Convergence of CD Algorithm in Exponential Family

Tung-Yu Wu[*], **Bai Jiang**[*], **Yifan Jin** and **Wing H. Wong**[†]

Department of Statistics
390 Serra Mall
Stanford, CA 94305
e-mail: whwong@stanford.edu
url: http://web.stanford.edu/group/wonglab/

Abstract: This paper studies the convergence properties of contrastive divergence algorithm for parameter inference in exponential family, by relating it to Markov chain theory and stochastic stability literature. We prove that, under mild conditions and given a finite data sample $X_1, \dots, X_n \sim p_{\theta^*}$ i.i.d. in an event with probability approaching to 1, the sequence $\{\theta_t\}_{t \geq 0}$ generated by CD algorithm is a positive Harris recurrent chain, and thus processes an unique invariant distribution π_n. The invariant distribution concentrates around the Maximum Likelihood Estimate at a speed arbitrarily slower than $\sqrt{n}$, and the number of steps in Markov Chain Monte Carlo only affects the coefficient factor of the concentration rate. Finally we conclude that as $n \to \infty$,

$$\limsup_{t \to \infty} \left\| \frac{1}{t} \sum_{s=1}^{t} \theta_s - \theta^* \right\| \xrightarrow{p} 0.$$

MSC 2010 subject classifications: Primary 68W48, 60J20; secondary 93E15.
Keywords and phrases: CD Algorithm, Exponential Family, Harris Recurrent Chain, Convergence Rate.

1. Introduction

Contrastive Divergence (CD) algorithm [1] has been widely used for parameter inference of MRFs. This first example of application is given by Hinton [1] to train RBMs, the essential building blocks for Deep Belief Networks [2, 3, 4]. The key idea behind CD is to approximate the computationally intractable term in the likelihood gradient by running a small number (m) of steps of a MCMC run. Thus it is much faster than the conventional MCMC methods that run a large number to reach equilibrium distributions.

Despite of CD's empirical success, theoretical understanding of its behavior is far less satisfactory. Both computer simulation and theoretical analysis show that CD may fail to converge to the correct solution [5, 6, 7]. Studies on theoretical convergence properties have thus been motivated. Yuille relates the

[*]Contributing equally to this paper
[†]Corresponding author

algorithm to the stochastic approximation literature and gives very restrictive convergence conditions [8]. Others show that for Restricted Boltzmann Machines the CD update is not the gradient of any function [9], but that for full-visible Boltzmann Machines the CD update can be viewed as the gradient of pseudo-likelihood function if adopting a simple scheme of Gibbs sampling [10]. In any case, the fundamental question of why CD with finite m can work asymptotically in the limit of $n \to \infty$ has not been answered.

This paper studies the convergence properties of CD algorithm in exponential families and gives the convergence conditions involving the number of steps of Markov kernel transitions m, spectral gap of Markov kernels, concavity of the log-likelihood function and learning rate η of CD updates (assumed fixed in our analyses). This enables us to establish the convergence of CD with a fixed m to the true parameter as the sample size n increases.

Section 2 describes the CD algorithm for exponential family with parameter θ and data X. Section 3 states our main result: denoting $\{\theta_t\}_{t\geq 0}$ the parameter sequence generated by CD algorithm from an i.i.d. data sample $X_1, \ldots, X_n \sim p_{\theta^*}$, a sufficiently large m can guarantee

$$\limsup_{t\to\infty} \left\| \frac{1}{t}\sum_{s=1}^{t} \theta_s - \theta^* \right\| \xrightarrow{p} 0 \text{ as } n \to \infty,$$

under mild conditions. Section 4 shows that $\{\theta_t\}_{t\geq 0}$ is a Markov chain under $\mathbb{P}^{\mathbf{X}}$, the conditional probability measure given any realization of data sample $\mathbf{X} = (X_1, \ldots, X_n)$, and impose three constraints on $\mathbf{X}$, which hold asymptotically with probability 1. Thereafter Sections 5-8 studies $\{\theta_t\}_{t\geq 0}$ under $\mathbb{P}^{\mathbf{X}}$ in the framework of Markov chain theory, and show that the chain is *positive Harris recurrent* and thus processes a unique invariant distribution π_n. The invariant distribution π_n concentrates around the MLE $\hat{\theta}_n$ at a speed arbitrarily slower than $\sqrt{n}$, and m only affects the coefficient factor of the concentration rate. Section 9 completes the proof of the main result.

For convenience of the reader, we assume throughout Sections 3-9 that the exponential family under study is a set of continuous probability distributions and show in Section 10 how to get a similar conclusion for the case of discrete probability distribution. We also provide two numerical experiments to illustrate the theories in Section 11.

2. CD in Exponential Family

Consider an exponential family over $\mathcal{X} \subseteq \mathbb{R}^p$ with parameter $\theta \in \Theta \subseteq \mathbb{R}^d$

$$p_\theta(x) = c(x)e^{\theta\cdot\phi(x)-\Lambda(\theta)}$$

where $c(x)$ is the carrier measure, $\phi(x) \subseteq \mathbb{R}^d$ is the sufficient statistic and $\Lambda(\theta)$ is the cumulant generating function

$$\Lambda(\theta) = \log \int_{\mathcal{X}} c(x)e^{\theta\cdot\phi(x)}dx.$$

We assume $\phi(\mathcal{X})$ is bounded, then the natural parameter domain $\{\theta \in \mathbb{R}^d : \Lambda(\theta) < \infty\} = \mathbb{R}^d$ (if it is not empty). $\Lambda(\theta)$ is convex and differentiable at any interior point of the natural parameter domain, and both the gradient and Hessian of cumulant generating function $\Lambda(\theta)$ exist

$$\mu(\theta) \triangleq \nabla\Lambda(\theta) = \mathbb{E}_\theta[\phi(X)]$$
$$\Sigma(\theta) \triangleq \nabla^2\Lambda(\theta) = \mathbb{C}\mathrm{ov}_\theta[\phi(X)]$$

Given an i.i.d. sample $\mathbf{X} = (X_1, ..., X_n)$ generated from a certain underlying distribution p_{θ^*}, the log-likelihood function is

$$l(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log c(X_i) + \theta \cdot \frac{1}{n}\sum_{i=1}^{n} \phi(X_i) - \Lambda(\theta),$$

and the gradient

$$g(\theta) \triangleq \nabla l(\theta) = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i) - \mu(\theta).$$

Assuming the positive definiteness of $\Sigma(\theta)$, the Maximum Likelihood Estimate (MLE) $\hat{\theta}_n$ uniquely exists and satisfies $g(\hat{\theta}_n) = 0$ or equivalently

$$\mu(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i).$$

ML learning can be done by gradient ascent

$$\theta^{\mathrm{new}} = \theta + \eta g(\theta) = \theta + \eta \left[\frac{1}{n}\sum_{i=1}^{n} \phi(X_i) - \mu(\theta)\right]$$

where learning rate $\eta > 0$.

When computing the gradient $g(\theta)$, the first term $\frac{1}{n}\sum_{i=1}^{n} \phi(X_i)$ is easy to compute. But it is usually difficult to compute the second term $\mu(\theta)$, which involves a complicated integral over $\mathcal{X}$. MCMC methods may generate a random sample from a Markov chain with the equilibrium distribution $p_\theta(x)$ and approximate $\mu(\theta)$ by the sample average . However , Markov chains take a large number of steps to reach the equilibrium distributions.

To address this problem, Hinton proposed the CD method [1]. The idea of CD is to replace $\mu(\theta)$ and $g(\theta)$ with

$$\mu_{\mathrm{cd}}(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} \phi\left(X_i^{(m)}\right), \quad g_{\mathrm{cd}}(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} \phi(X_i) - \hat{\mu}(\theta)$$

respectively, where $X_i^{(m)}$ is obtained by a small number (m) of steps of an MCMC run starting from the observed sample X_i. Formally, denote by $k_\theta(x, y)$

the Markov transition kernel with $p_\theta(x)$ as equilibrium distribution. CD first run Markov chains for m steps

$$X_i \xrightarrow{k_\theta} X_i^{(1)} \xrightarrow{k_\theta} X_i^{(2)} \xrightarrow{k_\theta} \dots \xrightarrow{k_\theta} X_i^{(m)}, \text{ for } i = 1, ..., n,$$

and makes update

$$\theta^{\text{new}} = \theta + \eta g_{\text{cd}}(\theta) = \theta + \eta \left[\frac{1}{n} \sum_{i=1}^{n} \phi(X_i) - \frac{1}{n} \sum_{i=1}^{n} \phi\left(X_i^{(m)} \right) \right].$$

Denote by K_θ the Markov operator associated with $k_\theta(x, y)$, i.e.

$$\boxed{(K_\theta f)(x) = \int_{\mathcal{X}} f(y) k_\theta(x, y) dy,}$$

and by $\alpha(\theta)$ the second largest absolute eigenvalue of K_θ. Markov kernel K_θ is said to have $\mathcal{L}_2$-spectral gap $1 - \alpha(\theta)$ if $\alpha(\theta) < 1$. The convergence rate of MCMC depends on $\mathcal{L}_2$-spectral gap [11].

Throughout the paper, k_θ^m denotes the m-step transition kernel of k_θ, $k_\theta^m p_{\theta'}(\cdot)$ denotes the distribution of Markov chain after m-step transition starting from initial distribution $p_{\theta'}$, and K_θ^m denotes the m-step Markov operator of K_θ. We also let $\|\cdot\|$ denote the l_2-norm $\|\cdot\|_2$.

3. Main Result

We base the convergence properties of CD algorithms for for exponential family of continuous distributions on the assumptions (A1), (A2), (A3), (A4), (A5), (A6). Theorem 3.1 states our main result, whose proof is presented in Sections 4-9. We later show in Section 10 a similar conclusion for the case of discrete distribution.

(A1) $\phi(x)$ is bounded, i.e. there exists some constant C such that $\phi(x) \subseteq [-C, C]^d$ for any $x \in \mathcal{X}$.

(A2) $\Theta \subseteq \mathbb{R}^d$ is convex and compact, and the true parameter θ^* is an interior point of Θ.

