論文97-34C-7-9

Fokker-Plank 방정식의 해석을 통한 Langevine 경쟁학습의 동역학 분석

(Analysis of the Fokker-Plank Equation for the Dynamics of Langevine Cometitive Learning Neural Network)

石 鎭 旭 *, 趙 成 元 *

(Jinwuk Seok and Seongwon Cho)

요 약

본 논문에서는 Langevine 경쟁학습 신경회로망의 Fokker-Plank방정식에 바탕을 둔 동역학 분석을 논한다. 확률미분방정식적 관심에서 Langevine 경쟁학습의 학습알고리즘은 Langevine 확률미분방정식의 한 종류이며 확률측도를 가진 위상공간 $(\Omega, \mathcal{F}, P)$ 위의 확산 방정식을 가지고 있다. 제안한 알고리즘에서 먼저 학습방정식의 분석을 통해 학습방정식의 Fokker-Plank 방정식을 유도하며 Markov 준군위의 극소 작용소를 도입하여 특정 단체(simplex)내의 가중치 벡터가 목적함수의 볼록, 혹은 의사 볼록(pseudo-convex) 조건하에서 대역 최소점을 찾을 수 있음을 증명한다. 원격탐사 데이터에 대한 패턴 인식 실험을 통하여 기존의 경쟁학습 신경회로망보다 제안한 Langevine 경쟁학습 신경회로망의 성능이 우수함과 본 논문에서 제기한 대역 기능이 가능함을 증명하였다.

Abstract

In this paper, we analyze the dynamics of Langevine competitive learning neural network based on its Fokker-Plank equation. From the viewpoint of the stochastic differential equation (SDE), Langevine competitive learning equation is one of Langevine stochastic differential equation and has the diffusion equation on the topological space $(\Omega, \mathcal{F}, P)$ with probability measure. We derive the Fokker-Plank equation from the proposed algorithm and prove by introducing a infinitestimal operator for Markov semigroups, that the weight vector in the particular simplex can converge to the globally optimal point under the condition of some convex or pseudo-convex performance measure function. Experimental results for pattern recognition of the remote sensing data indicate the superiority of Langevine competitive learning neural network in comparison to the conventional competitive learning neural network.

I. 서 론

현재까지 신경회로망에 대한 연구는 대부분 특정 적용대상에 대한 직관적인 신경회로망 모델 개발에 중점이 맞추어져 이루어져 왔다. 따라서 범용적으로 쓰이는 몇몇 모델(오류 역전파 모델이나 자기 조직화 형상지도 모델 혹은 홉필드, 유전 알고리즘 등)을 제외하고 신경회로망 모델들의 안정성이나 수렴성 혹은 동역학적 연구는 거의 이루어지지 않은 형편이다. 그러므로 각 신경회로망 모델에 대한 동역학(Dynamics)을 파악하여 각 알고리즘이 어떠한 경우에 안정적인 동작을 하는가에 대한 정보와 각 신경회로망 모델들이 어떠한 경우 국소 수렴성을 피할 수 없는 가를 파악하는 것은 공학적으로 매우 중요한 문제라 하겠다.

* 正會員, 弘益大學校 電子·電氣 工學部

(School of Electronic and Electrical Engineering, Hong Ik University)

接受日字:1996年7月22日, 수정완료일:1997年7月4日

본 연구에서는 기존의 경쟁학습 신경회로망보다 실험적으로 학습성능이 우수한 것으로 알려지고 하드웨어 구현에 편리한 Langevine 경쟁학습 신경회로망의 확산성과 대역 최소화 성질을 분석한다[8]. 확률 미분 방정식(SDE)의 관점에서 볼 때, Langevine 경쟁학습 신경회로망은 두가지 방법으로 분석 가능하다. 첫 번째 방법은 학습 방정식의 확률적 요소를 추출해 내어 학습 방정식의 Fokker-Plank 방정식을 유도, 분석하는 방법이며[3], 또 하나의 방법은 학습 방정식을 미분 다양체위의 상 미분 방정식의 해석 기법들을 사용하여 분석하는 방법이다.

첫 번째 방법은 Transition확률이 준군(Semi-group)을 이루는 특징을 사용, 시간과 위치에 관한 2차 편미분 방정식인 Fokker-Plank 방정식을 유도, 선형 준군동형사상(Linear Semigroup Operator)에 대한 해석을 통해 Weight Vector의 확산성, 안정성, 수렴성 등의 동역학적 특징을 분석하는 방법이며 두 번째 방법은 Langevine 경쟁학습 신경회로망 학습 방정식에 확률적 요소(이진 강화함수)가 있음을 통해 학습 방정식을 확률 미분 방정식으로 놓고 일반적인 상 미분 방정식의 해석 방법을 사용하여 동역학을 파악하는 방법이 있다[2] [4].

본 논문에서는 먼저 Langevine 경쟁학습 신경회로망의 학습 알고리즘이 기존의 경쟁학습 알고리즘에 Gaussian 분포를 가진 Browinian Motion Process가 결합된 것임을 밝히고 이를 통해 Langevine 경쟁학습 신경회로망의 학습 알고리즘의 Fokker-Plank 방정식을 유도한다. 유도된 Fokker-Plank 방정식은 Markov 준군위에 정의된 2차 편 미분 방정식이므로 선형 준군동형사상을 도입하여 Transition 확률의 축소(Contraction)가 어떤 조건 위에 성립하는가를 밝혀 대역 최소성의 조건을 밝힌다. 마지막으로 세가지 종류의 원격탐사 데이터에 대한 패턴인식실험을 통해 단위 단체(Simplex)내에 기존의 확률적 최적화 알고리즘으로 구성된 경쟁학습 (Simple Competitive Learning : SCL) 신경회로망과 성능 비교를 통해 Langevine 경쟁학습 신경회로망의 우수성을 확인하였다.

II. Langvine 경쟁학습 신경회로망 알고리즘

Langevine 경쟁학습 신경회로망의 알고리즘은 다음과 같다[1].

step 0 : Weight vector초기화

step 1 : Input vector x^p에 대하여 다음 조건을 만족하는 Weight vector index r을 선택

$$r = \arg_r \min_{w_r(t)} d(v(t), w_r(t))$$

step 2 : Input vector x^p과 Weight vector index r 에 대하여 Directional Derivation h_t 계산

$F(t)$: = Final Epoch - t 사이의 임의의 값을 취하는 확률변수

$P(t)$: = 0에서 t 사이의 임의의 값을 취하는 확률변수

$$\delta(t) = 1 \quad \textit{with probability } \frac{1}{2} P_\delta(F(t) > P(t))$$

$$\eta(\lambda, \delta, t, v_t, w_t^r) = \begin{cases} \delta(t) = 1 & -\lambda\,\delta(t)\,sgn(v(t) - w_r(t)) \\ \delta(t) = 0 & 0 \end{cases}$$

$$h_t = \varepsilon_L \nabla J_t + \eta(\lambda, \delta, t, v_t, w_t^r) = -\varepsilon_L (v(t) - w_r(t)) - \lambda\,\delta(t)\,sgn(v(t) - w_r(t))$$

step 3 : Weight vector w^r_t 을 다음 식에 의해 갱신

$$w_r(t+1) = w_r(t) + \varepsilon_L (v(t) - w_r(t)) + \delta\rho(t)\,sgn(v(t) - w_r(t))$$

step 4 : t를 t+1로 변환

step 5 : 한계 Epoch (Final Epoch)까지 진행되었으면 끝. 그렇지 않으면 Step 1.

