entropy

MDPI

Article

Dynamics of Coordinate Ascent Variational Inference: A Case Study in 2D Ising Models

Sean Plummer *, Debdeep Pati and Anirban Bhattacharya

Department of Statistics, Texas A&M University, College Station, TX 77843, USA; debdeep@stat.tamu.edu (D.P.); anirbanb@stat.tamu.edu (A.B.)

* Correspondence: snplmmr@stat.tamu.edu

Received: 3 September 2020; Accepted: 3 November 2020; Published: 6 November 2020

Abstract: Variational algorithms have gained prominence over the past two decades as a scalable computational environment for Bayesian inference. In this article, we explore tools from the dynamical systems literature to study the convergence of coordinate ascent algorithms for mean field variational inference. Focusing on the Ising model defined on two nodes, we fully characterize the dynamics of the sequential coordinate ascent algorithm and its parallel version. We observe that in the regime where the objective function is convex, both the algorithms are stable and exhibit convergence to the unique fixed point. Our analyses reveal interesting discordances between these two versions of the algorithm in the region when the objective function is non-convex. In fact, the parallel version exhibits a periodic oscillatory behavior which is absent in the sequential version. Drawing intuition from the Markov chain Monte Carlo literature, we empirically show that a parameter expansion of the Ising model, popularly called the Edward–Sokal coupling, leads to an enlargement of the regime of convergence to the global optima.

Keywords: bifurcation; dynamical systems; Edward–Sokal coupling; mean-field; Kullback–Leibler divergence; variational inference

1. Introduction

Variational Bayes (VB) is now a standard tool to approximate computationally intractable posterior densities. Traditionally this computational intractability has been circumvented using sampling techniques such as Markov chain Monte Carlo (MCMC). MCMC techniques are prone to be computationally expensive for high dimensional and complex hierarchical Bayesian models, which are prolific in modern applications. VB methods, on the other hand, typically provide answers orders of magnitude faster, as they are based on optimization. Introduction to VB can be found in chapter 10 of [1] and chapter 33 of [2]. Excellent recent surveys can be found in [3,4].

The objective of VB is to find the best approximation to the posterior distribution from a more tractable class of distributions on the latent variables that is well-suited to the problem at hand. The best approximation is found by minimizing a divergence between the posterior distribution of interest and a class of distributions that are computationally tractable. The most popular choices for the discrepancy and the approximating class are the Kullback–Leibler (KL) divergence and the class of product distributions, respectively. This combination is popularly known as mean field variational inference, originating from mean field theory in physics [5]. Mean-field inference has percolated through a wide variety of disciplines, including statistical mechanics, electrical engineering, information theory, neuroscience, cognitive sciences [6] and more recently deep neural networks [7]. While computing the KL divergence is intractable for a large class of distributions, reframing the minimization problem for maximizing the evidence lower bound (ELBO) leads to efficient algorithms. In particular, for conditionally conjugate-exponential family models, the optimal distribution for mean

field variational inference can be computed by iteration of closed form updates. These updates form a coordinate ascent algorithm known as coordinate ascent variational inference (CAVI) [1].

Research into the theoretical properties of variational Bayes has exploded in the last few years. Recent theoretical work focuses on statistical risk bounds for variational estimate obtained from VB [8–11], asymptotic normality of VB posteriors [12] and extension to model misspecification [8,13]. While much of the recent theoretical work focuses on statistical optimality guarantees, there has been less work studying the convergence of the CAVI algorithms employed in practice. Convergence of CAVI to the global optima is only known in special cases that depend heavily on model structure for normal mixture models [14,15]; stochastic block models [16–19]; topic models [20]; and under special restrictions of the parameter regime, Ising models [21,22]. The convergence properties of the CAVI algorithm still largely constitute an open problem.

The goal of this work is to suggest a general systematic framework for studying convergence properties of CAVI algorithms. By viewing CAVI as a discrete time dynamical system, we can leverage dynamical systems theory to analyze the convergence behavior of the algorithm and bifurcation theory to study the types of changes that solutions can undergo as the various parameters are varied. For sake of concreteness, we focus on the 2D Ising model. While dynamical systems theory possesses the tools [23–25] necessary to analyze higher dimensional systems, they were mainly developed for non-sequential systems. The general theory for n-dimensional discrete dynamical systems is dependent on having the evolution function in the form $x_{n+1} = F(x_n)$. Deriving this F is typically not possible for densely connected higher dimensional sequential systems. The 2D Ising model has the special property that both the sequential and parallel updates in the two variables case can be written as two separate one variable dynamical systems, allowing for a simplified analysis. Our contributions to the literature are as follows: We provide a complete classification of the dynamical properties of the the traditional sequential update CAVI algorithm, and a parallelized version of the algorithm using dynamical systems and bifurcation theory on the Ising models. Our findings show that the sequential CAVI algorithm and the parallelized version have different convergence properties. Additionally, we numerically investigated the convergence of the CAVI algorithm on the Edward–Sokal coupling, a generalization of the Ising model. Our findings suggest that couplings/parameter expansion may provide a powerful way of controlling the convergence behavior of the CAVI algorithm, beyond the immediate example considered here.

2. Mean-Field Variational Inference and the Coordinate Ascent Algorithm

In this section, we briefly introduce mean-field variational inference for a target distribution in the form of a Boltzmann distribution with potential function Ψ,

$$p(\mathbf{x}) = \frac{\exp\{\Psi(\mathbf{x})\}}{\mathcal{Z}}, \quad \mathbf{x} \in \mathcal{X},$$

where $\mathcal{Z}$ denotes the intractable normalizing constant. The above representation encapsulates both posterior distributions that arise in Bayesian inference, where Ψ is the log-posterior up to constants, and probabilistic graphical models such as the Ising and Potts models. For instance, $\Psi(x) = \beta \sum_{u \sim v} J_{uv} x_u x_v + \beta \sum_u h_u x_u$ for the Ising model; see the next section for more details. Many of the complications in inference arise from the intractability of the normalizing constant $\mathcal{Z}$, which is commonly referred to as the free energy in probabilistic graphical models, and the marginal likelihood or evidence in Bayesian statistics. Variational inference aims to mitigate this problem by using optimization to find the best approximation q^* to the target density p from a class $\mathcal{F}$ of variational distributions over the parameter vector $\mathbf{x}$,

$$q^* = \arg\min_{q \in \mathcal{F}} D(q \,||\, p) \tag{1}$$

where $D(q \,||\, p)$ denotes the Kullback–Leibler (KL) divergence between q and p. The complexity of this optimization problem is largely determined by the choice of variational family $\mathcal{F}$. The objective function of the above optimization problem is intractable because it also involves the evidence $\mathcal{Z}$. We can work around this issue by rewriting the KL divergence as

$$D(q \,||\, p) = \mathbb{E}_q[\log q] - \mathbb{E}_q[\Psi] + \log \mathcal{Z} \tag{2}$$

where $\mathbb{E}_q$ denotes the expectation with respect to $q(\mathbf{x})$. Rearranging terms,

$$\begin{aligned} \log \mathcal{Z} \quad &= D(q \,||\, p) + \mathbb{E}_q[\Psi] - \mathbb{E}_q[\log q] && (3) \\ &\geq \mathbb{E}_q[\Psi] - \mathbb{E}_q[\log q] := \mathrm{ELBO}(q). && (4) \end{aligned}$$

The acronym ELBO stands for evidence lower bound and the nomenclature is now apparent from the above inequality. Notice from Equation (2) that maximizing the ELBO is equivalent to minimizing the KL divergence. By maximizing the ELBO we can solve the original variational problem while by-passing the computational intractability of the evidence.

As mentioned above, the choice of variational family controls both the complexity and accuracy of approximation. Using a more flexible family achieves a tighter lower bound but at the cost of having to solve a more complex optimization problem. A popular choice of family that balances both flexibility and computability is the mean-field family. Mean-field variational inference refers to the situation when q is restricted to the product family of densities over the parameters,

$$\mathcal{F}_{\mathrm{MF}} := \{q(\mathbf{x}) = q_1(x_1) \otimes \cdots \otimes q_n(x_n) \text{ for probability measures } q_j, j = 1, \ldots, n\}, \tag{5}$$

The coordinate ascent variational inference (CAVI) algorithm (refer to Algorithm 1) is a learning algorithm that optimizes the ELBO over the mean-field family $\mathcal{F}_{\mathrm{MF}}$. At each time step $t \geq 1$, the CAVI algorithm iteratively updates the current mean field marginal distribution $q_j^{(t)}(x_j)$ by maximizing the ELBO over that marginal while keeping the other marginals $\{q_\ell^{(t)}(x_\ell)\}_{\ell \neq j}$ fixed at their current values. Formally, we update the current distribution $q^{(t)}(\mathbf{x})$ to $q^{(t+1)}(\mathbf{x})$ by the updates,

$$\begin{aligned} q_1^{(t+1)}(x_1) &= \arg\max_{q_1} \mathrm{ELBO}(q_1 \otimes q_2^{(t)} \otimes \cdots \otimes q_n^{(t)}) \\ q_2^{(t+1)}(x_2) &= \arg\max_{q_2} \mathrm{ELBO}(q_1^{(t+1)} \otimes q_2 \otimes q_3^{(t)} \otimes \cdots \otimes q_n^{(t)}) \\ &\vdots \\ q_n^{(t+1)}(x_n) &= \arg\max_{q_n} \mathrm{ELBO}(q_1^{(t+1)} \otimes \cdots \otimes q_{n-1}^{(t+1)} \otimes q_n). \end{aligned}$$

Algorithm 1 Coordinate ascent variational inference (CAVI).

Input: Model $p(\mathbf{x}) = \exp(\Psi(\mathbf{x}) - \log \mathcal{Z})$
Output: A variational density $q(\mathbf{x}) = \prod_{j=1}^n q_j(x_j)$
Initialize: variational densities $q_j(x_j)$
while *ELBO*(q) *not converged* **do**
  for $j \in \{1, \ldots, n\}$ **do**
    $q_j(x_j) \propto \exp\{\mathbb{E}_{-j}[\Psi(\mathbf{x})]\}$
  end
  Compute $\mathrm{ELBO}(q) = \mathbb{E}_q[\Psi(\mathbf{x})] - \mathbb{E}_q[\log q(\mathbf{x})]$
end
return $q(\boldsymbol{x})$

The objective function $\text{ELBO}(q_1 \otimes \cdots \otimes q_n)$ is concave in each of the arguments individually (although it is rarely jointly concave), so these individual maximization problems have unique solutions. The optimal update for the jth mean field variational component of the model has the closed form,

$$q_j^*(\mathbf{x}_j) \quad \propto \quad \exp\left\{\mathbb{E}_{-j}\left[\Psi(\mathbf{x})\right]\right\}$$

where the expectations $\mathbb{E}_{-j}$ are taken with respect to the distribution $\prod_{i \neq j} q_i(\mathbf{x}_i)$. Furthermore, the update step of the algorithm is monotonous, as each step of the CAVI increases the objective function

$$\begin{aligned}\text{ELBO}(q_1^{(t+1)} \otimes q_2^{(t+1)} \otimes \cdots \otimes q_n^{(t+1)}) \quad &\geq \text{ELBO}(q_1^{(t+1)} \otimes q_2^{(t+1)} \otimes \cdots \otimes q_{n-1}^{(t+1)} \otimes q_n^{(t)}) \geq \cdots \geq \\ &\text{ELBO}(q_1^{(t)} \otimes q_2^{(t)} \otimes \cdots \otimes q_n^{(t)}).\end{aligned}$$

For parametric models, the sequential updates of the variational marginal distributions in the CAVI algorithm is done by a sequential update of the variational parameters of these distributions. The CAVI algorithm updates for parametric models induce a discrete time dynamical system of the parameters. Clearly, convergence of the CAVI algorithm can be framed in terms of this induced discrete time dynamical system. As discussed before, the ELBO is generally a non-convex function, and hence the CAVI algorithm is only guaranteed to converge to a local optimum of the system. It is also not clear how many local optima (or fixed points) the system has, nor whether the algorithm always settles on a single fixed point, diverges away from the fixed point or cycles between multiple fixed points. These questions translate to questions about the existence and stability of fixed points of the induced dynamical system. We are also interested in how the behavior of the CAVI algorithm could possibly change as we vary the parameters of the model. This translates to questions about the possible bifurcations of the induced dynamical system. In Section 3, we formally introduce the Ising model and its mean-field variational inference.

3. CAVI in Ising Model

We first briefly review the definition of an Ising model. The Ising model was first introduced as a model for magnetization in statistical physics, but has found many applications in other fields; see [26] and references therein. The Ising model is a probability distribution on the hypercube $\{\pm 1\}^n$ given by

$$p(\mathbf{x}) \quad \propto \quad \exp\left[\beta \sum_{u \sim v} J_{uv} x_u x_v + \beta \sum_u h_u x_u\right], \tag{6}$$

where the interaction matrix J is a symmetric real $n \times n$ matrix with zeros on the diagonal, h is a real n-vector that represents the external magnetic field, and β is the inverse temperature parameter. The model is said to be ferromagnetic if $J_{uv} \geq 0$ for all u, v and anti-ferromagnetic if $J_{uv} < 0$ for all u, v. The normalizing constant or the partition function of the Ising model is

$$\mathcal{Z} \quad = \quad \sum_{\mathbf{x} \in \{\pm 1\}^n} \exp\left[\beta \sum_{u \sim v} J_{uv} x_u x_v + \beta \sum_u h_u x_u\right].$$

Refer to Chapter 31 of [2] for an excellent review of Ising models.

Mean Field Variational Inference in Ising Model

Here we provide a derivation of the CAVI update function for the Ising model, focusing on the two nodes ($n = 2$) case for simplicity and analytic tractability.

Notice $\log p(\mathbf{x}) := \beta\mathcal{H}(\mathbf{x}) = \beta\sum_{u\sim v} J_{uv}x_u x_v + \beta\sum_u h_u x_u$. In this case, we have the Ising model on two spins with $\mathbf{x} = (x_1, x_2)$ and influence matrix J with off diagonal term J_{12} and external magnetic field $h = (h_1, h_2) = (0, 0)$. From the general framework in Section 2, the CAVI updates are given by,

$$q_j^*(x_j) \propto \exp\left\{\mathbb{E}_{-j}\left[\beta(J_{12}x_1x_2 + h_1x_1 + h_2x_2)\right]\right\}.$$

Equivalently, the same updates are obtained by setting the gradient of the ELBO as a function of (x_1, x_2) equal to the $(0, 0)'$ vector. Illustrations of the ELBO and the gradient functions for various values of β are in Figures 1 and 2 respectively.

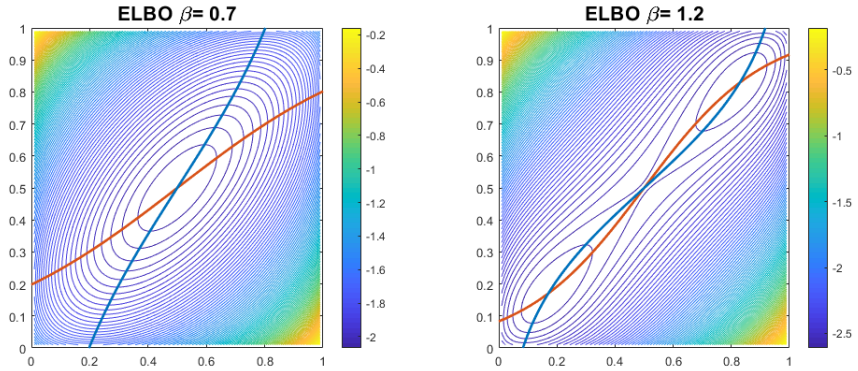


Figure 1. A contour plot of the ELBO as a function of x_1 and x_2 for $\beta = 0.7$ (**left**) and $\beta = 1.2$ (**right**) together with the optimal update functions for x_1 (**orange**) and x_2 (**blue**) given in Equation (8). For $\beta = 0.7$ the ELBO is a convex function and has exactly one optima, the global maximum, at $(0.5, 0.5)$. For $\beta = 1.2$ the ELBO is now a nonconvex function and has three optima at $(0.5, 0.5)$, $(0.17071, 0.17071)$ and $(0.82928, 0.82928)$.

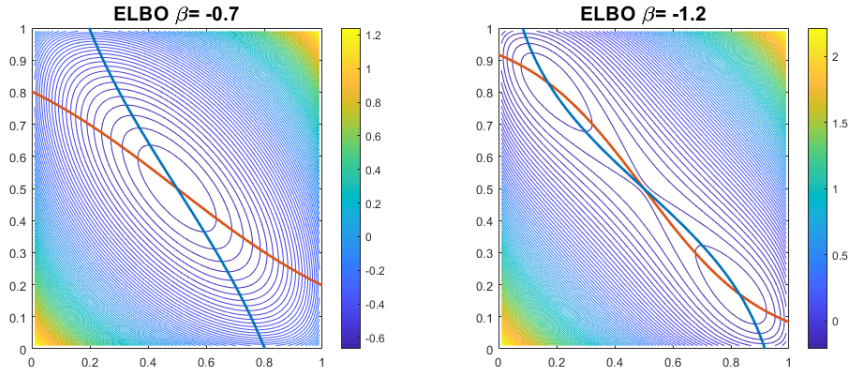


Figure 2. A contour plot of the ELBO as a function of x_1 and x_2 for $\beta = -0.7$ (**left**) and $\beta = -1.2$ (**right**) together with the optimal update functions for x_1 (**orange**) and x_2 (**blue**) given in Equation (8). For $\beta = -0.7$ the ELBO is a convex function and has exactly one optima, the global maximum, at $(0.5, 0.5)$. For $\beta = -1.2$ the ELBO is now a nonconvex function and has three optima at $(0.5, 0.5)$, $(0.17071, 0.82928)$ and $(0.82928, 0.17071)$.