(A3) For any $\theta \in \Theta$, $\phi_j(X), 1 \leq j \leq d$ are linearly independent under p_θ, and thus $\Sigma(\theta)$ is positive definite. This assumption together with continuity of $\Sigma(\theta)$ and (A2) immediately implies the existence of the bounds for smallest and largest eigenvalues of $\Sigma(\theta)$

$$\lambda_{\min} \triangleq \inf_{\Theta} \lambda_{\min}(\theta) > 0, \quad \lambda_{\max} \triangleq \sup_{\Theta} \lambda_{\max}(\theta) < \infty.$$

(A4) Define a metric ρ on the set of Markov operators $\{K_\theta : \theta \in \Theta\}$ as

$$\rho(K_\theta, K_{\theta'}) \triangleq \sup_{f:\ |f| \leq 1} \sup_{x \in \mathcal{X}} |(K_\theta f)(x) - (K_{\theta'} f)(x)|,$$

and assume the Lipchitz continuity of K_θ in sense that

$$\rho(K_\theta, K_{\theta'}) \leq \zeta \|\theta - \theta'\| \text{ for any } \theta, \theta' \in \Theta.$$

(A5) Markov operators K_θ have $\mathcal{L}_2$-spectral gap $1-\alpha(\theta)$ and

$$\alpha \triangleq \sup_{\Theta} \alpha(\theta) < 1.$$

(A6) $\phi(\mathcal{X})$ is a convex set. Using MCMC transition kernel $k_\theta(x,y)$, $\phi(y)$ has a conditional pdf $p(\phi|\theta,x)$ conditioning on θ and x. For any $x \in \mathcal{X}$, $\inf_{\theta\in\Theta}$ $\inf_{\phi\in\phi(\mathcal{X})} p(\phi|\theta,x) > 0$.

Theorem 3.1. *Assume (A1), (A2), (A3), (A4), (A5), (A6), and the data sample $X_1,\ldots,X_n \sim p_{\theta^*}$ i.i.d.. There exists some constant $L>0$. For any m and learning rate η satisfying*

$$\lambda_{min}^2 - \sqrt{d}CL\alpha^m\lambda_{max} - \frac{\eta\lambda_{max}}{2}\left(\lambda_{max} + \sqrt{d}CL\alpha^m\right)^2 > 0,$$

CD algorithm generates a sequence $\{\theta_t\}_{t\geq 0}$ such that for any $\epsilon>0$

$$\lim_{n\to\infty} \mathbb{P}\left(\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=1}^{t}\theta_s - \theta^*\right\| > \epsilon\right) = 0.$$

4. Conditioning on Data Sample

It is not hard to see that CD generates a Markov chain $\{\theta_t\}_{t\geq 0}$ in the parameter space Θ given any realization of the data sample $\mathbf{X}=\mathbf{x}$. Indeed, denote by $\mathbf{X}_t^{(m)} = \left(X_{t,1}^{(m)}, X_{t,2}^{(m)}, ..., X_{t,n}^{(m)}\right)$ the m-step MCMC random sample to estimate the CD gradient from θ_{t-1} to θ_t. The filtration

$$\mathcal{G}_t \triangleq \sigma\text{-algebra}\left(\mathbf{X}, \theta_0, \mathbf{X}_1^{(m)}, \theta_1, \mathbf{X}_2^{(m)}, ..., \theta_{t-1}, \mathbf{X}_t^{(m)}, \theta_t\right)$$

contains all historical information until t^{th} step of CD. The CD update

$$\theta_t = \theta_{t-1} + \eta\left[\frac{1}{n}\sum_{i=1}^{n}\phi(x_i) - \frac{1}{n}\sum_{i=1}^{n}\phi\left(X_{t,i}^{(m)}\right)\right]$$

is merely function of data $\mathbf{X}=\mathbf{x}$, current parameter θ_{t-1} and m-step MCMC sample $\mathbf{X}_t^{(m)} = \left(X_{t,1}^{(m)}, X_{t,2}^{(m)}, ..., X_{t,n}^{(m)}\right)$. Conditioning on data sample $\mathbf{x}$ and current parameter θ_t, $X_{t,i}^{(m)}$ is independent to the history of CD updates. Thus $\{\theta_t\}_{t\geq 0}$ is indeed a homogeneous $\mathcal{G}_t$-adapted Markov chain under $\mathbb{P}^{\mathbf{x}}$, the conditional probability measure given $\mathbf{X}=\mathbf{x}$.

Thereafter the remaining of the paper studies CD path $\{\theta_t\}_{t\geq 0}$ in the framework of Markov chain theory. From now on $\mathbb{P}_\theta^{\mathbf{x}}$ denotes the conditional probability measure given data sample $\mathbf{x}$ and parameter θ. And $\mathbb{E}_\theta^{\mathbf{x}}$ and $\mathbb{C}\mathrm{ov}_\theta^{\mathbf{x}}$ denote the expectation and covariance under $\mathbb{P}_\theta^{\mathbf{x}}$.

Next we impose three constraints (4.1), (4.2), (4.3) on data sample $\mathbf{X}$, which are shown in Lemma 4.1 to hold asymptotically with probability 1 as $n\to\infty$. We later show that the Markov chain $\{\theta_t\}_{t\geq 0}$ converges to an invariant distribution under $\mathbb{P}^{\mathbf{x}}$ if the data sample $\mathbf{X} = \mathbf{x}$ satisfies these constraints.

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \phi(X_i) - \mu(\theta^*) \right\| < n^{-1/2+\gamma_1} \tag{4.1}$$

$$\|\hat{\theta}_n(X_1, ..., X_n) - \theta^*\| < n^{-1/2+\gamma_1} \tag{4.2}$$

$$\sup_{\theta\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^{n} \int \phi(y) k_\theta^m(X_i, y) dy - \int \phi(y) k_\theta^m p_{\theta^*}(y) dy \right\| < n^{-1/2+\gamma_1}. \tag{4.3}$$

where θ^* is the true parameter and γ_1 is any number between 0 and 1/2).

Lemma 4.1. *Assume (A1), (A2), (A3), (A4), and* $X_1, ..., X_n \overset{i.i.d.}{\sim} p_{\theta^*}$.

$$\lim_{n\to\infty} \mathbb{P}\left((4.1), (4.2), (4.3)\right) = 1$$

for any $\gamma_1 \in (0, 1/2)$.

The result that (4.1) and (4.2) hold asymptotically with probability 1 follows from standard theorems in large sample theory [12]. Therefore it suffices to show (4.3) holds asymptotically with probability 1. To this end, we define

$$f_\theta : x \mapsto \int_{\mathcal{X}} \phi(y) k_\theta^m(x, y) dy$$

and bound the tail of empirical process

$$\begin{aligned} v_n(f_\theta) &\triangleq n^{-1/2} \sum_{i=1}^{n} \left[f_\theta(X_i) - \int_{\mathcal{X}} f_\theta(x) p_{\theta^*}(x) dx \right] \\ &= n^{-1/2} \sum_{i=1}^{n} \left[\int_{\mathcal{X}} \phi(y) k_\theta^m(X_i, y) dy - \int_{\mathcal{X}} \phi(y) k_\theta^m p_{\theta^*}(y) dy \right] \end{aligned}$$

by Theorem 2.14.9 in [13], which relates the tail of empirical process to *covering number* of function class. Details of proof are provided in Appendix.

5. Gradient Approximation Error

Our study on the Markov chain $\{\theta_t\}_{t\geq 0}$ starts from appropriately bounding bias and variance of CD gradient $g_{\text{cd}}(\theta)$ under $\mathbb{P}^{\mathbf{x}}$. Lemma 5.1 shows that the bias of $g_{\text{cd}}(\theta)$ is $\mathcal{O}(n^{-1/2+\gamma_1}) + \mathcal{O}(\alpha^m \|\theta - \hat{\theta}_n\|)$ depending on the mixing rate of chains in MCMC α, the MCMC step number m, sample size n and the distance between θ and the MLE $\hat{\theta}_n$, and that the covariance of $g_{\text{cd}}(\theta)$ is $\mathcal{O}(1/n)$ depending on the sample size n.

Write the gradient approximation error

$$\Delta g(\theta) = g_{\text{cd}}(\theta) - g(\theta).$$

Lemma 5.1. *Assume (A1), (A2) and (A5) and data sample* $\mathbf{x}$ *satisfies (4.2) and (4.3). Then*

$$\|\mathbb{E}_\theta^{\mathbf{x}}[\Delta g]\| \le \left(1+\sqrt{d}CL\alpha^m\right)n^{-1/2+\gamma_1}+\sqrt{d}CL\alpha^m\|\theta-\hat{\theta}_n\|$$

for some constant $L>0$*, where* $1-\alpha$ *is the* $\mathcal{L}_2$*-spectral gap of Markov operators* K_θ *in MCMC and* γ_1 *is introduced by inequalities (4.2) and (4.3). And*

$$trace\left[\mathbb{C}ov_\theta^{\mathbf{x}}\Delta g(\theta)\right] \le \frac{dC^2}{n}.$$

Proof. For simplicity of notations, we abbreviate $\Delta g(\theta)$ as Δg.

$$\begin{aligned}\mathbb{E}_\theta^{\mathbf{x}}[\Delta g] &= \mu(\theta)-\frac{1}{n}\sum_{i=1}^n\int_{\mathcal{X}}\phi(y)k_\theta^m(x_i,y)dy\\ &= \mu(\theta)-\int_{\mathcal{X}}\phi(y)k_\theta^m p_{\theta^*}(y)dy+\int_{\mathcal{X}}\phi(y)k_\theta^m p_{\theta^*}(y)dy-\frac{1}{n}\sum_{i=1}^n\int_{\mathcal{X}}\phi(y)k_\theta^m(x_i,y)dy\end{aligned}$$

implying

$$\|\mathbb{E}_\theta^{\mathbf{x}}[\Delta g]\| \le \left\|\mu(\theta)-\int_{\mathcal{X}}\phi(y)k_\theta^m p_{\theta^*}(y)dy\right\|+\left\|\frac{1}{n}\sum_{i=1}^n\int_{\mathcal{X}}\phi(y)k_\theta^m(x_i,y)dy-\int_{\mathcal{X}}\phi(y)k_\theta^m p_{\theta^*}(y)dy\right\|.$$

