A Minimum Velocity Approach to Learning
MPLab TR 2007.1

Javier R. Movellan
Institute for Neural Computation
University of California San Diego
La Jolla, CA 92093-0515
2007

This paper investigates the relationship between CD, SM, and minimization of probability velocity fields in stochastic diffusion networks. By doing so it reveals a practical way for training stochastic diffusions and a theoretical way of thinking about adaptive computing in terms of manipulation of probability fields and of the resulting probability currents.

Learning Deterministic Equilibria

x(t+1) = xt + f(xt, w)

Consider a deterministic recurrent neural network model of the form $dx_t/dt = w(\theta - x_t)$, where x_t is a vector of neural activations, w is a fixed positive definite matrix of synaptic connections, and θ is an adaptive bias vector. Since the network is linear it is easy to find an analytical solution to the network activation process. In particular, $\lim_{t\to\infty} x_t = \theta$, i.e., as time progresses the network activations converge to θ. Suppose we want for this network to exhibit a pattern of activation ξ at equilibrium. The standard approach would be to minimize the difference between the desired and obtained equilibrium conditions, i.e., $\|\theta - \xi\|^2$. The gradient of this cost function is proportional to $(\theta - \xi)$ and thus gradient descent learning would move θ in the direction of ξ, converging to $\theta = \xi$. A disadvantage of this approach is that the training signals depend on equilibrium statistics. For linear networks this is not a problem because they can be obtained analytically. However, for the general case we would have to simulate the network numerically until equilibrium, a process that may be time consuming.

The previous approach is analogous to teaching a child to ride a bicycle by letting her fall (the equilibrium condition before learning has occurred) and providing a training signal every time she falls. An alternative approach, which here we refer to as *minimum velocity*, is to provide a training signal every time the child deviates from the desired equilibrium condition, without letting her fall at all. The approach avoids the costly process of converging to equilibrium, thus potentially maximizing the speed of learning. Going back to our recurrent network problem, a minimum velocity approach, would initialize the network with the desired pattern ξ, and use the initial velocity as the cost function, i.e.,

$$\Big\| \frac{dx_t}{dt}\Big|_{x_t=\xi} \Big\|^2 = (\theta - \xi)' w' w (\theta - \xi). \tag{1}$$

GD of the squared initial velocity moves θ towards ξ and converges when $\theta = \xi$, at which point the desired pattern has been learned.

Learning Stochastic Equilibria: BMs

The idea of learning by minimizing velocity can also be applied for stochastic neural networks to exhibit a desired equilibrium probability distribution, e.g., to learn the statistics of natural images. Here we study how the minimum velocity approach applies to *diffusion networks*, a stochastic version of continuous time, continuous state version of RNNs (Movellan, 1998; Movellan and McClelland, 1993; Movellan et al., 2002). A particularly useful type of diffusion network has activation dynamics that, on average, follow the gradient of a *potential function*

$$dX_t = \nabla_x \phi(X_t)dt + \sigma dB_t, \tag{2}$$

where $-\phi(x)$ is the *potential* of the activation state. The negative of the potential, i.e. $\phi(x)$ can be interpreted as the degree of match between the pattern x and the type of patterns the the network "likes" to generate. The term σ is a fixed parameter controlling the degree of randomness in the network, and dB_t is a stochastic *Brownian motion differential* (Movellan, 2006; Oksendal, 1992). This equation can be interpreted as the limit, as $\Delta t \to 0$ of:

$$\Delta X_t = \nabla_x \phi(X_t)\Delta t + \sigma \sqrt{\Delta t} Z_t, \tag{3}$$

where $Z_t \sim N(0,1)$. It is useful to analyze the diffusion network dynamics in probability space, rather than activation space. Due to the fact that probability conserves, i.e., it integrates to unity at all time steps, the distribution of activations obeys the Fokker-Planck-Kolmogorov (FPK) equation

$$\frac{\partial p_t(x)}{\partial t} = -\nabla_x \cdot J_t(x) \tag{4}$$

$$J_t(x) \stackrel{\text{def}}{=} p_t(x) V_t(x) \tag{5}$$

$$V_t(x) \stackrel{\text{def}}{=} \nabla_x \Big(\phi(x) - \frac{\sigma^2}{2} \log p_t(x)\Big), \tag{6}$$

where p_t is the distribution of activations at time t, "$\nabla \cdot$" is the *divergence operator* (see Appendix), J is the *probability current*, and V is the *probability velocity*. The probability velocity plays a critical role in this paper: for each time t and each activation state x, the probability velocity $V_t(x)$ is a vector whose magnitude represents the rate at which probability flows out of the state and whose orientation represents the direction towards which it flows. In diffusion networks probability behaves as a substance moving about the different network states according to standard fluid dynamic equations.

Under mild conditions that include the fact that the potential function shall be bounded from below and shall grow sufficiently fast as $|x|$ increases (Movellan, 1998), the FPK equation converges to a unique probability distribution, known as the *equilibrium distribution*. The equilibrium solution can be obtained by setting the velocity field $V_t = 0$

$$p_\infty(x) \stackrel{\text{def}}{=} \lim_{t\to\infty} p_t(x) = \frac{e^{\frac{2}{\sigma^2}\phi(x)}}{Z}, \tag{7}$$

Thus the equilibrium distribution is *Boltzmann* on the potential $-\phi$.