Since q_1^* and q_2^* are two point distributions, it is sufficient to keep track of the mass assigned to 1. Simplifying,

$$\begin{aligned} q_1^*(x_1) &\propto \exp\left\{\mathbb{E}_2\left[\log p(x_1, x_2)\right]\right\} \\ &= \exp\left\{\beta\mathcal{H}(x_1, x_2 = 1)q_2(x_2 = 1) + \beta\mathcal{H}(x_1, x_2 = -1)q_2(x_2 = -1)\right\} \\ &= \exp\left\{(\beta J_{12}x_1 + \beta h_1 x_1 + \beta h_2)\xi + (-\beta J_{12}x_1 + \beta h_1 x_1 - \beta h_2)(1 - \xi)\right\} \\ &= \exp\left\{(2\xi - 1)(\beta J_{12}x_1 + \beta h_2) + \beta h_1 x_1\right\}, \end{aligned}$$

where $\xi = q_2(x_2 = 1)$. Therefore

$$\begin{aligned} q_1^*(x_1 = 1) &= \frac{\exp\{(2\xi - 1)(\beta J_{12} + \beta h_2) + \beta h_1\}}{\exp\{(2\xi - 1)(\beta J_{12} + \beta h_2) + \beta h_1\} + \exp\{(2\xi - 1)(-\beta J_{12} + \beta h_2) - \beta h_1\}} \\ &= \frac{1}{1 + \exp\{-2\beta J_{12}(2\xi - 1) - 2\beta h_1\}}. \end{aligned}$$

Similarly denoting $\zeta = q_1(x_1 = 1)$,

$$\begin{aligned} q_2^*(x_2 = 1) &= \frac{\exp\{(2\zeta - 1)(\beta J_{12} + \beta h_1) + \beta h_2\}}{\exp\{(2\zeta - 1)(\beta J_{12} + \beta h_1) + \beta h_2\} + \exp\{(2\zeta - 1)(-\beta J_{12} + \beta h_1) - \beta h_2\}} \\ &= \frac{1}{1 + \exp\{-2\beta J_{12}(2\zeta - 1) - 2\beta h_2\}}. \end{aligned}$$

Let ζ_k (resp. ξ_k) denote the kth iterate of $q_1(x_1 = 1)$ (resp. $q_2(x_2 = 1)$) from the CAVI algorithm. To succinctly represent these updates, define the logistic sigmoid function

$$\sigma(u, \beta) = \frac{1}{1 + e^{-\beta u}}, \quad u \in [0, 1], \quad \beta \in \mathbb{R}. \tag{7}$$

With this notation, we have, for any $k \in \mathbb{Z}_+$,

$$\begin{aligned} \zeta_{k+1} &= \sigma(J_{12}(2\xi_k - 1) + h_1, 2\beta) \\ \xi_{k+1} &= \sigma(J_{12}(2\zeta_{k+1} - 1) + h_2, 2\beta). \end{aligned} \tag{8}$$

Without loss of generality we henceforth set $J_{12} = 1$. Under this choice the model is in the ferromagnetic regime for $\beta > 0$ and the anti-ferromagnetic regime for $\beta < 0$.

4. Why the Ising Model: A Summary of Our Contributions

There are exactly two cases of the Ising model that have a full analytic solution for the free energy. They are (i) the one dimensional line graph solved by Ernst Ising in his thesis [27] and (ii) the two dimensional case on the anisotropic square lattice when the magnetic field $h = 0$ by [28]. Comparison with the mean field solution for the same models highlights the poor approximation quality of the mean field solution in low dimensions. To the best knowledge of the authors, there are no results in the literature detailing the properties of the mean field solution to the anti-ferromagnetic Ising model. Readers not familiar with the physics may wonder why this is the case. To explain this, there are two cases in the anti-ferromagnetic regime: one of the two regions is equivalent to the ferromagnetic case and in the other the mean field approximation is not a good approximation of the system. The first case occurs in a bipartite graph where a transformation of variables makes the antiferromagnetic regime equivalent to the ferromagnetic one [29]. The other case can be seen on the triangle graph. By fixing the spin of one vertex as 1 and the other as -1, the third vertex becomes geometrically frustrated and neither choice of spin lowers the energy level of the system and the two configurations are equivalent [30]. In this case the mean field approximation gives a completely incorrect answer and does not merit further investigation from a qualitative point of view. The physics literature is primarily concerned with using the mean field solutions to the Ising model to estimate important physical constants of the systems. These constants are only meaningful when the mean field solution provides a good approximation to the behavior of the system in large dimensions. It is known, however, that under certain conditions the mean field approximation does indeed converge to the true free energy of the system as the dimension increases [21,31].

Our work is focused on providing a rigorous methodology to analyze dynamics of the CAVI algorithm that can be applied to any model structure. All of the interesting behaviors exhibited by the CAVI algorithm fit into the classical mathematical framework of discrete dynamical systems

and bifurcation theory. Specifically, we use the Ising model as a simple and yet rich example to illustrate the potential of dynamical systems theory to analyze CAVI updates for mean field variational inference. The bifurcation of the ferromagnetic Ising model at the boundary of the Dobrushin regime is known [2,26]; however, a rigorous proof in terms of dynamical systems theory is missing in the literature.

There are several features that make the CAVI algorithm on the Ising model a nontrivial example worth investigating. The optimization problem arising from mean field variational inference on the Ising model is, in general, non-convex [21]. However, it is straightforward to obtain sufficient conditions to guarantee the existence of a global optima. One such condition is that the inverse temperature β is inside the Dobrushin regime, $|\beta| < 1$ [21]. Inside the Dobrushin regime, the CAVI update equations form a contraction mapping guaranteeing a unique global optima [21]. Outside of this regime the behavior of the CAVI algorithm is nontrivial. The CAVI solution to the Ising model with zero external magnetic field exhibits multiple local optima outside of the Dobrushin regime [2].

Our contributions to the literature are as follows. We utilize tools from dynamical systems theory to rigorously classify the full behavior of Ising model for the full parameter regime in dimension $n = 2$ for both the sequential and parallel versions of CAVI algorithm. We show that the dynamical behavior of the sequential CAVI is not equivalent to the behavior of the parallel CAVI. Lastly we derive a variational approximation to the Edward-Sokal parameter expansion of the Potts and Random Cluster models and numerically study its convergence behavior under the CAVI algorithm. Our numerical results reveal that the parameter expansion leads to an enlargement of the regime of convergence to the global optima. In particular the Dobrushin regime is strictly contained in the expanded regime. This is compatible with the analogous results in Markov chain literature. See the introduction of [32] for a well written summary of Markov chain mixing in the Ising model.

Statistical Significance of Our Results

Although mean-field variational inference has been routinely used in applications [3] for computational efficiency, it may not yield statistically optimal estimators. A statistically optimal estimator should correctly recover the statistical properties of the true distribution. Ideally, we would like the estimate to recover the true mean and true covariance of the distribution. It is well known that mean-field variational inference produces estimators that underestimate the posterior covariance [14]. More recently, it was shown that the mean-field estimators for certain topic models and stochastic block models may not even be correlated with the true distribution [17,20]. For these reasons, it is important to see if the mean field estimators can at least recover the true mean for various $\beta \in \mathbb{R}$.

Mean field inference approximates the joint probability mass function in (6) for $n = 2$ by product of two distributions on $\{-1, 1\}$ in the sense of Kullback–Leibler divergence. As discussed in Section 3, minimizing this divergence is equivalent to maximizing an objective function, called the Evidence Lower Bound (ELBO). Our objective is to better understand the relation between the CAVI estimate and the global maximum of ELBO in (6) when $n = 2$ and $h = 0$. Ideally, we want the global maximum of the ELBO to be a statistically reliable estimate. To understand this, let us denote $2 \times \text{Bernoulli}(p) - 1$ by $\langle 1, -1; p \rangle$. As the marginal distributions of (6) are both equal to $\langle 1, -1; 0.5 \rangle$, we want the ELBO to be maximized at this value. From an algorithmic perspective, we would like to ensure that the CAVI iterates converge to this global maximum. The synergy of these two phenomena leads to a successful variational inference method. We show in this article that both these conditions can be violated in a certain regime of the parameter space in the context of Ising model on two nodes. Inside the Dobrushin regime ($-1 \leq \beta \leq 1$), the global optima of the ELBO obtained from a mean field inference occurs at $(\langle 1, -1; 0.5 \rangle, \langle 1, -1; 0.5 \rangle)$ which is qualitatively the optimal solution. In this regime, the CAVI system converges to this global optimum irrespective of where the system is initialized. Thus, in the Dobrushin regime, the mean field inference yields the statistically optimal estimate. Additionally, the CAVI algorithm is stable and convergent at this value. Unfortunately, this property deteriorates outside of the Dobrushin regime. Outside of the regime, the global maxima occur at

two symmetric points which are different from $(\langle 1, -1; 0.5\rangle, \langle 1, -1; 0.5\rangle)$. These two symmetric points are equivalent under label switching. For example, when $\beta = 1.2$ one of the optima is $(\langle 1, -1; 0.17071\rangle, \langle 1, -1; 0.17071\rangle)$ and the other is $(\langle 1, -1; 0.82928\rangle, \langle 1, -1; 0.82928\rangle)$. Notice this second optima is equivalent to the sign swapped version $(\langle -1, 1; 0.17071\rangle, \langle -1, 1; 0.17071\rangle)$.

The original optima $(\langle 1, -1; 0.5\rangle, \langle 1, -1; 0.5\rangle)$ is actually a local minimum of the ELBO outside the Dobrushin regime. We illustrate in our theory that the CAVI system returns one of two global maxima of the objective function depending on the initialization of the algorithm. Although it is widely known that the statistical quality of the mean field inference is poor outside the regime, we show in addition that the algorithm itself exhibits erratic behavior and may not converge to the global maximizer of the ELBO for all initializations. Interestingly, outside the Dobrushin regime, the statistically optimal solution $(\langle 1, -1; 0.5\rangle, \langle 1, -1; 0.5\rangle)$ is a repelling fixed point of the CAVI system. This means that as the system is iterated, the current value of the system is pulled away from $(\langle 1, -1; 0.5\rangle, \langle 1, -1; 0.5\rangle)$ to the global maximum.

A common technique to further improve computational time is the use of block updates in the CAVI algorithm, meaning groups of parameters are updated simultaneously. We refer to this as the parallelized CAVI algorithm. This has been shown to work well in certain models [17], but has not been investigated in a general setting. However, it turns out that block updating in the Ising model can lead to new problematic behaviors. Outside the Dobrushin regime, block updates can exhibit non-convergence in the form of cycling. As the system updates, it eventually switches back and forth between two points that yield the same value in the objective function.

Parameter expansions (coupling) is another method of improving the convergence properties of algorithms. In the Markov chain theory for Ising models, it is well-known that mixing and convergence time are typically improved by using the Edward–Sokal coupling, a parameter expansion of the ferromagnetic Ising model [33]. Our preliminary investigation reveals that the convergence properties of the CAVI algorithm also exhibit a similar phenomenon.

5. Main Results

In this section, we analyze the behavior of the dynamical systems that one can form using the CAVI update equations and show that the behaviors of the systems differ. Our results are heavily dependent on well-known techniques in dynamical systems. For readers unfamiliar with some of technical terminology below, we have included a primer on the basics of dynamical systems in Appendix A.

Recall the system of sequential updates, which are the updates used in CAVI:

$$\zeta_{k+1} = \sigma(2\xi_k - 1, 2\beta), \quad \xi_{k+1} = \sigma(2\zeta_{k+1} - 1, 2\beta), \tag{9}$$

and the parallel updates:

$$\zeta_{k+1} = \sigma(2\xi_k - 1, 2\beta), \quad \xi_{k+1} = \sigma(2\zeta_k - 1, 2\beta). \tag{10}$$

We will show that these two systems are not topologically conjugate. We first state and prove some results on the dynamics of the sigmoid function (7). These results will be used as building blocks to study the dynamics of (9) and (10). Phase change behavior of dynamical systems using the sigmoid and RELU activation functions are known in the literature in the context of generalization performance of deep neural networks [34,35]. In this section we present a complete proof of the bifurcation analysis of non-linear dynamical systems involving sigmoid activation function despite its connections with [34,35]. Our results in Section 5.1 provide a more complete picture of the behavior of the dynamics in all regimes and can be readily exploited to analyze the dynamics of (9) and (10).

5.1. Sigmoid Function Dynamics

In this section we provide a full classification for the dynamics of the following sigmoid function and its second iterate,

$$\sigma(2x-1,2\beta), \tag{11}$$
$$\sigma(2\sigma(2x-1,2\beta)-1,2\beta). \tag{12}$$

To the best of our knowledge, we could not find a formal proof of the full classification of the dynamics of the sigmoid function (or its second iterate) for all $\beta \in \mathbb{R}$ in the literature. Additionally, it provides an introductory example to demonstrate the concepts and techniques of dynamical systems. We begin by using numerical techniques to determine the number of fixed points in the system and its possible periodic behavior. We then proceed by providing a formal proof of the full dynamical properties of (11) in Theorem 1 and the full dynamical properties of (12) in Theorem 2.

Using numerical techniques, we solve for the number of fixed points of the system. The number of fixed points the function (11) depends on the magnitude of the parameter. For $\beta > 0$, there is no periodic behavior, so there are no additional fixed points in (12) that are not fixed points in (11). For $-1 \leq \beta \leq 1$, there is a single fixed point at $x_* = 1/2$ and for $\beta > 1$, there are 3 fixed points $c_0(\beta), 1/2, c_1(\beta)$ in the interval $[0,1]$. These fixed points satisfy $0 \leq c_0(\beta) < 1/2 < c_1(\beta) \leq 1$, $c_0(\beta) \to 0$ and $c_1(\beta) \to 1$ as $\beta \to \infty$. For $\beta < 0$, we see periodic behavior in the system; there are fixed points of (12) that are not fixed points of (11). For $\beta < -1$, the function (11) has one fixed point at $x_* = 1/2$ and a periodic cycle $C = \{c_0(\beta), c_1(\beta)\}$. Both $c_0(\beta), c_1(\beta)$ are fixed points of (12) and these points are the same fixed points from the $\beta > 0$ regime as (12) is an even function with respect to β.

Table 1 denotes the values of the derivatives at the fixed point $1/2$ for $\beta = \pm 1$.

Table 1. Partial derivatives of (11) and (12) at fixed point $x_* = 1/2$ for parameter value $\beta = \pm 1$. The derivatives of the the function (11) are denoted using σ and the derivatives for (12) are denoted using σ^2.

	σ_x	σ_{xx}	σ_{xxx}	σ_β	$\sigma_{\beta x}$	σ^2_x	σ^2_{xx}	σ^2_{xxx}	σ^2_β	$\sigma^2_{\beta x}$
$\beta = 1$	1	0	-8	0	1/2	1	0	-16	0	1
$\beta = -1$	-1	0	8	0	1/2	1	0	-16	0	-1

We now have enough information to provide a complete classification of the dynamics of the sigmoid function.

Theorem 1 (Dynamics of sigmoid function)**.** *Consider the discrete dynamical system generated by (11)*

$$x \mapsto \sigma(2x-1,2\beta) = \frac{1}{1+e^{-2\beta(2x-1)}}.$$

The full dynamics of the system (11) are as follows

1. *For $-1 \leq \beta \leq 1$, the system has a single hyperbolic fixed point $x_* = 1/2$ which is a global attractor and there are no p-periodic points for $p \geq 2$.*
2. *For $\beta > 1$, the system has one repelling hyperbolic fixed point $x_* = 1/2$ and two hyperbolic stable fixed points c_0, c_1, with $0 < c_0 < 1/2 < c_1 < 1$, and stable sets $W^s(c_0) = [0, 1/2)$, $W^s(c_1) = (1/2, 1]$. There are no p-periodic points for $p \geq 2$.*
3. *For $\beta < -1$, the system has one unstable hyperbolic fixed point $x_* = 1/2$, and a stable 2-cycle $\mathcal{C} = \{c_0, c_1\}$ with stable set $W^s(\mathcal{C}) = [0, 1/2) \cup (1/2, 1]$, with $0 < c_0 < 1/2 < c_1 < 1$. There are no p-periodic points for $p > 2$.*

4. *For $|\beta| = 1$, the system has one non-hyperbolic fixed point at $x_* = 1/2$ which is asymptotically stable and attracting.*

   The system undergoes a PD bifurcation at $\beta = -1$ and a pitchfork bifurcation at $\beta = 1$.

