Stochastic Gradient Descent in Continuous Time

Justin Sirignano
University of Illinois at Urbana Champaign

with Konstantinos Spiliopoulos (Boston University)
We consider a diffusion $X_t \in \mathcal{X} = \mathbb{R}^m$:

$$dX_t = f^*(X_t) dt + \sigma dW_t.$$ 

- The goal is to statistically estimate a model $f(x, \theta)$ for $f^*(x)$ where $\theta \in \mathbb{R}^n$.
- $f(x, \theta)$ and $f^*(x)$ may be non-convex
- $W_t$ is a standard Brownian motion.
- The diffusion term $W_t$ represents any random behavior of the system or environment.
The parameter update satisfies the SDE:

\[ d\theta_t = \alpha_t \left[ \nabla_{\theta} f(X_t; \theta_t) (\sigma \sigma^T)^{-1} dX_t - \nabla_{\theta} f(X_t, \theta_t) (\sigma \sigma^T)^{-1} f(X_t, \theta_t) dt \right] \]

- \( \alpha_t \) is the learning rate
- Can be used for both:
  - Statistical estimation given previously observed data
  - Online learning (i.e., statistical estimation in real-time as data becomes available)
- If \( m = 1 \) and \( \sigma = 1 \):

\[ d\theta_t = \alpha_t \left[ \nabla_{\theta} f(X_t; \theta_t) dX_t - \nabla_{\theta} f(X_t, \theta_t) f(X_t, \theta_t) dt \right] \]
Why is Stochastic Gradient Descent in Continuous Time useful?

- Physics and engineering models are typically in continuous time. It therefore makes sense to also develop the statistical learning updates in continuous time.

- Continuous-time dynamics are oftentimes much simpler than discrete dynamics at longer time intervals.
Although stochastic gradient descent in continuous time must ultimately be discretized for numerical implementation, the continuous-time framework still has significant numerical advantages.

Continuous-time stochastic gradient descent allows for the control and reduction of numerical error due to discretization.

Example 1: Higher-order numerical schemes for numerical solution of SDE

Example 2: Non-uniform time step sizes. If convergence is slow, the time step size may be adaptively decreased.

In contrast, discrete-time stochastic gradient descent uses fixed discrete steps and cannot do this.
Assume $X_t$ is ergodic and has a unique invariant measure $\pi(dx)$.

Define:

$$\tilde{h}(\theta) = \int_{X} h(x, \theta)\pi(dx)$$

Define the natural objective function:

$$g(x, \theta) = \frac{1}{2} \|f(x, \theta) - f^*(x)\|_{\sigma\sigma^T}^2$$

We show that

$$\lim_{t \to \infty} \|\nabla \tilde{g}(\theta_t)\| = 0, \text{ almost surely.}$$
Assumptions

- Assume that $\int_0^\infty \alpha_t dt = \infty$, $\int_0^\infty \alpha_t^2 dt < \infty$ and that $\int_0^\infty |\alpha'_s| ds < \infty$.

- The condition $\int_0^\infty |\alpha'_s| ds < \infty$ follows immediately from the other two restrictions for the learning rate if it is chosen to be a monotonic function of $t$.

- A standard choice is $\alpha_t = \frac{1}{C + t}$ for some constant $0 < C < \infty$.

- Polynomial bounds on $g$ and $f$ is Lipschitz (see our paper for details)
Extensive research on stochastic gradient descent in discrete time.

Relatively little research for continuous time.


The $X$ term introduces correlation across times, and this correlation does not disappear as $t \to \infty$.

Unlike in Bertsekas and Tsitsiklis (2000) where parameter updates are unbiased and noise is i.i.d., the $X$ process causes parameter updates to be biased and correlated across times. This complicates the analysis.
“ODE method”: proves discrete-time stochastic gradient descent converges to the solution of an ODE which itself converges to a limiting point, Kushner and Yin (2003), Benveniste, Metivier and Priouret (2012)

Requires the strong assumption that the iterates (i.e., the model parameters which are being learned) remain in a bounded set with probability one.

