UNIVERSITY OF CALIFORNIA
Los Angeles

A Dance with the Langevin Equation

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Erik Lennart Nijkamp

2018

ABSTRACT OF THE THESIS

A Dance with the Langevin Equation

by

Erik Lennart Nijkamp
Master of Science in Statistics
University of California, Los Angeles, 2018
Professor Song-Chun Zhu, Chair

In this thesis, we shall discuss the Langevin equation. While the equation is well known in Statistical Physics, we show novel adaptions to the field of Machine Learning. In particular, we elaborate briefly on a recipe in an unsupervised learning scheme. That is, we introduce an irreversible Langevin sampler to accelerate Contrastive Divergence. This adjustment to the Langevin equation, as such, improves the convergence behavior towards the equilibrium by injecting a vector field $C(x)$ satisfying some conditions derived from the Fokker-Planck equation. In particular, $C(x)$ motivates the particles in the negative phase to dance such that convergence of system towards its thermal equilibrium state is accelerated. We illustrate an application in the form of learning a Gaussian-Bernoulli Restricted-Boltzmann machine.

The thesis of Erik Lennart Nijkamp is approved.

Chad J. Hazlett

Ying Nian Wu

Song-Chun Zhu, Committee Chair

University of California, Los Angeles

2018

To my parents for their unconditional love and support.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisors Dr. Song-Chun Zhu and Dr. Ying Nian Wu for their continuous support to my M.Sc. and Ph.D. research, for their patience, motivation, enthusiasm, and immense knowledge. Their guidance was invaluable to me with respect to research and the writing of this thesis.

Besides my advisors, I would like to thank Dr. Chad J. Hazlett for his continuous support and acting as a committee member. Further, I would like to show my gratitude to Mitchell Hill. Our discussions and research efforts were instrumental to this thesis. Also, I would like to thank Anshu Wang for her unconditional support.

I would like to thank Glenda Jones for her outstanding support in administrative and general matters.

I sincerely acknowledge the support and facilities offered by the Department of Statistics at UCLA.

CHAPTER 1

Introduction

In the past decade, models in the field of Machine Learning have made tremendous leaps towards generalizations, despite their overwhelming capacity. While we are seemingly on the brink of understanding inherent interplay between model architectures and learning dynamics, the phenomenon of emerging, desirable generalization characteristics in vastly over-parameterized models is still an unsolved mystery.

In this thesis, we leave such considerations behind and put our emphasis on much older, and arguably, better-understood models and learning methods with their foundations in Physics. As such, we will discuss Paul Langevin's description of Brownian motion and it's application the Geoffrey Hinton's Restricted Boltzmann machines. We show how one may accelerate diffusion in the form of Langevin dynamics by injecting a carefully crafted vector field $C(x)$ while satisfying Fokker-Planck's equation. These considerations give rise to new questions, such as, which properties should $C(x)$ satisfy such that desirable convergence properties emerge.

As a logical consequence of this preamble, the thesis is structured in 6 Chapters in the following manner.

In Chapter 2, we will illuminate the history of Brownian motion starting with Albert Einstein, and his contemporary, Paul Langevin in the early 20th century. We draw connections to the works of Adriaan Fokker, Max Planck, Ernst Ising, Ludwig Boltzmann, John Hopfield, Paul Smolensk, Geoffrey Hinton, and Ulf Grenander.

In Chapter 3, we will briefly revise the relation between the Langevin equation and the Fokker-Planck equation following [14, 18, 19]. In particular, we are interested in the desired behavior of the stochastic noise term in the description of Brownian motion such that the

the noise drives the system to it's equilibrium at a given temperature.

In Chapter 4, we consider diffusion $X(t)$ in the form of the overdamped Langevin equation,

$$dX(t) = -\nabla U(X(t))dt + \sqrt{2T}dW(t), \tag{1.1}$$

where U is a given potential, $W(t)$ is a Wiener process and T is temperature. We may use (1.1) to sample the Boltzmann-Gibbs measure with density

$$\pi(x) \propto \exp(-U(x)/T), \tag{1.2}$$

which is the invariant measure of (1.1) under some conditions. Then, we introduce a family of altered diffusions

$$dX(t) = -\nabla U(X(t))dt + C(X(t))dt + \sqrt{2T}dW(t) \tag{1.3}$$

where the invariant measure is maintained if the irreversible force $C(x)$ satisfies

$$div(C) = T^{-1}C \cdot \nabla U. \tag{1.4}$$

We show how to derive (1.4) by the Fokker-Planck equation and discuss particular choices of $C(x)$ satisfying (1.4). The efficiency of the discussed irreversible Langevin sampler is illustrated in sampling from the classical double-well potential and estimation of the covariance of a Gaussian.

In Chapter 5, we interpret Contrastive Divergence as a stochastic variant of the Maximum Likelihood method. Then, we discuss the Restricted-Boltzmann machine and substitute the Bernoulli visible units with Gaussian ones. The irreversible Langevin sampler is recruited in the negative phase of Contrastive Divergence learning and the effect of accelerated diffusion will be evaluated empirically.

In Chapter 6, the thesis is concluded with open questions which haven't been addressed and future directions for research.

CHAPTER 2

Prior Art

In 1905, Albert Einstein sparked the discussion of random processes with his pioneering study on Brownian motion [5]. Three years after Einstein's ground breaking work, the French Physicist Paul Langevin, a contemporary of Einstein, devised a very different, but equally successful formulation of Brownian motion [14].

Remarkably, in Langevin's own words, his approach is "infinitely more simple" than Einstein's derivations. While Einstein solved a partial differential equation governing the time-evolution of the probability density of a particle in Brownian motion, which in 1914 and 1917 was conceptualized by Adriaan Fokker and Max Planck in their works [6, 18], respectively, and became known as the Fokker-Planck equation. Instead, Langevin referred to Newton's second law of motion to describe a Brownian particle. As such, we may argue that Langevin introduced a stochastic reinterpretation of Newtonian physics.

Today, we realize that the elegant simplicity of Langevin's description is not without caveats as he forced unseen mathematical objects with unusual properties into existence [16]. While Langevin was capable of manipulating objects such as Gaussian white noise and stochastic differential equations by intuitively, their formal properties were not known at this time.

The behavior of a continuous, Markov process can be characterized by both, the Langevin and the Fokker-Planck equation. In fact, both Einstein and Langevin used their respective equations to derive a fundamental result in Physics: the root-mean-squared displacement of a Brownian particle increases with the square root of time [16].

One may cautiously argue that Langevin's treatment of Brownian motion is slightly more rigorous and general than Einstein's derivation. That is, Langevin described the velocity of

a Brownian particle as Ornstein-Uhlenbeck process and the position of the partial as the time integral of its velocity. Over time, the process tends to drift towards its long-term mean. Whereas Einstein described the position as a driftless Wiener process. The former is a covering theory for the latter and reduces to it in a special "coarsegraining" limit [16].

In 1924, Ernst Ising formulated [13] ferromagnetism as discrete variables that represent magnetic dipole moments of atomic spins which are arranged in a lattice such that each spin may interact with its neighbors. In 1975, the treatment of weights in the Ising models in a stochastic manner was introduced in [24]. In 1982, John Hopfield [9] revealed connections between the Spin-Glass model and "associative" memory. In 1985, Geoffrey Hinton proposed Boltzmann machines (RBM) [1] as a particular form of log-linear Markov random fields or stochastic, generative variant of Hopfield networks with hidden units. The name stems from the fact in thermal equilibrium the relation between energy and probability in Boltzmann machines is in the form of the the Boltzmann distribution.

Learning a general Boltzmann machine is impractical as the time the machine must be run to gain equilibrium statistics grows exponentially with the machine's size and magnitude of connections strength. In 1986, Paul Smolensky invented harmony theory [25] which formulated restricted Boltzmann machines under the name Harmonium, and rose to prominence after Hinton proposed Contrastive Divergence (CD) in 2002. In the field of Statistics, we may interpret CD as a stochastic variant of the Maximum Likelihood method. The restriction is defined such that intra-layer connections are omitted such that the neurons must form a bipartite graph. The negative phase of Contrastive Divergence involves intractable sampling from the model as corresponding densities are only known up to normalizing constants. One has to resort to approximations. A Markov process with the underlying distribution as its equilibrium is often used to generate an approximation. The factorization into a bipartite graph allows for efficient Block Gibbs sampling of particles in the negative phase. Related work was conducted by Amit and Grenander in 1991 [2], Frigessi, Hwang and Younes in 1992 [12], Gilks and Roberts in 1996 [7], Barbu and Zhu in 2003 [3], Roberts and Rosenthal in 2004 [22].

