In *Proceedings of the 1993 Connectionist Models Summer School*, M.C. Mozer, P. Smolensky, D.S. Touretzky, J.L. Elman and A.S. Weigend (eds.), Erlbaum Associates, 1993.

# Momentum and Optimal Stochastic Search

**Genevieve B. Orr, Todd K. Leen**
Department of Computer Science and Engineering
Oregon Graduate Institute of Science & Technology
20000 Northwest Walker Road
PO Box 91000
Portland, OR 97291-1000
orr@cse.ogi.edu, tleen@cse.ogi.edu

## INTRODUCTION

The rate of convergence for gradient descent algorithms, both batch and stochastic, can be improved by including in the weight update a "momentum" term proportional to the previous weight update. Several authors [1, 2] give conditions for convergence of the mean and covariance of the weight vector for momentum LMS with *constant learning rate*. However stochastic algorithms require that the learning rate decay over time in order to achieve true convergence of the weight (in probability, in mean square, or with probability one).

This paper uses the dynamics of weight space probabilities [3, 4] to address stochastic gradient algorithms with learning rate annealing and momentum. This theoretical framework provides a simple, unified treatment of asymptotic convergence rates and asymptotic normality. The results for algorithms without momentum have been previously discussed in the literature. Here we gather those results under a common theoretical structure and extend them to stochastic gradient descent with momentum.

## DENSITY EVOLUTION AND ASYMPTOTICS

We consider stochastic optimization algorithms with weight $\omega \in R^N$. We confine attention to a neighborhood of a local optimum $\omega_*$ and express the dynamics in terms of the *weight error* $v \equiv \omega - \omega_*$. For simplicity we treat the continuous time algorithm [1]

$$\frac{dv(t)}{dt} = \mu(t)\, H[\, v(t), x(t)\,] \tag{1}$$

where $\mu(t)$ is the learning rate at time $t$, $H : R^N \times R^K \to R^N$ is the weight update function and $x(t)$ is the data fed to the algorithm at time $t$. For stochastic gradient algorithms $H = -\nabla_v \mathcal{E}(v, x(t)\,)$, the gradient of the instantaneous cost function.

[1] Although algorithms are executed in discrete time, continuous time formulations are often advantagous for analysis. The passage from discrete to continuous time is treated in various ways depending on the needs of the theoretical exposition. Kushner and Clark [5] define continous time functions that interpolate the discrete time process in order to establish an equivalence between the asymptotic behavior of the discrete time stochastic process, and solutions of an associated deterministic differential equation. Heskes [6] draws on the results of Bedeaux *et al.* [7] that link (discrete time) random walk trajectories to the solution of a (continuous time) master equation. Heskes' master equation is equivalent to our Kramers-Moyal expansion (3) below.

Convergence (in mean square) to $\omega_*$ is characterized by the average squared norm of the weight error $E[\,|\,v\,|^2\,] =$ Trace $C$ where

$$C \;\equiv\; \int d^N v \;\; v\, v^T \; P(v,t) \tag{2}$$

is the weight error correlation matrix and $P(v,t)$ is the probability density at $v$ at time $t$. In [3] we show that the probability density evolves according to the Kramers-Moyal expansion

$$\frac{\partial P(v,t)}{\partial t} \;=$$

$$\sum_{i=1}^{\infty} \frac{(-1)^i}{i!} \sum_{j_1,\ldots j_i=1}^{N} \frac{\partial^i}{\partial v_{j_1}\,\partial v_{j_2}\ldots\partial v_{j_i}} \;\Big\{\, \langle\, \mu H_{j_1}\,\mu H_{j_2}\ldots\,\mu H_{j_i}\rangle_x \; P(v,t)\,\Big\}\;, \tag{3}$$

where $H_{j_k}$ denotes the $j_k^{th}$ component of the $N$-component vector $H$, and $\langle \ldots \rangle_x$ denotes averaging over the density of inputs. Differentiating (2) with respect to time, using (3) and integrating by parts, we obtain the equation of motion for the weight error correlation

$$\begin{aligned}\frac{d\,C}{dt} \;&=\; \mu(t) \int d^N v \;\; P(v,t)\,\big[\, v\;\big\langle\, H(v,x)^T\,\big\rangle_x \;+\; \big\langle\, H(v,x)\,\big\rangle_x \; v^T\,\big] \\ &\qquad +\; \mu(t)^2 \int d^N v \;\; P(v,t)\;\big\langle\, H(v,x)\,H(v,x)^T\,\big\rangle_x \quad .\end{aligned} \tag{4}$$