Proof. We will break the proof up into three parts. The first part of the proof is a linear stability analysis of the system, the second part is a stability analysis of the periodic points in the system and the third part is an analysis of the bifurcations of the system. We begin with a linear stability analysis of the system at each fixed point. For $\beta \leq 1$ the system has one fixed point $x_* = 1/2$ and for $\beta > 1$ the system has three fixed points $c_0, 1/2, c_1$. The derivative of $\sigma(2x - 1, 2\beta)$ is $\sigma_x(2x - 1, 2\beta) = -4\beta\sigma(2x - 1, 2\beta)(1 - \sigma(2x - 1, 2\beta))$.

Fixed point $x_* = 1/2$: The Jacobian of the system at the fixed point $x_* = 1/2$ is $\sigma_x(2x_* - 1, 2\beta) = \beta$. For $\beta \neq 1$, the fixed point $x_* = 1/2$ is hyperbolic and for $\beta = \pm 1$ the fixed point is non-hyperbolic. We classify the stability of the hyperbolic fixed point $x_* = 1/2$ using Theorem A2. For $|\beta| < 1$ the fixed point $x_* = 1/2$ is globally attracting as $|\sigma_x(2x_* - 1, 2\beta)| < 1$ and for $|\beta| > 1$ the fixed point $x_* = 1/2$ is globally repelling as $|\sigma_x(2x_* - 1, 2\beta_*)| > 1$. For $\beta = \pm 1$ we invoke Theorem A3 to check for stability of the fixed point. At $\beta = -1$ we have $\sigma_x(2x_* - 1, 2\beta) = -1$ and we need to check the Schwarzian derivative. The fixed point $x_* = 1/2$ is asymptotically stable for $\beta = -1$ by Theorem A3, as $\mathcal{S}\sigma(2\sigma(2x - 1, 2\beta) - 1, 2\beta) \mid_{x=x_*} = -8$. For $\beta = 1$ we have $\sigma_x(2x_* - 1, 2\beta) = 1$ and we need to check the second and third derivatives at the fixed point. The fixed point $x_* = 1/2$ is asymptotically stable when $\beta = 1$ by Theorem A3 as $\sigma_{xx}(2x_* - 1, 2\beta) = 0$ and $\sigma_{xxx}(2x_* - 1, 2\beta) = -8$.

Fixed points c_0, c_1 : These fixed points have the same behavior so we have grouped them together in the analysis. When $\beta > 1$ there are two additional fixed points c_0, c_1 of the system, both are attracting fixed points by Theorem A2 as $|\sigma_x(2c_i - 1, 2\beta)| < 1$ for each $i = 0, 1$ and all $\beta > 1$. The stable sets are $W^s(c_0) = [0, 1/2)$ and $W^s(c_1) = (1/2, 1]$.

Periodic points: For $\beta < -1$ we see the two cycle $\mathcal{C} = \{c_0, c_1\}$. Notice $\sigma(2c_0 - 1, 2\beta) = c_1$ and $\sigma(2c_1 - 1, 2\beta) = c_0$. This two cycle is stable since c_0 and c_1 are both stable fixed points of (12). The stable set is $W^s(\mathcal{C}) = [0, 1/2) \cup (1/2, 1]$, $\quad 0 < c_0 < 1/2 < c_1 < 1$.

At $(x_*, \beta_*) = (1/2, 1)$ the system under goes a pitchfork bifurcation as it satisfies the conditions in Theorem A5:

$$\sigma(2x_* - 1, 2\beta_*) = 1/2 \quad \sigma_x(2x_* - 1, 2\beta_*) = 1 \quad \sigma_{xx}(2x_* - 1, 2\beta_*) = 0,$$
$$\sigma_\beta(2x_* - 1, 2\beta_*) = 0 \quad \sigma_{x\beta}(2x_* - 1, 2\beta_*) \neq 0 \quad \sigma_{xxx}(2x_* - 1, 2\beta_*) \neq 0.$$

Similarly at $(x_*, \beta_*) = (1/2, -1)$ the system under goes a period doubling bifurcation as it satisfies the conditions in Theorem A4

$$\sigma(2x_* - 1, 2\beta_*) = 1/2 \quad \sigma_x(2x_* - 1, 2\beta_*) = -1 \quad \sigma_{xx}(2x_* - 1, 2\beta_*) = 0,$$
$$\sigma_\beta(2x_* - 1, 2\beta_*) = 0 \quad \sigma_{x\beta}(2x_* - 1, 2\beta_*) \neq 0 \quad \sigma_{xxx}(2x_* - 1, 2\beta_*) \neq 0.$$

□

We can fully classify the dynamics of (12) using the above theorem. We omit the proof as it is similar to the proof of Theorem 1.

Theorem 2. *The full dynamics of the system (12) are as follows*

1. *For $-1 \leq \beta \leq 1$, the system has a single hyperbolic fixed point $x_* = 1/2$ which is a global attractor and there are no p-periodic points for $p \geq 2$.*
2. *For $|\beta| > 1$, the system has one repelling hyperbolic fixed point $x_* = 1/2$ and two hyperbolic stable fixed points c_0, c_1, with $0 < c_0 < 1/2 < c_1 < 1$, and stable sets $W^s(c_0) = [0, 1/2)$, $W^s(c_1) = (1/2, 1]$.*
3. *For $|\beta| = 1$, the system has one non-hyperbolic fixed point at $x_* = 1/2$ which is asymptotically stable and attracting.*

The system undergoes a pitchfork bifurcation at $\beta = \pm 1$. There are no p-periodic points for $p \geq 2$.

5.2. Sequential Dynamics

To fully understand the dynamics of the equations defining the updates to q_1^* and q_2^* it suffices to track the evolution of the points $q_1^*(1) = \zeta$ and $q_2^*(1) = \xi$. The CAVI algorithm updates terms sequentially, using the new values of the variables to calculate the others. We initialize the CAVI algorithm at points ζ_0, ξ_0. The CAVI algorithm is a dynamical system formed by sequential iterations of $\sigma(2x - 1, 2\beta)$ starting from ζ_0, ξ_0. We can decouple the CAVI updates for ξ_k and ζ_k by looking at the second iterations. This decoupling is visualized in the diagram (14) below. The system formed the sequential updates is equivalent to the following decoupled system

$$\begin{aligned} \zeta_1 &= \sigma(2\xi_0 - 1, 2\beta), \\ \zeta_{k+1} &= \sigma(2\sigma(2\zeta_k - 1, 2\beta) - 1, 2\beta), \quad k \geq 1, \\ \xi_{k+1} &= \sigma(2\sigma(2\xi_k - 1, 2\beta) - 1, 2\beta), \quad k \geq 0. \end{aligned} \tag{13}$$

We propose to investigate the dynamics of the sequential system (9) by studying the dynamics of individual subsequences ζ_{k+1} and ξ_{k+1} of the decoupled system (13). The dynamical properties of the individual subsequences follow from a combination of Theorem 1, Theorem 2 and other methods from Appendix A.

$$\begin{array}{ccccccc} \zeta_0 & & \zeta_1 & \xrightarrow{\sigma^2} & \zeta_2 & \xrightarrow{\sigma^2} & \zeta_3 \cdots \\ & \overset{\sigma}{\nearrow} & \downarrow \sigma & \overset{\sigma}{\nearrow} & \downarrow \sigma & \overset{\sigma}{\nearrow} & \\ \xi_0 & \xrightarrow{\sigma^2} & \xi_1 & \xrightarrow{\sigma^2} & \xi_2 & \xrightarrow{\sigma^2} & \xi_3 \cdots \end{array} \tag{14}$$

Illustrations of the evolution of the dynamics of the sequential updates for various initializations and values of β are in Figures 3–6.

Theorem 3 (CAVI dynamics)**.** *The Dynamics of the CAVI System (9) Are Given by*

1. *For $\beta < -1$, the system has the system has one locally asymptotically unstable fixed point $(1/2, 1/2)$ and two locally asymptotically stable fixed points (c_0, c_1) and (c_1, c_0), with stable sets $W^s((c_0, c_1)) = [0, 1] \times [0, 1/2)$ and $W^s((c_1, c_0)) = [0, 1] \times (1/2, 1]$ respectively.*
2. *For $|\beta| \leq 1$, the system has a global asymptotically stable fixed point $(1/2, 1/2)$.*
3. *For $\beta > 1$ the system has the system has one locally asymptotically unstable fixed point $(1/2, 1/2)$ and two locally asymptotically stable fixed points (c_0, c_0) and (c_1, c_1), with domains of attraction $W^s((c_0, c_0)) = [0, 1] \times [0, 1/2)$ and $W^s((c_1, c_1)) = [0, 1] \times (1/2, 1]$ respectively.*

where $0 \leq c_0 < 1/2 < c_1 \leq 1$ are the fixed points of (11) in $[0, 1]$. The system undergoes a super-critical pitchfork bifurcation at $\beta = -1$ and again at $\beta = 1$. Furthermore the system has no p-periodic points for $p \geq 2$.

Proof. We will proceed to construct the dynamics of the system (9) by tracing the behavior of the dynamics in the equivalent system (13). The dynamics of each of these subsequences is governed by the Functions (11) and (12) and dependent on the initialization ξ_0. The behavior for each of the subsequence ξ_{k+1}, for $k \geq 0$ is governed by Theorem 2. Similarly the behavior of the subsequence ζ_{k+1}, for $k \geq 1$ is governed by Theorem 2 with the additional point $\zeta_1 = \sigma(2\xi_0 - 1, 2\beta)$ dependent on Theorem 1. For $|\beta| < 1$, (11) has a globally stable fixed point at $x_* = 1/2$ and thus for all ξ_0, $\zeta_1 = \sigma(2\xi_0 - 1, 2\beta) \in W^s(1/2)$. It now follows from Theorem 2 that the only fixed point in the sequential system is $(1/2, 1/2)$ which must be globally stable. For $\beta = \pm 1$, the fixed point $x_0 = 1/2$ is asymptotically stable by Theorem A3. The system undergoes a super-critical pitchfork bifurcation at $\beta = -1$ and again at $\beta = 1$ as a consequnece from its relation to (12). For $\beta > 1$, (11) bifurcates. We have the unstable fixed point $x_* = 1/2$, and the two locally stable fixed points, c_0 with stable set

$W^s(c_0) = [0, 1/2)$, and c_1 with stable set $W^s(c_1) = (1/2, 1]$. For $\xi_0 \in W^s(c_0)$ we have $\zeta_1 \in W^s(c_0)$ and $\xi_1 \in W^s(c_0)$. It now follows from Theorem 2 that the system will converge to (c_0, c_0) and that $W^s((c_0, c_0)) = [0, 1] \times [0, 1/2)$. A similar argument shows the system converges to (c_1, c_1) for $\xi_0 \in W^s(c_1)$ and $W^s((c_1, c_1)) = [0, 1] \times (1/2, 1]$. Lastly, $(1/2, 1/2)$ is a repelling fixed point of the systems since $x_* = 1/2$ is a repelling fixed point for both (11) and (12). For $\beta < -1$, (11) bifurcates. We have the unstable fixed point $x_* = 1/2$, and the stable two cycle, $\mathcal{C} = \{c_0, c_1\}$ with stable set $W^s(\mathcal{C}) = [0, 1/2) \cup (1/2, 1]$. For any $\xi_0 < 1/2$ we have, $\zeta_1 > 1/2$ and $\xi_1 < 1/2$. It now follows from Theorem 2 that the system will converge to (c_1, c_0) and that $W^s((c_1, c_0)) = [0, 1] \times [0, 1/2)$. A similar argument shows the system converges to (c_0, c_1) for $\xi_0 > 1/2$ and $W^s((c_0, c_1)) = [0, 1] \times [0, 1/2)$. Lastly, $(1/2, 1/2)$ is a repelling fixed point of the systems since $x_* = 1/2$ is a repelling fixed point for both (11) and (12). The dynamics of (13) lack any p-period point and cycles for $p > 2$ as a consequence of its construction from (12). □

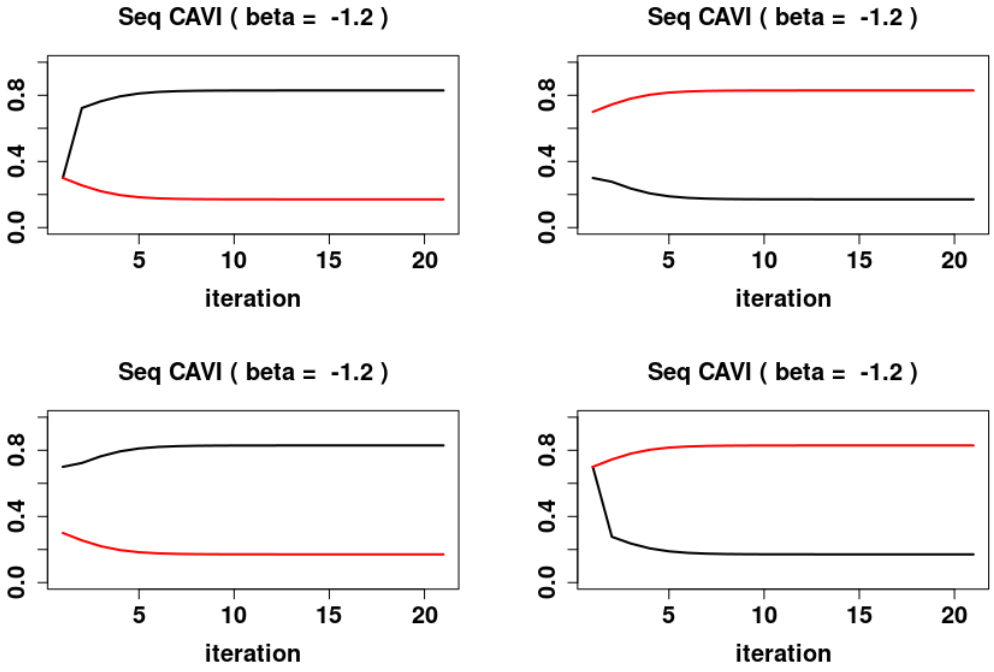


Figure 3. A plot of the first 20 iterations of the CAVI algorithm at various initializations for $\beta = -1.2$. In each of the plots the ζ updates are black and the ξ updates are red. The upper left plot is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.3$; we see that ζ_k converges to the local fixed point $c_1(1.2) = 0.82928$ and ξ_k converges to the local fixed point $c_0(1.2) = 0.17071$. The upper right is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.7$; we see that ζ_k converges to the local fixed point $c_0(1.2) = 0.17071$ and ξ_k converges to the local fixed point $c_1(1.2) = 0.82928$. The lower left is is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.3$; we see that ζ_k converges to the local fixed point $c_1(1.2) = 0.82928$ and ξ_k converges to the local fixed point $c_0(1.2) = 0.17071$. The upper left plot is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.7$; we see that ζ_k converges to the local fixed point $c_0(1.2) = 0.17071$ and ξ_k converges to the local fixed point $c_1(1.2) = 0.82928$.

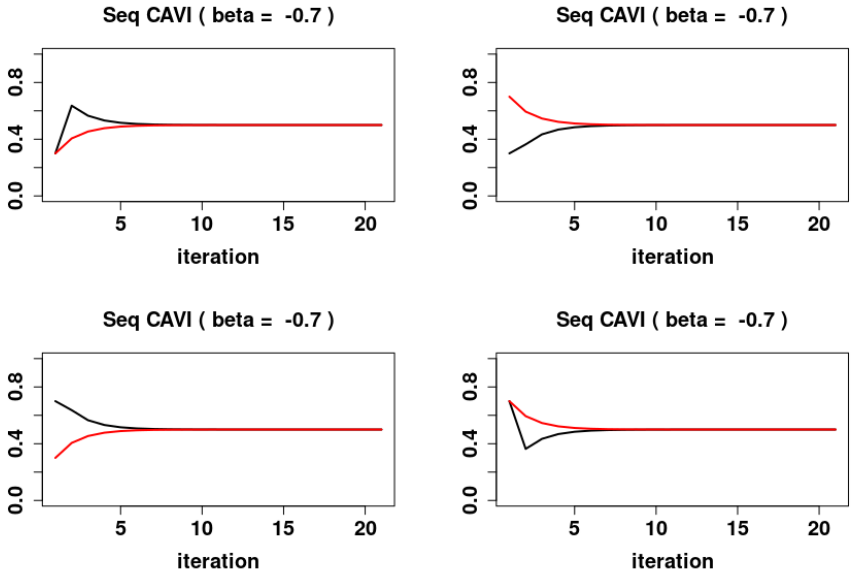


Figure 4. A plot of the first 20 iterations of the CAVI algorithm at various initializations for $\beta = -0.7$. In each of the plots the ζ updates are black and the ξ updates are red. The upper left plot is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.3$; we see that both of these converge to the global fixed point $1/2$. The upper right is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.7$; we see that this initialization converges to the global fixed point $1/2$. The lower left is is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.3$; we see that this initialization converges to the global fixed point $1/2$. The upper left plot is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.7$; we see that both of these converge to the global fixed point $1/2$.