Since the Restricted Boltzmann machine as a physical system is equipped with the notion

of free energy, while training the model, we may consider Monte-Carlo sampling in the negative phase as simulating the diffusion of particles in the form of the Langevin dynamics. A natural question one may pose is whether the diffusion towards the equilibrium state can be accelerated? Convergence properties of irreversible Gibbs samplers are studied in Amit and Grenander in 1991 [2], Geman and Geman in 1997 [4]. The acceleration of diffusion in the Langevin form was proposed by Hwang and Sheu in 2005 [11], and applied by Rey-Bellet Spiliopoulos in 2015 [20], Lu and Spiliopoulos in 2016 [17].

To the best of our knowledge, a study on training of Gaussian-Bernoulli Restricted Boltzmann (GRBM) machine with irreversible Langevin sampling has not been published. The training of Gaussian-Bernoulli Restricted Boltzmann machines is known to be notoriously difficult to train and studies refer to Hamiltonian Monte Carlo samplers with Metropolis-Hastings correction.

CHAPTER 3

Langevin's Equation

In this Chapter, we will review Paul Langevin's equation. Paul Langevin, Adriaan Fokker and Max Planck were foremost known as Physicists. To pay our respect, we shall borrow the notation and terminology of the field of Physics. Specifically, we borrow from [14, 18, 19] and rewrite for brevity.

3.1 Langevin's description

Brownian motion. Consider the Langevin description of Brownian motion in the form of the Langevin equation

$$m\ddot{x} = -6\pi\tilde{\eta}a\dot{x} + X, \tag{3.1}$$

where m denotes position, x mass, a radius of a particle, $\tilde{\eta}$ viscosity of the medium, X a random force resulting from thermal fluctuations, and $\dot{x} = dx/dt$, $\ddot{x} = d^2x/dt^2$ the partial derivatives in time. Langevin solved (3.1) assuming $\langle X\rangle = 0$, and, by stationary of the process, $\langle xX\rangle = 0$. Remarkably, while the particle moves within a medium of temperature T, the temperature only when rewriting the left-hand side of (3.1) as $m\ddot{x} = m\ddot{x^2}/2 - m\dot{x}^2$ and replacing $\langle m\dot{x}^2\rangle$ by the average kinetic energy k_BT where k_B is known as Boltzmann's constant. Examining the derivation carefully, one can see that $m\ddot{x^2}/2$ only plays a role in determining the relaxation to the long-time asymptote $\langle x^2\rangle = 2Dt$. Therefore, we may neglect the inertial term and consider overdamped motion of the particle

$$0 = -6\pi\tilde{\eta}a\dot{x} + X, \tag{3.2}$$

Will we still obtain $\langle x^2\rangle = 2Dt$? If so, what are the assumptions on X to imitate the fluctuations of a T-temperature medium?

Motion in a potential. Let us imagine the particle is attached to the origin by a spring, i.e. it is moving in a potential $U(x) = kx^2/2$. Then, we must introduce $-dU/dx = -kx$ in (3.1),

$$0 = -6\pi\tilde{\eta}a\dot{x} - kx + X. \tag{3.3}$$

Then, what properties does X have to satisfy so that the system relaxes to it's thermal equilibrium. That is, the long-time limit of the probability of the particle at position x is given by the Boltzmann distribution,

$$P_{eq}(x) = \exp(-kx^2/2k_BT)/Z, \tag{3.4}$$

where Z denotes the partition function

$$Z = \int dx \exp(-kx^2/2k_BT). \tag{3.5}$$

In a more general setting, we pose the question of which properties does X have to satisfy such that for a motion in a potential $U(x)$

$$0 = -6\pi\tilde{\eta}a\dot{x} - \frac{dU}{dx} + X, \tag{3.6}$$

the system relaxes to the thermal equilibrium,

$$P_{eq}(x) = \exp(-U/k_BT)/Z. \tag{3.7}$$

Rewrite (3.6) in the form of

$$\dot{x}(t) = -\mu\frac{dU}{dx} + \eta(t), \tag{3.8}$$

where $\mu = 1/(6\pi\tilde{\eta}a)$ and $\eta(t) = \mu X(t)$. Then, may we construct $\eta(t)$ such that (3.7) holds in the long-time limit?

Brownian Motion Limit. Let $(U = 0)$, then integrate (3.8)

$$x(t) = \int_0^t \eta(\tau)d\tau. \tag{3.9}$$

Then,

$$\langle x\rangle = \int_0^t \langle\eta(\tau)\rangle d\tau, \tag{3.10}$$

and

$$\langle x^2 \rangle = \int_0^t \int_0^t \langle \eta(\tau)\eta(\tau') \rangle d\tau d\tau'. \tag{3.11}$$

Assume, (1) zero-centered, and, (2) time-independent noise, that is,

$$\langle \eta(t) \rangle = 0, \tag{3.12}$$

and,

$$\langle \eta(t)\eta(t') \rangle = 2D\delta(t - t'), \tag{3.13}$$

respectively, where $\delta(t)$ denotes the delta-function and D is the amplitude of the noise. Then, substitution of (3.12) in (3.10) gives

$$\langle x \rangle = \int_0^t \langle \eta(\tau) \rangle d\tau = 0, \tag{3.14}$$

while (3.13) in (3.11) gives

$$\langle x^2 \rangle = \int_0^t \int_0^t \langle \eta(\tau)\eta(\tau') \rangle d\tau d\tau' = 2Dt. \tag{3.15}$$

Contrary to the Langevin description, it appears that the diffusion coefficient D is not determined?

Linear oscillator. Consider the equation of motion for a linear oscillator (3.12) in (3.10) gives

$$\dot{x} = -\mu k x + \eta(t), \tag{3.16}$$

and the solution of (3.17) in the form of

$$x(t) = x(0) \exp(-\mu k t) + \int_0^t \exp(-\mu k(t - \tau))\eta(\tau) d\tau, \tag{3.17}$$

where $x(0)$ is the initial value of x at $t = 0$. Since $\langle \eta(t) \rangle = 0$, then rewriting (3.17) reveals that the average position of the particle relaxes to the mechanical equilibrium position where the net force on the particle is zero,

$$\langle x(t) \rangle = x(0) \exp(-\mu k t) \to 0. \tag{3.18}$$

We calculate the mean-square fluctuations $\langle x^2(t) \rangle$ by rewriting (3.17) where the cross terms are linear in η and their average is zero

$$\langle x^2(t) \rangle = x^2(0) \exp(-2\mu k t) + \int_0^t \int_0^t \exp(-\mu k((t - \tau) + (t - \tau')) \langle \eta(\tau)\eta(\tau') \rangle d\tau d\tau'. \tag{3.19}$$

We simplify the double integral by (3.13) such that

$$\langle x^2(t)\rangle = x^2(0)\exp(-2\mu k t) + \frac{D}{\mu k}(1-\exp(-2\mu k t)). \tag{3.20}$$

Then,

$$\langle x^2(t\to\infty)\rangle = \frac{D}{\mu k}. \tag{3.21}$$

Assuming this limit equals the equilibrium limit, we may relate the amplitude of noise D and temperature T,

$$\langle x^2(t\to\infty)\rangle = \langle x^2\rangle_{eq} = \frac{D}{\mu k} = \frac{k_B T}{k}. \tag{3.22}$$

Then, the diffusion coefficient in Brownian motion or amplitude of noise is

$$D = \mu k_B T = \frac{k_B T}{6\pi\tilde{\eta} a}, \tag{3.23}$$

which coincides with the Langevin description. Thus, considering (3.12) and (3.13) as properties of the noise, then the Brownian motion result agrees with known results from linear oscillators. But, will (3.12) and (3.13) yield the equilibrium distribution $P_{eq}(x)$?

