Continuous-time Models for Stochastic Optimization Algorithms

Antonio Orvieto, Aurelien Lucchi

Accepted as a poster at NeurIPS | 2019
Unconstrained non-convex optimization

For some regular $f : \mathbb{R}^d \to \mathbb{R}$, find $x^* := \arg \min_{x \in \mathbb{R}^d} f(x)$.

Training loss of ResNet-110, no skip connections on CIFAR-10 (for more details, check [Li et al., 2018])

Usual Assumption: $f$ is $L$-smooth, i.e. $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.

A recent trend is to model the dynamics of iterative gradient-based optimization algorithms with differential equations.
Tutorial: how is an ODE model constructed?

\[ x_{k+1} = x_k - h \nabla f(x_k) \]  \hspace{1cm} \text{(GD)}

Define curve \( y(t) \) as smooth interpolation: \( y(kh) = x_k \)

What is the law for \( y(t) \)?

1) by construction : \( y(t + h) = y(t) - h \nabla f(y(t)) \);

2) thanks to smoothness: \( y(t + h) = y(t) + h \dot{y}(t) + O(h^2) \).

In the limit \( h \to 0 \), \( \dot{y}(t) = - \nabla f(y(t)) \)

*Solution \( y \in C^1(\mathbb{R}_+, \mathbb{R}^d) \) exists unique in since \( \nabla f \) is globally Lip.
Some important ODE/SDE models

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Model</th>
<th>Perturbed model (stochastic grads)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_{k+1} = x_k - h\nabla f(x_k) ) (GD)</td>
<td>( \dot{X} = -\nabla f(X) )</td>
<td>( dY = -\nabla f(Y) dt + \sigma dB )</td>
</tr>
<tr>
<td>[\text{Mertikopoulos and Staudigl, 2016}]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_{k+1} = x_k + \beta(x_k - x_{k-1}) - h\nabla f(x_k) ) (HB)</td>
<td>( \ddot{y} = -\alpha \dot{y} - \nabla f(y) ) or ( \begin{cases} \dot{v} = -\alpha v - \nabla f(y) \ \dot{y} = v \end{cases} )</td>
<td>( \begin{cases} dV = -\alpha V dt - \nabla f(Y) dt + \sigma dB \ dY = Y dt \end{cases} )</td>
</tr>
<tr>
<td>[Polyak, 1964]</td>
<td>[Polyak, 1964]</td>
<td>[Orvieto et al., 2019]</td>
</tr>
<tr>
<td>( \begin{cases} x_{k+1} = u_k - h\nabla f(u_k) \ u_{k+1} = x_{k+1} + \frac{k}{k+3}(x_{k+1} - x_k) \end{cases} ) (NAG)</td>
<td>( \ddot{y} = -\frac{3}{t} \dot{y} - \nabla f(y) ) or ( \begin{cases} \dot{v} = -\frac{3}{t} v - \nabla f(y) \ \dot{y} = v \end{cases} )</td>
<td>( \begin{cases} dV = -\frac{3}{t} V dt - \nabla f(Y) dt + \sigma dB \ dY = V dt \end{cases} )</td>
</tr>
<tr>
<td>[Nesterov, 1983]</td>
<td>[Su et al., 2016]</td>
<td>[Krichene and Bartlett, 2017]</td>
</tr>
</tbody>
</table>

and many more: primal-dual algorithms, adaptive methods, etc.
why should we care about SDE models? do we really need to introduce these objects? what’s the gain?

- GD-ODE/SDE is the basis for many seminal contributions to the theory of SGD:
  1. asymptotic behavior [Kushner and Yin, 2003];
  2. connection to Bayesian inference [Mandt et al., 2017];
  3. generalization, width of minimas [Jastrzębski et al., 2017].

- NAG-ODE recently provided us with some novel insights of the acceleration phenomenon\(^1\):
  1. [Su et al., 2016] studied accel. with Bessel functions;
  2. [Wibisono et al., 2016] connected NAG to meta-learning and physics via the minimum action principle;
  3. [Krichene and Bartlett, 2017] studied the non-trivial interplay between noise and acceleration in NAG using stochastic analysis on the NAG-SDE;
  4. [Orvieto et al., 2019] showed NAG is equivalent to a linear gradient averaging system after the time-stretch \( \tau = t^2/8 \).

