Stochastic Manhattan learning: Time-evolution operator for the ensemble dynamics

Todd K. Leen* and John E. Moody†
Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, P.O. Box 91000, Portland, Oregon 97291-1000
(Received 6 January 1997)

Typical theoretical descriptions of the ensemble dynamics of stochastic learning algorithms rely on a truncated expansion to approximate the time-evolution operator appearing in the master equation. In this paper we give an exact expression for the time-evolution operator for Manhattan learning, a variant of stochastic gradient-descent learning in which the weights are updated in proportion to the *sign* of the cost function gradient. This closed form for the time evolution captures the full nonlinearity of the problem without approximation, allowing exact study of the ensemble dynamics. [S1063-651X(97)07207-3]

PACS number(s): 87.10.+e, 02.50.−r, 05.40.+j, 07.95.Mh

ENSEMBLE DYNAMICS OF STOCHASTIC LEARNING

Stochastic learning algorithms provide recursively refined estimates of optimal model parameters in machine learning, neural networks, adaptive signal processing, and control. Most algorithms are of the form

$$w(n+1)=w(n)+\mu(n)H(w(n),x(n)), \tag{1}$$

where $w(n)\in\mathbb{R}^N$ (with components denoted w_i) is the parameter estimate at the nth iteration of the recursion, $\mu(n)$ is called the learning rate, H embodies the learning rule, and $x(n)\in\mathbb{R}^M$ is the datum input to the algorithm at the nth iteration. In supervised learning (e.g., regression or classification), this datum is an input-target pair $\{\xi(n),t(n)\}$. In unsupervised learning (e.g., clustering), there is no target. The sequence of inputs $\{x(1),x(2),\dots\}$ results from sampling from an empirical distribution of training data. In most theoretical treatments, and as considered here, this sampling is independent and identically distributed. Through an initial choice $w(0)$ and recursive application of Eq. (1), the sequence of inputs generates a sequence of parameter estimates $\{w(0),w(1),\dots\}$. This sequence of parameter estimates is a Markov chain whose probability law is specified by the master equation

$$P(w,n+1)-P(w,n)=\int dw'[P(w',n)W(w|w') - P(w,n)W(w'|w)], \tag{2}$$

where $W(w|w')$ is the single time-step transition probability

$$W(w|w')=\langle\delta(w-w'-\mu H(w',x))\rangle_x \tag{3}$$

and $\langle\ \rangle_x$ indicates the ensemble average with respect to x.

The usual theoretical approach to Eq. (2) involves approximating the integrals in the time-evolution operator. Expanding the transition probability (3) as a power series in μ leaves the Kramers-Moyal expansion [1,2]. In a region close to the optimal parameters [zeros of $\langle H(w,x)\rangle_x$ with negative definite Jacobian $\langle DH(w,x)\rangle_x$] and for small learning rate, the dynamics induced by the Kramers-Moyal expansion can be described by the diffusion approximation to a small noise expansion [1].

Global phenomena, e.g., transitions between basins of different optimal parameter values, have been treated with some success [3–5]. However, for appreciable learning rates, the theory is encumbered by the contribution of jump moments beyond the drift and diffusion coefficients. The lack of a solvable model has left theoretical expositions reliant entirely on low-order expansions.

We give an example of an algorithm of practical interest for which the integrals in the master equation can be solved in closed form. (Equivalently, the Kramers-Moyal expansion can be summed exactly.) This obviates the need for approximating truncations, allowing an exact treatment of the dynamics.

MANHATTAN LEARNING

Many supervised and unsupervised learning algorithms are cast as function minimization tasks. One writes a cost function $E(w)=\langle E(x,w)\rangle_x$ that is a functional of the parametrized map being learned $f(x;w)$ and the distribution of data x. Learning corresponds to adjusting the parameters w so as to minimize $E(w)$.

Commonly, the learning algorithm is derived as a gradient descent on $E(w)$. In on-line or stochastic learning, one performs gradient descent on the *instantaneous* cost $E(w,x)$ rather than on the average cost. The corresponding learning equations are Eq. (1) with

$$H(w,x)=-\nabla_w E(w,x). \tag{4}$$

Such algorithms are known as stochastic gradient descent algorithms. They are a specific instance of stochastic approximation procedures. Unlike batch gradient descent, which uses the gradient of the average cost $E(w)=\langle E(w,x)\rangle_x$, the stochastic algorithm uses a noisy estimate of the gradient. The noise can help avoid trapping in poor local optimum for nonlinear optimization problems. The stochastic algorithms also offer a speed advantage for large, redundant datasets. Since parameter updates are based on a single datum, the stochastic algorithm avoids computing the average over x required for the batch algorithm.