Gaussian noise. Let us discretize (3.8) as

$$x(t+\Delta t) = x(t) - \mu\frac{dU}{dx}\Delta t + \eta_{\Delta t}(t). \tag{3.24}$$

While we understand the deterministic part of this equation, how shall we interpret $\eta_{\Delta t}(t)$? Again, assume $\langle\eta_{\Delta t}(t)\rangle = 0$ and the values of $\eta_{\Delta t}(t)$ are independent with respect to t. Then, how shall we determine the amplitude of the noise in the discretized case? Let us neglect the external potential, and discretize Brownian motion as

$$x(t+\Delta t) = x(t) + \eta_{\Delta t}(t). \tag{3.25}$$

After n steps the elapsed time is equal to $n\Delta t$ while the position of the particle is

$$x(t+n\Delta t) = x(t)\eta_{\Delta t}(t) + \eta_{\Delta t}(t+\Delta t) + \ldots + \eta_{\Delta t}(t+(n-1)\Delta t). \tag{3.26}$$

Then, the average of the mean square displacement is

$$\begin{aligned}\langle[x(t+n\Delta t)]^2\rangle &= \langle[\eta_{\Delta t}(t) + \eta_{\Delta t}(t+\Delta t) + \ldots + \eta_{\Delta t}(t+(n-1)\Delta t]^2\rangle \\ &= \langle\eta^2_{\Delta t}(t)\rangle + \langle\eta^2_{\Delta t}(t+\Delta t)\rangle + \ldots + \langle\eta^2_{\Delta t}(t+(n-1)\Delta t)\rangle \\ &= \langle\eta^2_{\Delta t}\rangle n.\end{aligned} \tag{3.27}$$

As the elapsed time is $n\Delta t$, we know for the mean square displacement of Brownian motion it holds

$$\langle [x(t + n\Delta t) - x(t)]^2 \rangle = 2Dn\Delta t. \tag{3.28}$$

Then, (3.27) and (3.28) reveal

$$\langle \eta_{\Delta t}^2 \rangle = 2D\Delta t. \tag{3.29}$$

Thus the displacement of a Brownian particle is not proportional to the elapsed time, but rather to its square root. Therefore, we assume the noise shall be (1) zero-centered $\langle \eta_{\Delta t} \rangle = 0$, (2) with amplitude $\langle \eta_{\Delta t}^2 \rangle = 2D\Delta t$, and, (3) be of Gaussian nature. Thus, at each iteration of (3.24), we draw $\eta_{\Delta t}$ from

$$P(\eta_{\Delta t}) = \frac{1}{\sqrt{4\pi D\Delta t}} \exp(-\eta_{\Delta t}^2 / 4D\Delta t). \tag{3.30}$$

We hope to prove that noise in this form will drive the system towards it's equilibrium at temperature T.

3.2 Fokker-Planck Equation

Consider the time evolution of the probability distribution $P(x, t)$ defined as the probability of the particle being at location x at time t. Then, the $t \to \infty$ limit of $P(x, t)$ gives the stationary thermal distribution at temperature T. Rewrite (3.24) as

$$x(t + \Delta t) = x(t) + v(x)\Delta t + \eta_{\Delta t}(t), \tag{3.31}$$

with velocity

$$v(x) = -\mu \frac{dU}{dx}. \tag{3.32}$$

$P(x, t)$ can be derived by the time-discretized Chapman-Kolmogorov equation,

$$P(x, t + \Delta t) = \int_{-\infty}^{\infty} W(x, y; \Delta t) P(y, t) dy \tag{3.33}$$

where $W(x, y; \Delta t)$ is the probability of a particle moving from y at time t to x at time $t + \Delta t$. Then, (3.31) will only move the particle from y to x in time Δt, if $\eta_{\Delta t}$ satisfies

$$x = y + v(x)\Delta t + \eta_{\Delta t}(t). \tag{3.34}$$

Substitution of (3.34) in (3.33) gives

$$W(x,y;\Delta t) = P(\eta_{\Delta t} = x - y - v(y)\Delta t) = \frac{1}{\sqrt{4\pi D\Delta t}} \exp(-(x-y-v(y)\Delta t)^2/4D\Delta t). \quad (3.35)$$

Substitution of (3.35) in (3.33) gives

$$P(x,t+\Delta t)+\partial_t P(x,t)\Delta t = \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi D\Delta t}} \exp(-(x-y-v(y)\Delta t)^2/4D\Delta t)P(y,t)dy. \quad (3.36)$$

Considering the right-hand side of (3.36), when $\Delta t \to 0$, the Gaussian reduces to the delta function and the integral equals $P(x,t)$. We omit the expansion of Δt. Collecting the order Δt terms gives the Fokker-Planck equation

$$\frac{P(x,t)}{\partial t} = -\frac{\partial[v(x)P(x,t)]}{\partial x} + D\frac{\partial^2 P(x,t)}{\partial x^2}, \quad (3.37)$$

or, substituting $v(x) = -\mu\partial U(x)/\partial x$,

$$\frac{\partial P(x,t)}{\partial t} = \mu\frac{\partial}{\partial x}\left(\frac{\partial U}{\partial x}P(x,t)\right) + D\frac{\partial^2 P(x,t)}{\partial x^2}. \quad (3.38)$$

Then, we obtain the stationary solution by

$$0 = \mu\frac{d}{dx}\left(\frac{dU}{dx}P_s(x)\right) + D\frac{d^2 P_s(x)}{dx^2}. \quad (3.39)$$

It can be seen that a solution to (3.39) emerges in the form of

$$P_s(x) = C\exp(-\mu U(x)/D), \quad (3.40)$$

where C denotes the normalization constant. Recall, by (3.23) we know the relation $D = \mu k_B T$, and obtain the stationary distribution function as the equilibrium Boltzmann distribution

$$P_s(x) = \frac{1}{Z}\exp(-U(x)/k_B T). \quad (3.41)$$

CHAPTER 4

Irreversible Langevin Sampler

At this point, we characterized the noise term in (3.1) such that the system is driven towards the equilibrium at temperature T, and the time-evolution of the probability distribution satisfies the appropriate Fokker-Planck equation. Now, let us consider Markov-Chain Monte-Carlo (MCMC) sampling in the form of a Langevin dynamics in the field of Statistics.

4.1 Markov-Chain Monte-Carlo

Reversibility. A Markov chain $X(t)$ on state space Ω is determined by the transition probability

$$K(x,y) = P(X_{t+1} = y \mid X_t = x, X_{t-1}, \ldots, X_0) = P(X_{t+1} = y \mid X_t = x). \tag{4.1}$$

Let $K = (K(x,y)$ and $p^{(t)} = (p^{(t)}(x))$. Then, $p^{(t+1)} = p^{(t)}K$. By induction, $p^{(t)} = p^{(0)}K^t$. Under mild conditions, $p^{(t)} \to \pi$ where π is the stationary distribution for which $\pi = \pi K$ holds. In the special case of reversibility, the Markov chain satisfies

$$\pi(x)K(x,y) = \pi(y)K(y,x) \tag{4.2}$$

for each $(x,y) \in \Omega^2$. (4.2) is know as the detailed balance condition. If a chain is reversible with respect to π, then π is the stationary distribution.

Stationary distribution. If $X(0)$ is given by π, then $X(t > 0)$ is stationary such that it has distribution π for all $t \geq 0$. Such π exists, if and only if it has a positive recurrent state. As such, an invariant distribution is not guaranteed to be unique. A Markov chain is said to be irreducible, if any state is accessible from any other state. A state is said to be positive recurrent, if it's mean return time is finite. A state is said to be aperiodic, when the

period of occurrence of the state is 1. If all states in an irreducible Markov chain are ergodic, then the chain is said to be ergodic. A positive recurrent Markov chain converges to π via $\pi_j = \lim_{t\to\infty} P(X_t = j), j \in \Omega$, if and only if the chain is aperiodic. A positive recurrent and aperiodic Markov chain is sometimes called an ergodic chain. For For an ergodic Markov chain, there is a unique stationary distribution.

4.2 Irreversible Langevin

Reversible Langevin. Consider diffusion $X(t)$ in the form of the overdamped Langevin equation,

$$dX(t) = -\nabla U(X(t))dt + \sqrt{2T}dW(t), \tag{4.3}$$

where U is a given potential, $W(t)$ is a Wiener process and T is temperature. One may consider a Wiener process as the Gaussian noise with a vanishing mean and a variance of dt. We may use (4.3) to sample the Boltzmann-Gibbs measure with density

$$\pi(x) \propto \exp(-U(x)/T), \tag{4.4}$$

which is the invariant measure of (4.3) under some conditions. The infinitesimal generator of (4.3) is symmetric with respect to the invariant measure, and thus the dynamics (4.3) is reversible in time, i.e., it satisfies detailed balance (4.2) [17].

