Dept. of Math. University of Oslo
Statistical Research Report No. 3
ISSN 0806–3842 February 2005

How Online Learning Approaches Ornstein Uhlenbeck Processes

Fredrik A. Dahl

16th February 2005

Abstract

We show that under reasonable conditions, online learning for a nonlinear function near a local minimum is similar to a multivariate Ornstein Uhlenbeck process. This implies that the parameter state oscillates randomly around the minimum point, with a Gaussian limiting distribution.

1 Introduction

Our setting is this: We have a (large) set of training data (x_i, y_i), which we want to approximate with a parameterized function $f(w; x)$. We consider the algorithm of drawing random training patterns, and performing small gradient descent steps to minimize the error sequentially. In the neural net litterature, this algorithm is often referred to as *online learning* [1].

Online learning was first proposed as a model of learning in a biological context of neural networks [2], where the so called backpropagation of errors is the biological mechanism believed to compute the gradient.

Online learning is often used for computation problems as well, where artificial neural nets are used as a nonlinear regression function, without any particluar interest in biological interpretations. Considered as a computational tool, online learning has been critisized by some authors for introducing unnecesary random noise, and more powerful optimization algorithms exist that use second order derivatives [3]. However, the algorithm is still in use, as it is extremely simple to implement, scales favourably with the size of the training set and net size, and behaves reasonably well in most cases. The random noise of the algorithm may also have the beneficial effect of moving the parameter state away from suboptimal local minima. A different argument for the relevance of online learning is that many reinforcement learning algorithms rely on its incremental mode, see e.g. the famous backgammon application by Tesauro [4].

We do not attempt to make any strong argument in favour of online learning, but simply point out that as long as it is considered a relevant biological model and is also in use as a computational tool, it is important to understand its behavior.

2 Definitions

Let the error function for pattern $i \in \{1, ..., N\}$ and parameter state $w \in \mathbf{R}^n$ be given by

$$Err(i, w) = \frac{1}{2}(f(w, x_i) - y_i)^2 + \lambda w^T w$$

where f is three times continuously differentiable in w and $\lambda \geq 0$. The quadratic term $\lambda w^T w$ is often used in order to prevent the parameter vector from growing indefinitely, a problem often seen with neural networks that feature close to flat error surfaces for large parameter values. Also, a $\lambda > 0$ works as a penalty for overfitting, similar to ridge regression shrinkage used in linear regression models [5]. Note, however, that we allow $\lambda = 0$.

The total error function we define as:

$$Err(w) = \frac{1}{N} \sum_{1}^{N} Err(i, w)$$

For $\alpha > 0$ the online learning algorithm produces a random sequence w_k^α by

$$w_{k+1}^\alpha = w_k^\alpha - \alpha \nabla Err(I_k, w_k^\alpha)$$

where $\alpha > 0$ is the step length, I_k is a sequence of independent uniform random variables in $\{1, ..., N\}$, and the gradient is taken with respect to w.

For convenience, we define $\phi^k(w) = \nabla Err(I_k, w)$, so that

$$w_{k+1}^\alpha = w_k^\alpha - \alpha \phi^k(w_k^\alpha)$$

Given $w \in \mathbf{R}^n$, $\phi^k(w)$ is an iid sequence of discrete random variables, while for a given k, $\phi^k : \mathbf{R}^n \mapsto \mathbf{R}^n$ is twice differentiable.

Without loss of generality, we let 0 be a local minimum of Err. We assume that $Err(0) > 0$ and the Hessian H of Err at 0 is positive definite.

Our goal is to characterize the paths of w close to the minimum point, and therefore take $w_0 = 0$.

We will often be using a Taylor expansion of ϕ^k around 0:

$$\phi^k(w) = \phi^k(0) + H^k w + R^k(w)$$

where H^k is the Hessian of $Err(k,0)$ and the remainder term $R^k(w)$ is of the order w^2. Because 0 is a local minimum of Err, we have $E(\phi^k(0)) = 0$. Averaging this over k therefore gives

$$E(\phi^k(w)) = Hw + R_1(w)$$

where $R_1(w)$ is of the order w^2.

Let Σ be the covariance matrix of $\phi^k(0)$, which we assume to be invertible. The Taylor expansion gives $Var(\phi^k(w)) = \Sigma + Var(R^k(w) + H^k w) - 2Cov(\phi^k(0), R^k(w) + H^k w)$, which implies

$$Var(\phi^k(w)) = \Sigma + R_2(w)$$

where the matrix R_2 is of the order w.

3 The multivariate Ornstein Uhlenbeck process

The well-known onedimensional O-U process is a diffusion of this form, see [6]:

$$dX_t = -aX_t dt + sdB_t$$

where $a > 0, s \neq 0$ and B_t is standard Brownian motion. Its stationary distribution is normal with mean zero and variance $s^2/2a$.

The n-dimensional generalization of the O-U process has $a, s \in \mathbf{R}^n$. Schach [7] defines the multivariate O-U process as a continuous Markov process with multinormal limiting distribution. This implies that the eigenvalues of a must have positive real parts, and that s be nonsingular.