*Electronic address: tleen@cse.ogi. edu
†Electronic address: moody@cse.ogi.edu

Gradient descent algorithms suffer from a number of difficulties: Convergence is terribly slow when the condition number of the Hessian of $E(w)$ is large and progress along the cost surface is slow where the gradients are small. To alleviate the latter, several researchers (see [6,7] and references therein) have suggested Manhattan learning, where the weights are updated in proportion to the *sign* of the components of the gradient

$$H_i(w,x) = -\mathrm{sgn}\left[\frac{\partial E(w,x)}{\partial w_i}\right]. \tag{5}$$

Peterson and Hartman [7] found Manhattan learning to be beneficial, presuming that its advantage lies in the fact that its weight changes are fixed and bounded. Manhattan learning prevents the stallout observed in gradient descent where the cost function has nearly flat plateaus.

EVOLUTION OF THE PROBABILITY DENSITY

With parameter updates based on the sign of the gradient components, $H_i(w,x)$ is limited to $0,\pm 1$ and is piecewise constant on x. Consequently, the averages required to calculate the transition probability (3) can be completed in closed form. For clarity of exposition, we first derive the master equation for a one-dimensional parameter space $w \in R^1$ and then give the results for the multidimensional case.

To begin, we partition the data space into three disjoint regions

$$\begin{aligned} S_+(w) &= \{x | H(w,x) = +1\}, \\ S_0(w) &= \{x | H(w,x) = 0\}, \\ S_-(w) &= \{x | H(w,x) = -1\}. \end{aligned} \tag{6}$$

In terms of this partition, the average in the single time-step transition probability (3) becomes

$$\begin{aligned} W(w|w') &= \langle \delta(w-w'-\mu)\rangle_{S_+(w')} + \langle \delta(w-w')\rangle_{S_0(w')} \\ &\quad + \langle \delta(w-w'+\mu)\rangle_{S_-(w')} \\ &\equiv F_+(w')\,\delta(w-w'-\mu) + F_0(w')\,\delta(w-w') \\ &\quad + F_-(w')\,\delta(w-w'+\mu), \end{aligned} \tag{7}$$

where F_+, F_0, and F_- are the measure of x on the sets S_+, S_0, and S_-, respectively. Finally, the measures F_+, F_0, and F_- can be rewritten in terms of the first two jump moments (normalized by the step size μ)

$$\begin{aligned} D^{(1)}(w) &= \langle H(w,x)\rangle_x = F_+(w) - F_-(w), \\ D^{(2)}(w) &= \langle H^2(w,x)\rangle_x = F_+(w) + F_-(w). \end{aligned} \tag{8}$$

Solving Eqs. (8) for the F's in terms of the jump moments and substituting the resulting transition probabilities (7) into the master equation (2) leaves our closed form for the evolution of the probability density

$$\begin{aligned} &P(w,n+1) - P(w,n) \\ &\quad = -\frac{1}{2}[D^{(1)}(w+\mu)P(w+\mu,n) \\ &\qquad - D^{(1)}(w-\mu)P(w-\mu,n)] \\ &\qquad + \frac{1}{2}[D^{(2)}(w+\mu)P(w+\mu,n) - 2D^{(2)}(w)P(w,n) \\ &\qquad + D^{(2)}(w-\mu)P(w-\mu,n)]. \end{aligned} \tag{9}$$

Equation (9) is an exact expression for the time evolution of the probability density $P(w,n)$ that involves only the first two jump moments (8). The higher-order jump moments are explicitly absent because they are all trivially related to the first two. The first jump moment $D^{(1)}(w)$ is just the average of the update function and is the same update used in the batch version of the algorithm. The second moment $D^{(2)}(w)$ is zero at those w for which the update function is zero *for every possible pattern* x, i.e., it is zero for weights that perfectly solve the problem. At all other weights, the second jump moment lies in the range $0 < D^{(2)}(w) \leqslant 1$.

Curiously, the right-hand side of Eq. (9) is just a finite-difference approximation to the Fokker-Planck equation on a grid with spacing equal to the stepsize μ. This is reasonable since the learning rule restricts parameter changes to $\pm\mu$. However, we stress that Eq. (9) describes the *complete* dynamics and in no way represents a diffusion approximation to the master equation.

