The Annals of Statistics
2018, Vol. 46, No. 6A, 3067–3098
https://doi.org/10.1214/17-AOS1649

CONVERGENCE OF CONTRASTIVE DIVERGENCE ALGORITHM IN EXPONENTIAL FAMILY[1]

BY BAI JIANG, TUNG-YU WU, YIFAN JIN AND WING H. WONG[2]

Stanford University

The Contrastive Divergence (CD) algorithm has achieved notable success in training energy-based models including Restricted Boltzmann Machines and played a key role in the emergence of deep learning. The idea of this algorithm is to approximate the intractable term in the exact gradient of the log-likelihood function by using short Markov chain Monte Carlo (MCMC) runs. The approximate gradient is computationally-cheap but biased. Whether and why the CD algorithm provides an asymptotically consistent estimate are still open questions. This paper studies the asymptotic properties of the CD algorithm in canonical exponential families, which are special cases of the energy-based model. Suppose the CD algorithm runs m MCMC transition steps at each iteration t and iteratively generates a sequence of parameter estimates $\{\theta_t\}_{t\geq 0}$ given an i.i.d. data sample $\{X_i\}_{i=1}^n \sim p_{\theta_\star}$. Under conditions which are commonly obeyed by the CD algorithm in practice, we prove the existence of some bounded m such that any limit point of the time average $\sum_{s=0}^{t-1}\theta_s/t$ as $t\to\infty$ is a consistent estimate for the true parameter $\theta_\star$. Our proof is based on the fact that $\{\theta_t\}_{t\geq 0}$ is a homogenous Markov chain conditional on the data sample $\{X_i\}_{i=1}^n$. This chain meets the Foster–Lyapunov drift criterion and converges to a random walk around the maximum likelihood estimate. The range of the random walk shrinks to zero at rate $\mathcal{O}(1/\sqrt[3]{n})$ as the sample size $n\to\infty$.

1. Introduction.

1.1. *Exponential family and maximum likelihood learning.* Consider a canonical exponential family over the sample space $\mathcal{X}\subseteq\mathbb{R}^p$ with the parameter $\theta\in\Theta\subseteq\mathbb{R}^d$

$$p_\theta(x) = c(x)e^{\theta^T\phi(x)-\Lambda(\theta)}, \tag{1.1}$$

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

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

Received July 2016; revised September 2017.
[1]Supported by NSF Grant DMS-1407557.
[2]Corresponding author.
MSC2010 subject classifications. Primary 68W99, 62F12; secondary 60J20.
Key words and phrases. Contrastive Divergence, exponential family, convergence rate.

$\Lambda(\theta)$ is convex and differentiable at any interior point of the natural parameter domain. Denote by $\nabla\Lambda(\theta)$ and $\nabla^2\Lambda(\theta)$ the gradient vector and the Hessian matrix of $\Lambda(\theta)$, respectively. They are the expectation and the covariance of the sufficient statistic $\phi(X)$ under p_θ. That is,

$$\nabla\Lambda(\theta) = \mathbb{E}_\theta\phi(X) = \int_{\mathcal{X}} \phi(x)p_\theta(x)\,dx, \tag{1.2}$$

$$\begin{aligned}\nabla^2\Lambda(\theta) &= \operatorname{Cov}_\theta \phi(X)\\ &= \int_{\mathcal{X}} [\phi(x) - \nabla\Lambda(\theta)][\phi(x) - \nabla\Lambda(\theta)]^T p_\theta(x)\,dx.\end{aligned} \tag{1.3}$$

Given an i.i.d. sample $\mathbf{X} = \{X_i\}_{i=1}^n$ following a certain underlying distribution $p_{\theta_\star}$, the log-likelihood function is given by

$$l(\theta) := \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) = \frac{1}{n}\sum_{i=1}^n \log c(X_i) + \theta^T\left[\frac{1}{n}\sum_{i=1}^n \phi(X_i)\right] - \Lambda(\theta).$$

Denote by $g(\theta)$ the gradient of $l(\theta)$:

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

The concavity of $l(\theta)$ follows from the convexity of $\Lambda(\theta)$. Successively iterating the update equation (1.4) of the gradient ascent algorithm will generate a sequence $\{\theta_t\}_{t\geq 0}$ indexed by the iteration number t. This sequence converges to the Maximum Likelihood Estimate (MLE) $\hat{\theta}_n := \arg\max_\theta l(\theta)$ if the learning rate η is suitably chosen:

$$\theta^+ = \theta + \eta g(\theta) = \theta + \eta\left[\frac{1}{n}\sum_{i=1}^n \phi(X_i) - \nabla\Lambda(\theta)\right]. \tag{1.4}$$

In many cases, the term $\nabla\Lambda(\theta)$, essentially an integral as the expectation of the sufficient statistic in light of (1.2), is neither available in a simple closed form nor computationally tractable due to the complexity of the sample space $\mathcal{X}$ and/or the sufficient statistic ϕ. An example is the fully visible Boltzmann Machine (FVBM) model [1]. Its probability mass function is given by

$$p_{W,b}(x) \propto \exp\left(\frac{1}{2}x^T Wx + b^T x\right), \tag{1.5}$$

where x is a p-dimensional vector of binary variables being either -1 or $+1$, W is a $p \times p$ symmetric matrix with zero diagonal entries called "weight matrix" and b is a p-dimensional vector called "bias vector." This model (1.5) is indeed an

exponential family (1.1) with

$$\begin{aligned}\mathcal{X} &= \{-1,+1\}^p,\\ \phi(x) &= (x_i x_j, 1 \le i < j \le p; x_i, 1 \le i \le p) \in \mathbb{R}^{p(p+1)/2},\\ \theta &= (W_{ij}, 1 \le i < j \le p; b_i, 1 \le i \le p) \in \mathbb{R}^{p(p+1)/2},\\ \Lambda(\theta) &= \log \sum_{x\in\mathcal{X}} e^{\theta^T \phi(x)},\end{aligned}$$

and $c(x)$ being the counting measure on the sample space $\mathcal{X}$. The gradient of the cumulant generating function

$$\nabla\Lambda(\theta) = \frac{\sum_{x\in\{-1,+1\}^p} \phi(x) e^{\theta^T\phi(x)}}{\sum_{x\in\{-1,+1\}^p} e^{\theta^T\phi(x)}} = \sum_{x\in\{-1,+1\}^p} \phi(x) p_\theta(x)$$

involves sums of exponentially many terms, which are computationally prohibitive even for a moderately high-dimensional x.

Markov Chain Monte Carlo (MCMC) is the standard approach to approximate $\nabla\Lambda(\theta)$. An MCMC run takes a large number of transition steps to reach the equilibrium, and gradient ascent algorithms iterate the update equation (1.4) hundreds or thousands of times. Thus, it is computationally costly to implement a long MCMC run at each iteration of the gradient ascent algorithm (1.4).

1.2. *Contrastive divergence.* In an influential paper [15], Hinton attempted to alleviate the long MCMC run time by first doing just a small number (say $m = 1, 2$ or 3) of transitions from the data sample $\{X_i\}_{i=1}^n$ as the initial values of the MCMC chains and then using the m-step MCMC sample $\{X_i^{(m)}\}_{i=1}^n$ to approximate $\nabla\Lambda(\theta)$. This is known as Hinton's Contrastive Divergence algorithm, hereafter abbreviated as CD, or CD-m to also specify the fixed number m of transitions. Formally, denote by $k_\theta(x, y)$ the MCMC transition kernel for the equilibrium distribution p_θ. The CD-m algorithm first runs n Markov chains from $\{X_i\}_{i=1}^n$ independently for m steps

$$\begin{aligned} X_1 &\xrightarrow{k_\theta} X_1^{(1)} \xrightarrow{k_\theta} X_1^{(2)} \cdots \xrightarrow{k_\theta} X_1^{(m)},\\ X_2 &\xrightarrow{k_\theta} X_2^{(1)} \xrightarrow{k_\theta} X_2^{(2)} \cdots \xrightarrow{k_\theta} X_2^{(m)},\\ &\vdots\\ X_n &\xrightarrow{k_\theta} X_n^{(1)} \xrightarrow{k_\theta} X_n^{(2)} \cdots \xrightarrow{k_\theta} X_n^{(m)}\end{aligned}$$

and then uses $\{X_i^{(m)}\}_{i=1}^n$ to approximate $\nabla\Lambda(\theta)$ in the update equation (1.4) with

$$\nabla\Lambda(\theta) \approx \frac{1}{n}\sum_{i=1}^n \phi\big(X_i^{(m)}\big).$$

To sum up, the CD-m algorithm replaces $g(\theta)$ in the update equation (1.4) of the gradient ascent algorithm with the CD gradient approximation

$$g_{\mathrm{cd}}(\theta) := \frac{1}{n}\sum_{i=1}^{n}\phi(X_i) - \frac{1}{n}\sum_{i=1}^{n}\phi\big(X_i^{(m)}\big)$$

and iterates the following update equation:

$$\theta^+ = \theta + \eta g_{\mathrm{cd}}(\theta) = \theta + \eta\left[\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) - \frac{1}{n}\sum_{i=1}^{n}\phi\big(X_i^{(m)}\big)\right]. \tag{1.6}$$

The CD algorithm has been widely used by machine learners to train energy-based models of the form $p_\theta(x) = e^{-E(x,\theta)-\Lambda(\theta)}$, where $E(x,\theta)$ is called "energy function" and $\Lambda(\theta)$ is called "log-partition function." A notable example of these energy-based models is the Restricted Boltzmann Machines (RBM), which is the building block on each layer of deep belief network. The CD algorithm has performed well in layer-wise RBM pre-training stage of deep belief network and in this way played an key role in the emergence of deep learning [8, 16, 17, 30]. Applications of CD and RBM also include collaborative filtering [39], classification [24], topic modeling [18] and feature learning [10]. Apart from RBM, the CD algorithm has found practical applications in training energy-based models for acoustic modeling [41], image modeling [14, 37] and coarse-grained protein forcefield learning [42]. Exponential families (1.1) are special cases of energy-based models with $E(x,\theta) = -\theta^T\phi(x) - \log c(x)$. The CD algorithm has also been used to approximate MLEs for exponential-family random graph models (ERGM) [3, 19].

1.3. *Theoretical studies on contrastive divergence.* There are a few open questions concerning the theoretical properties of the CD algorithm. First is whether or under what conditions the sequence $\{\theta_t\}_{t\ge 0}$ generated by the CD algorithm given a data sample $\mathbf{X} = \{X_i\}_{i=1}^n$ converges to some limit points as $t \to \infty$. If the answer is yes, we can regard these limit points as the CD estimates. Next, questions of interest are how close these CD estimates are to the MLE, and whether they are asymptotically consistent for the true parameter $\theta_\star$ as $n \to \infty$.

Many eminent scholars in machine learning have attempted to answer these questions. MacKay [26] provided examples in which the CD-1 algorithm does not converge to the MLE as the iteration number $t \to \infty$. The reason is arguably the bias of the CD gradient approximation. The CD gradient has been proven biased in many models like the Gaussian Boltzmann Machine and RBM, and the bias tends to decrease as m increases in simulation studies [7, 9, 40, 43]. Yuille [44] stated formal conditions for the CD algorithm to converge to the true parameter (rather than the MLE) as $t \to \infty$. But he did not clearly distinguish the behavior of $\{\theta_t\}_{t\ge 0}$ in the limits of the iteration number $t \to \infty$ from that of the sample size $n \to \infty$; thus, his conditions are not satisfied in even the simplest examples such as a bivariate Gaussian model (see more discussion in Section 5). For the FVBM model

(1.5), Hyvärinen [20] showed that a specific CD-1 algorithm with the random-scan Gibbs sampler as the MCMC transition kernel is a stochastic version of an maximum pseudo-likelihood learning process, as the conditional expectation of the CD-1 gradient approximation given the data sample $\{X_i\}_{i=1}^n$ and the current parameter estimate θ is in the direction of the gradient of a certain pseudo-likelihood function. In this way, the author gave a heuristic argument for the consistency of the CD-1 algorithm in the specific setting. However, as the conditional expectation is not actually used in his CD-1 algorithm, this connection to the maximum pseudo-likelihood estimate cannot directly establish the consistency of the CD-1 algorithm.

This paper is devoted to answer the open question whether and why the CD algorithm with some bounded m can yield an asymptotically consistent estimate. We restrict our focus to exponential families rather than general energy-based models for two reasons. First, the convexity of the cumulant generating function $\Lambda(\theta)$ in an exponential family guarantees the uniqueness of the MLE, enabling a transparent comparison of the CD algorithm to the maximum likelihood learning. The bias of the CD gradient approximation and the comparison of the CD estimates and MLE are of primary interest, because the idea of the CD algorithm is to replace the exact gradient with a computationally-cheap but biased CD gradient approximation when doing maximum (log-)likelihood learning. Second, the exponential family itself is a central statistical model. Yet except in the special case when the cumulant generating function is analytically tractable, there is no estimation method known to be asymptotically consistent and computationally efficient. As the CD algorithm appears to be a solution to these cases [3, 19, 23], its consistency for exponential families is of importance.