We may discretize (4.3) considering the displacement of a Brownian particle being proportional to the square root of time as revealed in (3.29)

$$X(t + \Delta t) = X(t) - \frac{dU}{dx}\Delta t + \sqrt{2T\Delta t}\epsilon(t), \tag{4.5}$$

where $\langle \epsilon(t) \rangle = 0$ and $\langle \epsilon(t)_i \epsilon(t)_j \rangle = \delta_{ij}$. The Langevin dynamics may be considered as a stochastic variant of gradient descent $X(t+\Delta t) = X(t) - (dU/dx)\Delta t$ on U. If $X(t) \sim \pi(x)$, then the distribution of $X(t + \Delta t)$ will be shifted towards basins of low energy. We may recover π by smoothing with $\sqrt{2T\Delta t}\eta(t)$. One may interpret (4.5) as the ballet of energy and entropy where the term $-dU/dx$ decreases energy and the term $\sqrt{2T\Delta t}\epsilon(t)$ increases entropy.

Irreversible Langevin. If diffusion is applicable as a means of sampling, then one may consider a family of diffusions [10, 20]

$$dX(t) = -\nabla U(X(t))dt + C(X(t))dt + \sqrt{2T}dW(t) = D(X(t))dt + \sqrt{2T}dW(t) \tag{4.6}$$

where the invariant measure is maintained if the irreversible force $C(x)$ satisfies

$$div(C(x)\exp(-U(x)/T)) = 0, \tag{4.7}$$

or, equivalently,

$$div(C) = T^{-1}C \cdot \nabla U. \tag{4.8}$$

One may derive (4.7) by plugging $C(x)$ into the Fokker-Planck equation

$$\frac{\partial P_t(x)}{\partial t} = -\frac{\partial}{\partial x}\left(-\frac{\partial U(x)}{\partial x} + C(x) - T\frac{\partial}{\partial x}\right)P_t(x). \tag{4.9}$$

Then, (4.9) holds if the probabilistic flow vanishes in the steady state of the system

$$-\frac{\partial}{\partial x}\left(C(x)P_t(x)\right) = div(C(x)\exp(-U(x)/T)) = 0. \tag{4.10}$$

A trivial choice of $C(x)$ satisfying (4.7) is

$$C(x) = 0. \tag{4.11}$$

A non-trivial choice of $C(x)$ is

$$C(x) \propto \exp(U(x)/T). \tag{4.12}$$

Another non-trivial choice of $C(x)$ is

$$C(x) = S(\nabla U(x)/T), \tag{4.13}$$

where S can be any skew symmetric matrix satisfying $S^T = -S$. If $C(x)$ is not zero, then the corresponding diffusion, in the form of a Markov process, is irreversible.

Obviously, (4.11) and (4.12) satisfy (4.7). For (4.13), we may write the curl operator $\nabla\times$ of vector field F and observe a skew-symmetric matrix.

$$\nabla \times F = \begin{bmatrix} 0 & -\frac{\partial}{\partial z} & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} & 0 & -\frac{\partial}{\partial x} \\ -\frac{\partial}{\partial y} & \frac{\partial}{\partial x} & 0 \end{bmatrix} F = \begin{bmatrix} \hat{x} & \hat{y} & \hat{z} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_x & F_y & F_z \end{bmatrix}, \tag{4.14}$$

where $\hat{x}$, $\hat{y}$, and $\hat{z}$ denote the unit vectors for the x-, y-, and z-axes, respectively.

In this sense, skew-symmetric matrices can be thought of as infinitesimal rotations. The divergence of the curl of any vector field is equal to zero $\nabla \cdot (\nabla \times F) = 0$,

$$\nabla \cdot \nabla \times F = \nabla \cdot \begin{bmatrix} \hat{x} & \hat{y} & \hat{z} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_x & F_y & F_z \end{bmatrix} = \frac{\partial}{\partial x}\left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F_z}{\partial x} - \frac{\partial F_x}{\partial z}\right) + \ldots = 0. \quad (4.15)$$

While (4.11) is not of particular use, and (4.12) suffers from numerical instability, we will put our emphasis on (4.13). We are interested in how $C(x)$ influences the convergence of the diffusion (4.3) to equilibrium.

Violation of Detailed Balance. Consider $C(x) = \beta S \nabla U(x)$ where $S = (s_{ij})$ with $s_{ij} = 2 \cdot 1(i \leq j) - 1$ such that $S^T = -S$ holds. It appears the strength β of the rotational force can be interpreted as the degree to which one violates the detailed balance condition (4.2). Then, an open endeavour is the rigorous analysis of the relation between β and (4.2).

Double-Well Example. Consider a double-well potential $U(x) = -x^2 + \frac{1}{2}x^4$ where we average over $1,000$ particles and discretize (4.6) with $dt = 1\text{e}{-4}$ depicted in Figure 4.1 and 4.2. The effect of the rotational force in $C(x)$ is apparent and the convergence towards the equilibrium state appears accelerated.

Gaussian Example. We may employ the irreversible Langevin sampler to estimate the mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}$ of a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with energy

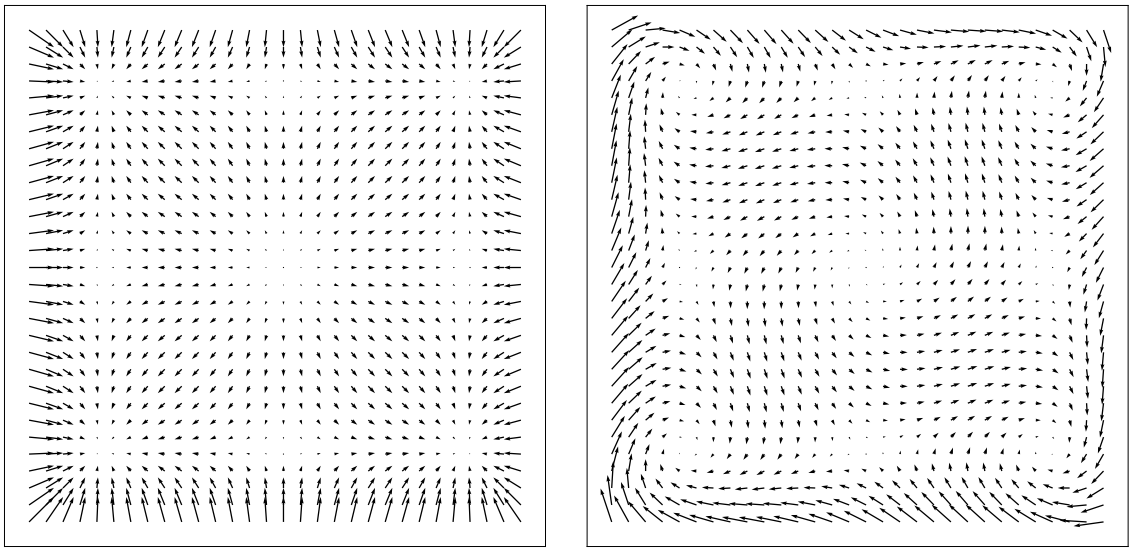
$$U(x) = \frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu). \quad (4.16)$$

We compute the unbiased estimator of Σ is

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T. \quad (4.17)$$

We shall measure some notion of divergence from the underlying distribution

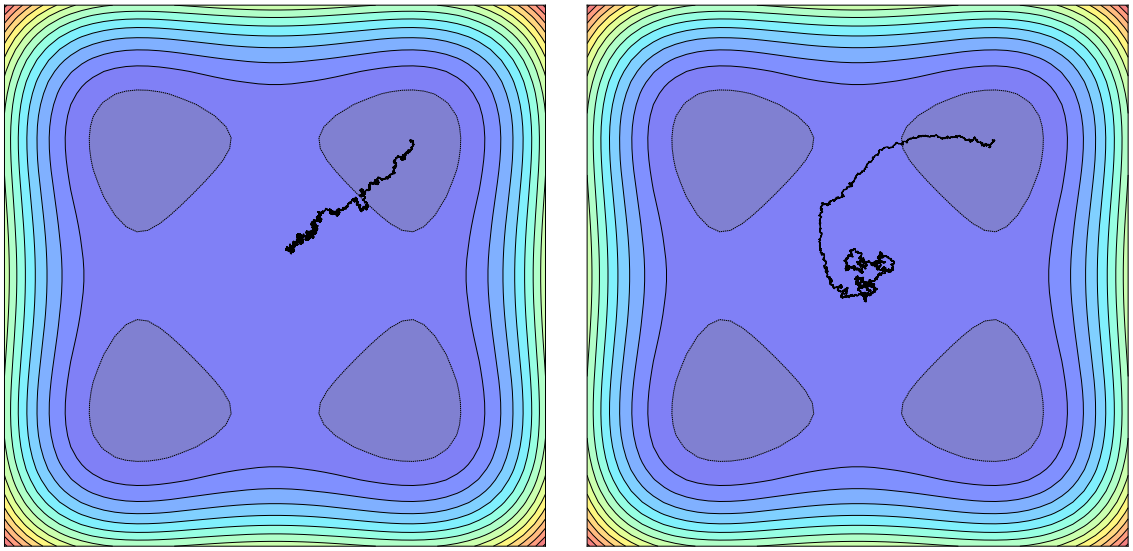
$$\|\mu - \bar{x}\|_2 = \sqrt{\sum_i |\mu_i - \bar{x}_i|^2}, \quad (4.18)$$



(a) $-\nabla U(x)$.

