{B}_n(\beta^\mathbf{o},\, \varepsilon)} V\big[p(\cdot \mid \beta^\mathbf{o}, \mathbf{X}) \,\big|\big|\, p(\cdot \mid \beta, \xi, \mathbf{X})\big]\, \tilde{q}(\beta)\, d\beta \le n\, \varepsilon^2.$$

Putting pieces together, applying Chebyshev's inequality, we obtain

$$\begin{aligned}
&\mathbb{P}_{\beta^\mathbf{o}}\bigg\{\int_\beta \tilde{q}(\beta)\, \log \frac{p_l(y \mid \beta, \xi, \mathbf{X})}{p(y \mid \beta^\mathbf{o}, \mathbf{X})}\, d\beta \le -D\, n\, \varepsilon^2\bigg\} \\
&\le \mathbb{P}_{\beta^\mathbf{o}}\bigg\{\int_\beta \tilde{q}(\beta)\, \log \frac{p_l(y \mid \beta, \xi, \mathbf{X})}{p(y \mid \beta^\mathbf{o}, \mathbf{X})}\, d\beta - \mathbb{E}_{\beta^\mathbf{o}}\Big[\int_\beta \tilde{q}(\beta)\, \log \frac{p_l(y \mid \beta, \xi, \mathbf{X})}{p(y \mid \beta^\mathbf{o}, \mathbf{X})}\, d\beta\Big] \le -(D-1)\, n\, \varepsilon^2\bigg\} \\
&\le \frac{\mathrm{Var}_{\beta^\mathbf{o}}\big[\int_\beta \tilde{q}(\beta)\, \log \frac{p_l(y \,|\, \beta, \xi, \mathbf{X})}{p(y \,|\, \beta^\mathbf{o}, \mathbf{X})}\, d\beta\big]}{(D-1)^2\, n^2\, \varepsilon^4} \le \frac{1}{(D-1)^2\, n\, \varepsilon^2}.
\end{aligned}$$

It follows with probability $1-1/\{(D-1)^2\, n\, \varepsilon^2)\}$, the first term of (33) evaluated at $q=\tilde{q}$ satisfies

$$-\alpha \int_{\beta} \tilde{q}(\beta)\, \log \frac{p_l(y \mid \beta, \xi, \mathbf{X})}{p(y \mid \beta^{\mathbf{o}}, \mathbf{X})}\, d\beta \leq Dn\alpha\, \varepsilon^2.$$

ii) Upper bound for the negative $\log$*-prior concentration term* $-\log\big\{\pi_\beta[\mathcal{B}_n(\beta^{\mathbf{o}},\, \varepsilon)]\big\}$. We first obtain an upper bound for the log-pseudo-likelihood ratio

$$\begin{aligned}\log \frac{p_l(y \mid \beta, \xi, \mathbf{X})}{p(y \mid \beta^{\mathbf{o}}, \mathbf{X})} &= y^{\mathrm{T}}\mathbf{X}(\beta-\beta^{\mathbf{o}}) + \beta^{\mathrm{T}}\left\{\mathbf{X}^{\mathrm{T}}\mathrm{diag}\{A(\xi)\}\mathbf{X}\right\}\beta + 0.5\mathbb{1}_n^{\mathrm{T}}\mathbf{X} \\ &\quad + \mathbb{1}_n^{\mathrm{T}}C(\xi) + \mathbb{1}_n^{\mathrm{T}}\log\{1+\exp(\mathbf{X}\beta^{\mathbf{o}})\}.\end{aligned}$$

To obtain the lower bound $p_l(y \mid \xi, \beta, \mathbf{X})$ of $p(y \mid \beta, \mathbf{X})$, Jaakkola and Jordan (2000) used $-\log\{1+\exp(-x)\} = x/2 - \log(e^{x/2}+e^{-x/2})$ and noted $f(x) = -\log(e^{x/2}+e^{-x/2})$ is a convex function in the variable x^2. Since a tangent surface to a convex function is a global lower bound for the function, we can bound $f(x)$ globally with a first order Taylor expansion in the variable of x^2 around ξ^2 as

$$\begin{aligned}f(x) &\geq f(\xi) + \frac{df(\xi)}{d\xi^2}(x^2-\xi^2) \\ &= -\log\{1+\exp(-\xi)\} - \xi/2 - \frac{1}{4\xi}\tanh(\xi/2)(x^2-\xi^2). \qquad (36)\end{aligned}$$

To quantify the gap $\Delta(\beta,\beta^{\mathbf{o}}) := \log p_l(y \mid \beta, \xi, \mathbf{X}) - \log p(y \mid \beta^{\mathbf{o}}, \mathbf{X})$, observe that

$$\begin{aligned}\Delta(\beta,\beta^{\mathbf{o}}) &= \log p(y \mid \beta, \mathbf{X}) - \log p(y \mid \beta^{\mathbf{o}}, \mathbf{X}) + \log p_l(y \mid \beta, \xi, \mathbf{X}) - \log p(y \mid \beta, \mathbf{X}) \qquad (37) \\ &:= y^{\mathrm{T}}\mathbf{X}(\beta-\beta^{\mathbf{o}}) + \mathbb{1}_n^{\mathrm{T}}[\log(1+\exp(\mathbf{X}\beta^{\mathbf{o}}) - \log(1+\exp(\mathbf{X}\beta)] + \Delta\end{aligned}$$

where Δ is the *Jensen-Gap* in (36). To estimate Δ, we perform a second order Taylor-expansion around ξ^2

$$f(x) = f(\xi) + \frac{df(\xi)}{d\xi^2}(x^2-\xi^2) + \frac{1}{2}\frac{d^2 f(\xi)}{d\xi^4}|_{\xi=\hat{\xi}}(x^2-\xi^2)^2.$$

Observe further,

$$\frac{df(\xi)}{d\xi^2} = -\frac{1}{4\sqrt{\xi^2}}\tanh\frac{\sqrt{\xi^2}}{2}, \quad \frac{d^2 f(\xi)}{d\xi^4} = -\Big[\frac{0.0625\,\mathrm{sech}^2(\sqrt{\xi^2}/2)}{\xi^2} - \frac{0.125\tanh(\sqrt{\xi^2}/2)}{(\xi^2)^{1.5}}\Big].$$

Moreover, $d^2 f(\xi)/d\xi^4$ is a decreasing function of ξ^2 and $0 < d^2 f(\xi)/d\xi^4 < 1$. Hence $\Delta \leq \sum_{i=1}^n\{(\mathbf{x}_i^{\mathrm{T}}\beta)^2 - \xi_i^2\}^2$. Setting $\xi_i = \mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}}$ for all $i=1,\ldots,n$, we have

$$\begin{aligned}\Delta &\leq \sum_{i=1}^n\{\mathbf{x}_i^{\mathrm{T}}(\beta-\beta^{\mathbf{o}})\}^2\{\mathbf{x}_i^{\mathrm{T}}(\beta-\beta^{\mathbf{o}}) + 2\mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}}\}^2 \\ &\leq 2\sum_{i=1}^n\{\mathbf{x}_i^{\mathrm{T}}(\beta-\beta^{\mathbf{o}})\}^4 + 8\sum_{i=1}^n\{\mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}}\}^2\{\mathbf{x}_i^{\mathrm{T}}(\beta-\beta^{\mathbf{o}})\}^2. \\ &\leq 2n\|\mathbf{X}\|_{2,\infty}^4\|\beta-\beta^{\mathbf{o}}\|^4 + 8n\|\mathbf{X}\|_{2,\infty}^4\|\beta^{\mathbf{o}}\|_2^2\|\beta-\beta^{\mathbf{o}}\|^2,\end{aligned}$$

where the final inequality follows from $\mathbf{x}_i^{\mathrm{T}}(\beta - \beta^{\mathbf{o}}) \leq \|\mathbf{X}\|_{2,\infty}\|\beta - \beta^{\mathbf{o}}\|$. Plugging in the bound obtained above in $\Delta(\beta, \beta^{\mathbf{o}})$, we get

$$\begin{aligned}\Delta(\beta, \beta^{\mathbf{o}}) &\leq y'\mathbf{X}(\beta - \beta^{\mathbf{o}}) + \mathbb{1}_n^{\mathrm{T}}[\log(1 + \exp(\mathbf{X}\beta^{\mathbf{o}}) - \log(1 + \exp(\mathbf{X}\beta)] + \sum_{i=1}^n \{(\mathbf{x}_i^{\mathrm{T}}\beta)^2 - (\mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}})^2\}^2 \\ &\leq \sum_{i=1}^n \{y_i + 1\}\mathbf{x}_i^{\mathrm{T}}(\beta - \beta^{\mathbf{o}}) + 2n\|\mathbf{X}\|_{2,\infty}^4\|\beta - \beta^{\mathbf{o}}\|^4 + 8n\|\mathbf{X}\|_{2,\infty}^4\|\beta^{\mathbf{o}}\|_2^2\|\beta - \beta^{\mathbf{o}}\|^2 \\ &\leq 2n\|\mathbf{X}\|_{2,\infty}\|\beta - \beta^{\mathbf{o}}\| + 2n\|\mathbf{X}\|_{2,\infty}^4\|\beta - \beta^{\mathbf{o}}\|^4 + 8n\|\mathbf{X}\|_{2,\infty}^4\|\beta^{\mathbf{o}}\|_2^2\|\beta - \beta^{\mathbf{o}}\|^2.\end{aligned}$$

where the last inequality is obtained by noting that $\log(1 + e^x)$ is a 1-Lipschitz function. Recall that $L(\beta^{\mathbf{o}}, \mathbf{X}) = \max\{4\|\mathbf{X}\|_{2,\infty}, 8\|\mathbf{X}\|_{2,\infty}^2\|\beta^{\mathbf{o}}\|_2\}$. If $\|\beta - \beta^{\mathbf{o}}\| < \varepsilon^2/L(\beta^{\mathbf{o}}, \mathbf{X})$, then $\Delta(\beta, \beta^{\mathbf{o}}) \leq n\varepsilon^2$ which implies

$$n^{-1}\,\widetilde{\mathrm{D}}\big[p(\cdot \mid \beta^{\mathbf{o}}, \mathbf{X}) \,\big|\big|\, p_l(\cdot \mid \beta, \xi, \mathbf{X})\big] \leq \varepsilon^2.$$