MLE/ min KLD

Many problems of interest in machine learning and statistics are formally equivalent to the problem of training diffusion networks to exhibit desired equilibrium distributions. An approach for doing so is to let the network achieve stochastic equilibrium, define a cost function that represents the *divergence* between the desired and the obtained equilibrium distributions, and change the network parameters θ via GD on the divergence

D(X0, X∞)

SDE: X0 — > X∞ (φ: potential)

$$D(X_\infty, \xi) \stackrel{\text{def}}{=} \lim_{t\to\infty} D(X_t, \xi) \tag{9}$$

where ξ is a target rv whose distribution we want to learn, and D is the *K-L divergence*.

$$\boxed{D(\xi, X_\infty) = \frac{2}{\sigma^2} E[\phi(\xi)] - \log Z + H(\xi),} \tag{10}$$

==> the gradient of the K-L divergence wrt parameters θ:

$$\nabla_\theta D(X_\infty, \xi) = \frac{2}{\sigma^2} \Big(E[\nabla_\theta \phi(\xi)] - E[\nabla_\theta \phi(X_\infty)] \Big). \tag{13}$$

Which is the continuous time version of the BM learning algorithm. The main practical difficulty with this approach: requires computing an equilibrium statistic, $E[\nabla_\theta \phi(X_\infty)]$, which in general may come at significant time cost. This is the main reason why BMs have not been competitive in practical applications

Minimum Velocity and Score Matching

Vt=0(X0)

MV/SM

Alternatively the principle of learning by minimizing velocity can be applied. The network can be initialized to the desired distribution, i.e., $X_0 = \xi$, and then the network parameters can be trained to minimize the resulting probability velocity field:

$$\min \quad E[\|V_0(\xi)\|^2]$$

$$\begin{aligned} E[\|V_0(\xi)\|^2] &= E[\|\nabla_x \phi(\xi) - \frac{\sigma^2}{2} \nabla_x \log p(\xi)\|^2] \\ &= \frac{\sigma^4}{4} E[\|\nabla_x \log p_\infty(\xi) - \nabla_x \log p(\xi)\|^2]. \end{aligned} \tag{14}$$

Thus, minimizing velocity is equivalent to matching the gradients of the desired distribution and the equilibrium distribution, *score matching* (Hyvarinen, 2005, 2006). One difficulty with this formula is that it requires to compute $\nabla_x \log p(\xi)$, the gradient of the desired distribution. This may be tricky since in most practical cases we know how to sample from the desired distribution but we do not know its gradient.

SM can be solved without the gradient of the desired distribution. Applying integration by parts (See Appendix, Corollary 2) and taking gradients CONFIRM EQ BELOW

$$\nabla_\theta E[\|V_0(\xi)\|^2] = \nabla_\theta \Big(E[\|\nabla_x \phi(\xi)\|^2] + \sigma^2 E[\nabla^2 \cdot \phi(\xi)] \Big). \tag{15}$$

Minimum Velocity and Exact Contrastive Divergence

Free energy statistic of X_t,

$$F_t := -\left(E[\phi(X_t)] + \frac{\sigma^2}{2} H_t\right). \tag{16}$$

Free energy is uniquely minimized by the network's equilibrium distribution (See Appendix, Lemma 1). It can be shown that the rate of change in the free energy is always negative and proportional to the norm of the probability velocity field (See Appendix, Theorem 4)

$$\boxed{\frac{dF_t}{dt} = -E[\|V_t(X_t)\|^2].} \tag{17}$$

Thus, the Free Energy of the network never increases minimizing the rate of change in free energy <==> minimizing the probability velocity field.

Contrastive divergence (Hinton, 2002) is an approach to learning equilibrium distributions based on minimization of a difference between two K-L divergences, i.e., a **contrastive divergence**

$$\begin{aligned} C_t := D(\xi, X_\infty) - D(X_t, X_\infty) &= \frac{2}{\sigma^2} E[\phi(\xi)] + H(\xi) \\ &\quad - \frac{2}{\sigma^2} E[\phi(X_t)] - H(X_t) \\ &= \frac{2}{\sigma^2}\left(F_0(\xi) - F_t(X_t)\right), \end{aligned} \tag{19}$$

proportional to a difference of free Energies,

as $t \to 0$

$$\boxed{\left.\frac{dC_t}{dt}\right|_{t=0} = -\frac{2}{\sigma^2} \left.\frac{dF_t}{dt}\right|_{t=0} = \frac{2}{\sigma^2} E[\|V_0(\xi)\|^2].} \tag{20}$$

Thus minimizing the initial contrastive divergence <==> to minimizing the initial probability velocity field.

Minimum Velocity and Standard Contrastive Divergence

the gradient of the CD can be approximated :

$$\nabla_\theta \Big(E[\phi(X_t^{\theta'}, \theta)] - E[\phi(\xi, \theta)] \Big) \tag{21}$$

Where in discrete time systems, t stands for a integer number of time steps. The analogous for a continuous time system:

$$\frac{1}{\Delta t} \nabla_\theta \Big(E[\phi(X_{\Delta t}^{\theta'}, \theta)] - E[\phi(\xi, \theta)] \Big) \tag{22}$$

Turns out, in continuous time systems as $\Delta \quad t \to 0$ this converges to the exact gradient of the initial velocity, and thus the exact gradient of the initial divergence

$$\nabla_\theta E[\|V_0(\xi)\|^2] = \frac{\sigma^2}{2} \frac{d}{d_t} E[\nabla_\theta \phi(X_t^{\theta'}, \theta)] \Big|_{t=0, \theta'=\theta} \tag{23}$$

Partially Observable Case LVM

In many cases of interest the state X_t can be partitioned into observable units Z_t and hidden units, i.e $X_t = (Z_t, H_t)'$. In this case the goal is to learn a probability distribution over the observable units. One way to approach this problem is to initialize so they are at stochastic equilibrium, given the desired observable states, i.e. $Z_0 = \xi$ and $p_0(h|givenz) = p_\infty(h \,|\, z)$ and make the goal of learning to minimize the joint probability velocity over the initial joint state of observable and hidden variables. As in the fully observable case, the probability velocity measures the distance between two distributions

$$\begin{aligned} J(\theta', \theta) &= \frac{\sigma^4}{4} E[\|\boxed{V_0^\theta(\xi, H_0^{\theta'})}\|^2] \\ &= \int p_\xi(z) \int p_\infty^{\theta'}(h \mid z) \|\nabla_{z,h} \log p_\infty^\theta(z, h) - \nabla_{z,h} \log p_\xi(z) p_\infty^{\theta'}(h \mid z)\|^2 dh dz \end{aligned} \tag{24}$$

$$\lim_{\theta \to \theta'} J(\theta', \theta) = \int p_\xi(z) \|\nabla_z \log p_\infty^\theta(z) - \nabla_z \log p_\xi(z)\|^2 dz \tag{25}$$