The second term is bounded by $n^{-1/2+\gamma_1}$ in inequality (4.3). For the first term, consider each $j=1,...,d$,

$$\begin{aligned}&\left|\mu_j(\theta)-\int_{\mathcal{X}}\phi_j(y)k_\theta^m p_{\theta^*}(y)dy\right|\\ =&\left|\int_{\mathcal{X}}\phi_j(y)k_\theta^m p_\theta(y)dy-\int_{\mathcal{X}}\phi_j(y)k_\theta^m p_{\theta^*}(y)dy\right|\\ =&\left|\int_{\mathcal{X}}K_\theta^m\phi_j(y)\left(\frac{p_{\theta^*}(y)}{p_\theta(y)}-1\right)p_\theta(y)dy\right|\\ =&\left|\int_{\mathcal{X}}K_\theta^m\left(\phi_j(y)-\mu_j(\theta)\right)\left(\frac{p_{\theta^*}(y)}{p_\theta(y)}-1\right)p_\theta(y)dy\right|\\ \le&\,\alpha(\theta)^m\sqrt{\int_{\mathcal{X}}\left(\phi_j(y)-\mu_j(\theta)\right)^2p_\theta(y)dy}\sqrt{\int_{\mathcal{X}}\left(\frac{p_{\theta^*}(y)}{p_\theta(y)}-1\right)^2p_\theta(y)dy}\\ \le&\,\alpha^m C\sqrt{\int_{\mathcal{X}}\left(\frac{p_{\theta^*}(y)}{p_\theta(y)}-1\right)^2p_\theta(y)dy}\\ =&\,\alpha^m C\sqrt{e^{-2\Lambda(\theta^*)+\Lambda(\theta)+\Lambda(2\theta^*-\theta)}-1}\\ \le&\, CL\alpha^m\|\theta-\theta^*\|\end{aligned}\tag{5.1}$$

where L is the Lipchitz constant of the continuously differentiable function $f:\theta\in\Theta\mapsto\sqrt{e^{-2\Lambda(\theta^*)+\Lambda(\theta)+\Lambda(2\theta^*-\theta)}-1}$, and the last step follows from

$$|f(\theta)|=|f(\theta)-0|=|f(\theta)-f(\theta^*)|\le L\|\theta-\theta^*\|.$$

Putting (4.2), (4.3) and (5.1) together yields

$$\begin{aligned}\|\mathbb{E}_\theta^{\mathbf{x}}\Delta g\| &\leq n^{-1/2+\gamma_1} + \sqrt{d}CL\alpha^m\|\theta-\theta^*\| \\ &\leq n^{-1/2+\gamma_1} + \sqrt{d}CL\alpha^m\|\hat{\theta}_n-\theta^*\| + \sqrt{d}CL\alpha^m\|\theta-\hat{\theta}_n\| \\ &\leq \left(1+\sqrt{d}CL\alpha^m\right)n^{-1/2+\gamma_1} + \sqrt{d}CL\alpha^m\|\theta-\hat{\theta}_n\|\end{aligned}$$

Also, noting that $X_i^{(m)}|\mathbf{x},\theta \sim k_\theta^m(x_i,\cdot)$ are conditional independent (but not identically distributed) since we run n Markov Chains independently starting from different $x_i, i=1,...,n$, write

$$\begin{aligned}&\operatorname{trace}\left[\mathbb{C}\mathrm{ov}_\theta^{\mathbf{x}}\Delta g\right] \\ &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^d\int_{\mathcal{X}}\left(\phi_j(y)-\int_{\mathcal{X}}\phi_j(z)k_\theta^m(x_i,z)dz\right)^2 k_\theta^m(x_i,y)dy \\ &\leq \frac{dC^2}{n}.\end{aligned}$$

□

6. Drift Towards MLE

When studying the convergence of CD method, we hypothesize that starting at some θ_0 far away from $\hat{\theta}_n$, the exact gradient $g(\theta_t)$ is large enough to dominate the approximation error of m-step MCMC sampling, and bring a positive *drift* in log-likelihood, until θ_t is close to $\hat{\theta}_n$ and $g(\theta_t)$ fails to suppress the MCMC error. To precisely characterize this phenomenon, we give the definitions of *drift* and establish the *drift condition* with *Lyapunov function*

$$u(\theta) \triangleq l(\hat{\theta}_n) - l(\theta)$$

being the log-likelihood gap at θ compared to the MLE $\hat{\theta}_n$.

Definition 6.1. *(drift) Let $V: S \to \mathbb{R}_+$ be some function on the state space of a Markov chain $\{Z_t\}_{t\geq 0}$. The one-step drift of V is defined as*

$$\mathbb{E}_z V(Z_1) - V(z)$$

Definition 6.2. *(drift condition) V satisfies the drift condition if*

$$\mathbb{E}_z V(Z_1) - V(z) \leq -\delta \text{ for } z \in B^c$$

with some $\delta > 0$ and some subset of the state space B. V is called a Lyapunov function for the Markov chain $\{Z_t\}_{t\geq 0}$.

Lemma 6.1 shows that function $u(\theta) \triangleq l(\hat{\theta}) - l(\theta)$ satisfies *drift condition* outside of closed balls $B_\beta = \{\theta \in \Theta : \|\theta - \hat{\theta}\| \leq \beta r_n\}$ with $\beta > 1$ and $r_n = \mathcal{O}(n^{-1/2+\gamma_1})$.

Lemma 6.1. *Assume (A1), (A2), (A3), (A5), and data sample* $\mathbf{x}$ *satisfies (4.2) and (4.3). For any* m *and learning rate* η *satisfying*

$$a \triangleq \lambda_{min}^2 - \sqrt{d}CL\alpha^m\lambda_{max} - \frac{\eta\lambda_{max}}{2}\left(\lambda_{max} + \sqrt{d}CL\alpha^m\right)^2 > 0,$$

the chain $\{\theta_t\}_{t\geq 0}$ *has Lyapunov function* $u(\theta) = l(\hat{\theta}_n) - l(\theta)$ *which satisfies drift condition outside closed ball*

$$B_\beta = \{\theta \in \Theta : \|\theta - \hat{\theta}_n\| \leq \beta r_n\}$$

for any $\beta > 1$ *with*

$$\delta \geq \eta(\beta^2 - 1)c_n, \quad r_n = \frac{b_n + \sqrt{b_n^2 + 4ac_n}}{2a} \asymp n^{-1/2+\gamma_1}$$

where

$$b_n = \lambda_{max}\left(1 + \sqrt{d}CL\alpha^m\right)\left(1 + \eta\lambda_{max} + \eta\sqrt{d}CL\alpha^m\right)n^{-1/2+\gamma_1}$$
$$c_n = \frac{\eta\lambda_{max}}{2}\left[dC^2 n^{-2\gamma_1} + \left(1 + \sqrt{d}CL\alpha^m\right)^2\right]n^{-1+2\gamma_1}$$

Proof. For simplicity of notations, we abbreviate $g(\theta)$ as g, $g_{\mathrm{cd}}(\theta)$ as g_{cd}, and $\Delta g(\theta)$ as Δg. The difference $u(\theta_1) - u(\theta)$ when moving from θ to $\theta_1 = \theta + \eta g_{\mathrm{cd}}$ is bounded from above as

$$\begin{aligned} u(\theta_1) - u(\theta) &= -\eta\langle g, g_{\mathrm{cd}}\rangle + \frac{1}{2}\eta^2\langle\Sigma(\theta')g_{\mathrm{cd}}, g_{\mathrm{cd}}\rangle \\ &\leq -\eta\langle g, g_{\mathrm{cd}}\rangle + \frac{1}{2}\eta^2\lambda_{\max}\langle g_{\mathrm{cd}}, g_{\mathrm{cd}}\rangle \\ &= -\eta\langle g, g\rangle - \eta\langle g, \Delta g\rangle + \frac{1}{2}\eta^2\lambda_{\max}\langle g_{\mathrm{cd}}, g_{\mathrm{cd}}\rangle \end{aligned}$$

where θ' is a convex combination of θ and θ_1. The first term $\|g\|^2$ is constant. Expectation of the second term is

$$\mathbb{E}_\theta^{\mathbf{x}}\left[-\langle g, \Delta g\rangle\right] = -\langle g, \mathbb{E}_\theta^{\mathbf{x}}[\Delta g]\rangle \leq \|g\|\,\|\mathbb{E}_\theta^{\mathbf{x}}[\Delta g]\|, \tag{6.1}$$

and expectation of the third term is

$$\begin{aligned} \mathbb{E}_\theta^{\mathbf{x}}\left[\langle g_{\mathrm{cd}}, g_{\mathrm{cd}}\rangle\right] &= \mathrm{trace}\left[\mathbb{C}\mathrm{ov}_\theta^{\mathbf{x}} g_{\mathrm{cd}}\right] + \|\mathbb{E}_\theta^{\mathbf{x}}[g_{\mathrm{cd}}]\|^2 \\ &\leq \mathrm{trace}\left[\mathbb{C}\mathrm{ov}_\theta^{\mathbf{x}}\Delta g\right] + \left(\|g\| + \|\mathbb{E}_\theta^{\mathbf{x}}[\Delta g]\|\right)^2. \end{aligned} \tag{6.2}$$

Since $g(\theta) = g(\theta) - g(\hat{\theta}_n) = \mu(\hat{\theta}_n) - \mu(\theta)$,

$$\lambda_{\min}\|\theta - \hat{\theta}_n\| \leq \|g(\theta)\| \leq \lambda_{\max}\|\theta - \hat{\theta}_n\|. \tag{6.3}$$

Putting (6.1), (6.2) and (6.3) with Lemma 5.1 together yields

$$\mathbb{E}_\theta^{\mathbf{x}}\left[u(\theta_1) - u(\theta)\right] \leq -\eta(a\|\theta - \hat{\theta}_n\|^2 - b_n\|\theta - \hat{\theta}_n\| - c_n). \tag{6.4}$$

The RHS of (6.4) has a quadratic form of $\|\theta - \hat{\theta}_n\|$, and it is clear that $a > 0$ for sufficiently large m and sufficiently small η. Then

$$\|\theta - \hat{\theta}_n\| \geq r_n = \frac{b_n + \sqrt{b_n^2 + 4ac_n}}{2a} \implies \mathbb{E}_\theta^{\mathbf{x}}\left[u(\theta_1) - u(\theta)\right] \leq 0$$

can guarantee $\mathbb{E}_\theta^{\mathbf{x}}\left[u(\theta_1) - u(\theta)\right] \leq 0$. In particular, the drift condition holds outside any closed ball B_β centering at MLE of radius βr_n with $\beta > 1$, i.e.

$$\|\theta - \hat{\theta}_n\| > \beta r_n \implies \mathbb{E}_\theta^{\mathbf{x}}\left[u(\theta_1) - u(\theta)\right] \leq -\delta < 0$$

with

$$\begin{aligned}
\delta &= \eta(a\beta^2 r_n^2 - b_n\beta r_n - c_n) \\
&= \eta[a(\beta^2 - 1)r_n^2 - b_n(\beta - 1)r_n] \\
&\geq \eta(\beta^2 - 1)(ar_n^2 - br_n) \\
&= \eta(\beta^2 - 1)c_n
\end{aligned}$$

□

Remark. A function $h(\cdot)$ is called *supharmonic* for a transition probability $p(x, \cdot)$ at x if $\int h(y)p(x, dy) \leq h(x)$. And it is called *strong supharmonic* if

$$\int h(y)p(x, dy) \leq h(x) - \delta$$

for some positive δ. We actually prove in Lemma 6.1 that $u(\theta)$ is strong supharmonic at B_β^c. We later see in Theorem 7.1 a nice connection between strong supharmonic functions, positive recurrence of Markov chains, and supmartingales.