Several features of the dynamics are illuminated by inspection of Eq. (1) with Eq. (5) and the ensemble behavior (9). First, if the ensemble is initialized such that $P(w,0)$ has support only on the grid $w = \pm i\mu$, $i = 0,1,\ldots,$ then $P(w,n)$ will have support on this grid for all n. The dynamics are then easily developed by matrix multiplication. The evolution matrix is sparse since Eq. (9) involves only nearest-neighbor interactions on the grid. Indeed, for a one-dimensional parameter space, the evolution matrix is tridiagonal. With these simple dynamics, equilibria can be calculated by finding the null space of a matrix and first-passage time calculations reduce to the solution of a linear system.

Our second observation concerns absorbing states. For some problems, e.g., supervised learning problems with zero-error solutions, there exist parameter vectors $w_*^{[i]}$, $i = 1,\ldots q$, such that $\nabla E(w_*^{[i]},x) = 0 \ \forall x$. Then the master equation has equilibria consisting of δ functions at the w_*,

$$P_0(w) = \sum_{i=1}^{q} a_i \delta(w - w_*^{[i]}),$$

where q is the number of distinct absorbing states. The a_i are dependent on the initial distribution $P(w,0)$. For learning driven by the sign of the gradient, these solutions are only accessible if the initial distribution $P(w,0)$ is restricted to points such that $w - w_{*i}$ are integer multiples of μ, that is, all the initial density and all of the w_* must lie on a grid of size μ. If this is not the case, then at late times $P(w,n)$ contains peaks that execute oscillations about these absorbing states.

A type of oscillatory behavior can also occur when an occupied state has unoccupied nearest neighbors, for example, if the initial density is confined to a *single* grid point. Suppose w_0, not an absorbing state, is occupied at time n and the adjacent states $w_0 \pm \mu$ are unoccupied, that is, $P(w_0,n) \neq 0$ but $P(w_0 \pm \mu, n) = 0$. Then by Eq. (9)

$$P(w_0,n+1) = P(w_0,n)[1 - D^{(2)}(w_0)].$$

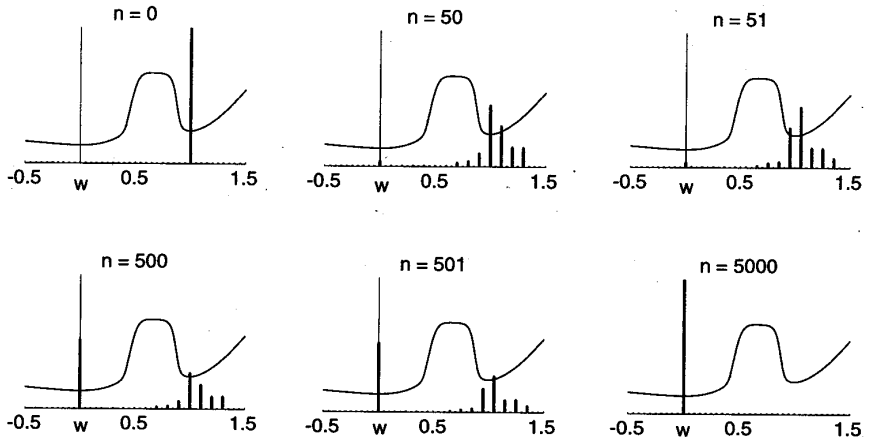

FIG. 1. Cost function $E(w)$ (curve) overlaid with $P(w,n)$ (histogram) on a grid with $\mu=0.05$. There is an absorbing state at $w=0$ and a local minimum at $w=1$. The ensemble is initialized in the local minimum (top left frame) and the evolution of $P(w,n)$ is portrayed at 50, 51, 500, 501, and 5000 iterations.

By assumption, w_0 is not an absorbing state, so $H^2(w_0,x)$ is nonzero for some of the inputs x. Hence by Eqs. (5) and (8), $0<D^{(2)}(w_0)\leqslant 1$. When $D^{(2)}(w_0)<1$, some of the density transitions to the states $w_0\pm\mu$ at $n+1$. In general, some of this density will transition back to w_0 at $n+2$. When $H^2(w_0,x)$ is nonzero for *all* of the inputs x, then $D^{(2)}(w_0)=1$ and all of the density leaves w_0 at time $n+1$. When this condition is met over a set of adjacent states, the density switches between even- and odd-numbered grid points on successive iterations. (See the example in Fig. 1.)