(b) $-\nabla U(x) - 2\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \nabla U(x)$.

Figure 4.1: Vector fields $D(x)$.



(a) $C(x) = 0$.

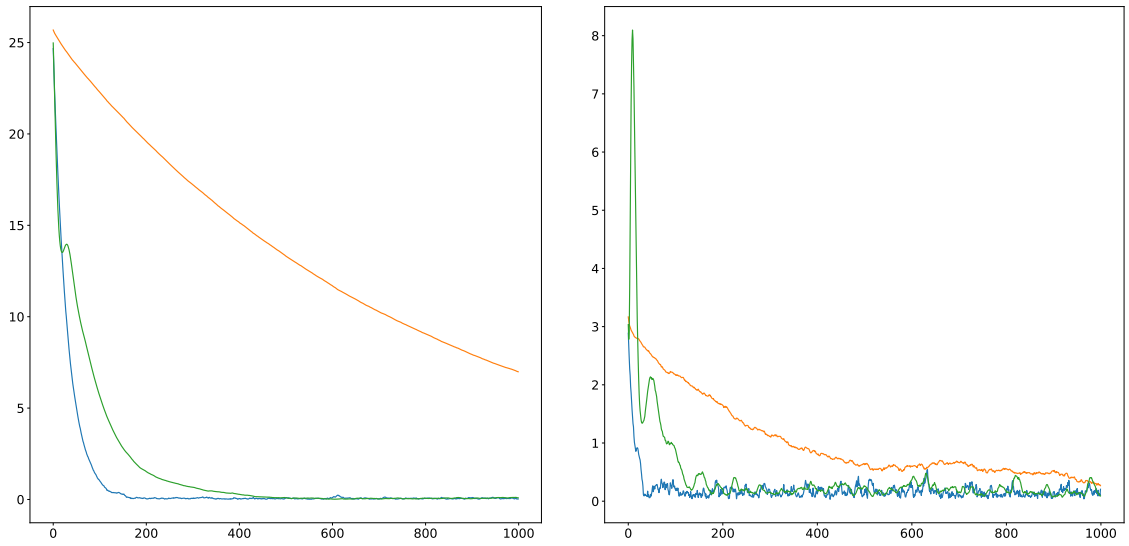
(b) $C(x) = 2\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \nabla U(x)$.

Figure 4.2: Average Langevin trajectory of $1,000$ particles towards the equilibrium (a) without injected vector field and (b) with injected vector field.

and,

$$\|\Sigma - \hat{\Sigma}\|_F = \sqrt{\sum_i \sum_j |(\Sigma - \hat{\Sigma})_{ij}|^2}. \tag{4.19}$$

The dynamics is simulated in the form of $1,000$ Markov chains for $1,000$ steps omitting the burn-in phase at time discretization $dt = 1\text{e}{-1}$ in $d = 5$ dimensions. In Figure (4.3), we compare the estimation based on particles driven by (1) the Hamiltonian Monte-Carlo (HMC) sampler with 5 leap-frog steps and Metropolis-Hastings correction, (2) the reversible Langevin dynamics (4.3), and, (3) the irreversible Langevin dynamics (4.6) where $C(x) = 2S\nabla U(x)$ with $S = (s_{ij})$ such that $s_{ij} = 2 \cdot 1(i \leq j) - 1$. Remarkably, the irreversible



(a) $\|\mu - \bar{x}\|_2$. (b) $\|\Sigma - \hat{\Sigma}\|_F$.

Figure 4.3: Estimation of (μ, Σ) where particles are sampled by HMC (blue), reversible Langevin (orange), irreversible Langevin (green) versus number of sampling iterations.

Langevin sampler appears on par with HMC. In computational terms, integrating Hamiltonian dynamics with discretized time is performed in the leap-frog integration scheme. Note, with a single leap-frog step, the dynamics of HMC reduce to Langevin. Note, that a single MCMC sampling iteration requires 5 leap-frog steps whereas the irreversible Langevin sampler requires a single step.

CHAPTER 5

Accelerated Contrastive Divergence

In this Chapter, since we are now equipped with an understanding of Brownian motion and an efficient sampler in the form of the irreversible Langevin dynamics, we shall put these means to good use and accelerate the learning of energy-based models.

5.1 Contrastive Divergence

Consider a Gibbs distribution with parameters θ in the form of

$$P(X \mid \theta) = \frac{1}{Z(\theta)} \exp(-U(X \mid \theta)), \tag{5.1}$$

where $Z(\theta)$ denotes the intractable partition function and $U(X \mid \theta)$ is the energy function. If the energy is composed as a sum of terms, then we are considering the product of probability distributions. Let $\{X_i, i = 1, \ldots, n\}$ denote a sample. Then the empirical data distribution is

$$P_0(X) = \frac{1}{n} \sum_{i=1}^{n} \delta(X - X_i), \tag{5.2}$$

where $\delta()$ is the Dirac delta function. The average log-likelihood is

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log P(X_i \mid \theta) = -\langle U(X \mid \theta) \rangle_0 - \log Z(\theta). \tag{5.3}$$

We desire to choose θ such that

$$\theta^* = \arg\max_{\theta} L(\theta) = \arg\min_{\theta} KL(P_0 || P_\infty), \tag{5.4}$$

where $P_\infty(x \mid \theta) = P(x \mid \theta)$. Differentiation of the log-likelihood with respect to θ

$$\frac{\partial \log P(X \mid \theta)}{\partial \theta} = \sum_i \frac{\partial}{\partial \theta} \log \left(\frac{1}{Z(\theta)} \exp(-U(X_i \mid \theta)) \right) \tag{5.5}$$

$$= \sum_i \left(-\frac{1}{Z(\theta)} \frac{\partial Z(\theta)}{\partial \theta} - \frac{\partial U(X_i \mid \theta)}{\partial \theta} \right) \tag{5.6}$$

$$= \sum_i \left(\frac{1}{Z(\theta)} \int \frac{\partial U(X \mid \theta)}{\partial \theta} \exp(-U(X \mid \theta)) dX - \frac{\partial U(X_i \mid \theta)}{\partial \theta} \right) \tag{5.7}$$

$$\propto \langle \frac{\partial U(X \mid \theta)}{\partial \theta} \rangle_\infty - \langle \frac{\partial U(X \mid \theta)}{\partial \theta} \rangle_0, \tag{5.8}$$

where $\langle \cdot \rangle_\infty$ denotes the average with respect to the model distribution and $\langle \cdot \rangle_0$ denotes the average using the sample. The two terms in (5.8) are referred two as positive and negative phase as the first term increases the probability of the data, while the second term decreases the probability of samples drawn from the model. While $\langle \cdot \rangle_0$ is readily available, $\langle \cdot \rangle_\infty$ involves the partition function.

That is, the log-likelihood gradient (5.5) performs traditional maximum likelihood (ML) learning, but it is intractable. To avoid this computational difficulty, Hinton proposed Contrastive Divergence (CD) [8] learning where the gradient (5.12) approximately follows the gradient of a different function. CD does noisy gradient ascent (noise due to averaging over a finite number of samples) on an objective function which approximates the log-likelihood.

$$\frac{\partial L(\theta)}{\partial \theta} = \langle \frac{\partial}{\partial \theta} U(X \mid \theta) \rangle_\infty - \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} U(X_i \mid \theta) \tag{5.9}$$

$$= \langle \frac{\partial}{\partial \theta} U(X; \theta) \rangle_\infty - \langle \frac{\partial}{\partial \theta} U(X \mid \theta) \rangle_0 \tag{5.10}$$

$$\approx \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \frac{\partial}{\partial \theta} U(\tilde{X}_i \mid \theta) + \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} U(X_i \mid \theta) \tag{5.11}$$

$$= \langle \frac{\partial U(X \mid \theta)}{\partial \theta} \rangle_n - \langle \frac{\partial U(X \mid \theta)}{\partial \theta} \rangle_0. \tag{5.12}$$

We transform p_0 into p_n by starting a Markov chain from the data distribution p_0 and run the chain for n number of steps. That is, $p_n = T^n p_0$ where T is the Markov operator.