7. Positive Harris Recurrence

Tweedie [15] connected the *drift condition* in Definition 6.2 to positive recurrence of sets in the state space by a Markov chain. We restate this result in Theorem 7.1 and provide a proof based on sup-martingales and sup-harmonic functions in Appendix. Next, Corollary 7.1 combines Lemma 6.1 with Theorem 7.1 and concludes that the closed balls B_β centering at the MLE $\hat{\theta}_n$ of radius βr_n are positive recurrent by the chain $\{\theta_t\}_{t\geq 0}$.

Theorem 7.1. *(Theorem 6.1 in [15]) Suppose a Markov chain $\{Z_t\}_{t\geq 0}$ has a non-negative function V on the state space satisfying the drift condition in Definition 6.2 with some $\delta > 0$ and set B. Let $T = \min\{t \geq 1 : Z_t \in B\}$ be the first hitting time of B if starting from $Z_0 \in B^c$ or the first returning time otherwise, then*

$$\mathbb{E}_z T \leq \begin{cases} V(z)/\delta & \text{for } z \in B^c \\ 1 + \frac{1}{\delta}\int_{B^c} V(z_1)p(z, dz_1) & \text{for } z \in B \end{cases}$$

where $p(z, dz_1)$ is the transition probability of the chain. Thus, if

$$\sup_{z \in B} \int_{B^c} V(z_1) p(z, dz_1) < \infty$$

also holds, then the set B is positive recurrent.

Lyapunov function is widely used in stochastic stability or optimal control study [16]. As we have seen in Theorem 7.1, a suitably designed *Lyapunov function* can determine the positive recurrence of sets of a Markov chain. We proceed to apply Theorem 7.1 to the Markov chain $\{\theta_t\}_{t \geq 0}$, for which $u(\theta)$ satisfies the *drift condition* outside of any closed ball B_β in Lemma 6.1, and conclude in Corollary 7.1 that closed balls B_β centering at MLE are positive recurrent by the chain $\{\theta_t\}_{t \geq 0}$.

Corollary 7.1. *Following Theorem 7.1 and Lemma 6.1, B_β for each $\beta > 1$ are positive recurrent by the chain $\{\theta_t\}_{t \geq 0}$.*

Proof. Let $T \triangleq \min\{t \geq 1 : \theta_t \in B_\beta\}$ be the first hitting or returning time of B_β by the chain $\{\theta_t\}_{t \geq 0}$. Lemma 6.1 establishes the drift condition for the likelihood gap function $u(\theta)$ outside of B_β, i.e.

$$\mathbb{E}_\theta^{\mathbf{x}} u(\theta_1) - u(\theta) \leq -\delta \text{ for } \theta \in B_\beta^c.$$

The compactness of Θ and continuity of $u(\theta)$ follow the boundedness of $u(\theta)$, implying

$$\sup_{\theta \in B_\beta} \int_{B_\beta^c} u(\theta_1) p(\theta, d\theta_1) < \sup_{\Theta} u(\theta_1) < \infty.$$

Both conditions of Theorem 7.1 are satisfied, thus B_β is positive recurrent. $\square$

Next we prove the positive Harris recurrence of the chain $\{\theta_t\}_{t \geq 0}$, which further implies the distribution convergence of Markov chain in total variation.

Definition 7.1. *An accessible set B is called a small set of a Markov chain $\{Z_t\}_{t \geq 0}$ if*

$$\mathbb{P}_z(Z_1 \in \cdot) \geq \epsilon \mathbb{I}_B(z) q(\cdot)$$

for some positive $\epsilon > 0$ and probability measure $q(\cdot)$ over the state space.

Definition 7.2. *A Markov chain $\{Z_t\}_{t \geq 0}$ is called Harris recurrent if there exists a set B s.t.*

(a) B is recurrent.
(b) B is a small set.

If B is positive recurrent in addition, then the chain is called positive Harris recurrent.

Lemma 7.1. *Assume (A1), (A2), (A3), (A4), (A5), (A6). For sufficiently large n and for any m and learning rate η satisfying*

$$a \triangleq \lambda_{min}^2 - \sqrt{d} C L \alpha^m \lambda_{max} - \frac{\eta \lambda_{max}}{2} \left(\lambda_{max} + \sqrt{d} C L \alpha^m \right)^2 > 0,$$

data sample $\mathbf{x}$ *satisfying (4.1), (4.2) and (4.3), the chain* $\{\theta_t\}$ *generated by CD updates is positive Harris recurrent.*

Proof. Since Corollary 7.1 ensures the positive recurrence of B_β, it suffices to show B_β is a small set by checking Definition 7.1. Since θ^* is an interior point of Θ and $\mu : \Theta \to \phi(\mathcal{X})$ a continuous mapping, $\mu(\theta^*)$ is an interior point of $\phi(\mathcal{X})$. Denoting by $\partial\phi(\mathcal{X})$ the boundary of $\phi(\mathcal{X})$,

$$\inf_{\phi \in \partial\phi(\mathcal{X})} \|\mu(\theta^*) - \phi\| > 0.$$

If (4.1) $\|\frac{1}{n}\sum_{i=1}^n \phi(x_i) - \mu(\theta^*)\| < n^{-1/2+\gamma_1}$ holds,

$$\begin{aligned}\inf_{\phi \in \partial\phi(\mathcal{X})} \left\| \frac{1}{n}\sum_{i=1}^n \phi(x_i) - \phi \right\| &\geq \inf_{\phi \in \partial\phi(\mathcal{X})} \|\mu(\theta^*) - \phi\| - \left\| \frac{1}{n}\sum_{i=1}^n \phi(x_i) - \mu(\theta^*) \right\| \\ &\geq 2\beta r_n/\eta\end{aligned}$$

for sufficiently large n. Then for any $\theta \in B_\beta$

$$B_\beta \subseteq \theta + \eta \left[\frac{1}{n}\sum_{i=1}^n \phi(x_i) - \phi(\mathcal{X}) \right].$$

Assumption (A6) implies $\frac{1}{n}\sum_{i=1}^n \phi\left(X_i^{(m)}\right)$ has positive density over $\phi(\mathcal{X})$, which is strictly bounded away from 0 for any θ, so does

$$\theta_1 = \theta + \eta \left[\frac{1}{n}\sum_{i=1}^n \phi(x_i) - \frac{1}{n}\sum_{i=1}^n \phi\left(X_i^{(m)}\right) \right]$$

over $\theta + \eta\left[\frac{1}{n}\sum_{i=1}^n \phi(x_i) - \phi(\mathcal{X})\right]$. Denote by $p(\theta, \theta_1)$ the transition kernel of the chain $\{\theta_t\}$, then

$$\inf_{\theta,\theta_1 \in B_\beta} p(\theta, \theta_1) \geq \inf_{\theta \in B_\beta} \inf \left\{ p(\theta, \theta_1) : \theta_1 \in \theta + \eta \left[\frac{1}{n}\sum_{i=1}^n \phi(x_i) - \phi(\mathcal{X}) \right] \right\} > 0.$$

Let $q(\cdot)$ be the uniform measure on B_β. There exists some constant $\epsilon > 0$ such that

$$\int_A p(\theta, \theta_1) d\theta_1 \geq \epsilon I_{B_\beta}(\theta) q(A)$$

for any Borel set $A \subseteq \Theta$, completing the proof. □

As stated in Theorems 6.8.5, 6.8.7, 6.8.8 in [17], any aperiodic, positive Harris recurrent chain $\{Z_t\}_{t\geq 0}$ processes an unique invariant distribution π, and the chain $\{Z_t\}_{t\geq 0}$ converges to the invariant distribution π in total variation for π-a.e. starting point z. We strengthen these results for chain $\{\theta_t\}_{t\geq 0}$ in Corollary 7.2.

Corollary 7.2. *Following Lemma 7.1,*

(a) *The positive Harris recurrent chain* $\{\theta_t\}_{t\geq 0}$ *has an unique invariant probability measure* π_n

(b) *let* Θ_1 *be the set of* $\theta \in \Theta$ *s.t.*

$$\lim_{t\to\infty} \|\mathbb{P}_\theta^{\mathbf{x}}(\theta_t \in \cdot) - \pi_n(\cdot)\|_{total\ variation} = 0 \tag{7.1}$$

(c) π_n *and the Lesbegue measure are absolutely continuous w.r.t. each other.*

(d) π_n *has a positive density function over* Θ.

(e) *(7.1) holds for almost all* $\theta \in \Theta$ *(in sense of Lesbegue measure).*

(f) *For any* f *such that* $\int_\Theta |f(\theta)|\pi_n(\theta)d\theta < \infty$,

$$\lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} f(\theta_s) = \int_\Theta f(\theta)\pi_n(\theta)d\theta \tag{7.2}$$

Proof. Clearly the chain $\{\theta_t\}_{t\geq 0}$ is aperiodic. Then parts (a)(b) are immediate consequences of Lemma 7.1 and Theorems 6.8.5, 6.8.7, 6.8.8 in [17]. Proceed to prove part (c). On the first hand, part (b) and absolutely continuity of $\mathbb{P}_\theta^{\mathbf{x}}(\theta_t \in \cdot)$ w.r.t. the Lesbegue measure imply absolutely continuity of π_n w.r.t. the Lesbegue measure. On the other hand, the invariant measure π_n is a maximal irreducibility measure (See Definition 7.3) by Theorems 10.0.1 and 10.1.2 in [16]. Hence the Lesbegue measure is absolutely continuous w.r.t. π_n, completing the proof of part (c). Further, parts (b)(c) imply (d)(e). Part (f) is another consequence of part (a) (See details in Section 17.1 in [16]). ☐

Definition 7.3. *Let* π *be a positive measure on the state space of chain* $\{Z_t\}$. *If for any* z *and set* A *in the state space* $\pi(A) > 0$ *implies the accessibility of set* A *by the chain from any start point* z, *we say* π *is an irreducible measure for the chain (or the chain is* π*-irreducible). An irreducible measure* π^* *specifying the minimal family of null sets, i.e.* $\pi^*(A) = 0 \implies \pi(A) = 0$ *for any irreducible measure, is called a maximal irreducibility measure.*

8. Concentration of the Invariant Distribution

Lemma 8.1 shows that the invariant distribution π_n concentrates on positive recurrent ball B_β.