Proving that the iterates remain in a bounded set with probability one can be challenging to show and, moreover, may not necessarily be true for all models.
Proof Approach

Consider the cycles of random times

\[ 0 = \sigma_0 \leq \tau_1 \leq \sigma_1 \leq \tau_2 \leq \sigma_2 \leq \ldots \]

where for \( k = 1, 2, \ldots \)

\[ \tau_k = \inf\{ t > \sigma_{k-1} : \| \nabla \bar{g}(\theta_t) \| \geq \kappa \} \]

\[ \sigma_k = \sup\{ t > \tau_k : \frac{\| \nabla \bar{g}(\theta_{\tau_k}) \|}{2} \leq \| \nabla \bar{g}(\theta_s) \| \leq 2 \| \nabla \bar{g}(\theta_{\tau_k}) \| \} \]

for all \( s \in [\tau_k, t] \) and \( \int_{\tau_k}^{t} \alpha_s ds \leq \lambda \}

The purpose of these random times is to control the periods of time where \( \| \nabla \bar{g}(\theta.)\| \) is close to zero and away from zero. Let us next define the random time intervals \( I_k = [\tau_k, \sigma_k) \) and \( J_k = [\sigma_{k-1}, \tau_k) \). Notice that for every \( t \in J_k \) we have \( \| \nabla \bar{g}(\theta_t) \| < \kappa \).
Suppose that there are an infinite number of intervals $I_k = [\tau_k, \sigma_k]$.

There is a fixed constant $\gamma = \gamma(\kappa) > 0$ such that for $k$ large enough, one has

$$\bar{g}(\theta_{\sigma_k}) - \bar{g}(\theta_{\tau_k}) \leq -\gamma$$

Then, $\bar{g}(\theta_t) \to -\infty$.

However, $\bar{g} \geq 0$. Therefore (by contradiction) there are a finite number of intervals $I_k$. 
\[
\bar{g}(\theta_{\sigma_k}) - \bar{g}(\theta_{\tau_k}) = - \int_{\tau_k}^{\sigma_k} \alpha_s \| \nabla \bar{g}(\theta_s) \|^2 \, ds \\
+ \int_{\tau_k}^{\sigma_k} \alpha_s \left\langle \nabla \bar{g}(\theta_s), \nabla_{\theta} f(X_s, \theta_s)\sigma^{-1} dW_s \right\rangle \\
+ \int_{\tau_k}^{\sigma_k} \frac{\alpha_s^2}{2} \text{tr} \left[ (\nabla_{\theta} f(X_s, \theta_s)\sigma^{-1})(\nabla_{\theta} f(X_s, \theta_s)\sigma^{-1})^T \nabla_{\theta} \nabla_{\theta} \bar{g}(\theta_s) \right] \, ds \\
+ \int_{\tau_k}^{\sigma_k} \alpha_s \left\langle \nabla_{\theta} \bar{g}(\theta_s), \nabla_{\theta} \bar{g}(\theta_s) - \nabla_{\theta} g(X_s, \theta_s) \right\rangle \, ds
\]

Recall that \( \int_0^\infty \alpha_t dt = \infty \) and \( \int_0^\infty \alpha_t^2 dt < \infty \) (Ex: \( \alpha_t = \frac{1}{1+t} \)).
\[
\int_{\tau_k}^{\sigma_k} \alpha_s \left\langle \nabla_{\theta} \tilde{g}(\theta_s), \nabla_{\theta} \tilde{g}(\theta_s) - \nabla_{\theta} g(X_s, \theta_s) \right\rangle \, ds
\]

Rewrite this term using an associated Poisson equation. Assume:

\[
\int_{\mathcal{X}} G(x, \theta) \pi(dx) = 0
\]

Let \( \mathcal{L}_x \) be the generator for the \( X \) process. Then the Poisson equation

\[
\mathcal{L}_x u(x, \theta) = -G(x, \theta)
\]

has a unique solution (with some nice properties).
Numerical Examples

- Ornstein-Uhlenbeck (OU) process
- Multi-dimensional OU process
- Burger’s equation
- Reinforcement learning
The Ornstein-Uhlenbeck (OU) process $X_t \in \mathbb{R}$ satisfies the stochastic differential equation:

$$dX_t = c(m - X_t)dt + dW_t.$$ 

We use continuous stochastic gradient descent to learn the parameters $\theta = (c, m) \in \mathbb{R}^2$. 