여기에서 w^r_t는 시간 t일 때 r번째 Weight vector를 의미하며 ε_L은 시간에 대하여 불변인 일정적응 이득, v_t는 시간 t일 때 입력 Vector, 그리고 $\eta(\lambda, \delta, t, v_t, w^r_t)$는 이진 강화함수 항으로서 다음과 같이 구성된다.

$$\eta(\lambda, \delta, t, v_t, w^r_t) = -\lambda\delta(t)sgn(v_t - w^r_t) \tag{1}$$

식 (1)에서 λ는 이진 강화함수의 이득이며, $\delta(t)$는 1 혹은 0을 반환하는 Random Variable로서 본 논문에서는 Epoch에 종속인 P_δ의 확률로 1을 반환하며, $sgn(v_t - w^r_t)$은 다음을 만족하는 함수이다.

$$sgn(v_t - w^r_t) = \begin{cases} 1 & v_t - w^r_t \geq 0 \\ -1 & v_t - w^r_t < 0 \end{cases} \tag{2}$$

이때 제안한 경쟁학습 신경회로망의 성능지표 J_t는 다음과 같다[5].

$$J_t = \frac{1}{2} \int_{B_t} \sum_r^{M} \| v_t - w^r_t \|^2 dP(v_t) \tag{3}$$

한편 시간 t에서의 성능지표 J_t에 대한 Directional Derivation h_t를 J_t의 Gradient로 놓으면

$$h_t = -\nabla J_t = (v_t - w^r_t)$$

따라서 Directional Derivation h_t의 Theorem Descent Direction $h^D{}_t$를 Theorem 1과 같이 놓으면 $h^D{}_t$는 J_t를 최소화 하는 방향으로 $w^r{}_t$를 보낸다.

Theorem 1 : Descent Direction $h^D{}_t$가 다음과 같이 정의되면

$$h^D{}_t = \varepsilon_L h_t + \lambda\delta(t) sgn(v_t - w^r{}_t) \qquad \varepsilon_L > 0$$

학습 방정식 $w^r{}_{t+1} = w^r{}_t + h^D{}_t$는 J_t를 최소화 하는 방향으로 $w^r{}_t$를 보낸다.

proof :

이진 강화 함수와 J_t의 Weight vector $w^r{}_t$에 대한 Gradient ∇J_t와 $h^D{}_t$의 스칼라 적이 0보다 작다는 것과 위 명제는 동치이다. 따라서 이진 강화함수와 ∇J_t의 스칼라 적을 살펴보면 이진 강화함수의 정의에 의해 이진 강화함수 n차원 각 Component들은 크기가 1 혹은 0이며 부호가 ∇J_t의 대응되는 Component와 반대이므로

$$-n \le \langle \nabla J_t, \lambda\delta(t) sgn(v_t - w^r{}_t) \rangle \le 0$$

그러므로

$$\langle \nabla J_t, \varepsilon_L h_t + \lambda\delta(t) sgn(v_t - w^r{}_t) \rangle = -\varepsilon_L \|\nabla J_t\|^2 + \langle \nabla J_t, \lambda\delta(t) sgn(v_t - w^r{}_t) \rangle < 0 \quad \blacksquare$$

Ⅲ. Gaussian Process로서의 이진 강화함수

앞 절 Theorem 1 에서 살펴보았듯이 이진 강화함수는 기존의 경쟁학습이 Weight vector $w^r{}_t$를 ∇J_t의 반대 방향으로 갱신하는데 비하여 이진 강화함수가 결합된 경우 갱신거리가 보다 크고 방향이 ∇J_t방향에서 약간 틀어진 방향으로 보내게 됨을 알 수 있다. 이진 강화함수는 또한 무작위로 발생하게 되므로 확률로서 학습 방정식의 동역학을 분석할 수 밖에 없는데 이를 위해 본 절에서는 먼저 이진 강화함수가 어떤 프로세스인지를 알아보도록 한다.

Assumption 1 :

ⅰ. 입력벡터 v_t는 최적 Weight vector $\widehat{w^r}$의 Neighborhood 주위에 충분히 조밀하게 분포하고 있다. i.e.

$$\forall v_t \in R^n \ \& \ \rho_t > 0 \ \ s.t. \ v_t \in B^o(\widehat{w^r}, \rho_t)$$
$$= \{ v_t \in R^n | d(v_t, \widehat{w^r}) < \rho_t \} \ \ as \ t \to \infty$$

ⅱ. 입력벡터 v_t 는 Gaussian 분포를 따른다.

ⅲ. 입력벡터 v_t는 variance $\|D\| > 0$를 가지면서 Identically, Independent Distribution이다.

Assumption 2의 첫 번째 가정은 논의의 대상을 충분히 많은 입력벡터 v_t에 대하여 제한한 것이고 두 번째 가정은 매우 많은 데이터의 경우 일반적으로 나타나는 분포가 Gaussian 분포이기 때문이다[6]. 세 번째 가정은 입력 데이터의 분포가 White Noise에 의해 교란되어 나타나는 것임을 의미한다. 이것은 첫 번째와 두 번째 가정이 성립할 때 일반적으로 나타나는 성질이기도 하다.

증명에 들어가기 앞서 제안한 알고리즘이 가지는 일반적인 성질에 따르는 가정을 두자.

Assumption 2 :

ⅰ. $v_t - w^{r_t} \ge 0$일 확률은 $t \to \infty$에서다음과 같다.

$$P\{ v_t - w^{r_t} \ge 0 \} = 0.5$$

ⅱ. $t \to \infty$에서 최적 Weight vector w^* 주변에는 평균적으로 강화함수의 영향이 없다.

$$\text{i.e.} \quad \lim_{M \to \infty} E \sum_{t=0}^{M} \eta(\lambda, \delta, t, x, w^r{}_t) < \infty$$

Assumption 1과 Assumption 2에서 다음의 정리 2가 유도된다.

Theorem 2 :

이진 강화함수항 $\eta(\lambda, \delta, t, v_t, w^r{}_t) = \lambda\delta(t) sgn(v_t - w^r{}_t)$는 최적 Weight vector $\widehat{w^r}$의 Neigh -borhood 주위에서 Independent Incremental Gaussian Process이다.

proof :

이진 강화함수항

$\eta(\lambda, \delta, t, v_t, w^r{}_t) = \lambda\delta(t) sgn(v_t - w^r{}_t)$에 의해 생성되는 확률 프로세스를 X_t라 하고 초기 시간 t_0 이후, 시간 t에서 이진 강화 함수항에 의한 변위를 $BR(t)$라 놓으면 $BR(t)$는 다음과 같다.

$$BR(t) = \lambda(X_{t_0+1} + X_{t_0+2} + \cdot \cdot \cdot \cdot + X_{t-1} + X_t)$$

임의 k번째 $sgn(v_{t_0+k} - w^r{}_{t_0+k})$의 값이 1 혹은 -1을

가질 확률은 Assumption 2에서 각각 다음과 같다.

$$sgn(v_{t_0+k} - w^r_{t_0+k}) = \begin{cases} 1 & with\ P(v_{t_0+k} - w^r_t \geq 0 \mid \delta(t_0+k)=1) = 0.5 \\ -1 & with\ P(v_{t_0+k} - w^r_t \geq 0 \mid \delta(t_0+k)=1) = 0.5 \end{cases}$$

위 식에서 $\delta(t_0+k)=1$일 확률을 P_δ라 하면 $\delta(t_0+k)=0$일 확률은 $1-P_\delta$이다. 따라서

$$P(v_{t_0+k} - w^r_t \geq 0 \mid \delta(t_0+k)=1) = \frac{P(v_{t_0+k} - w^r_t \geq 0, \delta(t_0+k)=1)}{P(\delta(t_0+k)=1)} = \frac{1}{2}$$

$$P(v_{t_0+k} - w^r_t \geq 0, \delta(t_0+k)=1) = \frac{1}{2}P(\delta(t_0+k)=1) = \frac{1}{2}P_\delta$$

마찬가지로

$$P(v_{t_0+k} - w^r_t < 0, \delta(t_0+k)=1) = \frac{1}{2}P(\delta(t_0+k)=1) = \frac{1}{2}P_\delta$$

따라서 X_{t_0+k}은 각각 다음의 확률을 가진다.

$$X_{t_0+k} = \begin{cases} 1 & with\ probability\ \frac{1}{2}p_\delta \\ -1 & with\ probability\ \frac{1}{2}p_\delta \\ 0 & with\ probability\ 1-p_\delta \end{cases}$$

Assumption 1의 iii과 Assumption 2의 i, 그리고 X_{t_0+k}의 확률에 의하여, X_{t_0+k}는 $\forall k \in Z^+$에 대하여 Independent Process이며 따라서 $BR(t)$는 Independent Increment성질을 만족한다.