1.4. *Organization of paper.* In practice, the CD algorithm iterates the update equation (1.6) many times ($t \to \infty$) to obtain an estimate given a particular data sample of size n. Thus, we first study the behavior of $\{\theta_t\}_{t\geq 0}$ in the limit of $t \to \infty$ given a data sample of fixed size n, and then let the sample size $n \to \infty$. The details of our approach can be stated as follows.

Conditional on a data sample of size n, the sequence $\{\theta_t\}_{t\geq 0}$ is a homogenous Markov chain. This chain has two phases: "quick move" and "random walk." When the chain moves from θ which is far away from the MLE $\hat{\theta}_n$, the exact gradient $g(\theta)$ is relatively large compared to the approximation error resulting from the m-step MCMC sampling. The update equation (1.6) keeps pushing θ to quickly move toward $\hat{\theta}_n$. When θ is so close to $\hat{\theta}_n$ that $g(\theta)$ fails to suppress the MCMC approximation error, the "quick move" phase ends and the chain starts a "random walk" in the neighborhood around $\hat{\theta}_n$. This intuition is mathematically formalized as a *Foster–Lyapunov drift criterion* with $V(\theta) = \|\theta - \hat{\theta}_n\|^2/2$ as a *Foster–Lyapunov function* (see Definition 3.2), where $\|z\|$ denotes the l_2-norm of vector z. This idea of a quick move to a random walk neighborhood and different types of Lyapunov drift conditions have been intensively explored in Markov

chain theory [13, 28, 29]. Finally, we show that the random walk neighborhood centering at $\hat{\theta}_n$ shrinks to the true parameter $\theta_\star$ as the sample size $n \to \infty$.

Section 2 states our main result: under six conditions (A1), (A2), (A3), (A4), (A5) and (A6), there exists a bounded m for which the limiting time average $\frac{1}{t}\sum_{s=0}^{t-1}\theta_t$ converges to the true parameter $\theta_\star$ in probability

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

as the sample size $n \to \infty$ at a rate of $1/\sqrt[3]{n}$. Section 3 presents our proof in several stages. First, we show that $\{\theta_t\}_{t\geq 0}$ is a homogenous Markov chain under $\mathbb{P}^{\mathbf{x}}$, the conditional probability measure given any realization of the data sample $\mathbf{X} = \mathbf{x}$, and impose three constraints on $\mathbf{x}$ (and its sample size n). These constraints are proven to hold with probability approaching 1 as $n \to \infty$. Hereafter, we study the chain $\{\theta_t\}_{t\geq 0}$ under $\mathbb{P}^{\mathbf{x}}$ in the framework of the Markov chain and supermartingale theories and demonstrate that a neighborhood around the MLE $\hat{\theta}_n$ is positively recurrent. The key is to establish the Foster–Lyapunov drift criterion with $V(\theta) = \|\theta - \hat{\theta}_n\|^2/2$ as a Foster–Lyapunov function. From the Foster–Lyapunov drift criterion, it follows that

$$\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s - \hat{\theta}_n\right\| = \mathcal{O}\big(1/\sqrt[3]{n}\big) \qquad \mathbb{P}^{\mathbf{x}}\text{-a.s.}$$

Putting this $\mathbb{P}^{\mathbf{x}}$-a.s. convergence result and the fact that $\hat{\theta}_n$ is $\mathcal{O}_p(1/\sqrt{n})$-close to $\theta_\star$ together yields

$$\begin{aligned}\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s - \theta_\star\right\| &\leq \limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s - \hat{\theta}_n\right\| + \|\hat{\theta}_n - \theta_\star\| \\ &= \mathcal{O}_p\big(1/\sqrt[3]{n}\big).\end{aligned}$$

Section 4 uses a bivariate Gaussian model, a FVBM model (1.5) and a exponential-family random graph model (ERGM) as examples to illustrate the theories. Section 5 briefly discusses related works, novelties of our paper in theoretical aspects and guidance to practitioners of the CD algorithm.

2. Main result. We base the asymptotic properties of the CD algorithm in canonical exponential families[3] on the assumptions (A1), (A2), (A3), (A4), (A5) and (A6). These assumptions can be directly verified in many applications of the CD algorithm in canonical exponential families. See three examples in Section 4.

[3] An exponential family is canonical if the d-dimensional sufficient statistic $\phi(X)$ does not satisfy any linear constraint. If so, $\nabla^2\Lambda(\theta) = \mathrm{Cov}_\theta\,\phi(X)$ is positive definite.

(A1) The parameter space of interest Θ is a convex and compact subset of the natural parameter domain $\mathcal{D} = \{\theta \in \mathbb{R}^d : \Lambda(\theta) < \infty\}$, and the true parameter $\theta_\star$ is an interior point of Θ.

The successive iterations of the update equation (1.6) may lead θ^+ to leave compact Θ. If it happens, we project θ^+ onto Θ. The remainder of this paper studies the modified update equation (2.1),

$$\theta^+ = \Pi_\Theta\left(\theta + \eta\left[\frac{1}{n}\sum_{i=1}^n \phi(X_i) - \frac{1}{n}\sum_{i=1}^n \phi\big(X_i^{(m)}\big)\right]\right), \tag{2.1}$$

where Π_Θ denotes the projection mapping onto Θ and is the proximal mapping associated to the convex function $h(\theta) = 0$ if $\theta \in \Theta$ or ∞ otherwise. This proximal mapping trick has been well studied by researchers focusing on the proximal gradient algorithm (see, e.g., [6, 31]).

Let $\lambda_{\min}(\theta), \lambda_{\max}(\theta)$ be the smallest and largest eigenvalues of $\nabla^2\Lambda(\theta)$, and let $\lambda_{\text{sum}}(\theta)$ be the trace (sum of eigenvalues) of $\nabla^2\Lambda(\theta)$. The compactness of Θ in (A1) together with the positive definiteness and continuity of $\nabla^2\Lambda(\theta)$ in a canonical exponential family imply the existence of the following constants:

$$\lambda_{\min} := \inf_{\theta\in\Theta} \lambda_{\min}(\theta) \in (0, \infty), \tag{2.2}$$

$$\lambda_{\max} := \sup_{\theta\in\Theta} \lambda_{\max}(\theta) \in (0, \infty), \tag{2.3}$$

$$\lambda_{\text{sum}} := \sup_{\theta\in\Theta} \lambda_{\text{sum}}(\theta) \in (0, \infty). \tag{2.4}$$

We next explain how to quantify the difference of two distributions p_θ and $p_{\theta_\star}$ in an exponential family. We define χ^2-contrast as follows. This χ^2-contrast is commonly seen in the studies on the MCMC approximation error [33, 38].

DEFINITION 2.1 (χ^2-contrast). Let ν, π be two distributions on $\mathcal{X}$. If there exists a density of ν with respect to π, then denote it by $\frac{d\nu}{d\pi}(x)$. The χ^2-contrast of ν and π is given by

$$\chi^2(\nu, \pi) = \int_{\mathcal{X}} \left[\frac{d\nu}{d\pi}(x) - 1\right]^2 \pi(x)\, dx = \int_{\mathcal{X}} \frac{[\nu(x) - \pi(x)]^2}{\pi(x)}\, dx.$$

Let $\chi(\nu, \pi)$ be the square root of $\chi^2(\nu, \pi)$.

(A2) There exists some positive constant L such that

$$\chi(p_{\theta_\star}, p_\theta) \le L\|\theta - \theta_\star\| \qquad \forall \theta \in \Theta.$$

(A2) is not very restrictive. Indeed, the function $f : \theta \in \Theta \mapsto \chi(p_{\theta_\star}, p_\theta) = \sqrt{e^{-2\Lambda(\theta_\star)+\Lambda(\theta)+\Lambda(2\theta_\star-\theta)} - 1}$ is continuously differentiable, and thus Lipschitz

continuous in $\theta \in$ compact Θ as long as $\Lambda(2\theta_\star - \theta) < \infty$ for any $\theta \in \Theta$. Denote by L the Lipschitz constant then

$$\chi(p_{\theta_\star}, p_\theta) = f(\theta) = f(\theta) - 0 = f(\theta) - f(\theta_\star) \le L\|\theta - \theta_\star\|.$$

This condition holds in exponential families with $\Lambda(\theta) < \infty$ for any $\theta \in \mathbb{R}^d$. These exponential families include the FVBM model (1.5) and the Gaussian model with unknown mean θ and known covariance.

We also need a regularity condition on the MCMC transition kernel. Denote by $k_\theta(x, y)$ the MCMC transition kernel for the equilibrium distribution p_θ, and by K_θ its associated Markov operator. That is, for any function $h : \mathcal{X} \to \mathbb{R}$,

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

The Markov operator K_θ admits an $\mathcal{L}_2$-spectral gap, if

$$\alpha(\theta) := \sup_{h \ne 0} \left\{ \frac{[\int_{\mathcal{X}} |K_\theta h(x)|^2 p_\theta(x)\, dx]^{1/2}}{[\int_{\mathcal{X}} |h(x)|^2 p_\theta(x)\, dx]^{1/2}} : \int_{\mathcal{X}} h(x) p_\theta(x)\, dx = 0 \right\} < 1,$$

where the $\mathcal{L}_2$-spectral gap is given by $1 - \alpha(\theta)$. A larger $\mathcal{L}_2$-spectral gap indicates faster convergence rate of the MCMC chain [38]. (A3) requires that the Markov operators $\{K_\theta\}_{\theta \in \Theta}$ converge to their corresponding equilibrium $\{p_\theta\}_{\theta \in \Theta}$ uniformly fast.

(A3) The Markov operator K_θ admits an $\mathcal{L}_2$-spectral gap $1 - \alpha(\theta)$ and

$$\alpha := \sup_{\theta \in \Theta} \alpha(\theta) < 1.$$

Here, we call $1 - \alpha$ the "uniform $\mathcal{L}_2$-spectral gap" of the Markov operators $\{K_\theta\}_{\theta \in \Theta}$. The spectrum theory for Markov operators has been elegantly established and studied [2, 11, 12]. This assumption is generally obeyed by popular MCMC transition kernels like Metropolis–Hastings algorithms and random-scan Gibbs samplers. These kernels usually generate reversible, φ-irreducible and aperiodic chains in practice [34], and admit $\mathcal{L}_2$-spectral gaps if and only if they are geometrically ergodic (Proposition 1.2 in [22]). See [27, 35, 38] for more detailed discussions on $\mathcal{L}_2$-spectral gap and geometric ergodicity for MCMC algorithms. This "uniform $\mathcal{L}_2$-spectral gap" condition is equivalent to the "uniform geometric ergodicity" condition (H5) assumed in [4], which studies other MCMC-based estimation scheme.

Denote by $k_\theta^m(x, y)$ the m-step transition kernel

$$k_\theta^m(x, y) = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} k_\theta\big(x, x^{(1)}\big) k_\theta\big(x^{(1)}, x^{(2)}\big) \cdots k_\theta\big(x^{(m-1)}, y\big)\, dx^{(1)} \cdots dx^{(m-1)},$$

and by $k_\theta^m \nu$ the m-step transition of a (signed) measure ν on $\mathcal{X}$

$$k_\theta^m \nu(y) = \int_{\mathcal{X}} \nu(x) k_\theta^m(x, y)\, dx.$$

Then $k_\theta^m p_{\theta_\star}$ is the m-step transition of $p_{\theta_\star}$. In the CD algorithm, the m-step MCMC sample $\{X_i^{(m)}\}_{i=1}^n$ can in fact be regarded as i.i.d. draws from $k_\theta^m p_{\theta_\star}$. (A4) assumes that $\phi(X_i^{(m)})$ is sub-exponential.

DEFINITION 2.2 (Sub-exponential random variable). A d-dimensional random variable Y is sub-exponential with parameters (σ, ζ) if

$$\mathbb{E}e^{z^T(Y-\mathbb{E}Y)} \le e^{\sigma^2\|z\|^2/2} \qquad \forall z \in \mathbb{R}^d \text{ s.t. } \|z\| \le \zeta.$$

When $d = 1$, the one-dimensional random variable Y is said to be sub-exponential with parameters (σ, ζ) if

$$\mathbb{E}e^{z(Y-\mathbb{E}Y)} \le e^{\sigma^2 z^2/2} \qquad \forall z \in \mathbb{R} \text{ s.t. } |z| \le \zeta.$$

Apparently, each component of a d-dimensional sub-exponential random variable is a one-dimensional sub-exponential random variable.

(A4) For any $\theta \in \Theta$, if $X^{(m)} \sim k_\theta^m p_{\theta_\star}$ then $\phi(X^{(m)})$ is sub-exponential with some constants (σ_m, ζ_m).