Maximum likelihood learning minimizes the Kullback-Leibler divergence

$$KL(p_0||p_\infty) = \sum_X p_0(X) \log \frac{p_0(X)}{p_\infty(X)} \tag{5.13}$$

$$= \sum_X p_0(X) \log p_0(X) - \sum_X p_0(X) \log p_\infty(X) \tag{5.14}$$

$$= -H(p_0(X)) - \sum_X p_0(X) \log p_\infty(X) \tag{5.15}$$

$$\propto -\sum_X p_0(X) \log p_\infty(X) \tag{5.16}$$

$$= -\sum_X \left[\frac{1}{n} \sum_{i=1}^{n} \delta(X - X_i) \right] \log p_\infty(X) \tag{5.17}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \log p(X_i \mid \theta) = L(\theta). \tag{5.18}$$

Contrastive divergence learning approximately follows the gradient of two divergences

$$CD_n = KL(p_0||p_\infty) - KL(p_n||p_\infty). \tag{5.19}$$

The first term minimizes divergence between p_∞ and p_0 which increases the likelihood of the sample $\{X_i, i = 1, \ldots, n\}$ (by reducing the corresponding energy), while the second term maximizes the divergence between p_n and p_∞ which decreases the likelihood of MCMC samples generated by p_∞. We may recover the gradient (5.12) by neglecting the third term

$$\frac{\partial}{\partial \theta} CD_n = \frac{\partial}{\partial \theta} \left(KL(p_0||p_\infty) - KL(p_n||p_\infty) \right) \tag{5.20}$$

$$= \langle \frac{\partial}{\partial \theta} U(X \mid \theta) \rangle_0 - \langle \frac{\partial}{\partial \theta} U(X \mid \theta) \rangle_n \tag{5.21}$$

$$- \frac{\partial}{\partial \theta} p_n(X) \frac{\partial}{\partial p_n(X)} KL(p_n||p_\infty) \tag{5.22}$$

$$\approx \langle \frac{\partial}{\partial \theta} U(X \mid \theta) \rangle_0 - \langle \frac{\partial}{\partial \theta} U(X \mid \theta)) \rangle_n. \tag{5.23}$$

5.2 Restricted Boltzmann Machine

Bernoulli-Bernoulli RBM. Consider a restricted Boltzmann Machine with weights $W = (w_{ij}) \in \mathbb{R}^{m \times n}$ connecting n stochastic hidden units $\{h_j\} \in \{0, 1\}^n$ and m visible units

$\{v_i\} \in \{0,1\}^m$, with bias terms $\{a_i\}$ and $\{b_j\}$

$$P(v,h) = \frac{1}{Z}\exp(-U(v,h)) \tag{5.24}$$

with energy of state $\{v,h\}$ given parameters $\theta = \{a,b,W\}$

$$U(v,h \mid \theta) = -a^T v - b^T h - v^T W h. \tag{5.25}$$

Then, the conditional probabilities of configurations are given by

$$P(v \mid h) = \prod_{i=1}^{m} P(v_i \mid h), \tag{5.26}$$

$$P(h \mid v) = \prod_{j=1}^{n} P(h_j \mid v). \tag{5.27}$$

Further, the activation probabilities are given by

$$P(h_j = 1 \mid v) = \sigma\left(\sum_{i=1}^{m} w_{ij} v_i + b_j\right), \tag{5.28}$$

$$P(v_i = 1 \mid h) = \sigma\left(\sum_{j=1}^{n} w_{ij} h_j + a_i\right), \tag{5.29}$$

where

$$\sigma(x) = \frac{1}{1+\exp(-x)}. \tag{5.30}$$

Given a training set V, we may learn (5.24) by maximum likelihood

$$\arg\max_W \prod_{v \in V} P(v) = \arg\max_W \mathbb{E}(\log P(v)), \tag{5.31}$$

in the form of Contrastive Divergence and employ Block Gibbs sampling.

Gaussian-Bernoulli RBM. The energy function of a RBM with Gaussian visible units assumes the form

$$U(v,h \mid \theta) = \frac{1}{2}\sum_i (v_i - a_i)^2 - \sum_j h_j (b_j + \sum_i v_i W_{ij}), \tag{5.32}$$

with parameters $\theta = \{a,b,W\}$. The conditions distributions are given by

$$P(h_j \mid v) = Bernoulli\left(h_j \mid \sigma(b_j + \sum_i v_i W_{j,i})\right) \tag{5.33}$$

$$P(v_i \mid h) = Normal\left(v_i \mid a_i + \sum_j h_j W_{j,i}\right) \tag{5.34}$$

As potential energy U we will use thee free-energy U_f

$$P(v \mid \theta) = \frac{1}{Z(\theta)} \exp(-U_f(v \mid \theta)) \tag{5.35}$$

$$= \sum_h P(v, h \mid \theta) \tag{5.36}$$

$$= \sum_h \frac{1}{Z(\theta)} \exp(-U(v, h \mid \theta). \tag{5.37}$$

Then,

$$U_f(v \mid \theta) = -\log \sum_h \exp(-U(v, h \mid \theta)) \tag{5.38}$$

$$= \frac{1}{2} \sum_i (v_i - a_i)^2 - \sum_j \log \left(1 + \exp\left(b_j + \sum_i v_i W_{j,i}\right)\right). \tag{5.39}$$

Training. While the factorization of the restricted Boltzmann machine allows for efficient training by the Block Gibbs sampling were we alternate the sampling of visible and hidden units, in the case of the Gaussian-Bernoulli RBM a Hamilitonian Monte-Carlo with Metropolis-Hasting correction or the irreversible Langevin sampler may be more adequate. We simulate particles in the negative phase of persistent Contrastive Divergence by irreversible Langevin dynamics (4.6) given the potential in form of the free energy U_f. Note, that in the training phase there is an interplay between the particle dynamics and the changing energy landscape. When a particle is get trapped within a ravine of the energy landscape, eventually, may get dislodged as the parameters change. This phenomenon is said to improve mixing of the Markov chains [23].

The training was performed on the MNIST dataset of $70,000$ handwritten digits [15]. Block Gibbs sampling, Hamiltonian Monte-Carlo sampling with Metropolis-Hastings correction, the reversible and irreversible Langevin samplers were tried. While training the model with HMC appears feasible, adjustment of the hyper-paremeters such that desirable synthesis is achieved, remains non-trivial.

It is noteworthy, that training by HMC required to lower the temperature T of the system such that low kinetic energy does not cause high potential energy with constant Hamiltonian.

Synthesis. The images depicted in Figure (5.2) are sampled by irreversible Langevin dynamics with small rotational force. In the notion of simulating annealing, we gradually lowered the temperature of the system such that the energy gradient term becomes dominate in the dynamics while entropy vanishes. This recipe appears to improve the visual fidelity of the sampled images.