Thus

$$J(\theta_1, \theta_2) < SM(\theta_1) \tag{26}$$

$$J(\theta_2, \theta_3) < SM(\theta_2) \tag{27}$$

$$J(\theta_3, \theta_4) < SM(\theta_3) \cdots \tag{28}$$

Thus, minimizing the joint probability velocity ends up minimizing the Fisher score distance between the desired distribution and the obtained distribution for the observable units Since the problem now is fully observable, the gradient of the probability velocity wrt θ can be obtained using Hinton's learning rule

$$\nabla_\theta J(\theta, \theta') = \nabla_\theta E[\phi(X_t^{\theta'}, H_t^{\theta'}, \theta)] - \nabla_\theta E[\phi(\xi, H_0^{\theta'}, \theta)] \tag{29}$$

or, equivalently the score matching learning rule .

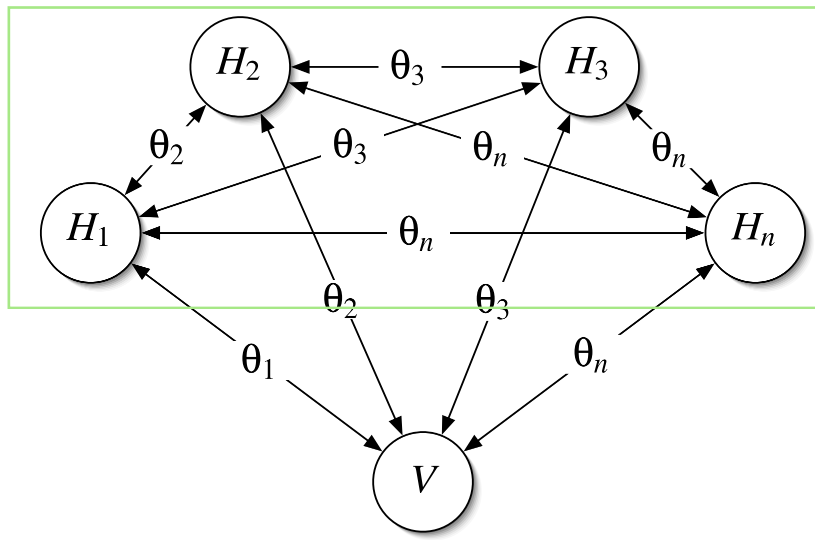

Figure 1: *A Cascaded Network.*

Sequential Minimum Velocity Learning

Here we show that the deep belief network learning algorithm is a special case of minimum velocity learning. In particular as more networks are added to the layer the velocity of the visible velocity decreases.

Let $\theta = (\theta_0, \theta_n)$ were θ_i is the parameter vector for the i^{th} layer (see Figure **??**). We let V represent the visible units and H_i the hidden units of the i^{th} layer. Assume we have an energy model :

$$\phi(v, h_1, \cdots, h_n, \theta) = \phi_0(v, \theta_0) + \phi_1(v, h_1, \theta_1) + \phi_2(v, h_1, h_2, \theta_2) + \cdots + \phi_n(v, h_1, h_2, \cdots, h_{n-1}, h_n, \theta_n) \quad (30)$$

In networks like the BM this is always possible to do.

Here we show that these networks can be efficiently trained in a sequential manner. On each stage a different set of weights is trained, leaving the previously trained parameters fixed. An additional advantage is that it is not necessary to resample the hidden units at equilibrium, i.e., the sample of hidden units from previous layers are treated as observable units for training the next stage of weights.

$$\theta^{(k)} = (\theta_0^{(k)}, \theta_1^{(k)}, \cdots, \theta_n^{(k)}) \quad (31)$$

represent the network parameter after k learning cycles.

To begin with we set all the connections to zero, i.e.,

$$\theta^{(0)} = (0, 0, \cdots, 0) \quad (32)$$

On the first stage we vary θ_0, to minimize the velocity of the observable units. To this effect we collect s samples ~ target rv ξ:

$$\{v^{(1)}, v^{(2)}, ..., v^{(s)}\} \tag{33}$$

and chose a value $\hat{\theta}_0$ of θ_0 that minimizes the velocity of the observable units

$$J(\theta^{(0)}, \theta^{(1)}) < J(\theta^{(0)}, \theta^{(0)}) \tag{34}$$

where

$$\theta^{(1)} = (\hat{\theta}_0, 0, \cdots, 0) \tag{35}$$

We then fix θ_0 to $\hat{\theta}_0$ and vary the connections in the first layer of hidden units θ_1. To this effect we need a sample from the equilibrium distribution of ξ, H_1. This can be efficiently obtained as follows. For each sample $v(i)$ we obtain a random of hidden units vector $h(i)$ with probability $\propto$

$$e^{\phi_1(v^{(i)}, h_1, \theta_1)} \tag{36}$$

This result on a sample

$$\{(v^{(1)}, h_1^{(1)}), (v^{(2)}, h_1^{(2)}), \cdots, (v^{(n)}, h_1^{(n)})\} \tag{37}$$

This becomes the target distribution which is treated as if it were fully observable. Training results on a new of $\hat{\theta}_1$ of θ_1 such that

$$J(\theta^{(1)}, \theta^{(2)}) < J(\theta^{(1)}, \theta^{(1)}) \tag{38}$$

where

$$\theta^{(2)} = (\hat{\theta}_0, \hat{\theta}_1, 0, \cdots, 0) \tag{39}$$

With θ_1,θ_2, fixed to $\hat{\theta}_1$, $\hat{\theta}_2$ we then train the parameters of the second layer of hidden units. To this effect we need a sample from the equilibrium distribution of V, H_1, H_2. This can be obtained efficiently as follows: For each sample $(v^{(i)}, h_1^{(i)})$ we collect a sample ${h_2}^{(i)}$ with probability $\propto$

$$e^{\phi_2(v^{(i)}, h_1^{(i)}, h_2, \theta_1)} \tag{40}$$

We treat this as if it were an observable sample and use it to find $\hat{\theta}_2$ such that

$$J(\theta^{(2)}, \theta^{(3)}) < J(\theta^{(2)}, \theta^{(2)}) \tag{41}$$

where

$$\theta^{(3)} = (\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, 0, \cdots, 0) \tag{42}$$