Also, since

$$V\big[p(y \mid \beta^{\mathbf{o}}, \mathbf{X}) \,\big|\big|\, p(y \mid \beta, \xi, \mathbf{X})\big] = nV\big[p(y_1 \mid \beta^{\mathbf{o}}, \mathbf{x}_1) \,\big|\big|\, p(y_1 \mid \beta, \xi, \mathbf{x}_1)\big],$$

following the same argument as before but with one observation, we have

$$\begin{aligned}\Delta_1(\beta, \beta^{\mathbf{o}}) &:= \log p_l(y_1 \,|\, \beta, \xi, \mathbf{x}_1) - \log p(y_1 \,|\, \beta^{\mathbf{o}}, \mathbf{x}_1) \\ &\leq 2\|\mathbf{X}\|_{2,\infty}\|\beta - \beta^{\mathbf{o}}\| + 2\|\mathbf{X}\|_{2,\infty}^4\|\beta - \beta^{\mathbf{o}}\|^4 + 8\|\mathbf{X}\|_{2,\infty}^4\|\beta^{\mathbf{o}}\|_2^2\|\beta - \beta^{\mathbf{o}}\|^2,\end{aligned}$$

which implies if $\|\beta - \beta^{\mathbf{o}}\| < \varepsilon^2/L(\beta^{\mathbf{o}}, \mathbf{X})$, then $n^{-1}\,\mathrm{V}\big[p(\cdot \mid \beta^{\mathbf{o}}, \mathbf{X}) \,\big|\big|\, p_l(\cdot \mid \beta, \xi, \mathbf{X})\big] \leq \varepsilon^2$. Hence

$$-\log\big\{\pi_\beta[\mathcal{B}_n(\beta^{\mathbf{o}}, \varepsilon)]\big\} \leq -\log \pi_\beta\big\{\|\beta - \beta^{\mathbf{o}}\| < \varepsilon^2/L(\beta^{\mathbf{o}}, \mathbf{X})\big\} \leq p\log\left\{\frac{L(\beta^{\mathbf{o}}, \mathbf{X})}{\varepsilon^2}\right\} + \frac{1}{2}\beta^{\mathbf{o}\mathrm{T}}\Sigma_\beta^{-1}\beta^{\mathbf{o}},$$

where the final inequality holds using using multivariate Gaussian concentration through Anderson's inequality.

A.2 Proof of Theorem 4

We start by rewriting the log-likelihood ratio as

$$\begin{aligned}\log\frac{p(y \mid \beta, \mathbf{X})}{p(y \mid \beta^{\mathbf{o}}, \mathbf{X})} &= (\beta - \beta^{\mathbf{o}})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}y - \sum_{i=1}^n [a(\mathbf{x}_i^{\mathrm{T}}\beta) - a(\mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}})] \\ &= \langle y - \mathbb{E}y, \mathbf{X}(\beta - \beta^{\mathbf{o}})\rangle - n\,D(\beta^{\mathbf{o}}, \beta).\end{aligned}$$

Since $a(t) = \log(1+e^t)$ satisfies $a(t+h) \geq a(t) + h\,a^{(1)}(t) + \mathrm{r}(|h|)\,a^{(2)}(t)/2$ for all t, h, where $r(h) = h^2/(\mathrm{r}_1 h + 1)$ for $\mathrm{r}_1 > 0$, we have

$$n\,\mathrm{D}(\beta^{\mathbf{o}}, \beta) = \sum_{i=1}^n \big\{a(\mathbf{x}_i^{\mathrm{T}}\beta) - a(\mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}}) - a^{(1)}(\mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}})\mathbf{x}_i^{\mathrm{T}}(\beta - \beta^{\mathbf{o}})\big\} \geq \sum_{i=1}^n \mathrm{r}(|\mathbf{x}_i^{\mathrm{T}}(\beta - \beta^{\mathbf{o}})|)a^{(2)}(\mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}}).$$

Defining $\mathrm{k}(h) = h^2/\mathrm{r}(h)$ and $W = \mathrm{diag}[a^{(2)}(\mathbf{x}_1^{\mathrm{T}}\beta^{\mathbf{o}}), \dots, a^{(2)}(\mathbf{x}_n^{\mathrm{T}}\beta^{\mathbf{o}})]$,

$$\begin{aligned}
n\ \mathrm{D}(\beta^{\mathbf{o}}, \beta) &\geq (\beta - \beta^{\mathbf{o}})^{\mathrm{T}}\Big[\sum_{i=1}^{n} \frac{a^{(2)}(\mathbf{x}_i^{\mathrm{T}}\beta^{\mathbf{o}})}{\mathrm{k}(|\mathbf{x}_i^{\mathrm{T}}(\beta - \beta^{\mathbf{o}})|)}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\Big](\beta - \beta^{\mathbf{o}}), \\
&\geq \frac{(\beta - \beta^{\mathbf{o}})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}W\mathbf{X}(\beta - \beta^{\mathbf{o}})}{1 + \mathrm{r}_1\|\mathbf{X}\|_\infty\sqrt{p}\|\beta - \beta^{\mathbf{o}}\|}, \\
&\geq \frac{n\kappa_2\,\|\beta - \beta^{\mathbf{o}}\|^2}{1 + \mathrm{r}_1\|\mathbf{X}\|_\infty\sqrt{p}\|\beta - \beta^{\mathbf{o}}\|},
\end{aligned}$$

where the second last inequality follows $|\mathbf{x}_i^{\mathrm{T}}(\beta - \beta^{\mathbf{o}})| \leq \|\mathbf{X}\|_\infty\sqrt{p}\|\beta - \beta^{\mathbf{o}}\|$. Define Ω_n be the set

$$\max_{1\leq j\leq p}\Big|\sum_{i=1}^{n}(y_i - \mathbb{E}y_i)\,x_{ij}\Big| \leq \|\mathbf{X}\|_\infty(n\log p)^{1/2}/2.$$

Setting $\zeta_p := \|\mathbf{X}\|_\infty\sqrt{np\log p}/2$, it follows that in Ω_n, $\langle y - \mathbb{E}y, \mathbf{X}(\beta - \beta^{\mathbf{o}})\rangle \leq \zeta_p\|\beta - \beta^{\mathbf{o}}\|$. Set $\widetilde{\varepsilon} = (n\kappa_2 + 2\zeta_p r_1\sqrt{p}\|\mathbf{X}\|_\infty)/\{r_1\|\mathbf{X}\|_\infty\sqrt{p}\,(n\kappa_2 - 2\zeta_p r_1\|\mathbf{X}\|_\infty\sqrt{p})\}$, then whenever $\sqrt{n} > r_1\|\mathbf{X}\|_\infty^2 p\sqrt{\log p}/\kappa_2$, for any β such that $\|\beta-\beta^{\mathbf{o}}\| \geq \widetilde{\varepsilon}$, we have, $n\ \mathrm{D}(\beta^{\mathbf{o}}, \beta) > 2\langle y-\mathbb{E}y, \mathbf{X}(\beta-\beta^{\mathbf{o}})\rangle$ inside Ω_n which further leads to

$$\log\frac{p_l(y \mid \beta, \xi, \mathbf{X})}{p(y \mid \beta^{\mathbf{o}}, \mathbf{X})}\mathbb{1}_{\Omega_n} \leq \log\frac{p(y \mid \beta, \mathbf{X})}{p(y \mid \beta^{\mathbf{o}}, \mathbf{X})}\mathbb{1}_{\Omega_n} \leq -n\ \mathrm{D}(\beta^{\mathbf{o}}, \beta)/2.$$

Let us define the following probability measure on Ω_n

$$\widetilde{p}(y \mid \beta^{\mathbf{o}}, \mathbf{X}) = \frac{p(y \mid \beta^{\mathbf{o}}, \mathbf{X})\,\mathbb{1}_{\Omega_n}}{\int_{\Omega_n} p(y \mid \beta^{\mathbf{o}}, \mathbf{X})dy}$$