Figure 1 depicts the evolution of $P(w,n)$ overlaid on the cost function for a problem with a global minimum at the origin and a local minimum at $w=1.0$. The global minimum is an absorbing state. The grid, of spacing $\mu=0.05$, contains both the local and global minima. The ensemble is initialized with all the density at the local minimum. The oscillatory behavior is evident by comparing frames $n=50$ with $n=51$ and $n=500$ with $n=501$.

In most problems, absorbing states are not present. To obtain convergence (in mean square or with probability one) one employs learning rate annealing in which the step size $\mu(n)$ is decayed according to some fixed schedule [8] or using an adaptive scheme to obtain more rapid convergence [9]. In Manhattan learning, annealing corresponds to shrinking the grid size.

MULTIDIMENSIONAL CASE

For the one-dimensional algorithm, we were able to express the averages in the transition probability (3) in a finite number of terms because the update function $H(w,x)$ is piecewise constant in x. This holds for the multidimensional case as well, and stands in sharp contrast to stochastic gradient descent where the averages lead to the infinite Kramers-Moyal expansion.

For the multidimensional case, we consider two forms of stochastic Manhattan learning. In the first form, at each iteration *a single weight* is chosen at random and only the chosen weight is updated. In this case, like the one-dimensional algorithm, the evolution is expressed in terms of the first two jump moments. In the second form of the algorithm, *all* weights are updated at each iteration. Here the evolution cannot be expressed in terms of the first two jump moments. Nevertheless the averages are computed in a finite number of terms.

SINGLE WEIGHT UPDATE

Let $w\in R^N$ with components w_j. At each learning iteration, we choose one input x and one component w_j to increment in proportion to the sign of the corresponding gradient component. We explicate the random choice of component by introducing indicator variables $\xi_j\in\{0,1\}$ equal to 1 with probability $1/N$. (Only a single ξ_j is nonzero at each iteration.) Then

$$H_j(w,x,\xi)=-\xi_j\,\mathrm{sgn}\left[\frac{\partial E(w,x)}{\partial w_j}\right]. \tag{10}$$

The transition probability (3) now contains an average over the indicator variables, explicitly,

$$\begin{aligned}
W(w|w') &= \sum_{\xi} P(\xi)\langle \delta(w-w'-\mu H(w',x,\xi))\rangle_x \\
&= \sum_{\xi} P(\xi)\left\langle \prod_{i=1}^{N} \delta\left(w_i-w_i' + \mu\xi_i\,\mathrm{sgn}\,\frac{\partial E(w',x)}{\partial w_i}\right)\right\rangle_x \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\prod_{j\neq i}\delta(w_j-w_j')\right) \\
&\quad\times\left\langle \delta\left(w_i-w_i'+\mu\,\mathrm{sgn}\,\frac{\partial E(w',x)}{\partial w_i}\right)\right\rangle_x \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\prod_{j\neq i}\delta(w_j-w_j')\right)[F_{i+}(w') \\
&\quad\times\delta(w_i-w_i'-\mu) \\
&\quad+F_{i0}(w')\,\delta(w_i-w_i') \\
&\quad+F_{i-}(w')\,\delta(w_i-w_i'+\mu)],
\end{aligned} \tag{11}$$

where the F are defined in obvious analogy to those in Eq. (7).

Substituting the transition probability (11) into the master equation, one obtains an expression for the evolution of the density in terms of the F. As in the one-dimensional case, the latter can be written in terms of the first two jump moments

$$\begin{aligned} D_j^{(1)}(w) &= N\langle \xi_j H_j(w,x)\rangle_{x,\xi} \\ &= \langle H_j(w,x)\rangle_x = F_{i+}(w) - F_{i-}(w), \\ D_{jk}^{(2)}(w) &= N\langle \xi_j \xi_k H_j H_k\rangle_{x,\xi} = \delta_{jk}\langle H_j^2(w,x)\rangle_x \\ &= [F_{i+}(w) + F_{i-}(w)]\delta_{jk}. \end{aligned} \tag{12}$$

Using these expressions and the fact that $F_{i-} + F_{i0} + F_{i+} = 1$, the evolution of the density can be written as

$$\begin{aligned} &P(w,n+1) - P(w,n) \\ &= -\frac{1}{2N}\sum_{j=1}^{N} [D_j^{(1)}(w+\mu_j)P(w+\mu_j,n) \\ &\quad - D_j^{(1)}(w-\mu_j)P(w-\mu_j,n)] \\ &\quad + \frac{1}{2N}\sum_{j=1}^{N} [D_{jj}^{(2)}(w+\mu_j)P(w+\mu_j,n) \\ &\quad - 2D_{jj}^{(2)}(w)P(w,n) + D_{jj}^{(2)}(w-\mu_j)P(w-\mu_j,n)], \end{aligned} \tag{13}$$

where μ_j is an increment of length μ along w_j. As in the one-dimensional case, this is of the form of a finite-difference approximation to a Fokker-Planck equation, but carries the full dynamics implicit in the original master equation.