Sub-exponentiality is a commonly-seen condition in many statistics problems nowadays [32]. While many previous theoretical studies [20, 44] on the CD algorithm focused on cases with bounded $\phi(X^{(m)})$, this assumption covers the unbounded cases as long as the tail probability of $\phi(X^{(m)})$ is not heavy. Intuitively, (A4) is expected to holds as $\phi(X)$ is sub-exponential under both p_θ and $p_{\theta_\star}$ in regular exponential families (see Lemma 3.1 in the [21]), and $k_\theta^m p_{\theta_\star}$ lies between the initial distribution $p_{\theta_\star}$ and the equilibrium distribution p_θ. We also directly verify (A4) for the Gaussian, FVBM and ERGM examples in Section 4.

Let $k_\theta^m(x, \cdot)$ denote the m-step distribution of the chain starting from x. Then $X^{(m)}|x, \theta \sim k_\theta^m(x, \cdot)$. We assume that $\phi(X^{(m)})$ is conditionally square integrable given any $x \in \mathcal{X}$ and any $\theta \in \Theta$, and satisfies (A5) and (A6).

(A5) For any $\theta \in \Theta$, $f_\theta : x \mapsto \mathbb{E}[\phi(X^{(m)})|x; \theta] = \int_{\mathcal{X}} \phi(y) k_\theta^m(x, y)\, dy$ is a function of x. We assume that the mapping $\theta \mapsto f_\theta$ is Lipschitz continuous in the sense that there exists some positive constant $C_{1,m}$ (depending on m) such that

$$\sup_{x\in\mathcal{X}} \big\| f_{\theta_1}(x) - f_{\theta_2}(x) \big\| \le C_{1,m}\|\theta_1 - \theta_2\| \qquad \forall \theta_1, \theta_2 \in \Theta.$$

(A6) There exists some positive constant $C_{2,m}$ (depending on m) such that

$$\operatorname{Cov}\big[\phi(X^{(m)})|x, \theta\big] \preceq C_{2,m} I_d \qquad \forall \theta \in \Theta,$$

where I_d denotes the $d \times d$ identity matrix, and $A \preceq B$ means $B - A$ is positive semi-definite for two symmetric matrices A and B.

The intuition behind (A5) is that for two MCMC kernels using similar θ_1 and θ_2, the m-step transitions of the sufficient statistic $\phi(x)$ are similar. (A6) assumes the square integrability of $\phi(X^{(m)})$ under the m-step distribution $X^{(m)}|x;\theta \sim k_\theta^m(x,\cdot)$. They are commonly obeyed by the CD algorithm in practice. See examples in Section 4.

Now Theorem 2.1 states our main result.

THEOREM 2.1. *Assume* (A1), (A2), (A3), (A4), (A5) *and* (A6). *If the CD-m algorithm* (2.1) *generates a sequence* $\{\theta_t\}_{t\geq 0}$ *given an i.i.d. data sample* $X_1,\dots,X_n \sim p_{\theta_\star}$, *then for any* m *and learning rate* η *satisfying*

$$\lambda_{\min} - \sqrt{\lambda_{\text{sum}}}L\alpha^m - \frac{\eta}{2}\big(\lambda_{\max} + \sqrt{\lambda_{\text{sum}}}L\alpha^m\big)^2 > 0, \tag{2.5}$$

one has

$$\lim_{n\to\infty}\mathbb{P}\left(\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s - \theta_\star\right\| > A_m n^{-\gamma/3}\right) = 0$$

for any $\gamma \in (0,1)$ *and some constant* A_m *depending on* m. *Here*, $\lambda_{\min}, \lambda_{\max}, \lambda_{\text{sum}}$ *are defined by* (2.2), (2.3) *and* (2.4), d *is the dimension of* θ *and* $\phi(x)$, L *is the Lipschitz constant introduced by* (A2), $1-\alpha$ *is the uniform* $\mathcal{L}_2$*-spectral gap defined in* (A3) *and* η *is the learning rate of the update equation* (2.1).

The left-hand side of (2.5) goes to $\lambda_{\min} > 0$ as $m \uparrow \infty$ and $\eta \downarrow 0$. There exist bounded m and η to satisfy (2.5). For such m and η, the CD algorithm will give a consistent estimate for $\theta_\star$.

3. Proof. This section presents our proof in several stages. Section 3.1 shows that $\{\theta_t\}_{t\geq 0}$ is a homogenous Markov chain under $\mathbb{P}^{\mathbf{x}}$, the conditional probability measure given any realization of the data sample $\mathbf{X} = \mathbf{x}$, and imposes three constraints on $\mathbf{x}$ (and its sample size n). These constraints are proven to hold with probability approaching 1 as $n \to \infty$. The following subsections analyze the behaviors of the chain $\{\theta_t\}_{t\geq 0}$ under $\mathbb{P}^{\mathbf{x}}$ in the framework of Markov chain and supermartingale theories. Section 3.2 bounds the bias and the variance of the CD gradient approximation. With these bounds, we establish in Section 3.3 the Foster–Lyapunov drift criterion for the chain $\{\theta_t\}_{t\geq 0}$ with $V(\theta) = \|\theta - \hat{\theta}_n\|^2/2$ as a Foster–Lyapunov function. Section 3.4 follows to show that a neighborhood around the MLE $\hat{\theta}_n$ is positively recurrent, and further that

$$\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s - \hat{\theta}_n\right\| = \mathcal{O}(1/\sqrt[3]{n}) \qquad \mathbb{P}^{\mathbf{x}}\text{-a.s.}$$

Putting this $\mathbb{P}^{\mathbf{x}}$-a.s. convergence result and the fact that $\hat{\theta}_n$ is $\mathcal{O}_p(1/\sqrt{n})$-close to $\theta_\star$ together, Section 3.5 yields

$$\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s - \theta_\star\right\| \leq \limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s - \hat{\theta}_n\right\| + \|\hat{\theta}_n - \theta_\star\| = \mathcal{O}_p(1/\sqrt[3]{n}).$$

3.1. *Conditioning on the data sample*. We claim that the CD algorithm (2.1) generates a homogenous Markov chain $\{\theta_t\}_{t\geq 0}$ in the state space Θ conditional on any realization of the data sample $\mathbf{X}=\mathbf{x}$.

Indeed, denote by $\mathbf{X}_t^{(m)}=\{X_{t,i}^{(m)}\}_{1\leq i\leq n}$ the m-step MCMC sample generated at iteration t of the CD algorithm. The filtration

$$\mathcal{F}_t := \sigma\text{-algebra}\big(\mathbf{X},\theta_0,\mathbf{X}_1^{(m)},\theta_1,\mathbf{X}_2^{(m)},\ldots,\theta_{t-1},\mathbf{X}_t^{(m)},\theta_t\big)$$

contains all historical information until iteration t. At each iteration t, the CD update

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

is merely a function of the data sample $\mathbf{X}=\mathbf{x}$, the current parameter estimate θ_{t-1} and the m-step MCMC sample $\mathbf{X}_t^{(m)}$. Given the data sample $\mathbf{X}=\mathbf{x}$ and the current parameter estimate, θ_{t-1}, $\mathbf{X}_t^{(m)}$ is conditionally independent to the past history of CD updates. Thus, $\{\theta_t\}_{t\geq 0}$ is a homogeneous $\mathcal{F}_t$-adapted Markov chain under $\mathbb{P}^{\mathbf{x}}$, the conditional probability measure given $\mathbf{X}=\mathbf{x}$.

Next, we impose three constraints (3.1), (3.2) and (3.3) on the data sample $\mathbf{X}=\mathbf{x}$. Lemma 3.1 proves that they hold with probability approaching 1 as $n\to\infty$. In the following sections, we study the chain $\{\theta_t\}_{t\geq 0}$ under $\mathbb{P}^{\mathbf{x}}$ with $\mathbf{x}$ satisfying these constraints.

LEMMA 3.1. *Assume* (A1), (A4), (A5) *and* $X_1,\ldots,X_n\overset{i.i.d.}{\sim}p_{\theta_\star}$. *Denote by* $\partial\Theta$ *the boundary of compact* Θ, *by* $\hat{\theta}_n$ *the MLE, and by* $f_\theta: x\mapsto\mathbb{E}[\phi(X^{(m)})|x,\theta]$ *the function defined in* (A5). *For any* $\gamma\in(0,1)$,

$$\inf_{\theta\in\partial\Theta}\|\theta-\theta_\star\|>n^{-\gamma/2}, \tag{3.1}$$

$$\|\hat{\theta}_n-\theta_\star\|<n^{-\gamma/2}, \tag{3.2}$$

$$\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^n f_\theta(X_i)-\mathbb{E}f_\theta(X_1)\right\|<n^{-\gamma/2} \tag{3.3}$$

hold with probability approaching 1 *as* $n\to\infty$.

PROOF. (3.1) holds for sufficiently large n since the true parameter $\theta_\star$ is an interior point of compact Θ as assumed in (A1). The standard theorem for the MLE [25] asserts that (3.2) holds with probability approaching 1 as $n\to\infty$. Only left is to show (3.3) holds with probability approaching 1 as $n\to\infty$.

To this end, consider $N=\mathcal{O}(\varepsilon^{-d})$ ε-balls to cover Θ, which center at $\{\theta_l\}_{1\leq l\leq N}$. Any $\theta\in\Theta$ is ε-close to at least one θ_l. (A5) implies

$$\sup_{x\in\mathcal{X}}\|f_\theta(x)-f_{\theta_l}(x)\|\leq C_{1,m}\varepsilon$$

and further

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n} f_\theta(X_i) - \mathbb{E} f_\theta(X_1)\right\| &\le \left\|\frac{1}{n}\sum_{i=1}^{n} f_\theta(X_i) - \frac{1}{n}\sum_{i=1}^{n} f_{\theta_l}(X_i)\right\| \\ &\quad + \left\|\frac{1}{n}\sum_{i=1}^{n} f_{\theta_l}(X_i) - \mathbb{E} f_{\theta_l}(X_1)\right\| \\ &\quad + \left\|\mathbb{E} f_{\theta_l}(X_1) - \mathbb{E} f_\theta(X_1)\right\| \\ &\le \left\|\frac{1}{n}\sum_{i=1}^{n} f_{\theta_l}(X_i) - \mathbb{E} f_{\theta_l}(X_1)\right\| + 2C_{1,m}\varepsilon.\end{aligned}$$

It follows that

$$\begin{aligned}&\mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n} f_\theta(X_i) - \mathbb{E} f_\theta(X_1)\right\| \ge 3C_{1,m}\varepsilon\right) \\ &\quad\le \mathbb{P}\left(\max_{l=1}^{N}\left\|\frac{1}{n}\sum_{i=1}^{n} f_{\theta_l}(X_i) - \mathbb{E} f_{\theta_l}(X_1)\right\| \ge C_{1,m}\varepsilon\right) \\ &\quad\le \sum_{l=1}^{N}\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} f_{\theta_l}(X_i) - \mathbb{E} f_{\theta_l}(X_1)\right\| \ge C_{1,m}\varepsilon\right) \\ &\quad\le \sum_{l=1}^{N}\sum_{j=1}^{d}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} f_{\theta_l,j}(X_i) - \mathbb{E} f_{\theta_l,j}(X_1)\right| \ge \frac{C_{1,m}\varepsilon}{\sqrt{d}}\right).\end{aligned}$$

$\phi(X_i^{(m)})$ is sub-exponential with (σ_m, ζ_m) as assumed in (A4), so is its conditional expectation $f_\theta(X_i) = \mathbb{E}[\phi(X_i^{(m)})|X_i, \theta]$ by Lemma 3.2 in the [21]. Let $f_{\theta,j}$ be the jth component of f_θ then one-dimensional random variables $\{f_{\theta,j}(X_i)\}_{i=1}^n$ are i.i.d. and sub-exponential with (σ_m, ζ_m). By Lemma 3.3 in the [21],

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} f_{\theta_l,j}(X_i) - \mathbb{E} f_{\theta_l,j}(X_1)\right| \ge \frac{C_{1,m}\varepsilon}{\sqrt{d}}\right) \le 2\exp\left(-\frac{nC_{1,m}^2\varepsilon^2/d}{2\sigma_m^2}\right)$$

if $C_{1,m}\varepsilon/\sqrt{d} < \sigma_m^2\zeta_m$. Putting together with the fact that $N = \mathcal{O}(\varepsilon^{-d})$ yields

$$\begin{aligned}\mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n} f_\theta(X_i) - \mathbb{E} f_\theta(X_1)\right\| \ge 3C_{1,m}\varepsilon\right) &\le 2Nd\exp\left(-\frac{nC_{1,m}^2\varepsilon^2/d}{2\sigma_m^2}\right) \\ &= \mathcal{O}(\varepsilon^{-d})\exp\left(-\frac{nC_{1,m}^2\varepsilon^2/d}{2\sigma_m^2}\right)\end{aligned}$$

if $C_{1,m}\varepsilon/\sqrt{d} < \sigma_m^2\zeta_m$. Let $\varepsilon = n^{-\gamma/2}/3C_{1,m}$ then $C_{1,m}\varepsilon/\sqrt{d} < \sigma_m^2\zeta_m$ if n is sufficiently large. Thus,

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n} f_\theta(X_i) - \mathbb{E}f_\theta(X_1)\right\| \geq n^{-\gamma/2}\right) \leq \mathcal{O}\big(n^{\gamma d/2}\big)\exp\left(-\frac{n^{1-\gamma}}{18d\sigma_m^2}\right) \to 0$$

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

3.2. *Gradient approximation error.* Denote by $\mathbb{P}_\theta^{\mathbf{x}}$ the conditional probability measure given the data sample $\mathbf{x}$ and the current state $\theta_t = \theta$ of the chain. And $\mathbb{E}_\theta^{\mathbf{x}}$ and $\mathbb{C}\mathrm{ov}_\theta^{\mathbf{x}}$ denote the expectation and covariance under $\mathbb{P}_\theta^{\mathbf{x}}$. Lemma 3.2 bounds the approximation error of the CD gradient $g_{\mathrm{cd}}(\theta)$ under $\mathbb{P}_\theta^{\mathbf{x}}$. Specifically, the bias of $g_{\mathrm{cd}}(\theta)$ is $\mathcal{O}(n^{-\gamma/2}) + \mathcal{O}(\alpha^m\|\theta - \hat{\theta}_n\|)$, which depends on the uniform $\mathcal{L}_2$-spectral gap $1-\alpha$ of the MCMC kernels, the number m of transition steps in MCMC, the sample size n and the distance between θ and the MLE $\hat{\theta}_n$. This result agrees with Bengio and Delalleau's [7] finding that the CD gradient approximation error decreases at a rate depending on the mixing rate of the MCMC kernels.