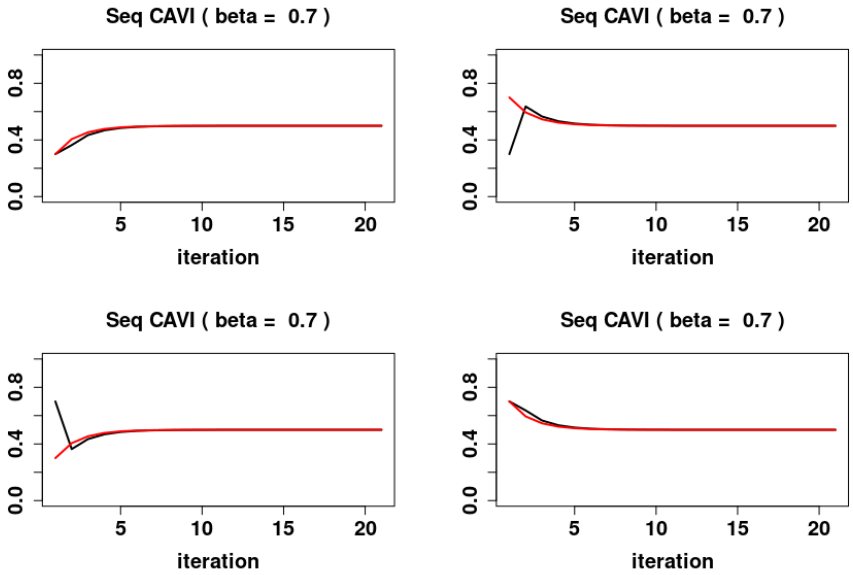


Figure 5. A plot of the first 20 iterations of the CAVI algorithm at various initializations for $\beta = 0.7$. In each of the plots the ζ updates are black and the ξ updates are red. The upper left plot is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.3$; we see that both of these converge to the global fixed point $1/2$. The upper right is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.7$; we see that this initialization converges to the global fixed point $1/2$. The lower left is is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.3$; we see that this initialization converges to the global fixed point $1/2$. The upper left plot is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.7$; we see that both of these converge to the global fixed point $1/2$.

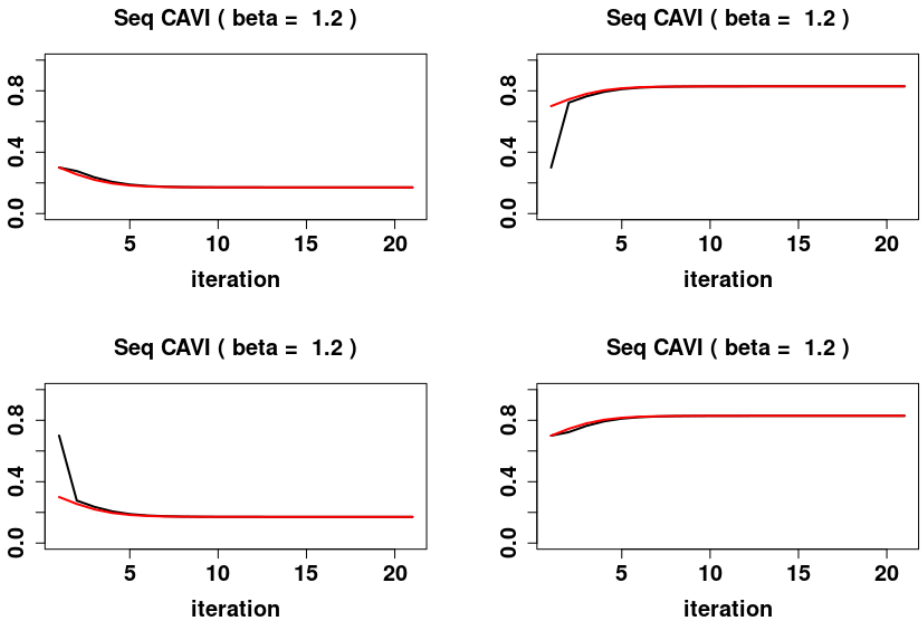


Figure 6. A plot of the first 20 iterations of the CAVI algorithm at various initializations for $\beta = 1.2$. In each of the plots the ζ updates are black and the ξ updates are red. The upper left plot is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.3$; we see that both of these converge to the local fixed point $c_0(1.2) = 0.17071$. The upper right is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.7$; we see that this initialization converges to the local fixed point $c_1(1.2) = 0.82928$. The lower left is is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.3$; we see that this initialization converges to the local fixed point $c_0(1.2) = 0.17071$. The upper left plot is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.7$; we see that both of these converge to the local fixed point $c_1(1.2) = 0.82928$.

5.3. Parallel Updates

The system of parallel updates is defined by the one-step map $F : \mathbb{R}^2 \to \mathbb{R}^2$

$$\begin{pmatrix} \zeta \\ \xi \end{pmatrix} \mapsto F(\zeta, \xi) = \begin{pmatrix} \sigma(2\xi - 1, 2\beta) \\ \sigma(2\zeta - 1, 2\beta). \end{pmatrix} \tag{15}$$

The dynamics of the parallel system are similar to the system studied in [36]. As we shall show below, the parallel system exhibits periodic behavior that the sequential system does not and it follows as a corollary that the systems are not locally topologically conjugate.

The parallelized CAVI algorithm is a dynamical system formed by iterations of F defined in (15). We shall decouple the parallelized CAVI updates for sequences ξ_k and ζ_k by looking at iterations of (12) acting on the sequences individually. This decoupling is visualized in diagram form

$$\begin{array}{ccccccc} \zeta_0 & & \zeta_1 & & \zeta_2 & & \zeta_3 \cdots \\ & \searrow\!\!\nearrow & & \searrow\!\!\nearrow & & \searrow\!\!\nearrow & \\ \xi_0 & & \xi_1 & & \xi_2 & & \xi_3 \cdots \end{array} \tag{16}$$

where each cross is an application of F. The system formed the parallel updates is equivalent to the following decoupled systems of even subsequences and odd subsequences. The even subsequences are

$$\zeta_{2k} = \sigma(2\sigma(2\zeta_{2(k-1)} - 1, 2\beta) - 1, 2\beta), \quad k \geq 1 \tag{17}$$

$$\xi_{2k} = \sigma(2\sigma(2\xi_{2(k-1)} - 1, 2\beta) - 1, 2\beta), \quad k \geq 1. \tag{18}$$

The odd subsequences are

$$\zeta_{2k+1} = \begin{cases} \sigma(2\xi_0, 2\beta) & k = 0 \\ \sigma(2\sigma(2\zeta_{2k-1}, 2\beta), 2\beta) & k \geq 1 \end{cases} \tag{19}$$

$$\xi_{2k+1} = \begin{cases} \sigma(2\zeta_0, 2\beta) & k = 0 \\ \sigma(2\sigma(2\xi_{2k-1}, 2\beta), 2\beta) & k \geq 1. \end{cases} \tag{20}$$

Following a similar approach to the one used to study the sequential dynamics, we investigate the dynamics of the parallel system (15) by studying the dynamics of four individual subsequences (17)–(20) of the decoupled system given by diagram (16). The dynamical properties of the individual subsequences follow from a combination of Theorem 1, Theorem 2 and other methods from Appendix A. Illustrations of the evolution of the dynamics of the parallel updates for various initializations and values of β are in Figures 7–12.

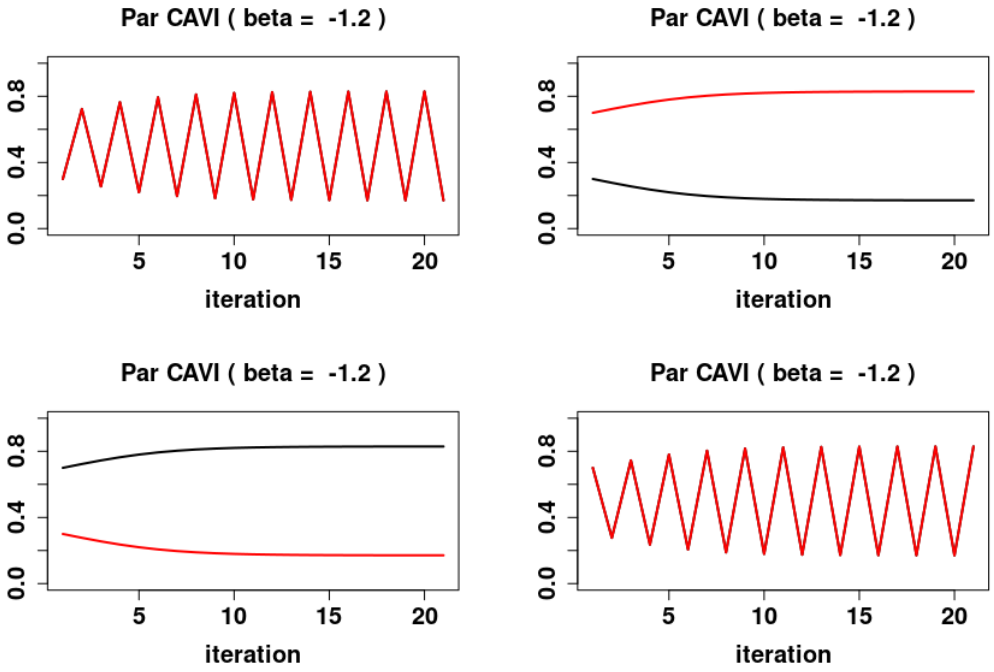


Figure 7. A plot of the first 20 iterations of the parallel update CAVI algorithm at various initializations for $\beta = -1.2$. In each of the plots the ζ updates are black and the ξ updates are red. The upper left is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.7$; we see that this initialization converges to the two cycle $\mathcal{C}_0 = \{(c_0, c_0), (c_1, c_1)\}$. The upper right plot is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.7$; we see that both of these converge to $c_0(1.2) \approx 0.17071$. The lower left plot is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.7$; we see that both of these converge to $c_1(1.2) \approx 0.82928$. The lower right is is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.3$; we see that this initialization converges to the two cycle $\mathcal{C}_0 = \{(c_0, c_0), (c_1, c_1)\}$.

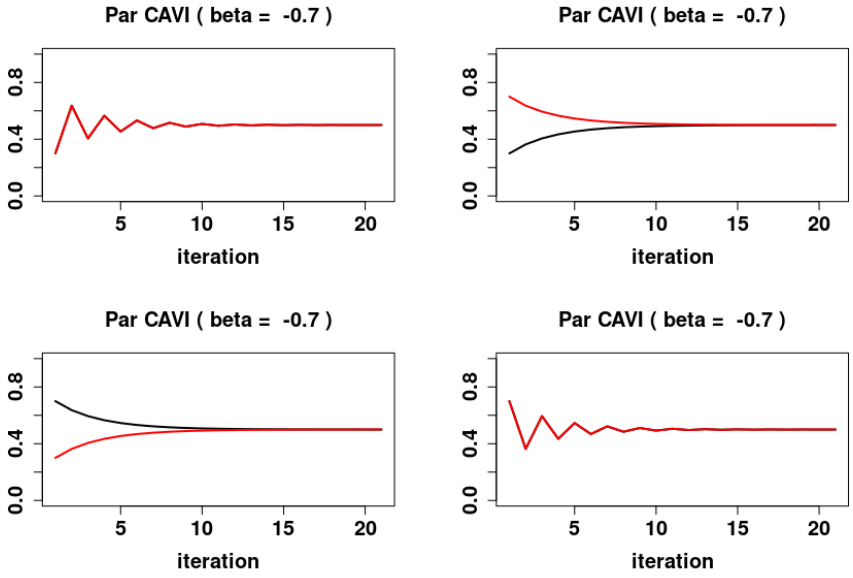


Figure 8. A plot of the first 20 iterations of the parallel update CAVI algorithm at various initializations for $\beta = -0.7$. In each of the plots the ζ updates are black and the ξ updates are red. The upper left plot is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.3$; we see that both of these converge to the global fixed point $1/2$. The upper right is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.7$; we see that this initialization converges to the global fixed point $1/2$. The lower left is is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.3$; we see that this initialization converges to the global fixed point $1/2$. The upper left plot is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.7$; we see that both of these converge to the global fixed point $1/2$.

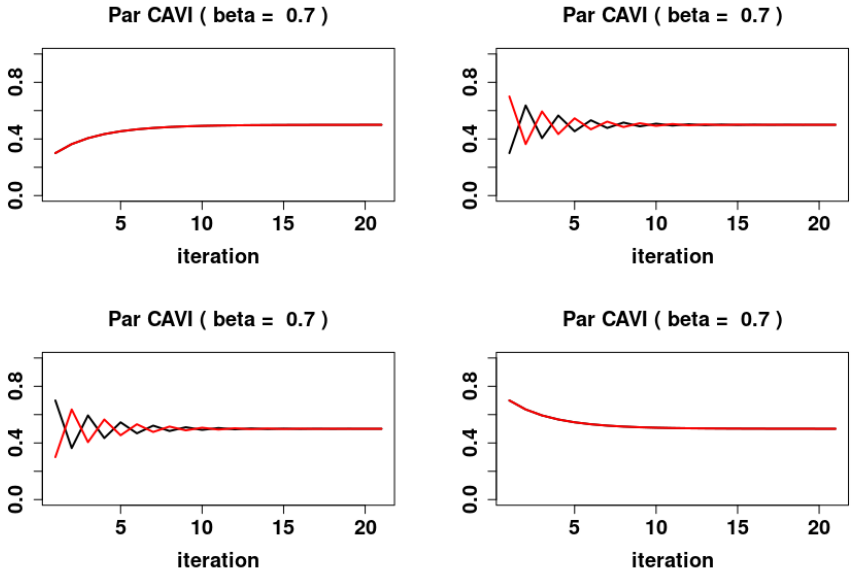


Figure 9. A plot of the first 20 iterations of the parallel update CAVI algorithm at various initializations for $\beta = 0.7$. In each of the plots the ζ updates are black and the ξ updates are red. The upper left plot is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.3$; we see that both of these converge to the global fixed point $1/2$. The upper right is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.7$; we see that this initialization converges to the global fixed point $1/2$. The lower left is is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.3$; we see that this initialization converges to the global fixed point $1/2$. The upper left plot is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.7$; we see that both of these converge to the global fixed point $1/2$.

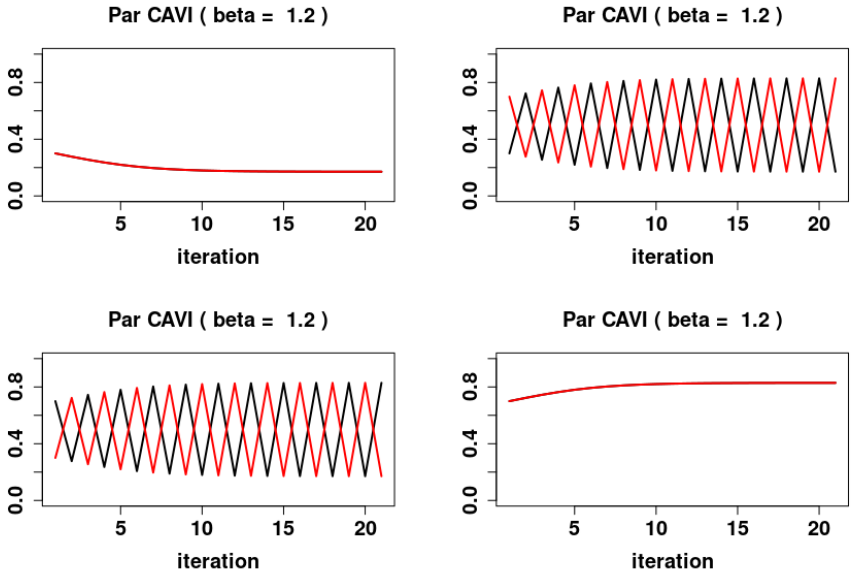


Figure 10. A plot of the first 20 iterations of the parallel update CAVI algorithm at various initializations for $\beta = 1.2$. In each of the plots the ζ updates are black and the ξ updates are red. The upper left plot is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.3$; we see that both of these converge to $c_0(1.2) \approx 0.17071$. The upper right is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.7$; we see that this initialization converges to the two cycle $\mathcal{C}_1 = \{(c_1, c_0), (c_0, c_1)\}$. The lower left is is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.3$; we see that this initialization converges to the two cycle $\mathcal{C}_1 = \{(c_1, c_0), (c_0, c_1)\}$. The lower right plot is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.7$; we see that both of these converge to $c_1(1.2) \approx 0.82928$.