Lemma 8.1. *Following Corollary 7.2, the invariant probability measure* π *concentrates on* B_β *as*

$$\pi_n(B_\beta^c) \asymp n^{-2\gamma_2}$$

for any $\gamma_2 \in (0, 1/2 - \gamma_1)$ *and* $\beta \asymp n^{\gamma_2}$ *increasing with* n, *while the ball* B_β *shrinks with radius*

$$\beta r_n \asymp n^{-1/2+\gamma_1+\gamma_2}.$$

Proof. By (6.4) in Lemma 6.1,

$$\begin{aligned} 0 &= \int_\Theta \mathbb{E}_\theta^{\mathbf{x}} u(\theta_1)\pi_n(d\theta) - \int_\Theta u(\theta)\pi_n(d\theta) \\ &\leq \int_\Theta -\eta\left(a\|\theta - \hat{\theta}_n\|^2 - b_n\|\theta - \hat{\theta}_n\| - c_n\right)\pi_n(d\theta) \end{aligned}$$

Rearranging terms yields

$$\int_{B_\beta^c} \left(a\|\theta - \hat{\theta}_n\|^2 - b_n\|\theta - \hat{\theta}_n\| - c_n\right) \pi_n(d\theta)$$
$$\leq \int_{B_\beta} - \left(a\|\theta - \hat{\theta}_n\|^2 - b_n\|\theta - \hat{\theta}_n\| - c_n\right) \pi_n(d\theta).$$

At $\theta \in B_\beta^c$,

$$\begin{aligned} a\beta^2\|\theta - \hat{\theta}_n\|^2 - b_n\beta\|\theta - \hat{\theta}_n\| - c_n &\geq a\beta^2 r_n^2 - b_n\beta r_n - c_n \\ &\geq a(\beta^2 - 1)r_n^2 - b_n(\beta - 1)r_n \\ &\geq (\beta^2 - 1)\left(ar_n^2 - b_n r_n\right) \\ &\geq (\beta^2 - 1)c_n \end{aligned}$$

At $\theta \in B_\beta$,

$$-\left(a\|\theta - \hat{\theta}_n\|^2 - b_n\|\theta - \hat{\theta}_n\| - c_n\right) \leq c_n + \frac{b_n^2}{4a}$$

Thus

$$\frac{\pi_n(B_\beta^c)}{\pi_n(B_\beta)} \leq \frac{1}{\beta^2 - 1} \frac{c_n + b_n^2/4a}{c_n}$$

Noting that $b_n, c_n^{1/2} \asymp n^{-1/2+\gamma_1}$, letting $\beta \asymp n^{\gamma_2}$ increase with n yields

$$\frac{\pi_n(B_\beta^c)}{\pi_n(B_\beta)} \leq \frac{1}{\beta^2 - 1} \frac{c_n + b_n^2/4a}{c_n} \asymp n^{-2\gamma_2}$$

while the ball B_β has shrinking radius

$$\beta r_n \asymp n^{\gamma_2} \times n^{-1/2+\gamma_1} = n^{-1/2+\gamma_1+\gamma_2}.$$

□

9. Convergence of CD Estimator in Probability

Now we can complete the proof of the main result in Theorem 3.1: the estimator $\frac{1}{t}\sum_{s=1}^t \theta_t$ converges to the true parameter θ^* in probability as the sample size $n \to \infty$ in sense that for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(\limsup_{t\to\infty} \left\| \frac{1}{t}\sum_{s=1}^t \theta_s - \theta^* \right\| > \epsilon\right) = 0.$$

Proof. With Lemma 4.1, it suffices to show

$$\lim_{n\to\infty} \mathbb{P}^{\mathbf{x}}\left(\limsup_{t\to\infty} \left\| \frac{1}{t}\sum_{s=1}^t \theta_s - \theta^* \right\| > \epsilon\right) = 0$$

for any realization of data sample $\mathbf{x}$ satisfying (4.1), (4.2), (4.3). Write

$$\left\|\frac{1}{t}\sum_{s=1}^{t}\theta_s-\theta^*\right\|\le\left\|\frac{1}{t}\sum_{s=1}^{t}\theta_s-\int_\Theta\theta\pi_n(d\theta)\right\|+\left\|\int_\Theta\theta\pi_n(d\theta)-\hat\theta_n\right\|+\|\hat\theta_n-\theta^*\|.$$

The first term, by part (f) of Corollary 7.2, converges to 0, i.e.

$$\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=1}^{t}\theta_s-\int_\Theta\theta\pi_n(d\theta)\right\|=0\ \mathbb{P}^{\mathbf{x}}\text{-a.s.}.$$

The second term, by Jensen's inequality and Lemma 8.1, vanishes as n increases.

$$\begin{aligned}\left\|\int_\Theta\theta\pi_n(d\theta)-\hat\theta_n\right\|&\le\int_\Theta\|\theta-\hat\theta_n\|\pi_n(d\theta)\\&=\int_{B_\beta}\|\theta-\hat\theta_n\|\pi_n(d\theta)+\int_{B_\beta^c}\|\theta-\hat\theta_n\|\pi_n(d\theta)\\&\le\pi_n(B_\beta)\times\beta r_n+\pi_n(B_\beta^c)\times\max_{\theta\in\Theta}\|\theta-\hat\theta_n\|\\&\asymp n^{-\min\{1/2-\gamma_1-\gamma_2,2\gamma_2\}}\end{aligned}$$

The third term $\|\hat\theta_n-\theta^*\|<n^{-1/2+\gamma_1}$ as in (4.2). Therefore as $n\to\infty$

$$\begin{aligned}&\mathbb{P}^{\mathbf{x}}\left(\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=1}^{t}\theta_s-\theta^*\right\|>\epsilon\right)\\\le&\mathbb{P}^{\mathbf{x}}\left(\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=1}^{t}\theta_s-\int_\Theta\theta\pi_n(d\theta)\right\|+\left\|\int_\Theta\theta\pi_n(d\theta)-\hat\theta_n\right\|+\|\hat\theta_n-\theta^*\|>\epsilon\right)\\\le&\mathbb{P}^{\mathbf{x}}\left(\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=1}^{t}\theta_s-\int_\Theta\theta\pi_n(d\theta)\right\|>\epsilon/3\right)\\&+\mathbb{P}^{\mathbf{x}}\left(\left\|\int_\Theta\theta\pi_n(d\theta)-\hat\theta_n\right\|>\epsilon/3\right)+\mathbb{P}^{\mathbf{x}}\left(\|\hat\theta_n-\theta^*\|>\epsilon/3\right)\\\to&0.\end{aligned}$$

□

10. Results for Discrete Exponential Family

If the sufficient statistic $\phi(X)$ in the exponential family is discrete, we have a similar conclusion as stated in Theorem 10.1. In contrast to positive Harris recurrence of $\{\theta_t\}_{t\ge0}$ in Theorem 3.1 for the continuous case, $\{\theta_t\}_{t\ge0}$ for the discrete case is a Markov chain with countable state space, and may admit multiple invariant distributions.

Theorem 10.1. *Consider an exponential family of discrete probability distributions. Assume (A1), (A2), (A3), (A4), (A5) and*

(A7) $\phi(\mathcal{X})$ is finite, and for each $j = 1, ..., d$, elements in $\phi_j(\mathcal{X})$ have rational ratios.

The conclusion in Theorem 3.1 is also true.

Proof. If the sufficient statistic $\phi(X)$ in the exponential family is discrete and has finite possible values, the CD gradient $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^n \phi(x_i) - \frac{1}{n}\sum_{i=1}^n \phi(X_i^{(m)})$ has finite possible values $g_1, g_2, ..., g_s$. Starting from any initial parameter guess $\theta_0 \in \Theta$, the chain

$$\theta_t = \theta_0 + \eta \sum_{j=1}^t \hat{g}(\theta_j) = \theta_0 + \eta \sum_{k=1}^s \left[\sum_{j=1}^t \mathbb{I}\left(\hat{g}(\theta_j) = g_k\right) \right] g_k$$

lies in a countable state space $\tilde{\Theta} \subset \Theta$. And $\tilde{\Theta} \cap B_\beta$ is a finite set due to (A7). By decomposition theorem, $\tilde{\Theta}$ can be partitioned uniquely as

$$\tilde{\Theta} = \tilde{\Theta}_0 \cup \tilde{\Theta}_1 \cup \tilde{\Theta}_2 \cup ...$$

where $\tilde{\Theta}_0$ is the set of transient states and the $\tilde{\Theta}_i, i \geq 1$ are disjoint, irreducible closed set of recurrent states. The chain either remains forever in $\tilde{\Theta}_0$, or it lies eventually in some $\tilde{\Theta}_i, i \geq 1$.

We argue by contradiction that the first of these possibility does not occur. Suppose for the sake of contradiction that the chain $\{\theta_t\}$ forever lies in $\tilde{\Theta}_0$. By Corollary 7.1, B_β is positively recurrent. Thus the chain visits $\tilde{\Theta}_0 \cap B_\beta$ infinitely many times, and thus visits at least one state in $\tilde{\Theta}_0 \cap B_\beta$ infinitely many times for reason that $\tilde{\Theta}_0 \cap B_\beta \subseteq \tilde{\Theta} \cap B_\beta$ is finite. Such a state is recurrent, contradicting the fact that it belonging to the set of transient states $\tilde{\Theta}_0$.

Therefore the chain will eventually lies in the first irreducible set of recurrent states it enters, and converges to the corresponding invariant distribution π_n. Every invariant distribution π_n concentrates in B_β with

$$\frac{\pi_n(B_\beta^c)}{\pi_n(B_\beta)} \leq \frac{1}{\beta^2 - 1} \frac{c_n + b_n^2/4a}{c_n} \asymp n^{-2\gamma_2}$$

The same convergence rate with that in Theorem 3.1 can be obtained. □

Theorem 10.1 establishes the convergence property of CD algorithm for relative simple exponential family of discrete distributions, which satisfy the assumption (A7). It suffices to guaranteed the success of CD algorithm for Restricted Boltzmann Machines (RBM) which has $\phi(\mathcal{X}) = \{0, 1\}^d$.

Also, it is noteworthy that CD algorithm converges to the MLE even for more complicated cases in which $\phi(\mathcal{X})$ is infinite and/or elements in $\phi(\mathcal{X})$ have irrational ratios, if one takes the finite precision of computation into account. Due to the finite precision of any computer, numbers are always rounded or truncated. In real world, the update rule of CD algorithm is

$$\tilde{\theta}^{\text{new}} = \theta^{\text{new}} + \varepsilon = \theta + \eta \left[\frac{1}{n} \sum_{i=1}^n \phi(X_i) - \frac{1}{n} \sum_{i=1}^n \phi\left(X_i^{(m)}\right) \right] + \varepsilon$$

where ε is the numerical error incurred such that θ^{new} is substituted by its nearby grid point $\tilde{\theta}^{\text{new}}$. As

$$u(\tilde{\theta}_1) - u(\theta) = \left[u(\tilde{\theta}_1) - u(\theta_1)\right] + [u(\theta_1) - u(\theta)] \leq \mathcal{O}(\|\varepsilon\|) + [u(\theta_1) - u(\theta)]\,,$$

we have the following drift condition akin to (6.4)

$$\mathbb{E}_\theta^{\mathbf{x}}\left[u(\tilde{\theta}_1) - u(\theta)\right] \leq -\eta(a\|\theta - \hat{\theta}_n\|^2 - b_n\|\theta - \hat{\theta}_n\| - (c_n + \mathcal{O}(\|\varepsilon\|))).$$

If the precision of computation copes well with the sample size n, i.e. $\|\varepsilon\| = \mathcal{O}(c_n)$, the chain $\{\tilde{\theta}_t\}_{t\geq 0}$ is positive recurrent to the ball B_β centering at the MLE $\hat{\theta}_n$. Noting the ball B_β contains finitely many grid points, a similar argument to Theorem 10.1 can prove that the chain $\{\tilde{\theta}_t\}_{t\geq 0}$ admits invariant distributions concentrating around the MLE.