Also, let us define the following finite measure $\widetilde{p}_l(y \mid \beta, \xi, \mathbf{X}) = p_l(y \mid \beta, \xi, \mathbf{X})\mathbb{1}_{\Omega_n}$. Then, for any $\delta > 0$, define $B(\beta_1;\delta) = \{\widetilde{p}_l(\cdot \mid \beta, \xi, \mathbf{X}) : \|\beta - \beta_1\| < \delta\}$. Denote by $\mathrm{conv}\{B(\beta_1;\delta)\}$ the convex hull of $B(\beta_1;\delta)$. Pick any β_1 such that $\|\beta_1 - \beta^{\mathbf{o}}\| = r$. Then, from Lemma 2 of Bhattacharya and Pati (2020) and the assumption $\sqrt{n} \geq r_1 p\sqrt{\log p}\|\mathbf{X}\|_\infty^2/\kappa_2$, there exists measurable functions $0 \leq \Phi_n \leq 1$ such that for every $n \geq 1$ and $\alpha \in (0,1)$

$$\sup_{\widetilde{p}_l(\cdot|\beta,\xi,\mathbf{X})\in\mathrm{conv}\{B(\beta_1;r/2)\}} \mathbb{E}_{\widetilde{p}(\cdot|\beta^{\mathbf{o}},\mathbf{X})}\Phi_n + \mathbb{E}_{\widetilde{p}_l(\cdot|\beta,\xi,\mathbf{X})}(1 - \Phi_n) \quad\leq\quad \exp\Big\{-\frac{\alpha\,n\kappa_2\|\beta - \beta^{\mathbf{o}}\|}{8r_1\,\sqrt{p}\,\|\mathbf{X}\|_\infty}\Big\}. \tag{38}$$

Now, for any event $\mathcal{A} \in Y^{(n)}$, one can write

$$\begin{aligned}
\mathbb{P}_{\beta^{\mathbf{o}}}(\mathcal{A}^c) &= \mathbb{P}_{\beta^{\mathbf{o}}}(\mathcal{A}^c \mid \Omega_n)\mathbb{P}_{\beta^{\mathbf{o}}}(\Omega_n) + \mathbb{P}_{\beta^{\mathbf{o}}}(\mathcal{A}^c \mid \Omega_n^c)\mathbb{P}_{\beta^{\mathbf{o}}}(\Omega_n^c) \\
&\leq \mathbb{P}_{\beta^{\mathbf{o}}}(\mathcal{A}^c \mid \Omega_n) + \mathbb{P}_{\beta^{\mathbf{o}}}(\Omega_n^c)
\end{aligned} \tag{39}$$

We know that $\mathbb{P}_{\beta^{\mathbf{o}}}(\Omega_n^c) \leq 2/p$ and show in Lemma 14 that with high probability (w.r.t. $\mathbb{P}_{\beta^{\mathbf{o}}}(\cdot \mid \Omega_n)$),

$$\int\Big[\exp\{\ell_n(\beta, \beta^{\mathbf{o}}) + (\gamma/4)n\ \mathrm{D}(\beta^{\mathbf{o}}, \beta)\}\Big]\pi_\beta(\beta)d\beta \leq 3e^{n\gamma\kappa_1\varepsilon^2/16}, \tag{40}$$

where $\ell_n(\beta, \beta^{\mathbf{o}}) = \log\{p_l(y \mid \beta, \xi, \mathbf{X})/p(y \mid \beta^{\mathbf{o}}, \mathbf{X})\}$. And hence provide a high probability bound w.r.t. $\mathbb{P}_{\beta^{\mathbf{o}}}$. Next, we use the variational inequality (31) with $\mu = \pi_\beta$, $\rho = \hat{q}_\beta$ to show with high probability

$$\frac{\gamma}{4}\int n\ \mathrm{D}(\beta^{\mathbf{o}}, \beta)\, q^*_\beta(\beta)\, d\beta \leq -\int_\beta \log\frac{p_l(y \mid \beta, \xi, \mathbf{X})}{p(y \mid \beta^{\mathbf{o}}, \mathbf{X})}\, q^*_\beta(\beta)\, d\beta + \mathrm{D}(q^*_\beta \,||\, \pi_\beta) + n\kappa_1\gamma\varepsilon^2/16 + \log 3.$$

This brings us back to the proof of Theorem 2 and the remaining part of the proof of Theorem 4 follows verbatim from the proof of Theorem 2.

A.3 Auxiliary Results for Proofs in Section 3

Lemma 13 *Let u and v denote two continuous random vectors with joint density function $p(u, v)$. The maximum value of*

$$\int q(u)\log\left\{\frac{p(u, v)}{q(u)}\right\}du$$

over all density functions q is attained by $q^*(u) = p(u \mid v)$.

The following lemma is derived similar to Lemma 14 in Atchadé (2017) and Theorem 2 in Kleijn and van der Vaart (2006).

Lemma 14 *Fix any $\gamma \in (0, 1)$ and $\varepsilon \geq 2\widetilde{\varepsilon}$, also set $\epsilon = \kappa_2\varepsilon/(8\mathrm{r}_1\|X\|_\infty\sqrt{p})$. If $\sqrt{n} \geq r_1 p\sqrt{\log p}\|\mathbf{X}\|^2_\infty/\kappa_2$, then with probability $1 - e^{-n\gamma\epsilon/8} - e^{-n\gamma\kappa_1\varepsilon^2/32}$ in $\mathbb{P}_{\beta^{\mathbf{o}}}(\cdot \mid \Omega_n)$,*

$$\int \exp\left\{\ell_n(\beta, \beta^{\mathbf{o}}) + (\gamma/4)n\ D(\beta^{\mathbf{o}}, \beta)\right\}\pi_\beta(\beta)d\beta \leq 3e^{n\gamma\kappa_1\varepsilon^2/16}. \tag{41}$$

Proof Writing $\eta(\beta, \beta^{\mathbf{o}}) = \exp\left\{\ell_n(\beta, \beta^{\mathbf{o}}) + (\gamma/4)n\ \mathrm{D}(\beta^{\mathbf{o}}, \beta)\right\}$ and

$$U := \{\beta : \|\beta - \beta^{\mathbf{o}}\| > \varepsilon\} = \bigcup_{j=1}^{\infty} U_{j,n} \tag{42}$$

where $U_{j,n} = \{\beta : j\varepsilon < \|\beta - \beta^{\mathbf{o}}\| < (j+1)\varepsilon\}$, we express

$$\begin{aligned}\int \eta(\beta, \beta^{\mathbf{o}})\pi_\beta(\beta)d\beta &= \int_{U^c}\eta(\beta, \beta^{\mathbf{o}})\pi_\beta(\beta)d\beta + \int_U \tilde{\Phi}_n\eta(\beta, \beta^{\mathbf{o}})\pi_\beta(\beta)d\beta \\ &\quad + \sum_{j=1}^{\infty}\int_{U_{j,n}}(1 - \tilde{\Phi}_n)\eta(\beta, \beta^{\mathbf{o}})\pi_\beta(\beta)d\beta \\ T &= T_1 + S_1, \quad S_1 = T_2 + \sum_{j=1}^{\infty} T_{2j}\end{aligned}$$

for any sequence of test functions $\{\tilde{\Phi}_n : n \geq 1\}$. Then

$$\mathbb{E}_{\widetilde{p}(\cdot|\beta^{\mathbf{o}}, \mathbf{X})}T_1 \leq \int_{U^c} e^{(\gamma/4)n\ \mathrm{D}(\beta^{\mathbf{o}}, \beta)}\pi_\beta(\beta)d\beta \leq \int_{U^c} e^{(\gamma/4)n\kappa_1\|\beta - \beta^{\mathbf{o}}\|^2/8}\pi_\beta(\beta)d\beta \leq e^{n\gamma\kappa_1\varepsilon^2/32}.$$

The above ineqaulity follows from the fact that $n\mathrm{D}(\beta^{\mathbf{o}}, \beta) \leq n\kappa_1\|\beta - \beta^{\mathbf{o}}\|^2/8$, since $a^{(2)}(x) \leq 1/4$ for all $x \in \mathbb{R}$. By Markov's inequality, $T_1 \leq e^{n\gamma\kappa_1\varepsilon^2/16}$ with probability $1 - e^{-n\gamma\kappa_1\varepsilon^2/32}$. To bound T_2 and $T_{2j}, j = 1, \ldots, \infty$, we detail the construction of $\tilde{\Phi}_n$. Let $N_{j,n} := N(j\varepsilon/2, U_{j,n}, \|\cdot\|)$ denote the $j\varepsilon/2$-covering number of $U_{j,n}$ with respect to $\|\cdot\|$. For each $j \geq 1$, let S_j be a maximal $j\varepsilon/2$-separated points in $U_{j,n}$ and for each point $\tilde{\beta}_k \in S_j$ we can construct a test function $\Phi_{n,\tilde{\beta}_k}$ as in (38), with $r = j\varepsilon$. Then we set $\tilde{\Phi}_n$ to $\tilde{\Phi}_n = \sup_{j\geq 1} \max_{\tilde{\beta}_k \in S_j} \Phi_{n,\tilde{\beta}_k}$. Note also that

$$\sup_{\widetilde{p}_l(\cdot|\beta,\xi,\mathbf{X}) \in \mathrm{conv}\{B(\beta_1; r/2)\}} \mathbb{E}_{\widetilde{p}(\cdot|\beta^{\mathbf{o}},\mathbf{X})}\Phi_n + \mathbb{E}_{\widetilde{p}_l(\cdot|\beta,\xi,\mathbf{X})}(1 - \Phi_n) \quad \leq \quad \exp\Big\{ - \frac{\alpha\, n\kappa_2\|\beta - \beta^{\mathbf{o}}\|}{8r_1\sqrt{p}\,\|\mathbf{X}\|_\infty}\Big\}.$$