---

\(^1\)For convex functions, there is a method (NAG) strictly faster than GD.
In this paper, inspired by this success,

- we build SDE models for SVRG and mini-batch SGD, which include the effect of **decaying learning rates and increasing batch-sizes**.

- We derive **convergence rates for our models**. We focus on non-convex functions relevant for machine learning.

- We derive equivalent **novel results for the algorithmic counterparts**, using the same Lyapunov functions. This proves the effectiveness of our SDE models.

- We provide a new interpretation for the distribution induced by SGD with decreasing stepsizes, which reveals an underlying **time warping** that can be used for designing Lyapunov functions.

- We provide a dual interpretation of this last phenomenon as **landscape stretching**.
Below are the two SDEs corresponding to mini-batch SGD (MB-PGF) and SVRG (VR-PGF).

\[
dX(t) = -\psi(t)\nabla f(X(t)) \, dt + \psi(t)\sqrt{h/b(t)} \, \sigma_{\text{MB}}(X(t)) \, dB(t) \tag{MB-PGF}
\]

\[
dX(t) = -\psi(t)\nabla f(X(t)) \, dt + \psi(t)\sqrt{h/b(t)} \, \sigma_{\text{VR}}(X(t), X(t - \xi(t))) \, dB(t) \tag{VR-PGF}
\]

where

- \(\xi: \mathbb{R}_+ \rightarrow [0, \Xi]\) is the \textit{staleness function} (linked to the pivot update frequency \(m\) in SVRG);
- \(\psi(\cdot) \in C^1(\mathbb{R}_+, [0, 1])\) is the \textit{adjustment function} (encodes the relative decrease in the learning rate);
- \(b(\cdot) \in C^1(\mathbb{R}_+, \mathbb{R}_+)\) is the \textit{mini-batch size function};
- \(\{B(t)\}_{t \geq 0}\) is a \(d\)-dimensional Brownian Motion on some filtered probability space.
<table>
<thead>
<tr>
<th>Cond.</th>
<th>Rate (Continuous-time)</th>
</tr>
</thead>
</table>
| $(\sim), (H^{-}), (H_{\sigma})$ | \[
\frac{f(x_0) - f(x^*)}{\varphi(t)} + \frac{h d L \sigma^2_*}{2 \varphi(t)} \int_0^t \frac{\psi(s)^2}{b(s)} ds
\] |
| $(\sim), (H^{-}), (H_{\sigma}), (HWQC)$ | \[
\frac{\|x_0 - x^*\|^2}{2 \tau \varphi(t)} + \frac{h d \sigma^2_*}{2 \tau \varphi(t)} \int_0^t \frac{\psi(s)^2}{b(s)} ds
\] |
| $(H^{-}), (H_{\sigma}), (HWQC)$ | \[
\frac{\|x_0 - x^*\|^2}{2 \tau \varphi(t)} + \frac{h d \sigma^2_*}{2 \tau \varphi(t)} \int_0^t \frac{(L \tau \varphi(s) + 1) \psi(s)^2}{b(s)} ds
\] |
| $(H^{-}), (H_{\sigma}), (HP_L)$ | \[
e^{-2\mu \varphi(t)}(f(x_0) - f(x^*)) + \frac{h d L \sigma^2_*}{2} \int_0^t \frac{\psi(s)^2}{b(s)} e^{-2\mu(\varphi(t) - \varphi(s))} ds
\] |
| $(H^{-}), (HRSI)$ | \[
\left(\frac{1 + 2hL^2m}{m(\mu - 2hL^2)}\right)^j \|x_0 - x^*\|^2 (\text{under variance reduction})
\] |