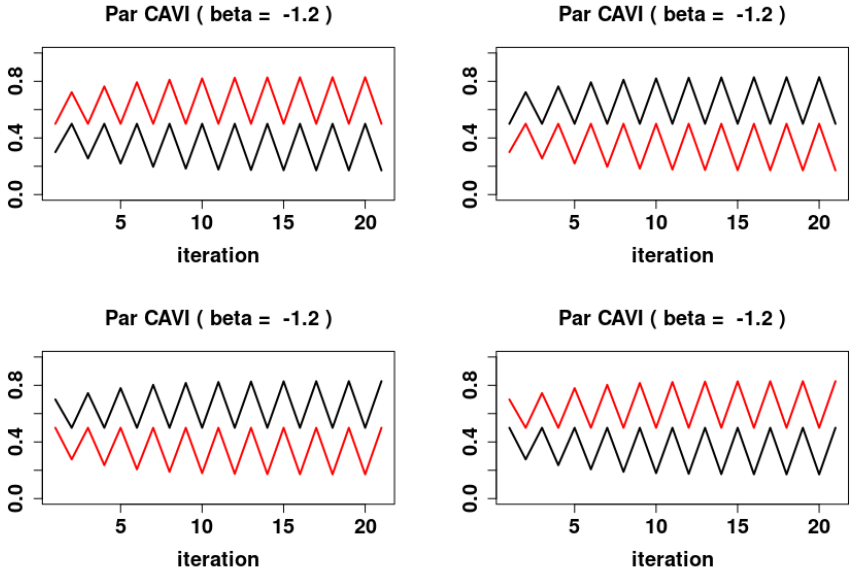


Figure 11. A plot of the first 20 iterations of the parallel update CAVI algorithm at various initializations for $\beta = -1.2$. In each of the plots the ζ updates are black and the ξ updates are red. The upper left plot is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.5$; we see that this converges to the two-cycle $\mathcal{C}_2 = \{(c_0, 1/2), (1/2, c_1)\}$. The upper right is an initialization of $\zeta_0 = 0.5$ and $\xi_0 = 0.3$; we see that this initialization converges to the two cycle $\mathcal{C}_3 = \{(c_1, 1/2), (1/2, c_0)\}$. The lower left is is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.5$; we see that this initialization converges to the two cycle $\mathcal{C}_3$. The lower right plot is an initialization of $\zeta_0 = 0.5$ and $\xi_0 = 0.7$; we see that this converges to the two-cycle $\mathcal{C}_2$.

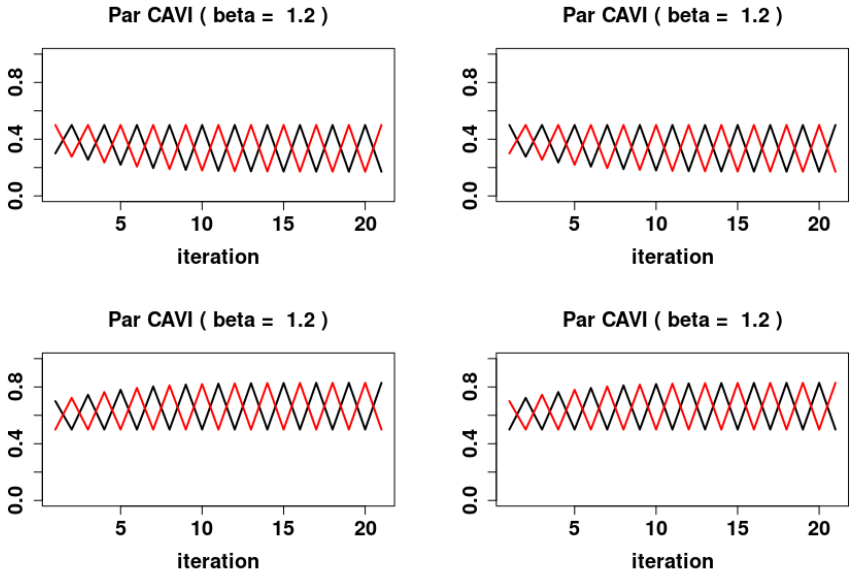


Figure 12. A plot of the first 20 iterations of the parallel update CAVI algorithm at various initializations for $\beta = 1.2$. In each of the plots the ζ updates are black and the ξ updates are red. The upper left plot is an initialization of $\zeta_0 = 0.3$ and $\xi_0 = 0.5$; we see that this converges to the two-cycle $\mathcal{C}_4 = \{(c_0, 1/2), (1/2, c_0)\}$. The upper right is an initialization of $\zeta_0 = 0.5$ and $\xi_0 = 0.3$; we see that this initialization converges to the two cycle $\mathcal{C}_4$. The lower left is is an initialization of $\zeta_0 = 0.7$ and $\xi_0 = 0.5$; we see that this initialization converges to the two cycle $\mathcal{C}_5 = \{(c_1, 1/2), (1/2, c_1)\}$. The lower right plot is an initialization of $\zeta_0 = 0.5$ and $\xi_0 = 0.7$; we see that this converges to the two-cycle $\mathcal{C}_5$.

We now present the main result for the parallel dynamics.

Theorem 4 (Parallel Dynamics). *The Dynamics of the Parallel System (10) Are As Follows*

1. *For $\beta < -1$, the system has two locally asymptotically stable fixed points (c_1, c_0) and (c_0, c_1), and one locally asymptotically unstable fixed point $(1/2, 1/2)$, where c_0 and c_1 are the fixed points of (11). Furthermore the system exhibits periodic behavior in the form of 2-cycles. The asymptotically stable 2-cycle, $\mathcal{C}_1 = \{(c_0, c_0), (c_1, c_1)\}$ and asymptotically unstable 2-cycles,*

$$\mathcal{C}_2 = \{(1/2, c_1), (c_0, 1/2)\} \text{ and } \mathcal{C}_3 = \{(1/2, c_0), (c_1, 1/2)\}.$$

The stable sets are

$$\begin{aligned} W^s(c_0, c_1) &= [0, 1/2) \times (1/2, 1] \\ W^s(c_1, c_0) &= (1/2, 1] \times [0, 1/2) \\ W^s(\mathcal{C}_1) &= ([0, 1/2) \times [0, 1/2)) \cup ((1/2, 1] \times (1/2, 1]) \\ W^s(\mathcal{C}_2) &= ([0, 1/2) \times \{1/2\}) \cup (\{1/2\} \times (1/2, 1]) \\ W^s(\mathcal{C}_3) &= ([0, 1/2) \times \{1/2\}) \cup (\{1/2\} \times (1/2, 1]) . \end{aligned}$$

2. *For $-1 \leq \beta \leq 1$, the system has a global attracting fixed point $(1/2, 1/2)$.*
3. *For $\beta > 1$, the system has two locally asymptotically stable fixed points (c_0, c_0) and (c_1, c_1), and one locally asymptotically unstable fixed point $(1/2, 1/2)$, where c_0 and c_1 are the fixed points of (11). Furthermore the system exhibits periodic behavior in the form of 2-cycles. The asymptotically stable 2-cycle, $\mathcal{C}_3 = \{(c_0, c_0), (c_1, c_1)\}$ and asymptotically unstable 2-cycles, $\mathcal{C}_4 = \{(1/2, c_0), (c_1, 1/2)\}$ and $\mathcal{C}_5 = \{(1/2, c_1), (c_1, 1/2)\}$. The stable sets are*

$$\begin{aligned} W^s(c_0, c_1) &= [0, 1/2) \times (1/2, 1] \\ W^s(c_1, c_0) &= (1/2, 1] \times [0, 1/2) \\ W^s(\mathcal{C}_3) &= ([0, 1/2) \times [0, 1/2)) \cup ((1/2, 1] \times (1/2, 1]) \\ W^s(\mathcal{C}_4) &= ([0, 1/2) \times \{1/2\}) \cup (\{1/2\} \times [0, 1/2)) \\ W^s(\mathcal{C}_5) &= (\{1/2\} \times (1/2, 1]) \cup ((1/2, 1] \times \{1/2\}) . \end{aligned}$$

The system has no p-periodic points for $p > 2$. The system under goes a PD bifurcation at $\beta = -1$ and a pitchfork bifurcation at $\beta = 1$.

Proof. The dynamics of the system defined by F in (15) are equivalent to the dynamics of the system generated by the subsequences (17)–(20). The dynamics of each of these subsequences are governed by the functions (11) and (12). By Theorem 1, we have the behavior for each of the subsequences (17)–(20). For $|\beta| < 1$, (11) has a globally stable fixed point at $x_* = 1/2$ and thus the only fixed point in the parallel system is $(1/2, 1/2)$ which must be globally stable. For $\beta = \pm 1$, the fixed point $x_0 = 1/2$ is asymptotically stable by Theorem A3.

For $\beta > 1$, (11) bifurcates. We have the unstable fixed point $x_* = 1/2$, and the two locally stable fixed points, c_0 with stable set $W^s(c_0) = [0, 1/2)$, and c_1 with stable set $W^s(c_1) = (1/2, 1]$. Returning to the system generated by F, if we consider the initialization $(\zeta_0, \xi_0) = (c_0, c_0)$ then by the sequence construction of ζ_n, given in (17) and (19), we see that $\zeta_n = c_0$ for $n \geq 1$, as c_0 is a fixed point of (11) for $\beta > 1$. Similarly, using the sequence construction of ξ_n, given in (18) and (20), we see that $\xi_n = c_0$ for $n \geq 1$, as c_0 is a fixed point of (11) for $\beta > 1$. Therefore, (c_0, c_0) is a fixed point. An analogous argument shows that (c_1, c_1) is also a fixed point. The parallel system has the stable fixed points (c_0, c_0) with stable set $W^s(c_0, c_0) = W^s(c_0) \times W^s(c_0)$ and (c_1, c_1) with stable set $W^s(c_1, c_1) = W^s(c_1) \times W^s(c_1)$. After the bifurcation at $\beta = 1$ the parallel system also contains 2-cycles. Using the sequence construction we see that $\mathcal{C}_3 = \{(c_1, c_0), (c_0, c_1)\}$ is an asymptotically stable 2-cycle in the parallel system, with stable subspace $W^s(\mathcal{C}_3) = (1/2, 1] \times [0, 1/2) \cup [0, 1/2) \times (1/2, 1]$. Additionally, we have two asymptotically unstable 2-cycles $\mathcal{C}_4 = \{(c_0, 1/2), (1/2, c_0)\}$ and $\mathcal{C}_5 = \{(c_1, 1/2), (1/2, c_1)\}$. Perturbing the $1/2$ coordinate in the unstable cycle pushes it into the basin of attraction for one of the fixed points or the asymptotically stable 2-cycle. The stable sets are $W^s(\mathcal{C}_4) = ([0, 1/2) \times \{1/2\}) \cup (\{1/2, 1\} \times [0, 1/2))$, $W^s(\mathcal{C}_5) = (\{1/2\} \times (1/2, 1]) \cup ((1/2, 1] \times \{1/2\})$. The dynamics of F lack any p-period point and cycles for $p > 2$ as a consequence of its construction from (12).

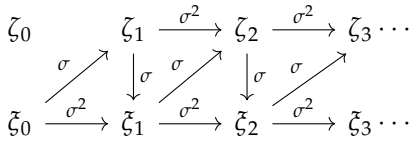
For $\beta < -1$, (11) bifurcates. We have the unstable fixed point $x_* = 1/2$, and the stable two cycle, $\mathcal{C} = \{c_0, c_1\}$ with stable set $W^s(\mathcal{C}) = [0, 1/2) \cup (1/2, 1]$. Returning to the system generated by F, if we consider the initialization $(\zeta_0, \xi_0) = (c_0, c_1)$ then by the sequence construction of ζ_n, given in (17) and (19), we see that $\zeta_n = c_0$ for $n \geq 1$, as $\mathcal{C}$ is a 2-cycle of (11) for $\beta < -1$. Similarly, using the sequence construction of ξ_n, given in (18) and (20) we see that $\xi_n = c_1$ for $n \geq 1$, as $\mathcal{C}$ is a 2-cycle of (11) for $\beta < -1$. Therefore, (c_0, c_1) is a fixed point. An analogous argument shows that (c_1, c_0) is also a fixed point. The parallel system has the stable fixed points (c_0, c_1) with stable set $W^s(c_0, c_1) = W^s(c_0) \times W^s(c_1)$ and (c_1, c_0) with stable set $W^s(c_1, c_0) = W^s(c_1) \times W^s(c_0)$, where $W^s(c_0) = [0, 1/2)$ and $W^s(c_1) = (1/2, 1]$. After the bifurcation at $\beta = -1$ the parallel system also contains 2-cycles. Using the sequence construction we see that $\mathcal{C}_1 = \{(c_0, c_0), (c_1, c_1)\}$ is an asymptotically stable 2-cycle in the parallel system, with stable subspace $W^s(\mathcal{C}_1) = W^s(c_0) \times W^s(c_0) \cup W^s(c_1) \times W^s(c_1)$. Additional we have two asymptotically unstable 2-cycles $\mathcal{C}_2 = \{(c_0, 1/2), (1/2, c_1)\}$ and $\mathcal{C}_3 = \{(c_1, 1/2), (1/2, c_0)\}$. Perturbing the $1/2$ coordinate in the unstable cycle pushes it into the basin of attraction for one of the fixed points or the asymptotically stable 2-cycle. The stable sets are $W^s(\mathcal{C}_3) = ([0, 1/2) \times [0, 1/2)) \cup ((1/2, 1] \times (1/2, 1])$, $W^s(\mathcal{C}_4) = ([0, 1/2) \times \{1/2\}) \cup (\{1/2\} \times [0, 1/2))$, $W^s(\mathcal{C}_5) = (\{1/2\} \times (1/2, 1]) \cup ((1/2, 1] \times \{1/2\})$.

The dynamics of F lack any p-period point and cycles for $p > 2$ as a consequence of its construction from (12).

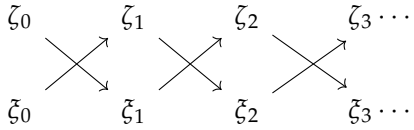
This completes the characterization of the dynamics of F for $\beta \in \mathbb{R}$. □

5.4. A Comparison of the Dynamics

We end the section by providing a comparison of the dynamical properties of the sequential system in Theorem 3 and the parallel system in Theorem 4. The main difference between the sequential system and the parallel system is the presence of two-cycles that can be found in the parallel system when $|\beta| > 1$. This behavior stems from the difference between the sequential and parallel implementations of the CAVI. Looking closely at the update diagrams for the two systems reveals the key difference that produces these two-cycles. The decoupled sequential system is



and the decoupled parallel system is



The major difference between these diagrams is how the individual update sequences begin. Notice ζ_0 plays no role in updating the sequential system as both the ζ_k update sequence and the ξ_k update sequence are dependent only on the choice of ξ_0. Even after rewriting the sequential updates in terms of individual sequences the system is not truly decoupled as both sequences depend on a common starting point. This precisely prescribes the behavior that we see in the system relative to the sigmoid function dynamics in Theorem 1 and Theorem 2. Compare this to the parallel system. Here ζ_0 is involved in updating both the odd ξ_{2k+1} subsequence and the even ζ_{2k} subsequence. Furthermore, ξ_0 remains involved by controlling the updates for the even ξ_{2k} subsequence and the odd ζ_{2k+1} subsequence. This additional flexibility allows the parallel system to develop periodic behavior outside of the Dobrushin regime $(1 \leq \beta \leq 1)$.

As an example, we will consider initializing the sequential algorithm to the parallel algorithm for $\beta = 1.2$. We begin with the sequential algorithm. For $\beta = 1.2$, consider initializing the sequential system at $(\zeta_0, \xi_0) = (0.7, 0.3)$. The sequential system updates are fully determined by ξ_0, so for $\xi_0 = 0.3$ it follows from Theorem 1 that an application of the function (11) will cause $\zeta_1 \in W^s(c_0)$. At this point, the system can be evolved by applying (12) to the independent sequences for ζ and ξ as given in (13). The dynamics of the system are now controlled by the function (12). From this initialization the system will converge to the fixed point $(c_0, c_0) = (0.17071, 0.17071)$ as shown in Figure 6.

Contrast this with the behavior of the parallel system in which the updates are determined by both ξ_0 and ζ_0. For $\beta = 1.2$, consider initializing the parallel system at $(\zeta_0, \xi_0) = (0.7, 0.3)$. It follows from Theorem 1 that an application of the function (11) will cause $\zeta_1 \in W^s(c_0)$ and $\xi_1 \in W^s(c_1)$. Successive updates will cause the sequences ζ_k and ξ_k to flip back and forth between the domains $W^s(c_0)$ and $W^s(c_1)$, until the system settles into the two cycle $\mathcal{C}_1 = \{(c_0, c_1), (c_1, c_0)\} = \{(0.17071, 0.82928), (0.82928, 0.17071)\}$ as seen in Figure 10.

This simple example highlights the danger of naively parallelizing the CAVI algorithm. The convergence properties of a parallel version of the CAVI algorithm will heavily depend on the models CAVI update equations. In the case of the Ising model we have demonstrated that for

certain parameter regimes the parallel implementation of the algorithm can fail to converge due to the dependence of the algorithm on both ζ_0 and ξ_0.

6. Edward–Sokal Coupling

One method of improving convergence in Markov chains is through the use of probabilistic couplings. The Edward–Sokal (ES) coupling is a coupling of two statistical physics models, the random cluster model and the Potts model (a generalization of the Ising model) [37]. Running a Markov chain on the ES coupling leads to improved mixing properties compared to the equivalent Potts model and random cluster models [33]. Motivated by these findings in the Markov chain literature, we ask a similar question: Can the convergence properties of mean-field VI be improved by using the ES coupling in place of the Ising model? In this section we investigate this idea numerically. We first introduce the Edward–Sokal coupling following [37]. We introduce a variational family for the Edward–Sokal coupling and derive the variational updates for this model. Our findings suggests the variational updates converge to a unique solution in a larger range than the equivalent Dobrushin regime for the corresponding Ising measure.