The process is iterated every time fixing the parameters of the previous layers and the samples obtained from the previous layers, and training a new layer. It can be shown (See Appendix Corollary XXX) that on each iteration the probability velocity of the observable units given the target distribution ξ decreases

1. Formulate a parameterized energy function $\phi(v,h,\theta)$ with known gradients $\Psi_v(v,h,\theta) = \nabla_v \phi(v,h,\theta)$, $\Psi_h(v,h,\theta) = \nabla_h \phi(v,h,\theta)$.
2. Choose a dispersion constant $\sigma > 0$ and time step size $\Delta_t > 0$ and a learning step size $\epsilon > 0$.
3. Choose a method to sample from the target rv ξ. data
4. Initialize network to parameter θ.
5. Choose a sample v from the target rv ξ
6. Generate a sample h from the equilibrium distribution of the hidden units given the observed v. (Depending on the problem, the equilibrium distribution may be obtained analytically or by repeated iteration of the process)

h ~ p∞(h|v)
if it is easy

$$H_{t+\Delta_t} = H_t + \Psi_h(v, H_t, \theta)\Delta_t + \sigma\sqrt{\Delta_t} Z_t \tag{43}$$

where Z_t ~ N(0,1). ~ p∞(h|v) with potential φ(h,v), v is fixed

7. Unclamp the observable units and generate a sample v', h' from

$$\begin{pmatrix} V_{\Delta_t} \\ H_{\Delta_t} \end{pmatrix} = \begin{pmatrix} v \\ h \end{pmatrix} + \Delta_t \begin{pmatrix} \Psi_v(v,h,\theta') \\ \Psi_h(v,h,\theta') \end{pmatrix} + \sigma\sqrt{\Delta_t} \begin{pmatrix} Z \\ W \end{pmatrix} \tag{44}$$

were Z, W ~ N(0,1).

8. Update θ based on Hinton's learning rule

$$\theta \leftarrow \theta + \epsilon\Big(\nabla_\theta \phi(z',h') - \nabla_\theta \varphi(z,h)\Big) \tag{45}$$

9. Go to Step 5.

Figure 2: A Minimum Probability Velocity Algorithm.

1 Non Gradient Systems

Say that directed graphs can be converted into a gradient diffusion by considering $\phi(x) = \log p(x)$ so that we must know how to get gradient wrt x and wrt θ and example of this is ICA.

Analytical Example: Learning Orstein-Ullembach Processes

Consider a single neuron network with a quadratic energy function

$$\boxed{\phi(x,\theta) = -\frac{1}{2}\theta x^2} \tag{46}$$

$$dX_t^\theta = \nabla_x \phi(X_t^\theta, \theta)dt + \sigma dB_t = -\theta X t dt + \sigma dB_t \tag{47}$$

This defines an Ornstein-Hullenbeck diffusion process with a Gaussian equilibrium distribution

$$\boxed{p_\infty^\theta(x) \propto e^{-\theta x^2/\sigma^2}} \tag{48}$$

Suppose we want for this network to learn a target rv ξ that is Gaussian with 0 mean and variance σ_ξ^2

We show that indeed the different approaches reviewed in the paper: score matching, exact CD, standard CD, are up to a proportionality constant,identical versions of minimum velocity.

Minimum Velocity/Score Matching

The objective function in minimum velocity

ESM

$$E[\|V_0^\theta(\xi)\|^2] = E[\|\nabla_x \phi(\xi,\theta) - \frac{\sigma^2}{2}\nabla_x \log p(\xi)\|^2] \tag{49}$$

And considering

$$\nabla_x \phi(x,\theta) = -\theta x \tag{50}$$

$$\nabla_x \log p_\xi(x) = -\frac{x}{\sigma_\xi^2} \tag{51}$$

==>

$$E[\|V_0^\theta(\xi)\|^2] = E[(\theta\xi - \frac{\sigma^2}{2}\frac{\xi}{\sigma_\xi^2})^2] = \sigma_\xi^2\Big(\theta - \frac{\sigma^2}{2\sigma_\xi^2}\Big)^2 \tag{52}$$

$$\nabla_\theta E[\|V_0^\theta(\xi)\|^2] = 2\sigma_\xi^2\Big(\theta - \frac{\sigma^2}{2\sigma_\xi^2}\Big) = 2\sigma_\xi^2\theta - \sigma_2 \tag{53}$$

Thus GD on the probability velocity would converge to

$$\hat\theta = \frac{\sigma^2}{2\sigma_\xi^2} \tag{54}$$

and at equilibrium

$$p_\infty(x) \propto \exp\left(\frac{2\phi(x,\theta)}{\sigma^2}\right) = \exp\left(-\frac{\hat\theta x^2}{\sigma^2}\right) = \exp\left(-\frac{x^2}{2\sigma_\xi^2}\right) \tag{55}$$

Indicating that the desired equilibrium distribution has been learned.

The gradient of the velocity can also be computed as follows, which has the advantage of not requiring the gradient of p_ξ

ISM

$$\nabla_\theta E[\|V_0(\xi)\|^2] = \nabla_\theta\Big(E[\|\nabla_x \phi(\xi)\|^2] + \sigma^2 E[\nabla^2 \cdot \phi(\xi)]\Big). \tag{56}$$

In our case

$$\nabla_x \phi(x,\theta) = -\theta x \tag{57}$$

$$\nabla_x^2 \phi(x,\theta) = -\theta \tag{58}$$

and,

$$\nabla_\theta E[\|V_0(\xi)\|^2] = \nabla_\theta(\theta^2 E[\xi^2] - \sigma^2\theta) = 2\theta{\sigma_\xi}^2 - \sigma^2 \tag{59}$$