LEMMA 3.2. *Assume* (A1), (A2), (A3), (A6) *and that the data sample* $\mathbf{x} = \{x_i\}_{i=1}^n$ *satisfies* (3.2) *and* (3.3). *Let* $\Delta g = g_{\mathrm{cd}}(\theta) - g(\theta)$ *be the gradient approximation error. Then*

$$\|\mathbb{E}_\theta^{\mathbf{x}}\Delta g\| \leq \big(1 + \sqrt{\lambda_{\mathrm{sum}}}L\alpha^m\big)n^{-\gamma/2} + \sqrt{\lambda_{\mathrm{sum}}}L\alpha^m\|\theta - \hat{\theta}_n\|,$$

where λ_{sum} *is defined in* (2.4), L *denotes the Lipschitz constant introduced by* (A2), $1-\alpha$ *is the uniform* $\mathcal{L}_2$*-spectral gap defined in* (A3), *and* $\gamma \in (0,1)$ *is introduced by constraints* (3.2) *and* (3.3). *Also*,

$$\mathbb{C}\mathrm{ov}_\theta^{\mathbf{x}}\,\Delta g \preceq \frac{C_{2,m}}{n}I_d,$$

where $C_{2,m}$ *is defined in* (A6).

PROOF. From the fact that

$$\Delta g = g_{\mathrm{cd}}(\theta) - g(\theta) = \nabla\Lambda(\theta) - \frac{1}{n}\sum_{i=1}^{n}\phi\big(X_i^{(m)}\big),$$

it follows that

$$\begin{aligned}-\mathbb{E}^{\mathbf{x}}\Delta g &= \mathbb{E}_\theta^{\mathbf{x}}\left[\frac{1}{n}\sum_{i=1}^{n}\phi\big(X_i^{(m)}\big)\right] - \nabla\Lambda(\theta)\\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}}\phi(y)k_\theta^m(x_i,y)\,dy - \nabla\Lambda(\theta) \qquad [X_i^{(m)}|x_i,\theta \sim k_\theta^m(x_i,\cdot)]\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} f_\theta(x_i) - \nabla\Lambda(\theta) && \text{[Definition of } f_\theta \text{ in (A5)]}\\
&= \left[\frac{1}{n}\sum_{i=1}^{n} f_\theta(x_i) - \mathbb{E} f_\theta(X_1)\right]\\
&\quad + \left[\mathbb{E} f_\theta(X_1) - \nabla\Lambda(\theta)\right].
\end{aligned}
$$

Constraint (3.3) bounds the length of the first term (vector) by $n^{-\gamma/2}$. Proceed to consider the second term $\mathbb{E} f_\theta(X_1) - \nabla\Lambda(\theta)$. Write

$$
\begin{aligned}
\mathbb{E} f_\theta(X_1) &= \int_{\mathcal{X}} f_\theta(x) p_{\theta_\star}(x)\,dx && [X_1 \sim p_{\theta_\star}]\\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \phi(y) k_\theta^m(x,y)\,dy\right) p_{\theta_\star}(x)\,dx && \text{[Definition of } f_\theta \text{ in (A5)]}\\
&= \int_{\mathcal{X}} \phi(y) \left(\int_{\mathcal{X}} k_\theta^m(x,y) p_{\theta_\star}(x)\,dx\right) dy && \text{[Fubini's theorem]}\\
&= \int_{\mathcal{X}} \phi(y) k_\theta^m p_{\theta_\star}(y)\,dy. && \text{[Definition of } k_\theta^m p_{\theta_\star}]
\end{aligned}
$$

The fact that

$$
\nabla\Lambda(\theta) = \int_{\mathcal{X}} \nabla\Lambda(\theta) k_\theta^m p_{\theta_\star}(x)\,dx = \int_{\mathcal{X}} \nabla\Lambda(\theta) p_\theta(x)\,dx
$$

and (1.2) imply

$$
\begin{aligned}
\mathbb{E} f_\theta(X_1) - \nabla\Lambda(\theta) &= \int_{\mathcal{X}} \phi(x) k_\theta^m p_{\theta_\star}(x)\,dx - \int_{\mathcal{X}} \phi(x) p_\theta(x)\,dx\\
&\quad - \int_{\mathcal{X}} \nabla\Lambda(\theta) k_\theta^m p_{\theta_\star}(x)\,dx + \int_{\mathcal{X}} \nabla\Lambda(\theta) p_\theta(x)\,dx\\
&= \int_{\mathcal{X}} \left[\phi(x) - \nabla\Lambda(\theta)\right]\left[k_\theta^m p_{\theta_\star}(x) - p_\theta(x)\right] dx.
\end{aligned}
$$

For each $j = 1, \ldots, d$, let $f_{\theta,j}(x), \phi_j(x)$ and $\nabla_j\Lambda(\theta) = \partial\Lambda(\theta)/\partial\theta_j$ be the jth component of $f_\theta(x), \phi(x)$ and $\nabla\Lambda(\theta)$, respectively. Let $\nabla_{jj}^2\Lambda(\theta) = \partial^2\Lambda(\theta)/\partial\theta_j^2$ be the jth diagonal entry of $\nabla^2\Lambda(\theta)$:

$$
\begin{aligned}
&\left|\mathbb{E} f_{\theta,j}(X_1) - \nabla_j\Lambda(\theta)\right|\\
&\quad = \left|\int_{\mathcal{X}} \left[\phi_j(x) - \nabla_j\Lambda(\theta)\right]\left[k_\theta^m p_{\theta_\star}(x) - p_\theta(x)\right] dx\right|\\
&\quad = \left|\int_{\mathcal{X}} \left[\phi_j(x) - \nabla_j\Lambda(\theta)\right]\left[\frac{k_\theta^m p_{\theta_\star}(x)}{p_\theta(x)} - 1\right] p_\theta(x)\,dx\right|\\
&\quad \le \sqrt{\int_{\mathcal{X}} \left[\phi_j(x) - \nabla_j\Lambda(\theta)\right]^2 p_\theta(x)\,dx}
\end{aligned}
$$

$$\times \sqrt{\int_{\mathcal{X}} \left[\frac{k_\theta^m p_{\theta_\star}(x)}{p_\theta(x)} - 1\right]^2 p_\theta(x)\,dx} \qquad \text{[Cauchy–Schwarz]}$$

$$= \sqrt{\nabla^2_{jj}\Lambda(\theta)} \times \chi\big(k_\theta^m p_{\theta_\star}, p_\theta\big). \qquad \text{[(1.3), Definition 2.1]}$$

Noting that $\lambda_{\text{sum}}(\theta) = \text{trace}[\nabla^2\Lambda(\theta)]$ and that $\chi(k_\theta^m p_{\theta_\star}, p_\theta) \le \alpha(\theta)^m \chi(p_{\theta_\star}, p_\theta)$ due to Lemma 3.4 in the Supplementary Material [21] (also part of Theorem 2.1 in [33] and Proposition 3.12 in [38]), we further have

$$\big\|\mathbb{E} f_\theta(X_1) - \nabla\Lambda(\theta)\big\| \le \sqrt{\lambda_{\text{sum}}(\theta)} \times \alpha(\theta)^m \chi(p_{\theta_\star}, p_\theta). \tag{3.4}$$

Hence

$$\begin{aligned}
\big\|\mathbb{E}^{\mathbf{x}}_\theta \Delta g\big\| &\le \left\| \frac{1}{n}\sum_{i=1}^n f_\theta(x_i) - \mathbb{E} f_\theta(X_1)\right\| \\
&\quad + \big\|\mathbb{E} f_\theta(X_1) - \nabla\Lambda(\theta)\big\| \\
&\le n^{-\gamma/2} + \sqrt{\lambda_{\text{sum}}(\theta)} \times \alpha(\theta)^m \chi(p_{\theta_\star}, p_\theta) && \text{[(3.3), (3.4)]} \\
&\le n^{-\gamma/2} + \sqrt{\lambda_{\text{sum}}} L\alpha^m \|\theta_\star - \theta\| && \text{[(2.4), (A2), (A3)]} \\
&\le n^{-\gamma/2} + \sqrt{\lambda_{\text{sum}}} L\alpha^m \|\hat\theta_n - \theta_\star\| \\
&\quad + \sqrt{\lambda_{\text{sum}}} L\alpha^m \|\theta - \hat\theta_n\| \\
&\le \big(1 + \sqrt{\lambda_{\text{sum}}} L\alpha^m\big) n^{-\gamma/2} + \sqrt{\lambda_{\text{sum}}} L\alpha^m \|\theta - \hat\theta_n\| && \text{[(3.2)].}
\end{aligned}$$

$X_i^{(m)} | \mathbf{x}, \theta \sim k_\theta^m(x_i, \cdot)$ are conditionally independent (but not identically distributed) since n chains independently starts from different x_i. By (A6),

$$\mathbb{C}\text{ov}^{\mathbf{x}}_\theta \Delta g = \frac{1}{n^2}\sum_{i=1}^n \mathbb{C}\text{ov}\big[\phi\big(X_i^{(m)}\big) | x_i, \theta\big] \preceq \frac{C_{2,m}}{n} I_d. \qquad \square$$

3.3. *Foster–Lyapunov drift criterion.* This subsection uses the error bounds of CD gradient approximation in Lemma 3.2 to establish the *Foster–Lyapunov drift criterion* with

$$V(\theta) := \|\theta - \hat\theta_n\|^2/2$$

as a *Foster–Lyapunov function*. See Definitions 3.1 and 3.2 for drift, Foster–Lyapunov drift criterion and the Foster–Lyapunov function.

DEFINITION 3.1 (drift). Let $V : \Theta \to \mathbb{R}_+$ be some nonnegative function on the state space of a Markov chain $\{\theta_t\}_{t\ge 0}$. The one-step drift of V is defined as $\mathbb{E}_\theta V(\theta^+) - V(\theta)$, which is the expected value change of V when the chain moves from θ to θ^+.

DEFINITION 3.2 (Foster–Lyapunov drift criterion). A Markov chain $\{\theta_t\}_{t\le 0}$ satisfies the Foster–Lyapunov drift criterion if

$$\mathbb{E}_\theta V(\theta^+) - V(\theta) \le -\delta_1 \mathbb{I}(\theta \notin B) + \delta_2 \mathbb{I}(\theta \in B)$$

with some $\delta_1, \delta_2 > 0$ and some subset B of the state space Θ. V is called a Foster–Lyapunov function.

Lemma 3.3 shows that the chain $\{\theta_t\}_{t\ge 0}$ satisfies the *Foster–Lyapunov drift criterion*.