where the second last inequality follows since $\sqrt{n} \geq r_1 p\sqrt{\log p}\|\mathbf{X}\|_\infty^2/\kappa_2$ and $\|\beta - \beta^{\mathbf{o}}\| > j\varepsilon/2 \geq \widetilde{\varepsilon}$ for $j \geq 1$. Then

$$\mathbb{E}_{\widetilde{p}(\cdot|\beta^{\mathbf{o}},\mathbf{X})}\{\tilde{\Phi}_n\} \quad \leq \quad \sum_{j=1} N_{j,n} \exp\Big\{ - \alpha\frac{n\kappa_2 j\varepsilon}{8\mathrm{r}_1\|\mathbf{X}\|_\infty\sqrt{p}}\Big\} \leq \exp\Big\{ - C\frac{n\kappa_2\varepsilon}{8\mathrm{r}_1\|\mathbf{X}\|_\infty\sqrt{p}}\Big\} := e^{-Cn\epsilon},$$

for some constant $C > 0$. From Markov's inequality we obtain $\tilde{\Phi}_n \leq e^{-Cn\epsilon/2}$ with probability at least $1 - e^{-Cn\epsilon/2}$. Hence $T_2 \leq S_1 e^{-Cn\epsilon/2}$ with probability atleast $1 - e^{-Cn\epsilon/2}$. Finally note that

$$\mathbb{E}_{\widetilde{p}(\cdot|\beta^{\mathbf{o}},\mathbf{X})} \sum_{j=1}^{\infty} T_{2j} \leq \int_U e^{-(\gamma/4)n\ \mathrm{D}(\beta^{\mathbf{o}},\beta)}\pi_\beta(\beta)d\beta \leq \exp\Big\{ - \frac{\gamma n\kappa_2\varepsilon}{16\mathrm{r}_1\|\mathbf{X}\|_\infty\sqrt{p}}\Big\} := \exp\{-\gamma n\epsilon/2\}.$$

The penultimate inequality follows from the fact that $\widetilde{\phi} \in (0,1)$ and for any $\gamma \in (0,1)$, one can write, $\ell_n(\beta, \beta^{\mathbf{o}}) + \gamma n\mathbb{D}(\beta^{\mathbf{o}}, \beta)/4 \leq -\gamma n\mathbb{D}(\beta^{\mathbf{o}}, \beta)/4$ inside Ω_n. The last inequality follows because whenever $\sqrt{n} > r_1\|\mathbf{X}\|_\infty^2 p\sqrt{\log(p)}/\kappa_2$ and $\|\beta - \beta^{\mathbf{o}}\|_2 \geq \epsilon/2 \geq \widetilde{\epsilon}$ we have $n\mathbb{D}(\beta^{\mathbf{o}}, \beta) \geq n\kappa_2\|\beta - \beta^{\mathbf{o}}\|_2/\{2r_1\sqrt{p}\|\mathbf{X}\|_\infty\}$. Then, $E_{\beta^{\mathbf{o}}}S_1 \leq e^{-Cn\epsilon/2}E_{\beta^{\mathbf{o}}}S_1 + e^{-n\gamma\epsilon/2}$, whence $E_{\beta^{\mathbf{o}}}S_1 \leq 2e^{-n\gamma\epsilon/4}$ for sufficiently large n. Hence $S_1 \leq 2e^{-n\gamma\epsilon/8}$ with probability at least $1 - e^{-n\gamma\epsilon/8}$ in $\mathbb{P}_{\beta^{\mathbf{o}}}(\cdot \mid \Omega_n)$. ■

Appendix B. Review of Dynamical Systems & Notion of Stability

Dynamical systems theory is a classical technique that deals with stability and convergence of complex iterative methods. We call a dynamical system to be discrete-time if the system is observed on discrete time points $\{t_0, t_1, t_2, \ldots\}$. Usually, we consider the time-points to be evenly placed, i.e. $t_{j+1} = t_j + h$ for some $h > 0$. Moreover, a system is considered autonomous if the function is independent of time and non-autonomous otherwise. In this section, we will discuss the notion of stability for discrete time autonomous systems. Let us consider the following discrete-time autonomous system given by,

$$\psi^{t+1} = f(\psi^t), \quad t \in \mathbb{N} \tag{43}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$(or, $f : \mathbb{D} \to \mathbb{R}^n, \mathbb{D} \subseteq \mathbb{R}^n$) is a diffeomorphism, i.e. a smooth function with smooth inverse and $\psi^t \in \mathbb{R}^n$. $\psi^* \in \mathbb{R}^n$ is called a fixed point to this system if $\psi^* = f(\psi^*)$. We recall the following definition from Bof et al. (2018).

Definition 15 *A fixed point ψ^* of a system given by* (43) *is called*

(a) locally stable if given any $\epsilon > 0$, there exists $\delta = \delta(\epsilon)$ such that, whenever $\|\psi^0 - \psi^*\| < \delta$, we have $\|f(\psi^t) - \psi^*\| < \epsilon$ for all t.

(b) locally asymptotically stable if it is stable and δ can be chosen such that, whenever $\|\psi^0 - \psi^*\| < \delta$, we have $\psi^t \to \psi^*$ as $t \to \infty$.

(c) locally unstable if it is not locally stable.

The *locality* in the definition is used to denote the fact that we are initializing the system in a δ-ball around the fixed point. We say the stability is *global* if the system converges to the fixed point independent of the initialization, i.e. we can initialize at any point in the function domain.

Lemma 16 *Consider a system $x^{t+1} = g(x^t)$ where $g : \mathbb{D} \to \mathbb{R}$ $(\mathbb{D} \subseteq \mathbb{R})$ with a fixed point x^* such that, $|g'(x)| \leq \delta$ for all $x \in \mathbb{D} - \{x^*\}$, for some $\delta < 1$. Then, x^* is globally asymptotically stable.*

Proof Given a fixed point $x^* = g(x^*)$, use the Mean Value Theorem to get, $|x^{t+1} - x^*| = |g'(x)||x^t - x^*|$, for some $x \in (x^t, x^*)$. Since, $|g'(x)| \leq \delta$, we have, $|x^{t+1} - x^*| \leq \delta^{t+1}|x^0 - x^*|$. This implies $|x^{t+1} - x^*| \to 0$ as $t \to \infty$. ■

Let β_k be the k-th coordinate of a vector $\beta \in \mathbb{R}^n$. Consider the system in (43) with a fixed point ψ^*. Using generalized Taylor's theorem we get,

$$\psi_k^{t+1} - \psi_k^* = f_k(\psi^t) - f_k(\psi^*) = \nabla f_k(\psi^*)(\psi^t - \psi^*) + h(\psi^t)|\psi^t - \psi^*|,$$

where $\nabla f_k(\psi)$ is the gradient vector with i^{th} entry given by $\partial f_k(\psi)/\partial \psi_i$ and $h : \mathbb{R}^n \to \mathbb{R}$ such that, $\lim_{\psi \to \psi^*} h(\psi) = 0$. If ψ_t is close to ψ^*, the convergence of the system depends on $\nabla f_k(\psi^*)$ by the following approximation,

$$(\psi^{t+1} - \psi^*) \approx \mathbf{J}(\psi^t - \psi^*), \tag{44}$$

where $\mathbf{J}$ is the $n \times n$ Jacobian matrix evaluated at ψ^* with i^{th} row given by $\nabla f_i^{\mathrm{T}}(\psi^*)$. Thus the behavior of the dynamical system (43) around a small neighbourhood of ψ^* is exactly same as that of the linearization in (44). This is formalized in the Hartman-Grobman theorem.

Definition 17 *A fixed point ψ^*, for a map $\psi \to f(\psi)$, $\psi \in \mathbb{R}^n$ is called* hyperbolic *if none of the eigenvalues of* $\mathbf{J}$ *has magnitude 1.*

Theorem 18 (Hartman & Grobman) *In a neighborhood of a hyperbolic fixed point, a diffeomorphism is topologically conjugate to the derivative at that fixed point.*

The theorem above asserts that the behavior of a system around a hyperbolic fixed point is essentially same as the linearization near this point. Refer to Quandt (1986) for a complete review. This motivates us to check stability of a fixed point using Lemma 20 . Refer to Wiggins (2003); Barbarossa (2011) for a proof and for further reading on this topic.

Definition 19 *For a square matrix A, the spectral radius $\rho(A)$ is defined by*

$$\rho(A) := \max\{|\lambda| : \lambda \textit{ is eigenvalue of } A\}$$

Lemma 20 *Let ψ^* be a fixed point solution to the discrete-time autonomous system given by $\psi_{t+1} = f(\psi_t)$. Suppose, $f : \mathbb{D} \to \mathbb{R}^n (\mathbb{D} \subseteq \mathbb{R}^n)$ is a twice continuously differentiable function around a neighbourhood $\mathbb{D}$ of ψ^*. Let $\mathbf{J} = [\partial_i f(\psi)/\partial \psi_j]_{\psi=\psi^*}$ be the Jacobian matrix of f evaluated at ψ^*. Then,*

(a) ψ^* is locally asymptotically stable if $\rho(\mathbf{J}) < 1$.

(b) ψ^* is locally unstable if at least one eigenvalue of $\mathbf{J}$ is greater than one in absolute value.