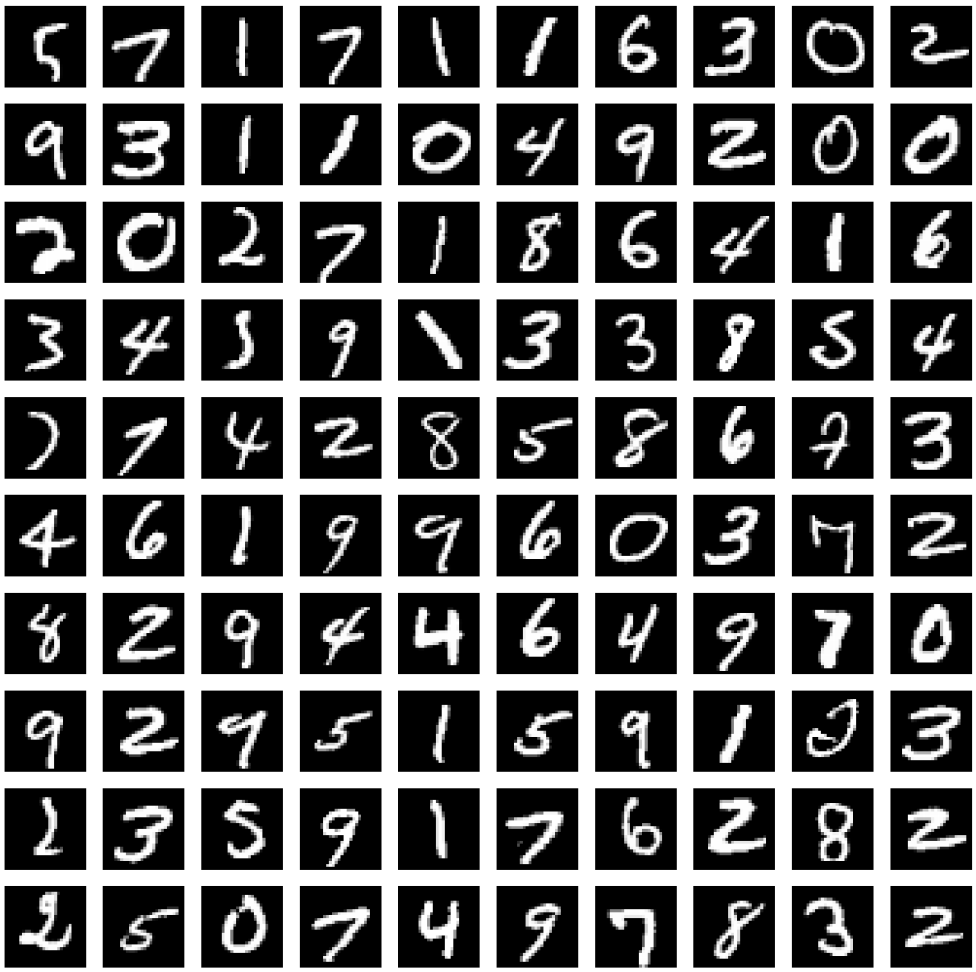


Figure 5.1: Subset of the MNIST dataset.

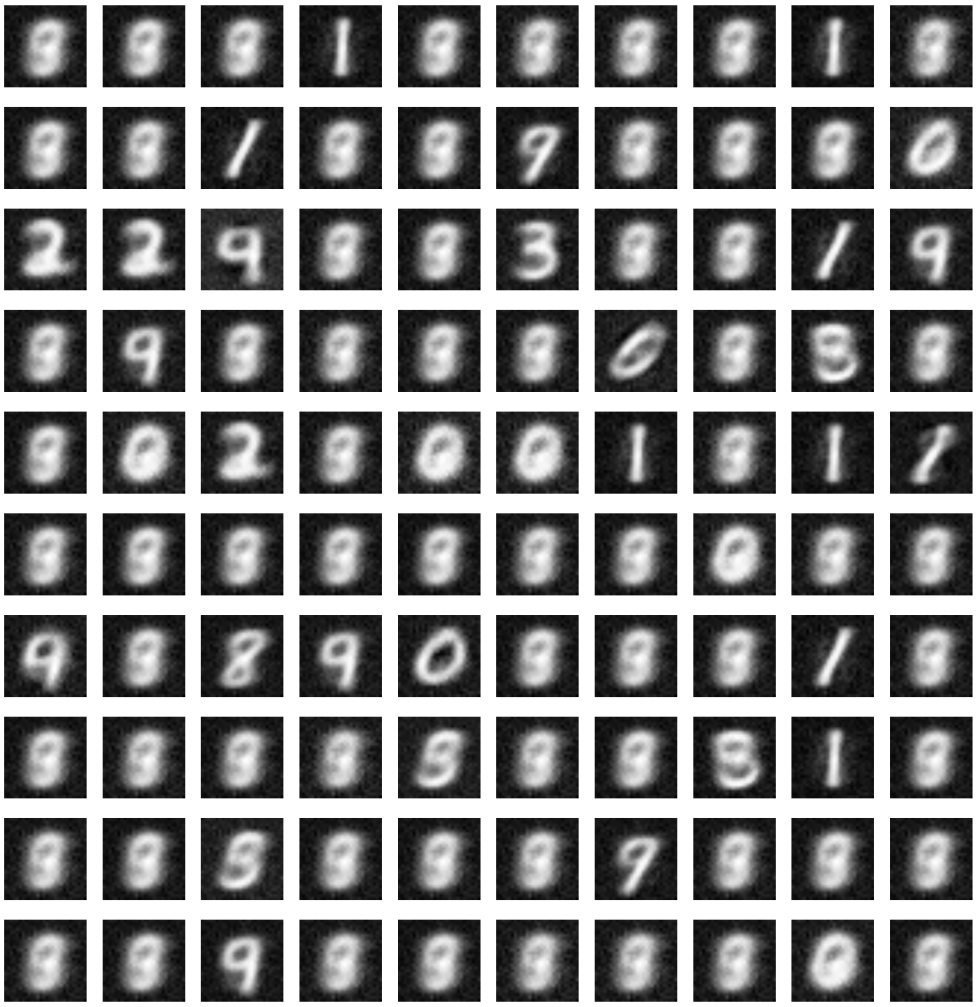


Figure 5.2: Synthesized samples of single hidden-layer Gaussian-Bernoulli Restricted Boltzmann Machine trained on MNIST.

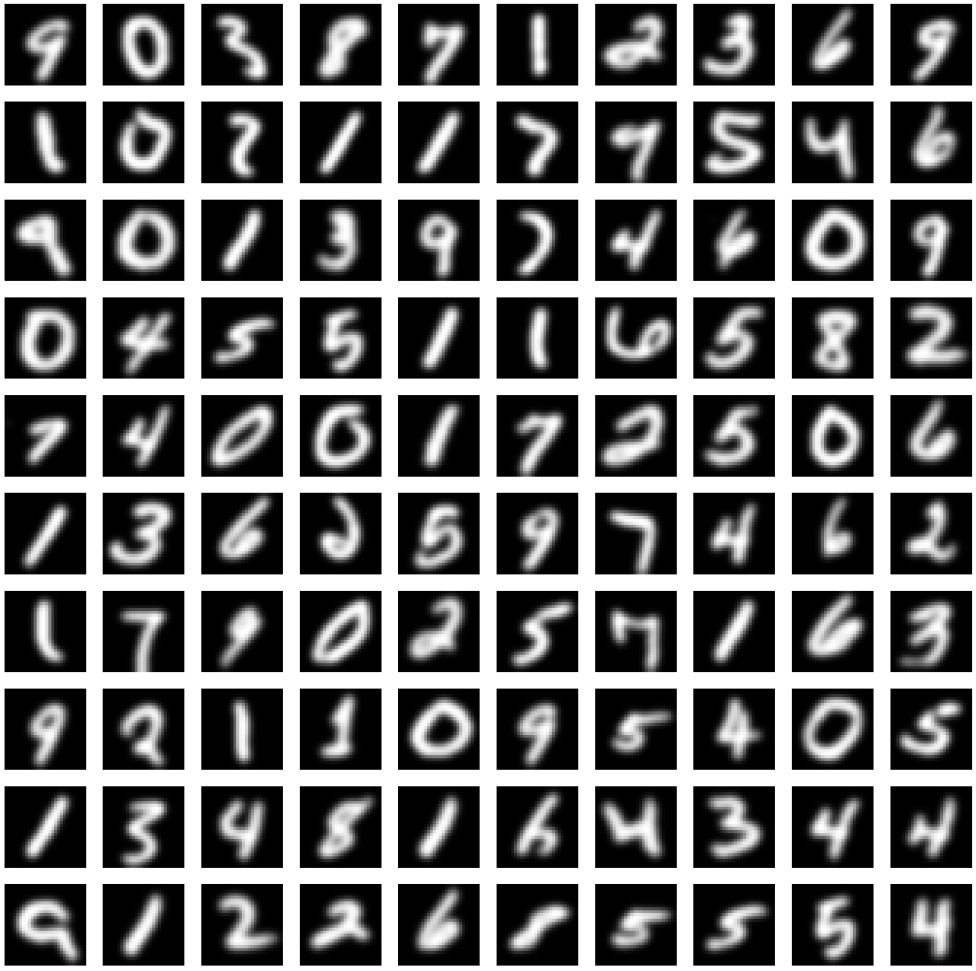


Figure 5.3: Synthesized samples of two hidden-layer Gaussian-Bernoulli Restricted Boltzmann Machine trained on MNIST.