**Asymptotics of the Weight Error Correlation**

We wish to study the late time behavior of (4). Since the update function $H(v,x)$ is in general non-linear in $v$, the time evolution of the correlation matrix $C_{ij}$ is coupled to higher moments $E[\,v_i\,v_j\,v_k\ldots]$ of the weight error. However, the learning rate is assumed to follow a schedule $\mu(t)$ that satisfies the requirements for convergence in mean square to a local optimum. Thus at late times the density becomes sharply peaked about $v = 0$[2]. This suggests that we expand $H(v,x)$ in a power series about $v = 0$ and retain the lowest order non-trivial terms in (4) leaving:

$$\frac{dC}{dt} \;=\; -\,\mu(t)\,\big[\;(R\,C) \;+\; (C\,R^T)\;\big] \;+\; \mu(t)^2\;D \quad , \tag{5}$$

where $R$ is the Hessian of the average cost function $\langle\,\mathcal{E}\,\rangle_x$, and

$$D \;\equiv\; \big\langle\, H(0,x)\,H(0,x)^T\,\big\rangle_x \tag{6}$$

is the diffusion matrix, both evaluated at the local optimum $\omega_*$. (Note that $R^T = R$.) We use (5) with the understanding that it is valid for large $t$. The solution to (5) is

$$C(t) \;=\; U(t,t_0)\,C(t_0)\,U^T(t,t_0) \;+\; \int_{t_0}^{t} d\tau\;\mu(\tau)^2\;U(t,\tau)\;D\;U^T(t,\tau) \quad . \tag{7}$$

where the evolution operator $U(t_2,t_1)$ is

$$U(t_2,t_1) \;=\; \exp\left[\; -R\int_{t_1}^{t_2} d\tau\;\mu(\tau)\;\right] \quad . \tag{8}$$

[2]In general the density will have nonzero components outside the basin of $\omega_*$. We are neglecting these, for the purpose of calculating the second moment of the the *local* density in the vicinity of $\omega_*$.

We assume, without loss of generality, that the coordinates are chosen so that $R$ is diagonal ($D$ won't be) with eigenvalues $\lambda_i, \;\; i = 1 \ldots N$. Then with $\mu(t) = \mu_0/t$ we obtain

$$E\left[\,|\,v\,|^2\right] \;=\; \text{Trace}\left[\,C(t)\,\right] \;\;=\;\; \sum_{i=1}^{N} \left\{ C_{ii}\,(t_0) \;\left(\frac{t_0}{t}\right)^{2\,\mu_0\,\lambda_i} \right.$$

$$\left. +\; \frac{\mu_0^2\, D_{ii}}{(2\,\mu_0\,\lambda_i - 1)} \;\left[\; \frac{1}{t} \;-\; \frac{1}{t_0}\left(\frac{t_0}{t}\right)^{2\,\mu_0\,\lambda_i} \;\right] \right\} . \tag{9}$$

We define

$$\mu_{crit} \equiv \frac{1}{2\,\lambda_{min}} \tag{10}$$

and identify two regimes for which the behavior of (9) is fundamentally different:

1. $\mu_0 > \mu_{crit}$: $E\left[\,|v|^2\,\right]$ drops off asymptotically as $1/t$.
2. $\mu_0 < \mu_{crit}$: $E\left[\,|v|^2\,\right]$ drops off asymptotically as

$$\left(\frac{1}{t}\right)^{(2\,\mu_0\;\lambda_{min}\;)}$$

   i.e. *more slowly than* $1/t$.

Figure 1 shows results from simulations of an ensemble of 2000 networks trained by LMS, and the prediction from (9). For the simulations, input data were drawn from a gaussian with zero mean and variance $R = 1.0$. The targets were generated by a noisy teacher neuron. The upper two curves in each plot (dotted) depict the behavior for $\mu_0 < \mu_{crit} = 0.5$ . The remaining curves (solid) show the behavior for $\mu_0 > \mu_{crit}$.