Lemma 20 along with Theorem 18 provides sufficient conditions for the local convergence of a system. Consider the linear system given by $\alpha^{(t+1)} = A\alpha^{(t)}$, with a fixed point $\alpha^* = 0$. Let us consider, $A\nu_i = \lambda_i \nu_i \, (i = 1, 2, \ldots, n)$ and $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$. Suppose A has a complete set of eigenvectors, i.e. the set of eigenvectors $\{\nu_1, \nu_2, \ldots, \nu_n\}$ form a basis of $\mathbb{R}^n$. Then it can be easily seen that a solution to the system is $\alpha^{(t)} = c_1\lambda_1^t\nu_1 + c_2\lambda_2^t\nu_2 + \ldots + c_n\lambda_n^t\nu_n$ for some arbitrary constants $c_1, c_2, \ldots, c_n$. Also, $c_1\lambda_1^t\nu_1 + c_2\lambda_2^t\nu_2 + \ldots + c_n\lambda_n^t\nu_n \to 0$ iff $|\lambda_1| < 1$. This illustrates the Lemma above, in the most simplistic scenario, can be extended to the case where A does not have a complete set of eigenvectors using *Jordan Canonical form* of A (refer to Wood and O'Neill (2003)).

Appendix C. Proofs of Algorithmic Convergence Results in Section 4

Definition 21 *Consider two $n \times n$ real symmetric matrices A & B. Then we write,*

(a) $B \precsim A$ if for any $a \in \mathbb{R}^n$ such that, $a \neq 0$; we have, $a^{\mathrm{T}}(A - B)a \geq 0$, i.e. $A - B$ is a positive semi-definite matrix.

(b) $B \prec A$ if for any $a \in \mathbb{R}^n$ such that, $a \neq 0$; we have, $a^{\mathrm{T}}(A - B)a > 0$, i.e. $A - B$ is a positive definite matrix.

In the following, we first provide the proof of Theorems 7 in §C.1, calculation of spectral radius for $p = 1$ in §C.2 and then provide the proofs of some of the auxiliary results used in subsequent §C.3.

C.1 Proof of Theorem 7

For some fixed $\alpha \in (0, 1]$, the update equation in (8) can be rewritten as,

$$(\xi^{t+1})^2 = \mathrm{diag}[\mathbf{X}\{\Sigma_\alpha(\xi^t)/\alpha + \Sigma_\alpha(\xi^t)B_\alpha B_\alpha^{\mathrm{T}}\Sigma_\alpha(\xi^t)\}\mathbf{X}^{\mathrm{T}}].$$

where $B_\alpha = [\mathbf{X}^{\mathrm{T}}(Y - 1/2\,\mathbb{1}_n) + \Sigma_\beta^{-1}\mu_\beta/\alpha]$ and $\Sigma_\alpha(\xi) = [\Sigma_\beta^{-1}/\alpha - 2\mathbf{X}^{\mathrm{T}}\mathrm{diag}\{A(\xi)\}\mathbf{X}]^{-1}$. We calculate the partial derivatives in order to get the Jacobian matrix,

$$\frac{\partial \left(\xi_i^{t+1}\right)^2}{\partial \left(\xi_j^t\right)^2} = \frac{A'(\xi_j^t)}{\xi_j^t}\mathbf{x}_i^T \left[\Sigma_\alpha\left(\xi^t\right)\mathbf{x}_j\mathbf{x}_j^{\mathrm{T}}\Sigma_\alpha\left(\xi^t\right)/\alpha + 2\Sigma_\alpha\left(\xi^t\right)\mathbf{x}_j\mathbf{x}_j^{\mathrm{T}}\Sigma_\alpha\left(\xi^t\right)B_\alpha B_\alpha^{\mathrm{T}}\Sigma_\alpha\left(\xi^t\right)\right]\mathbf{x}_i.$$

Then Jacobian Matrix($\mathbf{J}_\alpha$) at $\xi = \xi^t$ is given by,

$$\mathbf{J}_\alpha = \left[\mathbf{X}\Sigma_\alpha\left(\xi^t\right)\mathbf{X}^{\mathrm{T}} \circ \mathbf{X}\left\{\Sigma_\alpha\left(\xi^t\right)/\alpha + 2\mu_\alpha\left(\xi^t\right)\mu_\alpha^{\mathrm{T}}\left(\xi^t\right)\right\}\mathbf{X}^{\mathrm{T}}\right]\mathrm{diag}\left(\frac{A'\left(\xi^t\right)}{\xi^t}\right), \tag{45}$$

where $\circ$ denotes the Hadamard Product. Let us denote the maximum eigenvalue of a matrix A by $\lambda_1(A)$. Our objective is to show that $\lambda_1(\mathbf{J}_\alpha)|_{\xi=\xi^*} < 1$. We call $D = \operatorname{diag}(A'(\xi)/\xi)$. By Lemma 22 $D^{1/2}\left[\mathbf{X}\Sigma_\alpha(\xi^*)\mathbf{X}^{\mathrm{T}} \circ \mathbf{X}\left\{\Sigma_\alpha(\xi^*)/\alpha + 2\mu_\alpha(\xi^*)\mu_\alpha^{\mathrm{T}}(\xi^*)\right\}\mathbf{X}^{\mathrm{T}}\right]D^{1/2}$ has the same set of eigenvalues as with $\mathbf{J}_\alpha|_{\xi=\xi^*}$. $\mathbf{X}\Sigma_\alpha(\xi^*)\mathbf{X}^{\mathrm{T}}$ and $\mathbf{X}\left\{\Sigma_\alpha(\xi^*)/\alpha + 2\mu_\alpha(\xi^*)\mu_\alpha^{\mathrm{T}}(\xi^*)\right\}\mathbf{X}^{\mathrm{T}}$ are positive semi-definite matrices which imply $D^{1/2}\left[\mathbf{X}\Sigma_\alpha(\xi^*)\mathbf{X}^{\mathrm{T}} \circ \mathbf{X}\left\{\Sigma_\alpha(\xi^*)/\alpha + 2\mu_\alpha(\xi^*)\mu_\alpha^{\mathrm{T}}(\xi^*)\right\}\mathbf{X}^{\mathrm{T}}\right]D^{1/2}$ is positive semi-definite as well as symmetric. Since the eigenvalues of a real symmetric positive semi-definite matrix are real and non-negative, the eigenvalues of $\mathbf{J}_\alpha|_{\xi=\xi^*}$ are real and non-negative. We denote ξ^* by ξ in the following discussion for notational simplicity.

From the assumptions of the theorem and fixed point equation, it is clear that $\xi_i > 0$ for all $i \in \{1, 2, \ldots, n\}$. We begin with the fact that, $A(x) + xA'(x) < 0$ for all $x \in \mathbb{R}$. Then, recalling the Definition 21, we have the following,

$$2[\mathbf{X}^{\mathrm{T}}\operatorname{diag}\{A(\xi) + \xi A'(\xi)\}\mathbf{X}] \prec \Sigma_\beta^{-1}/\alpha,$$

Since for all non-zero $a \in \mathbb{R}^p$, we have $a^{\mathrm{T}}(\Sigma_\beta^{-1}/\alpha - 2[\mathbf{X}^{\mathrm{T}}\operatorname{diag}\{A(\xi) + \xi A'(\xi)\}\mathbf{X}])a > 0$, assuming Σ_β to be a positive definite matrix. Then,

$$2\mathbf{X}^{\mathrm{T}}\operatorname{diag}\{\xi A'(\xi)\}\mathbf{X} \prec \Sigma_\beta^{-1}/\alpha - 2\mathbf{X}^{\mathrm{T}}\operatorname{diag}\{A(\xi)\}\mathbf{X}. \tag{46}$$

Now, $\mathbf{X}\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}} = \mathbf{X}[\Sigma_\beta^{-1}/\alpha - 2\mathbf{X}^{\mathrm{T}}\operatorname{diag}\{A(\xi)\}\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}$ is a positive semi-definite matrix. This implies $\mathrm{D}^{1/2}\left[\mathbf{X}\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}} \circ \mathbf{X}\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}}\right]\mathrm{D}^{1/2}/\alpha$ is positive semi-definite by Schur product theorem. Then we have the following,

$$\begin{aligned}&\mathrm{D}^{1/2}\left[\mathbf{X}\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}} \circ \mathbf{X}\left\{\Sigma_\alpha(\xi)/\alpha + 2\mu_\alpha(\xi)\mu_\alpha^{\mathrm{T}}(\xi)\right\}\mathbf{X}^{\mathrm{T}}\right]\mathrm{D}^{1/2}\\ &\precsim 2\,\mathrm{D}^{1/2}\left[\mathbf{X}\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}} \circ \mathbf{X}\left\{\Sigma_\alpha(\xi)/\alpha + \mu_\alpha(\xi)\mu_\alpha^{\mathrm{T}}(\xi)\right\}\mathbf{X}^{\mathrm{T}}\right]\mathrm{D}^{1/2}.\end{aligned} \tag{47}$$