LEMMA 3.3. *Assume* (A1), (A2), (A3), (A6) *and that the data sample* $\mathbf{x}$ *satisfies* (3.1), (3.2) *and* (3.3) *for some* $\gamma \in (0, 1)$. *Then for any* m *and learning rate* η *satisfying*

$$a := \lambda_{\min} - \sqrt{\lambda_{\text{sum}}} L\alpha^m - \frac{\eta}{2}\big(\lambda_{\max} + \sqrt{\lambda_{\text{sum}}} L\alpha^m\big)^2 > 0,$$

the chain $\{\theta_t\}_{t\ge 0}$ *satisfies Foster–Lyapunov drift criterion*

$$\mathbb{E}_\theta^{\mathbf{x}} V(\theta^+) - V(\theta) \le -\delta_1 \mathbb{I}(\theta \notin B) + \delta_2 \mathbb{I}(\theta \in B),$$

with the Foster–Lyapunov function

$$V(\theta) = \|\theta - \hat{\theta}_n\|^2/2$$

and

$$\begin{aligned} B &= \{\theta \in \Theta : \|\theta - \hat{\theta}_n\| \le \beta r_n\}, \\ \delta_1 &= \eta(\beta^2 - 1)c_n, \\ \delta_2 &= \eta(c_n + b_n^2/4a). \end{aligned}$$

Here, $\beta > 1$ *is arbitrary and* b_n, c_n, r_n *are defined in the following way*:

$$\begin{aligned} b_n &:= \big(1 + \sqrt{\lambda_{\text{sum}}} L\alpha^m\big)\big(1 + \eta\lambda_{\max} + \eta\sqrt{\lambda_{\text{sum}}} L\alpha^m\big) n^{-\gamma/2}, \\ c_n &:= \frac{\eta}{2}\big[dC_{2,m} n^{-1+\gamma} + \big(1 + \sqrt{\lambda_{\text{sum}}} L\alpha^m\big)^2\big] n^{-\gamma}, \\ r_n &:= \frac{b_n + \sqrt{b_n^2 + 4ac_n}}{2a} \asymp n^{-\gamma/2}. \end{aligned}$$

PROOF. Let Π_Θ denote the projection mapping onto Θ. From (3.1) and (3.2), it follows that $\hat{\theta}_n \in \Theta$, further implying that

$$\hat{\theta}_n = \Pi_\Theta(\hat{\theta}_n).$$

Let $\Delta g = g_{\mathrm{cd}}(\theta) - g(\theta)$ be the approximation error of the CD gradient. We analyze the one-step drift of the update equation (2.1). Write

$$\begin{aligned}
V(\theta^+) &= \frac{1}{2}\|\theta^+ - \hat{\theta}_n\|^2 \\
&= \frac{1}{2}\|\Pi_\Theta(\theta + \eta g_{\mathrm{cd}}(\theta)) - \Pi_\Theta(\hat{\theta}_n)\|^2 \\
&\le \frac{1}{2}\|\theta + \eta g_{\mathrm{cd}}(\theta) - \hat{\theta}_n\|^2 \\
&= V(\theta) + \eta(\theta - \hat{\theta}_n)^T g_{\mathrm{cd}}(\theta) + \frac{\eta^2}{2}\|g_{\mathrm{cd}}(\theta)\|^2 \\
&= V(\theta) + \eta(\theta - \hat{\theta}_n)^T g(\theta) + \eta(\theta - \hat{\theta}_n)^T \Delta g + \frac{\eta^2}{2}\|g_{\mathrm{cd}}(\theta)\|^2,
\end{aligned}$$

implying the one-step drift

$$\begin{aligned}
&\mathbb{E}^{\mathbf{x}}_\theta V(\theta^+) - V(\theta) \\
&\le \eta(\theta - \hat{\theta}_n)^T g(\theta) + \eta(\theta - \hat{\theta}_n)^T \mathbb{E}^{\mathbf{x}}_\theta \Delta g + \frac{\eta^2}{2}\mathbb{E}^{\mathbf{x}}_\theta[\|g_{\mathrm{cd}}(\theta)\|^2] \\
&\le \eta(\theta - \hat{\theta}_n)^T g(\theta) + \eta(\theta - \hat{\theta}_n)^T \mathbb{E}^{\mathbf{x}}_\theta \Delta g \\
&\quad + \frac{\eta^2}{2}\|\mathbb{E}^{\mathbf{x}}_\theta g_{\mathrm{cd}}(\theta)\|^2 + \frac{\eta^2}{2}\operatorname{trace}[\operatorname{Cov}^{\mathbf{x}}_\theta \Delta g] \\
&\le \eta(\theta - \hat{\theta}_n)^T g(\theta) + \eta\|\theta - \hat{\theta}_n\|\|\mathbb{E}^{\mathbf{x}}_\theta \Delta g\| \\
&\quad + \frac{\eta^2}{2}(\|g(\theta)\| + \|\mathbb{E}^{\mathbf{x}}_\theta \Delta g\|)^2 + \frac{\eta^2}{2}\operatorname{trace}[\operatorname{Cov}^{\mathbf{x}}_\theta \Delta g].
\end{aligned} \tag{3.5}$$

From the facts that $g(\hat{\theta}_n) = 0$ and that $\nabla g(\theta) = -\nabla^2 \Lambda(\theta)$, it follows that

$$g(\theta) = g(\theta) - g(\hat{\theta}_n) = -\nabla^2 \Lambda(\theta')(\theta - \hat{\theta}_n)$$

for some θ' between θ and $\hat{\theta}_n$. In the first term of the right-hand side of (3.5),

$$\begin{aligned}
(\theta - \hat{\theta}_n)^T g(\theta) &= -(\theta - \hat{\theta}_n)^T \nabla^2 \Lambda(\theta')(\theta - \hat{\theta}_n) \\
&\le -\lambda_{\min}\|\theta - \hat{\theta}_n\|^2.
\end{aligned}$$

In the third term of the right-hand side of (3.5),

$$\begin{aligned}
\|g(\theta)\| &= \|\nabla^2 \Lambda(\theta')(\theta - \hat{\theta}_n)\| \\
&= \sqrt{(\theta - \hat{\theta}_n)^T [\nabla^2 \Lambda(\theta')]^2 (\theta - \hat{\theta}_n)} \\
&\le \sqrt{\lambda^2_{\max}\|\theta - \hat{\theta}_n\|^2} \\
&= \lambda_{\max}\|\theta - \hat{\theta}_n\|.
\end{aligned}$$

Plugging them and results in Lemma 3.2 together into (3.5) yields

$$\begin{aligned}
&\mathbb{E}^{\mathbf{x}}_{\theta} V(\theta^+) - V(\theta) \\
&\quad \le -\eta \lambda_{\min} \|\theta - \hat{\theta}_n\|^2 \\
&\qquad + \eta\big[(1 + \sqrt{\lambda_{\text{sum}}} L \alpha^m) n^{-\gamma/2} \\
&\qquad + \sqrt{\lambda_{\text{sum}}} L \alpha^m \|\theta - \hat{\theta}_n\|\big] \|\theta - \hat{\theta}_n\| \\
&\qquad + \frac{\eta^2}{2}\big[\lambda_{\max} \|\theta - \hat{\theta}_n\| + (1 + \sqrt{\lambda_{\text{sum}}} L \alpha^m) n^{-\gamma/2} \\
&\qquad + \sqrt{\lambda_{\text{sum}}} L \alpha^m \|\theta - \hat{\theta}_n\|\big]^2 + \frac{\eta^2}{2} \times \frac{d C_{2,m}}{n} \\
&\quad = -\eta\big(a \|\theta - \hat{\theta}_n\|^2 - b_n \|\theta - \hat{\theta}_n\| - c_n\big), \qquad (3.6)
\end{aligned}$$

whose right-hand side is quadratic in $\|\theta - \hat{\theta}_n\|$. If $a > 0$, then large $\|\theta - \hat{\theta}_n\|$ guarantees a negative drift. Specifically,

$$\|\theta - \hat{\theta}_n\| \ge r_n := \frac{b_n + \sqrt{b_n^2 + 4ac_n}}{2a} \quad \Longrightarrow \quad \mathbb{E}^{\mathbf{x}}_{\theta} V(\theta^+) - V(\theta) \le 0.$$

For any $\beta > 1$, let

$$B := \{\theta \in \Theta : \|\theta - \hat{\theta}_n\| \le \beta r_n\}.$$

If $\theta \notin B$, that is, $\|\theta - \hat{\theta}_n\| \ge \beta r_n$ then

$$\begin{aligned}
\mathbb{E}^{\mathbf{x}}_{\theta} V(\theta^+) - V(\theta) &\le -\eta\big(a\beta^2 r_n^2 - b_n \beta r_n - c_n\big) \\
&= -\eta\big[a(\beta^2 - 1) r_n^2 - b_n(\beta - 1) r_n\big] \\
&\le -\eta(\beta^2 - 1)(a r_n^2 - b r_n) \\
&= -\eta(\beta^2 - 1) c_n \\
&= -\delta_1.
\end{aligned}$$

On the other hand, if $\theta \in B$,

$$\begin{aligned}
\mathbb{E}^{\mathbf{x}}_{\theta} V(\theta^+) - V(\theta) &\le \max_{\theta \in \Theta} -\eta\big(a \|\theta - \hat{\theta}_n\|^2 - b_n \|\theta - \hat{\theta}_n\| - c_n\big) \\
&= \eta(c_n + b_n^2/4a) \\
&= \delta_2,
\end{aligned}$$

completing the proof. □

Figure 1 illustrates the intuition of Lemma 3.3 that the drift of $V(\theta) = \|\theta - \hat{\theta}_n\|^2/2$ is upper bounded by a quadratic function of $\|\theta - \hat{\theta}_n\|$. This bound later implies the drift criterion that V decreases at least δ_1 after a move from B^c, and increases at most δ_2 after a move from B.

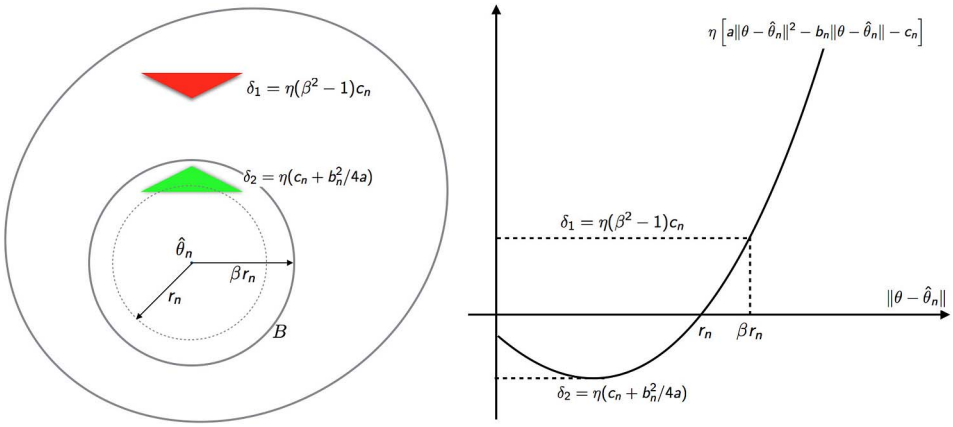

FIG. 1. *Equation* (3.6) *asserts that the drift of* $V(\theta) = \|\theta - \hat{\theta}_n\|^2/2$ *is upper bounded by a quadratic function of* $\|\theta - \hat{\theta}_n\|$. *It implies the drift criterion that* V *decreases at least* δ_1 *after a move from* B^c, *and increases at most* δ_2 *after a move from* B.

3.4. *Positive recurrence around the MLE under the chain.* Lemma 3.4 follows to show that B is positive recurrent under the chain $\{\theta_t\}_{t\geq 0}$ in the sense that the proportion of time that the chain stays inside of B in the long term is at least $\delta_1/(\delta_1 + \delta_2)$. The proof of Lemma 3.4 uses the Azuma–Hoeffding inequality [5] for supermartingales with bounded differences.

LEMMA 3.4. *Following Lemma* 3.3, $B = \{\theta \in \Theta : \|\theta - \hat{\theta}_n\| \leq \beta r_n\}$ *is positive recurrent under the chain* $\{\theta_t\}_{t\geq 0}$ *in the sense that*

$$\liminf_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} \mathbb{I}(\theta_s \in B) \geq \frac{\delta_1}{\delta_1 + \delta_2} = \frac{\beta^2 - 1}{\beta^2 + b_n^2/4ac_n},$$

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} \mathbb{I}(\theta_s \notin B) \leq \frac{\delta_2}{\delta_1 + \delta_2} = \frac{1 + b_n^2/4ac_n}{\beta^2 + b_n^2/4ac_n}.$$

PROOF. First, construct two supermartingales with bounded differences. Let

$$Y_t = V(\theta_{t+1}) - V(\theta_t) + \delta_1,$$
$$Z_t = V(\theta_{t+1}) - V(\theta_t) - \delta_2.$$

By the Foster–Lyapunov drift criterion established in Lemma 3.3 and the Markov property of $\{\theta_t\}_{t\geq 0}$,

$$\sum_{s=0}^{t-1} Y_s \mathbb{I}(\theta_s \notin B), \qquad \sum_{s=0}^{t-1} Z_s \mathbb{I}(\theta_s \in B)$$

are supermartingales under $\mathbb{P}^{\mathbf{x}}$. They have bounded differences $|Y_t| < D, |Z_t| < D$ for some D since $V(\theta)$ is bounded on compact Θ. By the Azuma–Hoeffding inequality (Lemma 3.5 in the [21]), for any $\varepsilon > 0$,

$$\mathbb{P}^{\mathbf{x}}\left(\frac{1}{t}\sum_{s=0}^{t-1} Y_s \mathbb{I}(\theta_s \notin B) \geq \varepsilon\right) = \mathbb{P}^{\mathbf{x}}\left(\sum_{s=0}^{t-1} Y_s \mathbb{I}(\theta_s \notin B) \geq t\varepsilon\right)$$
$$\leq \exp\left(-\frac{t\varepsilon^2}{2D^2}\right).$$