6.1. Random Cluster Model

Let $G = (V, E)$ be a finite graph. Let $e = \langle x, y \rangle \in E$ denote an edge in G with endpoints $x, y \in V$. $\Sigma = \{1, 2, \ldots, q\}^V$, $\Omega = \{0, 1\}^E$ and $\mathcal{F}$ denotes the powerset of Ω. The random cluster model is a 2 parameter probability measure with an edge weight parameter $p \in [0, 1]$ and a cluster weight parameter $q \in \{2, 3, \ldots\}$ on $(\Omega, \mathcal{F})$ given by

$$\phi_{p,q}(\omega) \propto \left\{ \prod_{e \in E} p^{\omega(e)} (1-p)^{(1-\omega(e))} \right\} q^{\kappa(\omega)},$$

where $\kappa(\omega)$ denoted the number of connected components in the subgraph corresponding to ω. The partition function for the random cluster model is

$$\mathcal{Z}_{RC} = \sum_{\omega \in \Omega} \left\{ \prod_{e \in E} p^{\omega(e)} (1-p)^{(1-\omega(e))} \right\} q^{\kappa(\omega)}.$$

For $q = 2$ the the random cluster model reduces to the Ising model on G.

The Edward–Sokal Coupling is a probability measure μ on $\Sigma \times \Omega$ given by

$$\mu(\sigma, \omega) \propto \prod_{e \in E} \left[(1-p)\delta_{\omega(e),0} + p\delta_{\omega(e),1}\delta_e(\sigma) \right], \tag{21}$$

where $\delta_{a,b} = 1(a = b)$, and $\delta_e(\sigma) = 1(\sigma_x = \sigma_y)$, for $e = (x, y) \in E$.

It is well known that in the special case, $p = 1 - e^{-\beta}$ and $q = 2$ the Σ-marginal of the ES coupling is the Ising model, the Ω-marginal is the random cluster model [37]. We are interested in better understanding how the convergence of the CAVI algorithm on the ES coupling compares to the convergence of the CAVI algorithm on the Ising model.

6.2. VI Objective Function

To calculate the VI updates for each variable we may need to make use of the alternative characterization of the ES coupling

$$\mu(\sigma, \omega) \propto \psi(\sigma)\phi_{p,1}(\omega)1_F(\sigma, \omega)$$

where ψ is uniform measure on Σ and $\phi_{p,1}(\omega)$ is a product measure on Ω

$$\phi_{p,1}(\omega) = \prod_{e \in E} p^{\omega(e)}(1-p)^{(1-\omega(e))} \tag{22}$$

and

$$F = \{(\sigma, \omega) : \delta_\omega(e) = 1 \implies \delta_e(\sigma) = 1\} \tag{23}$$

The variational family that we will be optimizing over is

$$q(\sigma, \omega) = q_1(\sigma_1) q_2(\sigma_2) q_0(\omega) 1_F(\sigma, \omega). \tag{24}$$

We have added the indicator on the set F to eliminate the configurations (σ, ω) that are not well defined in the variational objective. We will use the convention that $0 \log(0) = 0$.

6.3. VI Updates

The ELBO that corresponds to the variational family (24) is

$$\begin{aligned}
\text{ELBO}(x_1, x_2, y, p) &= x_1 x_2 y \log(x_1 x_2 y) - x_1 x_2 y \log(1-p) \\
&+ (1-x_1) x_2 y \log((1-x_1) x_2 y) - (1-x_1) x_2 y \log(1-p) \\
&+ x_1 (1-x_2) y \log(x_1 (1-x_2) y) - x_1 (1-x_2) y \log(1-p) \\
&+ (1-x_1)(1-x_2) y \log((1-x_1)(1-x_2) y) - (1-x_1)(1-x_2) y \log(1-p) \\
&+ x_1 x_2 (1-y) \log(x_1 x_2 (1-y)) - x_1 x_2 (1-y) \log(p) \\
&+ (1-x_1)(1-x_2)(1-y) \log((1-x_1)(1-x_2)(1-y)) - (1-x_1)(1-x_2)(1-y) \log(p).
\end{aligned}$$

Taking the derivative with respect to x_1 and simplifying gives us

$$\begin{aligned}
\text{ELBO}_1(x_1, x_2, y, p) &= y \log\left(\frac{x_1}{1-x_1}\right) + (1-y) \log\left(\frac{1}{1-x_1}\right) \\
&+ x_2 (1-y) \log(x_1 (1-x_1)) + x_2 (1-y) \log\left(\frac{x_2 (1-x_2)(1-y)^2}{p^2}\right) \\
&+ \log\left(\frac{p}{(1-x_2)(1-y)}\right) + (2x_2 - 1)(1-y).
\end{aligned}$$

Taking the derivative with respect to x_2 and simplifying gives us

$$\begin{aligned}
\text{ELBO}_2(x_1, x_2, y, p) &= y \log\left(\frac{x_2}{1-x_2}\right) + (1-y) \log\left(\frac{1}{1-x_2}\right) \\
&+ x_1 (1-y) \log(x_2 (1-x_2)) + x_1 (1-y) \log\left(\frac{x_1 (1-x_1)(1-y)^2}{p^2}\right) \\
&+ \log\left(\frac{p}{(1-x_1)(1-y)}\right) + (2x_1 - 1)(1-y).
\end{aligned}$$

Taking the derivative with respect to y and simplifying gives us

$$\begin{aligned}
\text{ELBO}_y(x_1, x_2, y, p) &= x_1 x_2 \log\left(\frac{y}{1-y}\right) + x_1 x_2 \log\left(\frac{p}{1-p}\right) \\
&+ (1-x_1)(1-x_2) \log\left(\frac{y}{1-y}\right) + (1-x_1)(1-x_2) \log\left(\frac{p}{1-p}\right) \\
&+ (1-x_1) x_2 \log\left(\frac{(1-x_1) x_2 y}{1-p}\right) + x_1 (1-x_2) \log\left(\frac{x_1 (1-x_2) y}{1-p}\right) \\
&+ (1-x_1) x_2 + x_1 (1-x_2).
\end{aligned}$$

Absence of closed form updates for any of the variables limits our ability to study the convergence of the system with classical dynamical systems techniques. Instead we look at the long evolution

behavior of the system by plotting 100 iterations of the CAVI updates which are generated from the following system

$$\begin{aligned} x_1(t+1) &= \text{argmin}_{z\in(0,1)}|\text{ELBO}_1(z, x_2(t), y(t), p)|, \\ x_2(t+1) &= \text{argmin}_{z\in(0,1)}|\text{ELBO}_2(x_1(t+1), z, y(t), p)|, \\ y(t+1) &= \text{argmin}_{z\in(0,1)}|\text{ELBO}_y(x_1(t+1), x_2(t+1), z, p)|. \end{aligned}$$

We generate the argmin of the free variable z from a line search with a step size of $\Delta = 10^{-6}$. Running these simulations we find that the iterations of $x_1(t), x_2(t), y(t)$ converge to a global solution within about $T = 20$ time steps from any initialization $x_1(0), x_2(0), y(0) \in (0,1)$ and any $\beta > 0$. It is evident that using the ES coupling, we get global convergence of the algorithm outside of the Dobrushin regime of the corresponding paramagnetic Ising model. The figures depicting the simulation results of convergence of the variational inference algorithm in the Edward–Sokal coupling can be found below in Figures 13–16.

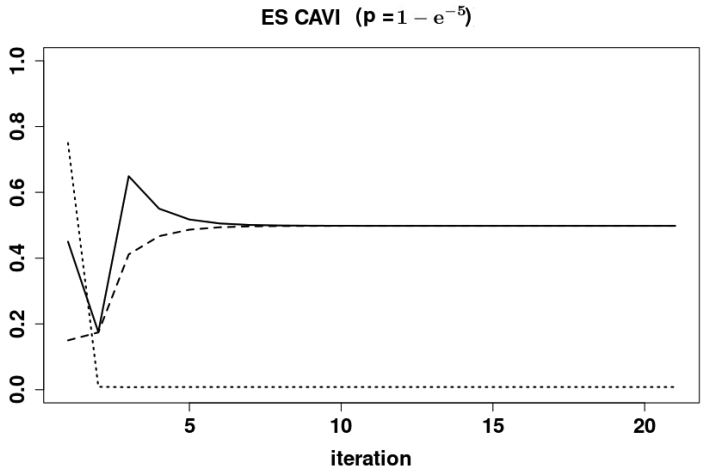


Figure 13. A plot of the 20 iterations of the ES updates for $p = 1 - e^{-5}$ from a uniformly random initialization. Each of the lines represents a different parameter. The solid line is x_1, the dashed line is x_2 and the dotted line is y. We see convergence to a unique fixed point for each of the variables.

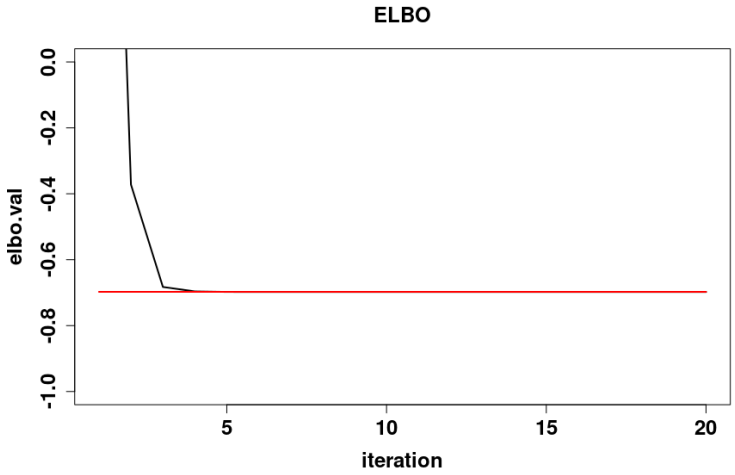


Figure 14. A plot of the ELBO of the ES coupling for $p = 1 - e^{-5}$. The red line denotes the global minimum ELBO value.

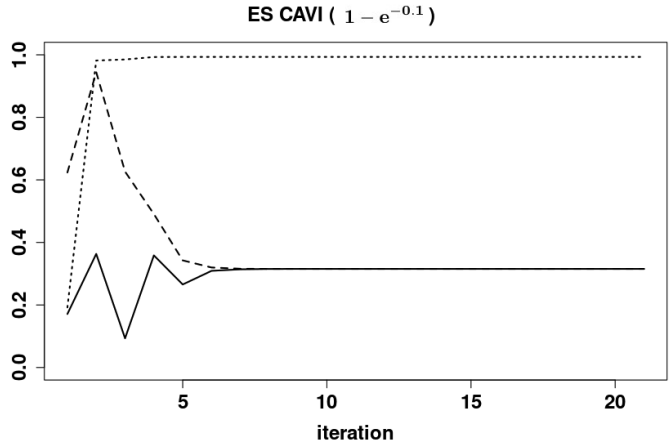


Figure 15. A plot of the 20 iterations of the ES updates for $p = 1 - e^{-0.1}$ from a uniformly random initialization. Each of the lines represents a different parameter. The solid line is x_1, the dashed line is x_2 and the dotted line is y. We see convergence to a unique fixed point.

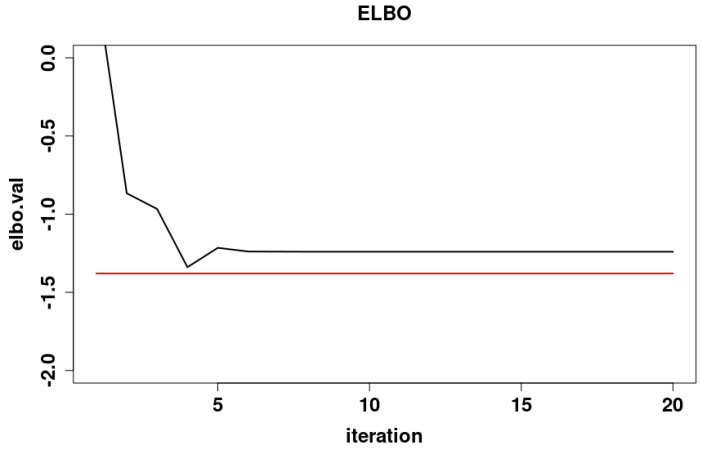


Figure 16. A plot of the ELBO of the ES coupling for $p = 1 - e^{-0.1}$. The red line denotes the global minimum ELBO value.

7. Conclusions

This paper demonstrates the use of classical dynamical systems and bifurcation theory to study the convergence properties of the CAVI algorithm of the Ising model on two nodes. In our simple model we are able to provide the complete dynamical behavior for the Ising model on two nodes. Interestingly, we find that the sequential CAVI algorithm and parallelized CAVI algorithm are not topologically conjugate owing to the presence of periodic behavior in the parallelized CAVI. This behavior originates from the added flexibility of the initialization in the parallelized CAVI when compared to the sequential CAVI. The erratic behavior we see in the Ising model for $|\beta| > 1$ is due to a combination of the existence of multiple fixed points of the systems update function and the instability of these fixed points. In this parameter regime, the fixed point that produces the optimal solution (0.5, 0.5) is a repelling fixed point. Unless we initialize the algorithm exactly at (0.5, 0.5), the CAVI system cannot converge to this point. The other two suboptimal fixed points are both asymptotically stable. This suggests that the main problem that the CAVI algorithm experiences is centered around the existence of multiple fixed points. Recent work on stochastic block models (SBM) and topic models (TM) models shows that mean field VI leads to suboptimal estimators [17–20]. It is not clear if this property comes from the mean field variational inferences construction using product distributions or if this is a consequence of structure among latent variables. A minor difference of the stochastic block model (SBM) or topic

model (TM) with the Ising model is that the former contain parameters (e.g., the cluster labels) that are identifiable only up to permutations. That being said, in the SBM or TM, if the cluster means are not well-separated, then it is not possible to identify the labels even up to permutations. This is somewhat related to having multiple fixed points of the objective function and we conjecture similar behavior to what we have found in the Ising model will be exhibited in the SBM or TM outside the Dobrushin regime. Interestingly, a close look at the BCAVI updates in [17,18] reveals a similar sigmoid update function $1/(1+e^{-x})$. Applying the tools and techniques from dynamical systems theory to study the CAVI algorithm in the SBM, TM and other models will provide a better understanding of the issues that come with using mean field variational inference and is important to developing better variational inference techniques.

Most of the research into the theoretical properties of variational inference has focused on the mean field family due to its computational simplicity. This computational simplicity comes at the cost of limited expressive power. Can we make due with this limited expressive power in practical applications? More specifically, is there an equivalent parameter regime to the Dobrushin regime ($1 \leq \beta \leq 1$) for other similar models like the SBM and TM inside which the CAVI produces statistically optimal estimators? The answer to this question provides researchers with stable parameter regimes for the model. The non-existenceof such a region would indicate the need for more expressive variational methods for the model beyond mean field methods. Recent work [19,20] suggests that this adding some structure to algorithms may fix the problems that arise from mean field VI. How much structure is needed to recover statistically optimal estimators? Could adding in a simple structure of pair-wise dependence to the mean field VI in the Ising model, similarly to [19], be enough to recover the optimal estimator outside of the Dobrushin regime? Is the amount of additional structure that is needed somehow related to the latent structure of the models? Tools from dynamical systems theory can be used to study these questions.

Using dynamical systems to study the convergence properties of the CAVI algorithm is not without its challenges. While dynamical systems theory can provide the answers to many of the above questions, applying these tools to higher dimensional sequential systems is a challenging problem. As mentioned previously, the general theory for n-dimensional discrete dynamical systems is dependent on writing the evolution function in the form $x_{n+1} = F(x_n)$. Deriving this F is typically not possible for densely connected higher dimensional sequential systems like the n-dimensional Ising model CAVI. This is not the only challenging aspect to the problem. These systems typically possess multiple fixed points which can only be found numerically. Multiple fixed points will lead to more complicated partitions of the space into domains of attraction. Furthermore, higher dimensional systems can possess bifurcations of multiple codimensions, which as significantly more difficult to study. Bifurcations of codimension 3 are so exotic that they are not well studied [23,24]. Software to handle such calculations has only recently been developed [24]. In practical terms this means that the convergence properties can only be studied numerically for models with a small number of parameters. Furthermore, most of the numerical techniques work under the assumption of differentiability of the evolution operator and will fail to be applicable to many systems of practical interest in statistics such as the Edward–Sokal CAVI. Applying tools from dynamical systems to the study of variational inference algorithms will require developing new theory for high dimensional and well connected sequential dynamical systems.

Author Contributions: Conceptualization, S.P., D.P. and A.B.; formal analysis, S.P.; supervision, D.P. and A.B.; writing—original draft, S.P., D.P. and A.B. All authors have read and agreed to the published version of the manuscript.