Recall $(\xi)^{\mathrm{T}}_{1\times n} = [\xi_1, \xi_2, \ldots, \xi_n]$ and denote $[\Lambda_\alpha(\xi)]_{p\times p} = \Sigma_\alpha(\xi)/\alpha + \mu_\alpha(\xi)\mu_\alpha^{\mathrm{T}}(\xi)$ and $[Q_\alpha]_{n\times n} = \mathbf{X}\Lambda_\alpha\mathbf{X}^{\mathrm{T}}$. Then $Q_\alpha = \Delta(\xi) \circ \Gamma_\alpha$ where, $\Delta(\xi) = \xi\,\xi^{\mathrm{T}}$ and $\Gamma_\alpha = \operatorname{diag}(1/\xi)\,Q_\alpha\operatorname{diag}(1/\xi)$. Now, Γ_α is positive definite because Q_α is positive definite and $1/\xi > 0$ for all $\xi \in \mathbb{R}^+$. Note that, $[\Gamma_\alpha]_{ii} = 1$ for all $i \in \{1, 2, \ldots, n\}$ because at the fixed point solution $\xi_i^2 = [Q_\alpha]_{ii}$ for all $i \in \{1, 2, \ldots, n\}$. Using the above expression and properties of Hadamard product, we rewrite (47) as

$$\begin{aligned}&2\,\mathrm{D}^{1/2}[\mathbf{X}\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}} \circ \mathbf{X}\left\{\Sigma_\alpha(\xi)/\alpha + \mu_\alpha(\xi)\mu_\alpha^{\mathrm{T}}(\xi)\right\}\mathbf{X}^{\mathrm{T}}]\mathrm{D}^{1/2}\\ &= 2\,\mathrm{D}^{1/2}\left\{\mathbf{X}\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}} \circ \Delta(\xi)\right\}\mathrm{D}^{1/2} \circ \Gamma_\alpha.\end{aligned} \tag{48}$$

Next we can write,

$$\mathrm{D}^{1/2}\left\{\mathbf{X}\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}} \circ \Delta(\xi)\right\}\mathrm{D}^{1/2} = \operatorname{diag}\Big[\{\xi A'(\xi)\}^{1/2}\Big]\mathbf{X}\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}}\operatorname{diag}\Big[\{\xi A'(\xi)\}^{1/2}\Big].$$

The above equality follows from the fact that the $(i,j)^{th}$ entry of the matrices on the both side of the equation is given by, $\{\xi_i A'(\xi_i)\}^{1/2}\left[\mathbf{X}\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}}\right]_{ij}\{\xi_j A'(\xi_j)\}^{1/2}$. Let us call $\mathrm{R}_\alpha = 2\operatorname{diag}[\{\xi A'(\xi)\}^{1/2}]\mathbf{X}\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}}\operatorname{diag}[\{\xi A'(\xi)\}^{1/2}]$. Then R_α has the same set of non-zero eigenvalues with $\mathrm{R}_\alpha^{(1)} = 2\,\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}}\operatorname{diag}\{\xi A'(\xi)\}\,\mathbf{X}$. Using Lemma 24 with $B = 2\,\mathbf{X}^{\mathrm{T}}\operatorname{diag}\{\xi A'(\xi)\}\,\mathbf{X}$ and $A = \Sigma_\beta^{-1}/\alpha - 2\mathbf{X}^{\mathrm{T}}\operatorname{diag}\{A(\xi)\}\mathbf{X} = \Sigma_\alpha^{-1}(\xi)$, along with (46) we have, $\lambda_1(\mathrm{R}_\alpha^{(1)}) < 1$. Hence we can write, $\lambda_1(2\,\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}}\mathrm{diag}\,\{\xi A'\,(\xi)\}\,\mathbf{X}) < 1$ which implies $\lambda_1(\mathrm{R}_\alpha) < 1$. Also, we can rewrite (48)

$$2\,\mathrm{D}^{1/2}\,\{\mathbf{X}\Sigma_\alpha(\xi)\mathbf{X}^{\mathrm{T}} \circ \Delta(\xi)\}\,\mathrm{D}^{1/2} \circ \Gamma_\alpha = \mathrm{R}_\alpha \circ \Gamma_\alpha. \tag{49}$$

Finally we use Lemma 23 on (49) with $A = \Gamma_\alpha$ and $B = \mathrm{R}_\alpha$ for $k = 1$. This concludes the proof.

C.2 Calculation of spectral radius for $p = 1$

For a fixed $\alpha \in (0, 1]$, One can write the updtaes from (8) for $p = 1$

$$\left(\xi_i^{t+1}\right)^2 = x_i^2[\sigma_\alpha(\xi^t)/\alpha + \{c\,\sigma_\alpha(\xi^t)\}^2] \quad (i = 1, 2, \ldots, n), \tag{50}$$

where $\sigma_\alpha(\xi^t) = \{\sigma_\beta^{-2}/\alpha - 2\sum_{i=1}^n x_i^2 A(\xi_i^t)\}^{-1}$ and $c = x'(y - 1/2\mathbb{1}_n) + \mu_\beta/\{\alpha\,\sigma_\beta^2\}$. We calculate $\partial(\xi_i^{t+1})^2/\partial(\xi_j^t)^2$ to get the Jacobian Matrix.

$$\frac{\partial\left(\xi_i^{t+1}\right)^2}{\partial(\xi_j^t)^2} = 2x_i^2x_j^2\{A'(\xi_j^t)/2\xi_j^t\}\,\{\sigma_\alpha^2(\xi^t)/\alpha + 2c^2\sigma_\alpha^3(\xi^t)\}.$$

Denote $\eta_\alpha(\xi^t) = \sigma_\alpha^2(\xi^t)/\alpha + 2c^2\sigma_\alpha^3(\xi^t)$ and $a_{ij} = 2x_i^2x_j^2A'(\xi_j^t)/2\xi_j^t$. Then the $(i, j)^{th}$ entry of the Jacobian Matrix $\mathbf{J}_\alpha^t$ at $\xi = \xi^t$ is given by $[\mathbf{J}_\alpha^t]_{ij} = \eta^t a_{ij}$. Since $[\mathbf{J}_\alpha^t]_{*j} = [\mathbf{J}_\alpha^t]_{*1} \times x_j^2/x_1^2 \times A'(\xi_j^t)/A'(\xi_1^t) \times \xi_1^t/\xi_j^t$ it follows Rank($\mathbf{J}_\alpha$)=1. Here, $[\mathbf{J}_\alpha^t]_{*j}$ is the j^{th} column of $\mathbf{J}_\alpha^t$. Order the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Then assuming $\mathrm{tr}(\mathbf{J}_\alpha^t) \neq 0$ and using Lemma 25 we obtain that the non-zero eigenvalue of $\mathbf{J}_\alpha^t$ is given by,

$$\lambda_1 = \sum_i^n \eta_\alpha(\xi^t)a_{ii} = \sum_i^n x_i^4A'(\xi_i^t)/\xi_i^t\{\sigma_\alpha^2(\xi^t)/\alpha + 2c^2\sigma_\alpha^3(\xi^t)\}. \tag{51}$$

Further Simplification at $\xi^t = \xi^*$. In case of $p = 1$, the self-consistency or the fixed point equation (9) for (50) is given by

$$(\xi_i^*)^2 = x_i^2[\sigma_\alpha(\xi^*)/\alpha + \{c\,\sigma_\alpha(\xi^*)\}^2]. \tag{52}$$

From (51), we can calculate λ_1 at $\xi = \xi^*$,

$$\lambda_1 = \sum_{i=1}^n x_i^4\{A'(\xi_i^*)/\xi_i^*\}\{\sigma_\alpha^2(\xi^*)/\alpha + 2c^2\sigma_\alpha^3(\xi^*)\}.$$

Substituting (52) into the (53) gives us,

$$\begin{aligned}\lambda_1 &= \sum_{i=1}^n x_i^4\{A'(\xi_i^*)/\xi_i^*\}\sigma_\alpha(\xi^*)[2(\xi_i^*)^2/x_i^2 - \sigma_\alpha(\xi^*)/\alpha], \\ &\leq \frac{\sum_{i=1}^n 2x_i^2A'(\xi_i^*)\xi_i^*}{\sigma_\beta^{-2}/\alpha - \sum_{i=1}^n 2x_i^2A(\xi_i^*)}.\end{aligned} \tag{53}$$

The above inequality follows from the fact that $\sum_{i=1}^{n}\{x_i^4 A'(\xi_i^*)/\xi_i^*\}\{\sigma_\alpha^2(\xi^t)/\alpha\} > 0$ since $A(\xi_i^*)/\xi_i^* > 0$ for all $\xi_i^* \in \mathbb{R}^+$. From (56) as $\sigma_\beta > 0$, we obtain

$$2\sum_{i}^{n}\left\{A'(\xi^*)\xi^* + A(\xi^*)\right\} x_i^2 < 0 < \sigma_\beta^{-2}/\alpha.$$

By rearranging the terms it follows that $\lambda_1 < 1$.