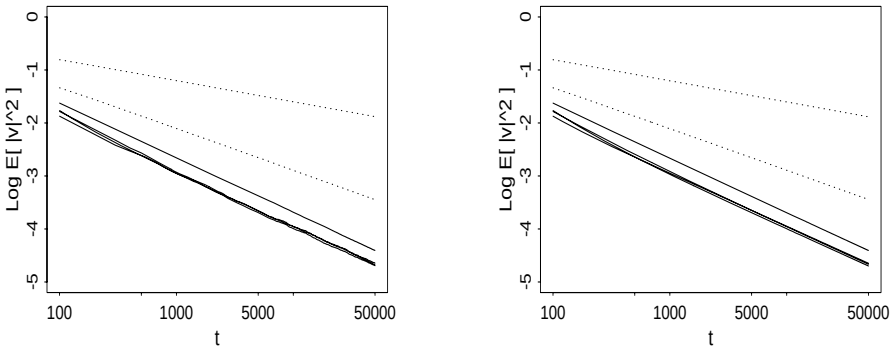


Fig.1: LEFT – Simulation results from an ensemble of 2000 one-dimensional LMS algorithms with $R = 1.0$, $D = 1.0$ and $\mu = \mu_0/t$. RIGHT – Theoretical predictions from equation (9). Curves correspond to (top to bottom) $\mu_0 = 0.2, 0.4, 0.6, 0.8, 1.0, 1.5$ .

**Asymptotic Normality**

This formalism yields asymptotic normality rather simply. At late times, $\mu(t)$ becomes very small and one can truncate the Kramers-Moyal expansion (3) to second order in $\mu$, leaving a Fokker-Planck equation. As the density becomes peaked up about $v = 0$, the linear part of the drift and the constant part of the diffusion dominate and the dynamics is governed by

$$\frac{\partial P(v,t)}{\partial t} \;=\; \mu(t)\, R_{ij}\, \frac{\partial}{\partial v_i}\, \left(\, v_j\; P(v,t)\,\right) \;+\; \frac{1}{2}\,\mu(t)^2\, D_{ij}\, \frac{\partial^2 P(v,t)}{\partial v_i\, \partial v_j} \tag{11}$$

where repeated indices are summed over[3]. Next, we define a time-dependent coordinate transformation

$$y_i \;=\; h_{ij}(t)\; v_j \tag{12}$$

$$P(v,t) \;=\; \tilde{P}(y,t)\; h \;\;, \tag{13}$$

where $h = \mathrm{Det}\,[h_{ij}(t)]$. The Fokker-Planck equation, expressed in terms of the $y$ variables, reads

$$\begin{aligned}\frac{\partial \tilde{P}}{\partial t} \;&=\; \frac{\partial}{\partial y_j}\, \left\{\, \left[\, \mu(t)\, h_{ji}\, R_{il}\, h^{-1}_{lk} \;-\; h^{'}_{jl}\, h^{-1}_{lk}\,\right]\;\; y_k\, \tilde{P}\,\right\} \\ &\quad +\; \frac{\mu(t)^2}{2}\, D_{ij}\, h_{ki}\, h_{lj}\; \frac{\partial^2}{\partial y_k\, \partial y_l}\; \tilde{P} \;\;\;. \end{aligned} \tag{14}$$

Depending on $\mu_0$ and $R$, different choices for $h_{ij}$ are required to obtain stationary solutions. Summarizing the results (in one dimension for simplicity)

1. $\mu_0 > \mu_{crit}$:
   $y \;=\; v\,\sqrt{t}$ is asymptotically normal with variance $\mu_0^2\; D/(2\mu_0 R - 1)$.

2. $\mu_0 = \mu_{crit}$:
   $y \;=\; v\,\sqrt{t/\mathrm{Ln}(t)}$ is asymptotically normal with variance $\mu_0^2\; D$.

3. $\mu_0 < \mu_{crit}$:
   Using the new variable $y \;=\; v\, t^{(\mu_0\, R)}$ the drift term in the Fokker-Planck equation (14) vanishes, leaving a pure diffusion process. The diffusion coefficient goes to zero as $t^{2(\mu_0\, R-1)}$. Thus the density $\tilde{P}(y)$ freezes at late times and we expect $y$ to approach some non-trivial random variable. Though not rigorous, this argument is consistent with the results of Major and Revesz [9, and references therein].

The conditions for "optimal" (i.e. $1/t$) convergence of the weight error correlation, and the related results on asymptotic normality were previously discussed in the stochastic approximation literature [10, 11, and references therein]. The present formal structure provides the results with relative ease and facilitates the extension to stochastic gradient descent with momentum.