UPDATING ALL WEIGHTS

In this case, all the indicator variables ξ_i are equal to one at every iteration. The transition probability becomes

$$W(w|w') = \left\langle \prod_{i=1}^{N} \delta\left(w_i - w_i' + \mu \ \mathrm{sgn}\frac{\partial E(w',x)}{\partial w_i}\right)\right\rangle_x . \tag{14}$$

It is convenient to define the quantity

$$\gamma_i \equiv -H_i(w',x) \equiv \mathrm{sgn}\frac{\partial E(w',x)}{\partial w}.$$

Since each component of γ has three possible values $0, \pm 1$, there are 3^N possible vectors γ. Next define $F_\gamma(w')$ to be the measure of x corresponding to the specific vector γ when the system is in state w'. With these definitions, the transition probability is

$$W(w|w') = \sum_\gamma F_\gamma(w') \prod_{i=1}^{N} \delta(w_i - w_i' + \mu\gamma_i) \tag{15}$$

and the density evolves according to

$$\begin{aligned} P(w,n+1) &= \int dw' W(w|w')P(w',n) \\ &= \sum_\gamma F_\gamma(w+\mu\gamma)P(w+\mu\gamma,n). \end{aligned} \tag{16}$$

In general, the evolution in Eq. (16) cannot be rewritten in terms of the first two jump moments. However, unlike the Kramers-Moyal expansion, only a finite number (3^N) of terms are required, though the computation rapidly becomes cumbersome for large N. (The F can be written in terms of the jump moments, but will involve higher-order moments. For example, for $w \in R^2$, there are nine distinct F. It is straightforward to verify that these can be rewritten in terms of the jump moments $D_1^{(1)}$, $D_2^{(1)}$, $D_{11}^{(2)}$, $D_{22}^{(2)}$, $D_{12}^{(2)}$, $D_{112}^{(3)}$, $D_{122}^{(3)}$, and $D_{1122}^{(4)}$.)

SUMMARY

Stochastic learning algorithms are ubiquitous in the machine learning literature, with stochastic gradient-descent methods the most commonly used in practice. Previous theoretical treatments of the ensemble weight space dynamics based on the master equation have relied entirely on low-order approximations, usually diffusion equations. If the usual gradient descent is modified so that parameter updates are based on the *sign* of the gradient, then the dynamics can be developed without approximation.

For Manhattan learning, we are able to write the dynamics of the density without approximation because the averages appearing in the single step transition probability (3) can be evaluated in closed form. This is possible because the update function $H(w,x)$ is piecewise constant on x. This condition restricts the class of algorithms for which one will be able to evaluate the ensemble dynamics without approximation.

[1] C.W. Gardiner, *Handbook of Stochastic Methods*, 2nd ed. (Springer-Verlag, Berlin, 1990).

[2] Todd K. Leen and John E. Moody, *Advances in Neural Information Processing Systems, Vol. 5*, edited by S. J. Hanson, J. D. Cowan, and C. L. Giles (Kaufmann, San Mateo, CA, 1993), p. 451.

[3] G. Radons, H.G. Schuster, and D. Werner, in *Processing in Neural Systems and Computers*, edited by R. Eckmiller, G. Hartmann, and G. Hausek (Elsevier, Amsterdam, 1990).

[4] Genevieve B. Orr and Todd K. Leen, *Advances in Neural Information Processing Systems, Vol. 5* (Ref. [2]), p. 507.

[5] Tom M. Heskes, Eddy T. P. Slijpen, and Bert Kappen, Phys. Rev. A **46**, 5221 (1992).

[6] Martin Riedmiller and H. Braun, in *Proceedings of the IEEE International Conference on Neural Networks*, edited by H. Ruspini (IEEE, San Francisco, 1993).

[7] Carsten Peterson and Eric Hartman, Neural Networks **2**, 475 (1989).

[8] L. Ljung, IEEE Trans. Autom. Control. **22**, 551 (1977).

[9] Christian Darken and John Moody, in *Advances in Neural Information Processing Systems, Vol. 4*, edited by J. E. Moody, S. J. Hanson, and R. P. Lipmann (Kaufmann, San Mateo, CA, 1992).