By Borel–Cantelli lemma, $\sum_{t=0}^{\infty} \exp(-\frac{t\varepsilon^2}{2D^2}) < \infty$ implies

$$\mathbb{P}^{\mathbf{x}}\left(\frac{1}{t}\sum_{s=0}^{t-1} Y_s \mathbb{I}(\theta_s \notin B) \geq \varepsilon \text{ finitely often}\right) = 0.$$

That is,

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} Y_s \mathbb{I}(\theta_s \notin B) \leq \varepsilon \qquad \mathbb{P}^{\mathbf{x}}\text{-a.s.}$$

$\varepsilon > 0$ can be arbitrarily small. It follows that

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} Y_s \mathbb{I}(\theta_s \notin B) \leq 0 \qquad \mathbb{P}^{\mathbf{x}}\text{-a.s.}$$

Similarly,

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} Z_s \mathbb{I}(\theta_s \in B) \leq 0 \qquad \mathbb{P}^{\mathbf{x}}\text{-a.s.}$$

Combining them with the fact that

$$\frac{1}{t}\sum_{s=0}^{t-1}(\delta_1 + \delta_2)\mathbb{I}(\theta_s \in B)$$
$$= \frac{1}{t}\sum_{s=0}^{t-1}(Y_s - Z_s)\mathbb{I}(\theta_s \in B)$$
$$= \frac{1}{t}\sum_{s=0}^{t-1} Y_s - \frac{1}{t}\sum_{s=0}^{t-1} Y_s \mathbb{I}(\theta_s \notin B) - \frac{1}{t}\sum_{s=0}^{t-1} Z_s \mathbb{I}(\theta_s \in B)$$

yields

$$\liminf_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} \mathbb{I}(\theta_s \in B) \geq \liminf_{t\to\infty} \frac{1}{(\delta_1 + \delta_2)t}\sum_{s=0}^{t-1} Y_s$$

$$= \liminf_{t\to\infty} \frac{V(\theta_t) - V(\theta_0) + \delta_1 t}{(\delta_1 + \delta_2)t}$$

$$= \frac{\delta_1}{\delta_1 + \delta_2} = \frac{\beta^2 - 1}{\beta^2 + b_n^2/4ac_n} \qquad \mathbb{P}^{\mathbf{x}}\text{-a.s.}$$

It is equivalent to say

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \mathbb{I}(\theta_s \notin B) \le \frac{\delta_2}{\delta_1 + \delta_2} = \frac{1 + b_n^2/4ac_n}{\beta^2 + b_n^2/4ac_n} \qquad \mathbb{P}^{\mathbf{x}}\text{-a.s.},$$

completing the proof. □

This lemma shows that the proportion of time that the chain stays inside of B in the long term is at least

$$\liminf_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \mathbb{I}(\theta_s \in B) \ge \frac{\delta_1}{\delta_1 + \delta_2} = \frac{\beta^2 - 1}{\beta^2 + b_n^2/4ac_n} = \frac{\beta^2 - 1}{\beta^2 + \mathcal{O}(1)}.$$

Note that the radius of the closed ball is βr_n. Letting $\beta \asymp n^{\gamma'/2}$ for any $\gamma' \in (0, \gamma)$, we have

$$\liminf_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \mathbb{I}\big(\|\theta_s - \hat{\theta}_n\| \le n^{-(\gamma-\gamma')/2}\big) \ge 1 - \mathcal{O}\big(n^{-\gamma'}\big)$$

for sufficiently large n. As $n \to \infty$, the chain will gradually concentrate at the MLE. Choosing an appropriate γ', Lemma 3.5 shows that every limit point of the time average $\frac{1}{t}\sum_{s=0}^{t-1} \theta_s$ is $\mathcal{O}(1/\sqrt[3]{n})$-close to the MLE $\hat{\theta}_n$.

LEMMA 3.5. *Following Lemma 3.4,*

$$\limsup_{t\to\infty} \left\| \frac{1}{t} \sum_{s=0}^{t-1} \theta_s - \hat{\theta}_n \right\| = \mathcal{O}\big(n^{-\gamma/3}\big) \qquad \mathbb{P}^{\mathbf{x}}\text{-a.s.}$$

PROOF. Write

$$\left\| \frac{1}{t} \sum_{s=0}^{t-1} \theta_s - \hat{\theta}_n \right\| \le \frac{1}{t} \sum_{s=0}^{t-1} \|\theta_s - \hat{\theta}_n\|$$

$$= \frac{1}{t} \sum_{s=0}^{t-1} \|\theta_s - \hat{\theta}_n\| \mathbb{I}(\theta_s \in B) + \frac{1}{t} \sum_{s=0}^{t-1} \|\theta_s - \hat{\theta}_n\| \mathbb{I}(\theta_s \notin B)$$

$$\le \beta r_n + \max_{\theta,\theta' \in \Theta} \|\theta - \theta'\| \times \frac{1}{t} \sum_{s=0}^{t-1} \mathbb{I}(\theta_s \notin B),$$

where the first step is due to the convexity of l_2-norm. Putting it together with the result

$$\limsup_{t\to\infty}\frac{1}{t}\sum_{s=0}^{t-1}\mathbb{I}(\theta_s\notin B)\le\frac{\delta_2}{\delta_1+\delta_2}=\frac{1+b_n^2/4ac_n}{\beta^2+b_n^2/4ac_n}$$

in Lemma 3.4 yields

$$\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s-\hat{\theta}_n\right\|\le\beta r_n+\underbrace{\max_{\theta,\theta'\in\Theta}\|\theta-\theta'\|}_{\text{not depend on }n}\times\frac{1+b_n^2/4ac_n}{\beta^2+b_n^2/4ac_n}.$$

Recall that $r_n\asymp n^{-\gamma/2}$, $b_n^2/4ac_n\asymp 1$. If $\beta\asymp n^{\gamma'/2}$ increases with n, then

$$\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s-\hat{\theta}_n\right\|=\mathcal{O}\big(\max\{n^{-(\gamma-\gamma')/2},n^{-\gamma'}\}\big).$$

The bound is minimized when $\gamma'=\gamma/3$ such that

$$(\gamma-\gamma')/2=\gamma'.$$

That is,

$$\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s-\hat{\theta}_n\right\|=\mathcal{O}\big(n^{-\gamma/3}\big).$$

The coefficient constant depends on m, since $\max_{\theta,\theta'\in\Theta}\|\theta-\theta'\|$ depends on the size of parameter space Θ only, and the ratio of $b_n^2/4ac_n$ in the limit of n is determined by m. □

3.5. *Proof of the main theorem.* Now we can complete the proof of the main result in Theorem 2.1.

PROOF OF THEOREM 2.1. In the light of Lemma 3.1, it suffices to show

$$\lim_{n\to\infty}\mathbb{P}^{\mathbf{x}}\left(\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s-\theta_\star\right\|>A_m n^{-\gamma/3}\right)=0$$

for any data sample $\mathbf{x}$ satisfying (3.1), (3.2) and (3.3). To this end, Lemma 3.5 asserts that

$$\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s-\hat{\theta}_n\right\|=\mathcal{O}\big(n^{-\gamma/3}\big)\qquad\mathbb{P}^{\mathbf{x}}\text{-a.s.}$$

and constraint (3.2) ensures $\|\hat{\theta}_n - \theta_\star\| < n^{-\gamma/2}$. Combining them yields

$$\limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s - \theta_\star\right\| \le \limsup_{t\to\infty}\left\|\frac{1}{t}\sum_{s=0}^{t-1}\theta_s - \hat{\theta}_n\right\| + \|\hat{\theta}_n - \theta_\star\|$$
$$= \mathcal{O}\big(n^{-\gamma/3}\big) \qquad \mathbb{P}^{\mathbf{x}}\text{-a.s.}$$

as desired; and the coefficient constant does not depend on $\mathbf{x}$. $\square$

4. Examples. We provide three examples: a bivariate Gaussian model, a 2×2 Fully visible Boltzmann Machine (FVBM) and an exponential family random graph model to illustrate our theories. For each example, we first verify the assumptions (A1)–(A6) one by one, and then show two phases, namely "quick move" and "random walk," of the CD learning process by plotting the sequence $\{\theta_t\}_{t\ge 0}$.

4.1. *Bivariate Gaussian.* We take a bivariate Gaussian model with unknown mean $\theta \in \mathbb{R}^2$ but known covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$$

as our first example. We abuse the notation $X^{(1)}$ and $X^{(2)}$ to denote the components of the Gaussian variable $X \sim \mathcal{N}(\theta, \Sigma)$, and $\theta^{(1)}$ and $\theta^{(2)}$ to denote the components of the parameter θ. This Gaussian model is a canonical exponential family with

$$\phi(X) = \Sigma^{-1}X, \qquad \Lambda(\theta) = \theta^T\Sigma^{-1}\theta/2.$$

We run the CD algorithm with a random-scan Gibbs sampler

$$X^{(1)}|X^{(2)} \sim \mathcal{N}\Big(\theta^{(1)} + \rho\frac{\sigma_1}{\sigma_2}\big(X^{(2)} - \theta^{(2)}\big), \big(1-\rho^2\big)\sigma_1^2\Big),$$
$$X^{(2)}|X^{(1)} \sim \mathcal{N}\Big(\theta^{(2)} + \rho\frac{\sigma_2}{\sigma_1}\big(X^{(1)} - \theta^{(1)}\big), \big(1-\rho^2\big)\sigma_2^2\Big)$$

on three data samples of size $n = 10^2, 10^3, 10^4$ for evaluation purposes. The data samples are generated with the true mean parameter $\theta_\star = (0,0)^T \in \Theta = [-0.5, +0.5] \times [-0.5, +0.5]$. For each data sample, CD starts from $\theta_0 = (0.5, 0.5)^T$, and iterates $T = 2000$ times with learning rate $\eta = 0.01$.

Let us first verify (A1)–(A6). (A1) trivially holds since $\theta_\star = (0,0)^T$ is a interior point of $\Theta = [-0.5, +0.5] \times [-0.5, +0.5]$, and for any $\theta \in \Theta$, $\nabla^2\Lambda(\theta) = \Sigma^{-1}$:

$$\lambda_{\min}(\theta) = \lambda_{\min} = 0.67,$$
$$\lambda_{\max}(\theta) = \lambda_{\max} = 2.00,$$
$$\lambda_{\text{sum}}(\theta) = \lambda_{\text{sum}} = 2.67.$$

For (A2), solving an optimization problem yields

$$L = \max_{\theta\in\Theta} \frac{\chi(p_{\theta_\star}, p_\theta)}{\|\theta - \theta_\star\|} = \max_{\theta\in\Theta} \frac{\sqrt{e^{\theta^T\Sigma^{-1}\theta} - 1}}{\|\theta\|} \approx 1.84.$$

For (A3), using the explicit expression of $\alpha(\theta)$ in [2] we have

$$1 - \alpha(\theta) = 1 - \alpha = \text{smallest eigenvalue of } \Sigma^{-1}/d = 0.33.$$

The m-step distribution $k_\theta^m(x, \cdot)$ of the chain starting from any $x \in \mathbb{R}^2$ is essentially Gaussian, and

$$f_\theta(x) = \int_{\mathcal{X}} \phi(y) k_\theta^m(x, y)\, dy = A(m, \Sigma)\theta + B(m, \Sigma)x$$

is linear in θ and x with coefficient matrices A and B depending on m and Σ. Thus, (A4), (A5), (A6) hold. When $m \geq 4$ and $\eta = 0.01$, condition (2.5) is satisfied.

By our theories, the CD-4 algorithm will generate a sequence $\{\theta_t\}_{t\geq 0}$, which quickly moves to $\hat{\theta}_n$ and then randomly walks around $\hat{\theta}_n$. The range of the random walk decreases as n increases. Figure 2 illustrates this phenomenon by plotting $\{\theta_t\}_{t=0}^{2000}$. Figure 3 plots $\|\theta_t - \hat{\theta}_n\|$ versus iteration t and clearly shows two phases: "quick move" and "random walk" of the CD learning process. The random walk phase starts at $t \approx 500, 750, 1000$ for $n = 10^2, 10^3, 10^4$, respectively. Larger samples result in smaller random walk neighborhoods.