Exact Contrastive Divergence: As shown before the instanteneous contrastive divergence== the time derivative of the free energy

$$-F(X_t^\theta) = E[\phi(X_t^\theta,\theta)] + \frac{\sigma^2}{2}H(X_t^\theta) \tag{60}$$

First we will compute the temporal derivative of the entropy. First note $X_0^{\theta} = \xi$ thus, considering that $\xi \sim$ Gaussian 0 mean and variance σ^2

$$H(X_0^\theta) = -E[\log p(\xi)] = \frac{1}{2}(1 + \log 2\pi\sigma_\xi^2) \tag{61}$$

Now note to first order

$$\boxed{\begin{aligned} X_{\Delta_t}^\theta &= X_0^\theta + \Delta_t \nabla_x^{\overset{\varphi}{\theta}}(X_0^\theta) + \sigma^2\sqrt{\Delta_t} Z_t \\ &= \xi - \theta\xi\Delta_t + \sigma\sqrt{\Delta_t} Z_t \\ &= \xi(1-\theta\Delta_t) + \sigma\sqrt{\Delta_t} Z_t \end{aligned}} \tag{62}$$

where Z_t~N(0,1). Thus, to first order, the variance at time Δ_t is

$$E[(X_{\Delta_t}^\theta)^2] = \sigma_\xi^2(1-\theta\Delta_t)^2 + \sigma^2\Delta_t \tag{63}$$

and since, ${X_{\Delta}}_t^{\theta}$ is Gaussian, its entropy is

$$H(X_{\Delta_t}^\theta) = -E[\log p(X_{\Delta_t}^\theta)] = \frac{1}{2}(1 + \log 2\pi(\sigma_\xi^2(1-\theta\Delta_t)^2 + \sigma^2\Delta_t)) \tag{64}$$

From which the temporal derivative of the entropy follows

$$\begin{aligned} \boxed{\left.\frac{dH(X_t^\theta)}{dt}\right|_{t=0}} &= \lim_{\Delta_t \to 0} \frac{H(X_{\Delta_t}^\theta) - H(X_0^\theta)}{\Delta_t} \\ &= \lim_{\Delta_t \to 0} \frac{1}{2} \frac{\log(\sigma_\xi^2(1-\theta\Delta_t)^2 + \sigma^2\Delta_t) - \log\sigma_\xi^2}{\Delta_t} \\ &= \lim_{\Delta_t \to 0} \frac{1}{2\Delta_t}\Big(\log((1-\theta\Delta_t)^2 + \frac{\sigma^2}{\sigma_\xi^2}\Delta_t\Big) = \frac{\sigma^2}{2\sigma_\xi^2} - \theta \end{aligned} \tag{65}$$

Second we compute the temporal derivative of the expected potential

$$E[\phi(X_0^\theta, \theta)] = E[\phi(\xi, \theta)] = -\frac{1}{2}\theta\sigma_\xi^2 \tag{66}$$

$$E[\phi(X_{\Delta_t}^\theta, \theta)] = -\frac{\theta}{2}E[(X_{\Delta-t}^\theta)^2] = -\frac{\theta}{2}\Big(\sigma_\xi^2(1-\theta\Delta_t)^2 + \sigma^2\Delta_t\Big) \tag{67}$$

Thus

$$\boxed{\left.\frac{dE[\phi(X_t^\theta, \theta)}{dt}\right|_{t=0}} = \lim_{\Delta_t \to 0} \frac{E[\phi(X_{\Delta_t}^\theta, \theta)] - E[\phi(X_0^\theta, \theta)]}{\Delta_t} = \theta^2\sigma_\xi^2 - \theta\frac{\sigma^2}{2} \tag{68}$$

and

$$\begin{aligned} \frac{\sigma^2}{2}\left.\frac{dF(X_t^\theta, \theta)}{dt}\right|_{t=0} &= -\frac{\sigma^2}{2}\left.\frac{dE[\phi(X_t^\theta, \theta)}{dt}\right|_{t=0} - \left.\frac{dH(X_t^\theta)}{dt}\right|_{t=0} \\ &= \frac{2}{\sigma^2}\Big(\theta\frac{\sigma^2}{2} - \theta^2\sigma_\xi^2\Big) - \frac{\sigma^2}{2\sigma_\xi^2} + \theta = 2\theta - 2\theta^2\frac{\sigma_\xi^2}{\sigma^2} - \frac{\sigma^2}{2\sigma_\xi^2} \end{aligned} \tag{69}$$

Taking the gradient

$$\nabla_\theta \frac{\sigma^2}{2}\left.\frac{dF(X_t^\theta, \theta)}{dt}\right|_{t=0} = 2 - 4\theta\frac{\sigma_\xi^2}{\sigma^2} \tag{70}$$

Thus GD on the free energy velocity which equals GD on the contrastive divergence would converge to

$$\hat\theta = \frac{\sigma^2}{2\sigma_\xi^2} \tag{71}$$

Standard Contrastive Divergence

Given a fixed parameter θ' we have to first order

$$X^{\theta'}_{\Delta_t} = \xi - \theta'\xi\Delta_t + \sigma\sqrt{\Delta_t}Z_t = \xi(1-\theta'\Delta_t) + \sigma\sqrt{\Delta_t}Z_t \tag{72}$$

Thus

$$\phi(X^{\theta'}_{\Delta_t}, \theta) = -\frac{\theta}{2}(X^{\theta'}_{\Delta_t})^2 \tag{73}$$

$$E[\phi(X^{\theta'}_{\Delta_t}, \theta)] = -\frac{\theta}{2}E[(X^{\theta'}_{\Delta_t})^2] = -\frac{\theta}{2}\sigma^2_\xi(1-\theta'\Delta_t)^2 + \sigma^2\Delta_t \tag{74}$$

and

$$\lim_{\Delta_t \to 0} \frac{E[\phi(X^{\theta'}_{\Delta_t}, \theta)] - E[\phi(X^{\theta'}_0, \theta)]}{\Delta_t} = -\frac{\theta}{2}(\sigma^2 - 2\theta'\sigma^2_\sigma) \tag{75}$$