## STOCHASTIC SEARCH WITH MOMENTUM

The discrete time algorithm for stochastic optimization with momentum is:

$$v(t+1) \;=\; v(t) + \mu(t)\; H[v(t), x(t)] + \beta\; \mathbf{\Omega}(t) \tag{15}$$

$$\begin{aligned}\mathbf{\Omega}(t+1) \;&\equiv\; v(t+1) - v(t) \\ &=\; \mathbf{\Omega}(t) + \mu(t)\; H[v(t), x(t)] + (\beta - 1)\; \mathbf{\Omega}(t),\end{aligned} \tag{16}$$

[3]Equation (11) can be derived rigorously by writing the weight error as the sum of deterministic and stochastic contributions $v \;=\; \phi + \sqrt{\mu}\,\xi$ and using these coordinates in the Kramers-Moyal expansion (3). This prescription corresponds to Van Kampen's system size expansion [6, 8]. The deterministic piece $\phi$ evolves by descent along the average or true gradient, approaching zero exponentially. The lowest order (in $\mu$) piece of the Kramers-Moyal expansion for the density of the fluctuations $P(\xi, t)$ is equivalent to (11).

or in continuous time,

$$\frac{dv(t)}{dt} = \mu(t)\ H[v(t), x(t)] + \beta\ \mathbf{\Omega}(t) \tag{17}$$

$$\frac{d\mathbf{\Omega}(t)}{dt} = \mu(t)\ H[v(t), x(t)] + (\beta - 1)\ \mathbf{\Omega}(t). \tag{18}$$

**Weight Error Correlation**

As before, we are interested in the late time behavior of $E\,[\,|v|^2\,]$. To this end, we define the $2N$-dimensional variable $Z \equiv (v, \mathbf{\Omega})^T$ and, following the arguments of the previous sections, expand $H[v(t), x(t)]$ in a power series about $v = 0$ retaining the linear part of the drift, and the constant part of the diffusion matrix. In this approximation the correlation matrix $\overline{C} \equiv E[ZZ^T]$ evolves according to

$$\frac{d\overline{C}}{dt} = K\overline{C} + \overline{C}K^T + \mu(t)^2\ \overline{D} \tag{19}$$

with

$$K \equiv \begin{pmatrix} -\mu(t)\ R & \beta I \\ -\mu(t)R & (\beta - 1)I \end{pmatrix}, \quad \overline{D} \equiv \begin{pmatrix} D & D \\ D & D \end{pmatrix}, \tag{20}$$

$I$ is the $N \times N$ identity matrix, and $R$ and $D$ are defined as before.

The evolution operator is now

$$\overline{U}(t_2, t_1) \equiv \exp\left[\int_{t_1}^{t_2} d\tau\ K(\tau)\right] \tag{21}$$

and the solution to (19) is

$$\overline{C} = \overline{U}(t, t_0)\ \overline{C}(t_0)\ \overline{U}^T(t, t_0) + \int_{t_0}^{t} d\tau\ \mu^2(\tau)\ \overline{U}(t, \tau)\ \overline{D}\ \overline{U}^T(t, \tau) \tag{22}$$

The squared norm of the weight error is the sum of first $N$ diagonal elements of $\overline{C}$. In coordinates for which $R$ is diagonal and with $\mu(t) = \mu_0/t$, we find that for $t \gg t_0$

$$E[|v|^2] \approx \sum_{i=1}^{N} \left\{ \overline{C}_{ii}(t_0)\ \left(\frac{t_0}{t}\right)^{\frac{2\mu_0\lambda_i}{1-\beta}} + \frac{\mu_0^2\ D_{ii}}{(1-\beta)\ (2\mu_0\lambda_i - 1 + \beta)} \left(\frac{1}{t} - \frac{1}{t_0}\left(\frac{t_0}{t}\right)^{\frac{2\mu_0\lambda_i}{1-\beta}}\right) \right\}. \tag{23}$$

This reduces to (9) when $\beta = 0$. Equation (23) defines two regimes of interest:

1. $\mu_0/(1-\beta) > \mu_{crit}$: $E[|v|^2]$ drops off asymptotically as $1/t$.
2. $\mu_0/(1-\beta) < \mu_{crit}$: $E[|v|^2]$ drops off asymptotically as

$$\left(\frac{1}{t}\right)^{\frac{2\mu_0\lambda_{min}}{1-\beta}},$$

   i.e. *more slowly than* $1/t$.

The form of (23) and the conditions following it show that the asymptotics of gradient descent with momentum are governed by the *effective learning rate*

$$\mu_{eff} \equiv \frac{\mu}{1-\beta}.$$

Figure 2 compares simulations with the predictions of (23) for fixed $\mu_0$ and various $\beta$. The simulations were performed on an ensemble of 2000 networks trained by LMS as described previously but with an additional momentum

term of the form given in (15). The upper three curves (dotted) show the behavior of $E[|v|^2]$ for $\mu_{eff} < \mu_{crit}$. The solid curves show the behavior for $\mu_{eff} > \mu_{crit}$.