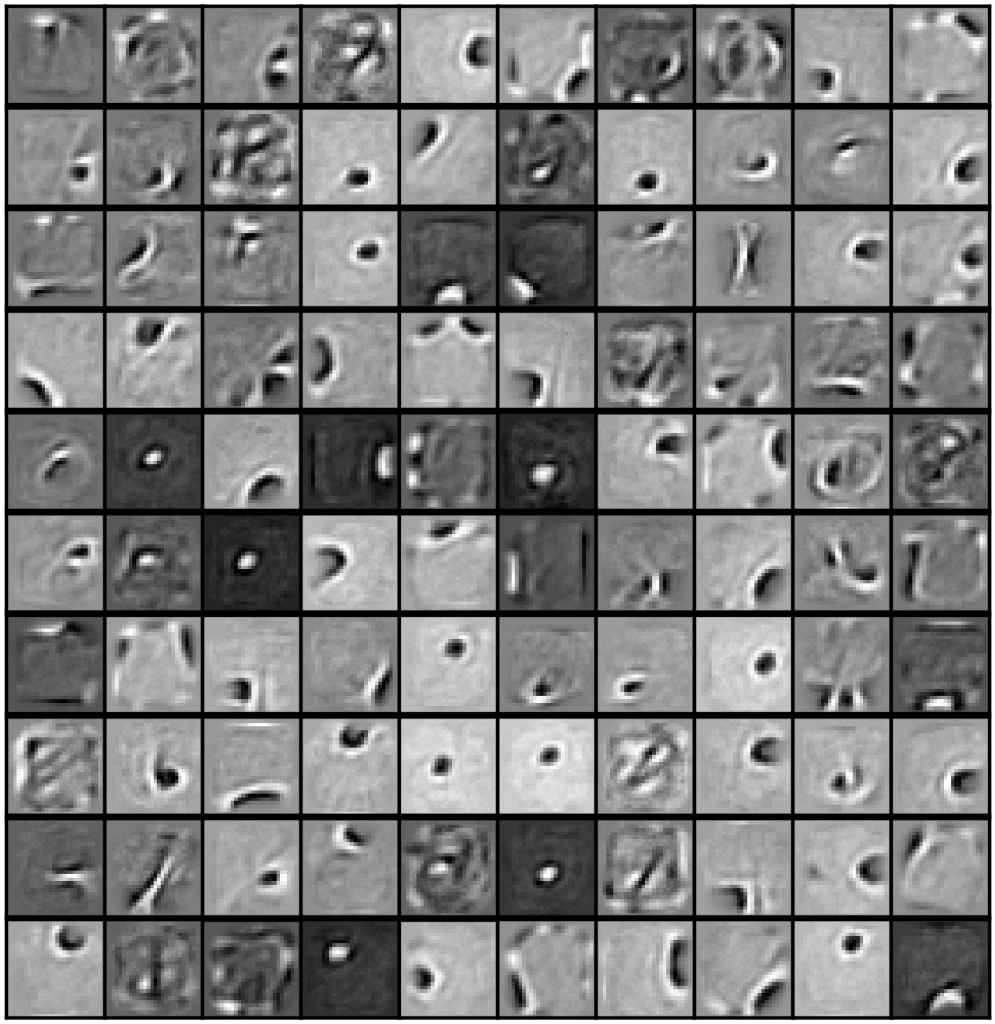


Figure 5.4: Filters learned on MNIST.

CHAPTER 6

Conclusion

We conclude the thesis with briefly revisiting the Chapters 3-5 and discuss several research opportunities with emphasis on energy-based models and learning dynamics.

In Chapter 3, we will briefly revised the relation between the Langevin equation and the Fokker-Planck equation and discussed the properties of the noise term such that the system is driven towards it's equilibrium.

In Chapter 4, we considered diffusion in the form of the overdamped Langevin equation and introduced a family of altered diffusions where the invariant measure is maintained if the irreversible force $C(x)$ satisfies $div(C) = T^{-1}C \cdot \nabla U$. We showed how to derive (1.4) by the Fokker-Planck equation and discuss particular choices of $C(x)$ satisfying (1.4). We illustrated the accelerations of the diffusion towards it's equilibrium state against a Hamiltonian Monte-Carlo sampler with Metropolis-Hastings correction. The Chapter was concluded illustrating how a rotational force in the altered Langevin diffusion can be utilized to accelerate the estimation of the covariance of a simple Gaussian distribution.

In Chapter 5, the irreversible Langevin sampler was employed in the negative phase of Contrastive Divergence learning of a Gaussian-Bernoulli Restricted Boltzmann machine. These machines are known to be notoriously difficult to train, but preliminary results obtained by sampling with the irreversible Langevin dynamics appear to be promising. We depicted synthesized samples from one- and two-layer Gaussian-Bernoulli Restricted Boltzmann machines trained by irreversible Langevin sampling.

From a theoretical view, it remains an open question how the rate of convergence depends on $C(x)$. From a practical view, one may try the irreversible Langevin sampler on a variety of energy-based models. Empirically, investigate the convergence of the training pro-

cess of Gaussian-Bernoulli Restricted Boltzmann machines under varying sampling schemes. Further, the properties such as smoothness of the learned energy landscape might be of interest. Finally, it might be worth considering to introduce Metropolis-Hasting correction as proposed in the Metropolis-adjusted Langevin algorithm [21].

In this sense, the thesis may be understood as a first step towards the application of accelerated diffusion in energy-based models. In the future, we hope to revive active research in energy-based models such these models gain the attention they deserve.

REFERENCES

[1] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A Learning Algorithm for Boltzmann Machines*. *Cognitive Science*, 9(1):147–169, January 1985.

[2] Yali Amit and Ulf Grenander. Comparing sweep strategies for stochastic relaxation. *Journal of multivariate analysis*, 37(2):197–222, 1991.

[3] A. Barbu and S. C. Zhu. Graph partition by swendsen-wang cuts. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 320–327 vol.1, Oct 2003.

[4] P Charbonnier, G Aubert, L Blanc-Feraud, and M Barlaud. Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. *IEEE Trans. Image Processing*, 6(2):298–311, 1997.

[5] A. Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Ann. Phys.*, 322(8):549–560, January 1905.

[6] A. D. Fokker. Die mittlere Energie rotierender elektrischer Dipole im Strahlungsfeld. *Annalen der Physik*, 348(5):810–820, 1914.

[7] Walter R Gilks and Gareth O Roberts. Strategies for improving mcmc. *Markov chain Monte Carlo in practice*, 6:89–114, 1996.

[8] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, August 2002.

[9] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[10] C.-R. Hwang, S.-Y. Hwang-Ma, and S.-J. Sheu. Accelerating diffusions. *ArXiv Mathematics e-prints*, May 2005.

[11] Chii-Ruey Hwang, Shu-Yin Hwang-Ma, Shuenn-Jyi Sheu, et al. Accelerating diffusions. *The Annals of Applied Probability*, 15(2):1433–1444, 2005.

[12] Chii-Ruey Hwang and Shuenn-Jyi Sheu. A remark on the ergodicity of systematic sweep in stochastic relaxation. In *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis*, pages 199–202. Springer, 1992.

[13] Ernst Ising. Contribution to the Theory of Ferromagnetism. *Z. Phys.*, 31:253–258, 1925.

[14] Paul Langevin. Sur la théorie du mouvement brownien. *C. R. Acad. Sci. Paris*, 146:530–533, 1908.

[15] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[16] Don S. Lemons and Anthony Gythiel. Paul langevin's 1908 paper "on the theory of brownian motion" ["sur la théorie du mouvement brownien," c. r. acad. sci. (paris) 146, 530–533 (1908)]. *American Journal of Physics*, 65(11):1079–1081, 1997.

[17] J. Lu and K. Spiliopoulos. Analysis of multiscale integrators for multiple attractors and irreversible Langevin samplers. *ArXiv e-prints*, June 2016.

[18] M. Planck. Über einen satz der statistischen dynamik und seine erweiterung in der quantentheorie. *Sitzungsberichte der Preussischen Akademie der Wissenschaften zu Berlin*, 1917.

[19] Zoltán Rácz. From langevin to fokker-planck equation.

[20] L. Rey-Bellet and K. Spiliopoulos. Irreversible Langevin samplers and variance reduction: a large deviations approach. *Nonlinearity*, 28:2081, July 2015.

[21] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

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

[23] Ruslan Salakhutdinov and Hugo Larochelle. Efficient learning of deep boltzmann machines. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 693–700, 2010.

[24] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Phys. Rev. Lett.*, 35:1792–1796, Dec 1975.

[25] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE, 1986.