11. Numerical Experiments

11.1. Bivariate Normal Distribution

We conduct numerical experiments on the bivariate normal model

$$p(\theta) = \frac{1}{\sqrt{(2\pi)^2\,|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{11.1}$$

with unknown mean μ (parameter θ) and known covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

.

Figure 1-3 shows CD path $\{\mu_t\}_{t\geq 0}$ given a sample $\mathbf{X}$ of size $n = 50, 100, 500$ with true parameter $\mu^* = (0,0)$, respectively. For each data sample, CD-3 (i.e. CD with m=3) is applied to learn the parameter μ.

In each of Figure 1-3, subplot (a) shows the CD paths $\{\mu_t\}_{t\geq 0}$ in the parameter space with different start points $\mu = (3,3), (-3,3), (3,-3), (-3,-3)$. They illustrate that the estimated parameter initially directly moves towards to true parameter but eventually randomly walks around the true parameter. Furthermore, comparison of Figures 1(a), 2(a), 3(a) shows that the region of the random walk decreases in size as the sample size n increases.

Subplot (b) shows the true gradient of the likelihood function in each case. Subplot (c) presents the approximate gradient used by CD. For each grid point in the subplot (c), we run CD-3 5 times and draw 5 estimated gradients to reveal the randomness in CD approach. Subplot (d) reveals the directions of these gradients by normalizing the magnitude of these gradients. According to the plots (b) and (c), it can be observed that the magnitude and direction of the gradient become smaller and more stochastic when the point moves closer to the true parameter. Comparing the three figures we can see that the range of randomness decreases as the sample size increases. This is exactly what our theory predicts.

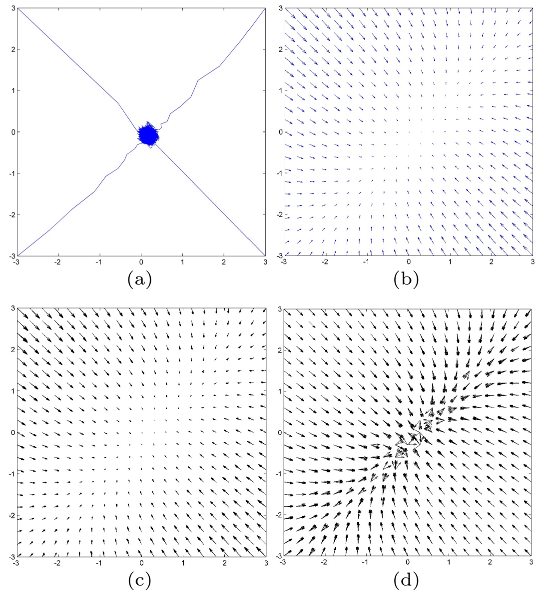

FIG 1. *Simulation results of the bivariate normal distribution with $N = 50$.*

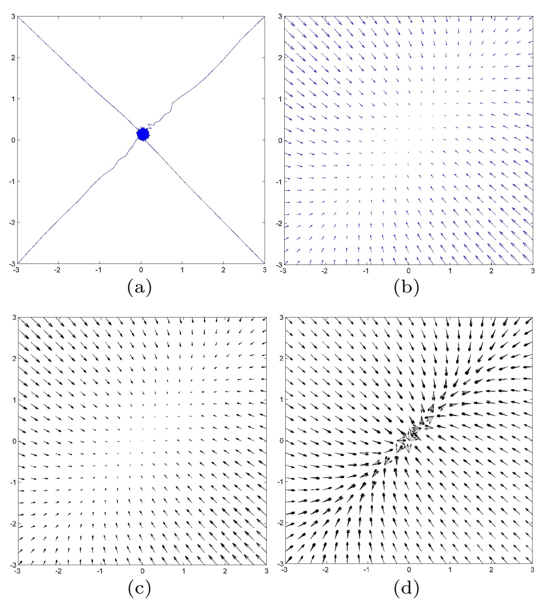

FIG 2. *Simulation results of the bivariate normal distribution with $N = 100$.*

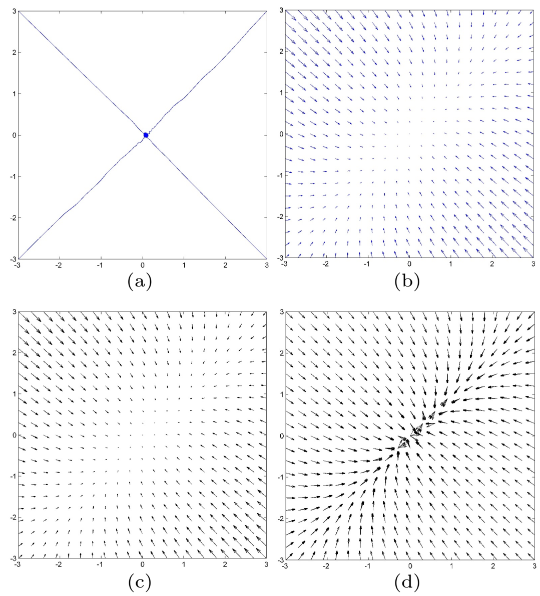

FIG 3. *Simulation results of the bivariate normal distribution with $N = 500$.*

11.2. RBMs

In our next experiment , we simulate data from the RBM . The CD method is a standard way for inferring RBM during the training of deep belief neural network [2]. There are m visible units $\boldsymbol{V}=(V_1,...,V_m)$ to represent observable data and n hidden units $\boldsymbol{H}=(H_1,...,H_n)$ to capture dependencies between observed variables . In the simulation , we focus on the binary RBM which the random variables $(\boldsymbol{V},\boldsymbol{H})$ take values $(\boldsymbol{v},\boldsymbol{h})\in\{0,1\}^{m+n}$ and $p(\boldsymbol{v},\boldsymbol{h})=\frac{1}{Z}e^{-E(\boldsymbol{v},\boldsymbol{h})}$ with the energy function

$$E(\boldsymbol{v},\boldsymbol{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}h_i v_j - \sum_{j=1}^{m} b_j v_j - \sum_{i=1}^{n} c_i h_i \tag{11.2}$$

In the simulation, the data sets is generated from a RBM with the weight matrix $\boldsymbol{w} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$ and $c_i = b_j = 0$ for $i, j = 1, 2$.Figure 4-6 shows the approximate gradients for $N = 10^2, 10^4$ and 10^6. In each figure, subplots in the lower triangular part show the approximate gradients. Let $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}^T = \begin{bmatrix} w_{11} & w_{21} & w_{12} & w_{22} \end{bmatrix}^T$. Subplot (i, j) in the lower triangular part gives the projections of the gradient onto the plane (x_i, x_j), at those points $\tilde{\mathbf{x}}$ satisfying x_k equal to 0.5 approximately, for k not equal to i or j. The corresponding directions of these gradients are given in the upper triangular part of each figure.

Again,the behaviour of the CD approximate gradients is in agreement with our theory.

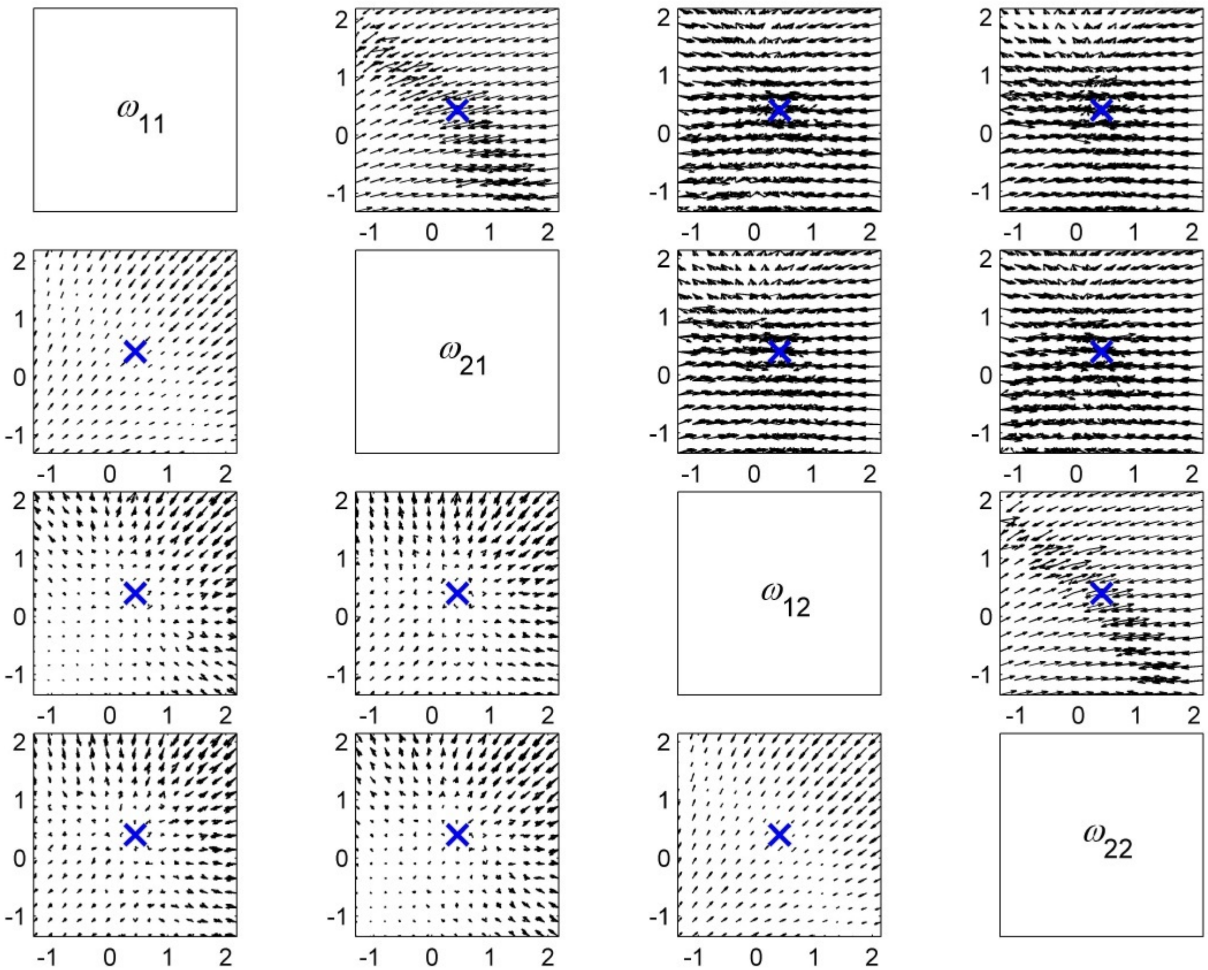

FIG 4. *Simulation results of RBM with $N = 10^2$.*

Appendix A: Proof of Lemma 2.1 and Theorem 6.1

A.1. Proof of Lemma 2.1

In Lemma 4.1, The result that (4.1) and (4.2) hold asymptotically with probability 1 follows from standard theorems in large sample theory [12]. Therefore it suffices to show inequality (4.3) holds asymptotically. Therefore it suffices to show inequality (4.3) holds with asymptotically probability 1. To this end, we applied limit theorem for empirical processes [13], which relates the tail of empirical process to *covering number* of function class.