Taking the gradient wrt θ

$$\boxed{\nabla_\theta \lim_{\Delta_t \to 0} \frac{E[\phi(X^{\theta'}_{\Delta_t}, \theta)] - E[\phi(X^{\theta'}_0, \theta)]}{\Delta_t}\Bigg|_{\theta=\theta'}} = \frac{1}{2}(\sigma^2 - 2\theta\sigma^2_\sigma) \tag{76}$$

Thus GD converges to the minimum velocity solution

$$\hat{\theta} = \frac{\sigma^2}{2\sigma^2_\xi} \tag{77}$$

Previous Work

The original *Boltzmann machine* paper (Ackley et al., 1985) and *Harmony Theory* paper (Smolensky , 1986), introduced the notion of training stochastic networks to exhibit desired equilibrium distributions The application of this idea to diffusion networks was presented in Movellan and McClelland (1993), Movellan (1998), and Movellan et al. (2002). The idea of minimizing contrastive divergence first appeared in Hinton (2002) where an algorithm was presented to approximate the gradient of the contrastive divergence. Score matching was developed by Hyvärinen (2005) as a method for training unnormalized probability models . The relationship between score matching and standard contrastive divergence was first pointed out by Hyvärinen (2006). Here we show that in continuous time systems as $t \to 0$ exact contrastive divergence, standard contrastive divergence and score matching become identical We also present the connection between minimization of probability velocity fields, contrastive divergence, and score matching .

Summary of Results

We introduced the idea of minimum velocity as a general principle for training equilibrium distributions We showed that:

- The probability velocity field of stochastic diffusions can be minimized using the score matching algorithm (Hyvärinen, 2005), providing a novel interpretation for that algorithm and a practical way to train diffusion networks.
- The rate of change of the free energy is proportional to the norm of the velocity field.
- Exact and stnadard CD are proportional to the derivative of the temporal derivative of the free energy.
- Exact CD, standard CD, score matching, minimum free energy velocity, and minimum probability velocity are equivalent in the limit as t → 0.

Appendix

Unless otherwise stated, capital letters are used for rvs, small letters for specific values taken by rvs, and Greek letters for fixed parameters. The exceptions are H, and F, that stand for the entropy and free energy of rvs. We leave implicit the properties of the probability space $(\Omega, \mathcal{F}, P)$ in which the rvs are defined.

The symbol $\nabla_x \cdot f(x)$ is the divergence of f, i.e.,

$$\nabla_x \cdot f(x) = \sum_i \frac{\partial f(x)}{\partial x_i} \tag{79}$$

The symbol $\nabla_x^2 \cdot f(x)$ is the Laplacian of f, i.e.,

$$\nabla_x^2 \cdot f(x) = \sum_i \frac{\partial^2 f(x)}{\partial x_i^2} \tag{80}$$

If X is a rv the term $\frac{\partial f(X)}{\partial x_i}$ is short notation for a rv Y such that for each outcome $\omega \in \Omega$

$$Y(\omega) \stackrel{\text{def}}{=} h(X(\omega)) \quad h(x) = \frac{\partial f(x)}{\partial x_i} \tag{81}$$

For simplicity in the main body of the paper we suppress the dependencies on the network parameter θ. In the Appendix we need to make this dependency explicit. Thus here we work with a collection of random processes X^θ parameterized by the network parameter θ. Each process is a solution to the following SDE

$$dX_t^\theta = \nabla_x \phi(X_t^\theta, \theta) dt + \sigma dB_t \tag{83}$$

$$X_0 = \xi \tag{84}$$

Where ξ is a target rv whose distribution we want to match. We use the following shorthand notation

$$p_t^\theta(x) \stackrel{\text{def}}{=} p_{X_t^\theta}(x) \tag{85}$$

$$X_\infty^\theta \stackrel{\text{def}}{=} \lim_{t \to \infty} X_t^\theta \tag{86}$$

$$p_\infty^\theta(x) \stackrel{\text{def}}{=} \lim_{t \to \infty} p_{X_t^\theta}(x) \tag{87}$$

The goal of learning is to find values of θ such that the distribution of X_∞^θ approximates as best as possible the distribution of ξ, i.e. $p_\xi(x) \approx p_\infty^\theta(\xi)$

Lemma 1 (**Botlzmann Equilibrium**). *The Boltzmann equilibrium distribution* $p(x) \propto \exp(-2\phi(x)/\sigma^2)$ *minimizes the free energy,*

$$F(p) = -\int p(x)\phi(x)dx - \frac{\sigma^2}{2}\int p(x)\log p(x)dx. \tag{88}$$

~ H(p, p_φ)

Lemma 2 (**Ito's Lemma**). *Let the n-dimensional process X satisfy the following stochastic differential equation*

$$dX_t = \mu(X_t)dt + \sigma dB_t \tag{91}$$

where $\sigma > 0$. *Let* $f : \Re^n \times \Re \to \Re$. *Then the process* $f(X_t, t)$ *satisfies the following SDE*

$$df(X_t, t) = \frac{\partial f(X_t, t)}{\partial t}dt + \nabla_x f(X_t) \cdot \mu(X_t)dt + \sigma\nabla_x f(X_t) \cdot dB_t + \frac{\sigma^2}{2}\nabla^2 \cdot f(X_t)dt \tag{92}$$

d/dt Ef(Xt) = E D(f . μ) + σ2/2 △ Ef

□

Lemma 3 (**Score Matching**). *Let p be the pdf of the rv X and q a pdf, such that*

$$\lim_{u\to-\infty} E\Big[\frac{\partial \log q(X)}{\cancel{\partial x_i}} \mid X_i = u\Big] = \lim_{u\to\infty} E\Big[\frac{\partial \log q(X)}{\cancel{\partial x_i}} \mid X_i = u\Big] = 0 \tag{93}$$

for $i = 1 \cdots n$. *Then*

$$\boxed{\begin{aligned} \frac{1}{2}E\|\nabla_x \log p(X) - \nabla_x \log q(X)\|^2 = {} & \frac{1}{2}E\|\nabla_x \log q(X)\|^2 \\ & + E\,{\nabla_x}^2 \cdot \log q(X) + \frac{1}{2}E\|\nabla_x \log p(X)\|^2. \end{aligned}} \tag{94}$$