이진 강화 함수항에 의한 변위 $BR(t)$가 Gaussian Process임을 증명하기 위하여 먼저 $BR(t)$의 평균을 구하면 X_{t_0+k}가 1을 가질 확률과 -1을 가질 확률이 같으므로 $BR(t)$의 평균은 0이고 다음과 같이 표시한다.

$$\boldsymbol{E}BR(t)=0$$

$BR(t)$의 평균은 0이므로 각 k에 대한 X_{t_0+k}의 Variance는

$$VAR(X_{t_0+k}) = \boldsymbol{E}\ X^2_{t_0+k} = p_\delta$$

따라서 $BR(t)$의 Variance는 다음과 같다.

$$VAR(BR(t)) = \boldsymbol{E}BR^2(t) = \lambda^2(t-t_0)p_\delta$$
$$VAR(BR(t))|_{t_0=0} = \lambda^2 t p_\delta$$

한편 Assumption 1의 ii에서 입력 v_{t_0+k}는 $v_{t_0+k} \in B^o(\widehat{w^r}, \rho_{t_0+k})$이며 평균 $\widehat{w^r}$인 Gaussian분포를 가지므로

$$P(v_{t_0+k} - w^r_t \geq 0 \mid \delta(t_0+k)=1) = \int_{v_{t_0+k} - w^r_t \geq 0}^{\infty} \frac{1}{\sqrt{2\pi\sigma_{\rho_t}}} e^{-\frac{(v_{t_0+k} - w^r_t)^2}{2\sigma^2_{\rho_t}}} dv_{t_0+k}$$

따라서 $P(v_{t_0+k} - w^r_t \geq 0 \mid \delta(t_0+k)=1)$의 확률밀도 함수

$$f(v_{t_0+k} - w^r_t \geq 0 \mid \delta(t_0+k)=1) = d_{v_{t_0+k}} P(v_{t_0+k} - w^r_t \geq 0 \mid \delta(t_0+k)=1)$$

는 Gaussian 분포이며, 각 X_{t_0+k}이 Gaussian 분포를 따르므로 Chain Rule에 의해

$$f(BR(t)=B_t|B_t) = J^{-1}\left(\frac{\partial BR(t)}{\partial X_{t_0+k}}\right) J^{-1}\left(\frac{\partial X_{t_0+k}}{\partial v_{t_0+k}}\right) \cdot f(v_{t_0+k} - w^r_t \geq 0 | \delta(t_0+k)=1)$$

$BR(t)$는 평균 0, Variance $\lambda^2 t p_\delta$를 따르는 Gaussian 분포를 가진다.

그러므로 $BR(t)$는 Gaussian Process 이다. ■

Corollary : $BR(t)$의 Co-variance는 $s \leq t$에 대하여 $COV(BR(s), BR(t)) = \lambda^2 p_\delta \delta(t-s)$이다.

Proof :

$$\begin{aligned} &COV(BR(s), BR(t)) \\ &= COV(BR(s), BR(s) + BR(t) - BR(s)) \\ &= COV(BR(s), BR(s)) + COV(BR(s), BR(t) - BR(s)) \\ &= COV(BR(s), BR(s)) = \lambda^2 p_\delta \delta(t-s) \end{aligned}$$ ■

Ⅳ. Langvine 경쟁학습 신경회로망의 Fokker-Plank 방정식 유도

Theorem 2에서 이진 강화함수 항은 Gaussian Process임이 밝혀졌으므로, 이진 강화함수와 일정적응 이득을 가진 경쟁학습 신경회로망의 학습 방정식은 다음과 같이 일반적인 Langevine 방정식의 형태를 가지게 된다.

$$\partial w^r_t = -\nabla J(w^r_t)dt + \eta(t, \cdot)dt \qquad (4)$$

Theorem 2에서 이진 강화함수 항 $\eta(t, \cdot)$ 는 "Gaussian Process"이므로 다음의 성질을 만족한다.

$$E(\eta(t, \cdot)) = \langle \eta(t, \cdot) \rangle = 0$$
$$E(\eta(t, \cdot)\eta(t, \cdot)) = \langle \eta(t, \cdot)\eta(s, \cdot) \rangle = \lambda^2 p_\delta \delta(t-s)$$
$$E\left(\prod_{i=1}^{n} \eta(t_i, \cdot)\right) = \left\langle \prod_{i=1}^{n} \eta(t_i, \cdot) \right\rangle = \begin{cases} 0 & : l\ is\ odd \\ \sum\prod \lambda^2 p_\delta \delta(t-s) & : l\ is\ even \end{cases}$$

시간 t에서 r번째 Weight vector w^r_t이 열린구간 $(w^r_t, w^r_t + dw^r_t)$사이에 있을 확률에 대한 확률밀도 함수를 $\rho(w^r_t, t)$라 놓고 $\rho(w^r_t, t)$의 시간에 대한 변화율을 고려하면,

$$\frac{\partial \rho(w^r_t, t)}{\partial t} = -\frac{\partial}{\partial w^r_t}\left[\frac{\partial w^r_t}{\partial t}\rho(w^r_t, t)\right] \tag{5}$$

식 (1)을 식 (2)에 대입하면

$$\frac{\partial \rho(w^r_t, t)}{\partial t} = -\frac{\partial}{\partial w^r_t}[(-\nabla J(w^r_t) + \eta(t, \cdot))]\rho(w^r_t, t) \tag{6}$$

여기서 변위에 대한 Foward 선형 미분 연산자 L_f를 다음과 같이 정의한다.

Definition 1 : 변위에 대한 Foward 선형 미분 연산자 L_f는 다음과 같이 정의된다.

$$L_f\rho(x, t) \equiv \frac{\partial}{\partial x}[A(x)\rho(x, t)] = \left[\frac{\partial A(x)}{\partial x} + A(x)\frac{\partial}{\partial x}\right]\rho(x, t)$$

Definition 1에 의해 정의된 연산자 L_f를 식 (6)에 도입하면

$$\begin{aligned} L_f\rho(w^r_t, t) &= \frac{\partial}{\partial w^r_t}[\nabla J(w^r_t)\rho(w^r_t, t)] \\ &= \left[\frac{\partial \nabla J(w^r_t)}{\partial w^r_t} + \nabla J(w^r_t)\frac{\partial}{\partial w^r_t}\right]\rho(w^r_t, t) \end{aligned} \tag{7}$$

식 (7)을 식 (6)에 대입하면

$$\frac{\partial \rho(w^r_t, t)}{\partial t} = L_f\rho(w^r_t, t) - \frac{\partial \rho(w^r_t, t)}{\partial w^r_t}\eta(t, \cdot) \tag{8}$$

편미분 방정식 (8)을 시간에 대하여 풀기 위해 식 (8)의 해의 한 후보를 다음과 같이 놓는다.

$$\rho(w^r_t, t) = e^{tL_f}\xi(w^r_t, t) \tag{9}$$

식 (9)에서 $\xi(w^r_t, t)$는 시간 t에 종속인 함수이므로 w^r_t에 대한 미분 값은 0이다. 식 (9)을 t에 대하여 편미분하면

$$\begin{aligned} \frac{\partial \xi(w^r_t, t)}{\partial t} &= -L_f(w^r_t)e^{-tL_f}\rho(w^r_t, t) + e^{-tL_f}\frac{\partial \rho(w^r_t, t)}{\partial t} \\ &= e^{-tL_f}[-L_f(w^r_t)\rho(w^r_t, t) + L_f(w^r_t)\rho(w^r_t, t) - \eta(t, \cdot)\frac{\partial \rho(w^r_t, t)}{\partial w^r_t}] \\ &= -\eta(t, \cdot)e^{-tL_f}[\frac{\partial e^{-tL_f}}{\partial w^r_t}\xi(w^r_t, t) + \frac{\partial \xi(w^r_t, t)}{\partial w^r_t}e^{-tL_f}] \end{aligned} \tag{10}$$