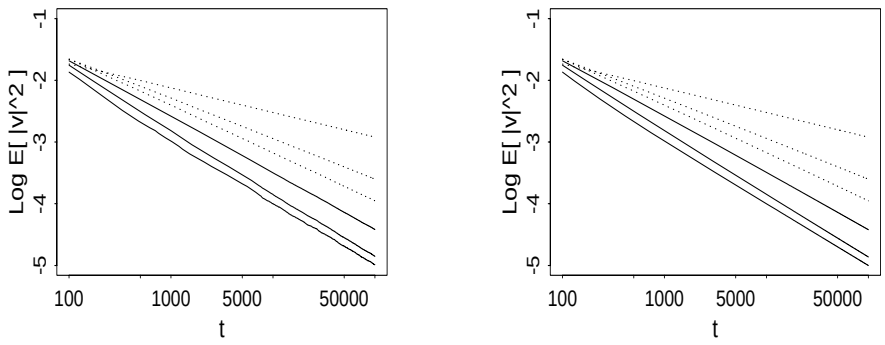


Fig.2: LEFT – Simulation results from an ensemble of 2000 one-dimensional LMS algorithms with momentum with $R = 1.0$, $D = 1.0$, and $\mu_0 = 0.2$. RIGHT – Theoretical predictions from equation (23). Curves correspond to (top to bottom) $\beta = 0.0, 0.4, 0.5, 0.6, 0.7, 0.8$ .

**Asymptotic Normality**

The derivation of asymptotic normality proceeds similarly to the case without momentum. One first makes a (time-dependent) transformation into coordinates that diagonalize the drift $K$. The coefficients of the resulting Fokker-Planck equation are expanded in powers of $(1/t)$ and terms dominant at large $t$ are retained. One finds that if

$$\mu_{eff_0} \equiv \frac{\mu_0}{1-\beta} > \frac{1}{2\lambda_{min}}$$

then $\sqrt{t}\, v$ is asymptotically normal with variance

$$\frac{{\mu_{eff_0}}^2\, D}{2R\,\mu_{eff_0} - 1} \quad .$$

## DISCUSSION

We have used the dynamics of the weight space probabilities to derive the asymptotic behavior of densities and weight error correlation for annealed stochastic gradient algorithms with momentum. The late time behavior is governed by the *effective* learning rate $\mu_{eff} \equiv \mu/(1-\beta)$. For learning rate schedules $\mu_0/t$, if $\mu_{0\,eff} > 1/(2\,\lambda_{min})$, then the squared norm of the weight error $v \equiv \omega - \omega_*$ falls off as $1/t$ and $\sqrt{t}\, v$ is asymptotically normal.

**Acknowledgments** This work was supported by grants from the Air Force Office of Scientific Research (F49620-93-1-0253) and the Electric Power Research Institute (RP8015-2).

**References**

[1] Mehmet Ali Tugay and Yalcin Tanik. Properties of the momentum LMS algorithm. *Signal Processing*, 18:117–127, 1989.

[2] John J. Shynk and Sumit Roy. The LMS algorithm with momentum updating. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, pages 2651–2654. IEEE, 1988.

[3] Todd K. Leen and John E. Moody. Weight space probability densities in stochastic learning: I. Dynamics and equilibria. In Giles, Hanson, and Cowan, editors, *Advances in Neural Information Processing Systems, vol. 5*, San Mateo, CA, 1993. Morgan Kaufmann.

[4] Genevieve B. Orr and Todd K. Leen. Weight space probability densities in stochastic learning: II. Transients and basin hopping times. In Giles, Hanson, and Cowan, editors, *Advances in Neural Information Processing Systems, vol. 5*, San Mateo, CA, 1993. Morgan Kaufmann.

[5] H.J. Kushner and D.S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York, 1978.

[6] Tom M. Heskes, Eddy T.P. Slijpen, and Bert Kappen. Learning in neural networks with local minima. *Physical Review A*, 46(8):5221–5231, 1992.

[7] D. Bedeaux, K. Laktos-Lindenberg, and K. Shuler. On the relation between master equations and random walks and their solutions. *Journal of Mathematical Physics*, 12:2116–2123, 1971.

[8] C.W. Gardiner. *Handbook of Stochastic Methods, 2nd Ed.* Springer-Verlag, Berlin, 1990.

[9] Larry Goldstein. Mean square optimality in the continuous time Robbins Monro procedure. Technical Report DRB-306, Dept. of Mathematics, University of Southern California, LA, 1987.

[10] Christian Darken and John Moody. Towards faster stochastic gradient search. In J.E. Moody, S.J. Hanson, and R.P. Lipmann, editors, *Advances in Neural Information Processing Systems 4*. Morgan Kaufmann Publishers, San Mateo, CA, 1992.

[11] Halbert White. Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1:425–464, 1989.