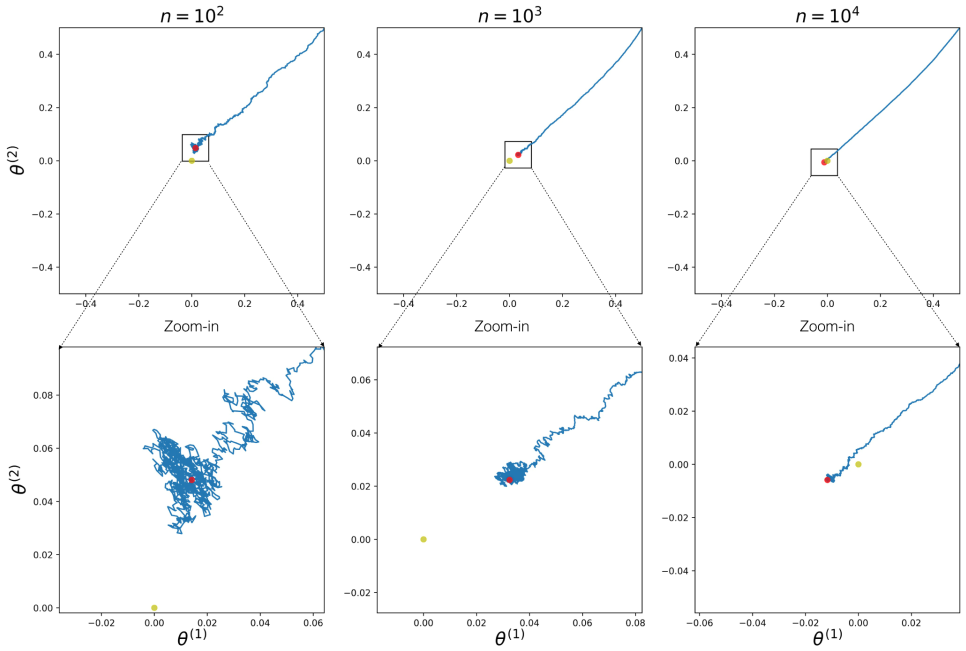

FIG. 2. $\{\theta_t\}_{0\leq t\leq 2000}$ *generated by CD-4 using Gaussian samples of size* $n = 10^2$ (*left*), 10^3 (*middle*), 10^4 (*right*). *Red dots are the MLE* $\hat{\theta}_n$, *and yellow dots are the true parameter* $\theta_\star$.

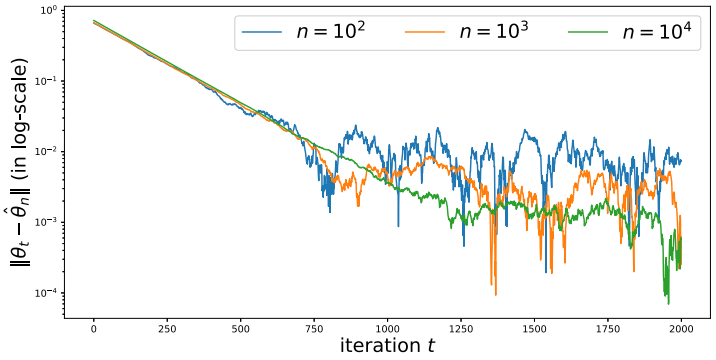

FIG. 3. *$\|\theta_t - \hat{\theta}_n\|$ versus iteration t for CD-4 using Gaussian samples.*

4.2. *Fully visible Boltzmann machine.* We analyze CD-1 in a 2×2 FVBM (i.e., x is bivariate) as another example to illustrate our theories. Let $\theta = (b_1, b_2, W_{12}) \in \mathbb{R}^3$, and write 2×2 FVBM in the canonical form of exponential family

$$p_\theta\big(x^{(1)}, x^{(2)}\big) \propto \exp\big(\theta^{(1)}x^{(1)} + \theta^{(2)}x^{(2)} + \theta^{(3)}x^{(1)}x^{(2)}\big).$$

Here, we abuse the notation $X^{(1)}$ and $X^{(2)}$ to denote the components of the FVBM variable X, and $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}$ to denote the components of the parameter θ. The data samples are generated with the true parameter $\theta_\star = (0, 0, 0.5)^T \in \Theta = [-1, +1] \times [-1, +1] \times [-0.5, +1.5]$. For each data sample, CD-1 starts from $\theta_0 = (1, 1, 1)^T$, and iterates $T = 2000$ times with learning rate $\eta = 0.01$. Each iteration uses a random-scan Gibbs sampler:

$$X^{(1)}|X^{(2)} \sim 2 \times \text{Bernoulli}\big(p_\theta\big(X^{(1)} = +1|X^{(2)}\big)\big) - 1,$$

$$X^{(2)}|X^{(1)} \sim 2 \times \text{Bernoulli}\big(p_\theta\big(X^{(2)} = +1|X^{(1)}\big)\big) - 1.$$

We can verify (A1)–(A6) one by one. (A1) holds since $\theta_\star$ is an interior point of Θ. (A2) holds since $\Lambda(\theta) < \infty$ for any $\theta \in \mathbb{R}^3$. For (A3), $\alpha(\theta)$ is the second largest absolute eigenvalue of the transition probability matrix, which is less than 1 and continuous in $\theta \in$ compact Θ, and thus has an upper bound $\alpha < 1$. (A4) and (A6) hold because components of $\phi(X^{(m)})$ are bounded random variables. $k_\theta^m(x, y)$ can be represented as a transition probability matrix whose entries are continuously differentiable functions of θ. Then for any $x \in \mathcal{X} = \{-1, +1\}^p$, $f_\theta(x)$ is continuously differentiable, and thus Lipschitz continuous in $\theta \in$ compact Θ. In addition, $\mathcal{X}$ is a finite set, thus (A5) holds.

Figure 4 plots the sequence $\{\theta_t\}_{t=0}^{2000}$ generated by CD-1 on FVBM samples. At the beginning, θ_t moves quickly toward the MLE and then randomly walks around it. The range of the random walk decreases as n increases.

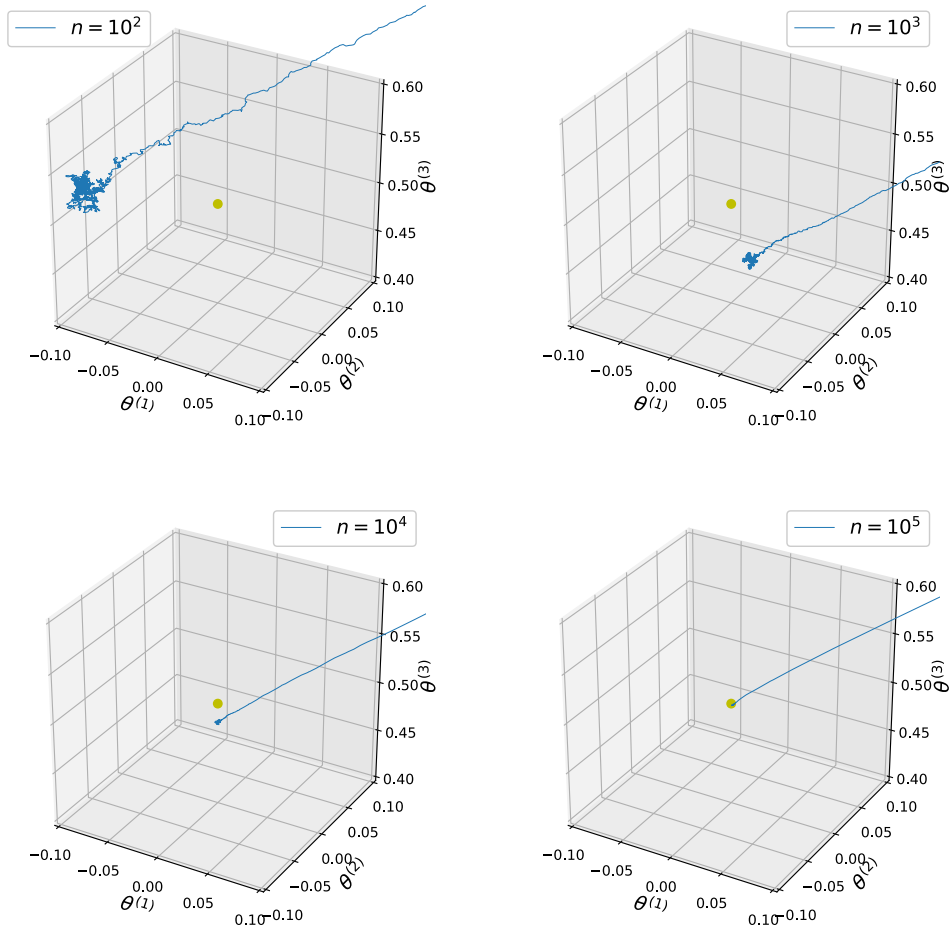

FIG. 4. *For each FVBM sample of size* $n = 10^2$ (*upper-left*), 10^3 (*upper-right*), 10^4 (*lower-left*), 10^5 (*lower-right*), *CD-1 generates* $\{\theta_t\}_{0 \le t \le 2000}$. *The yellow dots are the true parameter* $\theta_\star = (0, 0, 0.5)^T$.

4.3. *Exponential-family random graph model.* We analyze CD-5 in the exponential family random graph model (ERGM) with 10 nodes as the third example. ERGM is widely used in social network analysis [36]. The CD algorithm has performed well in these models [3, 19, 23].

Assume we have a undirected graph x with $x_{ij} = 1$ indicating the existence of an edge between ith node and jth node, and $x_{ij} = 0$ otherwise. The probability mass function is given by $p_\theta(x) \propto \exp[\theta^T \phi(x)]$, where $\phi(x) = [\phi_1(x), \ldots, \phi_d(x)]$ are the global features of the network. Like [3], our experiment sets $\phi(x)$ to be three network statistics, namely, the number of edges, the number of stars and the number of triangles. The data samples are generated with the true parameter $\theta_\star = (-2, 0, 0)^T \in \Theta = [-5, +5] \times [-5, +5] \times [-5, +5]$. For each data sample,

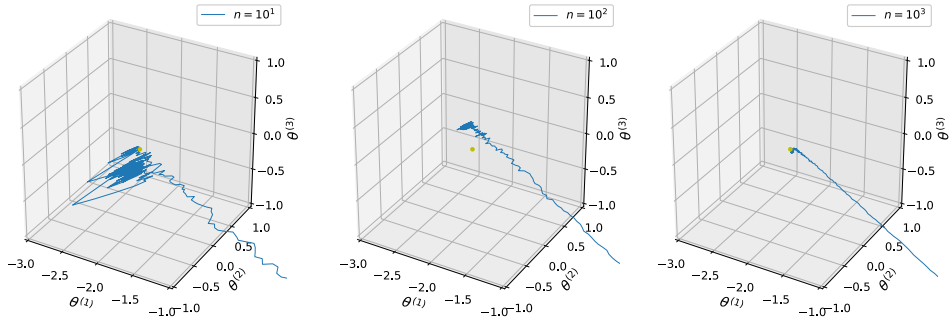

FIG. 5. *For each ERGM sample of size* $n = 10^1$ (*left*), 10^2 (*middle*) *and* 10^3 (*right*), *CD-5 generates* $\{\theta_t\}_{t=0}^{500}$. *The yellow dots are the true parameter* $\theta_\star = (-2, 0, 0)^T$.

CD-5 starts from $\theta_0 = (5, 5, 5)^T$, and iterates $T = 500$ times with learning rate $\eta = 0.05$. Each iteration uses a Metropolis–Hastings sampler.

We can verify (A1)–(A6) one by one. (A1) holds since $\theta_\star$ is an interior point of Θ. (A2) holds since $\Lambda(\theta) < \infty$ for any $\theta \in \mathbb{R}^3$. For (A3), $\alpha(\theta)$ is the second largest absolute eigenvalue of the transition probability matrix, Each entry of the transition probability matrix is a continuous function in $\theta \in$ compact Θ. It follows that $\alpha(\theta)$ is less than 1 and continuous in $\theta \in$ compact Θ. Thus, it has an upper bound $\alpha < 1$. (A4) and (A6) hold because components of $\phi(X^{(m)})$ are bounded. Then for any $x \in \mathcal{X} = \{0, 1\}^{10}$, $f_\theta(x)$ is continuously differentiable, and thus Lipschitz continuous in $\theta \in$ compact Θ. In addition, the sample space $\mathcal{X}$ is a finite set, thus (A5) holds.

Figure 5 plots the sequence $\{\theta_t\}_{t=0}^{500}$ generated by CD-5 on ERGM samples. At the beginning, θ_t moves quickly toward the MLE and then randomly walks around it. The range of the random walk decreases as n increases.

5. Discussion. The CD algorithm was proposed to train energy-based models (e.g., Restricted Boltzmann Machine) and later used in the layer-wise pre-training step of deep belief network. It has played a key role in the emergence of deep learning. The algorithm approximates the gradient of the log-likelihood function by running short MCMC chains to save computational cost. Although using biased gradient approximations in the iteration of the gradient ascent update, the CD algorithm gives satisfactory parameter estimates in many practical applications. On the other hand, Mackay [26] provides Gaussian examples in which the CD-1 algorithm does not converge to the MLE. Many eminent scholars in machine learning including Yuille [44], Carreira-Perpinan [9], Bengio and Delalleau [7] and sutskever and Tieleman [40] had attempted to theoretically analyze the CD algorithm. But whether and why this algorithm is asymptotically consistent is still an open question.