□

Corollary 1 (SM).

$$\begin{aligned} \frac{1}{2}E[\|\nabla_x \log \overset{q}{p^\theta_\infty}(X^\theta_t) - \nabla_x \log \overset{p}{p^\theta_t}(X^\theta_t)\|^2] &= \frac{1}{2}E\|\nabla_x \log {p_t}^\theta({X_t}^\theta)\|^2 \\ &+ \frac{1}{2}E\|\nabla_x \ \frac{2}{\sigma^2}\phi({X_t}^\theta, \theta)\|^2 + \frac{2}{\sigma^2}E[\nabla^2_x \cdot \phi(X^\theta_t, \theta)]. \end{aligned} \tag{101}$$

Corollary 2 (**Probability Velocity**).

$$\begin{aligned} \frac{1}{2}E[\|V^\theta_t(X^\theta_t)\|^2 &= \frac{1}{2}E[\|\frac{\sigma^2}{2}\nabla_x \log p^\theta_t(X^\theta_t)\|^2] \\ &+ \frac{1}{2}E[\|\nabla_x\phi(X^\theta_t, \theta)\|^2] + \frac{\sigma^2}{2}E[\nabla^2_x \cdot \phi(X^\theta_t, \theta)]. \end{aligned} \tag{102}$$

where

$$V^\theta_t(x) \overset{def}{=} \nabla_x\Big(\overset{q}{\phi(x, \theta)} - \frac{\sigma^2}{2}\log \overset{p}{p^\theta_t}(x)\Big) \tag{103}$$

is the probability velocity at state x, and time t, on a network with parameter θ.

Theorem 1 (**Velocity of the Average Potential**).

$$\frac{dE[\phi(X_t^\theta,\theta)]}{dt} = \frac{1}{2}\Big(E[\|V_t^\theta(X_t^\theta)\|^2] + E[\|\nabla_x\phi(X_t^\theta,\theta)\|^2] - E[\|\nabla_x\frac{\sigma^2}{2}\log p_t^\theta(X_t^\theta)\|^2]\Big). \tag{105}$$

$$= E[\|\nabla_x\phi(X_t^\theta,\theta)\|^2] + \frac{\sigma^{\ 2}}{2}E[\nabla_x^2\cdot\phi(X_t^\theta,\theta)].$$

□

Theorem 2 (**Velocity of the Entropy**). *Let $\phi, {p_t}^\theta$ such that*

$$\lim_{t\to-\infty}\phi(x,\theta)\frac{\partial}{\partial x_i}p_t^\theta(x) = \lim_{t\to\infty}\phi(x,\theta)\frac{\partial}{\partial x_i}p_t^\theta(x) = 0 \tag{108}$$

for $i = 1,\cdots,n$. Then

$$\frac{dH(X_t^\theta)}{dt} = E[\nabla_x^2\cdot\phi(X_t^\theta,\theta)] - \frac{\sigma^2}{2}E[\nabla_x^2\cdot\log p_t^\theta(X_t^\theta)], \tag{109}$$

where $H(X_t^\theta)$ is the entropy of X_t^θ.

Proof. Applying Ito's Lemma (Lemma 2) with $X_t \equiv {X_t}^\theta$, $f(x,t) = \log {p_t}^\theta(x)$ and $\mu(x) = \nabla\phi(x,\theta)$ ==>

$$d\log p_t^\theta(X_t^\theta) = \left(\frac{\partial}{\partial t}\log p_t^\theta(X_t^\theta)\right)dt + \boxed{\sum_i\frac{\partial}{\partial x_i}\phi(X_t^\theta,\theta)\frac{\partial}{\partial x_i}\log p_t^\theta(X_t^\theta)dt}$$

$$+\sum_i\frac{\partial}{\partial x_i}\phi(X_t^\theta,\theta)\frac{\partial}{\partial x_i}\log p_t^\theta(X_t^\theta)dB_{i,t} + \frac{\sigma^2}{2}\nabla_x^2\cdot\log p_t^\theta(X_t^\theta)dt. \tag{110}$$

==>

$$\frac{dH(X_t^\theta)}{dt} \overset{\text{def}}{=} -\frac{dE[\log p_t^\theta(X_t^\theta)]}{dt} = \boxed{-\sum_i E[\frac{\partial}{\partial x_i}\phi(X_t^\theta,\theta)\frac{\partial}{\partial x_i}\log p_t^\theta(X_t^\theta)]}$$

$$-\frac{\sigma^2}{2}E[\nabla_x^2\cdot\log p_t^\theta(X_t^\theta)]. \tag{111}$$

used the *fact*:

$$E[\frac{\partial}{\partial t}\log p_t^\theta(X_t^\theta)] \overset{\text{def}}{=} \int p_t^\theta(x)\frac{\partial}{\partial t}\log p_t^\theta(x)dx = 0 \tag{112}$$

Integrating by parts on the first term in the RHS of (111)

$$E[\frac{\partial}{\partial x_i}\phi(X_t^\theta,\theta)\frac{\partial}{\partial x_i}\log p(X_t^\theta)] = \int p_t^\theta(x)\frac{\partial}{\partial x_i}\phi(x,\theta)\frac{\partial}{\partial x_i}\log p_t^\theta(x)dx$$

$$= \int\frac{\partial}{\partial x_i}\phi(x,\theta)\frac{\partial}{\partial x_i}p_t^\theta(x)dx = \left[\phi(x,\theta)\frac{\partial}{\partial x_i}p_t^\theta(x)\right]_{x=-\infty}^{x=\infty} - \int p_t^\theta(x)\frac{\partial^2}{\partial x_i^2}\phi(x,\theta)dx$$

$$= -E[\frac{\partial^2}{\partial x_i^2}\phi(X_t^\phi,\theta)] \tag{113}$$

□

Velocity of F : Velocity of average potential + Velocity of entropy

Theorem 3 (**Velocity of the Free Energy**).

$$\frac{dF_t^\theta(X_t^\theta)}{dt} = -E[\|V_t^\theta(X_t^\theta)\|^2], \tag{114}$$

where F_t^θ is the free energy of X_t^θ.