$f(x, \theta) = c(m - x)$ and $f^*(x) = f(x, \theta^*)$
We study 10,500 cases. For each case, a different $\theta^*$ is randomly generated in the range $[1, 2] \times [1, 2]$. For each case, we solve for the parameter $\theta_t$ over the time period $[0, T]$ for $T = 10^6$. To summarize:

- For cases $n = 1$ to 10,500
  - Generate a random $\theta^*$ in $[1, 2] \times [1, 2]$
  - Simulate a single path of $X_t$ given $\theta^*$ and simultaneously solve for the path of $\theta_t$ on $[0, T]$
Figure: Mean error in percent plotted against time. Time is in log scale.
Figure: Mean squared error plotted against time. Time is in log scale.
The multidimensional Ornstein-Uhlenbeck process $X_t \in \mathbb{R}^d$ satisfies the stochastic differential equation:

$$dX_t = (M - AX_t)dt + dW_t.$$ 

We use continuous stochastic gradient descent to learn the parameters $\theta = (M, A) \in \mathbb{R}^d \times \mathbb{R}^{d \times d}$. 
Figure: Mean error in percent plotted against time. Time is in log scale.
Figure: Mean squared error plotted against time. Time is in log scale.
The stochastic Burger’s equation is:

\[
\frac{\partial u}{\partial t}(t, x) = \theta \frac{\partial^2 u}{\partial x^2} - u(t, x) \frac{\partial u}{\partial x}(t, x) + \sigma \frac{\partial^2 W(t, x)}{\partial t \partial x},
\]

where \( x \in [0, 1] \) and \( W(t, x) \) is a Brownian sheet.
### Table: Error at different times for the estimate $\theta_t$ of $\theta^*$ across 525 cases.

The “error” is $|\theta^*_t - \theta^*_n|$ where $n$ represents the $n$-th case. The “error in percent” is $100 \times |\theta^*_t - \theta^*_n| / |\theta^*_n|$.

<table>
<thead>
<tr>
<th>Error/Time</th>
<th>$10^{-1}$</th>
<th>$10^0$</th>
<th>$10^1$</th>
<th>$10^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Error</td>
<td>.1047</td>
<td>.106</td>
<td>.033</td>
<td>.0107</td>
</tr>
<tr>
<td>99% quantile of error</td>
<td>.08</td>
<td>.078</td>
<td>.0255</td>
<td>.00835</td>
</tr>
<tr>
<td>Mean squared error</td>
<td>$1.00 \times 10^{-3}$</td>
<td>$9.25 \times 10^{-4}$</td>
<td>$1.02 \times 10^{-4}$</td>
<td>$1.12 \times 10^{-5}$</td>
</tr>
<tr>
<td>Mean Error in percent</td>
<td>1.26</td>
<td>1.17</td>
<td>0.4</td>
<td>0.13</td>
</tr>
<tr>
<td>Maximum error in percent</td>
<td>37.1</td>
<td>37.5</td>
<td>9.82</td>
<td>4.73</td>
</tr>
<tr>
<td>99% quantile of error in percent</td>
<td>12.6</td>
<td>18.0</td>
<td>5.64</td>
<td>1.38</td>
</tr>
</tbody>
</table>
Figure: Mean error in percent plotted against time. Time is in log scale.
Figure: Mean squared error plotted against time. Time is in log scale.
We consider the classic reinforcement learning problem of balancing a pole on a moving cart.

The goal is to balance a pole on a cart and to keep the cart from moving outside the boundaries via applying a force of ±10 Newtons.

The position $x$ of the cart, the velocity $\dot{x}$ of the cart, angle of the pole $\beta$, and angular velocity $\dot{\beta}$ of the pole are observed.

The dynamics of $(x, \dot{x}, \beta, \dot{\beta})$ satisfy a set of ODEs
<table>
<thead>
<tr>
<th>Reward/Episode</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Reward</td>
<td>-20</td>
<td>981</td>
<td>$2.21 \times 10^4$</td>
<td>$6.64 \times 10^5$</td>
<td>$9.22 \times 10^5$</td>
</tr>
<tr>
<td>90% quantile of reward</td>
<td>-63</td>
<td>184</td>
<td>760</td>
<td>8354</td>
<td>$1.5 \times 10^4$</td>
</tr>
<tr>
<td>Mean reward</td>
<td>-78</td>
<td>67</td>
<td>401</td>
<td>5659</td>
<td>$1.22 \times 10^4$</td>
</tr>
<tr>
<td>10% quantile of reward</td>
<td>-89</td>
<td>-34</td>
<td>36</td>
<td>69</td>
<td>93</td>
</tr>
<tr>
<td>Minimum reward</td>
<td>-92</td>
<td>-82</td>
<td>-61</td>
<td>-46</td>
<td>-23</td>
</tr>
</tbody>
</table>

**Table:** Reward at the $k$-th episode across the 525 cases using continuous stochastic gradient descent to learn the model dynamics.