식 (10)에서 $\xi(w^r_t, t)$의 t에대한 미분연산자 $K(w^r_t, t)$를 다음과 같이 놓는다.

$$K(w^r_t, t) = -\eta(t, \cdot)e^{-tL_f(w^r_t)}g(w^r_t, t)\frac{\partial}{\partial w^r_t}e^{tL_f(w^r_t)}$$

$\xi(w^r_t, t)$를 적분형 Taylor급수로 놓으면

$$\begin{aligned} \xi(w^r_t, t+\Delta t) &= \xi(w^r_t, t) + \int_t^{t+\Delta t} dt_1 K(w^r_t, t)\xi(w^r_t, t) \\ &= [1 + \int_t^{t+\Delta t} dt_1 K_1(w^r_t, t) + \frac{1}{2!}\int_t^{t+\Delta t} \\ &\quad dt_1\int_t^{t+\Delta t} dt_2 K_1(w^r_t, t)K_2(w^r_t, t) \\ &\quad + o(w^r_t, t)]\xi(w^r_t, t) \end{aligned} \tag{11}$$

$\xi(w^r_t, t+\Delta t)$의 평균을 식 (11)에서 구하게 되면

$$\begin{aligned} E\xi(w^r_t, t+\Delta t) &= [1 + \int_t^{t+\Delta t} dt_1 EK_1(w^r_t, t) + \frac{1}{2!}\int_t^{t+\Delta t} dt_1\int_t^{t+\Delta t} \\ &\quad dt_2 E(K_1(w^r_t, t)K_2(w^r_t, t)) \\ &\quad + Eo(w^r_t, t)]E\xi(w^r_t, t) \end{aligned} \tag{12}$$

식 (12)에서 $K(w^r_t, t)$의 2차 모멘트 까지의 평균을 구하면

$$E(K(w^r_t, t)\xi(w^r_t, t)) = E\eta(t, \cdot)e^{-tL_f(w^r_t)}\frac{\partial}{\partial w^r_t}e^{tL_f(w^r_t)}E\xi(w^r_t, t) = 0$$

$$\begin{aligned} E(K(w^r_t, t)K(w^r_t, s)g(w^r_t, t)) &= E\eta(t, \cdot)\eta(s, \cdot))e^{-tL_f(w^r_t)}\frac{\partial}{\partial w^r_t}e^{tL_f(w^r_t)} \\ &\quad \cdot e^{-sL_f(w^r_t)}\frac{\partial}{\partial w^r_t}e^{sL_f(w^r_t)}E\xi(w^r_t, t) \\ &= \lambda^2 p_s\delta(t-s)e^{-2tL_f(w^r_t)}\frac{\partial^2}{\partial^2 w^r_t}e^{tL_f(w^r_t)}E\xi(w^r_t, t) \end{aligned}$$

위 결과를 통해 Talyor급수의 2차항에 대한 평균을 구하면

$$\begin{aligned} &\frac{1}{2!}\int_t^{t+\Delta t} dt_1\int_t^{t+\Delta t} dt_2 E(K_1(w^r_t, t)K_2(w^r_t, t)\xi(w^r_t, t)) \\ &= \frac{1}{2!}\int_t^{t+\Delta t} dt_1\int_t^{t+\Delta t} \lambda^2 p_s\delta(t_1 - t_2)e^{-t_1L_f(w^r_t)} \\ &\quad \cdot \frac{\partial}{\partial w^r_t}e^{(t-t_2)L_f(w^r_t)}\frac{\partial}{\partial w^r_t}e^{-t_2L_f(w^r_t)}E\xi(w^r_t, t) \\ &= \frac{1}{2}\lambda^2 p_s\Delta t e^{-tL_f(w^r_t)}\frac{\partial^2}{\partial^2 w^r_t}e^{tL_f(w^r_t)}E\xi(w^r_t, t) \end{aligned}$$

한편 고차항에서는 앞에서 살펴본 바와 같이 차수 n이 홀수이면 0, 2n이면 $(\Delta t)^n$에 비례한다. 따라서 $\Delta t \to 0$의 극한의 경우 2차항까지만 고려한다면

$$\begin{aligned} \lim_{\Delta t \to 0}\frac{E\xi(w^r_t, t+\Delta t) - E\xi(w^r_t, t)}{\Delta t} &= \frac{\partial E\xi(w^r_t, t)}{\partial t} \\ &= \frac{1}{2}\lambda^2 p_s e^{-tL_f(w^r_t)}\frac{\partial^2}{\partial^2 w^r_t} \\ &\quad e^{tL_f(w^r_t)}E\xi(w^r_t, t) \end{aligned}$$

$P(w^r_t, t)$를 $\rho(w^r_t, t)$의 ensemble에서 취한 평균으로 놓으면

$$P(w^r_t, t) = E\rho(w^r_t, t) = e^{tL_f(w^r_t)}E\xi(w^r_t, t)$$

따라서 다음의 Fokker-Plank방정식을 얻을 수 있다.

$$\begin{aligned} \frac{\partial P(w^r_t, t)}{\partial t} &= L_f e^{tL_f}E\xi(w^r_t, t) + e^{tL_f}\frac{\partial E\xi(w^r_t, L_f)}{\partial t} \\ &= L_f P(w^r_t, t) + \frac{1}{2}\lambda^2 p_s\frac{\partial^2}{\partial^2 w^r_t}e^{tL_f(w^r_t)}E\xi(w^r_t, t) \\ &= \frac{\partial}{\partial w^r_t}[\nabla J(w^r_t)P(w^r_t, t)] + \frac{1}{2}\lambda^2 p_s\frac{\partial^2}{\partial^2 w^r_t}P(w^r_t, t) \end{aligned}$$

V. Fokker-Plank 방정식을 사용한 Langevine 경쟁학습 신경회로망의 대역 최소화 근사 해석

앞에서 살펴본 바 Assumption 1에서 "최적 Weight vector $\widehat{w}^r$의 Neighborhood을 도입함에 따라 실제로 4절까지의 모든 분석은 국소적인 분석에 지나지 않는다. 따라서 4절에 유도된 Fokker-Plank 방정식의 해를 통한 고찰도 실제로는 국소적인 특성만을 분석한 것이 된다. 그러나 경쟁학습 신경회로망에서는 다수의 Weight Vector들에 의해 입력 벡터공간을 다수의 단체(Simplex)들로 분할하는 것이므로 하나의 단체를 덮을 수 있는 Neighborhood을 도입하여, Neighborhood내의 대역 최소점을 찾을 수 있는가를 분석하여야 한다. 이를 위해 먼저 Fokker-Plank방정식에 대한 Infinitestimal Operator $\mathcal{L}_t^\sigma$ 도입하고, $\mathcal{L}_t^\sigma$의 Quasi-Norm을 정의하여 Weight vector w_t^r ^. Transition 확률에 대한 방정식 (Langevine 경쟁학습 신경회로망의 Fokker-Plank방정식) $\partial_t P(w_t^r, t) = \mathcal{L}_t^\sigma P(w_t^r, t)$이 Contraction되어 있는가를 살펴 보아야 한다. 일반적으로 확산 방정식(Diffusion Equation)혹은 Fokker-Plank방정식의 해(Transition 확률)가 어떤 조건아래 Contraction 되어 있고 Transition 확률이 성능지표 $f(w_t^r)$의 미분동상 사상(Diffeomorphism)이면 Contraction 조건을 만족하는 범위내에서 대역 최소점을 찾을 수 있음이 알려져 있다[7].