C.3 Auxiliary Results for the Proofs in Section 4

Lemma 22 *For a symmetric matrix $M_{n\times n}$ and a invertible diagonal matrix $N_{n\times n}$, the set of eigenvalues of MN and $N^{1/2}MN^{1/2}$ are the same.*

Proof The characteristic equation for MN is given by, $|\mathrm{MN}-\lambda\mathbb{I}| = 0$. Since, N is invertible we can write, $|\mathrm{N}^{1/2}||\mathrm{MN}-\lambda\mathbb{I}||\mathrm{N}^{-1/2}| = 0$ which implies $|\mathrm{N}^{1/2}\mathrm{MN}^{1/2}-\lambda\mathbb{I}| = 0$. Hence the proof. ■

Lemma 23 *Let A, B be $n \times n$ given positive semidefinite Hermitian matrices. Arrange the eigenvalues of $A \circ B$ and B and the main diagonal entries $d_i(A)$ of A in decreasing order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and $d_1(A) \geq d_2(A) \geq \ldots \geq d_n(A)$. Then,*

$$\sum_{i=1}^{k}\lambda_i(A\circ B) \leq \sum_{i=1}^{k} d_i(A)\lambda_i(B), \quad k = \{1, 2, \ldots n\},$$

Proof See Theorem 5.5.12 in Horn and Johnson (1994). ■

Lemma 24 *For two $n\times n$ symmetric matrices A and B, such that $B \prec A$ where A is positive definite and B is positive semi-definite. Then the largest eigenvalue of $A^{-1/2}\,B\,A^{-1/2}$ given by $\lambda_1(A^{-1/2}\,B\,A^{-1/2})$ is less than 1.*

Proof Since $\mathrm{A}-\mathrm{B}$ is positive definite and A is invertible, it is easy to see that $\mathrm{A}^{-1/2}(\mathrm{A}-\mathrm{B})\mathrm{A}^{-1/2}$ is also positive definite. Then the smallest eigen value of $I_n - \mathrm{A}^{-1/2}\,\mathrm{B}\,\mathrm{A}^{-1/2}$ is bigger than 0. This implies $\lambda_1(\mathrm{A}^{-1/2}\,\mathrm{B}\,\mathrm{A}^{-1/2}) < 1$. ■

Lemma 25 *For an $n \times n$ matrix with rank 1, the number of non-zero eigenvalues is at most 1. If trace of the matrix (denoted tr(A)) is non-zero then a non-zero eigenvalue exists and equal to trace of the matrix.*

Proof Suppose an $n\times n$ matrix A has two non-zero eigenvalue λ_1, λ_2 with non-zero linearly independent eigenvectors v_1, v_2. Then $Av_1 = \lambda_1 v_1$ and $Av_2 = \lambda_2 v_2$. This contradicts the fact $\mathrm{rank}(A) = 1$. Now assuming that $\mathrm{tr}(A) \neq 0$, and using the fact $\mathrm{tr}(A) = \sum_{i=1}^{n}\lambda_i$, we claim that a non-zero eigenvalue exists and $\lambda_1 = \mathrm{tr}(A)$. ■

C.4 Global Convergence Rate in a Semi-Orthogonal Case: Proof of Theorem 9

Recall that throughout this proof we are going to assume $\sigma_\beta = 1$. We consider two separate cases, given by $n = 1$ and $n \geq 2$.
Case $n = 1$: For notational convenience, let us call $h'_{1,n}(z) = h'_n(z)$. For $n = 1$ we have,

$$h'_1(z) = \frac{A'(\sqrt{z})}{\sqrt{z}}(1 - 2A(\sqrt{z}))^{-2}\left[1 + \frac{1}{2}(1 - 2A(\sqrt{z}))^{-1}\right].$$

From Proposition 26 we have $1 \leq [1 - 2A(\sqrt{z})] \leq 5/4$ which $h'_1(z) \leq 3A'(\sqrt{z})/2\sqrt{z}$. In the following, we derive an upper bound for $A'(x)/x$ for $x \in \mathbb{R}^+$.

$$\begin{aligned}\frac{A'(x)}{x} &= \frac{(e^x - x)^2 - (1 + x^2)}{4x^3(1 + e^x)^2},\\ &= \frac{2\sum_{n=3}^{\infty} x^{n-3}/n! + \sum_{n=4}^{\infty} x^{n-3}[1/\{2!(n-2)!\} + 1/\{3!(n-3)!\} + \cdots + 1/\{(n-2)!2!\}]}{4(1 + e^x)^2},\\ &\leq \frac{2e^x}{4(1 + e^x)^2} + \frac{8e^{2x}}{4(1 + e^x)^2}. \qquad (54)\end{aligned}$$

The inequality in (54) is due to $\sum_{n=3}^{\infty} x^{n-3}/n! < e^x$ and $x^{n-3}[1/\{2!(n-2)!\} + 1/\{3!(n-3)!\} + \cdots + 1/\{(n-2)!2!\}]$ and $\sum_{n=1}^{\infty} x^{n-3}\, 2^n/n! \leq 2^3 \exp(2x)$. Using $\exp(x) + \exp(-x) \geq 2$ for all $x \in \mathbb{R}$, we further obtain,

$$\frac{2e^x}{4(1 + e^x)^2} + \frac{8e^{2x}}{4(1 + e^x)^2} \leq \frac{1}{2(e^{-x} + e^x + 2)} + \frac{8}{4(e^{-x} + e^x)^2} \leq \frac{1}{8} + \frac{8}{16} = 5/8.$$

Hence $\|h'_1\|_\infty \leq 15/16$.
Case $n \geq 2$: We begin with the function $h'_n(z)$ which is given by,

$$h'_n(z) = \frac{A'(\sqrt{z})}{\sqrt{z}}\sigma_n^{-2}\left[\frac{1}{n} + \frac{1}{2\,\sigma_n}\right], \qquad (55)$$

where, $\sigma_n = \{1/n - 2\,A(\sqrt{z})\}$. In Lemma 27 we show that for any $z \in \mathbb{R}^+$, $h'_n(z)$ is a monotonically increasing function of n, provided $n \geq 2$. And also h'_n converges pointwise to $h'(z) := -A'(\sqrt{z})/\{16\sqrt{z}A^3(\sqrt{z})\}$ and $h'(z) < 1$ for $z \in \mathbb{R}^+$. So for any fixed $z \in \mathbb{R}^+$ and $n \in \{2, 3, 4, \ldots\}$ we have $h'_n(z) \leq h'(z) < 1$. Hence $\|h'_n\|_\infty < 1$ for any fixed n.

C.5 Auxiliary Results for the Global Convergence Rate Result

The function $A(\cdot)$ plays a crucial role in studying the convergence of the EM. The following proposition provides some properties of $A(\xi)$.

Proposition 26 *The following are true for the function $A(\xi) := -\tanh(\xi/2)/4\xi$, defined on $\mathbb{R}^+$. $A : \mathbb{R}^+ \to \mathbb{R}^-$ is monotonically increasing and twice continuously differentiable with $A(0) = -1/8$.*

Proof It is easy to see that the range of $A(\cdot) \subseteq \mathbb{R}^-$ since $\tanh(\xi/2) > 0$ for all $\xi \in \mathbb{R}^+$. $A(0) = -1/8$ follows from the fact that, $\lim_{\xi \to 0}\{\exp(\xi) - 1\}/\xi = 1$ and $A(\xi) =$

$-\{\exp(\xi)-1\}/4\xi\{\exp(\xi)+1\}$. Differentiating $\xi\, A(\xi)$ gives the following for all $\xi \in \mathbb{R}^+$,

$$A(\xi) + \xi A'(\xi) = -\frac{1}{2}\frac{e^{\xi}}{(e^{\xi}+1)^2} < 0. \tag{56}$$

It follows immediately that,

$$A'(\xi) = \frac{(e^{\xi}-\xi)^2 - (1+\xi^2)}{4\xi^2(1+e^{\xi})^2}. \tag{57}$$

Since $(e^{\xi}-\xi)^2 - (1+\xi^2) > 0$ for all $\xi \in \mathbb{R}^+$, $A'(\xi) > 0$ for all $\xi \in \mathbb{R}^+$. Also, $A''(\xi) = -2A'(\xi)/\xi + 4A(\xi)\big[A(\xi)+\xi A'(\xi)\big]$ is a continuous function, thus completing the claim. ■

Lemma 27 *For any $n \geq 2$, the following claims are true for the function $h'_n : \mathbb{R}^+ \to \mathbb{R}^+$,*
(a) For any fixed $z \in \mathbb{R}^+$, $h'_n(z)$ is an increasing function of n.
(b) For any fixed $z \in \mathbb{R}^+$, define $h'(z) = -A'(\sqrt{z})/\{16\sqrt{z}A^3(\sqrt{z})\}$. Then, $h'_n(z)$ converges pointwise to $h'(z)$. Also, $h'(z) < 1$ for all $z \in \mathbb{R}^+$.