We are interested in the special case where a is symmetric and positive definite, so that all eigenvalues are real and positive, and the eigenvectors $\{e_i\}$ form an orthogonal basis. In this case, X can be decomposed to a sum of 1-dimensional O-U processes:

$$X_t = \sum_{i=1}^{n} \xi_t^i$$

Here, $\xi_t^i = e_i^T X_t$ solves

$$d\xi_t^i = -a_i \xi_t^i dt + \sigma_i dB_t$$

for some positive constants a_i, σ_i.

Let $Y_t = b^T X_t$, for some constant vector b. Then Y solves

$$dY_t = a(\omega, t) Y_t dt + \tilde{\sigma} dB_t$$

where $a(\omega, t)$ is continuous in t almost surely, and takes values in the range $[\min(a_i), \max(a_i)]$. The process $a(\omega, t)$ is adapted to the filtration of X_t, but not necessarily to that of

Y_t, so the latter will typically be non-Markov. However, for a large t, $E[a(\omega, t)]$ will approach a constant $\bar{a}$, regardless of the initial state X_0.

By the general theory of O-U processes, the limiting distribution of X is multinormal, so the limiting distribution of Y is also normal.

It should be noted that some authors [8] refer to the diffusion X_t defined above as the time derivative of the Ornstein Uhlenbeck process, while other authors refer to X_t as the O-U process itself [7]. This, of course, is only a matter of definition, and we choose to follow the latter tradition.

4 Transformation

We tranform the w^α process, depending on the step length α.

$\{w^\alpha\}_k \mapsto \{v^\alpha\}_t$, where v^α is defined for real times.

Let

$$v_t^\alpha = \alpha^{-1/2} w_{\lfloor t/\alpha \rfloor}$$

where $\lfloor \rfloor$ rounds downward to the closest integer.

This implies $v_{t+\alpha}^\alpha = \alpha^{-1/2} w_{\lfloor t/\alpha \rfloor + 1}^\alpha = \alpha^{-1/2} w_{\lfloor t/\alpha \rfloor}^\alpha - \sqrt{\alpha}\phi^{\lfloor t/\alpha \rfloor}(w_{\lfloor t/\alpha \rfloor}^\alpha)$, giving

$$v_{t+\alpha}^\alpha = v_t^\alpha - \sqrt{\alpha}\phi^{\lfloor t/\alpha \rfloor}(\sqrt{\alpha} v_t^\alpha)$$

with $v_0 = 0$.

Let X_t solve:

$$dX_t = -HX_t dt + \Sigma^{1/2} dB_t$$

with $X_0 = 0$. The process X is the multivariate Ornstein Uhlenbeck (O-U) process.

If we fix the value of v_t^α and let α fall toward zero, then by a Taylor approximation, $E[v_{t+\alpha}^\alpha - v_t^\alpha | v_t^\alpha] = -E[\sqrt{\alpha}\phi^{\lfloor t/\alpha \rfloor}(\sqrt{\alpha} v_t^\alpha)]$ approaches $-\alpha H v_t^\alpha$. Also, $Var[v_{t+\alpha}^\alpha - v_t^\alpha | v_t^\alpha]$ approaches $\alpha\Sigma$. If we regard α as an infinitesimal, then these properties are equivalent to those of the O-U process. We formalize this idea in the following section.

5 Main result

We will prove that the transformed process v^α converges weakly toward a multivariate Ornstein Uhlenbeck process, when α falls toward zero.

In order to prove the main result below, we first give a lemma stating "non-explosion in probability" of v^α.

LEMMA 1 *For $\epsilon > 0$ there exist $M, \delta > 0$ such that $P(\sup_{t\in[0,1]} ||v_t^\alpha|| > M) \leq \epsilon$ when $\alpha < \delta$.*

Proof:

Let $\bar{v}_t^\alpha$ be the drift compensated process:

$$\bar{v}_{t+\alpha}^\alpha = \bar{v}_t^\alpha + (v_{t+\alpha}^\alpha - v_t^\alpha) - E[v_{t+\alpha}^\alpha - v_t^\alpha | v_t^\alpha]$$

Now fix an $M > 0$.

For each point $v \in \mathbf{R}^n$, we have

$$lim_{\alpha\to 0} E[v_{t+\alpha}^\alpha - v_t^\alpha | v_t^\alpha = v] = -Hv$$

This linear Taylor approximation converges uniformly on the compact set $\{v : ||v|| \leq M\}$. This, combined with the fact that H is positive definite, implies the existence of an $e > 0$ such that $\alpha < e$ implies

$$P(||\bar{v}_t^\alpha|| \geq ||v_t^\alpha|| \ | sup_{s\in[0,t]}\{||\bar{v}_s^\alpha|| \leq M\}) = 1$$

for each $t \in [0, 1]$. Less formally, for small α, the correction term $E[v_{t+\alpha}^\alpha - v_t^\alpha | v_t^\alpha]$ pushes $\bar{v^\alpha}$ away from 0. Therefore, it suffices to show the lemma statement for $\bar{v^\alpha}$.