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

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. An Overview of One Dimensional Dynamical Systems

The main focus of discrete dynamical systems is the asymptotic behavior of iterated systems (8). Bifurcation theory studies how the dynamical behavior of a system changes as the parameter J_{12} changes. We study the behavior of convergence of the CAVI algorithm by studying the autonomous discrete time dynamical system formed by the update Equation (8). This allows us to utilize tools from dynamical systems theory to study the behavior of the algorithm with respect to its parameters. In this section we provide a brief overview of the necessary dynamical systems and bifurcation theory in dimension 1 used in Section 5.

Appendix A.1. Notation

Our focus will be on parametric dynamical systems defined by a functions $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$. We will call elements $\mathbf{x} \in \mathbb{R}^n$ elements in the state space (phase space) and elements $\alpha \in \mathbb{R}^p$ as parameters. We denote real numbers $x \in \mathbb{R}$ and real vectors in $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ with bold. We denote the inverse of an invertible function f by f^{-1}. The k-fold composition of a function f with itself at a point $(\mathbf{x}, \alpha)$ will be denoted by $f^k(\mathbf{x}, \alpha)$. The k-fold composition of the inverse function f^{-1} will be denoted f^{-k}. The identity function will be denoted id. We use the convention $f^0 = \text{id}$. We denote the tensors of derivatives of f by $f_{\mathbf{x}}(\mathbf{x}, \alpha) = (\partial f_i / \partial x_j)$, $f_{\mathbf{xx}}(\mathbf{x}, \alpha) = (\partial^2 f_i / \partial x_j \partial x_k)$, $f_{\mathbf{xx}}(\mathbf{x}, \alpha) = (\partial^2 f_i / \partial x_j \partial x_k)$, $f_{\mathbf{xxx}}(\mathbf{x}, \alpha) = (\partial^3 f_i / \partial x_j \partial x_k \partial x_\ell)$, $f_\alpha(\mathbf{x}, \alpha) = (\partial f_i / \partial \alpha_j)$.

Appendix A.2. Dynamical Systems

Dynamical systems is a classical approach to studying the convergence properties of non-linear iterative systems. These systems can be continuous in time, for example a differential equation, or discrete in time, for example iterations of a function from an initial point. A dynamical system is called autonomous if the function governing the system is independent of time and non-autonomous otherwise. The coordinate ascent variational inference for the Ising model is a discrete-time autonomous dynamical system. Before giving a complete proof of the dynamical properties of the CAVI algorithm for the Ising model in dimension 2, we first give a basic introduction to the theory of discrete time dynamical systems and bifurcations following [23–25,38].

Formally, a dynamical system is triple $\{T, X, \phi^t\}$ where T is a time set, X is the state space and $\phi^t : X \to X$ is a family of evolution operators parameterized by $t \in T$ satisfying $\phi^0 = id$ and $\phi^{s+t} = \phi^t \circ \phi^s$ for all $x \in X$. For a discrete time system the evolution operator is fully specified by the one-step map $\phi^1 = f$, since the composition rule then defines $\phi^k = f^k$ for $k \in \mathbb{Z}$. We restrict the further discussion to discrete time dynamical systems defined by the one-step map

$$\mathbf{x} \mapsto f(\mathbf{x}, \alpha), \quad \mathbf{x} \in \mathbb{R}^n, \alpha \in \mathbb{R}^p, \tag{A1}$$

where f is a diffeomorphism, a smooth function with smooth inverse, of the state space $\mathbb{R}^n$ and α are the parameters of the system.

The basic geometric objects of a dynamical system are orbits in the state space and the phase portrait, defined as follows. The phase portrait is the partition of the state space induced by the orbits. The orbit starting at a point $\mathbf{x}$ is an ordered subset of the state space $\mathbb{R}^n$ denoted $\text{orb}(\mathbf{x}) = \{f^k(\mathbf{x}) : k \in \mathbb{Z}\}$. There are two special types of orbits, fixed points and cycles, defined below.

A fixed point $\mathbf{x}_*$ of the system are points that remain fixed under the evolution of the system, ones that satisfies $\mathbf{x}_* = f(\mathbf{x}_*)$. We can classify fixed points of the system by studying the local behavior of the system near the fixed point. To do this we consider small perturbations of the system near the fixed point. A fixed point $\mathbf{x}_*$ is said to be locally stable if points that are near the fixed point do not move too far away from the fixed point as the system evolves. Formally, if for any $\varepsilon > 0$ there exists $\delta > 0$ such that for all x with $|\mathbf{x} - \mathbf{x}_*| < \delta$ we have $|f^k(\mathbf{x}) - \mathbf{x}_*| < \varepsilon$ for all $k > 0$. A fixed point is called semi-stable from the right if for any $\varepsilon > 0$ there exists $\delta > 0$ such that for all x with $0 < \mathbf{x} - \mathbf{x}_* < \delta$ we have $|f^k(\mathbf{x}) - \mathbf{x}_*| < \varepsilon$ for all $k > 0$ (semi-stable from the left is defined analogously). It is said to be

locally unstable otherwise. A fixed point $\mathbf{x}_*$ is locally attracting if all points in a small neighborhood converge to the fixed point as we let the system evolve. Formally, if there exists an $\eta > 0$ such that $|\mathbf{x} - \mathbf{x}_*| < \eta$ implies $f^n(\mathbf{x}) \to \mathbf{x}_*$ as $n \to \infty$. A fixed point $\mathbf{x}_*$ is locally asymptotically stable if it is both locally stable and attracting. A fixed point $\mathbf{x}_*$ is locally semi-asymptotically stable from the right if it is both locally semi-stable from the right and $\lim_n f^n(x) = x_*$ for $0 < \mathbf{x} - \mathbf{x}_* < \eta$ for some η. It is globally asymptotically stable if the point is attracting for all $\mathbf{x}$ in the state space.

A cycle is a periodic orbit of distinct points $C = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{K-1}\}$, where $\mathbf{x}_0 = f(\mathbf{x}_{K-1})$ for some $K > 0$. The minimal K generating the cycle is called the period of the cycle. A subset $S \subset \mathbb{R}^n$ is called invariant if $f^k(S) \subset S, \quad k \in \mathbb{Z}$. An invariant set S is called asymptotically stable if there exists a neighborhood U of S such that for any point in U is eventually inside the set S. The stable set of $S \subset \mathbb{R}^n$ is $W^s(S) = \left\{\mathbf{x} \in \mathbb{R}^n : \lim_{k\to\infty} f^k(\mathbf{x}) \in S\right\}$. If f is invertible, we define the unstable set of $S \subset \mathbb{R}^n$ is $W^u(S) = \left\{\mathbf{x} \in \mathbb{R}^n : \lim_{k\to\infty} f^{-k}(\mathbf{x}) \in S\right\}$. The unstable set of S for the forward system $f^k, k > 0$ is the stable set of S for the backward system $f^{-k}, k > 0$. It is possible to study the behavior of points that diverge by studying points that converge under the inverse map. We can also classify the stability of K-cycles. We classify the stability of the cycle as a fixed point in the map f^K.

Consider a discrete time dynamical system defined by a diffeomorphism $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Let x_* be a fixed point of $f(x, \alpha)$ and consider a nearby point x, $|x - x_*| = \epsilon$. Taking a Taylor expansion of the system about the fixed point gives us

$$f(x, \alpha) - x_* \quad = \quad f_x(x_*, \alpha)(x - x_*) + f_{xx}(x_*, \alpha)(x - x_*)^2 + O(|x - x_*|^3).$$

If the Jacobian does not have modulus one and ϵ is small enough, then the contribution by the terms of $O(|x - x_*|^2)$ will be negligible, in which case the behavior of the system is governed by the the behavior of the linearization of the system $f_x(x_*, \alpha)$. We now introduce the idea of a hyperbolic fixed point. Assume that the Jacobian $A := f_x(x_*, \alpha)$ of the system (A1) at a fixed point x_* is non-singular. The fixed point x_* is called hyperbolic if $|f_x(x_*, \alpha)| \neq 1$ and non-hyperbolic if $|f_x(x_*, \alpha)| = 1$. The notion of hyperbolic fixed and non-hyperbolic fixed points generalizes to higher dimensions where it involves the eigenvalues of the Jacobian; see [23,25,38] for more details.

Near a hyperbolic fixed point a non-linear dynamical system behaves its first order Taylor approximation (also known as the linearization of the system). To make this argument rigorous we need to discuss what it means for two dynamical systems to be equivalent. Two systems are topologically equivalent if we can map orbits of one system to orbits of another system in a continuous way that preserves the order of time. The dynamical system (A1) is called topologically equivalent to the system

$$\mathbf{y} \mapsto g(\mathbf{y}, \beta), \quad \mathbf{y} \in \mathbb{R}^n, \quad \beta \in \mathbb{R}^p, \tag{A2}$$

if there exists a homeomorphism of the parameter space $h_p : \mathbb{R}^p \to \mathbb{R}^p$, $\beta = h_p(\alpha)$ and a parameter dependent state space homeomorphism, continuous in the first argument, $h : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ such that, $\mathbf{y} = h(\mathbf{x}, \alpha)$, mapping orbits of the system (A1) at parameter value α onto orbits of the system (A2) at parameter $\beta = h_p(\alpha)$ preserving the direction of time. If h is a diffeomorphism then the systems are called smoothly equivalent.

Let (A1) and (A2) be two topologically equivalent invertible dynamical systems. Consider the orbit of the system under the mapping $f(\mathbf{x}, \alpha)$, $\text{orb}(\mathbf{x}; f, \alpha)$ and the orbit of the system $g(\mathbf{y}, \beta)$, $\text{orb}(\mathbf{y}; g, \beta)$. Topological equivalence means that the homeomorphism $(h(\mathbf{x}, \alpha), h_p(\alpha))$ maps $\text{orb}(\mathbf{x}; f, \alpha)$ to $\text{orb}(\mathbf{y}; g, \beta)$ preserving the order of time. This gives us the following commutative diagram

$$\begin{array}{ccccccccc} \cdots & \xrightarrow{f} & f^{-1}(\mathbf{x},\alpha) & \xrightarrow{f} & \mathbf{x} & \xrightarrow{f} & f(\mathbf{x},\alpha) & \xrightarrow{f} & \cdots \\ & & \downarrow h & & \downarrow h & & \downarrow h & & \\ \cdots & \xrightarrow{g} & g^{-1}(\mathbf{y},\beta) & \xrightarrow{g} & \mathbf{y} & \xrightarrow{g} & g(\mathbf{y},\beta) & \xrightarrow{g} & \cdots . \end{array}$$

The orbits being topologically equivalent means that orbit $\mathbf{x}$ under the mapping h should produce the same orbit as mapping $\mathbf{x}$ to $\mathbf{y} = h(\mathbf{x},\alpha)$ computing the orbit of $\mathbf{y}$ under $g(\cdot,\beta)$ and mapping back to $f(\mathbf{x},\alpha)$ by h^{-1}, $f(\mathbf{x},\alpha) = h^{-1} \circ g \circ h(\mathbf{x},\alpha)$. We shall primarily be interested in the behavior of the system in a small neighborhood of an equilibrium point. A system (A1) is called locally topologically equivalent near an equilibrium $\mathbf{x}_*$ to a system (A2) near an equilibrium $\mathbf{y}_*$ if there exists a homeomorphism $h : \mathbb{R}^n \to \mathbb{R}^n$ defined in a small neighborhood U of $\mathbf{x}_*$ with $\mathbf{y}_* = h(\mathbf{x}_*)$ that maps orbits of (A1) in U onto orbits of (A2) in $V = h(U)$, preserving the direction of time.

We now have enough terminology to introduce the following theorem, which shows that the dynamics of a smooth system in the neighborhood of a hyperbolic fixed point are equivalent to the dynamics of the linearization of the system,

Theorem A1 (Grobman–Hartman)**.** *Consider a smooth map*

$$\boldsymbol{x} \mapsto A\boldsymbol{x} + F(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^n, \tag{A3}$$

where A is an $n \times n$ matrix and $F(\boldsymbol{x}) = O(\|\boldsymbol{x}\|^2)$. If $\boldsymbol{x}_* = 0$ is a hyperbolic fixed point of (A3), then (A3) is topologically equivalent near this point to its linearization

$$\boldsymbol{x} \mapsto A\boldsymbol{x}, \quad \boldsymbol{x} \in \mathbb{R}^n.$$

Note Theorem A1 is true for a general n-dimensional system. Theorem A1 provides sufficient conditions to determine the stability of a hyperbolic fixed point of a general discrete time system,

Theorem A2. *Consider a discrete time dynamical systems (A1) where f is a smooth map. Suppose for a fixed point x_* that the eigenvalues of Jacobian $f_x(x_*,\alpha)$ all satisfy $|\lambda| < 1$ then the fixed point is stable. Alternatively, suppose for a fixed point x_* that the eigenvalues of Jacobian $f_x(x_*,\alpha)$ all satisfy $|\lambda| > 1$ then the fixed point is unstable.*

The linearization of the system near a non-hyperbolic fixed point is not sufficient to determine stability of the fixed point and we need to investigate higher order terms. The following theorem provides sufficient condition to check the stability of a smooth one dimensional system at a non-hyperbolic fixed point,

Theorem A3. *Let $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Suppose that $f(\cdot,\alpha) \in C^3(\mathbb{R};\mathbb{R})$ and x_* is a non-hyperbolic fixed point of f, $x_* = f(x_*,\alpha)$. We have the following cases:*

Case 1: If $f_x(x_*,\alpha) = 1$, then

1. *If $f_{xx}(x_*,\alpha) \neq 0$ then x_* is semi-asymptotically stable from the left if $f_{xx}(x_*,\alpha) > 0$ and semi-asymptotically stable from the right if $f_{xx}(x_*,\alpha) < 0$;*
2. *if $f_{xx}(x_*,\alpha) = 0$ and $f_{xxx}(x_*,\alpha) < 0$ then x_* is asymptotically stable;*
3. *if $f_{xx}(x_*,\alpha) = 0$ and $f_{xxx}(x_*,\alpha) > 0$ then x_* is unstable.*

Case 2: If $f_x(x_*,\alpha) = -1$, then

1. *If $\mathcal{S}f(x_*,\alpha) < 0$, then x_* is asymptotically stable;*

2. *If* $\mathcal{S}f(x_*;,\alpha) > 0$, *then* x_* *is unstable.*

where $\mathcal{S}f(x)$ *is the Schwarzian derivative of* f

$$\mathcal{S}f(x,\alpha) = \frac{f_{xxx}(x,\alpha)}{f_x(x,\alpha)} - \frac{3}{2}\left[\frac{f_{xx}(x,\alpha)}{f_x(x,\alpha)}\right]^2.$$

The Schwarzian derivative controls the higher order behavior in oscillatory systems.

Appendix A.3. Codimension 1 Bifurcations

Until now we have kept the parameter of the system fixed. The study of the change in behavior of a dynamical system as the parameters are varied is called bifurcation theory. A bifurcation occurs when the dynamics of the system at a parameter value α_1 differ from the dynamics of the system at a different parameter value α_2. Changing the parameter in a system may cause a stable fixed point to become unstable, the fixed point may split into multiple fixed points, or a new orbit may form. Each of these is an example of a bifurcation, although these are not the only things that can happen. The point at which a bifurcation occurs is called a bifurcation point. More formally, the parameter α_* is called a bifurcation point if arbitrarily close to it there is α such that $\mathbf{x} \mapsto f(\mathbf{x},\alpha), \mathbf{x} \in \mathbb{R}^n$ is not topologically equivalent to $\mathbf{x} \mapsto f(\mathbf{x},\alpha_*), \mathbf{x} \in \mathbb{R}^n$ in some domain $U \subset \mathbb{R}^n$.

A necessary, but not sufficient condition for bifurcation of a fixed point to occur is for the fixed point to be nonhyperbolic. Theorem A1 together with the implicit function theorem show that in a sufficiently small neighborhood of a hyperbolic fixed point $(\mathbf{x}_*,\alpha_*)$, for each α there is another unique fixed point with the same stability properties as $(\mathbf{x}_*,\alpha)$. So hyperbolic fixed points do not undergo local bifurcations. In the context of discrete systems, a local bifurcation can occur only at a fixed point $(\mathbf{x}_*,\alpha_*)$ when the Jacobian of the system at $(\mathbf{x}_*,\alpha_*)$ has an eigenvalue with modulus one.

Perhaps surprisingly, there are only three types of generic bifurcations that can happen in a discrete system with one parameter. They are the limit point (LP), period doubling (PD) and Neimark–Sacker (NS) bifurcations. The reason for this is fairly simple. It turns out that there is a generic system, called the topological normal form, which undergoes this bifurcation at the origin in the $(\mathbf{x},\alpha)$-plane. For any other system that undergoes the same bifurcation and satisfies certain non-degeneracy conditions there is a local change of coordinates that transforms the system into the topological normal form.

In general the types of bifurcations that can occur are connected to the number of parameters in the system. The minimal number of parameters that must be changed in order for a particular bifurcation to occur in $f(\mathbf{x},\alpha)$ is called the codimension of the bifurcation. A bifurcation is called local if it can be detected in any small neighborhood of the fixed point, otherwise its called global. Global bifurcations are much harder to analyze and since we do not attempt to investigate them in this paper we will not expand upon them further. More detailed results on bifurcations in codimension 1 and 2 can be found in [23,24].