Proof *Part (a):* From Proposition 26 it is clear that $h'_n(z) > 0$ for all $z \in \mathbb{R}^+$. For any fixed $z \in \mathbb{R}^+$, it is easy to see σ_n^{-2} increases with n. Next we show that for $n \geq 2$,

$$\frac{1}{n} + \frac{1}{2\,\sigma_n} < \frac{1}{n+1} + \frac{1}{2\,\sigma_{n+1}}. \tag{58}$$

We begin with the fact that for $n \geq 2$, $\{(n+1)^{-1}+1/4\}(1/n+1/4) < 1/2$. From Proposition 26 we know that $-1/8 \leq A(\sqrt{z}) < 0$ for all $z \in \mathbb{R}^+$. Then for any fixed z on $\mathbb{R}^+$ we have $\sigma_{n+1}\,\sigma_n < 1/2$ which when multiplied on the both sides by $1/n - 1/(n+1)$ yields (58). This proves the first part of Lemma 27.
Part (b): For a fixed $z \in \mathbb{R}^+$, it is easy to see that, $\sigma_n \to -2A(\sqrt{z})$ as $n \to \infty$. This leads to

$$\lim_{n\to\infty}\left[\frac{1}{n} + \frac{1}{2\,\sigma_n}\right] = -\frac{1}{4A(\sqrt{z})}. \tag{59}$$

Multiplying (59) with $A'(\sqrt{z})/\sqrt{z}$ for a fixed $z \in \mathbb{R}^+$, we get $\lim_{n\to\infty} h'_n(z) = h'(z)$. Next we show that for $h'(z) < 1$ for any $z \in \mathbb{R}^+$. From Proposition 26,

$$\frac{-\,A'(\sqrt{z})}{16\sqrt{z}\,A^3(\sqrt{z})} = \frac{\big(e^{2\sqrt{z}} - 2\sqrt{z}e^{\sqrt{z}} - 1\big)\big(1+e^{\sqrt{z}}\big)}{\big(e^{\sqrt{z}}-1\big)^3} > 0. \tag{60}$$

Next, write $\psi(x) = 2(e^x-1) - x(e^x+1)$. Then $\psi'(x) = e^x - xe^x - 1$ and $\psi''(x) = -xe^x$. Hence $\psi(0) = 0$, $\psi'(0) = 0$ and ψ' is decreasing, which entails ψ is decreasing for $x > 0$ and $\psi(x) < 0$ for $x > 0$. Hence $2(e^{\sqrt{z}}-1) - \sqrt{z}(e^{\sqrt{z}}+1) < 0$ for $z \in \mathbb{R}^+$ and the numerator of the right hand side of (60) is

$$\big(e^{2\sqrt{z}} - 2\sqrt{z}e^{\sqrt{z}} - 1\big)\big(1+e^{\sqrt{z}}\big) - \big(e^{\sqrt{z}}-1\big)^3 = 2e^{\sqrt{z}}\big\{2(e^{\sqrt{z}}-1) - \sqrt{z}(e^{\sqrt{z}}+1)\big\} < 0.$$

This proves the second part of Lemma 27. ■

References

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *arXiv:2110.11216*, 2021.

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *Annals of Statistics*, 48(3):1475–1497, 2020.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1): 8374–8414, 2016.

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

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45 (1):77–120, 2017.

Maria Barbarossa. Stability of discrete dynamical systems. *Matrix*, 21:a22, 2011.

Anirban Bhattacharya and Debdeep Pati. Nonasymptotic Laplace approximation under model misspecification. *arXiv preprint arXiv:2005.07844*, 2020.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian fractional posteriors. *The Annals of Statistics*, 47(1):39–66, 2019.

Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.

David M Blei and John D Lafferty. A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

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.

Nicoletta Bof, Ruggero Carli, and Luca Schenato. Lyapunov Theory for Discrete Time Systems. *arXiv preprint arXiv:1809.05289*, 2018.

Guillaume Bouchard. Efficient Bounds for the Softmax Function and Applications to Approximate Inference in Hybrid models. In *Proceedings of the Presentation at the Workshop For Approximate Bayesian Inference in Continuous/Hybrid Systems at Neural Information Processing Systems (NIPS), Meylan, France*, volume 31, 2008.

Michael Braun and Jon McAuliffe. Variational Inference for Large-Scale Models of Discrete Choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

Trevor Campbell and Xinglong Li. Universal Boosting Variational Inference. In *Advances in Neural Information Processing Systems*, pages 3484–3495, 2019.

Olivier Catoni. A PAC-Bayesian approach to adaptive classification. *preprint*, 840, 2003.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12 (2):2995–3035, 2018.

Daniele Durante and Tommaso Rigon. Conditionally Conjugate Mean-Field Variational Bayes for Logistic Models. *Statistical Science*, 34(3):472–485, 2019.

Mohammad Emtiyaz Khan, Aleksandr Y Aravkin, Michael P Friedlander, and Matthias Seeger. Fast Dual Variational Inference for Non-Conjugate LGMs. *arXiv*, pages arXiv–1306, 2013.

Behrooz Ghorbani, Hamid Javadi, and Andrea Montanari. An Instability in Variational Inference for Topic Models. In *International Conference on Machine Learning*, 2018.

Alex Graves. Practical Variational Inference for Neural Networks. In *Advances in neural information processing systems*, pages 2348–2356, 2011.

Shunsuke Hirose, Tomotake Kozu, Yingzi Jin, and Yuichi Miyamura. Hierarchical Relevance Determination based on Information Criterion Minimization. *SN Computer Science*, 1 (4):1–19, 2020.

Roger A Horn and Charles R Johnson. *Topics in matrix analysis*. Cambridge university press, 1994.

David R Hunter and Kenneth Lange. A Tutorial on MM Algorithms. *The American Statistician*, 58(1):30–37, 2004.

Tommi S Jaakkola and Michael I Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

Tommi Sakari Jaakkola. *Variational methods for inference and estimation in graphical models*. PhD thesis, Massachusetts Institute of Technology, 1997.

Tony Jebara and Anna Choromanska. Majorization for CRFs and Latent Likelihoods. In *Advances in Neural Information Processing Systems*, pages 557–565, 2012.

James E. Johndrow, Aaron Smith, Natesh Pillai, and David B. Dunson. MCMC for Imbalanced Categorical Data. *Journal of the American Statistical Association*, 114(527):1394–1403, 2019. doi: 10.1080/01621459.2018.1505626. URL `https://doi.org/10.1080/01621459.2018.1505626`.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An Introduction to Variational Methods for Graphical Models. *Machine learning*, 37(2): 183–233, 1999.

Bas JK Kleijn and Aad W van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34(2):837–877, 2006.

Kenta Konagayoshi and Kazuho Watanabe. Minimax Online Prediction of Varying Bernoulli Process under Variational Approximation. In *Asian Conference on Machine Learning*, pages 141–156, 2019.

Neil D Lawrence, Marta Milo, Mahesan Niranjan, Penny Rashbass, and Stephan Soullier. Reducing the variability in cDNA microarray image processing by Bayesian inference. *Bioinformatics*, 20(4):518–526, 2004.

Francesco Locatello, Rajiv Khanna, Joydeep Ghosh, and Gunnar Ratsch. Boosting Variational Inference: an Optimization Perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 464–472, 2018.

David JC MacKay. Ensemble Learning for Hidden Markov Models. Technical report, Citeseer, 1997.

Andreas Maurer. A note on the PAC Bayesian theorem. *arXiv preprint cs/0411099*, 2004.

David A McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

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.

Hannes Nickisch and Matthias W Seeger. Convex Variational Bayesian Inference for Large Scale Generalized Linear Models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 761–768, 2009.

G. Parisi. *Statistical Field Theory*. Frontiers in Physics. Addison-Wesley, 1988.

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.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

Juergen Quandt. On the Hartman-Grobman Theorem for Maps. *Journal of differential equations*, 64(2):154–164, 1986.

R Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1970.

Orlando Romero, Sarthak Chatterjee, and Sérgio Pequito. Convergence of the Expectation-Maximization Algorithm Through Discrete-Time Lyapunov Stability Theory. In *2019 American Control Conference (ACC)*, pages 163–168. IEEE, 2019.

Orlando Romero, Subhro Das, Pin-Yu Chen, and Sérgio Pequito. A Dynamical Systems Approach for Convergence of the Bayesian EM Algorithm. *arXiv preprint arXiv:2006.12690*, 2020.

Matthias Seeger. PAC-Bayesian Generalisation Error Bounds for Gaussian Process Classification. *Journal of machine learning research*, 3(Oct):233–269, 2002.

Weishi Shi and Qi Yu. Integrating Bayesian and Discriminative Sparse Kernel Machines for Multi-class Active Learning. In *Advances in Neural Information Processing Systems*, pages 2285–2294, 2019.

Nathan Srebro and Tommi Jaakkola. Weighted Low-Rank Approximations. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 720–727, 2003.

Nicholas Syring and Ryan Martin. Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486, 2019.

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

Martin J Wainwright and Michael I Jordan. Variational inference in graphical models: The view from the marginal polytope. In *proceedings of the annual Allerton conference on communication control and computing*, volume 41, pages 961–971. The University; 1998, 2003.

Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. A New Class of Upper Bounds on the Log Partition Function. *IEEE Transactions on Information Theory*, 51 (7):2313–2335, 2005.

Martin J Wainwright, Michael I Jordan, et al. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Stephen Walker and Nils Lid Hjort. On Bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821, 2001.

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

Yixin Wang and David Blei. Variational Bayes under Model Misspecification. In *Advances in Neural Information Processing Systems*, pages 13357–13367, 2019a.

Yixin Wang and David M Blei. Frequentist Consistency of Variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019b.

Stephen Wiggins. *Introduction to applied nonlinear dynamical systems and chaos*, volume 2. Springer Science & Business Media, 2003.

RJ Wood and MJ O'Neill. An always convergent method for finding the spectral radius of an irreducible non-negative matrix. *ANZIAM Journal*, 45:474–485, 2003.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-Variational Inference with Statistical Guarantees. *Annals of Statistics*, 48(2):886–905, 2020.

Mingzhang Yin, YX Rachel Wang, and Purnamrita Sarkar. A Theoretical Case Study of Structured Variational Inference for Community Detection. In *International Conference on Artificial Intelligence and Statistics*, pages 3750–3761, 2020.

Anderson Y Zhang and Harrison H Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. *Annals of Statistics*, 48(5): 2575–2598, 2020.

Fengshuo Zhang and Chao Gao. Convergence Rates of Variational Posterior Distributions. *Annals of Statistics*, 48(4):2180–2207, 2020.