Definition 2 : Fokker-Plank방정식에 대한 Infinitestimal Operator $\mathcal{L}_t^\sigma$를 다음과 같이 정의한다.

$$\mathcal{L}_t^\sigma = -\frac{\partial}{\partial w_t^r} a(w_t^r, v_t) + \frac{1}{2} \sigma(t) \frac{\partial^2}{\partial^2 w_t^r}$$

$a(w_t^r, v_t) = -\varepsilon \nabla f(w_t^r)$, $\sigma(t) = \lambda^2 p_\delta(t)$, $t > 0$이며 Fokker-Plank방정식은

$$\frac{\partial P(w^r_t, t)}{\partial t} - \mathcal{L}_t^\sigma P(w_t^r, t) = 0$$

Definition 3 : Infinitestimal Operator $\mathcal{L}_t^\sigma$는 $P(\cdot) \in L[0,1]$인 임의의 함수 f에 대하여 다음과 같이 Quasi-Norm을 정의한다.

$$\mathcal{L}_t = \|\mathcal{L}_t^\sigma\| := \sup_{t_0 \le \tau \le t} \mathcal{L}_\tau^\sigma P(w_\tau^r, t)$$

Infinitestimal Operator $\mathcal{L}_t^\sigma$의 Quasi-Norm $\mathcal{L}_t$는 오직 다음의 Norm 성질만 만족한다.

$$\mathcal{L}_t = 0 \quad \forall P(w_t^r, t) \in L[0,1] \cdot \sup_{t_0 \le \tau \le t} \mathcal{L}_\tau^\sigma P(w_t^r, t) = 0$$

Theorem 3 : Langevine 경쟁학습 신경회로망의 Fokker-Plank방정식 $\partial_t P(w_t^r, t) = \mathcal{L}_t^\sigma P(w_t^r, t)$ 은 Infinitestimal Operator $\mathcal{L}_t^\sigma$의 Quasi-Norm의 시간에 대한 적분이 다음을 만족하면 $L[0,1]$에서 Contraction 되어 있다

$$\exists t > t_0,\ t \in R^+ \quad s.t. \quad \int_{t_0}^t \mathcal{L}_s ds \le \log\left(1 - \frac{\varepsilon}{P(w^r_{t_0}, t_0)}\right)$$
$$\forall \varepsilon \in (0, P(w^r_{t_0}, t_0))$$

Proof :

Langevine 경쟁학습 신경회로망의 Fokker- Plank방정식은 다음과 같다.

$$\frac{\partial P(w^r_t, t)}{\partial t} = \mathcal{L}_t^\sigma P(w_t^r, t) \le \mathcal{L}_t P(w_t^r, t) \tag{13}$$

$\mathcal{L}_s$를 다음과 같이 정의하고

$$\mathcal{L}_s = \sup_{s-\varepsilon < \tau < s+\varepsilon} \mathcal{L}_s^\sigma P(w_s^r, s) \quad \forall P(w_s^r, s) \in L[0,1]$$

식 (13)의 양변에 $\exp(-\int_{t_0}^t \mathcal{L}_s ds)$을 곱한다.

$$\frac{\partial P(w^r_t, t)}{\partial t} \exp\left(-\int_{t_0}^t \mathcal{L}_s ds\right) - \mathcal{L}_t P(w_t^r, t) \exp\left(-\int_{t_0}^t \mathcal{L}_s ds\right) \le 0$$
$$\frac{\partial}{\partial t} P(w^r_t, t) \exp\left(-\int_{t_0}^t \mathcal{L}_s ds\right) \le 0 \tag{14}$$

식 (14)를 Lebesgue 적분하면

$$\int_{P(w^r_{t_0}, t_0)}^{P(w^r_t, t)\exp(-\int_{t_0}^t \mathcal{L}_s ds)} \partial P(w^r_t, t) \exp\left(-\int_{t_0}^t \mathcal{L}_s ds\right) \le 0 \Rightarrow$$
$$P(w^r_t, t) \exp\left(-\int_{t_0}^t \mathcal{L}_s ds\right) \le P(w^r_{t_0}, t_0)$$

$$P(w^r_t, t) \le P(w^r_{t_0}, t_0) \exp\left(\int_{t_0}^t \mathcal{L}_s ds\right) \tag{15}$$

식 (15)의 양변에 초기 Transient 확률 $P(w^r_{t_0}, t_0)$를 빼고 Norm을 취하면 다음의 관계식이 성립한다.

$$P(w^r_t, t) - P(w^r_{t_0}, t_0) \le P(w^r_{t_0}, t_0) \exp\left(\int_{t_0}^t \mathcal{L}_s ds\right) - P(w^r_{t_0}, t_0)$$
$$\|P(w^r_t, t) - P(w^r_{t_0}, t_0)\| \le \|P(w^r_{t_0}, t_0)[\exp\left(\int_{t_0}^t \mathcal{L}_s ds\right) - 1]\|$$
$$\le \|P(w^r_{t_0}, t_0)\| \cdot \|1 - \exp\left(\int_{t_0}^t \mathcal{L}_s ds\right)\|$$

가정에서 $\int_{t_0}^t \mathcal{L}_s ds \le \log\left(1 - \frac{\varepsilon}{P(w^r_{t_0}, t_0)}\right)$ 이므로

$$\|P(w^r_t, t) - P(w^r_{t_0}, t_0)\| \le \|P(w^r_{t_0}, t_0)\| \cdot \|1 - \exp\left(\int_{t_0}^t \mathcal{L}_s ds\right)\|$$
$$\le \|P(w^r_{t_0}, t_0)\| \cdot \|1 - \left(1 - \frac{\varepsilon}{P(w^r_{t_0}, t_0)}\right)\|$$
$$\le \varepsilon \tag{16}$$

가정에서 ε은 $\varepsilon \in (0, P(w^r_{t_0}, t_0))$를 만족하는 임의의 양수이므로 Langevine 경쟁학습 신경회로망의 Fokker-Plank방정식 $\partial_t P(w^r_t, t) = \mathcal{L}^{\sigma}_t P(w^r_t, t)$은 $L[0,1]$에서 Contraction되어 있다. ■

Lemma 1 :

$$\forall \xi > 0 \ , \ \exists t_\xi > t_0 \quad s.t. \quad \int_{t_0}^{t} \frac{(w^r_\tau - w^r_{t_0})}{\|w^r_\tau - w^r_{t_0}\|} \eta(\tau, \cdot) P(w^r_\tau, t) d\tau < \xi \tag{17}$$

proof :

Transition확률 $P(w^r_\tau, t)$는 1 보다 작으므로 다음이 성립한다.

$$\int_{t_0}^{t} \frac{(w^r_\tau - w^r_{t_0})}{\|w^r_\tau - w^r_{t_0}\|} \eta(\tau, \cdot) P(w^r_\tau, t) d\tau \le \int_{t_0}^{t} \frac{(w^r_\tau - w^r_{t_0})}{\|w^r_\tau - w^r_{t_0}\|} \eta(\tau, \cdot) d\tau$$

논의를 간편히 하기 위해 $y_\tau = \dfrac{(w^r_\tau - w^r_{t_0})}{\|w^r_\tau - w^r_{t_0}\|}$이라 놓으면, y_τ의 상한은 1 이다. 따라서

$$\int_{t_0}^{t} \frac{(w^r_\tau - w^r_{t_0})}{\|w^r_\tau - w^r_{t_0}\|} \eta(\tau, \cdot) d\tau = \int_{t_0}^{t} y_\tau \eta(\tau, \cdot) d\tau \tag{18}$$

Assumption 1과 2, 그리고 Theorem 2에서

$$\eta(\tau, \cdot) = \begin{cases} \lambda & v_\tau - w^r_t \ge 0 \ \& \ \delta(\tau) = 1 \\ -\lambda & v_\tau - w^r_t < 0 \ \& \ \delta(\tau) = 1 \end{cases}$$