Definition A.1. *(covering number) Let $(\mathcal{F}, D)$ be an arbitrary semi-metric space. Then the covering number $N(\epsilon, \mathcal{F}, D)$ is the minimal number of balls of radius $\epsilon > 0$ needed to cover $\mathcal{F}$. Formally,*

$$N(\epsilon, \mathcal{F}, D) = \min\{k : \exists f_1, \dots, f_k \in \mathcal{F} \ s.t. \ \sup_{f\in\mathcal{F}} \min_{1\le j\le k} D(f, f_k) < \epsilon\}.$$

Theorem A.1. *(Theorem 2.14.9 in [13]) Let $X_1, ..., X_n$ i.i.d. and $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to [0, 1]$. If*

$$\sup_Q N(\epsilon, \mathcal{F}, \mathcal{L}_{2,Q}) \le \left(\frac{C_1}{\epsilon}\right)^s, \ \forall \ 0 < \epsilon < C_1$$

where s, C_1 are constants, Q is a probability measure on $\mathcal{X}$, and

$$\mathcal{L}_{2,Q}(f_1, f_2) \triangleq \sqrt{\int_{\mathcal{X}} (f_1(x) - f_2(x))^2 Q(dx)},$$

then for every $t > 0$

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left|\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [f(X_i) - \mathbb{E}f(X_1)]\right| > t\right) \le \left(\frac{C_2 t}{\sqrt{s}}\right)^s e^{-2t^2},$$

where constant C_2 only depends on C_1.

We proceed to bound the tail $\sup_\Theta \|v_n(f_\theta)\|$ by using Theorem A.1.

Lemma A.1. *Assume (A1), (A2) and (A4) and $X_1, ..., X_n \overset{i.i.d.}{\sim} p_{\theta^*}$. Then*

$$\mathbb{P}\left(\sup_{\theta\in\Theta} \left\|\frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{X}} \phi(y) k_\theta^m(X_i, y)dy - \int_{\mathcal{X}} \phi(y) k_\theta^m p_{\theta^*}(y)dy\right\| < n^{-\frac{1}{2}+\gamma_1}\right) \to 1$$

as $n \to \infty$, for any $\gamma_1 \in (0, 1/2)$.

Proof. Let $f_\theta : x \mapsto \int \phi(y) k_\theta^m(x, y)dy$, then

$$\begin{aligned}
\int_{\mathcal{X}} \phi(y) k_\theta^m(X_i, y)dy &= f_\theta(X_i) \\
\int_{\mathcal{X}} \phi(y) k_\theta^m p_{\theta^*}(y)dy &= \int_{\mathcal{X}} \phi(y) \left(\int_{\mathcal{X}} k_\theta^m(x, y) p_{\theta^*}(x)dx\right) dy \\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \phi(y) k_\theta^m(x, y)dy\right) p_{\theta^*}(x)dx \\
&= \int_{\mathcal{X}} f_\theta(x) p_{\theta^*}(x)dx.
\end{aligned}$$

For $j = 1, ..., d$, let

$$\begin{aligned} v_n(f_\theta^j) &\triangleq n^{-1/2}\sum_{i=1}^{n}\left[f(X_i) - \mathbb{E}f(X)\right] \\ &= n^{-1/2}\sum_{i=1}^{n}\left[\int \phi_j(y)k_\theta^m(X_i, y)dy - \int \phi_j(y)k_\theta^m p_{\theta^*}(y)dy\right]. \end{aligned}$$

and view $v_n(f_\theta^j)$ as a stochastic process indexed by $\theta \in \Theta$.

$$\begin{aligned} f_\theta^j(x) - f_{\theta'}^j(x) &= \int_{\mathcal{X}} \phi_j(y)k_\theta^m(x, y)dy - \int_{\mathcal{X}} \phi_j(y)k_{\theta'}^m(x, y)dy \\ &= \sum_{i=0}^{m-1}\left[\int_{\mathcal{X}} \phi_j(y)k_\theta^{m-i}k_{\theta'}^i(x, y)dy - \int_{\mathcal{X}} \phi_j(y)k_\theta^{m-i-1}k_{\theta'}^{i+1}(x, y)dy\right] \\ &= \sum_{i=0}^{m-1}\int_{\mathcal{X}}\left[K_\theta K_\theta^{m-i-1}\phi_j - K_{\theta'}K_\theta^{m-i-1}\phi_j\right](y)k_{\theta'}^i(x, y)dy \end{aligned}$$

implying

$$\begin{aligned} |f_\theta^j(x) - f_{\theta'}^j(x)| &\le \sum_{i=0}^{m-1}\rho(K_\theta, K_{\theta'}) \times \sup_{y\in\mathcal{X}}|K_\theta^{m-i-1}\phi_j(y)| \times \int_{\mathcal{X}} k_{\theta'}^i(x, y)dy \\ &\le mC\zeta\|\theta - \theta'\| \end{aligned}$$

where ρ is the metric and ζ is the Lipchitz constant introduced by Assumption (A4). It concludes that

$$\sup_Q \mathcal{L}_{2,Q}(f_\theta^j, f_{\theta'}^j) \le mC\zeta\|\theta - \theta'\|.$$

Denoting by $\mathcal{F}^j$ the function class of $\{f_\theta^j, \theta \in \Theta\}$, it follows that

$$\sup_Q N(\epsilon, \mathcal{F}^j, \mathcal{L}_{2,Q}) \le N(\epsilon/mC\zeta, \Theta, \|\cdot\|) = \mathcal{O}(\epsilon^{-d}).$$

Applying Theorem A.1 to function class $\mathcal{F}^j$ yields

$$\mathbb{P}\left(\sup_\Theta |v_n(f_\theta^j)| > \frac{n^{\gamma_1}}{\sqrt{d}}\right) \to 0$$

as $n \to \infty$. Further,

$$\mathbb{P}\left(\sup_\Theta \|n^{-1/2}v_n(f_\theta)\| > n^{-1/2+\gamma_1}\right) \le \sum_{j=1}^{d}\mathbb{P}\left(\sup_\Theta |v_n(f_\theta^j)| > \frac{n^{\gamma_1}}{\sqrt{d}}\right) \to 0$$

as $n \to \infty$, completing the proof. □

A.2. Proof of Theorem 6.1

Proof. We first show that, if $Z_0 \in B^c$, $M_t = V(Z_{t\wedge T}) + (t\wedge T)\delta_1$ is a super-martingale adapted to Z_t's canonical filtration $\mathcal{G}_t$. The adaptedness of M_t to $\mathcal{G}_t$ follows T being a $\mathcal{G}_t$-stopping time. It suffices to show $\mathbb{E}_z\left[M_{t+1} - M_t|\mathcal{G}_t\right] \leq 0$, then we also have integrability of non-negative M_t by induction $\mathbb{E}_z M_t \leq \mathbb{E}_z M_{t-1} \leq ... \leq \mathbb{E}_z M_0 = V(z) < \infty$. Indeed,

$$\begin{aligned}
(M_{t+1} - M_t)\mathbb{I}(T \leq t) &= \left[(V(Z_T) + T\delta_1) - (V(Z_T) + T\delta_1)\right]\mathbb{I}(T \leq t) \\
&= 0 \\
(M_{t+1} - M_t)\mathbb{I}(T \geq t+1) &= \left[(V(Z_{t+1}) + (t+1)\delta_1) - (V(Z_t) + t\delta_1)\right]\mathbb{I}(T \geq t+1) \\
&= \left[V(Z_{t+1}) - V(Z_t) + \delta\right]\mathbb{I}(T \geq t+1)
\end{aligned}$$

implying for $z \in B^c$

$$\begin{aligned}
\mathbb{E}_z\left[M_{t+1} - M_t|\mathcal{G}_t\right] &= \mathbb{E}_z\left[(M_{t+1} - M_t)\mathbb{I}(T \leq t)|\mathcal{G}_t\right] + \mathbb{E}_z\left[(M_{t+1} - M_t)\mathbb{I}(T \geq t+1)|\mathcal{G}_t\right] \\
&= \mathbb{E}_z\left[(V(Z_{t+1}) - V(Z_t) + \delta)\,\mathbb{I}(T \geq t+1)|\mathcal{G}_t\right] \\
&\overset{(i)}{=} \mathbb{E}_z\left[(V(Z_{t+1}) - V(Z_t) + \delta)\,|\mathcal{G}_t\right]\mathbb{I}(T \geq t+1) \\
&\overset{(ii)}{=} \mathbb{E}_z\left[(V(Z_{t+1}) - V(Z_t) + \delta)\,|Z_t\right]\mathbb{I}(T \geq t+1) \\
&\overset{(iii)}{\leq} \left[-\delta + \delta\right]\mathbb{I}(T \geq t+1) \\
&= 0
\end{aligned}$$

where (i) follows T is a $\mathcal{G}_t$-stopping time, and thus $\{T \geq t+1\} \in \mathcal{G}_t$, (ii) is due to the Markov property of $\{Z_t\}$ and (iii) follows $Z_t \in B^c$ (given $T \geq t+1$ and $z \in B^c$) and the drift condition in Definition **??**. Consequently, $\mathbb{E}_z M_t \leq \mathbb{E}_z M_0 = V(z)$ for $z \in B^c$. That is

$$\mathbb{E}_z V(Z_{t\wedge T}) + \mathbb{E}_z(t\wedge T)\delta \leq V(z).$$

implying with non-negativeness of V

$$\mathbb{E}_z(t\wedge T)\delta \leq V(z).$$

Taking $t \to \infty$, the monotone convergence theorem yields

$$\mathbb{E}_z T \leq V(z)/\delta \text{ for } z \in B^c$$

Furthermore, one step analysis gives

$$\begin{aligned}
\mathbb{E}_z T &= \mathbb{P}_z(Z_1 \in B) + \int_{B^c}(1 + \mathbb{E}_{z_1}T)\,p(z, z_1)dz_1 \\
&= 1 + \int_{B^c}(\mathbb{E}_{z_1}T)p(z, z_1)dz_1 \\
&\leq 1 + \frac{1}{\delta}\int_{B^c}V(z_1)p(z, z_1)dz_1
\end{aligned}$$

for $z \in B$, completing the proof. □

Acknowledgements

Lester Mackey, Dr. Rachel Wang and Weijie Su .