In order to fill in the gap, this paper is devoted to a theoretical analysis of the CD algorithm in exponential families. Exponential families are special cases of the

energy-based models with convex log-partition functions. We narrow the scope of analyses down to these special cases, because (1) the convexity ensures the uniqueness of the MLE so that we could compare the CD estimates to the unique MLE, and (2) the CD algorithm has been used to fit exponential families with unavailable gradient of log-partition function, for example, exponential-family random graph models. For nonconvex energy-based models like Restricted Boltzmann Machines, we conjecture that the CD algorithm converges to a random walk around one or more local maximum points of the likelihood function. Intuitively, the gradient ascent algorithm using exact gradients usually does not converge to the MLEs of nonconvex models. We do not expect that the CD algorithm using approximate gradients does better than the gradient ascent algorithm.

We find that $\{\theta_t\}_{t\geq 0}$ is a homogenous Markov chain conditional on the data sample. And the chain quickly moves toward the MLE $\hat{\theta}_n$ and then randomly walk around it. This phase transition can be explained as follows. When the chain moves from θ, which is far away from the MLE $\hat{\theta}_n$, the exact gradient $g(\theta)$ is relatively large compared to the approximation error resulting from the m-step MCMC sampling. The CD update equation keeps pushing θ to quickly move toward $\hat{\theta}_n$. When θ enters a small neighbor around $\hat{\theta}_n$, $g(\theta)$ fails to suppress the MCMC approximation error. The "quick move" phase ends and the chain starts a "random walk" in this neighborhood. In addition to the plots of the sequence $\{\theta_t\}_{t\geq 0}$ in Section 4, we plot the gradient fields of the bivariate Gaussian and the FVBM examples and put them in the [21].

Our theories can explain Mackay's [26] Gaussian examples in which the CD-1 algorithm does not converge to the MLE. MacKay found that the CD-1 algorithm with different kernels has one or multiple fixed points but these fixed points may not be the MLEs. However, the author reported that "in the special case of an infinite amount of data that come from an axis-aligned Gaussian, the noisy swirl operator's one-step algorithm (CD-1) does converge to the maximum-likelihood parameters." According to our theories, the fixed points or the limit points of the sequence $\{\theta_t\}_{t\geq 0}$ are not exactly the MLEs. But they are $\mathcal{O}(1/\sqrt[3]{n})$-close to MLEs with high probability and the gap shrinks to 0 as the amount of data goes to infinity. In this way, our theories give an explanation for MacKay's Gaussian examples. We think the phenomenon that CD does not converge to the MLE should not be considered as a failure of CD. We note that the $\mathcal{O}_p(1/\sqrt[3]{n})$ rate is probably not the best rate. We believe that a rate closer to or equals to $\mathcal{O}_p(1/\sqrt{n})$ should be obtainable with a more refined analysis.

We would like to highlight three other novelties of the theoretical analyses in this paper. First, we let the iteration number $t \to \infty$ and then let $n \to \infty$ when analyzing the CD algorithm, while many previous works had not clearly distinguished the two limits and led to unreasonable convergence conditions. For example, two convergence conditions given in the remarks of Result 4 in [44] (translated to the

form using our notation) are

$$\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) = \nabla\Lambda(\theta_\star),$$

$$\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}}\phi(y)k_{\theta_\star}^{m}(X_i, y)\,dy = \nabla\Lambda(\theta_\star).$$

These conditions are satisfied with probability zero in most models of continuous distributions. It is because their left-hand sides are functions of $\{X_i\}_{i=1}^n$, and thus random, while their right-hand sides are nonrandom. Second, conditioning on a finite data sample, we can see the sequence $\{\theta_t\}_{t\geq 0}$ as a Markov chain. Hence, many nice results in the Markov chain, martingale and stability analysis theories like Foster–Lyapunov drift criterion can be used. We believe Lyapunov drift conditions might be used to analyze other stochastic gradient descent schemes with approximate gradients. Third, the sub-exponentiality assumption (A4) allows to analyze models with unbounded sufficient statistics. It is an improvement over existing theoretical analyses, which are restricted to bounded cases.

Our theoretical results also provide some guidance for the practitioners of the CD algorithm. First, since the CD algorithm converges to a random walk around the MLE, a single estimate θ_t is not as reliable as the average $\frac{1}{t}\sum_{s=0}^{t-1}\theta_s$. Thus, an averaging scheme should be taken. Second, the CD learning process typically has two phases: "quick move" and "random walk." This phase division suggests practitioners to stop the iterations of the algorithm when θ_t starts the random walk. Third, the success of the CD algorithm highly relies on the speed of MCMC kernels in use. One may design and test a few candidate MCMC kernels, and use the fastest kernel in the CD algorithm. Last but not the least, one could do mini-batch sampling at each iteration of the CD algorithm, as a smaller sample size like $n/10$ would change neither the bias of the CD gradient approximation nor the order of its variance. Thus, the asymptotical consistency of the CD algorithm still holds if the mini-batch sampling scheme is in use.

Acknowledgments. The authors would like to thank Dr. Weijie Su, Dr. Rachel Wang, Dr. Lester Mackey and Dr. Percy Liang for valuable advice.

SUPPLEMENTARY MATERIAL

Appendix: Other simulation results and lemmas (DOI: 10.1214/17-AOS1649SUPP; .pdf). This supplementary material contains other simulation results and five lemmas.

REFERENCES

[1] ACKLEY, D. H., HINTON, G. E. and SEJNOWSKI, T. J. (1985). A learning algorithm for Boltzmann machines. *Cogn. Sci.* **9** 147–169.

[2] AMIT, Y. (1996). Convergence properties of the Gibbs sampler for perturbations of Gaussians. *Ann. Statist.* **24** 122–140. MR1389883

[3] ASUNCION, A., LIU, Q., IHLER, A. and SMYTH, P. (2010). Learning with blocks: Composite likelihood and contrastive divergence. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 33–40.

[4] ATCHADÉ, Y. F., FORT, G. and MOULINES, E. (2017). On perturbed proximal gradient algorithms. *J. Mach. Learn. Res.* **18** Paper No. 10, 33. MR3634877

[5] AZUMA, K. (1967). Weighted sums of certain dependent random variables. *Tôhoku Math. J.* (2) **19** 357–367. MR0221571

[6] BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. MR2486527

[7] BENGIO, Y. and DELALLEAU, O. (2009). Justifying and generalizing contrastive divergence. *Neural Comput.* **21** 1601–1621. MR2527797

[8] BENGIO, Y., LAMBLIN, P., POPOVICI, D., LAROCHELLE, H. (2007). Greedy layer-wise training of deep networks. In *Proceedings of the* 19*th International Conference on Neural Information Processing Systems* (*NIPS'*06) 153–160. MIT Press, Cambridge, MA.

[9] CARREIRA-PERPINAN, M. A. and HINTON, G. E. (2005). On contrastive divergence learning. In *AISTATS* **10** 33–40.

[10] COATES, A., NG, A. and LEE, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (G. Gordon, D. Dunson and M. Dudik, eds.). *Proceedings of Machine Learning Research* **15** 215–223.

[11] CONWAY, J. B. (1990). *A Course in Functional Analysis*, 2nd ed. *Graduate Texts in Mathematics* **96**. Springer, New York. MR1070713

[12] DIACONIS, P. (2009). The Markov chain Monte Carlo revolution. *Bull. Amer. Math. Soc.* (*N.S.*) **46** 179–205. MR2476411

[13] HAIRER, M. and MATTINGLY, J. C. (2011). Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis*, *Random Fields and Applications VI*. *Progress in Probability* **63** 109–117. Birkhäuser, Basel. MR2857021

[14] HE, X., ZEMEL, R. S. and CARREIRA-PERPIÑÁN, M. Á. (2004). Multiscale conditional random fields for image labeling. In *Proceedings of the* 2004 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR* 2004) **2** 692–702. IEEE, New York.

[15] HINTON, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14** 1771–1800.

[16] HINTON, G. E., OSINDERO, S. and TEH, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* **18** 1527–1554.

[17] HINTON, G. E. and SALAKHUTDINOV, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* **313** 504–507. MR2242509

[18] HINTON, G. E. and SALAKHUTDINOV, R. R. (2009). Replicated softmax: An undirected topic model. In *Advances in Neural Information Processing Systems* 1607–1614.

[19] HUNTER, D. R., HANDCOCK, M. S., BUTTS, C. T., GOODREAU, S. M. and MORRIS, M. (2008). ERGM: A package to fit, simulate and diagnose exponential-family models for networks. *J. Stat. Softw.* **24** 1–29 nihpa54860.

[20] HYVÄRINEN, A. (2006). Consistency of pseudolikelihood estimation of fully visible Boltzmann machines. *Neural Comput.* **18** 2283–2292.

[21] JIANG, B., WU, T.-Y., JIN, Y. and WONG, W. H. (2018). Supplement to "Convergence of contrastive divergence algorithm in exponential family." DOI:10.1214/17-AOS1649SUPP.

[22] KONTOYIANNIS, I. and MEYN, S. P. (2012). Geometric ergodicity and the spectral gap of non-reversible Markov chains. *Probab. Theory Related Fields* **154** 327–339.

[23] KRIVITSKY, P. N. (2017). Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models. *Comput. Statist. Data Anal.* **107** 149–161. MR3575065

[24] LAROCHELLE, H. and BENGIO, Y. (2008). Classification using discriminative restricted Boltzmann machines. In *Proceedings of the* 25*th International Conference on Machine Learning* 536–543. ACM, New York.

[25] LEHMANN, E. L. and CASELLA, G. (1991). *Theory of Point Estimation*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA. Reprint of the 1983 original. MR1138212

[26] MACKAY, D. (2001). Failures of the one-step learning algorithm. Technical report, Available at http://www.inference.phy.cam.ac.uk/mackay/abstracts/gbm.html.

[27] MENGERSEN, K. L. and TWEEDIE, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24** 101–121. MR1389882

[28] MEYN, S. P. and TWEEDIE, R. L. (1992). Stability of Markovian processes I: Criteria for discrete-time chains. *Adv. in Appl. Probab.* **24** 542–574. MR1174380

[29] MEYN, S. P. and TWEEDIE, R. L. (1993). Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes. *Adv. in Appl. Probab.* **25** 518–548.

[30] MOHAMED, A.-R., DAHL, G. E. and HINTON, G. (2012). Acoustic modeling using deep belief networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **20** 14–22.

[31] PARIKH, N., BOYD, S. P. et al. (2014). Proximal algorithms. *Found. Trends Optim.* **1** 127–239.

[32] RIGOLLET, P. (2015). Lecture notes in high dimensional statistics.

[33] ROBERTS, G. O. and ROSENTHAL, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Commun. Probab.* **2** 13–25. MR1448322

[34] ROBERTS, G. O. and ROSENTHAL, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv.* **1** 20–71. MR2095565

[35] ROBERTS, G. O. and TWEEDIE, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83** 95–110.

[36] ROBINS, G., PATTISON, P., KALISH, Y. and LUSHER, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Soc. Netw.* **29** 173–191.

[37] ROTH, S. and BLACK, M. J. (2005). Fields of experts: A framework for learning image priors. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR* 2005) **2** 860–867. IEEE, New York.

[38] RUDOLF, D. (2011). Explicit error bounds for Markov chain Monte Carlo. ArXiv preprint. Available at arXiv:1108.3201.

[39] SALAKHUTDINOV, R., MNIH, A. and HINTON, G. (2007). Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the* 24*th International Conference on Machine Learning* (*ICML*'07) 791–798. ACM, New York.

[40] SUTSKEVER, I. and TIELEMAN, T. (2010). On the convergence properties of contrastive divergence. In *International Conference on Artificial Intelligence and Statistics* 789–795.

[41] TEH, Y. W., WELLING, M., OSINDERO, S. and HINTON, G. E. (2004). Energy-based models for sparse overcomplete representations. *J. Mach. Learn. Res.* **4** 1235–1260. MR2103628

[42] VÁRNAI, C., BURKOFF, N. S. and WILD, D. L. (2013). Efficient parameter estimation of generalizable coarse-grained protein force fields using contrastive divergence: A maximum likelihood approach. *J. Chem. Theory Comput.* **9** 5718–5733.

[43] WILLIAMS, C. K. I. and AGAKOV, F. V. (2002). An analysis of contrastive divergence learning in Gaussian Boltzmann machines. Working paper, Institute for Adaptive and Neural Computation, Edinburgh.

[44] YUILLE, A. L. (2005). The convergence of contrastive divergences. In *Proceedings of the* 17*th International Conference on Neural Information Processing Systems* (*NIPS*'04) 1593–1600. MIT Press, Cambridge, MA.

B. JIANG
T.-Y. WU
Y. JIN
WONG LAB
JAMES H. CLARK CENTER
STANFORD UNIVERSITY
318 CAMPUS DRIVE
STANFORD, CALIFORNIA 94305
USA
E-MAIL: baijiang@stanford.edu
tungyuwu@stanford.edu
yifanj@stanford.edu

W. H. WONG
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
390 SERRA MALL
STANFORD, CALIFORNIA 94305
USA
E-MAIL: whwong@stanford.edu
URL: http://web.stanford.edu/group/wonglab/