혹은

$$\eta(\tau, \cdot) = \begin{cases} \lambda & \textit{with Probability } 0.5 E\delta(t) \\ -\lambda & \textit{with Probability } 0.5 E\delta(t) \end{cases}$$

이므로

$$\begin{aligned} \int_{t_0}^{t} y_\tau \eta(\tau, \cdot) d\tau &= \lambda \int_{t_0}^{t} y_\tau d\tau P(v_\tau - w^r_\tau \ge 0, \ \delta(\tau) = 1) \\ &\quad - \lambda \int_{t_0}^{t} y_\tau d\tau P(v_\tau - w^r_\tau < 0, \ \delta(\tau) = 1) \end{aligned} \tag{19}$$

그런데 $P(\delta(\tau) = 1)$ 은 $\tau \to \infty$에 따라 $P(\delta(\tau) = 1) \to 0$ 이므로

"Weak Law of Large Numbers"에 의해[8]

$$\begin{aligned} \exists t_\xi > t_0 \ s.t. \ P[y_t | y_t \ge \xi] &= \int_{t_0}^{t} y_\tau d\tau P(v_\tau - w^r_\tau \ge 0, \delta(\tau) = 1) \\ &\quad - \int_{t_0}^{t} y_\tau d\tau P(v_\tau - w^r_\tau < 0, \delta(\tau) = 1) \\ &= \frac{1}{2} \int_{t_0}^{t} y_\tau d\tau P(\delta(\tau) = 1) - \frac{1}{2} \int_{t_0}^{t} y_\tau d\tau P(\delta(\tau) = 1) \\ &= \frac{1}{\lambda} \int_{t_0}^{t} y_\tau \eta(\tau, \cdot) d\tau < \xi \end{aligned} \tag{20}$$

따라서 $\xi > 0$ 인 임의의 ξ에 대하여 $\int_{t_0}^{t} \dfrac{(w^r_\tau - w^r_{t_0})}{\|w^r_\tau - w^r_{t_0}\|} \eta(\tau, \cdot) P(w^r_\tau, t) d\tau < \xi$을 만족하는 $t_\xi > t_0$인 t_ξ가 존재한다. ■

Assumption 3 : 성능지표 $J(w^r_t)$는 Lipschitz 조건을 만족한다 i.e.

$$\exists K > 0 \ s.t \ \|J(w^r_{t_1}) - J(w^r_{t_2})\| \le K \|w^r_{t_1} - w^r_{t_2}\| \ \forall J(w^r_t) \in R, \forall w^r_t \in R^n$$

Lemma 2 : $R^+ \ \forall \varepsilon \in (0, P(w^r_{t_0}, t_0))$에 대하여 성능지표 $J(w^r_t)$가 Convex함수일 때 다음을 만족하는 $t > t_0$이며 R^+의 한 원소인 t가 존재한다.

$$\int_{t_0}^{t} \mathcal{L}_s ds \le \log\left(1 - \frac{\varepsilon}{P(w^r_{t_0}, t_0)}\right) \tag{21}$$

proof :

위 명제가 거짓이다 즉, $\forall t > t_0$, $\int_{t_0}^{t} \mathcal{L}_s ds > \log\left(1 - \dfrac{\varepsilon}{P(w^r_{t_0}, t_0)}\right)$이라 가정하자.

Langevine 경쟁학습 신경회로망의 확률미분방정식(SDE)형식에서

$$\begin{aligned} \partial w^r_t &= -\varepsilon_L \nabla J(w^r_t) dt + \eta(t, \cdot) dt \\ w^r_t - w^r_{t_0} &= -\int_{t_0}^{t} \varepsilon_L \nabla J(w^r_t) - \eta(t, \cdot) dt \end{aligned} \tag{22}$$

양변에 $(w^r_t - w^r_{t_0}) P(w^r_t, t)$을 곱하면

$$\begin{aligned} \|w^r_t - w^r_{t_0}\|^2 P(w^r_t, t) &= -\int_{t_0}^{t} [\varepsilon_L (w^r_\tau - w^r_{t_0}) \nabla J(w^r_\tau) \\ &\quad - (w^r_\tau - w^r_{t_0}) \eta(\tau, \cdot)] P(w^r_\tau, \tau) d\tau \end{aligned} \tag{23}$$

Theorem 1과 Convex조건에 의해 Langevine 경쟁학습에 의해 생성되는 수열 $\{w^r_\tau\}_{t_0}^{t}$에 대하여 $\forall \tau > t_0$ $J(w^r_{t_0}) - J(w^r_\tau) \ge (w^r_{t_0} - w^r_\tau) \nabla J(w^r_\tau)$ 이므로

$$\begin{aligned} \|w^r_t - w^r_{t_0}\|^2 P(w^r_t, t) &= \int_{t_0}^{t} [\varepsilon_L (w^r_{t_0} - w^r_\tau) \nabla J(w^r_\tau) + (w^r_\tau - w^r_{t_0}) \eta(\tau, \cdot)] P(w^r_\tau, \tau) d\tau \\ &\le \int_{t_0}^{t} [\varepsilon_L (J(w^r_{t_0}) - J(w^r_\tau)) + (w^r_\tau - w^r_{t_0}) \eta(\tau, \cdot)] P(w^r_\tau, \tau) d\tau \\ &\le \int_{t_0}^{t} [\varepsilon_L \|J(w^r_{t_0}) - J(w^r_\tau)\| + (w^r_\tau - w^r_{t_0}) \eta(\tau, \cdot)] P(w^r_\tau, \tau) d\tau \end{aligned} \tag{24}$$

또한 Lipschtz 조건에서

$$\begin{aligned} &\int_{t_0}^{t} [\varepsilon_L \|J(w^r_{t_0}) - J(w^r_\tau)\| + (w^r_\tau - w^r_{t_0}) \eta(\tau, \cdot)] P(w^r_\tau, \tau) d\tau \\ &\qquad \le \int_{t_0}^{t} [\varepsilon_L K \|w^r_\tau - w^r_{t_0}\| + (w^r_\tau - w^r_{t_0}) \eta(\tau, \cdot)] P(w^r_\tau, \tau) d\tau \end{aligned} \tag{25}$$

그러므로

$$\begin{aligned} \|w^r_t - w^r_{t_0}\|^2 P(w^r_t, t) \le \int_{t_0}^{t} [&\varepsilon_L K \|w^r_\tau - w^r_{t_0}\| \\ &+ (w^r_\tau - w^r_{t_0}) \eta(\tau, \cdot)] P(w^r_\tau, \tau) d\tau \end{aligned}$$

따라서

$$\|w_t^r - w_{t_0}^r\| P(w_t^r, t) \le \int_{t_0}^{t} \left[\varepsilon_L K + \frac{(w_\tau^r - w_{t_0}^r)}{\|w_\tau^r - w_{t_0}^r\|} \eta(\tau, \cdot) \right] P(w_\tau^r, \tau) d\tau \tag{26}$$

식 (15)에서 $P(w^r{}_t, t) \le P(w^r{}_{t_0}, t_0)\exp(\int_{t_0}^{t} \mathcal{L}_s ds)$이고 $P(w^r{}_{t_0}, t_0)\exp(\int_{t_0}^{t} \mathcal{L}_s ds) > 0$ 이므로

$$\begin{aligned} \|w_t^r - w_{t_0}^r\| P(w_\tau^r, t) &\le \int_{t_0}^{t} \varepsilon_L KP(w^r{}_{t_0}, t_0)\exp(\int_{t_0}^{t} \mathcal{L}_s ds) d\tau \\ &+ \int_{t_0}^{t} \frac{(w_\tau^r - w_{t_0}^r)}{\|w_\tau^r - w_{t_0}^r\|} \eta(\tau, \cdot) P(w_\tau^r, t) d\tau \\ &\le \varepsilon_L KP(w^r{}_{t_0}, t_0) \int_{t_0}^{t} \exp(\int_{t_0}^{\tau} \mathcal{L}_s ds) d\tau \\ &+ \int_{t_0}^{t} \frac{(w_\tau^r - w_{t_0}^r)}{\|w_\tau^r - w_{t_0}^r\|} \eta(\tau, \cdot) P(w_\tau^r, t) d\tau \end{aligned} \tag{27}$$