Proof. Using Theorems 1 and 2

$$-\frac{dF_t^\theta(X_t^\theta)}{dt} = \frac{dE\phi(X_t^\theta,\theta)}{dt} + \frac{\sigma^2}{2}\frac{dH_t^\theta(X_t^\theta)}{dt} = \frac{1}{2}E[\|V_t^\theta(X_t^\theta)\|^2] + \frac{1}{2}E[\|\nabla_x\phi(X_t^\theta,\theta)\|^2]$$
$$-\frac{1}{2}E[\|\nabla_x\frac{\sigma^2}{2}\log p_t^\theta(X_t^\theta)\|^2] + \frac{\sigma^2}{2}E[\nabla_x^2\cdot\phi(X_t^\theta)] - \frac{\sigma^4}{4}E[\nabla_x^2\cdot\log p_t^\theta(X_t^\theta)]. \tag{115}$$

Using Corollary 2 ==>

$$\frac{1}{2}E[\|\nabla_x\phi(X_t^\theta,\theta)\|^2] + \frac{\sigma^2}{2}E[\nabla_x^2\cdot\phi(X_t^\theta,\theta)] = \frac{1}{2}E[\|V_t^\theta(X_t^\theta)\|^2] - \frac{1}{2}\frac{\sigma^4}{4}E[\|\nabla_x\log p_t^\theta(X_t^\theta)\|^2]. \tag{116}$$

Applying Lemma 3 with $X = X_t^\theta$, and $p(x) = q(x) = p_t^\theta(x)$ we get

$$-\frac{\sigma^4}{4}\left(E[\nabla_x^2\cdot\log p_t^\theta(X_t^\theta)] + \frac{1}{2}E[\|\nabla_x\log p_t^\theta(X_t^\theta)\|^2]\right) = \frac{\sigma^4}{4}\frac{1}{2}E[\|\nabla_x\log p_t^\theta(X_t^\theta)\|^2]. \tag{117}$$

□

Theorem 4 (**Exact Contrastive Divergence**). *Exact Contrastive Divergence is a Minimum Velocity Algorithm.*

Proof. Contrastive divergence == contrastive differential free energy == velocity

□

Theorem 5 (**Standard Contrastive Divergence**). *As $t \to 0$ standard contrastive divergence converges to exact contrastive divergence.*

Given a fixed network parameter θ'

$$\nabla_\theta \left.\frac{dE[\phi(X_t^{\theta'},\theta)]}{dt}\right|_{t=0,\theta=\theta'} = \nabla_\theta\frac{1}{2}E[\|V_0^\theta(\xi)\|^2] \tag{118}$$

Proof. Applying the Ito rule

$$d\phi(X_t^{\theta'},\theta) = \nabla_x\phi(X_t^{\theta'},\theta)\cdot\nabla_x\phi(X_t^{\theta'},\theta')dt$$
$$+\frac{\sigma^2}{2}\nabla_x^2\cdot\phi(X_t^{\theta'},\theta)dt + \sum_i\frac{\partial\theta(X_t^{\theta'})}{\partial x_i}dB_{i,t} \tag{119}$$

==>

$$\frac{dE[\phi(X_t^{\theta'},\theta)]}{dt} = E[\nabla_x\phi(X_t^{\theta'},\theta')\cdot\nabla_x\phi(X_t^{\theta'},\theta)] + \frac{\sigma^2}{2}E[\nabla_x^2\cdot\phi(X_t^{\theta'},\theta)] \tag{120}$$

Thus

$$\frac{\partial}{\partial\theta_i}\left.\frac{dE[\phi(X_t^{\theta'},\theta)]}{dt}\right|_{t=0,\theta=\theta'} = E[\nabla_x\phi(\xi,\theta)\cdot\frac{\partial}{\partial\theta_i}\nabla_x\phi(\xi,\theta)]$$
$$+\frac{\sigma^2}{2}\frac{\partial}{\partial\theta_i}E[\nabla_x^2\cdot\phi(\xi,\theta)]$$
$$=\frac{1}{2}\frac{\partial}{\partial\theta_i}E[\|\nabla_x\phi(\xi,\theta)\|^2]+\frac{\sigma^2}{2}\frac{\partial}{\partial\theta_i}E[\nabla_x^2\cdot\phi(\xi,\theta)] \quad (122)$$

Thus, using Corollary 2 with $t=0$, and considering $X_t^\theta(\theta)=\xi$

$$\nabla_\theta\left.\frac{dE[\phi(X_t^{\theta'},\theta)]}{dt}\right|_{t=0,\theta=\theta'} = \nabla_\theta\Big(\frac{1}{2}E[\|\nabla_x\phi(\xi,\theta)\|^2]+\frac{\sigma^2}{2}E[\nabla_x^2\cdot\phi(\xi,\theta)]\Big)$$
$$=\frac{1}{2}\nabla_\theta E[\|V_0^\theta(\xi)\|^2]. \quad (123)$$

□

References

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(2):147–169, 1985.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8), 2002.

A. Hyvärinen. Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

A. Hyvärinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *Technical Report: http://www.cs.helsinki.fi/u/ahyvarin/papers/unsuper.shtml*, 2006.

J. R. Movellan. A learning theorem for networks at detailed stochastic equilibrium. *Neural Computation*, 10(5):1157–1178, 1998.

J. R. Movellan. Tutorial on stochastic differential equations. *MPLab Tutorials. http://mplab.ucsd.edu*, 2006.

J. R. Movellan and J. L. McClelland. Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science*, 17:463–496, 1993.

J. R. Movellan, P. Mineiro, and R. J. Williams. A Monte-Carlo EM approach for partially observable diffusion processes: Theory and applications to neural networks. *Neural Computation*, 14(7):1507–1544, 2002.

B. Oksendal. *Stochastic Differential Equations*. Springer Verlag, Berlin, 1992.

P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, chapter 6, pages 194–281. MIT Press, Cambridge, MA, 1986.