Define the stopping time $\tau = \inf\{t : t > 1 \ or \ ||\bar{v}_t^\alpha|| > M\}$, and let $\hat{v}^\alpha$ be $\bar{v}^\alpha$ stopped at time τ:

$$\hat{v}_t^\alpha = \bar{v}_{t\wedge\tau}^\alpha$$

By construction $\bar{v}^\alpha$ and $\hat{v}^\alpha$ are (non-Markov) martingales with respect to the filtration generated by v^α.

The event $sup_{t\in[0,1]}||\bar{v}_t^\alpha|| > M$ is identical to the event $||\hat{v}_1^\alpha|| > M$. We procede to show that the probability of this event can be made arbitrarily small, by choosing M large and letting α tend to zero.

When α falls toward 0, the covariance matrices of $(\hat{v}_{t+\alpha}^\alpha - \hat{v}_t^\alpha)/\sqrt{\alpha}$ converge uniformly toward Σ when $||\hat{v}^\alpha|| < M$. Therefore, $\limsup_\alpha ||Var(\hat{v}_1^\alpha)|| \leq ||\Sigma||$.

This bound does not depend on M. The result follows from the Markov inequality.

THEOREM 1 *Let $T > 0$. Then $\lim_{\alpha\downarrow 0} v_t^\alpha = X_t$ for $t \in [0, T]$ (weak convergence).*

Proof:

It suffices to show convergence for $T = 1$.

For $M > 0$ define the process $v^{M,\alpha}$ by:
$v_{t+\alpha}^{M,\alpha} = v_t^{M,\alpha} - \sqrt{\alpha}\phi^{\lfloor t/\alpha \rfloor}(\sqrt{\alpha} v_t^{M,\alpha})$ for $||v_t^{M,\alpha}|| < M$
$v_{t+\alpha}^{M,\alpha} = v_t^{M,\alpha} + \xi_t$ otherwise,
where $\xi_t \sim N(-\alpha H v_t^{M,\alpha}, \alpha\Sigma)$.

Let $f^\alpha(x) = E[v_{t+\alpha}^{M,\alpha} - v_t^{M,\alpha} | v_t^{M,\alpha} = x]/\alpha$ and $C^\alpha(x) = Cov[v_{t+\alpha}^{M,\alpha} - v_t^{M,\alpha} | v_t^{M,\alpha} = x]/\alpha$ and

We will prove that $v^{M,\alpha}$ converges to X (for a fixed M) through a theorem given in [9]. This result was originally stated for one-dimensional processes, but easily extends to multidimensional processes, as pointed out by [10].

Let $f(x) = -\alpha H x$ and $C(x) = \alpha\Sigma$, which are Lipschitz continuous. In order to use the theorem, we must show:

$$E[\sum_{k=0}^{1/\alpha-1} ||f^\alpha(v_{k\alpha}^{M,\alpha}) - f(v_{k\alpha}^{M,\alpha})||^2 + ||C^\alpha(v_{k\alpha}^{M,\alpha}) - C(v_{k\alpha}^{M,\alpha})||^2]\alpha \to 0$$

The terms of this sum with $||v_{k\alpha}^{M,\alpha}|| > M$ are zero by construction. From a Taylor approximation argument, it follows that f^α converges toward f, uniformly on the compact set where $||x|| \leq M$, which implies the desired convergence.

From our lemma we know that for $\epsilon > 0$ there are $M, \delta > 0$ so that
$P(v_{k\alpha}^{M,\alpha} = v_{k\alpha}^{\alpha}, \quad k = 0, ..., \alpha^{-1} - 1) < \epsilon$.

It follows that v^α converges weakly toward X.

References

[1] Hassoun, M. H., 1995: *Fundamentals of artificial neural networks*, MIT Press.

[2] Widrow, B., Hoff, M. E., 1960: Adaptive switching circuits. *IRE WESCON Convention record* 4, 94-104.

[3] Ripley, B. D., 1996: *Pattern recognition and neural networks*, Cambridge university press.

[4] Tesauro, G., 1992: Practical issues in temporal difference learning, *Machine learning*, 8, 257-277.

[5] Hastie, T., Tibshirani, R., Friedman, J., 2001: *The elements of statistical learning*, Springer.

[6] Øksendal, B., 1998: *Stochastic differential equations, an introduction with applications*, Springer.

[7] Schach, S., 1971: Weak convergence results for a class of multivariate markov processes *The annals of mathematical statistics*, 42 (2) pp. 451-465.

[8] Protter, P., 1990: *Stochastic integration and differential equations, a new approach*, Springer.

[9] Gikhman, I. I., Skorokhod A. V. 1969: *Introduction to the theory of stochastic processes*, Saunders, Philadelphia.

[10] Kushner, H. J., 1974: On the weak convergence of interpolated Markov chains to a diffusion, *The annals of Probability*, Vol.2, No. 1, pp. 40-50.