가정에서 $\forall t > t_0$, $\int_{t_0}^{t} \mathcal{L}_s ds > \log(1 - \frac{\varepsilon}{P(w^r{}_{t_0}, t_0)})$이므로 $\int_{t_0}^{t} \exp(\int_{t_0}^{\tau} \mathcal{L}_s ds) d\tau$ 의 하한은

$$\int_{t_0}^{t} \exp(\int_{t_0}^{\tau} \mathcal{L}_s ds) d\tau > \int_{t_0}^{t} (1 - \frac{\varepsilon}{P(w^r{}_{t_0}, t_0)}) d\tau = (1 - \frac{\varepsilon}{P(w^r{}_{t_0}, t_0)})(t - t_0) \tag{28}$$

이다. Lemma 1에서 $\lim_{t \to \infty} \int_{t_0}^{t} \frac{(w_\tau^r - w_{t_0}^r)}{\|w_\tau^r - w_{t_0}^r\|} \eta(\tau, \cdot) P(w_\tau^r, t) d\tau \to 0$ 이므로

$$\begin{aligned} \|w_t^r - w_{t_0}^r\| P(w_\tau^r, t) &> \int_{t_0}^{t} \frac{(w_\tau^r - w_{t_0}^r)}{\|w_\tau^r - w_{t_0}^r\|} \eta(\tau, \cdot) P(w_\tau^r, t) d\tau = H(t_\xi) \\ &> \int_{t_0}^{t_\xi + t_o} \frac{(w_\tau^r - w_{t_0}^r)}{\|w_\tau^r - w_{t_0}^r\|} \eta(\tau, \cdot) P(w_\tau^r, t) d\tau \quad \forall t_o > 0 \end{aligned}$$

되는 t_ξ를 취할 수 있고, 식 (27)은 다음과 같이 쓸 수 있다.

$$\frac{\|w_t^r - w_{t_0}^r\| P(w_\tau^r, t)}{\varepsilon_L KP(w^r{}_{t_0}, t_0)} \le \int_{t_0}^{t} \exp(\int_{t_0}^{\tau} \mathcal{L}_s ds) d\tau + \frac{H(t_\xi + t_o)}{\varepsilon_L KP(w^r{}_{t_0}, t_0)}$$

그런데 $\varepsilon \to P(w^r{}_{t_0}, t_0)$ 에 따라 $inf \int_{t_0}^{t} \exp(\int_{t_0}^{\tau} \mathcal{L}_s ds) d\tau \to 0^+$ $\forall t > t_0$, $t \in R^+$ 이므로

$$\frac{\|w_t^r - w_{t_0}^r\| P(w_\tau^r, t)}{\varepsilon_L KP(w^r{}_{t_0}, t_0)} - \frac{H(t_\xi + t_o)}{\varepsilon_L KP(w^r{}_{t_0}, t_0)} > \int_{t_0}^{t} \exp(\int_{t_0}^{\tau} \mathcal{L}_s ds) d\tau \tag{29}$$

되는 t가 존재한다. 이는 가정에 모순. 따라서 다음을 만족하는 $t > t_0$인 t가 존재한다.

$$\int_{t_0}^{t} \mathcal{L}_s ds \le \log(1 - \frac{\varepsilon}{P(w^r{}_{t_0}, t_0)}). \quad \blacksquare$$

Corollary 2 : R^+ $\forall \varepsilon \in (0, P(w^r{}_{t_0}, t_0))$에 대하여 성능지표 $J(w^r{}_t)$가 다음과 같은 Pseudo- Convex함수일 때 식 (21)을 만족하는 $t > t_0$ 이며 R^+의 한 원소인 t 가 존재한다.

$$\forall \tau > t_0 \quad J(w^r{}_{t_0}) - J(w_\tau^r) \ge (w^r{}_{t_0} - w_\tau^r) \nabla J(w_\tau^r) - o(\|w_\tau^r\|) \tag{30}$$

여기서 $o(\|w_\tau^r\|)$은 $(w_t^r + dw^r)$ 범위에서 (30)을 만족할 수 있는 $o(\|w_t^r\|)$의 절대값의 하한값을 가지며 다음의 두 특성을 만족한다.

$$\int_{t_0}^{t} o(\|w_\tau^r\|) P(w_\tau^r, \tau) d\tau \le 0 \qquad o(\|w_t^r\|) \to 0 \ as \ w_t^r \to w^r{}_*,$$

$w^r{}_*$는 대역 최소점(Globally Optimal point)이다.

proof :

Lemma 2의 증명에서 $o(\|w_\tau^r\|)$항을 포함하여 정리하면

$$\begin{aligned} \|w_t^r - w_{t_0}^r\| P(w_\tau^r, t) &\le \varepsilon_L KP(w^r{}_{t_0}, t_0) \int_{t_0}^{t} \exp(\int_{t_0}^{\tau} \mathcal{L}_s ds) d\tau \\ &+ \int_{t_0}^{t} o(\|w_\tau^r\|) P(w_\tau^r, t) d\tau \\ &+ \int_{t_0}^{t} \frac{(w_\tau^r - w_{t_0}^r)}{\|w_\tau^r - w_{t_0}^r\|} \eta(\tau, \cdot) P(w_\tau^r, t) d\tau \end{aligned} \tag{31}$$

가정에서 $\int_{t_0}^{t} o(\|w_\tau^r\|) P(w_\tau^r, \tau) d\tau \le 0$의 최대값은 0이므로 (31)은 (27)과 동일해 진다. 이후의 증명은 Lemma 2의 증명과 동일 ■

Theorem3, Lemma 2, Corollary 2의 결론에서 성능지표 $J(w^r{}_t)$가 단위 단체(Simplex)에서 Convex이거나, Corollary 2의 제한조건을 만족하는 Pseudo-Convex함수의 경우 단위 단체(Simplex)내의 대역 최소점(Globally Optimal point)을 취할 수 있음을 알 수 있다.

VI. 실험결과

실험에 사용한 데이터는 Flight Line C1(FLC1)으로 불리는 다중분광 지상관측 원격탐사 데이터(Multispectral Earth Observational Remote Sensing Data)로서 미국 인디애나주 Tippecanoe Country 남부지역의 농작물 재배지역을 촬영한 것이다. 이 데이터들은 256 Gray Level로 표현되고 4개의 클래스가 4개의 주요 농산물을 대표하도록 선택된 하나의 데이터와 8개의 클래스가 8개의 주요 농산물을 대표하도록 선택된 데이터 두가지 종류가 있다. 각 클래스는 학습을 위한 200개의 데이터 벡터들과 테스트를 위한 375

개의 테스트 벡터를 가지며 하나의 데이터 혹은 테스트 벡터는 8차원으로 구성되어 있으며 추정된 확률 분포와 실제 확률분포가 근사하여 분류 결과가 비교적 우수한 특징을 지닌다. 실험에 사용한 또하나의 데이터 집합은 미국 Colorado주의 산악 지역의 지형을 촬영한 것으로서 모두 13개의 서로 다른 지형을 나타내며 256 Gray Level로 표현되고 각 지형에 따라 서로 다른 벡터들의 개수를 가진다.