References

[1] Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation* **14(8)** 1771–1800.

[2] Hinton, G., S. Osindero and Y. Teh (2006). A fast learning algorithm for deep belief nets. *Neural Computation* **18(7)** 15271554.

[3] Hinton, G. and R. Salakhutdinov (2006). Reducing the dimensionality of data with neural networks. *Science* **313(5786)** 504-507.

[4] Bengio, Y., P. Lamblin, D. Popovici, H. Larochelle, and U. Montreal (2007). Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems* **19** 153.

[5] MacKay, D. (2001). Failures of the one-step learning algorithm. In Available electronically at http://www.inference.phy.cam.ac.uk/mackay/abstracts/gbm.html.

[6] Teh, Y., M. Welling, S. Osindero and G. Hinton (2003). Energy-based models for sparse overcomplete representations. *The Journal of Machine Learning Research* **4** 1235-1260.

[7] Williams, C. K. and F. V. Agakov (2002). An analysis of contrastive divergence learning in gaussian boltzmann machines. *Institute for Adaptive and Neural Computation.*

[8] Yuille (2005). The convergence of contrastive divergence. In *Advances in neural information processing systems* **17** 1593-1600.

[9] Sutskever, I. and Tieleman, T. (2010). On the Convergence Properties of Contrastive Divergence. In *International Conference on Artificial Intelligence and Statistics* 789-795.

[10] Hyvärinen, A. (2006). Consistency of pseudolikelihood estimation of fully visible Boltzmann machines. *Neural Computation* **18(10)** 2283-2292.

[11] Rudolf, D. (2011). Explicit error bounds for Markov chain Monte Carlo. *arXiv preprint arXiv:1108.3201.*

[12] Lehmann, E.L. and Casella, G. (1998). *Theory of point estimation* **31**. Springer Science and Business Media.

[13] Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence.* Springer, New York.

[14] Foster, F. G. (1953). On the stochastic matrices associated with certain queuing processes. *The Annals of Mathematical Statistics* 355–360.

[15] Tweedie, R. L. (1976). Criteria for classifying general Markov chains. *Advances in Applied Probability* 737–771.

[16] Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability.* Springer Science and Business Media.

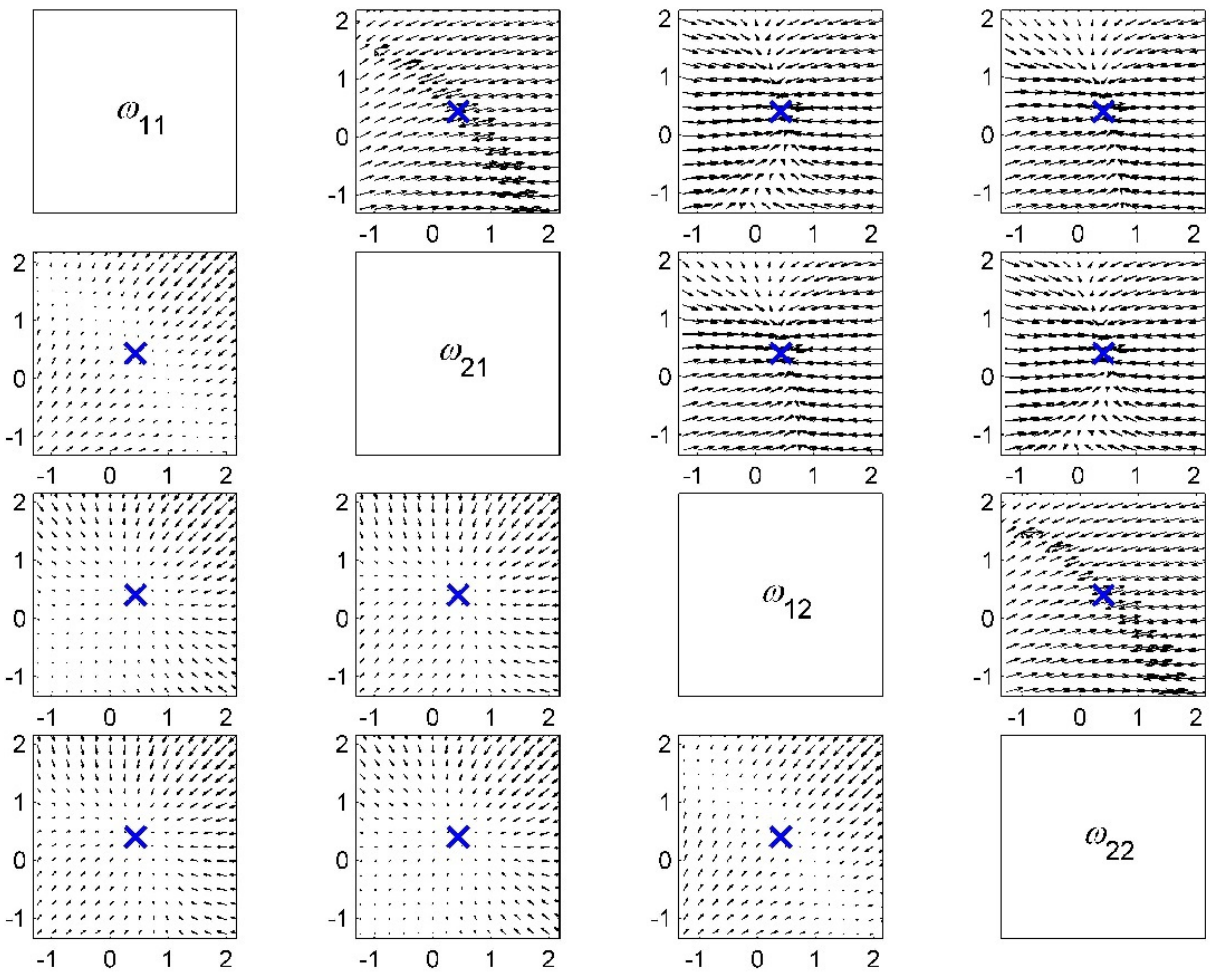

FIG 5. *Simulation results of RBM with $N = 10^4$.*

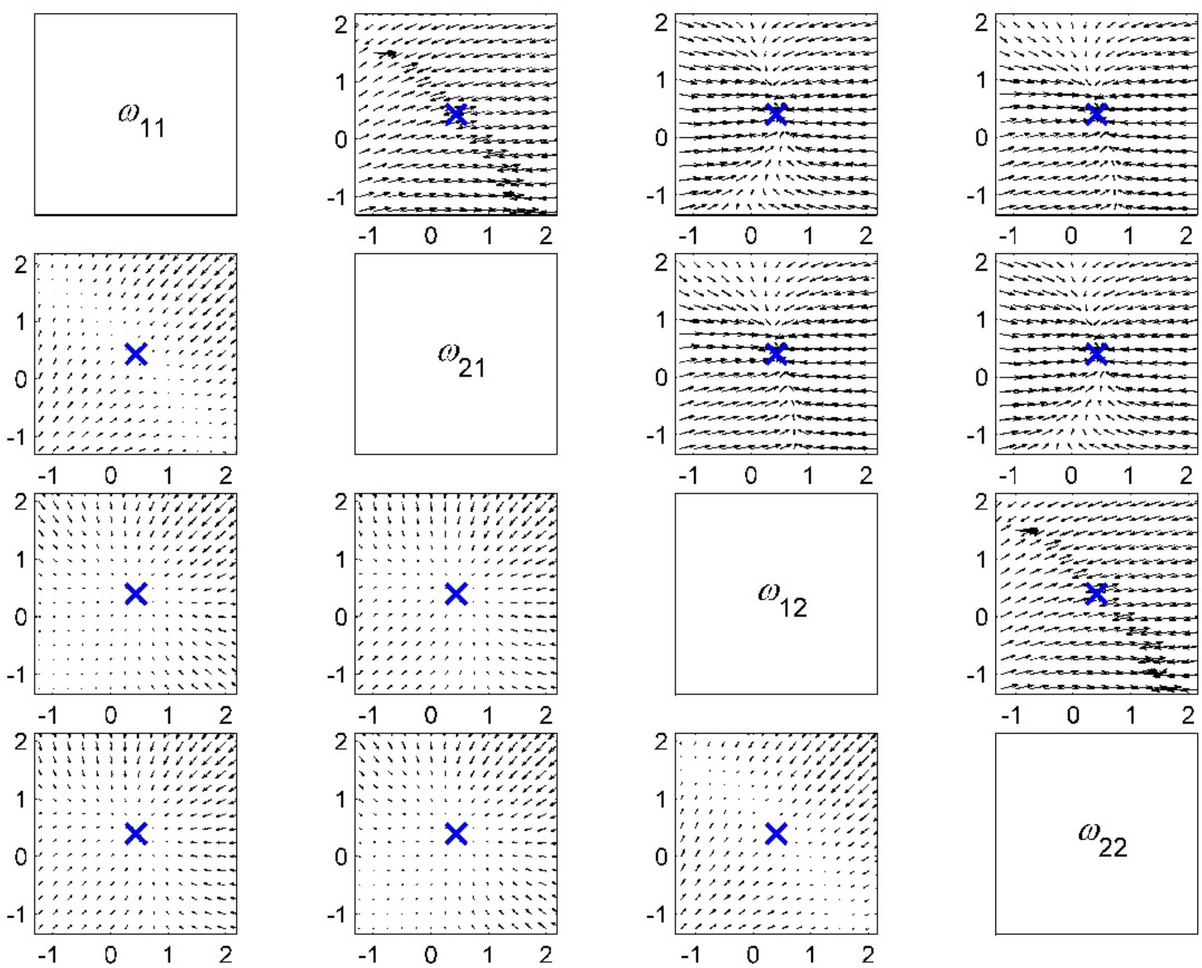

FIG 6. *Simulation results of RBM with $N = 10^6$.*