<table>
<thead>
<tr>
<th>Cond.</th>
<th>Rate (Discrete-time, no Gaussian assumption)</th>
</tr>
</thead>
</table>
| $(\sim), (H^{-}), (H_{\sigma})$ | \[
\frac{2 (f(x_0) - f(x^*))}{(h \varphi_{k+1})} + \frac{h d L \sigma^2_*}{(h \varphi_{k+1})} \sum_{i=0}^k \frac{\psi_i^2 h}{b_i}
\] |
| $(\sim), (H^{-}), (H_{\sigma}), (HWQC)$ | \[
\frac{\|x_0 - x^*\|^2}{\tau (h \varphi_{k+1})} + \frac{d h \sigma^2_*}{\tau (h \varphi_{k+1})} \sum_{i=0}^k \frac{\psi_i^2 h}{b_i}
\] |
| $(H^{-}), (H_{\sigma}), (HWQC)$ | \[
\frac{\|x_0 - x^*\|^2}{2 \tau (h \varphi_{k+1})} + \frac{h d \sigma^2_*}{2 \tau (h \varphi_{k+1})} \sum_{i=0}^k (1 + \tau \varphi_{i+1} L) \frac{\psi_i^2 h}{b_i}
\] |
| $(H^{-}), (H_{\sigma}), (HP_L)$ | \[
\prod_{i=0}^k (1 - \mu h \psi_i)(f(x_0) - f(x^*)) + \frac{h d L \sigma^2_*}{2} \sum_{i=0}^k \frac{\prod_{\ell=0}^k (1 - \mu h \psi_{\ell}) \psi_i^2 h}{b_i}
\] |
| $(H^{-}), (HRSI)$ | \[
\left(\frac{1 + 2L^2 h^2 m}{hm(\mu - 2L^2 h)}\right)^j \|x_0 - x^*\|^2 (\text{under variance reduction})
\] |

We derive matching convergence rates in continuous- and discrete-time, using the same Lyapunov functions. This proves the effectiveness of our SDE models.
Insight 1: time stretching

Using the SDE models, we can transform an algorithm to an equivalent one which is easier to study.

**Theorem.** Let \( \{X(t)\}_{t \geq 0} \) satisfy PGF and define \( \tau(\cdot) = \varphi^{-1}(\cdot) \), where \( \varphi(t) = \int_0^t \psi(s)ds \). For all \( t \geq 0 \), \( X(\tau(t)) = Y(t) \) in distribution, where \( \{Y(t)\}_{t \geq 0} \) has the stochastic differential

\[
dY(t) = -\nabla f(Y(t))dt + \sqrt{h(\psi(\tau(t)) / b(\tau(t)))}\sigma(\tau(t)) dB(t).
\]

**Example.**

\( b(t) = 1, \sigma(s) = \sigma I_d \) and \( \psi(t) = 1 / (t + 1) \); we have \( \varphi(t) = \log(t + 1) \) and \( \tau(t) = e^t - 1 \).

\[
dX(t) = -\frac{1}{t+1} \nabla f(X(t))dt - \frac{\sqrt{h\sigma}}{t+1} dB(t)
\]

is s.t. the sped-up solution \( Y(t) = X(e^t - 1) \) satisfies

\[
dY(t) = -\nabla f(X(t))dt + \sqrt{h\sigma e^{-t}} dB(t).
\]

Verification of the Thm. on a 1d quadratic (100 samples): empirically \( X(t) \overset{d}{=} Y(\varphi(t)) \).
Insight 2: landscape stretching

For the sake of simplicity, let \( f(x) = \frac{1}{2} \| x \|^2 \). PGF with \( b(t) = 1, \sigma(s) = \sigma I_d, \psi(t) = \frac{1}{t+1} \) is

\[
dX(t) = -\frac{1}{t+1} X(t) dt + \frac{h\sigma}{t+1} dB(t).
\]

Using solution feedback (only possible with a continuous time formulation), we find that in expectation

\[
E[dX] = CX^2 dt \rightarrow \frac{dE[X]}{dt} = C \nabla \left( \frac{X^3}{3} \right).
\]

Hence, PGF on the quadratic \( \frac{1}{2} \| x \|^2 \) with learning rate decreasing as \( 1/t \) behaves in expectation like PGF with constant learning rate on a cubic.

i.e., we lose strong convexity hence we converge slower!
References
Anosov, D. V. (1967). 
Geodesic flows on closed riemannian manifolds of negative curvature.


Three Factors Influencing Minima in SGD.
*ArXiv e-prints*.

Acceleration and Averaging in Stochastic Mirror Descent Dynamics.
*ArXiv e-prints*. 


A method for solving the convex programming problem with convergence rate $o \left( \frac{1}{k^2} \right)$.


The role of memory in stochastic optimization.


Some methods of speeding up the convergence of iteration methods.


A differential equation for modeling nesterovs accelerated gradient method: theory and insights.


A Variational Perspective on Accelerated Methods in Optimization.

*ArXiv e-prints.*