제안한 알고리즘에서 조정해주어야 할 파라메타는 경쟁학습 신경회로망의 경우 식 (32)와 같이 놓았으며 Langevine 경쟁학습의 경우 식 (33)과 같이 놓았다. 초기 Weight vector 선정은 각 클래스 당 같은 개수의 데이터 벡터들로 이루어 지도록 했으며 Weight vector 수는 20개이다. 실험환경은 Pentium-100 IBM-PC에서 Visual C++ 1.5 컴파일러로 행하였으며 MFC함수를 사용, Windows Application으로 프로그램 하였다.

$$\varepsilon(t) = 0.9 \cdot \left(1 - \frac{t}{Number\ of\ Iteration}\right) \tag{32}$$

$$\varepsilon = 2^{-4} = \frac{1}{16} = 0.0625 \qquad \lambda = 2^{-5} = \frac{1}{32} = 0.03125 \tag{33}$$

식 (33)에서 ε은 시간에 불변인 학습계수이며 λ는 이진 강화함수항의 이득이다.

표 1. FLC1 4-클래스 데이터에 대한 패턴인식 실험결과

Table 1. Experimental results of pattern recognition for FLC1 4 class Data.

Epoch	경쟁학습		Langevine 경쟁학습	
	학습 (%)	테스트 (%)	학습 (%)	테스트 (%)
100	93.525	92.09	94.625	92.93
200	94.0	90.6	96.125	94.40
300	94.25	93.0	94.75	93.73
400	94.625	91.13	95.875	93.53
500	94.625	92.4	96.125	94.87
600	95.875	91.93	95.875	94.73
700	94.625	88.47	96.375	94.60
800	95.375	88.0	96.125	94.20
900	94.75	83.13	95.0	93.93
1000	94.625	92.27	94.25	94.07
평균	94.625	92.07	95.51	94.10

표 2. FLC1 8-클래스 데이터에 대한 패턴인식 실험결과

Table 2. Experimental results of pattern recognition for FLC1 8 class Data.

Epoch	경쟁학습		Langevine 경쟁학습	
	학습 (%)	테스트 (%)	학습 (%)	테스트 (%)
100	89.88	86.67	90.06	89.07
200	91.125	88.33	89.25	86.93
300	90.8125	88.0	90.06	88.4
400	89.625	86.57	90.0	88.1
500	90.6875	88.6	90.38	88.77
600	90.3125	86.80	90.50	89.07
700	89.625	86.57	90.88	89.17
800	89.625	86.57	90.50	89.43
900	89.9375	86.7	90.31	89.3
1000	89.88	86.67	90.13	88.5
평균	90.15	87.35	90.21	88.67

표 3. Colorado 13-클래스 데이터에 대한 패턴인식 실험결과

Table 3. Experimental results of pattern recognition for Colorado 13 class Data.

Epoch	경쟁학습		Langevine 경쟁학습	
	학습 (%)	테스트 (%)	학습 (%)	테스트 (%)
100	48.31	49.46	62.40	61.42
200	47.52	48.66	62.10	61.23
300	48.31	49.46	62.0	61.33
400	48.31	49.36	60.22	60.04
500	48.31	49.46	63.49	62.61
600	47.52	48.67	62.80	61.82
700	48.31	49.46	61.51	60.53
800	48.31	49.46	59.52	59.94
900	47.52	48.67	60.81	59.45
1000	48.31	49.46	62.80	62.12
평균	48.07	49.21	61.77	61.05

실험결과 제안한 알고리즘이 학습 및 테스트 양면에서 기존의 경쟁학습 신경회로망 알고리즘보다 평균적으로 더 나은 성능을 보였다. 신경회로망에 의한 추정 확률분포와 실제 확률분포가 근사할 경우(FLC1 데이터), 즉 성능지표의 Convexity가 클 경우, 기존의 알고리즘보다 크게 분류 정확도가 향상되지 않으나, 성능지표의 Convexity가 떨어질 경우 (Colorado 13 클래스

데이터) Corollary 2의 결론을 통해 어느정도의 국소 최소점(Locally Optimal point)은 Langevine 경쟁학습 신경회로망이 극복 할 수 있음을 알 수 있다.

Ⅶ. 결 론

본 논문에서는 먼저 Langevine 경쟁학습 신경회로망의 학습 알고리즘이 기존의 경쟁학습 알고리즘에 Gaussian 분포를 가진 Browinian Motion Process가 결합된 것임을 밝히고 이를 통해 Langevine 경쟁학습 신경회로망의 학습 알고리즘의 Fokker-Plank 방정식을 유도하였다. 유도된 Fokker-Plank 방정식은 Markov 준군위에 정의된 2차 편 미분 방정식이므로 선형 준군동형사상을 도입하여 Transition 확률의 축소(Contraction)가 어떤 조건 위에 성립하는가를 밝혀 대역 최소성의 조건을 밝혔으며. 실험을 통해 해석 결과를 살펴 보고 증명된 해석 조건들의 타당성을 실험을 통해 검증하였다.

본 논문에서 유도된 단위 단체(Simplex)들에 대한 대역 최소화 충분조건들은, 그러나 상당히 강한 조건들로서 $\int_{t_0}^{t} \mathcal{L}_s ds \to -\infty$가 만족 되려면 $t \to \infty$임을 의미한다. 따라서 현재까지 알려진 거의 유일한 대역 최소화 알고리즘인 "Simulated Annealing"의 경우에서와 마찬가지로 유한번의 Epoch로 대역 최소점을 찾아내기는 사실상 어렵다는 것을 의미한다. 앞으로의 과제로는 본 논문에서 논한 분석이 매우 약한 위상에서의 분석임에 따라 좀 더 강한 위상에서, 전통적인 확률 미분 방정식의 해석방법으로 재 해석하여 본 알고리즘이 구체적으로 어떤 Nonconvex 조건까지를 만족 하면서 대역 최소점에 도달 할 수 있는지를 밝히는 것과, Homology 분석을 통해 초기 Weight Vector들의 입력공간 분할에 대한 해석을 하는 것이 요구된다.

참 고 문 헌

[1] J. Seok, S. Cho, "Self-Organizing Feature Map with Binary Reinforcement and Constant Adaptation Gain : For an easier Hardware Implementation", *Proc. ICONIP '94*, vol 2, 966-971 1, 1994.

[2] I.I. Gihman, A.V. Skorohod, *The theory of stochastic process I*, Springer-Verlag, 1974.

[3] H. Ritter, K. Schulten, "Convergence Properties of Kohonen's Topology Conserving Maps : Fluctuations, Stability, and Dimension Selection", *Biological Cybernetics*, vol 60, pp. 59-71, 1988.

[4] I. I. Gihman, A. V. Skorohod, *The Theory of Stochastic Process II*, Springer-Verlag, 1974.

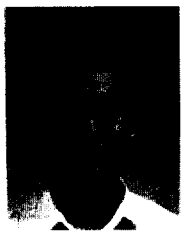
[5] Yu. Ermoliev, R. J-B Wets (Eds.), 'Numerical Techniques for Stochastic Optimization', *Springer-Verlag*, 1980.

[6] 석 진욱, 조 성원, 홍 윤광, "시불변 학습계수와 이진 강화함수를 가진 경쟁학습 신경회로망에 의한 고장진단 퍼지 전문가 시스템의 자동구축", *Proc '96 FAN 춘계종합학술대회 논문집*, pp 153-158, 1996

[7] T. S. Chiang, C. R. Hwang, S. J. Sheu, "Diffusion for Global Optimization in R^n", *SIAM J. Cont. and Optimization*, vol 25, no 3, pp 737-753, May 1987,

[8] P. Billingsley, *Probability and Measure*, Wiley, 3rd ED., 1994.

저 자 소 개



石 鎭 旭(正會員)

1969년 6월 26일생. 1993년 2월 홍익대학교 전기제어 공학과 졸업. 1995년 동대학원 전기 공학과 졸업(석사). 1997년 7월 현재 홍익대학교 대학원 전기 공학과 박사과정 재학중. 주관심분야는 신경회로망, 확률과정, 비선형 최적화, 비선형제어, 미분기하, 미분위상등임.

趙 成 元(正會員) 第 32卷 B編 第 1號 參照