We will now formally define the sufficient conditions for a system to undergo a period doubling or a pitchfork bifurcation. The period doubling bifurcation occurs when a system with a non-hyperbolic fixed point with multiplier $\lambda_1 = -1$ satisfies certain non-degeneracy conditions. There are two types of PD bifurcations. In the super-critical case, a stable 2-cycle is generated when a fixed point becomes unstable. In the sub-critical case, a stable fixed point turns unstable when it coalesces with an unstable 2-cycle (This is true for a general k-cycle. In the super-critical case, a stable $2k$-cycle is generated when a k-cycle becomes unstable. In the sub-critical case, a stable k-cycle turns unstable when it coalesces with an unstable $2k$-cycle). The conditions for a PD bifurcation to occur are given as follows

Theorem A4 (Period Doubling Bifurcation)**.** *Suppose That A One-Dimensional System*

$$x \mapsto f(x,\alpha), \quad x,\alpha \in \mathbb{R},$$

with smooth f, has at $\alpha = 0$ the fixed point $x_* = 0$, and let $\lambda = f_x(0,0) = -1$. Assume the following non-degeneracy conditions are satisfied

1. $1/2(f_{xx}(0,0))^2 + 1/3 f_{xxx}(0,0) \neq 0$
2. $f_{x\alpha}(0,0) \neq 0$

Then there are smooth invertible coordinate and parameter changes transforming the system into

$$\eta \mapsto -(1+\beta) \pm \eta^3 + O(\eta^4). \tag{A4}$$

An classical example of a period doubling bifurcation can be seen in the logistic map $f(x,\mu) = \mu x(1-x)$, for $x \in [0,1]$. The bifurcation occurs at the point $(x_*, \mu_*) = (2/3, 3)$. The logistic map has two fixed points. One fixed point is at $x = 0$ and the other is at $x = (\mu - 1)/\mu$. We will ignore the fixed point at $x = 0$ since it is repelling for $\mu > 1$. We look at the behavior of the system in a small neighborhood of $\mu_* = 3$. For $\mu = 2.9$, the fixed point $x_* = (\mu - 1)/\mu$ is a hyperbolic attracting fixed point since $|f_x(x_*, 2.9)| = |2 - \mu| < 1$. For $\mu = 3$ the fixed point $x_* = (\mu - 1)/\mu$ is a non-hyperbolic fixed point since $f_x(x_*, 2.9) = 2 - \mu = -1$. Checking the Schwarzian derivative shows that the fixed point is asymptotically stable. For $\mu = 3.1$, $x_* = (\mu - 1)/\mu$ becomes a repelling fixed point. The points in $(0, x_*) \cup (x_*, 1)$ converge to the attracting 2-cycle $C = \{0.558014, 0.7645665\}$. A super-critical period doubling bifurcation has occurred in the system formed by the logistic map. As the parameter μ increases we see a stable fixed point degenerate and a stable 2-cycle is formed.

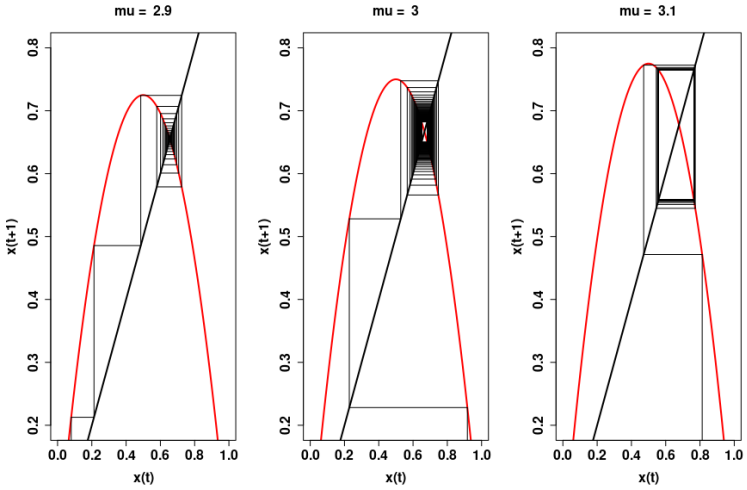


Figure A1. The above plots are cobweb diagrams for the logistic map $f(x,\mu) = \mu x(1-x)$, for $x \in [0,1]$, with parameters $\mu = 2.9$, $\mu = 3$ and $\mu = 3.1$, respectively. For $\mu = 2.9$ the system has one stable fixed point $x_* = (\mu - 1)/\mu$. For $\mu = 3$, the system has one non-hyperbolic fixed point $x_* = (\mu - 1)/\mu$ which is asymptotically stable attracting; the plot was not iterated long enough to see convergence. For $\mu = 3.1$, the system has a hyperbolic repelling fixed point $x_* = (\mu - 1)/\mu$ and an asymptotically stable attracting two cycle $C = \{0.558014, 0.7645665\}$.

The second iterate of a map that undergoes a PD bifurcation undergoes a bifurcation know as the pitchfork bifurcation. A system that undergoes a super-critical pitchfork bifurcation when a stable fixed point becomes unstable and two stable fixed points appear in the system. A system that undergoes a sub-critical pitchfork bifurcation when two stable fixed points coalesce with an unstable fixed point, the unstable fixed point becomes stable as the parameter crosses the bifurcation point. Below we present extra details pertaining to the period doubling bifurcation and its relation to the pitchfork bifurcation.

Consider the one-dimensional system

$$x \mapsto -(1+\alpha)x + x^3 = f(x,\alpha).$$

The map $f(x,\alpha)$ is invertible in a small neighborhood of $(0,0)$. The system has a fixed point at $x_* = 0$ for all α, with eigenvalue $-(1+\alpha)$. For small $\alpha < 0$ the fixed point is hyperbolic stable and for $\alpha > 0$ is it hyperbolic unstable. For $\alpha = 0$ the fixed point is non-hyperbolic, but is asymptotically stable.

Consider the second iterate of $f(x,\alpha)$

$$\begin{aligned} f^2(x,\alpha) &= -(1+\alpha)f(x,\alpha) + (f(x,\alpha))^3 \\ &= (1+\alpha)^2 x - \left[(1+\alpha)(2+2\alpha+\alpha^2)\right] x^3 + O(x^5). \end{aligned}$$

The second iterate has a trivial fixed point at $x_* = 0$ and for $\alpha > 0$ it has two non-trivial stable fixed points $x_1 = (\sqrt{\alpha} + O(\alpha))$, $x_1 = -(\sqrt{\alpha} + O(\alpha))$ that form a two cycle

$$x_2 = f(x_1,\alpha), \quad x_1 = f(x_2,\alpha).$$

The conditions for a generic pitchfork bifurcation can be found in [25]

Theorem A5 (Pitchfork Bifurcation)**.** *For A System*

$$x \mapsto f(x,\alpha), \quad x,\alpha \in \mathbb{R}$$

having non-hyperbolic fixed point at $x_* = 0$, $\alpha_* = 0$ *with* $f_x(0,0) = 1$ *undergoes a pitchfork bifurcation at* $(x_*,\alpha_*) = (0,0)$ *if*

$$f_\alpha(0,0) = 0, \quad f_{xx}(0,0) = 0, \quad f_{xxx}(0,0) \neq 0, \quad f_{x\alpha}(0,0) \neq 0.$$

A pitchfork bifurcation is super-critical if $-f_{xxx}(x_*,\alpha_*)/f_{\alpha x}(x_*,\alpha_*) > 0$ *and sub-critical if* $-f_{xxx}(x_*,\alpha_*)/f_{\alpha x}(x_*,\alpha_*) < 0$

An example of a pitchfork bifurcation can be seen in the second iteration of the logistic map $f^2(x,\mu) = \mu^2 x(1-x)(1-\mu x(1-x))$, for $x \in [0,1]$. The bifurcation occurs at the point $(x_*,\mu_*) = (2/3,3)$. For $\mu \leq 3$, the second iteration of the logistic map has the same fixed points as the first iteration. One fixed point is at $x = 0$ and the other is at $x = (\mu-1)/\mu$. We will ignore the fixed point at $x = 0$ since it is repelling for $\mu > 1$. We look at the behavior of the system in a small neighborhood of $\mu_* = 3$. For $\mu = 2.9$, the fixed point $x_* = (\mu-1)/\mu$ is a hyperbolic attracting fixed point since $|f_x^2(x_*,2.9)| < 1$. For $\mu = 3$ the fixed point $x_* = (\mu-1)/\mu$ is non-hyperbolic since $f_x^2(x_*,2.9) = 2-\mu = 1$. Checking the higher order derivative shows that the fixed point is asymptotically stable. For $\mu = 3.1$, $x_* = (\mu-1)/\mu$ becomes a repelling fixed point. Using numerical methods we find two additional fixed points, $x_1 = 0.558014$ and $x_2 = 0.7645665$, both of which are attracting. A super-critical pitchfork bifurcation has occurred in the system formed by the logistic map. As the parameter μ increases we see a stable fixed point degenerates to an unstable fixed point and two stable fixed points.

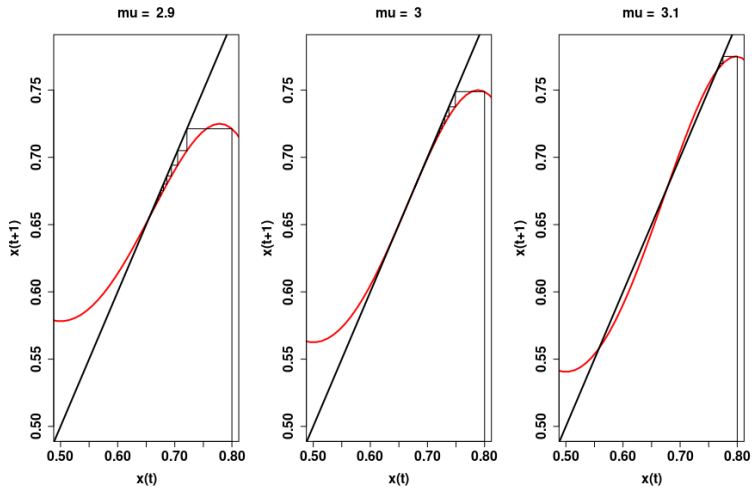


Figure A2. The above plots are cobweb diagrams for the second iterate of the logistic map $f(x,\mu) = \mu x(1-x)$, for $x \in [0,1]$, with parameters $\mu = 2.9$ and $\mu = 3.1$, respectively. For $\mu = 2.9$ the system has one stable fixed point $x_* = (\mu-1)/\mu$. For $\mu = 3.1$, the system has a hyperbolic repelling fixed point $x_* = (\mu-1)/\mu$ and two asymptotically stable attracting fixed points $x_1 = 0.0558014$ and $x_2 = 0.7645665$.

References

1. Bishop, C. *Pattern Recognition and Machine Learning*; Information Science and Statistics; Springer: Berlin/Heidelberger, Germany, 2006.
2. MacKay, D.J.; Mac Kay, D.J. *Information Theory, Inference and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
3. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
4. Zhang, C.; Bütepage, J.; Kjellström, H.; Mandt, S. Advances in variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2008–2026. [CrossRef] [PubMed]
5. Parisi, G. *Statistical Field Theory*; Frontiers in Physics; Addison-Wesley: Boston, MA, USA, 1988.
6. Opper, M.; Saad, D. *Advanced Mean Field Methods: Theory and Practice*; MIT Press: Cambridge, MA, USA, 2001.
7. Gabrié, M. Mean-field inference methods for neural networks. *J. Phys. A Math. Theor.* **2020**, *53*, 223002. [CrossRef]
8. Alquier, P.; Ridgway, J.; Chopin, N. On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.* **2016**, *17*, 1–41.
9. Pati, D.; Bhattacharya, A.; Yang, Y. On statistical optimality of variational Bayes. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Canary Islands, Spain, 9–11 April 2018; pp. 1579–1588.
10. Yang, Y.; Pati, D.; Bhattacharya, A. α-Variational inference with statistical guarantees. *Ann. Stat.* **2020**, *48*, 886–905. [CrossRef]
11. Chérief-Abdellatif, B.E.; Alquier, P. Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electron. J. Stat.* **2018**, *12*, 2995–3035. [CrossRef]
12. Wang, Y.; Blei, D.M. Frequentist consistency of variational Bayes. *J. Am. Stat. Assoc.* **2019**, *114*, 1147–1161. [CrossRef]
13. Wang, Y.; Blei, D.M. Variational Bayes under Model Misspecification. *arXiv* **2019**, arXiv:1905.10859.
14. Wang, B.; Titterington, D. *Inadequacy of Interval Estimates Corresponding to Variational Bayesian Approximations*; AISTATS; Citeseer: Princeton, NJ, USA, 2005.
15. Wang, B.; Titterington, D. Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Anal.* **2006**, *1*, 625–650. [CrossRef]
16. Zhang, A.Y.; Zhou, H.H. Theoretical and Computational Guarantees of Mean Field Variational Inference for Community Detection. *arXiv* **2017**, arXiv:math.ST/1710.11268.

17. Mukherjee, S.S.; Sarkar, P.; Wang, Y.R.; Yan, B. Mean field for the stochastic blockmodel: Optimization landscape and convergence issues. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2018; pp. 10694–10704.
18. Sarkar, P.; Wang, Y.; Mukherjee, S.S. When random initializations help: A study of variational inference for community detection. *arXiv* **2019**, arXiv:1905.06661.
19. Yin, M.; Wang, Y.X.R.; Sarkar, P. A Theoretical Case Study of Structured Variational Inference for Community Detection. *Proc. Mach. Learn. Res.* **2020**, *108*, 3750–3761.
20. Ghorbani, B.; Javadi, H.; Montanari, A. An Instability in Variational Inference for Topic Models. *arXiv* **2018**, arXiv:stat.ML/1802.00568.
21. Jain, V.; Koehler, F.; Mossel, E. The Mean-Field Approximation: Information Inequalities, Algorithms, and Complexity. *arXiv* **2018**, arXiv:cs.LG/1802.06126.
22. Koehler, F. Fast Convergence of Belief Propagation to Global Optima: Beyond Correlation Decay. *arXiv* **2019**, arXiv:cs.LG/1905.09992.
23. Kuznetsov, Y. *Elements of Applied Bifurcation Theory*; Applied Mathematical Sciences; Springer: New York, NY, USA, 2008.
24. Kuznetsov, Y.; Meijer, H. *Numerical Bifurcation Analysis of Maps*; Cambridge Monographs on Applied and Computational Mathematics; Cambridge University Press: Cambridge, UK, 2019.
25. Wiggins, S. *Introduction to Applied Nonlinear Dynamical Systems and Chaos*; Texts in Applied Mathematics; Springer: New York, NY, USA, 2003.
26. Friedli, S.; Velenik, Y. *Statistical Mechanics of Lattice Systems: A Concrete Mathematical Introduction*; Cambridge University Press: Cambridge, UK, 2017.
27. Ising, E. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik* **1925**, *31*, 253–258. [CrossRef]
28. Onsager, L. Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition. *Phys. Rev.* **1944**, *65*, 117–149. [CrossRef]
29. Toda, M.; Toda, M.; Saito, N.; Kubo, R.; Saito, N. *Statistical Physics I: Equilibrium Statistical Mechanics*; Springer Series in Solid-State Sciences; Springer: Berlin/Heidelberg, Germany, 2012.
30. Moessner, R.; Ramirez, A.P. Geometrical frustration. *Phys. Today* **2006**, *59*, 24. [CrossRef]
31. Basak, A.; Mukherjee, S. Universality of the mean-field for the Potts model. *Probab. Theory Relat. Fields* **2017**, *168*, 557–600. [CrossRef]
32. Blanca, A.; Chen, Z.; Vigoda, E. Swendsen-Wang dynamics for general graphs in the tree uniqueness region. *Random Struct. Algorithms* **2019**, *56*, 373–400. [CrossRef]
33. Guo, H.; Jerrum, M. Random cluster dynamics for the Ising model is rapidly mixing. *Ann. Appl. Probab.* **2018**, *28*, 1292–1313. [CrossRef]
34. Oostwal, E.; Straat, M.; Biehl, M. Hidden Unit Specialization in Layered Neural Networks: ReLU vs. Sigmoidal Activation. *arXiv* **2019**, arXiv:1910.07476.
35. Çakmak, B.; Opper, M. A Dynamical Mean-Field Theory for Learning in Restricted Boltzmann Machines. *arXiv* **2020**, arXiv:2005.01560.
36. Blum, E.; Wang, X. Stability of fixed points and periodic orbits and bifurcations in analog neural networks. *Neural Netw.* **1992**, *5*, 577–587. [CrossRef]
37. Grimmett, G. *The Random-Cluster Model*; Grundlehren der Mathematischen Wissenschaften; Springer: Berlin/Heidelberg, Germany, 2006.
38. Elaydi, S. *Discrete Chaos: With Applications in Science and Engineering*; CRC Press: New York, NY, USA, 2007.